**Flight Delay**

CISC 372 - Advanced Data Analytics

Runze Yi 20073329

Chengxu Li 20073324

**Abstract**

The primary goal of this project is to predict flight delays caused by some factors. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. To carry out the predictive analysis, which encompasses a range of statistical techniques from exploratory data analysis, supervised machine learning and, data mining, that studies current and historical data to make predictions or just analyze about the future delays, with help of Decision Tree Classifier using Python 3. This study will be helpful for giving a detailed analysis on the performance of individual airlines, and help passengers to have insight on days, months, hour of departure, distance of destination and airline companies that may influence their decision on booking a flight.

**Introduction**

In simple terms, data mining is characterized as a procedure used to extract useful information from a large set of any raw data. It suggests breaking down large data into abstract information by utilizing mathematics, statistics and computer science. Data mining has applications in different fields, similar to science, medicine, finance, business and research. As a utilization of data mining, organizations can get familiar with their users and grow progressively viable techniques identified with different business capacities and thus influence decision making. This causes organizations be nearer to their target and settle on better decisions. Date mining includes powerful data collection and storing as well as computer processing.

Data mining is a branch of science that combines mathematics, statistics, computer science and intuition to dig into large data sets with the aim of discovery knowledge. For example; The CEO of Amazon (Jeff Bezos) may be interested in doing a reward package for loyal customers in his database and a promotional package for semi or disloyal customers in his database. To have insight into how the customers in his database are distributed, he will need to analyze the data. Because of the volume of the customers in his database, human approach will be impossible. This where data mining helps. With the help of data mining, he can easily dig into the large data, derive insights on the distribution of his customers and perform a customer segmentation analysis on the data. (customer segmentation is the division of customers into different segments based on some underlying pattern and similarities).

Data Mining is being used worldwide by many sectors to discover knowledge and make inferences. In the Medical sector, data mining is being used to discover pattern in cancer patients and infer on who may survive or not survive after a clinical trial. In Credit Industry, Data Mining is used to discover borrowers that may default on their credit based on similar patters with borrowers that have defaulted in the past.

In this paper, data mining will be used to provide insight and make inference on flight delay. For example; a traveller may want to know which carrier have less probability of delaying their flight. Another traveller may be interested in known in which day in a week, which day of the month is good for travelling in order to avoid his/her flight being delayed.

Knowledge discover from this study will help travellers have insight into factors that have influence on flight being delayed. The study will help travellers have insights on which carrier have the highest record of delay, which days, month, etc. does delay occur the most. The paper will try as much as possible to give predict the likelihood of a flight being cancelled using the available information. Although, information like reason for delay, weather of the day, etc are not provided in the data set being used. However, the study will give detailed insight on the problem using the available information.

**Problem Statement**

Flight delay has been a bottleneck for many travellers and flight carriers. Most times, the reason for delaying a flight is beyond this carrier.

In 2013, it was estimated that about 36% of flights were delayed by more than five minutes in Europe, 32% of flights were delayed by more than 15 minutes in the US, and 16% of flights were cancelled or delayed greater than 30-40 minutes in Brazil. Therefore, this indicates how important this indicator is and how it acts no matter how wide the scale of airline is. Hence, the insight and prediction model retrieved from this project can contribute in the form of a prototype in helping travellers to identify operational variables that contribute to delays in any airline company. To provide solution to the stated problem, the following questions will be asked and answered;

## How is the delay of flights distributed?

## How is flight delay distributed among the days of the week

## How is flight delay distributed over period of the day?

## How are delayed flights distributed over the months?

## How are delayed flights distributed over the Unique Carriers?

## What is the monthly average hour of the delayed flights?

## what is the hourly average distance of destination delayed flights?

## What is the daily average hour of the delayed flights?

* + How accurately will the features predict whether a flight gets delayed for more than 15mins or not

**Proposed Method**

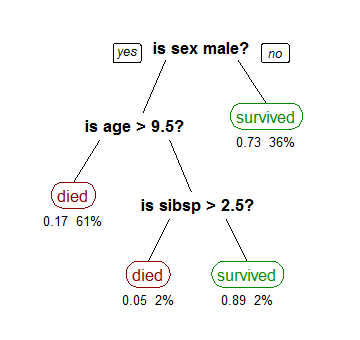
The data mining methods are exploratory data analysis (EDA) and Decision Tree. EDA will be used to explore the underlying knowledge in the data. It will be used to answer the above questions. The aspect of EDA that will be used is data visualization. Visualization techniques such as histogram, bar plot and line graph will be used. Decision Tree will be used to build a predictive model for predicting whether a flight will be delayed or not based on some factors. Decision Tree is an algorithm under the supervise machine learning models.

**Supervised machine learning model**

Supervised learning is the machine learning task of learning a function that maps some input(s) to an output based on example pairs of the input and output features. The inputs are often called features while the outputs are called targets. The inputs are used to learn a pattern in the data and are mapped to the outputs. In other hand, unsupervised machine learning algorithms are those that have only inputs. The features are then used to search for patterns, the patterns are then used to group the features together into different clusters based on similarities in patterns. Some of this algorithm include K-Means, Hierarchical clustering, etc.

**Decision Tree**

In data mining, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. It is a commonly used tool in data mining for deriving a model. A decision tree is drawn upside down with its root at the top. In the image below, if we are predicting whether a cancer patient dies after a clinical trial. The bold text in black represents a condition/internal node, based on which the tree splits into branches. The end of the branch that doesn’t split anymore is the decision/leaf, in this case, whether the cancer patient died or survived after a clinical trial, represented as red and green text respectively.



**How to Build a Decision Tree**

*Step 1*

Determine the root of the **Tree**: Since decision trees are commonly used for classification, you need to determine the target classes which are the basis for the decision.

*Step 2*

for every attribute/feature:

* calculate entropy for all categorical value
* take average information entropy for the current attribute
* calculate gain for the current attribute

*Step 3*

pick the highest gain attribute.

*Step 4*

Repeat until we get the tree we desired.

**EXPERIMENTAL RESULT**

**Data Description**

The data used for this study is collected from kaggle and can be found in <https://www.kaggle.com/c/flight-delays-spring-2018/data>

1. Month, Day of Month, Day Of Week
2. Dep-Time – departure time
3. UniqueCarrier – code of a company-career
4. Origin – flight origin
5. Dest – flight destination
6. Distance, distance between Origin and Dest airports
7. dep\_delayed\_15min – targe

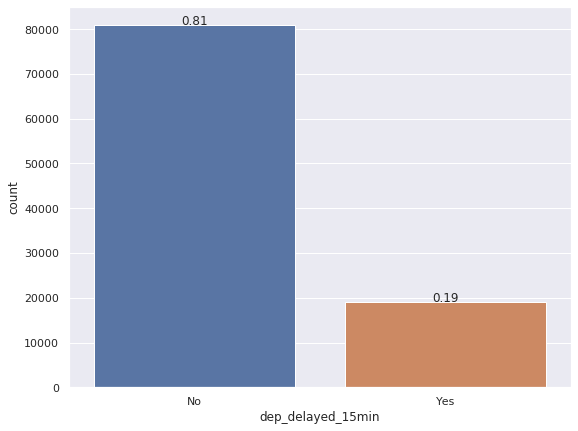
**Data Summary**

The table below shows that there are 100000 observations of flights recorded which also implies that there are no missing observations. It also shows the descriptive statistics of the numerical features. It shows that the average hour of all flights is around the 13th hour of the day (1300)

|  | count |  | mean |  | std |  | min | 25% |  | 50% | 75% | max |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day of Month | 100000.0 |  | 15.70385 |  | 8.793931 |  | 1.0 | 8.0 |  | 16.0 | 23.0 | 31.0 |
| Dep Time | 100000.0 |  | 1341.52388 |  | 476.378445 |  | 1.0 | 931.0 | 1330.0 |  | 1733.0 | 2534.0 |
| Distance | 100000.0 |  | 729.39716 |  | 574.616860 |  | 30.0 | 317.0 |  |  | 957.0 | 4962.0 |
| Hour | 100000.0 |  | 13.11958 |  | 4.771657 |  | 0.0 | 9.0 | 13.0 | 575.0 | 17.0 | 25.0 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

**Exploratory Data Analysis (EDA)**

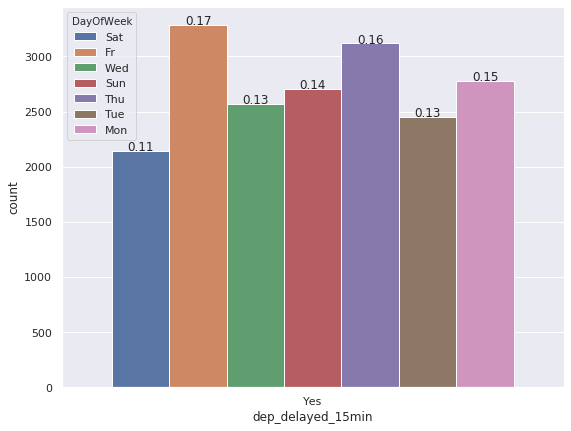
## How is the delay of flights distributed

**

The bar plot shows that count of the delayed and non-delayed flights. It shows that that about 19% of the total flights were delayed for more than 15mins

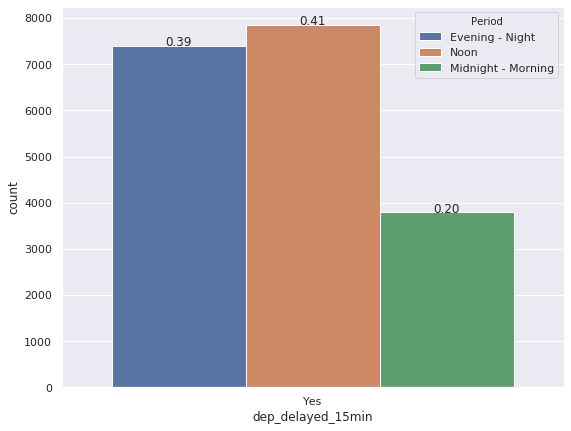
## 

## *How is flight delay distributed among the days of the week*



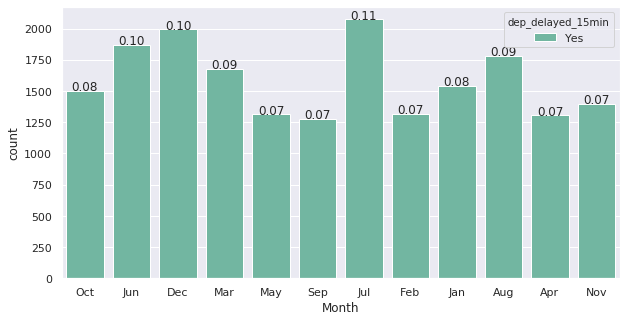
The figure above shows how the delayed flights were distributed across the days of the week. The figure shows that Flights on Fridays has the highest proportion of delay then Thursdays. This implies that flights on the later part of working days tend to get delayed for more than 15 mins compared to other days

***How is flight delay distributed over period of the day***

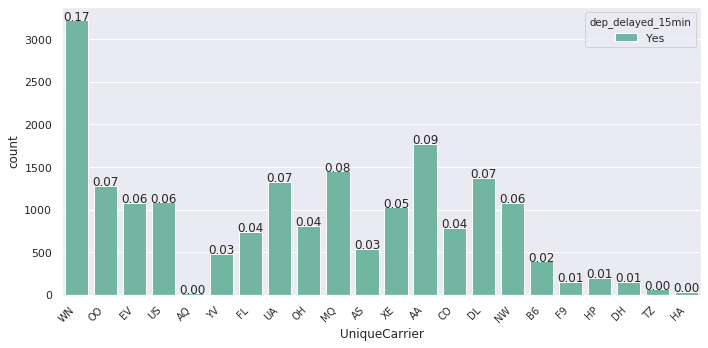


The figure above explores how the delayed flights were spread across the periods of a day. It showed that the delayed flights were common at noon (between 1200 and 1700). It further shows that flights in the earliest part of the day have lower proportion of delay (0000 and 1100).

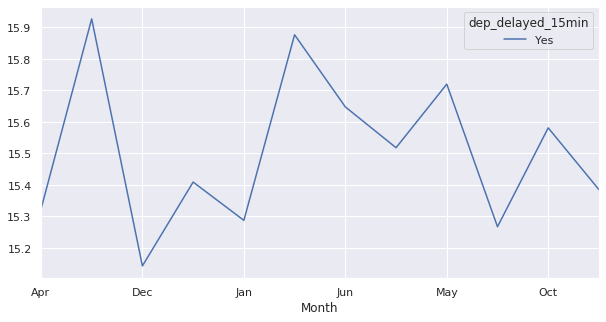
## *How are delayed flights distributed over the months*

The plot above shows how the delayed flights are spread over each months of the year. It showed that the month with the highest delay was July while September has the least delayed flights. This implies that the flights tend to get delayed more from the 6th month to the 8th month of the year

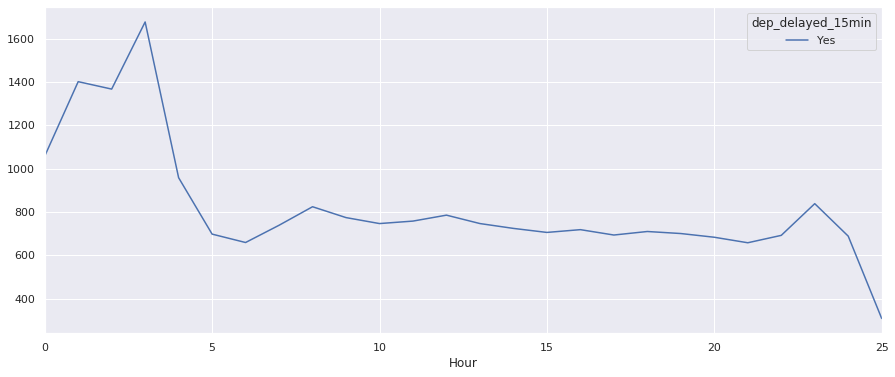
## *How are delayed flights distributed over the Unique Carriers*

The above plot shows how the delayed flights were distributed among the airline companies using their code. It showed that WN (Southwest airlines) has the highest proportion of delayed flights which are over 15mins with 17% of the total delayed flights coming from them.

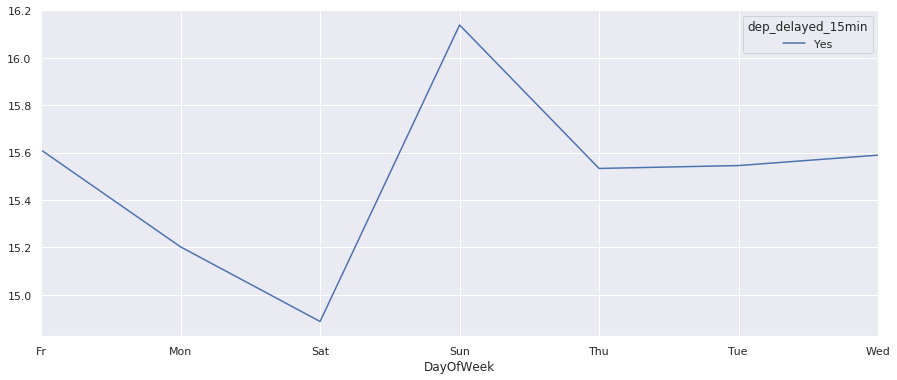
## *What is the monthly average hour of the delayed flights*

The plot above shows the monthly average hour of departure of the delayed flights. Given that the months are arranged alphabetically, it shows that the average hour for the delayed flights in each month falls between the 15th hour and the 16th hour (1500 and 1600). This corroborates with our discovery that most of the delayed flights occurred at noon.

## *what is the hourly average distance of destination delayed flights*

The above figure shows the average distance of destination of the delayed flights based on their hour of departure. It shows that the highest average distance of the delayed flight falls between 00:00 and 05:00 hour. This implies that flights of long distance in the midnight tends to get delayed for more than 15mins.

## *What is the daily average hour of the delayed flights*

 The above plot shows the average hour of departure of the delayed flights in each day. It shows that the average flights for each day falls between 14:30 and 16:30 hours of each day. This discovery supports our above discovery that most flights get delayed at noon.

***How accurately will the features predict whether a flight gets delayed for more than 15mins or not***

In order to evaluate the model created using the Decision Tree, we will split the dataset into two parts (Training set and Test set). The training set will be used to fit the model while the testing set will be used test the accuracy of the fitted model. This is to justify how well the fitted model makes prediction for new features. The table below shows the confusion matrix of the predictions made on the test set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Observed** | | |
| Classes | No | Yes | Total |
| **Predicted** | No | 26502 | 221 | 26723 |
| Yes | 5939 | 338 | 6277 |
| Total | 32441 | 559 | 33000 |

The above table shows that of the observed 32441 non-delayed flights, 26502 were correctly classified while 5939 were wrongly classified. Also, of the observed 559 delayed flights, 221 were wrongly classified while 338 were correctly classified. The below table shows the explicit report of the model.

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Precision | Recall | f1-score |
| No | 0.82 | 0.99 | 0.90 |
| Yes | 0.60 | 0.05 | 0.10 |
| accuracy |  |  | 0.81 |

The table above shows that the model has 0.82 probability of predicting non-delay correctly and 0.60 probability of predicting delay correctly. It further shows that the model has an overall accuracy of 0.81.

**Conclusion**

The study examined how accurately features like month, day, hour of departure respectively and distance of destination predict the likelihood of a flight being delayed. The data used for the study was collected from Kaggle. Some questions about the distribution of the features with respect to the target variable were asked.

EDA was used to answer most of the questions, while Decision Tree Classifier was used to examine how accurately the features predicts the target variable. The EDA shows that 19% of the total observation were delayed for more than 15mins. It also showed that flights are likely to be delayed for more than 15mins on fridays than other days. Furthermore, it showed that flights at Noons (12:00 – 17:00) have more likelihood to be delayed. It further shows that flights in June – August were mostly delayed and Southwest Airlines has the highest proportion of delay. The analysis also shows that flights with average high distance and falls between 00:00 and 05:00 have high likelihood of being delayed.

Using the Decision tree to build a classifier which tried to classify the target variable using the features. The model had an overall accuracy of 0.81 while it has a precision of 0.82 for non-delayed flights and a precision of 0.60 for delayed flights. The recall which is the measure of how many truly relevant results are returned especially when classes are imbalanced over the observations was shown to be 0.99 for non-delayed classes and 0.05 for delayed class. The model seems accurate overall and predict non-delay flights well. However, it may not accurately predict the non-delay well. A good solution to this problem would be to increase the amount of flights that were delayed.