Team 9 Project 2 Report

Airline Satisfaction Dataset

For

OPIM 5604

Predictive Modeling

Team Members:

Abhinav Dubey

Emily Maher

Jiaxin Wu

Ping-Hsuan Chiang

Rumeysa Maras

**Table of Contents**

[**Executive Summary**](#_heading=h.gjdgxs) **[2](#_heading=h.gjdgxs)**

[**Problem Statement**](#_heading=h.30j0zll) **[3](#_heading=h.30j0zll)**

[**Methodology**](#_heading=h.1fob9te) **[3](#_heading=h.1fob9te)**

[Sample](#_heading=h.3znysh7) [3](#_heading=h.3znysh7)

[Explore](#_heading=h.2et92p0) [4](#_heading=h.2et92p0)

[Modify](#_heading=h.tyjcwt) [4](#_heading=h.tyjcwt)

[Model](#_heading=h.tyjcwt) 5

[Assess](#_heading=h.tyjcwt) 8

[**Results**](#_heading=h.3dy6vkm) **[9](#_heading=h.3dy6vkm)**

[**Conclusions and Recommendations**](#_heading=h.2s8eyo1) **[10](#_heading=h.2s8eyo1)**

[**References**](#_heading=h.17dp8vu) **11**

# **Executive Summary**

Measuring customer satisfaction is a crucial aspect for modern businesses as it can significantly contribute to a continuing effort of service quality improvement. In this paper, we show how data from an airline can be used to measure passengers’ satisfaction from a large set of dimensions and indicate those dimensions that need to be improved.

The airline dataset contains about 130,000 survey entries with details about passenger, flight, and satisfaction ratings for different flight elements. In total, there are 24 columns out of which 5 columns are about the passenger, 3 columns have information on the flight. Around 14 columns are survey entries where passengers have rated the flight experience elements on a scale of 1 to 5. Our target variable has binary values that state if the customer is satisfied or not and it is a categorical type. After applying the data pre-processing concepts, we have excluded some rows with NaN values. Also, in some of the columns, indicator columns were created so that the complexity of the model can be reduced.

Before starting with the models, we split the data into a ratio of 70:15:15. We have used the training set to train the model, the validation set to tune the model and test set parameters used for selecting the final model. The models we created and tuned on the dataset are Logistic regression, Neural Net, Naive Bayes, KNN, Ensemble Trees, and Decision Trees. While evaluating all the models we found that the highest performing model is Decision Tree. It has the highest total accuracy of 95.9% and the lowest RMSE of 17.3%. The most highly correlated predictors are type of travel, inflight Wi-Fi service and online boarding. And the lowest correlated predictors are Seats, Delays in arrival and departure from our finding through models. Gender variable was found to have little or no impact on the target variable. The actionable recommendations are to offer complimentary Wi-Fi to all the passengers and get a more user-friendly online boarding process.

To conclude the results from the analysis, reveal useful findings regarding satisfaction criteria that value the most and in which dimensions the airline can improve their service. Overall, this analysis will help the airline to excel in the competitive market and be the best.

# **Problem Statement**

Today we are living in a competitive market and the airline industry is characterized by an oligopoly market structure. The oligopoly market structure has imperfect competition in which a limited number of firms dominate the industry. To be the best in the market, it is necessary to know about the customer and the customer's level of satisfaction from the service. We have entries from different customers in the dataset and their satisfaction rating response on various flight features. We will use this data to create a model that predicts which all flight elements are in high correlation with the satisfaction level and if a customer is satisfied or not. With this knowledge, the airline can improve the areas that have the lowest correlation with the target variable. The modeling results will help improve the airline's business as the airline will better understand the customer requirements and areas of improvement in their service.

# **Methodology**

## Sample

We found this dataset which is large enough to contain the significant information from many passengers and small enough to have the essential predictors to build our models. There are 129,880 rows and 25 columns in the dataset. All columns are listed below:

* **#**: counting number of rows.
* **Id**: id number of each row of record
* **Gender**: Gender of the passengers (Female/ Male)
* **Customer Type**: The customer type (Loyal customer/ disloyal customer)
* **Age**: The actual age of the passengers
* **Type of Travel**: Travel purposes of the passengers (Personal/ Business Travel)
* **Class**: Travel class in the plane of the passengers (Business/ Eco/ Eco Plus)
* **Flight Distance**: The flight distance of this journey
* **Inflight Wi-Fi Service**: Satisfaction level of the service (0: Not Applicable; 1-5)
* **Departure/Arrival Time Convenient**: Satisfaction level of the time convenient
* **Ease of Online Booking**: Satisfaction level of online booking
* **Gate Location**: Satisfaction level of Gate location
* **Food and Drink**: Satisfaction level of Food and drink
* **Online Boarding**: Satisfaction level of online boarding
* **Seat Comfort**: Satisfaction level of Seat comfort
* **Inflight Entertainment**: Satisfaction level of inflight entertainment
* **On-board Service**: Satisfaction level of On-board service
* **Leg room Service**: Satisfaction level of Leg room service
* **Baggage Handling**: Satisfaction level of baggage handling
* **Check-in Service**: Satisfaction level of Check-in service
* **Inflight Service**: Satisfaction level of inflight service
* **Cleanliness**: Satisfaction level of Cleanliness
* **Departure Delay in Minutes**: Minutes delayed when departure
* **Arrival Delay in Minutes**: Minutes delayed when Arrival
* **Satisfaction**: Airline satisfaction level (Satisfaction/ neutral or dissatisfaction)

## Explore

We understand the importance of customer satisfaction because more satisfied customers mean getting better chances to become loyal customers. The high level of customer satisfaction helps boost revenue and make an airline company stand out in the competition. While exploring the data, we found that Type of Customer and the variables associated with rating appear important. We also found that the higher scores a passenger rated, the greater probability he is satisfied with an airline company.

## Modify

We changed the types of several variables from numerical to classificational variables. For Inflight Wi-Fi Service, Departure/Arrival Time Convenient, Ease of Online Booking, Gate Location, Food and Drink, Online Boarding, Seat Comfort, Inflight Entertainment, On-board Service, Leg room Service, Baggage Handling, Check-in Service, Inflight Service and Cleanliness, we changed the type from the numerical to ordinal variables. For Id, we changed from the numerical to nominal variable.

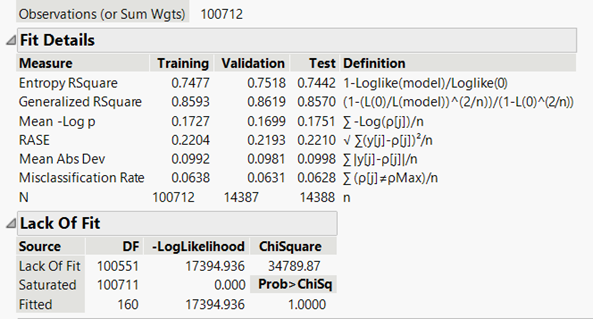
There are no outliers and only 393 missing values for Arrival Delay in Minutes excluded. We binned some categorical columns into dummy variables and deleted the useless columns. The binned columns are Satisfaction (Satisfied = 1, Unsatisfied = 0), Gender (Female =1, Male=0), Customer Type (Loyal customer =1, Disloyal=0), Type of Travel (Business = 1, Personal = 0), and Class (kept Business and Eco). Finally, we transformed the data of Flight Distance to SHASH normal distribution.

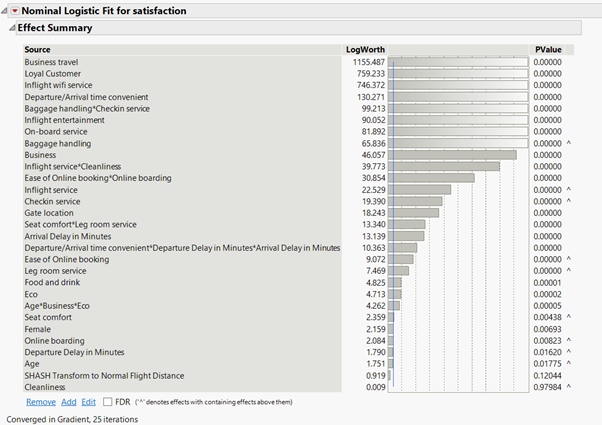
## Model

All models we built are listed below:

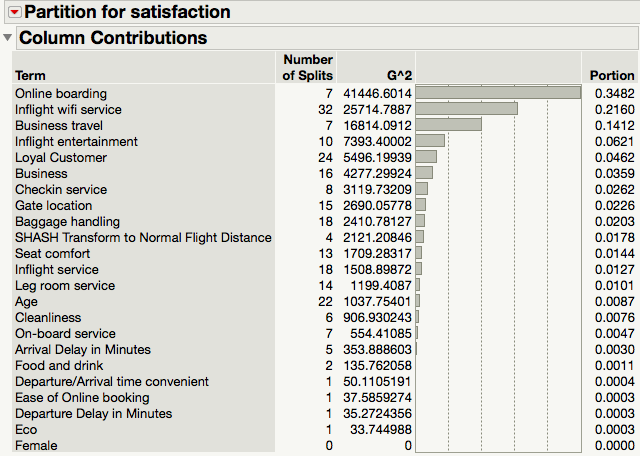
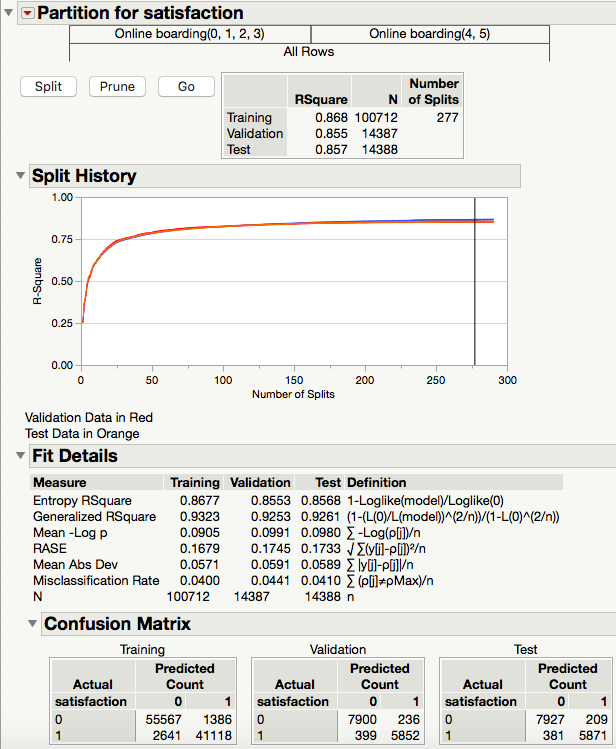
* **Logistic Regression**: The accuracy is 93.7% and RASE is 22.1%. In the effect summary, the three most important variables are Business Travel, Loyal Customer, and Inflight Wi-Fi Service.

Before we get the optimal results, we pruned it with several processes. Firstly, we added “Crosses” to the original variables to reduce the error that variables with similar information bring into. Cross: 1. Baggage handling \* check in service; 2. Inflight services \* cleanliness; 3. Seat comfort \* leg room service. Then, we deleted the variables with high P Values.

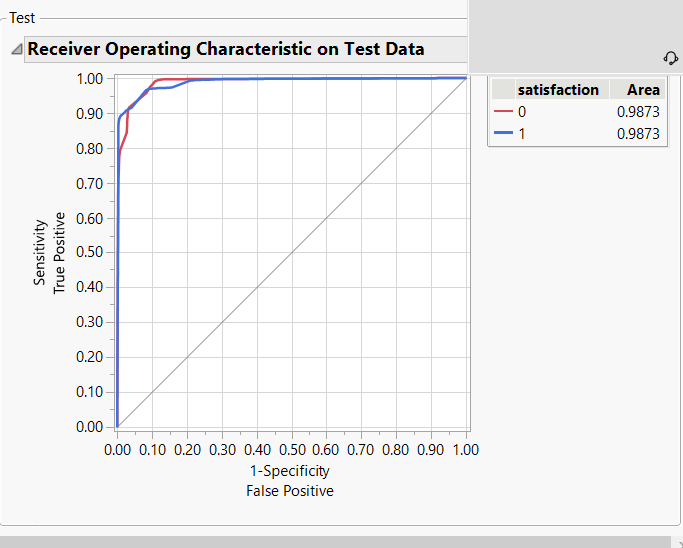
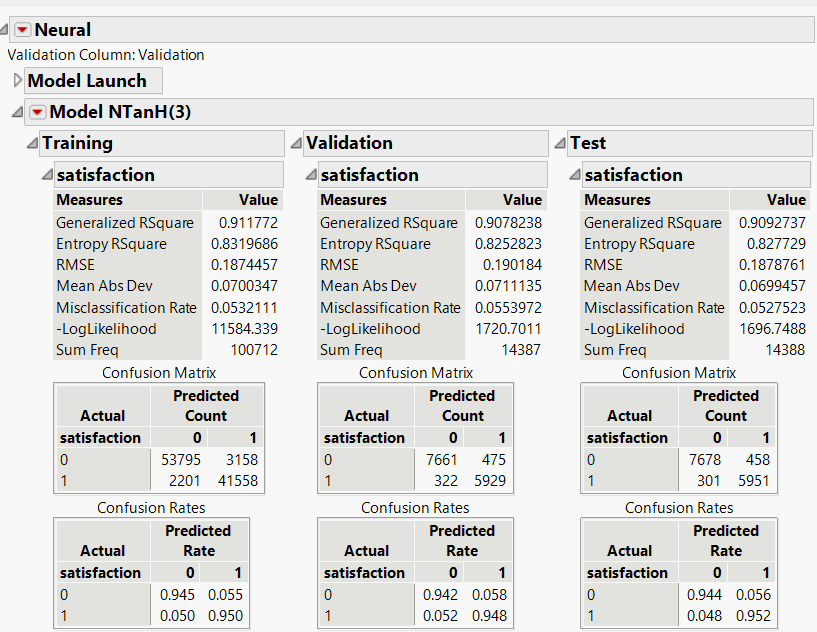


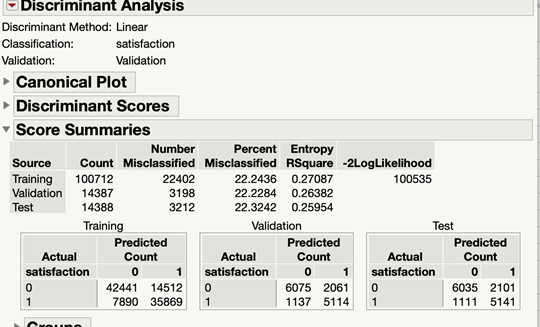


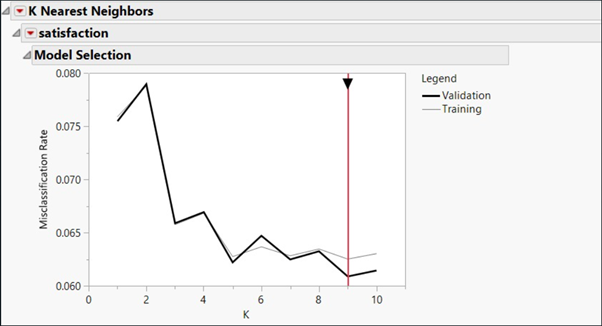
* **Decision Trees**: We found that using all variables for predictors except for Id gets the best performance. The accuracy is 95.9% and RASE is 17.3% while trees are pruned from 280 to 277 splits. The table of the column contributions shows that Online Boarding impacts customer satisfaction the most. The AUC is 99.3%. The performance is relatively strong.

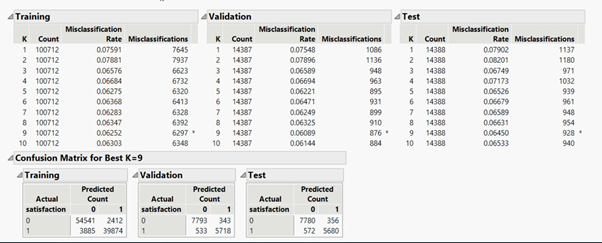


* **Bootstrap Forests**: The accuracy is 91.9% and RASE is 26.2%. The AUC is 96.5%.
* **Boosted Trees**: The accuracy is 93.8% and RASE is 21.6% while the model was using 120 layers and setting 10% of the learning rate. Its AUC is 98.4%.
* **Neural Nets**: We set the random seed reset to 123 and get a total accuracy of 94.7% and RMSE of 18.9%. The AUC is 98.7%.

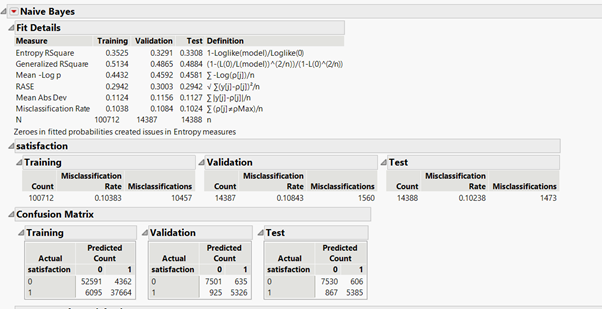


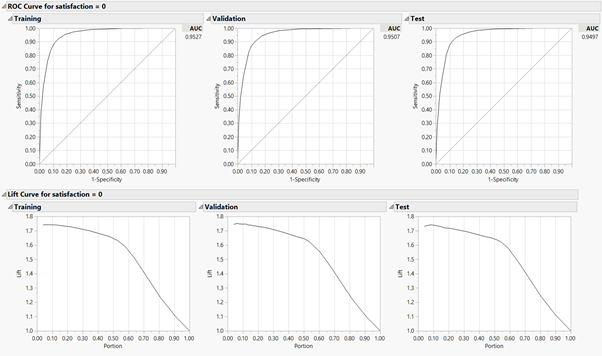
* **Discriminant Analysis**: This model does not work well with our data because most of the data is ordinal. The percent misclassified of 22.3% is too high.
* **K-Nearest Neighbor (KNN)**: The optimal k-value we got is 9. The accuracy is 93.55%

K = 10

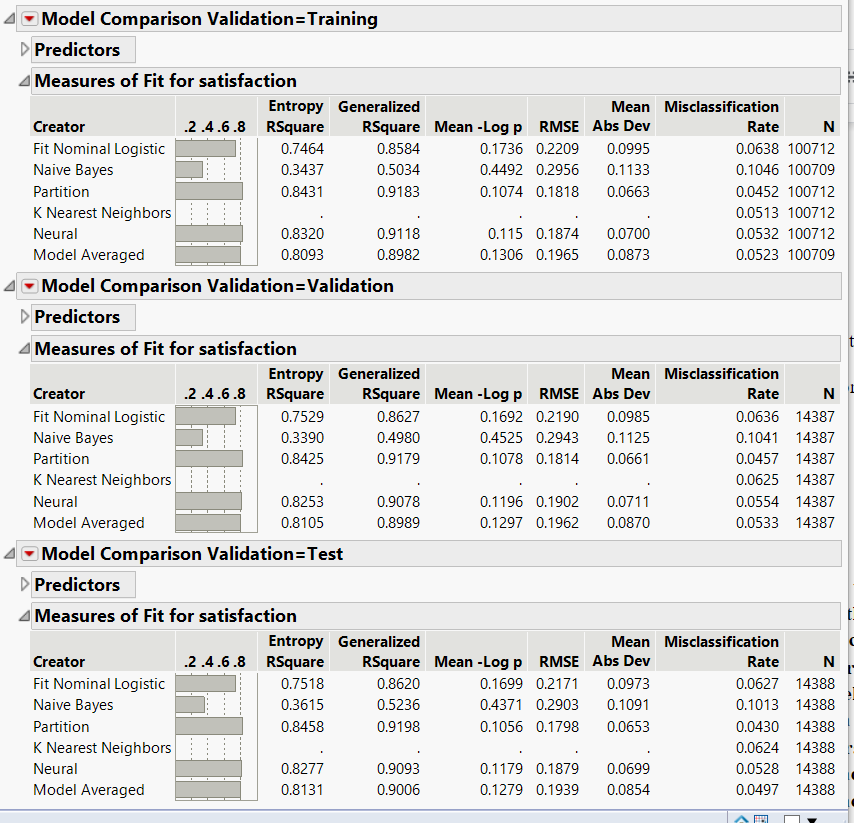


* **Naive Bayes**: The accuracy is 89.77%, and RASE is 29.42% (AUC = 94.97%).





* **Ensemble Models**: We used Simple Model Averaging (Model Comparison) to get the result of Model Averaging (Weighted). The performance of Model Averaging is not better than the performance of Decision Trees. The accuracy of Model Averaging is 95.1%, and RASE is 19.4%.



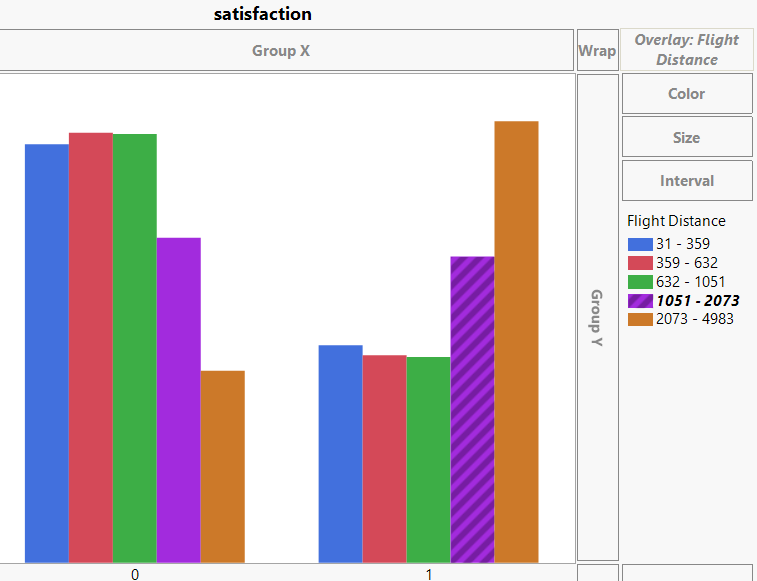
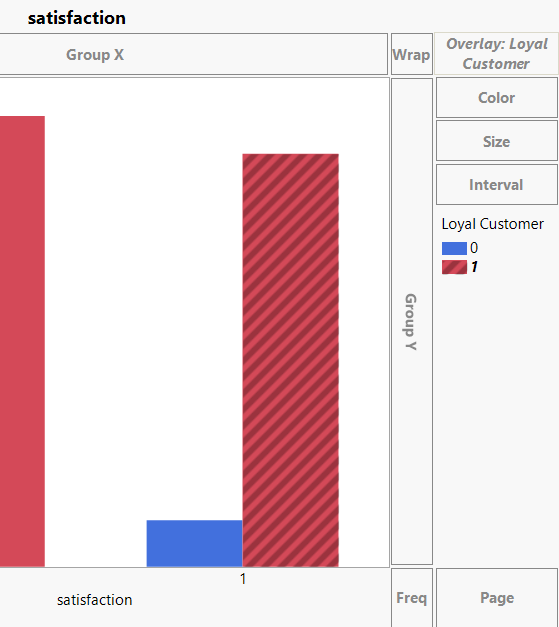
## Assess

After comparing the results of the models, we figured out that Decision Trees performed the best. Through Simple Model Averaging, the result presents that the model of Decision Trees has the highest accuracy and the lowest RMSE.

# **Results**

We assessed multiple models. There were a few that we could not utilize for this project. First, the discriminant analysis model was not compatible with our dataset because most of our variables were ordinal, and the discriminant prefers continuous predictor variables. The following models were developed: Decision Tree, K Nearest Neighbors, Neural Net, Linear Regression and Naive Bayesian. Several of the models had a relatively high accuracy rate of 95% or higher. The least accurate model was the Naive Bayes, with a 10% misclassification rate. Our two highest performing models were the Decision Tree and Neural Network. The decision tree model had a total accuracy of 95.9%, while the neural network had a total accuracy of 94.7%. We then combined the models to create an ensemble model. The accuracy of the ensemble model was lower than the accuracy of the Decision tree. The Decision Tree was chosen as the highest performing and final model.

The factors most highly correlated with passenger satisfaction were found to be online boarding, inflight Wi-Fi service, type of travel, inflight entertainment and loyal customers. These results are consistent with some of our other models, including the logistical regression. Customers on business travel were thirteens times more likely to be satisfied with their flight than those on personal travel. An interesting note to add to this was that reason for travel was higher correlated to satisfaction than travel class a passenger was flying in. This is an interesting differentiation. Loyal customers were ten times more likely to be satisfied than non-loyal customers. (Figure 1) Cleanliness and leg room were discovered to have a low correlation value. This was surprising because one might think that cleanliness or physical comfort would be important to customer satisfaction, particularly if on a long flight. Some other interesting patterns to note were that younger passengers, ages 7-26, were the most likely to be unsatisfied. Another interesting distinction was that flight distance had little effect on satisfaction. The highest number of unsatisfied customers were on short flights (1051 miles or less), while the highest number of satisfied customers was on a long flight (>2073.) (Figure 2)





# Conclusions and Recommendations

Based on the modeling results, there are several recommendations to be made to increase passenger satisfaction. These recommendations are based on our highest-performing model. The online boarding process was the most significant element to passenger satisfaction. Passengers want to be able to check in online or use their phone to check their seat and boarding status. Whether this be through a website or an app on their phone, a passenger’s experience with online boarding needs to be intuitive and user-friendly. Your airline should focus on web and app development to ensure customers have a highly rated experience. Another recommendation is to ensure you have strong Wi-Fi service available for passengers. Their satisfaction with the Wi-Fi service was a significant input into their overall experience. Ensure that you have free, accessible and reliable Wi-Fi service available on flights. The type of travel was highly correlated with passenger satisfaction. Type of travel depicts business or personal and should not be confused with travel class. Analyze what services you provide to business travelers and what it is that may make their experience different. Are there specific characteristics of staff interaction, services offered or financial incentives with business travelers? Something about their experience is leading to high satisfaction.

While slightly less significant, there are several other factors that can boost a passenger’s experience. Inflight entertainment and check in service are important in satisfying travelers. Just as with Wi-Fi, ensure you are providing a reliable service. Services such as free headphones or a variety of movie genres are a few ways of creating favorable inflight entertainment. Check in service is another element that should not be overlooked. Ensure an online check-in is available, as the data has proven significance in online options. Any staff that meet customers at check in should be providing optimum customer service. Our last recommendation is to not forget the value in loyal customers. Any incentives or benefits being offered to loyal customers are clearly effective. To summarize the recommendations based on the findings, customers want a user-friendly online boarding experience, high quality inflight Wi-Fi and entertainment and a positive check in experience. Loyal customers and customers traveling on business have the highest satisfaction rates, so attempt to reproduce any aspects of those flights that you can.

# **References**

Benchmarking was completed to assess our results against other industry studies. It is significant to note that this dataset was developed before COVID-19. The reason this is important is because if you look at recent research, cleaning protocols is the number one concern today that airline passengers have. With the recent pandemic, passengers are more concerned with the cleaning protocols airlines have in place to keep them safe. In our analysis, cleanliness was a less significant variable. Type of travel and whether a customer was a returning customer (loyal customer in our dataset) were both contributing factors in industry studies we reviewed and in line with our results.

**Source of Dataset chosen for project:**

[**https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction**](https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction)

**Textbook:**

Bruce, Peter, et al. *Data Mining for Business Analytics: Concepts, Techniques and Applications with JMP PRO.* John Wiley and Sons, Inc., 2017.

**Sources for Benchmarking Analysis:**

[**http://www.iosrjournals.org/iosr-jbm/papers/Vol20-issue8/Version-5/A2008050106.pdf**](http://www.iosrjournals.org/iosr-jbm/papers/Vol20-issue8/Version-5/A2008050106.pdf)

[**https://www.theacsi.org/news-and-resources/customer-satisfaction-reports/reports-2019/acsi-travel-report-2018-2019/acsi-travel-report-2018-2019-download**](https://www.theacsi.org/news-and-resources/customer-satisfaction-reports/reports-2019/acsi-travel-report-2018-2019/acsi-travel-report-2018-2019-download)

[**https://www.jdpower.com/business/press-releases/2020-north-america-airline-satisfaction-study4**](https://www.jdpower.com/business/press-releases/2020-north-america-airline-satisfaction-study4)