Travel Time Prediction in Ride-Sourcing Networks: A Case Study for Machine Learning Applications

Sina Shokoohyar
Saint Joseph's University

Ahmad Sobhani
Oakland University

Rashmi Malhotra
Saint Joseph's University

Weimin Liang
Saint Joseph's University

ABSTRACT

This paper explores the applications of machine learning for predicting the travel time in the ride-sourcing networks using the Uber movement dataset. Using the Python programming environment, a case study is presented to analyze the travel time of the ride-sourcing services from the central Washington D.C. to the given specific destinations by considering the distance, railway/subway and street density in different destination zones (areas) and also weather conditions. To this end, in the first step, a descriptive analytics is completed to include potential features (attributes) affecting the travel times of Uber (ride-sourcing) services. Then, machine learning techniques such as random forest and robust regressions are applied to identify key attributes (features) for the prediction of the average travel times. The findings and accuracy of the robust regression models are compared with the random forest to select the best model in predicting the mean travel time. This case study provides opportunities in data preparation, descriptive and predictive analytic topics covered in applied machine learning, data science and decision support system courses using data mining programming environments like Python and R. Students are also able to change the study area (city) for this case study based on their interest.

Keywords: Ride-sourcing, Uber Movement, Travel time prediction, Machine Learning, Random forest, Huber robust regression, Python programming.

1. INTRODUCTION

Ride-sourcing industry faces a booming development in recent years, occupying quite share of vehicle usage on daily travelling in big cities. Uber and Lyft are two key players of ride-sourcing industry in the USA. In 2017, Uber and Lyft owned 54% and 37% of the United States ride-sourcing market, respectively (Certify, 2017). In 2017, Uber launched "Uber Movement", a website that employs Uber's riding data to help urban planners in improvising urban and traffic decisions (Gilbertson, 2017; Pearson etal., 2018). The Uber Movement website provides zone-to-zone travel time data (the arithmetic and geometric mean and standard deviations) of Census Tracts and Traffic Analysis Zones (TAZs) in cities such as Mania, London, Boston and Washington, D.C. A census Tract is a geographic region defined for the purpose of taking a census while TAZ is the unit of geography measure index commonly used in conventional transportation planning models. This data is open to the public and can be downloaded as the comma-separated values (CSV) format. With respect to the Uber movement data, this study was motivated to predict the mean travel time of Uber services and identify major factors (attributes) affecting that by using machine learning approach. In this approach, the required steps are: data preparation, data modeling, descriptive and predictive analytics. These steps are usually covered in applied machine learning, data science, and decision support system courses enabling students to practice and understand the effective ways using machine learning techniques to deliver accurate data-driven decisions.

As machine learning applications is fueling with full impetus, there is a popular and growing trend in business schools to launch machine learning and business analytics courses in their curriculum. These courses mainly focus on applied machine learning and data mining methods that address business questions. During the course, students usually learn machine learning algorithms including supervised and unsupervised techniques. Working on a case study helps students to better understand how to apply the machine learning approach in a real business world. However, small and fabricated data sets are usually used as teaching tools business analytics courses to teach machine learning (data mining) methods. While these teaching tools provide great opportunities for students to be exposed to different techniques, they may not provide a comprehensive experience for them to face with challenges of employing all steps of the machine learning approach in an integrated case study. Therefore, as a new teaching opportunity, the presented case study in this paper provides students with real data in order to practice the application of machine learning steps for a focused business problem. With this teaching tool, students face with challenges for data collection, processing and interpretation of findings occurred in the real business world. As the experience of authors, students' learning curve will be dramatically improved by having opportunities to deal with such challenges during the course under the supervision of the instructor. The next section will introduce the objective and implication of this case in details.

2. CASE STUDY OBJETIVES AND IMPLICATION

This case study employs the machine learning approach in order to predict the mean travel time of Uber services by considering the effects of distance between origin and destinations, railway/subway and street density, and daily weather conditions. For the purpose of this study, the origin of all Uber trips was set at Washington D.C. city center. Destination locations include holistic areas (zones) that Uber covers in D.C. The case study provides

opportunities for students to collect data from different web-based databases, prepare data for analytic evaluations, run supervised machine learning techniques, and designate major (key) factors (attributes) affecting the Uber travel time services by completing different prescriptive analytics. With respect to the identified attributes, students are able to train predictive machine learning methods to estimate the mean travel time of Uber services. According to the methodology developed for this case study, the accuracy of trained models is compared to select the best one. Note that Washington D.C is only used as an example here and other cities and different locations can be used depending on the students' interest. In this case study, the Python programming language is used to derive results. The case study can be adapted such that other data mining environments such R, SAS, and JMP PRO are used for data analysis.

The learning objectives of this case study are as follows:

- How to formulate a business question (problem) and consider the machine learning approach for solving such a problem.
- How to prepare a ready-to-analysis dataset for a study by integrating different datasets from different sources.
- How to determine (extract) potential features (factors) affecting Uber travel time services by employing an appropriate analytic algorithm while utilizing multiple corresponding libraries in the Python environment.
- How to visualize data using Python data visualization libraries such as "Seaborn" and "Matplotlib".
- How to identify the important (key) features (factors) impacting the Uber travel time using robust regression.
- How to build, tune and compare machine learning prediction models using "Scikit-learn" library.

The developed case study is originally designed for graduate (on-campus or online) courses in the area of applied machine learning (data mining) and data science in business schools. In overall, the case is flexible, and instructors are able to customize the content for the course requirements. For instance, instructors can separate analytical parts of the case to be used as a set of programing homework (or in class assignments) throughout the term. These separated analytical parts can be adapted based on the machine learning approach (Figure 1) letting students gradually practice data collection, data preparation, descriptive analytics, predictive analytics and model evaluations. The developed case study can also be customized as a topic for an independent study letting student apply different analytic and machine learning techniques and algorithms to predict the travel time in a ride-sourcing network in different cities under the supervision of the instructor. Section 3 provides the details of case structure in collecting and analyzing data. A basic knowledge of database management and a programing language is required to effectively incorporating this case study as a course project or homework assignment. Prior knowledge of a programming language is required to complete this case study. The Python programming language is taught in most business schools as an elective course and it is used in this case to explain steps followed in this case.

3. CASE STRUCTURE

The following chart presents the required steps in applying machine learning approach in this case study. It includes the following stages:

- Data collection and processing stage: In this stage data are collected from different web-based sources such as Uber movement and Climate Data Online (CDO) in National Ocean and Atmosphere Association (NOAA) websites. The collected data are prepared to be processed for selecting potential features (factors) affecting the mean travel time of Uber services. Feature selection are completed in the Python environment using "OpenStreetMap (OSM)", "Networkx", "Pandas/Geopandas", "Shapely", and "Fiona" libraries.
- Data modeling (descriptive analytics stage): This stage is completed for a better understanding of data content from numerical and statistical points of view. Libraries such as "Matplotlib", "Seaborn", "Pandas/Geopandas", "Scipy", and "Mapclassify" are used in this stage to visualize data and derive major descriptive analysis.
- Data modeling (predictive analytics stage): In this stage, a regression based predictive analysis is completed using Huber robust regression. These techniques enable the users to determine the key features impacting on the prediction of the mean travel time of Uber services. "Statsmodels" library in Python is used for setup and running such a regression-based analysis. Huber robust regression model developed according to sequential forward feature selection methodology and random forest are compared for estimating the travel time and also their prediction accuracy. The parameters of the models are tuned separately using "GridsearchCV" library for reaching the highest prediction accuracy. "Mlxtend" and "Scikit-learn" libraries are used in the Python environment for setting up and running these machine learning models. The above stages and corresponding results are explained in the following sections (As indicated in Figure 1 Appendix).

3.1. Data Collection

The required data used to measure the shortest path distance for a given origin and destination as well as the road network in Washington, D.C. was collected from the Uber Movement website in regards to Washington, D.C. Figure 2 demonstrates the Traffic Analysis Zone and Census Tracts coverage areas collected from Uber Movement websites. Several studies have shown the significant impact of weather on the travel time in different transportation networks (Brodeur & Nield, 2017; Sina Shokoohyar, Sobhani, & Sobhani, 2019). To incorporate the impact of weather on the travel time, the daily weather condition data was collected from Climate Data Online (CDO) in the National Ocean and Atmosphere Association (NOAA) website. For the purpose of this paper, this data covers all weather-related information for January, February and March 2018.

--As indicated in Figure 2 (Appendix)--

Table 1 shows two samples of data downloaded from the Uber Movement website. This table shows "mean travel time", "geometric" and "geographic" of a given Uber service in Washington D.C. The first data sample includes the IDs of origin and destination, date of the service, the corresponding mean travel time and its upper- and lower-time boundaries. The second sample includes "Destination census Tracts (area)" data for an Uber trip from Washington D.C., as well as the trip ID and the destination geometry polygons. The geometry polygons are

nodes that specify the border of a given census tract (census zone). In this case study, the origins of all trips are set in the city center of Washington D.C. Note that the case study can be modified by considering any other census tracts as the origin.
--As indicated in Table 1 (Appendix)--

Table 2 presents the samples of weather data collected from the NOAA website. Daily Precipitation (PRCP) and average daily temperature (TAVG) are weather related data used in this case study. For instance, on January 4th, 2018, there was 0.1 inches precipitation on the day and the daily temperature was 26 Fahrenheit degrees on average.
--As indicated in Table 2 (Appendix)--

Linking weather data to Uber trips enables the users to have weather information for the date a given trip was taken place.

3.2. Data Processing and Preparation

The shortest path distances of Uber trips are estimated with respect to the corresponding geometry polygons of origins and destinations. The road network information used for estimating such distances is getting accessed by using "OpenStreetMap (OSM)" and "Networkx" libraries in the Python environment. These libraries are also used in this case study to determine the density of street and railway/subway stations in different Uber destination areas. OSM is built in 2004 by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations and etc. It is a free access library which is compiled with Python and R. "NetworkX" is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks (NetworkX, 2019). The following syntax is used to compute the shortest path between a given 2 nodes:

NetworkX.shortest_path() (1)

As discussed above, the Uber movement website only collects the geometry polygons of origins and destinations (borders of census tracts) of trips. Since the road network in each zone has multiple intersections, the intersections with the max betweenness centrality (Max BC node) in the origin and destination are selected as the start and the end point of a path to calculate the shortest path distance. Betweenness centrality measures the extent to which a node lies on the paths between other nodes. Betweenness centrality has been used extensively in the literature on transportation research, natural language processing and graph theory (Brandes, 2008; Newman, 2005; Pearson et al., 2018; Shokoohyar, 2018; Shokoohyar, 2019; Sobhani et al., 2019). The specified origin and destination points can be used to generate the corresponding shortest path by using the "NetworkX.shortest_path()" syntax.

Figure 3 demonstrates the road network, shortest path and BC nodes (origin and destination points) for a given Uber trip. The travel was started from the city center. The ID of the city center census tract is 186. The destination census tract ID is 2. The corresponding destination point (Max BC node) is shown in the middle picture. Using "NetworkX.shortest_path()" syntax with respect to the origin and destination points results the corresponding shortest path (right side picture in Figure 3) . The output of running the shortest path syntax gives us the distance of 10226.499 meters for the Uber trip between the city center census tract (ID 196) and the census tract with an ID of 2.
--As indicated in Figure 3 (Appendix)--

The density of street, railway and subway in census tracts (areas) of destinations are estimated by applying the following syntaxes:

```
OSM.stats.basic_stats ()                                                    (1)
OSM.core.graph_from_polygon (infrastructure='way["railway"~"subway"]')      (2)
```

The street density is measured by the total length of street divided by area in square kilometer. For example, the street density ('street_density_km') of the ID 2 destination is 15488.186421599354 meters per square kilometer. The density of the railway/subway density is measured as the total number of stations in a destination census tract divided by the area in square kilometer. For example, the destination ID of 7 (virtual address: 4500 Ohio Drive Southwest, Southwest Washington, Washington) has 1 railway/subway station with areas of 15.799 square kilometers. So, the corresponding railway/subway density is 0.06329 per square kilometer.

3.3. Data Modeling (Descriptive Analytics Stage)

In this section, first, the descriptive statistics of the prediction variables (features) used in this study is presented. Then, the relations (correlations) between prediction variables (features) and the target variable (mean travel time) are explored by visualizing corresponding data. Table 3 presents the descriptive statistics of major prediction variables used in this case study.
--As indicated in Table 3 (Appendix)--

To better understand the reasons for selecting the potential variables (features) in predicting the mean travel time, the corresponding data is visualized to explore the correlations between the potential variables (features) and the travel time of Uber services. Figures 4 to 7 demonstrate the examples of such data visualizations. Figure 4 shows the correlation between the mean travel time and the shortest path of Uber trips. The data is distributed in a CORN shape, indicating that the mean travel time intends to increase while the travel distances become longer. This correlation may be because more potential barriers would appear in a long path, such as traffic lights, traffic jam and detour.
--As indicated in Figure 4 (Appendix)--

Figure 5 shows the correlation between the mean travel time with street and railway/subway densities. Findings show that street density (at destination census areas) has a negative association with the mean travel time of Uber services. That is, the average travel time decreases in denser street areas. This is mainly due to a better (more easily) accessibility for the Uber drivers to arrive to the destination points. However, the existence of the railway/subway stations around destination points does not have a significant impact on the average travel time of Uber trips (Right-side graph in Figure 5).
--As indicated in Figure 5 (Appendix)--

Figure 6 shows the correlations between Uber trips during weekdays or weekends/holidays and the mean travel time. The results conclude that there is a significant difference between the average travel times during weekdays and weekends (Independent T-test for the mean travel time between the workday and weekend/holiday: p-Value=0.0000). The mean travel time in workdays is 1358.64 seconds on average; 153.48 seconds more than the one on weekends/holidays. In addition, as shown by the right-side graph in Figure 6, the distribution of the mean travel time in workdays covers a wider range in comparing with the one on weekends/holidays. These findings imply that the traffic congestion in workdays has significant impact on the Uber mean travel time.
--As indicated in Figure 6 (Appendix)—

--As indicated in Figure 7 (Appendix)--

Figure 7 presents the heat map plots of the mean travel time during workday and weekends/holidays. The x-axis is the longitude of the area while the y-axis is the corresponding latitude. The plots map the Uber average travel time in different destinations by changing the density of blue color in the plot. Lighter blues relate to destinations where the average travel time is shorter. Darker blue areas belong to destinations where the average travel time from Washing D.C. city center to those destination areas are longer. Comparing the heat map plots demonstrates that workday also has obvious impacts on the mean travel time of the center areas of Washington D.C. Furthermore, trips to Western, Eastern and Southeastern D.C. only have higher the mean travel times during workdays in compared with ones during weekends/holidays. This indicates that the most residents in these locations, usually commute to central D.C. to work in workdays but hardly go to the central area on weekends/holidays.

Figure 8 pictures the relation between the mean travel times of Uber services with changes on weather condition indexes (PRCP and TAVG) during both weekdays and weekends/holidays. The findings show slight positive correlations between the travel time and both precipitation and average weather temperature during work days (red line in both graphs). TAVG also has an obvious positive association with the mean travel time during weekends/holidays.

--As indicated in Figure 8 (Appendix)--

Figure 9 shows the mean travel time during different days of a week. In overall, the mean travel time has a concave-downward pattern from Monday to Sunday in a week (Figure 9). Changing the date from Monday to Thursday, the mean travel time maintains a steady level. After that the mean travel time declines between Friday and Sunday. Therefore, the mean travel time has an obvious variation on each workday and weekend. Thus, days of the week are added in prediction models to explore their impact.

--As indicated in Figure 9 (Appendix)--

Figure 10 presents that January, February and March have specific patterns on the mean travel time, all of which are distinct enough to add them as variables in the machine learning models to explore their effects.

--As indicated in Figure 10 (Appendix)--

3.4. Data Modeling (Predictive Analytics Stage)

In this section, first, the machine learning models are developed to predict travel times based on the features described in section 3.3. In particular, shortest path distance, rail/subway density, street density, precipitation and temperature average are considered as explanatory variables. Then, the results derived from these methods are presented and evaluated.

3.4.1. Machine Learning Models

Shortest path travel distance, street density and railway/subway stations density, precipitation, average daily temperature, date of travels including corresponding months and days of the week are considered as potential factors (attributes) affecting the mean travel times of Uber services. Equation (4) and (5) demonstrate the corresponding Hubert Robust regression models to predict the mean travel time with respect to the above attributes in workdays and non-working days. Two models are developed as travel behaviors in working days and non-working days are shown to be different in several studies (Koetse & Rietveld, 2009; Stover & McCormack, 2012; Cools & Creemers, 2013; Singhal et al., 2014; Sobhani & Wahab, 2017; Shokoohyar et al., 2019). Therefore, the following regression models consider the effects of the attributes on the mean travel time of Uber services for weekdays and weekend travels separately.

$$Y_{workday} = \beta_0 + \beta_1 ShorPth + \beta_2 DSTRailDen + \beta_3 DSTStrDen + \beta_4 PRCP + \beta_5 TAVG + \beta_6 Month\_Feb + \beta_7 Month\_Mar + \beta_8 DoW\_Mon + \beta_8 DoW\_Tue + \beta_9 DoW\_Wed + \beta_{10} DoW\_Tur + \beta_{11} DoW\_Fri + \epsilon \quad (3)$$

$$Y_{weekend} = \beta_0 + \beta_1 ShorPth + \beta_2 DSTRailDen + \beta_3 DSTStrDen + \beta_4 PRCP + \beta_5 TAVG + \beta_6 Month\_Feb + \beta_7 Month\_Mar + \beta_8 DoW\_Sat + \beta_9 DoW\_Sun + \beta_{10} DoW\_Holiday + \epsilon \quad (4)$$

The Huber Robust method is applied to estimate the coefficients. R squared metric is used for evaluating the fitness (accuracy) of the models (Seabold & Perktold, 2010). Variables (features) with statistically significant coefficients are selected as key factors to predict the mean travel time.

To enhance the prediction accuracy of the Uber travel time and also identify the key features, a Sequential Feature Selection algorithm is considered to develop the regression models. Sequential Feature Selection algorithms are a family of greedy search algorithms that are used to reduce an initial d-dimensional feature space to a k-dimensional feature subspace (k<d) (Raschka, 2014). Their objective is to automatically select a subset of features that is most relevant to the model. For the sake of this case study, Sequential Forward Selection (SFS) method is used to select a subset of features. This method initializes the Huber Robust regression model with an empty set and adds an additional feature that maximize the performance of the criterion function at each step until the termination criterion is satisfied (Sina Shokoohyar, Sobhani, & Ramezanpour Nargesi, 2020). The following Python syntax is used in this case study to develop the regression models introduced by Equation (4) and (5).

```
mlxtend.feature_selection.SequentialFeatureSelector(forward=True, floating=False)
```
(5)

At the end, Huber Robust regression models are compared with Random Forest (RF) using “Scikit” library to find the best model (the most accurate one) predicting the Uber mean travel time with respect to the given features. This section provides great opportunities for students to getting familiar with steps in setting up machine learning methods.

3.4.2. Predictive Model results (Huber Robust Linear Regression)

Summary of the Huber Robust Linear Regression (HRLM) models for predicting the mean travel time of Uber services during the workday and weekend/holiday are presented in Table 4. R-squared is presented for each model by using "Sklearn.metrics" syntax. It is approximately 0.457 in workday model and 0.435 in the weekend/holiday model. Variance Inflation Factor (VIF) is a statistical parameter that measures the collinearity of variables. It is less than 2 for all variables in both workday and weekends/holidays models indicating that the variables in the models are well independent and the models are not affected by collinearity. (As indicated in Table 4 Appendix)

The results in Table 4 provide five major highlights: First, precipitation and the density of railway/subway stations in destination census tracts do not have statistical significant relationships with the mean travel time (p-Values are greater than 0.05) during weekdays and weekends. Second, the coefficients for Shortest Path in the models are positive and significant, indicating that the longer a shortest path is, the more time an Uber driver takes to travel. Third, the negative and significant coefficients of DSTStrDen (density of streets) imply that the well-developed road network in the destination areas helps Uber drivers to shorten their travel times. By increasing the street density, there are more alternative routes for Uber drivers to choose to arrive destinations. Four, the average weather temperature (TAVG) prolongs the mean travel time of Uber services during workdays. Fifth, Tuesday and Thursday have more positive impact on the mean travel time implying that business activities are usually arranged on Tuesday to Thursday, so the traffic congestion on Tuesday to Thursday is heavier, increasing the times of Uber services.

Table 5 also concludes three main points from the weekend/holiday model. First, DSTStrDen (street density) and Shortest Path have the same coefficient sign in the weekend/holiday model, showing that they have similar influence on the mean travel time of Uber services. Second, the positive and significant coefficient of TAVG implies that the travel time increases by increasing the weather temperature. Third, the positive and significant coefficients of February and March implies that in the spring season, the better weather condition leads to a higher travel demand which in turn increases traffic congestion. The traffic congestion in turn will lead to a longer travel time.

3.4.3. Predictive Model Results (Comparing HRLM and Random Forest)

Huber Robust Linear Regression (HRLM) and Random Forest (RF) models were set up according to the following steps:

1. Identifying the most relevant features:
   - HRML Method: Sequential Forward Selection + sklearn.linear_model.HuberRegressor
   - RF Method: Sequential Forward Selection + sklearn.ensemble.RandomForestRegressor
2. Tuning the key parameters, based on the step 1 models using the GridsearchCV method

First, Sequential Forward Selection (SFS) is applied to develop models with the most relevant features using for travel time prediction. Second, "GridsearchCV" syntax is applied to complete the cross validation approach that utilizes and tunes the parameters of the machine learning models such that the highest accuracy of the travel time prediction is reached for each of the models. Finally, the developed HRML and Random Forest models are compared in terms of accuracy. Cross validation is a resampling procedure employed to assess and compare machine

learning models in terms of their prediction/classification accuracy. In general cross validation approach applies the following steps to reach the most accurate machine learning models:

1. Randomly shuffle the dataset used for the prediction.
2. Split the dataset into k subsets, called k-fold cross validation.
3. Take a subset of data as test set
4. Take the remaining subsets as a training data set
5. Fit the machine learning model on the training set and evaluate its accuracy on the test set
6. Retain the evaluation score, Mean Square Error in this case study
7. Take another subset of data from k subsets as a test set and repeat steps 4 -7.
8. At the end, the model with the highest accuracy is selected as the best machine learning model.

With regard to HRLM models, Figure 11 shows that including 3 features are leading to the highest accuracy in predicting the mean travel time of Uber services on weekdays. Note that Figure 11 presents the negative mean squared error, and therefore a higher negative mean squared error is preferred. In the non-working-days model, 5 variables should be included for the best accuracy. Table 5 demonstrates those key variables (features) for both HRLM models (i.e. working days and non-working-days models).

--As indicated in Figure 11 (Appendix)--

--As indicated in Table 5 (Appendix)--

With the same approach, Figure 12 and Table 6 demonstrate the results of random forest regression in predicting the average travel time of Uber services for weekday (weekday) and weekend trips after tuning the models to reach the maximum prediction accuracy. For the workday model, 8 key features (DSTStrDen, DSTRailDen, ShorPth and 5 days of the week) are added in the model to reach the highest accuracy, highest MSE (Neg.MSE = -0.0083), since the MSE decreases when features are more than 8 in the model.

--As indicated in Figure 12 (Appendix)--

In the weekend/holiday random forest model, 6 key features (DSTStrDen, DSTRailDen, ShorPth, Month_Jan, Month_Feb, Month_Mar) are added to the model to reach the highest accuracy (Neg.MSE = -0.0054). With more than 6 features, the negative MSE goes downward.

--As indicated in Table 6 (Appendix)--

Table 7 summarizes the prediction performance of HRML and random forest regression models. For workdays, Random Forest performs better in comparison with Huber Robust Regression. The Random Forest model has a much smaller prediction error (Neg.MSE: -0.0081) than the Huber Robust Regression model (-0.0909) due to its higher predicted negative MSE. Second, Random Forest contains more useful and effective information (more key variables) in comparing with Huber Robust Regression. Similarly, the Random Forest model performs better in weekend/holiday prediction, since the Random Forest has more precise prediction with lower predicted MSE (-0.0053) and contains more useful and effective features.

--As indicated in Table 7 (Appendix)--

4. IMPLEMENTATION GUIDANCE

The case study presented in this paper gives a real-life example for graduate students to learn and practice how to identify and address a business question by applying machine learning methods. Specifically, first, this case study enhances the students' capability of building hypothesis of a prediction model, based on different prediction variables. Building hypothesis is the start-line of all quantitative models in either business world or academic research. Second, the case improves their proficiency in Python programming and applying related machine learning and data visualization libraries. In particular, several important Python libraries such as "Pandas/Geopandas", "Matplotlib", "Seaborn", "Networkx", and "OpenStreetMap" are used in this case study.

This case study is used both as a course project for graduate level machine learning course and independent study project. It can be used in two phases after the required topics are covered following the flow chart in Figure 1. After covering basic data visualization and descriptive statistics libraries, the case can be introduced and students would be asked to present the first course project report. The report should include the dataset collected, processed, visualized, and a short interpretations of observations. Following this phase, a class discussion helps students to discuss their findings. Following questions can be considered for the class discussion:

- Does travel time depend on the distance traveled? What is the best way to calculate the distance between an origin and a destination? Note that there are several approaches in calculating the distance between an origin and a destination. The one used in this study is the most popular one (i.e. shortest path), and other methods such as Haversine distance can be used similarly.
- How does weather impact travel time?

In the second phase, the data modeling phase, the prediction model should be developed and important features should be selected. Following this phase, an oral project presentation can be scheduled. In the current case study, we have shown the findings using Huber Robust and Random Forest regression models. Students may be asked to at least use a third model (e.g. Support Vector Machines or Neural Network) and compare their results. During the project presentation, the advantages and disadvantages of using these additional models can be discussed.

REFERENCES

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, *30*(2), 136–145.

Brodeur, A., & Nield, K. (2017). *Has uber made it easier to get a ride in the rain?*

Certify. (2017). Uber Declines, Lyft Picks Up in the Certify SpendSmart™ Report for Q3 2017. Retrieved from Certify website: https://www.certify.com/2017-10-24-Uber-Declines-Lyft-Picks-Up-in-the-Certify-SpendSmart-Report-for-Q3-2017

Cools, M., & Creemers, L. (2013). The dual role of weather forecasts on changes in activity-travel behavior. *Journal of Transport Geography*, *28*, 167–175.

Gilbertson, J. (2017). Introducing Uber Movement | Uber Newsroom US.

Koetse, M. J., & Rietveld, P. (2009). The impact of climate change and weather on transport: An overview of empirical findings. *Transportation Research Part D: Transport and Environment*, *14*(3), 205–221.

NetworkX. (2019). NetworkX.

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, *27*(1), 39–54.

Pearson, M., Sagastuy, J., & Samaniego, S. (2018). Traffic Flow Analysis Using Uber Movement Data. *Palo Alto, CA: Stanford University), Available at Http://Web. Stanford. Edu/Class/Cs224w/Projects/Cs224w-11-Final. Pdf (Last Accessed March 2018)*.

Raschka, S. (2014). Sequential Feature Selector.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Shokoohyar, S. (2018). Ride-sharing platforms from drivers' perspective: Evidence from Uber and Lyft drivers. *International Journal of Data and Network Science*, *2*(4), 89–98.

Shokoohyar, Sina. (2019). Determinants of Backpackers Perceptions of Security? A WOM-based Approach. *E-Review of Tourism Research*, *16*(4).

Shokoohyar, Sina, Sobhani, A., & Ramezanpour Nargesi, S. R. (2020). On the Determinants of Uber Accessibility and Its Spatial Distribution: Evidence from Uber in Philadelphia. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge discovery*.

Shokoohyar, Sina, Sobhani, A., & Sobhani, A. (2019). Impacts of trip characteristics and weather condition on ride-sourcing network: evidence from Uber and Lyft. *Research in Transportation Economics*.

Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice*, *69*, 379–391.

Sobhani, A. & Wahab, M.I.M. (2017) The effect of working environment-ill health aspects on the carbon emission level of a manufacturing system, *Computers & Industrial Engineering*, 113, 75-90.

Sobhani, A., Wahab, M.I.M., & Jaber, M.Y. (2019) The effect of working environment aspects on a vendor–buyer inventory model. *International Journal of Production Economics*, 208, 171-183.

Stover, V. W., & McCormack, E. D. (2012). The impact of weather on bus ridership in Pierce County, Washington. *Journal of Public Transportation*, *15*(1), 6.

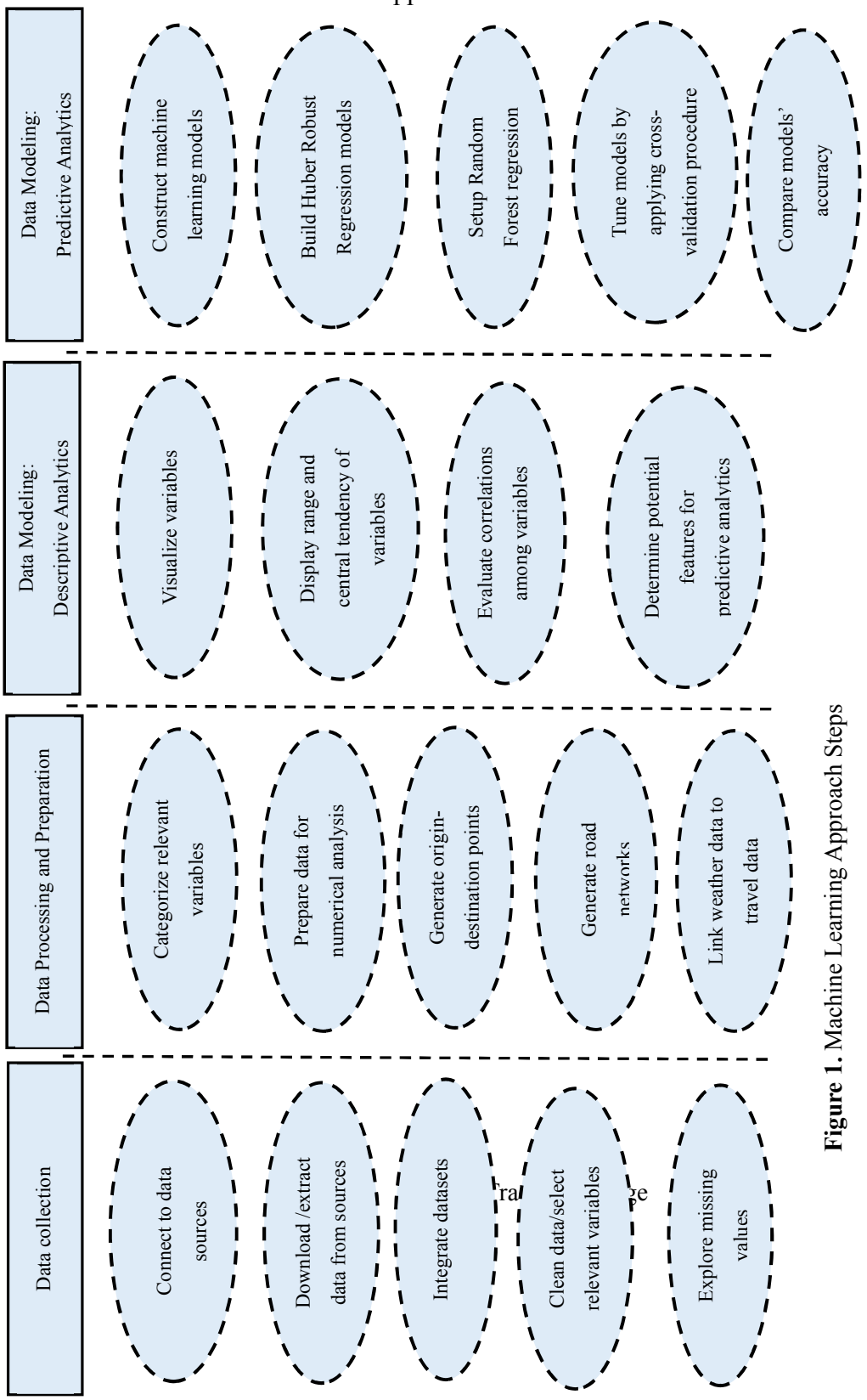


Figure 1. Machine Learning Approach Steps

Traffic Analysis Zones Coverage

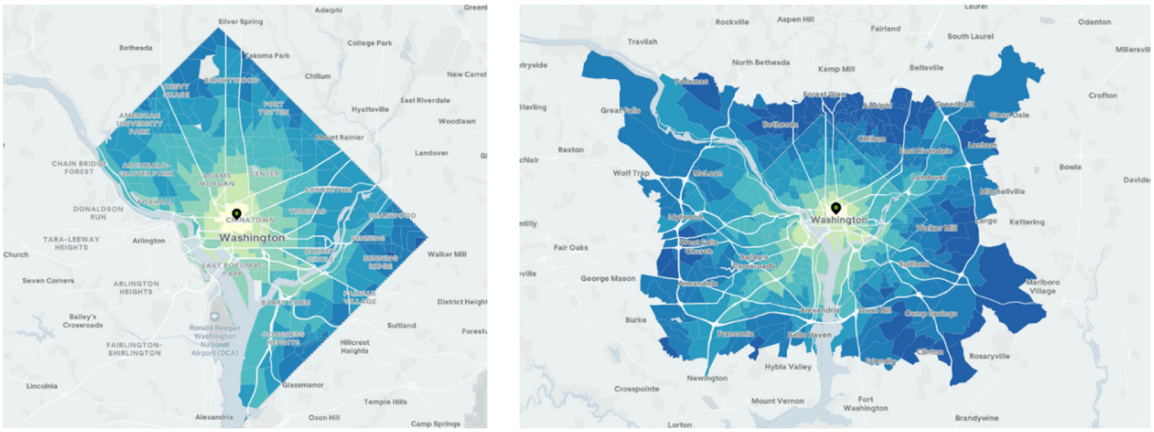


Figure 2. Data Coverage of Census Tracts and Traffic Analysis Zones (TAZs) in Washington, DC

Origin Zone (ID 186) Destination Zone (ID 2) The Shortest Path From ID 186 to ID 2

Figure 3. The Max BC Node in the Origin ID of 186 and the Destination ID of 2 and its Shortest Path

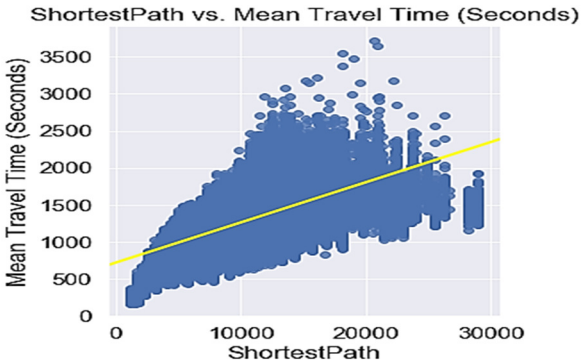


Figure 4. Association of Mean Travel Time with Shortest Path between Source and Destination

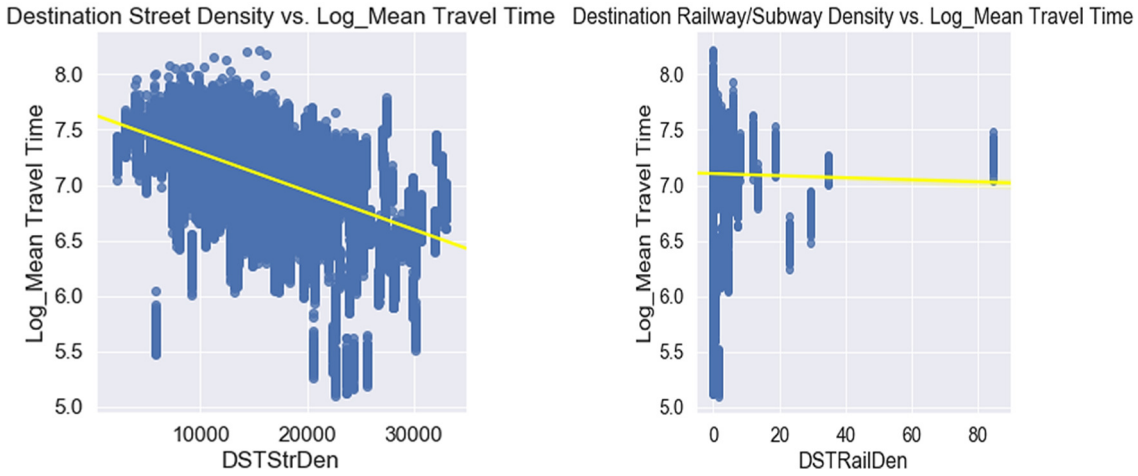


Figure 5. Association of Mean Travel Time with Destination Street and Railway/Subway Density

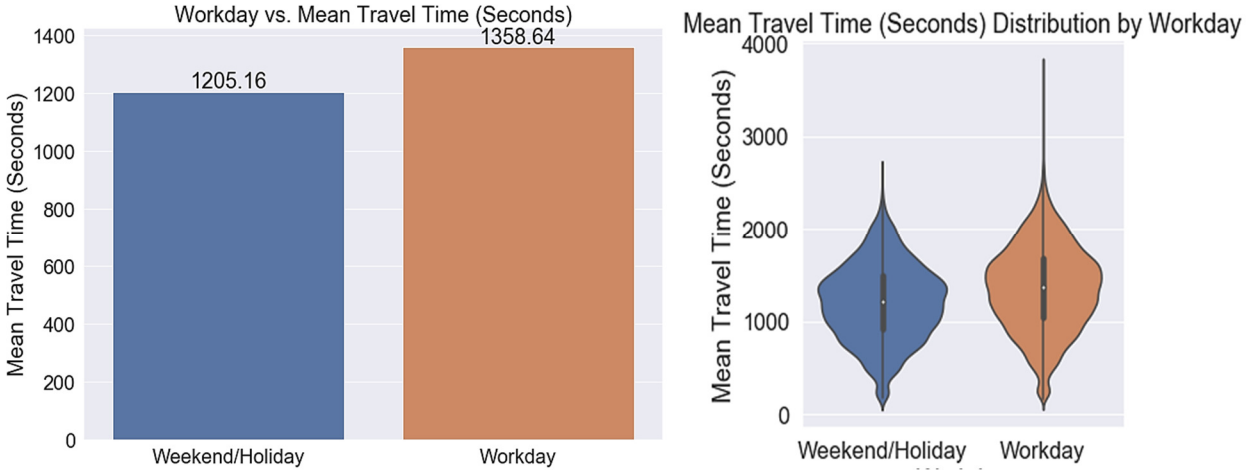


Figure 6. Mean Travel Time and Distribution During Weekdays

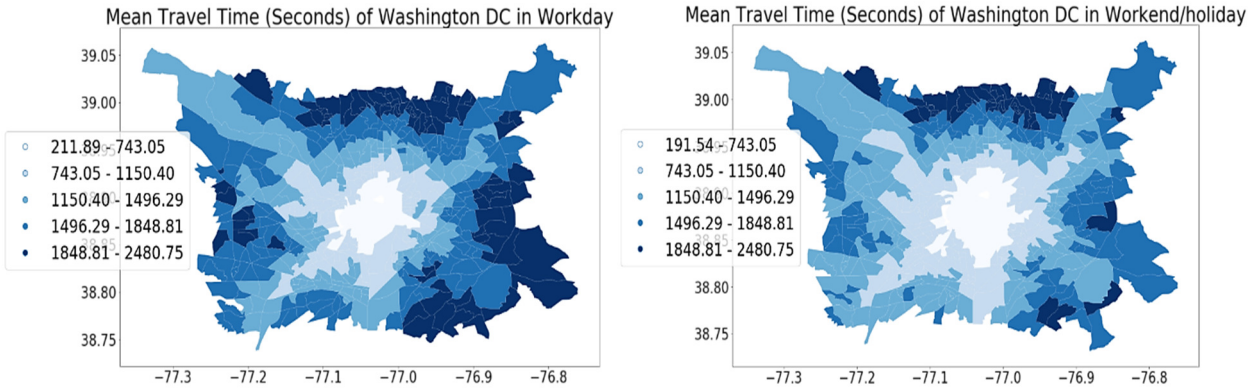


Figure 7. Mean Travel Time Heat-map by Workday and Weekend

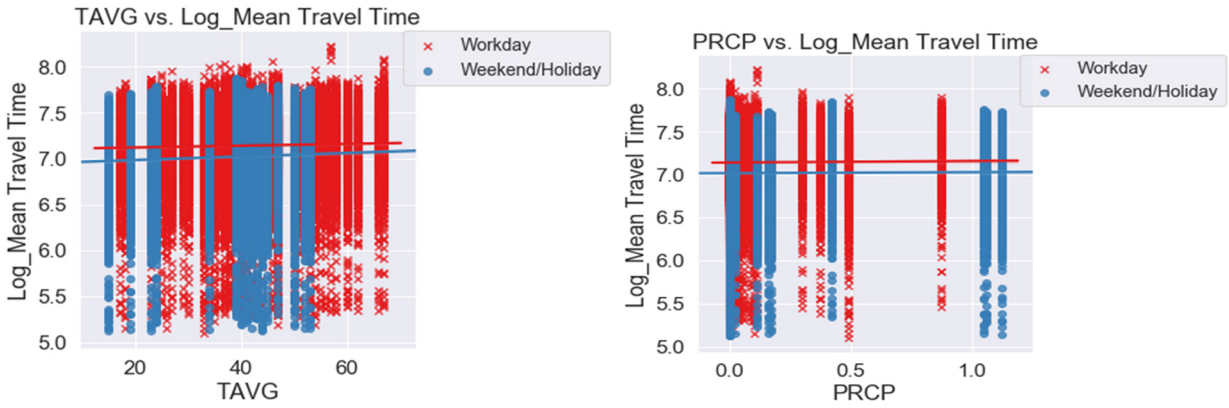


Figure 8. Association of Mean Travel Time with Different Weather Types During Weekdays

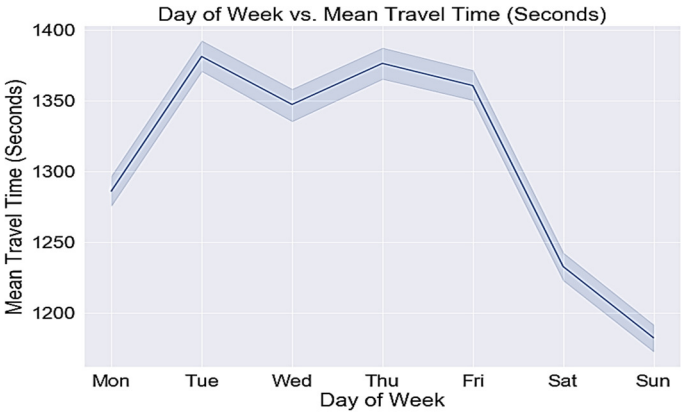
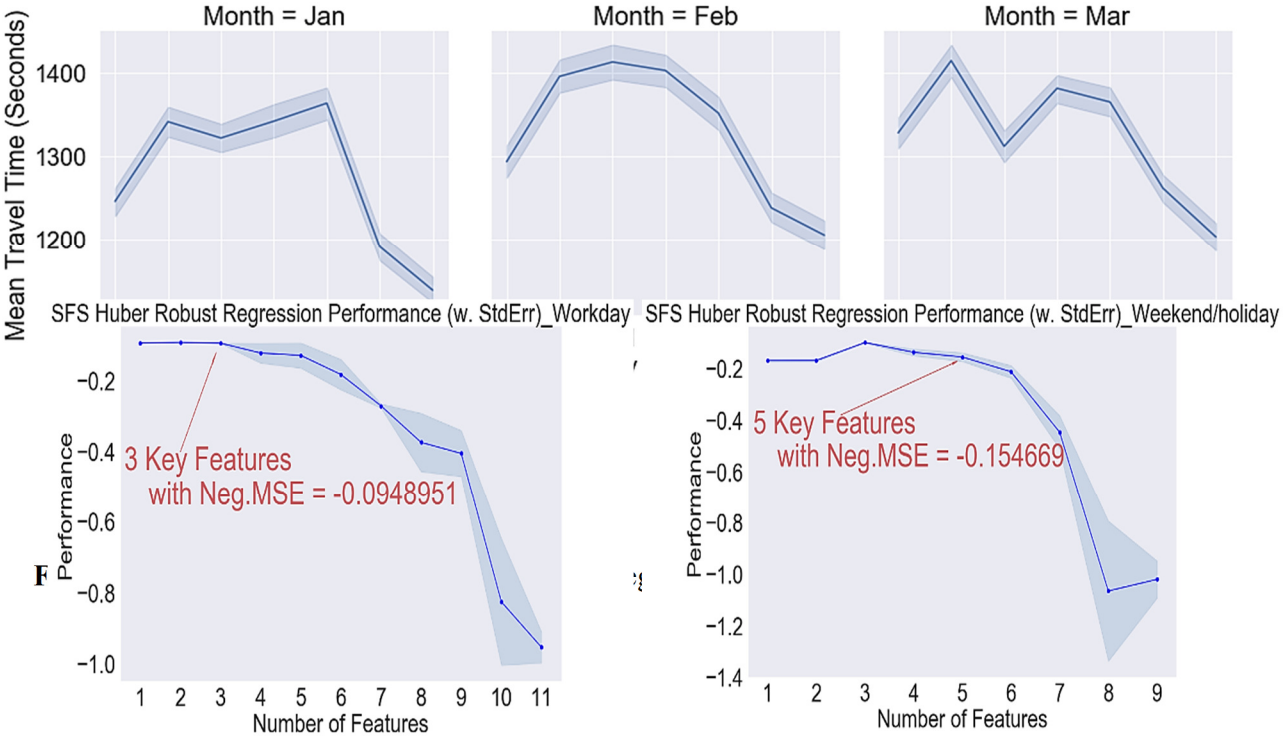


Figure 9. Mean Travel Time Changes during Weekdays



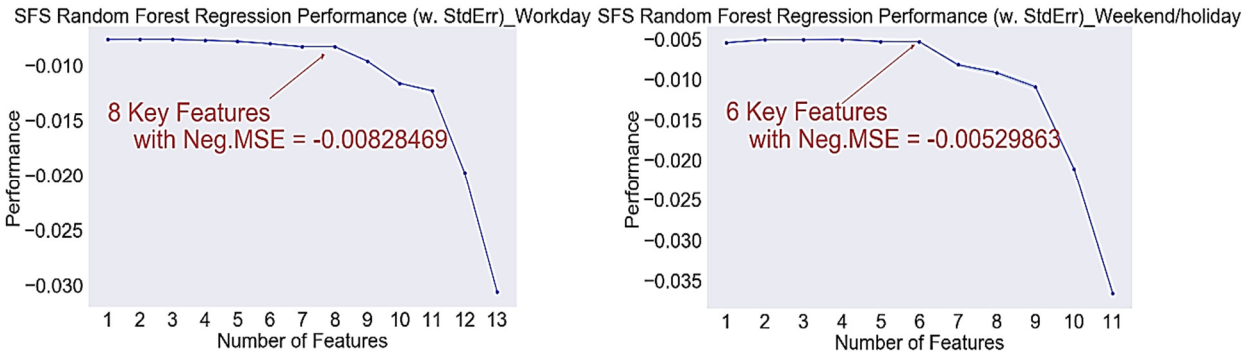


Figure 12. Negative MSE in SFS Random Forest Regression Outcome

Table 1. Sample Dtasets Downloaded from Ubermovement Website

Sample 1

Origin Movement ID	Destination Movement ID	Date Range	Mean Travel Time (Seconds)	Range - Lower Bound Travel Time (Seconds)	Range - Upper Bound Travel Time (Seconds)
186	1	12/31/2017 - 12/31/2017, Every day, Daily Average	652	493	860
186	2	12/31/2017 - 12/31/2017, Every day, Daily Average	1560	1273	1911
186	3	12/31/2017 - 12/31/2017, Every day, Daily Average	1241	972	1583
186	4	12/31/2017 - 12/31/2017, Every day, Daily Average	1685	1475	1924

Sample 2

MOVEMENT_ID	DISPLAY_NAME	Destination Geometry Data
1	5400 Arnold Avenue Southwest, Southwest Washin...	(POLYGON ((-77.04800899999999 38.841266, -77.0...
2	1400 Juniper Street Northwest, Northwest Washi...	(POLYGON ((-77.05162300000001 38.987145, -77.0...
3	4800 Canal Road Northwest, Northwest Washingto...	(POLYGON ((-77.11975099999999 38.93435, -77.11...
4	2700 Unicorn Lane Northwest, Northwest Washing...	(POLYGON ((-77.071647 38.971786, -77.071250000...
5	4500 Q Place Northwest, Northwest Washington, ...	(POLYGON ((-77.100859 38.911209, -77.099577 38..

Table 2. Sample Datasets Downloaded from NOAA

DATE	(Daily Precipitation)PRCP	(Average daily temperature) TAVG
1/1/2018	0	19
1/2/2018	0	18
1/3/2018	0	23
1/4/2018	0.1	26
1/5/2018	0	17

Table 3. Variables and descriptive statistics

Variable	Description	Min	Mean	Max	Std. Error
PRCP	Precipitation (inches)	0	0.085529	1.12	0.22804
TAVG	The average of MAX and TMIN temperature for the day	15	40.8008	67	10.8147
DSTStrDe	The total street length divided by area (square kilometres) of Destination area	2186.370	15359.020	32944.964	5782.5176
DSTRailDen	The total railway/subway station nodes divided by area (square kilometres) of Destination area	0	0.622992	84.605391	4.522131
ShorPth	The shortest-path route of the node with max betweenness_centrality from origin and destination (meters)	1138.114	10886.88	28964.231	5316.7340

Table 4. HRLM Regression Results (Dependent Variables: Log-Transformed Mean Travel Time)

	Workday model				Weekend/holiday model			
Variables	Coef.	Std. err.	*p*-Value	VIF	Coef.	Std. err.	*p*-Value	VIF
const	6.8327	0.011	0	52.05	6.6625	0.017	0	47.079
PRCP	0.0089	0.011	0.421	1.101	0.0053	0.009	0.573	1.694
TAVG	0.0005	0	0.002	1.2	0.001	0	0	1.333
DSTStrDen	-1.02E-05	3.41E-07	0	1.608	-8.84E-06	5.36E-07	0	1.627
DSTRailDen	-0.0003	0	0.316	1.006	5.19E-05	0.001	0.924	1.006
ShorPth	3.96E-05	3.72E-07	0	1.602	3.79E-05	5.75E-07	0	1.621
DoW_Tue	0.0456	0.005	0	1.924				
DoW_Wed	0.0195	0.005	0	1.915				
DoW_Thu	0.0361	0.005	0	1.837				
DoW_Fri	0.0206	0.005	0	1.847				
DoW_HDY					0.0383	0.009	0	1.325
DoW_Sat					0.0434	0.005	0	1.13
Month_Feb	0.0234	0.004	0	1.543	0.0378	0.007	0	1.968
Month_Mar	0.0147	0.004	0	1.487	0.0536	0.007	0	1.59
No. Obs.	0.4567				No. Obs.	14924		
R squared	32,279				R squared	0.4350		

Table 5. Best SFS Robust Regression Model After Completing Cross Validation

Workday Regression Model				Weekend/holiday Regression Model			
# of Key Features	Feature Name	GridCV_Score (neg. MSE)	Predicted MSE	# of Key Features	Feature Name	GridCV_Score (neg. MSE)	Predicted MSE
3	PRCP DSTStrDen ShorPth	-0.0924	0.0909	5	PRCP ShorPth DoW_Sat Month_Feb Month_Mar	-0.09798	0.0982

Table 6. Best Random Forest Regression Models After Tuning

Workday Random Forest Model				Weekend/holiday Random Forest Model			
# of Key Features	Feature Name	GridCV_Score (neg. MSE)	Predicted MSE	# of Key Features	Feature Name	GridCV_Score (neg. MSE)	Predicted MSE
8	DSTStrDen DSTRailDen ShorPth DoW_Mon DoW_Tue DoW_Wed DoW_Thu DoW_Fri	-0.0082	0.0081	6	DSTStrDen DSTRailDen ShorPth Month_Jan Month_Feb Month_Mar	-0.0052	0.0053

Table 7. Prediction Comparison

	Model	# of Key Features	GridCV_Score (neg. MSE)	Pred. MSE
Workday	Huber Robust Regression	3	-0.0924	0.0909
	Radom Forest	8	-0.0082	0.0081
Weekday/holiday	Huber Robust Regression	5	-0.0979	0.0982
	Radom Forest	6	-0.0052	0.0053