Data Analytics Bootcamp by EDX.org

Airline Passenger Satisfaction

Project 4

Rodriguez, Fatima

White, Henry

Jyles, Abrea

DATA-PT-EAST-APRIL-041524-MTTH

Instructor: Booth, Alexander

October 3rd, 2024

Table of contents

Introduction

In today's rapidly evolving transportation industry, understanding and predicting passenger satisfaction has become crucial for improving service quality and customer experience. This project leverages machine learning techniques to analyze various factors influencing passenger satisfaction. By using this dataset, we aim to develop a predictive model that can accurately forecast passenger satisfaction levels. This model will assist transportation companies in identifying key areas for improvement, ultimately leading to enhanced service delivery and increased customer satisfaction.

To achieve this, we collected a comprehensive dataset of airline passenger feedback, including various factors such as service quality, seat comfort, in-flight entertainment, and overall satisfaction ratings. We processed the data to handle any missing values and standardized the features to ensure consistency. Using this cleaned dataset, we trained several machine learning models including logistic regression, decision tree, random forest, XGBoost and others. to predict passenger satisfaction. After evaluating the performance of these models, we selected the one with the highest accuracy and fine-tuned it parameters to optimize its predictive capabilities. The final model was then validated using a separate test set to ensure its reliability in predicting passenger satisfaction.

Our exploration included analyzing the importance of various features to understand which factors most significantly impact passenger satisfaction. We also performed cross-validation to ensure the model's performance was consistent across different subsets of the data. Additionally, we visualized the results using graphs and charts to provide a clear insight into the model's predictions and the underlying trends in passenger feedback. This comprehensive approach allowed us to build a robust model capable of accurately predicting airline passenger satisfaction.

Parallel to this, we load the dataset into Tableau to gain a deeper understanding of passenger satisfaction. By visualizing the results, we can effectively identify areas that are performing well and those that require improvement. through interactive dashboards, Our goal is to leverage these insights to implement targeted strategies that address any shortcomings and elevate service quality.

This project will be delivered through a web application deployed on PythonAnywhere, where end users can interact with the model to predict passenger satisfaction and analyze survey results. The user-friendly interface will allow users to input various parameters related to their travel experience, such as flight duration, amenities, and service quality, to generate real-time predictions of satisfaction levels.

Additionally, the application will feature visualizations that summarize key insights from the dataset, enabling users to explore trends and correlations. By incorporating interactive elements, users can filter data and view results based on different criteria, facilitating a comprehensive understanding of passenger sentiments.

Data Cleaning/Data base creation

After selecting the data set, we proceed to download the csv file and use jupyter notebooks to perform the data cleaning. The first step was to load the data into a panda's data Frame for easy manipulation. We then inspected the dataset for any missing values, outliers, or inconsistencies. Missing values were then handled by either filling them with appropriate statistics. Outliers were identified and treated to prevent them from skewing the model's performance. We also standardized and normalized numerical features to ensure they were on a comparable scale. Categorical variables were encoded using techniques such as one-hot encoding to make them suitable for machine leaning algorithms. Finally, the cleaned and preprocessed data was split into training and testing sets to facilitate model building and evaluation.

Color design considerations

For our website, we utilized the minty template, renowned for its neutral background and structured layout that effectively showcases our visualizations. To ensure clarity and visual appeal, we selected a color gradient ranging from teal to green for these elements, complementing the overarching theme of our presentation.

For the presentation and dashboard, we selected a palette featuring light tones of green, yellow, and brown. This color scheme was chosen deliberately to evoke feelings of warmth and positivity, creating an inviting atmosphere for users interacting with the data.

Comparing Model Predictions

When comparing model predictions, it's important to look at several key metrics to understand their performance. In this instance, we are comparing model predictions through their ROC curves, specifically the area under curve score. With the dataset being relativley balanced between positive and negative target outcomes, this score is a good summary of recall, precision, and f1 scores. While several models were tested, we specifically call out the logistic regression, random forest, gradient boost, and XGBoost models below:

Model	**Area Under Curve Score**
Logistic Regression	0.87
Random Forest	0.99
Gradient Boost	0.98
XGBoost	0.99

While at first glance, Random Forest appears tied for the best forest, there were clear signs of overfitting. This was not the case with the gradient boost or XGBoost models, which is why they became our top two selections. The full metrics for each are:

- Gradient Boost: Precision, Recall, and f1 score of 0.94
- XGBoost: Precision, Recall, and f1 score of 0.96

Dashboard design concepts

Using the dataset, we designed two comprehensive dashboards in Tableau Public to gain deeper insights into passenger satisfaction.

The first dashboard is dedicated to showcasing overall passenger satisfaction while highlighting key demographic factors that may influence the flight experience. We considered variables such as age, gender, loyalty program participation, type of travel, and class of service. By examining these demographics, we can identify trends and correlations that may shed light on how different groups of passengers perceive their travel experience.

The second dashboard takes a different approach, focusing on categorizing the various services provided by the airline. Our goal here is to identify which services are performing better than others, allowing us to pinpoint strengths and weaknesses in the airline's offerings. Additionally, we aimed to explore the potential impact of flight delays on passenger satisfaction.

Website architecture

Our website architecture is designed to provide a seamless and user-friendly experience. The landing page serves as the main entry point, offering an overview of our project and easy navigation to other sections. The machine learning experiment page showcases our data analysis and predictive modeling efforts, allowing users to easily navigate and explore the methodologies. We have dedicated two pages for tableau, each featuring a unique dashboard that visualizes different aspects of our data, providing insightful and interactive visual representations, lastly the about us page offers information about our team, our mission, and the goal of our model, ensuring visitors to understand the context and purpose behind our work.

Bias and Limitations

Initially, we chose XGBoost as the machine learning model for our analysis due to its renowned robustness and exceptional performance in handling complex datasets. However, as we progressed toward deployment on PythonAnywhere, we encountered several technical challenges with integrating XGBoost into the web application environment.

These issues prompted us to explore alternative options that would better suit our deployment needs. Ultimately, we replaced XGBoost with its more flexible counterpart, Gradient Boost. This transition allowed us to maintain a high level of predictive performance while addressing the deployment constraints we faced. Gradient Boost provides similar advantages to XGBoost but is designed to run with the scikit-learn python libary, making it easier to implement in various environments, including web applications.

With both XGBoost and Gradient Boost being tree-based models, some interesting phenomena were witnessed in testing different parameters with model. While common sense would lead one to assume that boosting the score in any one satisfaction category should boost the overall probability of satisfaction, there are instances where an increase to one category may slightly drop the probability of satisfaction (though not enough to change the overall result). This is a result of using a tree-based model instead of a regression-based model; given that there are over a trillion combinations of possibile satisfaction survey results, the model is not going to account for every possible combination (hence why machine learning is used). It is hypothesized that with a larger dataset, some of these counterintuitive phenomena may cease.

It is also worth noting that the source data utilized treated gender as a binary, which may not reflect the desired response of all possible respondents. Thankfully, gender's feature importance in the overall model was relatively insignificant

To enhance the efficiency of our analysis in Tableau, we created new columns derived from the existing dataset. This strategic decision was aimed at improving the clarity and depth of our visualizations, allowing us to present more insightful results. By generating these new columns, we were able to transform raw data into more meaningful metrics that could reveal underlying trends and patterns.

In our analysis, we opted to treat null values as zero, a decision that could significantly impact the overall satisfaction rate. By interpreting missing data in this way, we assumed that a lack of response or data points indicates a corresponding lack of satisfaction. However, this approach carries inherent risks and may not accurately reflect the true sentiments of passengers.

Treating null values as zero can lead to a skewed perception of satisfaction rates, as it implies that every instance of missing data equates to a negative experience. This assumption overlooks the possibility that passengers may not have provided feedback due to various reasons, such as time constraints or indifference, rather than outright dissatisfaction.

Conclusions

The analysis reveals a concerning trend, with 56.55% of passengers reporting dissatisfaction with their airline experience. This statistic is not just a number; it underscores a critical need for airlines to take immediate action in addressing customer concerns. The substantial dissatisfaction rate suggests that many passengers feel their expectations were not met during their journey, which can stem from a variety of factors, including service quality, timeliness, and overall comfort. With the overall service rating falling below 3.5 out of 5, it becomes evident that there are significant areas requiring improvement.

The study identified several key features that significantly influence passenger satisfaction, including inflight Wi-Fi service, online boarding, leg room, and seat comfort. These elements emerged as critical determinants of the overall travel experience, highlighting specific areas where airlines can make impactful improvements.

The development of a prediction model utilizing XGB, achieving an impressive 98% accuracy, underscores the remarkable effectiveness of machine learning in analyzing and understanding passenger satisfaction. This high level of accuracy indicates that the model can reliably identify patterns and relationships within the data, providing valuable insights into the factors that influence customer experiences.

Although we considered variables such as gender, customer type, age, type of travel, class, flight distance, and delays in our analysis, these factors showed no significant weight in predicting customer satisfaction. This finding is particularly noteworthy as it suggests that passenger satisfaction is more strongly influenced by service quality and operational factors rather than demographic characteristics.

Future work

- Leveraging data analytics to understand passenger preferences and behaviors can help airlines offer more personalized services. Implementing loyalty programs that reward frequent flyers with tailored benefits, such as extra legroom or priority boarding, can enhance customer satisfaction.

- Investing in advanced digital solutions can significantly enhance passenger experience. This includes improving inflight wi-fi connectivity, streamlining online boarding processes, and offering user-friendly mobile apps that provide real-time updates and personalized travel information.

- By continuously refining their prediction models with real-time data (such as flight delays, customer feedback, and service usage patterns), airlines can proactively adjust staffing, resources, and service offerings.

Works Cited

- Teejmahal20. *Airline Passenger Satisfaction Dataset*. Kaggle. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.
- XGBoost. *Python Package Introduction*. https://xgboost.readthedocs.io/en/stable/python/python_intro.html.
- Scikit-learn. *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/stable/.
- PythonAnywhere. *PythonAnywhere: Python in the Cloud*. https://www.pythonanywhere.com/.
- Tableau Public. *Tableau Public: Discover and Share Data Visualizations*. https://public.tableau.com/app/discover.