Ridesharing in Chicago, IL:
Exploring community- and trip-level factors influencing willingness to pool

Valeria Balza
University of Chicago
vbalza@uchicago.edu

Michelle Orden
University of Chicago
morden@uchicago.edu

Abstract

The emergence of ridesourcing providers such as Uber and Lyft have prompted numerous questions around urban mobility, the gig economy, and environmental sustainability—among other topics. While recent research focuses on the benefits and limitations of ridesharing, we explore the factors affecting customers' willingness to share rides with other customers (i.e., ridesplitting). In particular, we integrate ridesourcing trip data featuring cost, time and spatial variables with socioeconomic and demographic data for Chicago, IL and train Lasso Regression and Random Forest models to predict customers' willingness to share rides with other customers. After accounting for multicollinearity in our data, we find that trip-level factors—namely cost and distance—and community area-level factors—namely percentage of white population and median income—are among the top features affecting customers' willingness to split rides. The identified patterns can help the city's policymakers identify community areas with the greatest potential to promote ridesplitting as the debate on ridesourcing behavior evolves.

1 Introduction

The emergence of ridesourcing providers such as Uber and Lyft have prompted numerous questions around urban mobility, the gig economy, and environmental sustainability—among other topics. While some argue such services provide first- and last-mile solutions to transit, others suggest the access to and use of ridesourcing services have been geographically and socially uneven (Jin et al., 2019, Rayle et al., 2016; Shaheen and Cohen, 2018; Su and Wang, 2019; Tarabay and Abou-Zeid 2019; Yan et al., 2019). Several studies, for instance, reveal demand for ridesourcing is concentrated in medium and large urban areas with younger, more educated, and wealthier populations—highlighting technological and financial barriers among low-income groups (Goodspeed et al., 2019; Grahn et al., 2019; King, Conway, and Salon, 2020; Spurlock et al., 2019; Young and Farber, 2019).

Recent research also points to the environmental effects of ridesourcing, noting ridesplitting—in which customers share rides with other customers heading in the same direction—can mitigate traffic congestion, reduce fuel consumption and greenhouse gas emissions, as well as curb parking demand (Wang and Yang, 2019; Cramer and Krueger, 2016; Tirachini and Gomez-Lobo, 2019; Xue et al. 2018). Still, a parallel body of research demonstrates ridesourcing providers have introduced more idle vehicles on the road, increased traffic congestion, and contributed to air and noise pollution (Rayle et al., 2016; Wei at al., 2017; Wenzel et al. 2019). Despite increasing evidence on the positive and negative effects of ridesourcing, limited research exists characterizing the factors affecting customers' willingness to share/pool rides with other customers. This characterization is important to better understand how the benefits and costs of ridesourcing are spatially distributed across cities.

We leverage data containing time, cost, and location features for ridesourcing trips for the city of Chicago and integrate it with community area-level socioeconomic and demographic variables. We then train two classification algorithms—Lasso Regression and Random Forest models—to predict a customer's willingness to share their ride with another customer.

Specifically, given a new rider's pickup and drop-off community area and the demographic and socioeconomic information associated with said community areas, what is the predicted outcome for whether the rider authorizes a shared ride? Further, what trip- and community-level features associated with a given ride contribute to whether or not a rider agrees to a pooled trip? Exploring such questions will allow us to see if and how ridesplitting behavior varies with socioeconomic and demographic variables related to a given community area—providing additional context to the debate on ridesourcing.

2 Data

2.1 Data Sources

This paper analyzes publicly-available ridesourcing trip data for the city of Chicago from January 2019 to December 2019 from Uber, Lyft, and Via. The original dataset contains 21 features, including trip start and end time, distance, duration, pickup and dropoff community areas, trip costs (disaggregated into fare, tip, additional charges, and total costs)—and a binary variable indicating whether the customer agreed to a shared trip with another customer.

We also draw on the 2014-2018 American Community Survey to incorporate demographic and socioeconomic features for each of the 77 community areas in Chicago, including: (i) median income; (ii) White population (percentage of total); (iii) Black/African American population (percentage of total); (iv) Hispanic population (percentage of total); (v) age distribution; (vi) employment levels; (vii) education levels; and (viii) and household sizes, among other features.

2.2 Data Pre-processing

Since the original ridesourcing data set contains 112 million rows and is approximately 30GB, we decided to use a 0.003 % random sample comprising 335,552 observations for our analysis due to computational power constraints.

To account for variations in ridesplitting behavior as a function of time, we created new features based on each trip's timestamp: $trip\_start\_month$, $trip\_start\_weekday$, and $trip\_start\_hour$. The feature $trip\_total\_per\_mile$ standardizes the trip cost by the miles traveled and the feature $trip\_total\_per\_min$ standardizes the trip cost by the minutes traveled.

Further, we dropped observations with missing values for trip duration, trip distance, pickup location, and drop-off location. To exclude outliers, we also dropped observations for which the trip distance was equal to 0, where the trip total is equal to $\$0$ dollars or greater than $\$100$ dollars, and where the trip duration is equal to 0 minutes or greater than 90 minutes. After cleaning, our data set comprised 287,763 observations to train and test our models. Finally, our target variable of interest, $shared\_trip\_authorized$, is a boolean variable equal to -1 if the customer did not authorize a shared trip and 1 if a shared trip was authorized.

The data set was merged with community area-level features from the 2014-2018 American Community Survey described above. We merged the community area data set on the pickup and drop-off location coordinates to account for demographic and socioeconomic variations across both pickup and drop-off locations in our models. Our final data set contains 89 numeric features.

3 Exploratory Analysis

Figures A1-A4, which visualize aggregate ride characteristics (total trip count, proportion of authorized pooled rides, average cost, and average distance) by community area, suggest ride characteristics are not evenly distributed across the city. At a high level, trip counts tend to be concentrated in the downtown Chicago and O'Hare airport areas. By contrast, the proportion of ridesourcing trips in which riders authorize pooling tend to be concentrated in the south and west parts of the city. Average trip cost and distance tend to be relatively higher for community areas in the city north and south edges of the city, presumably because these riders are traveling longer distances. Finally, Figure A5 shows trip demand and cost also vary by time—peaking at morning and evening rush hours during each week day.

4 Machine Learning Models

To better understand what community area- and trip-level characteristics affect the uneven distribution of ridesplitting behavior, we draw on the data sets described in Section 2 and train two classification models—via Lasso Regression and Random Forest—to predict whether or not a customer will agree to a shared ride. It should be noted that this does not mean the customer was actually matched for a pooled trip but is rather a reflection of a customer's willingness to share their ride.

4.1 Lasso Regression

4.1.1 Model

A least squares regression model allows us to take a set of n training features in a p-dimensional space and estimate a weight vector $\hat{w}$ such that $y = X\hat{w}$ where $X \epsilon \mathbb{R}^{n \times p}$. This weight vector can then be used to formulate a decision rule and predict new labels y_{new} for new observations x_{new}. The goal of such a model is to find the weight vector $\hat{w}$ that minimizes the error between the predicted labels and true labels.

In class, we explored Ridge Regression, a regularized version of least squares. In least squares, any error associated with the weight vector is amplified when the training data matrix X has small singular

values. Ridge Regression is one solution to this problem, as it finds $\hat{w} = \arg\min_w ||y - Xw||_2^2 + \lambda||w||_2^2$, which forces unimportant singular values to be close to zero, avoiding the amplification of error.

Similar to Ridge Regression, Lasso Regression is another regularized form of Least Squares. The difference between Ridge and Lasso is that Lasso's regularization term uses the L_1 norm of the weight vector instead of the square of the L_2 norm (Géron, 2019). One advantage of Lasso regression over Ridge regression is that it produces simpler and more interpretable models that incorporate only a reduced set of the predictors, aiding in feature selection. The equation for the weight vector $\hat{w}$ using Lasso Regression is as follows:

$$\hat{w} = \arg\min_w ||y - Xw||_2^2 + \lambda||w||_1$$

We first split our data set into training (80% of the data) and testing (%20 of the data) sets. We then use Lasso Regression on our training data set to estimate the weight vector (w_{Lasso}) using all 89 features. We do so by first standardizing the features. One decision that needs to be made when performing Lasso Regression is which value of the λ parameter to include in the model. If $\lambda = 0$, then the Lasso Regression is the same as Least Squares. When $\lambda > 0$, smaller values of λ prioritize data fit while larger values of λ prioritize suppressing noise.

To find the optimal value of λ, we employ a Grid Search, which uses cross-validation to evaluate various values of λ (or more generally, any set of hyperparameters) on the training set and compares the root mean squared error of the associated models. We perform Grid Search and determine that $\lambda = 0.1$ is the best parameter for our data. We then fit the model on our training data set using $lambda = 0.1$ and test the model's performance predicting the target variable, $shared\_trip\_authorized$, with our test data set.

Although Lasso Regression will regularize the weights of features with little importance, we find that some of the 89 features are highly correlated and/or are multiples of each other (see the correlation matrix in Figure A6). For example, the features $fare$, tip, $additional\_charges$, and $trip\_total$ represent very similar information, as $trip\_total$ is simply the sum of the previous three features. Additionally, $trip\_total\_per\_mile$ is highly correlated with $trip\_total$ and $trip\_miles$. We know from class that features (columns in X) that are linearly dependent do not provide additional information to our model. Based on this issue, we decided to perform Lasso Regression a second time on a smaller set of 78 linearly independent features.

4.1.2 Results and Performance

Figure A8 shows the coefficients (greater than 0.001 or less than -0.001) produced by Lasso Regression model using all 89 features. These coefficients specify how important the given feature is in predicting the outcome ($shared\_trip\_authorized$) as well as their sign. It is worth noting that the features representing the percent of the community area population that is employed and the percent of the community area population that has a bachelor's degree have a negative sign, suggesting that the higher these percentages are, the less likely the Lasso Regression model will predict that a given rider will agree to share their ride.

Figure A7 shows the absolute values of the coefficients produced by Lasso Regression in descending order using all 89 features. These coefficients specify how important the given feature is in predicting the outcome ($shared\_trip\_authorized$). As the figure shows, the three features with the highest rates are related to ride characteristics: $additional\_charges$, $trip\_seconds$, and $trip\_total\_per\_mile$. Interestingly, we also see four community related features with high importance. This result suggests that in addition to the features related to the actual ride (trip seconds and miles), pickup and dropoff community areas and their demographic information (education and employment information) also help us predict whether or not a rider will authorize to share their ride.

For the second model, Figure A10 shows the coefficients (greater than 0.001 or less than -0.001) produced by Lasso Regression using the smaller set of 78 features. Again, we see features such as

the percent of the community area that identifies as white and the percentage of the population that commutes for pickup and drop-off community areas have a negative sign, meaning that the higher these percentages are, the less likely our model will predict that a given customer will authorize to share their ride. Finally, Figure A9 shows the absolute values of the coefficients produced by the second model. The feature with the largest weight is $trip\_total\_per\_mile$.

The Lasso mean squared error (for all 89 features) is 74.67% and the mean squared error for the second model with the 78 features is 56.80%. Although Lasso Regression is a regularized model, carefully choosing features that are relatively independent of one another resulted in a better mean squared error.

4.2 Random Forest Model

To further explore our classification problem in the context of the data set's multicollinearity, we train a Random Forest model with the same set of features as above. Random Forest classification is a tree-based machine learning algorithm that draws random samples of data with replacement and generates decision trees for each sample. Once the model is trained, the algorithm passes new observations through each decision tree and outputs the class selected by most of the decision trees as the predicted target class for the new observation.

Decision tree splits are determined based on impurity measures like Gini, which measures the frequency at which any element of the data set will be mislabelled when it is randomly labeled:

$$GiniIndex = 1 - \sum_j p_j^2,$$

where p_j is the probability of class j. The minimum value of the Gini Index is 0, which occurs when all the contained observations in a given tree node are of one unique class (e.g., $shared\_trip\_authorized = -1$). In this case, the algorithm will not split the corresponding node. By contrast, the maximum value of the Gini Index is 0.5, which occurs when the probability of the two target classes occurring are the same. The variables with the most predictive power will be highest in the tree and have the largest mean decrease in Gini values while the least important variables will be lowest in the tree and will have the smallest mean decrease in Gini values.

4.2.1 Model and Results

First, we conduct a Random Forest model using the entire feature set using a gini impurity measure and three different metrics to evaluate model performance: recall, accuracy, and precision. The accuracy metric evaluates the fraction of predictions a model labeled correctly, recall evaluates the proportion of actual positives that were identified correctly, and precision evaluates the proportion of positive identifications that were actually correct.

The initial Random Forest model results in an accuracy score of 91.38%, a precision score of 87.86%, and a recall score of 62.95%. As Figure A11 shows, trip-level factors—namely $trip\ total\ per\ mile$, $trip\ miles$, and $trip\ seconds$ appear to have the highest predictive power. Community-level factors like $percent\ white\ population$ and $median\ income$, and also appear to affect ridesplitting behavior. These results align with the earlier results from the Lasso Regression model.

However, when data sets with correlated features are used, any of these correlated features can predict the target class, with no concrete preference of one over the others. Once one of the correlated features is used, the importance of the other features decreases since the impurity they can remove is already removed by the first feature (Saabas, 2014).

While Random Forest models avoid overfitting and can handle correlated features, interpreting Random Forest models with correlated features could lead to the incorrect conclusion that one of the variables is a strong predictor while the others in the same correlated group are unimportant—while in reality they have a similar effect on the dependent variable.

One possible solution for feature selection is to permute the values of each feature and measure how much the permutation decreases the accuracy of the model. Figure A12 shows the results from computing the permutation importance. As the figure shows, several community-level variables are correlated with each other (e.g., employment and education levels). Based on this observation, we then perform hierarchical clustering on the features' Spearman rank-order correlations, pick a threshold, and keep a single feature from each of the identified cluster and train a new Random Forest model with the selected features. Upon finding the clusters, we keep the following features from each cluster: $trip\_seconds$, $trip\_start\_weekday$, $trip\_start\_hour$, med_age, $median\_income$, $perc\_white$, $perc\_black$, $perc\_hisp$, $perc\_other\_race$, and $perc\_in\_lbfrc$.

The new model yields an accuracy score of 76.29%, a precision score of 32.43%, and a recall score of 22.09%. As Figure A13 shows, trip-level factors—namely $trip\_start\_hour$, $trip\ miles$, and $trip\ seconds$ appear to have the highest predictive power. Community-level factors like $percent\ white$ and $median\ income$ also appear to affect ridesplitting behavior.

5 Discussion

Overall, the Lasso Regression and Random Forest models reveal that both trip- and community-level factors are important in determining a customer's willingness to split rides. Importantly, a ride's cost and distance was consistently important across both models. This result makes sense intuitively, as customers expecting longer, higher cost rides may be more willing to split rides as a means to reduce associated costs. At the same time, both models also suggest there are relevant community-level factors influencing customers' ridesplitting behavior—namely the percentage of the total population that is White, median income and education levels. Ultimately, further research characterizing customers' ridesharing behavior is needed in this area to better inform the ongoing debate surrounding ridesourcing.

References

[1] Cramer, J., Krueger, A. B. (2016). Disruptive change in the taxi business: The case of Uber. American Economic Review, 106(5), 177-82.

[2] Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (Second edition.). Sebastopol, CA: O'Reilly.

[3] Goodspeed, R., Xie, T., Dillahunt, T. R., Lustig, J. (2019). An alternative to slow transit, drunk driving, and walking in bad weather: An exploratory study of ridesourcing mode choice and demand. Journal of transport geography, 79, 102481.

[4] Grahn, R., Harper, C. D., Hendrickson, C., Qian, Z., Matthews, H. S. (2020). Socioeconomic and usage characteristics of transportation network company (TNC) riders. Transportation, 47(6), 3047-3067.

[5] Jin, S. T., Kong, H., Sui, D. Z. (2019). Uber, public transit, and urban transportation equity: a case study in New York City. The Professional Geographer, 71(2), 315-330.

[6] King, D. A., Conway, M. W., Salon, D. (2020). Do For-Hire Vehicles Provide First Mile/Last Mile Access to Transit?. Transp. Find, 1-7.

[7] Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S. (2016). Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transport Policy, 45, 168-178.

[8] Shaheen, S., Cohen, A. (2019). Shared ride services in North America: definitions, impacts, and the future of pooling. Transport reviews, 39(4), 427-442.

[9] Spurlock, C. A., Sears, J., Wong-Parodi, G., Walker, V., Jin, L., Taylor, M., ... Todd, A. (2019). Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the San Francisco Bay Area. Transportation Research Part D: Transport and Environment, 71, 283-301.

[10] Su, Q., Wang, D. Z. (2019). Morning commute problem with supply management considering parking and ride-sourcing. Transportation Research Part C: Emerging Technologies, 105, 626-647.

[11] Tarabay, R., Abou-Zeid, M. (2020). Modeling the choice to switch from traditional modes to ridesourcing services for social/recreational trips in Lebanon. Transportation, 47(4), 1733-1763.

[12] Tirachini, A., Gomez-Lobo, A. (2020). Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. International journal of sustainable transportation, 14(3), 187-204.

[13] Wang, H., Yang, H. (2019). Ridesourcing systems: A framework and review. Transportation Research Part B: Methodological, 129, 122-155.

[14] Wei, H., Zuo, T., Liu, H., Yang, Y. J. (2017). Integrating land use and socioeconomic factors into scenario-based travel demand and carbon emission impact study. Urban rail transit, 3(1), 3-14.

[15] Wenzel, T., Rames, C., Kontou, E., Henao, A. (2019). Travel and energy implications of ridesourcing service in Austin, Texas. Transportation Research Part D: Transport and Environment, 70, 18-34.

[16] Xue, M., Yu, B., Du, Y., Wang, B., Tang, B., Wei, Y. M. (2018). Possible emission reductions from ride-sourcing travel in a global megacity: the case of Beijing. The Journal of Environment Development, 27(2), 156-185.

[17] Yan, X., Levine, J., Zhao, X. (2019). Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. Transportation Research Part C: Emerging Technologies, 105, 683-696.

[18] Young, M., Farber, S. (2019). The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. Transportation Research Part A: Policy and Practice, 119, 383-392.