Introduction

What determines a successful airline? While many believe that the travel experience is a primary factor influencing customer ratings, one must consider whether there are hidden elements that also impact these evaluations. How can airlines enhance their services, and on what improvements should companies concentrate? In the competitive landscape of the airline industry, understanding passenger satisfaction is crucial for enhancing service quality and achieving strategic differentiation. This report employs a variety of analytical methods—including Logistic Regression, Decision Trees, LASSO, RIDGE Regression, Random Forest, and Boosted Trees—to tackle the main question: "What factors are associated with passenger satisfaction for an Airline and to predict if a passenger is satisfied with the providing airline's services?" Based on our study, each method is carefully analyzed to determine the most effective approach for evaluating and interpreting the factors that influence customer satisfaction, thus guiding airlines on how to best tailor their improvement strategies.

Data Description

The Airline Passenger Satisfaction is sourced from Kaggle. TJ Klein - the owner of our dataset, published the dataset with two different sets: test dataset and train dataset. The test dataset includes 20% of the full dataset to use for testing and the train dataset includes 80% of the full dataset to use for training. Each of the train and test dataset provides 25 variables for 103,904 observations and 25,976 observations. Airline passengers who completed this survey are our observations. Since data is expensive to collect, we decided to utilize the train dataset with 103,904 observations. The response variable is the binary satisfaction variable with two levels 0 and 1. The 0 level stands for neutral or dissatisfied while the 1 level stands for satisfaction. We remove the number order (X) and passenger id numbers since they are currently not useful for our purpose of data analysis. The 23 possible predictor variables include gender, customer type, age, type of travel, travel class, flight distance, departure delay in minutes, arrival delay in minutes, and other 14 airline related services with the assigned satisfaction level from 0 to 5: 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, cleanliness. There are no overall quality issues, only around 0.3% of the arrival delay in minutes is missing. Since our chosen dataset is large with 103,904 observations, we can afford to drop the 310 missing rows of arrival delay in minutes information. As a result, the final dimension of our dataset is 103,594 rows with 23 variable columns. The Figure 1 below shows histograms of all the potential quantitative variables.

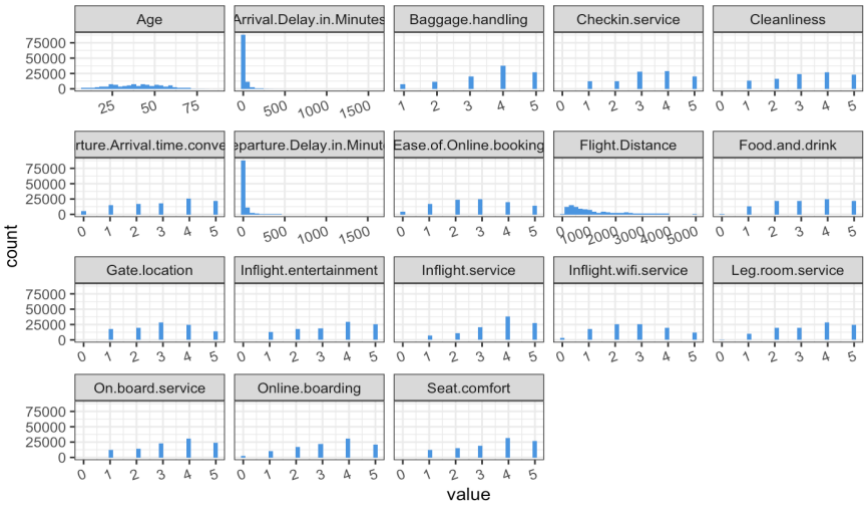

Figure 1, Distribution of Quantitative Variables

As shown above, Figure 1 shows that the airlines have minimal delay. It also includes the fourteen airline services that have 6 levels from 0 to 5. The 0, 1, 2, 3, 4, 5 levels in the Figure 1 stand for void response, very dissatisfied, dissatisfied, neutral, satisfied, and very satisfied, respectively. Figure 1 shows us that the seat comfort, inflight entertainment, on board service, leg room service, baggage handling, and inflight service receive a high

percentage of satisfaction. On the other hand, inflight Wi-Fi service and ease of online booking receive a high percentage of dissatisfaction.

Table 1 below provides detail of summary statistics for some of key quantitative variables in our dataset. Age, indicating the actual age of passengers, averages 39 years old with a standard deviation 15 years old. Departure delay in minutes, indicating minutes delayed when departure, have an average of around 15 minutes with a standard deviation of 38 minutes. Similarly, arrival delay in minutes, indicating minutes delayed when arrival, have an average of 15 minutes with a standard deviation of around 39 minutes. Lastly, the flight distance, indicating the flight distance of the airplane providing journey, has an average of 1,189.33 miles with a standard deviation of 997.30 miles.

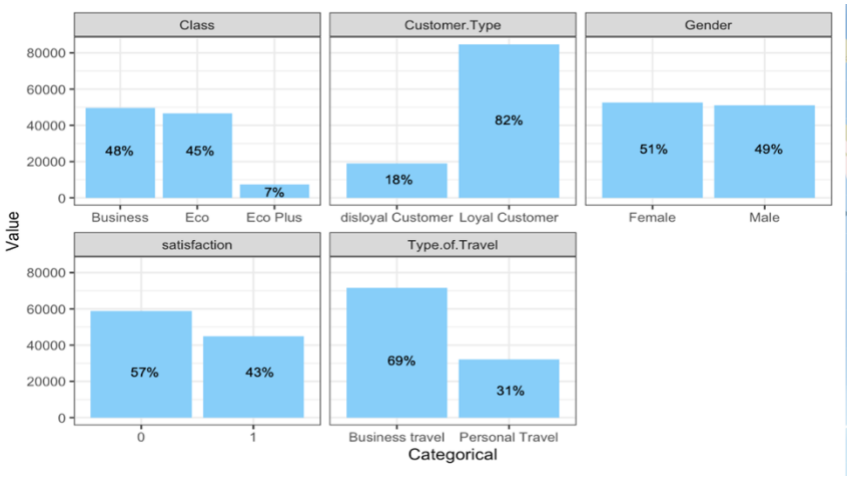

Figure 2, Distribution of Categorical Variables

Characteristic	N = 103,594[1]
Age	39.38 (15.11)
Departure.Delay.in.Minutes	14.75 (38.12)
Arrival.Delay.in.Minutes	15.18 (38.70)
Flight.Distance	1,189.33 (997.30)

[1] Mean (SD)

Table 1, Quantitative Variables Summary Statistic

Figure 2 above displays the bar charts and summary statistics for our key categorical variables in our dataset. Satisfaction 's bar chart, which is our binary response variable, indicating airline satisfaction levels of 0 – neutral or dissatisfaction and 1 – satisfaction, shows that the percentage for level of neutral or dissatisfaction is 57% while the percentage for level of satisfaction is 43%, reveal that the passengers are not satisfied with the airlines. Class, indicating travel class in the plane of the passengers has three levels for business, economy, and economy plus, has the percent of 48%, 45%, 7%, respectively. Business class are the most common passengers, followed by economy class. Majority of the airplane passengers fall into business class and economy class while economy plus class is the least common. The customer type 's bar chart compares the counts of disloyal and loyal customers. Loyal customers are the most common, accounting for 82% of the passengers compared to 18% of disloyal passengers. Gender, indicating the gender of the passengers either female or male, having a similar percentage which are 51% and 49%, respectively. Type of travel, indicating purpose of the flight of the passengers for either personal travel or business travel, showing that the percentage of business travel is more common than personal travel. Business travel includes 69% of the passengers while personal travel only includes 31% of the passengers.

Models

Multiple Logistic Description

To assess airline passenger satisfaction, we first perform a logistic regression analysis. Logistic regression analysis can help us to predict the probability whether a customer is satisfied with the airline. The model operates on odds, meaning that the ratio of the probability satisfaction over the probability of the customer dissatisfied.

We fit the model by using the glm( ) function, which helps us to predict a binary outcome. The summary of the glm( ) function results shows that there is only one variable, Flight Distance, that has a high P value of 0.1140, which is greater than the traditional 0.05 alpha value. Therefore, we used the backward elimination method to remove the variable. We refit the model and the result shows that all the predictors are significant. The refined model includes 22 predictors with one-unit higher AIC score than the previous model. Therefore, we choose to use the reduced model because the reduced model has similar performance but with fewer variables.

Further, we evaluate the reduced model by using the confusion matrix and ROC curve with the testing data set. A confusion matrix (Table 2) describes the performance of our logistic model. We use the result from the matrix to calculate the model's accuracy, sensitivity, and specificity. The model's accuracy is at 87.0%, which reflects the overall likelihood that the model correctly predicts whether passengers are satisfied or not. Sensitivity, also known as the true positive rate, was calculated at 82.9%, indicating how well the model identifies actual satisfied customers in the test data set. Specificity, the true negative rate, stood at 90.2%, showing a high accuracy the model identifies the neutral or dissatisfied passengers.

		Actual	
Prediction	*Neutral or dissatisfied*	*Satisfied*	**Sum**
Neutral or dissatisfied	13146	1940	15086
Satisfied	1427	9463	10890
Sum	14573	11403	25976

Table 2, Confusion Matrix for Logistic Regression

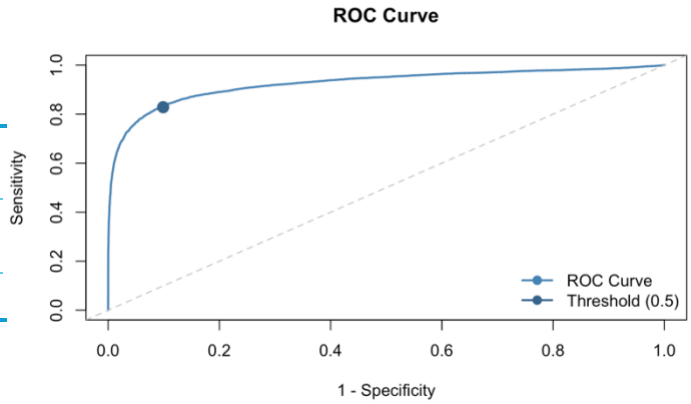

Figure 3, ROC Curve for Logistic Regression

We also use the ROC (Receiver Operating Characteristic) curve to evaluate the performance of our model as its discrimination threshold is varied. In Figure 3, the curve quickly rises towards the top-left corner and then flattens out, which indicates a high area under the curve (AUC) and suggests a good model performance. The model shows a high true positive rate while maintaining a low false positive rate. The Area Under the Curve (AUC) was conducted to measure the ability of the model to avoid false classification. With a high AUC value of 92.55%, our model reflects a high level of discriminative ability for the logistic regression model in distinguishing between the two categories of passenger satisfaction. Therefore, our final logistic regression model comes below:

$$log(\frac{p(x)}{1-p(x)}) = -7.88 - 2.72 \times Type\ of\ Travel\ Personal - 0.84 \times Class\ Eco\ plus - 0.73 \times Class\ Eco - 0.14 \times Ease\ of\ Online\ booking - 0.12 \times Departure\ Arrival\ time\ convenient - 0.028 \times Food\ \&\ drink - 0.0094 \times Arrival\ Delay\ in\ Minutes - 0.0082 \times Age + 2.02 \times Customer\ Type\ Loyal + 0.61 \times Online\ boarding + 0.40 \times Inflight\ WiFi\ service + 0.29 \times Gate\ location + 0.32 \times Check\ in\ service + 0.30 \times On\ board\ service + 0.25 \times Leg\ room\ service + 0.22 \times Cleanliness + 0.13 \times Baggage\ handling + 0.12 \times Inflight\ service + 0.065 \times Inflight\ entertainment + 0.065 \times Seat\ comfort + 0.0047 \times Departure\ Delay\ in\ Minutes + 0.043 \times Gender\ (Male)$$

This logistic regression model suggests that loyalty status, service quality across multiple touchpoints (like boarding, inflight service, cleanliness), and convenience features (like gate location and leg room) play significant roles in enhancing passenger satisfaction. However, variables like personal travel, traveling in economy classes, certain perceptions about the booking process, convenience, and food, as well as older age and longer delays, are all associated with lower satisfaction.

Decision tree

Decision trees are flowchart-like structures that are widely used in both regression and classification problems. The tree automatically selects the best rules for finding the most suitable attributes to make decisions regarding customers' satisfaction. Our decision nodes are based on online boarding, inflight Wi-Fi service, and type of travel.

The box at the top of Figure 4 shows the root node of the decision tree. The first decision node splits on "Online boarding < 4". If the online boarding rating is less than 4, it further considers the type of travel to the right side of the tree. Conversely, ratings of 4 or higher lead to the left side and toward an evaluation of inflight Wi-Fi service. On the left side branch, the tree shows that if the inflight Wi-Fi service rating is between 1 and 4 customers will be neutral or dissatisfied. If the inflight Wi-Fi service rating is bigger than 4 or less than 1 the customer will be satisfied. On the right-side branch if the customer type of travel is not personal travel, the

customer will be satisfied. If it is for personal travel, satisfaction depends on the inflight Wi-Fi service: a rating below 5 typically results in dissatisfaction. The decision tree analysis indicates that online boarding quality and inflight Wi-Fi service are significant predictors of passenger satisfaction. Higher ratings in these categories generally lead to higher satisfaction, especially among passengers traveling for personal reasons.

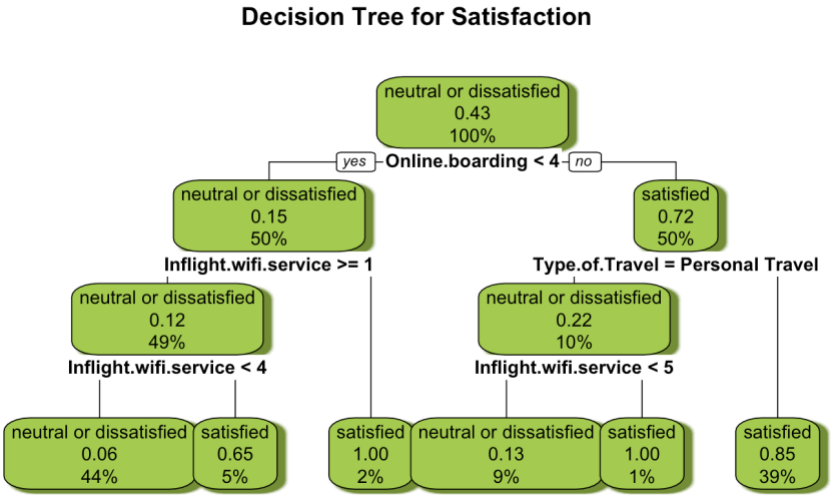

Figure 4, Decision Tree

LASSO

Regression regularization is a technique used to overcome deficiencies in regression models by adding a penalty term to the error function. This penalty discourages overly complex models by penalizing large coefficients or shrinking the coefficients to zero. The LASSO model is part of the regression regularization family that adds a penalty term, so-called the L1 regularization, proportional to the absolute value of the coefficients. It tends to shrink coefficients all the way to zero, effectively performing feature selection and handling multicollinearity issues.

Our LASSO model is fitted using R's **glmnet** engine with the L1 term equals to 0.001. The L1 term is found by fitting the model with a range of L1 input then extracting the L1 term corresponding to the largest accuracy. Referring to Figure 8, the accuracy stands at about 87.54% on the training set and 87.14% on the testing set. It indicates that the overfitting is minimal. Automatic feature selection can be seen in the LASSO model that "Departure Delay in Minutes" has a coefficient of zero, effectively being dropped out due to a strong correlation with the predictor "Arrival Delay in Minutes". The top five predictors by absolute coefficients include the type of travel, customer type (loyal), flight class (eco or eco plus) and online boarding.

Referring to Figure 5, the accuracy stands at about 87.54% on the training set and 87.14% on the testing set. It indicates that the overfitting is minimal. Automatic feature selection can be seen in the LASSO model that "Departure Delay in Minutes" has a coefficient of zero, effectively being dropped out due to a strong correlation with the predictor "Arrival Delay in Minutes". The top five predictors by absolute coefficients include the type of travel, customer type (loyal), flight class (eco or eco plus) and online boarding. The LASSO model in Figure 5 below indicates that loyal passengers traveling for business purposes are more likely to be satisfied. Passengers also prefer higher flight class with use of online boarding.

RIDGE

Ridge regression is another popular model under the regression regularization family that adds a penalty term, so-called the L2 regularization, proportional to the square of the coefficients. It shrinks the coefficients toward zero but rarely all the way to zero, keeping all features in the model but with reduced impact, as to prevent overfitting.

The RIDGE is fitted using R's **glmnet** engine with the L2 term equals to 0.001. The L2 term is found by fitting the model with a range of L2 input then extracting the L2 term corresponding to the largest accuracy. Referring to 5, the accuracy stands at about 87.29% on the training set and 87.08% on the testing set. It indicates that the overfitting is minimal. The top five predictors by absolute coefficients include the type of travel, customer type (loyal), flight class (eco or eco plus) and online boarding, which are the same as the LASSO model with

slightly different coefficients. The RIDGE model in Figure 5 also indicates that loyal passengers traveling for business purposes are more likely to be satisfied. Passengers also prefer higher flight class with use of online boarding.

LASSO (87.14%)

TERM	ESTIMATE
(Intercept)	-7.74
Type of travel Personal Travel	-2.67
Customer type Loyal Customer	1.95
class- Eco Plus	-0.79
class- Eco	-0.71
Online boarding	0.6
.	
.	
.	
Flight distance	0
Departure delay in minutes	0

RIDGE (87.08%)

TERM	ESTIMATE
(Intercept)	-6.57
Type of travel Personal Travel	-1.85
Customer type Loyal Customer	1.3
class- Eco Plus	-0.79
class- Eco	-0.71
Online boarding	0.47
.	
.	
.	
Flight distance	-0.000937202
Departure delay in minutes	8.27751E-05

Figure 5, Regression Regularization model output.

Random Forest

Ensemble trees are a type of ensemble learning method that combines multiple decision trees to create a more powerful model. Random Forests is part of the regression regularization family that combines the predictions of multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the training data and a random subset of the features. When making predictions, the random forest averages the predictions of all the individual trees (for regression) or uses a majority vote (for classification).

Our Random Forest is fitted with R's **ranger** with the number of trees set to 50. The accuracy stands at about 98.97% on the training set and 96.29% on the testing set. We can see that there is a small overfitting as the difference between the training and testing accuracy is about 2.7%. Referring to Figure 6, the feature importance chart from the ranger package shows that WiFi service, online boarding, type of travel, flight class, inflight entertainment, seat comfort, leg room, ease of online booking and on board service are the top 10 most important factors. The Random Forest model indicates that these top 10 factors from the feature importance chart contribute the most to the model to classify a satisfied passenger, but we do not have the directions and magnitudes of how these factors impact the satisfaction of a passenger.

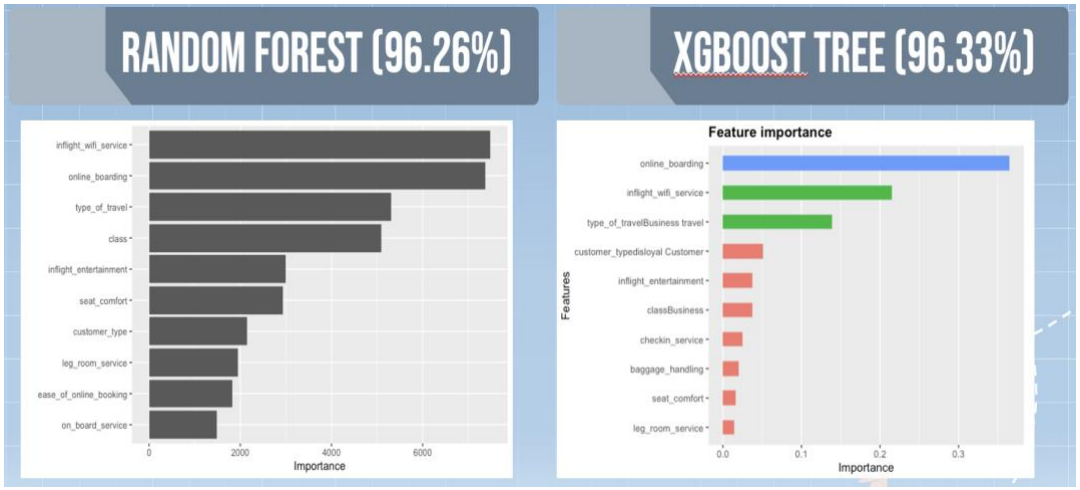

Figure 6, Ensemble Trees output

Boosting Tree

Boosting trees is part of the regression regularization family that uses an ensemble learning technique that combines the predictions of several base estimators to improve robustness and accuracy. In essence, each Boosted

tree works by iteratively training decision trees on the residuals of the previous trees, combining their predictions to improve accuracy, with each subsequent tree focusing more on the difficult-to-predict cases.

Our Boosting Tree model is fitted with R's **xgboost** with the number of trees set to 50. The accuracy stands at about 96.77% on the training set and 96.34% on the testing set. Referring to Figure 6, the feature importance chart from the **xgboost** package shows that online boarding, WiFi service, type of travel, inflight entertainment, flight class, check-in service, baggage handling, seat comfort and legroom are the top 10 most important factors. The Boosting Tree model indicates that these top 10 factors from the feature importance chart contribute the most to the model to classify a satisfied passenger, but we do not have the directions and magnitudes of how these factors impact the satisfaction of a passenger.

Model Comparison

As mentioned above, logistic regression (LR) is a statistical model used for binary classification that estimates the probability of an instance belonging to a certain class based on its features, fitting a linear decision boundary. Decision trees, on the other hand, are non-linear models that recursively split the data into subsets based on the features, making them suitable for both classification and regression tasks, with the ability to capture complex interactions in the data. While LR is efficient, robust, highly interpretable, and theoretically sound, it has its shortcomings applying to the subject dataset. Firstly, the subject dataset is moderately large in predictor dimensions, LR may not be ideal handling large dimensions and needs multiple steps to achieve a parsimonious model. Moreover, the subject dataset is survey-based information, as a result the questions (i.e. the predictors of LR) could be linearly or non-linearly correlated. Again, extra analysis and/or transformations may have to be performed to achieve a parsimonious model, while LR could be deemed to be an inappropriate model in the end if the violations of assumptions cannot be fixed after all the analysis. On top, such analysis/transformations could require manual interventions that does not make LR to be suitable for an automatic classification system.

As a result, Regression Regularization and Ensemble Trees become a reasonable approach, which have better interoperability, high dimensionality and efficiency. Both Regression Regularization and Ensemble Trees offer models that are relatively interpretable compared to traditional logistic models. In Regression Regularization, the regularization term helps to control the complexity of the model, leading to more interpretable coefficient estimates. Ensemble Trees provide insights into feature importance since each decision tree learns to make predictions based on the feature available. Regression Regularization, LASSO and RIDGE, automatically handle multicollinearity and prevent overfitting when dealing with big datasets by adding the additional penalized terms. Ensemble Trees, meanwhile, can handle high dimensionality by selecting subsets of features at each split, effectively reducing the search space and mitigating the curse of dimensionality. Logistic regression, LASSO, and RIDGE models offer nearly closed-form solutions, making them relatively efficient. However, Random Forests and Boosted Trees involve multiple decision trees, thus requiring more time to fit given the same computing resources. Unlike Random Forests construct multiple trees independently, Boosted Trees construct them sequentially. This implies that Boosted Trees cannot utilize parallel computing, making them somewhat less efficient than Random Forests.

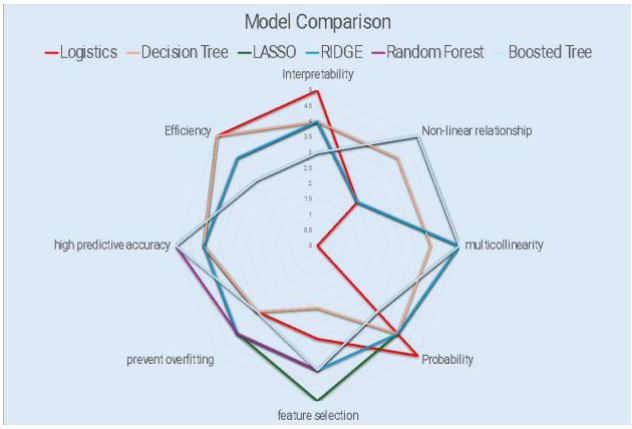

Figure 7, model features comparison.

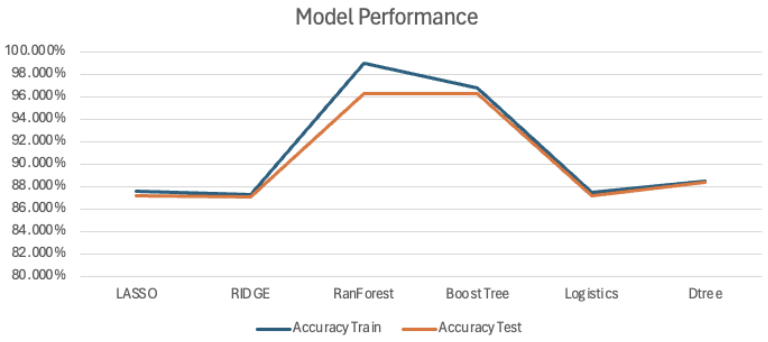

Figure 8, Train vs Test Accuracy

Figure 7 above is a strength radar chart that compares the strength of all the classifiers to be used in this project. For interpretability, logistics and decision trees generally help decision makers to arrive in a transparent decision. In terms of non-linear relationship and multicollinearity, one must be aware that logistic models may not be well suited. Random Forest and Boosting Trees could better handle these issues and perform automatic feature selection and provide better accuracy in the cost of computing efficiency. Automatic feature selection is a process of selecting the most relevant features from a large set of features in a dataset. It helps reduce the dimensionality of a model and reduce the risk of overfitting.

The test accuracies of all the models are high, ranging from 87% to 96%. The regression based models that include Logistic, LASSO and RIDGE perform similarly having both train and test accuracies just below 88%. They have similar accuracies since all three models have similar significant predictors and coefficients. The decision tree model has a slightly higher accuracy slightly above 88%. The small improvement could be due to the ability of handling potential non-linear relationships in the dataset. The ensemble tree family, including Random Forests and Boosted Trees, perform significantly better with test accuracy above 96%. The improvement could be due to the models' strong ability to deal with non-linear relationships and multicollinearity by resampling the original dataset. To check if there exists any overfitting on our fitted models (i.e. high accuracy in training set but low accuracy in testing set), we compare the accuracies between training set and testing set for each model. Per Figure 8 above, we see that there is no significant drop in accuracy in the testing set, as the largest drop with Random Forest is relatively small, about only 2%.

Conclusion

To conclude, the first objective of this project is to find the factors that are important to a passenger's satisfaction. We have analyzed the strength of different models and employed six models (Logistic, Decision Tree, LASSO, RIDGE, Random Forest, and Boosting Tree) to come up with the top five important factors by averaging the model outputs: Online Booking, WiFi availability, Type of travel (business or pleasure), Customer types (loyal member) and Flight Class (business/economy). These findings can be used by airlines to improve their passengers' satisfaction. For instance, an airline could refine its online booking system, add WiFi hotspots, promote its loyalty membership, and improve economy class experience.

With the second objective, we switch gears to focus on the predictive power of the model to correctly classify if a passenger would be satisfied given we have some information on a passenger. This classification could help an airline to better locate resources. For instance, different marketing or advertising strategies could be given to a passenger to maximum revenue. We found that the ensemble trees class of classifiers (Random Forests and Boosted Trees) generally perform better in correctly classifying a satisfied passenger, in the cost of computing efficiency.

Through answering the two main questions in this project, we have also analyzed extensively the important factors of a good classifier and the positions in these factors of the six models we employed in this project. In short, when selecting a model, we will face the tradeoffs between interpretability and accuracy among other important factors such as the ability to handle non-linear relationships, multicollinearity, automatic feature selection. Generally speaking, logistics and decision trees are preferred if interpretability is important while Random Forest and Boost Trees perform well in predictive accuracy. However, there is no one-model-fits-all solution. One must thoroughly analyze the purpose of the classification task, the data structure, expected outcome and computing resources when choosing a preferable classifier.

Potential improvements of this project include performing model diagnostics and data transformations if required. For other classifiers we deployed in this project, we could perform Grid Searches on the hyperparameters such as the L1/L2 penalty terms for Regularization Regressions and the depth of level and number of trees for Ensemble Tress, to find the best combination of hyperparameters to achieve the best predictive accuracy on our dataset. Furthermore, we could consider employing other classifiers such as the KNN, Naïve Bayes and FNN for our comparison.