Airline Customer Satisfaction Prediction

A Binary Classification Problem

Flight Crew

DATA ANALYSTS

- Dreama Wang - 261112206
- Nishi Nishi - 261078870

DATA SCIENTISTS

- Riley Zhu - 261094733
- ShanShan Lao - 261072808

BUSINESS ANALYSTS

- Micheal Murphy - 261060598
- Vibhu Bhardwaj - 261113187

PRODUCT MANAGERS:

- Darin Zlatarev - 261081234
- Utkarsh Nagpal - 261071466

CUSTOMER SATISFACTION

1.4 Billion

USD Per Year

Revenue each US airline leaves on table by failing to improve their customer experience

Source: Forrester

NET PROMOTER SCORE PROGRAMS

200,000

USD Per Year

Amount spent on running NPS programs for a company with 1000 employees and $100 million in revneue

Source: CustomerGauge

Destination

Objective:

1) Predict Customer Satisfaction

2) Identify features that contribute most to customer satisfaction

3) Use semi-supervised learning to create labels for data

Outcome:

1) Increased Revenue (upto USD 1.4B)

2) Cost savings in running NPS programs

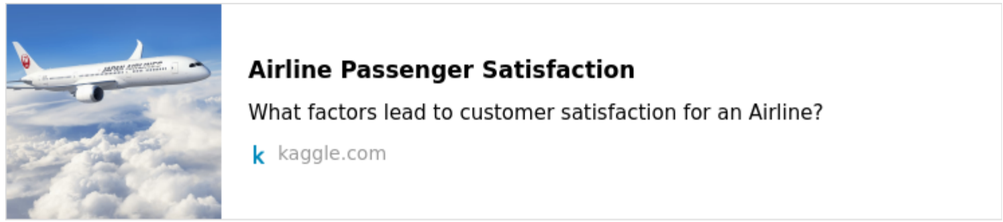
3) Improved brand image and perception

Dataset:

1) Balanced Classification Dataset

2) Binary labels in target column named "satisfaction"

3) Source: Kaggle



https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv

Flight Path

01	02	03	04	05
EDA	**Preprocess**	**Supervised Learning**	**Semi-Supervised Learning**	**Results & Interpretation**
Descriptive statistics, data exploration, correlation tests etc.	Null values impute, Outlier removal, test train validation split etc.	log reg, tree, boosting models on labelled dataset	log reg, tree, boosting models on unlabelled dataset	Results comparison, insight generation etc.

Performance Measure - F1 score

Test/Train/Validation Split - 10/80/10%

Assumptions - No data leakage

EDA Results

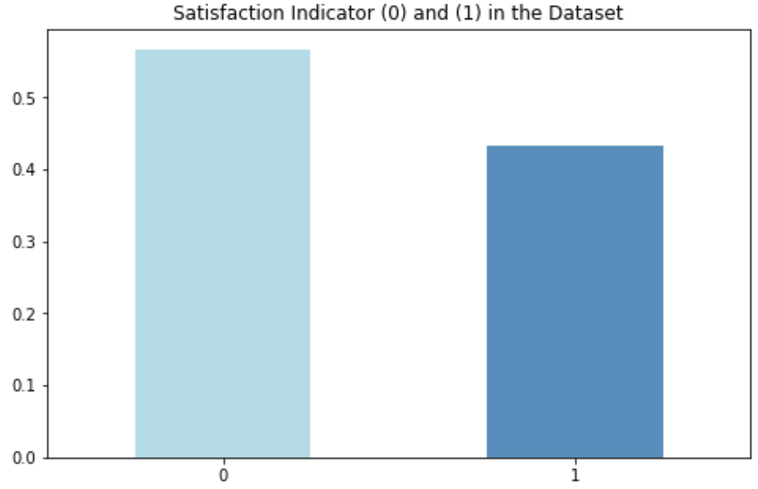
Dataset Overview

- Target Column - satisfaction (*satisfied: 1, neutral or dissatisfied: 0*)
- Passengers' ratings for flight services such as check-in, in-flight wifi, food & drinks, and seat comfort.
- Passengers' attributes such as age, gender, loyal/disloyal customer, business/personal travel.

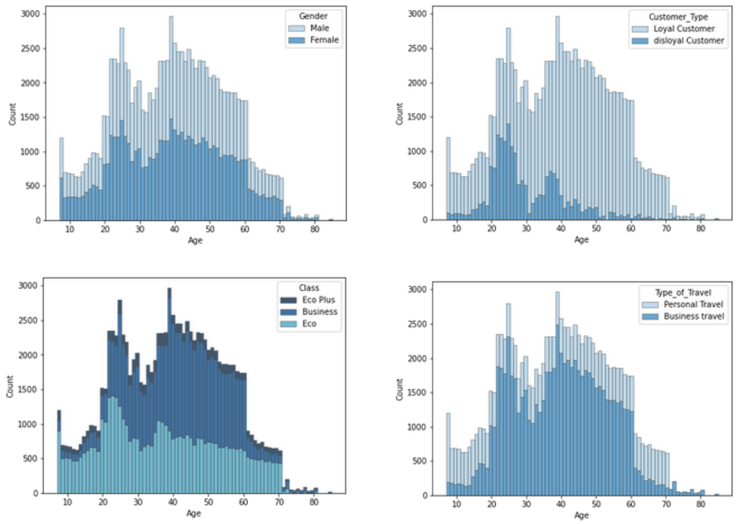
Data Preparation

- Treating categorical variables
  - Ordinal encoding for travel class
  - One-Hot Encoding - gender, customer type, type of travel
- Removed outliers using Isolation Forest
- Treated missing age values using SimpleImputer with median strategy

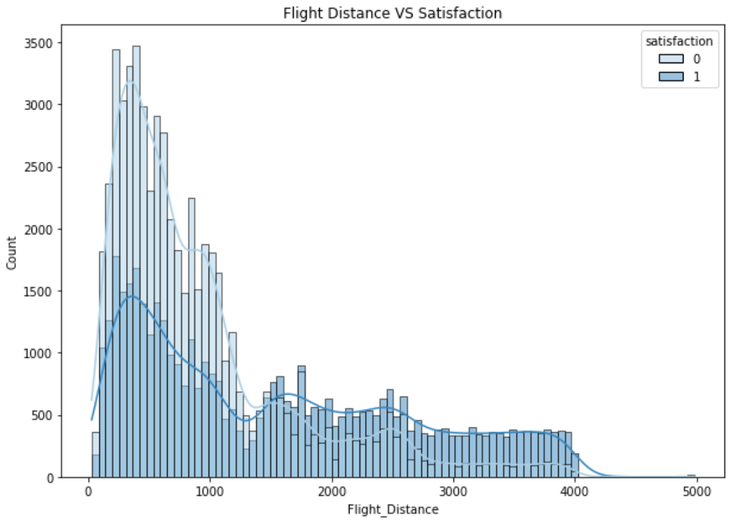
EDA - Visualizations



Distribution of Target Variable



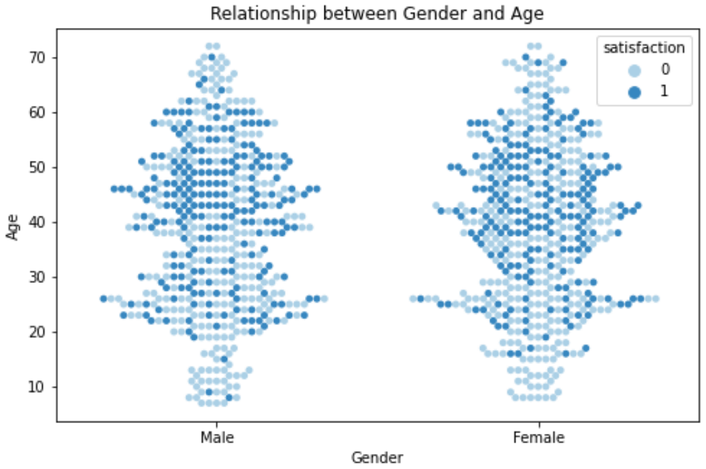
Relationship among categorical variables



Flight Distance vs Satisfaction

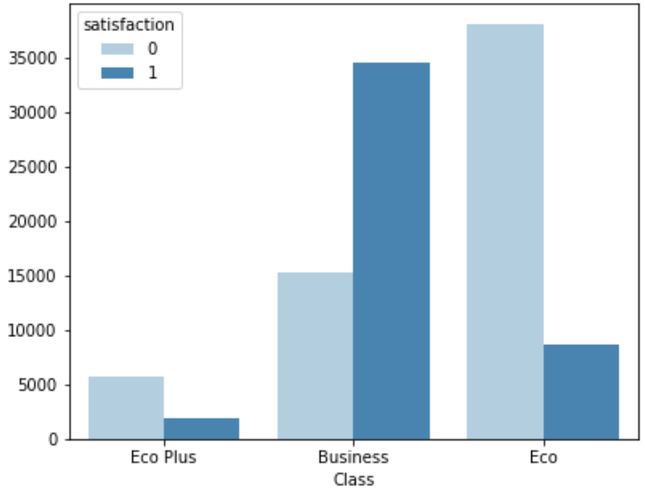
Data Insights

- Average satisfaction can vary based on age and gender
- Females are on average less satisfied than males with flying
- Idem for the oldest age group compared to others



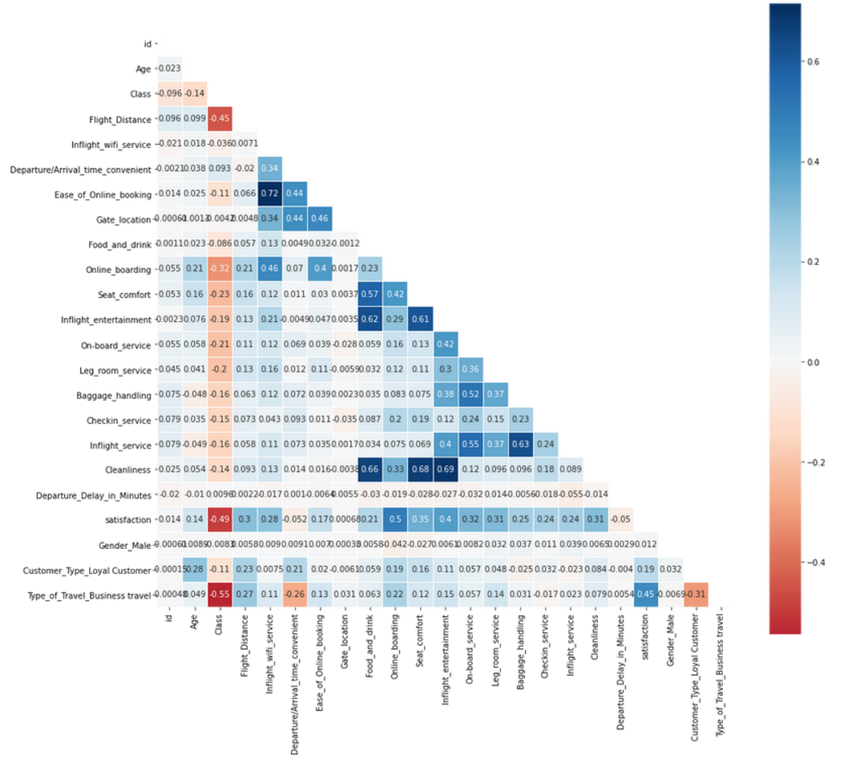
Females had an average satisfaction rate of 42.74% compared to 43.95% for males. Individuals aged between 40 and 60 years old are most satisfied.

- Idem for flying class and purpose of travel.
- Passenger satisfaction rates significantly increase when compared with the class they are flying in.
- One possible factor is that for business travel, flights are often subsidized by the company who chooses business class (hence the name) and thus passengers are more satisfied as they are essentially flying for free.



Business Class travelers were satisfied 69.43% of the time.
Passengers who flew for business similarly had a 58.26% satisfaction rate.

Correlation Matrix



- Negative correlation between distance and class
- Cleanliness being positively correlated with food, inflight entertainment and seat comfort
- Idem for inflight handling and baggage service
- Loyal customers being more likely to travel for business and upgrade their class
- Users who book online are more likely to use in flight Wi-fi and be dissatisfied by it

Modelling Approach

Pre-processing

- Scaling - Standard Scaler
- Missing Value Imputation - Custom Function
- Test Train Validation Split - 10/80/10% split
- **Label Propagation KNN**
- **Label Spreading KNN**

Feature Importances

- Recursive Feature Elimination (RFE)
- Lasso Regression

Model Building

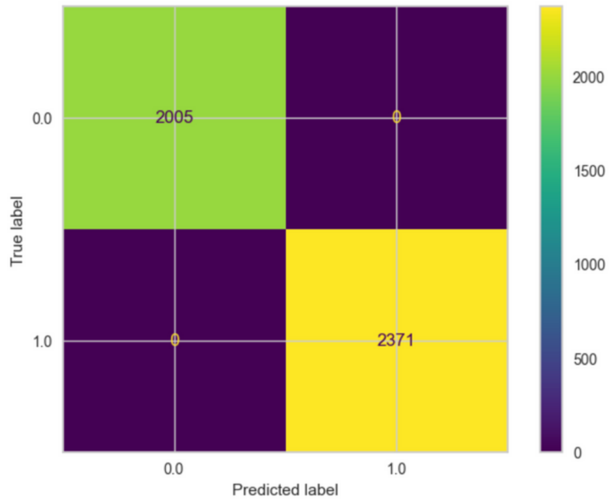
- Logistic Regression
- Random Forest
- Light GBM
- XGBoost
- AdaBoost
- Naive Bayes

Results Validation

- ROC curve, PR curve
- Classification score
- Confusion Matrix
- Precision, Recall, F1 Score

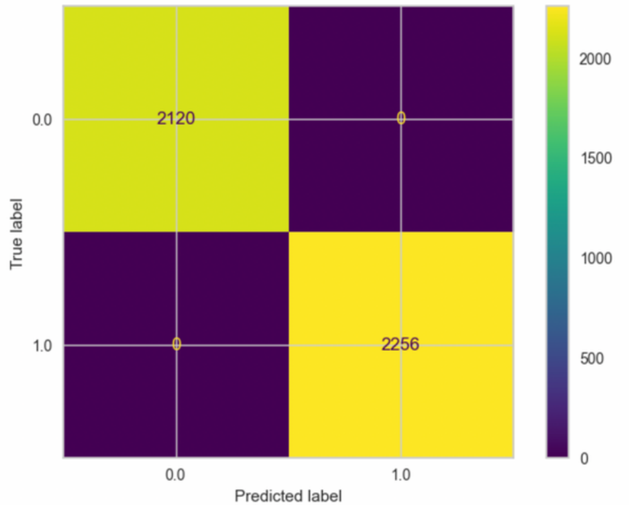
Results

Logistic Regression



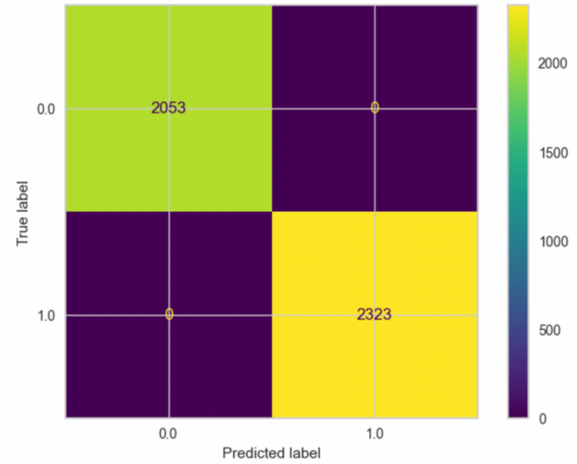
Precision : 93.71%
Recall: 94.41%

Naive Bayes



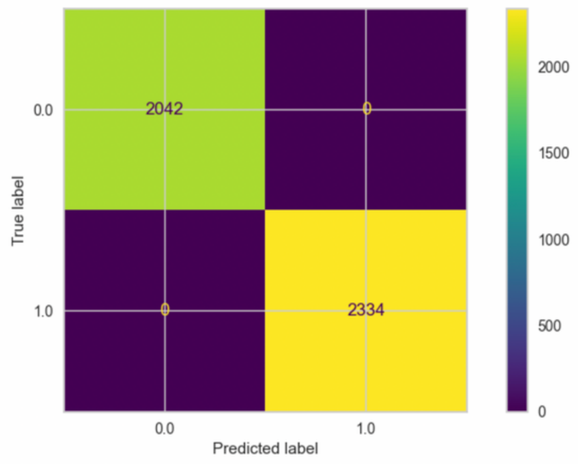
Precision : 94.01%
Recall: 90.44%

Random Forest



Precision : 98.14%
Recall: 97.22%

Light GBM



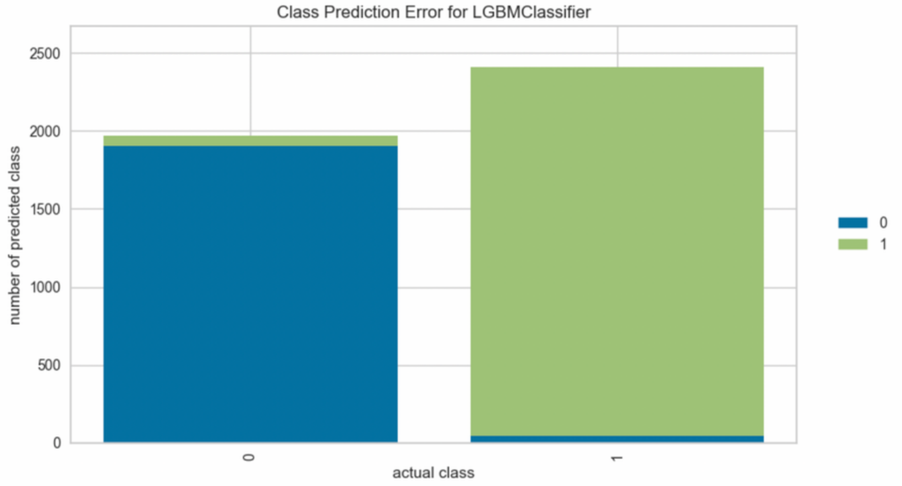
Precision : 98.11%
Recall: 97.65%

Best Model - Supervised Learning

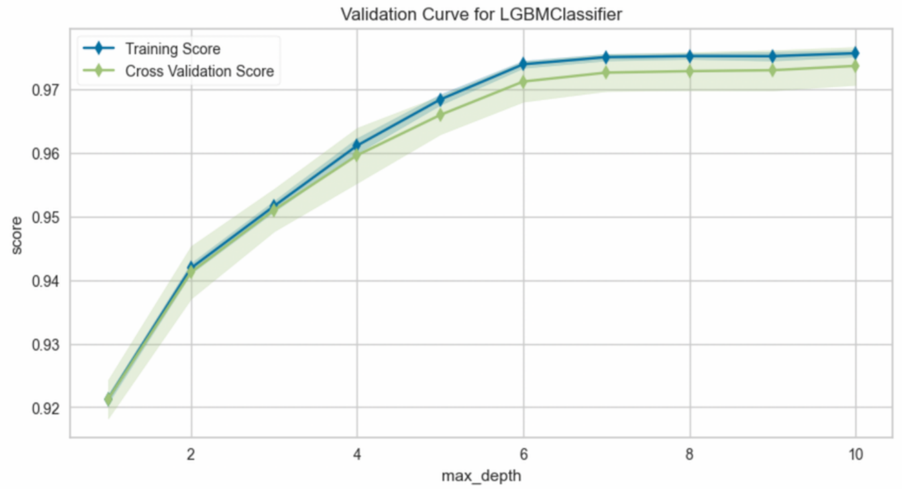
Results - LightGBM

	precision	recall	f1-score	support
DummyClassifier	0.535878	1.000000	0.697813	2345
LogisticRegression	0.933783	0.944136	0.938931	2345
KNN	0.968830	0.967591	0.968210	2345
Naive Bayesian	0.940160	0.904478	0.921973	2345
Decision Tree	0.973493	0.971002	0.972246	2345
Random Forest	0.981489	0.972281	0.976864	2345
LightGBM	0.981148	0.976546	0.978842	2345
SVM	0.961961	0.970576	0.966249	2345
AdaBoost	0.934379	0.965458	0.949664	2345

Based on the F1 score, LightGBM is the best model



LightGBM also performs really well on the test set



Supervised vs Semi-Supervised Learning

Semi-supervised Learning

	precision	recall	f1-score	support
DummyClassifier	0.535878	1.000000	0.697813	2345
LogisticRegression	0.933783	0.944136	0.938931	2345
KNN	0.968830	0.967591	0.968210	2345
Naive Bayesian	0.940160	0.904478	0.921973	2345
Decision Tree	0.973493	0.971002	0.972246	2345
Random Forest	0.981489	0.972281	0.976864	2345
LightGBM	0.981148	0.976546	0.978842	2345
SVM	0.961961	0.970576	0.966249	2345
AdaBoost	0.934379	0.965458	0.949664	2345
LabelPropagation	0.923297	0.965032	0.943703	2345
Label Spreading	0.941752	0.944563	0.943155	2345

Label Propogation is the best semi-supervised technique

Supervised Learning

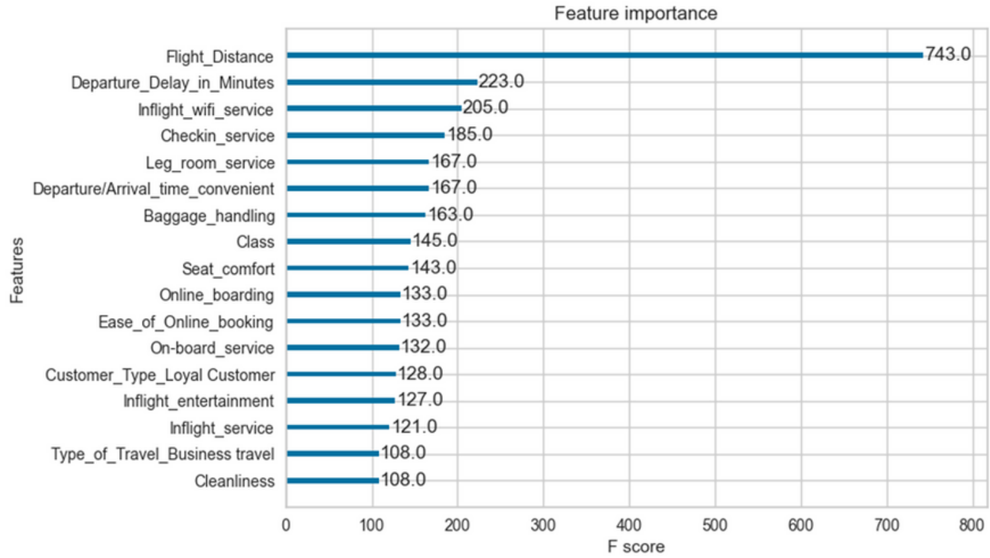
	precision	recall	f1-score	support
DummyClassifier	0.535878	1.000000	0.697813	2345
LogisticRegression	0.933783	0.944136	0.938931	2345
KNN	0.968830	0.967591	0.968210	2345
Naive Bayesian	0.940160	0.904478	0.921973	2345
Decision Tree	0.973493	0.971002	0.972246	2345
Random Forest	0.981489	0.972281	0.976864	2345
LightGBM	0.981148	0.976546	0.978842	2345
SVM	0.961961	0.970576	0.966249	2345
AdaBoost	0.934379	0.965458	0.949664	2345

LightGBM is the best supervised technique

Final Insights

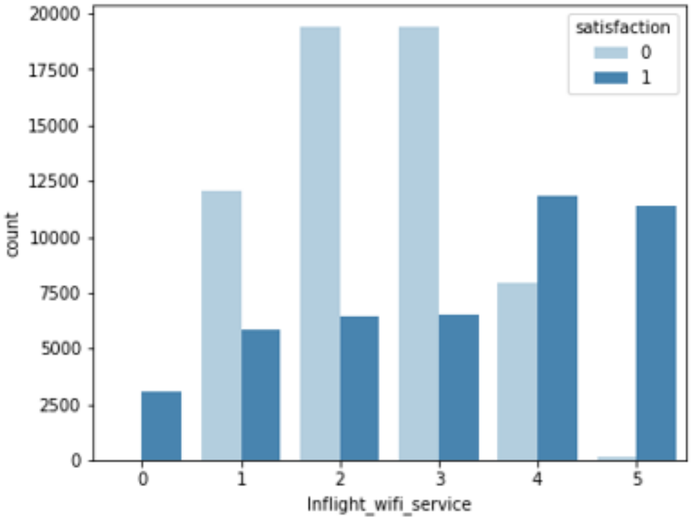
Feature Importance

We observe the feature importance of the most accurate model (LightGBM) and conclude that Flight_Distance, Departure_Delay_in_Minutes and Inflight_wifi_service are among the most important features to accurately predict the customer satisfaction.



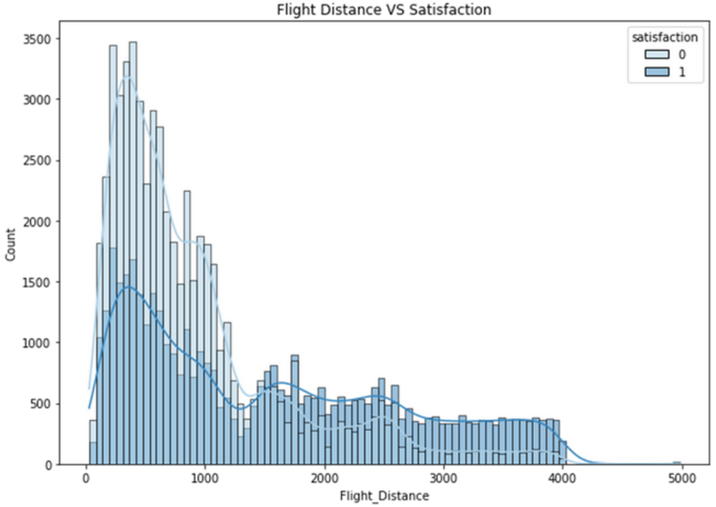
In-Flight Wifi Service

The customer satisfaction is increasing and dissatisfaction is decreasing as the WIFI service rating increase.



Flight Distance

Lower distance flights are where the most focus on increasing satisfaction should be made as these are where the most complaints are being generated.



Landing

- Quick Win - Provide better inflight WIFI
- Semi-supervised learning can save cost of expensive NPS programs
- Focus on operational efficiency increasing satisfaction on short distance flights