Springboard – DSC
Capstone Project 2

What Separates “Satisfied” Airline Passengers From “Neutral or Dissatisfied” Passengers?

Lauren Sweeney
July 2022

Introduction

Problem Statement

Our client, Super Important Airline, Inc. (hereinafter, "SIA"), is one of the major airlines of the United States and is one of the world's oldest airlines in operation. SIA operates over 5,400 flights daily and serves 325 destinations in 52 countries in six continents.

SIA is ranked third among the world's largest airlines by number of scheduled passengers carried and fleet size (Delta Air Lines, one of SIA's main competitors, is ranked second).

SIA is looking to increase revenue, and as a result, their board of directors has requested that we analyze survey data from their airline passengers. What separates SIA's "satisfied" airline passengers from their "dissatisfied or neutral" passengers? This project will focus on finding ways to characterize the drivers of the two satisfaction categories being studied using machine learning models and by analyzing the results the models will yield.

Criteria for Success

SIA wants to know which variables impact the satisfaction of passengers the most. While some variables that negatively impact satisfaction may ultimately be unavoidable (for example, flight distance, departure delays, and arrival delays), the other listed variables may have a positive impact on passenger satisfaction. We will determine which variables are the most important for passenger satisfaction by calculating the highest Accuracy, Precision, and Recall scores by building machine learning models.

Data Sources

The dataset can be found via this link: Kaggle

It includes data from over 120,000 airline passengers, including additional information about each passenger, their flight, and type of travel. It also considers each individual passenger's evaluation of different factors like the cleanliness of the airplane, comfort, service provided, experience, and overall satisfaction.

Raw Data

Target variable – Satisfaction

Variable Name:	Information re Variable:
Gender	Male | Female
Customer Type	First-Time | Returning
Type of Travel	Business | Personal
Class	Business | Economy | Economy Plus
Age	Numerical ranging from 7 to 85
Flight Distance	Numerical

Departure Delay	Numerical
Arrival Delay	Numerical

Passenger Survey Questions:

The passenger survey questions had numerical answers ranging from 0 (least satisfied) to 5 (most satisfied).

Variable Name:	Information re Variable:
Departure and Arrival Time Convenience	Numerical ranging between 0 and 5
Ease of Online Booking	Numerical ranging between 0 and 5
Check-in Service	Numerical ranging between 0 and 5
Online Boarding	Numerical ranging between 0 and 5
Gate Location	Numerical ranging between 0 and 5
On-board Service	Numerical ranging between 0 and 5
Seat Comfort	Numerical ranging between 0 and 5
Leg Room Service	Numerical ranging between 0 and 5
Cleanliness	Numerical ranging between 0 and 5
Food and Drink	Numerical ranging between 0 and 5
In-flight Service	Numerical ranging between 0 and 5
In-flight Wifi Service	Numerical ranging between 0 and 5
In-flight Entertainment	Numerical ranging between 0 and 5
Baggage Handling	Numerical ranging between 0 and 5

Brief Summary of Results

We reviewed and cleaned the dataset, performed exploratory data analysis, and created four models. Based on this analysis, we were able to discern the best model for our client's needs. Since SIA can better tolerate false negatives than false positives, we decided to focus on the precision of each model. As a result, we determined that the best model for our client was the Random Forest with random under sampling. The most important features for that model included online boarding, in-flight WIFI service, class_business, type of travel_personal, type of travel_business, in-flight entertainment, ease of online booking, seat comfort, class_economy, and leg room service.

For further information and details please see the notebooks developed for this project. The link can be found here: https://github.com/LCSDataSci/Airline-Customer-Satisfaction-Capstone

Data Wrangling

The first step of this project is to clean the data to make it easier to analyze and understand. Fortunately, our data has only unique values. As a result, we did not have to drop any duplicate records.

Next, we want to determine whether our dataset had any missing values. Unfortunately, Arrival Delay had 393 missing values. We could either remove those values, or fill them in. In order to determine what we want to do with our missing values in Arrival Delay, I decided to determine the median value:

	count	mean	std	min	25%	50%	75%	max
Age	129880.0	39.427957	15.119360	7.0	27.0	40.0	51.0	85.0
Flight Distance	129880.0	1190.316392	997.452477	31.0	414.0	844.0	1744.0	4983.0
Departure Delay	129880.0	14.713713	38.071126	0.0	0.0	0.0	12.0	1592.0
Arrival Delay	129487.0	15.091129	38.465650	0.0	0.0	0.0	13.0	1584.0

The above figure demonstrates that the median value for Arrival Delay is 0.0 minutes. As a result, it is safe to fill in the 393 missing values with 0.

Next, I wanted to determine what percentage of passengers were Satisfied versus Neutral or Dissatisfied:

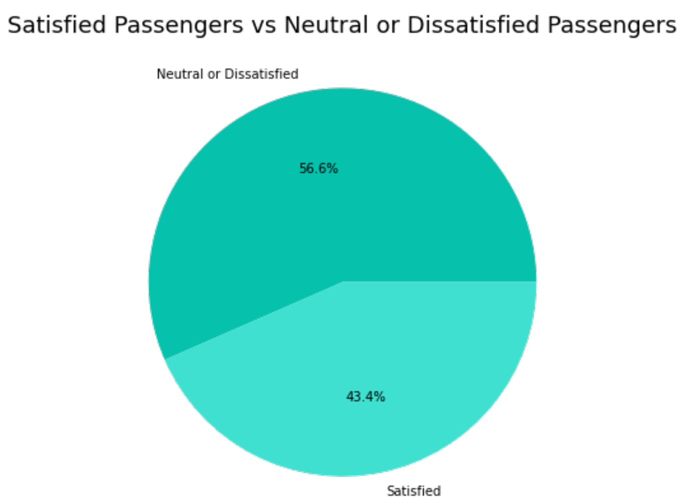
```
Neutral or Dissatisfied    73452
Satisfied                  56428
```

As you can see, of the 129,880 passengers surveyed, approximately 43.4% were Satisfied and 56.6% were Neutral or Dissatisfied.

Exploratory Data Analysis

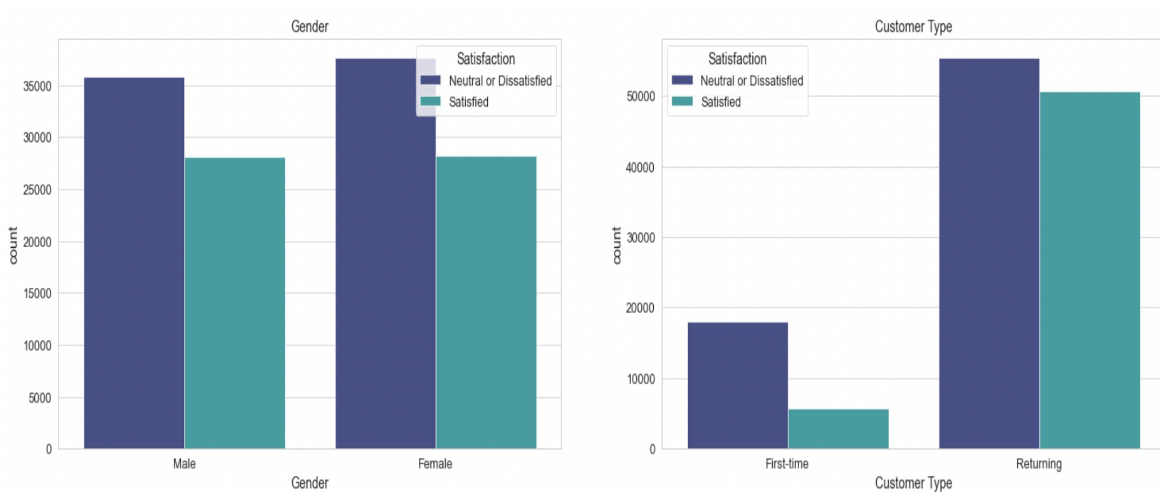
After cleaning and wrangling the dataset, the next step is to perform exploratory data analysis. This means that we want to perform initial investigations on the cleaned data to discover patterns and check hypotheses with the help of statistics and graphical representations.

In particular, we want to determine the relationship, if any, between the target "Satisfaction" and the predictor variables.



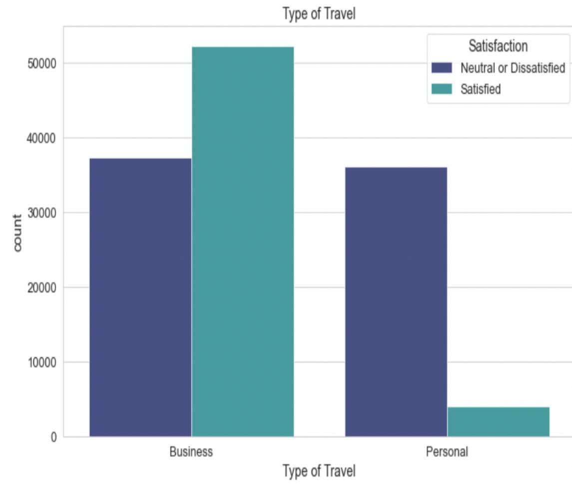
As previously discussed, the dataset is split fairly evenly between Satisfied and Neutral or Dissatisfied passengers. But what separates the two? We need to dig deeper. We decided to create countplots to help visualize the number of passengers who were Satisfied versus Neutral or Dissatisfied.

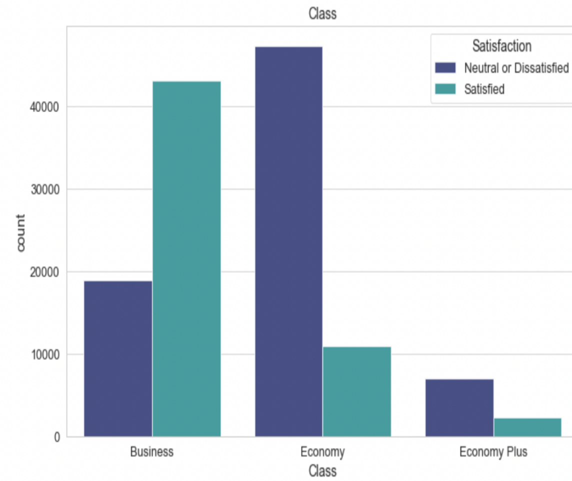
The first two countplots show the differences between Gender and Customer Type:



Here, we can see that the satisfaction rates between male and female passengers were similar. The above also illustrates that returning passengers were more satisfied as a whole than first-time passengers.

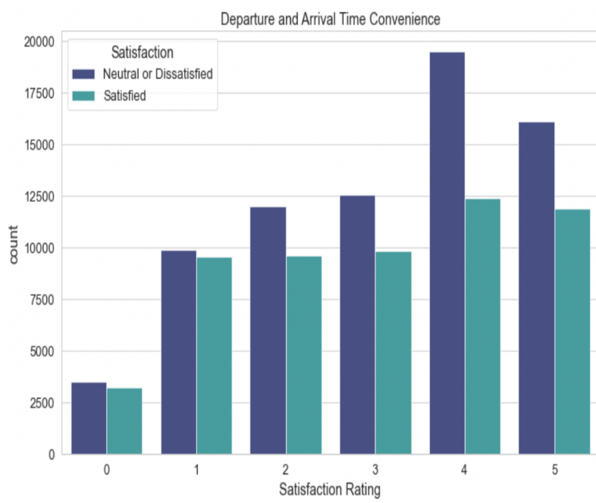
The next two countplots show the differences between Type of Travel and Class:

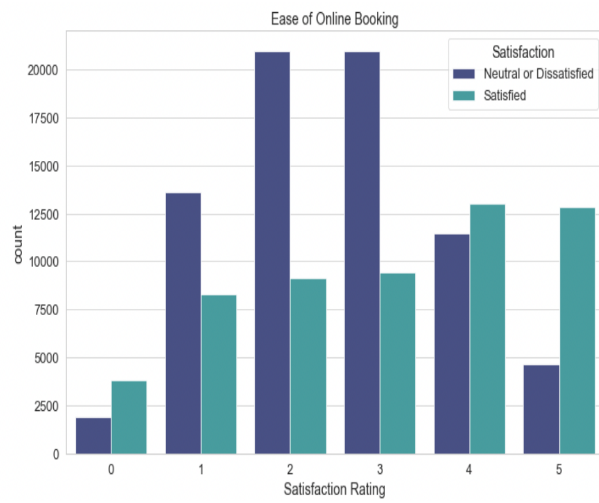




Here, we can see business travelers were more satisfied than personal travelers. Similarly, those traveling in business class were more satisfied than those traveling in economy class or economy plus class.

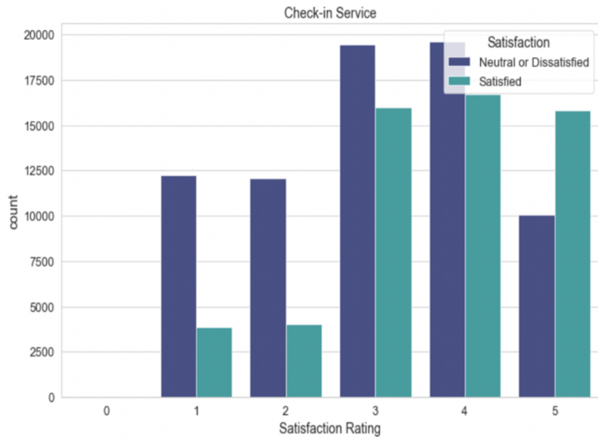
The next two countplots show the differences in survey results for the variables Departure and Arrival Time Convenience and Ease of Online Booking:

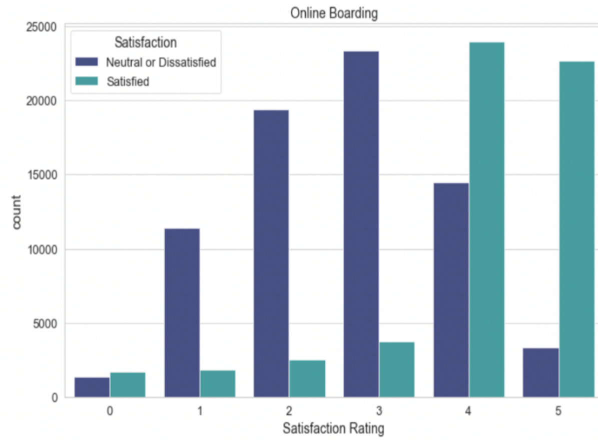




Here, we can see that more satisfied passengers rated the departure and arrival time convenience highly. This is also shown in the next plot, where again more satisfied passengers rated the ease of online booking highly. Surprisingly, neutral or dissatisfied passengers also tended to rate the departure and arrival time convenience highly. Unsurprisingly, those same passengers gave lower ratings for ease of online booking.

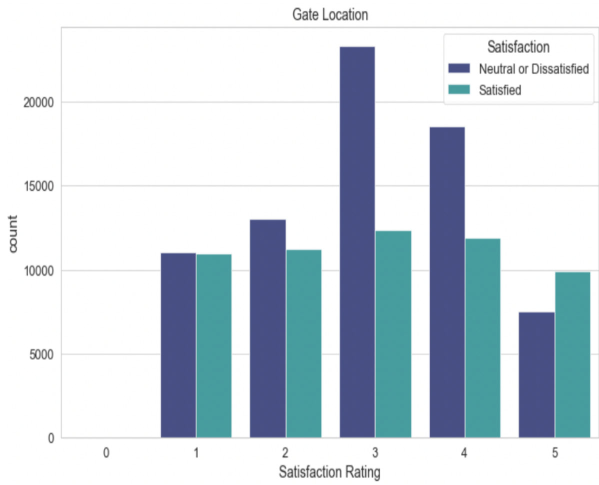
The next two countplots show the differences in survey results for the variables Check-in Service and Online Boarding:

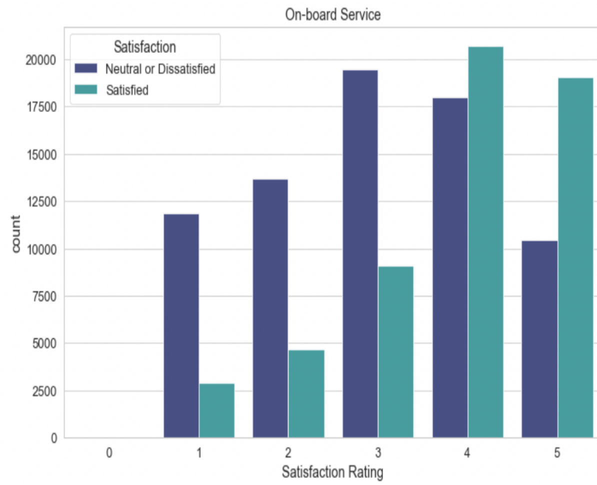




Here, we can see that satisfied passengers rated the check in service highly. Almost all of the satisfied passengers gave high ratings to the online boarding process as well.

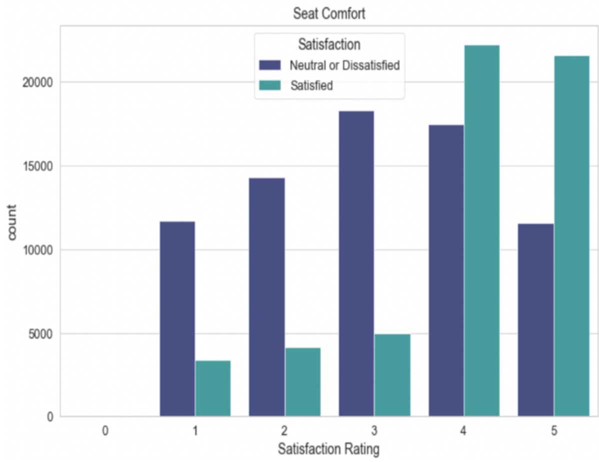
The next two countplots show the differences in survey results for the variables Gate Location and On-board Service:

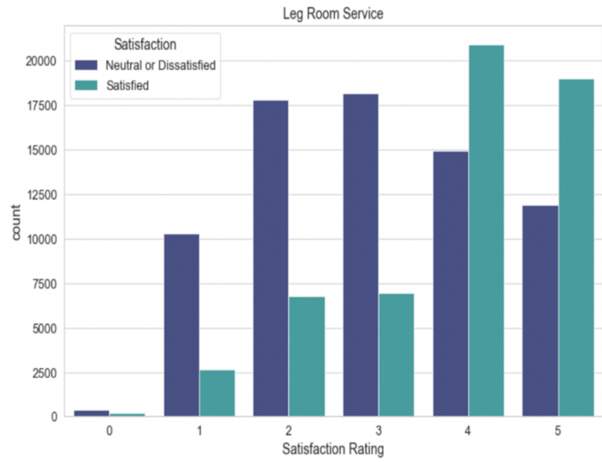




Here, we can see that satisfied passengers were split evenly in the view of the gate location, which ratings fairly evenly split between ratings one through five. However, satisfied passengers gave very high ratings to the on-board service.

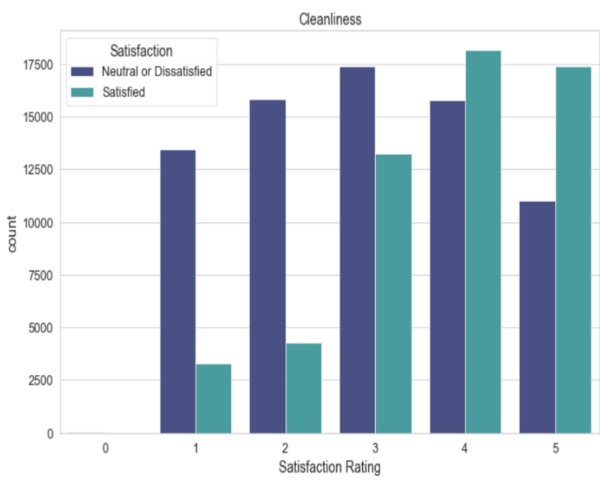
The next two countplots show the differences in survey results for the variables Seat Comfort and Leg Room Service:

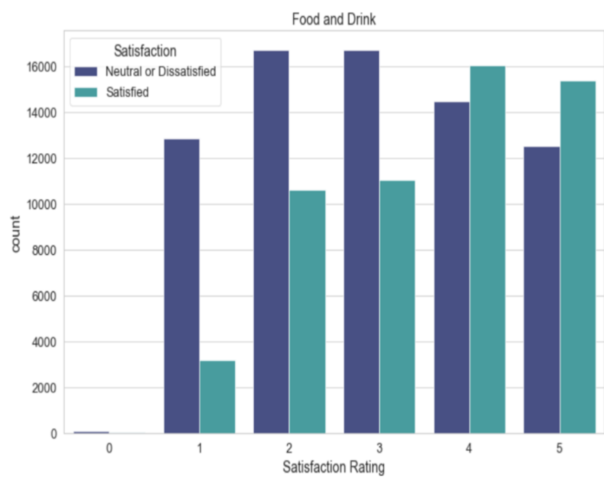




Here, we can see that satisfied passengers rated their seat comfort highly. Satisfied passengers also rated the leg room service highly. Neutral or dissatisfied passengers had their satisfaction ratings evenly split amongst both countplots.

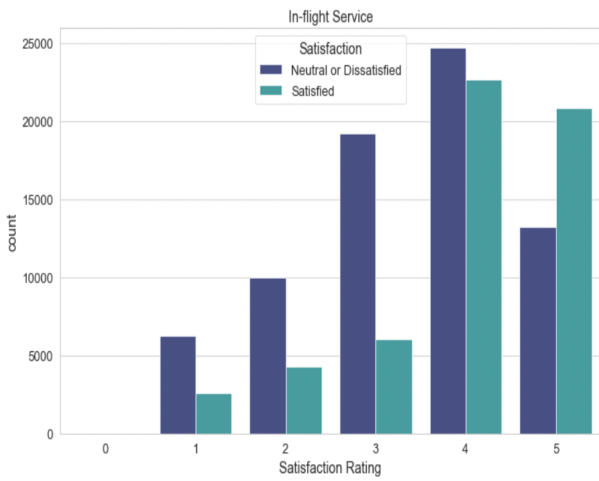
The next two countplots show the differences in survey results for the variables Cleanliness and Food and Drink:

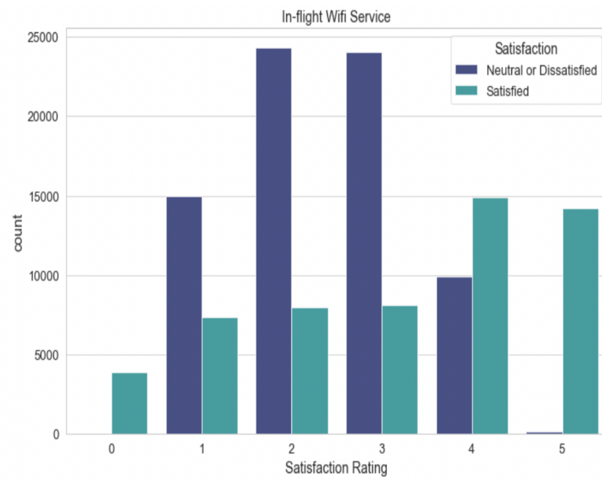




Here, we can see that satisfied passengers rated the cleanliness of the planes highly. They also rated the food and drink highly. And like before, neutral or dissatisfied passengers had their satisfaction ratings evenly split amongst both countplots.

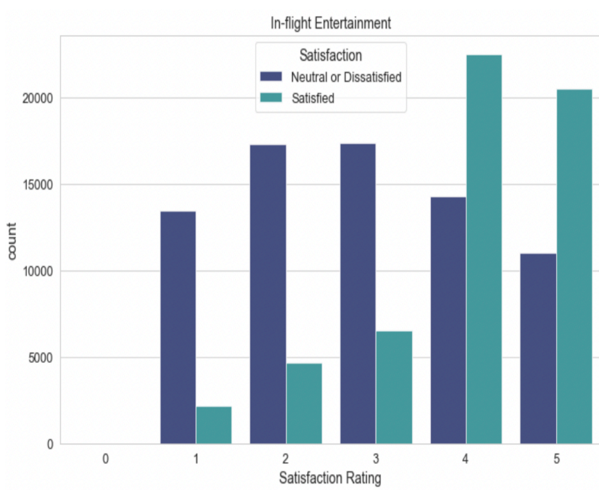
The next two countplots show the differences in survey results for the variables In-flight Service and In-flight Wifi Service:

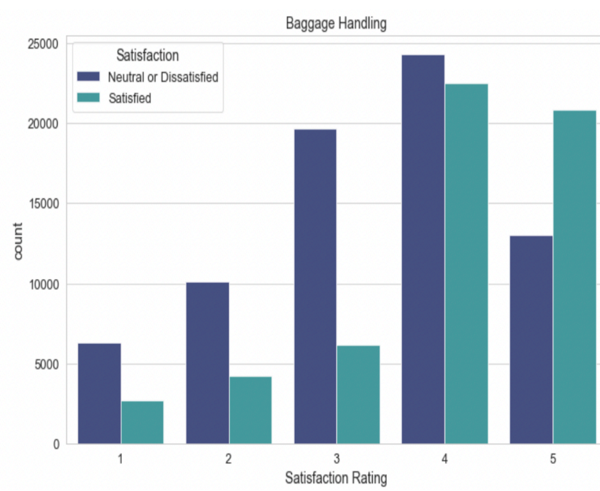




Here, we can see that satisfied passengers rated the in-flight service highly. However, satisfied passengers were split fairly evenly in their ratings of the in-flight WIFI service. Interestingly, neutral or dissatisfied passengers rated the in-flight service highly and the in-flight WIFI service poorly.

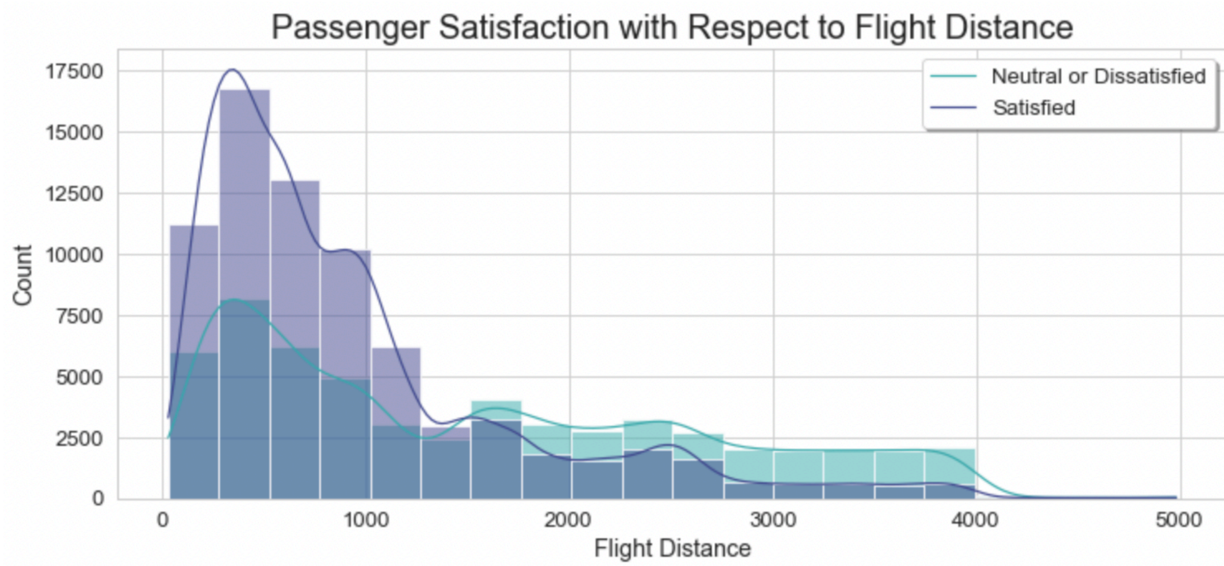
Finally, the last two countplots show the differences in survey results for the variables In-flight Entertainment and Baggage Handling:





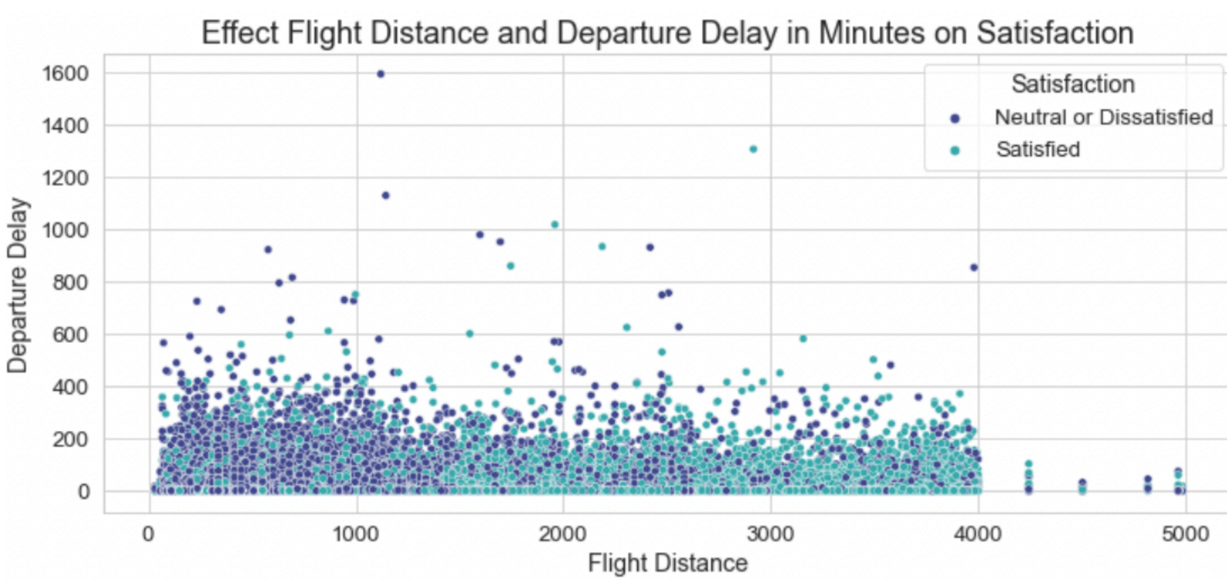
Here, we can see that satisfied passengers rated the in-flight entertainment highly. Satisfied passengers also rated the baggage handling highly. Neutral or dissatisfied passengers had their satisfaction ratings evenly split regarding the in-flight entertainment countplot, but rated the baggage handling highly.

We also wanted to use a scatterplot to determine whether a passenger's satisfaction was affected by the distance of their flight:



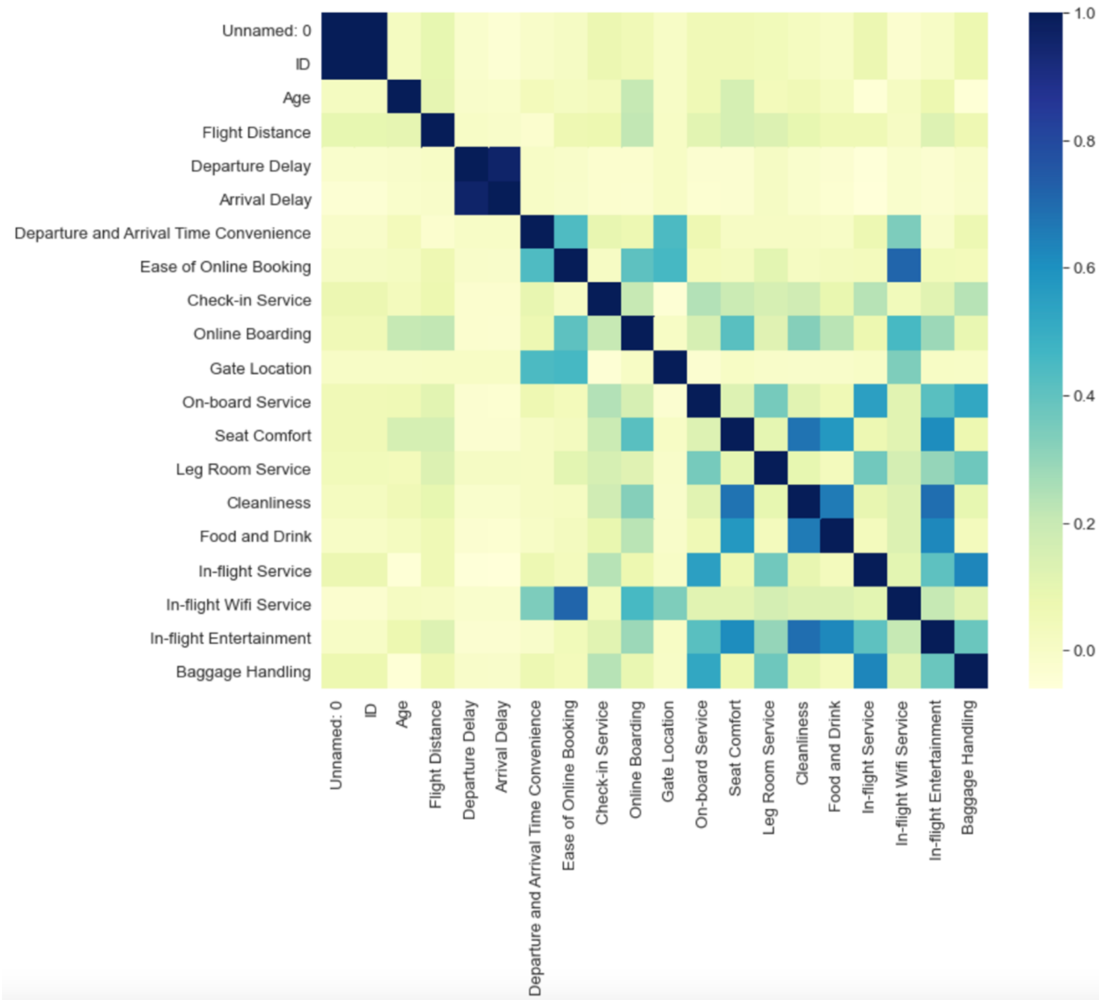
Here, we can see that as the flight distance increases, the satisfaction of passengers decreases. The plot is represented by a normal, right-skewed distribution.

We can also use a scatterplot to determine the effect the flight distance and departure delay has on passenger satisfaction:



Here, we can see that as the flight distance increases, passengers are still overwhelmingly satisfied as long as the departure delay is minimal.

Finally, we decided to create a heatmap. Heatmaps are graphical illustrations of data. This heatmap shown below is a correlation heatmap, meaning that it graphically depicts the strength of relationships between features. Darker colors will depict a stronger correlation, while lighter colors depict a weaker correlation.



Based on the above heatmap, it appears that the variables which led to the highest numbers of satisfied passengers were the following:

- Check-in service
- Online boarding
- On-board service
- Seat comfort
- Leg room service
- Cleanliness
- Food and drink
- In-flight service
- In-flight entertainment
- Baggage handling

Model Selection

Pre-Processing and Training the Data

Before we can select the best model for our client's needs, we must first pre-process and train the data.

To use categorical variables in a machine learning model, we first need to represent them in a quantitative way. We will need to use dummy variables. There are four categorical columns in our data: Gender, Customer Type, Type of Travel, and Class. Each will need to be converted to a dummy variable before we split the data into a training and testing set.

After creating the dummy variables, we were ready to perform a stratified train/test split of the dataset. Stratification is done for classification machine learning problems to avoid overfitting or underfitting, as it helps to ensure that the target variable (in this case, "Satisfaction") will have the same distribution in our training data.

Now we constructed a Logistic Regression model on both the testing and training sets. After doing so, we determined the accuracy score of each:

```
[Test] Accuracy score (y_predict_test, y_test): 0.7737277009807707


[Test] Accuracy score: (y_test, y_predict_test) 0.7737277009807707


[Training] Accuracy score: (y_train, y_predict_training) 0.777327890556045
```

Both the training accuracy and testing accuracy were very close, meaning that there was no "variance." However, the model's training accuracy was below 100%, indicating that there was some bias in this model.

We also created a classification report for both the training and testing sets:

```
[Training Classification Report]
                        precision    recall  f1-score   support

Neutral or Dissatisfied       0.81      0.80      0.80     51257
              Satisfied       0.74      0.75      0.75     39383

               accuracy                           0.78     90640
              macro avg       0.77      0.77      0.77     90640
           weighted avg       0.78      0.78      0.78     90640

[Test Classification Report]
                        precision    recall  f1-score   support

Neutral or Dissatisfied       0.81      0.79      0.80     21968
              Satisfied       0.73      0.75      0.74     16879
```

accuracy			0.77	38847
macro avg	0.77	0.77	0.77	38847
weighted avg	0.77	0.77	0.77	38847

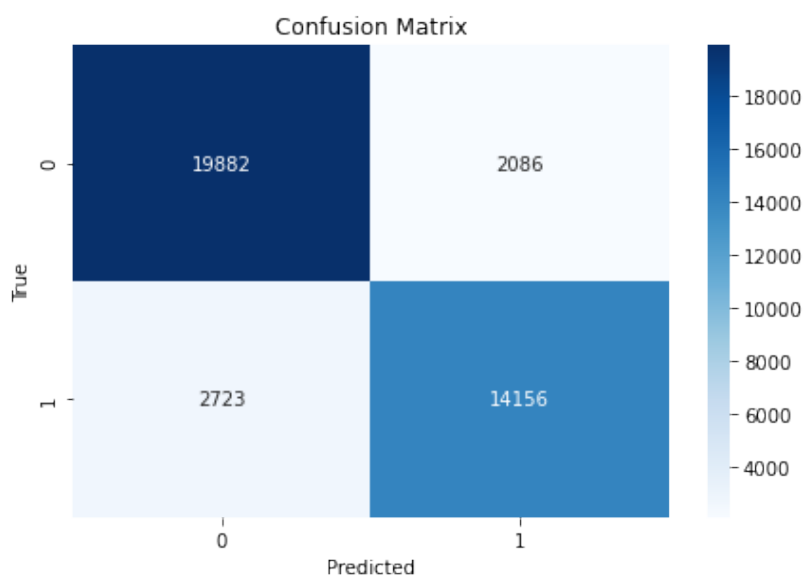
The classification report details the precision, recall, and f1-scores for both the training and testing datasets. Precision refers to the number of positive class predictions that belong to the positive class, while recall refers to the number of positive class predictions made out of all the positive examples in the entire dataset. The f1-score refers to the mean between precision and recall, and generally speaking, higher f1-scores are better.

So what is the above illustrating? The precision between both datasets is the same for Neutral or Dissatisfied, while the training dataset did a little better with our target variable, Satisfied. Furthermore, we can also see that the recall is slightly better for the training dataset with respect to Neutral or Dissatisfied.

Additionally, we can see that the precision and recall drop within each class. The training classification report details that the precision of Neutral or Dissatisfied is 0.81, while the precision of Satisfied is only 0.74. similarly, the test classification report details that the precision of Neutral or Dissatisfied is 0.81, while the precision of Satisfied is 0.73.

This report also shows how much support there is for each variable. As expected, the training dataset is much larger than the testing dataset.

Next, we created a confusion matrix:



Confusion matrices are useful for measuring recall, precision, and accuracy. It is a visual representation of actual versus predicted values. There are four elements of a confusion matrix:

True Positive – the values which were actually positive and predicted positive
False Positive – the values which were actually negative but falsely predicted as positive, also known as a Type I Error
False Negative – the values which were actually positive but falsely predicted as negative, also known as a Type II Error
True Negative – the values which were actually negative and were predicted negative

So knowing that, the above confusion matrix illustrates the following:

True Positives = 14,156
False Positives = 2,086
False Negatives = 2,723
True Negatives = 19,882

The accuracy of the confusion matrix is approximately 87.6%.

Creating Models

Now it's time to create some models with our dataset! In total, we created four models: (1) a Logistic Regression model using random under sampling, (2) a Logistic Regression model using SMOTE, (3) a Random Forest model using random under sampling, and (4) a Random Forest model using SMOTE.

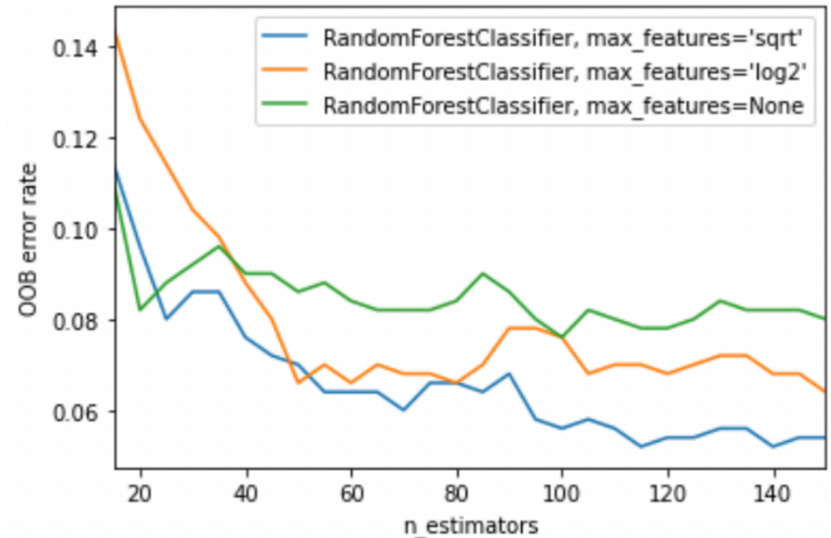
In order to help summarize the four models that were created, we created a table:

Model	Test Precision	Test Recall	Test F1-Score
Logistic Regression - RUS	0.68	0.58	0.63
Logistic Regression - SMOTE	0.68	0.59	0.63
Random Forest - RUS	0.97	0.96	0.97
Random Forest - SMOTE	0.96	0.98	0.97

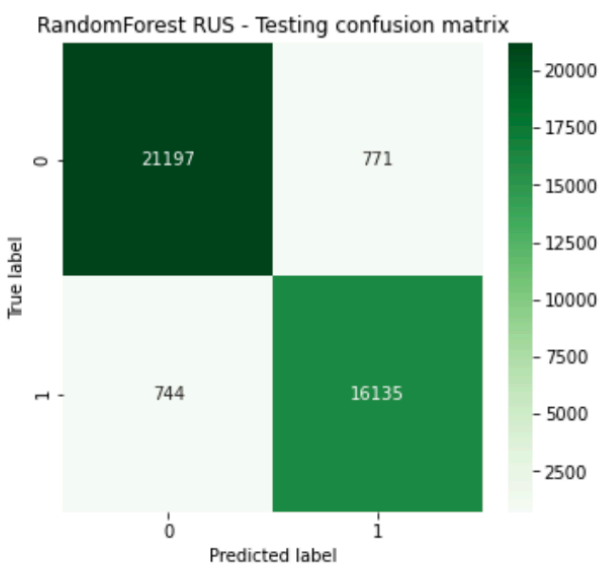
The above table clearly illustrates the four models that were created and their accompanying precision, recall, and f1-scores. For the needs of our client, we decided it would be ideal to optimize precision, as our client can best tolerate false negatives than false positives. Based on the above, we can see that the two random forest models performed the best with respect to precision. The top two models were Random Forest with random under sampling and Random Forest with SMOTE.

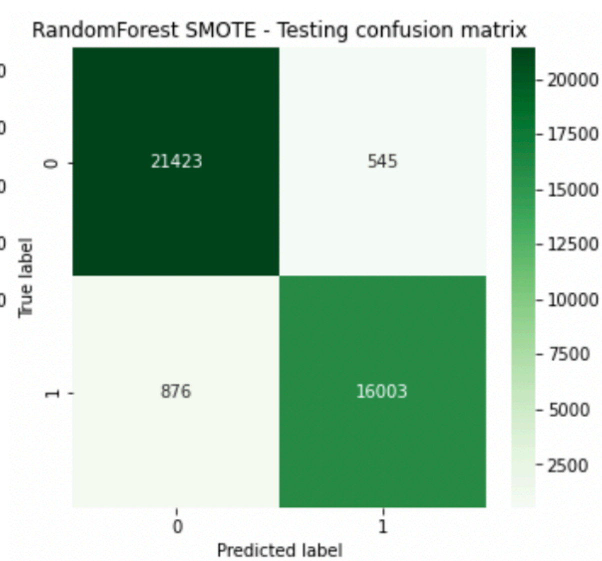
Hyperparameter Tuning

We now want to apply hyperparameter tuning to the top two models to see if we could decide on which model was best for our client's needs:



Per the above graph, it appeared that we should use n_estimators = 118 and max_features = sqrt. We did so for the two Random Forest models, and obtained the following confusion matrices:



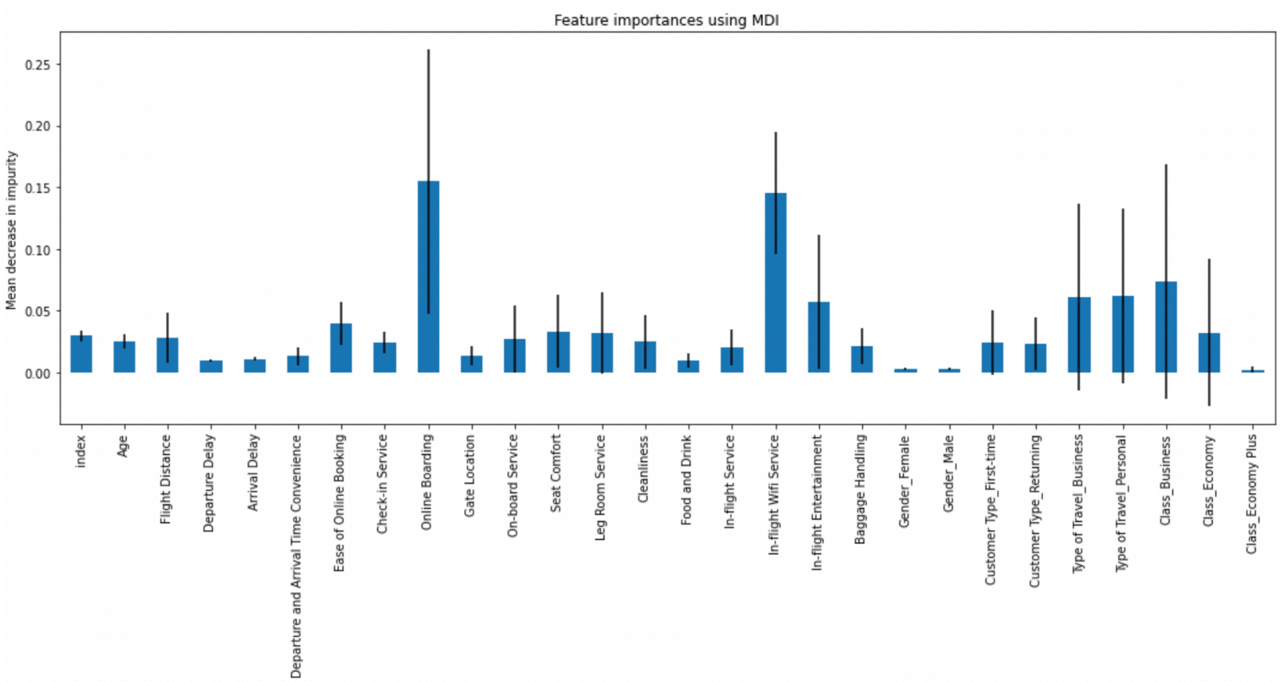


After hyperparameter tuning the two Random Forest models, we created another chart to show each model with their associated precision, recall, and f1-scores:

Model	Test Precision	Test Recall	Test F1-Score
Logistic Regression - RUS	0.68	0.58	0.63
Logistic Regression - SMOTE	0.68	0.59	0.63
Random Forest - RUS	0.97	0.96	0.97
Random Forest - SMOTE	0.96	0.98	0.97

Although we performed hyperparameter tuning on the two best models, we can see that this did not change the results. The Random Forest model using random under sampling still slightly outperformed the Random Forest model using SMOTE. As a result, we have determined that the best model for the needs of our client was the Random Forest model using random under sampling. This model had a test precision of 0.97, a test recall of 0.96, and a test f1-score of 0.97.

Since we identified the best model for our client, we were then able to use that model to determine which features were the most important.



From the above, we can see that the ten most important features (in descending order) for our chosen model were as follows:

- Online Boarding
- In-flight Wifi Service
- Class_Business
- Type of Travel_Personal
- Type of Travel_Business
- In-flight Entertainment
- Ease of Online Booking
- Seat Comfort
- Class_Economy
- Leg Room Service

Conclusions & Future Work

In conclusion, our client wanted to know what separated their "satisfied" passengers from their "neutral or dissatisfied" passengers. In order to make that determination, we cleaned the provided dataset, performed exploratory data analysis (which helped to showcase which features were rated highly amongst SIA passengers), and created four models.

During exploratory data analysis, we made several discoveries:

- Returning passengers reported higher satisfaction than first-time passengers.
- Similarly, business travelers reported higher satisfaction than personal travelers.
- Passengers who were seated in business class were overwhelmingly more satisfied than those in economy or economy plus classes.
- Online boarding rated very highly amongst SIA's satisfied passengers, as did SIA's on-board service, seat comfort, leg room service, in-flight service, in-flight entertainment, and baggage handling.
- Alternatively, with respect to SIA's neutral or dissatisfied passengers, SIA's on-board service, leg room service, cleanliness, food and drink, and in-flight WIFI service all rated poorly.

After creating models with the dataset, and performing hyperparameter tuning, we were able to discern the best model for our client's needs. Since SIA can better tolerate false negatives than false positives, we decided to focus on the precision of each model. As a result, we determined that the best model for our client was the Random Forest with random under sampling. The most important features for that model included online boarding, in-flight WIFI service, class_business, type of travel_personal, type of travel_business, in-flight entertainment, ease of online booking, seat comfort, class_economy, and leg room service.

Should SIA desire more information before they enact additional changes, we recommend the following for future work:

- Performing a similar analysis on competitor airlines to determine whether there are differences in passenger satisfaction.
- After enacting changes, performing analysis on a future passenger survey to determine whether those changes increased overall passenger satisfaction.

Recommendations

SIA wanted to know what separated their "satisfied" passengers from their "neutral or dissatisfied" passengers. As a result of this analysis, we recommend the following:

- SIA continue to do what they're doing with respect to the features that were shown to perform highly. As passengers were shown to already view those features favorably, SIA will not need to enact many changes.
- However, we also recommend that SIA enact changes to the features that performed poorly. The above analysis clearly illustrated which features were viewed unfavorably amongst passengers, and as a result, changes will need to be made if SIA wants to improve their overall reputation amongst current and future passengers.
- Finally, we recommend that SIA create another survey for their passengers at least one year after changes have been enacted, in order to determine whether those changes have increased satisfaction rates amongst their passengers.