MScA 31008 Data Mining Project Proposal

Group Member: Guangze Wang, Kaicheng Zhang, Meng Yang, Wanxuan Zhang, Zhiyi Zhao
(* in alphabetical order)

Project Title

Airline Passenger Satisfaction

What factors lead to customer satisfaction for an Airline?

Project Problem Statement

Since the invention of Aircraft, the corresponding development of the Airline Industry was anticipated as technology breakthroughs can arrive at a sustainable phase only if it goes for marketization. Generally speaking, the Airline Industry provided two main types of service, passenger transportation and cargo transportation. Cargo transportation has a stable and systematic management procedure, and given goods have no sentiment or behavior change, we anticipate the future development of this segment would experience a standardized gradual process, which only needs to consider a matter of transportation volume. In comparison, passenger transportation requires long-term and sophisticated operations just like others in the service sector.

While we might hear about those aspects the most, the customer experience is not about just the flight itself. It's everything from purchasing the ticket on the company's website or mobile app to checking bags in at the airport or via a mobile app to waiting in the terminal. Self-service has been top-of-mind for airlines since the introduction of airport kiosks that enable passengers to check-in, upgrade their seats, and even make flight changes. This mindset has been, and continues to be, adapted to the post-security, onboard, and post-flight experience. And for Airline companies, they share the same focus. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. It's important that customers have an excellent experience every time they travel. On-time flights, good in-flight entertainment, more (and better) snacks, and more legroom might be the obvious contributors to a good experience and more loyalty.

In this project, we stand in the position of data scientist for the Airline Companies with the aim to understand what factors in the whole passenger journey are most significant in determining their satisfaction rating. Given the response variable (satisfaction) is categorical, we would first run several classification approaches (Logistic Regression, SVM, Random Forests) on the training set, and use the best models to predict whether a customer was **Satisfied** or **Unsatisfied** with the experience and/or service which an airline provided. The result of this project would land a

tradeoff between optimizing customer experience and cost spent on areas to improve upon for Airline companies, and actionable recommendations would be given to the operation team of the companies.

Data Description

Sample Size:

Train dataset: Total of 104,000 samples

Test dataset: Total of 26,000 samples

Data Cleanliness: Data has no missing value

Columns and description:

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Expected results (Deliverables, Outputs)

Through this project, our team expects the final model to be able to accurately classify a customer's satisfaction based on his/her survey. The evaluation matrix we will use is F-score. Since airlines value both satisfaction and non-satisfaction results, we hope our model can achieve the highest F-score as possible.

At the same time, we also hope to help the airlines to understand which factor influences the customers' satisfaction the most, by providing a list of factors with decreasing influence power. So that the airlines continue focusing on improving their services from all aspects and put more effort on certain areas to optimize the result.

Proposed Procedure

There will be mainly four stages of our project, and each stages will be conducted for designed purposes as described below:

- Initial stage (Week 6)
    - Project Value Evaluation
        - Identifying the most crucial factors affecting passengers satisfaction with airlines, and therefore, improving such factors to increase sales and generate more revenues for Airline companies
    - Check The Source Data
        - Familiarizing with source data and data dictionary
- Data Exploration (Week 7)
    - Exploratory Data Analysis
        - Clustering
            - K-Means
            - DBSCAN
            - Hierarchical Clustering
    - Data Cleaning including clearing missing values, outliers engineering, feature scaling, and checking for imbalanced data problem
    - Dimension Reduction (PCA, t-SNE)
- Data Modeling (Week 8)
    - Algorithm

        - Generalized Linear Models
            - Logistic Regression
        - Decision Trees
            - Random Forest
        - SVM
        - Light Gradient Boosting Machine/XGBoost
- Summary (Week 9)
    - Model Selection
        - Based on different performance metrics
    - Model Interpretation
        - Identifying important factors for model prediction
    - Future Work
        - Model optimization