# The University of Texas at Austin

# Academic Year 2022/23 Spring Semester

ECO 395M: Data Mining and Statistical Learning

**Flight Delay Prediction for *Allianz Travel Insurance* Company**

Chen-Yen Liu (lc42838)
Yuzhu Liu (yl39726)
Ziyue Wang (zw7669)

**Word Count:** 2960
**Page Count:** 10

## 1. Problem Introduction

Airplanes are among the most advanced forms of transportation and have become an essential part of everyone's life, be it for business, travel, visiting, or other purposes. With the increasing popularity of air travel, it has become a critical mode of transportation for both personal and commercial purposes. However, flight delays have become a common issue that affects passengers and airlines alike. One key stakeholder impacted by this issue is the insurance company. Suppose we are three business analysts commissioned by Allianz Travel Insurance, a major insurance company in the United States that offers flight delay insurance, our task is to predict flight delays. The price of insurance sold to passengers varies according to factors such as flight time and departure airports. However, the insurance industry faces limited risk control capabilities and high payout rates for delay insurance, leading to reduced profits and growing concerns among customers. Therefore, accurately forecasting flight delays, controlling payout risks, and adjusting insurance delay policies to minimize payouts and increase profits have become crucial issues for insurers. As key stakeholders in this industry, it is important to address these challenges through predictive modeling and data analysis.

Given the limited data available from the US Bureau of Transportation, we have simplified the scenario for our analysis. We assume that the price of insurance offered by Allianz Travel Insurance is directly related to the duration of flight delays, with compensation offered to customers who experience delays exceeding 15 minutes. However, it is important to note that the actual business model for the insurance company is much more complex than what we can deliver in this report. Despite this complexity, accurately predicting flight delays and controlling payout risks remains a critical concern for insurers, and we will seek to address these challenges in our analysis.

In this report, we will start by introducing the dataset and explaining where and how we obtained it, as well as the information it contains. Next, we will select the input variables to determine the most relevant and practical variables for our models. Then, we will perform data cleaning to ensure the accuracy and completeness of the dataset. After preparing the data, we will introduce four models: KNN, Random Forest, Naïve Bayes and Logistic Regression. Based on our evaluation, we will recommend the best protected model to Allianz Travel Insurance. Finally, we will offer some practical business recommendations and guidance on how to use this model effectively.

## 2. Data Exploration

### 2.1 Overview of the Dataset

The dataset used for this project was obtained from the Bureau of Transportation Statistics (BTS), which provides data related to the transportation system in the United States. We pick up data in 2019. However, the raw data obtained from the BTS website was not in a clean, usable

format for analysis. Therefore, in order to create a clean dataset for analysis, we had to download data airport by airport and airline by airline from the BTS website.

To obtain information about the weather at each airport, we also accessed data from the National Weather Service (NWS) website. This weather data was merged with the airline and airport data obtained from the BTS to create a comprehensive dataset.

The process of obtaining this data required a significant amount of effort, including downloading and processing large amounts of data from multiple sources. However, this effort was necessary to ensure that the resulting dataset was accurate, complete, and ready for analysis. By merging data from multiple sources, we were able to create a dataset that provided a comprehensive view of the flights across the United States.

In summary, there are 50,205 observations with 22 input variables and 1 output variables in the dataset. The details of variables and data will be shown in the following parts.

**2.2 Input Variable Selection**

There are 22 input variables and 1 output. Output is DEP_DEL15, which is a target binary of a departure delay over 15 minutes and 1 indicates yes.

When it comes to selecting inputs, we first consider the relevance of the variables and whether or not an insurance company can obtain the necessary data when using the predictive model. For instance, the variable SEGMENT_NUMBER, which measures the number of flights a specific airplane takes in a single day and numbers the flights sequentially based on time, is difficult for insurance companies to obtain since the exact number of flights for a given airplane is uncertain and depends on the carrier's decision. Similarly, the variable CONCURRENT_FLIGHTS, which shows the number of flights departing from the airport at the same departure time, is difficult to predict since flight schedules change frequently. Regarding the weather-related variables, such as precipitation, temperature, snow, and wind speed, well-performing prediction models already exist. Therefore, we consider these variables as known inputs when predicting flight delays. In other words, we assume that the values of these weather-related variables are already known and available to us when making predictions.

We consider the variables NUMBER_OF_SEATS, FLT_ATTENDANTS_PER_PASS, and PLANE_AGE to be irrelevant to the output as they do not directly impact flight delays. For our purposes, one airplane is considered one unit, and the number of passengers and crew members aboard is not a significant factor. Additionally, as long as the age of the aircraft falls within the normal age range for flying, we do not believe it will have a significant impact on delays. In this case, we shrink the input variables to 18.

Next, we examine the correlations between the 17 variables.

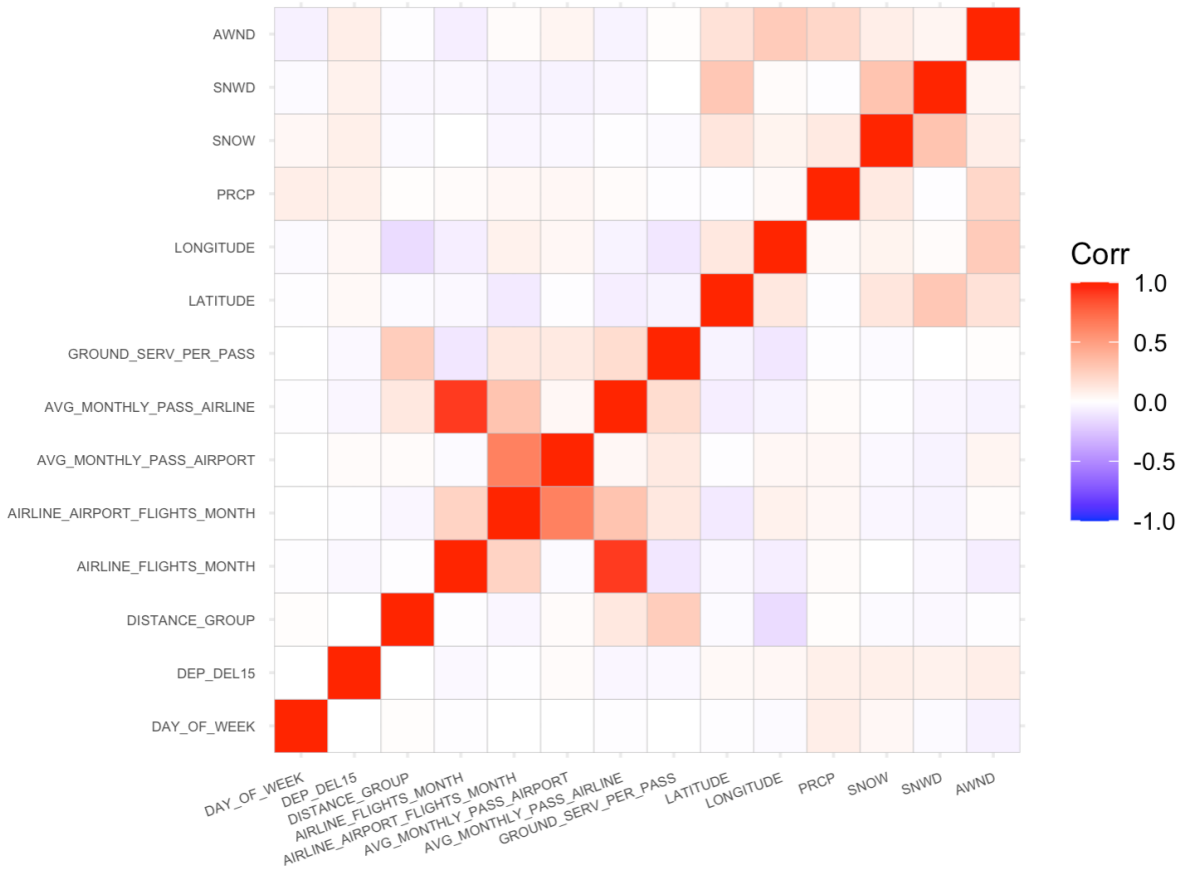


Figure 1

We observed a high positive correlation of 0.91 between Average Airline Flights per Month(AIRLINE_FLIGHTS_MONTH) and Average Passengers for the airline for the month (AVG_MONTHLY_PASS_AIRLINE). In this instance, we decided to keep the former since it can be predicted using previous years' records.

In addition, there is a 0.64 correlation between Average Flights per month for Airline AND Airport (AIRLINE_AIRPORT_FLIGHTS_MONTH) and Average Passengers for the departing airport for the month (AVG_MONTHLY_PASS_AIRPORT). As the latter variable contains less redundant information compared to the former one, we have decided to keep the variable AVG_MONTHLY_PASS_AIRPORT in our model.
Lastly, the LATITUDE and LONGITUDE variables serve as indicators for airports. Since we will already include the departing airports in our models, these variables are redundant, and we have decided to remove them from our input variables.

Finally, there are 13 input variables and 1 output variable in our dataset.

| No. | Name | Type | Feature Description |
|---|---|---|---|
| Input | | | |
| 1 | DAY_OF_WEEK | Categorical | Day of Week |
| 2 | DISTANCE_GROUP | Numerical | This is a number to indicate how far the aeroplane is going. Low numbers indicate a shorter distance. |
| 3 | DEP_BLOCK | Categorical | It indicates what time the flight is leaving from the airport in the form of blocks. |
| 4 | CARRIER_NAME | Categorical | Carrier |
| 5 | AIRLINE_FLIGHTS_MONTH | Numerical | Average Airline Flights per Month |
| 6 | AVG_MONTHLY_PASS_AIRPORT | Numerical | Average Passengers for the departing airport for the month |
| 7 | GROUND_SERV_PER_PASS | Numerical | Ground service employees (service desk) per passenger for airline |
| 8 | DEPARTING_AIRPORT | Categorical | Departing Airport |
| 9 | PREVIOUS_AIRPORT | Categorical | Previous airport that aircraft departed from |
| 10 | PRCP | Numerical | Inches of precipitation for the day |
| 11 | SNOW | Numerical | Inches of snowfall for the day |
| 12 | SNWD | Numerical | Inches of snow on the ground for the day |
| 13 | AWND | Numerical | Max wind speed for the day |
| Output | | | |
| 1 | DEP_DEL15 | Numerical | TARGET Binary of a departure delay over 15 minutes (1 is yes) |

Figure 2

### 2.3 Description of Categorical Variables

There are 5 categorical inputs: DAY_OF_WEEK, DEP_BLOCK, CARRIER_NAME, DEPARTING_AIRPORT, and PREVIOUS AIRPORT. To gain a better understanding of their distribution and relationships with the target variables, we have included a detailed analysis in figure 3. X axis are the number of delays and the y axis are the categorical inputs.

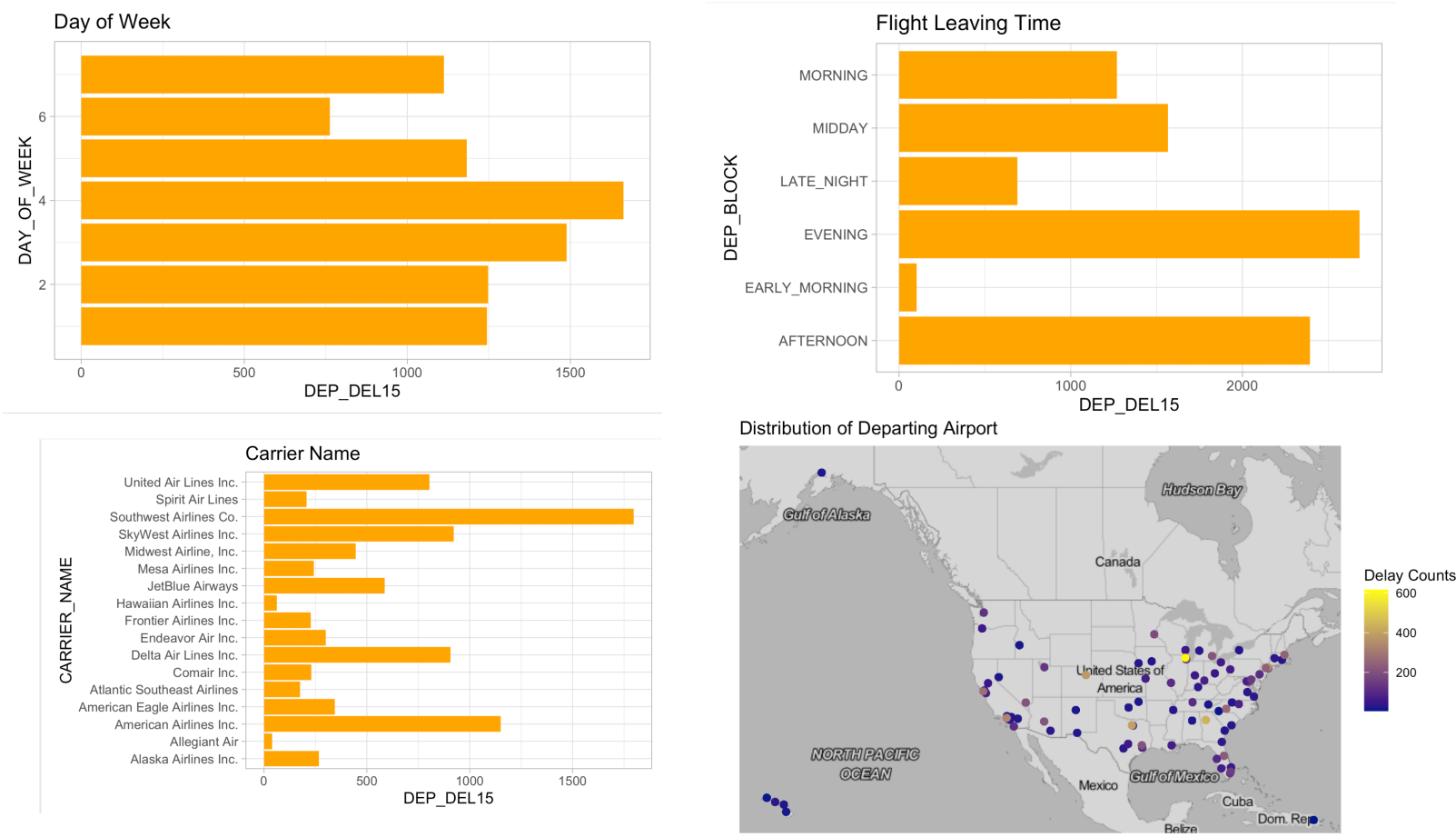


Figure 3

The 'PREVIOUS_AIRPORT' category represents the airport from which the aircraft departed before arriving at the current airport. It has two categories: 'None' indicating that there was no previous airport, and the name of the previous airport. For instance, if an airplane travels from Austin to Houston and then to New York, for Austin, the previous airport is "None," while for Houston, it is Austin. This situation not only occurs in interline ticketing but also when the airline decides the number of flights to be performed by one aircraft per day and its departure and destination. In this case, we can consider the airport with the previous airport as a stopover. There is time restriction in terms of these variables; however, due to the ambiguous definition,we cannot precisely determine the time range. We assume that if the airplane stays at the departing

airport for less than 2 hours before departing, we can consider it a stopover and from the previous airport. For example, if an airplane arrives at Austin airport and performs its next flight the following day, we believe that the previous airport is "None." In this case, we focus only on the effect of having a previous airport or not, not the specific airports. Therefore, we make it a binary variable with 0 (no previous airport) and 1 (have previous airport).

### 2.4 Description of Numerical Variables

There are 8 numerical inputs and their basic statistical features are presented in Figure 2. The standard deviations of 2 inputs, AIRLINE_FLIGHTS_MONTHS and AVG_MONTHLY_PASS_AIRPORT are extremely high, which indicates they enjoy a large degree of dispersion.

The means of PRCP, SNOW, and SNWD are significantly small compared to the maximum value, which may indicate they are uneven distributions and there are outliers.

| | Count | Min | Mean | Max | Standard Deviation |
|---|---|---|---|---|---|
| DISTANCE_GROUP | 50205 | 1 | 3.828 | 11 | 2.388371 |
| AIRLINE_FLIGHTS_MONTH | 50205 | 6713 | 59565 | 107363 | 33055.68 |
| AVG_MONTHLY_PASS_AIRPORT | 50205 | 89733 | 1600441 | 4365661 | 1119144 |
| GROUND_SERV_PER_PASS | 50205 | 0.00000713 | 0.0001359 | 0.000229 | 4.62349E-05 |
| PRCP | 50205 | 0 | 0.09882 | 4.6 | 0.2903781 |
| SNOW | 50205 | 0 | 0.09051 | 17.2 | 0.5318305 |
| SNWD | 50205 | 0 | 0.3743 | 25.2 | 1.458462 |
| AWND | 50205 | 0 | 8.577 | 26.62 | 4.381145 |

Figure 4

### 2.5 Missing Value

It is checked that there are no missing values in the data set.

### 2.5 Data Cleaning and Outliers

We use the standard, the data above 'Q3+3IQR' or below 'Q1-3IQR' as an outlier. However, we only deleted outliers for the following numerical variables,“DISTANCE_GROUP”, “AIRLINE_FLIGHTS_MONTH”,“AVG_MONTHLY_PASS_AIRPORT”,“GROUND_SERV_PER_PASS”, and “AWND”. We did not apply this method to PRCP, SNOW, and SNWD these three variables because as you can see from Figure 4, their mean and standard deviation is extremely small. If we apply this method to eliminate outliers, there are only a bunch of 0 observations left in these three variables. We are concerned that this may drop a lot of observations that are informative. Therefore, after we delete the outliers, there are 41,057 observations left.

## 3. Model Building and Evaluation

### 3.1 Model Selection and Variable Pre-processing

Four models will be tested in this report: logistic regression, random forest, Naive bayes and KNN.

As previously mentioned, the dataset contains four categorical variables, which cannot be directly used in the four models as they require numerical inputs. To address this, we used dummy variable encoding for each category. For instance, we used dummy variables for the day of the week, which has seven possible values, and created seven dummies: week_day_2 through week_day_7, with one of the days as the omitted variable. We applied the same method to the other three categorical variables. Consequently, we ended up with 120 input variables and one output variable. Since there are 41,057 data points in the dataset, we did not consider the issue of overfitting.

Besides categorical transformation, we also conduct standardization to non-binary numerical variables.

### 3.2 Methodology

We first evaluate the model performance by setting a threshold of our predicted probabilities at 0.5, which means if the predicted values are larger than 0.5, we assume the flight will delay. Then we calculate the confusion matrix, which tabulates predicted values versus actual values. Accuracy rate measures the ratio of correctly predicted values and total numbers. Higher accuracy rate is preferred.

The second selection criterion for models is RMSE, and we consider the model with the smallest RMSE to be better

When assessing the performance of a model on new data, there is a possibility of random variation due to the selection of data points for the train or test split. To mitigate this, we use a loop to obtain an average estimate of the out-of-sample RMSE and accuracy rate for 20 times. The results are presented in Figure 5.

Since the results of RMSE and accuracy rate are quite close, additional selection criteria are necessary to determine the superior model. As our client is an insurance company, profit estimation based on the model's predictions is also a crucial factor in selecting the winning model.

| | Accuracy Rate | RMSE |
|---|---|---|
| Logistic Regression | 0.8214 | 2.013 |
| Random Forest | 0.8248 | 0.3710731 |
| KNN | 0.8292 | 0.378 |
| Naïve Bayes | 0.4375 | 1.58892 |

Figure 5

### 3.3 Best Model Selection

#### 3.3.1 The 'Winning' Classifier

According to our client's official website, our team has assumed and simplified the cost for flight delay insurance and premium. If the flight is predicted not to delay, the fixed cost for insurance is $5 per person, and the purchase rate in this scenario is estimated as 100%. If the flight is predicted to delay, the fixed cost keeps unchanged but the price increases to $35 per person due to increased risk of premium, which is assumed to decrease the purchase rate to 50%.

| | Price | Purchase Rate |
|---|---|---|
| Predicted non-delay flight | $25 | 100% |
| Predicted delay flight | $35 | 50% |
| Insurance Premium | $100 | |

Figure 6: Insurance Price and Purchase Rate

Therefore, the cost/benefit matrix is calculated. We used the 20 loop samples to get 20 the confusion matrix and we get a averaged confusion matrix by averaging them. The expected return is calculated based on the averaged confusion matrix.

| | Non-delay | Delay |
|---|---|---|
| Predict_nondelay | 20 | -80 |
| Predict_delay | 30 | -70 |

| Logistic Regression | 0 | 1 |
|---|---|---|
| 0 | 6739 | 1453 |
| 1 | 14 | 6 |
| Expected Profit | 18540 | |

| KNN | 0 | 1 |
|---|---|---|
| 0 | 6663 | 1242 |
| 1 | 148 | 159 |
| Expected Profit | 30555 | |

| Naïve Bayes | 0 | 1 |
|---|---|---|
| 0 | 2889 | 367 |
| 1 | 3937 | 1019 |
| Expected Profit | 51810 | |

| Random Forest | 0 | 1 |
|---|---|---|
| 0 | 6564 | 1203 |
| 1 | 236 | 209 |
| Expected Profit | 31265 | |

Figure 7: Cost/Benefit Matrix and the Expected Return

Because our client concerns more about maximizing profit, and the index of accuracy is unsuitable in the imbalanced classifications (Luo, Qiao and Zhang, 2021), we assign a higher weight of 60% to the expected value, and 40% to the result of Accuracy. After conducting the normalization for both indexes and comparing the final result, the 'winning' model was Naive Bayes.

| | Accuracy | Normalized-Accuracy | Expected Profit | Normalized-EP | Weighted Sum |
|---|---|---|---|---|---|
| Logistic Regression | 0.8214 | 0.4789 | 18540 | -1.0503 | -0.43862 |
| KNN | 0.8302 | 0.5243 | 30555 | -0.1801 | 0.10166 |
| Naïve Bayes | 0.4375 | -1.4997 | 51810 | 1.3592 | 0.21564 |
| Random Forest | 0.8248 | 0.4964 | 31265 | -0.1287403 | 0.12131582 |

Figure 8: Evaluation and Comparison of the models

### 3.3.2. Best Model

Based on Figure 8, Naïve Bayes has the highest weighted sum and is considered the best prediction model, despite having a relatively low accuracy rate. The confusion matrix in Figure 7 shows that Naive Bayes performs well in predicting delays (1), but its accuracy decreases when predicting non-delays (0). Additionally, the false negative probability is high, indicating that the model predicts a delay when there is actually none. However, in this scenario, insurance companies can still benefit from not having to pay out for a delay, which explains why Naive Bayes has the highest profits. The same applies to false positive predictions, where the rate of predicted non-delays and actual delays is low, helping companies reduce the cost of insurance premiums.

## 4. Model Guidance and Business Recommendations

### 4.1 How to use the model and Business Recommendations

To use the model, insurance companies will need to input 13 variables. The output of the model will be a value between 0 and 1, and a threshold of 0.5 will be used to classify flights as delayed or not delayed. To obtain the input values for the model, the airline and airport flow indicators such as Average Passengers for the departing airport for the month can be estimated using data from previous years. For weather-related variables, estimation data can be derived from weather prediction models.

Our client, Allianz Travel Insurance, can adjust their insurance policies based on the forecast results. If a delay is predicted, the insurance price can be increased. Conversely, if no delays are predicted, the strategy can be adjusted to lower the insurance price to attract more passengers and maximize profits, as shown in the cost and benefit matrix and the expected return in 3.3.1.

### 4.2 Limitation and Improvement in the model

The model we developed is a simplified prediction model that has its limitations compared to real-world business scenarios. One of the limitations is that we may not have access to all the necessary data and there may be omitted variables that could influence the flight delay. Moreover, we only considered the input variables from the departing airport but did not take into account the arrival airport, which could potentially impact the prediction accuracy. To improve the accuracy of the model, more detailed data should be included.

Another limitation is that the data used in our model is from 2019, and since then, the pandemic has significantly changed the way we travel and there are more uncontrollable factors that may influence the flight. Therefore, insurers should focus on creating personalized protection, using individual variables that give confidence and comfort to travelers.

Additionally, the use of R studio has limitations when it comes to precisely adjusting parameters for each model. One of the reasons for this is that testing and running a model and obtaining predictions can take a considerable amount of time in R. This inefficiency can have a negative impact on companies when they put the model into real business and rely on it to predict.

The final limitation of our study pertains to the selection of the best prediction model. Since we had limited information and input variables, the RMSE and accuracy rate of the four models were found to be quite close. We also made several assumptions when calculating expected profits. However, these assumptions may not hold in the real business environment, and as such, it would be necessary to obtain more information from insurance companies to accurately calculate profits and select the best model.

Reference

Luo, J., Qiao, H. and Zhang, B., (2021) 'A Minimax Probability Machine for Non Decomposable Performance Measures'. *IEEE transaction on neural networks and learning systems*, PP, pp.1–13.