Ideation Phase
Literature Survey

Date	17 September 2022
Team ID	PNT2022TMID30209
Project Name	DEVELOPING A FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

1. Flight Delay Prediction based on Aviation Big Data and Machine Learning (Author: Rahul Garg et.al., 2022)

The dataset includes the scheduled and actual departure and arrival times for non-stop flights recorded by different airways. Information on delayed and canceled flights, actual travel time and non-stop distances is also available in the dataset. Airline origin and destination are also included. With this dataset, a predictive model was used to solve the flight delay cases. The flight dataset includes data for 161 airports. Flights arriving after the scheduled arrival time are considered delayed. In addition, the flight under attack is deleted from the dataset. , time of departure, time of boarding are applied. Unnecessary features are deleted from the flight data. This study aims to predict flight delays for airlines. Three methods are used to predict flight delay, that is, Random Forest, Support Vector Machine, K-nearest neighbor. The dataset is limited to only flight and weather data of the USA. The datasets from Other International Countries and the flight data for the domestic flights are not included.

2. Predicting Flight Delays with Error Calculation using Machine Learned Classifiers (Author Prof. S B Wani et.al., 2021)

For predicting the flight delays and to train the models, the data assembled by the organization of Transportation, U.S. Statistics of all the domestic flights taken in 2015 is collected and used. This Model is capable of filling the absent values which is crucial for refining data for the model. Supervised learning technique to gather the advantages of having the schedule and real arrival time. Algorithms are light computation costs. We develop a system that predicts for a delay in flight departure based on certain parameters.The mathematical models used in this are Logistic regression, Random Forest Regression, Decision Tree Regression. In the rest of the metrics, the value of error of Random Forest Regressor is even though not minimum but still gives a low value comparatively. In maximum metrics, it is found out that Random Forest Regressor gives the best worth and thus should be the model selected.

3. Machine Learning Model - based Prediction of Flight Delay (Author: N Lakshmi Kalyani et.al., 2020)

The Paper aims at predicting the arrival delay of a scheduled individual flight at the destination airport by utilizing available data. The predictive model presented in this work is to foresee airline arrival delays by employing supervised machine learning algorithms. XGBoost and linear regression algorithms were applied to develop the predictive model that aims at predicting flight delays. The performance of each algorithm was analyzed. XGBoost is a decision-tree-based machine learning algorithm that is implemented using a gradient boosting framework.Linear regression is one among the most popular machine learning algorithms for predicting values given a set of values. Linear regression is a linear method used to model the relationship between independent and dependent data. Flight data along with the weather data was given to the model.

Using this data, binary classification was carried out by the XGBoost trained model to predict whether there would be any arrival delay or not, and then the linear regression model predicted the delay time of the flight.

4. Assessing Strategic Flight Schedules at an Airport using Machine Learning based Flight Delay And Cancellation Prediction (Author: Miguel Lambelho et.al., 2020)

To migrate air traffic demand-capacity imbalances, demonstrate an approach for strategic flight schedules in the period 2013-2018. Machine learning approach to predict whether strategic, scheduled arrival/departure flights are delayed or canceled. These predictions are based on strategic flight schedules from LHR and assume a 6-month prediction horizon, i.e., we predict whether flights are delayed or canceled 6 months prior to the day of the flight execution. LightGBM is a tree-based machine learning algorithm that uses Gradient-based One-Side Sampling, which excludes data instances with small gradients, and Exclusive Feature Building, which bundles mutually exclusive variables, thus, reducing the number of features. We are considering extending the set of features for the prediction algorithms to improve the accuracy of the predictions and will evaluate the impact of considering flight delay and cancellation predictions in the flight scheduling optimization models, at the strategic phase.

5. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm (Author: M F Yazdi et.al., 2020)

The Levenberg-Marquart algorithm is applied to find weight and bias proper values, and finally the output has been optimized to produce high accurate results. To investigate the three models, we apply the proposed model on the U.S flight dataset that is an imbalanced dataset. Algorithm

used in this is min max normalization and denoising autoencoder training. In order to evaluate the model, the number of denoising autoencoders and neurons must be determined based on the values for precision, accuracy and time consuming. At the end, to evaluate the validity of the proposed model and the results from training, we evaluate the standard deviation of all the parameters after the 30 times repetition. Comparing the three models for two of imbalanced and balanced datasets shows that accuracy of SDA-LM model with imbalanced dataset respectively is greater by 8.2 and 11.3% Than SAE-LM and SDA models. On the other hand, these values for balanced datasets are respectively 10.4 and 7.3%. At the next stage, the model has been evaluated and computed for subjects of discarding with a standard deviation for all evaluation parameters during 30 times of model run. Finally, we compared the accuracy of the proposed Model against SAE-LM, SDA and RNN models .

6. Predicting flight delays using data from US Domestic flights (2019)

The objective of the project is to "Design a Model that predicts flight delays before they are announced on the departure boards". The dataset comes from Kaggle, and it consists of a multi-year data ranging from 2009 to 2019 separated in 10 different files. The data preprocessing and cleaning was done in two separate parts, documented in two notebooks to make it easier to follow up due to their length. The first section is a standard cleaning involving minimal feature engineering, and the second is driven after the 20 most common arrival destinations were defined based on the number of flights and is the one that contains the most feature engineering done.The same way the data cleaning and preprocessing was done in two separate notebooks, the EDA was done in two as well, however the difference here is that the visualizations done on each of the EDAs were done with different libraries. The first was done using matplotlib and Seaborn, and the second with plotly. Six type of ML Algorithms were tested, they were:

- Bagged Trees
- Random Forest
- AdaBoost
- Gradient Boosted Trees
- XGBoost
- Deep Neural Network (MLP)

It is quite hard to create a ML model for flight delay prediction before you even know that the flight is delayed on the departure board. Neural Networks responded a lot better under these conditions with an average difference in accuracy, precision and recall of over 15%. Maybe an even more thorough feature analysis could raise these metrics to close to 90%.

7. Flight Delay Prediction (Author: Bhuvan Bhatia, 2019)

The paper titled “Flight Delay Prediction” by Bhuvan Bhatia concentrated mainly on predicting flight delays for a particular airport over a specific period of time. First, they used a regression model to examine the significance of each feature and then, a feature selection approach to examine the impact of feature combination. These two techniques determined the features to retain in the model. Instead of using the whole set, we sampled 5,000 records at a time to run through different machine learning models. The machine learning models implemented here were Random Forest classifier and Support Vector Machine (SVM) classifier. Further, we applied an approach called One-Hot-Encoder to create a variant of the model for evaluating potential prediction.