YourDrive

Changing the Game: Providing better data to ride-share drivers

December 4, 2021

Introduction and Motivation

Ride-sharing companies such as Uber, Lyft and Didi are reshaping the modern transportation industry. In 2018, three million drivers operated with Uber, serving 75 million riders in over 600 cities worldwide [1]. An increasing concern is the tension between drivers and companies due to their differing objectives [2]. Furthermore, current research and applications focus on minimizing system-wide costs, but there is a lack of data insights available to the drivers [3].

Problem definition

Today, ride-share drivers are provided limited information about their upcoming drives, which hinders their ability to develop personalized shift strategies. Without laws to regulate operation and flow of information, ride-share giants have complete authority in assigning drivers to customers and little information is shared regarding their data and matching algorithms [4]. According to Muller,

> "Uber leads drivers to believe that the interests of riders, drivers, and the firm are aligned, when in fact they are divergent and often opposed. Uber enjoys total control of its algorithm and faces strong incentives to design it in such a way as to maximize its own growth and earnings at drivers' expense." [5]

Although there are some studies on drivers' estimated time of travel, pricing strategies, enhancing matching algorithms to reduce income inequality, etc., an interactive tool empowering drivers to make data-based decisions to maximize their income has not been explored. Based on our literature survey, our project will be the first.

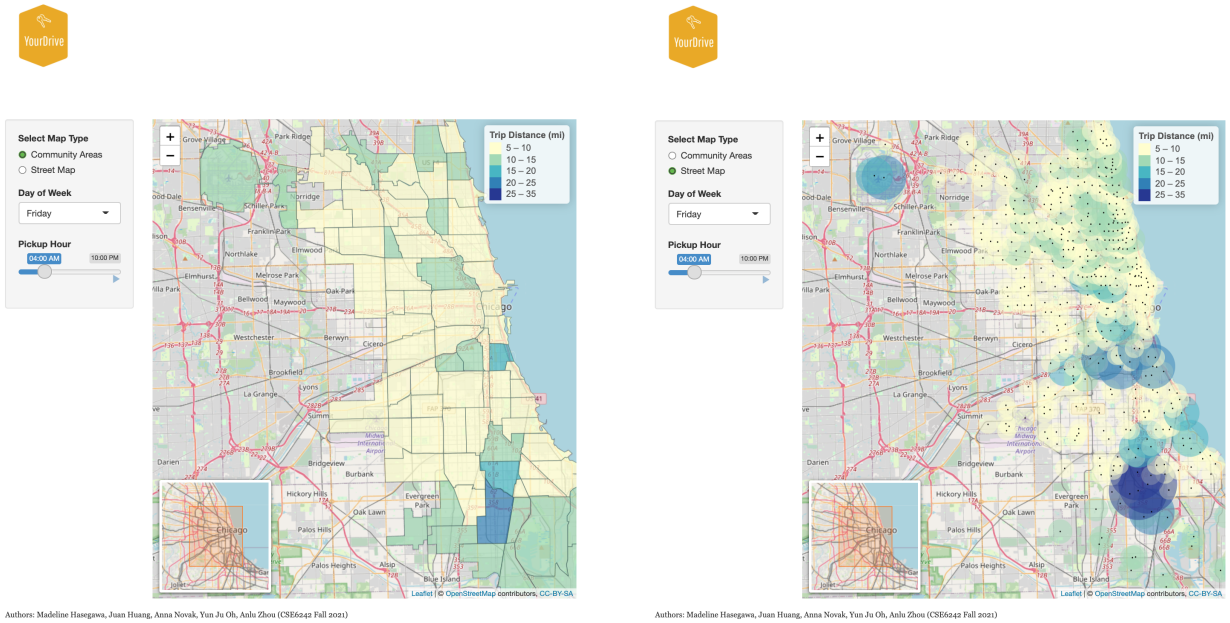
Literature Survey

Most drivers are financially motivated [6], and drivers in both developed and developing countries are usually from economically disadvantaged groups [7]. Ride-share companies developed expensive and complex algorithms to increase profits and drive market growth, not to maximize the income of their drivers. There is some evidence that data is intentionally withheld from drivers until customers are on board to prevent drivers from making strategic ride cancellations [8]. Currently, drivers can only use company-generated algorithms which assign drivers to customers based on location [4]. These algorithms are known to direct drivers to areas of high customer density via density heat maps. Uber also introduced surge pricing, which increases ride prices in times and locations where there is high demand [9]. While this helps supply drivers to meet customer demand, studies show that blindly chasing surge prices does not yield significant or consistent increases in driver income. Instead, it introduces opportunity costs and may waste drivers' time [10]. Matching algorithms also result in significant inequalities in driver income between drivers with comparable qualities and inconsistencies in day-to-day income, which results in poor economic outcomes over time [4].

For individual drivers, optimizing technologies, surge pricing and profitable trips, flexible working hours, and customer service and relationships can all factor into their income [11]. Drivers can increase their income if they strategically select areas to frequent at specific times of the day. Compared to a naive strategy, this type of driving has been found to increase driver profits by up to 50% [10]. Since Uber drivers work substantially fewer hours than taxi-drivers [12], data insights can help ride-share drivers make decisions on when and where to start their shift to increase income during their limited time.

Proposed Method

"YourDrive", is an interactive visualization that aids drivers in creating a custom shift strategy to meet their highest income potential. The tool allows the driver to open the application on any browser and select between two (2) map types: a map sectioned by community area and another with a detailed street view. The landing page shows the "Community Area" map with current day and hour data insights. The driver can easily distinguish by an intuitive color gradient which pickup areas of the city are likely to yield ride requests with the longest and shortest trip distances in miles. Drivers can also see the average predicted trip distances and top three (3) probable drop-off areas for each community area by selecting the day of week and pickup hour and clicking on the area shapes. The street map illustrates, per day and hour, the historical customer density and relative trip distances using marker sizing. By clicking on each location marker, the driver can view the predicted trip distances at historical pickup locations. Each map has an animation feature that allows the user to play through the changes in a day's predicted trip distances over pickup hours for the entire city. All data insights are provided in Chicago time and visualized using **R** (Leaflet and Shiny packages).



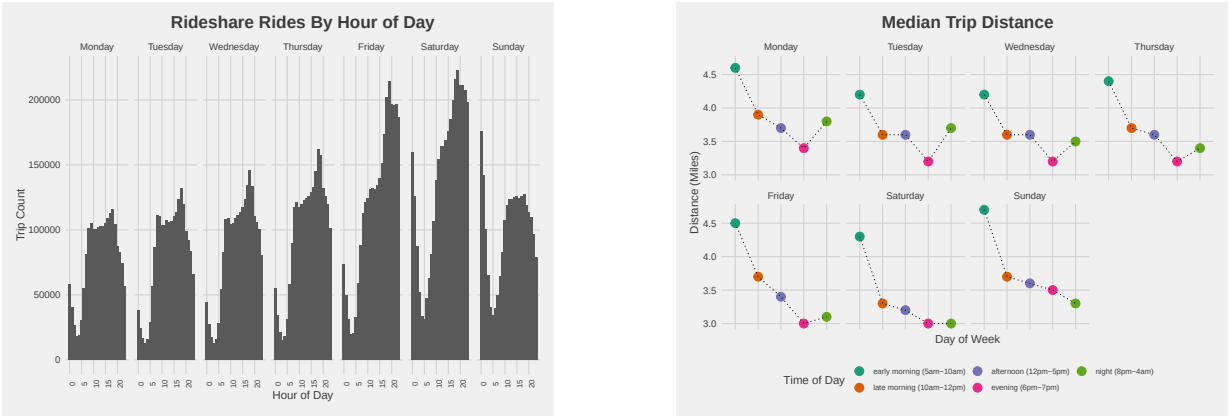
(a) Community Areas Map

(b) Street Map

Figure 1: YourDrive Application

Using a five month subset of the 211 million ride-share trip records provided by the City of Chicago, we pre-processed the data to remove all data points with missing information and narrowed down the attributes to the ones needed for the following analyses: predicting drop-off location, predicting trip distances, and predicting tip probability, all based off of pick-up location. In addition, we explored the pre-processed dataset to check the relationships between variables. Figure 1(a) below

displays the number of trips taken per hour of the day across the days of the week. We can see that Fridays and Saturdays tend to have more trips and each day sees a peak of rides early in the afternoon. We also see that on Friday and Saturday, the number of trips increases more gradually, compared to the other days, which show a more sudden increase around early evening. This may be explained by the usual commute times for work in a city. Figure 1(b) shows median trip distance for each day of the week. We can see that for every day, the early mornings tend to have longer trips with decreasing trip lengths throughout the day and a slight increase at night on the weekdays. It is interesting to see that although the number of trips increases throughout the day, the median trip distance trends downward.



(a) Rides by Hour of Day

(b) Median Trip Distance

Figure 2: Data exploration

Compared to more common analyses on market share and customer wait times, we studied the data from the drivers' perspective. By viewing the dataset from different angles and consolidating the insights into one comprehensive tool, we offer key features that are valuable to any driver aiming to proactively increase their income.

Prong 1: Probable Drop-off Location

The first prong of YourDrive provides probable drop-off locations for drivers' upcoming rides. If drivers know where their next trip may be taking them, they may be willing to accept additional trips, increasing their income. Today, the direction of the trip is not known by the driver until they are in transit with a customer. We computed the top three (3) probable drop-off areas for each Chicago community area based on historical pickup and drop-off location data, the day of the week and time of the day to account for variation in ride-share use throughout and within each day [13]. We preformed a directed graph analysis on the historical utilization to identify drop-off areas with the highest in-degrees and therefore highest probability of a drop-off. The probable drop-off areas are included in the Community Areas map as a click event popup to each area, for each day and hour. YourDrive users can use this information to gain insight on where their next trip is most likely headed.

Prong 2: Predicting Trip Distances

The second prong of our platform provides predicted trip distances to complement the drop-off locations so drivers can plan the number of trips to fit in one shift. For example, for a high-volume shift strategy, the driver will likely want to operate at the optimal days, time and geographic areas known to have shorter trips. We predicted trip distances using regression, with predictors being the day of the week, time of the day, and pickup area, and the response being trip miles. We were

mindful to only include predictors that would be known to a ride-share diver prior to picking up a passenger.

$$Gini = 1 - \sum_{j=1}^{c} p_j^2$$

Each predictor is categorical, therefore requiring translation to indicator variables. One-hot encoding yielded over one hundred attributes. We used regularization techniques to limit the variables but still used random forest regression since tree-based models handle highly dimensional datasets more appropriately, and its default split criteria is Gini. We then used cross-validation for tuning the following hyperparameters: max tree depth, minimum branch split, and minimum leaf size and used Root Mean Square Error (RMSE) and R-squared values as our hyperparameter scoring methods. The predicted trip distances are visualized with an intuitive color gradient (lighter indicating shorter distances and darker indicating longer distances) on the Community Areas map and marker sizing (smaller indicating shorter and larger indicating longer distances) and the same color gradient in the Street Map. These methods of visualizing distances enable the user to quickly interpret which areas of the city are more likely to have ride requests with shorter versus longer distances.

Prong 3: Predicting Tip Probabilities

Finally, YourDrive had planned to include a prediction of whether riders picked up in the driver's current location are likely to tip, since in previous studies, tips have been somewhat neglected. After data discovery, features engineering, and model evaluations, we created a logistic regression model to predict "tip pay" (0-negative, "no"; 1-positive, "yes") using day of the week, time of the day, pickup area, and drop-off area as predictors. Again, one-hot encoding was used to handle the categorical predictors. We used the accuracy as our scoring method during model Ridge regularization. Unexpected results arose from our modeling, so the tip probability predictions were not included in the final visualization.

$$\hat{\theta} = \arg\min_{\theta} -\frac{1}{n} \sum \left(y_i \phi(x_i)^T \theta + \log(\sigma(-\phi(x_i)^T \theta)) \right)$$

Experiments and Evaluation

There were minor challenges to overcome in Prong 2, the trip distance prediction model. When comparing the In-sample and Out-sample RMSE of the basic random forest regression, this model overfit the data. As expected, properly tuning hyperparameters fixed our overfitting problem (see Table 1).

Table 1: Trip Distance Model Performance

Model	**In-sample RMSE**	**Out-sample RMSE**
Basic Random Forest	0.8704	3.9516
Best Random Forest	2.2216	2.3629

Modeling for Prong 3 was more challenging, however. During our initial investigation we noted that the original dataset is highly imbalanced with respect to tip classification. That is, the dataset includes a much higher proportion of trips without tips than with tips (approximately 90% negative, 10% positive for tips). Although we used the weighting method to improve model performance, this imbalance suggested that we need to be cautious when evaluating the success

of our model. If we only care about the model's prediction accuracy, a naïve model could likely yield zero correctly classified positives. The model's Receiver Operating Characteristic (ROC) would likely suggest performance close to random guessing. Instead, we considered the quality and completeness of our prediction via precision (true positive rate) and recall (sensitivity):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

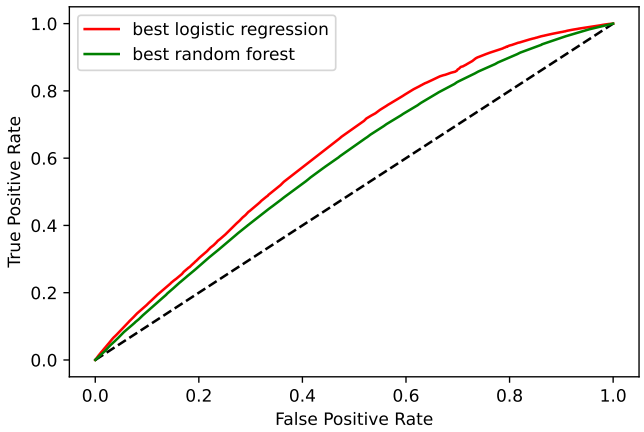


Figure 3: Tip Probability Model ROC

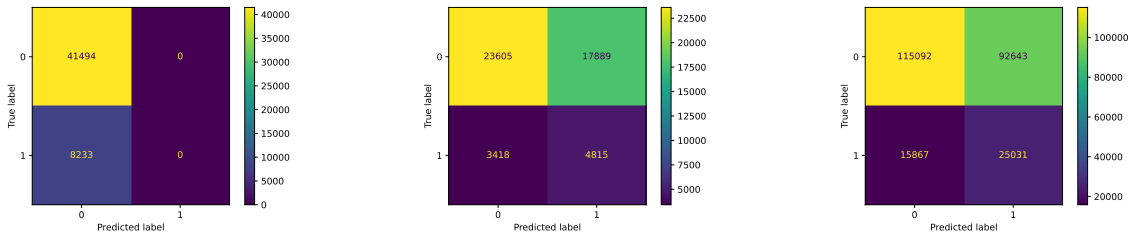
We experimented with several classification models for tip prediction, including a support vector machine (SVM), a random forest, and a logistic regression with L2 regularization. As shown in Figure 3, we found that logistic regression outperformed random forest. Logistic regression was ultimately selected because it yielded the highest F1-score and had the highest accuracy. These models' Receiver Operating Characteristic (ROC) are shown as below.

When using a re-weighting method with parameter class_weights = 'balanced' to deal with imbalanced classification, the model automatically assigns the class weights inversely proportional to their respective frequencies which increased the predicting power of the small-group class and improved the logistic regression's prediction.

$$w_j = \frac{\text{n\_samples}}{\text{n\_classes} * \text{n\_samples}_j}$$

Additionally, accuracy improved after hyperparameter tuning as shown in Figure 4. Unfortunately, the highly imbalanced nature of tips data in the historical dataset still yielded inaccurate predictions, even after re-balancing techniques. Our best precision and recall values achieved were 0.6121 and 0.2127, respectively (F1 score of 0.3157). These were not deemed to be of high enough quality to include in YourDrive, as the false positive rate was too high.

It is not unexpected that tip prediction models did not yield a successful outcome. Our research suggests that we do not have enough information to appropriately predict the likelihood of a future tip. For example, some customers may tip by cash which would not be recorded on the platform, resulting in missing data. Previous studies also suggest that there are many factors that influence tipping, including customer-driver interactions and driver gender and age [14]. While we were able to build a prediction model for this prong, we decided to omit results from YourDrive



(a) Vanilla Logistic Regression (b) Re-weight Logistic Regression (c) Best Logistic Regression

Figure 4: Tip Probability Model - Confusion Matrix of Validation Data

because the impact of false positive tip predictions on drivers could be frustrating and mislead their shift strategy.

Conclusions and Discussion

Before discussing the results of our modeling, there are some societal considerations necessary when utilizing YourDrive. Providing more information to drivers may yield an increase in wait times for riders looking for rides on less profitable routes [5]. This may lead to riders of a protected class being systemically under-served [15]. More broadly, our impact on the supply chain may ultimately decrease the ride-share supply in certain populous areas, causing customer attrition from the existing platforms and a decline in the overall quality of service in the share-ride market. Despite these risks, we expect YourDrive will put more data in drivers' hands, empowering them to plan how they can increase income each shift instead of relying on the baseline strategy provided by ride-share companies [16].

Based on the results of our analysis and modeling, ride-share drivers in the Chicago area can observe the following patterns (and more) using YourDrive:

1. Drivers may employ a strategy of fewer-but-longer trips by serving southern Chicago communities. These trips tend to arrive near downtown and have a predicted distance of 10-15 miles.
2. Rides in the early mornings are predicted to be longer than rides in the rest of the day.
    - Trips are predicted to be relatively short during the mid-day hours on weekdays.
3. Rides from O'Hare International Airport tend to be of similar length in the evenings (predicted to be approximately 10-20 miles) and similar during the day (approximately 5 miles).

These insights, among others, can empower drivers to customize their shifts instead of relying on ride-share platforms which are motivated to increase market share and income for the company. We expect that drivers who utilize these data insights will earn more income over time as they experiment with different strategies and witness changes in outcomes which was not possible before. This is the key appeal of our innovation - providing better data to ride-share drivers - so that the average individual can affect change and take control of their income stream and work satisfaction.

Distribution of Effort

All members have contributed to the project equally with effective communication and collaboration. Everyone has shown flexibility in task assignments and has been willing to help as needed.

References

[1] S. Eisenmeier, "Ride-sharing platforms in developing countries: Effects and implications in mexico city," *Pathways for Prosperity Commission*, no. 3, 2018.

[2] P. Ashkrof, G. Homem de Almeida Correia, O. Cats, and B. Arem, "Ride acceptance behaviour of ride-sourcing drivers," 07 2021.

[3] W. Lu and L. Quadrifoglio, "Fair cost allocation for ridesharing services – modeling, mathematical programming and an algorithm to find the nucleolus," *Transportation Research Part B: Methodological*, vol. 121, p. 41–55, Mar 2019.

[4] E. Bokányi and A. Hannák, "Understanding inequalities in ride-hailing services through simulations," *Scientific Reports*, no. 1, p. 10, 2020.

[5] Z. Muller, "Algorithmic harms to workers in the platform economy: The case of uber," *Colum. JL Soc. Probs*, vol. 53, p. 167, 2019.

[6] R. M. Berliner and G. Tal, "What drives your drivers: An in-depth look at lyft and uber drivers," in *Transportation Research Board 97th Annual Meeting*, 2018.

[7] T. Berger, C. Frey, G. Levin, and S. Danda, "Uber happy? work and well-being in the "gig economy"," *Economic Policy*, vol. 1, no. 1, pp. 1–54, 2019.

[8] M. Möhlmann and L. Zalmanson, "Hands on the wheel: Navigating algorithmic management and uber drivers," in *38th ICIS Proceedings*, 2017.

[9] M. Shapiro, "Density of demand and the benefit of uber," *Research Collection School Of Economics*, pp. 1–74, 2018.

[10] H. A. Chaudhari, J. W. Byers, and E. Terzi, "Putting data in the driver's seat: Optimizing earnings for on-demand ride-hailing," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, (New York, NY, USA), p. 90–98, Association for Computing Machinery, 2018.

[11] R. A. S. Hart, *Strategies Uber Drivers Use to Enhance Competitive Advantage for Increased Profits and Incomes*. PhD thesis, Walden University, 2021.

[12] J. Hall and A. Krueger, "An analysis of the labor market for uber's driver-partners in the united states," *Ilr Review*, vol. 71, no. 3, pp. 705–732, 2018.

[13] S. Jiang, L. Chen, A. Mislove, and C. Wilson, "On ridesharing competition and accessibility," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018.

[14] B. Chandar, U. Gneezy, J. List, and I. Muir, "The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment," Natural Field Experiments 00680, The Field Experiments Website, 2019.

[15] V. Nanda, P. Xu, K. Sankararaman, J. Dickerson, and A. Srinivasan, "Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2210–2217, 4 2020.

[16] S. Shokoohyar, "Ride-sharing platforms from drivers' perspective: Evidence from uber and lyft drivers," *International Journal of Data and Network Science*, vol. 2, no. 4, pp. 89–98, 2018.