# Predicting Airline Passenger Satisfaction

Loreana Oluić SW-60/2018

## 1. Motivation

Passenger satisfaction plays a crucial role in the airline company's improvement. It's significant to retain existing customers and add new customers. Unsatisfied passengers, on the other hand, will decide not to fly with the same airline again in the future [1], or they may launch a negative word-of-mouth campaign (which may be electronic) that harms the company's credibility and image [2]. Therefore, companies require a database that will store completed surveys and questionnaires.

## 2. Research questions

An effective customer satisfaction data analysis represents a challenge and can be solved using machine learning. Dataset which will be used contains attributes:

- Gender (*Male, Female*)
- Age
- Customer Type (*Returning, First-time*)
- Type of Travel (*Business, Personal*)
- Class (*Business, Economy, Economy Plus*)
- Flight Distance
- Departure Delay
- Arrival Delay
- Departure and Arrival Time Convenience (*1-5*)
- Ease of Online Booking (*1-5*)
- Check-in Service (*1-5*)
- Online Boarding (*1-5*)
- Gate Location (*1-5*)
- On-board Service (*1-5*)
- Seat Comfort (*1-5*)
- Leg Room Service (*1-5*)
- Cleanliness (*1-5*)
- Food and Drink (*1-5*)
- In-flight Service (*1-5*)
- In-flight Wi-Fi Service (*1-5*)
- In-flight Entertainment (*1-5*)
- Baggage Handling (*1-5*)

The dataset contains 23 different attributes about airline passenger satisfaction survey. All these attributes give us useful information about the passengers and how they rated the different services of the flights. Based on these attributes, we will be able to predict if the passenger is:

- 1 - Satisfied
- 0 - Neutral or unsatisfied

## 3. Related work

- Source: https://www.kaggle.com/code/tarunag1506/airline-passenger-satisfaction-logistic-regression
  Here we have an example of using Logistic regression for predicting passenger satisfaction.
  **Logistic regression** is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on [3].
  First, the dataset was preprocessed – checked for null values, removed extra column ("Arrival Delay"), split to train and test dataset (train size is 0.8), categorical data was converted to int64 by using label encoding. Accuracy was 87% and metrics: MAE (`18.676509085309515`), MSE (`0.12715583615645212`), RMSE (`0.356589169993218`).

Many other machine learning models were used like Naïve Bayes, Nearest Neighbor, Neural Networks...

## 4. Methodology

The first column, "Id", is unnecessary. Dataset contains null values in "Arrival Delay" column. Therefore, that column has to be removed. Columns "Gender", "Customer Type", "Type of Travel" and "Class" are object types. Using Label Encoding types will be converted to int64.
The prediction task was addressed as a classification problem. For classification is used SVM model and Random Forest.
A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. [4]

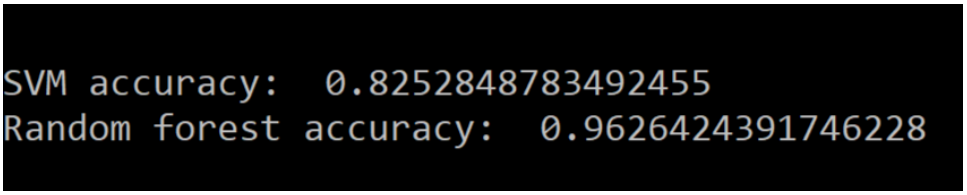
5. ## Discussion

From full dataset column “Satisfaction” is dropped and it represents x dataset. On the other hand, column “Satisfaction” is y dataset. Before splitting the dataset, Label Encoding has to be done on x dataset. For Label Encoding was used preproccessing.LabelEncoder() (from `sklearn` library) which converts, for example, Gender: Male – 1, Female – 0.

Dataset is split into train and test set. The size of the test set is 75% of the total dataset and the size of the train set is 25%.

At first, it seemed better to use SVM technique for data classification in this case, because here we have binary classification (satisfied or neutral/unsatisfied). Besides that, we also have a very large dataset of 129880 rows. SVM is not suitable for classification of large data sets, because the training complexity of SVM is highly dependent on the size of the data set. [6] It creates NxN matrix and uses a large amount of memory. Non-linear SVM (SVC from `sklearn` library) took too long to process so I tried to use linear (LinearSVC), set C=0.2 and with that I achieved best accuracy which is 82%.

As if this is not enough, I started to research other models and found out that Random Forest is giving very good results with big datasets. On this dataset accuracy was 96%.




For measuring accuracy, I used f1 micro score. Micro F1-score (short for micro-averaged F1 score) is used to assess the quality of multi-label binary problems. [7]

6. ## References

[1] J. Namukasa, “The influence of airline service quality on passenger satisfaction and loyalty the case of Uganda airline industry,” *TQM J.*, vol. 25, no. 5, pp. 520–532, 2013, doi: 10.1108/TQM-11-2012-0092.

[2] J. Blodgett and H. Li, “Assessing the Effects of Post-Purchase Dissatisfaction and Complaining Behavior on Profitability: A Monte Carlo Simulation,” *J. Consum. Satisf. Dissatisfaction Complain. Behav.*, vol. 20, pp. 1–14, 2007.

[3] https://www.sciencedirect.com/topics/computer-science/logistic-regression

[4] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[5] https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[6] https://www.sciencedirect.com/science/article/abs/pii/S0925231207002962

[7] https://peltarion.com/knowledge-center/documentation/evaluation-view/classification-loss-metrics/micro-f1-score