Flight Delay Prediction

SOE532C Course Project

Presenting By,
Nilotpal Pramanik (IRM2016501)
Shaik Rumaan (BIM2016004)
Anubhav Shrivastava (ITM2016006)
Adarsh Agarwal (IIT2016516)

Supervisors
Dr. Abhishek Vaish
Vishesh Middha (T.A)

Abstract

Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of the air transportation system, the number of methods for prediction, and the deluge of flight data. In this context, we are trying to use Weather dataset combined with Flight Dataset including data about previous flights and their delay information.

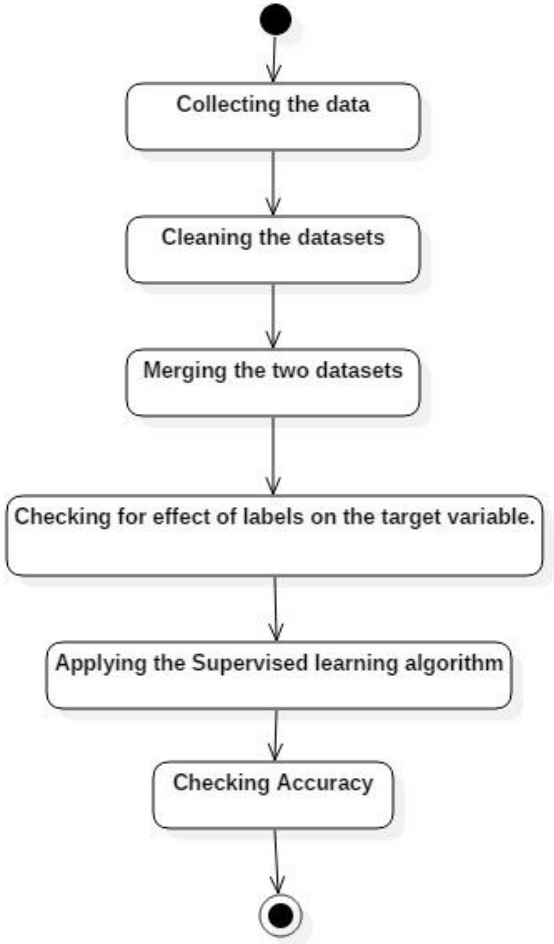
We will use the **Supervised machine learning algorithm** (Support Vector Machine in this case) to train our model and make a proper prediction.

Introduction

The main aim of this project is to create a machine learning model, which is able to forecast flight delays due to weather observations. The whole idea lies on the fact that when a user enters a flight destination, date, time, airline, origin and some weather observations, the system will respond without a time lapse with an answer that represents whether the flight entered may or not be delayed. The model is trained and tested against 2016 flight and weather records, which means that all the used data are facts and nothing was invented.

Flow Chart

Flow Chart


Collecting the data
Cleaning the datasets
Merging the two datasets
Checking for effect of labels on the target variable.
Applying the Supervised learning algorithm
Checking Accuracy

Data and Resources

Regarding the project dataset, we're dealing with the Airline data having attributes 'YEAR', 'MONTH', 'DAY_OF_MONTH', 'CARRIER', 'ORIGIN_AIRPORT_ID', 'DEST_AIRPORT_ID', 'CRS_ARR_TIME', 'ARR_DELAY', 'ARR_DEL15', 'SkyCondition'.

According to the description of the weather report, the attributes are "Visibility", "WindSpeed", "WindDirection", "StationPressure" and "DryBulbCelcius".

The dataset link::
https://github.com/goodday451999/Flight-delay-prediction-using-SVM/blob/master/dataset.csv

Methods

1. **Collecting the Data:** As finding dataset which contains information of both Flights as well as weather is extremely difficult and time consuming. Therefore for the purpose of this project we are using 2 separate datasets i.e. Airplane flights dataset and weather dataset.

2. **Cleaning the Datasets**: As publicly available datasets which contain information of a large period of time are very likely to contain large number of NAN values. This should be handled with care. For the purpose of this project we will remove some of the columns which have a very high percentage(%) of NAN values. However other NAN values are dealt by replacing it with the average of the values present in the column or by the most frequent value in the column(whichever suits for the Variable).

3. Merging the two datasets: This can be done by using “Date” information of airplane flight as well as using the weather information of that Date present in the Weather dataset. We now have both the informations available in a single dataset which is our main Database of all the information.

4. Applying the Machine Learning algorithm: Presently our aim is to apply SVM to our training dataset however we will use other algorithms as well to find the effect of other algorithms on this dataset.

Design and Implementation

In order to predict flight delay, we develop a system. The system includes the Dataset and the predictor. The Predictor here is **Support vector machine**. As we are predicting ranges thus this is a classification problem and SVMs are a very powerful tool for classification methods. The different classes are the different ranges of time by which a flight delay is predicted.

The whole algorithm is implemented in Python using Scikit-learn to use Support Vector machines and other algorithms. But before getting into the exact algorithm we need to know the importance of each labels and have some intuition of how it affects in flight performance. For this purpose we worked with the graphs of some of the labels and their effect on flight delay. The main **Python Libraries** and **APIs** used for all this are :-

1. **Scikit-learn (for models)**
2. **Numpy (for maths related functions and arrays)**
3. **Pandas (for dealing with Dataframe)**
4. **Matplotlib (for graphs)**
5. **Collections (Frequency of elements) and some others**

Conclusion

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system.

Screen Shots

Accuracy:0.788921071687

```
rumaan@rumaan-HP-Pavilion-Notebook:~/Desktop/flight_delay_prediction$ python3 flight_delay.py
/usr/local/lib/python3.6/dist-packages/sklearn/utils/fixes.py:313: FutureWarning: numpy not_equal will not check object identity in the future
. The comparison did not return the same result as suggested by the identity (`is`)) and will change.
  _nan_object_mask = _nan_object_array != _nan_object_array
/usr/local/lib/python3.6/dist-packages/sklearn/externals/joblib/externals/cloudpickle/cloudpickle.py:47: DeprecationWarning: the imp module is
 deprecated in favour of importlib; see the module's documentation for alternative uses
  import imp
sys:1: DtypeWarning: Columns (16) have mixed types. Specify dtype option on import or set low_memory=False.
0.788921071687
```

References

[1]https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial
[2]https://www.researchgate.net/publication/325034541_Airline_Delay_Predictions_using_Supervised_Machine_Learning
[3] https://github.com/AduraX/Flight-Delay-Prediction
[4]https://pdfs.semanticscholar.org/29e2/a5a6b72d6738c6feb41ee0f8a9b57f600e7d.pdf
[5]http://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198/pdf

Thank You