National College of Ireland

Taxi Trip Time and Trajectory Prediction Using Machine Learning

MSc Research Project
MSc in Data Analytics

Janvi Rajesh Rajani
Student ID: X20148712

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Janvi Rajesh Rajani
Student ID:	X20148712
Programme:	MSc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	31/01/2022
Project Title:	Taxi Trip Time and Trajectory Prediction Using Machine Learning
Word Count:	6902
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Janvi Rajesh Rajani
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Taxi Trip Time and Trajectory Prediction Using Machine Learning

Janvi Rajesh Rajani
X20148712

Abstract

Predicting route time is critical for both commercial traffic and trip planning. When making judgments, riders will benefit from accurate journey time estimations. As a result, traffic conditions will be improved. Machine learning methods are used to estimate cab travel time and trajectory in this research. Baseline Model, Decision Tree Regression, Lasso Regression, Random Forest Regression, XGBoost Regression, and KNN Regression are used to forecast cab travel time. Multiple Linear Regression and Gradient Boosting Regression are used to forecast cab journey trajectory. The cab ID, timestamp, call type, GPS co-ordinates, and Day type are in the dataset is used for this research. Following that, the performance metrices are used to analyze and compare the outputs of these algorithms. The model with the highest accuracy and lowest error is chosen. Following the evaluation of the models, it was discovered that Lasso regression having MAE=1.01 beats other models in predicting taxi trip prediction, while Gradient Boosting regression having MAE=0.008 outperforms Multiple Linear Regression. The models are also compared to state-of-the-art approaches.

Keywords — taxi time prediction, trajectory prediction, regression, Machine Learning, Lasso Regression, Gradient Boosting Regression

1 Introduction

People in big cities have a stressful lifestyle and attempt to manage their time well. The majority of people use taxis to get to and from various locations, such as the office, home, a short journey, or supermarkets. Taxi trip time and trajectory[1] are critical to anticipate, mainly in metro regions, so that riders can plan their journey effectively. In order to better plan and arrange the journey, it is critical to forecast the trip time ahead of time. Most taxis nowadays are GPS-enabled, allowing it to extract the taxi's GPS routes and anticipate real-time trajectories. It is simple to maintain track of any moving object using GPS sensors attached to mobile devices. The duration of a taxi ride and its trajectory are determined by a variety of factors, including real-time traffic conditions, the day of the week (weekday/ weekend), and historic GPS co-ordinate data.
Many essential professional meetings or family occasions are missed or delayed because people are unable to arrive on time. This is due to their inability to precisely forecast travel time from point A to point B. Taxi time predictions are based on a variety of circumstances, such as access to small lanes that only taxis have or any short cuts that only

[1]The next possible path a driver can take.

the taxi driver knows about. Passengers usually want to know when they will arrive at their destination and the route the driver will choose. Riders and drivers need accurate trip time predictions in order to make optimal travel decisions. It can be accomplished, after reviewing past data. The data should include previous ride travel record, as well as the pick-up and drop-off locations for each trip. This research also aims to predict the trajectory, therefore previous information about the path's co-ordinate is required. Taxi trajectory prediction is required to anticipate potential road hazards, trip delays, and traffic congestion.
Multiple machine learning regression algorithms is used to forecast taxi time and trajectory. With little human assistance, machine learning algorithms can spot trends in past data and make comprehensive choices. It requires GPS co-ordinates of the course at predetermined intervals. All taxis are equipped with GPS sensors that allow the co-ordinates to be recorded at regular intervals. The order in which location co-ordinates are gathered at regular intervals can help anticipate the trajectory. Automatic vehicles and taxis can also benefit from travel time and trajectory prediction. Intelligent vehicles have multiple sensors attached to them so that they can collect the essential data and assist the vehicle in learning about its environment.

1.1 Research Question and Objectives

To successfully manage travel, it is vital to know the traffic scenario ahead of time. This research question addresses the concerns that every taxi rider has while taking a taxi ride for short as well as long distance.
RQ:*"How well can Machine Learning predict the total journey time and trajectory a taxi driver can take to reach the destination based on criteria (pick-up location, a list of Global Positioning System (GPS) co-ordinates, and the day of the week (weekday/weekend)) that can help the riders arrive on time at their desired destination?"*

This research is done to predict the taxi trip time and trajectory using machine learning regression algorithms. Below are the Objectives for the research:

Obj 1: A thorough examination of the work done in the current state of the field in order to ensure that the lessons learned will be incorporated into this study, as well as the identification of any significant research gaps.

Obj 2: Identifying the characteristics and elements that influence the length of a taxi ride and the route a taxi driver will follow.

Obj 3: Implementation, evaluation and results of machine learning algorithms to predict taxi ride time of taxi journey.
Sub-Obj 3.1: Implementation, evaluation and results of KNN regression.
Sub-Obj 3.2: Implementation, evaluation and results of Decision tree regression.
Sub-Obj 3.3: Implementation, evaluation and results of Random Forest regression.
Sub-Obj 3.4: Implementation, evaluation and results of Baseline model.
Sub-Obj 3.5: Implementation, evaluation and results of XGBoost algorithm.
Sub-Obj 3.6: Implementation, evaluation and results of Lasso Regression algorithm.

Obj 4: Implementation, evaluation and results of machine learning algorithms to pre-

dict taxi ride trajectory of taxi journey.
Sub-Obj 4.1: Implementation, evaluation and results of Multiple Linear Regression.
Sub-Obj 4.2: Implementation, evaluation and results of Gradient Boosting Regression.

Obj 5: Compare and contrast the results based on evaluation metrices like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared.

The research's main contribution is using prediction models that can be used to anticipate cab travel time and trajectory ahead of time. For this, machine learning approaches are applied and assessed. Riders will be able to plan their route accordingly and arrive at their preferred destination on time because to the cab trip time and trajectory prediction.

The document is structured and follows : Section 2 discusses related work in the same field, which is further divided into subsections based on the technique used, section 3 describes the methodology used, the design decisions made, design architecture, data pre-processing and transformation, section 4 focuses on the implementation of machine learning regression models to predict taxi trip time and trajectory, section 5 provides the analysis of the results and its evaluation.The research is concluded in section 6 with a discussion of the findings and future work.

2 Related Work

The discussion in this section is based on previous work in the same area to forecast cab trip duration and trajectory. This will aid in a better understanding of the approach being used, as well as the evaluation of previous techniques and approaches adopted as a solution for this research.

2.1 Research Based on Travel Time Estimation Using Neural Networks

Tang et al. (2016) introduces a new fuzzy neural interface system-based technique for predicting taxi travel time. It takes into account a number of criteria, including traffic flow and cab speed. These are taken as input for the fuzzy neural system and the output is the estimated travel time. Two strategies are proposed for training the neural network: The first technique involves classifying the input samples into multiple clusters and measuring the membership degree of each cluster center using a Gaussian function. As a result, if the cluster size alters, the Gaussian function is likely to vary as well. The second method uses a weighted recursive least square estimator to optimize Takagi-Sugeno type fuzzy rules in linear function parameters. The proposed method is tested utilizing testing datasets that include both real and simulated data. The proposed model is assessed using performance metrics and compared to existing taxi journey time prediction methods. Similarly, in the study by Gholami et al. (2021), traffic related information is collected from detectors located on freeways and then used to anticipate freeway trip time. During peak hours, when the freeway is likely to be busier, these detectors provide rough estimates of trip time. This study employs detectors that provide directional volume counts and occu-

pancy, allowing for more accurate journey time estimation. Then, using the information collected from the detectors, the Adaptive Neural Fuzzy Inference System (ANFIS) is employed to forecast real-time journey times. ANFIS was chosen for the investigation because it can predict time even if the data is inaccurate or distorted. This makes it a useful and effective tool for estimating travel time on busy highways. The model was implemented in the software program FTTE (Freeway Travel Time Estimator) for better and easier use, and the results were compared between the congested and uncongested freeways.
The Counter Propagation Neural (CPN) network is employed in the research Dharia and Adeli (2003)to forecast journey time on highway connectors. The data for this model is gathered through the use of mass media or a satellite-based navigation system. It then compares the output to a forecasting model that uses the backpropagation (BP) neural network technique to predict trip duration. The paper goes on to say that the proposed freeway time prediction model is best suited for real-time advanced travel information and management systems.
To forecast taxi trip travel time in New York City, the study de Araujo and Etemad (2019) employs a deep neural network. Along with the XGBoost model, it uses a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. The DenseNet design is then compared to the outcomes of this model. The DenseNet architecture outperforms the SVM model, according to the results. In the study Ciskowski et al. (2018), real time data is used to predict travel time using neural nets. After collecting the real time data, a traffic simulator as training data generator is used to train the neural networks.

2.2 Research Based on Travel Time Estimation Using Machine Learning

The importance of taxi time prediction is demonstrated in the study Lee et al. (2015), as well as how it will save time and save fuel use. This research is being conducted in order to increase the overall efficiency of airport surface traffic. A simulation method called Linear Optimized Sequencing (LINOS) is developed to anticipate taxi time, which will improve real-time airport operations. In addition to LINOS, a machine learning-based analytical method is proposed for determining the model's correctness. Using several performance metrics, the anticipated time is compared to the actual data provided by Charlotte Douglas International Airport. There is also discussion regarding how the fast simulation model can be used to improve efficiency.
The importance of journey time estimation for public transportation is emphasized in the research Tran et al. (2020), as well as how it may be a useful tool for transportation agencies in assisting and planning public transit. This research presents a DeepTRANS network created by modifying the Deep Learning-based Estimated Time of Arrival (ETA) model. It also demonstrates that the changes made to the previous model improved its efficiency and accuracy by 21%
Another study Bahuleyan and Vanajakshi (2017)uses data acquired from vehicle GPS to forecast trip time on urban arterial networks. The study's shortcoming is that it only collects data from vehicles that are equipped with GPS. The trip time is predicted using historical data and real-time GPS data. The paths are divided into mid links and intersection links. K- nearest neighbor algorithm is used for travel time at the mid links and the random forest prediction is used at the intersection points. A comparison is made

between the travel time by separating the path links and without separating the links. It was observed that the model was accurate when the path links were separated. Similar study Hofleitner et al. (2012) collects GPS data from the vehicles and predicts the travel time through arterial network. It also uses the historical data along with the real time GPS data. It uses a hybrid model since it is robust to noisy data. Then the model is evaluated and compared with the data driven base line algorithms. The evaluation shows that the hybrid model is 16% more accurate than the baseline models. Both the studies use different models and performs better than the baseline models.
In the study Masiero et al. (2011), historic data and semantic characteristics were combined to calculate trip time within a particular region. It takes the trajectory from the data, pre-processes and aggregates it, and then uses the aggregated data to understand the behavior of the vehicles. This research is confined to a limited area. It just considers a small part of the city. The trip time between source and destination is predicted using a time estimation model. For this, the Support Vector Machine (SVM) is used. In the same way, Support Vector Machine (SVM) is employed in the study Vanajakshi and Rilett (2007) to forecast journey duration using real-time data. On real-time data, along with SVM, Artificial Neural Network, and time series analysis are also conducted. After comparing the data mining models, it was discovered that SVM is more accurate, can be utilized on noisy data, and is a clear alternative for short-term prediction problems.
The gradient boosting regression tree approach is utilized to forecast highway trip time in the study Zhang and Haghani (2015). To overcome the limitations of conventional machine learning models, the gradient boosting regression model employs a tree-based ensemble approach. The model's efficiency and performance are improved by adopting a tree-based method. When the model is compared to the baseline models, it is discovered that the Gradient boosting model outperforms the others.

2.3 Research Based on Trajectory Prediction using Machine Learning

According to the study Wang et al. (2019) trajectory prediction is necessary in order to be prepared for traffic circumstances ahead of time. A multi user multi step trajectory prediction is performed in this study using a deep learning prediction framework. The data is gathered from the vehicles' real-time GPS sensors. On the data, long short-term memory is used. The model's accuracy is then improved using sequence to sequence (Seq2Seq) learning.
The airplane trajectory is predicted using aircraft trajectory and meteorological data in the study De Leege et al. (2013). Multiple elements are properly considered, including the aircraft's starting point, wind direction, and altitude winds. The airplane trajectory is predicted using a regression machine learning algorithm. Using the prediction model, it was estimated how long the interval between aircrafts must be in order to prevent runway collisions. When the model is compared to the baseline models, it is clear that the regression model used in this study increases throughput. Similarly, the study Wang et al. (2017) considers 4D trajectory prediction to improve aircraft traffic. The research is divided into two sections. The first step is to prepare the data. The study takes into account a 4D trajectory dataset. Principle Component Analysis (PCA) is used on the dataset to minimize the vector variable dimensions. The second step is applying the machine learning algorithm. The trajectories are then clustered using a clustering algorithm.

On each cluster, the neural network machine learning method is used. The same is used to calculate an Estimated Time of Arrival (ETA). The model is then compared to the baseline models and evaluated using MSE and RMSE.
The Gan et al. (2016) uses Clustering and Artificial Neural Networks (ANN) to forecast ship trajectory in this study . It talks about the major traffic bottleneck caused by incorrectly projected trajectories. The ships on the existing Yangtze River go in a straight line down the middle of the river. This leads to several traffic problems. To solve this, the K-means clustering technique and ANN models are built taking into account the ship's speed, weight, maximum power, and water level in mind. The models are evaluated after they have been applied to historical data collected over time. The precision of the new model is 70 percent better than the old one.According to the study Altché and de La Fortelle (2017), automated vehicles have a hard time performing activities that humans can generally execute, such as automatic braking. Long short-term memory (LSTM) neural networks are utilized in this study to estimate the longitude and latitude of roadway vehicles. A total of 800 hours of trajectories with varied densities and data from 6000 individual drivers were used to forecast the trajectories. When the model is compared to existing state-of-the-art approaches, it is discovered that the LSTM model outperforms the existing model by 50%. Another study that shows the LSTM model is Wiest et al. (2012) uses Gaussian mixture models to forecast vehicle trajectory a few seconds ahead of time. The data utilized to make this forecast is probability distributed and takes previous motion patterns into account. Then, using this distribution, future trajectories are predicted. The outcome displays the entire trajectory distribution of future predicted trajectories. The variance is calculated and used to predict the accuracy of the forecasts. Similarly, in the study Lin et al. (2021) spatiotemporal attention long short-term memory (STA-LSTM) model is used for vehicle trajectory prediction. The model explains the influence of historical trajectories and nearby cars on the target vehicle using STA-LSTM. Different environmental factors and vehicle factors is considered to predict the vehicle trajectories.For efficient vehicle trajectory prediction, a study Kim et al. (2017) uses a Recurrent Neural Network (RNN). The vehicle's sophisticated behavior is decided with the use of trajectory data and a deep neural network. The RNN-based LSTM prediction approach was utilized in the study. It also generates probability data for a future trajectory on a map grid.

2.4 A Critical Review of the Related Work and Identified Gaps

Following a review of the literature, it is concluded that data attributes have a crucial role in predicting taxi ride time and trajectory. The amount of the data used and the techniques performed in data pre-processing and transformation affect the research results. Table 1 gives the summary of the research papers with their strengths and limitations.

Table 1: Summary of Related Work

Author	**Method**	**Strength**	**Limitation**
Masiero et al. (2011)	Travel Time Prediction Using Machine Learning-SVM	Takes into account a set of semantic variables to predict time 70% more accurately	The prediction can only be done in a limited area.

Table 1 continued from previous page

Wiest et al. (2012)	Probabilistic trajectory prediction with Gaussian mixture models (Variational Mixture Model, Gaussian Mixture Model)	Prevents overfitting. Variational Mixture Model outperforms Gaussian Mixture Model by 40%.	The response is delayed and it does not consider the confidence level of the model.
Bahuleyan and Vanajakshi (2017)	Arterial Path-Level Travel-Time Estimation Using Machine-Learning Techniques	Takes into account Indian Traffic conditions with real time data.	The research splits the intersection and midways making it difficult to predict time on high-variation intersection links.
de Araujo and Etemad (2019)	Deep Neural Network For predicting Travel Time	Using Simple DenseNet2 Architecture with Minimum MAPE (25.70)	Considers trip time prediction in a particular city. Uses limited data. (5 months)
Wang et al. (2019)	Exploring Trajectory Prediction Through Machine Learning Methods. (Seq2Seq, LSTM,SVR, Linear Regression)	Very huge dataset containing 18,670 trajectories.	Does not consider traffic factors and real time data.
Tran et al. (2020)	DeepTRANS: a deep learning system for public bus travel time estimation using traffic forecasting	Uses DeepTRANS Architecture which increases accuracy by 21% than state-of-art-methods.	Complex architecture and less interpretability
Gholami et al. (2021)	An Adaptive Neural Fuzzy Inference System model for freeway travel time estimation based on existing detector facilities	The model is precise and not over-fitting with MAPE=1.51%	The model takes a lot to time for the prediction.

The literature assessment discovered some gaps; most research papers focus on a specific area of the city where the car is located and projecting the travel time. When a large amount of data is given as input, some models, such as Fuzzy Neural Networks, perform slowly. XGBoost, Neural Network, and Support Vector Machine (SVM) are the most prominent approaches used to forecast taxi travel time . XGBoost models give the best results and can handle large amounts of data.

3 Methodology, Design Specification

3.1 Introduction

Before beginning the travel, it is vital to know the estimated arrival time so that the journey can be planned. There have been numerous occasions where people have missed flights or critical conferences owing to a miscalculation of travel time. To estimate cab travel time and trajectory, this study uses the Knowledge Discovery in Database (KDD) methodology in a modified way. The below section covers the methodology, the design decisions taken to implement the study and the steps taken to pre-process the data along with the exploratory data analysis.

3.2 Taxi Trip Methodology Approach Used

The methodology to predict the taxi trip time and trajectory is inspired by KDD methodology. To begin, a dataset[2] is chosen that has the relevant data, such as past travel timestamps and GPS co-ordinates for the journeys. This information is needed to estimate the length of a taxi ride and its trajectory. The data is then pre-processed before being transformed. The data mining regression algorithms are then implemented. The models are then assessed using the evaluation metrics.
Figure 1 shows the methodology that is used to predict taxi travel time and trajectory. In Data Selection a dataset that meets the requirements is identified from the available

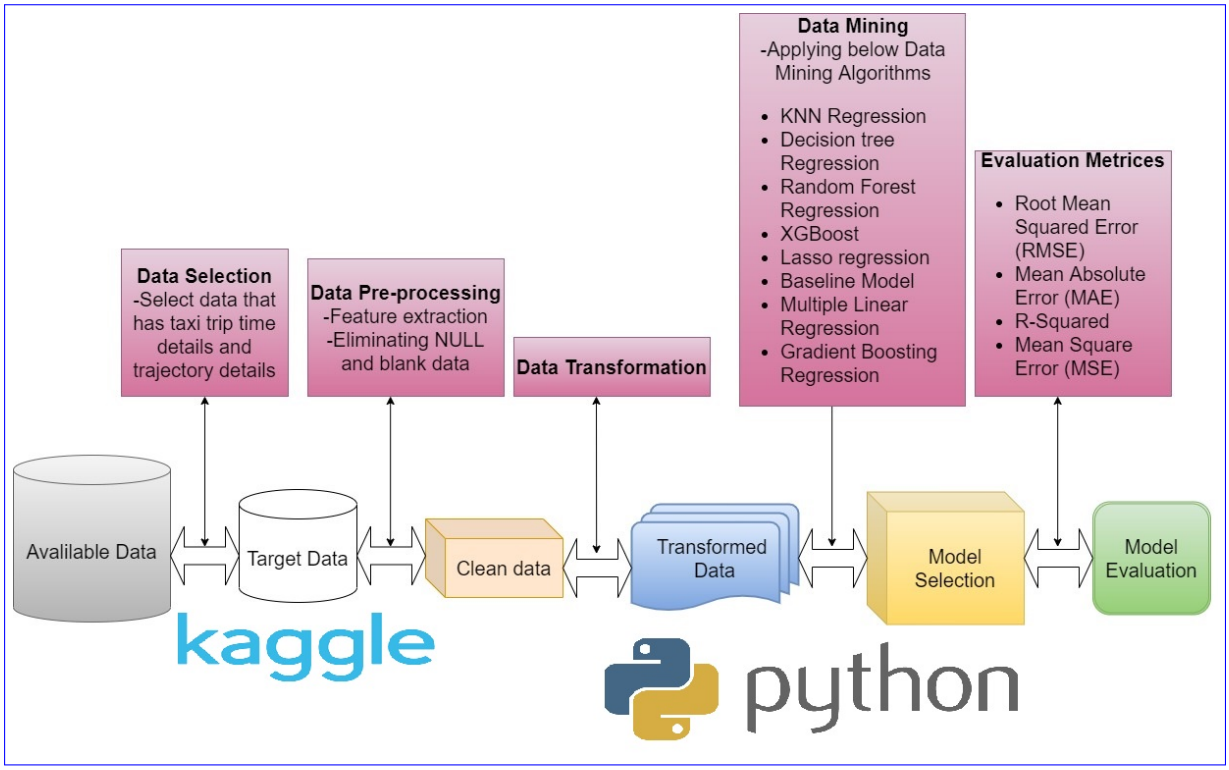

Figure 1: Methodology to predict taxi trip time and trajectory

datasets. The dataset for this study was chosen from Kaggle's public datasets. Before

[2]https://www.kaggle.com/crailtap/taxi-trajectory

being transformed, the data is cleaned and pre-processed. All NULL and missing values have been removed. The data pre-processing and transformation is described in detail in section 3.4. The dataset contains several columns that must be transformed. This transformation will make it easier to employ machine learning techniques. The cleaned and transformed dataset is subjected to data mining regression techniques. Finally, the models are assessed using metrics such as the Root Mean Squared Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE) and R-squared.

3.3 Design Specifications

The dataset chosen for this study is publicly available and it comprises nine features, including a unique identification, a timestamp that indicates the commencement of a cab ride, the type of day the trip was taken (weekday/weekend), and a list of GPS co-ordinates acquired every 15 seconds; it has multiple latitude and longitude values. The dataset has 1710670 records before pre-processing. There are total of 9 attributes in this dataset. One of the major features that is used to anticipate the future trajectory is the GPS co-ordinates captured every 15 seconds. Historical data will be valuable for estimating future latitude and longitude because GPS co-ordinates provide a precise location on the earth. For predicting the trip time, the type of day feature is used along with the timestamp feature.
The Design Architecture shown in the Figure 2 is applied in this research. It is a 2-tier architecture: Tier 1: Business layer, Tier 2: Presentation layer.

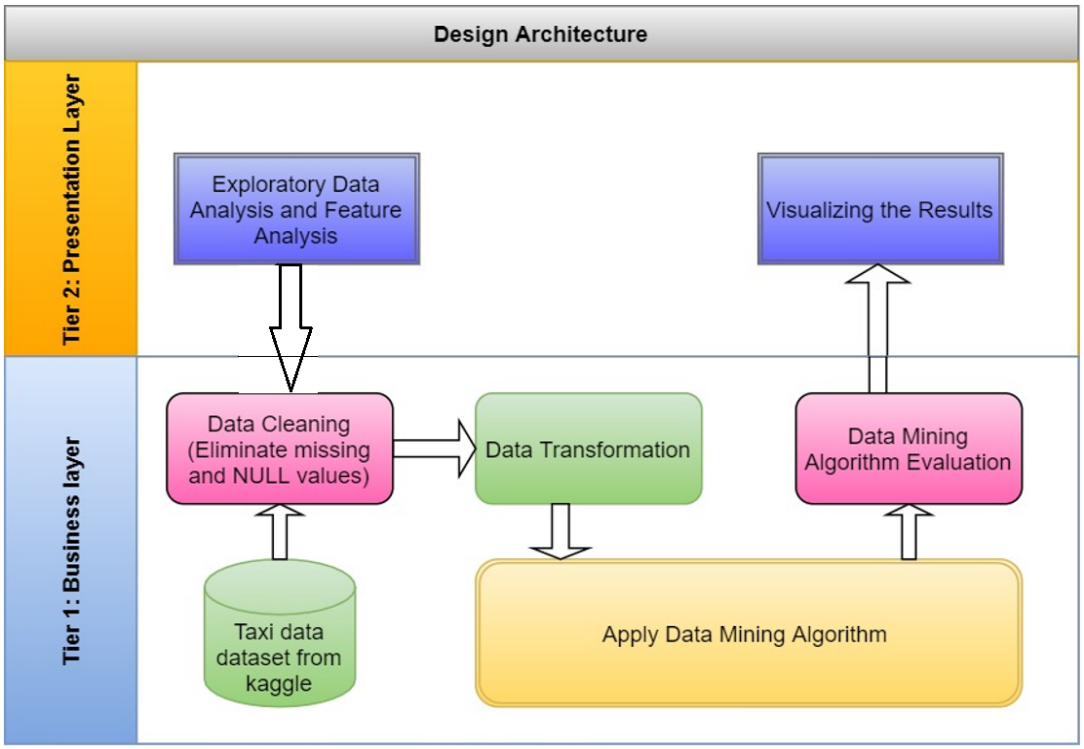

Figure 2: Design Architecture

Tier 1: Business layer: The business layer is where the actual logic is applied on the

dataset. The study will use a dataset that includes relevant taxi information such as timestamps and trajectory characteristics. The next operations carried out in the business layer are data cleansing and data transformation. Algorithms for data mining are used, and the findings are then analyzed.
Tier 2: Presentation layer: The presentation layer involves exploratory data analysis and feature selection once the data has been cleansed. The results are also visualized in the presentation layer after the data mining methods have been applied and the results have been assessed.

3.4 Data Pre-processing and Transformation

The dataset is open to the public and may be found on Kaggle. A continual process of data selection, pre-processing, and transformation guides data mining approaches and exploratory data analysis. The first step in pre-processing the data is to determine whether any attributes have missing values that could affect the outcome of the machine learning techniques. There are a lot of missing values in ORIGIN CALL and ORIGIN STAND attributes as shown in the heat-map in Figure 3. It is possible that none of the

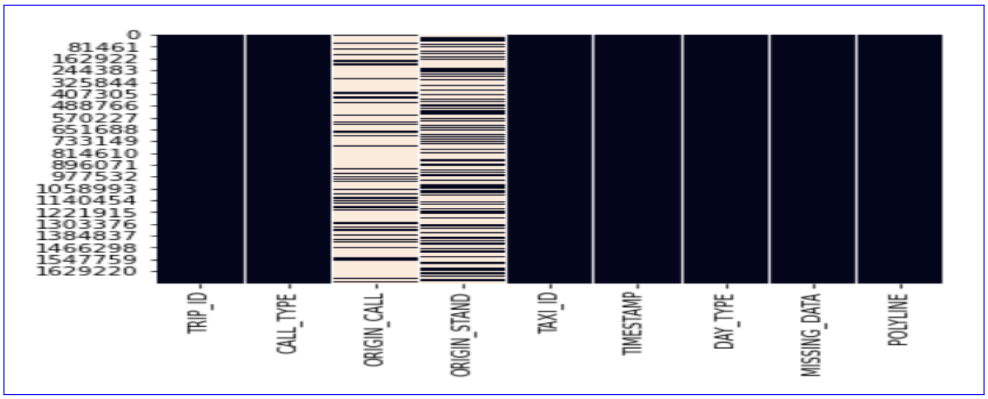

Figure 3: Heatmap showing NULL values

taxi passengers called and did not begin their journey at the taxi stand. Because there are many NULL values and not all rows can be deleted, the NULL values are replaced with 0.
The Figure 4 shows the categorical features of the data. The attribute DAY_TYPE has only 1 unique value and that is 'A' which means that all the trips are started on normal day or weekend. Also the 15 observations don't have the POLYLINE values means we cannot calculate the travel time for those trips. The trip's start time is identified by an

	CALL_TYPE	DAY_TYPE	POLYLINE
count	1710670	1710670	1710670
unique	3	1	1703650
top	B	A	[]
freq	817881	1710670	5901

Figure 4: Categorical features

attribute named TIMESTAMP. Figure 5 shows the individual aspects of the timestamp

attribute by transforming it into Year, Month, Day, Hour, and Weekday columns. Finally,

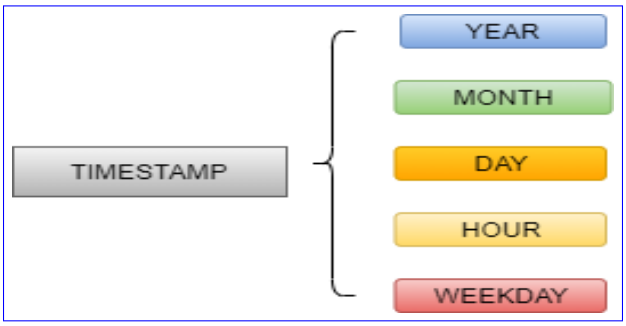

Figure 5: Dividing TIMESTAMP into other attributes

a lambda function is applied to the POLYLINE attribute to get the trip time. 1698888 columns and 24 features are finalized after data pre-processing and transformation, and data mining methods are applied to them.

3.5 Exploratory Data Analysis

To get a summarized information about the data the below exploratory data analysis is done on the pre-processed data. The data in the dataset spans two years, 2013 and 2014. Figure 6 is a time series visualization depicting the distribution of taxi rides in 2013 and 2014. It can be seen that there is little difference in taxi journeys between the two years.

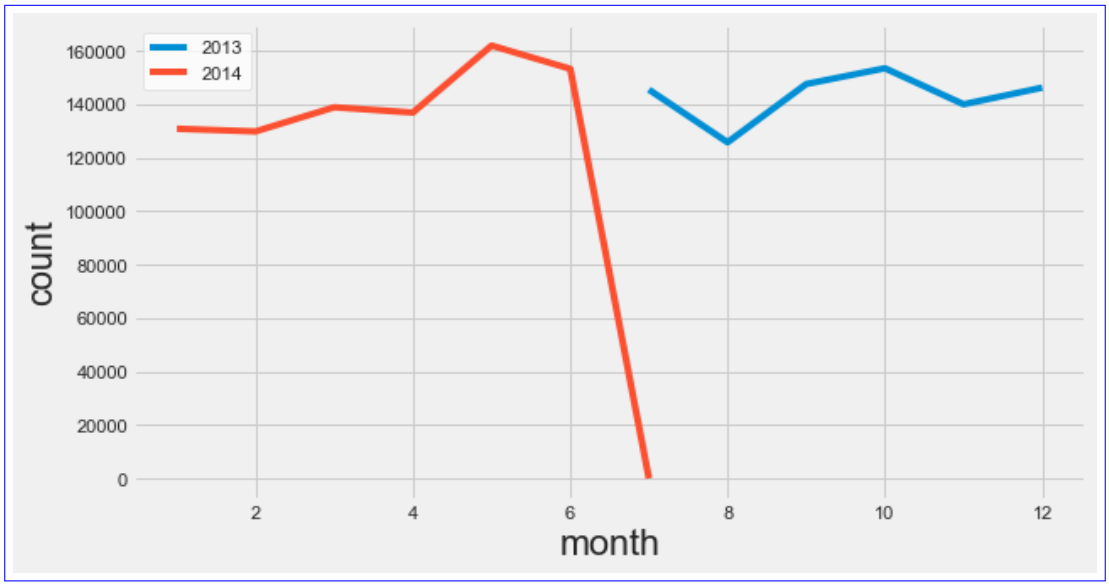

Figure 6: Time Series Visualization-taxi journey in each year

Figure 7 displays the number of taxi journeys that occur each day in a bar chart. The numbers 0 and 6 represent Sunday and Saturday, respectively. The most trips have been recorded on Friday, followed by Saturday.In addition, the number of journeys on Monday was lower. Another method to categorize the journeys by day is to say that the same

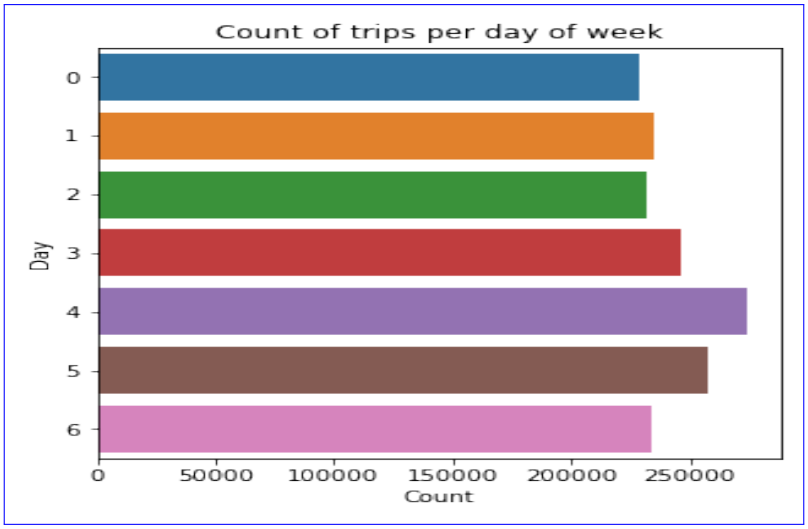

Figure 7: Bar Chart Showing Weekly Taxi Trips

number of taxi journeys were taken on the fourth and fifth days, and almost the same amount on the other days. Therefore, the taxi rides are not affected by the weekend or working days.
The Map in the Figure 8 shows the start and end of 5000 taxi trip respectively. The destination is highlighted in blue whereas red represents the start of the taxi trip. The map shows that the taxi trips are widely distributed and there are many numbers of entry and exit points of the city.

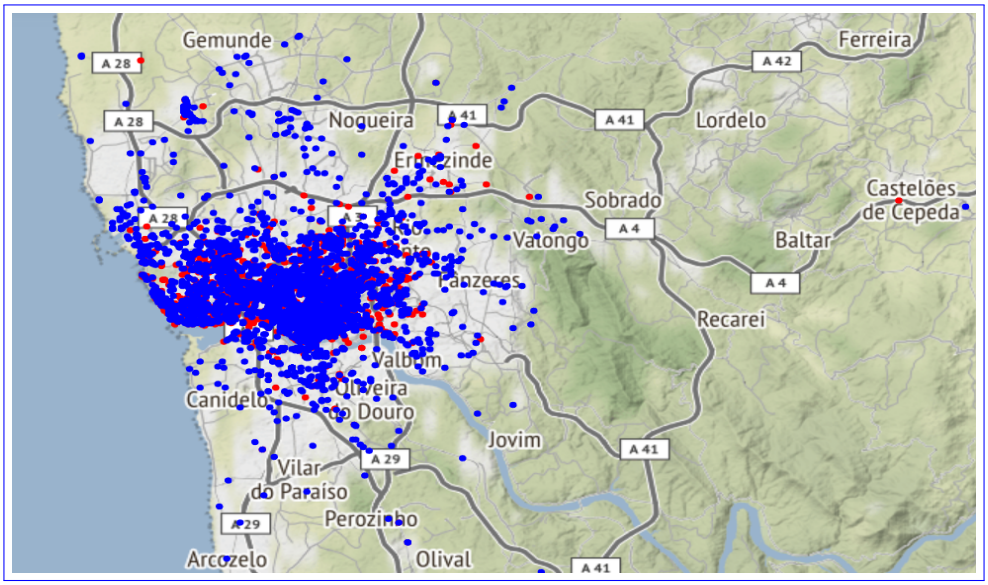

Figure 8: Start and End of the taxi trip

To predict the taxi trajectory, a delta longitude and latitude is calculated with the help of the source and destination co-ordinates. The Figure 9 shows the distribution of the delta latitude and delta longitude in scatter plot and histogram.

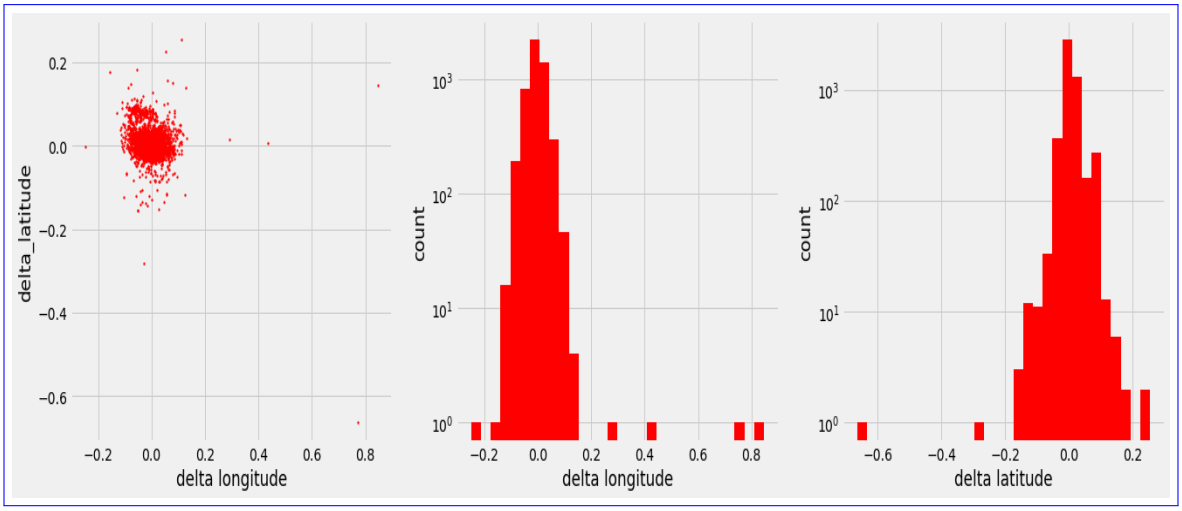

Figure 9: Delta Longitude and Latitude Distribution

3.6 Feature Selection

Using Feature Extraction, the attributes that have little or no link to the dependent variable are eliminated. In this research trip time and trajectory is predicted and both will have different independent variables and dependent variable. Two types of feature selection is performed in this research.

3.6.1 Feature Selection Using Linear Regression

To choose top features for predicting cab travel time, Linear Regression is employed. This method assigns a numerical value to each feature and assigns it importance. The features that aren't important are given a value of 0 by default. Some features have a negative relevance, yet they are still chosen because they may produce beneficial results when applied to the model. This feature selection strategy decreases model overfitting, increases model accuracy, and minimizes data training time. A bar plot for the same is shown in the figure Figure 10. Out of 18 features 12 features are selected based on the importance.

3.6.2 Co-relation Matrix Using Heatmap

The 2D matrix shows the co-relation between 2 variables using a monochromatic scale to represent the data in colored cells shown in Figure 11. Later, only those independent characteristics having a co-relation greater than 0.1 with the dependent variable are chosen based on the target variables (lat_last and lon_last). As a result, three features are chosen: lon_1st, delta_lon, and delta_lat.

3.7 Conclusion

The technique for implementing Machine Learning algorithms is described in this part, as well as the design considerations made in this regard. The feature selection process identifies the greatest qualities that will benefit the model and improve its efficiency. To

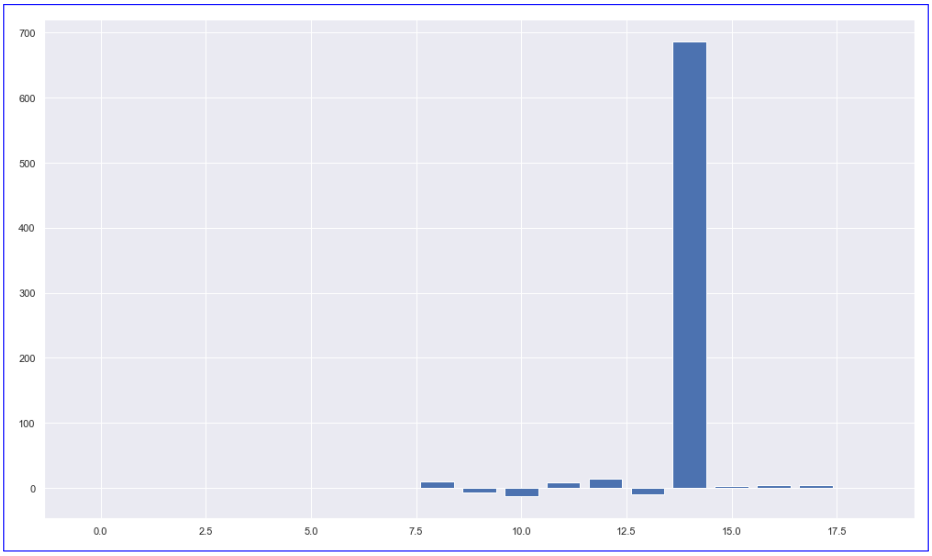

Figure 10: Feature Importance Using Linear Regression

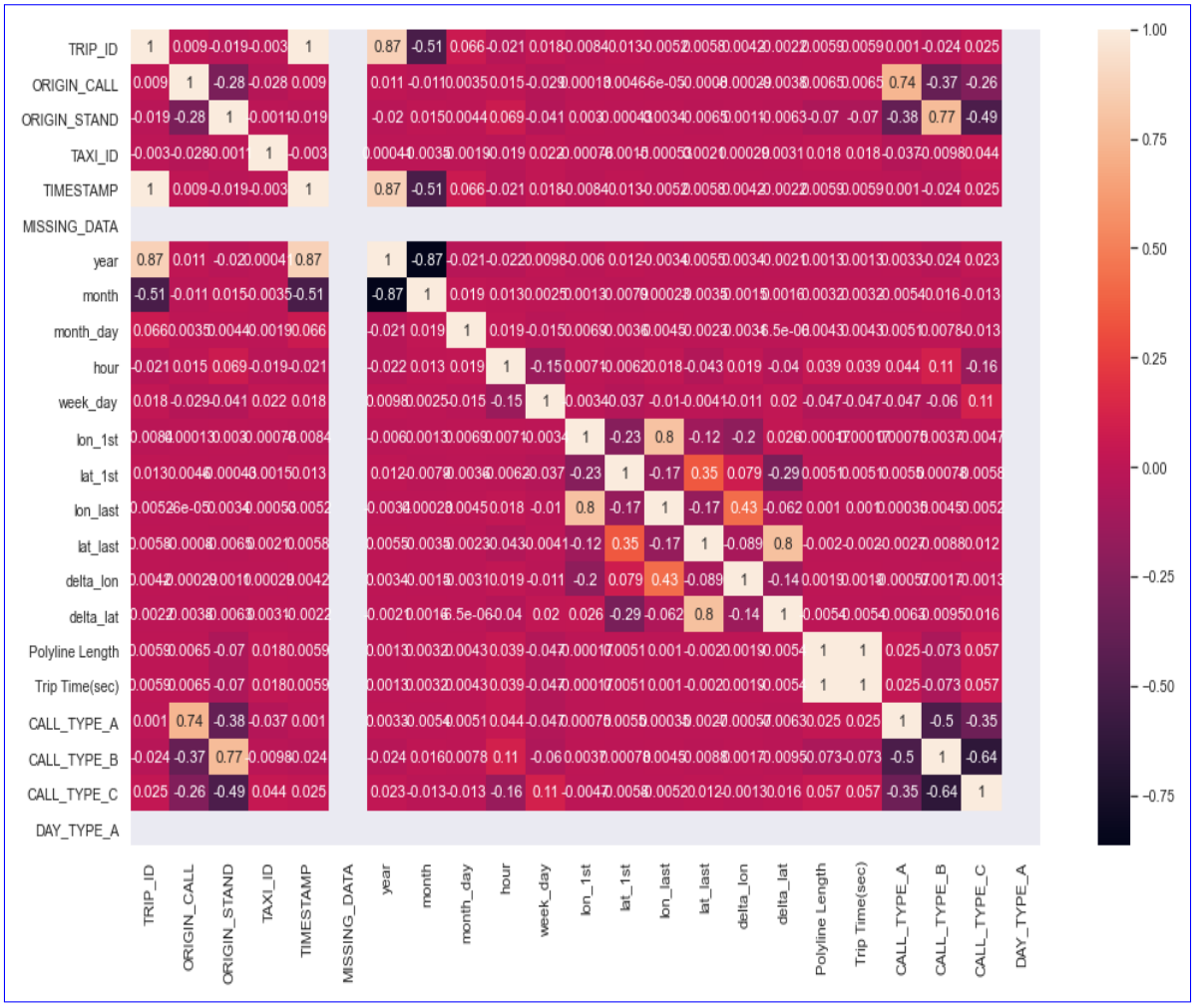

Figure 11: Co-relation Matrix Using Heatmap

forecast the cab trip time, the Linear regression approach selects the best 12 features. In the same way, three features are chosen to forecast the cab trajectory.

4 Implementation of Taxi Trip Prediction Models

This section presents the prediction models applied to the dataset after cleaning to forecast taxi journey time and trajectory. There are two parts to the implementation: 1. Estimate the duration of a taxi ride 2. To forecast the taxi trip's path. This study employs regression machine learning techniques to analyze the relationship between a dependent and independent variable (s). The research objective 3 and 4 is completed in this section.

4.1 Implementation of Prediction Models to Predict Taxi Trip Time

4.1.1 Baseline Model

The anticipated journey time under the baseline model would simply be the average of all trip times. It's one of the most basic models. After the use of the baseline model, the model's complexity is enhanced in order to get better results than the baseline model. When analyzing the other machine learning models, the scores from the baseline model give the necessary point of comparison. The performance of the other models presented in the following sections is compared to that of the baseline. The general concept is that the other machine learning algorithms must outperform the specified baselines.
Experimenting with the baseline model is also simple and inexpensive. It is less prone to overfit because it is a simple model. If the simple baseline model overfits, applying more complicated models is pointless because the complexity will degrade the model's performance. A simple baseline model should be straightforward to understand.

4.1.2 K-Nearest Neighbour Regressor (KNN)

The KNN algorithm predicts the values of new data points based on 'feature similarity.' As a result, the new point is given a value based on how similar it is to the points in the training set. So, in the case of taxi trip time prediction, the KNN algorithm will examine the values of independent variables in historical data and, if they are related to the one that has to be forecasted, the taxi journey time will be similar to the previous one. Therefore, it determines how similar the data points are and then predicts the outcome. The basic working of KNN regression algorithm is as follows:

- The first step in the KNN regression algorithm is to calculate the distance between the new point and the previous data points.
- The closest k (number of neighbors) data points are selected.
- The average of the selected data points is calculated.

The value of k is important while using the KNN regression. Very low value of k means the model can overfit on the training data. On the other hand, if the value of k is high, the model performs poorly on both train and validation sets. For regression, the KNN

regressor returns the mean of the k nearest neighbor. Figure 12. shows the elbow curve used to estimate the number of neighbours to be used in the KNN regressor.

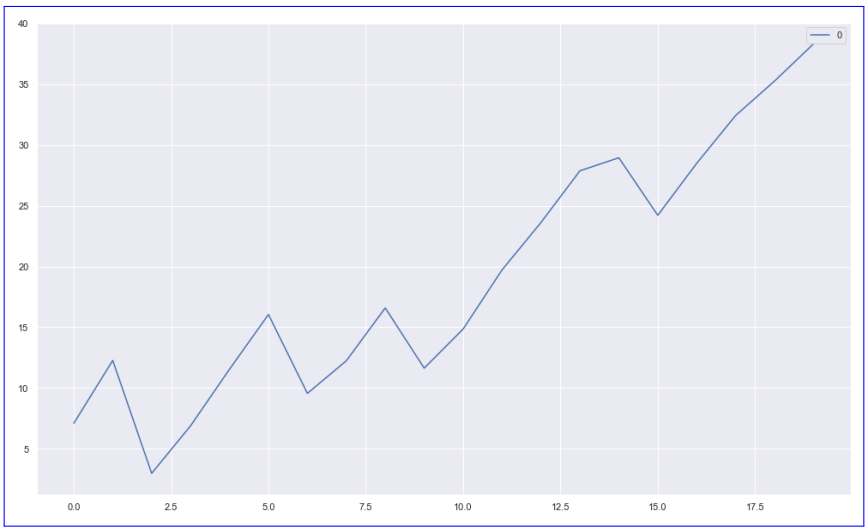

Figure 12: Elbow Curve to estimate k Neighbours

4.1.3 Lasso Regression

LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression enhances the model's prediction accuracy and interpretability as it performs variable selection and regularization. Lasso regression is used in this research since the dataset used here has high dimensionality and high co-relation. Lasso regression is a type of regularization technique which is used when the data is large and it reduces the chances of data overfitting.

4.1.4 Decision Tree Regression

The decision tree regression approach separates the dataset into smaller sets while concurrently developing an associated decision tree. The resulting tree has leaf nodes and decision nodes. J. R. Quinlan's ID3 technique for generating decision trees uses a top-down, greedy search through the set of feasible nodes with no backtracking. Since the decision tree involves partitioning the dataset into subsets based on the similar homogenous values. To calculate the homogeneity standard deviation is used. If the data is completely similar, standard deviation is zero.
When training the model, the model learns any relationship between the target and the other data attributes. Decision tree is used since it considers all possible outcome and traces the complete path to reach the conclusion.

4.1.5 XGBoost

XGBoost is a method of ensemble learning. Ensemble learning has the advantage of combining the predictive power of several learners, with the resultant model being the aggregate of numerous models. The following is a description of how the XGBoost algorithm works in general:

- To forecast the target variable, an initial model M0 is defined. There will be an error (residual) associated with this model.
- The error from the previous stage is fitted with a new model f0.
- M0 and f0 are now integrated to form M1, a new model. The M1 error rate will be lower than the M0 error rate.

XGBoost was introduced in 2014, and ever since it has been lauded for its incredible performance and speed. XGBoost is used in this research to predict taxi trip time since it is fast and its performance is better as compared to its other gradient boosting algorithm counterparts.

4.1.6 Random Forest Regression

Random forest, like the XGBoost method, is an ensemble learning algorithm. By merging different learning models, it creates an ensemble of decision trees that improves the overall result. The solution becomes more robust as the number of trees created increases, resulting in increased model accuracy. Random forest is a collection of decision trees in simple words, however there are some differences between random forest and decision tree. Because the random forest method generates more trees, the likelihood of model overfitting is reduced. Overfitting is more likely with decision trees than with random forests.
The Random Forest machine learning technique offers a high level of accuracy, can manage any missing values in the dataset, and can handle a huge amount of data with a significant dimensionality.

4.2 Implementation of Prediction Models to Predict Taxi Trajectory

4.2.1 Multiple Linear Regression

Multiple Linear Regression will help to determine if the target variable and the independent variable have a linear relationship. It appears to be more efficient than a simple linear regression model since it can predict the dependent variable using multiple independent factors. As a result, it reveals more details about the variable and its linearity.
Multiple linear regression is represented by the below equation y:

$$Y = mx1 + mx2 + mx3 + b$$

Where Y is the dependent variable, x1, x2, x3 are independent variables, m is the slope of regression and b is a constant.

4.2.2 Gradient Boosting Regression

When the data is real-time, the gradient boosting approach works well. Because the dataset comprises real-time data such as the sort of day the trip was made, the GPS co-ordinates, and so on, the gradient boosting regression approach was utilized in this study. Since the Gradient Boosting technique does not operate well when there is a lot

of noise in the data. As a result, as explained in section 3.3, the dataset is cleansed and pre-processed. Gradient Boosting algorithm is used to reduce the error of the model.

5 Evaluation and Results of Developed Taxi Trip Prediction Models

The R-squared (co-efficient of determination), Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE) are the evaluation metrics employed in this study. The model's performance will be measured using these metrics. The Research Objective 5 is completed in this section.

The standard deviation of prediction error, or RMSE, indicates how effectively the data is distributed around the focused line. The formula for calculating RMSE is as follows:

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (f_i - o_i)^2}$$

where n is the number of samples, f is the forecasts and o is the observes values.
The MSE indicates how close the regression line is to the points. It calculates the average of a group of errors. The formula for calculating MSE is as follows:

$$MSE = 1/n \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

where n is the number of data points, Y_i is observed values and $\widehat{Y}_i$ is predicted values.
R-squared is a relative fit metric. It demonstrates how well the model fits the data. The R-squared formula is as follows:

$$R^2 = 1 - RSS/TSS$$

where RSS is sum of square of residuals and TSS is total sum of squares.
The mean of absolute errors is known as MAE. It depicts the difference between the predicted and actual value. The MAE formula is as follows:

$$MAE = (\sum_{i=1}^{n} |y_i - x_i|) \backslash n$$

where y_i is predicted value, x_i is true value and n is the total number of data points.

5.1 Result of Prediction Models to Predict Taxi Trip Time

The best_estimator_ parameter was utilized for each model to identify the best parameter for the model without the need for a manual approach. After the parameters have been applied, the assessment metrics are chosen. The findings for the testing dataset is shown in Table 2. It indicates that Lasso Regression yields the best results in terms of model correctness, mean error, and relative fit. The model's MAE is 1.01, indicating the disparity in actual and predicted values. The lower the MSE, the better the model's

ability to predict actual values is. Similarly, the model's R-Squared value is 1.0, indicating that it correctly fits the data.

Table 2: Result for Testing Data-Taxi trip time prediction

Machine Learning Models	**Evaluation Metrices**			
	RMSE	**MSE**	**R^2**	**MAE**
Baseline Model	601.40	361688	0.0	335.71
KNN Regressor	2.97	1468.03	0.99	0.91
Lasso Regression	**1.81**	**3.28**	**1.0**	**1.01**
Decision Tree Regression	28.66	821.47	0.99	0.63
XGBoost	7.10	50.47	0.99	0.20
Random Forest Regression	3.16	10.03	0.99	0.11

Because it is critical to forecast travel time ahead of time, the most relevant statistic in terms of precision to handle the issue is MAE. The MSE score for Lasso Regression is the smallest, indicating that there is little discrepancy between the actual and predicted values.

5.2 Result of Prediction Models to Predict Taxi Trip Trajectory

There are two target columns to be predicted for taxi trip time prediction. The target columns are evaluated and accessed using histogram and probability scatter plot. The Figure 13 shows the data distribution for first target variable lon_last. The histogram plot shows the frequency of specified range of data. The range of data is limited and all gathered around one end. It means that the longitude of most of the taxi rides fall at the same place.

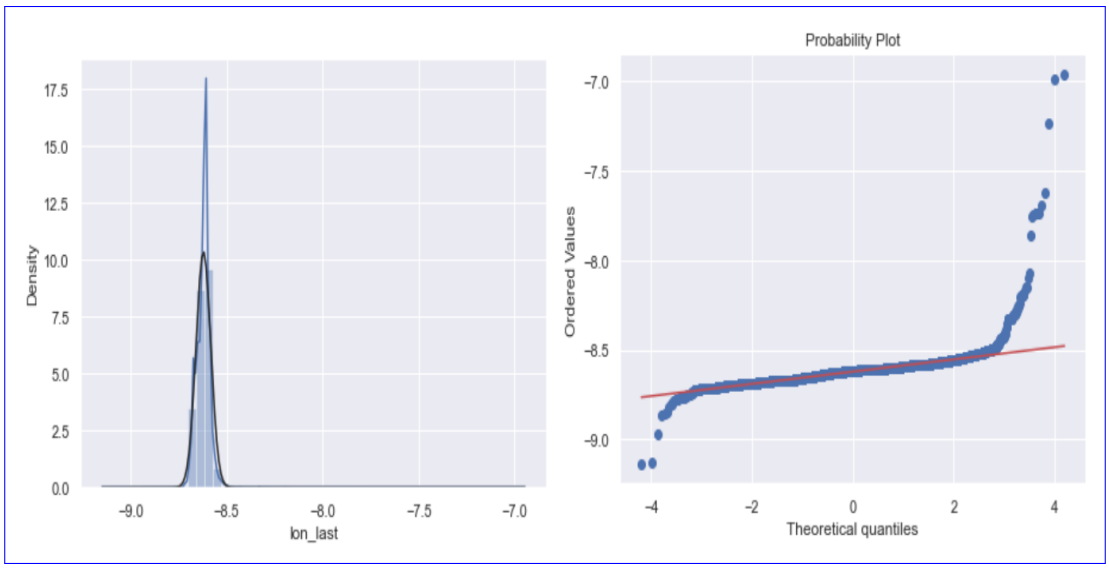

Figure 13: Histogram and Probability plot for lon_last

Similarly, the Figure 14 shows the data distribution for the second target variable lat_last. The histogram distribution is similar, a specific range of data is all gathered at the right

end which means that even the latitude falls at the same place.

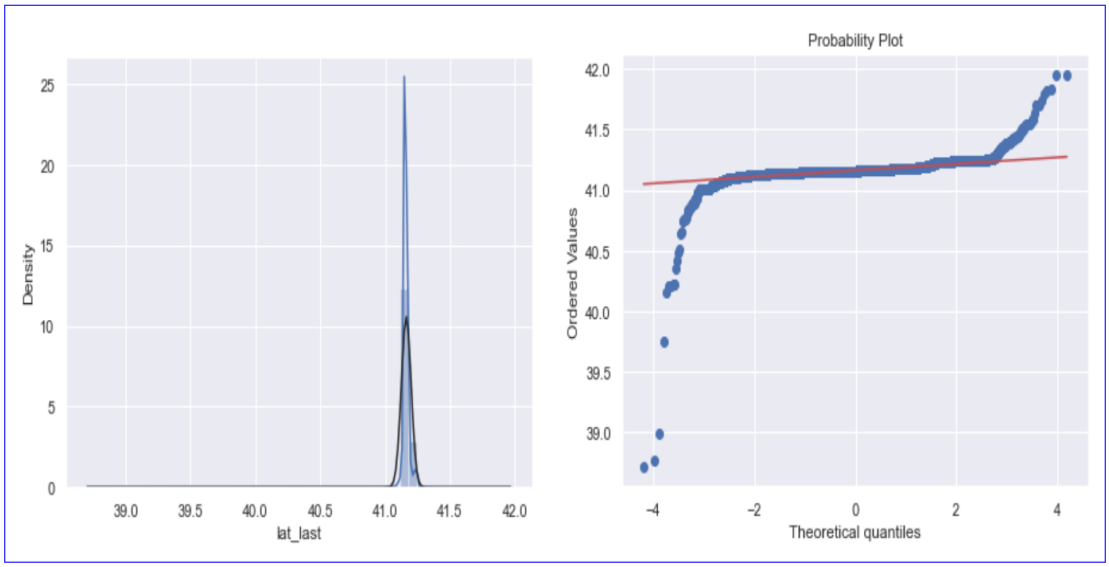

Figure 14: Histogram and Probability plot for lat_last

The results of the Multiple Linear Regression and Gradient Boosting Regression models used to forecast the cab journey trajectory are shown in Table 3. Gradient Boosting regression outperforms Multiple Linear Regression, according to the findings. R-squared equals 0.73, indicating that 73 percent of the data matches the regression model. Similarly, the model's MSE is very close to zero, indicating that there is little data disparity between the actual and projected values.

Table 3: Result for Testing Data- Taxi Trip Trajectory Prediction

Machine Learning Models	**Evaluation Metrices**			
	RMSE	**MSE**	**R^2**	**MAE**
Multiple Linear Regression	0.026	0.0006	0.50	0.008
Gradient Boosting Regression	**0.018**	**0.0003**	**0.73**	**0.008**

The discrepancy between the real and predicted values for Gradient Boosting Regression is shown in the Figure 15. There is a very small difference between the values, as can be observed.

5.3 Evaluation and Discussion

The expected models and technique were successfully implemented in this investigation. Baseline Model, XGBoost Regression, Lasso Regression, Decision Tree Regression, KNN Regression, and Random Forest Regression are used to predict taxi trip ride duration, whereas Gradient Boosting regression and Multiple Linear Regression are used to predict taxi trajectory. When it comes to predicting cab trip time trajectory, Lasso Regression outperforms other models, whereas Gradient Boosting Regression outperforms Multiple

	Actual Values	Predicted Values
0	-8.613738	-8.613826
1	41.167296	41.155174
2	-8.605188	-8.607922
3	41.125401	41.155677
4	-8.622792	-8.621324
5	41.165010	41.155677

Figure 15: Gradient Boosting Regression-Actual VS Predicted

Linear Regression. The proposed models are compared to the existing models in Table 4.

Table 4: Comparision with the Existing Models

Research	**Method**	**RMSE**	**MSE**	**MAE**	**R^2**	**MAPE**
Proposed Model	**Predicting taxi trip time using Machine Learning Algorithms-Lasso Regression**	**2.15**	**4.65**	**1.09**	**1.0**	
	Predicting taxi trip trajectory using Machine Learning Algorithm-Gradient Boosting Regression	**0.018**	**0.0003**	**0.0086**	**0.73**	
Dharia and Adeli (2003)	Neural network model for rapid forecasting of freeway link travel time		6.3		0.98	
Wang et al. (2017)	Short-term 4D Trajectory Prediction Using Machine Learning Methods	1.04		8.4		
Kim et al. (2017)	Probabilistic Vehicle Trajectory Prediction over Occupancy Grid Map via Recurrent Neural Network			0.29		

Table 4 continued from previous page

Gholami et al. (2021)	An Adaptive Neural Fuzzy Inference System model for freeway travel time estimation based on existing detector facilities				0.99	1.15%

Table 4 shows a comparison of the suggested method to the present state-of-the-art method. As described in sections 5.1 and 5.2, Lasso regression and Gradient Boosting Regression outperform the other proposed regression algorithms.

6 Conclusion and Future Work

It is critical for riders to be able to predict taxi travel time and trajectory so that they may plan their journey properly. This will allow them to get at their preferred locations on schedule. To do so, the data is pre-processed and transformed, and the key characteristics are discovered by visualizing the data. Later, feature selection techniques such as Linear Regression and Pearson correlation matrix are used to take into account specific features.Features: the type of day, the length of the travel ride and GPS co-ordinates are selected with the help of the feature selection techniques.These characteristics are subsequently taken into account, and prediction models are deployed. All of the objectives are satisfied, as mentioned in section 1.1, and this study offers a novel approach to predict taxi trip time and trajectory that outperforms existing methods.
The goal of this research is to predict the taxi trip time and trajectory that can help the travellers to reach their destination on time. The data is subjected to prediction techniques, and the outcomes are compared and assessed using the evaluation measure outlined in section 5. After a comparison, it can be said that Lasso regression outperforms other machine learning approaches when it comes to predicting taxi trip time, while Gradient Boosting regression outperforms other machine learning techniques when it comes to predicting taxi trip trajectory.
The data taken for this research is limited to taxi rides alone.The data can be expanded in the future to anticipate ties for public transportation as well. Deep learning techniques such as Convolutional Neural Network (CNN) can also be used to forecast ride time and trajectory. CNN is better at dealing with unstructured data and requires less human interaction to recognize inputs.

Acknowledgement

I would like to express my sincere thanks and gratitude to my supervisor Dr. Catherine Mulwa for her constant guidance, and feedback. I am also thankful to my parents and friends for their constant support and motivation.

References

Altché, F. and de La Fortelle, A. (2017). An lstm network for highway trajectory prediction, *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 353–359.

Bahuleyan, H. and Vanajakshi, L. D. (2017). Arterial path-level travel-time estimation using machine-learning techniques, *Journal of Computing in Civil Engineering* **31**(3): 04016070.

Ciskowski, P., Drzewiński, G., Bazan, M. and Janiczek, T. (2018). Estimation of travel time in the city using neural networks trained with simulated urban traffic data, *International Conference on Dependability and Complex Systems*, Springer, pp. 121–134.

de Araujo, A. C. and Etemad, A. (2019). Deep neural networks for predicting vehicle travel times, *2019 IEEE SENSORS*, IEEE, pp. 1–4.

De Leege, A., van Paassen, M. and Mulder, M. (2013). A machine learning approach to trajectory prediction, *AIAA Guidance, Navigation, and Control (GNC) Conference*, p. 4782.

Dharia, A. and Adeli, H. (2003). Neural network model for rapid forecasting of freeway link travel time, *Engineering Applications of Artificial Intelligence* **16**(7-8): 607–613.

Gan, S., Liang, S., Li, K., Deng, J. and Cheng, T. (2016). Ship trajectory prediction for intelligent traffic management using clustering and ann, *2016 UKACC 11th International Conference on Control (CONTROL)*, IEEE, pp. 1–6.

Gholami, A., Wang, D., Davoodi, S. R. and Tian, Z. (2021). An adaptive neural fuzzy inference system model for freeway travel time estimation based on existing detector facilities, *Case Studies on Transport Policy* .

Hofleitner, A., Herring, R. and Bayen, A. (2012). Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning, *Transportation Research Part B: Methodological* **46**(9): 1097–1122.

Kim, B., Kang, C. M., Kim, J., Lee, S. H., Chung, C. C. and Choi, J. W. (2017). Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network, *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 399–404.

Lee, H., Malik, W., Zhang, B., Nagarajan, B. and Jung, Y. C. (2015). Taxi time prediction at charlotte airport using fast-time simulation and machine learning techniques, *15th AIAA Aviation Technology, Integration, and Operations Conference*, p. 2272.

Lin, L., Li, W., Bi, H. and Qin, L. (2021). Vehicle trajectory prediction using lstms with spatial-temporal attention mechanisms, *IEEE Intelligent Transportation Systems Magazine* .

Masiero, L. P., Casanova, M. A. and de Carvalho, M. T. M. (2011). Travel time prediction using machine learning, *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, pp. 34–38.

Tang, J., Zou, Y., Ash, J., Zhang, S., Liu, F. and Wang, Y. (2016). Travel time estimation using freeway point detector data based on evolving fuzzy neural inference system, *PloS one* **11**(2): e0147263.

Tran, L., Mun, M. Y., Lim, M., Yamato, J., Huh, N. and Shahabi, C. (2020). Deeptrans: a deep learning system for public bus travel time estimation using traffic forecasting, *Proceedings of the VLDB Endowment* **13**(12): 2957–2960.

Vanajakshi, L. and Rilett, L. R. (2007). Support vector machine technique for the short term prediction of travel time, *2007 IEEE Intelligent Vehicles Symposium*, IEEE, pp. 600–605.

Wang, C., Ma, L., Li, R., Durrani, T. S. and Zhang, H. (2019). Exploring trajectory prediction through machine learning methods, *IEEE Access* **7**: 101441–101452.

Wang, Z., Liang, M. and Delahaye, D. (2017). Short-term 4d trajectory prediction using machine learning methods, *Proc. SID*, pp. 1–10.

Wiest, J., Höffken, M., Kreßel, U. and Dietmayer, K. (2012). Probabilistic trajectory prediction with gaussian mixture models, *2012 IEEE Intelligent Vehicles Symposium*, IEEE, pp. 141–146.

Zhang, Y. and Haghani, A. (2015). A gradient boosting method to improve travel time prediction, *Transportation Research Part C: Emerging Technologies* **58**: 308–324.