Software Project Report

Flight Delay Prediction

Group Members

ITM2016006 (Anubhav Shrivastava)
IIT2016516 (Adarsh Agrawal)
IRM2016501 (Nilotpal Pramanik)
BIM2016004 (Shaik Rumaan)

TABLE OF CONTENTS

1. 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.

2. 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.

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.

3. Literature Survey:-

The main papers and well as resources that we used in the making of this project include :

1. **Flight Delay Forecast due to Weather Using Data Mining by Adrian Alexander Arteche Simmons:**- The first step in the making of this project was to collect datasets. This seemed to be an easy task at first but it is not so easy. This paper helped us to learn how a proper dataset is generated and organized in a better way from 2 or more different kind of datasets. After learning how to merge different datasets we came to know that it is necessary to delete (or update) those rows/columns which contain any useless or garbage value. In our case all those columns which had NaN value, were updated by replacing them by mean value of those columns. Also all the rows which contain the information of flights which were eventually Cancelled were deleted as they would not help in the learning processes of the model.

2. **Flight Delay Prediction Using Random Forest by Rebollo:-**.

The first step is to clean the dataset means removing the Unwanted NAN values and also drop the unwanted columns which can reduce the accuracy of prediction.In this paper they have 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.

3.**Flight Delay Prediction by Lu et al :-**

They 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 authors noticed fast response time and easy, logical comprehension as the main advantages of their method.

4. A Review on Flight Delay Prediction [4]

Alice Sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara * CEFET/RJ Rio de Janeiro, Brazil March 20, 2017

Flight delays have a negative effect on 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 data related to such a system.

To build flight delay prediction models from the Data Science perspective. They propose a taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data and computational methods, giving special attention to an increased usage of machine learning methods. Besides, we also present a timeline of major works that depicts relationships between flight delay prediction problems and research trends to address them.

5. **Predicting flight delay based on multiple linear regression[5]**

In this paper we studied about prediction of delay in flights using multiple linear regression. In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression.

For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. This helped us in getting a basic idea of how to implement a model in linear regression.

4. Scope and Motivation

The problem which we are trying to solve in this project is to predict by how much a flight is delayed given the weather data of that day as well as other factors which affect the flight of any airplane.

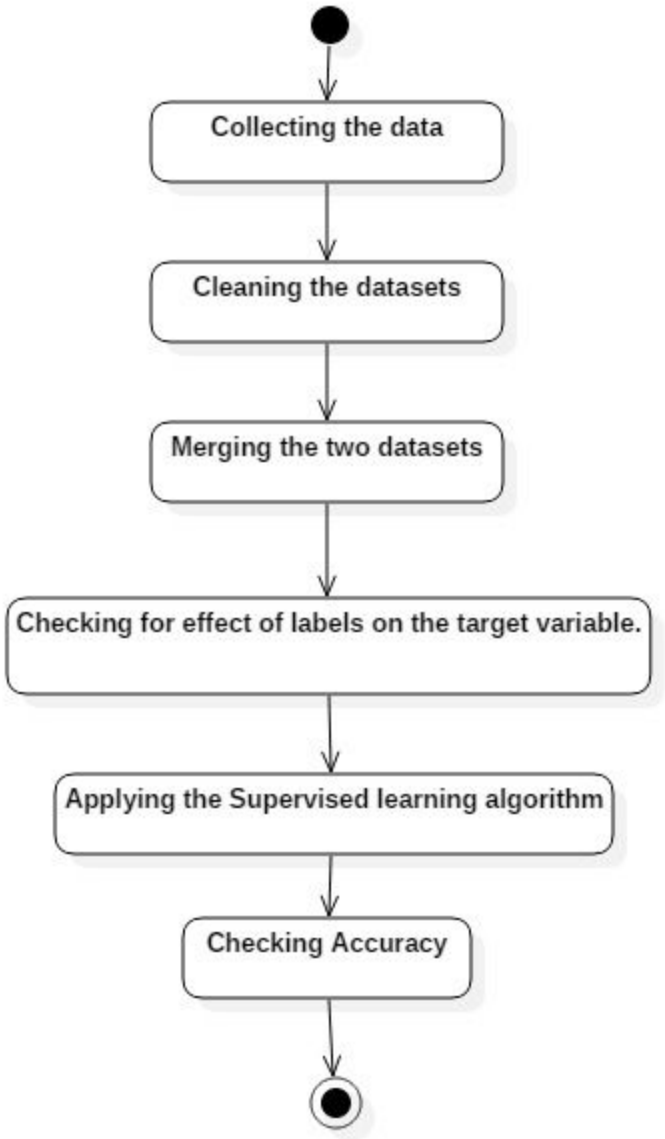
This project will be helpful in predicting flight delay with an approximate range by which it will be delayed thus serving a great deal to the Airline companies and Air Traffic control.

Our motivation to use SVM instead of other model is:

Why SVM? :

- Our dataset is high dimensional (9 dimensional) and it is known that SVM performs better than other common algorithms for a high dimensional dataset.
- Kernel trick of SVM can be utilized to obtain better accuracy.
- Because SVM finds the best separating hyperplane for a 2 class classification problem thus it works better for new test data.
- As the use of model depends on the problem, in our case we also tried with different algorithms and best accuracy was achieved with SVM (Support Vector Machines).

5. 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

6. 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

7. Method

The flight delay prediction problem can be modelled in many ways depending on the objectives or features present in the dataset. The whole problem solving process can be briefly divided into 4 parts.

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.

8. Design and Implementation of the System:

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**

9. 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.

10. 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