**CAS CS505 Mid-progress Report**

Title: Prediction of Flight’s Departure Performance

Tian Zhang, Lu Min

# Abstract

This project is mainly for predicting delay status of flights from two states’ airports, which are New York and New Jersey. At the meantime, the model would give a flight with less probability to delay with the same destination and departure time. Also, this project would analyze factors leading to delay and then obtain dominant factors.

# Introduction

Every year there are over 20% airline flights delayed or cancelled around the world. Everyone who is about to take a flight doesn’t want flights to be late, especially business people. Our purpose is to predict a flight which would have less probability to delay. Our model would give a record with a specific date, destination and other input variables, and this record would contain a unique flight record as a combination of a flight number and aircraft tail number.

To conduct our analysis and set up our model, we obtain the public-available dataset for major U.S. airports from website of Bureau of Transportation Statistics (BTS). This dataset contain 29 features of the flights information, and we may fully use about 9 features during our project. And this dataset also has details of factors due to delay or cancellation.

About approaches we use in this project, first, we build a Naïve Bayes Classifier for calculating metrics for future visualization and analysis. Second, we would apply our own logistic regression model by maximum likelihood estimates (MLE) to fit in our data, and figure out the probability we need by input variables. Third, we would use Apache Hydoop to perform various types of data pre-processing and feature engineering tasks, and then scikit-learn to add new and improved features on the resulting datasets. The last approach is Apache Spark, and we would use it to generate our feature matrix and also use ML-lib (Spark’s machine learning library).

Our predicted results are following: First, we expect to predict the delay status of flights from different airport with the same destination, departure date and other input variables. Second, we would analyze the factors which attribute to the delay, and figure out the dominant factors in which season or which quarter of a year. Finally, we can obtain a record with the lowest probability to delay, and each record contains a unique flight record as a combination of a flight number and aircraft tail number. Also, we would compare the models we built and analyze which one is more accurate and suitable based on the results.

# Technique (that will be) used in this project

## Naïve Bayes Classifier

We choose Naïve Bayes Classifier, because Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Also, it is important for our future visualization and analysis. The Naïve Bayes Classifier was implemented using scikit-learn, and in our project’s situation, Bayes theorem states:

By applying the model, we would like to figure out the probability of delay given a tuple of attributes. Python’s scikit-learn provides a fairly simple function for training, testing, and assessing the results of the model. Because the large size of our dataset, we would use a sample of our data.

## Logistic Regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. And we would use our own logistic regression model as:

The reason that we choose logistic regression model is that this model can iterate over the training data and get the global solution.

## Apache Hadoop

We would use a supervised learning model to predict airlines delays, and implement related variables: month, day of month, day of week, hour of the day, carrier, destination airport and distance.

//? Hadoop splits files into large blocks and distributes them across nodes in a cluster.

## Apache Spark

//? We have used Apache Spark on our HDP cluster to perform various types of data pre-processing and feature engineering tasks. We then applied a few ML-Lib machine learning algorithms such as Support Vector Machines and Decision Tree to the resulting datasets and showed how through iterations we continuously add new and improved features resulting in better model performance.

# Datasets and experiments

We collect dataset from the official website of Bureau of Transportation Statistics (BTS) which offer us publicly-available data on flight arrivals and departures for major U.S. airports. The downloaded dataset contains 29 variables that describe each flight in terms of departure/arrival date and time, carrier, taxi time, time spent in the air, as well as departure and arrival delays and their causes.

Here is a form of our dataset description with 29 features.

|  |  |  |
| --- | --- | --- |
|  | Feature | Description |
| 1 | YEAR | 2006-2016 |
| 2 | MONTH | 1-12 |
| 3 | DAY\_OF\_MONTH | 1-31 |
| 4 | DAY\_OF\_WEEK | 1 (Monday) - 7 (Sunday) |
| 5 | DEP\_TIME | actual departure time (local, hhmm) |
| 6 | CRS\_DEP\_TIME | scheduled departure time (local, hhmm) |
| 7 | ARR\_TIME | actual arrival time (local, hhmm) |
| 8 | CRS\_ARR\_TIME | scheduled arrival time (local, hhmm) |
| 9 | UNIQUE\_CARRIER | unique carrier code |
| 10 | FL\_NUM | flight number |
| 11 | TAIL\_NUM | plane tail number |
| 12 | ACTUAL\_ELAPSED\_TIME | in minutes |
| 13 | CRS\_ELAPSED\_TIME | in minutes |
| 14 | AIR\_TIME | in minutes |
| 15 | ARR\_DELAY | arrival delay, in minutes |
| 16 | DEP\_DELAY | departure delay, in minutes |
| 17 | ORIGIN | origin |
| 18 | DEST | destination |
| 19 | DISTANCE | in miles |
| 20 | TAXI\_IN | taxi in time, in minutes |
| 21 | TAXI\_OUT | taxi out time, in minutes |
| 22 | CANCELLED | if the flight is cancelled or not |
| 23 | CANCELLATION\_CODE | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | DIVERTED | 1 = yes, 0 = no |
| 25 | CARRIER\_DELAY | in minutes |
| 26 | WEATHER\_DEALY | in minutes |
| 27 | NAS\_DELAY | in minutes |
| 28 | SECURITY\_DELAY | in minutes |
| 29 | LATE\_AIRCRAFT\_DELAY | in minutes |

We build a simple Naïve Bayes Classifier based on sampling dataset, predicts labels for a subset of data. We use part of our data from year 2006 to year 2010. During processing, we choose 75% data for training and 25% data for testing. And the classifier also calculates metrics such as precision/recall and accuracy after classification. We convert the output is to pickle files which would be used later for visualization and analysis.

After getting preprocessed raw data, we transfer the data in dataframe into appropriate datatypes. And giving each data record an extra label: Define the departure of flight is late for more than 15 minutes as delay flight. Using '1’ to present ‘Delay’ and ‘0’ to present not ‘Delay’.

# Results and Discussion

# Conclusion