QUANTIFYING THE AIR TRAVEL EXPERIENCE:

A Comprehensive Analysis of Influential Factors in Passenger Satisfaction

Group 5

MUHAMMAD ZAHRIEL BIN ISMAIL	Leader	22085509
HAN XIANG	Maker	22093085
IZZUL ILHAM BIN YUSOF	Detective	22107573
TASSLIM BIN MANSOOR ALI	Oracle	23056322
ZHANG LEPING	Presenter	22098083

Content

1. Introduction
2. Data modeling
3. Evaluation
4. Deployment
5. Conclusion

Group 5

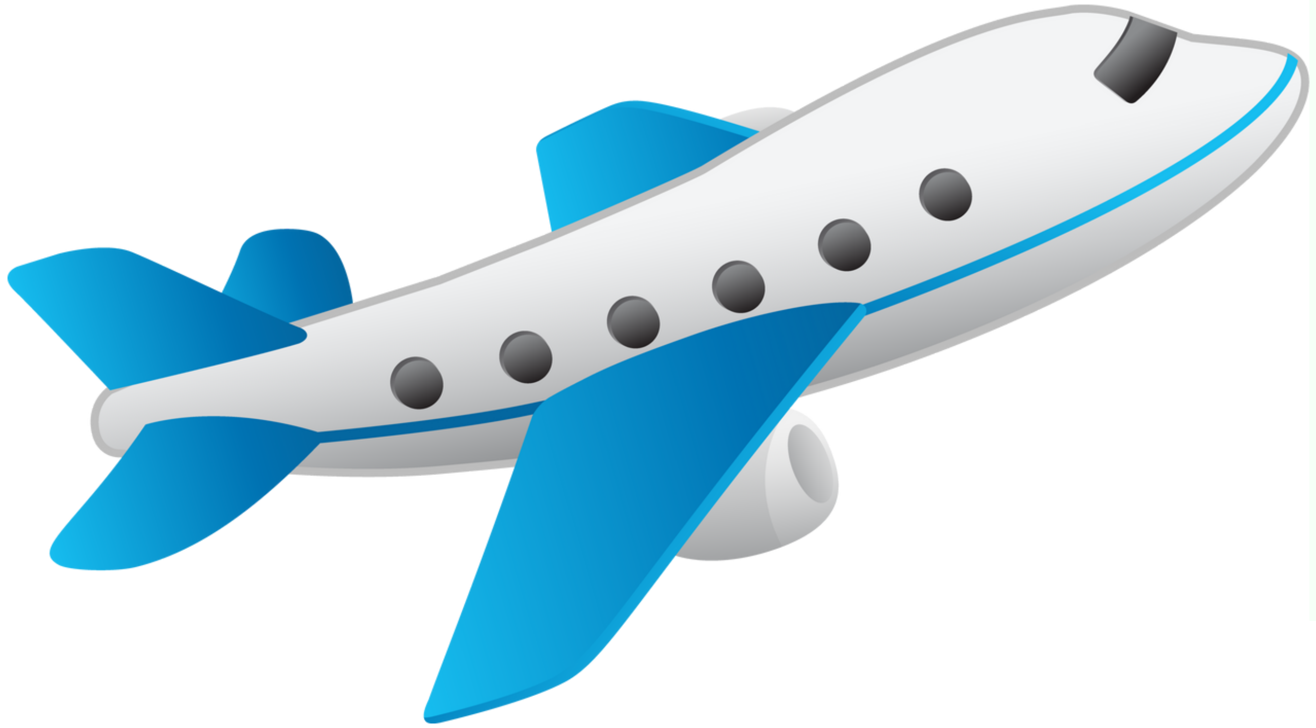
INTRODUCTION

Summary of Project

Project Background

Aviation Industry

- Major tourism driver
- Contributes 7.6% to global tourism GDP (WTC 2023)
- Highly competitive industry (focused on the journey)



Problem Statement

- What are the key contributing factors that impact the passengers overall experience while flying?
- What is the impact of flight delays and arrival delays on overall passenger’s experience?

Project Objective

- To identify and analyse the key contributing factors that impact the overall customer experience in the context of airline travel.
- To investigate the specific impacts of departure and arrival delays on overall customer experience.
- To predict customer satisfaction using machine learning models.

Mechanics

Hardware

- Minimum hardware requirement for R.
- Operating system: Windows 10
- CPU Architecture: Intel Core i3 or AMD Ryzen 3250u (64-bit)
- 1GB RAM, 2GB hard-disk space

Software

- The programming language used in this project is R.
- Chose R for familiarity and extensive libraries.
- Simplified syntax in R, beginner-friendly and easy to learn

Platform

- Employ Shiny for interactive web apps.
- Use GitHub for version control to ensure reproducible research.
- Use Team to hold meetings and share documents

Data Science Pipeline

Phase 1

- Define Problem & Goals
- Data Procurement
- Data Understanding
- Data Preparation
- Data Cleaning
- Exploratory Data Analysis

Phase 2

- Modeling
- Evaluation
- Deployment
- Results and Discussion

Group 5

DATA MODELING

Experiment & Results

Experiment & Results -Result Matrices

Machine Learning

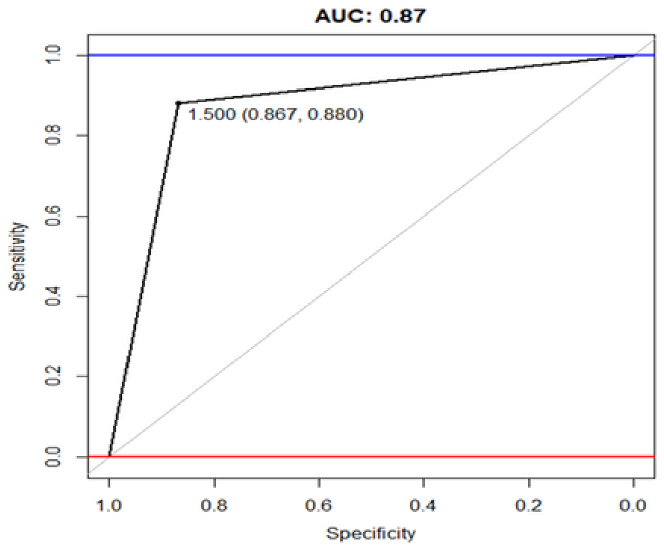
- Create data partition (75:25) using "set.seed(100)" for reproducibility
- Using caret package to streamline model training
- Employed 5 classification models
- Performed cross-validation for 5 times

	Accuracy	Sensitivity	Specificity	F1-Score	AUC
Logistic Regression	87%	90%	83%	89%	87%
Decision Tree	87%	87%	88%	89%	87%
Gradient Boosting	96%	97%	93%	96%	95%
K-Nearest L-Neighbour	92%	97%	87%	93%	92%
Random Forest	96%	98%	94%	97%	96%

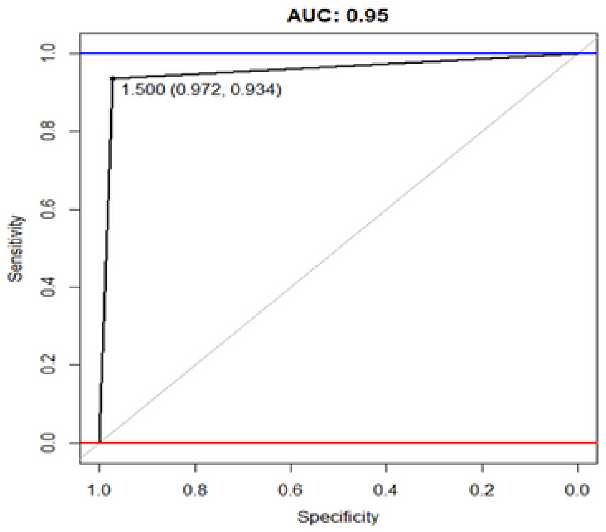
Result

Random Forest Model, the best performance model, is better to fit the data and predict passenger satisfaction.

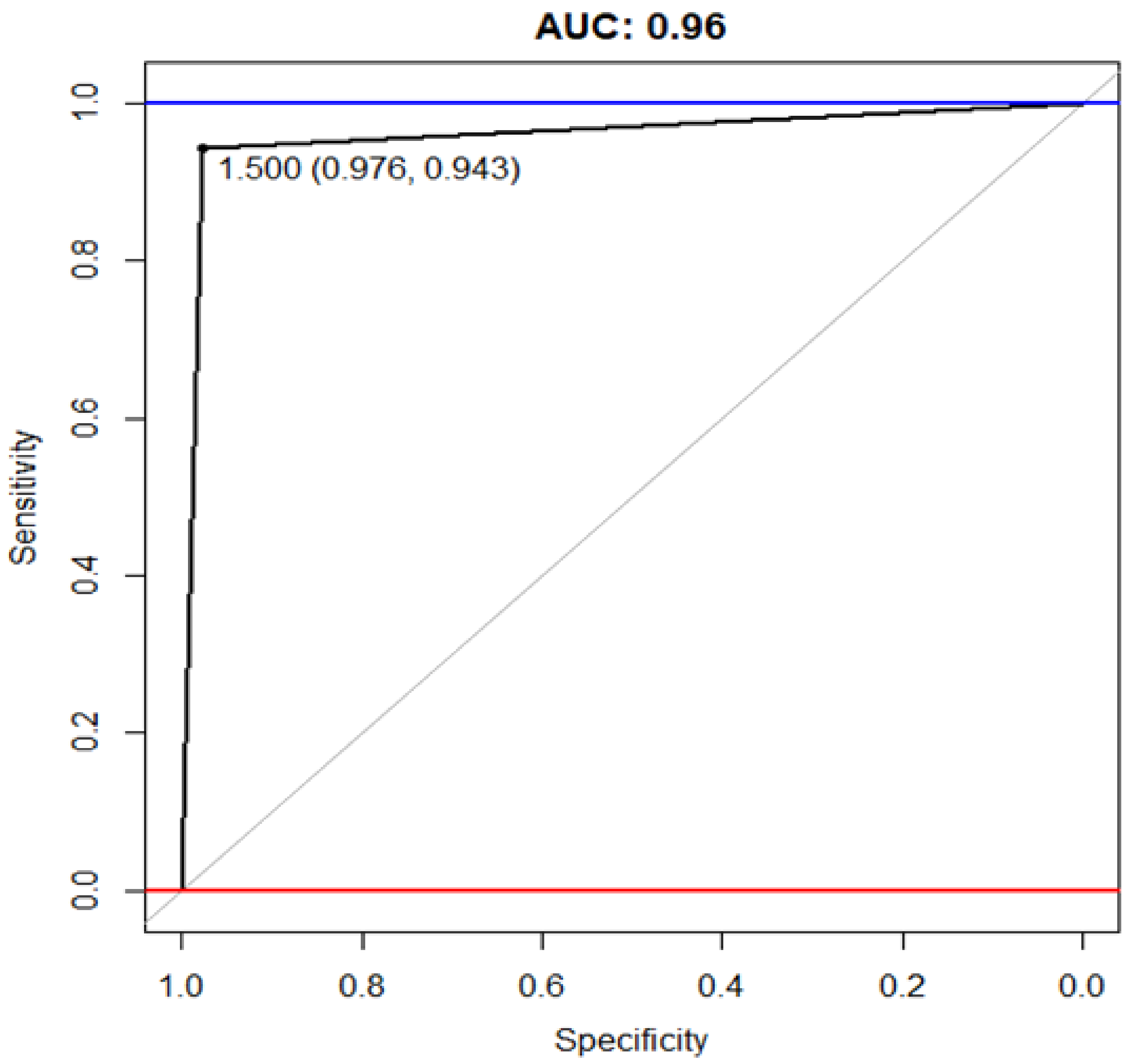
Experiment & Results -Result Matrices



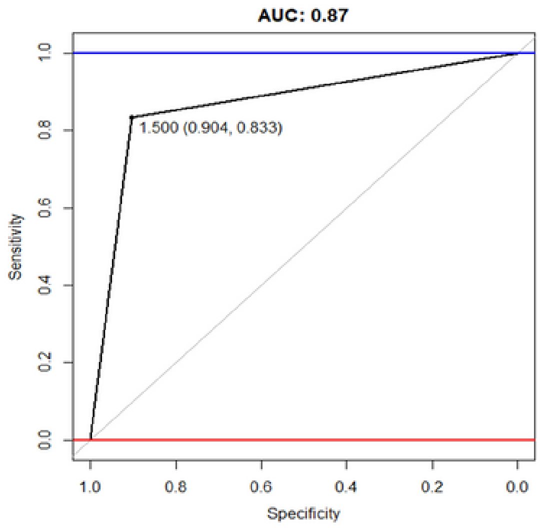
Logistic Regression



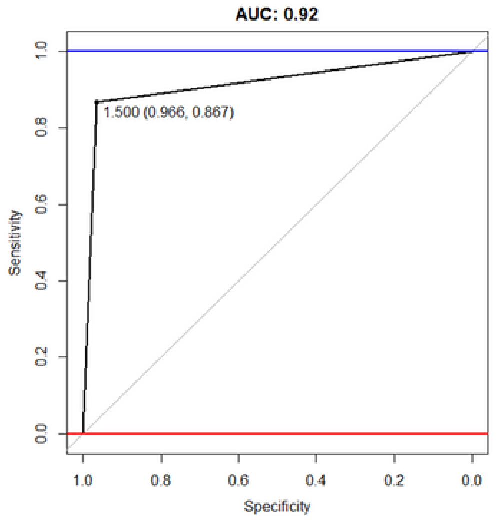
Gradient Boosting



Random Forest



Decision Tree



KNN

Experiment & Results - Rated Features

Based on the analysis conducted in the previous chapter we can do a cross analysis of the Highest rated features that is recognized by each model.

Logistic Regression	Decision Tree	Gradient Boosting	K-Nearest Neighbour	Random Forest
Type of travel	Class	Online Boarding	N/A	Online Boarding
Customer Type	Online Boarding	In-flight Wi-Fi	N/A	In-flight Wi-Fi
Online Booking	Type of Travel	Type of Travel	N/A	Class
Check In service	In-flight entertainment	Class	N/A	Type of Travel
In flight Wi-Fi	In-flight Wi-Fi	In-flight entertainment	N/A	In-flight entertainment

Result

In answering the problem statement highlighted earlier, the **top 3 features** based of the feature importance analysis across the models trained are:

- **Online Boarding**
- **In-Flight WiFi**
- **Type of Travel**

	Arrival Delay Ranked	Departure Delay Ranked
Logistic Regression	16/26	18/26
Decision Tree	29/29	29/29
Gradient Boosting	16/29	20/29
K-Nearest Neighbour	N/A	N/A
Random Forest	18/29	19/29

Group 5

EVALUATION

Comparison & Results
Future Studies

Comparative Study - Methods

Evaluation

Identify similar studies that utilized the same datasets and comparing their conclusions to this study forms the table below on the features

Characteristics	This Study	Study 1	Study 2
Preprocessing Methods	Binning (Ages, Flight Distance)	Normalization (Arrival Delay, Departure Delay)	Binning (Type of Travel), Dropped Columns (Departure Delay)
Handle Missing Data	MICE Imputation	MICE Imputation	Dropped Column
Models Used	Decision Tree, Logistic Regression, Random Forest, Gradient Boosting, K-Nearest Neighbour	Catboost Classification	K-Nearest Neighbour, Decision Tree, Random Forest, LASSO Regression

Comparative Study - Features

Result

Despite difference in ranking, the **overall features** that are present are the same.

Rank	Feature (This Study)	Feature (Study 1)	Feature (Study 2)
1	Online Boarding	In-Flight WiFi	Online Boarding
2	In-Flight WiFi	Type of Travel	In-Flight WiFi
3	Type of Travel	Online Boarding	Type of Travel

Further analysing the outcomes of the studies, the team has filtered the features for only **service-based features.**

Rank	Feature (This Study)	Feature (Study 1)	Feature (Study 2)
1	Online Boarding	In-Flight WiFi	Online Boarding
2	In-Flight WiFi	Online Boarding	In-Flight WiFi
3	In-Flight Entertainment	Check-In Service	In-Flight Entertainment

Future Studies

- **More airlines** for a clearer comparison
- **A bigger set of data** from different places to help understand what makes customers satisfied
- **Diverse models** to improve potentially better use-cases

Aspect	Improvement	Argument
Data Constraints	Data is limited to US Airlines, expanding the scope of data to include data from other countries/airlines.	Provides deeper insight into an overall view of airlines not limited to potential US related biases.
Data geo-temporal aspects	Expanding the data by adding a Geo-temporal aspect to the data such as location and date-time of the survey.	This helps by allowing the study of geographical cluster biases within a given area as well as showing the changes of sentiments over time via a temporal aspect.
Diverse Classification Models	Utilizing more robust models and exploring the potential usage of ensemble models.	Increasing the scope of models may identify potentially better use-cases as compared to the ones utilized in this study.

Group 5

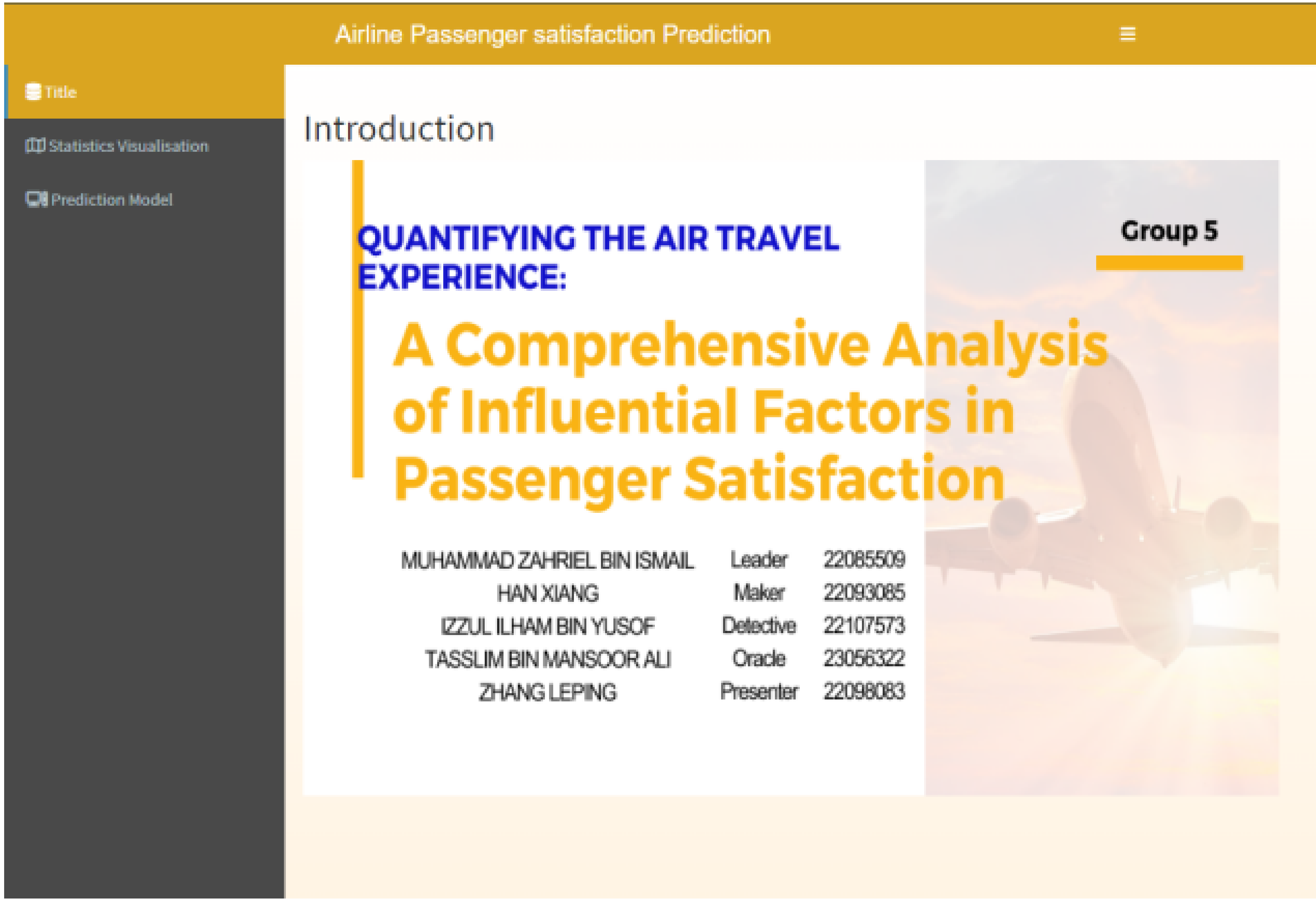
DEPLOYMENT

Data Product

Main Page

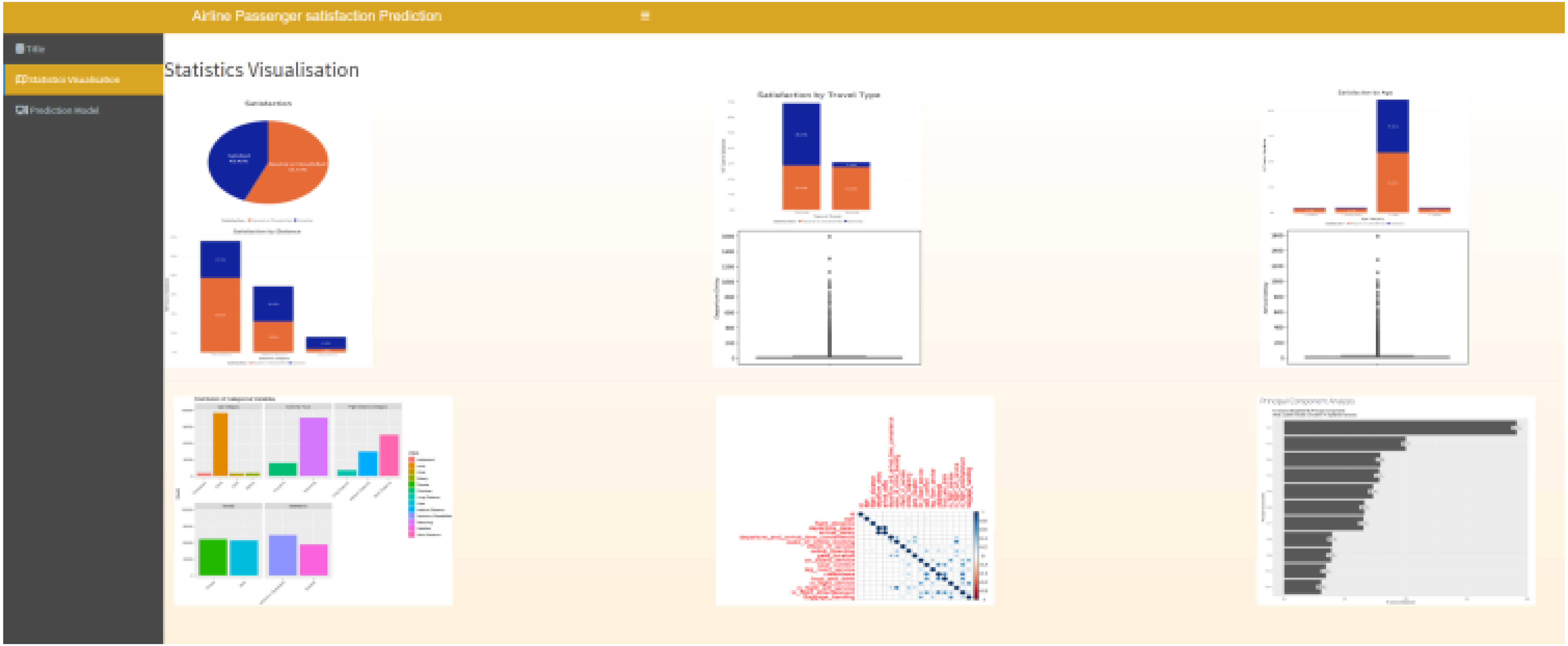
Data Product

An interactive web application on **Shiny** which can predict passenger satisfaction based on the facilities provided by the airline.



Visualization Page

Presenting insights from Exploratory Data Analysis (EDA)



Prediction Model Page

Model Used: Random Forest

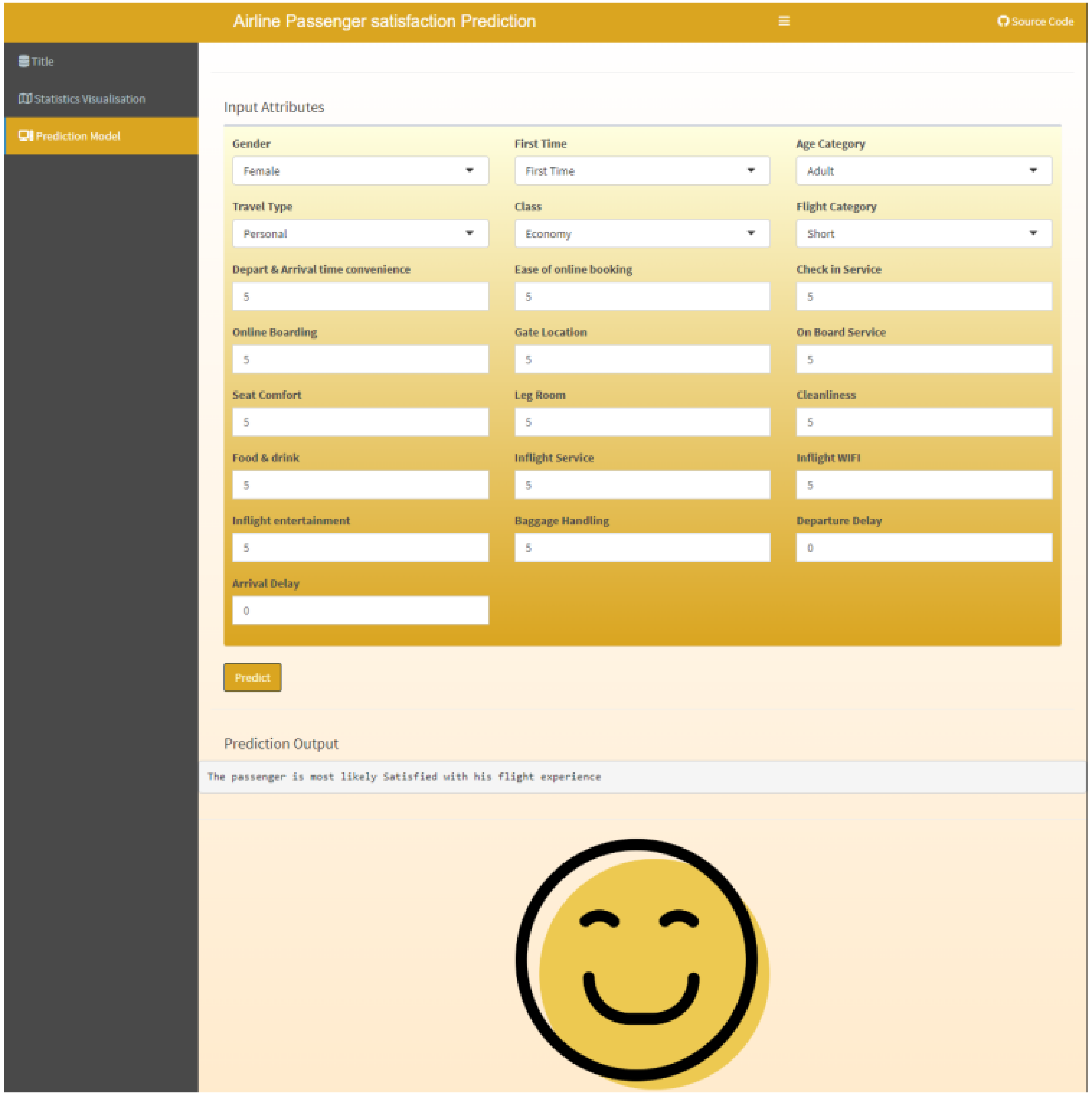
Instructions

Input

- Gender
- First-time flying status
- Age
- Category
- Travel type
- Class and flight category
- Ratings for specific facilities

Output

- Satisfaction level
- Relevant emoji



Group 5

CONCLUSION

Reproducible Research Insights

Plan for Reproducible Research

Reproducibility Validation

To ensure that the data is accessible as well as the code for this study, all the steps involved will be posted on GitHub and be publicly available for researchers to reproduce the study.

Code Document

Furthermore, to ensure that the model training will produce the same results, the team has preset the seed in R using “set.seed(100)” to ensure that no variation occurs.

Data Repository

Added on to this, documentation, datasets as well as dependencies (Libraries, etc.) will also be listed within the GitHub repository.

https://github.com/ZahrielIsmail/WQD7001_Group_5_Assignment

Conclusion

- According to customer reviews, Online Check-In, In-Flight WiFi, and In-Flight Entertainment services are areas that need improvement.
- Arrival/departure delays have a relatively low impact on overall satisfaction.
- These findings help airlines prioritize their efforts to improve customer experience and retention.

Reference

Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. *Scientific Reports*, *12*(1), 11174. https://doi.org/10.1038/s41598-022-14566-3

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets [Article]. *Journal of Big Data*, *6*(1), Article ARTN 62. https://doi.org/10.1186/s40537-019-0224-1

Noviantoro, T., & Huang, J. (2022). Investigating airline passenger satisfaction: Data mining method [Article]. *Research in Transportation Business and Management*, *43*, Article ARTN 100726. https://doi.org/10.1016/j.rtbm.2021.100726

Tahanisaz, S., & Shokuhyar, S. (2020). Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry [Article]. *Journal of Air Transport Management*, *83*, Article ARTN 101764. https://doi.org/10.1016/j.jairtraman.2020.101764

Herawan Hayadi B., & Khodijah, H. (2021). Predicting Airline Passenger Satisfaction with Classification Algorithms [Journal]. International Journal of Informatics and Information System *t*, *83*, Article ARTN 101764. http://ijiis.org/index.php/IJIIS/article/view/80

David L. (2022) Airline Passenger Satisfaction Catboost Model [Kaggle]

https://www.kaggle.com/code/jeongkyulim/catboost-model-over-96-4-accuracy)

Group 5

THANK YOU !