CAPSTONE PROJECT - DATA 69099

Title: Time-Based Patterns of Taxi Activity and Their Impact on Trip Durations in Peak and Off-Peak Hours

Kollu Sravan Kumar Reddy - 811295843 - skollu@kent.edu

Keerthi Akhila Pasam – 811304142 - kpasam@kent.edu

Archana Katta – 811298298 – akatta2@kent.edu

05/08/2024

1. An exact description of the research question or problem

How do temporal patterns of taxi pickups and drop-offs correlate with trip duration variations between peak and off-peak hours, considering the fare amount charged for the trips?

2. A discussion about why the solution to the problem would be valuable

It is important to comprehend the temporal patterns of taxi activity and how they affect the length of trips during peak and off-peak times in order to optimize resource allocation, improve service quality, and guide transportation policy. Authorities may more effectively deploy resources, lowering traffic and wait times, by pinpointing the hours when taxi demand is at its highest. Furthermore, by understanding the relationship between trip durations and taxi activity patterns, service providers can enhance client happiness and service reliability. Urban planners can use this information to guide future development efforts by learning how effective the current transportation regulations and infrastructure are. Furthermore, by comprehending how passenger behavior relates to taxi activity, tailored initiatives to support off peak travel and environmentally friendly transportation methods can be implemented. All things considered, examining time-based taxi activity patterns aids in evidence-based decision-making for more effective.

3. A discussion about previous attempts to solve the problem and what you learned from them

In previous attempts to solve the problem of predicting fare amounts in ride-sharing services, simpler linear regression models were often employed. These models struggled to capture the complexities of the data, resulting in poor predictive performance. However, through experimentation, it became evident that incorporating additional features beyond just distance and time of pickup could significantly enhance model accuracy. By including features such as peak hours, passenger count, and specific pickup and drop-off locations, the model gained a more nuanced understanding of the factors influencing fare amounts. Moreover, attempts to address overfitting by employing regularization techniques, such as Ridge Regression, were somewhat successful but still left room for improvement in predictive accuracy.

Random Forest Regression emerged as a promising alternative to traditional linear models. Its ability to capture non-linear relationships and interactions among features allowed for more accurate predictions. By leveraging the ensemble of decision trees, Random Forest Regression was better equipped to handle the complexities of the data, resulting in a substantial improvement in predictive performance compared to linear models. However, despite its advantages, Random Forest Regression introduced new challenges, such as the need for careful hyperparameter tuning to optimize model performance. Overall, these previous attempts underscored the importance of feature engineering, model selection, and tuning in improving the accuracy of fare prediction models in ride-sharing services.4. A discussion about the methods your team used to solve the problem.

6. A discussion about the methods your team used to solve the problem.

Our team used a structured strategy combining different techniques to address the issue of estimating fare prices in ride-sharing platforms. At first, a thorough investigation of the data was performed to comprehend the dataset's features and address any discrepancies or absent data efficiently. Feature selection was crucial in our role, as we chose important features like pickup time, distance traveled, peak hour indicators, and number of passengers to capture key factors affecting fare prices. This process required both domain expertise and experimentation to pinpoint the most valuable attributes for training the model.

We tested various regression algorithms, including linear models such as Ridge Regression and more sophisticated ensemble methods like Random Forest Regression and Gradient Boosting Regression, to create predictive models. Every model was trained using suitable training and validation methods, such as cross-validation, to guarantee durability and broad applicability. Adjusting hyperparameters was used to enhance model effectiveness, finding a middle ground between bias and variance. Additionally, we employed standardization methods to rescale characteristics, guaranteeing that all models were trained on standardized data for equitable comparison and enhanced convergence.

During the process, a thorough assessment was carried out using different performance metrics like R-squared, mean squared error, root mean squared error, and mean absolute error. This enabled us to impartially evaluate the forecasting precision and adaptability of each model. Furthermore, we performed in-depth examinations of model inaccuracies and remnants to detect any trends or consistent prejudices, allowing for continuous improvement of the models.

In general, our team utilized a mix of data preprocessing, feature engineering, model selection, training, validation, and evaluation, with guidance from both domain knowledge and empirical experimentation. Through this iterative process, we were able to continually improve our methodologies and create predictive models that accurately predict fare amounts in ride-sharing services.

5. A discussion of the results your team obtained from your work.

The results from our work with different regression models, including Ridge Regression, Random Forest Regression, and Gradient Boosting Regression, provide valuable insights into predicting fare amounts in ride-sharing services. Each model exhibited varying degrees of performance, with Gradient Boosting Regression demonstrating the highest accuracy among the three.

Gradient Boosting Regression achieved an R-squared value of 0.407 on the test set, indicating that it explains around 40.75% of the variance in fare amounts. This model also yielded the lowest mean squared error (158.55) and mean absolute error (3.41), indicating its ability to predict fare amounts with relatively high precision. These results highlight the effectiveness of Gradient Boosting Regression in capturing complex relationships within the data and making accurate predictions.

Comparatively, Random Forest Regression also performed well, with an R-squared value of 0.378 and slightly higher errors than Gradient Boosting Regression. However, it still outperformed traditional linear models such as Ridge Regression, which exhibited significantly lower R-squared values and higher errors. Overall, these findings underscore the importance of employing advanced regression techniques like Gradient Boosting Regression in accurately predicting fare amounts in ride-sharing services, ultimately enhancing user experience and operational efficiency**.**

6. A discussion about how we can determine if you have successfully solved the problem.

Evaluating the effectiveness of resolving the issue of forecasting fare prices in ride-sharing platforms can be analyzed from different perspectives. The primary metric is the predictive performance of the developed models. This includes assessing metrics like R-squared, mean squared error, root mean squared error, and mean absolute error on both validation and test sets. An effective outcome would show strong R-squared values, reflecting a strong alignment with the data, and at the same time reduce error metrics, indicating precise forecasts. Consistently achieving low error rates with various evaluation metrics would suggest that the model is accurately predicting fare amounts, thus fulfilling the main goal.

Furthermore, the solution's practical implications must be considered. Determining success involves evaluating if the created models are used in real-life situations and produce concrete advantages. If the estimated fare closely matches the actual charges in real ride-sharing transactions, it shows that the model is successful in its use. In addition, keeping track of important metrics like customer satisfaction, revenue generation, and operational efficiency can offer valuable insights into the effectiveness of the solution. If the models being used result in better customer experiences, higher revenue, and more efficient resource allocation in ride-sharing services, then it indicates that the issue has been effectively resolved. In the end, a successful solution is one that not only shows excellent predictive performance but also results in real advantages for service providers and users.

7. A discussion about how you might deploy the solution in the real world to create value for someone.

Several stakeholders in the transportation sector may benefit from the model's practical implementation. The model's insights might help taxi service operators make operational decisions, like fleet deployment optimization based on expected demand patterns during peak and off-peak hours. Service providers can improve overall service quality, reduce passenger wait times, and increase driver efficiency by carefully assigning vehicles to high-demand regions and times. Moreover, by offering more precise anticipated journey durations and fare estimates based on current demand and traffic circumstances, integrating the model into taxi-hailing services might enhance user experience. This can lessen the uncertainty around taxi rides and assist clients in making well-informed judgments regarding their travel options. The model's conclusions can be used by legislators and urban planners to guide transportation infrastructure.