Impacts of Rideshare: Predicting Inter-City Mobility Patterns in the Absence of Data

Frank Wang, Andualem Girma, Taylor Bixby

1. Introduction

Rideshare companies like Uber and Lyft have experienced tremendous growth over the last ten years. While some proponents believed this growth would result in reduced congestion by converting round trips to one way trips, it appears the opposite is true - rideshare seems to be increasing the vehicle miles driven in cities[1]. This raises an important question for key stakeholders in this space: where and when do we anticipate this increase in road usage to occur? For transportation planners and providers, public and private alike, this insight is highly impactful for several reasons. For the public sector, are high-flow times and places not well served by public transportation, and does the current infrastructure support rideshare services adequately? Changing bus routes, increasing the density of bus-stops, or frequency of car-waiting zones in these origin-destination (OD) pairs may be an effective policy choice. For private transportation companies, identification of these predicted geo-temporal peak volume routes could lead to further investigation of riderbase needs for these routes and uncover potential new market offerings for these peaks.

Predicting rideshare demand isn't a new problem though. Companies like Uber and Lyft do so everyday in order to price their trips; however, the optimal approach to modeling this demand is still very much a topic of research. Ke et al. developed a custom deep learning architecture, named the fusion convolutional long short-term memory network (FCL-Net), to better capture the spatio-temporal characteristics of rideshare demand.[2] It stacked multiple convolutional long short-term memory (LSTM) layers, standard LSTM layers, and convolutional layers - and showed significant improvement when compared to traditional benchmark algorithms like artificial neural networks and standard LSTMs. However, this class of problem - demand forecasting - isn't limited to ridesharing, and has been researched for many other applications. Chang et al. tested Gradient Tree boosting as a method of predicting bike demand as well as computer equipment demand.[3] The algorithm showed consistently better results than other standard prediction algorithms like support vector regression and multi-layer perceptron, across two very different datasets.

Our project aims to build on this work and provide rideshare insights to those without access to rideshare data by building a demand model for rideshare which can be employed in any US city. Covered in more detail below, our model for explaining variance in OD by time of day trips is trained on Chicago rideshare data taking into account the socio-geographic characteristics of the city and temporal information such as day-in-week, time-of-day and weather information.

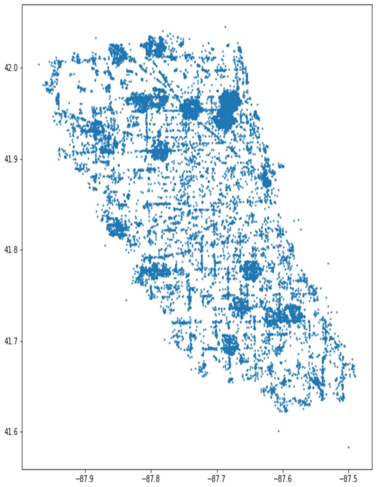
[1] "Disruptive Transportation: The Adoption, Utilization, and ...." 12 Oct. 2017, https://trid.trb.org/view/1485471.
[2] "Short-Term Forecasting of Passenger Demand under On ...." 20 Jun. 2017, https://arxiv.org/abs/1706.06279. Accessed 16 Dec. 2019.
[3] "Forecasting demand with limited information using Gradient ...." https://pdfs.semanticscholar.org/f170/60ab4315997dd027edbe1a9b0aa7538ab912.pdf. Accessed 16 Dec. 2019.

2. Methods and Results

Datasets

Our demand model relied on four primary sources of data; rideshare data from the Chicago Data Portal, census data from the US Census Bureau (USCB), weather data from Weather Underground, and *Place* data from Google Maps Places API.

Chicago Data Portal contains 100M+ records with 80+ columns about Uber and Lyft rides in the city of Chicago. The data was down-sampled from over a year of records to the month of September 2019 with roughly 4.3M records. The time range was selected to make sure the data is recent and relevant. Even with reducing the size of the data a GPU had to be used for data cleaning and merging. The trip pickup and drop-off latitude and longitudes and the pickup and drop-off were retained for all September 2019 ride records.



The latest census data (2010) was used to collect information on the population of Chicago such as car ownership, age statistics, population density, use of public transit and the average income of the area. This data was pulled at the tract level. Tracts, defined by the USCB, are small (in cities) contiguous geographic areas which contain approximately 5000 to 10,000 people.

The *Google Maps API* was used to pull Points Of Interest (POIs). POIs are geolocation tags which include the name a place and what type of place it is. Within Chicago there were 104 unique location types and 34k unique places of interest. These POIs are plotted across Chicago Tracts in the figure to the right.

Weather plays an important role in predicting the number of rides to expect between tracts. The weather data was sourced from World Weather Online. The temperature, precipitation, wind, wind-chill and humidity were retained for the hour across the city of Chicago (assuming consistency in conditions geographically but not temporally).

Feature Engineering

The assembled dataset consisted of relevant information that could be leveraged to predict the number of rides that can be expected to occur between a given origin-destination tract pare within a date-time period. For feature engineering we focused on POIs per area and key types of POIs as important features to express. We first one-hot encoded the tag converting the data from a list to 104 unique features. Next we grouped by track to get a total count of POIs by type and tract. While this gives good granularity, the 104 feature columns were quite sparse so we created a *Total POIs* feature. We also thought it was important to express access to public transit within the tract, so we created a summary feature for all transportation (bus, subway, train, etc.) within the tract. While tracts are relatively consistent at a population level, they vary widely in area. To reflect this we found the area of each tract and divided *Total_POI* and *Total_transport* features by this area to create density based metrics.

We performed a compute intensive merge of the POIs, census and weather tables at the many levels of temporal granularity shown in the table below. We then performed exploratory data analysis inspecting the average and distribution of record level counts at each of these levels of granularity. As granularity reduced, we saw a better spread on the distribution of rideshare demand, which could lead to more accurate predictions. However, there was also an important tradeoff to be considered between working with the lower levels of granularity that were more computationally manageable, and attempting to incorporate detailed information at the level at which the demand modelling would be useful for practical applications. For this reason we retained these varying levels of granularity and tested them throughout our modeling process.

Granularity	Example:	Records per OD pair
Day in Month by Hour	Sep 9th from 9-10PM	720
Day in Week by Hour	Mon from 9-10PM	168
Week days* by time of day	Tues from 6-10AM	16
All-weekdays* by Hour	Average of M,T,W,Th from 9-10PM	24
All-weekdays* by time of day	Average of M,T,W,Th from 6-10AM	4

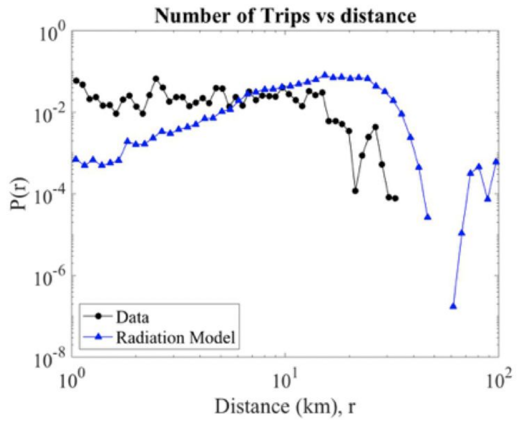
* Dropping weekend days from the dataset

We then normalized the data - bounding outliers, demeaning and dividing by the standard deviation per variable - for the purpose of doing PCA later, as well as for testing different regression models. After normalization, we used PCA to reduce feature dimensionality from the initial 315 features to 127, while preserving 95% of the predictive power.

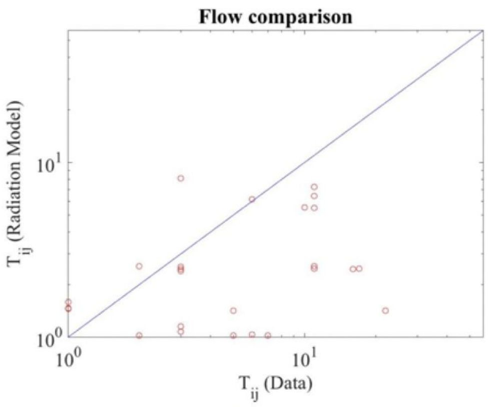
Modeling Results

When choosing how to measure our accuracy we considered standard accuracy metrics like the F1 score and Root Mean Squared Error (RMSE), but decided we should stick with the Common Part of Commuters (CPC) as a well accepted standard for benchmarking commuting network analysis models. Beyond its popularity in the field, we found that this metric placed penalty in accordance with our priorities. CPC places equal weight on misestimating rides across all predictions (does not disproportionately penalize misestimates in low volume or high volume instances). It also penalized overestimates and underestimates evenly.

For modeling, it was important to our team to develop a baseline to compare our boosted regressor and random forest models against. We used the Extended Radiation Model (ERM) for this purpose. For the ERM we used the population, POIs, distance and existing trip data as our input features. We used daily trip data and assumed an alpha value ranging from 0.1 to 2.0 from which we selected 0.1. We also experimented with taking different time aggregations of the data. Hourly aggregation resulted in unreasonable predictions, so we decided to aggregate by day. We also narrowed our selected days to weekdays only (Monday to Thursday) because of their relatively similar trends in distribution. This returned a CPC of 0.35.



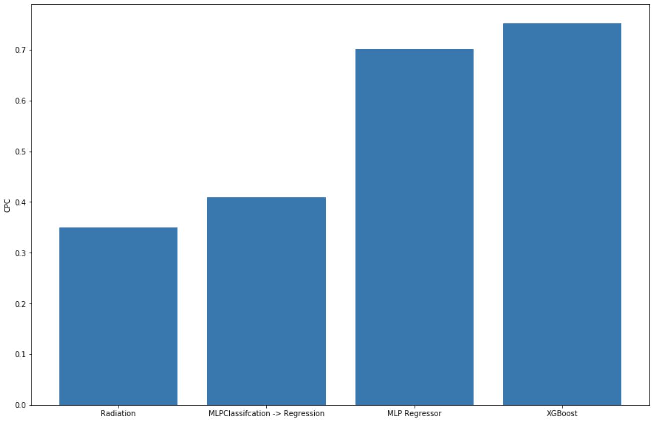
Distribution of trip counts given trip distance



ERM prediction vs data collected

After observing heavy skew in our data, with the vast majority of rideshare demand being equal to 1 for a numbers for an OD pair at a given time, we considered running two models in series; a classification model to separate few trips vs many trips and then a MPLRegressor to predict the count of trips for the many-trip classified subset. This was an attempt to further reduce prediction skew. Sadly this approach yielded an only slight improvement above baseline with a CPC of 0.41.

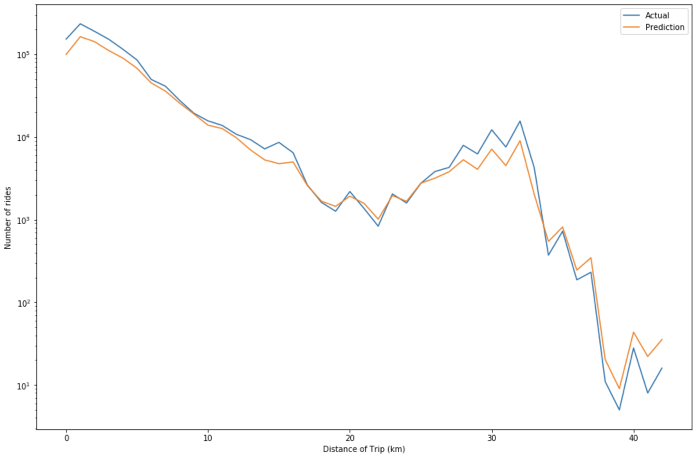
Running a MLPRegressor directly improved our results markedly yielding a CPC of 0.702. While we found this result initially surprising, review of the predictions rendered by this ensemble method provided insight into the observed increase in accuracy. The initial classification was assigning a boolean high/low value to all records. Because of the heavy left skew of the data and a simple mean ride count assignment label for all low-ride predicted records, the error created through this assignment and the relatively high count of records receiving this assignment significantly added to the final error.



To further improve results, we shifted to using a gradient boosted decision tree regressor, which has proven to be useful for a variety of demand modeling problems. After tuning, the model returned a CPC value of 0.752. The figure to the left show these improvements in CPC by model.

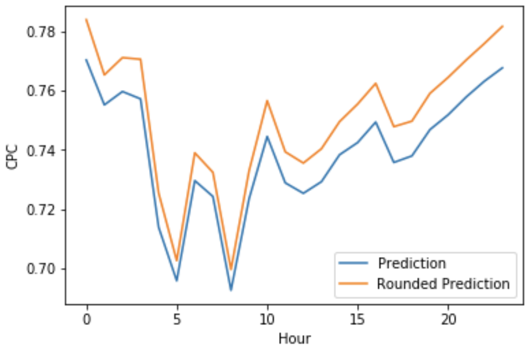
Final Model Results

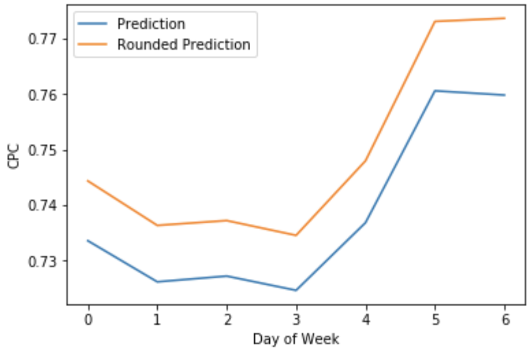
Overall the gradient boosted model performed the best of all 4 methods tested when evaluating by CPC. To better understand which aspects of the model worked well, and its pitfalls, we looked into various distributions of its results.



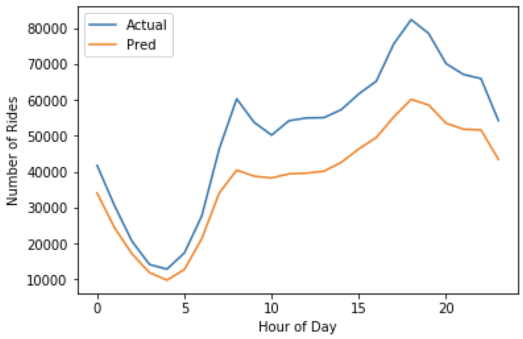
When predicting the relationship between the distance of trips and the log of the number of trips, the model performed very well - following the actual values closely. This would seem to indicate that the model was able to capture the spatial component to rideshare demand.

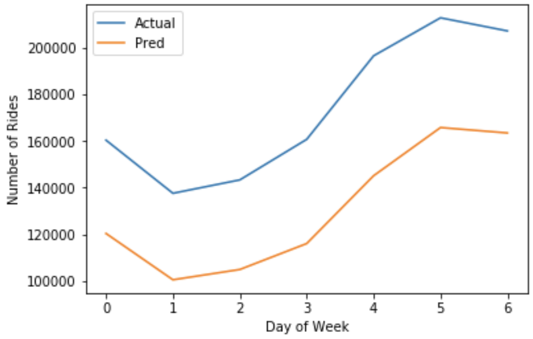
To better understand the temporal component, we graphed the CPC across different hours of the day and days of the week. The figures below indicate that our model had lower CPC values between 5 - 8a.m., as well as for all weekdays (days 0 - 4). Weekdays having lower CPC values makes intuitive sense, as those days are much more prone to having large fluctuations in demand. Addressing low CPC values during the morning would be another avenue for potential improvement, as they’re directly a result of 5a.m. being peak dropoff time at the Chicago airport, and 8a.m. being the time with the greatest spread of ride demand values. We also found that rounding the predicted value to the nearest integer slightly increased CPC values - as many estimates that should’ve been two rides or higher, oftentimes were in the 1.5 - 1.9 range.





To better understand how to improve predicted values in the future, we graphed the total number of rides per day-of-week and hour, against the predicted totals. In the figures below, it's clear that the gradient boosting model captures the overall shape of rideshare demand, but consistently predicts values too low. This is likely due to the fact that the number of rides for a given OD pair on some time/day is equal to one, 72.4% of the time. This means an area for great potential improvement is tuning the model to better recognize inputs that have more than one ride.





3. Future work and conclusions

Through this work we see that a Gradient Boosting regressor aggregated at the hour-weekday level most accurately predicts the number of rides. See Appendix A for the full table of results by testing method, feature transformation, and selection criteria. The various approaches to transform the assimilated data, including treating outliers, principal component analysis to reduce dimensionality, and log transformations were all contributing factors to increase model performance. Our model beat the baseline radiation model by a large margin, although more could be done to tune the hyperparameters of the gradient boosted regressor. We also found that the gains made from the PCA show that many of the features in the feature set we used were highly correlated with respect to the distribution patterns in the city of Chicago.

Real-world implications

The goal of this project was to construct a transportation demand model, using location and time characteristics to predict demand for rides in any US city. Our model was trained on data from the city of Chicago, but we see its value residing in the models ability to predict the number of rides in any date-time combination for any US city. Improving transportation network providers and planners understanding of the characteristics of rideshare activity in their city of interest enables more informed decision making about investments in infrastructure (roads, bus stations, etc.) and services (more buses, new mobility companies, etc.). This primary value is derived from the fact that the model can also be used in cities or regions where no historical ride-share data exists to model the demand.

While we already talked about which accuracy measurements were best to understand this model's performance, all measurements we considered focus on internal validity. We believe an exciting and necessary next step in this work is to test its generalizability and external validity. The model has already been refined for the Chicago dataset, so testing - and probably retraining - it on different cities is an impactful avenue for future work. We envision building a more robust, standardized model for all cities, with additional neural network layers to learn weights on the differences between cities. In summary, we see immediate real-world value this model could provide through directional insights to transportation decision makers. Beyond the current state of the model and noted areas for improvement, we also see great value in our final feature set and data transformation pipeline. Its ability to capture inter-city differences may be very useful in future transportation demand related analysis.

Appendix A

Model Results		Extended Radiation Model	Classifier + MLP Regressor*	MLP Regressor	Adaboost Regressor*	**Gradient Boosted Regressor**
Grouping (In addition to OD tract pairs)	Day in Week		X	X	X	**X**
	Weekdays	X				
	Time-of-day		X		X	
	Hour			X		**X**
Data Transformation	Outlier Limiting		X	X		**X**
	PCA		X	X		**X**
	Log(Y) (base e)		X	X		
Results	CPC (Avg)	0.35	0.41	0.702	0.61	**0.752**

* Ran models without Log transformation with lower CPC. Also ran models on Log e, log 2, log 10.