Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
06/11/2017

Par : **Xuxiang Hu**

Titre : **Models Comparing and Application for Flight Departure Delay Related Insurance**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Entreprise : MOONSHOT-INTERNET
Nom : Marie Huyghues-Beaufond
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Marie Huyghues-Beaufond
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*
Signature du responsable entreprise

Secrétariat:

Signature du candidat

Bibliothèque:

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
3, Avenue Pierre Larousse - 92245, MALAKOFF CEDEX, FRANCE

Résumé

Mots-Clefs

Approche binaire, approche discrète, GLM, GAM, CART, forêt aléatoire, GBM, ZIP, ZINB, ZIG, retard au départ, *Vuong's non-nested test*, Hosmer-Lemeshow Test, AUC, MSE, scénario

La technologie Internet permet de créer de nombreuses assurances nouvelles. La société MOONSHOT-INTERNET offre un service pour couvrir l'attente lors d'un retard d'avion au départ. Lorsque l'indemnité est fixée, la précision de l'estimation de la prime pure correspond à la précision de l'estimation de la fréquence de retard. Huit variables indépendantes sont sélectionnées dans la base de données. La fréquence est significativement influencée par les variables explicatives. En appliquant et en testant six modèles d'approche binaire et trois modèles d'approche discrète, la fréquence est estimée. Les résultats de *Vuong's Non-Nested Test*, Hosmer-Lemeshow test, critère AUC, et critère MSE prouvent que les modèles CART et ZINB sont les deux meilleurs modèles de chaque approche. Les performances de l'estimation de la prime pure vérifient aussi ce résultat. Les écarts agrégés de CART et de ZINB sont stables et faibles dans le scénario de test. La capacité d'adaptabilité des deux modèles est validée en utilisant d'autres seuils de retard, des données plus denses et des données d'autres aéroports. Grâce au résultat du test d'adaptabilité, le modèle ZINB est préféré au modèle CART, son estimation individuelle étant meilleure et son estimation globale restant précise sans sous-estimation.

Abstract

Keywords

Binary approach, discrete approach, GLM, GAM, CART, random forest, GBM, ZIP, ZINB, ZIG, departure delay, Vuong's non-nested test, Hosmer-Lemeshow test, AUC, MSE, scenario

The internet technology could create lots of new insurance products. MOONSHOT-INTERNET offers a service to guarantee the waiting process when there is departure delay. As the indemnity of this product is fixed, the precision of frequency estimation equals the precision of pure premium estimation. Eight independent variables are selected from the database. The frequency is significantly influenced by the independent variables. By applying six models of binary approach and three models of discrete approach to the database, the frequency could be estimated. The results of Vuong's Non-Nested Test, Hosmer-Lemeshow Test, criterion AUC, and criterion MSE have proven that CART model and ZINB model are the two best models of each approach. The performances of pure premium estimations also validate the conclusion. The aggregate deviations of CART model and ZINB model are stable and small in the result of the scenario test. The adaptability of the two models is validated by using other delay thresholds, by using more concentrated data, and also by using the data of other airports. In the result of adaptability test, the ZINB model is preferable to the CART model, as its individual estimation is better and its aggregate estimation is stable with no underestimation.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my manager Marie Huyghues-Beaufond for the continuous support of my research and writing, for her patience, motivation, and immense knowledge. Without her insightful comments and encouragement, this project would hardly have been completed.

I would like also thank all the team members in MOONSHOT Internet. The group has been a source of friendships as well as good advice and collaboration. All the team members are kind and give me a lot of helps during my professional time.

Furthermore, I would like to thank my academic supervisor Pierre Picard for his valuable comments on this project.

Last but not the least, I would like to thank my family and my friends for their delightful support and encouragement.

Table of Contents

Introduction

Nowadays, the application of new insurance technology improves the insurance product in different ways. Meanwhile, a great number of new insurance products become feasible, thanks to the technology revolution. The flight departure delay insurance is one of the new insurance products. The flight departure delay insurance covers the waiting process when there is flight departure delay. The coverage includes the need of supply, rest, etc. The existing insurances only cover the risk of large delays of three hours, four hours, etc. Unlike the traditional insurance products, the flight departure product of MOONSHOT-INTERNET offers the coverage for small departure delays and is realized via internet technology. This new product has characters of short life cycle, fast transaction process, and rapid validation. In this case, this product is operated in high frequency, and thus, the well discriminated price is important to make this product adaptable and stable.

This essay is focused on the precise estimation of flight departure delay frequency. The indemnity is fixed for the flight departure delay product. In this case, precise frequency estimation means a precise risk evaluation. Hence, the main aim is to study possible models for the flight departure delay frequency and to find the best one for estimating the pure premium of the flight departure delay insurance. The research is composed of following parts:

The first part **Background** introduces the internet insurance environment, MOONSHOT-INTERNET, and the premium calculation of the flight departure delay insurance.

The second part **Database and Variable** presents the database of the research and the explanatory variables to be used in the models. The explanatory variables are year, month, day of the week, vacation, scheduled departure time, and scheduled arrival time. The statistics of the variables are then introduced in this part.

The third part **Model and Analysis** studies and compares the nine models of the two approaches. The models of binary approach contain Generalized Linear Model (GLM), Generalized Additive Model (GAM), Classification And Regression Tree (CART), random forest, Bernoulli and Adaboost Gradient Boosting Model (GBM). The models of discrete approach are zero-inflated models with Poisson distribution (ZIP), Negative Binomial distribution (ZINB), and Geometric distribution (ZIG). After applying those models to the database, those models are compared by Vuong's Non-Nestd Test, Hosmer-Lemshow test, criterion AUC, and criterion MSE.

The fourth part **Evaluation** introduces the pricing evaluation, scenario test, and adaptability of the selected models in the third part. In pricing evaluation, the estimations of pure premium and technical premium are used to study the possible price distribution. The scenario test evaluates the model performance for different policy portfolios and estimates the risk level in aggregate. The adaptability evaluation validates the model performance for potential situations such as different delay thresholds, limited databases, and other airports.

Chapter 1

Background

This part introduces the research background of this essay, including introductions of internet-insurance, microinsurance, flignt departure delay insurance, MOONSHOT-INTERNET, and premium calculation.

1.1 Internet Insurance

Nowadays there are more and more Internet insurance products in the world (America, Europe, China, India, etc) such as travel insurance, car insurance, shipping return insurance, etc.

In total, those products are faster and simpler because of the automatic electrical treatment in the website or application. Besides, the insurance claims of those products are easier to be validated because of the availability of corresponding accident data. As the Internet is spreading widely, extensive real-time information can be retrieved quickly and easily, such as the arrival and departure time of flight, arrival of shipping, train time table, weather information, etc. Consequently, the insurance company can take advantage of that information to develop lots of new insurance products. With the increasing volume and increasing diversification of policies, those insurance products highly depend on the reliability and accuracy of the Internet treatment process. As a result, a higher level of Internet technology and Internet security is required.

In the time horizon, with faster inscription process and faster payment process, the selling of on-line insurance product is faster. Also, as the real-time accident data are available via internet, the corresponding claims can be verified faster, and the indemnities can also be paid faster by financial technology. In general, Internet insurance products are sold, checked and paid in a more rapid and automatic way. As a result, the life cycles of those insurance products are shorter. In this case, the insurance products which have shorter life cycle could be created by applying the internet technology.

In the cost horizon, the internet insurance products have higher fixed costs in the process of the system construction, etc. However, the automatic processes of checking and payment have lower variable costs. Thus, in the life cycle of Internet insurance product, the variable costs that happen in the operation processes such as checking and payment are relatively lower than traditional product. Hence, the insurance products which have smaller coverage are more feasible.

With a shorter life cycle and lower trading cost, the checking and payment frequency of

those new insurance products will be much higher. This high frequency will create huge historical data of the insurance products. Consequently, there will be a chance of the data analysis and also a challenge to the data treatment. Thus, a more flexible and accurate pricing strategy can be generated grace to the increasing data volume and data diversification.

The term "microinsurance" is introduced to present the insurance with small coverage and premium, which can be well developed by the internet technology.

1.2 Microinsurance

The term "microinsurance" was first published around 1999. The definition of the term "microinsurance" has been the subject of important debate and discussion within the development environment. The definition of microinsurance[1] is continually evolving:

- The protection of low-income people against specific perils in return for regular premium payments proportionate to the likelihood and cost of the risk involved (Preliminary Donor Guidelines, 2003).
- A risk transfer device, characterized by low premiums and low coverage limits, and designed for low-income people not served by typical social insurance schemes (Micro Insurance Academy, India, 2007).
- Insurance that is accessed by the low-income population, provided by a variety of different entities, but run in accordance with generally accepted insurance practices. This means that the risk insured under a microinsurance policy is managed based on insurance principles and funded by premiums (International Association of Insurance Supervisors, 2007).

To sum up, there are three main characteristics of microinsurance: the low coverage limit, the one designed for low-income people, and the low premium. By applying the internet technology, the trading speed increases and the trading cost decreases. Hence, a huger number of small risks can be insured. Thus, more small insurances, which have small amount of premium and indemnity, will be acceptable for insurance companies. In addition, because of the convenience of inscription and payment, more insurance products can be accepted by the public. Thus, from the aspect of demand and support, the insurances of small coverage have a good market. The character "designed for low-income people" is neglected and the term "microinsurance" used in this essay refers to the insurance of small coverage and low premium.

There already exist some microinsurance products of different companies, such as AXA Motor Policy, AXA Business Liability Offer in France (the average premiums less than 1 euro per day), AXA microinsurance products in India (Personal Accident Health), "AXA Contigo" card-based product (Fire, Assistance, Motor Robbery lump sum compensations) in Mexico, Allianz Personal Accident Plus Dental in Colombia, *Allianz Obsèques product* in Ivory Coast, etc. However, those products are all similar to transitional insurance products, without applying the internet technology. In this case, by applying the internet technology, more microinsurances could be created.

[1] cf. Microinsurance Network. (2017) *A brief history.* `http://www.microinsurancenetwork.org/brief-history`

1.3 Flight Departure Delay Insurance

As the micinsurance products introduced above, the flight departure delay insurance is also a microinsurance product. The premium of this product is expected at several dollars and the indemnity of this product is expected within one hundred dollars. Meanwhile this product has a short product life cycle, which starts from the purchase of flight ticket and ends after several hours of the flight.

This essay is mainly focused on the study of flight delay insurance product. According to EU Regulation 261/2004, air passengers have the right to get compensation in the following cases:

- Cancelled flight
- Delayed flight
- Denied boarding.

Those compensations only work for cancellation and large flight delay (larger than four hours or five hours), but not for small departure delay such as one hour or thirty minutes. Also, this compensation is limited to the European area.

Insurer	Product	Coverage	Indemnity
MasterTravel	Travel Accident Insurance	Delay longer than 4h	Take charge of basic necessary until 450 euros
Sainsbury	Travel Insurance	Delay longer than 12h	30£ for the first complete 12h flight delay and 20£ for each subsequent 12h delays.
Allianz	Travel Insurance	Delay longer than 6h	Reimburse for meals, accommodations and lost prepaid expenses
Berkshire Hathaway	AirCare Insurance	Delay longer than 2h	Reimburse of 50 dollars

Table 1: Flight Delay Related Insurance

There are the insurance products that guarantee the flight departure delay nowadays. Most flight insurance products insure great delays, but not small delays. Table 1 presents several flight delay related insurances. MasterTravel provides Travel Accident Insurance which covers flight delay longer than four hours by taking charge of basic necessary until 450 euros. However, Sainsbury provides Travel Insurance which only covers flight delay longer than twelve hours with small indemnities. Allianz provides Travel Insurance which covers fight delay longer than six hours by reimbursing for meals, accommodations, etc. Specially, Berkshire Hathaway provides AirCare Insurance to cover delays longer than two hours with an indemnity of 50 dollars. It is clearly that most insurances are focused on large flight delays. And the guarantee of the flight delays is normally included in a travel insurance.

1.4 MOONSHOT-INTERNET

MOONSHOT-INTERNET is an Insurtech company dedicated to E-merchants backed by SOCIETE GENERALE Insurance. The aim of MOONSHOT-INTERNET is to make insur-

ance and service simple, useful and accessible by the innovation technology. MOONSHOT-INTERNET wants to verify its business model in France first and to apply it to the entire Europe. It provides usage-based insurance with plug and play API (application programming interface) to E-commerce merchants, allowing any size of E-merchant to implement them. Those products include Shipping Return Insurance, Delivery Insurance, Travel Weather Insurance, and it also offers guarantee for flight/train delay, ticket cancellation, etc.

This type of insurance is new to the European Market, with great opportunities and challenges. In traditional insurance industry, the related service and process generate some costs and delays, and there is not enough user integration (complex customer path). Hence, MOONSHOT-INTERNET provides insurances in a better way. All the operations are executed on the computer or mobile phone by plug-in or API, following with on-line user portfolio creating, management, and advising. So it has the character of full-time subscription, full-time claim handling and also the real time insurance pricing. With the increase of market occupation and the augmentation of user histories, MOONSHOT-INTERNET can understand the market better, and then makes a better strategy. In this case, it is a new task to take full advantage of market information as well as individual history. Hence, the real-time precise pricing of those insurance products is important.

As introduced, the existing coverage of flight departure delay is mainly focused on insuring the loss of great delays. But the insured may experience losses caused by the need of rest, the need of food, and bad mood when the flight departure delay is in a small level. In this case, MOONSHOT-INTERNET comes up with the on-line flight departure delay insurance, which is fast, reliable, and has no area limit. The flight departure delay product of MOONSHOT-INTERNET can compensate the losses caused by the departure delay which is longer than thirty minutes. As the small flight departure delay has not been covered by the present insurance products in France, the flight departure delay product of MOONSHOT-INTERNET has an emerging market.

The departure delay insurance with delay threshold of thirty minutes could cover most losses that caused by departure delay. And the departure delay within thirty minutes is tolerable. Thus, the delay threshold is chosen as thirty minutes to make the coverage range as large as possible. This insurance uses the lump-sum indemnity. Once the flight has departure delay longer than thirty minutes, the indemnity is available for the insured. And the insured can take advantage of this indemnity to get access to the airport lounge to make the waiting process agreeable.

1.5 Premium Calculation

The pure premium is the premium that reflects the risk of coverage. Thus, the quality of pricing strategy largely relies on the precision of pure premium estimation.

$$PP = E[I * F]; \qquad \begin{array}{r} PP = \text{Pure premium} \\ I = \text{Indemnity} \\ F = \text{Accident Frequency} \\ E[\,] = \text{Expectation} \end{array} \tag{1.1}$$

According to formula (1.1), the pure premium can be calculated by the frequency and indemnity if those two are independent. As the airport lounge can supply the food, drink, rest, etc, the losses caused by the delay could be covered by the service of airport lounge. Thus,

the cost of airport lounge room is used to estimate the indemnity of flight departure delay. Another factor of pure premium calculation is the accident frequency. Unlike the indemnity, the accident frequency can not be estimated by a stable value such as the cost of airport lounge room. The accident frequency may be influenced by lots of parameters, thus it is important to make an accurate estimation of the accident frequency when different flight parameters are given.

In this essay, one of the most important work is to estimate the accident frequency, which is the probability that the flight departure delay is longer than thirty minutes. As departure delays shorter than thirty minutes are acceptable, delay threshold of thirty minutes is chosen to take most cases of accident into account. By the law of total probability theory, the total probability of an outcome can be realized via several distinct events. According to this theory, the probability that the flight departure delay is longer than thirty minutes can be composed by two events: the probability when the flight is canceled, the probability when the flight is not canceled:

In this case, this essay is mainly focused on the estimation of pure premium

$$P(D) = P(D|C) * P(C) + P(D|C^c) * P(C^c);$$
$$\begin{array}{r} D = \text{Departure delay longer than thirty minutes} \\ C = \text{Flight is canceled} \\ C^c = \text{Flight is not canceled} \end{array} \tag{1.2}$$

$P(D)$ is the probability of event D and $P(D|C)$ is the conditional probability of event D given that event C has occurred.

As the cancelled flight is included in the EU Regulation 261/2004 (cf 1.3 Flight Departure Delay Insurance), the cancellation of flight is not covered by the flight departure delay insurance. Thus, when the flight is canceled, the flight departure delay is not considered. In this case, $P(D|C) = 0$:

$$P(D) = P(D|C^c) * P(C^c); \qquad \begin{array}{r} D = \text{Departure delay longer than thirty minutes} \\ C^c = \text{Flight is not canceled} \end{array} \tag{1.3}$$

$$\text{Condition:} \qquad P(C^c) \approx 1; P(C) \approx 0 \qquad C^c = \text{Flight is not canceled} \tag{1.4}$$

If the condition (1.4) is satisfied, according to formula (1.3), the delay probability satisfies:

$$P(D) \approx P(D|C^c); \qquad \begin{array}{r} D = \text{Departure delay longer than thirty minutes} \\ C^c = \text{Flight is not canceled} \end{array} \tag{1.5}$$

If condition (1.4) is satisfied, according to formula (1.3), the departure delay probability is approximately the same of the conditional probability given that the flight is not cancelled. Therefore the frequency is estimated by the conditional probability given that the flight is not

cancelled. And this conditional probability can be modeled by the part of database which has no cancellation.

If condition (1.4) is not satisfied, both the cancellation probability and the departure delay probability in the condition that the flight is not cancelled should be estimated.

In summary, when the probability that the flight departure is longer than thirty minutes has been estimated, with airport lounge room price given, the pure premium can be estimated.

Chapter 2

Database and Variables

Data of Bureau of Transportation Statistics (BTS) contains various details of USA flight in the period 1995-2016. After treatment, these data are used as database to model the flight delay probability in different approaches.

2.1 Database

Bureau of Transportation Statistics (BTS) is one of the principal federal statistical agencies. This agency is an official transportation statistic organization which collects and issues the transportation data. After acquiring the Office of Airline Information (OAI) on June 1, 1995, the Bureau of Transportation Statistics has completed source of the airline data. (OAI is originated as the financial and operating statistics arm of the Civil Aeronautics Board (CAB).)

Among the data diffused by BTS, the On-Time Performance Data is chosen as the database. It contains complete flight information for the period 1995-2016, and is composed of those different variable groups:
Time Period, Airline, Origin, Destination, Departure Performance, Arrival Performance, Cancellations and Diversions, Flight Summaries, Cause of Delay, Gate Return Information at Origin Airport, etc.

As presented by the names of those groups, this database contains a complete set of flight information, from departure to arrival, delay to cancellation, which guarantees the good quality of modelling and application.

2.2 Variable

2.2.1 Dependent Variable

In Section 1.3, two probabilities are important in the pure premium calculation: conditional probability given that the flight is not canceled, probability that the flight is canceled.

The conditional probability can be studied using the data without cancellation. To estimate the departure delay probability, the departure delay variable is studied (as discrete variable or as binary index whether the delay is longer then thirty minutes). The aim is to model delay probability using other independent variables in the database, and to precisely estimate the probability for new data. The departure delay variable corresponds to the variable "DepDelay"

in the database (Difference in minutes between scheduled and actual departure time. Negative numbers mean early departures). To study the delay probability in a binary approach, this discrete variable in the database is transformed into a binary variable whether the delay is longer than thirty minutes or not.

As for the cancellation probability, the "Cancelled" variable in the database is studied. This variable is an indicator of the flight cancellation, and reflects whether the flight is cancelled or not[1].

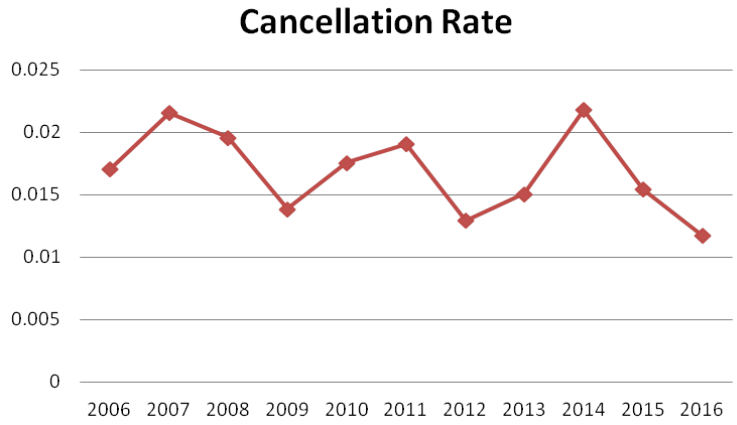

Figure 1: Flight Cancellation Rate In USA

The annual cancellation rate of all USA flight is presented in Figure 1. From 2006 to 2016, the annual cancellation rates are all below 2.2%. If the pure premium equals 5 dollars, the mean error caused by the cancellation rate is less than $2.2\% \times 5 = 0.11$ dollars, which is small enough to be neglected. In addition, the neglect of cancellation rate does not result in an underestimation. Hence, concerning the low flight cancellation rate, the cancellation probability is neglected, and the condition (1.4) (cf 1.5 Premium Calculation) is considered to be satisfied. Thus the departure delay probability is considered to be the conditional probability, which can be estimated by the database without cancellation.

2.2.2 Independent Variable

The independent variables are divided into three types: time variable, geography variable, and operation variable.

2.2.2.1 Time Variable

Six time variables are considered to be studied: year, month, weekday, vacation, scheduled departure time and scheduled arrival time. On one hand, concerning the date variables, year, month, weekday, and vacation are the four main variables. On the other hand, concerning the clock variables, departure time and arrival time are the two main variables.

The year variable is used to present a long-term tendency (thus should be treated as a quantitative variable). While month variable is used to represent periodicity or seasonality in

[1]The mathematical notation is $\mathbb{1}_{\{\text{flight is canceled}\}}$

one year (total twelve different values). Also the month variable is more precise than the season variable to study the cycle in one year. Thus, the month variable is used and is considered as a qualitative variable. Then weekday variable can represent the cycle in one week, and thus it is also considered as a qualitative variable.

In vacation days, people are more likely to travel around, thus the airport would have more passengers. As a result, the plane may have more chance to get a delay (either departure delay or arrival delay). In this case, it is necessary to utilize the vacation variable to take this influence into account.

The scheduled departure time is also considered as one independent variable, which may have a great effect on the delay rate. Similar as the scheduled departure time variable, scheduled arrival time may also have influence on the departure and arrival delay rate. As the the scheduled departure time variable is focused on the departure time while the scheduled arrival time variable is focused on the arrival time, the two variables have different influence on the departure and arrival delay rate.

2.2.2.2 Geography Variables

Three geography variables are considered to be studied: arrival airport, departure airport, and distance. The arrival airport variable and the departure airport variable decide the geography parameter of one flight. Given the arrival and departure airports, the distance of this flight is determined. Concerning airport capacity and airport installations, the airport variables have some influences on the arrival or departure delays. However, these variables may have too many modalities to be a good regression factor.

As distance can be determined by the arrival and departure airports, there is no need to use this variable if the departure and arrival airport variables have been used. Concerning that not all of those two variables are used, the distance variable is a good geography variable, which is numerical and can represent a part of the influence of the two airports.

2.2.2.3 Operation Variables

One operation variable is taken into account: the airline variable. As the operation processes of each flight may have influence on the flight departure delay, the airline variable is utilized to represent the influence of flight operations.

2.3 Statistics

The On-Time Performance database contains too many records to make the analysis. Only the last twenty years data of the flight departed from New York is used for this part.

The aim of this part is to study the statistics of the average probability that the departure delay is longer than thirty minutes. Moreover, we are going to know how the independent variables influence the dependent variable and in which way the independent variable should be treated. For each independent variable, the treatment method is decided by its characteristics as well as the relationship with other variables. Without specific explanation, the departure

delay in the following analysis means the delay longer than thirty minutes.

2.3.1 Time Variables

1) Year

This variable corresponds to the "Year" variable in the On-Time Performance database, which is used as the independent variable for the long term trend.

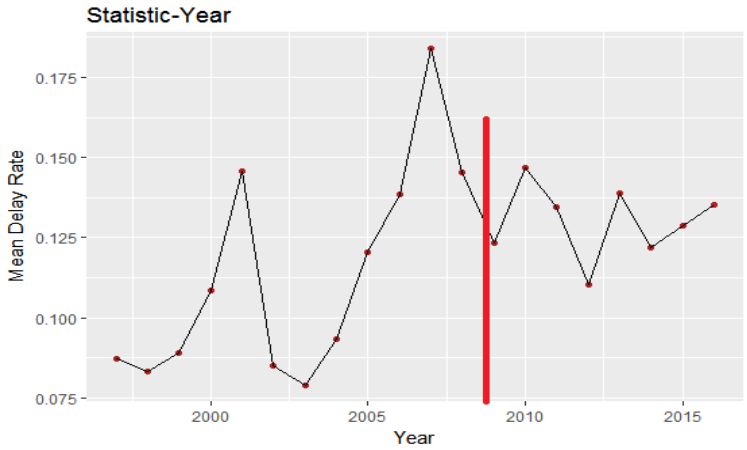

Figure 2: Statistic-Year

Figure 2 presents the relation between departure delay frequency and the year variable. The x-axe is the year variable, and the y-axe is the flight departure delay frequency. This graph shows a significant increase in the period 2003-2008, and shows a more stable curve in the period 2009-2016. The average delay rate increases from 0.080 to 0.185 in the period 2003-2008, and are between 0.110 and 0.150 in the period 2009-2016. Thus, to prevent the irregular effects, only the recent data in the period 2009-2016 is used for modelling. And the year variable is used to represent the long-term tendency of flight departure delay.

2) Month

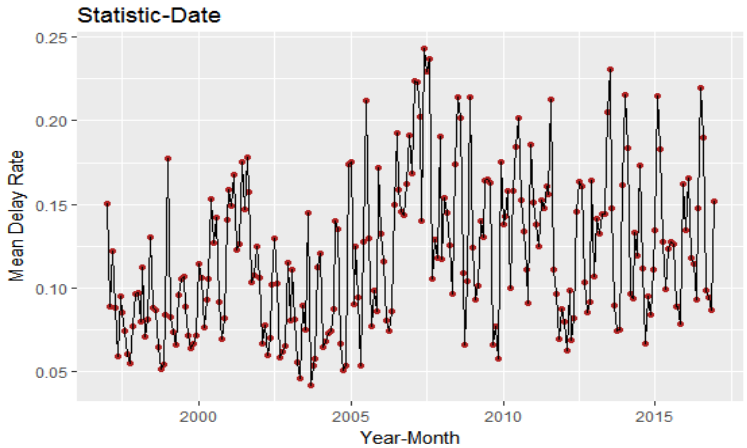

Figure 3: Statistic-Month by Year

There are two graphs for the month variable. Figure 3 shows a monthly average time series of the flight departure delay. The x-axe of this graph is time axe of month, the y-axe is the mean departure delay rate of all flights in the corresponding month. The monthly departure delay rate oscillates largely between 0.01 and 0.25. And the annual tendency is hard to be observed in the Figure 3. Hence, compared to the long-term tendency in the Figure 2, the influence of month variable is much bigger. And the cycle is also hard to be extracted from it.

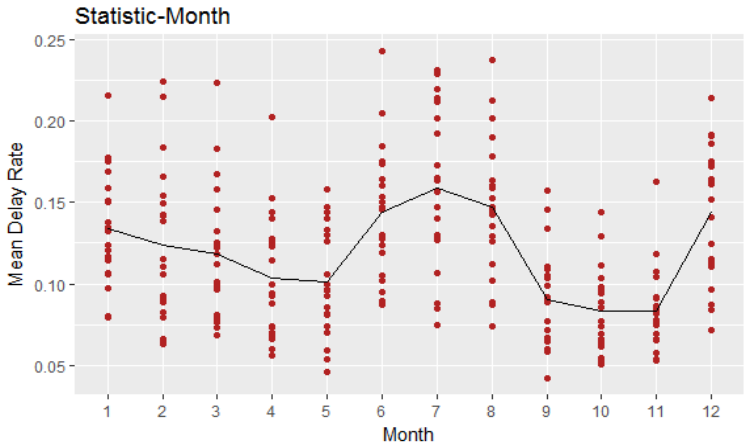

Figure 4: Statistic-Month

Figure 4 shows the monthly cycle. The x-axe of this graph is the month variable, and the y-axe represents the average departure delay rate. Those red points in the same x-axe value are the average departure delay rate of different years. The y value of the point in the curve is the average value of the point set in the same x-axe value. In other words, the corresponding value of y-axe is the average of mean departure delay rates of save month and different year. In this graph, the average delay rate decreases from January to February, and increases rapidly to the peak in July, then decreases to the lowest value in November, and finally increases to 0.14 at the twelfth month.

3) Day of the Week

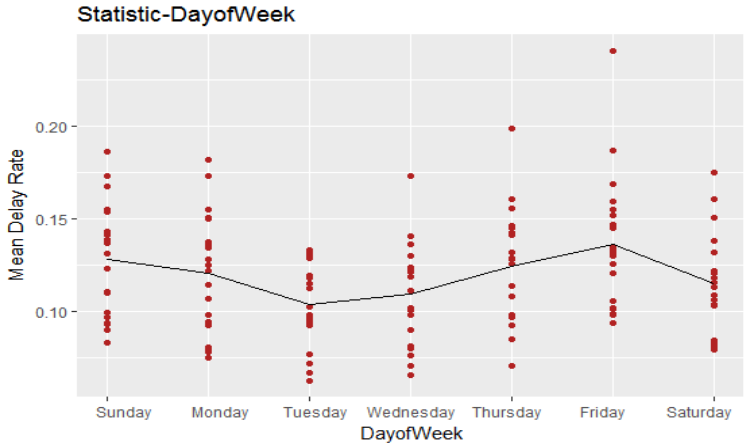

Figure 5: Statistic-Day of the Week

This graph shows the periodicity of the variable week. X-axe in this graph corresponds to day of the week variable, and y-axe corresponds to the mean departure delay rate. For each

x value, the points show the departure delay rates of the corresponding day of the week in different years, and the line is the average value of the point set. This line reaches a peak on Friday and Sunday, and gets the lowest value on Tuesday.

4) Vacation

"Vacation" is an important independent variable. To model the influence of vacation variable, this independent variable is calculated as the time difference between the date and the closest vacation calender day[2] (cf. Appendix H: Vacation calender day). Negative value of vacation variable means that the date is before the closest vacation calendar day, positive value means that the date is after the closest vacation calendar day. For example, given the two vacation calendar days 01/01/2017 and 25/12/2016, the value of vacation variable is 2 for the date 27/12/2016 and -2 for the date 30/12/2016

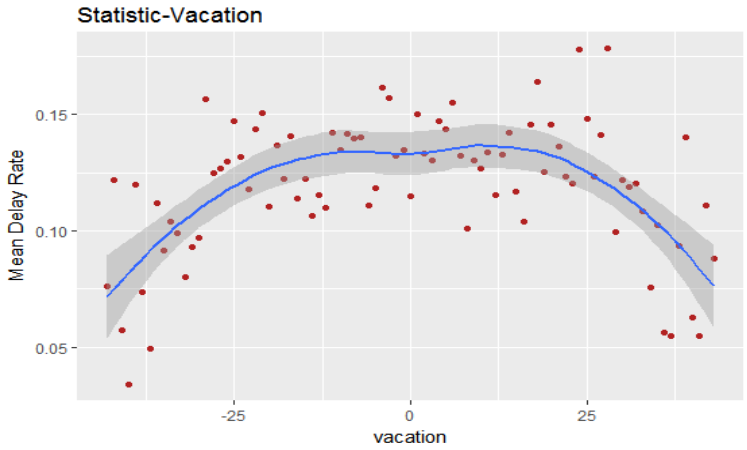

Figure 6: Statistic-Vacation

Figure 6 shows a quasiconcave curve relation between vacation and flight delay rate. The flight delay rate arrives almost maximum around zero, and arrives minimum around the two ends. To integrate the quasiconcave relation in the model, the treatment method of two or three degree polynomial smoothing[3] is used.

5) Scheduled Departure Time

[2] The vacation calender day is from the holiday calendar for the New York Stock Exchange.
[3] The polynomial smoothing uses a polynomial regressor to replace the simple one degree regressor

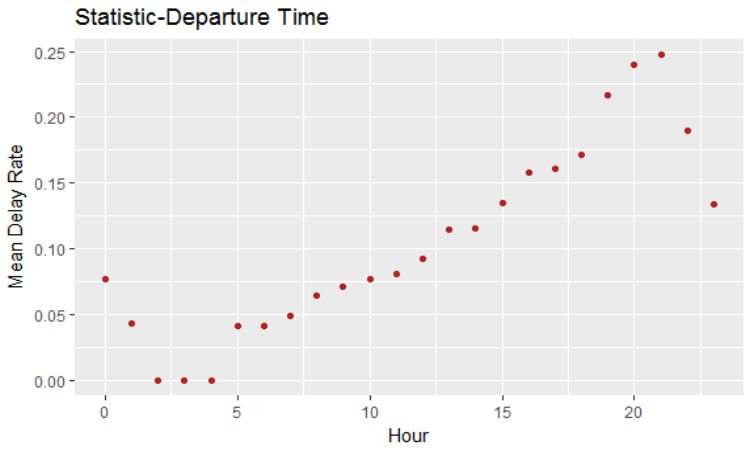

Figure 7: Statistic-Scheduled Departure Time

Figure 7 shows the relation between scheduled departure time and departure delay. The x-axe of this graph represents the hour of scheduled departure time. The y-axe value of each point represents the flight delay frequency in the corresponding hour. As this variable has a cycle of 24 hours, the left end of this graph is connected with the right end. The graph shows three different intervals and the two break points (the lowest point and highest point). One of the intervals is from five to twenty. In this interval, the flight delay frequency increases steadily. And in the interval of twenty one to twenty three and the interval of zero to four, the flight delay decreases steadily. In this case, to well integrate the influence of the scheduled departure time variable, a three interval treatment is used for this variable: [0,4], [5,20], [21,23].

6) Scheduled Arrival Time

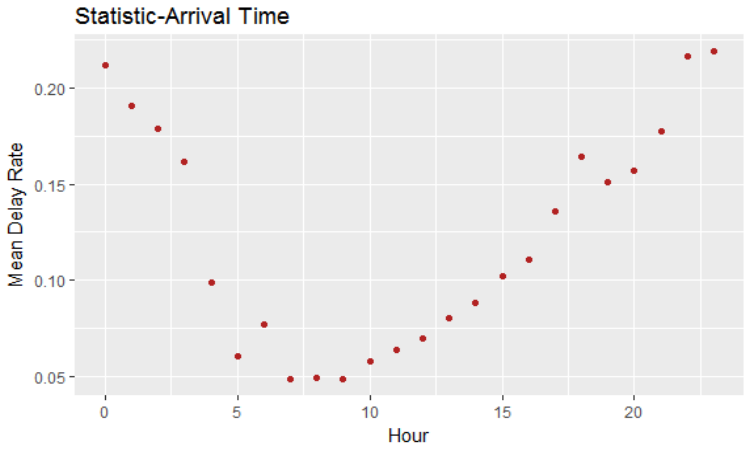

Figure 8: Statistic-Scheduled Arrival Time

Similarly to the scheduled departure time variable, the scheduled arrival time variable has two intervals, with the lowest and highest points as the breaking points. The first interval is from zero to seven, and the second interval is from seven to twenty three. In this graph, the curve of scheduled arrival time is presented. The curve decreases in the first interval and increases steadily in the second interval. And for this variable, the two interval treatment is used: [0,6], [7,23]

2.3.2 Geography Variables

1) Departure Airport.

There are 332 departure airports in the database. However, there are only thirty two departure airports that have more than fifty destinations. Considering the volume of the database is large, the data of the New York John F. Kennedy International Airport has been used to study the different models.

2) Arrival Airport.

There are eighty one arrival airports in this data of New York John F. Kennedy International Airport. In this graph, except several abnormal points, the flight delay frequencies of those destination airports are in the interval [0.1, 0.2]. Considering that there is not enough data for some arrival airports, this variable is not used.

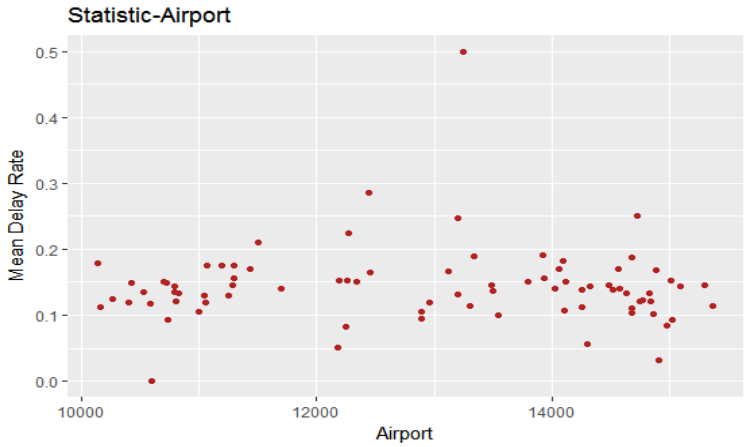

Figure 9: Statistic-Arrival Airport

3) Distance.

As the arrival airport variable and departure airport variable are not used, the distance variable is considered as the independent variable to represent the geography influence. The unit of this variable is kilometer, and the range of this variable is from ninety four kilometers to around five thousand kilometers. And the distribution of this variable is unbalanced. Then this variable is treated by the logarithm function so that its variance and average are comparable to that of other variables.

The graph following shows the relation between distance and mean flight delay rate. The x-axe is the logarithm result of distance, and y-axe is the flight delay frequency of corresponding distance. In this graph there is a small positive relation between distance and flight delay probability.

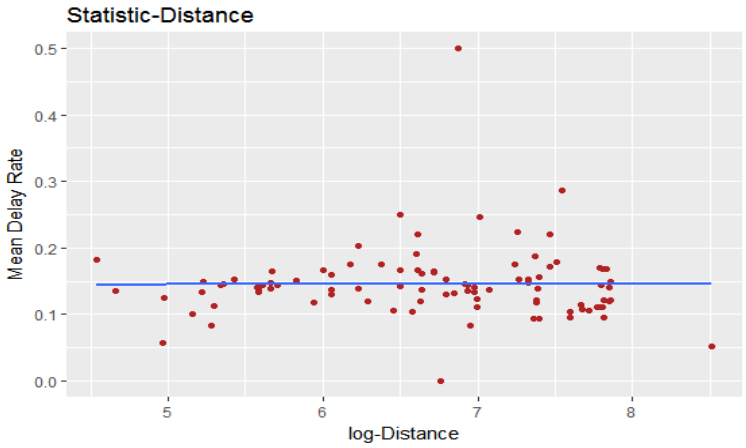

Figure 10: Statistic-Distance

2.3.3 Operation Variables

There is one important operation variable to be considered: airline variable. As the flights of different airlines have different processes and operations, the departure delay rate are also different. Consequently, the airline variable could have some influence on the flight departure delay. Figure 11 shows the flight delay frequencies of different airlines. And the delay rates of those airlines are quite different.

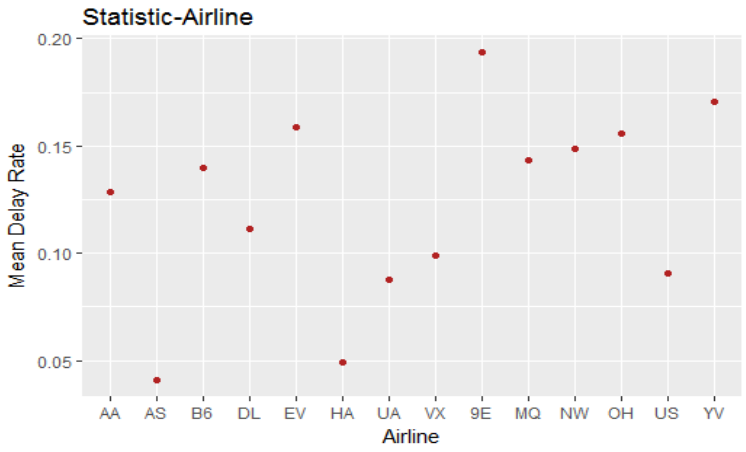

Figure 11: Statistic-Airline

Table 2 shows the explanation of the value of airline variable in the database (the relation between airline id and airline name).

AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
HA	Hawaiian Airlines Inc.
UA	United Air Lines Inc.
VX	Virgin America
9E	Endeavor Air Inc.
MQ	Envoy Air
NW	Northwest Airlines Inc.
OH	PSA Airlines Inc.
US	US Airways Inc.
YV	Mesa Airlines Inc.

Table 2: Airline ID - Airline Name

2.3.4 Variable Correlation

The Pearson correlation coefficient is a measure of the linear correlation between two variables. Given two random variables X and Y, the pearson correlation is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where

σ_X is standard deviation of X

σ_Y is standard deviation of Y

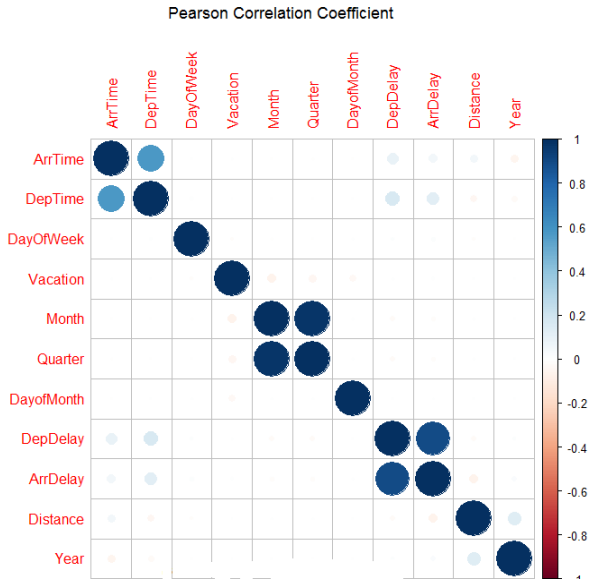

Figure 12: Correlation of explanatory variables

Figure 12 presents the pearson correlation coefficient of those quantitative explanatory variables. There are strong correlations between scheduled arrival time variable ("ArrTime") and

scheduled departure time variable ("DepTime"), between departure delay variable ("DepDelay") and arrival delay variable ("ArrDelay"). And there are also weak correlations between delay variables ("DepDelay", "ArrDelay") and scheduled time variables ("DepTime", "ArrTime").

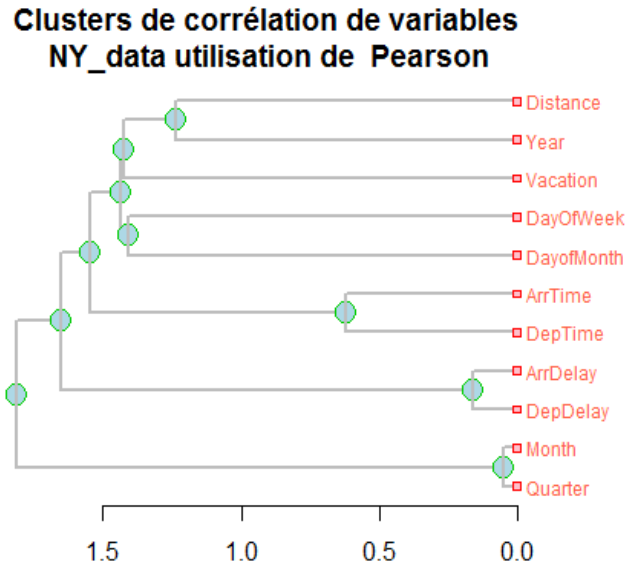

Figure 13: Clustering by Correlation Coefficient

Figure 13 presents clustering result by the Pearson correlation coefficient. As the quarter variable can be generated by month variable, those two variables are completely correlated and are the closest in clustering. By the clustering result, the arrival delay variable and departure delay variable are close, and the scheduled arrival time variable and the scheduled departure time variable are also close. For other clustering that is beyond 1.0 in the x-axe, the distance is relatively large. Small distance of two variable means that the correlations with other variables are similar. Consequently, the scheduled arrival time variable and the scheduled departure time variable have similar relationship with other variables. However, those two variables represent the different influences of departure airport and arrival airport. Thus both variables are considered in this study, regardless of the small pearson correlation coefficient.

Chapter 3

Model and Analysis

This part is about the modelling of the flight delay probability. The statistic graphs in part 2.3 present the relationship between variables and validate that independent variables have influences on dependent variables. To take advantage of all those influences in the pricing process, models are considered, configured, applied, and tested.

Two approaches exist to study the flight delay probability. The first one is the binary approach. In this approach, the binary variable that the departure delay is longer than thirty minutes is used as the dependent variable. The delay probability is the probability that the binary variable equals one (or the frequency of this binary variable). The departure delay probability is estimated by studying the frequency of this binary variable. The second approach is the discrete approach. In this approach, the discrete variable of the departure delay is studied. The value of this discrete variable is integer with unit of minute. By modelling the distribution of this variable, the departure delay probability can be estimated by the value of cumulative distribution function at thirty minutes.

Among the tested models, six models are in binary approach and three models are in discrete approach. The study of each modelling method is composed of mathematical theory, model application, analysis and evaluation. analysis. The data of New York John F. Kennedy International Airport in the period 2009-2016 serve as the database to train the models.

The database has 814 966 observations. It is divided into two parts: training set and test set. Seventy percent of the data compose the training set to build the model. To get the optimal model by the training set, the training set is also divided into two parts, first part for model training and second part for model configuration. The testing set has the other thirty percent observations. It is used to validate the optimal model by the training set.

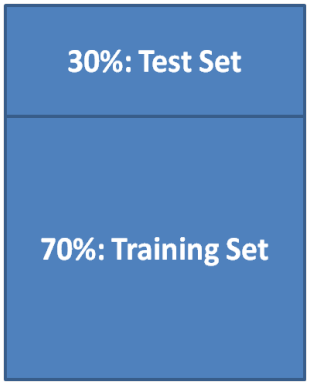

Figure 14: Data Division: training set and test set

3.1 Binary Approach

The binary approach is focused on the study of binary delay indicator that the flight delay is longer than thirty minutes or not. Six models are studied in this approach. They are generalized linear model, generalized additive model, classification and regression tree, random forest and gradient boosting models.

3.1.1 Generalized Linear Model

The generalized linear model (GLM) is an generalized method of the ordinary linear regression. Not limited by the normal distribution, GLM can treat the independent variable whose error distribution is one of the exponential distribution family, such as Poisson distribution, gamma distribution, etc.

The generalized linear model has three components[1]: random component, systematic component, link function (connect systematic component and random component). The random component is the error component. This means that the response variable $Y = [y_i]^T$ are independent random variables, with the same distribution of exponential family:

$$f(y|\theta, \phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

Where $a(\cdot), b(\cdot), c(\cdot)$ are functions and (θ, ϕ) are parameters.

The systematic component is the linear predictor of the explanatory variables. With n explanatory variables $X = \{1, x_1, x_2 ... x_n\}$ and n+1 unknown parameters $\beta = \{\beta_0, \beta_1, \beta_2, \beta_n\}$, the linear predictor is generated by the linear combination of X and β:

$$\eta = X\beta^T$$

Let $g(\cdot)$ be the link function, the GLM model is presented by the formula:

$$g(E(Y)) = \eta = X\beta^T$$

[1] cf. James K. Lindsey. 1997. *Applying Generalized Linear Models.* Springer-Verlag New York

In the case that the binary delay indicator is the response variable, the corresponding distribution of response variable is the Bernoulli distribution, which is also one of exponential distribution family:

$$X \sim B(1,p), \qquad p(x) = e^{x\log(\frac{p}{1-p})+log(1-p)} \qquad (x \in \{0,1\})$$

As the corresponding distribution is of the exponential distribution family, the binary delay indicator can be modeled by generalized linear model. And the logit function is used as the GLM link function to connect the probability and linear predictor.

$$logitfunction : f(x) = log(\frac{x}{1-x})$$

The linear predictor is generated from explanatory variables. But not all the variable influences are linear. To make the model more precise, some explanatory variables are additionally treated by some methods such as logarithm or polynomial smoothing[2].

3.1.2 Variable Selection

The variable selection process helps to achieve the optimal model by GLM. The model could be complex because of the redundant predictors. That is why they should be removed. And the unnecessary variable can also bring in additional noise which decreases the precision of model estimation. Thus, the variable selection process is necessary. The variable selection process distinguishes the variables that should not be used in the model, and then refines the model to the optimal one. The variable selection process from the initial model to the final model is executed in a recursive way, and has mainly two approachs: forward selection and backward selection[3].

[2] The polynomial smoothing uses a polynomial regressor to replace the simple one degree regressor

[3] cf. Julian J. Faraway. 2002. *Practical Regression and Anova using R.*

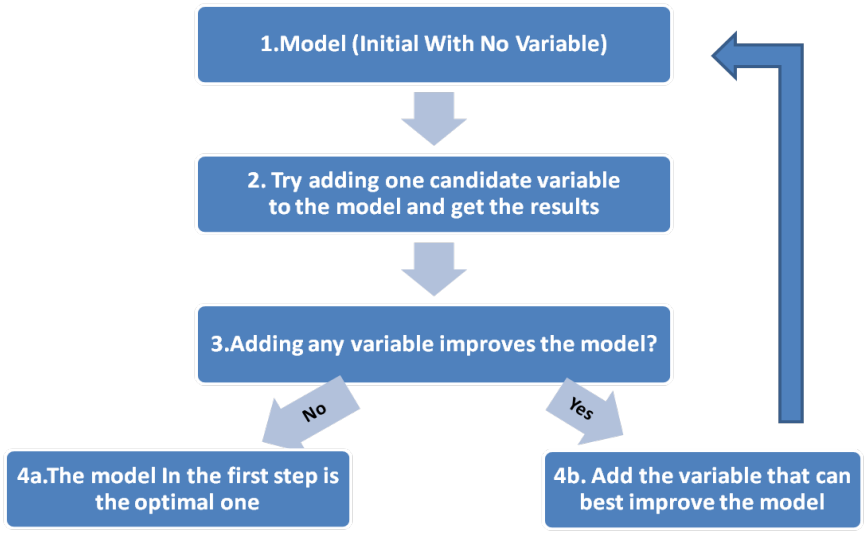

Figure 15: Variable Selection Approach: forward selection

Forward variable selection

Forward selection process starts with no variable in the model, and the best candidate variable that can be added to the model is selected by given criterion. Then, based on the model of one variable, another variable is tried to be added to this model. The best variable to be added is selected by the criterion and is added to the model. This step is executed recursively, until the model can not be improved by adding any variable. The terminal model is the optimal result of the forward variable selection process.

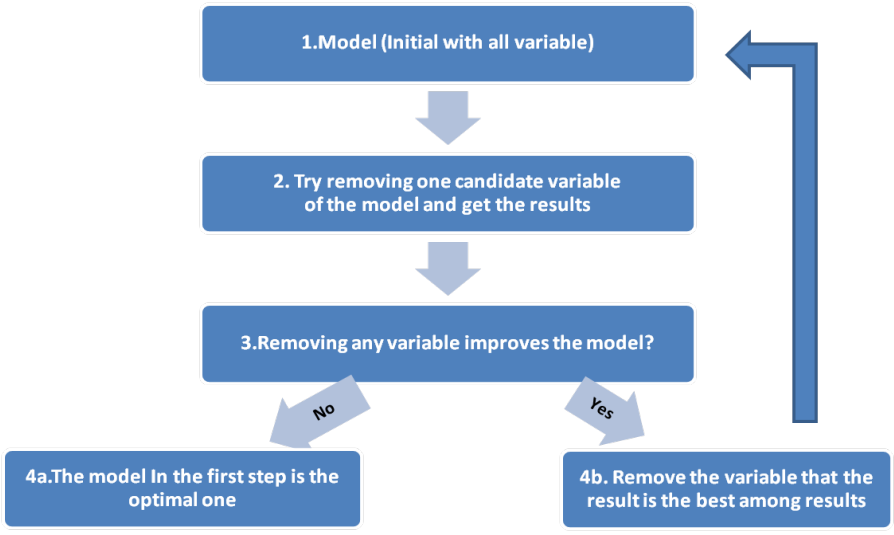

Figure 16: Variable Selection Approach: backward selection

Backward variable selection

Backward selection process starts with model which includes all candidate variables. Based on the model of all variables, this selection process tries removing one candidate variable in the model, and optimal variable is selected to be removed by the given criterion. This step is executed recursively until the model can not be improved by deleting any variable in the model. Then the terminal model is the optimal model of the backward variable selection.

The criteria used in the variable selection process include p-value, AIC, BIC, Mallows's Cp, etc. The p-value is generated for each item in the model, and the item is more significant when the p-value is smaller. The AIC, BIC, and Mallows's Cp are generated for each model, and the model is better if the criterion value is smaller.

A lower p-value means that the variable is more significant. In the forward selection process, the variable with lowest p-value is added if the p-value is lower than the given significance level. In the backward selection process, the variable with highest p-value is removed if the p-value is larger than the given significance level.

Akaike information criterion (AIC) is defined as

$$\boxed{AIC = -2lnL + 2p}$$

where L is the likelihood for an estimated model with p parameters (Akaike, 1973). The Akaike information criterion is a popular criterion for comparing the adequacy of multiple, possibly non-nested models (Eric-Jan Wagenmakers, 2004).

Similarly, Bayesian information criterion (BIC)[4] is defined as:

$$\boxed{BIC = -2lnL + 2plog(n)}$$

where L is the likelihood for an estimated model with p parameters and n observations.

Unlike criterion AIC and criterion BIC, the Mallows's Cp (Mallows 1973) is defined as:

$$\boxed{C_p = \frac{RSS_p}{\hat{\sigma^2}} + 2p - n}$$

where RSS_p is the residual sum square error of the p parameters model, $\hat{\sigma^2}$ is the estimated variation of the sample and n is the number of observations. This criterion trades off the explanation power and the complexity of the model.

The variable selection process that bases on the three criteria above is executed by minimizing the criterion of the model. Therefore, the forward approach process starts with no variable, and then adds step by step the variable that can minimize the criterion value, until adding any variable can not decrease the criterion value. And the backward approach starts with all candidate variables, and then removes step by step the variable that can minimize the criterion value, until removing any variable can not improve the result. The p-value criterion is focused

[4] cf. David Posada Thomas R. Buckley. 2004. *Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests.*. Syst Biol

on the detection of insignificant variables, while the AIC, BIC and Mallow's Cp are focused on the refinement of the model.

3.1.3 Variable Selection and Result

The backward variable process is used to select the variables for GLM model. All candidate variables are used to build up the GLM model in the initial stage. Using the p-value as criterion and 5% as the significance level, the variable with highest p-value is removed recursively. The "DepTime" variable and "Vacation" variable are removed from the model. And the result of the p-value backword selection process is generated.

Variable	*Coefficient*	*P-value*	*P-value>0.05*
(Intercept)	-0.130567	0.00761	TRUE
bs(Year)1	0.26886	3.24E-15	TRUE
bs(Year)2	-0.136818	3.89E-07	TRUE
bs(Year)3	0.206446	<2e-16	TRUE
bs(ArrTime, knots = 480, degree = 1)1	-2.038383	<2e-16	TRUE
bs(ArrTime, knots = 480, degree = 1)2	-0.083868	7.01E-06	TRUE
log(Distance)	-0.127542	<2e-16	TRUE
month2	0.044544	0.02696	TRUE
dayOfWeek2	-0.154189	<2e-16	TRUE
AirlineDL	-0.260185	<2e-16	TRUE

Table 3: Result of P-value Backward Variable Selection

In this GLM model, the quantitative year variable is smoothed by the three degree polynomial functions and the p-values of all the three degrees are significant. The scheduled arrival time variable is treated with two linear interval, which was introduced in the part 2.3.1 (cf. 2.3.1 Time Variables). The month variable, day of the week variable and airline variable are treated as qualitative variables. And the distance variable is treated by the logarithm function to transfer the variance similar to others variable. All variables in this model have p-values that are larger than 0.05, which means that the influence of each variable is significant. (The results of other modalities of "Month", "DayofWeek", "Airline" are attached in the Appendix A)

The GLM model only takes the influence of each solo variable into account. The interaction influences of the explanatory variables are not considered. Therefore, the cross terms of those variables are then added into the model to represent the interaction effect. To prevent too much interaction items caused by many modalities of qualitative variables, the "Month" variable and "DayofWeek" variable are transferred to quantitative variables to create the interaction items (if not, hundred of cross terms would be created, which would increase the model complexity). Then after adding the corresponding cross terms, the P-value variable backward process is used to select those cross terms.

Variable	*Coefficient*	*P-value*	*P-value>0.05*
Year:Month	-4.66E-03	2.00E-16	TRUE
log(Distance):Year	-5.31E-03	0.00919	TRUE
Month:DayOfWeek	-1.87E-03	0.00202	TRUE
log(Distance):DayOfWeek	6.34E-03	0.00386	TRUE
log(Distance):Month	2.78E-03	0.0402	TRUE
Year:ArrTime	-1.26E-05	0.00888	TRUE
log(Distance):ArrTime	-8.18E-05	2.25E-11	TRUE

Table 4: P-value Variable Backward Selection For Interaction Terms

Table 4 shows the result of second P-value backward selection process, the other seven interaction terms are all significant in significance level of 5%, and thus are added to the model.

Though each item in the second model is significant, the model can also be refined by other criterion. Then the step process is introduced to verify the model by the AIC criterion. Step process recursively returns the AIC values after dropping every item in the present model, and compares the AIC values with that of original model to decide which variable is to be removed. Thus, the step process is a backward AIC process.

MSE Value	*AUC Value*
0.1082247	0.681614

Table 5: Result of GLM Model

In the result of the step process of this GLM model, the AIC value (414 493.3) can not be reduced by removing any item in the model. This means by removing any one of those items, the model can not be better in the aspect of AIC criterion. Thus, this model is selected as the optimal model. Then this model is validated by the test set, and the Mean Square Error value and AUC value are generated: 0.1082247, 0.681614.

3.1.4 Generalized Additive Model

The Generalized Additive Model (GAM) is an extension of the generalized linear model. In GAM model, the linear predictor depends linearly on unknown smooth functions of the explanatory variables. And the study is focused on the generation of these smooth functions. In other word, the GAM adds additional smooth functions to improve the linear predictor part of GLM[5]:

$$g(E(y_i)) = X_i\theta + \sum_{j=1}^{n} f_j(x_j)$$

where

$g()$ is a smooth monotonic link function;

y_i are observations of response variable;

[5] cf. Simon N. Wood. 2007. *Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society.* Series B (Statistical Methodology)

X_i are the linear explanatory variable vectors;

θ is the linear explanatory parameter vector;

f_j are smooth functions of variable x_j;

x_j are explanatory variable, which may an explanatory variable vector.

The GAM uses the smoothing function to fit the non-linear effect of each independent variable. For the variable that has non-linear effect such as "vacation" variable, generalized addictive model is more precise than GLM (linear regression does not fit the non-linear influence well). The generalized additive model is more exact when there is non-linear explanatory variable.

MSE Value	AUC Value
0.1078653	0.6851962

Table 6: Result of GAM Model

The GAM model is generated by the training set after adding smooth functions to the quantitative explanatory variables. The test set is used to validate the model, and the mean square error and AUC value are generated, separately 0.1078653, 0.6851962. Comparing to the result of GLM model, either the MSE or AUC value is better than that of GLM model, which means that the smoothing function truly improves the regression.

3.1.5 CART: Classification And Regression Tree

Different from the GLM and GAM, the Classification And Regression Tree (CART) is a binary tree based on non-parameter model. The CART has two tree types: classification tree and regression tree. The classification tree is used for modelling qualitative response variable. In contrast, the regression tree is used for modelling quantitative response variable. However, for classification tree or regression tree, the binary tree is the basic structure.

Binary Tree

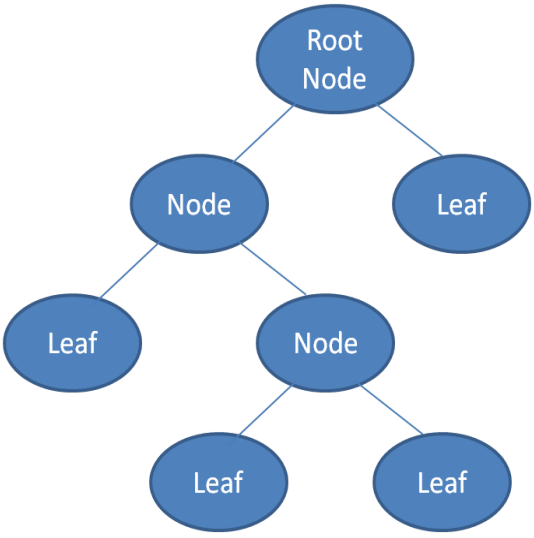

Figure 17: Exemple: Bianry Tree

Figure 17 is an example of binary tree, the binary tree has only one root node. In a binary tree, each node can only be divided into two child nodes. The terminal node that is not divided is also called leaf.

CART Model

The basic idea of CART is generating a binary tree to predict response variable Y from explanatory variables $X_1, X_2, ...X_p$. Each node in the binary tree corresponds to a subset of observations and a prediction of Y. The tree growing process begins at the root node, which contains all the observations. And the tree growing process stops when every leaf can not be divided anymore.[6]

For each undivided node, the CART algorithm analyses each input variable $X_1, X_2, ...X_p$, then the best variable and corresponding decision rule are selected to divide the node into two child nodes. By applying the selected decision rule, the observations of the node are divided into two sub sample. Simultaneously, two different predictions of Y are generated for the two child nodes. This process will repeat until all terminal nodes can not be divided anymore. The reason the node can not be divided is that all the response variable values in the terminal node are the same or the dividing process is stopped by the training rules.

The training rules are used to prevent that the binary tree is over developed. The training rules contain the minimum observation for each node or the minimum improving of each split. When the process terminates with a final binary tree, all terminal nodes are indivisible. Then all the observations are divided into different terminal nodes, or leaves. And the prediction of each leaf is the prediction of corresponding observations.

The prediction method is different for classification tree and regression tree. For classification tree, the prediction of each node is the major modality of the observations in this node. And the predictions of two child nodes must be different. While for classification tree, the prediction of each node is the average value of all the observations in this node.

The decision rule which is used in the prediction tree is specified by the variable type. For quantitative variable, the decision rule is a threshold which can divide this variable into two parts, such as the rule for the quantitative variable X_i: $X_i \geq 0$ or $X_i < 0$. For qualitative variable, the decision rule is a partition of the modalities of this variable. For example, the possible decision rules for a volume variable with modalities { big, median, small } are:

$$\{\text{big, median, small}\} \longrightarrow \{\text{big, small}\}, \{\text{median}\}$$
$$\{\text{big, median, small}\} \longrightarrow \{\text{big, median}\}, \{\text{small}\}$$
$$\{\text{big, median, small}\} \longrightarrow \{\text{median, small}\}, \{\text{big}\}$$

By CART model, the prediction process of new data starts from the root node. From each node, the sample will arrive at one child node by applying the corresponding decision rule. And by applying the decision rule of this child node, the sample will go to one child node of this child node. Recursively, the prediction process will stop when the data arrives at one terminal

[6]cf. Cosma Shalizi, (2009) Classification and Regression Trees. http://www.stat.cmu.edu/cshalizi/350/lectures/22/lecture-22.pdf

node or one decision rule is not applicable. Therefore, each new data has only one unique stopping node. And the prediction of each new data is the the prediction of the corresponding stopping node. The situation that the rule is not applicable is mainly caused by missing value or new modality of qualitative explanatory variable. Therefore, the CART model can also make predictions for the data with missing value or with new modality.

To sum up, the CART model uses the binary tree to present the recursive partition. The predicting process is to decide which node that the new data belongs to, and the prediction value of this node is the prediction of the new data. The CART model has following advantages (cf. Cosma Shalizi, Classification and Regression Trees, 2009):

1) The prediction is fast because there is no complex calculation but only decision making by the decision rule.

2) It is easy to understand the model, and the importance of each variable is clear in the prediction tree.

3) The prediction for the sample with missing data is also feasible. The prediction process may not stop with terminal node, but with some internal node, and the prediction can be made by prediction of this internal node.

4) The prediction of CART model is jagged, so this model works when the true regression surface is not smooth. If the true regression surface is smooth, the piece-wise-constant surface can also approximate it arbitrarily closely (with enough leaves)

5) There exist robust and fast algorithms to achieve the prediction tree of CART model.

Algorithm of Regression Tree

To evaluate the performance of one regression tree model T, the sum square error (SSE) is used as the evaluation parameter.

$$SSE = \sum_{L \in \{leaves(T)\}} \sum_{i \in L} (y_i - m_L)^2$$

Where leaves(T) is the set of terminal nodes.
and m_L is the average value of the leaf L.

Therefore, the aim of the algorithm is to find the CART model which has the minimum sum square error (SSE). The SSE does not increase when one node is divided into two child nodes. Thus, if there is no limitation in the tree developing, the prediction tree will develop into a saturated tree, in which each leaf has only one observation. But for the regression tree, the response variable is a random variable (from the error component), and the estimation should be the mathematical expectation of this random variable. In this case, the saturated regression tree is surely over fitted and a part of nodes is unnecessary. Then the prediction is not precise and the calculation capacity is wasted in the growing process of unnecessary nodes. Thus, the training rules in the tree growing process are mandatory, such as minimize observations for each node and minimum sum square error reducing for each node split. Assuming limitation of

the minimize observations of node is d and limitation of minimum SSE reducing is s, the tree growing algorithm of regression tree is:

1. Start with one single node, calculate the SSE.

2. If all values in the node are the same, stop the process. Otherwise, find all binary partitions of each input variable.

3. Calculate the SSE decreasing ΔSSE of each binary partition and find the largest one.

4. If the largest ΔSSE is larger than SSE reducing limitation d, use this partition and calculate the observation numbers of two parts. Otherwise, stop the process.

5. If the observation numbers of two parts are larger than observation limitation d, create two new child nodes. Otherwise, stop the process.

6. For each new child node, return to step 1.

This algorithm creates all possible partitions that do not break the training rules. In addition, if in one process, there are two partitions with same decreasing in SSE, either one can be used and the selection is random. The values of limitations d and s are very important for CART. Large limitation values of d and s will cause the tree growing process to stop too early. But with small limitation values, the CART model will have too many redundant nodes, which import additional noises.

Overfitting and Cross Validation

To solve the problem of overfitting, the cross validation is used in the CART model to find the optimal tree size. The training data is divided into two parts. One part is used for training and another part is used for test. Then the best node number is the one with minimum sum square error of prediction of the test data. Thus the idea of this method is to generate a big binary tree with small limitations (but not too complex to grow), and to prune the tree at the optimal node number, which has the minimum sum square error for the test data.

So as to estimate the departure delay probability, the regression tree is used instead of classification tree. When the algorithm is applied to the training data, small limitation values are configured to prevent the tree from becoming too big (or at the end, becoming saturated tree). And in the first step of modelling, 4 329 nodes are generated.

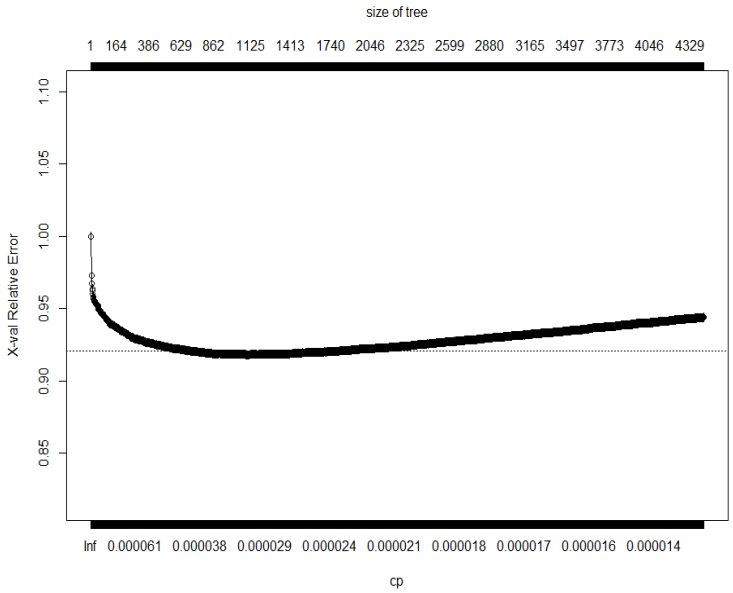

Figure 18: Relative Error Curve of Test data

To calculate the optimal tree size, the 5-fold cross validation process is executed. The relative error curve of test part is then presented in Figure 18, the X_val Relative Error means the sum square error of the test data by cross validation, which is standardized by the sum square error without split. The curve of X_val Relative Error decreases firstly and then increases as the number of tree increases. The minimum relative error of test data is achieved when the tree number equals 1 096.

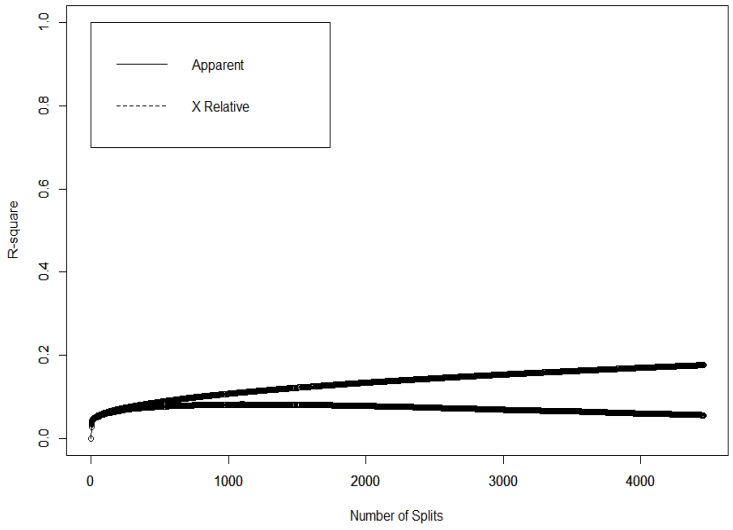

Figure 19: R Square Curve of Training Data and Test Data

The same conclusion can be concluded by the R-square curves in Figure 19. In this graph, the R square value of training data always increases, meaning that the model fits training data better as the node number increases. However the R-square value of test data increases firstly and then decreases as the node number increases. The difference between performances of training and test data reflects the appearing of overfitting when the tree number is larger than 1 096.

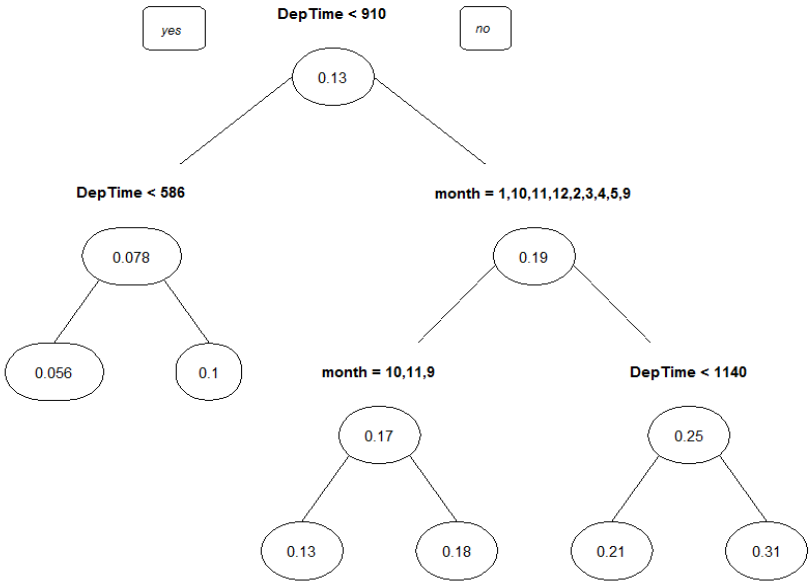

Figure 20: First Several Nodes of Optimal CART Model

As the CART model has many nodes, the first several nodes of the CART model are introduced in Figure 20. The average delay rate in the root node is 0.13, with 571 095 observations. The root node is divided by the rule "Whether scheduled departure time is early than 15:10 or not"[7]. The observations with positive answer belongs to the left child node, which has 302 966 observations and has 0.078 average delay rate. And the other observations with negative answer belongs to right child node, which has 268 129 observations and has 0.19 average delay rates. Similarly, its two child nodes are separately divided by rules "Whether scheduled departure time is early than 9:46" and "Whether the month of flight in between 6 and 8".

3.1.6 Random Forest

Random Forest is an aggregate method of the classification and regression tree. Despite the advantages of the CART model, it is not stable: the model may change a lot while the training data changes. The law of large numbers concludes that the observed average of one random variable will converge to the mathematical expectation when the sample size is big enough. Similarly, to eliminate the instability of CART model, the Random Forest method builds a set of random CART models, and utilizes the average result of the set of random tree as the prediction. In this aspect, the random forest method is more stable by using aggregate method.

[7] As the scheduled departure time is treated in minutes in the data, the value 910 means 15:10 and the value 586 means 9:46

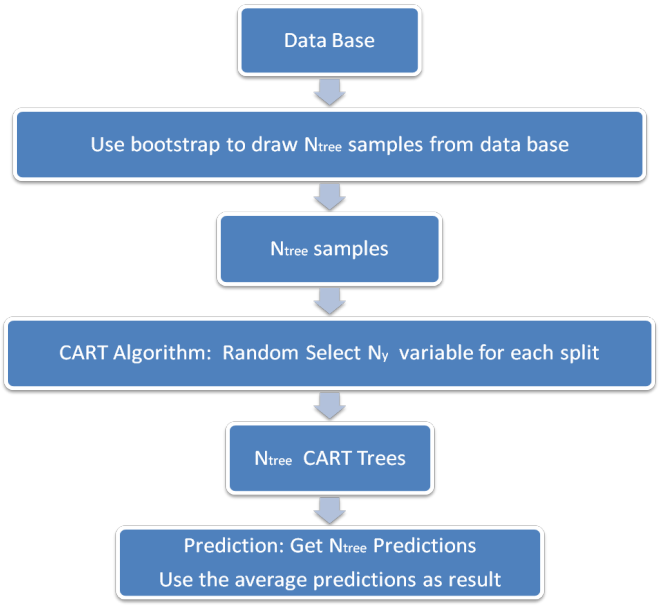

Figure 21: Algorithm of Random Forest

The algorithm of random forest is presented in Figure 21, the N_{tree} represents the number of binary trees in the random forest model, and the N_y is the number of explanatory variables that are random selected in the tree split process. The random forest algorithm starts with random sampling. It draws N_{tree} samples by bootstrap method and then uses the samples to generate N_{tree} binary trees by CART algorithm. However, the CART algorithm in Random Forest is modified. In the CART algorithm of random forest, the best split is not chosen from the splits of all explanatory variables, instead, N_y explanatory variables are randomly selected each time, and the best split is chosen from the possible splits of the N_y explanatory variables. Therefore, each binary tree is grown with the random variable selection. Then the prediction of Random Forest is the average of predictions of N_{tree} binary trees.

The estimation of "out of bag" error is used to evaluate the model prediction performance. For each bootstrap sample of random forest, there exist data that are not included in this bootstrap sample, which is called Out-Of-Bag data, or OOB data. Then, the OOB data can be used as test data, and the difference between true value and prediction of OOB data is called OOB error. Thus, the "out of bag" error of the random forest model is the mean square error of the predictions of OOB data. As the N_{tree} is big enough, the OOB error of random forest is quite accurate[8].

[8] cf. Andy Liaw and Matthew Wiener. Classification and Regression by Random Forest. R news 2(3), 18-22, 2002

Variable	%IncMSE	IncNodePurity
Year	41.92	32.82
month	66.28	333.30
dayOfWeek	29.38	15.36
Airline	46.28	124.01
Vacation	47.49	48.69
DepTime	48.43	1257.77
ArrTime	30.81	855.34
Distance	43.69	118.41

Table 7: Importance Table Of Random Forest

In the training process using training data set, N_{tree} is set as 1 000, N_y is set as two (one third of explanatory variables), node limit for CART is set as 100 (limited by the calculate capacity).

Then the Random Forest model is generated. The variable importance of the random forest model is presented in Table 7. The IncNodePurity[9] values of "DepTime" and "ArrTime" are much bigger than that of other variables, which means the two variables are more important in the random forest model.

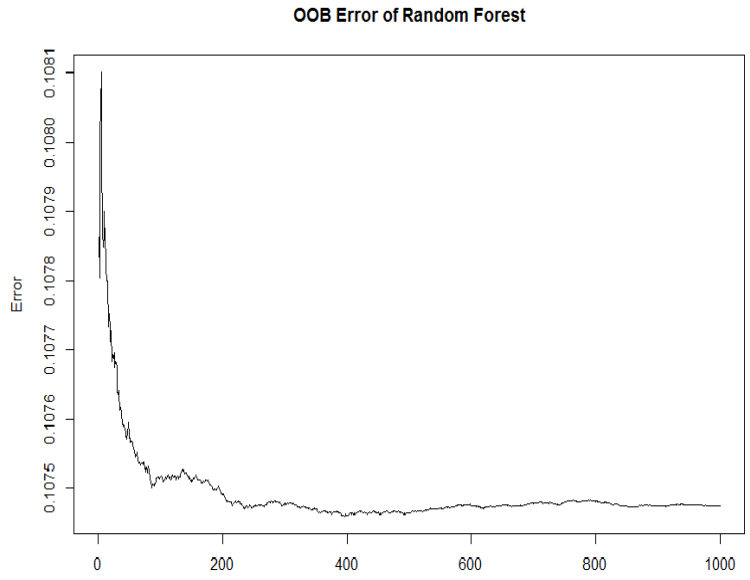

Figure 22: Out-of-bag Error Of Random Forest

The "out of bag" error is presented in Figure 22. The "out of bag" error is unstable when the tree number is small. However, as the tree number increases, the "out of bag" error becomes stable, and oscillates small around 0.1075[10].

[9] IncNodePurity is the sum SSR reducing of the nodes which are splitted by the corresponding variable

[10] The difference between probability and observation does not change too much, then the change in the error is relatively small

3.1.7 Gradient Boosting Model (GBM)

Similar to other machine learning methods, the aim of the gradient boosting model is to find the optimal predictive function $F^*(x)$ that maps the explanatory variables $x = \{x_i\}$ to the response variable y. Given the loss function: $\Psi(y, F(x))$, the aim is to find $F^*(x)$ which minimizes the loss function. Therefore the optimal function is defined in formula (3.1).

$$F^*(x) = arg\min_{F(x)} E_{y,x}(\Psi(y, F(x)) = arg\min_{F(x)} E_x[E_y(\Psi(y, F(x))|x] \tag{3.1}$$

The error function $\Psi(y, F(x))$ in the formula (3.1) includes square error $|y - F|^2$, absolute error $|y-F|$ for quantitative y, and also negative binomial log-likelihood for binary $y \in \{-1; 1\}$, etc.

Among different gradient boosting algorithms, the one of J.H. Friedman (2002)[11] is wildly accepted. This algorithm is also used by some machine learning packages of statistic software[12].

The predictive function $F(x)$ in formula (3.1) is in additive form and uses the parameters $P = \{\beta_m, \alpha_m\}_0^M$. The function h(x, α) used in following formula is a simple function with parameter α:

$$F(x, P) = \sum_{m=0}^{M} \beta_m h(x; \alpha_m)$$

The optimization problem of the predictive function in formula (3.1) is transferred to the optimization problem of the function parameters in formula (3.2), where P^* is in the form of $\sum_{m=0}^{M} P_m$.

$$P^* = arg\min_{P} \phi(P) = arg\min_{P} E_{y,x}(\Psi(y, F(x; P)) \tag{3.2}$$

Apart from the initial status P_O which has no function parameter, the new parameter is generated by solving (3.2) using the steepest-decent methods[13]. Recursively, the new parameters are generated one by one. And the generated new additive functions are added to the prediction function. Naturally, the model becomes more subtle if more additive functions are trained and integrated.

To estimate the probability of binary response variable, there are two possible error functions (or deviation function for estimation). The corresponding models are Bernoulli GBM and Adaboost GBM[14].

[11] cf J.H. Friedman (2002). *Stochastic Gradient Boosting.* Computational Statistics and Data Analysis 38(4):367-378.

[12] The package GBM of software R adopts this algorithm

[13] Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function

[14] cf. G. Ridgeway. 2007, Generalized Boosted Models: A guide to the gbm package

Deviance of Bernoulli GBM:

$$-2\frac{1}{\sum w_i}\sum w_i(y_i f(x_i) - log(1 + exp(f(x_i))))$$

Deviance of Adaboost GBM:

$$\frac{1}{\sum w_i}\sum w_i exp(-(2y_i - 1)f(x_i))$$

In this equation w_i is the weight of i^{th} observation, and y_i, x_i are the values of response variable and explanatory variables. Combining the deviance function and steepest-decent methods, the parameters of the corresponding GBM model are generated recursively[15].

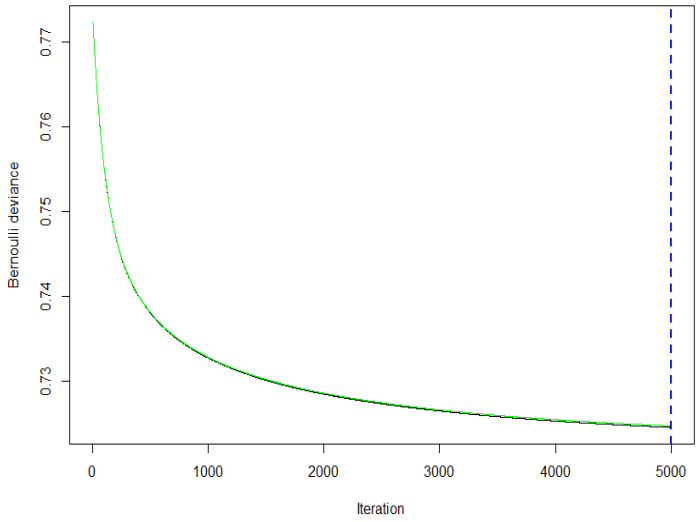

Figure 23: Cross Validation For Bernoulli GBM

Setting tree number as 5 000 and shrinkage parameter as 0.005, the Bernoulli GBM model is generated from the training set. Similar as the CART model, to prevent the overfitting effect, a 5-fold cross validation process is executed. Figure 23 presents the cross validation result of Bernoulli GBM. The cross validation result shows no overfitting effect, as either training error or test error as the tree number increases. In consequence, the model is optimal when tree number is 5 000.

[15] cf. G. Ridgeway. 2007, Generalized Boosted Models: A guide to the gbm package

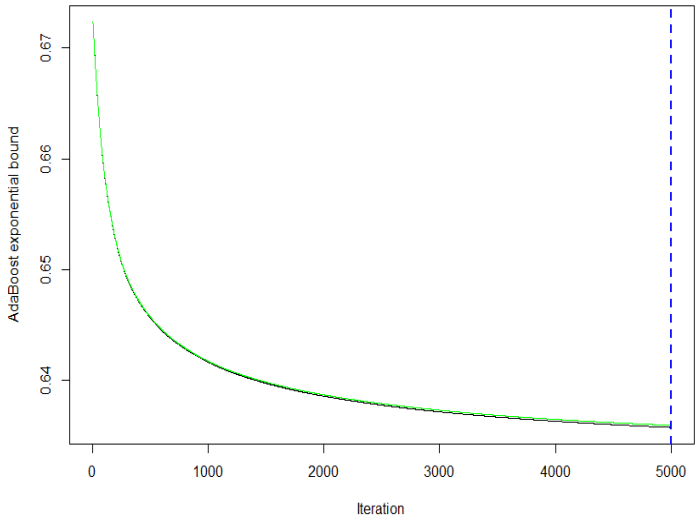

Figure 24: Cross Validation For Adaboost GBM

Same as the Bernoulli GBM model, the Adaboost GBM is generated from the training set. Figure 24 presents the 5-fold cross validation result of Adaboost GBM. Either training error or test error decreases as the tree number increases, therefore the cross validation result shows no overfitting effect. Also, the model is optimal when the tree number is 5 000.

By the article of J.H. Friedman (2002), the relative influence of one variable in the regression GBM is calculated by the formula (3.3).

$$J_j = \left(E_x \left[\frac{\partial \hat{F}(x)}{\partial x_j} \right]^2 \cdot var_x(x_j) \right)^{1/2} \tag{3.3}$$

where $\hat{F}(x)$ is estimated predictive function and x_j is j^{th} explanatory variable of x.

Variable	Name	Bernoulli	Adaboost
Departure Time	DepTime	77.7085	75.3660
Month	Month	14.290	14.8172
Distance	Distance	4.002	4.595
Career	Airline	3.214	3.672
Arrival Time	ArrTime	0.402	0.771
Year	Year	0.238	0.480
Day of the Week	DayOfWeek	0.145	0.299

Table 8: Variable Importance Of Bernoulli and Adaboost GBM

Table 8 presents the variable importance result of the two GBM models. In consequence, the "DepTime" variable and "Month" variable are the two most important variables for the regression. Although the"DayofWeek" variable has the least variable importance, the correspond value is larger than zero. This means that the "DayofWeek" variable has also small influence on the modelling. In this aspect, no variable is removed from the two models.

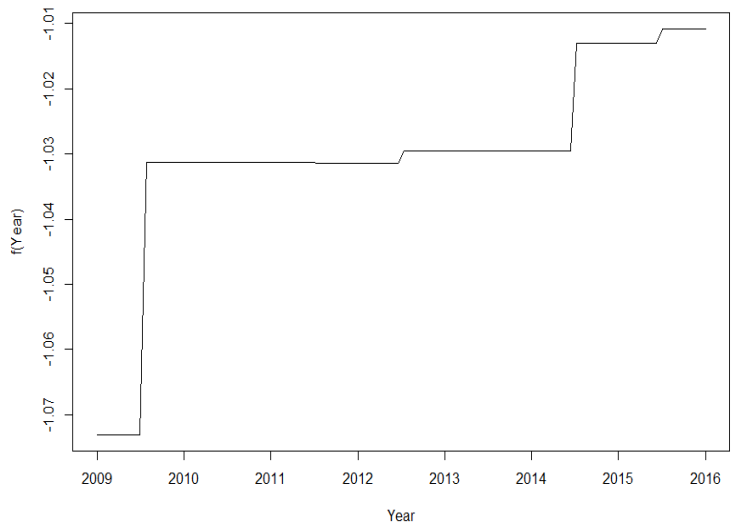

Figure 25: GBM Function for year variable

The generated function of year variable is a monotonic increasing function, which means that there exists an increasing trend of the departure delay probability. This function also shows that the changes in the period 2008-2009 and in the period 2014-2015 are large, but the changes in the period 2009-2014 is relatively stable.

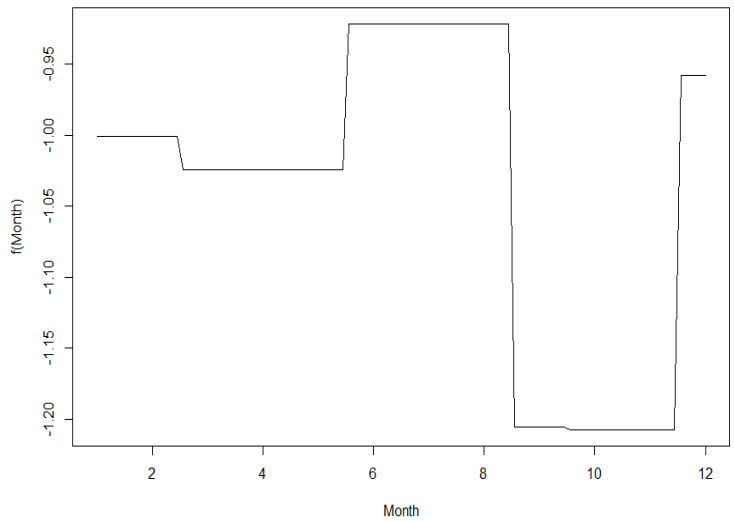

Figure 26: GBM Function for Month Variable

The generated function of month variable is not regular. This function is high at month number 8, 9 and 12, meaning that the delay probability is higher in the corresponding months. This cycle is mainly caused by the vacations in the summer and winter, which is similar to the cycle in the statistic graph.

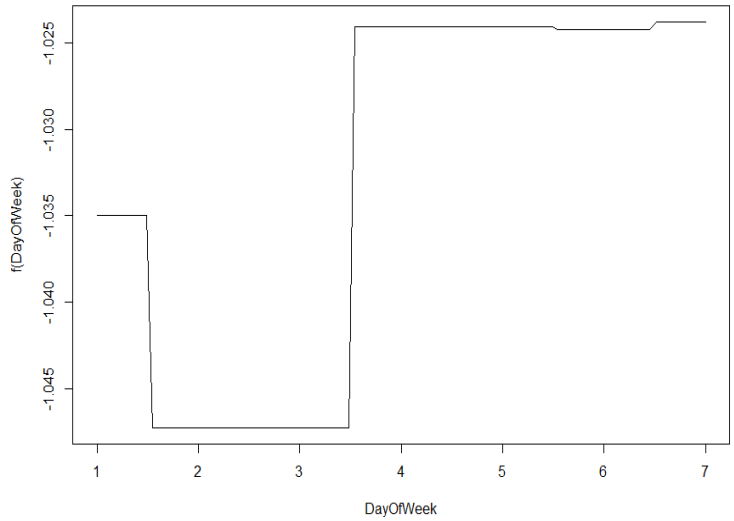

Figure 27: GBM Function for Day of the Week Variable

The generated function of day of the week variable is high from Thursday to Sunday, and then decreases from Monday to Wednesday. This cycle is also similar to the cycle of work day and weekend day in the week.

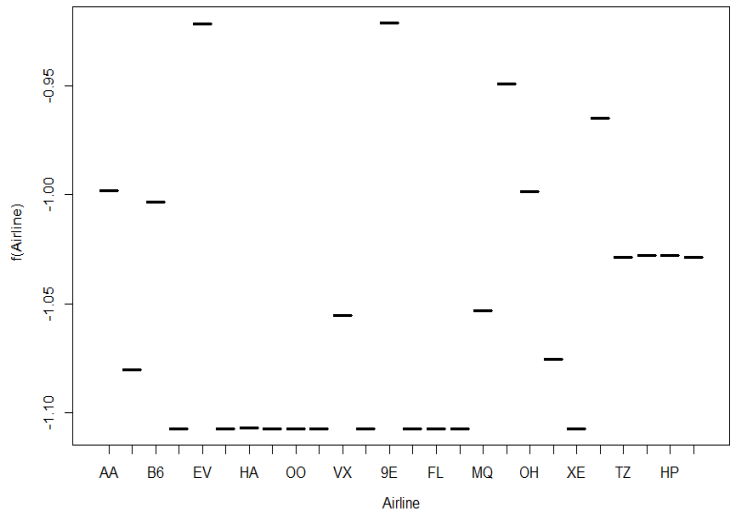

Figure 28: GBM Function for Airline Variable

The generated function of airline variable is well dispersed. This means that the influence of different airlines is well integrated in the GBM model as the Airline variable is qualitative.

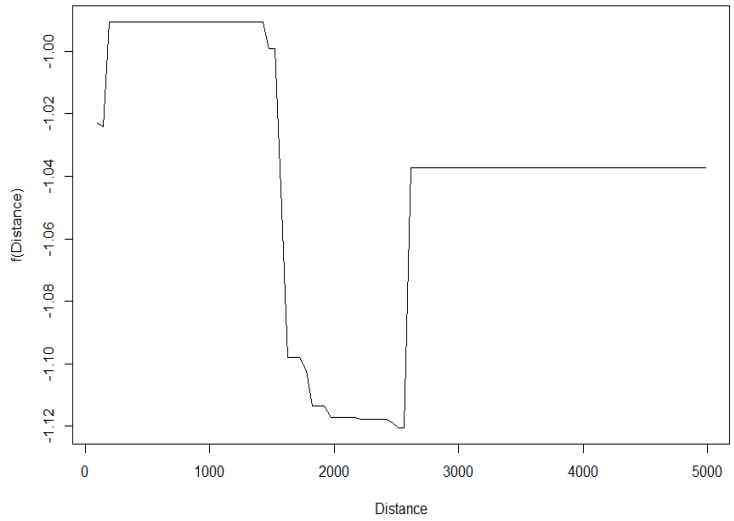

Figure 29: GBM Function for Distance Variable

The generated function of distance variable is irregular, and shows three different intervals. This function is not monotonic and reaches the lowest value at 2 500km.

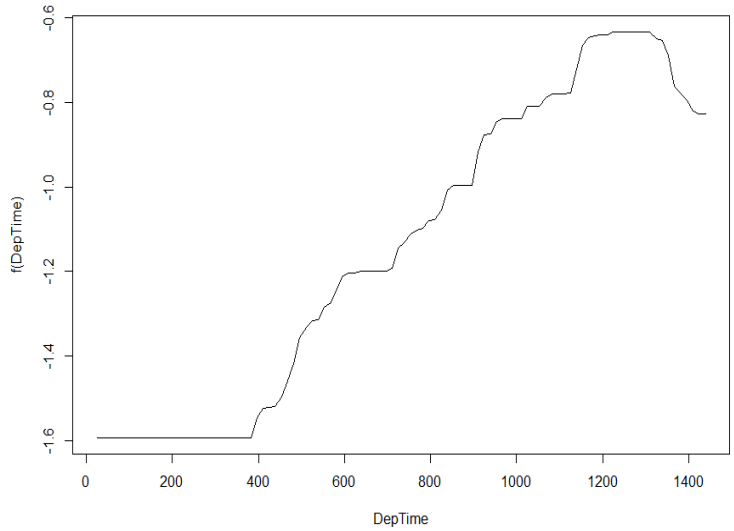

Figure 30: GBM Function for Scheduled Departure Time Variable

The generated function of scheduled departure time variable ("DepTime") and scheduled arrival time variable ("ArrTime") are presented in Figure 30 and Figure 31. The generated function of scheduled departure time variable integrates the time influence from six o'clock to midnight, while the function of scheduled arrival time variable integrates the time influence from midnight to six o'clock. The combination of those two functions thus performs the increasing trend in day time and decreasing trend in night time.

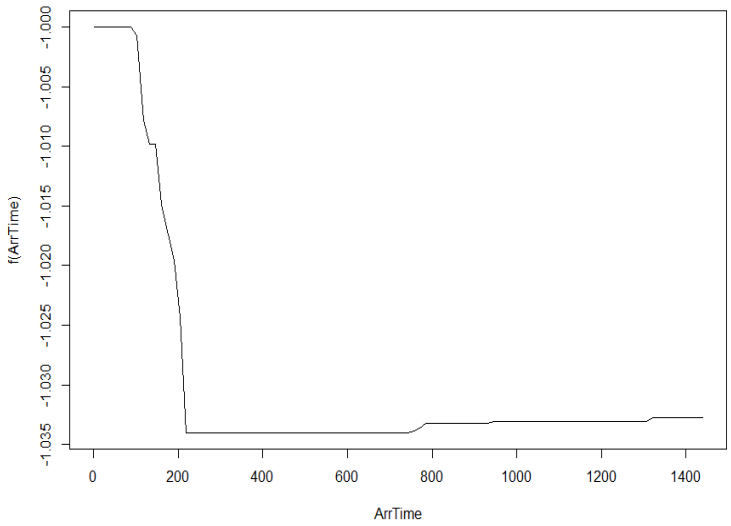

Figure 31: GBM Function for Scheduled Arrival Time Variable

Figure 25 to Figure 31 present the generalized additive functions of the seven variables in the Adaboost GBM model. The generalized addictive functions of Bernoulli GBM model are similar (therefore are not introduced here). Those functions show how the explanatory variable works on the response variable.

Model	MSE Value	AUC Value
Adaboost GBM	0.10811	0.6842
Bernoulli GBM	0.10807	0.6839

Table 9: Result of GBM Models

The two models are applied to the test set to evaluate the estimation performances, and the corresponding AUC and MSE are generated. As presented in the table, the AUC and MSE of Bernoulli GBM model are 0.6839 and 0.10807. The AUC and MSE of Adaboost GBM model are 0.6842 and 0.10811. The values of the two models are very close, which means that the performances of the two models are also similar.

3.2 Discrete Approach

As the database contains the departure delay information in minutes, the flight departure delay could be studied as discrete variable. The discrete departure delay variable has more information than the binary departure delay variable (binary delay variable can be produced by the discrete delay variable). Considering the additional information in the discrete departure delay variable, the delay probability estimation in the discrete approach may be more precise than the estimation in the binary approach.

3.2.1 Variable Censuring

Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. Left censoring is when the event of interest has already occurred before enrollment. In practice, one can be confronted to right-censoring (if X is the variable of interest, the observation of censoring C indicates that X is not less than C) or left-censoring (the observation of censoring C indicates that X is not larger than C), and the two types of censoring can be observed simultaneously[16].

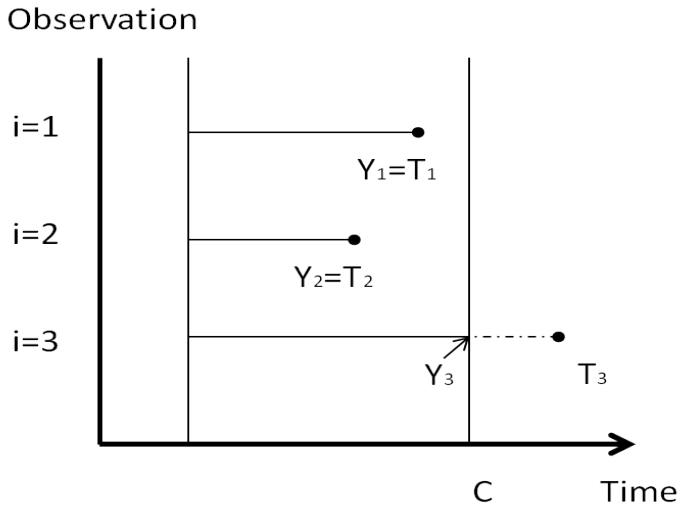

Figure 32: Right Censuring

For example, there are three observations for three events, real event times are T_1, T_2, T_3, and the censuring time is C. Let Y_1, Y_2, Y_3 be the observed values. Figure 32 presents the right censuring, for observation i: $Y_i = max\{T_i, C\}$. The first two observations are not censured, while the third observation in the graph is right censured. The dotted line part is censured and can not be observed, then the observed value Y_3 equals C. Similarly, Figure 33 presents the left censuring, for individual i: $Y_i = min\{T_i, C\}$. The third individual in the graph is left censured. The dotted line part is not observed and the observed value Y_3 equals C.

[16] cf. Frédéric PLANCHET. 2016. *Statistique des modèles paramétriques et semi-paramétriques* . http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance3.pdf?OpenElement

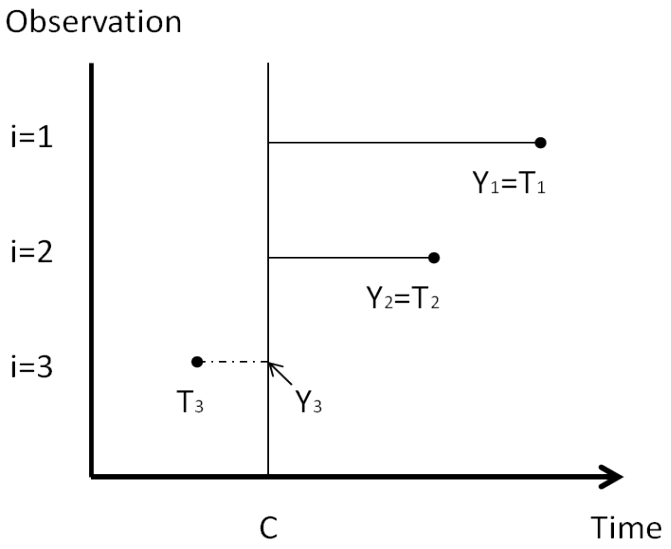

Figure 33: Left Censuring

If the plane takes off earlier than the scheduled time, there is no delay and the flight delay is zero. The flight departure delay is a non-negative variable. However, the case that the plane takes off earlier is not the case the plane takes off on-time. The departure time difference variable can be positive or negative (this variable is the time difference between the scheduled departure time and the real departure time, the value is negative when the plane takes off earlier). In this case, the flight delay variable is a left-censured variable, and the censuring variable equals zero (means the observation starts at scheduled departure time). And the original variable is the departure time difference variable.

In Figure 33, the T_i is the departure time difference (negative when the plane takes off earlier). As the early departure is taken as on-time departure, the censuring variable C equals zero, means that the flight has no departure delays. Then the flight departure delay is the observation value Y_i=min$\{T_i, C\}$=min$\{T_i, 0\}$. As a result, the negative departure time difference is censured to zero.

In the discrete approach, the left-censured discrete flight delay is considered as the dependant variable (if not censured, the negative part of departure delay will makes the model complex, therefore the discrete flight delay is left-censured to simplify the analysis). By modelling the flight delay distribution, we can estimate the probability of the flight delay which is longer than thirty minutes. In this case, the first step of this approach is to come up with all the possible distributions for the dependant variable. The second step is to study the relationship between the distribution parameter and the independent variables. Then by estimating the distribution parameter, the delay probability can be estimated.

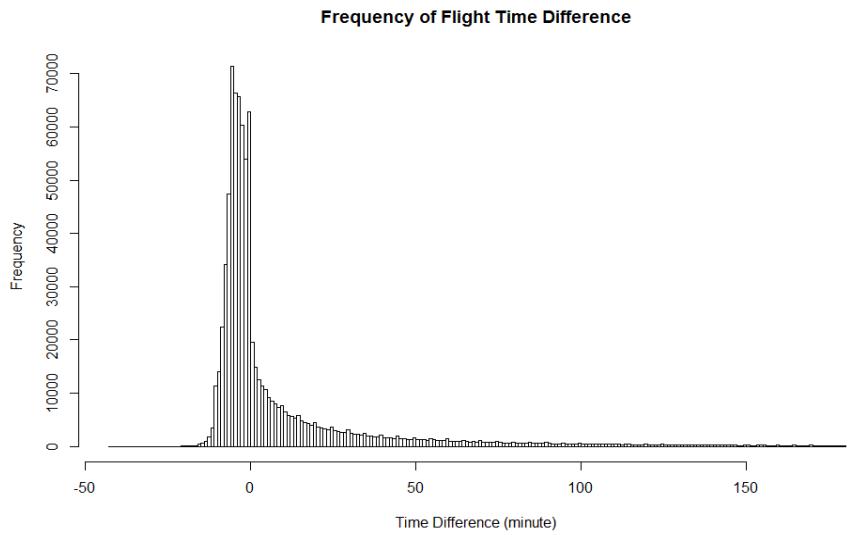

Figure 34: Frequency of Departure Time Difference (before censuring)

Figure 34 and Figure 35 are the histograms of the departure time difference and departure delay. Figure 34 shows the distribution of departure time difference, in which the negative part has higher frequencies than the positive part. And the frequency of the positive part decreases as the time difference increases. In positive part, the departure delay could arrive at 150 or larger, while in the negative part the departure delay is more concentrated and is not less than -25.

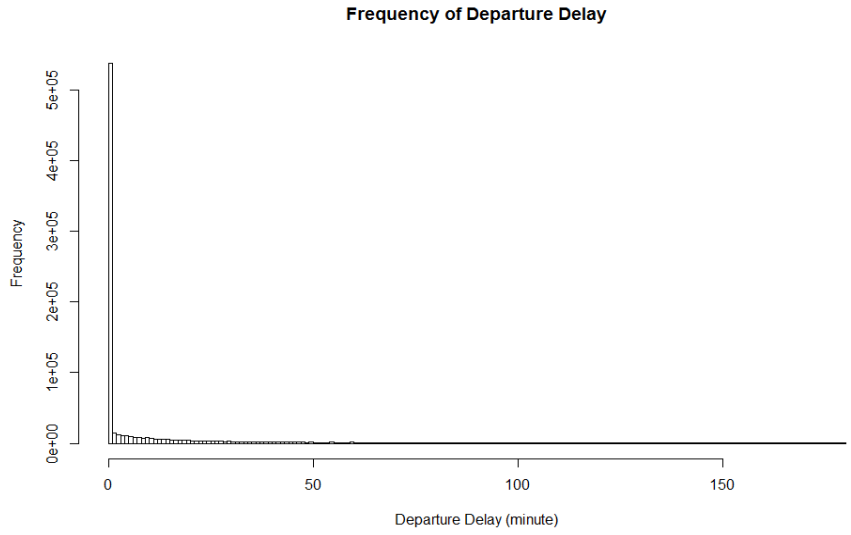

Figure 35: Frequency of Departure Delay (after censuring)

Figure 35 shows the distribution of the left-censored variable: departure delay variable. The frequency of zero departure delay is very high, equals 62.1% (518 297 out of 834 951). This means that most flights take off on time.

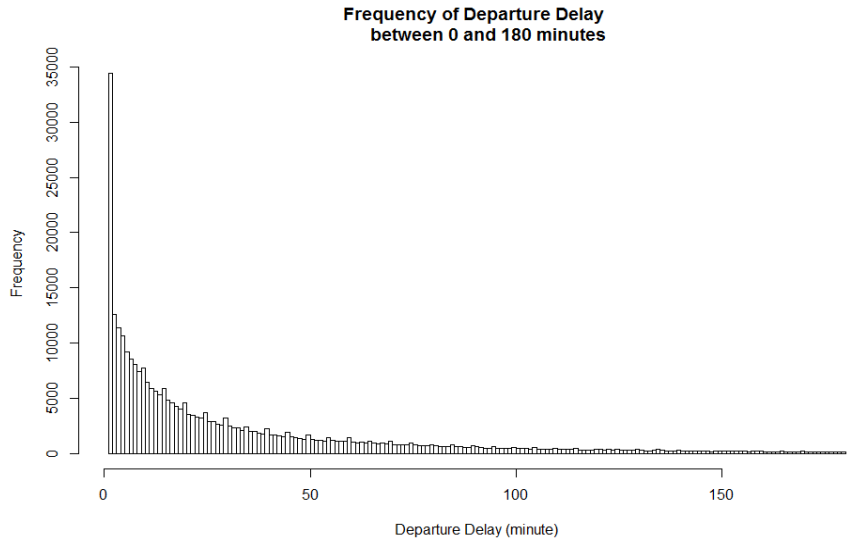

Figure 36: Departure Delay Distribution between 0 and 180 minutes

Figure 36 and Figure 37 are focused on the non-zero part to study the distribution of delay frequency frequency better. Those two graphs show an approximate concave function for the frequency of the left censured flight delay. The first graph only uses the non-zero delay data which are shorter than one hundred and eighty minutes, and the second graph uses the delay data that are longer than one hour.

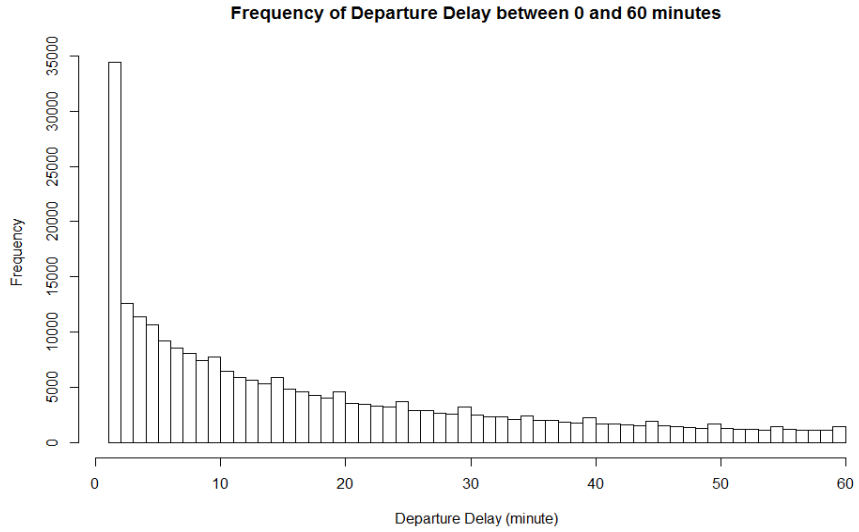

Figure 37: Departure Delay Distribution between 0 and 60 minutes

We can observe that there are small peaks for each five minutes. One possible reason is that the time of delay is more likely to be fixed at the multiple of five minutes. Without the five minutes peaks, this curve is almost a concave decreasing function, which decreases fast around zero and decreases slowly at large value. In this figure, those five minutes peaks are also proportional to the delay frequency, and the surplus part of each peak decreases as the delay frequency decreasing. Considering the complexity of the discrete model, our study does not take the peaks of every five minutes into account. And because of the proportional property of those five minutes peaks, those peaks have little influence on the final estimation of probability

that departure delay is longer than thirty minutes.

3.2.2 Zero-inflated Models

The zero-inflated models are used to cope with the count data with too many zeros. The first zero-inflated model is the zero-inflated Poisson model, which concerns a random event containing excess zero-count data in unit time (Lambert, Diane 1992). One could think of the zero-inflated Poisson model as the special mixture model of two parts: zero inflated part and Poisson part:

$$P(Y=i)=\begin{cases} w+(1-w)e^{-\lambda} & i=0; \\ (1-w)e^{-\lambda}\lambda^{y}\frac{1}{y!} & i=1,2,3...; \end{cases} \qquad \begin{array}{l} Y=\text{count variable} \\ w=\text{probability of zero inflated part} \\ \lambda=\text{Poisson parameter of second part} \end{array} \tag{3.4}$$

The zero-inflated negative binomial model takes the distribution of negative binomial as the second count part. Figure 38 present the two processes. When the zero inflated part is zero, the result equals zero, when the zero inflated part is not zero, the count result follows the negative binomial distribution.

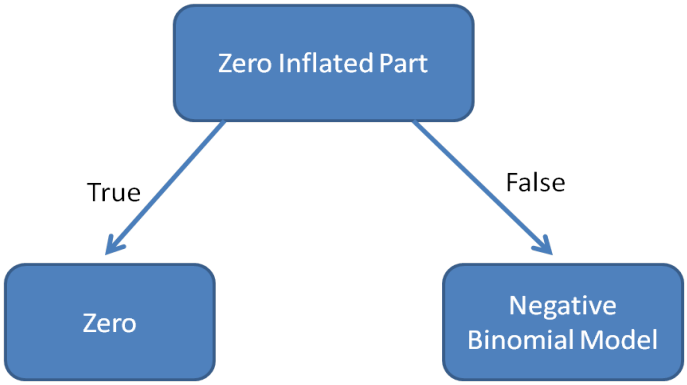

Figure 38: Two process for ZINB model

By replacing the Poisson mass density by the Negative Binomial mass density, the formula (3.5) presents the mix distribution formula for the zero-inflated model, where the parameter **r** is the shape parameter. When **r** = 1, this model becomes the zero-inflated geometric model.

$$P(Y=k)=\begin{cases} w+(1-w)(1-p)^{r} & k=0; \\ (1-w)\binom{k+r-1}{k}p^{k}(1-p)^{r} & k=1,2,3...; \end{cases} \qquad \begin{array}{l} Y=\text{count variable} \\ w=\text{probability of zero inflated part} \\ r=\text{negative binomial parameter} \end{array} \tag{3.5}$$

Considering the high frequency of zero delay and the concave property of non-zero delay, the zero-inflated model can fit the left censured departure delay well. In statistics, a zero-inflated model is a statistical model based on a zero-inflated probability distribution, which is like the

distribution of departure delay. For the distribution function of the non-zero part, simple exponential distribution is not suitable, because the frequency jump between one minute and two minutes or between two minutes and three minutes is not proportional to the frequency of one minutes and that of two minutes. And the first jump is much larger (the first jump is over 1/2 of the frequency of one mintues while the second jump is less than 1/10 of the frequency of two minutes.). Considering this characteristic, the geometric distribution could be one possible distribution. In this case, Poisson distribution, negative binomial distribution, and geometric distribution are the possible distribution models.

3.2.3 Zero Inflated Regression

Those three zero-inflated models are studied by the R software. Those zero-inflated models use the same database as in the binary approach: data of New York John F. Kennedy International Airport. There are two parts of estimation for the departure delay in the zero-inflated models. The first one is the zero part, which is a probability estimation. The second one is the count part, which uses the previously described distributions to study the departure delay distribution. And for each part, these independent variables "Year", "Month", "Day of the Week", "Airline", "Vacation", "Departure Time", and "Arrival Time" are used. The variables "Year" and "Vacation" are treated as a 3-degree polynomial smoothing function, and "Departure Time" variable and "Arrival Time" variable are treated by two intervals (for "Departure Time" variable, the third interval has no effect). And variables "Month", "Day of the Week", and "Airline" are also categorical variables in this approach. Because of the complexity of model, the interaction items are not considered[17].

Part	Name	Coefficient	P-value
count part	Month2	0.022493	0.117
count part	AirlineAS	0.093204	0.477
count part	AirlineVX	0.018387	0.354
count part	AirlineYV	0.02263	0.782

Table 10: Insignificant Items of ZIP

As a result of the zero-inflated Poisson model, all variables are significant for the zero part, while several variables are not significant for the count part. For the count part, the insignificant modalities of month variable, airline variable are listed in Table 10.

[17] 12 interaction items in each part(as in the result of GLM model) will cause the model too large to be used, and may also import the over-fitting to the model.

Part	Name	Coefficient	P-value
zero part	Month12	-0.002723	0.853819
zero part	Month4	-0.001736	0.912589
zero part	dayOfWeek5	-0.001379	0.903014
zero part	dayOfWeek7	0.005608	0.624643
zero part	AirlineNW	-0.071109	0.190472
count part	Month6	-0.016439	0.295106
count part	AirlineAS	0.135808	0.327328
count part	AirlineVX	0.00586	0.793332
count part	AirlineYV	0.070728	0.426545

Table 11: Insignificant Items of ZINB

As a result of zero-inflated negative binomial model, all the variables are significant except several modalities of categorical variable. For the zero and count part, the insignificant modalities of month variable, airline variable, and day of the week variable are listed in Table 11.

Part	Name	Coefficient	P-value
zero part	dayOfWeek5	-0.001202	0.881992
zero part	dayOfWeek7	0.003428	0.675822
zero part	AirlineNW	-0.069441	0.074829
zero part	Month12	-0.002499	0.813317
zero part	Month4	-0.001919	0.865445
count part	Month2	0.025527	0.0812
count part	AirlineAS	0.109343	0.4084
count part	AirlineVX	0.014465	0.4747
count part	AirlineYV	0.036699	0.6576

Table 12: Insignificant Items of ZIG

AS a result of zero-inflated geometric model, all the variables are also significant except the "departure time" variable in the count part and several modalites of categorical variable. For the categorical variables in the zero and count part, the insignificant modalities of month variable, airline variable, and day of the week variable are listed in Table 12.

In summary, in the aspect of significance, the result of zero-inflated negative binomial model and that of zero-inflated geometric model are similar. And the "Departure Time" variable is not significant in the count part for all three models.

3.3 Model Comparison

The nine models are evaluated by two tests and two criteria. The Vuong's Non-Nested Test allows comparing the three discrete models. The Hosmer-Lemeshow Test allows testing the goodness of fit for the six binary models. And the criteria of AUC (Area Under the Curve) and MSE (Mean Square Error) are used to quantitatively analyse the model performance. Combining the result of model comparing and qualitative analysis, the best model can be selected.

3.3.1 Vuong's Non-Nested Test

The Vuong's non-nested test is used to select the best model between two non-nested models. This test uses the likelihood ratio statistic as the test statistic. The conditional probability is considered so as to present the likelihood of response variable, such as

$$F_\theta = f(y|z;\theta); \theta \in \Theta$$

(y, z, θ, Θ) are response variable, explanatory variables, model parameters and parameter space. As measured by the minimum KLIC (Kullback-Leibler Information Criterion, 1951), the distance between conditional model and the true conditional density $h^0(y|z)$ is defined by the formula following:

$$Distance = E^0[\log h^0(y|z)] - E^0[\log f(y|z;\theta_*)]$$

$E^0[\cdot]$ denotes the expectation with respect to the true joint distribution of (y, z), and θ_* is the pseudo-true value of θ [18].

Thus an equivalent selection criterion can be based on the quantity $E^0[\log f(y|z;\theta_*)]$, and the "best" model is the one with the largest quantity. Meanwhile, the log-likelihood ratio (LR) statistic is a consistent estimator of the quantity[19]:

$$E^0[\log f(y|z;\theta_*)] - E^0[\log g(y|z;\gamma_*)]$$

Therefore, the log-likelihood ratio (LR) statistic can be used as the criterion of model comparing. Given two non-nested models $F_\theta = f(y|z;\theta) and G_\gamma = g(y|z;\gamma)$ with two pseudo-true parameter $(\theta_*;\gamma_*)$, the model selection equals to the validation of those three hypotheses[20] :

$$H_0 : \qquad E^0[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}] = 0 \tag{3.6}$$

[18] cf. White, H. (1982), *Maximum likelihood estimation of misspecified models*, Econometrica 50(1), 1–26

[19] cf. Gourieroux, C., Monfort, A. Trognon, A. (1984b), *Pseudo maximum likelihood methods: Theory*, Econometrica 52(3), 681–700.

[20] cf. VUONG, Q. H. 1989. *Likelihood ratio tests for model selection and non-nested hypotheses.* Econometrica 57:307–333.

meaning that F_θ and G_γ are equivalent, against

$$H_f: \qquad E^0[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}] > 0 \tag{3.7}$$

meaning that F_θ is better than G_γ

$$H_g: \qquad E^0[log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}] < 0 \tag{3.8}$$

meaning that F_θ is worse than G_γ.

With the basic statistic assumptions in VUONG, Q. H. 1989[21], the Vuong non-nested test is used to test those three hypotheses:

$$\begin{aligned}
(i)under H_0: &\qquad \frac{1}{\sqrt{n}}LR_n(\hat{\theta_n},\hat{\gamma_n})/\hat{\omega_n} \xrightarrow{D} N(0,1)\\
(ii)under H_f: &\qquad \frac{1}{\sqrt{n}}LR_n(\hat{\theta_n},\hat{\gamma_n})/\hat{\omega_n} \xrightarrow{D} +\infty\\
(iii)under H_g: &\qquad \frac{1}{\sqrt{n}}LR_n(\hat{\theta_n},\hat{\gamma_n})/\hat{\omega_n} \xrightarrow{D} -\infty\\
(iv)&\text{properties (i)-(iii) hold if } \hat{\omega_n} \text{is replaced by } \tilde{\omega_n}
\end{aligned}$$

$$where \qquad LR_n(\hat{\theta_n},\hat{\gamma_n}) = \sum_{t=1}^{n} log\frac{f(Y_t|Z_t;\hat{\theta_n})}{f(Y_t|Z_t;\hat{\gamma_n})} \tag{3.9}$$

$$\hat{\omega_n} = \frac{1}{n}\sum_{t=1}^{n}[log\frac{f(Y_t|Z_t;\hat{\theta_n})}{f(Y_t|Z_t;\hat{\gamma_n})}]^2 - \frac{1}{n}[\sum_{t=1}^{n} log\frac{f(Y_t|Z_t;\hat{\theta_n})}{f(Y_t|Z_t;\hat{\gamma_n})}]^2$$

$$\tilde{\omega_n} = \frac{1}{n}\sum_{t=1}^{n}[log\frac{f(Y_t|Z_t;\hat{\theta_n})}{f(Y_t|Z_t;\hat{\gamma_n})}]^2 = \hat{\omega_n} + (\frac{1}{n}LR_n(\hat{\theta_n},\hat{\gamma_n}))^2$$

When the sample size is big enough, the distribution of likelihood-ratio statistic converges to the normal distribution. Hence, the null hypothesis is rejected when the statistic is far away from zero. As the likelihood function is hard to be presented in nonparametric binary models, the Vuong's non-nested test is applied to the models of discrete approach. The result is presented in Table 10.

$$\begin{cases} ZIP: \text{Zero-inflated Poisson model} \\ ZINB: \text{Zero-inflated Negative Binomial Model} \\ ZIG: \text{Zero-inflated Geometric Model} \end{cases}$$

[21] cf. VUONG, Q. H. 1989. *Likelihood ratio tests for model selection and non-nested hypotheses.* Econometrica 57:307–333.

Model1	Model2	Vuong z-statistic	p-value	Conclusion
ZIP	ZINB	-52.13424	2.22e-16	ZINB > ZIP
ZIP	ZIG	-52.27174	2.22e-16	ZIG > ZIP
ZINB	ZIG	21.16899	2.22e-16	ZINB > ZIG

Table 13: Vuong's Non-nested Test for Discrete Models

The results of three tests reflect that the ZINB model is the best among the discrete models. In the first two tests, the statistics are negative and the corresponding p-values are relatively small. As a consequence, the hypothesis Hg is accepted for the two tests, meaning that ZIP model is worse than the other two models. The third test compares ZINB model and ZIG model. With positive statistic and small p-value, the hypothesis Hg is accepted. According to Vuong's non-nested test, ZINB model performs better than ZIG model.

3.3.2 Hosmer-Lemeshow Test

Hosmer-Lemeshow test[22] is a goodness of fit test for binary models such as logistic regression model. The null hypothesis H_0 and alternative hypothesis H_A of Hosmer-Lemeshow Test are defined as:

$$\begin{cases} H_0: & \text{the current model fits well} \\ H_A: & \text{the current model does not fit well} \end{cases} \tag{3.10}$$

The result of the Hosmer-Lemeshow test is generated from the observed event rates and estimate event rates of subgroup samples. If the observations are divided into m subgroups by the model predicted probabilities, the Hosmer-Lemeshow statistic is calculated by:

$$G^2_{HL} = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i(1 - E_i/n_j)} \sim \chi^2_{m-2}$$

where

$$\begin{cases} \mathrm{O}_i \text{ is the sum of observations in the } i^{th} \text{ group.} \\ \mathrm{E}_i \text{ is the sum of estimations in the } i^{th} \text{ group.} \\ \mathrm{n}_j \text{ is number of observations in the } i^{th} \text{ group.} \\ \chi^2_{m-2} \text{ is a Chi-square distribution with m-2 degrees of freedom.} \end{cases}$$

The p-value of this test equals to the probability $P(\chi^2_{m-2} > G^2_{HL})$. If the p-value is smaller than the significance level, the null hypothesis is rejected.

[22] cf. Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression.*

Model	X-squared	Df	p-value
GLM	138.48	98	0.004467
GAM	150.9	98	0.0004816
CART	429.65	98	<2.2e-16
Random Forest	831.92	98	<2.2e-16
Bernouli GBM	268.99	98	<2.2e-16
Adaboost GBM	275.39	98	<2.2e-16

Table 14: Hosmer-Lemeshow Test Result

Table 14 presents the Hosmer-Lemeshow test result of the six binary approach models. The subgroup number is set as 100 in the test to make the result more reliable. In the result, the six p-values are all below 0.5%, meaning that given any significance level larger than 0.5%, the Hosmer-Lemeshow test can not reject the null hypothesis: the current model fits well. In other words, by the results of 100 subgroups Hosmer-Lemeshow test, it can not be concluded that any of the six models does not fit the test data well. In the aspect of probability, for one of the six binary models, there is less than 0.5% chance that it does not fit the database.

3.3.3 AUC Criterion and MSE Criterion

Confusion matrix, or error matrix, presents the prediction performance by comparing observed value and prediction value. As presented in Figure 39, when one observation is true, this observation is counted as true positive if it is predicted as true, and this observation is counted as false negative if it is predicted as false. When one observation is false, if it is predicted as false, this observation is counted as true negative; if it is predicted as true, this observation is counted false positive.

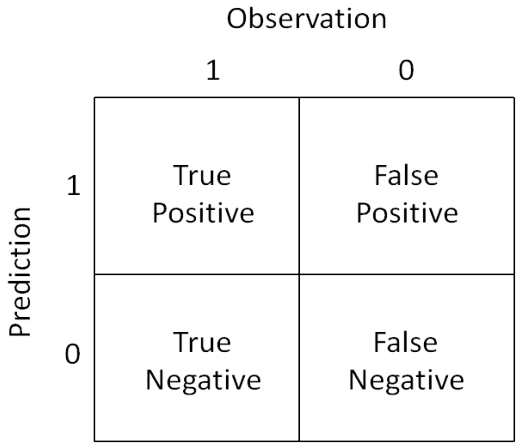

Figure 39: Confusion Matrix

The accuracy rate shows the prediction accuracy. By the confusion matrix, it is defined as:

$$\text{ACC} = \frac{\text{True Positive Number+ False Negative Number}}{\text{Number of Observations}}$$

The Receiver Operating Characteristics (ROC) curve presents the relationship between True Positive rate and False Positive rate. The True Positive rate (TPR) means the right prediction rate among true observations, and the False Positive rate (FPR) means the wrong prediction rate among false observations. The True Positive rate (TPR) and False Positive rate (FPR) are defined as[23]:

$$\text{True Positive rate } = \frac{\text{Number of True Positive}}{\text{Number of True Observations}}$$

$$\text{False Positive rate } = \frac{\text{Number of False Positive}}{\text{Number of False Observations}}$$

In this case, when all predictions are false, the True Positive rate and False Positive rate all equal zero. And when all predictions are true, both the True Positive rate and the False Positive rate equal one. Therefore, the ROC curve starts from (0,0) and stops at (1,1).

Given a series of Bernoulli observations y_i, a series of estimated probabilities x_i, and a fixed prediction threshold s, the predictions z_i are generated by the rule:

$$z_i = \mathbb{1}_{x_i > s}$$

Correspondingly, the True Positive rate and False Positive are generated by the formula:

$$TPR = \frac{\sum \mathbb{1}_{z_i=1, y_i=1}}{\sum \mathbb{1}_{y_i=1}}, \qquad FPR = \frac{\sum \mathbb{1}_{z_i=1, y_i=0}}{\sum \mathbb{1}_{y_i=0}}$$

Then by choosing different prediction thresholds s between zero and one, a series of True Positive rate and False Positive rate is generated in $[0,1]^2$. This series traces a curve in map $[0,1]^2$, which derives from (0,0) and arrives at (1,1). The AUC (Area Under the Curve) is defined as the area under the ROC curve. Obviously the AUC value is in the [0,1] interval. A point estimate of the AUC of the empirical ROC curve is introduced as the Mann-Whitney U estimator (DeLong et. al., 1988). And the confidence interval for AUC indicates the uncertainty of the estimate and uses the Wald Z large sample normal approximation (DeLong et al., 1998). As the sample size is large enough, the variance of AUC is relatively small and the point estimate of AUC is used.

The AUC value could be interpreted as:

- The probability that a randomly selected subject with the condition has a test result indicating greater suspicion than that of a randomly chosen subject without the condition[24].

[23] cf. T. Fawcett. 2006. *An introduction to ROC analysis.* Pattern Recognition Letters

[24] cf. Hanley, J. A., McNeil, B. J. 1982. *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology

If one ROC curve is at the top of another curve, the y-axe value (True Positive rate) of the first curve is larger in the same x-value (False Positive rate). This means that the prediction performance of the first curve is better at any False Positive rate. Therefore, the prediction performance of the first curve is better. To quantitatively compare the prediction performance in the general case, the AUC value is utilized. As all possible prediction situations are included in the AUC calculation, in general, a higher AUC value means a better prediction performance for the binary response variable.

When the predictions of six binary approach models are generated, the corresponding True Positive rate and False Positive rate are calculated. The corresponding ROC curves are presented in Appendix C.

Mean Square Error (MSE) is often used as the criterion of the estimation quality. A smaller MSE value means a better estimation. As the observation is binary while the estimation is the probability, the MSE can not be zero and a very small MSE value may be caused by the over-fitting.

Model	AUC	MSE
GLM	0.681614	0.1082247
GAM	0.6851962	0.1078653
CART	**0.7053183**	**0.1043134**
Random Forest	0.6890743	0.1079522
Bernoulli GBM	0.6839162	0.1080703
Adaboost GBM	0.6841997	0.1081064

Table 15: AUC and MSE of Binary Approach Models

Table 15 presents the AUC and MSE values of the binary approach models. Among the six models, the prediction performance of CART model is better according to AUC criterion and MSE criterion. The random forest model performs better than model CART. As node number of Random Forest is limited (set as 100) because of the limited calculation capacity, the performance of random forest model could be better. Considering that the number of observations in the training set is large enough, the randomness of CART model is relatively small. Compared to other models, the CART model is easier to understand, and the CART model can also be adapted to the data with missing values (cf 3.1.5 CART: Classification And Regression Tree). Thus, with the highest AUC value and lowest MSE value, the CART model is the best one among the six binary approach models.

Model	AUC	MSE
ZIP	0.6761196	0.1408480
ZINB	0.6776365	0.1085469
ZIG	0.6779710	0.1093668

Table 16: AUC and MSE of Discrete Approach Models

Table 16 presents the AUC and MSE values of the three discrete models. Among the three models, ZINB and ZIG are better than ZIP according to the criteria of AUC and MSE. Compared to the model ZIG, the model ZINB has smaller AUC value and smaller MSE value, meaning that the ZINB is better in the aspect of MSE and is worse in the aspect of AUC. And both the differences of AUC value and MSE value are small: 0.003 and 0.008. Thus, according

to the criteria of AUC and MSE, the ZINB model and ZIG model have similarly performances. Considering the result of Vuong's non-nested test, the model ZINB is selected as the best model of discrete approach.

As the nine models have been well built and evaluated, the probability estimations of each model could also be generated. In this case, the pure premium estimations of each model could be calculated by applying the estimated delay probabilities. Then by comparing the pure premium estimations and the real data of flight delays, the precision of the pure premium estimation could be evaluated. Then the best model could be selected and the corresponding premiums could also be generated.

Chapter 4

Evaluation

This part includes the applications and the evaluations of different mathematical models. The first point is the pricing application of the mathematical models. The technical premium is composed of the pure premium and pricing factors. The pure premium could be estimated by the trained mathematical models, and the pricing factors are decided by economic environment and corporate strategies. The second point is the model performance in different portfolio scenarios. The risk level of the pure premium estimation could be estimated by analysing scenarios of different simulated portfolios. The third point is the adaptability of models. By using different delay thresholds and different databases, the adaptability of the models could be evaluated. The fourth point is the approach comparison, which allows knowing the advantage and disadvantage of the two approaches.

4.1 Pricing

4.1.1 Pure Premium Calculation

The pure premium is used to evaluate the cost of risk. As introduced in Section 1.5 (cf. 1.5 Premium Calculation), the pure premium equals the product of accident frequency and severity.

$$PP = E[I * F];$$

PP = Pure premium
I = Severity
F = Accident Frequency
$E[\ ]$ = Expectation

The severity[1] is estimated by the cost of airport lounge. Referring to the price of a provider which has more than one thousand lounges worldwide, the cost of lounge is estimated at 27$. In this case, the severity equals 27$ and the indemnity for the departure delay is fixed at 27$.

The accident frequency is the departure delay probability, which can be estimated by the introduced mathematical models. Then the pure premium is generated by the estimated accident frequency and the severity.

[1] Severity refers to the cost of claim

4.1.2 Model Application

The pure premium calculation is applied to the flights in the first quarter of 2017. There are 22 375 flights which take off from New York John F. Kennedy International Airport in the first quarter of 2017, including 7 565 in January, 6 884 in February, and 7 926 in March.

As the historical data of New York John F. Kennedy International Airport is available, the mathematical models were trained by the eight years historical data. For each flight in the first quarter of 2017, the pure premium is estimated by the frequency predictions of nine models. The average accident cost, or average severity, equals the product of the average flight delay frequency and indemnity. Hence, compared to the average accident cost of the first quarter in 2017, the pure premium estimations of different models can be evaluated. If the average pure premium estimation is close to the average accident cost, the estimation is precise in aggregate. And the model with smallest positive deviation is the best model to calculate the pure premium.

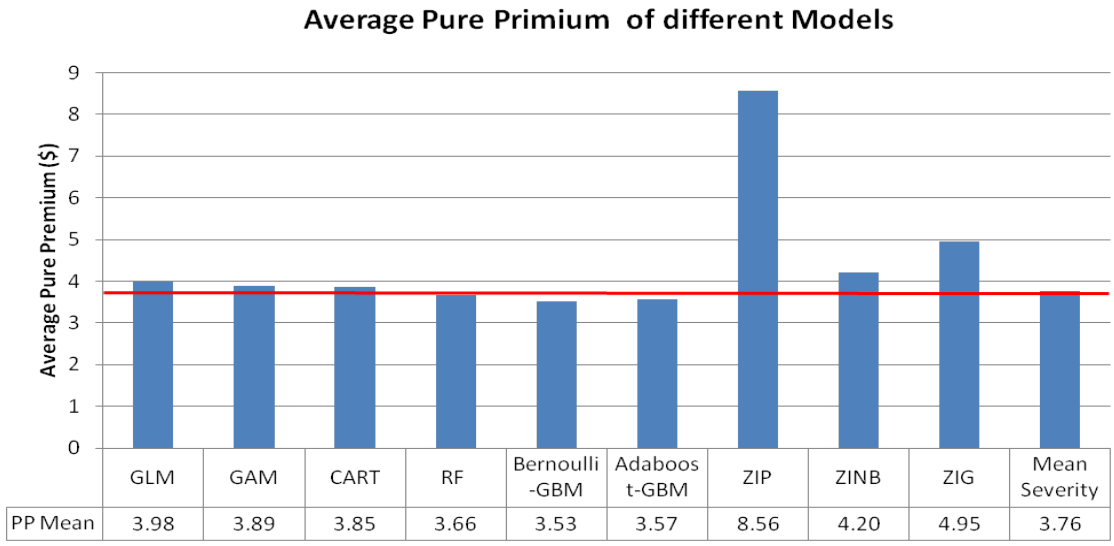

Figure 40: Average Pure Premiums and Average Accident Cost

Figure 40 presents the average pure premiums estimated by the nine models and the average severity. The average pure premium estimations of most models are close to the average severity, except the ZIP model and ZIG model. Among the pure premium estimations, the average estimations of random forest model and GBM models are lower than the average severity, meaning that the estimations are underestimated in aggregate.

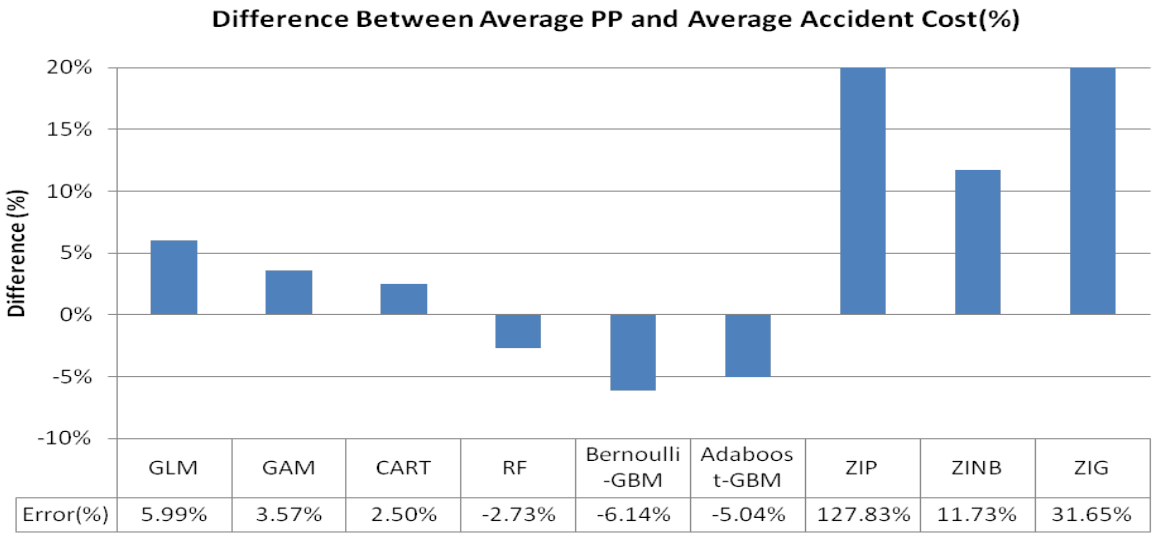

Figure 41: Difference Between Pure Premium Average and Accident Cost Average

Figure 41 presents the difference between average pure premium estimations and average accident cost by the percentage of average accident cost. The CART model has the closest average pure premium estimation without underestimation. And the model ZINB has the closest one among the discrete approach models. Therefore, the CART model and ZINB model are selected as the representations of discrete approach and binary approach for further analyses.

4.1.3 Technical Premium

The technical premium serves as the price of insurance product, therefore, it is calculated by two steps. The first step is to calculate the pure premium, which reflects the risk level of the coverage. The second step is to decide other pricing factors such as safety loading to calculate the technical premium. And the quality of technical premium estimation largely relies on the precision of pure premium estimation. The formula 4.1 presents the equation to calculate the technical premium.

$$TP = PP * \frac{(1 + S_L)}{(1 - C - K)}; \qquad \begin{aligned} TP &= \text{Technical Premium} \\ PP &= \text{Pure premium} \\ S_L &= \text{Safety loading} \\ C &= \text{Commission} \\ K &= \text{Cost of capital} \end{aligned} \tag{4.1}$$

The pure premiums of CART model and ZINB model are used to calculate the technical premiums of binary approach and discrete approach. The cost of capital, safety loading, and commission are set as x%, y% and x% in the calculation. And the distributions of the technical premiums are presented in the following figures.

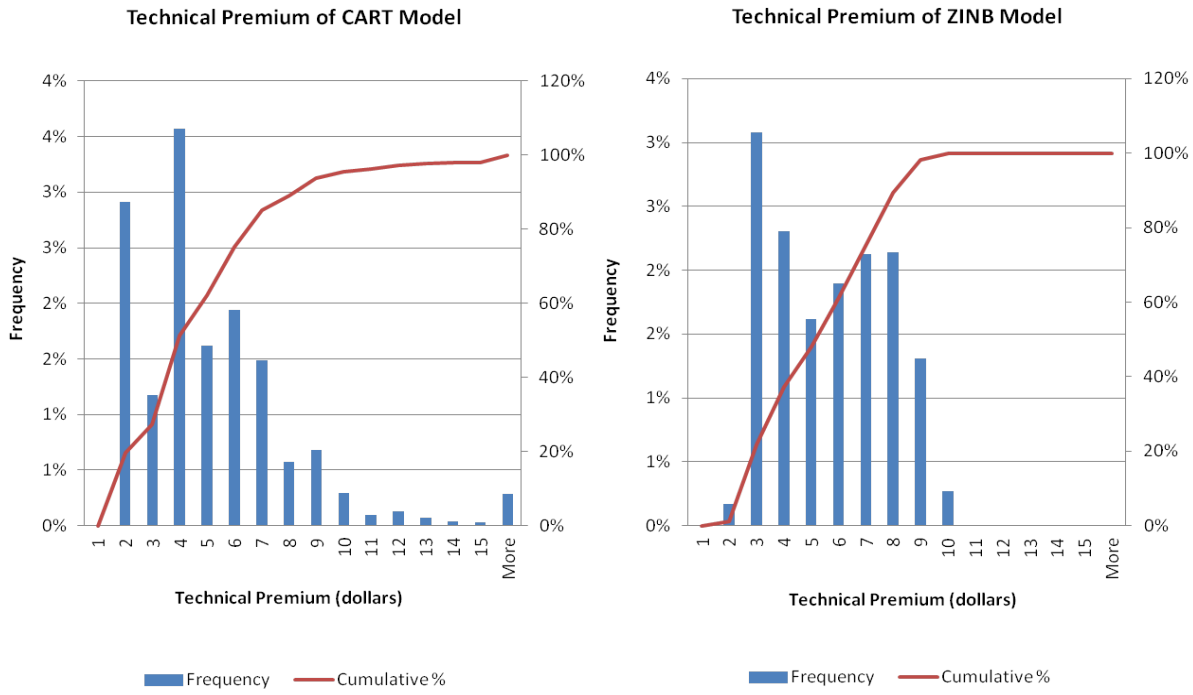

Figure 42: Technical Premium Distributions

The distribution of technical premium of CART model is more concentrated and more stable than the distribution of ZINB model. The technical premiums of both models are all larger than 1$. The first column shows the frequency of technical premium between one dollar and two dollars. There are more than 5% technical premiums of CART model is in this interval while there are less than 1% technical premiums of ZINB model in the interval. A small part of technical premium of CART model exceeds 15$ while no technical premium of ZINB model is more than 15$. To sum up, the technical premiums of CART model are more diversified and the technical premiums of ZINB model are more concentrated and are mostly in the interval [2,10].

Technical Premium of CART Model		*Technical Premium of ZINB Model*	
Mean	4.758	Mean	5.187
Median	3.846	Median	5.173
Standard Deviation	**3.569**	Standard Deviation	**2.099**
Kurtosis	17.289	Kurtosis	-1.246
Skewness	3.333	Skewness	0.149
Minimum	1.15	Minimum	1.418
Maximum	33.353	Maximum	10.045
Sum	106454	Sum	116049
Count	22375	Count	22375

Table 17: Technical Premium Statistic

The Kurtosis statistic of ZINB model premium equals -1.246, which is close to the value of uniform distribution. And the Skewness statistic of ZINB model is only 0.149, which is close to zero. It means that the distribution is almost symmetric. For the CART model, the Kurtosis statistic and the Skewness statistic are 17.289 and 3.333. The large Kurtosis statistic

indicates that there is an outlier problem. And the large Skewness statistic indicates that the distribution is asymmetric. In addition, the standard deviation of CART model is smaller than the standard deviation of ZINB model. Thus, the statistics in the table 17 also indicate that the distribution of ZINB model is more concentrated than the distrubution of CART model.

4.2 Scenario Test

4.2.1 Simulation

The result of the model application part presents the pure premium estimations of the nine models. However, the policy number of each flight is not always the same. To simulate the "true" situation, the simulated portfolios could be generated by the bootstrap method. In this part, the simulated sample of size 50 000 is drawn by bootstrap method from the monthly data of the first quarter 2017. The aggregation of those three samples is the simulated portfolio of the first quarter 2017. In the simulated portfolio, one flight can have several policies. As the delay frequency is unique for each flight, the technical premiums of the same flight are the same.

The mathematical models are trained by the last eight years data. The CART model is used for the discrete approach and the ZINB model is used for the discrete approach. The pure premiums are generated by applying the two models to the simulated portfolios.

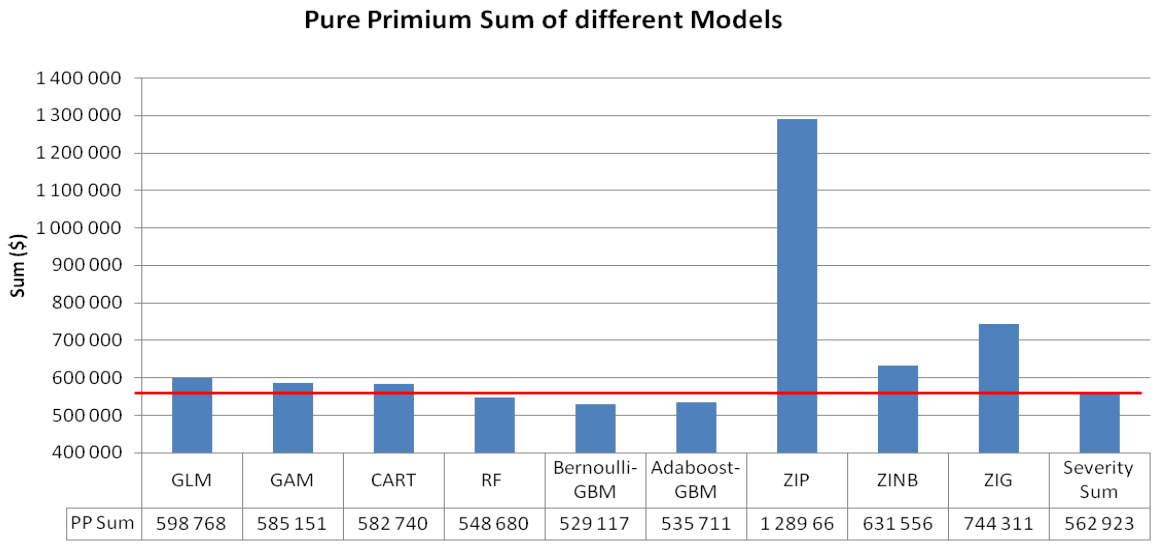

Figure 43: Pure Premium Sums of Different Models

Figure 43 presents the technical premium sums of different models. The technical premium estimations of ZIP model and ZIG model show significant overestimations. The estimations of other models are close to the total severity. Similar as the previous analysis, the pure premium sum is compared with the severity sum to evaluate the precision of pure premium estimation.

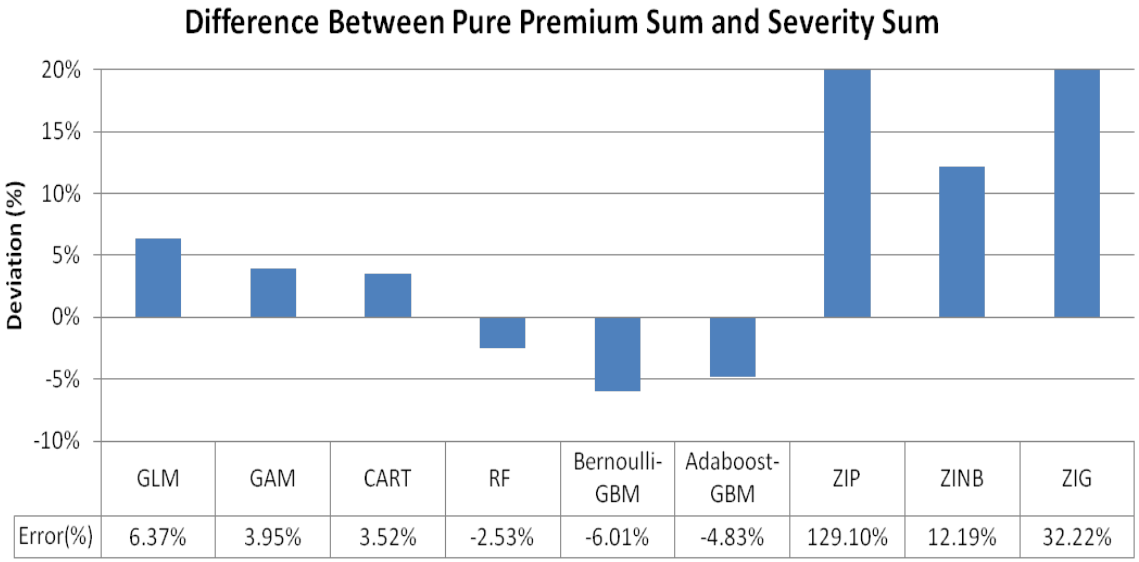

Figure 44: Difference Between Pure Premium Sum and Severity Sum

For the simulated portfolio, the performances of pure premiums of different models are similar to the conclusion of the previous analysis. CART model and ZINB model have separately the lowest deviations without underestimation among the models of binary approach and the models of discrete approach.

4.2.2 Evaluation

The two models are evaluated in aggregate. To better understand the precision of the pure premium estimation, the aggregate performances of subgroups by different variables are studied. As the portfolio is generated from data of the first quarter in 2017, the year variable has only one value and the month variable has only three modalities. Thus, the two variables are not used in the analysis. Then day of the week variable, airline variable, vacation variable, scheduled departure time variable, scheduled arrival time variable, and distance variable are used to divide the portfolio into subgroups. The precision of aggregate pure premium estimations of subgroups means that the predictive model performs well. As the severity is fixed at 27$, the precision of frequency estimation equals the precision of pure premium estimation, and is used to evaluate the mathematical models.

4.2.2.1 Day of the Week

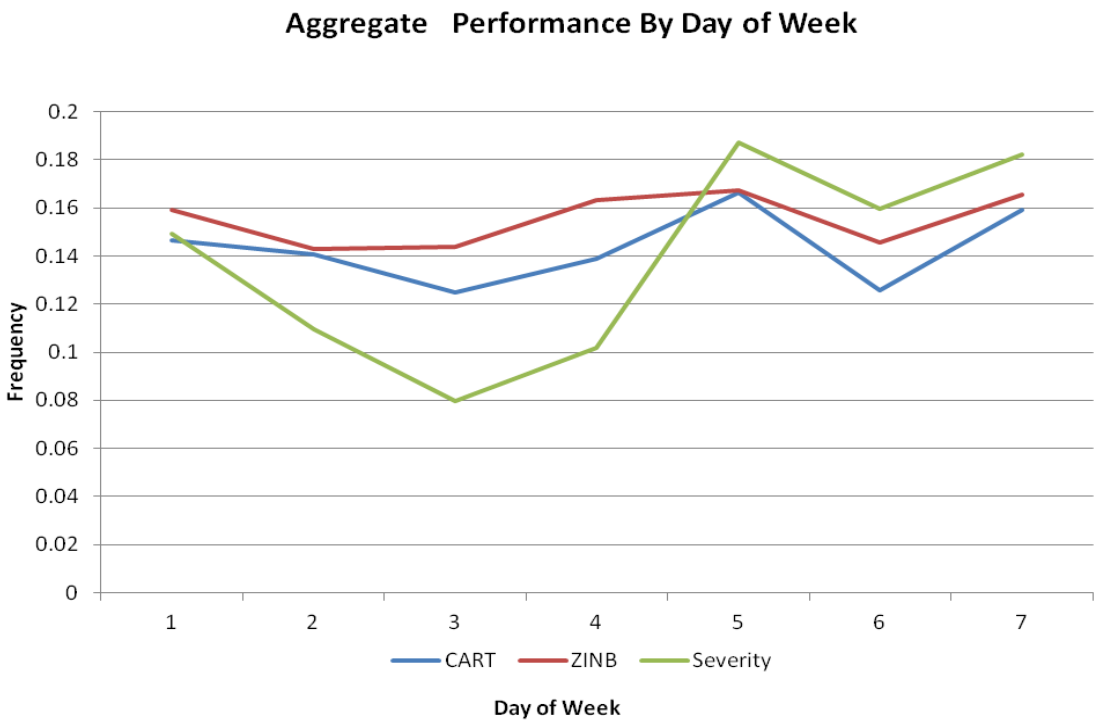

Figure 45: Aggregate Performance by Day of the Week Variable

This figure shows the predictive performance by day of the week variable. The true delay cycle of day of the week variable oscillates more than the cycle of model estimations. This means that the oscillation of day of the week variable in the first quarter of 2017 is larger than that of passing years. And the CART model performs better than the model ZINB because of the larger oscillations at the modalities 3 and 6 of the day of the week variable. As the day of the week variable is qualitative and the cycle is irregular, this difference is hard to be included by the cross term of year variable and day of the week variable. In contrast, the recent data can be used in model training to prevent the large cycle change of day of the week variable.

4.2.2.2 Scheduled Arrival Time

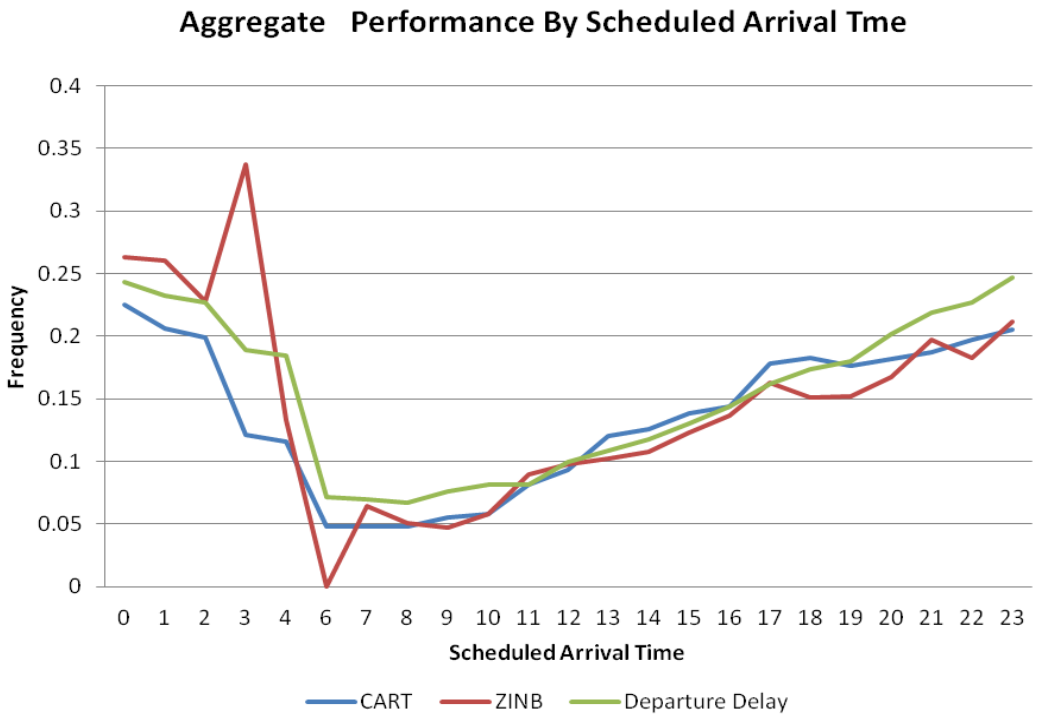

Figure 46: Aggregate Performance of Scheduled Arrival Time Variable

Figure 46 presents the aggregate predictive performance by scheduled arrival time variable. The three curves in Figure 46 are of the same cycle. And the two prediction curves are very close to the delay frequency curve. There are decreasing trend at the period of midnight to six o'clock and increasing trend at the period of six o'clock to midnight. This means that this cycle is well integrated in both models, and is also validated for the data of 2017.

4.2.2.3 Scheduled Departure Time

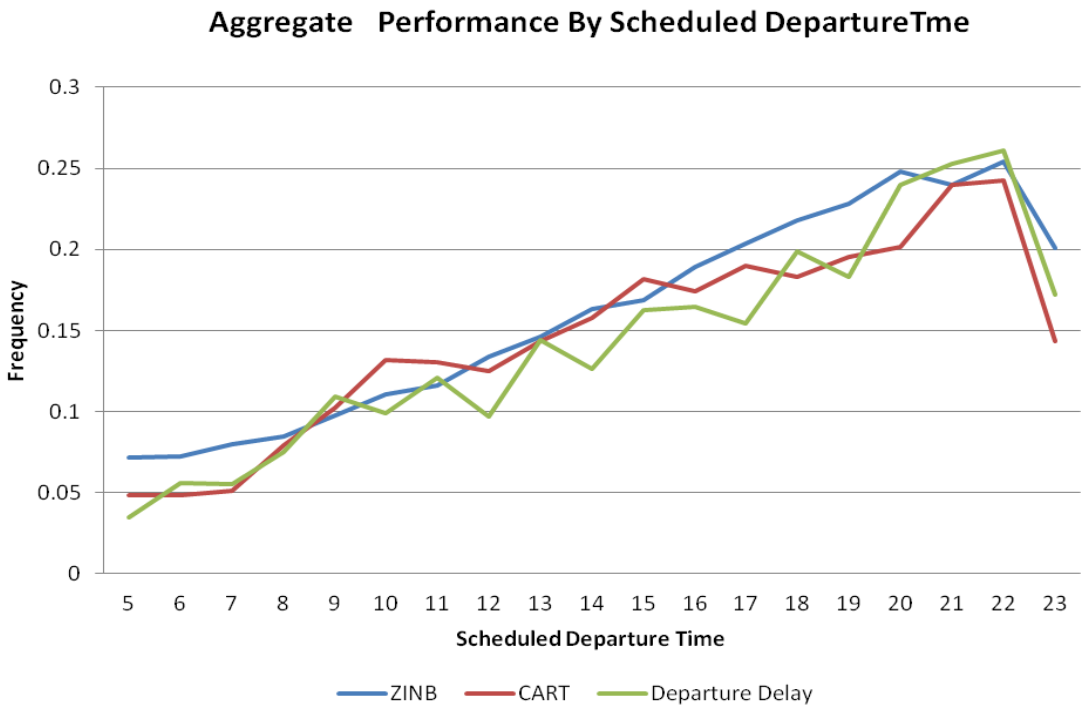

Figure 47: Aggregate Performance of Scheduled Departure Time Variable

Figure 47 presents the aggregate predictive performance by scheduled departure time variable. Similar as scheduled arrival time variable, the influence of this variable is well validated and integrated by both models.

4.2.2.4 Airline

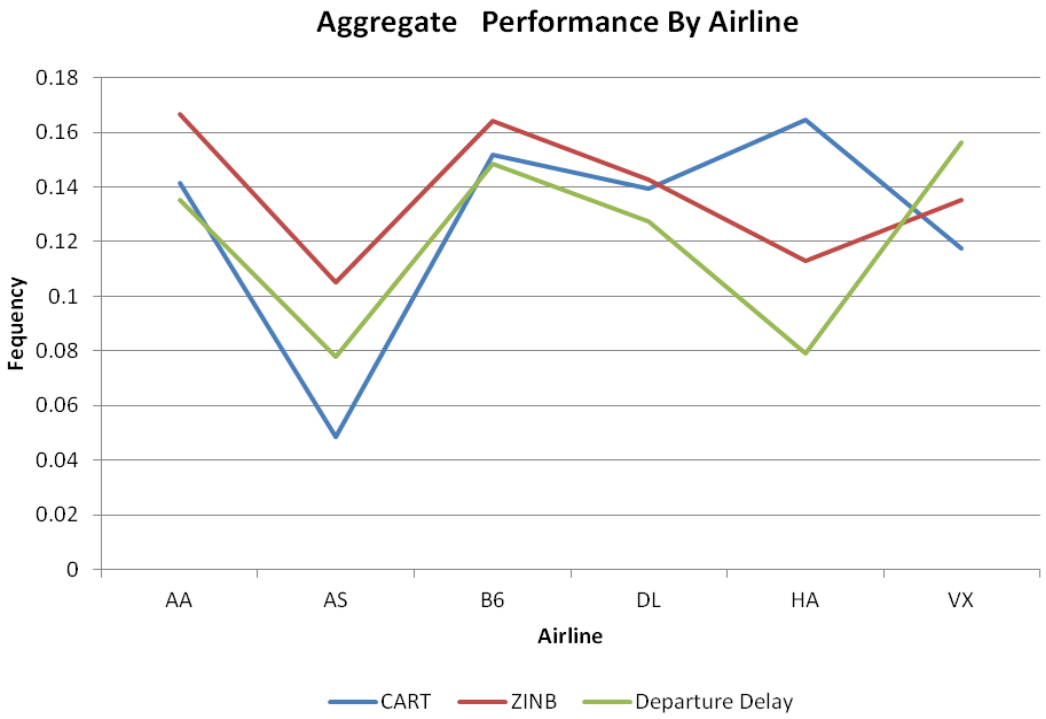

Figure 48: Aggregate Performance of Airline Variable

Figure 48 presents the aggregate predictive performance by airline variable. In general, the two models have predicted the frequency of the airline variable well. Considering that the curve of ZINB model has the same shape of delay frequency curve and the differences are almost the same, the ZINB model has well integrated the cycle of this variable, but this model also presents overestimation. In contrast, the CART model predicts well in general but has large prediction error for the modality "HA", which represents the Hawaiian Airline.

4.2.2.5 Distance

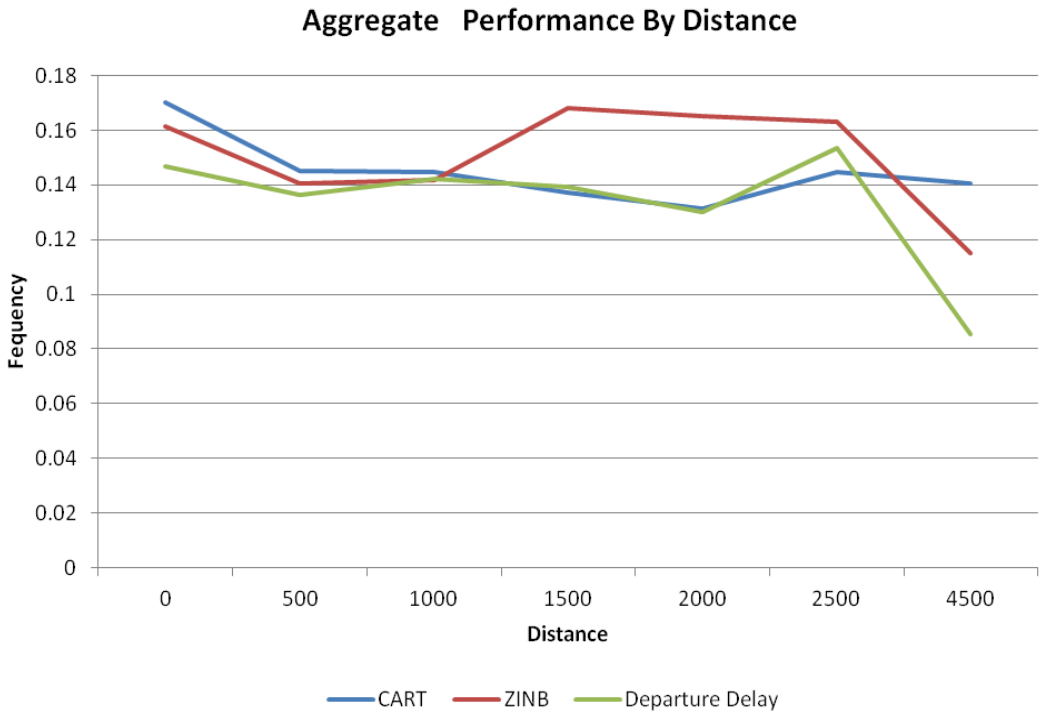

Figure 49: Aggregate Performance of Distance Variable

This figure shows the aggregate predictive performance by distance variable. As the distance variable is continue, the subgroups are divided by interval of 500, such as 0-500km, 500-1000km etc. The prediction of CART model is very precise. The curve of CART model is approximate to the curve of true delay frequency. While the ZINB model has overestimated the frequency when the distance is more than 1500km. The true delay frequency cycle of vacation variable is very irregular, thus this figure is presented in Appendix E and is not introduced there.

4.2.3 Risk Analysis

As presented in the previous part, the delay frequencies of subgroups are well estimated in general. In the section 4.2.1, the CART model has overestimation of 3.52% in aggregate while the ZINB model has overestimation of 12.19% in aggregate. As the simulated portfolio is randomly selected from the data of the first quarter 2017, the estimation errors are also randomly generated. To study the estimation errors in the general case, one thousand samples are drawn randomly as the "true" scenarios by bootstrap method. Then the difference between the pure premium sum and severity sum is calculated as the aggregate estimation error and is standardized by the percentage of servery sum.

CART Simulated Error (%)		ZINB Simulated Error (%)	
Mean	2.516	Mean	11.755
Median	2.497	Median	11.727
Standard Deviation	0.666	Standard Deviation	0.709
Kurtosis	0.118	Kurtosis	0.124
Skewness	0.037	Skewness	0.023
Minimum	0.618	Minimum	9.719
Maximum	4.929	Maximum	14.306
Count	1000	Count	1000

Table 18: Error Statistics of 1000 Simulated Portfolios

Table 18 shows the error statistics of one thousand simulated portfolios. The CART model has an average aggregate error of 2.516% and the ZINB model has an average aggregate error of 11.755%. And both the standard deviations of aggregate errors are small, separately 0.666 and 0.709. The two small standard deviations means that the aggregate error is relatively stable. The value of Kurtosis and Skewness are close to 0, which means that the distribution of aggregate error is close to the normal distribution.

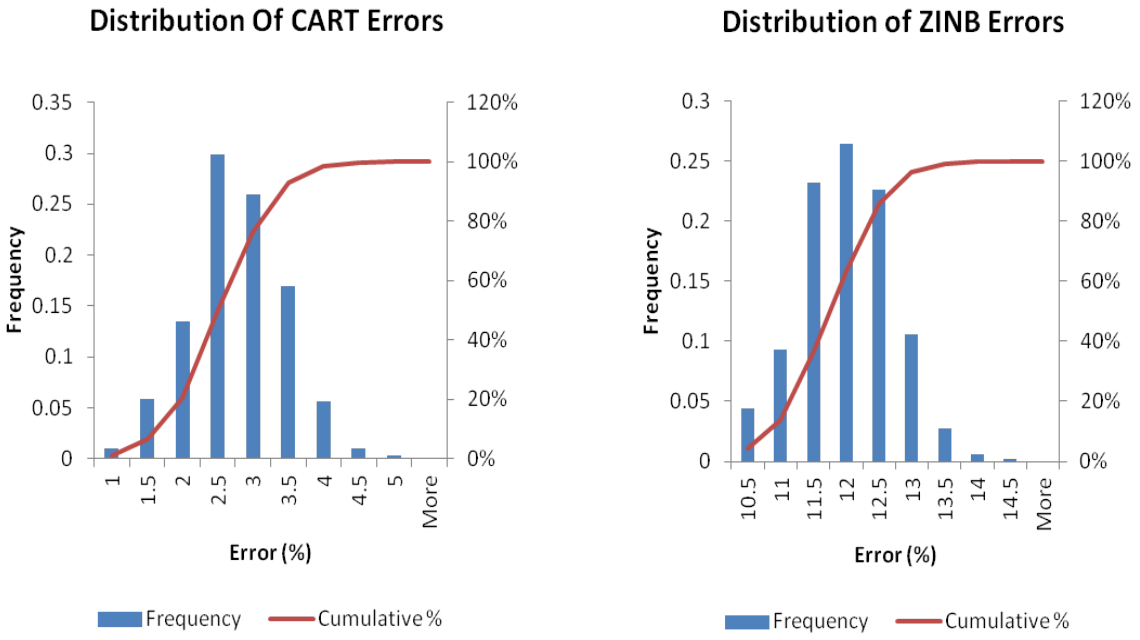

Figure 50: Error Distributions of 1000 Simulated Portfolios

Figure 50 presents the distributions of the aggregate estimation errors of one thousand simulated portfolios. Consistent with the statistics of the aggregate errors, the distributions of the aggregate errors are well symmetric and concentrated, and are mainly situated in the intervals ± 2% of the mean aggregate error.

4.3 Adaptability of Model

The CART model and ZINB model are validated for the eight years historical data of New York John F. Kennedy International Airport. But the two models are only validated for the delay threshold of thirty minutes. Thus, to study the adaptability of the two models, the cases of different delay thresholds, different available data, or different airports are generated and analysed.

4.3.1 Delay Threshold

The departure delay used in the previous analysis is defined as the delay that is longer than thirty minutes. The modelling and analysis are well validated for data of New York John F. Kennedy International Airport. However, the delay threshold of thirty minutes could be too short for the passenger who is very patient. And the delay threshold of one hour or longer may be more appropriate in this case.

With respect to the extended products of different delay thresholds, CART model and ZINB model are tested with different delay thresholds, separately 60 minutes, 90 minutes, 120 minutes and 180 minutes. The eight years historical data is used as training data set. And the first quarter data of 2017 are used as the test data set to evaluate the estimation performance of the two models.

The analysis is mainly composed of two parts. The first part is the comparison of prediction performances. The prediction performances are compared by the AUC criterion and MSE criterion. According to the result of comparison, the ZINB model has better prediction performances for any delay threshold. The second part is the aggregate deviation analysis. By comparing the pure premium averages of the two models with the average severity, the precision of the two models can be evaluated. According to the results of different delay thresholds, the estimation of CART model is more precise. In contrast, the ZINB model presents an overestimation at low delay threshold, but presents an underestimation at high delay threshold. In consequence, the CART model is relatively more stable than the ZINB model.

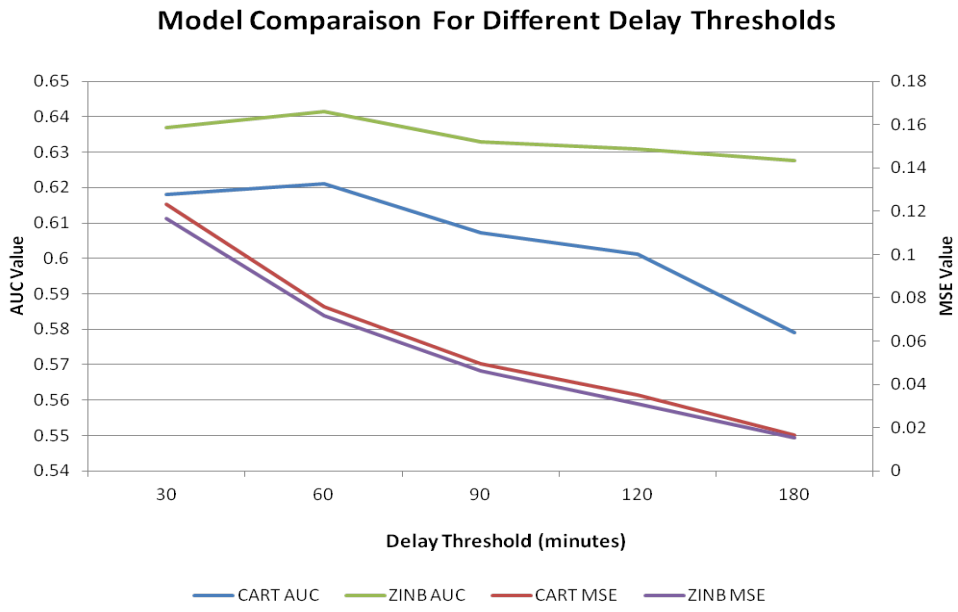

Figure 51: Predict Performances By Different Delay Thresholds

In Figure 51, the predictive performances are compared by the AUC criterion and by the MSE criterion. The AUC curve of CART model is all under the AUC curve of ZINB model. Thus, the AUC value of CART model is always smaller than the AUC value of ZINB model. It's concluded that ZINB model performs better than CART model by the criterion AUC. Similarly, the MSE curve of CART model is always above the MSE curve of ZINB. It means that ZINB model has smaller mean square errors, and therefore is better than CART model by the criterion MSE. In total, the predictive performances of ZINB model are better than CART model in any delay threshold by the criteria of AUC and MSE.

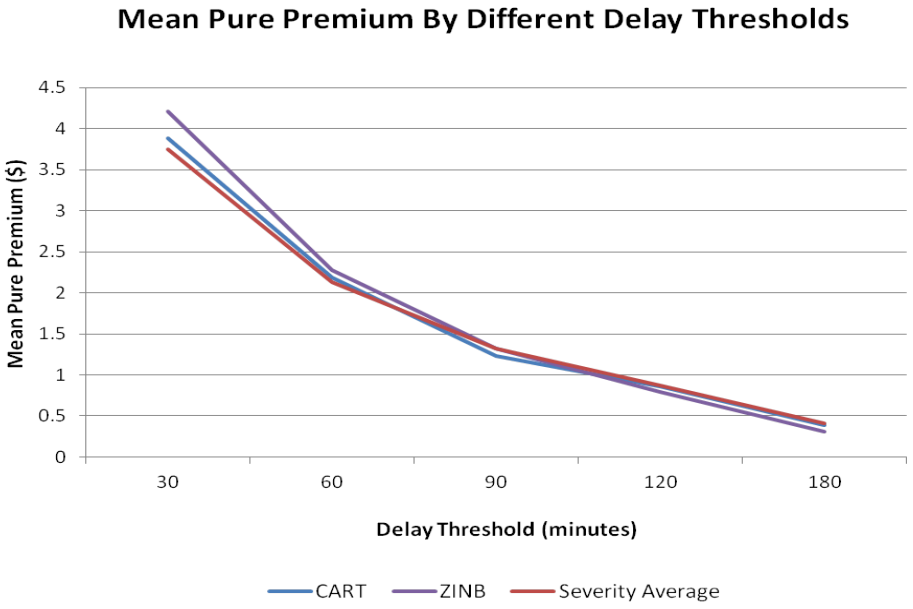

Figure 52: Mean Pure Premium By Different Delay Thresholds

The three curves in figure 52 are separately mean pure premium curve of CART model, mean pure premium curve of ZINB model, and the mean severity curve. The three curves are close to each other, meaning that the pure premiums are well estimated by CART model and ZINB model.

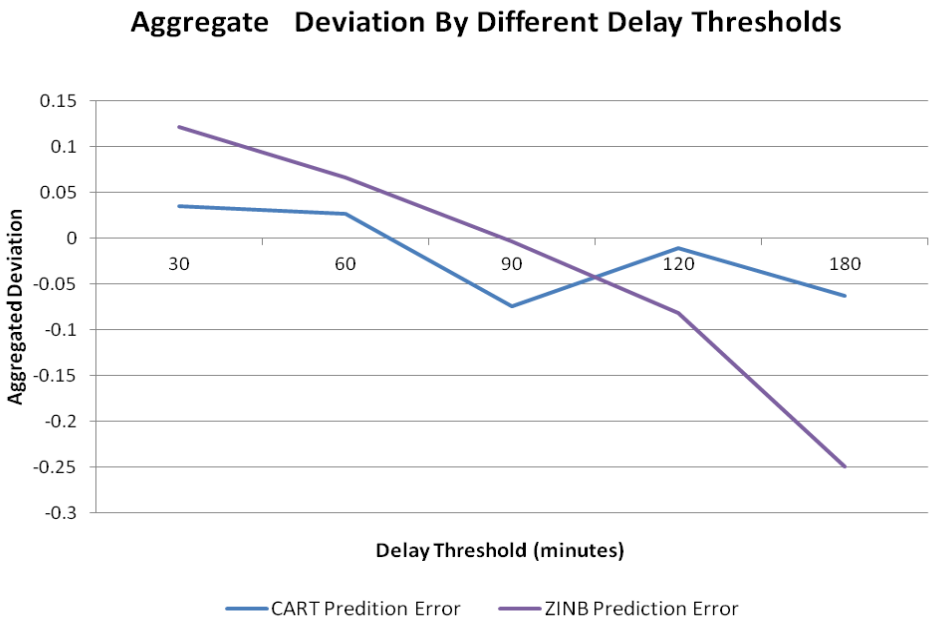

Figure 53: Aggregate Deviation By Different Delay Thresholds

Figure 53 shows the deviations of the mean pure premium, which is in percentage of the average severity. According to this figure, the deviations of CART model are always between -5% and 5%. However, the deviation curve of ZINB model decreases from 12% to -8% as the delay threshold increases from 30 minutes to 120 minutes, and reaches -24% at 180 minutes. Thus the deviation of average pure premium of the ZINB model decreases as the delay threshold increases. This means that the ZINB model shows overestimation using low delay threshold and shows underestimation using high delay threshold.

Although the CART model is more stable than ZINB model, ZINB has an unique advantage: one ZINB model can be applied to any delay threshold while CART model is limited to the threshold used in the model. Hence, when the product with several possible thresholds is studied, the ZINB model is more proper than the CART model.

4.3.2 Data Limitation

The two representative models are validated by using the eight years historical data. Although the historical data is available for any airport in the database, the historical data of other airports are not guaranteed. Thus, it is necessary to test the model performance when the historical data is limited. The CART model and the ZINB model are used in this part to evaluate the estimation performances of the discrete approach and the binary approach.

The different cases of insufficient historical data are studied by using different data volume for the modelling. To include the 12 modalities of month variable, the duration of the data is

set to be more than twelve months. Thus, the historical data of one year, two years, and three years serve as the limited database to study the model adaptability. In this case, the value of year variable is not enough to be a regressor. As a result, when the limited database is used, the year variable is removed from the CART model and ZINB model.

Last One Year Data

The historical data of 2016 serves as the training data. The first quarter data of 2017 is used to evaluate the estimation performance of the generated models. Based on the analysis of different delay thresholds, the estimation is evaluated in three aspects: criterion comparison, mean pure premium estimation, aggregate deviation.

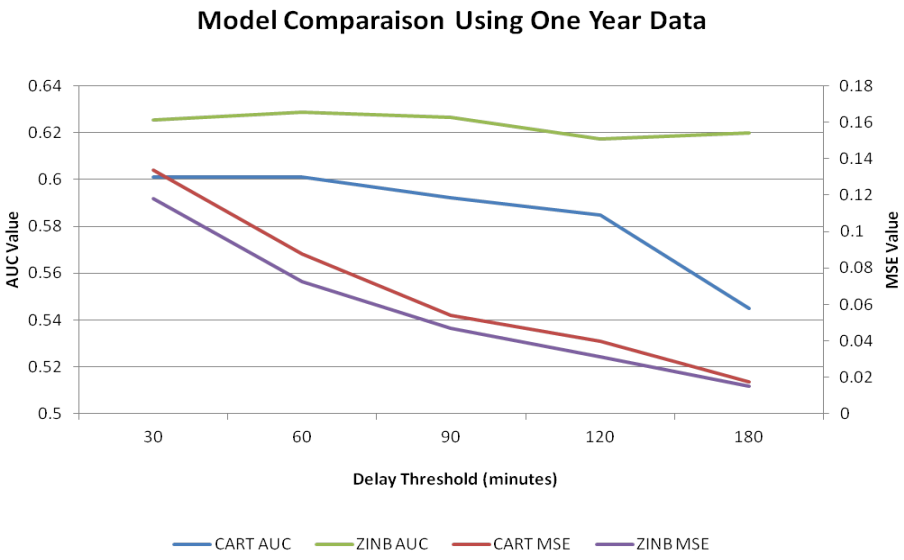

Figure 54: Adaptability: model performance using one year historical data

In Figure 54, the AUC curve of CART model is always below the AUC curve of ZINB model. The AUC values of the ZINB model are stable around 0.62 and the deviations do not exceed 0.01. However, the AUC value of the CART model changes largely and decreases as the delay threshold increases. The AUC value of the CART model even reaches 0.545 when the delay threshold is 180 minutes. The decreasing in AUC value means the decreasing estimation performance of the CART model. As the delay threshold increases, the delayed frequency decreases, then the delayed records decrease. As a result, the CART model could be trained incompletely, which results in the bad estimation performance. In this case, the estimation performance of large delay threshold is not as good as the performance of small delay threshold. For the ZINB model, as the delay variable is the same for different delay thresholds, the changes of threshold have small influences on the estimation performance. In conclusion, the estimation performance of CART model varies largely as the delay threshold differs. In contrast, the estimation performance of ZINB model is stable and consistent for all the delay thresholds.

The MSE curve of CART model is always above the MSE curve of ZINB model, meaning that the ZINB model has smaller MSE values and is more accurate. This result is consistent to the conclusion of AUC criterion. Therefore by the criterion of AUC or MSE, the CART model is not as good as ZINB model when the one year historical data is used as training data.

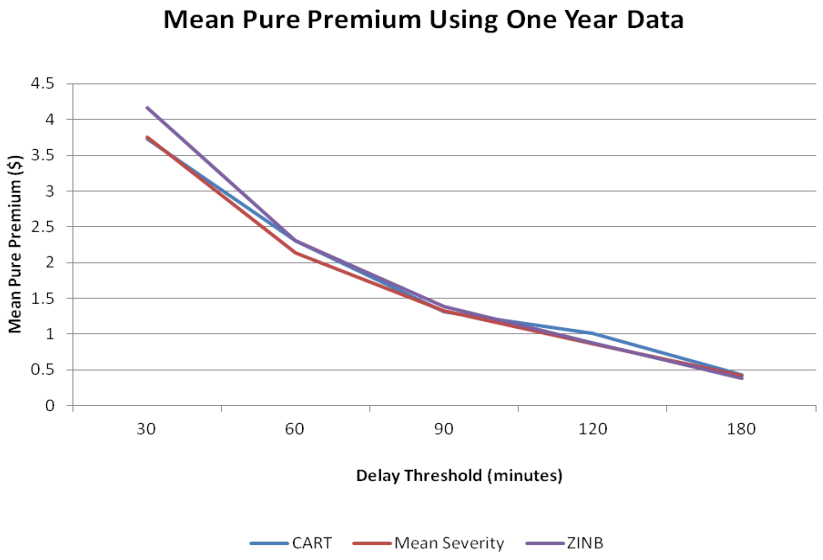

Figure 55: Adaptability: mean pure premium using one year historical data

The three lines graphed in Figure 55 present the mean pure premiums of the two models and the average severity. In this figure, the curve of ZINB model and the curve of CART model are close to the curve of average severity. This means that the estimations of CART model and ZINB model preform well. Furthermore, the difference between the mean pure premium and the mean severity does not exceed 0.4. And the difference decreases as the delay threshold increases. In this aspect, the estimations of the two models are both accurate and have small deviations.

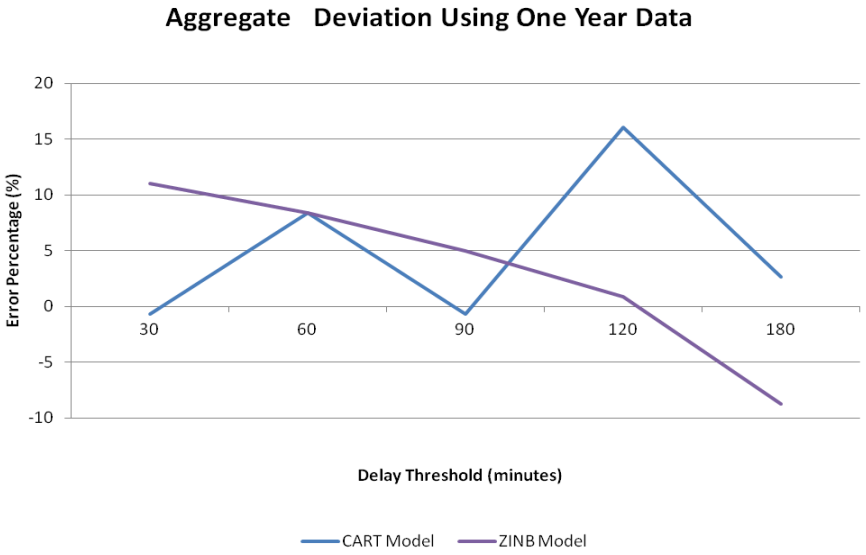

Figure 56: Adaptability: aggregate deviation using one year historical data

The aggregate deviations in Figure 56 are presented by the percentage of mean accident severity. The aggregate deviation of ZINB model decreases from 10.98% to -8.75% when the delay threshold increases from 30 minutes to 180 minutes. The curve of CART model is irregular and ranges from -1% to 8.4% except that the deviation equals 16% at 120 minutes. In total, the mean absolute deviations of ZINB model and CART model are separately 5.7% and

6.8%, which are acceptable.

Last Two Years Data

In this part, the historical data in the period 2016-2017 are used as the training data. Similarly, the first quarter data of 2017 is used to evaluate the estimation performance of the generated models. Compared to the result of one year historical data, the result of two years historical data has similar conclusions except that the estimations have larger overestimations.

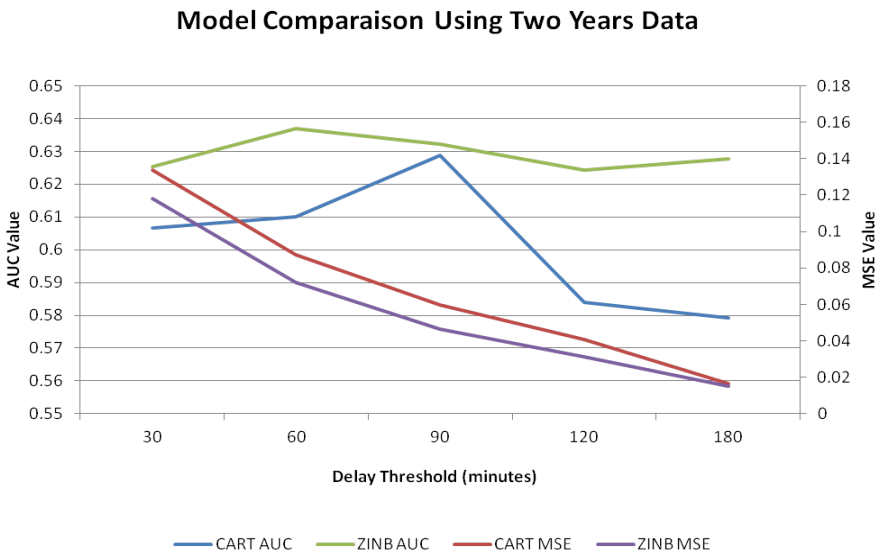

Figure 57: Adaptability: model performance using two years historical data

Same as the result of one year historical data, ZINB model performs better than CART model by either AUC criterion or MSE criterion. The AUC values of ZINB model are around 0.63 and the deviations do not exceed 0.01. In contrast, the AUC values of CART model are largely influenced by the delay threshold and the AUC values range from 0.58 to 0.63. Therefore, the performance of ZINB model is stable but the performance of CART model varies a lot by the AUC criterion. The MSE values of ZINB model are always smaller than the MSE values of CART model, regardless of the delay threshold change. This means that the ZINB model is more accurate by the criterion of MSE. In conclusion, the ZINB model is more stable and accurate than the CART model by the criterion of AUC or MSE.

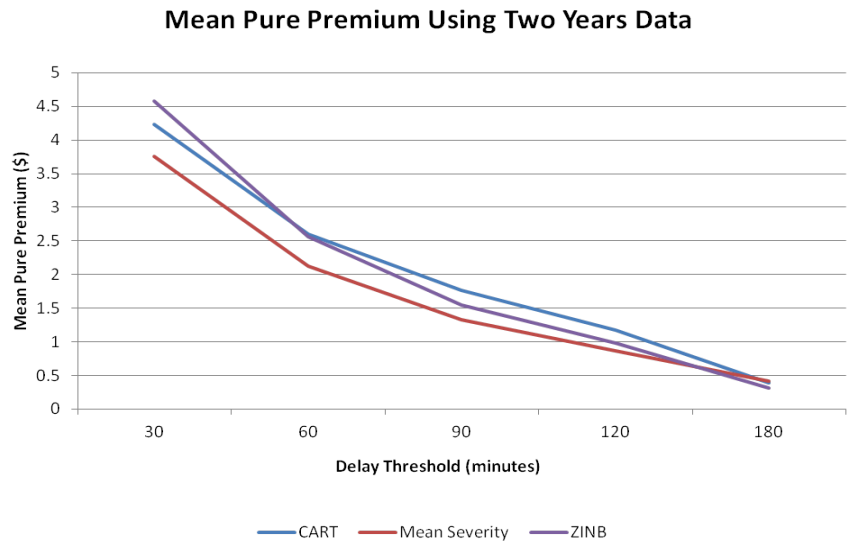

Figure 58: Adaptability: mean pure premium using two years historical data

This figure presents clearly the overestimations of ZINB model and CART model. The blue line (CART model) and purple line (ZINB model) are all above the red line (average severity) when the delay threshold is from 30 minutes to 120 minutes. Therefore, for both model, the mean pure premium is larger than the average severity at any delay threshold smaller than 180 minutes.

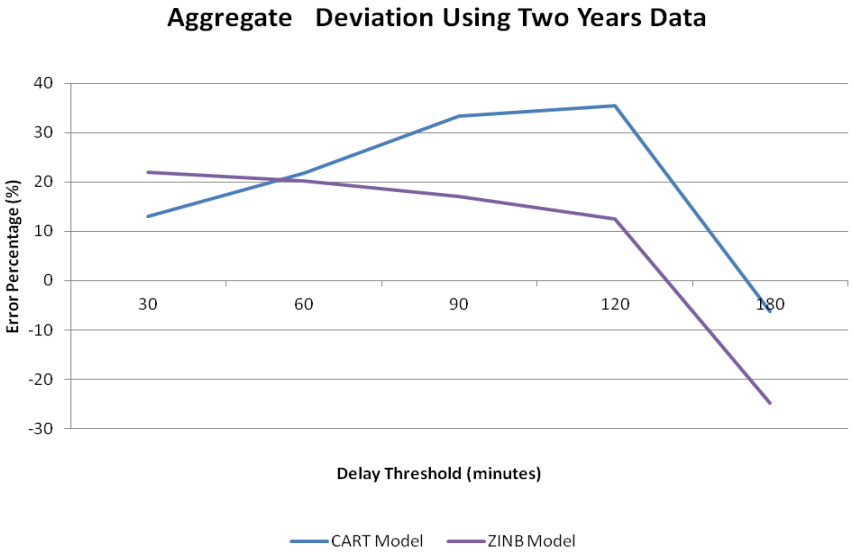

Figure 59: Adaptability: aggregate deviation using two years historical data

This figure presents the aggregate deviation in percentage. When the delay threshold is from 30 minutes to 120 minutes, the aggregate deviation is always larger than 10%. And in total, the mean absolute deviations of ZINB model and CART model are separately 21.93% and 19.30%, which are not precise and show large overestimations. Compared with the deviations using last one year data, those deviations are larger and show that the estimation is worse. Therefore it is concluded that the data of 2015 does not have similar effect as the data of 2016.

Last Three Years Data

In this part, the three years historical data in the period 2014-2016 serve as the training data. And the first quarter data of 2017 is used to evaluate the estimation performance of the generated models. The corresponding result shows also overestimation, and is similar to the result of two years historical data.

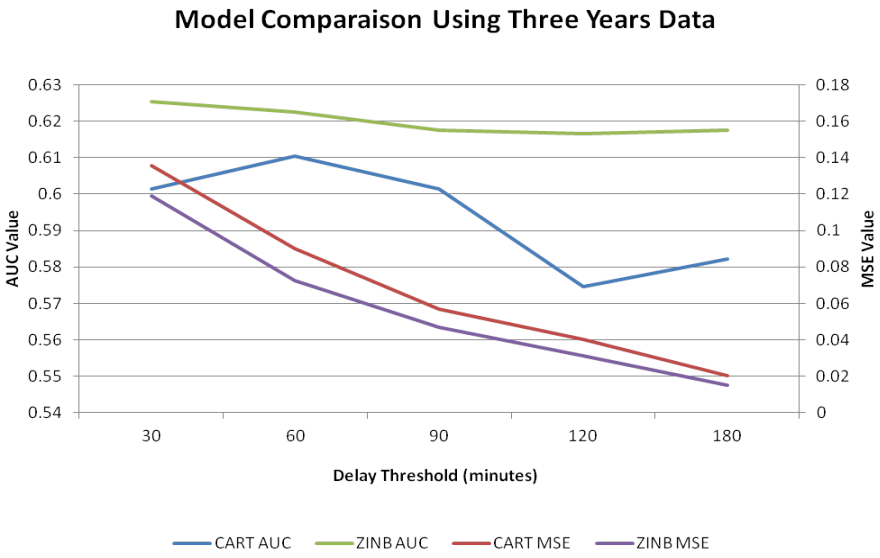

Figure 60: Adaptability: model performance using three years historical data

In Figure 60, the AUC values of ZINB model are all between 0.61 and 0.63, while the AUC values of CART model are unstable and are all below 0.61. Same as the results of one year historical data and two years historical data, the model ZINB model performs better than CART model by the criteria of AUC and MSE.

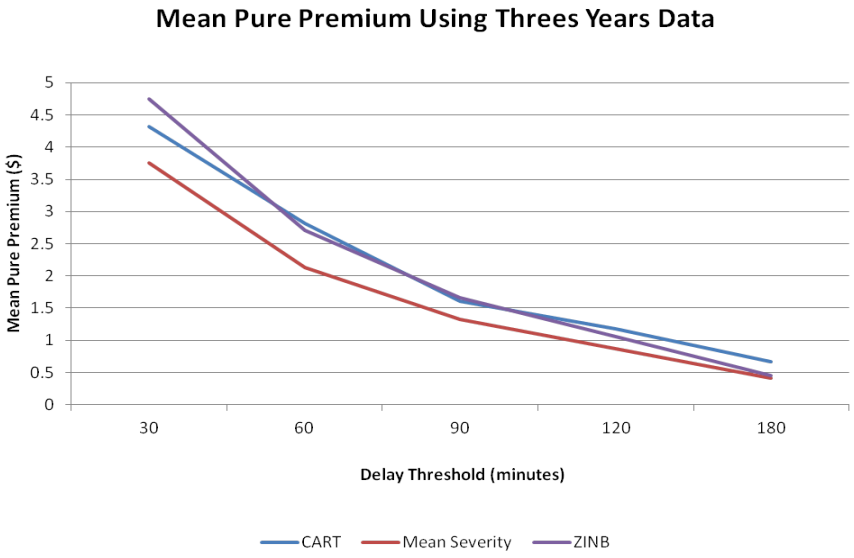

Figure 61: Adaptability: mean pure premium using three years historical data

In Figure 61, as the red line (Mean severity) is all below the blue line (CART) and purple line (ZINB), the pure premiums of CART model and ZINB model have overestimations in aggregate. Meanwhile, Figure 62 presents that the aggregate deviations are all over 10% for any delay threshold.

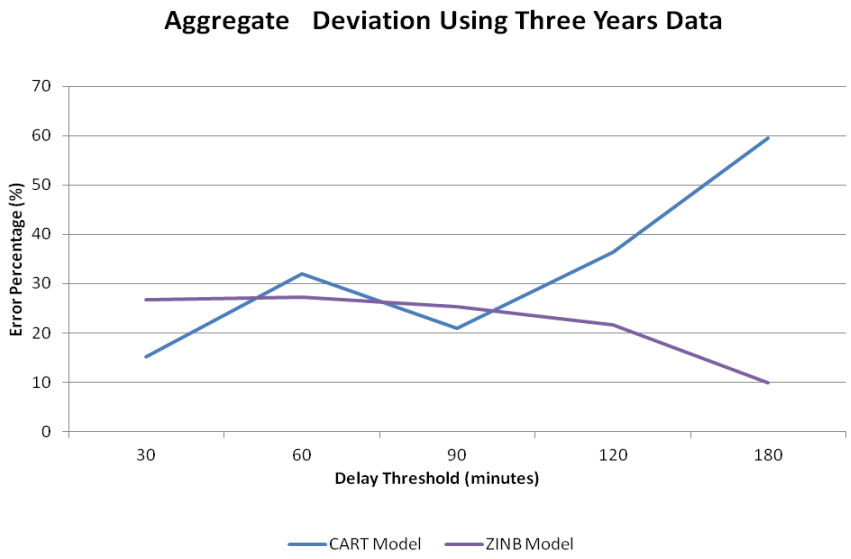

Figure 62: Adaptability: aggregate deviation using three years historical data

Figure 62 presents the aggregate deviations of CART model and ZINB model. The aggregate deviations of ZINB are stable, do not exceed 30%. But the aggregate deviations of CART model are unstable, and the value ranges from 10% to 60%. The aggregate deviations of CART model and ZINB model are separately 32.82% and 22.16%, meaning that the two models show significant overestimations.

To sum up, according to the AUC and MSE values of result, the ZINB model performs better than CART model at any delay threshold using the training data of one year, two years, or three years[2]. And in aggregate, the estimation using one year training data is more precise, and shows small overestimation. However the estimation using two years training data or three years data is not precise, and shows significant overestimations.

4.3.3 Other Airports

The previous study is focused on the database of the New York John F. Kennedy International Airport. To verify the adaptability of the model for the other airports, the databases of ten other airports are studied, including Chicago Midway International Airport, Boston Logan International Airport, etc. As the model trained by the last one year data have the best predictive performance (cf 4.3.2 Data Limitation), for each new airport, the data of 2016 is used to build the model for the departure delay. The delay threshold is set at thirty minutes in the analyses. Then the first quarter data of 2017 is used to validate the models. And the aggregate deviation, AUC criterion, and MSE criterion are used to evaluate the prediction performances.

[2] cf. The results of four years and five years have also large overestimation, and are included in Appendix F and Appendix G

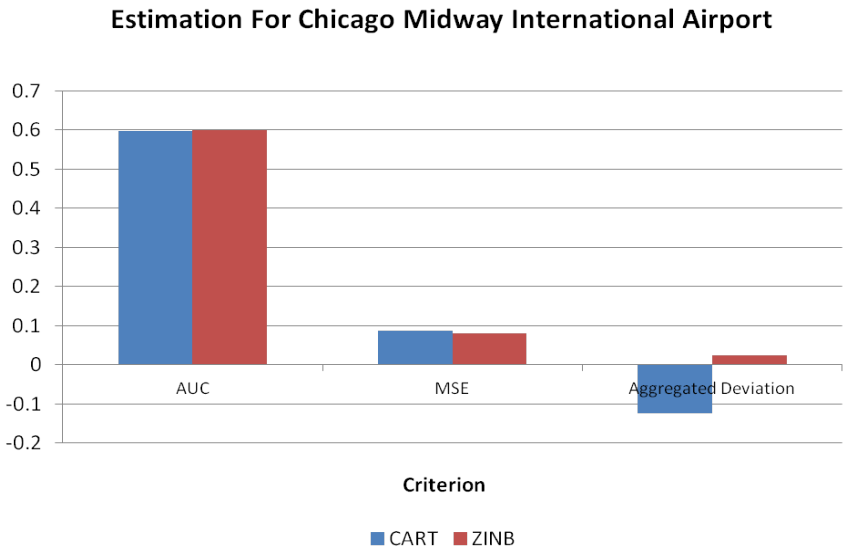

Figure 63: Prediction Performances-Chicago Midway International Airport

The result of Chicago Midway International Airport is presented in Figure 63. The CART model is not as good as the ZINB model by the criterion AUC or MSE. And the total deviations of CART model and ZINB model are -12.5% and 2.5%. Thus, the pure premium of CART model is underestimated. And the aggregate deviation of ZINB model is relatively small and is acceptable.

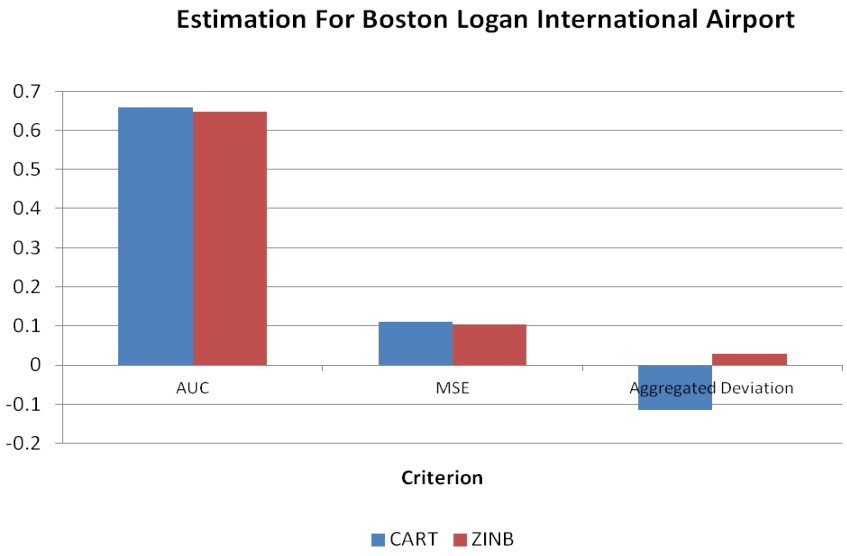

Figure 64: Prediction Performances-Logan International Airport

For Boston Logan International Airport, the AUC value of CART model is larger than the AUC value of ZINB model. Also, the pure premium of CART model is underestimated, and the aggregate deviation equals -11%. However, the estimation of ZINB model is more precise, and the aggregate deviation equals 2.8%, which is acceptable.

The previous analyses are used to compare the model performances using the data of one airport. To compare the model performance more generally, all the results of the ten airports

are analysed together.

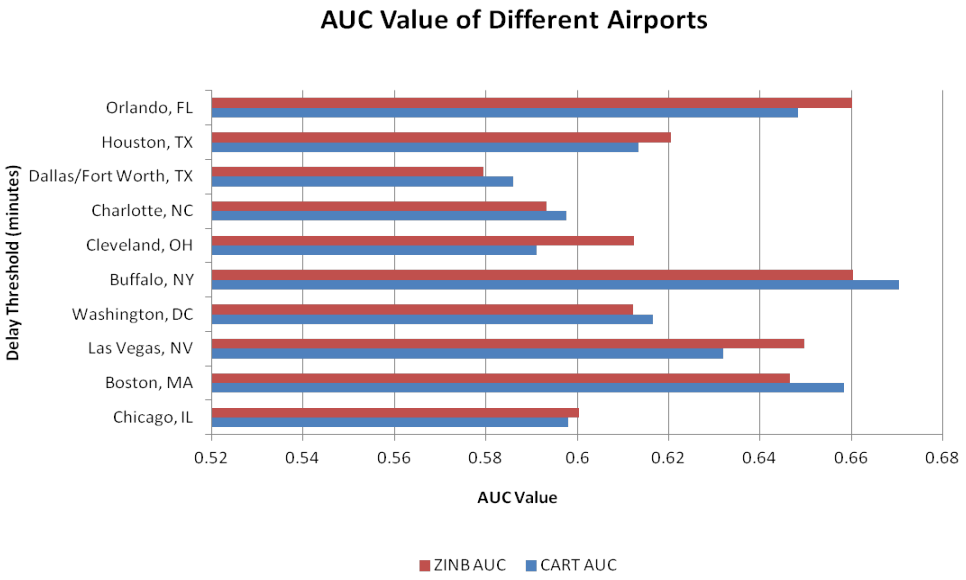

Figure 65: Adaptability of Other Airports-Criterion AUC

Figure 65 shows the AUC values of the frequency estimation of ten airports. In this figure, the AUC values of CART model and ZINB model are close. The AUC value of CART model is larger for five airports. While in the results of the other five airports, the AUC value of ZINB model is larger. Therefore, in the criterion of AUC, the estimation performances of CART model and ZINB model have no big difference. In the aspect of airport, the two models perform better for the airports Orlando FL, Buffalo NY, and Boston MA, of which the two AUC values are larger than 0.64. As the AUC values of New York John F. Kennedy International Airport are only 0.63 and 0.60, the CART model and ZINB model are not specified for New York John F. Kennedy International Airport.The two models can also be applied to other airports.

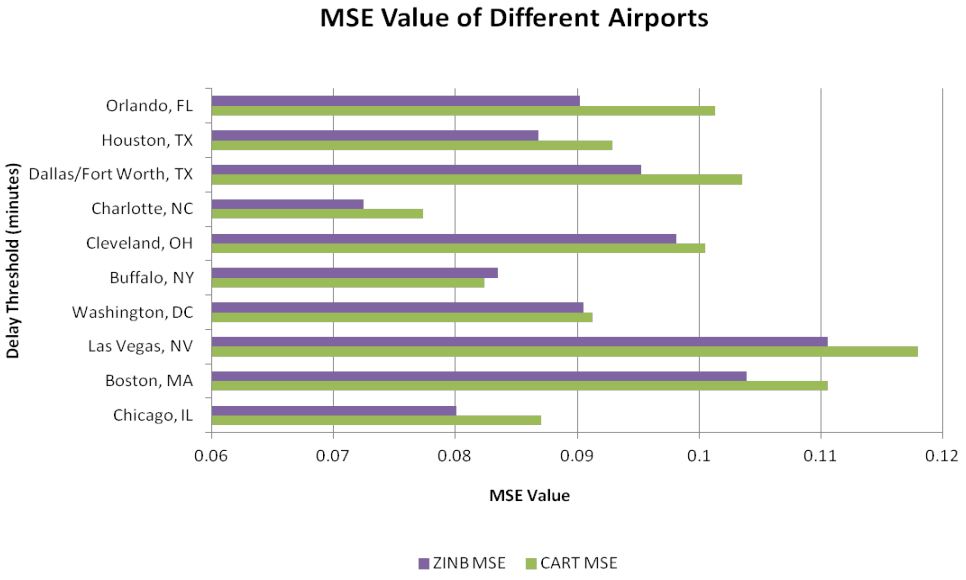

Figure 66: Adaptability of Other Airports-Criterion MSE

Figure 66 presents the MSE values of the frequnecy estimation of the same ten airports. In this figure, the MSE value of ZINB model is smaller than the MSE value of CART model for nine airports. Only for the airport Charlotte NC the ZINB model has higher MSE value. In the aspect of airport, the two models perform better for the airport Charlotte, NC, of which the two MSE values are smaller than 0.0775. And the two models perform worse for airport Boston MA and airport Las Vegas NV, of which the MSE values are all over 0.1.

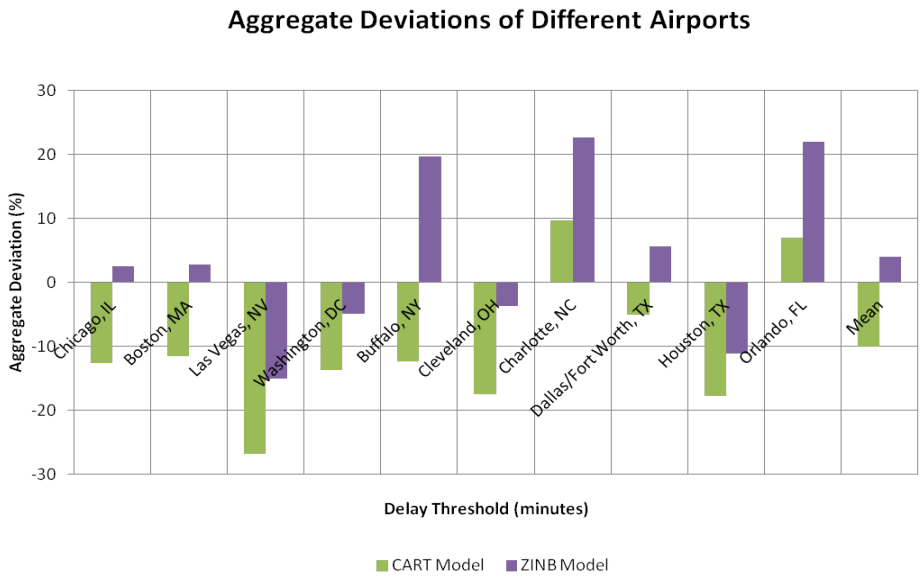

Figure 67: Adaptability of Other Airports- Aggregate Deviation

Figure 67 presents the aggregate deviations of the ten airports. The CART model underestimates the delay frequencies for 8 airports, and the average aggregate deviation is -10.0%. In contrast, the ZINB model underestimates the delay frequencies for only 4 airports, and the average aggregate deviation is 4.1%. Therefore, in the aspect of aggregate deviation, ZINB model is better than CART model.

To sum up, by the criterion of AUC, CART model and ZINB model have the same prediction performance for different airports. While by the criterion of MSE or by the aggregate deviation, ZINB model is better than CART model for different airports. In consequence, the ZINB model is preferable to CART model in general.

4.4 Comparison of Approaches

The two approaches study the independent variable in different ways, which result in different estimation performances. And the performance difference of the two approaches is studied by comparing the best models of each approach.

The CART model has a wider range of estimated technical premium while the ZINB model has more concentrated distribution of the estimated technical premium. Therefore, when a wide range pricing strategy is adopted, the model of discrete approach is more proper. It could well integrate the heterogeneity of database and then precisely specify the technical premium for the extreme situation. However, when a relatively concentrated price strategy is adopted, the model of ZINB model is better as the technical premium of discrete approach does not vary

largely. According to the criteria of AUC and MSE, the model of discrete approach performs better in general. In most situation, the technical premium estimations of discrete approach are more precise.

When the value of delay threshold is large, the ZINB model always presents underestimation of the flight departure delay frequency. When the value of delay threshold is not large and is like 30 minutes, 60 minutes, 90 minutes and 120 minutes, the ZINB model has more stable performances according to the results of limited historical data. Also, the AUC and MSE values of ZINB model are better than the values of CART model. In view of the above, it is concluded that the model of discrete approach is more stable and more consolidated when the value of delay threshold is not bigger than 120 minutes. When large delay threshold such as 160 minutes is applied, the model of binary approach is more proper. In addition, the model of discrete approach is preferable when a changeable delay threshold is applied as one model of discrete approach can be applied to all the delay thresholds.

By the results of section 4.3.3 (cf 4.3.3 Other Airports), the ZINB model has smaller underestimation probability than the CART model. In addition, for the Chicago Midway International Airport and Boston Logan International Airport, the CART model has aggregate underestimations larger than 10% . In contrast, there is no underestimation for the ZINB model. When the two models are applied to the ten airports, the CART models shows great underestimation while the ZINB works well. In this case, the model of discrete approach is preferable when there is a strong solvency requirement.

Chapter 5

Conclusion

To model the departure delay frequency, eight explanatory variables are selected from the database, including year, month, day of the week, vacation, scheduled departure time, scheduled arrival time, distance, and airline. The database is divided into training set and test set. For each model, the training set is used to configure the model to be optimal. Then the model performance for test set presents the true prediction level.

By the Vuong's non-neseted test, the ZINB model is the optimal model in the discrete approach. The six binary models have been validated by the one hundred groups Hosmer-Lemeshow Test. By the AUC criterion and MSE criterion, CART model is the best binary model and ZINB model is the best discrete model.

Consistently, the CART model and ZINB model are the best models when the nine models are applied to estimate the pure premium of the first quarter flights in 2017. The CART model performs better in aggregate estimation and the ZINB model performs better in individual estimation. The result of scenario tests for the two models concludes that the aggregate deviations are stable and limited.

For different delay threshold, CART model is stable while ZINB model has a small overestimation at low threshold and has a small underestimation at high threshold. The two models are validated by recent one year data, and show small overestimation when using recent two or three years data. For different airport, the ZINB model has precise aggregate estimation while CART model shows small estimation. In conclusion, the adaptability of CART model and ZINB model is validated, and the ZINB model is more adaptable with no aggregate underestimation and with better individual estimation performance. In general, the ZINB model is the best model for the pure premium estimation.

For the future work, other machine learning models such as svm and neural network can also be studied as possible models. With respect to that the ZINB model decreases by the delay threshold, a more plat distribution or model can make the zero-inflated model more precise.

In this essay, the model is specified for each airport. Therefore, a possible future study could be integrating the airport variables into the models and simplifying the airport based model to one general model. In addition, as more and more data are available nowadays, other pricing variables such as weather and airport flight number could also be integrated into the models.

Bibliography

[1] Andy Liaw and Matthew Wiener. (2002) *Classification and Regression by randomForest.* R news 2 (3), 18-22, 2002

[2] Cosma Shalizi, (2009) *Classification and Regression Trees.* `http://www.stat.cmu.edu/ cshalizi/350/lectures/22/lecture-22.pdf`

[3] David Posada Thomas R. Buckley. (2004) *Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests.*. Syst Biol 2004; 53 (5): 793-808.

[4] DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988) *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 837-845.

[5] Famoye, F., Singh, K.P. (2006) *Zero-inflated generalized Poisson model with an application to domestic violence data.* J. Data Sci. 4, 117–130.

[6] Frédéric PLANCHET. (2016) *Statistique des modèles paramétriques et semi-paramétriques* `http://www.ressources-actuarielles.net/`

[7] Gourieroux, C., Monfort, A. Trognon, A. (1984b). *Pseudo maximum likelihood methods: Theory.* Econometrica 52(3), 681–700.

[8] G. Ridgeway. (2007) *Generalized Boosted Models: A guide to the gbm package.* `http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf`

[9] Hanley, J. A., McNeil, B. J. (1982) *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 143(1), 29-36.

[10] Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. (2013) *Applied Logistic Regression.* 3rd ed. Hoboken, NJ: Wiley.

[11] James K. Lindsey. (1997) *Applying Generalized Linear Models.* Springer-Verlag New York

[12] J.H. Friedman (2002). *Stochastic Gradient Boosting* Computational Statistics and Data Analysis 38(4):367-378.

[13] Julian J. Faraway. (2002) *Practical Regression and Anova using R.*

[14] J. Friedman, T. Hastie, and R. Tibshirani. (2000) *Additive logistic regression: a statistical view of boosting.* Annals of Statistics, 28(2):337–407.

[15] Microinsurance Network. (2017) *A brief history.* `http://www.microinsurancenetwork.org/brief-history`

[16] Simon N. Wood. (2007) *Fast stable direct fitting and smoothness selection for generalized additive models*. [*Journal of the Royal Statistical Society*]. Series B (Statistical Methodology), 70: 495–518. doi:10.1111/j.1467-9868.2007.00646.x.

[17] T. Fawcett. (2006) *An introduction to ROC analysis*. Pattern Recognition Letters, (27): 861–874, 2006. 99, 100.

[18] VUONG, Q. H. (1989) *Likelihood ratio tests for model selection and non-nested hypotheses*. Econometrica 57:307–333.

[19] White, H. (1982). *Maximum likelihood estimation of misspecified models*. Econometrica 50(1), 1–26.

[20] Zhou, X. H., Obuchowski, N. A., McClish, D. K. (2011) *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach*. Statistical methods in diagnostic medicine. Wiley-Blackwell

List of Figures

List of Tables

Appendices

Appendix A: GLM Result

This table is the result of first p-value backward selection. (cf. 3.1.3 Variable Selection and Result)

Variable	***value***	***Pr(>|z|)***
(Intercept)	-0.13057	0.00761
bs(Year)1	0.26886	3.24E-15
bs(Year)2	-0.13682	3.89E-07
bs(Year)3	0.206446	<2.00E-16
month10	-0.45015	<2.00E-16
month11	-0.61126	<2.00E-16
month12	0.147328	2.28E-14
month2	0.044544	0.02696
month3	-0.06468	0.00105
month4	-0.16019	2.56E-15
month5	-0.08788	9.29E-06
month6	0.216104	<2.00E-16
month7	0.383659	<2.00E-16
month8	0.195231	<2.00E-16
month9	-0.39635	<2.00E-16
dayOfWeek2	-0.15419	<2.00E-16
dayOfWeek3	-0.12148	1.30E-15
dayOfWeek4	0.034357	0.01878
dayOfWeek5	0.040456	0.00561
dayOfWeek6	-0.15205	<2.00E-16
dayOfWeek7	0.026376	0.07244
AirlineAS	-0.57868	0.05965
AirlineB6	0.002327	0.85667
AirlineDL	-0.26019	<2.00E-16
AirlineEV	0.34195	4.49E-14
AirlineHA	-0.62588	1.02E-05
AirlineVX	-0.17297	2.95E-08
Airline9E	0.329836	<2.00E-16
AirlineMQ	-0.04479	0.03197
AirlineNW	0.349731	7.06E-06
bs(ArrTime, knots = 480, degree = 1)1	-2.03838	2.00E-16
bs(ArrTime, knots = 480, degree = 1)2	-0.08387	7.01E-06
log(Distance)	-0.12754	<2.00E-16

Table 19: GLM Result without Interaction Terms

Appendix B: GAM Result

This table is the result for GAM model in the Section 3.1.4.

Variable	***Parametric Anova***	***P-value***	***Nonparametric Anova***	***P-value***
s(Year)	0.5061	0.4768	264.81	2.20E-16
month	376.6634	2.20E-16		
dayOfWeek	69.984	2.20E-16		
Airline	164.0787	2.20E-16		
s(Vacation)	1.7381	0.1874	99.56	2.20E-16
s(DepTime)	10977.99	2.20E-16	743.5	2.20E-16
s(ArrTime)	63.1038	1.96E-15	178.98	2.20E-16
s(Distance)	441.93	2.20E-16	354.77	2.20E-16

Table 20: GAM Result

Appendix C: ROC Curves

The figures correspond to the ROC curves of six binary models.

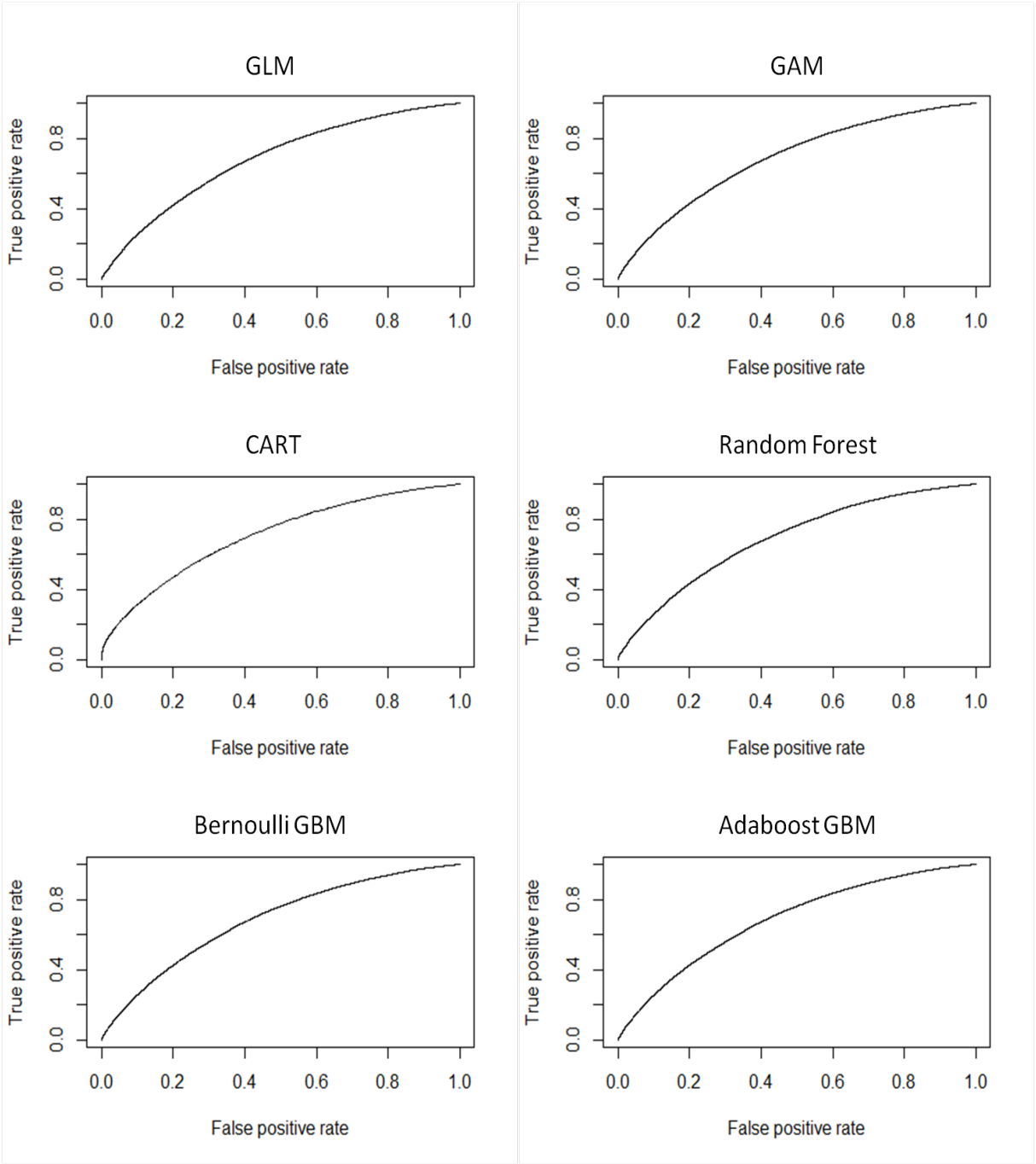

Figure 68: ROC Curve of Binary Approach Models

Appendix D: CART Model

This regression tree corresponds to the optimal one with 1096 nodes. (cf. 3.1.5 CART: Classification And Regression Tree)

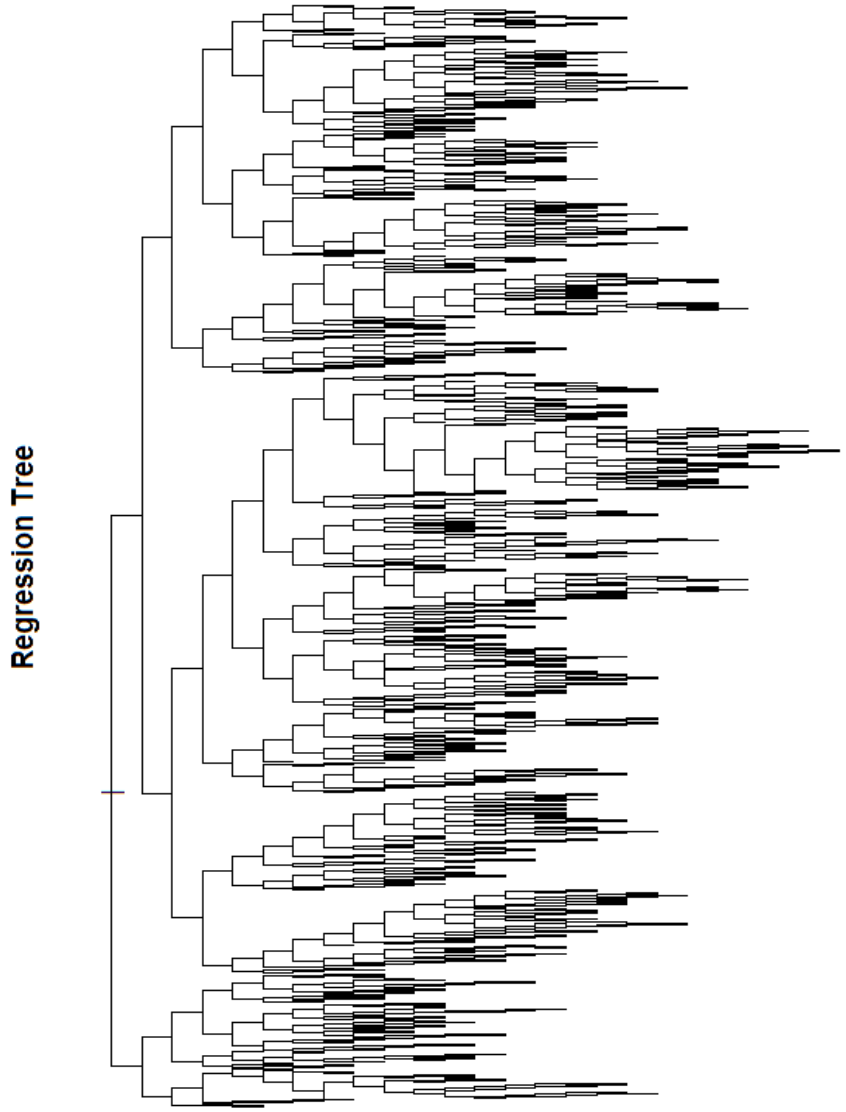

Figure 69: CART Model

Appendix E: Aggregate Performance of Vacation Variable

The true delay frequencies are irregular, and the CART model reflects a part of the influence. While the ZINB model integrate the influence in a concave way.

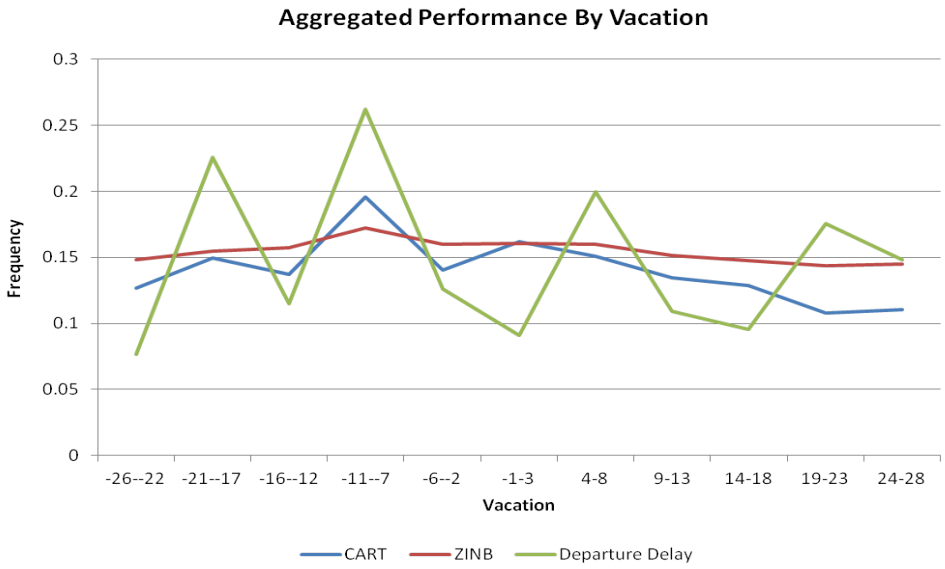

Figure 70: Aggregate Performance of Vacation Variable

Appendix F: Adaptability of Four Years Historical Data

Those figures present the model performance using four years historical data.

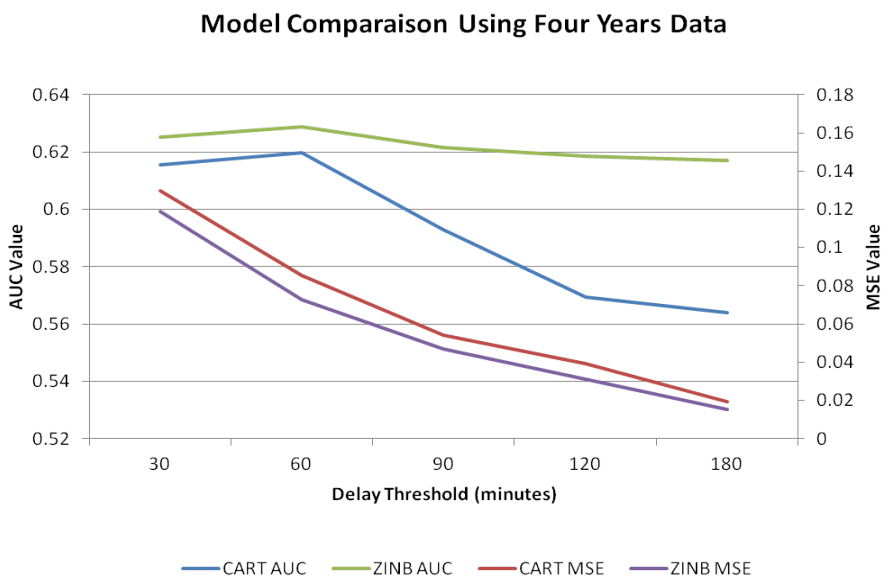

Figure 71: Adaptability: model performance using four years historical data

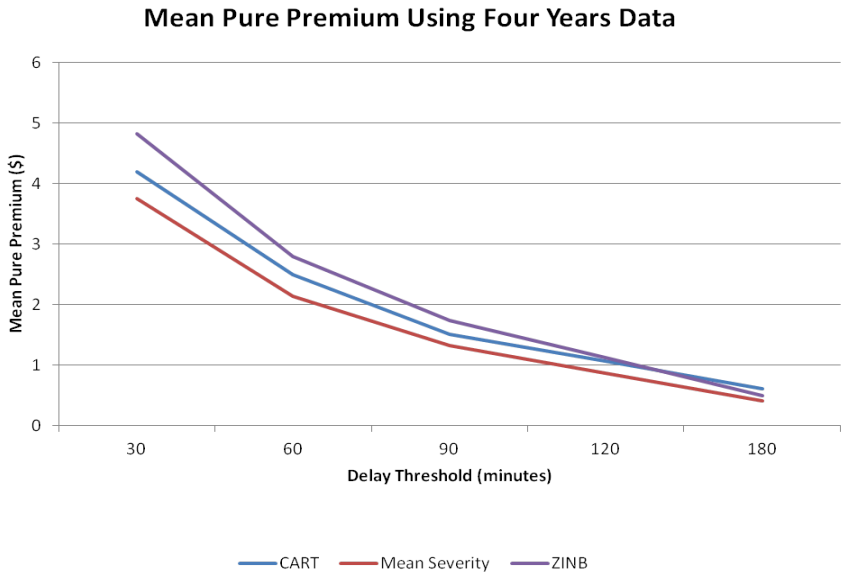

Figure 72: Adaptability: pure premium using four years historical data

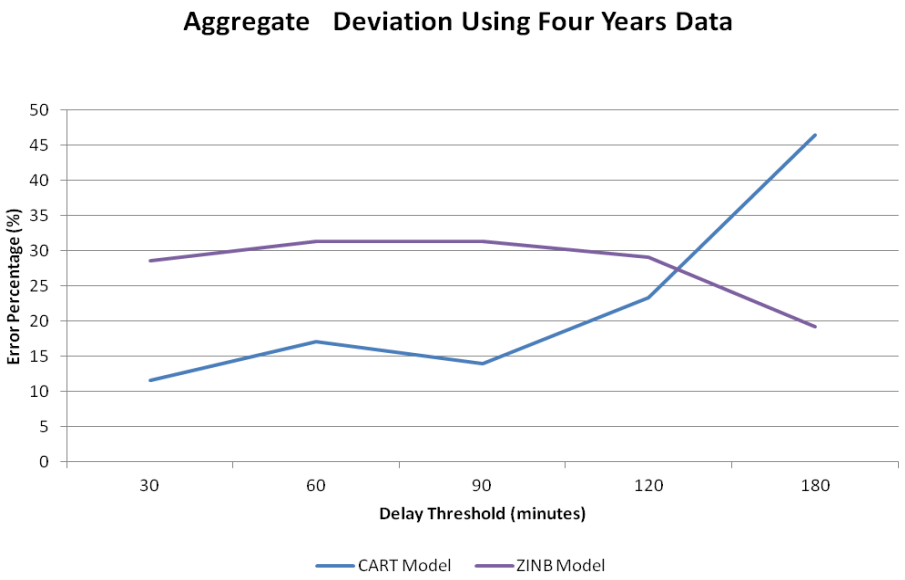

Figure 73: Adaptability: aggregate deviation using four years historical data

Appendix G: Adaptability of Five Years Historical Data

Those figures present the model performance using five years historical data.

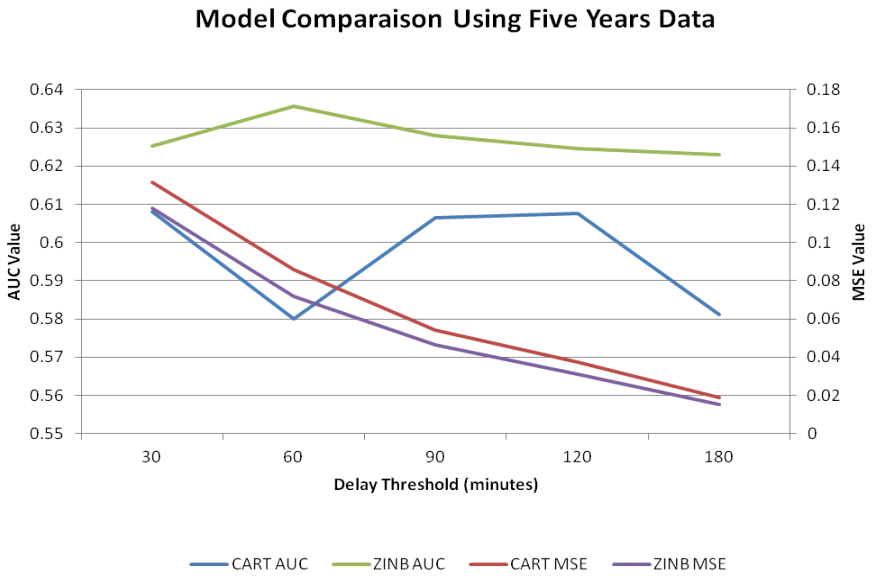

Figure 74: Adaptability: model performance using five years historical data

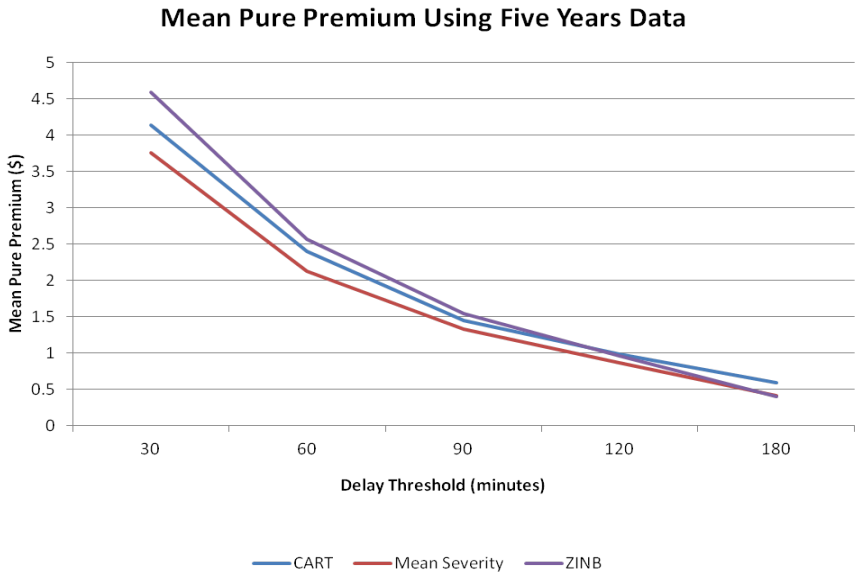

Figure 75: Adaptability: pure premium using five years historical data

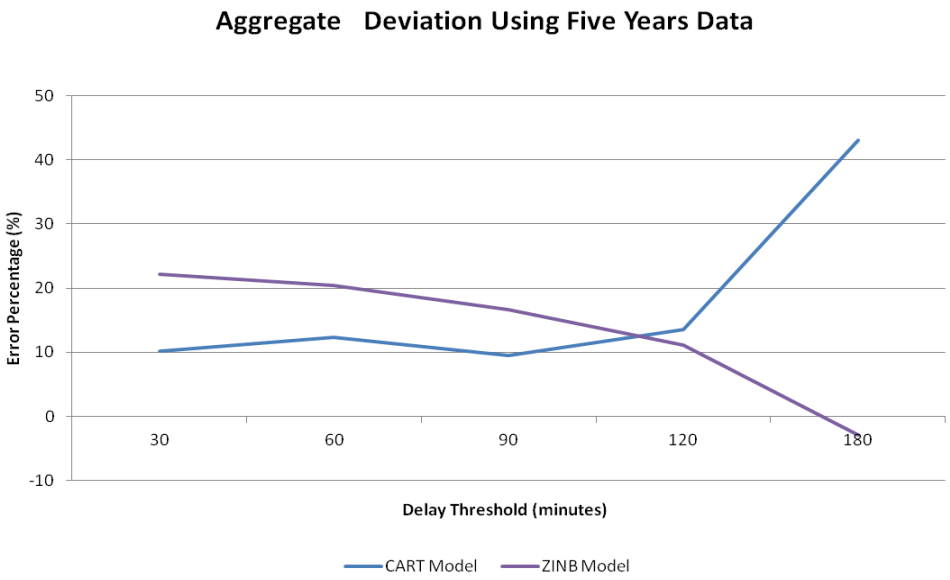

Figure 76: Adaptability: aggregate deviation using five years historical data

Appendix H: Vacation calender day

Table 21 presents the vacation calender days in the period 2009-2016.

2017 :	02/01/2017	16/01/2017	20/02/2017	14/04/2017	29/05/2017
2016 :	01/01/2016	18/01/2016	15/02/2016	25/03/2016	30/05/2016
2015 :	01/01/2015	19/01/2015	16/02/2015	03/04/2015	25/05/2015
2014 :	01/01/2014	20/01/2014	17/02/2014	18/04/2014	26/05/2014
2013 :	01/01/2013	21/01/2013	18/02/2013	29/03/2013	27/05/2013
2012 :	02/01/2012	16/01/2012	20/02/2012	06/04/2012	28/05/2012
2011 :	01/01/2011	17/01/2011	21/02/2011	22/04/2011	30/05/2011
2010 :	01/01/2010	18/01/2010	15/02/2010	02/04/2010	31/05/2010
2009 :	01/01/2009	19/01/2009	16/02/2009	10/04/2009	25/05/2009

2017 :	04/07/2017	04/09/2017	23/11/2017	25/12/2017
2016 :	04/07/2016	05/09/2016	24/11/2016	26/12/2016
2015 :	03/07/2015	07/09/2015	26/11/2015	25/12/2015
2014 :	04/07/2014	01/09/2014	27/11/2014	25/12/2014
2013 :	04/07/2013	02/09/2013	28/11/2013	25/12/2013
2012 :	04/07/2012	03/09/2012	22/11/2012	25/12/2012
2011 :	04/07/2011	05/09/2011	24/11/2011	26/12/2011
2010 :	05/07/2010	06/09/2010	25/11/2010	24/12/2010
2009 :	03/07/2009	07/09/2009	26/11/2009	25/12/2009

Table 21: Vacation calender day in the period 2009-2016