Airline Passenger Referral Prediction

Sayan Bandopadhyay
Data science trainee,
AlmaBetter, Bangalore

Abstract

Air travel is attractive because of its speed and range and also because, for business visitors, it offers status as well as saves valuable work time when travelling on a long- haul basis. Air transport comprises both scheduled and chartered categories and in some parts of the world, air taxies.

Here, I am provided with an airline dataset which includes reviews of passengers based on their experience of different services being provided by the airline company. Main objective here is to make a proper use of various machine learning algorithms so as to perform predictive analysis which may help the airline industry to make some rational decisions to boost up their business.

Keywords: machine learning, airline, predictive analysis

Problem Statement

Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. Data is scraped in Spring 2019. The main objective is to predict whether passengers will refer the airline to their friends.

Steps Involved

Following are the different steps been taken in the overall analysis process right from the beginning to the end:

- ✓ **Framing Questions**: This is the very first step of the analysis process. Going through the problem statement deeply and to extricate some hidden questions that problem statement is throwing to us is the main objective here as based on which the analysis will choose its path.
- ✓ **Data Inspection**: Once the objective of the problem statement becomes clear, now it's the time to go through the dataset thoroughly in order to understand different features that the dataset is comprised of and also to check some statistical scores and a bit of more information regarding the dataset based on which next steps of the process can be designed.
- ✓ **Data Scrubbing & Pre-processing**: Raw datasets usually have missing values or null values. To make sure that the models later does not malfunctions or provides some biased insights, data cleaning is essential and this is what is being

done in this particular step. This includes null or missing values treatment, outlier treatment and duplicate value check.

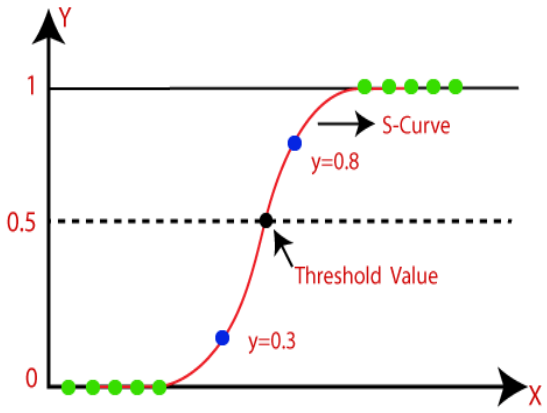
- ✓ **EDA**: Exploratory data analysis is the step where we just visualize data based on different features of the cleaned dataset so that we can come across some meaningful insights.
- ✓ **Data Preparation**: This step basically includes arrangement of data and feature engineering so that machine learning algorithms can be implemented without any hassle.
- ✓ **Model Implementation**: Sixth step is all about splitting data into training data and test data and fitting different machine learning algorithms for prediction of target variable or dependent variable.
- ✓ **Metric Evaluation**: This particular step has the objective of comparing the evaluation metrics of each of the models so as to find out the best fit model for this particular case.
- ✓ **Conclusions**: Last step is all about summing up the insights and model performances observed from analysis process.

Algorithms

For this particular case, 4 different supervised machine learning algorithms has been implemented. They are –

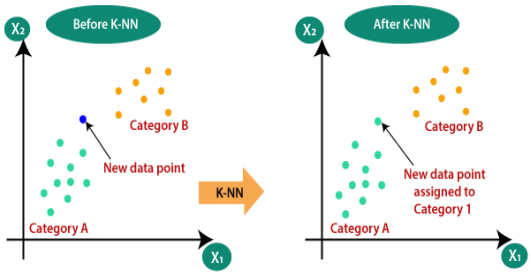
LOGISTIC REGRESSION

Its one of the most popular machine learning algorithms which comes under the supervised learning technique. It's used for predicting the categorical dependent variable using a given set of independent variables. Therefore, the outcome must be a categorical or discrete value. So, it's used for solving classification problems. It can be used to classify the observation using different types of data and can easily determine the most effective variables used for classification. Consider the below image:



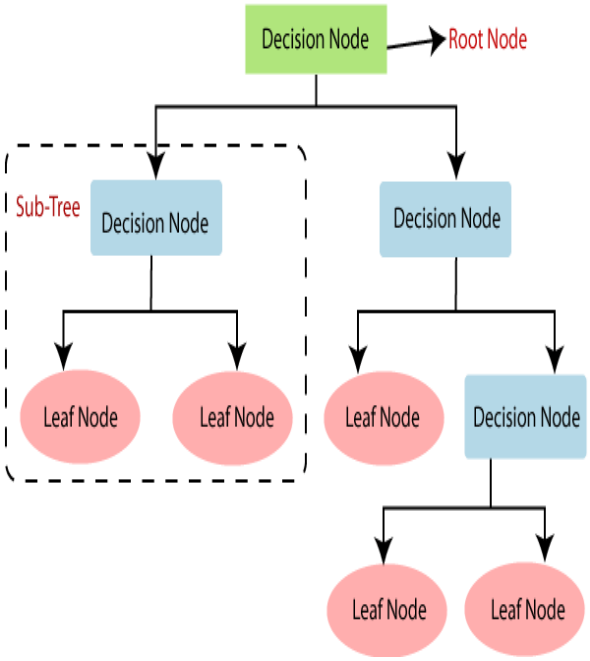
KNN

Stands for K-Nearest Neighbors is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.



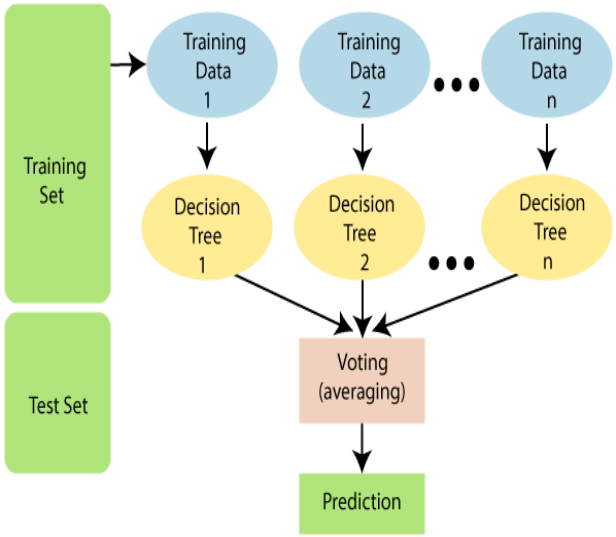
DECISION TREE

It's a supervised machine learning technique that can be used for both regression and classification problems, but mostly it's preferred for solving classification problems. It's a tree-structured classifier where internal nodes represent the features of the dataset, branches represent the decision rules and each leaf node represents the outcome. It's a graphical representation for getting all the possible solutions to a problem based on given conditions. Below diagram explains the general structure of a decision tree:



RANDOM FOREST

It's a supervised machine learning technique that can be used for both regression and classification problems. It's based on ensemble learning technique which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It's a classifier which contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The below diagram explains the working of random forest algorithm:



Challenges Faced

- Dataset had a lot number of null and missing values. So, to deal with that was a challenge as, while treating null and missing values, it was very important to take care of the fact

that the data does not lose its meaning.

- Data preparation took a lot of effort as there were many columns that was supposed to get removed and some of them also needed one hot encoding.
- Choosing best models to implement and tuning those models was a challenge so that it gets the best parameters for each model.
- Evaluation metrics for all models was very similar to each other which posed a challenge to choose the best one out of all the models implemented.

Conclusion

At the end of this long process, the results tells a story regarding the preference of the passengers on different services offered by the airlines and also machine learning algorithms managed to create some useful models for the purpose of predictive analysis where Random forest algorithm after hyperparameter tuning using 'GridSearchCV' performed the best out of all whereas Decision tree algorithm performed the worst.

References –

1. GeeksforGeeks

2. My Great Learning

3. Analytics Vidhya

4. JavaTpoint