Predicting Airline Customer Satisfaction using PySpark

Kunaal Garodia, Andres Soto Plaza, Alex Torres, Joseph Motta

Overview

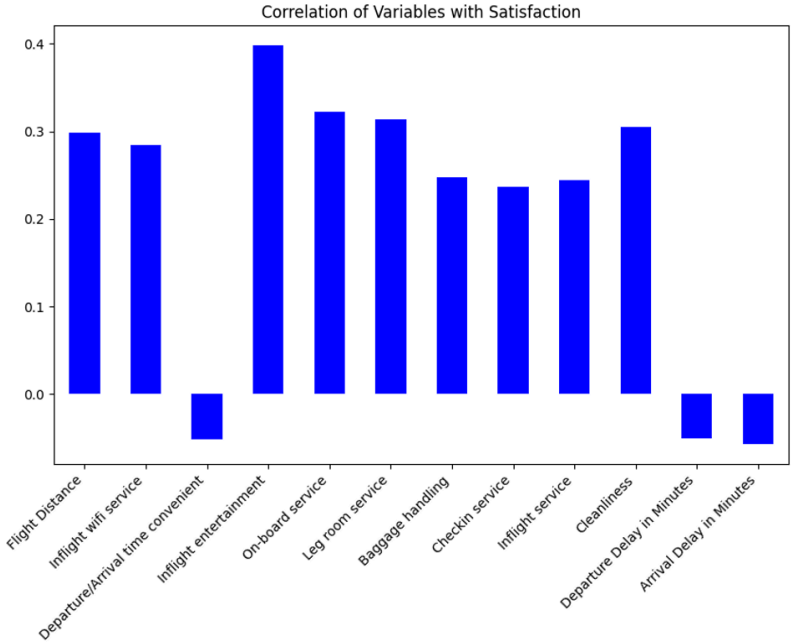
Understanding and predicting customer satisfaction is crucial for businesses aiming to retain customers and improve service quality. This is especially true for the competitive airline industry, where passengers have numerous options and often make decisions based on service reputation, comfort, and overall experience. To investigate this problem, we chose the Airline Passenger Satisfaction dataset from Kaggle, which included features such as demographics, flight details, and in-flight service ratings for over 100,000 passengers. Given the size of the dataset and the fact that airlines manage millions of passengers, this problem is a strong example of a real-world big data challenge.

Prediction Goals

Our goal for this project was to accurately predict customer satisfaction, a binary variable calculated based on the other features in the dataset. We also wanted to identify features passengers cared the most about and how they impacted overall satisfaction. This would help airlines identify pain points and enhance customer service to gain a competitive advantage in a crowded market.

Data Exploration

After loading our data into a PySpark dataframe, our first step was to get a high-level overview of our features. The categorical variables in our dataset included gender, type of travel, class, etc. A majority of our variables were numerical ratings on a scale of 1 to 5 for different aspects of the customer experience. This included everything from gate location to seat comfort and in-flight Wi-Fi service. We ran a simple correlation test to see how closely certain variables were related to customer satisfaction.



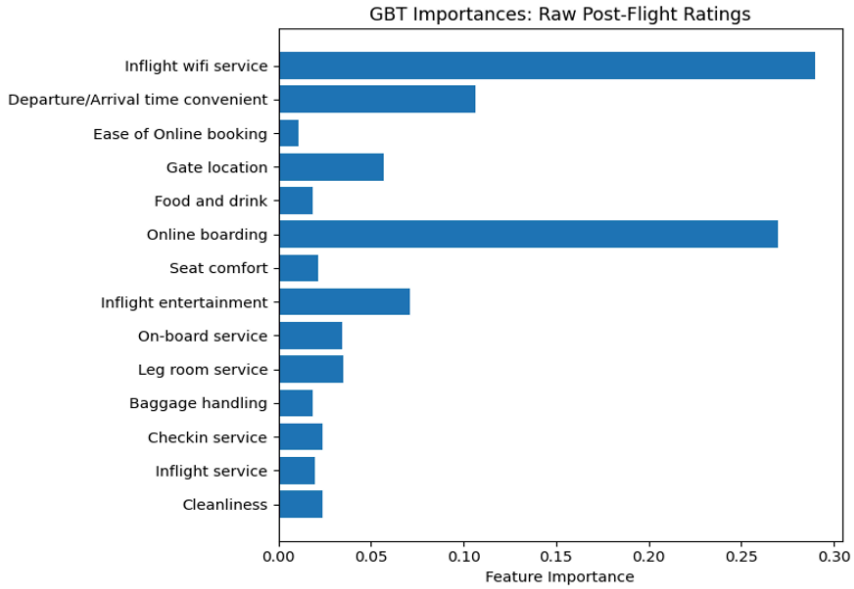
Methods

We performed some basic data preprocessing, such as string indexing and OneHotEncoding, to prepare our data for model building. Due to the large number of variables in our dataset, we decided to run a PCA analysis to see if we could cut them down without losing any important information. We found that the only viable feature reduction would involve cutting down 2-3 columns, as the majority of features proved to retain high levels of significance in determining our target variable (customer satisfaction).

The classification models we chose to use were logistic regression, random forest and a gradient boosted tree. We ran a handful of separate iterations using each classification model and saw promising results with each, however, we saw the best performance in the gradient boosted tree, followed by random forest, and ending with the comparatively worst performance using the logistic regression. Given these results, our final most effective model ended up being a gradient boosted tree.

Interesting Results

After running all three models, we discovered that the GBT classifier performed the best at predicting customer satisfaction with an 86% success rate. This was followed by the random forest, which gave us an 85% success rate and lastly the logistic regression classifier with 83%. It was interesting to look at which features customers valued the most and whether they were in line with what we would have guessed. The top four features were in-flight Wi-Fi service, online boarding, departure/arrival time convenience and in-flight entertainment. While we would not have guessed Wi-Fi to take the number one spot, it makes sense that customers value it highly, given the increase in Wi-Fi availability on planes and the need for people to be connected at all times, even 35,000 feet in the air. We were a little surprised by the low importance of some features, such as food and drink and seat comfort, but realized that these could be heavily influenced by how long the flight is.



Problems Encountered

By far the most significant problem we encountered was a complete pivot, which we had to conduct midway through our development process. Our pivot involved changing datasets from one that dealt with sepsis occurrence in a time-series format to the dataset presented in this report (airline satisfaction data). We ended up making the pivot decision on the basis that–even after thorough preprocessing and initial analysis–the time-series aspect of the initial dataset would prove to create an extremely difficult situation that would involve excessively advanced methods to achieve what we wanted to accomplish. While this was an unexpected and difficult roadblock to overcome, we were able to identify the airline satisfaction dataset and achieve our desired model efficacy on it using the methods that we intended to use on the initial sepsis data.

In terms of problems related to the airline satisfaction model specifically, the greatest difficulties dealt with were our PCA of the dataset's initial features and our effort to refine the final model through various iterations of classification. With that said, the airline dataset did not bring us any unusually severe problems, and all of the roadblocks we encountered with it were reasonably assumed to be possible obstacles at the start of our project.

Citations

Klein, TJ. "Airline Passenger Satisfaction." Kaggle, 20 Feb. 2020, www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.