Names: Naina Grover, Ronald Morris, Sarah Saltzman, Scott Slutsky, Mischa Subhan
Deliverable: Final Report
Deliverable Date: May 21, 2021
Project Title: Airline Passenger Satisfaction

Business Understanding

Business Problem

Airlines must analyze customer satisfaction to improve service quality. In a customer satisfaction survey for long-haul flights, Delta Air Lines scored 810 on a 1,000-point scale for customer satisfaction. Southwest Airlines scored 826 points, and JetBlue Airways scored 823 points. Improvements in customer satisfaction scores can increase consumer confidence and airline reputation. As a result, travelers will fly more on the airlines with the highest reputations.[1] Thus, our business problem is to analyze what factors contribute to customer satisfaction for people traveling through airlines by using data science techniques and best practices.

Moreover, customer satisfaction is always top of mind for airlines and is directly related to number of passengers, revenue, and service quality provided. As frequent airline travelers ourselves, we were interested in exploring how airlines contribute to customer satisfaction and which airlines are performing the best based on various factors. This project, in turn, is important for an airline, and any company in general, to understand which factors contribute the most to customer satisfaction. Once a company knows that, they can choose which points of the customer experience to focus more or less on to ensure satisfied customers. In the airline industry, satisfied customers could lead to repeat customers, referrals, and overall goodwill. These satisfied customers could also be targets for airline rewards credit cards, for example.

---

[1] David Kaplan. "What are the keys to airline passenger satisfaction now?" Kambr Media, 2020.
https://www.kambr.com/articles/the-keys-to-airline-passenger-satisfaction

Data Science Solution to Address Business Problem

With the help of data science, the airline industry keeps its customers up-to-date in real-time. By implementing predictive analysis, sentiment analysis, and travel journey analysis, airlines can promote special offers based on the needs, habits, and unique experiences of their passengers. Collecting and crunching customer data enables airlines to understand passengers' preferences and behaviour well enough to offer them transportation options they prefer and, more importantly, are ready to spend money on. Travel experience is getting extremely customized and customer-oriented using data science.[2]

To understand what variables contribute most heavily to passenger satisfaction, we implemented several models and evaluation techniques including but not limited to holdout validation, confusion matrices, ROC, and AUC. The high accuracies of our models portray how precisely a data science solution would address the business problem. Specifically, when utilizing an ROC curve to compare the accuracy of the classification tree model to the accuracy of the logistic regression model, the AUC score ranked the classification model slightly higher than the logistic regression model with an accuracy score of 92.5 percent and 96.2 percent, respectively.

We implemented logistic regression and classification trees to identify the features that most significantly influence airline satisfaction. To effectively perform this regression, we converted our response variable, airline satisfaction level, from text to binary form. In other words, we used 1 to represent a satisfied customer and 0 to represent an unsatisfied or neutral customer. Regression results will inform analysts of service areas the airline can modify to raise customer satisfaction levels and outperform competitors.

[2] Anastasiia Zamiatina. Datumize. 9 incredible ways data analytics is transforming airlines. May 15, 2019. Retrieved from https://blog.datumize.com/9-incredible-ways-data-analytics-is-transforming-airlines

The dataset that we downloaded from Kaggle was already split into a training (80%) and testing (20%) set.[3] The training set has 95,415 observations, and the test set has 23,789. One of the first things we did was drop the 300 rows with missing data. Then, we converted some categorical variables, such as gender, customer type, travel type, and satisfaction, to binary variables. The dataset contained three categories that pertained to travel class, so we constructed two dummy variables for that. After cleaning the data, we conducted descriptive analysis in Tableau and R. Our exploratory results will be found below. Our next step in the process was to build logistic regression and tree based classification models. Our objective was to predict 1 (satisfied) or 0 (neutral/dissatisfied).

Some variables included in this study are gender, customer type, age, type of travel, class, flight distance, inflight wifi service, and convenience of departure and arrival times. The response variable is customer satisfaction, which, as previously mentioned, was converted from character to binary format. Each instance represents a customer.

Data Understanding & Visualization

As previously introduced, data was sourced from Kaggle and was presented in the form of training (n = 103,904) and testing sets (n = 25,976). Data cleaning procedures reduced our training and testing sets to 95,415 and 23,789 observations, respectively. Preliminary descriptive analysis revealed that the vast majority of variables contained discrete values. Those that did not contain discrete values contained continuous values or characters, the latter of which were converted to binary form. Because our data was published in the form of training and testing sets, we rejoined these two subsets to maximize the number of observations in our dataset (n =

[3] Klein, TJ. (2020, June). *Airline Passenger Satisfaction* (Version 1) [Data set]. Kaggle. https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction/activity

119,204) for visualization accuracy. For organization and clarity, we categorized input variables into four logical groups:

1. Demographics (e.g. age, travel type, customer loyalty, etc.)
2. Convenience (e.g. ease of online booking, gate location, delays, etc.)
3. Service experience (e.g. online boarding, check-in service, inflight service, etc.)
4. Comfort, enjoyment and entertainment (e.g. seat comfort, inflight WiFi, cleanliness, etc.)

To obtain a greater understanding of variable distributions, we printed descriptive statistics, such as minimum, maximum, mean, and median values, among others, for all pertinent variables (Table 1B in Appendix B). While the majority of variables are relatively evenly distributed, departure and arrival delays are noticeably skewed. Specifically, the mean and median average departure and arrival delays are approximately 15 and 0 minutes, respectively, but both variables have maximum values of around 1600 minutes.

Dashboard 1A in Appendix A displays the distributions of several variables in each of the above categories. Variables are filtered to include individuals who are between 48 and 52 years old, as this age range was the most common among passengers. Factors, such as arrival and departure delays, that cannot be controlled by airlines or airports are omitted from the visualization.

Overall, we can see that of the individuals between the ages of 48 and 52 years, just over 95 percent were loyal customers and 76 percent were travelling for business purposes. Just under half of this population is male. The most popular rating across most categories is 4, suggesting that most passengers were pleased with their travel experience. However, some rating categories, such as Food and Drink and Inflight WiFi Service, exhibited subpar trends. Specifically, ratings for the Food and Drink category were fairly evenly dispersed over 2, 3, 4, and 5. The most

popular ratings for the Inflight WiFi Service category were 2 and 3, suggesting that on average, customers were only mildly pleased or displeased with this service.

Figure 1A in Appendix A portrays the distribution of the response variable. The percentage of satisfied passengers slightly exceeds that of unsatisfied passengers, but they are relatively evenly distributed and do not spark much concern. Figure 2A in Appendix A presents the relative proportions of satisfied and unsatisfied passengers between the ages of 48 and 52 who belong to a given demographic. While most of these results are as predicted, one particularly interesting finding is that just over 75% of passengers in the economy class were unsatisfied with their experience. Thus, subsequent analysis can aim to explain why this observation may exist. Further analysis can also focus on the rating patterns of different age groups, as older generations might have different preferences than younger generations and vice-versa. Understanding age differences in ratings, among other patterns and trends, would inform airport and airline employees about potential adjustments they can implement to existing services.

Modeling

As mentioned earlier, the main objective in building a model for this situation is to be able to be able to predict whether an airline customer will be satisfied or not satisfied with their experience. The output variable that we are trying to solve for only has those two outcomes, so we classify it as binary. Because the output we are trying to predict is binary, we have to use classification algorithms to build our model. Using a linear regression model, which seeks to predict a continuous output, would not be beneficial in our case.

There are four different models that we considered - classification trees, random forest, support vector machines (SVM) and logistic regression. Classification trees apply attribute selection to

find the best attribute to partition the data set. By doing this, classification trees are one of the easiest models to understand, interpret and visualize. A classification tree can be shown to a management team and understood conceptually. Additionally, it can build a model relatively quickly on a large data set. Two of the downsides are that trees can have high variance, and they can overfit easily. Because of the speed of building the model, as well as the ease of interpretation, we chose classification trees as one of the models to build.

The other model we chose to compare to our classification tree was logistic regression. Logistic regression is able to estimate the probability of membership in class 1. For example, given certain values for all of our input variables, logistic regression may estimate a probability of 0.70 that the person would be 1 - satisfied. We then determine a cutoff point where we say to classify anything above a certain probability as 1 and anything below it as 0. Similar to classification trees, logistic regression is relatively easy and fast to train. Additionally, the results and coefficients are easy to interpret.

The two other models that we considered were random forest and SVM. Because our data set had over 100,000 observations, neither of those models could run efficiently given our limited computing power. Random forest models can give good predictive performance, but are not as easy to explain and implement as classification trees. SVM applies a linear separator between the two classes, with the goal of maximizing the margin. SVM can be easy to interpret, but may not work as well if the data isn't easily linearly separable.

Once we decided on the two models we were going to use, we knew that holdout validation would be important. What this meant is that we were going to split the data into a training and testing set. The training set contained 80% of the data, and that is the data that we would build

our model on. Once we had the model, we used it to make predictions on the other 20% of the data, the test set. We could then compare how the model predictions did to the actual results. This holdout validation is helpful not only to assess the model across different metrics but also to prevent overfitting.

After splitting the data, we built an initial classification tree using the training data, and we set the complexity parameter (cp) to 0 to see the full tree. With a cp of 0, the tree was impossible to visually interpret, so we knew that we wanted to prune the tree to simplify it and get rid of parts that don't provide as much value. To prune the tree we wanted to see which cp value would give us the lowest x error. When we did that, the tree still had 450 nodes, which was excessive for our needs, given one of the benefits of the tree model is that it would be easy to visually comprehend. We decided to sacrifice a little model performance to make the model simpler, and chose a cp value of 8.33e-03. Now we were able to plot our tree and understand it visually, and then also test our model on the testing data set. After testing we built a confusion matrix where we could assess the model performance on a variety of metrics, which will be discussed shortly.

For building our logistic regression model, we used the same 80%/20% split for training and testing data set. We built an initial model using all of our predictor variables. We ran forward and backward stepwise regression, and they both recommended that the best logistic regression model included all of the predictor variables. We did delete the predictor variable arrival delay, because we found that it was highly correlated to departure delay, so it was unnecessary to include both. After building our logistic regression model, we tested it on the test data set and built a confusion matrix to assess the performance. We set the probability threshold at 0.5, so if the model predicted the probability of being satisfied was 0.5 or higher, the observation was considered satisfied.

Results

The classification tree shows that online boarding is the most significant predictor of satisfaction; inflight WiFi service, business travel, and business class are also significant predictors of satisfaction. The logistic regression model validated the predictors. This model also presents online boarding, inflight WiFi service, business travel, and business class as significant predictors at the 0.1 percent confidence level.

Evaluation

Model Performance

To assess the performance of the model, we implemented holdout validation, confusion matrices, and a ROC curve. First, we separated the dataset into training and testing data. We built the models initially on the training dataset and evaluated the model on the test data. Then, we implemented a confusion matrix for the unpruned classification tree. The model has a true positive rate of 95.2 percent and a true negative rate of 94.8 percent. The accuracy rate is 95 percent. Additionally, we applied a confusion matrix to the pruned classification tree model. The confusion matrix of the pruned tree had a true positive rate of 95.4 percent and a true negative rate of 83.4 percent. The accuracy rate of the matrix was 89.6 percent. We created a confusion matrix for the logistic regression model. The model has a true positive rate of 91.2 percent and a true negative rate of 88 percent. The accuracy of the model is 89.8 percent.

We utilized a ROC curve to compare the classification tree model and the logistic regression model. The AUC score ranked the classification model better than the logistic regression model slightly. The AUC score for the classification tree is .925 classification and the score for the logistic regression model is .962. Both models provide good predictions. Even though the

classification tree model scored lower, it is still a good model. Both scores are high, so both models perform well.

Business Case

Our models consistently predicted that online boarding, seat comfort, inflight entertainment, inflight WiFi, and business travel are predictors of satisfaction. For example, an airline that provides inflight WiFi may have higher satisfaction rates than an airline that does not provide this service. For airlines to improve satisfaction scores, they must develop a business case to project improved satisfaction. As airlines adjust the predictors of satisfaction (i.e. improve inflight entertainment, seat comfort, etc.) they must evaluate the impacts of these changes on satisfaction scores. They also must evaluate the predictors of the satisfaction scores. For example, airlines can evaluate which group of customers (i.e. gender, age, class, etc.) responds to the improvements.

Deployment

Now that airlines know which steps during the traveling process are most important to providing satisfactory customer service, airlines can better manage resources and increase consumer confidence and reputation. Specifically, an airline would do following: review the online boarding process on the company's website for improvement, reach out to either R&D or seat supplier to increase seat comfort, get the rights to more movies to increase inflight movie library, decide between increasing wifi capabilities and reducing the price, and finally increasing value of business class.

The model is very accurate and therefore ready for implementation into the travel process for constant feedback and data from surveys. There will be a satisfaction survey at the kiosk after

checking bags and online during inflight WiFi usage. This data will be ported to the data warehouse where it will be prepped and cleaned for the algorithm to be updated. The more data available for analysis the more accurate the model will become, although it is already very precise.

Should business class continue to prove as a significant variable, airlines would need to be careful not to increase its value so much that it creates a class divide to the point that economy class satisfaction drops. Ethically it would reflect poorly on the airlines. As for risks, the process of collecting and analysing the data should have a human component to it to make sure everything is working as planned. Also, customer satisfaction is a never ending struggle as customers will always find something for improvement; therefore, the airline needs to set a goal at which point the company will be happy with customer satisfaction.

Appendix A

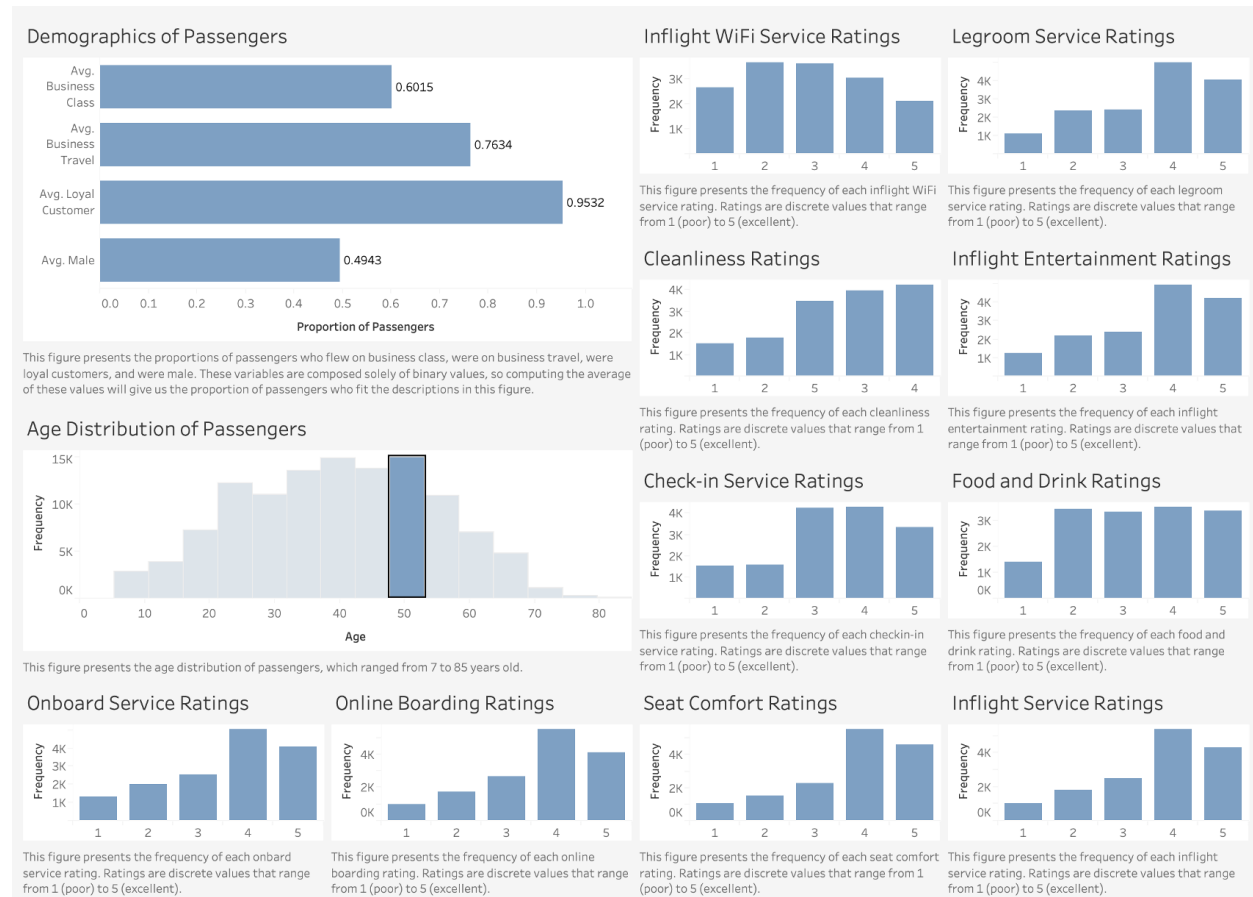## Dashboard 1A. Descriptive Analysis of Input Variables

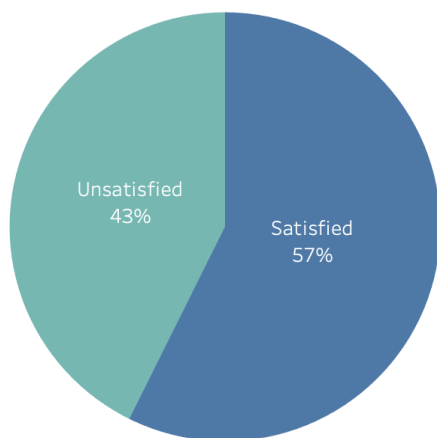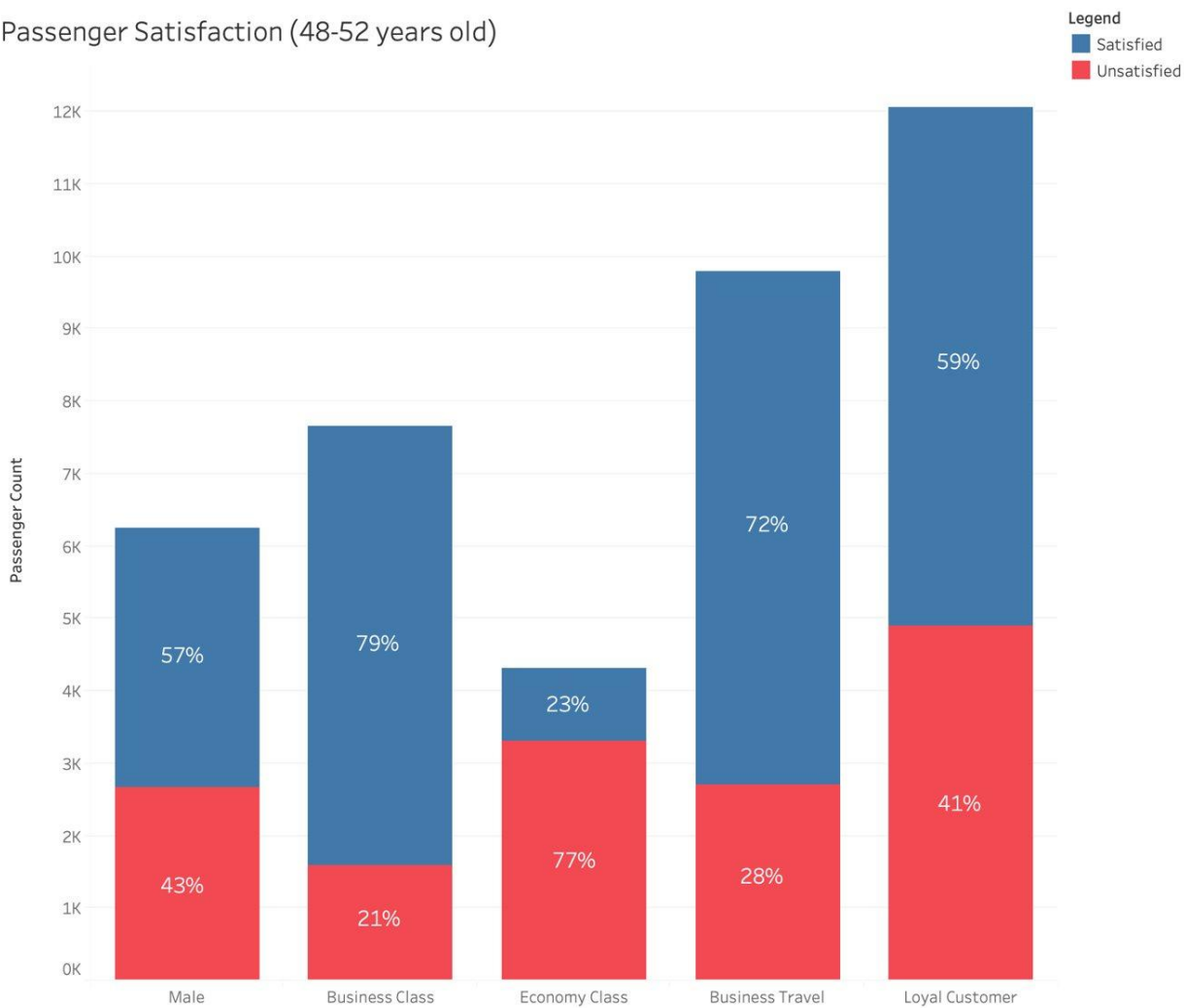
Figure 1A. Distribution of Response Variable

Figure 2A. Passenger Satisfaction Levels by Demographic

Passenger Satisfaction (48-52 years old)

Legend
■ Satisfied
■ Unsatisfied

Appendix B

Table 1B. Descriptive Statistics of All Relevant Variables[1]

	Age	Flight Distance	Inflight Wifi Service	Flight Time Convenience	Ease of Online Booking
Minimum	7	31	1	1	1
1st Quartile	28	440	2	2	2
Median	40	867	3	3	3
Mean	39.86	1224	2.818	3.207	2.879
3rd Quartile	51	1775	4	4	4
Maximum	85	4983	5	5	5

	Gate Location	Food and Drink	Online Boarding	Seat Comfort	Inflight Entertainment
Minimum	1	1	1	1	1
1st Quartile	2	2	2	2	2
Median	3	3	4	4	4
Mean	2.987	3.214	3.331	3.457	3.381
3rd Quartile	4	4	4	5	4
Maximum	5	5	5	5	5

	On Board Service	Leg Room Service	Baggage Handling	Check-in Service	Inflight Service
Minimum	1	1	1	1	1
1st Quartile	2	2	3	2	3
Median	4	4	4	3	4
Mean	3.386	3.381	3.637	3.295	3.647
3rd Quartile	4	4	5	4	5
Maximum	5	5	5	5	5

	Cleanliness	Departure Delay (minutes)	Arrival Delay (minutes)	Male	Loyal Customer
Minimum	1	0	0	0	0
1st Quartile	2	0	0	0	1
Median	3	0	0	0	1
Mean	3.294	14.84	15.28	0.4932	0.8391
3rd Quartile	4	13	13	1	1
Maximum	5	1592	1584	1	1

	Business Travel	Satisfied	Economy Class	Business Class
Minimum	0	0	0	0
1st Quartile	0	0	0	0
Median	1	0	0	0
Mean	0.6916	0.4268	0.4401	0.4865
3rd Quartile	1	1	1	1
Maximum	1	1	1	1

[1]Variables pertaining to observation number and identification were removed from the dataset.