Predicting New York City Taxi Fare Prices

Introduction

Within the metropolis of New York City, navigating the intricate web of transportation can be both a daily challenge and a financial puzzle. Our project addresses a ubiquitous problem: predicting taxi fares. In a city where prices for anything are impossible to predict, we created a tool to empower people in New York City, providing insights on how to navigate the city efficiently and economically. What may seem like an end of semester project, has high stakes: savings for individuals, reduced traffic, and a positive environmental impact.

Data

Our exploration began with the New York City Taxi Fare Prediction dataset, a dataset with over 90,000 instances of taxi rides spanning approximately four years. Acquired from Google for a Kaggle competition, this dataset encapsulates the essence of New York City's taxi ecosystem. Each entry does not only hold the fare (target feature), but also provides crucial information about time of day, pickup and drop-off locations, passenger counts, and the duration of each trip. However, to truly understand the dynamics, we needed to calculate the distance traveled (which was not provided) - a task that led us to the Haversine Formula.

Methods

Since our goal is to be able to make accurate predictions about taxi fare prices in New York City we focused on using different forms of regression to find the one that will make the most accurate prediction. We embarked on this journey armed with methods chosen for their ability to unravel the intricacies of the dataset. We used SimpleImputer to ensure a clean slate, eradicating any missing values within our dataset. We additionally extracted the different time expressions so that they were able to best fit our model. While working with the data, we utilized a variety of machine learning methods to tackle the complexity of predicting taxi fares. Our methods included Linear Regression and Random Forest Regression, each chosen for its unique strengths in handling the intricacies of the dataset. Linear Regression offers simplicity and interpretability, while Random Forest Regression harnesses the power of ensemble learning to capture complex patterns.

Results

Due to the nature of the data there is a pre-existing model that those participating in the Kaggle competition are working to beat. Using the standard room mean error squared the threshold to beat was $8 difference in fare prices. While our linear regression model did not surpass the threshold, currently standing at an RMSE of 9.642, the random forest regression seemed to have a much better improvement standing with a root mean square error 5.262. While this would not be the best results for a Kaggle competition it would still be considered significant enough to merit a submission.

Learnings

Throughout our time working on this project we gained valuable insights into a variety of aspects of the machine learning pipeline. More specifically, we have gained a deeper understanding of data preprocessing, feature engineering, and model selection. One unexpected challenge was dealing with the limitations of the data and understanding possible ethical qualms with our data set. For example we do not know if some of the surges during the day were due to inaccessible public transport, weather conditions, or special events. If we were to further dissect the time feature, we would have had more accurate results and factored in the surges at a specific time of day or time of year. We ran into trouble on that front, and, though we learned about time features, we were not able to successfully incorporate it. Ultimately, this was a fantastic opportunity to keep furthering our knowledge in Machine Learning and applying what we have learned this semester. With this project in our portfolio and under our belts, we will be able to tackle more problems in the real world with new skills.

Conclusion:

Our objective with this project was to better understand the concepts we learned throughout the semester by applying them to a real life scenario. While calculating taxi fares seems simple and straightforward, it is something that many top companies today continue to work upon and improve. Through a detailed analysis of this data set we were able to create a model that presents promising results with, of course, room for improvement. Future directions for this project could involve refining the model by incorporating real-time data streams, addressing bias through advanced algorithmic approaches. We could also consider taking into account external factors such as weather and traffic patterns. As we conclude this project it is important to consider that although we did not produce the most comprehensive and accurate model for predicting taxi fares, we learned a lot about taking a data driven approach and giving our best effort in completing an end to end machine learning project.

References:

You can find our data source here:

Andy Chavez, DJ Sterling, Julia Elliott, Lakshmanan V, Sagar, Will Cukierski. (2018). New York City Taxi Fare Prediction. Kaggle. https://kaggle.com/competitions/new-york-city-taxi-fare-prediction

Assistance with understanding and implementing the Haversine formula: https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128#:~:text=All%20of%20these%20can%20be,longitude%20of%20the%20two%20points

Help with using git:

https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128#:~:text=All%20of%20these%20can%20be,longitude%20of%20the%20two%20points.

Shoutout to ChatGPT for answering all questions big and small.