

Final Report

Predicting Airline Passenger Satisfaction



Background

A survey was conducted in 2015 among a hundred thousand respondents to track their level of satisfaction based on their inflight experience for a major airline carrier. For a long time, there has been a huge competition in the airline space to offer the best services at the lowest costs possible.

Problem Statement

Given that airlines need to balance budgets and cut costs to compete in this industry, the carrier conducting this study wants to identify the key variables that impact customer satisfaction and thereby optimize their marketing campaign to efficiently target consumer segments based on demographics, inflight experience, boarding experience and travel delays.

Criteria for Success

The CEO wants to target actionable variables that impact satisfaction the most. While airline delays do impact customer satisfaction, in some cases it's inevitable to avoid delays due to poor weather. The Criteria of success would be to classify the respondents' satisfaction with the highest Accuracy, Precision and Recall scores by building machine learning models and then predicting the out of sample dataset.

Data Sources

The train and test datasets are curated from the link here: [Kaggle](#)

The train dataset has values for the target variable, while the test dataset has all the variables in the dataset but the target. The goal is to build a Classification model on the Training dataset and predict the satisfaction of customers on the test dataset.

Raw Data

1. Target - Satisfaction

2. Categorical Variables

Variable Name	Notes
Gender	Female Male
Customer Type	Loyal Disloyal
Type of Travel	Business Travel Personal Travel
Class	Business Eco Eco Plus

3. Continuous Variables

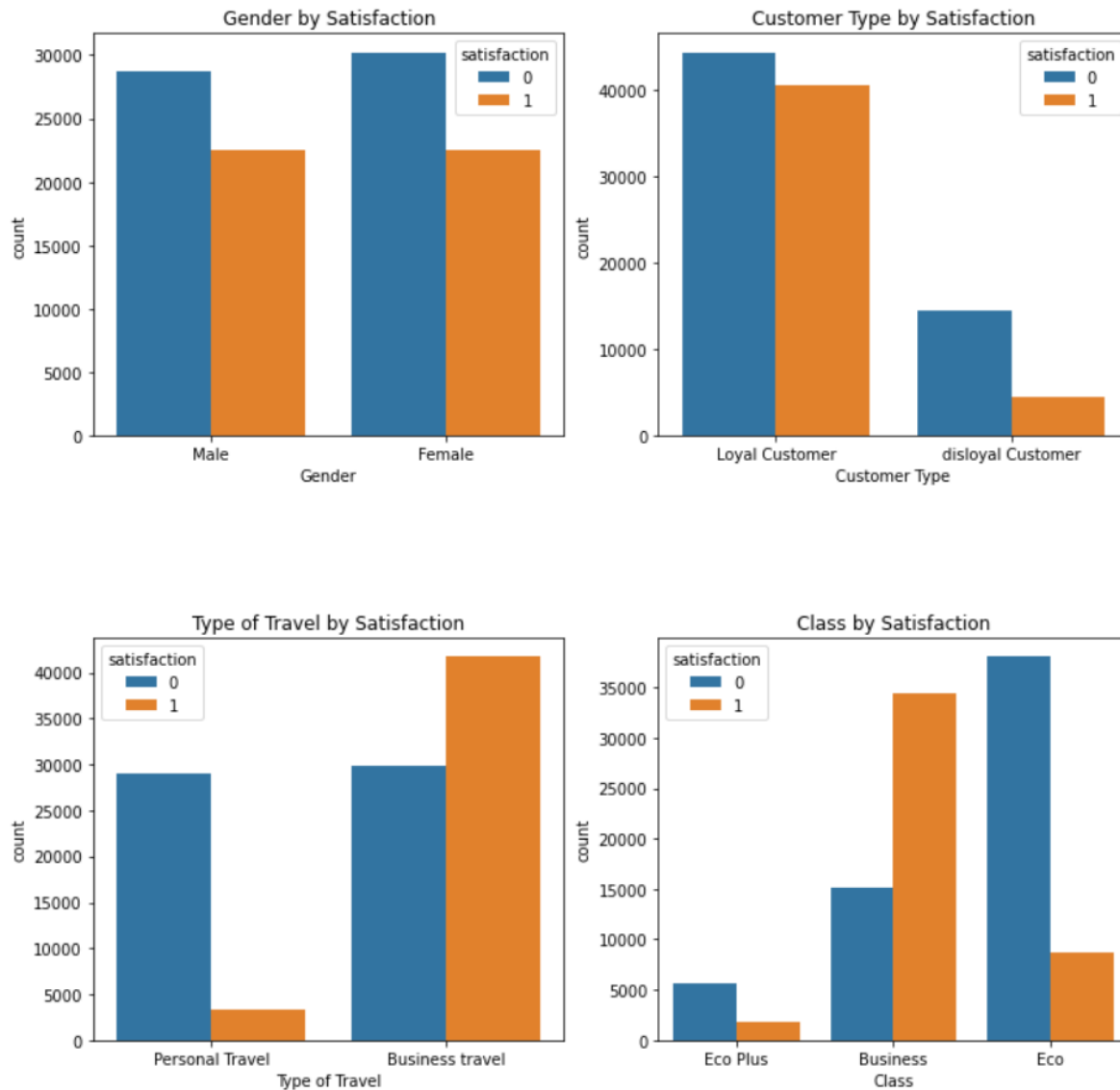
Variable Name	Notes
Age	Ranges from 7 years old to 85 years old
Flight Distance	Median Flight Distance is 414 miles
Departure Delay	75th Percentile Delay is 12 minutes
Arrival Delay	75th Percentile Delay is 13 minutes

4. Customer Experience - Survey Questions

	Very Unsatisfied	Unsatisfied	Nuetral	Satisfied	Very Satisfied
Inflight wifi service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Departure/Arrival time convenient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ease of Online booking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Gate location	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Food and drink	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Online boarding	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Seat comfort	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Inflight entertainment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On-board service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Leg room service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Baggage handling	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Checkin service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Inflight service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cleanliness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Exploratory Data Analysis

1. Categorical Variables



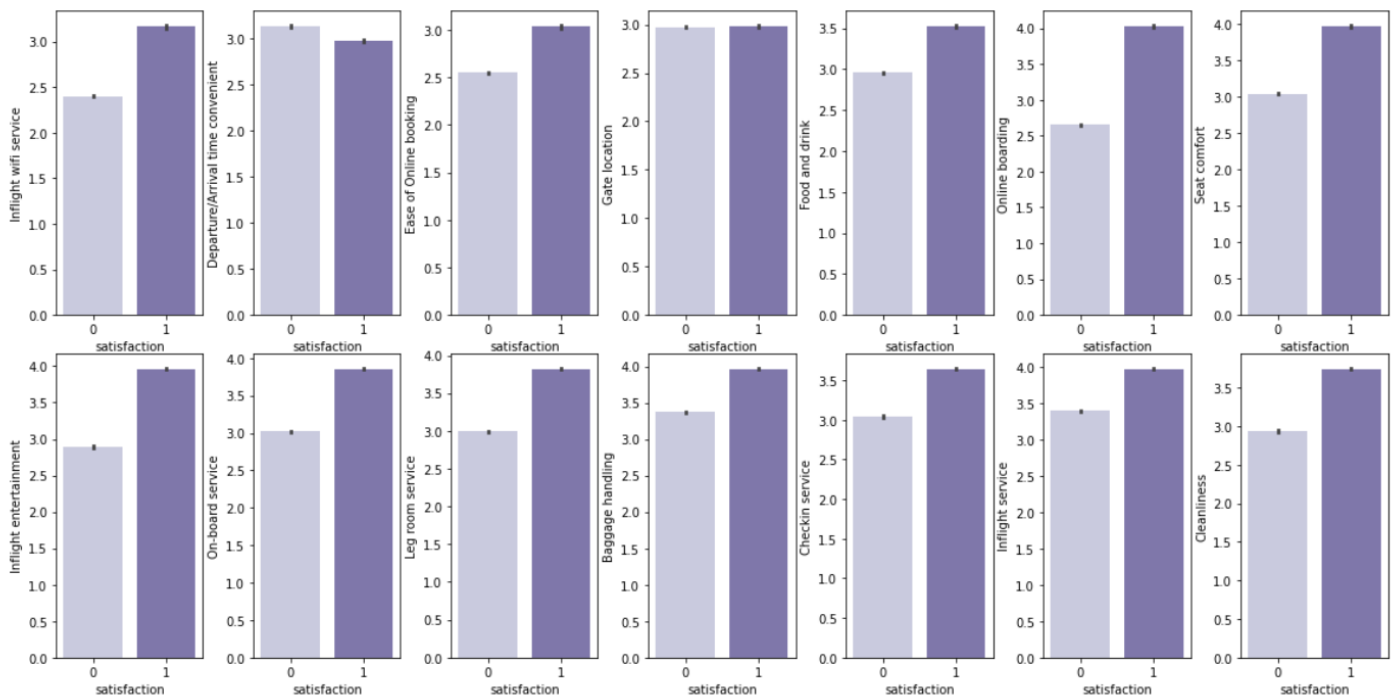
Gender - Satisfaction seems to be similar among males and females.

Customer Type - Loyal customers are much more satisfied than disloyal customers

Travel Type - Business class travellers are more satisfied than personal travellers

Class - Economy and economy plus passengers are not very satisfied.

2. Customer Experience:

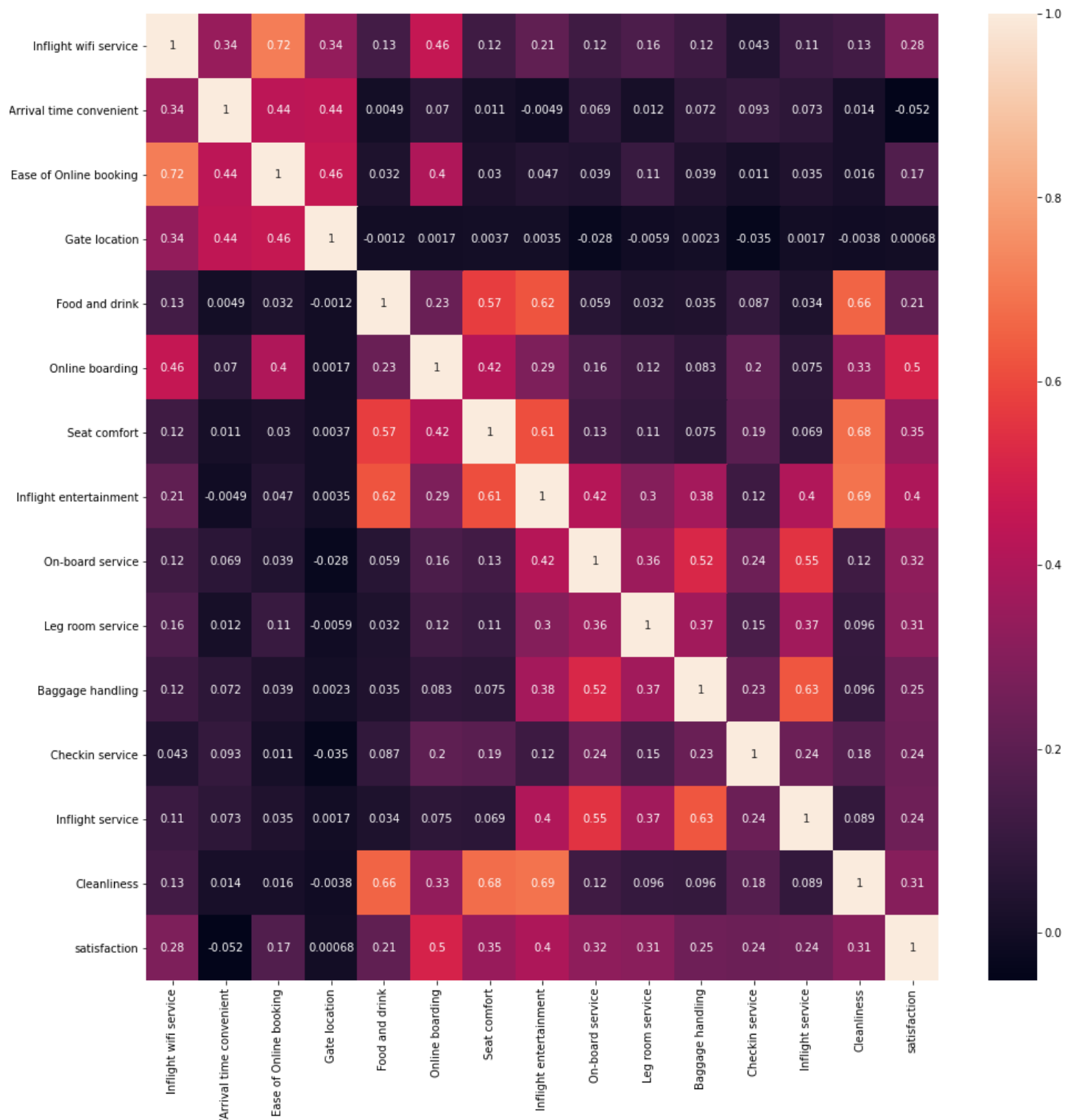


Analysing the results

There are some interesting findings here -- Most of the predictor variables seem to have a positive impact with increase in the survey response scores. **Gate location** does not seem to have any positive relation with satisfaction. **Departure/Arrival time convenient** is negatively correlated with satisfaction. That doesn't make much sense and it would seem to be a random chance that there is a skew towards the negative satisfaction class.

We should also be cognizant of the fact that some of these variables will be highly correlated with each other. So we should run some correlations to ensure that multicollinearity is not an issue.

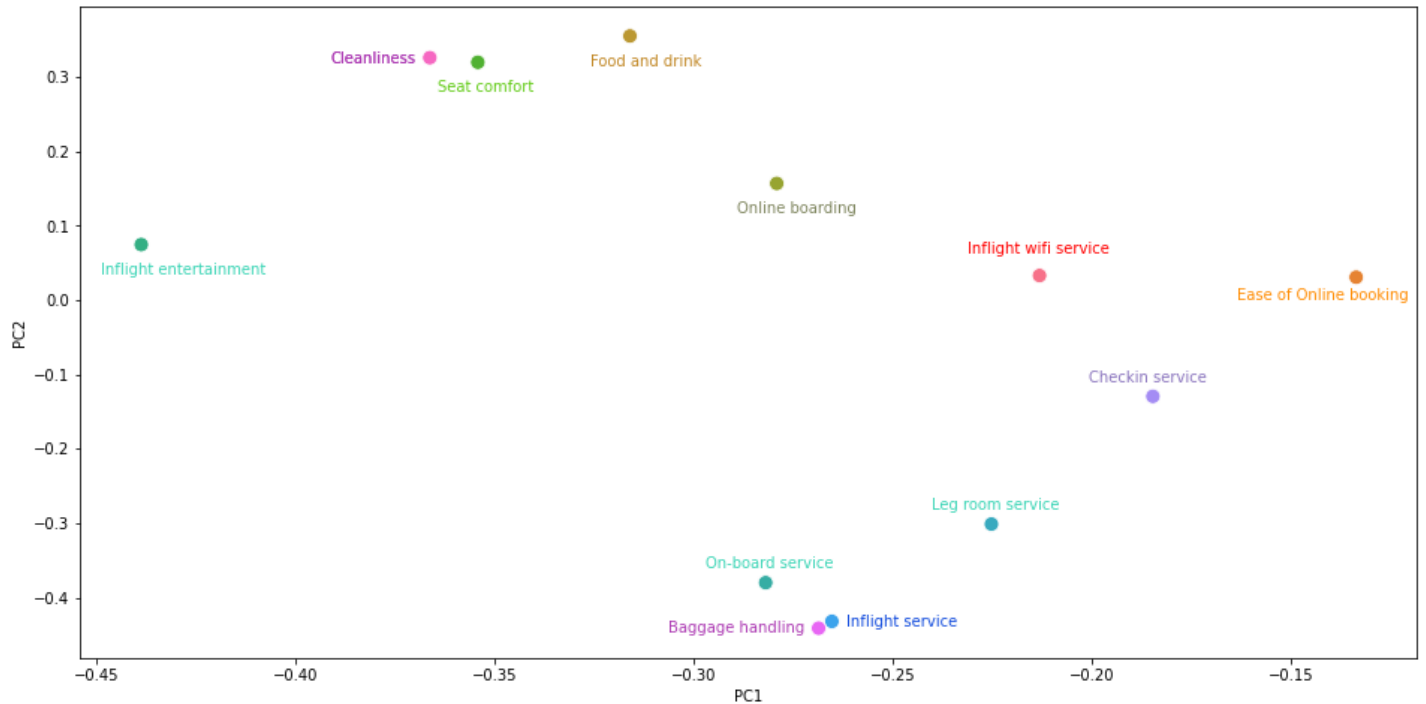
Correlations



Analysis of Correlations / Heatmap:

As you can see from the correlations, Gate location and Departure/Arrival time convenient have no correlation with customer satisfaction.

Principal Component Analysis



As you can see from the plot, Baggage handling and On-board service are close to each other. This makes sense. Cleanliness, Seat Comfort and Food and Drink appear to be close to each other

Summary of Exploratory Data Analysis

I've approached the EDA by analyzing Categorical, Ordinal and Continuous data separately.

- Categorical data
 - Based on the bar plots, the mean of Satisfaction by gender seems to be fairly same
 - Class, Type of Travel and Customer Type all have different satisfaction means so would be important predictors
- Ordinal Data
 - Most of the variables show a positive correlation with satisfaction. Gate location and Departure \ Arrival time convenient does not have a strong correlation with satisfaction.
 - There is some multi-collinearity between some of the variables. This can be taken care of after running a Random Forest algorithm during the modeling phase.
- Continuous Variables
 - Age seems to be normally distributed.
 - Flight Distance indicates that there are more short flights than long-haul / international flights
 - Most of the data indicates that most flights are on time. Some flights have extraordinarily long flight delays and some transformation such as log transformation should be done

Preprocessing Data

1. Data Engineering Pipeline - [Source](#)

❖ Categorical Pipeline

- ColumnSelector
- ModelImputer
- OneHotEncoder

❖ Ordinal Pipeline

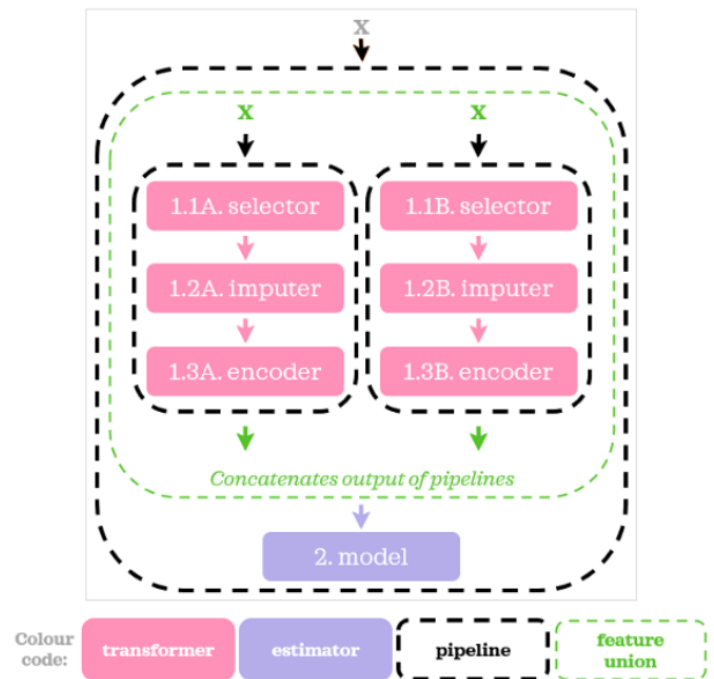
- ColumnSelector
- SimpleImputer → 0
- MinMaxScaler

❖ Age Distance Pipeline

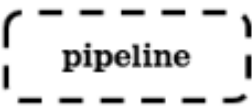

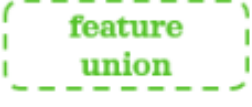
- ColumnSelector
- SimpleImputer → mean
- StandardScaler

❖ Wait Time Pipeline

- ColumnSelector
- SimpleImputer → 0
- PowerTransformer



2. Feature Union

	What does it do?	When to use it?
	Apply a series of transformers sequentially and then a final estimator.	When building machine learning pipeline that transforms the data then predicts.
	Apply different transformers to different subsets of columns in parallel and concatenate the output of these parallel transformations.	When different data transformations are to be applied to different subsets of columns . Use together with Pipeline.
	Apply different transformers on the same input data in parallel and concatenate the output of these parallel transformations.	When different data transformations are to be applied to on the same input data . Use together with Pipeline.

Modeling

- ❖ Dummy Classifier
 - Accuracy → 56.36%
- ❖ Simple Logistic Regression
 - Accuracy → 87.76%
- ❖ Tuned Logistic Regression
 - C: 1.0
 - Max_iter: 1000
 - Penalty: Ridge (L2)
 - Accuracy → 92.77%
- ❖ Tuned KNN Classifier
 - N = 3
 - Accuracy → 89.71%
- ❖ Tuned Random Forest
 - max_depth: 30
 - min_samples_split: 5
 - n_estimators: 500
 - Accuracy → 96.07%

Modeling Summary:

The classifiers used in this model are Logistic Regression, K-Nearest Neighbors and Random Forests. As you can see from the results above, all three classifiers have a test accuracy of more than 87% and the modeling variables capture the signal well. However, Random Forests are good at Feature Selection and as a result remove some redundant features during modeling.

Hyper parameter tuning is an important step, and the test accuracy score has improved by 5% for a tuned logistic regression model compared with the base model. The tuning is done using a Grid Search algorithm using Cross validation and is performed completely on the X_train dataset. To sum, we've fit our model on a 80% training dataset, tuned hyper parameters solely on the training sample and predicted accuracy scores on the hold out 20% of data.