Airline Satisfaction Report

Peter Kryspin, Shijie Li, Ke Wu, Yu Luan, Viona Xu
Stat 451 Project Group 28
12/14/22

Introduction

Commercial aviation is a critical economic engine and helps drive more than 10 million jobs and increase about 1.3 trillion GDP in the US. Thus, our group got interested in exploring an airline dataset from Kaggle[1] to predict whether a passenger feels satisfied during the trip, and determine which features are the most important factors.

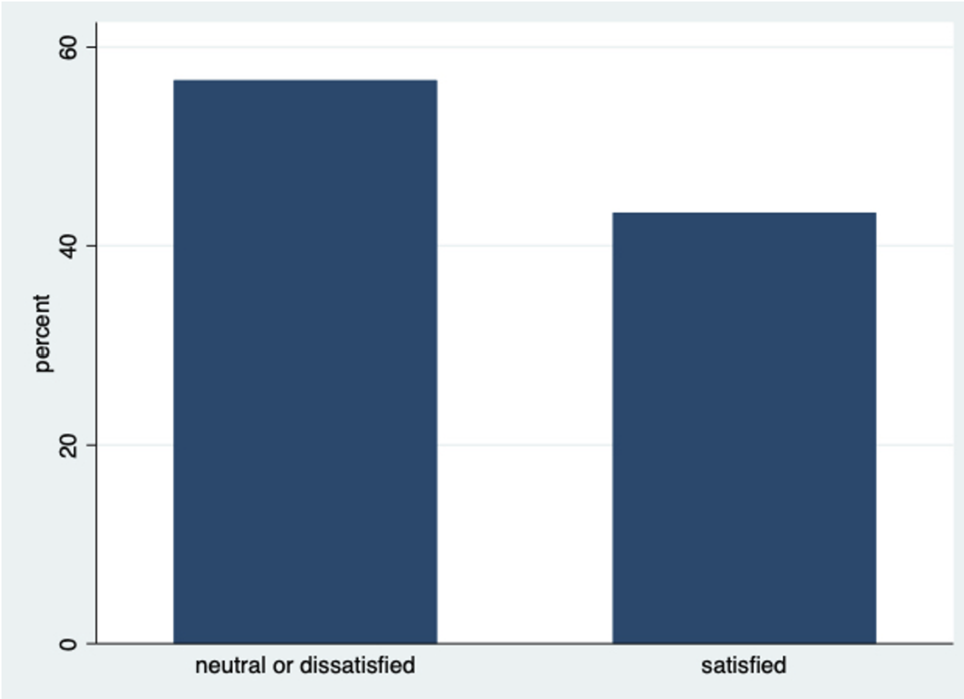
The data have been split into training (100k rows) and testing (25K rows). There are 22 features, consisting of numerical and categorical data. Our group used GUIDE and RandomForest to investigate feature importance and fit different models to the dataset.

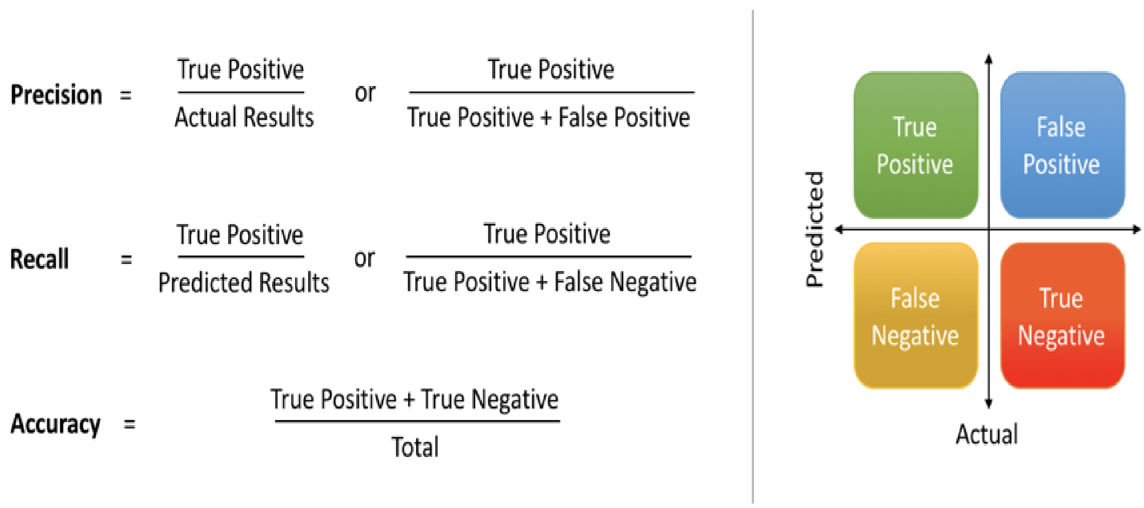
Decision tree and stacking model had the best performance on our data. Through feature importance and the coefficients of logistic regression, our results suggest improving online boarding and inflight wifi service may significantly improve passenger satisfaction. Other factors are important but some have a strong correlation with other features and may need further research.

Dataset

Our response variable was airline passenger satisfaction, a binary variable with levels "satisfied" and "neutral/dissatisfied" (coded as 1 and 0), respectively. Approximately 40% of the data contained "satisfied" customers so we do not need to worry about our data being imbalanced. For each model, we recorded their accuracy, precision, recall, and area under the ROC curve (AUC).

[1] https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction



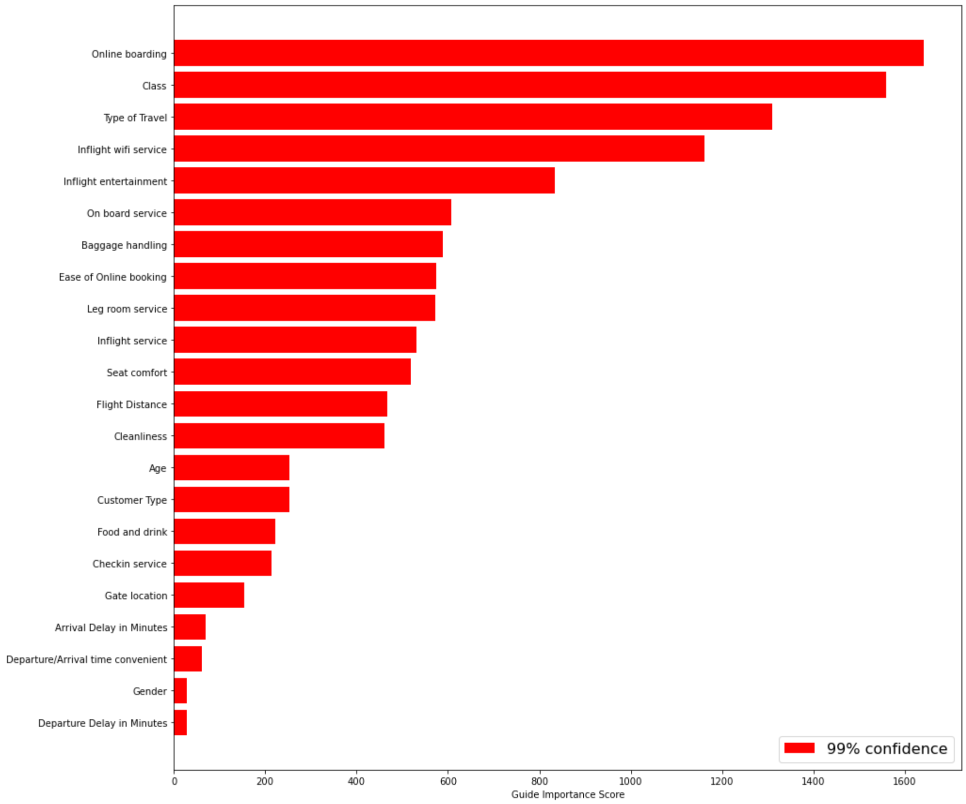


Feature Selection

GUIDE[2] importance score

Here, we employed it to first give a big picture of the feature importances. The results are shown below:

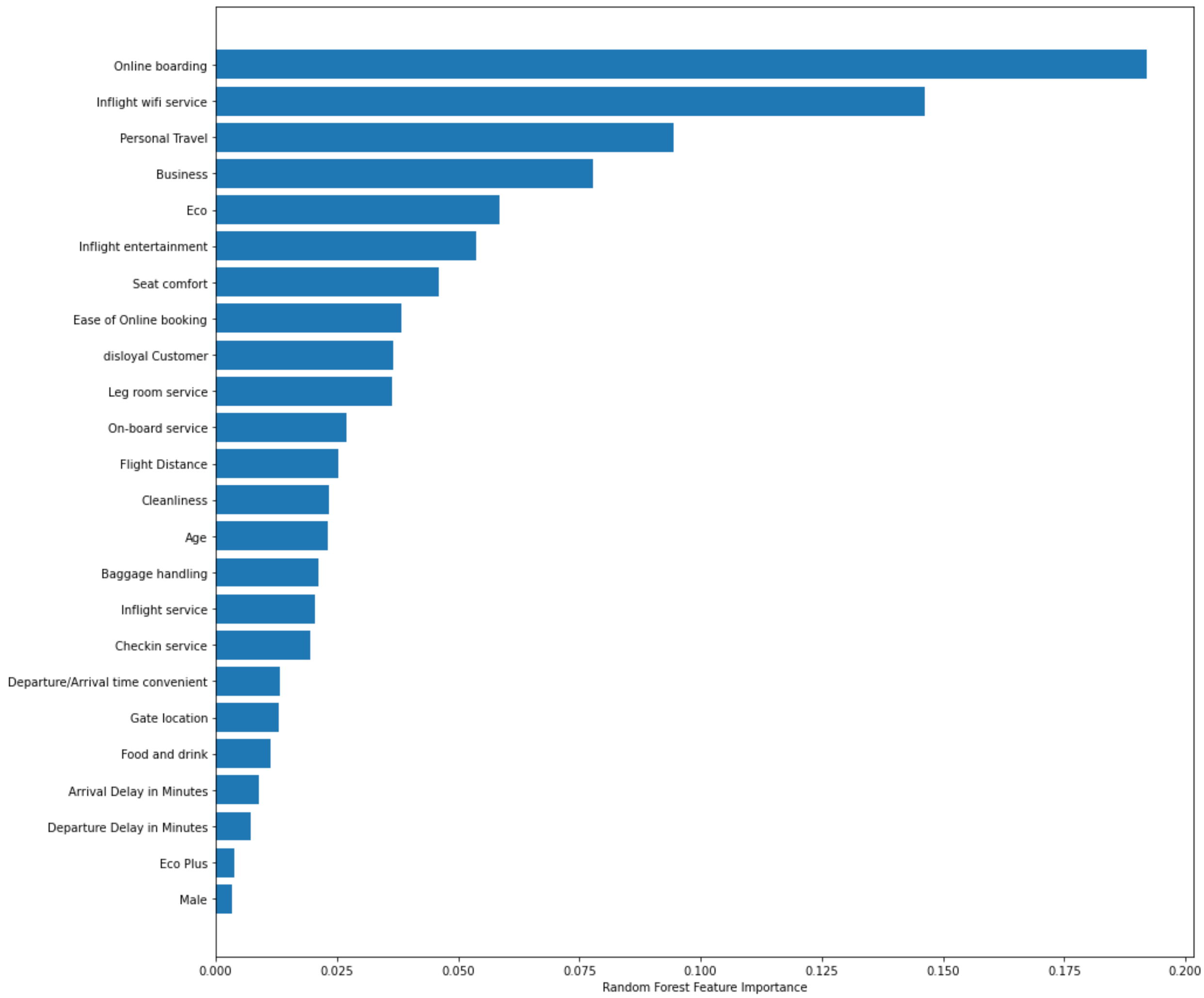
[2] https://pages.stat.wisc.edu/~loh/guide.html



The graph above shows that all features contribute to the results to a great extent.

RandomForest importance

To further investigate how features contribute to the response, we used RandomForest feature selection in advance. Below is the result:



Comparing two graphs, they agree to a great extent. The x-axis in the above graph indicates drop in accuracy if deleting that feature.

Example interpretation:

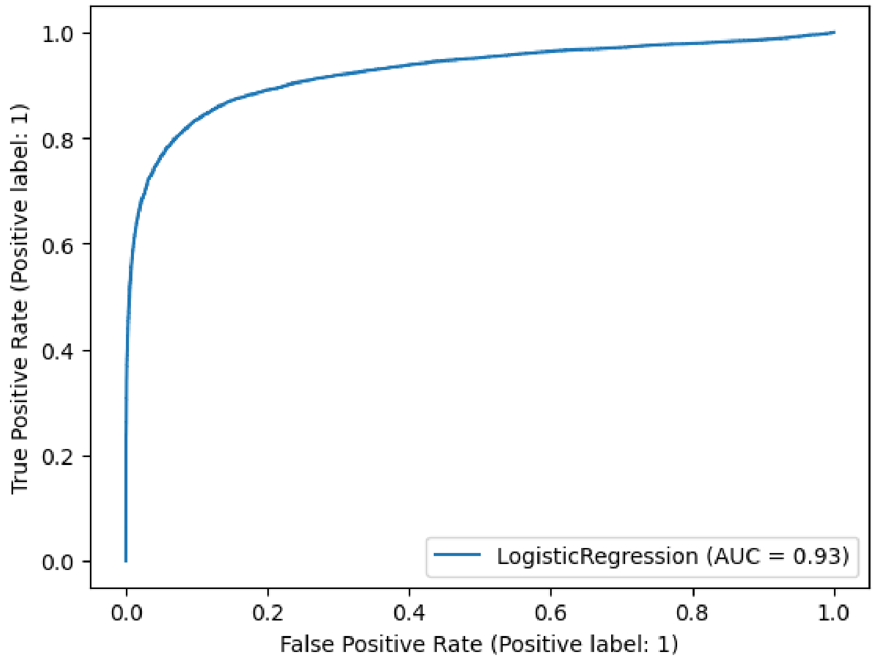
- Online boarding: each unit increase in online boarding satisfaction is associated with 14%[3] higher probability of satisfaction probability

Models

Logistic regression

Grid search found the parameters that produced the highest accuracy (.872 AUC=.926). “Arrival Delay in Minutes” and “Departure Delay in Minutes” were the most significant negative predictors for satisfaction while the most positive were “Seat Comfort” and “Inflight Wifi Service”.

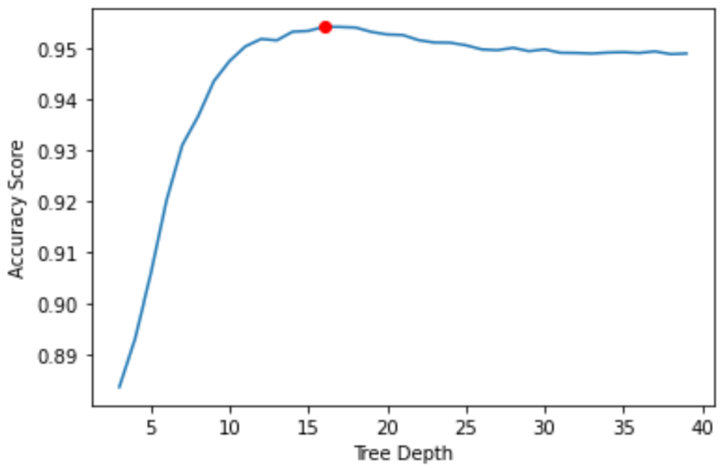
[3] e^0.131 - 100%

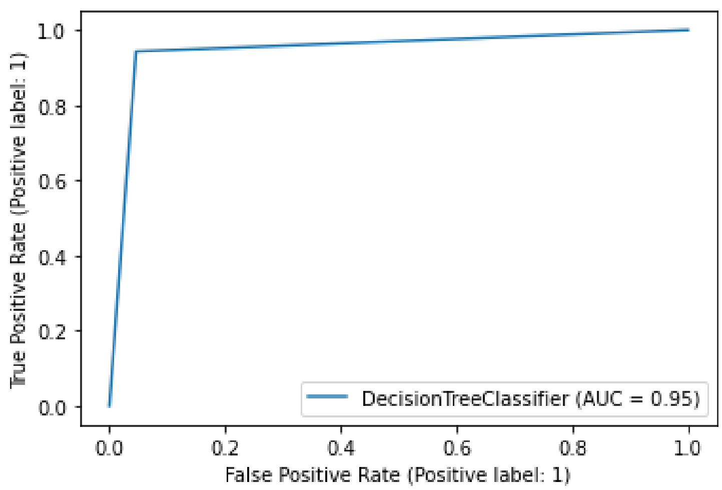


Decision Tree

By using cross-validated grid search, we found max-depth = 16, and the “Entropy” criterion is most accurate. With those hyperparameter settings, we got a very strong model to predict the airline satisfaction level. The accuracy, precision and recall were around .95 and had a high training speed.

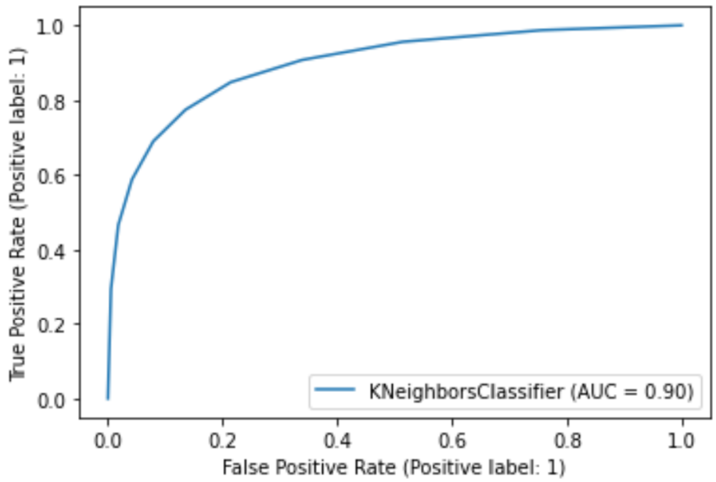
```
best depth=16
```





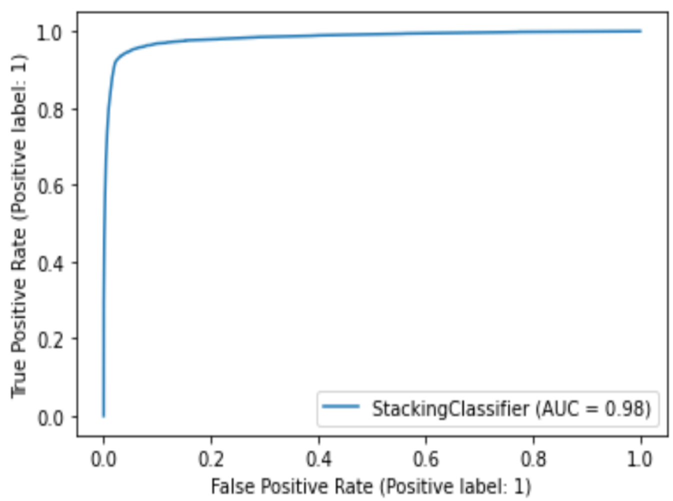
K Nearest Neighbors

With cross-validated grid search method, model performances were similar for n-neighbors between 5 and 20 and “manhattan” distance metric does the best performance. With n-neighbors=9, and “manhattan” metric, the model yields accuracy & precision of .82 and AUC of .9.



Stacked Model

Finally, we stacked all three models with the same parameter settings. The result of the stacked model is similar to the decision tree(Accuracy: .95; Precision: .95; Recall: 0.94; AUC: .98).



Conclusion

In summary, all the models have a good performance on our test dataset, but the decision tree and the stack model work much better. One possible reason might be that there are many descriptive features in our dataset, and some of the numeric features are categorical scales rather than pure numbers. Our results suggest that the decision tree model might work well on categorical data, which is consistent with entropy theory since average entropy calculation is not sensitive to the variables scaling. In addition, according to our random forest model and the coefficients of lasso logistic regression, we found that the online boarding, inflight wifi service, type of travel, and customer type might be the most relevant features of our model. The airline may invest more in improving online boarding and inflight wifi service to boost passengers' experience. However, the type of travel and customer type are two ambiguous features to improve. The result shows that business travel passengers and loyal customers are more likely to be satisfied with their flight experience. In contrast, personal travelers and disloyal customers are likely to make unsatisfactory comments about their flight journey. Both the travel type and customer type are closely correlated with other factors, which requires further research to analyze the correlation between independent variables.

Contributions

Member	Proposal	Coding	Presentation	Report
Peter	1	1	1	1
Shijie	.9	1	1	1
Ke	.9	1	1	1
Viona	.5	1	1	1
Yu	.5	.5	1	1