Final Project

John Bolgert

DSC441

Problem Statement

For any successful service-based industry, the ability to understand the customer experience is critical in the survival of a firm. For this study, I will specifically be looking at customer satisfaction of those in the airline industry. Not only will I be attempting to predict if a customer was satisfied with their experience on a flight, but also will be looking at major driving factors that are critical in the customer experience in determining satisfaction of a customer. Having a satisfied customer is essential in retaining customers and to continue to expand the business.

Airline firms have an added pressure to better understand the customer journey as this industry was hit especially hard during the Covid-19 pandemic. As travel became taboo, airlines were hit hard as individuals were not booking flights. As of April of 2021, 44% of the US entire fleet of airlines remained idle, according to Sarah Hansen of *Forbes*. As more and more vaccines become readily available, travel is seeing an uptick but is still a way away from returning to the pre-pandemic level. Therefore, in order to regain as much market share, as quickly as possible, having high level of customer satisfaction is incredibly important. This data was collected prior to the pandemic, so it will not be speaking to possible changes in preference in customer satisfaction but can still be insightful in understanding the customer experience.

The airline industry arguably operates in an oligopoly market structure, which increases the importance of the strategic pricing. This creates an environment where market share is closely tied to the pricing strategy of the respective airline. Since this industry has such an aggressive competition in pricing, the benefits and services provided that drive customer satisfaction can be overlooked in importance. The data obtained for this study will be focusing on the customer experience rather than the pricing of the flights.

To better understand customer satisfaction, I will be analyzing a data set that consists of a Likert survey of customers asking for them to rate their services on a scale of 1 to 5, and eventually say if they were satisfied with their experience. More specifically I am looking to determine what major factors are driving customer satisfaction so that airlines can use this insight to better drive their customer experience, especially at a time where customer experience is very critical.

Data

For this project I will be utilizing an airline passenger satisfaction dataset to categorize the satisfaction of the airline's customer. More specifically I am looking to categorize a customer as either 'satisfied' or 'neutral or dissatisfied'. Of the dataset chosen 57% were 'neutral or dissatisfied' while 43% were 'satisfied' about their flight.

Kaggle dataset link:

https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction

The dataset chosen contains 129,880 rows, with 24 columns. Of the 24 columns, there are 14 columns that are from a likert survey of the passengers to rate their approval from zero to five. These 14 likert survey variables include the following topics.

- inflight_wifi_service
- departure_arrival_time_convient
- ease_of_online_booking
- gate_location
- food_and_drink
- online_boarding
- seat_comfort
- inflight_entertainment

- onboard_service
- leg_room_service
- baggage_handling
- checkin_service
- inflight_service
- cleanliness

Another 4 columns of the 24 columns are dummy binary variables, describing certain attributes of each passenger. These variables are looking to describe the type of customer and the reason they are traveling, along with their age. These include the following variables.

- gender
- customer_type
- type_of_travel
- customer_class

Another 4 of the 24 columns are numeric variables. These variables are looking to measure performance of flight timeliness as well as the age and duration of the flight taken. These variables contain the following.

- Age
- flight_distance
- departure_delay_in_minutes
- arrival_delay_in_minutes.

The variable we are looking to predict is satisfaction, which has the response of either ‘neutral or dissatisfied’ or ‘satisified’. Of the 129,880 observations there are zero entries of missing data. This isn’t too surprising since this is fake Kaggle data. The only prepressing that needs to be done to this data is to assign values of zeros and ones to the binary variables.

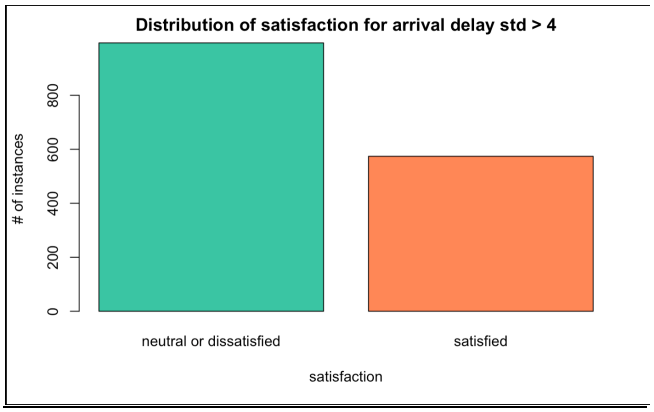
Data Preparation

In order to ensure the data was properly maintained, null values were searched for. Of all the rows within the dataset, only 393 null values were discovered for the variable *arrival delay in minutes.* Since the variable *arrival delay in minutes* has outliers, the null value was replaced with the median for that field.

For the Likert survey questions, the value of zero was used to represent null values. Since the value of zero was used for null values, they were originally not flagged. The null values for these fields were replaced with the median value. The fields this pertained to included, *inflight_wifi_service, departure_arrival_time_convenient, ease_of_online_booking, gate_location, food_and_drink, online_boarding, seat_comfort, inflight_entertainment, onboard_service, leg_room_service, baggage_handling, checkin_service, inflight_service.*

For the outliers that were discovered, they were kept within the model. Flight delays are a common occurrence, the outliers are kept in the model so it can handle future data that contains these instances. When looking at the label of interest of that of satisfaction, the results for instances of a standard deviation greater than four for the variable *arrival delay*, there is still about 36% of individuals who are still satisfied with their flight experience. This can be shown in figure 1 below.

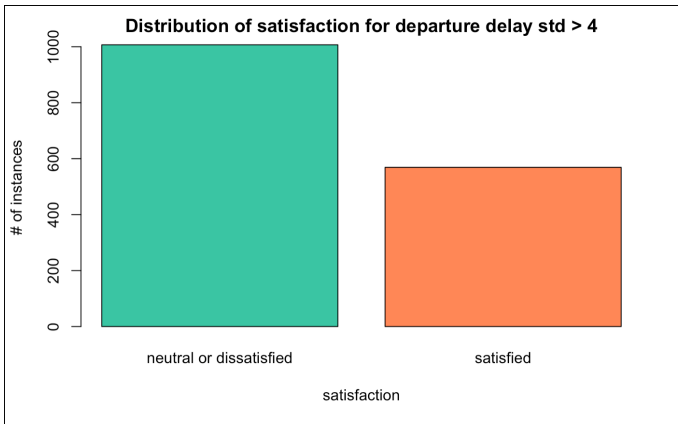
Figure 1



Since delays can be quite a pain point for consumers, I thought the imbalance of the label of interest would be greater, yet since it is not this adds more merit to keeping these outlier instances within our training and testing dataset, since there appears to be another driving force in customer satisfaction even at times of a long arrival delay of a flight.

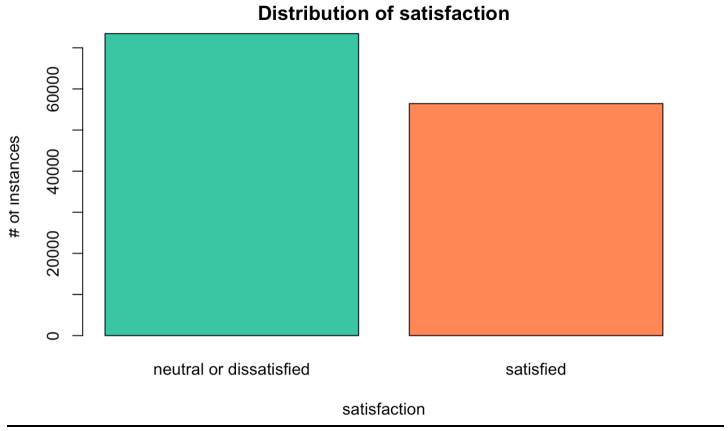
Another variable that had outliers was that of departure delay in minutes. Again, looking at the imbalance for the label satisfaction for instances where the departure delay std was greater than 4, the percentage of those who were satisfied was around 36%. This can be shown in figure 2 below. The outliers for this field were kept as well since departure delays can be a common occurrence and so our model can still predict future data with these occurrences.

Figure 2



Overall looking at the distribution of the label satisfaction, there appears to be a somewhat balanced distribution. There is a slight skew towards 'neutral or dissatisfied' customers but, this skew is not dramatic enough to cause concern. This can be shown in figure 3 below.

Figure 3



Other data transformations that occurred in the preprocessing step was the removal of the column of row number. This was removed since it provided no beneficial information that could be utilized within our model.

The variables *satisfaction, Gender, customer_type*, and *type_of_travel* were of type char, so they were transformed to type factor since they are categorical variables.

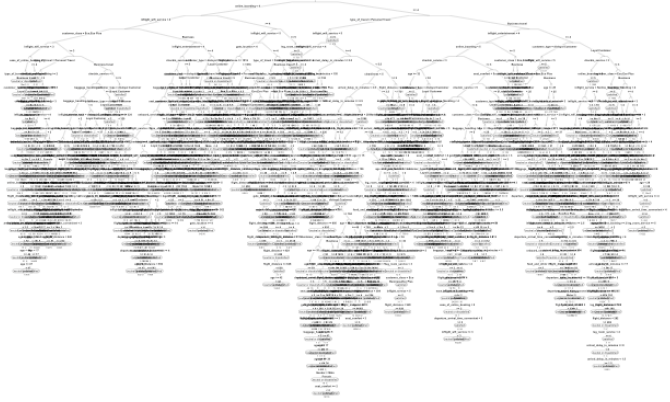
After these steps were completed, the data was split into a 70, 30 ratio of training, testing data subset respectively, to help avoid overfitting of the model.

Modeling

The modeling technique that was chosen to perform the classification was a decision tree algorithm. This technique was chosen not only because it is a robust but also because it provides practical significance in the range of values in the independent variables for interpretation. For example, say if delay in minutes have a great impact, what percentage above a certain threshold in x are where the split in satisfied vs non satisfied are found.

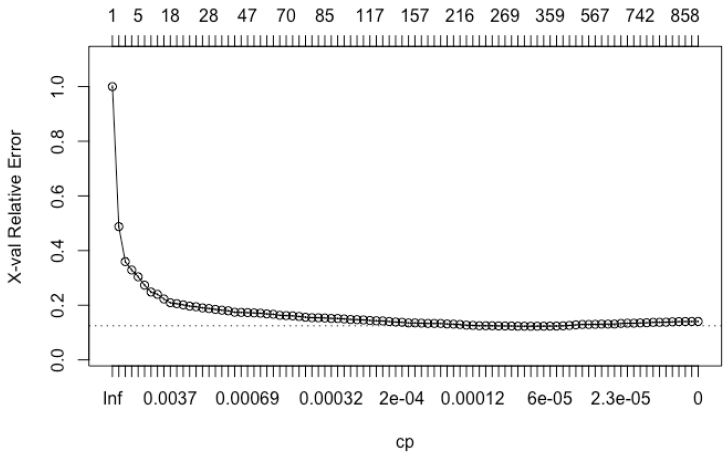
For the first decision tree that was built, the computation parameter was set to the value of zero with no other hyper parameters chosen. The resulting decision tree, which can be found in figure 4 below, produced an overly complex model.

Figure 4



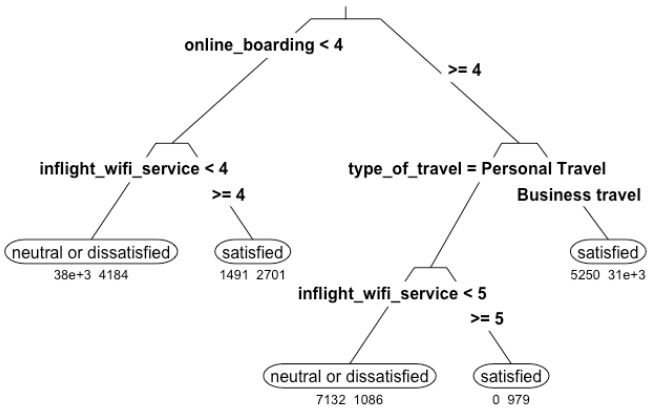
Since this model was found to be overly complex, pruning was then attempted in order to decrease the complexity of the model. Although when plotting the relative error against possible CP values, it was found that a clear cp value was hard to identify. This can be shown in figure 5 below.

Figure 5



Pruning was not delivering the expected results; therefore, another decision tree was built that contained no hyperparameters and no CP value. Since it contained no hyperparameters, R would be utilizing the already available optimization that the rpart package contains. Figure 6 below, produced the tree that was built.

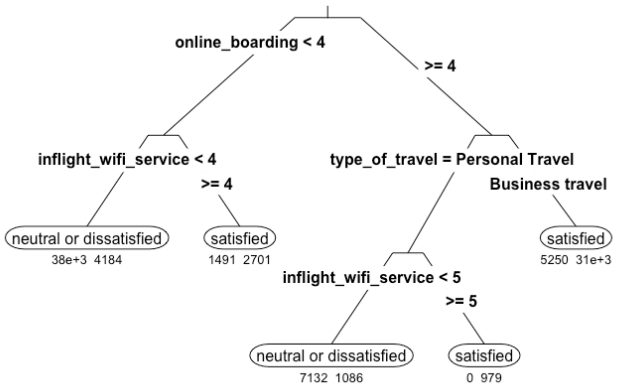
Figure 6



With the hyperparameters removed the decision tree that was built was very much reduced in the complexity. In order to attempt to find a better decision tree model, grid search was utilized to test accuracy of decision tree models that had different input values for some of the hyperparameters. The hyperparameters that were chosen to test on, were that of minsplit and maxdepth. The minsplit hyperparameter determines the required number of observations within a node in order for the split to occur. While the maxdepth determines the maximum number depth of nodes that can occur on the tree. The value for the hyperparameters being tested for minsplit ranged from 1-10 while the values for the hyperparameters for maxdepth being tested ranged from 1 to 8.

These combinations of hyperparameters were tested based on accuracy to deliver the best model. From this technique the resulting model can be shown in figure 7 below.

Figure 7

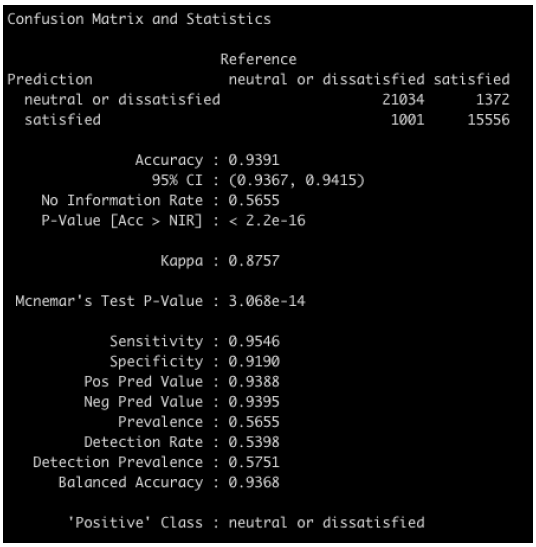


What is rather interesting about this final model, is that it produced the exact same decision tree that was produced when we had no hyperparameters inputted.

Evaluation

Since our data contained a somewhat balanced range for the label of interest, the measure of performance that was predominately used was that of accuracy. Of our three models that were produced, the one that had the highest accuracy score on the testing data set was the first model that was produced. This was rather shocking, since it was such an overly complex model that one would believe overfitting would occur. Overfitting would produce a high accuracy score on the training data, but if present should produce a lower accuracy score on the testing data. The fact that the model did extremely well on the testing data is cause for concern but since this is fabricated data, there could be something occurring in the creation process possibly causing the data to be relatively similar across the testing and training subsets. The overall accuracy for this model produced was about 93%. Although accuracy was our chosen level of performance it is interesting to note that this model performed better when using the measure sensitivity to evaluate, the model performed a sensitivity score of around 95%. A complete list of performance scores for this model can be shown in figure 8 below.

Figure 8



Although this model produced quite a high accuracy score, since it was overly complex it lacks interpretability for practical significance.

Looking at our models produced using no hyperparameters, and our model produces using grid search, they produced the same tree so therefore they have the same performance scores. The final model had an accuracy score of around 86%. Again, this is quite shocking that this model produced a lower accuracy score on the testing data, since this model produced a simpler decision tree that would most likely be more robust and less susceptible to overfitting. Although this final tree produced a lower accuracy score, it still produced a rather high accuracy score. Below in figure 9, a full list of the performance score indicators of this model can be show.

Figure 9

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   19027      2224
  satisfied                                  3008     14704

               Accuracy : 0.8657
                 95% CI : (0.8623, 0.8691)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7282

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8635
            Specificity : 0.8686
         Pos Pred Value : 0.8953
         Neg Pred Value : 0.8302
             Prevalence : 0.5655
         Detection Rate : 0.4883
   Detection Prevalence : 0.5454
      Balanced Accuracy : 0.8661

       'Positive' Class : neutral or dissatisfied
```

Discussion and Conclusion

For practical significance and conclusions of our models, only the second and third model produced insights. Since the first model was incredibly overcomplex, everything is important which leads to a lack of beneficial insight. From our final model produced, it was shown that some of the major variables of interest were, online boarding, inflight WIFI service, seat comfort. In figure 10 below, the full rankings of variables of interest can be shown.

Figure 10

```
 online_boarding  inflight_wifi_service          seat_comfort ease_of_online_booking
      15022.4929             9880.9940             6826.3420              5843.4752
  type_of_travel         customer_class inflight_entertainment                    age
       5786.7058             5772.8298             4657.2833                524.7486
```

Intuitively, it comes to no surprise that the online boarding process has such a high practical significance. This part of the user journey is where the customer has most interaction with the staff of the plane. After going through security, which can at time be frustrating, the last thing a customer wants to

experience is another confusing, inconvenient process to get on the plane. It is especially important that the boarding process is ideal for those who are traveling in business class. If someone is traveling to pitch new business to a potential client, the less stress they have in traveling appears to be a huge factor in determining their satisfaction. Of all the variables included in this data set, online boarding and delays would be the ones that could possibly induce the most stress on a customer. We however did not ask the customers about their stress level or what influenced it, so we cannot speak on that, but we could hypothesis this and develop further study to determine if that is indeed the case here. So therefore, we can only conclude that the online boarding process has a high practical significance especially for those traveling in business class. Since business class tickets tend to be one of the more expensive tickets offered besides first class, this could be a potential area for higher gains in profit. Therefore, airline should perform more studies focused on the online boarding process for customers, as this could be a potential opportunity for improvement of customer satisfaction.

Another variable that has high practical significance is that of inflight WIFI service. This variable having high practical significance also intuitively comes to no surprise. As more and more individuals become more connected and dependent on their digital devices, the ability to remain connected while on their flight would intuitively make sense to become more important to the customer. Since phones have been increasing in their digital capabilities it becomes not surprising that these devices' ability to connect rank higher for customers in importance than having great in-flight customer entertainment. This can be very insightful for airlines, as this could influence them to divert funds from adding more screens to the back of seats for flights to instead bolstering and increasing the WIFI capabilities of a plane. Currently of all the airline providers in the world, only 15 of them provide some form of free inflight WIFI service according to Dean at *point me to the plane*. As free drinks and snacks at one point were a hot commodity for travelers, free WIFI could possibly become that new essential service provided. Although the variable here is measuring their satisfaction of the WIFI provided, so we cannot determine if having free vs paid WIFI has an effect on customer satisfaction. But this could be a good potential follow up study to better understand this feature on customer satisfaction for them to optimize.

Overall, our final model produced areas that airlines could potentially focus more on in future studies to better understand how it affects the customer experience and satisfaction. As airlines continue to add more flights, the ability to better understand their customers could lead to a potential higher brand loyalty and therefore the possibility of higher market share which could allow them to generate better profits.

Citations

Dean, et al. “What Are The Airlines That Give You Free Inflight Wi-Fi?” Point Me to the Plane, 20 Apr. 2021, pointmetotheplane.boardingarea.com/airlines-free-wifi/.

Hansen, Sarah. “Six Numbers That Show How Hard The Travel Industry Is Being Hit By The Coronavirus Shutdown.” Forbes, Forbes Magazine, 17 Apr. 2020, www.forbes.com/sites/sarahhansen/2020/04/17/six-numbers-that-show-how-hard-the-travel-industry-is-being-hit-by-the-coronavirus-shutdown/?sh=19bf66ce417f.