Machine Learning Engineer Nanodegree Capstone Proposal
Predicting flight delays using supervised learning

Sirichandana Sambatur

Introduction:

Air-travel is very popular and many people choose air-travel because it is the fastest form of travel and people don't want to spend more amount of time traveling. But many airports and airlines are frequently prone to flight delays. These delays could be because of several reasons but predicting such delays ahead of time could help people plan their travel in an efficient manner. As there are several reasons that could cause a flight delay, manually assessing these reasons to predict a flight delay will be a tedious process. Machine Learning and historical data can be leveraged to build predictive models that predict flight delays[1].

Problem Statement:

The objective of this project is to build an efficient and generalized model that predicts flight delays. Machine Learning algorithms have the ability of understanding and detecting patterns in the historical data. Some of the machine learning algorithms such as Support Vector Machines, Logistic Regression, Decision Trees etc. can be used to learn the patterns of flight delays and build an effective predictive model. The project considers this problem as a classification problem and builds an identifier that predicts whether the departure time of the flight will be 'on-time' or 'delayed'. The input for these models will be a data instance that represents the information about a flight. This information includes airline carrier, origin airport, scheduled departure time, destination airport, scheduled arrival time etc. Provided an unlabelled instance which represents a flight, the trained models will identify if the flight will depart on-time or not.

Datasets and inputs:

The dataset that will be used for this project is gathered by The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics[2]. The dataset contains summary information of flights for the year 2015. The dataset represents information such as the airline carrier, scheduled departure time, actual departure time, origin airport, destination airport etc. This data will be provided as input for training the machine learning algorithms.

The dataset contains 1,048,576 instances and the dataset doesn't contain any class labels by default. During the data preprocessing stage, the class labels have to be assigned to the instances based on the features 'scheduled_departure_time' and 'departure_time'. If the scheduled departure time and the actual departure time are the same, then the instance will be labeled 'on-time' else 'delayed'. In the dataset, the majority of the instances represent ‘on-time’ flights resulting in an imbalanced class distribution. To avoid bias in the trained models, this imbalanced class distribution will be handled by using sampling techniques such as ‘upsampling’ or ‘downsampling’ while splitting the dataset. The following table provides a description of the features that will be used for training the models

Feature	Datatype
Year	Numeric
Month	Numeric
Day	Numeric
Day_Of_Week	Numeric
Airline	Categorical
Origin_Airport	Categorical
Destination_Airport	Categorical
Scheduled_Departure	Numeric
Departure_Time	Numeric
Departure_Delay	Numeric
Status	Categorical

Solution Statement:

The above-mentioned dataset contains information about the flights that were on-time, delayed, diverted or canceled for the year 2015. This information can be used to train classification algorithms and build predictive models. These models can be used to predict the status of a flight by providing the flight’s information to the model.

As the data does not contain any class labels by default, the class labels will be added during the data pre-processing stage. Exploratory Data Analysis, Feature Selection or Principal Component Analysis will be used to determine the most important features. Multiple classification algorithms such as Support Vector Machines, Decision Trees, Random Forests, Logistic Regression etc. will be trained and tested to identify the best model.

Benchmark Model:

A ZeroR Rule can be used as a benchmark model for this project. A ZeroR Rule does not consider any attributes or features. It uses prior probability and predicts the class label with the higher probability for all the instances. The area under ROC for the ZeroR Rule can be computed and the models trained on the data should outperform the ZeroR rule by having an area under ROC which is greater than that of the ZeroR Rule.

Evaluation Metrics:

The dataset will be split into training and testing datasets. Various models will be trained on the training dataset and will be tested on the testing dataset. The dataset is very imbalanced and contains a greater number of instances which represent the flight being on-time. Because of this imbalance in the class distribution, accuracy is not a good evaluation metric and different evaluation metrics such as precision, recall, and area under ROC will be computed to compare the performance of different models.

Project Design:

The dataset being used will be audited to look for missing values and inconsistencies. Any such missing values and inconsistencies should be resolved. The missing values can be filled with other properties of the distribution such as mean, median etc. If there are attributes with a majority of the values missing, such attributes can be removed as they do not provide much information.

Exploratory data analysis will be performed on the dataset to learn the statistics and behavior of the data. Univariate and multivariate relationships will be explored to look for important relationships and patterns that help in identifying the best attributes or features. Based on the results of the exploratory data analysis, feature engineering may be used to create new features.

The dataset will then be split into training and testing datasets. 80% of the data will be used as training dataset and the remaining 20% of the data will be used as the testing dataset. Multiple classifiers such as Support Vector Machine, Decision Tree, Random Forest and Logistic Regression will be trained and their parameters will be tuned using the training dataset. The trained models will be compared by evaluating their performance on the testing dataset. The model with the best performance will be identified as the final model.

References:

1. http://rstudio-pubs-static.s3.amazonaws.com/179496_64f977e866ee430484ec7c262f7583e4.html
2. https://www.kaggle.com/usdot/flight-delays/data

3. https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial/notebook
4. https://www.datasciencecentral.com/profiles/blogs/predicting-flights-delay-using-supervised-learning