Emily Hahn, Zack Horton, Mackenzie Lees, Ghazanfar Yezdan
March 10, 2024
15.S07 SSIM: Real-time Analytics for Digital Platforms

Real-Time Analytics for Rideshare App Optimization - ABSTRACT

Project Scope

We sought to utilize machine learning models to better understand the pricing strategies of Uber and Lyft in the Boston area. Our first analysis aimed to assist rideshare app users in better deciding which app to use given real-time inputs such as time of day, weather conditions, and vehicle type. The second analysis sought to assist rideshare drivers in maximizing their profits by determining which locations were the most profitable given their driving conditions. The models from analyses are connected to a dashboard through which users can assess their options.

Methods

We began by running various predictive models, including regression, neural networks, and tree-based approaches, to estimate the costs of rides for both Uber and Lyft given a fixed set of factors. We adapted the code to be used in our user-friendly dashboard. The dashboard asks users for their start location, end location, hour of day, month, weather conditions, and ride type (standard, XL, luxury, etc.). It then outputs a list of all the models' price estimates for each rideshare app and a one-sentence recommendation for which app to use. We also display more granular technical details of each model for fact-checking.

After being able to predict route prices, we utilized Q-learning to determine which routes were the most profitable given various time and weather constraints. For this model, we only looked at standard Uber rides for various hours of the day, days of the week, and weather conditions. The final dashboard shows a map of Boston with colored lines from the inputted start location to all possible end locations. If users hover their cursor over one of the lines, they can view the names of the end locations and the Q values for those trips. The optimal route is highlighted in green.

Results

Our models determined that the Financial District, Theater District, and Haymarket are generally the best places to operate. More specifically, the route from the Financial District to South Station is the best overall. Regardless of location, rides at 5-6 PM tend to be the most lucrative. This makes sense because rush hour is the heaviest. Rides at 10 PM also have high Q values on weekends, which makes sense considering nightlife popularity during these times. Drivers also benefit more from driving in the rain, which makes sense as people are less likely to walk or bike in these conditions.

Emily Hahn, Zack Horton, Mackenzie Lees, Ghazanfar Yezdan
March 14, 2024
15.S07 SSIM: Real-time Analytics for Digital Platforms

Real-Time Analytics for Rideshare App Optimization

INTRODUCTION

Ridesharing services have become integral to urban transportation, offering affordable and convenient options. Uber and Lyft are two of the leading companies in this industry and compete for users through Boston. Both companies employ complex pricing strategies that respond to demand, location, time, weather, and other factors. Given these services' significant role, a deeper understanding of their pricing mechanisms can help riders and drivers adjust their actions. However, since the exact nature of these pricing algorithms is unknown to the public, direct comparison is challenging for the average person.

Our project seeks to leverage advanced real-time analytics strategies to help rideshare users minimize expenditures and allow rideshare drivers to maximize profits. Through a user-friendly dashboard, riders can input information about their upcoming routes to estimate the projected costs for each service. With a separate application, rideshare drivers can determine which region of Boston to target for a given weather status and time to maximize expected profits. Through both of these use cases, we can better understand the factors that drive rideshare pricing strategies.

METHODS

Data Collection

We utilized a dataset from Kaggle containing Uber and Lyft rides data from November 26, 2018, to December 18, 2018. This dataset includes various attributes, with the primary focus being on the price of rides, which serves as the outcome variable we aimed to predict. To understand the dynamic pricing mechanisms of ridesharing services, we identified several factors potentially impacting the price, including weather conditions (categorized as rain, clear, snow, etc.), time of day (grouped by hour), month, and ride type (standard for up to 4 passengers, luxury for nicer cars, etc.).

In the preprocessing phase, we concentrated on refining the dataset by removing unnecessary or redundant variables that might not significantly influence the pricing of ridesharing services. For instance, the temperature was deemed too granular for the scope of our study and was therefore excluded. We applied one-hot encoding to categorical variables, such as weather conditions and ride type, to facilitate their use in our predictive models.

Predictive Modeling

To predict ridesharing prices, we employed a variety of models including regression, neural networks, and tree-based methods. We began our regression analyses using Ordinary Least Squares (OLS), Ridge, and Lasso to capture linear relationships between the factors and the price. We then implemented deep and feedforward neural networks to model complex, non-linear relationships in the data. Lastly, we explored decision trees, XGBoost, and Random Forest models because of their ability to handle non-linear data and provide robust predictions. Each model was trained to estimate prices for both Uber and Lyft rides.

We then developed a dashboard using Streamlit to serve as a practical interface for users to estimate rideshare costs. Users can input their pickup and dropoff locations, month, time, weather conditions, and ride type. The dashboard automatically filters out non-applicable options (e.g., routes with no available data) and presents a list of estimated prices across different models, offering a range of probable costs. The outcomes are then presented in a user-friendly format, sorting the estimated prices from highest to lowest to give users a comprehensive view of potential costs across different services. Additionally, the dashboard provides access to underlying data for transparency and fact-checking purposes.

Q-Learning

To assist rideshare drivers in maximizing profits, we implemented a Q-Learning algorithm to determine optimal driving locations in Boston based on weather conditions and time. Although our analysis primarily utilized Uber data, due to its abundance compared to Lyft, we posited that the findings would also apply to Lyft, given the similarities in rideshare pricing patterns.

Using expected prices from source to destination that were found in the predictive modeling phase, a Q-Learning model was trained to explore various routes and determine the rewards for

each locality in Boston. Rewards were normalized to a per mile basis to allow for like-to-like comparison. These results were compiled in a user-friendly dashboard that allows drivers to input various combinations of day of week, time of day, and weather (clear vs rainy) to determine which driving routes are optimal under their current conditions. The output is an interactive map showing all possible routes and their Q values. This approach allows us to provide drivers with actionable insights into where and when to drive under varying conditions to maximize earnings.

RESULTS

Our prediction models found that the prices for both apps are very similar. For example, when riders are leaving Back Bay to one of the 6 most popular locations on Friday around 5:00 PM, Uber is on average $0.56 more expensive compared to Lyft. However, when riders are leaving the Financial District in a similar situation, Uber is only $0.36 more expensive on average. Therefore, while there isn't a large difference in prices on the surface when you think about taking Uber or Lyft rides over the entire year, maybe even over multiple years, it adds up. Our tool offers real-time support for making cost-effective decisions across the entire Boston area. All of the models we utilized had high accuracies, showing that the pricing strategies of but companies are relatively learnable. Our model would likely benefit from more data, as our dataset's trips only span about 2 months. Furthermore, by looking at data from a wider span of time, we may be able to pick up on the impacts of seasonality on pricing.

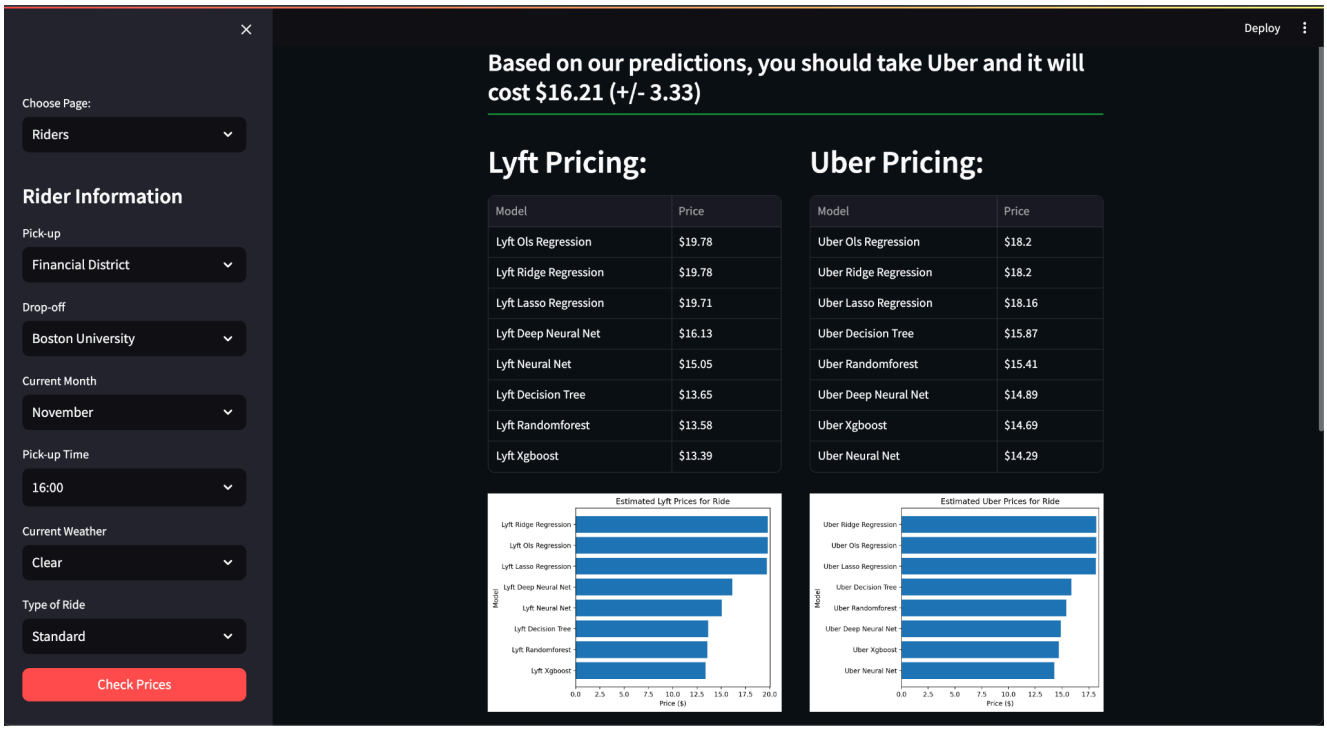
The Q-learning model identified the prime driver locations to be the Financial District, North/South End, and Haymarket regions. This might be due to the presence of offices leading to less price sensitive, professional commuters. Analyzing the Q-Values further revealed that optimal driving times are between 5PM and 7PM as well as 10PM on weekends. These times correspond to the end of the workday and party-hours. Lastly, as expected, Q-values (rewards for drivers) are higher for rainy days than they are on clear days.

CONCLUSION

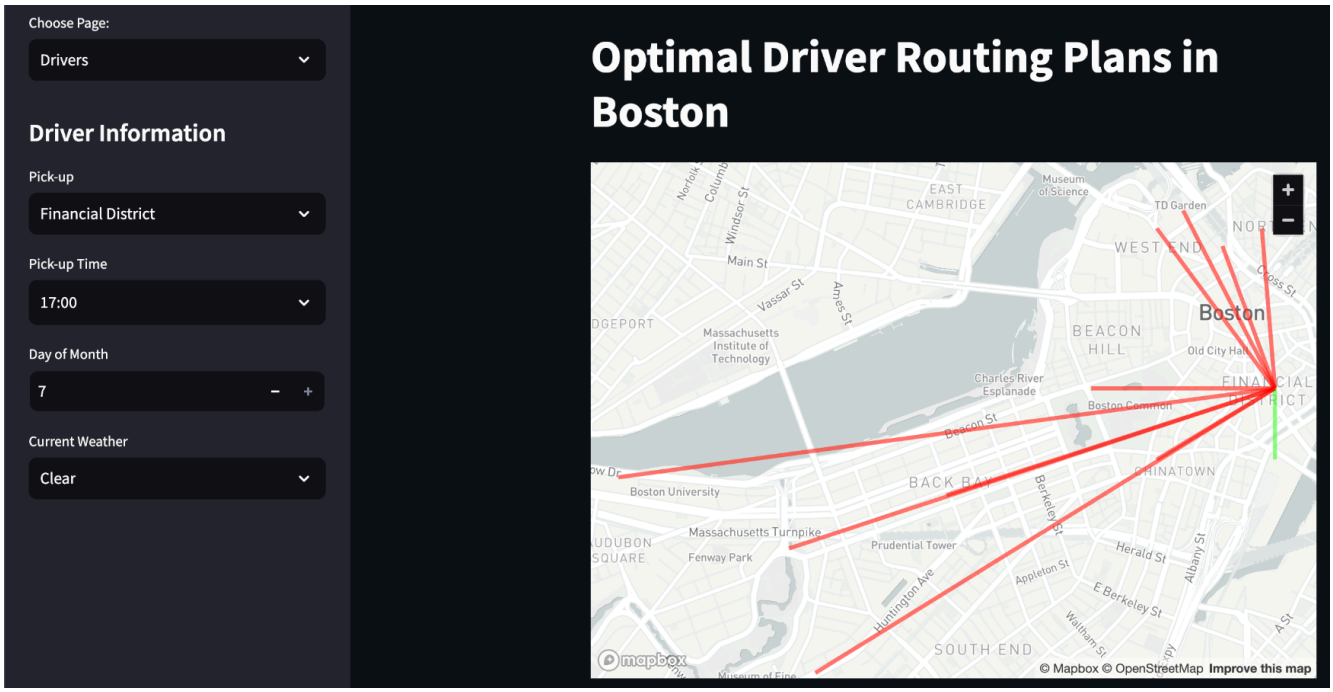
Real-time data analytics presents a complex challenge from which we can obtain valuable insights. We explored such potential within the context of dynamic pricing strategies employed by ridesharing services like Uber and Lyft. Through our study, we have demonstrated the

feasibility of leveraging advanced analytical methods to benefit both riders and drivers. Our project not only offers a practical tool for predicting rideshare costs but also aids drivers in identifying optimal routes for maximizing their earnings based on real-time conditions. By integrating data-driven models (namely regression, trees, neural networks, and Q-Learning) with user-friendly interfaces, we can bridge the gap between complex data analytics and everyday decision-making. This project underlines the potential of data science to enhance operational efficiency and customer satisfaction in the ridesharing industry and beyond. As technology and data availability evolve, further advancements in predictive analytics will continue to shape the future of urban mobility, making it more accessible, efficient, and responsive to the needs of its users.

APPENDIX A: Price Comparison Dashboard for Riders



APPENDIX B: Optimal Route Dashboard for Drivers



APPENDIX C: Q-Learning Reward Estimations by Location