**Predictive Analysis of Flight Delays**

**Venkat Sai Manoj Cheerla**

**Lakehead University**

**vcheerla@lakeheadu.ca**

**Abstract**

Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all commercial airline companies. Moreover, the development of accurate prediction models for flight delays became tough due to the complexity of air transportation system and the deluge of flight data. In this context, this paper presents a brief literature review of approaches used to build flight delay prediction models from the Data Science perspective. This project proposes to implement logistic regression for the predictive analysis and giving particular attention to usage of spark tools for data processing and also machine learning tools in Anaconda (3.5) platform using python 2.7.

**Keywords**: Airline Delay, machine learning, logistic regression, Python

1. **Introduction**
   1. **What we know so far about the problem?**
      1. What is the problem?

With the rapid development of the national economy, the demand for air transport has increased dramatically. The flight delay has become more and more serious, which directly causes serious damage to the image of civil aviation services. For passengers, flight delay caused the inconvenience of travel, bad mood, as well as the double loss of time and economy; for the airport, the delay of the flight seriously affects the normal operation of the airport; for airline, frequent flight delay not only bring huge economic losses to the airline, but also affect the reputation of the airline. Flight delay has become the shackles of the development of the aviation industry.

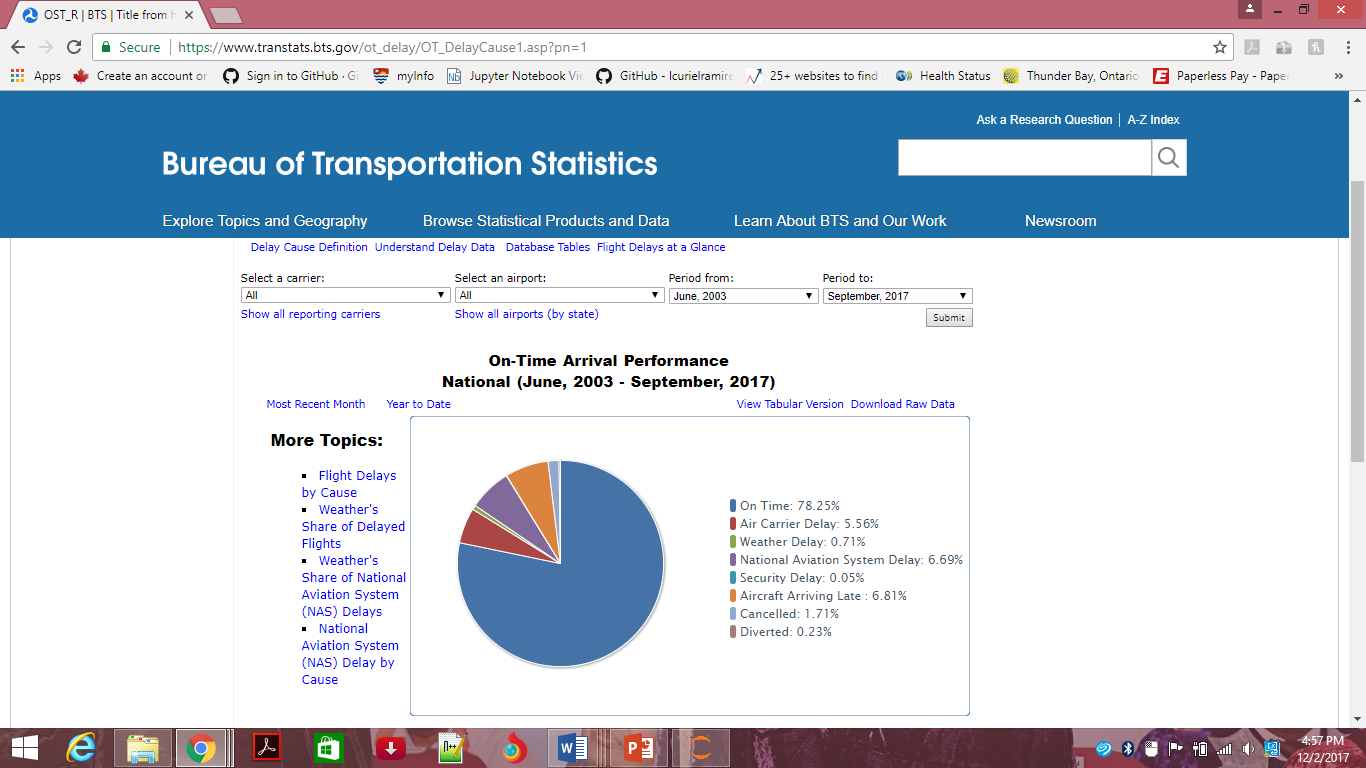


Figure 1 : On time arrival performance from Bureau of transportation statistics

* + 1. Why is it important?

Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip costs, to ensure their arrival on time. On the other hand, airlines suffer penalties, fines and additional operation costs, such as crew and aircrafts retentions in airports. Furthermore, from the sustainability point of view, delays may also cause environmental damage by increasing fuel consumption and gas emissions. If we could timely forecast the flight delay, we can take necessary measures to reduce the economic and credit losses due to it. Obviously, it is very important and necessary to predict the flight delay in real time.

* + 1. Are the current solutions not good enough? Why do we need your solution? What does your solution bring to the table?

At the end of the introduction section you must sate the organization of your paper “what are the next sections ad what do they contain?”

Following this introductory section, section 2 gives a background and description of flight delay along with a literature review of delay models, simulation methods and the statistical techniques used in the thesis. Section 3 provides descriptions of the data sources and definitions of the data used to calibrate the statistical models of the thesis along with the methods used and experiments performed. Section 4 consists of results and discussions which include the screenshot of the output obtained followed by section 5 and section 6 which provide the conclusion and references respectively.

1. **Literature Review**
   1. **Concise, critical, and chronological discussion of the related works**
      1. What are the related/similar previous works to your work?

Works that study ﬂight systems are increasing the usage of machine learning methods. The methods commonly used include k-Nearest Neighbor, neural networks, SVM, fuzzy logic, and random forests. They were mainly used for classiﬁcation and prediction.

* + 1. Clearly state the contribution of each work “what did they do to solve the problem that you investigated?”

Rebollo [[R1]](https://pdfs.semanticscholar.org/a0da/189ab8222a0d9b7178690dd116c5520b76a9.pdf) applied random forests to predict root delay. They compared their approach with regression models to predict root delay in airports of the United States considering time horizons of 2, 4, 6 and 24 hours. Their test errors grew as the forecast horizon increased. This paper presented new network-based air traffic delay prediction models that incorporated both temporal and network delay states as explanatory variables. The results obtained for the 100 most-delayed OD pairs in the NAS showed an average test error of 19% when classifying delays as above or below 60 minutes, for a 2-hour forecast horizon.

Khanmohammadi [[R2]](http://www.sciencedirect.com/science/article/pii/S1877050916324942) created an adaptive network based on fuzzy inference system to predict root delay. The predictions were used as an input for a fuzzy decision-making method to sequence arrivals at JFK International Airport in New York. One of the limitations of this study is the complexity of the proposed method (as the number of variables increases the number of connections also significantly increase). They need to consider the integration of the proposed method with fuzzy logic to expand the real-world applications of the proposed method.

Lu Whang Zhang [[R3]](http://ieeexplore.ieee.org/document/4732894/) built a recommendation system to forecast delays at some airports due to propagation eﬀects. The prediction was based on the k-Nearest Neighbor algorithm and used historical data to recognize similar situations in the past.

The methods performed by them are mostly very complex and require a variety of concepts to perform the statistical analysis. The Method which we have implemented in this paper is quite simple and requires only a basic knowledge of machine learning tools in python.

1. **Material and Methods**
   1. **Data**
      1. The dataset for this problem was obtained from the Bureau of Transportation Statistics which consists of all commercial flight operations from the year 1987 to 2008. The dataset consists of the following flight features:

• YEAR: Year of Flight Departure/Arrival

• CRS\_DEP\_TIME: Flight Departure time in Hours

• MONTH: Month of Flight Departure/Arrival (5-May and 12- December)

• DAY\_OF\_WEEK: Day of Week (1-7)

• CARRIER: Code assigned by IATA and commonly used to identify a carrier.

• ORIGIN\_CITY\_NAME: Origin City of flight

• DEST\_CITY\_NAME: Destination City of flight

• ARR\_DELAY\_NEW: Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.

• ARR\_DEL15: Arrival Delay Indicator, 15 Minutes or More (1=Yes)

• DISTANCE: Distance between origin and destination in miles.

• CARRIER\_DELAY: Carrier Delay, in Minutes

• WEATHER\_DELAY: Weather Delay, in Minutes

• NAS\_DELAY: National Air System Delay, in Minutes

• SECURITY\_DELAY: Security Delay, in Minutes

• LATE\_AIRCRAFT\_DELAY: Late Aircraft Delay, in Minutes

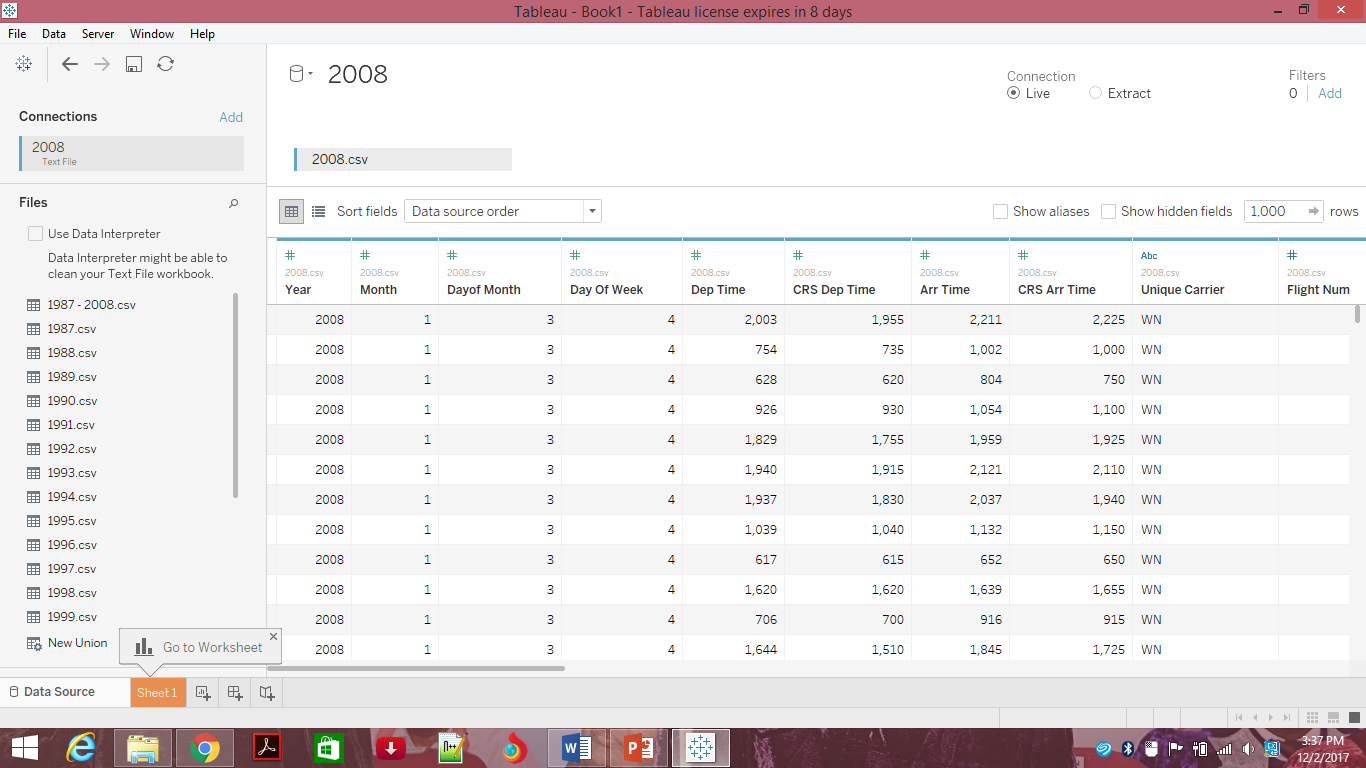


Figure 1 : An overview of the csv data set in tableau

* + 1. Data preprocessing

One of the main steps in the predictive analytics is data transformation. Data is never in the way you want them. One might have to do some kind of transformations to get it to the way we need them either because the data is dirty, not of the type we want, out of bounds, and a host of other reasons. This first transformation we’ll need to do is to convert the categorical variables into dummy variables.

We now create a SQL Data Frame, this entity is a distributed collection of data organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in Python, but with richer optimizations under the hood. We will utilize the recently created Spark RDD and use the Spark SQL context to create the desired data frame, We first create function that will allow to parse a record of our RDD into the desired format.

We add a new column to our data frame, **DepDelayed**, a binary variable:

* **True**, for flights that have > 15 minutes of delay
* **False**, for flights that have <= 15 minutes of delay

We will later use **Depdelayed** as the target/label column in the classification process.

We also add a new column, **Hour**, to determine the hour of flight (0 to 24). For this purpose we first define the following auxiliary function. Once created we will register it as a **user defined function (UDF).** This is useful when adding functions into the SparkSQL language.

Let's do some exploration of this dataset. Let's start by taking a look at airports that have the most delays.

* 1. **Methods and Experiments**
     1. What are the hypotheses behind your solution?

Now, it’s generally NOT a good idea to use your ENTIRE data sample to fit the model. What we want to do is to train the model on a sample of the data. Then we’ll see how it perform outside of our training sample. This breaking up of our data set to training and test set is to evaluate the performance of our models with unseen data. Using the entire data set to build a model then using the entire data set to evaluate how good a model does is a bit of cheating or careless analytics.

* + 1. What are the facts and/or established hypotheses that you are going to use to test your hypotheses?

We are first going to observe the characteristics of each of the variable and find out which one gives major contribution to flight delay.

* + 1. What are the different methods that you are going to use for the implementation of this work? What are special and/or unique about the methods and how are they going to help you develop your solution?

Logistic regression was chosen to model flight delay for multiple reasons. First, the weights of each feature trained by logistic regression are easily interpretable, as the sign of the weight indicates if a flight is more or less likely to be delayed if it has a high value for that feature. Second, logistic regression outputs a measure of confidence in its output through the probability of belonging to each class.

Using the logistic regression model above, we did not have any issues of overfitting, as judged by the roughly equivalent accuracy on the training and testing sets. Therefore, there was no need to optimize the regularization of the model. Additionally, there were no issues of model scalability, as training the model with all features using the entire training set took at most a few minutes on a laptop.

1. **Results and Discussion**

After the cleaning of the data (adding of additional column and removal of unwanted data) the next process of data analytics life cycle is the preprocessing of the data that is dividing of the data into testing and training datasets.

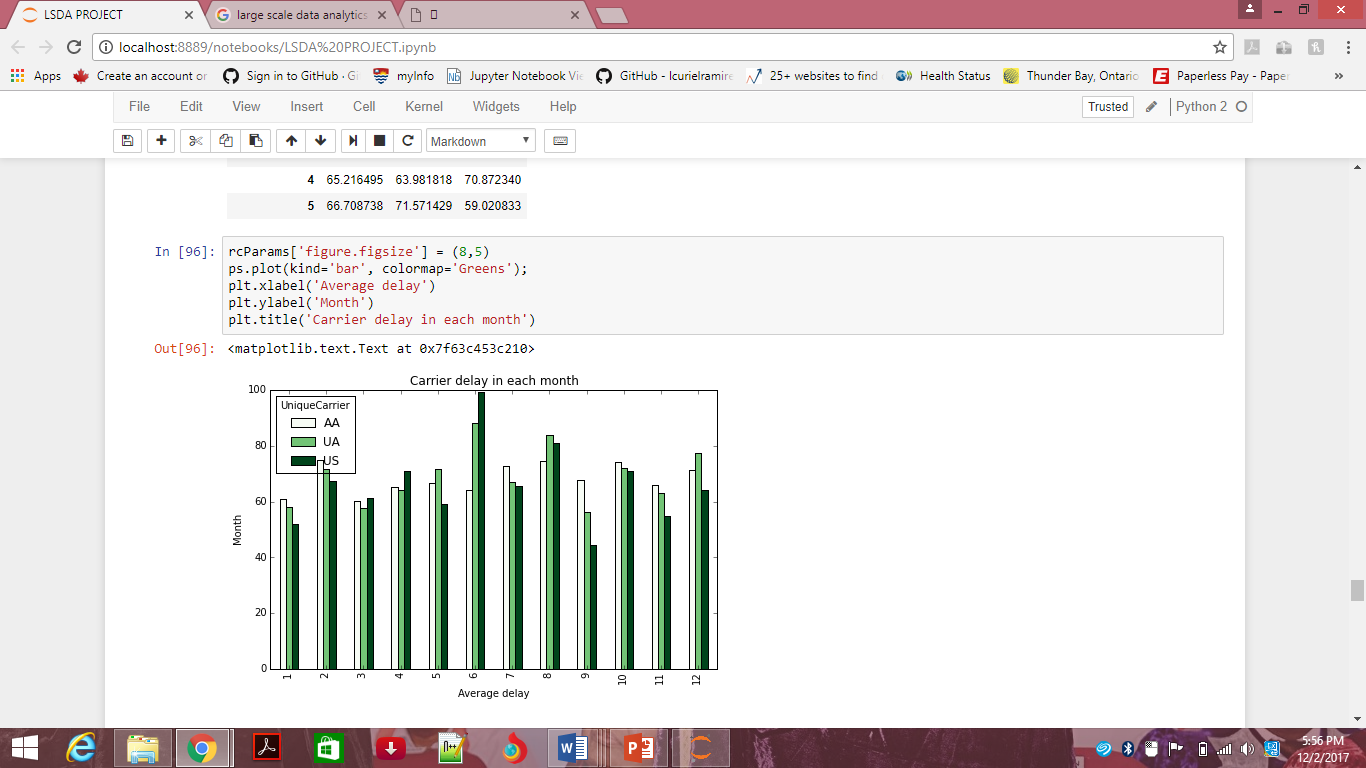


Figure 1 : Carrier delay in each month(Average delay vs month)

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

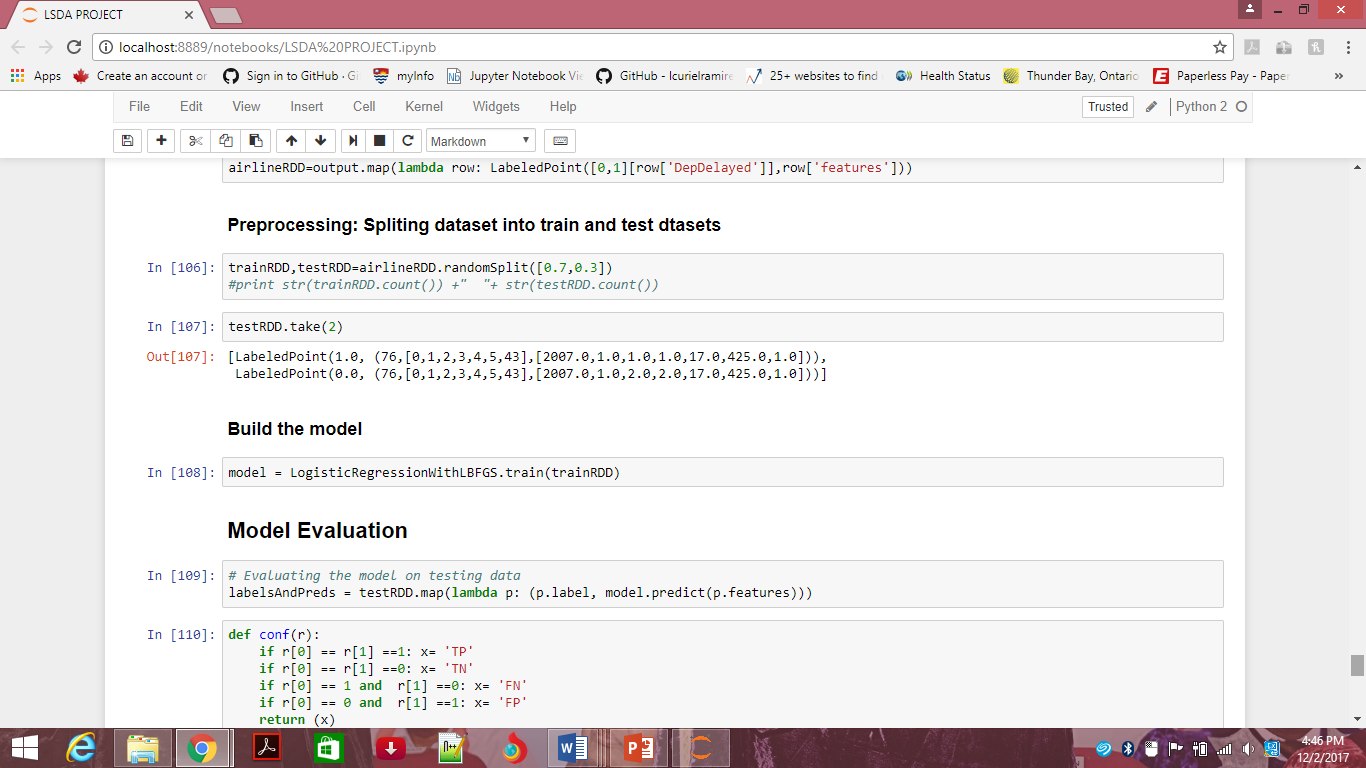


Figure 1 : Code snippet involving Preprocessing and model building stages of the analytic life cycle

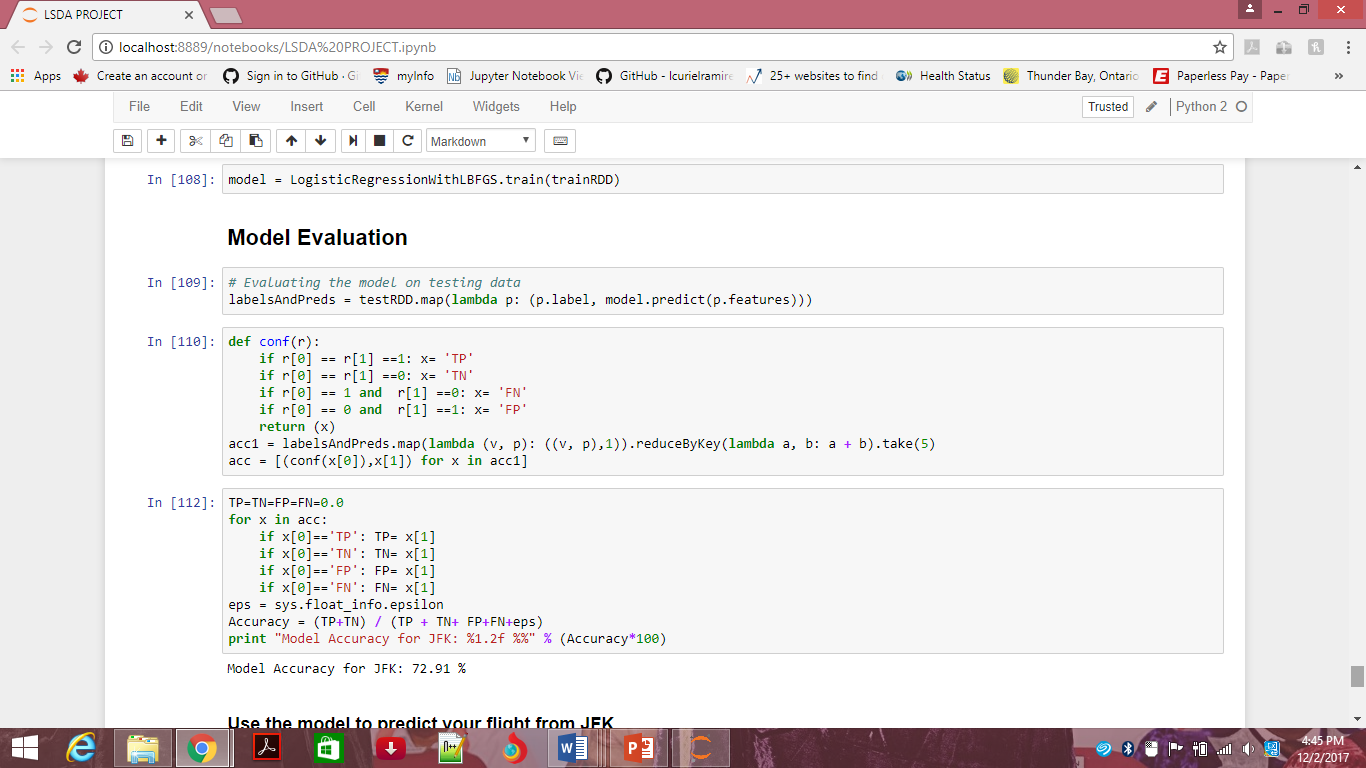


Figure 1 : Model Evaluation(this helps us know the accuracy of the model)

In the figure shown below we have given a test case and predicted that the flight doesn’t have a delay.The finally derived model always consists of only the relevant variable and their weights in logistic regression.

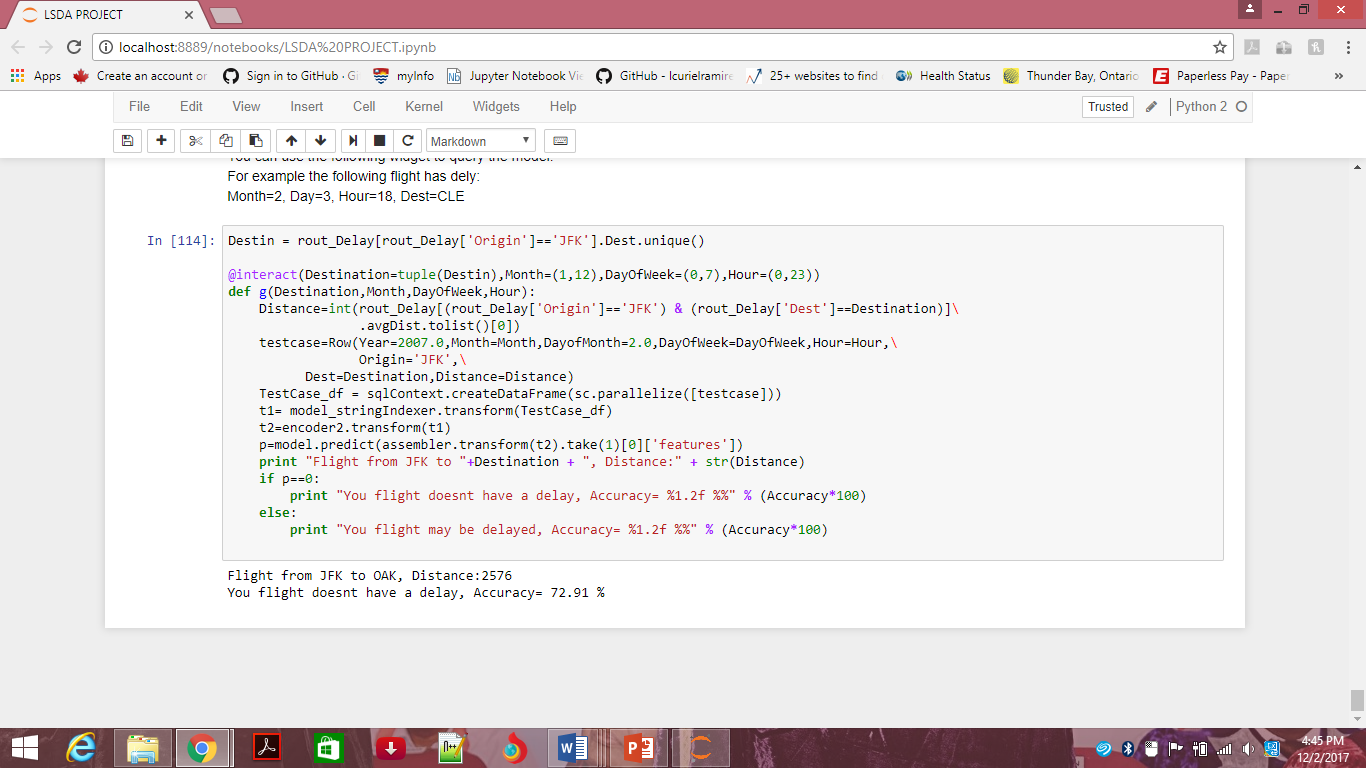


Figure 1 : Final output with the measure of accuracy

From the results shown above it can be observed that we could obtain an accuracy of 72% approximately by using Logistic regression. To increase the accuracy of the the system we need to increase the volume of the data. Due to technical constraints I could

Use only flight data belonging to the year 2008.But the data is available from the year 1987-2008.In order to combine all these data sets we can make use of Azure. I would like to implement it in my future work.

1. **Conclusion, Limitation, and Future work**
   1. Briefly, summarize the paper and state clearly “what the paper adds?” this is your contribution, be careful! Your contribution must be something beyond using existing blocks, simple lab experiments, restating facts, and/or an effort leading to obvious conclusions. It should be something concise, rational, realistic, and solid with interesting conclusion that adds to our understanding of the problem.

After the development of the module we have come to a conclusion that the model developed can be used in predicting the delay at the airports. The delay intensity of an airport and airport route can make it easier to understand the airport delay. The results of the research show that the delay is highly related to the originate delay. The model developed can be applied to predict occurrence of delay at airports. Such predictive capabilities can help the aviation agencies to prepare mitigation strategies for reducing delay. The models are calibrated using historical data and is hypothesized using python 2.7 in anaconda.

* 1. Interesting usage scenarios

Travelers could benefit from knowing the likelihood of a delay as it could help them prepare for the wait time. An app can be developed that predicts the impact of weather on airplane departure times.

* 1. What are the limitation of your results compared to others?

Overall, this model is only of limited utility since it wasn’t capable of correctly predicting flight delays with both precision. This seemingly low performance is likely due to the many causes of flight delays being outside the scope of our data. It is unclear if it is even possible to predict whether or not a flight will be delayed so far in advance, as we have so many of the causes of delays (e.g. mechanical issues and weather) which cannot be known in advance.

* 1. The future work section should contain a feasible rational solution to the work limitation that you want to pursue afterwards.

The model can be made better by adding factors such as weather. This can be done through the API of a weather service. Also I decided not to segregate by month as the computation was already intensive with so much data to process. Segregating by months according to season will give better insights as well.

1. **References**
2. L. Zonglei, W. Jiandong, and Z. Guansheng. A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In 2008 International Symposium on Knowledge Acquisition and Modeling, pages 589–592, Dec. 2008.
3. The data, American Statistical Association, Data Expo 2009

<http://stat-computing.org/dataexpo/2009/the-data.html>

1. Flight Delays by Cause, Bureau of Transportation Statistics

<https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp>

1. Railway passenger train delay prediction via neural network model

<https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwj6paGrhezXAhVKhlQKHWPxDUcQFgg9MAE&url=https%3A%2F%2Fpdfs.semanticscholar.org%2F006b%2F264d179adf5f9eaa65a4dc1153a44b17355e.pdf&usg=AOvVaw1mEqTtgZduXkP2sggCViqY>

1. <https://github.com/chrisalbon/code_py>
2. <https://github.com/jpatokal/openflights/tree/master/data>