Journal of Transportation Technologies, 2021, 11, 250-264
https://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Demand Prediction of Ride-Hailing Pick-Up Location Using Ensemble Learning Methods

Divine Carson-Bell[1], Mawutor Adadevoh-Beckley[1,2], Kendra Kaitoo[1,2]

[1]College of Transport and Communication, Shanghai Maritime University, Shanghai, China
[2]Department of Transport, Regional Maritime University, Accra, Ghana
Email: Carsonbell555@gmail.com, Mbheckley07@gmail.com, Kaitookendra09@gmail.com

How to cite this paper: Carson-Bell, D., Adadevoh-Beckley, M. and Kaitoo, K. (2021) Demand Prediction of Ride-Hailing Pick-Up Location Using Ensemble Learning Methods. *Journal of Transportation Technologies*, **11**, 250-264.
https://doi.org/10.4236/jtts.2021.112016

Received: March 16, 2021
Accepted: April 22, 2021
Published: April 25, 2021

Copyright © 2021 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Ride-hailing and carpooling platforms have become a popular way to move around in urban cities. Based on the principle of matching riders with drivers, with Uber, Lyft and Didi having the largest market share. The challenge remains being able to optimally match rider demand with driver supply, reducing congestion and emissions associated with Vehicle clustering, deadheading, ultimately leading to surge pricing where providers raise the price of the trip in order to attract drivers into such zones. This sudden spike in rates is seen by many riders as disincentive on the service provided. In this paper, data mining techniques are applied to ultimately develop an ensemble learning model based on historical data from City of Chicago Transport provider's dataset. The objective is to develop a dynamic model capable of predicting rider drop-off location using pick-up location data then subsequently using drop-off location data to predict pick-up points for effective driver deployment under multiple scenarios of privacy and information. Results show neural network algorithms perform best in generalizing pick-up and drop-off points when given only starting point information. Ensemble learning methods, Adaboost and Random forest algorithm are able to predict both drop-off and pick-up points with a MAE of one (1) community area knowing rider pick-up point and Census Tract information only and in reverse predict potential pick-up points using the Drop-off point as the new starting point.

Keywords

Ride-Hailing, Braess Paradox, Vehicle Clustering, Deadheading, Congestion, Predictive Modelling, Vehicle Deployment, Ensemble Learning

1. Introduction

In recent years, ride-hailing and carpooling platforms have become increasingly

popular and convenient way of moving around in most modern cities, matching riders with drivers, with Uber, Lyft and Didi being the biggest providers within the industry. In light of increased environmental awareness as well as concerns on minimizing carbon footprint, ridesharing and carpooling has become increasingly important.

Carpooling has numerous societal and individual benefits, including but not limited to reduction of Greenhouse-Gas emissions, cost savings in terms of shared travel costs for public agencies and employers [1].

In their paper, [2] present salient points in the understanding of the key aspects of the existing ridesharing system, going on to design a framework to identify challenges in the use of ridesharing thus fostering the development of mechanisms to overcome and promote widespread use.

Emerging studies [3] demonstrate psychological factors such as monetary and time benefits becoming more dominant factors in decisions to use ride-hailing and carpooling services. In relation to rider satisfaction, [4] found surge pricing not to bias Uber towards riders of higher income threshold, but rather, homophilous matching that is, matching riders to drivers of a similar age resulted in higher ratings and further went on to use these insights to predict driver and/or rider retention. Examining ridesharing platforms, [5] concluded moving forward, these platforms will do more good than harm, also, it was found that relatively little is known about their efficiency and equity but is likely to change with growing research interest. Using online reviews of drivers of popular ride-hailing companies, Uber and Lyft, [6] was able to demonstrate preference of Uber to Lyft. In addition, analysis show increased competition to attract more drivers, for which drivers counted job flexibility, and meeting new people as main advantages. In contrast, insufficient compensation, poor job security, poor rider behavior and poor customer service as impeding factors.

2. Problem Statement

The Braess Paradox [7] [8] is a network phenomenon in which it is observed that the addition of extra capacity reduces overall network performance over time with lack of cooperation of users being the ultimate culprit for network breakdown

Congestion & **Vehicle-Clustering:** Over the years, the number of vehicles engaged in ride-hailing has increased astronomically, surpassing taxis in many urban cities, [9]. A report from the (Union of Concerned Scientists, 2020) shows that ride hailing trips are responsible for 69 percent more emissions than the trips the service displaces with a significant amount of trips being Deadheading (Dead-mileage). This constitutes the period between drop-off and pick-up and is associated with increased costs [10]. Surge pricing i.e. where prices are adjusted upwards to meet acute driver shortage is viewed a disincentive to many riders, leading to lost revenue.

Solving the problem

In order to combat the problems above, it is necessary to develop a sound driver deployment strategy. Collective Intelligence (COIN) [11] was first suggested as a way of solving Braess paradox. This involves all networks users acting centrally for the benefit of all. [12], Observed that strategic repositioning is key to maximizing driver earnings as against surge chasing which increases Dead-mileage. First and foremost will be to be able to predict and deploy vehicles accordingly. [13], conclude that centralized fleet coordination offers substantial benefits towards sustainable growth and market share.

Research Purpose and Objective

The objective is to develop a city-wide prediction algorithm capable of predicting trip pick-up and drop-off points, as well as potential pick-up locations after each drop-off based on historical data using Data mining Techniques.

Case study: City of Chicago, Illinois.

3. Related Works

The growth of demand for ride-hailing services has disrupted urban transportation and is changing the way in which people travel. Modern ride-hailing services require the development of efficient recommendation systems in order to improve both riders and driver experience. In response, many researchers have conducted various experiments to help predict ride hailing demand in order to improve effective ride-hailing vehicle deployment.

In attempting to optimize the number of pick-ups whilst minimizing waiting time for taxi services, [14] developed a ride-hailing recommendation system. This is completed in 3 phases. The model starts by first effectively estimating future customer demand in different clusters within the area of interest. This is followed up with a taxi-to-region matching according to preset rules and conditions including driver preference and finally concluded with the design of an optimized geo-routing algorithm to help drivers minimize dead-mileage. The problem with this mainly lies with the instability of driver preference which changes frequently, making the approach difficult to deploy in real world situations.

Dead-mileage comprises a significant share of total travel covered by drivers within the ride-hailing industry in terms of miles travelled and number of trips overall. Accurate demand prediction within the ride-hailing industry can greatly improve vehicle utilization whilst reducing waiting time. Customers mainly desire minimization of waiting time whilst drivers on the other hand aim to minimize deadheading and idle time after trips. This subsection of the industry comprises another area of strong research interest.

[15], is one of the first to study this emerging field. He develops a model which predicts the gap between rider demand and driver supply within a given time period and specific geographic area using Point of Interest (POI), Traffic, Weather data as well as data from Car sharing orders. A data sampling techniques is used to determine patterns and generalizations which can be applied in real case sce-

narios forming the basis for future work. This concept of finding the supply and demand gap is important as it allows for the deployment of drivers to improve the level of service

Time based demand prediction is another research area fast gaining ground. This is based on the premise of predicting ride-hailing vehicle demand in the next hour.

3.1. Operational Research Mobility Optimization

The vast majority of human interaction takes place in one of two areas; home or work. In order to further understand mobility patterns of users of ridesharing services across home and work locations, as well as social ties between users, [16] developed an algorithm for matching users with similar mobility patterns under constraints and concluded, a decrease in social distance of as much as 31% when users shared rides with others. These findings indicate the importance of the study of mobility patterns and the benefits which can be derived from optimizing ride-hailing services at an operational level. Using a more flexible yet extendible mobility model representing ride-sharing users movement and habits, [17] deploy a Variable-Order Markov Model (VOMM) underplayed with a Partial Matching (PPM) algorithm for next location prediction, with a prediction accuracy ranging from 60% - 81%. A major limitation of the usage of the PPM algorithm hovers around the compression process which tends to limit performance over time. In comparing the use of privately owned vehicles and two Autonomous Mobility on-Demand (AMoD) simulated on a real transport network based on current situation, under different scenarios, [18] found the deployment of AMoD system resulted in a major decrease in both number of vehicles required in order to meet transport needs (that is, 43% in AMoD1 and 88% in AMoD2) and street parking space required (58% in AMoD1 and 83% in AMoD2). [19], also cite effective road utilization as another advantage of designing the matching algorithm. Comparing the use of privately owned vehicles and two autonomous mobility on-demand (AMoD) simulated on a real transport network based on current situation, under different scenarios. Autonomous Mobility on-Demand vehicles are viewed by many as the future of transport, however their effectiveness hinders largely on the ability to coordinate their movement and predict demand as accurately as possible using the vast quantity of data we have available at our disposal, for which this paper seeks to pursue further.

In an attempt to resolve the surge of homeward-bound persons during the holiday seasons, [20] proposed a large-scale ridesharing system called **Country-Roads**® using an online greedy matching algorithm to match drivers and passengers, recording a success rate of 23.2%. Online Greedy matching algorithms have a comparatively low performance threshold when applied in complex systems such as ride-hailing services as experienced by the authors this is largely due to the level of rigidity of process making it not ideal for location prediction. Based on the concept of space-time windows, [21], develop a unique approach based

on Lagrangian relaxation, and conclude that the adoption of flexible pickup and delivery will evidently reduce system-wide cost whilst improving service quality. This hypothesis although found to be true, defeats the purpose of ride hailing services. Flexible pickup and delivery have not been widely accepted even within the carpooling sphere as centralized pick-up location is yet to gather wide acceptance.

3.2. Linear Programming & Statistical Methods

In implementing optimization solutions based on linear programming, [22] deploy a Tabubased meta-heuristic algorithm with the aim of solving the mixed integer linear program (MILP) under differing scenarios. The algorithm is observed to have a higher computational accuracy than control, the introduction of meet points to the ridesharing system reduces total travel time by 2.7% - 3.8% for scaled tests. With meet-points not having been widely accepted within the ride-hailing and carpooling industry, the benefits of reduced travel time, and reduced travel costs associated with it cannot be fully quantified. Especially given Covid-19 social distancing protocols. This demonstrates the need to improve location prediction as a lasting solution.

From the domain of probability and statistics, [23] having collected data of taxi trips in New York, Singapore, San Francisco and Vienna compute shareability curves for each city, then through natural rescaling collapse them into a universal curve which is used to predict the potential of ridesharing in any given city based on a few qualities and parameters. The statistical methods employed here demonstrate the general overview of the potential of the growth of ride-hailing services in any given city. This is to help with city planning purposes and fails to examine rider-driver interaction.

Examining the relationship between the frequency and probability of ridesharing usage, and frequency of public transit usage, [24], develop a Zero-inflated negative binomial regression model.

Results show a positive relationship between ridesharing and public transit use particularly for people living in areas of high population density and comparatively fewer vehicles. The significance of this is to allow the measurement of ride-hailing service utilization across population densities across any given city taking into consideration anticipated demand and in the selection of the research Case study.

4. Research Framework and Design

To reduce the number of vehicles, alleviate traffic jams and curb pollution in transporting people in office hubs in Poland, [25] collected a representative sample of the population and used spatial data mining techniques to develop a set of parameters for the multi-agent system. Using the distributed model-free, system DeepPool® based on deep Q-network (DQN) techniques, [26] develop an algorithm able to learn the optimal dispatch policy through interaction with the en-

vironment, incorporating travel demand statistics and a dataset of taxi trips in New York to dispatch vehicles and anticipate future demand. Deploying a convolutional neural network (CNN) based on deep learning for multi-step ride-hailing demand prediction using trip request data in Chengdu, [27] showcase faster training and prediction of CNN models compared to the use of Long Short Term Memory (LSTM) models.

4.1. Data

In conducting this research, a large scale dataset of rideshare and taxi trips spanning 2018/2019 in Chicago is collected, as shown in **Table 1**, with each observation consisting of the following elements:

The data is processed and cleaned. As a first step, a comprehensive understanding of the individual features within the dataset is required, as well as knowledge of trip distribution across the city, from origin (O) to Destination (D). Numerous studies have demonstrated the importance of regional partitioning in location prediction. Research and experiments by [28] demonstrated that regional partitioning led to better forecast and demand prediction of geospatial data.

This is followed up with followed by scenario development. **Figure 1** shows a color-coded layout of the City of Chicago, detailing its community areas as well as census tracts.

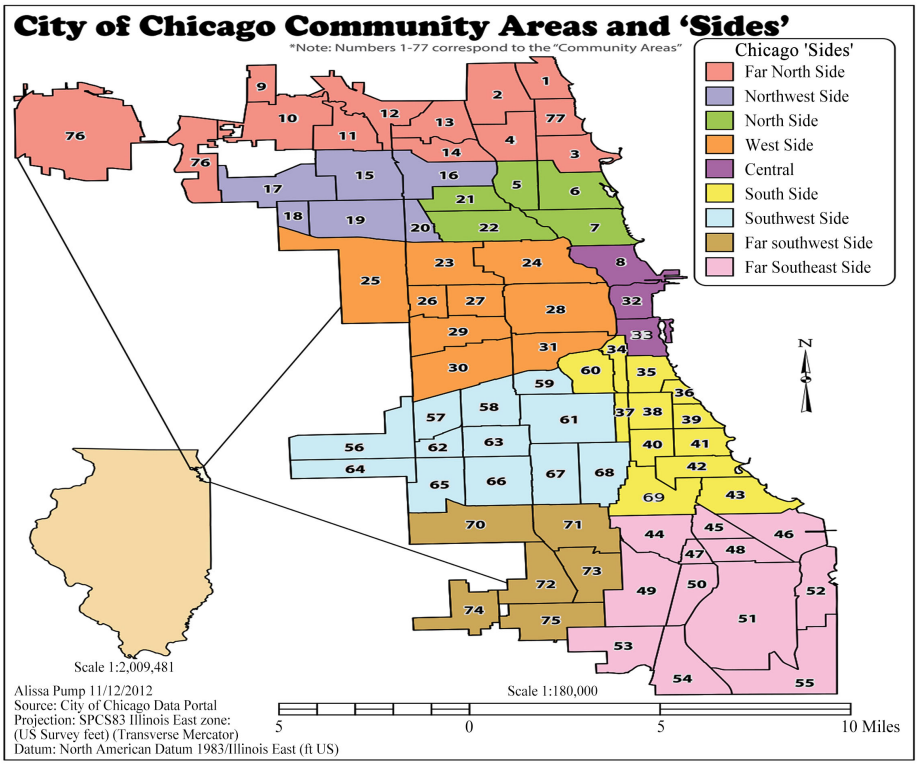

Figure 1. Map of City of Chicago including its community Areas and Census tracts. Visualization of potential pick-up and drop-off points across the city.

Table 1. Data points used for data mining and the development of the predictive algorithms.

Trip ID	DATA FEATURES AND ATTRIBUTES PER OBSERVATION		
Trip Start Timestamp	Drop-off Census Tract	Pickup Census Tract	Fare
Trip End Timestamp	Drop-off Community Area	Pickup Community Area	Shared Trip Authorized
Trip Seconds	Drop-off Centroid Longitude	Pickup Centroid Longitude	Additional Charges
Trip Miles	Drop-off Centroid Latitude	Pickup Centroid Latitude	Trips Pooled
Trip Total	Drop-off Centroid Location	Pickup Centroid Location	Tip

4.2. Research Framework and Scope

Multidimensional Scenario Formulation

Scenario performance analysis allows for measuring performance under varied rider privacy limitations.

Scenario 1

Location prediction with no information *i.e.* drop-off community area (destination) prediction with only pick-up (origin) data, and vice versa. This is in order to allow for riders with strict privacy concerns in information release, measuring ability to predict trip start and end points given rider privacy restrictions.

Scenario 2

Location prediction with partial information. That is, drop-off community area (destination) prediction with pick-up data and Census Tract (destination zone) information, vice versa. It is based on the idea of being able to predict trip start and end points under rider uncertainty.

Steps and Methodological process

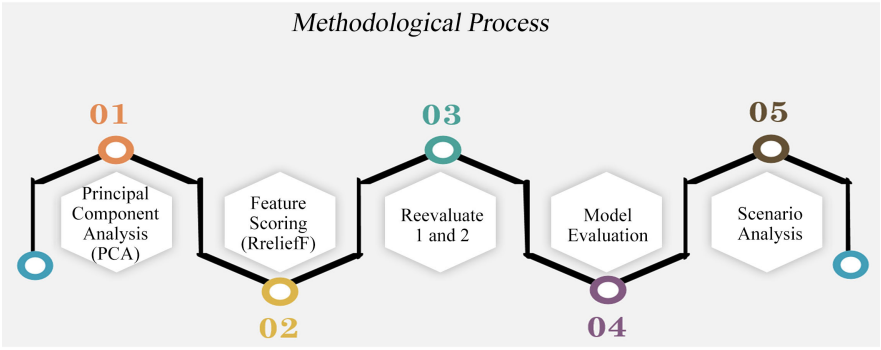

Figure 2. Step by step methodological process in designing and evaluating predictive models used.

Figure 2 shows the steps taken in the design, evaluation and interpretation of the research framework employed in carrying out this work.

1) Perform Principal Component Analysis (PCA) on trip dataset. Record and analyze results against degree of variance covered by each principal component.

2) Perform feature scoring and ranking using Relief metrics. Record and analyze results.

3) Reevaluate steps 1 and 2. Determine features and variables with largest

weight in designing and building the model.

4) Evaluation and scoring of prediction accuracy and error tolerance (MAE, MSE, and R2) under both scenario 1 and 2.

5) In-depth scenario analysis of both scenario 1 and 2, firstly on drop-off community area prediction and pick-up community area prediction.

6) Analyzing implications on surge pricing policy and ridesharing efficiency.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is based on the use of an orthogonal transformation to convert a set of observations with possibly correlated variables in a set of linearly uncorrelated principal components using eigenvalues to measure the total degree of variance explained by each factor.

FEATURE RANK USING RRELIEFF

The RReliefF algorithm estimates the quality of an attribute according to the degree with which it discriminates between instances near each other. Here, an instance R is randomly selected, then the K-nearest instances with respect to class value are selected. The difference between the value of A of R as well as the value of the same attribute for one of the K-instances is then compared with respect to the difference of their class values. This process is repeated and ultimately yields a weight for each attribute ranging between −1 and 1.

Cross Validation Model Evaluation and Scoring

The Leave-P-Out Cross Validation (CV) approach leaves "p" data points out of the training data, with a sample size of n-p being used as the validation set. This process is repeated for all possible combinations, with error being averaged for all trials in order to determine overall effectiveness.

To measure the degree of error of the developed models, error metrics will then be used to judge model quality and compare the different regression models. The Mean Average Error (MAE), Mean Squared Error (MSE), and R-Squared (R^2) will be used for evaluation.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y_i}\right|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2$$

where,

$\widehat{Y}$ —Predicted value of Y

$\overline{Y}$ —mean value of Y

$$\mathrm{R}^2 = \frac{(\mathrm{SSEM} - \mathrm{SSER})}{(\mathrm{SSEM})} = 1 - \frac{\mathrm{SSER}}{\mathrm{SSEM}}$$

where,

SSEM is the sum of Squared Errors by Mean line and

SSER is the sum of Squared Errors by Regression Line

Predictive Modelling using Ensemble Learning

Generally, ensemble learning is the term used to describe meta-algorithms that makes predictions based on inputs from different models, thus, by combining

multiple individual models, the ensemble model tends to have less bias, variance, and avoids overfitting culminating in improved predictions.

Adaboost and Random Forest are the most commonly used.

5. Framework and Results

5.1. Principal Component Analysis (PCA)

In analyzing the weights of the individual features within the data sample collected, PCA analysis is performed, measuring the degree of variance covered by each principal component within the data set.

Analysis of PCA results reveals an increase in the degree of variance explained by each of the data attributes within the dataset.

Figure 3 describes the results obtained from PCA analysis. Results show that certain attributes within the dataset are able to explain 55.9% of the recorded variance, with 5 attributes able to explain 73.7% of the variance and so on. This aids in selecting the most important data attributes which will effectively improve the models prediction accuracy. Analysis reveals that 9 attributes to be the optimal number of features to incorporate in building the models.

Feature Scoring and Rank

After PCA analysis, the features within the dataset are then ranked in order according to feature influence on prediction output. Figure 4 details the weight associated with each attribute used in designing the model, with some attributes being more critical to predictive performance than others.

RreliefF is used to rank and measure individual features by level of importance as shown above.

5.2. Re-Evaluation and Model Calibration

Scenario 1

Predicting drop-off community area (destination) with only pick-up (origin) data. Model results show an ability of linear regression models to predict potential Drop-off areas within a radius of 13 blocks (community area). This is in the absence of any information other than pick-up point (origin).

The results are shown in Figure 5:

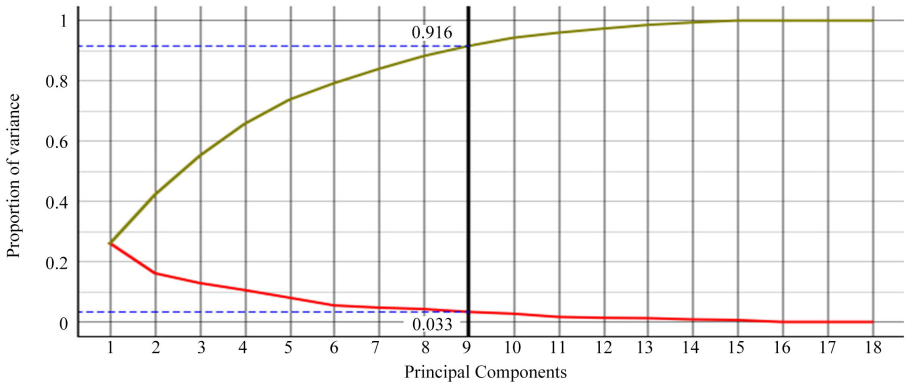

Figure 3. PCA analysis displaying the level of variance covered by each of the inputs.

	#	RReliefF
Dropoff Census Tract		0.204
Dropoff Centroid Latitude		0.147
Pickup Community Area		0.124
Fare		0.122
Trip End Timestamp		0.114
Trip Start Timestamp		0.114
Dropoff Centroid Longitude		0.106
Trip Miles		0.099
Trip Seconds		0.095
Pickup Centroid Latitude		0.095
Trip Total		0.088
Pickup Census Tract		0.083
Pickup Centroid Longitude		0.079
Additional Charges		0.048
Shared Trip Authorized	2	0.042
Tip		0.037
Trips Pooled		0.029

Figure 4. Feature rank displaying the weight of each Data point in prediction performance and drop-off Community area distribution graph.

Method	MSE	RMSE	MAE	R2
Linear Regression	328.828	18.134	13.060	0.088
Neural Network	323.470	17.985	13.182	0.103
SGD	336.336	18.339	13.206	0.068
Random Forest	392.224	19.805	14.358	-0.087
kNN	400.869	20.022	14.449	-0.111
SVM	391.029	19.774	14.689	-0.084
AdaBoost	516.937	22.736	15.256	-0.433
Tree	544.699	23.339	16.641	-0.510

	Linear Regression	Neural Network	SGD	Dropoff Community Area
1	8.6	15.0	6.4	1.0
2	22.3	22.4	24.2	24.0
3	20.9	21.5	22.1	24.0
4	20.7	21.9	23.3	8.0
5	17.0	19.8	18.2	76.0
6	19.6	21.0	20.4	28.0
7	29.1	27.0	29.4	24.0
8	24.8	22.6	28.2	24.0
9	15.8	12.6	14.0	7.0
10	24.4	23.9	26.2	8.0
11	20.1	16.4	18.1	6.0
12	21.3	22.1	23.6	6.0
13	20.7	21.4	22.1	24.0
14	29.5	31.5	26.2	34.0
15	19.8	15.5	18.7	24.0
16	26.0	21.6	28.7	7.0
17	20.0	18.0	23.8	3.0
18	22.0	22.7	22.8	32.0
19	30.7	37.0	31.8	28.0
20	22.4	22.4	24.2	24.0
21	29.4	27.1	30.2	28.0
22	31.1	28.1	27.9	24.0
23	18.6	17.6	21.9	1.0
24	20.6	17.6	19.3	3.0

Figure 5. Evaluation results of predictive accuracy of algorithms and results comparison under scenario 1.

This figure is divided into 2 parts, with the first part (Top) displaying results from model evaluation whilst the 2nd displays location predictive results against actual. The dark column above displays actual drop-off community areas as against predicted values on its left.

Scenario 2

Predicting drop-off community area (destination) with partial information, that is, (destination zone) information

Model results show an ability of ensemble learning models such as Adaboost and Random Forest to predict potential Drop-off areas precisely with error under 1 block (community area).

This is in the absence of any information other than pick-up point (origin). The results are shown below.

Figure 6 is divided into 2 parts, with the first part (Top) displaying results from model evaluation whilst the 2nd displays location predictive results against actual. The dark column above displays actual drop-off community areas as against predicted values on its left.

Evaluation Results

Method	MSE	RMSE	MAE	R2
AdaBoost	10.919	3.304	0.471	0.964
Tree	32.782	5.726	0.791	0.893
Random Forest	30.225	5.498	1.477	0.901
SVM	252.072	15.877	11.399	0.176
kNN	109.460	10.462	4.282	0.642
Linear Regression	159.293	12.621	7.830	0.479
Neural Network	149.562	12.230	8.325	0.511
SGD	197.649	14.059	9.254	0.354

	AdaBoost	Tree	Random Forest	Dropoff Community Area
1	1.0	1.0	1.0	1.0
2	24.0	24.0	22.0	24.0
3	24.0	24.0	24.0	24.0
4	8.0	8.0	8.0	8.0
5	76.0	76.0	76.0	76.0
6	28.0	28.0	28.0	28.0
7	24.0	24.0	24.0	24.0
8	24.0	24.0	24.0	24.0
9	7.0	7.0	7.0	7.0
10	8.0	8.0	8.0	8.0
11	6.0	6.0	6.0	6.0
12	6.0	6.0	12.6	6.0
13	24.0	24.0	24.0	24.0
14	34.0	34.5	34.1	34.0
15	24.0	24.0	24.0	24.0
16	7.0	7.0	13.2	7.0
17	3.0	3.0	17.8	3.0
18	32.0	32.0	32.5	32.0
19	28.0	28.0	28.8	28.0
20	24.0	24.0	24.0	24.0
21	28.0	28.0	28.0	28.0
22	24.0	24.0	24.0	24.0
23	1.0	1.0	1.0	1.0
24	3.0	3.0	3.0	3.0

Figure 6. Evaluation results of predictive accuracy of algorithms and results comparison under scenario 2.

Pick-up point Prediction after Drop-off

This part focuses on predicting demand centers within the city after the any given drop-off. The aim is to predict rideshare demand centers, anticipating demand and price surge before they happen.

In an effort to optimize rideshare vehicle distribution, it is imperative to be able to predict where demand will occur ahead of time, taking advantage of imbalance of supply and demand as well as revenue per trip, with the results displayed in Figure 7 below.

This figure is divided into 2 parts, with the first part (Top) displaying results from model evaluation whilst the 2nd part displays location predictive results against actual. The dark column above displays actual drop-off community areas as against predicted values on its left.

5.3. Discussion

Research into the field of mobility remains a hot topic amongst many researchers. Mobility-As-A-Service (MAAS) where vehicle trips are used to render services has come to stay in the era where we've experienced a boom in ride-hailing

Method	MSE	RMSE	MAE	R2
Random Forest	23.987	4.898	1.811	0.934
AdaBoost	25.317	5.032	1.110	0.930
Tree	44.073	6.639	1.722	0.878
kNN	151.521	12.309	5.240	0.580
Neural Network	242.069	15.559	10.924	0.329
Linear Regression	244.149	15.625	10.669	0.323
SGD	265.588	16.297	11.166	0.264
SVM	342.125	18.497	13.107	0.052

	Tree	Random Forest	AdaBoost	Pickup Community Area
1	1.0	6.8	1.0	1.0
2	8.0	8.0	8.0	8.0
3	8.0	8.0	8.0	8.0
4	8.0	8.0	8.0	8.0
5	6.0	6.0	6.0	6.0
6	13.5	14.5	14.0	14.0
7	28.0	27.9	28.0	28.0
8	28.0	29.3	28.0	28.0
9	6.0	6.0	6.0	6.0
10	32.0	32.0	32.0	32.0
11	8.0	8.0	8.0	8.0
12	8.0	8.0	8.0	8.0
13	8.0	8.0	8.0	8.0
14	30.0	29.1	28.0	28.0
15	8.0	8.0	8.0	8.0
16	24.0	24.0	24.0	24.0
17	77.0	77.0	77.0	77.0
18	8.0	8.0	8.0	8.0
19	76.0	76.0	76.0	76.0
20	8.0	8.0	8.0	8.0
21	26.0	26.6	28.0	28.0
22	31.0	31.0	31.0	31.0
23	77.0	74.7	77.0	77.0
24	18.0	13.4	8.0	8.0

Figure 7. Evaluation results of predictive accuracy of algorithms and results comparison under scenario.

services. The need to optimize the operations of these services remains of utmost importance. The results show neural network algorithms perform best in generalizing pick-up and drop-off points when provided with only starting point information. The significance of this is to allow for trip generalization in pooled trips, where riders are most likely to have a common drop-off point, e.g. co-worker's trip to work or trips to work or shared trip to a sporting event. Ensemble learning methods, Adaboost and Random forest algorithm are able to predict both drop-off and pick-up points with a MAE of 1 community area knowing rider pick-up point and Census Tract information only and in reverse predict potential pick-up points using the Drop-off point as the new starting point. This allows the algorithm to confidently predict the most likely pick-up point of potential riders following a drop-off in in so doing increasing supply of drivers into potential surge zones and thus being less reactive, more proactive in trip deployment. Here, it can be seen that the introduction of more data and ensemble learning techniques greatly increases the precision accuracy of the model. This demonstrates the influence of data management within the ride-hailing industry, especially in a time when privacy concerns and right to privacy have become a matter of safety and security, of which varies from rider to rider. Direct impacts on the ride-hailing industry and operations include:

Implications on ride-hailing Industry includes:

1) Improved vehicle utilization, and time efficiency.
2) Reduced dead-mileage and idle time after trips.
3) Improvement riders and driver experience.

6. Conclusions

In conclusion, results from the research indicate the ability to use predictive modelling and analytics to adequately maximize driver positioning and deployment by predicting surge zones before they occur irrespective of rider privacy settings.

The implications of these results on the transport industry includes:

- ✓ Reduced incidence of the surge and increasing rider satisfaction.
- ✓ Reduced transport costs.
- ✓ Increase in the ease of parking particularly in high-demand (downtown) areas.
- ✓ From a social and environmental point of view for fewer wasted miles would translate into less emissions overall.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Shaheen, S., Cohen, A. and Bayen, A. (2018) The Benefits of Carpooling. UC Berkeley. https://escholarship.org/uc/item/7jx6z631

[2] Furuhata, M., Dessouky, M., Ordóñez, F., Brunet, M.-E., Wang, X. and Koenig, S. (2013) Ridesharing: The State-of-the-Art and Future Directions. *Transportation Research Part B*: *Methodological*, **57**, 28-46. https://doi.org/10.1016/j.trb.2013.08.012

[3] Olsson, L.E., Maier, R. and Friman, M. (2019) Why Do They Ride with Others? Meta-Analysis of Factors Influencing Travelers to Carpool. *Sustainability*, **11**, 2414. https://doi.org/10.3390/su11082414

[4] Kooti, F., Grbovic, M., Aiello, L.M., Djuric, N., Radosavljevic, V. and Lerman, K. (2017) Analyzing Uber's Ride-Sharing Economy. *International World Wide Web Conference Committee* (*IW3C2*), Perth, 3-7 April 2017. https://doi.org/10.1145/3041021.3054194

[5] Hahn, R. and Metcalfe, R. (2017) The Ridesharing Revolution: Economic Survey and Synthesis. Volume IV: More Equal by Design: Economic Design Responses to Inequality. Oxford University Press, Oxford.

[6] Shokoohyar, S. (2018) Ride-Sharing Platforms from Drivers' Perspective: Evidence from Uber and Lyft Drivers. *International Journal of Data and Network Science*, **2**, 89-98. https://doi.org/10.5267/j.ijdns.2018.10.001

[7] Braess, D. (1968) Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, **12**, 258-268. https://doi.org/10.1007/BF01918335

[8] Braess, D., Nagurney, A. and Wakolbinger, T. (2005) On a Paradox of Traffic Planning. *Transportation Science*, **39**, 446-450. https://doi.org/10.1287/trsc.1050.0127

[9] Schneider, T. (2019) Taxi and Ride-Hailing Usage in Chicago. http://toddwschneider.com/dashboards/chicago-taxi-ridehailing-data

[10] Nair, G.S., Bhat, C.R., Batur, I., Pendyala, R.M. and Lam, W.H.K. (2020) A Model of Deadheading Trips and Pick-Up Locations for Ride-Hailing Service Vehicles. *Transportation Research Part A*: *Policy and Practice*, **135**, 289-308. https://doi.org/10.1016/j.tra.2020.03.015

[11] Tumer, K. and Wolpert, D. (2002) Collective Intelligence and Braess' Paradox. *Journal of Artificial Intelligence Research*, **16**, 359-387. https://doi.org/10.1613/jair.995

[12] Chaudhari, H.A., Byers, J.W. and Terzi, E. (2018) Putting Data in the Driver's Seat: Optimizing Earnings for On-Demand Ride-Hailing. 11*th Eleventh ACM International Conference on Web Search and Data Mining*, New York, 5-9 February 2018, 9 p. https://doi.org/10.1145/3159652.3159721

[13] Merlin, L.A. (2019) Transportation Sustainability Follows from More People in Fewer Vehicles, Not Necessarily Automation. *Journal of the American Planning Association*, **85**, 501-510. https://doi.org/10.1080/01944363.2019.1637770

[14] Wan, X., Ghazzai, H. and Massoud, Y. (2020) A Generic Data-Driven Recommendation System for Large-Scale Regular and Ride-Hailing Taxi Services. *Electronics*, **9**, 648. https://doi.org/10.3390/electronics9040648

[15] Wang, R. (2017) Supply-Demand Forecasting for a Ride-Hailing System. http://escholarship.org/uc/item/7hr5t5vv

[16] Cici, B., Markopoulou, A., Laoutaris, F.-M. and Nikolaos, E.A. (2017) Assessing the Potential of Ride-Sharing Using Mobile and Social Data: A Tale of Four Cities. https://doi.org/10.1145/2632048.2632055

[17] Roor, R., Karg, M., Liao, A., Lei, W. and Kirsch, A. (2017) Predictive Ridesharing Based on Personal Mobility Patterns. *Intelligent Vehicles Symposium* (*IV*), Los Angeles, 11-14 June 2017. https://doi.org/10.1109/IVS.2017.7995895

[18] Dia, H. and Javanshour, F. (2017) Autonomous Shared Mobility-on-Demand: Melbourne Pilot Simulation Study. *Transportation Research Procedia*, **22**, 285-292.

https://doi.org/10.1016/j.trpro.2017.03.035

[19] Sonet, K.M.H., Rahman, M.M., Mehedy, S.R. and Rahman, R.M. (2019) A Dynamic Ridesharing and Carpooling Solution Using Advanced Optimised Algorithm. *International Journal of Knowledge Engineering and Data Mining*, **6**, 1-31. https://doi.org/10.1504/IJKEDM.2019.097355

[20] Jiang, W., Dominguez, C.R., Zhang, P., Zhang, S., *et al.* (2018) Large-Scale Nationwide Ridesharing System: A Case Study of Chunyun. *International Journal of Transportation Science and Technology*, **7**, 45-59. https://doi.org/10.1016/j.ijtst.2017.10.002

[21] Zhao, M., Yin, J., An, S., Wang, J. and Feng, D. (2018) Ridesharing Problem with Flexible Pickup and Delivery Locations for App-Based Transportation Service: Mathematical Modeling and Decomposition Methods. *Journal of Advanced Transportation*, **2018**, Article ID: 6430950. https://doi.org/10.1155/2018/6430950

[22] Li, X., Hu, S., Fan, W. and Deng, K. (2018) Modeling an Enhanced Ridesharing System with Meet Points and Time Windows. *PLoS ONE*, **13**, e0195927. https://doi.org/10.1371/journal.pone.0195927

[23] Tachet, R., Sagarra, O., Santi, P., Resta, G., Szell, M., Strogatz, S.H. and Ratti, C. (2017) Scaling Law of Urban Ride Sharing. *Scientific Reports*, **7**, Article No. 42868. https://doi.org/10.1038/srep42868

[24] Zhang, Y. and Zhang, Y. (2018) Exploring the Relationship between Ridesharing and Public Transit Use in the United States. *International Journal of Environmental Research and Public Health*, **15**, 1763. https://doi.org/10.3390/ijerph15081763

[25] Olszewski, R., Pałka, P. and Turek, A. (2018) Solving "Smart City" Transport Problems by Designing Carpooling Gamification Schemes with Multi-Agent Systems: The Case of the So-Called "Mordor of Warsaw". *MDPI Sensors*, **18**, 141. https://doi.org/10.3390/s18010141

[26] Alabbasi, A., Ghosh, A. and Aggarwal, V. (2019) DeepPool: Distributed Model-Free Algorithm for Ride-Sharing Using Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, **20**, 4714-4727. https://doi.org/10.1109/TITS.2019.2931830

[27] Wang, C., Hou, Y. and Barth, M. (2019) Data-Driven Multi-Step Demand Prediction for Ride-Hailing Services Using Convolutional Neural Network. *Computer Vision Conference* (*CVC*), Las Vegas, 25-26 April 2019. http://www.researchgate.net/publication/329402005 https://doi.org/10.1007/978-3-030-17798-0_2

[28] Niu, K., Wang, C., Zhou, X. and Zhou, T. (2019) Predicting Ride-Hailing Service Demand via RPA-LSTM. *IEEE Transactions on Vehicular Technology*, **68**, 4213-4222. https://doi.org/10.1109/TVT.2019.2901284