AIRLINE PASSENGER SATISFICATION

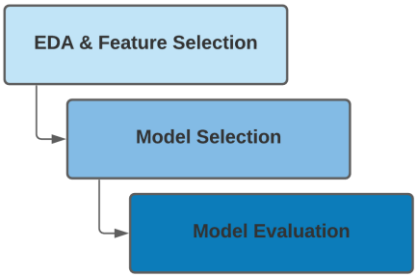
Abstract:

The results of the airline passenger survey are analysed to predict if or whether the passenger is satisfied or not. As it is very important to know if the passenger travelled are satisfied with the service provided or not and it helps the company to improve based on that. This study uses several classification models such as KNN, Logistic Regression, Gaussian NB, Decision Trees and Random Forest, Ensemble, unsupervised learning which will later be compared. The results of this study get the Random Forest Algorithm as the best one with an accuracy of 97 percent.

Introduction:

Before getting into the main content, it is very important to know about the data set. In our dataset includes approximately 129880 survey entries as well as passenger and flight information from US airlines. There are 21 function columns and one level goal column in all. Fourteen of the features are survey entries in which passengers score their flight experience on a scale of one to five. After which we have cleaned dataset by removing the null values and duplicates, which we call it as **data cleaning, EDA, model selection, model evaluation and Hyperparameter Tuning.**

Methods used:



Data Cleaning and Data Analysis:

- Removed the null values and duplicates from the dataset.
- After which we tried to label encode and decrease the number of features by combing two features. And changed the column names which makes easy to analyze.
- Data transformation in the class Satisfaction or satisfaction with, satisfied: 1, neutral or dissatisfied: 0.
- Remove Departure Delay In Minutes & Arrival Delay In Minutes features.
- Data transformation on class feature (Ticket type) with, Eco: Economy, Eco Plus: Economy, & Business: Business.

EDA & Feature Selection:

- In order to analyse the data, it is very important to select important features, so we will perform of KDE plots, LASSO lines, and heat maps. After careful evaluation and selection, we decided to delete 'Gender', 'Total Delay', 'Flight Distance', 'Age', 'Gate Location' and 'Departure / Arrival Time Convenience'.
- And then we split the dataset in to 80 % for 5-fold cross validation and 20 % as a test data. 60% data is training data and 20% is validation data and the rest 20% is testing data.

By running the GridSearchCV algorithm on Scikit-Learn, the optimum hyper models and parameters are:

- k-Nearest (k = 7)
- Logistics Regression (C = 0,04)
- Decision Tree (Max Depth= 12)
- Random Forest (Max Depth= 17)

Results and Evaluations:

- Models Compared:
    - Logistic Regression, KNN, Gaussian NB, Decision Tree, Random Forest, Ensemble, SVC, Neural Network, PCA (principal Component Analysis), K_Means Clustering.
- Finally, we test the predictions on the remaining 20% of the data set after we construct the model and retrain the selected model on 80 percent of the data set (Training + Set Validation) (Set Testing). The final model efficiency assessment yields an AUC of 99.03%, a Recall of 93.2 %, and a Precision of 96.1 %.
- The random forest gives best results than any other algorithm used which gives its accuracy as 96%. Here it was enough to use around 48 decision trees to build a random forest.

Conclusion:

We have finally built a mostly precise model which will perform the classification (satisfied / unsatisfied) properly and help the airlines to improve their services. We found that most of the passengers are unsatisfied/neutral with the airline services. As per the data we could see that most of the business travel passengers are satisfied. And it would have been better if we have classified the data in satisfied, unsatisfied and neutral, as the probability of getting unsatisfied/neutral is more than that of the satisfied.