# Crunching the Numbers: Identifying Top Opportunities for a Taxi App Upfront Pricing Precision Enhancement

**Prakhar Srivastava**

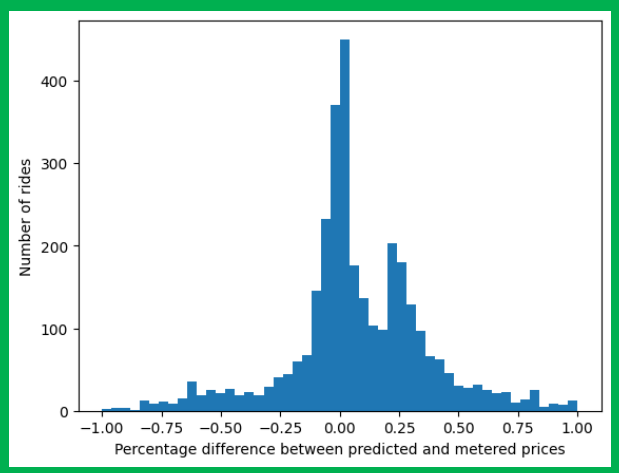
**"Data is the fuel that powers our decisions, and as a data analyst, my goal is to ignite that fuel and propel us forward towards informed insights and smarter choices."**

**Introduction**

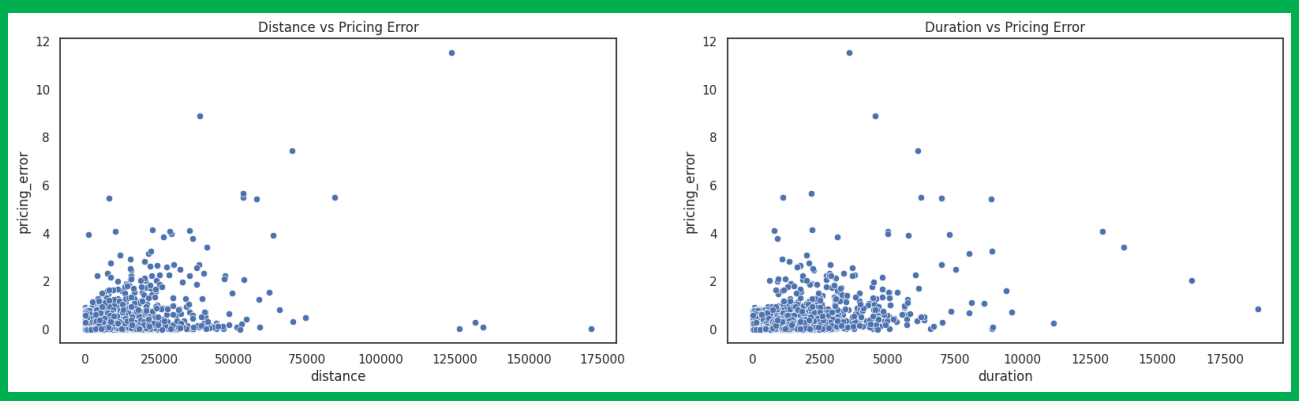
As a Data Analyst, I have been tasked with identifying the top opportunities for TaxiApp to improve its upfront pricing precision. This report aims to provide actionable insights based on the analysis of the provided dataset. In order to achieve this goal, I have examined the dataset to identify any patterns, trends, or anomalies that could impact the accuracy of the upfront pricing predictions.
TaxiApp's upfront pricing system plays a critical role in the ride-hailing experience for its customers. If TaxiApp's predictions are consistently off, it can lead to revenue loss and increased customer churn. Therefore, it is essential to ensure that the prices predicted before the ride are as close as possible to the actual metered prices. This not only helps to build trust with customers but also ensures that customers do not encounter any surprises or unpleasant experiences during their journey. Therefore, improving the upfront pricing precision is crucial for enhancing the overall customer experience and driving customer loyalty.
In this report, I present the findings of my analysis and provide recommendations for TaxiApp to improve its upfront pricing precision. My insights are based on a detailed exploration of the dataset, and I have used statistical techniques to identify potential opportunities for improvement. I believe that the recommendations presented in this report will help TaxiApp to enhance the accuracy of its upfront pricing predictions and provide a better experience to its customers, which can ultimately lead to increased revenue and market share.

**EDA**

I follow the CRISP-DM(Cross-Industry Standard Process for Data Mining) methodology for this project, I first undertook its data preprocessing stage that is EDA(Exploratory Data Analysis) that involved handling missing values by dropping NaN-value columns and replacing the NaN values within rows with zeros. This step streamlined the dataset's dimensions and facilitated better data quality by reducing the noise within the data. Subsequently, I visualized the percentage distribution histogram to demonstrate the difference between the predicted and metered prices.

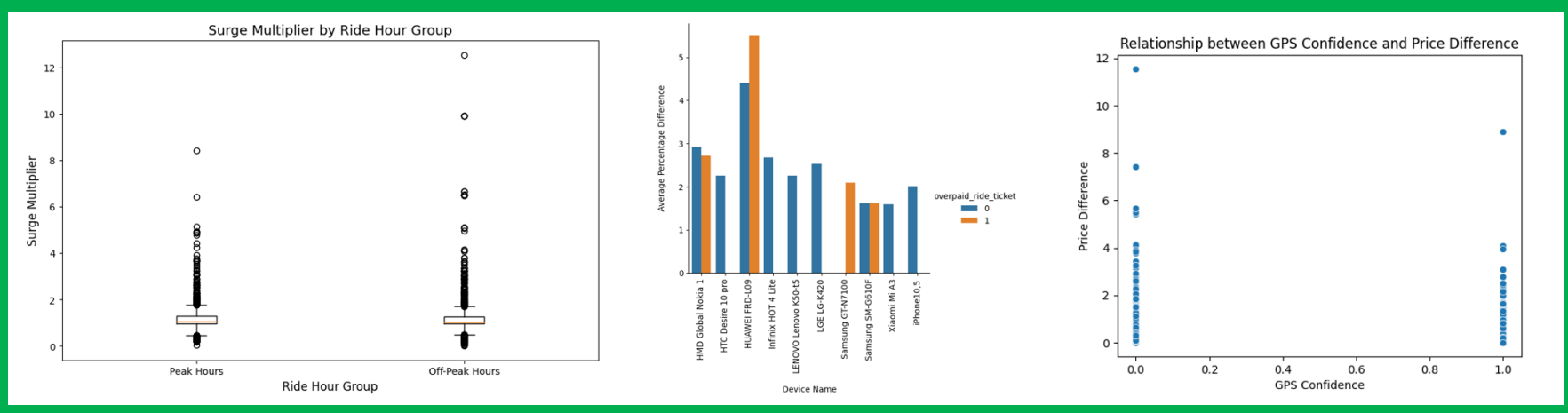



Further, I applied a percentage-based error metric to quantify the percentage of rides in which the predicted and metered prices differed by more than **20%**. My findings indicated that **62.25%** of the rides had significant differences between predicted and metered prices, which suggests that the current algorithm's predictive ability requires refinement.




As I examined the scatter plot for distance versus pricing error, I noticed a concentration of points at the lower end of the distance scale. This suggests that shorter trips tend to have less pricing error. As the distance increases, the scatter becomes more spread out, indicating that there may be more variability in pricing errors for longer trips. Similarly, when analyzing the scatter plot for duration versus pricing error, I observed a similar pattern, with a concentration of points at the lower end of the duration scale. This suggests that shorter trips also tend to have less pricing errors. As the duration increases, the scatter becomes more spread out, which could imply that there may be more variability in pricing errors for longer trips.

My analysis suggests that the factors that are affecting the upfront pricing are: **Surge pricing algorithm**, **GPS confidence**, and the problem with a particular **Device** model.



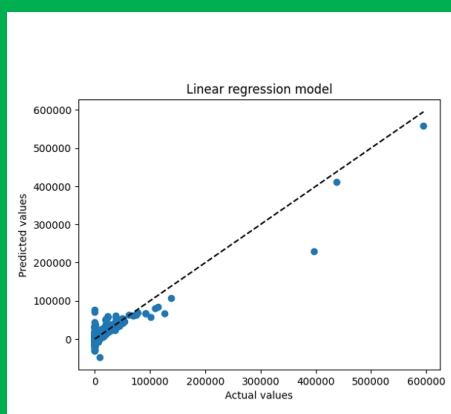

Based on my box plot analysis, the surge multiplier during off-peak hours has a higher mean of 1.196 and a higher maximum value of 12.547, compared to the peak hours group, which has a mean of 1.175 and a maximum value of 8.433. However, it's important to note that there were also some extreme outliers in both groups. These outliers indicate

anomalies or errors in the data. This unexpected finding suggests that TaxiApp may need to reevaluate its surge pricing algorithm during off-peak hours to ensure that it is accurately reflecting the demand for rides and providing customers with fair and transparent pricing.If surge pricing occurs more frequently during off-peak hours, it may have a significant impact on the precision of upfront pricing, potentially leading to customer dissatisfaction. By reviewing and adjusting the surge pricing algorithm and providing more transparency to customers about surge pricing, TaxiApp can improve the accuracy and fairness of upfront pricing and ultimately enhance customer loyalty and satisfaction. Furthermore, I have identified a major discrepancy with Huawei devices whereby there appears to be a greater variance between the predicted and metered prices compared to other devices. This has resulted in an increase in the number of complaints from customers regarding overpayment. Also, the cluster for gps_confidence of 1 appears to have a lower average price difference compared to the one for gps_confidence of 0. This indicates that rides with lower GPS confidence tend to have less accurate upfront pricing estimates.

To support my above analysis findings, I investigate the relationships between the various features in the dataset and the pricing error using a correlation matrix. By identifying which features are correlated with the pricing error, I gained insight into which factors may be contributing to inaccurate upfront pricing. This information can help to improve the overall accuracy of the pricing model and improve customer satisfaction.

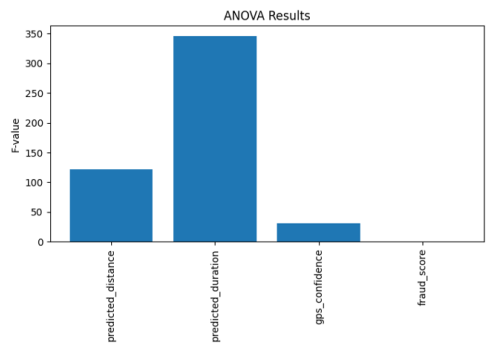

Above figure shows that, the metered price has a strong positive correlation with the price difference percentage, which indicates that the metered price has a significant impact on the final price charged to the customer. Similarly, distance and duration also have a moderate positive correlation with the price difference percentage, suggesting that longer trips tend to have higher pricing errors. In contrast, GPS confidence has a moderate negative correlation with the price difference percentage, indicating that higher GPS confidence may result in lower pricing errors. The fraud score and predicted distance have weak positive correlations, indicating that they may have an impact on the pricing errors.

However, this correlation matrix only measures the linear relationship between two variables, but it doesn't tell how one variable affects the other. This is why I applied **Statistical Linear regression** which helped me to build a model that shows how the predictor variables relate to the target variable and how much impact each predictor variable has on the target variable.




Linear regression model result

| Variable | Coefficient | Standard Error | t-value | p-value |
|---|---|---|---|---|
| Constant | -2665.737 | 499.581 | -5.336 | 1.013e-07 |
| price_diff_pct | 1849.019 | 217.694 | 8.494 | 2.952e-17 |
| metered_price | 0.644 | 0.012 | 55.810 | 0.000 |
| distance | -1.290 | 0.018 | -72.338 | 0.000 |
| duration | 0.139 | 0.121 | 1.145 | 0.252 |
| overpaid_ride_ticket | 1602.984 | 630.315 | 2.543 | 1.103e-02 |
| predicted_duration | 1.718 | 0.380 | 4.516 | 6.510e-06 |
| predicted_distance | 1.498 | 0.027 | 54.921 | 0.000 |
| gps_confidence | 315.291 | 392.518 | 0.803 | 4.219e-01 |
| ride_hour | -19.365 | 16.616 | -1.165 | 2.439e-01 |
| fraud_score | 0.081 | 0.115 | 0.699 | 4.845e-01 |
| R-squared Value | | | | 0.866 |




Based on the results of the Linear regression model, I can see that the variables predicted_distance, predicted_duration, and gps_confidence are all statistically significant predictors of the target variable upfront_price. This is indicated by their very low p-values and high t-values. Specifically, I can see that an increase in predicted_distance and predicted_duration leads to an increase in upfront_price, while an increase in gps_confidence leads to a decrease in upfront_price. Additionally, the R-squared value of 0.866 indicates that the model explains a high amount of the variance in the target variable.

After that, I applied the ANOVA approach to understand the significance of the predictor variables (predicted_distance, predicted_duration, gps_confidence, and fraud_score) in explaining the variability in the upfront_price. The ANOVA results helped me to identify which variables were significant and which were not and to assess the overall fit of the linear regression model. Based on the ANOVA results shown above, I can see that the variables 'predicted distance', 'predicted duration', and 'gps confidence' are all statistically significant predictors of the price difference between the upfront and metered prices.

Based on my analysis, I would recommend the following **opportunities** to help TaxiApp improve its upfront pricing precision:

1. **Evaluate and refine surge pricing algorithm during off-peak hours**: Based on the data, it appears that surge pricing may occur more frequently during off-peak hours, which can lead to inaccuracies in upfront pricing and customer dissatisfaction. TaxiApp could evaluate and refine its surge pricing algorithm during off-peak hours to ensure that it accurately reflects demand and provides fair and transparent pricing to customers. This could include adjusting the surge pricing multiplier or implementing different surge pricing rules during off-peak hours.
2. **Improve GPS data accuracy**: The confidence of GPS data, specifically the predicted distance, and duration, has a direct impact on upfront pricing accuracy. TaxiApp should explore ways to improve the accuracy of GPS data, such as investing in better GPS technology or leveraging other data sources to validate GPS data. By improving GPS data accuracy, TaxiApp can provide customers with more precise upfront pricing, which can lead to increased customer satisfaction and loyalty.

Here are some potential risks and solutions associated with these opportunities:

1. **Risk of reducing driver incentives**: Adjusting the surge pricing algorithm during off-peak hours could potentially reduce driver incentives if they are not compensated appropriately during these times. This risk can be mitigated by offering bonuses or other incentives to drivers who are available during these times.
2. **Risk of increased operating costs**: Improving GPS data accuracy may require investing in new technology or additional data sources, which could increase TaxiApp's operating costs. This risk can be mitigated by partnering with third-party providers or leveraging existing data sources within the TaxiApp company.

By prioritizing opportunities based on their feasibility and potential impact, TaxiApp can develop a strategy to improve upfront pricing precision while maximizing ROI(Return on Investment).

**References:**


1. https://medium.com/bolt-labs/fair-pricing-for-reliable-rides-6a6442a267b0
2. https://medium.com/@boltapp/introducing-upfront-pricing-ride-with-greater-confidence-75216362802c
3. https://blog.bolt.eu/en/debunked-the-three-most-popular-ride-hailing-myths/