Prediction of Dynamic Price of Ride-On-Demand Services Using Linear Regression

Kunal Arora, Sharanjit Kaur, Vinod Sharma
Amritsar Group of Colleges
Amritsar

ABSTRACT

Ride-on-demand (RoD) services are becoming more and more common, such as Uber and OLA cabs. To help both drivers and customers, RoD services use dynamic pricing to balance supply and demand in an attempt to increase service quality. Dynamic prices, however, often generate problems for passengers: often "unpredictable" prices prevent them from easily making fast decisions. In order to address this problem, it is therefore important to give passengers more detail, and forecasting dynamic prices is a feasible solution. Taking the Rapido dataset as an example in this paper, we focus on the estimation of dynamic prices, forecasting the price for each individual passenger order. Predicting prices will help passengers understand whether they could get a lower price in nearby locations or in a short period of time, thus alleviating their concerns. By learning the relationship between dynamic prices and features derived from the dataset, the prediction is carried out. As a representative, we train one linear model and test its output based on real service data from various perspectives. Furthermore, we view the contribution of features based on the model at different levels and find out what features contribute most to dynamic prices. Finally, we predict dynamic prices using an efficient linear regression model based on evaluation results. Our hope is that the study helps to make passengers happier as an accurate forecast.

Keywords – Machine learning, Prediction, Linear regression, Target, Correlation, Ride-on-demand services, and Dynamic price.

INTRODUCTION

Machine learning is a data processing tool that automates the creation of analytical models. It is a subset of artificial intelligence focused on the premise that, with minimal human interaction, computers can learn from data, recognize trends and make decisions. It is being widely used today, with the goal of allowing computers to learn automatically without human intervention or assistance and change their actions accordingly. Machine learning algorithms are used in a wide range of applications, for example, Educational data mining, Stock data mining, diagnosis of neurological disorders, power management, and sentiment analysis, etc.

In the area of education, Educational Data Mining (EDM) exploits statistics, computer learning and data mining to interpret and forecast educational data using different approaches. To better understand learners and their learning, EDM aims to use online learning modes and develop computational methods to evaluate the facts and figures in order to help learners. [1] On the other hand, nowadays, overwhelming stock details are available, and are only of value if correctly analyzed and mined. Using various mathematical and supervised learning methods, stock data may be significantly analyzed. [2] Moreover, neurological diseases are persistent and life-threatening disorders that have a bad effect on human life's overall routine. In the diagnosis of these diseases, deep learning strategies have attracted the enormous interest of scholars. [3]

Furthermore, in the background of smart cities, energy efficiency in the public sector is an important issue. To forecast energy consumption, machine learning models can be built to suggest the design of an intelligent machine learning-based public sector energy management framework that could be seen as part of the idea of a smart city. [4] Another application of machine learning is sentiment analysis using various kinds of techniques. Analysis of sentiment and viewpoint mining aid in the analysis of people's views, opinions, attitudes, thoughts and feelings. With the growth of social media like Twitter, Facebook, Quora, blogs, microblogs, Instagram and other social networks, the demand for sentiment analysis occupies the same space. [5] For the psychological study network, artificial learning can also be used to recognize the times and fields of most influential concern. Among the most important research topics, world’s highest affecting disorders such as depression (fourth largest disease) turn out to be excellent. [6]

Recent years are witnessing the rapidly rising worldwide ride-on-demand (RoD) services market, such as Uber and OLA cabs. RoD service attracts passengers by its comfort, affordable prices, and versatile service. In their daily transportation, a growing number of passengers now take RoD service as a standard option. Dynamic pricing is the central and distinctive characteristic of the RoD service and represents the attempt to balance supply (the number of cars on the road) and demand (the number of requests from passengers): higher prices decrease demand and increase supply in a busy area, whereas lower prices in a non-busy area do the opposite. For both drivers and passengers, this makes the service more responsive.

We extract relevant features and predict the dynamic prices based on the data, taking the Rapido dataset as an example. Rapido is an Indian online bike taxi RoD service, operating in over 75 cities across the country. Rapido was estimated to have over 15,000 registered riders in September 2018, with an average of 30,000 rides every day.The dynamic price (trip fare) of Rapido data may be influenced by many parameters such as pick- up location, drop location, travel distance, travel time and timestamp. We consider travel distance and travel time as the main features for predicting dynamic prices by analyzing the associations between dynamic prices and different data features. A linear regression model is trained to make predictions of the dynamic data price.

Linear Regression is one of the machine learning algorithms where the outcome is estimated by the use of known parameters that are associated with output. Instead of trying to classify values into different groups, it is used to predict values within a continuous range. The known parameters are used to predict the unknown parameter or the result. When known and unknown parameters are plotted on the X and Y axes, it forms a continuous and steady slope.

Our trained model predicts dynamic price (trip fare) with good precision and efficiency (i.e., 93.40%) by considering the features (travel distance & travel time) which are strongly correlated to dynamic price (trip fare).

RELATED WORK

The relationship between age and language has been studied as an example of one of these kinds of research. The model of Linear Regression can also be used to predict the age of the author of a text. Three separate data genres are used simultaneously to investigate the age prediction of the author: blogs, telephone conversations and online forum messages. The efficient features include both stylistic as well as content-oriented ones, after careful examination of different data characteristics. These characteristics are considered to be the main features that are going to be useful in age prediction. Correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years are obtained. In conclusion, content characteristics as well as stylistic features were discovered as strong indicators of the age of an individual.[7].On the other hand, multiple linear regression models were used in one more analysis to predict the most significant yarn parameters of ring spun cotton yarns. With linear multiple regression analysis, they utilized AFIS (Advanced Fiber Information System) fiber properties, roving and yarn properties. The correlations between dependent variables, i.e., yarn properties and independent variables, i.e., fiber, roving, and yarn properties, has been calculated one by one. The results suggested that for each yarn property, the relationships between variables and yarn properties are nearly linear.

Table 1: Goodness of fit statistics of models; * Standard Error of the Estimation

Para- meter	Tenaci-ty	Elongati-on	Unevennes-s	Hairines-s
R	0.969	0.844	0.942	0.894
R^2	0.938	0.712	0.887	0.799
Adj. R^2	0.936	0.702	0.882	0.790
SEEE*	0.967	0.457	0.809	0.367

Table 1 shows the goodness of fit statistics of their models: multiple R, R^2, adjusted R^2 and SEE (Standard Error of Estimation). It can be seen from the table that all the models have very high values of R^2 and low values of SEE. The goodness of the fit statistics table indicates that their models' predictive powers are very high. In describing yarn properties, the strong performance of linear regression models showed that the relationships between their variables (properties of fiber, roving properties, yarn count and twist) and yarn properties were very nearly linear. Their models have found out that roving properties have a significant influence on all properties of the yarn. Their work has shown that AFIS fiber properties can be used effectively for the prediction of yarn properties.[8]

Recently, linear regression modeling technique has been used in a study named "Prediction of new active cases of corona virus disease (COVID-19) pandemic using multiple linear regression models". The COVID-19 pandemic has had a significant effect on the health, socio-economic and financial problems around the world. An overview of the daily statistics of people affected by the disease is taken to forecast the pattern of active cases in Odisha (Indian State) and India over the next few days. Based on the correlations among key features (total confirmed, active, deceased, positive cases) of the considered dataset, regression models are trained to predict future active cases. The score of R^2 tends to be 0.99 and 1.0 indicating a strong predictive model.[9] In this article, the author suggested to predict real estate values using linear regression model. The real estate market is one of the prime fields to apply the ideas of machine learning due to fluctuations in pricing, to forecast costs with high accuracy. The key features to predict the price of a house are physical conditions, concepts and location. A linear regression model has been trained to predict house prices in Mumbai city. The breaking down of past business trends and value ranges, and future advances has been carried out. The outcome of this research has shown that linear regression gives a minimum prediction error of 0.3713. Moreover, this research may help customers bring money into a legacy without switching to a broker.[10]

Another similar research has also been found to estimate the resale value of cars. In this research, the price of the car is considered as dependent variable for target prediction and other parameters are considered as independent variables for target prediction. Most significant parameters have been considered by establishing correlations between intended parameter and price of the car using past data. Furthermore, linear regression modeling technique has been applied to this problem. The model takes training data to be ready to make predictions of car prices. It is providing an accuracy of 90% and giving an error of 10%. Author has observed that linear regression model is better suited for the prediction of target variable (car price) and it is performing very well. Moreover, different machine learning algorithms and approaches can be implemented to get better efficiency.[11] On the other hand, temperature prediction can be an application of linear regression modeling technique. The author considered temperature as the independent variable and pollution, population as dependent variables. After analysis of temperature, population and pollution, predictions of temperature have been made using multiple linear regression models on the basis of various factors in the years 2013-2016. As the outcome of study, the predicted value (temperature) comes out to be accurate. Moreover, measures must be taken to prevent this increasing temperature or it will increase to an uninhabitable extent.[12]

In this paper, author has used linear regression technique for stock price prediction. Since it depends on the demand of the stock, and there is no fixed variable that can accurately predict the demand of one stock each day, Stock price prediction is a difficult task. There are many features influencing demand of stock. One of the features may be people's opinion on social media about products from certain companies, whose analysis can be done through process of sentiment analysis. The results of the sentiment analysis are used to forecast the stock price of the firm. To construct the prediction model, the linear regression approach is used. The sentiment analysis model has been generated with 60.39% accuracy using the Random Forest algorithm classifying tweet data, and the other with 56.50% accuracy with the Naive Bayes algorithm. In price prediction, linear regression models have an R2 value close to 1, which means that a lot of data was fitted by the model.[13] In this study, the landscape position has been used to estimate soil properties using regression methods. This study has been carried out by taking soil samples from Nebraska. Upper interfluves, sampling depth and an irrigation code are the independent variables used to predict dependent variables, i.e., soil properties. Only eight models for pH, organic matter, electrical conductivity, exchangeable K, base saturation percent, and available P and K had major contributions out of the 100 models produced. Such models had R^2 values higher than 0.50. A comparison of the average values observed and projected for each soil property at each sampling depth showed that the values observed typically dropped over the predicted values within a 95% confidence interval.[14]

In this paper, the intended study established regression models for predicting peak hourly load conditional on all previous days during a given day. In forecasting hourly load, it utilizes a companion time series, namely, daily average load. To define the connections between peaks, loads, holidays and weather, the models use diagnostic tests. One that has present weather as well as past average load, past peak load, holiday and weekend dummy variables is the best overall model. Using a non-linear uni-variate model for weather, one reduced form model was also developed and it performed better than any other reduced form model. Another analysis of different models produced in the study showed that conditional models perform no better than those of the reduced form ones. The best overall model mentioned above has given the best efficiency among all.[15] The application in the food industry was disclosed in another study using linear regression modeling techniques. The purpose of the paper is to investigate purchasing trends at fast food restaurants and their relationship to restaurant features, customer characteristics, and calorie data usage. A cross- sectional survey has been conducted in fast food restaurants in New York State. Multiple regression techniques have been used to examine different characteristics including restaurant characteristics (type of fast-food chain, existence of calorie labels, and location poverty), participant characteristics (demographics, calorie information, understanding, and use), and consumer purchasing habits (ordering low-calorie or no beverage, small or no fries, or < 3 items) in order to predict total calories purchased. In order to predict total purchased calories, more than one model were created and the best was selected in the analysis.[16]

Another application of linear regression modeling has been observed in prediction of municipal solid waste generation in China. Municipal solid waste (MSW) management being a serious issue in China, has been provided with an idea of future planning by making predictions of future MSW generation quantity using multiple linear regression model. Urban population, GDP and resident consumption levels are selected as associated variables and their correlations with the quantity of MSW generation are evaluated. The model is established with historic data of 1981- 2011. Based on the estimation of three variables up to year 2030, the MSW generation quantity in China can be projected with the regression model for the next 20 years.[17] In addition to related work, a research has been come out with the impact of measurement errors in partials correlations among variables and multiple linear regression analysis. When documenting precise data observations, measurement errors may be due to a broad intra-individual variance and an adequate number of measurements, or to an incorrect measuring instrument. For estimating the possible attenuation of associations, quantitative methods are derived. The findings suggest that the attenuation of the partial correlation coefficient (or multiple regression coefficients) is higher than that of the simple correlation (or regression) coefficient when the correlated variables have measurement error, but the controlled variables do not. The partial correlation (or regression) coefficients can be either increased or decreased when both the correlated variables and the controlled variables have measurement errors.[18]

Linear regression models are constructed in this paper also to examine the evolving trend in the price of maize and the factors influencing maize prices. The univariate nonlinear and multivariate linear regression models are set up to predict the maize price, respectively, using the data and regression analysis. The univariate linear regression model, that took an independent variable as time and a dependent variable as maize price, passed the tests and was able to predict maize price to some degree. There is, however, a significant error in not recognizing the other influencing factors that may influence the price of maize. The key impact of the supply-demand relationship on the price change of maize has taken place in the multivariate linear regression prediction process, and the model produced will more accurately predict the future maize price.[19] On the other hand, the growing revolution in wind energy promotes more reliable wind speed forecasting models. This research uses a new integrated approach containing ANN (Artificial Neural Networks), Markov chains and linear regression due to very short-term wind speed prediction. First ANN is used in this approach for primary prediction of wind velocity. Then, in the first step, the second-order Markov chain is used to measure the transition probability matrix for the forecast wind speed. Finally, for the final prediction, a linear regression is used between ANN primary prediction and Markov chain determined likelihood. Author has got accurate predictions with fewer errors. The findings are based on real wind speed data with a 2.5 second resolution in a region of Denmark.[20]

In this study, regression techniques have been used to predict salary of a person after a certain year. The graphical representation of salary prediction is a method aimed at creating a computerized system to manage all the daily work in any area of the salary growth graph and can forecast salary after a certain period of time. It will import a graph that helps to observe the graphical representation by checking the salary fields. And then through the prediction algorithm it can forecast a certain time period salary. It can also be used in other powerful predictions as well.[21] In a number of applications, machine learning algorithms have been used. In data mining contexts containing massive datasets and where the environment is poorly known and thus hard to model by humans, they have been found to be of particular value. Such approaches are capable of dealing with significant data sizes, the synthesis of data from multiple databases, and the introduction of context information into the analysis. They are also potentially the most important applicants for broad patient databases to be reviewed. [22]

The topic of how many independent predictor variables can be used in a model of multivariable linear regression has been tested and findings have been obtained. This question is one that is faced in different fields of study by statistical analysts and applied researchers. For specific data, the result of this study is that only two SPV (Subjects per Variable) are required for linear regression models to adequately estimate regression coefficients, standard errors, and confidence intervals. [23] Another analysis arrives with a modern linear regression method called Modal Linear Regression. Provided a set of predictors x, modal linear regression models the conditional mode of an answer Y as a linear function of x. As standard linear regression models the conditional mean (as opposed to mode) of Y as a linear function of x, modal linear regression differs from standard linear regression. [24]

The Multiple Linear Regression research shows that a multilinear regression analysis assumes normality, linearity, no extreme values, and incomplete value analysis. Regression analysis is a mathematical method used to approximate the association between variables with a cause and a consequence relationship. Univariate regression models the relationship between one dependent variable and one independent variable, while multivariate regression models the relationship between one dependent variable and more than one independent variable. [25] Another research shows the power of a possible linear regression analysis with fuzzy data. It has recently become essential to deal with fuzzy data in link of expert expertise. The merits of the suggested formulations are to be able to conveniently obtain fuzzy parameters in possible linear models and to incorporate additional restriction conditions that could be extracted from specialist knowledge of fuzzy parameters. [26]

OBJECTIVES

- To find correlations between each variable and the target variable.
- To build a linear regression model.
- To calculate efficiency of the model and errors in predictions.

PROPOSED METHODOLOGY

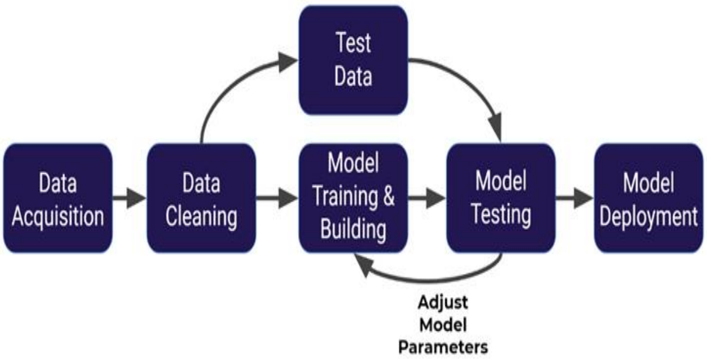


Figure 1: Workflow Pipeline

The various phases of workflow pipeline of data analytics is shown in Figure 1 and is described as follows:

1. **Data Acquisition** – It is the primary stage of data analytics which includes a process of obtaining data. There may be more than one source to obtain the data such as downloading the data files from web, obtaining a real world data through sensors or surveys, or getting real world data from any organization or industry.
2. **Data Cleaning**– It is the process of preparing data for analysis by deleting or altering data which is inaccurate, incomplete, irrelevant, duplicate, or in improper format. Data cleaning

include a variety of tasks such as dealing with missing values, identifying duplicity and removing it, coping with absurd values, and dealing with invalid formats, etc. It is a tedious task which can become interesting and amusing with the help of visualizations of data. Visualizations include various plots like histogram, line plot, bar graph, scatter plot, heat map, strip plot, regression plot, join plot, pair plot and many more.

3. **Test Data**– This stage can also be called as data splitting. There are two common approaches in which data split can take place. One approach is train- test split, in which data split takes place in two portions, training and testingdataset. In fact, the first part is a larger subset of data (such as 80% of the original data) and the second part is typically a smaller subset of data (the remaining 20% of the data). A predictive model is developed using the training set and such a trained model is then applied to the test set to make predictions. Figure 2 shows the pictorial representation of Train- Test split.

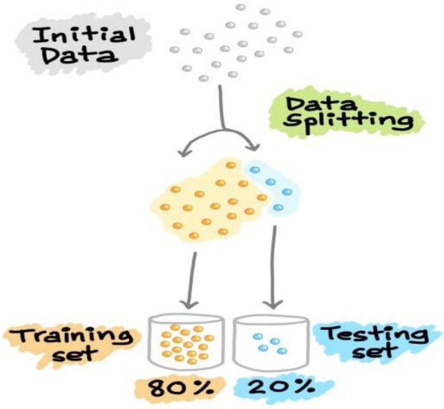
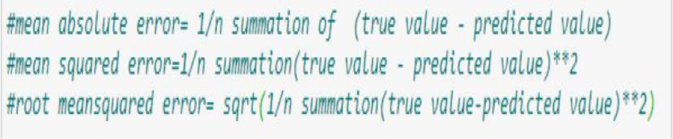


Figure2: Train –Test Split

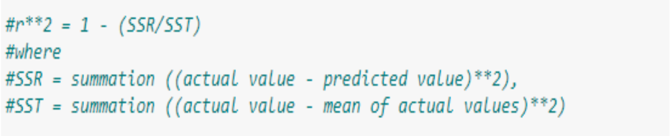
Another approach is train- validation- test split, in which data split is done in 3 portions, training, validation and testing dataset. The training set is used to construct a predictive model and is also tested on the validation set by which predictions are created, model tuning can be made and the best performing model can be chosen based on the validation set results. In model building and planning, the testing dataset is not especially used, but is used to make final predictions.

4. **Model Training & Building**– This stage is where the machine learning algorithms come into existence. Based on the type (i.e., continuous, categorical etc.) of target variable and many other circumstances of the analytical problem, one or more machine learning algorithms are chosen to build one or more models for the same problem. Some basic machine learning algorithms include linear regression, logistic regression, etc. Finally, the built model is trained using the training dataset.
5. **Model Testing**– In this stage, the trained model is tested by identifying errors in predictions. The errors in predictions are calculated by the difference between the actual value and the predicted value (using trained model) of target variable with the help of testing dataset. If Train – Validation – Test split approach has been followed while data splitting then validation set is also applied on the trained model. Formulae for calculation of various kinds of errors are shown in Figure 3.Also, the formula for calculation of R- squared value of model is shown in Figure 4. The value of R- squared lies between 0 and 1. A model is considered to be a good one if the R- squared value is greater than 0.8.



```
#mean absolute error= 1/n summation of  (true value - predicted value)
#mean squared error=1/n summation(true value - predicted value)**2
#root meansquared error= sqrt(1/n summation(true value-predicted value)**2)
```

Figure 3: Formulae for error calculation



```
#r**2 = 1 - (SSR/SST)
#where
#SSR = summation ((actual value - predicted value)**2),
#SST = summation ((actual value - mean of actual values)**2)
```

Figure 4: Formula for R- squared calculation

6. **Model Deployment**– Finally, the tested model is deployed. This model is able to predict the target or the expected variable beforehand.

Linear Regression Model

Linear Regression is a technique of statistical/supervised- machine learning that tries to model the linear relationship between the independent predictor variables X and the dependent variable Y of quantitative response. It is important that the predictor and response variables are numeric. Mathematically, a general linear regression model is depicted in Figure 5.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

$$where\ \beta_i\ is\ coefficient\ and\ \epsilon\ is\ a\ mean - zero\ random\ error\ term$$

Figure 5: Equation of Linear Regression Model

Based on number of predictor variables, linear regression can be classified into two categories:

1. Simple Linear Regression – One predictor variable
2. Multiple Linear Regression – Two or more predictor variable

Software Specification

Tools used - Anaconda Navigator, Jupyter Notebooks, and Microsoft Excel 2013.

Packages used – numpy, pandas, time, datetime, matplotlib, seaborn, sklearn.

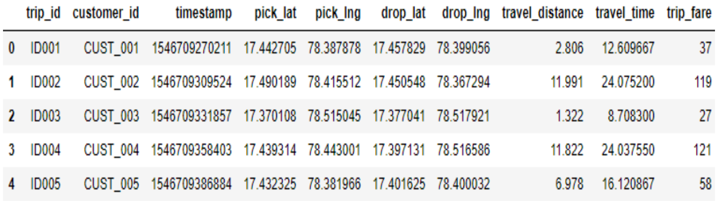
IMPLEMENTATION AND ANALYSIS

In Rapido Dataset's review and implementation process, we divide the whole process into three parts.

1. **Exploratory Data Analysis (EDA) -** It refers to the essential method of conducting initial data investigations in order to uncover patterns, detect anomalies, and use summary statistics and graphical representations to verify conclusions.
2. **Metric Calculation**– It is basically a calculation based question, "What is the average duration between the 1st trip and the 2nd trip of customers?" Only those customers who have taken more than one trip are included here.
3. **Model Building**– It refers to building a model to predict trip fare using travel distance and travel time, measuring the accuracy of the model and using the model to predict trip fare for a trip with travel distance of 3.5 KMs and travel time of 15 minutes.

RESULTS AND DISCUSSION

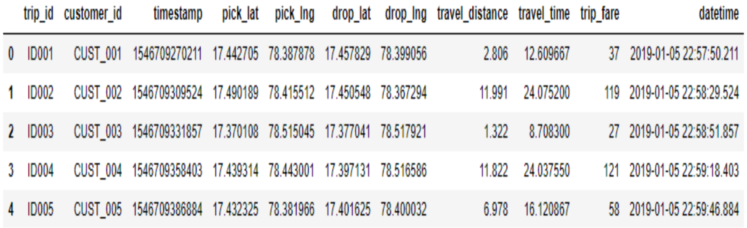
Initially, the dataset is in form of rows and columns with CSV (comma separated value) format. The dataset has been loaded into jupyter notebooks in a dataframe of pandas library. Figure 6 shows a sample of dataset, wherein 'trip_id' and 'customer_id' are unique identifiers for customer, timestamp is the time stamp of the trip in Epoch format, 'pick_lat' and 'pick_lng' are pick up latitude and longitudes, 'drop_lat' and 'drop_lng' are latitude and longitude of drop location, 'travel_distance' is distance of trip measured in KMs, 'travel_time' is the duration of trip measured in minutes, and 'trip_fare' is trip fare calculated in Indian Rupees. Here 'trip_fare' is our target variable.



	trip_id	customer_id	timestamp	pick_lat	pick_lng	drop_lat	drop_lng	travel_distance	travel_time	trip_fare
0	ID001	CUST_001	1546709270211	17.442705	78.387878	17.457829	78.399056	2.806	12.609667	37
1	ID002	CUST_002	1546709309524	17.490189	78.415512	17.450548	78.367294	11.991	24.075200	119
2	ID003	CUST_003	1546709331857	17.370108	78.515045	17.377041	78.517921	1.322	8.708300	27
3	ID004	CUST_004	1546709358403	17.439314	78.443001	17.397131	78.516586	11.822	24.037550	121
4	ID005	CUST_005	1546709386884	17.432325	78.381966	17.401625	78.400032	6.978	16.120867	58

Figure 6: Initial Dataset

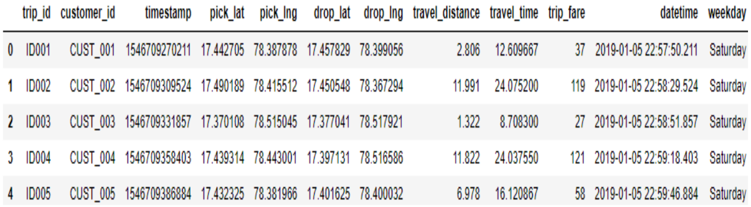
In the EDA of dataset, we explore the data collectively as well as variable by variable in order to check the trends and correlativity of each variable with target variable. While analyzing 'timestamp' variable, we extracted date and time of the trip from it. A new column 'datetime' can be seen in Figure 7.



	trip_id	customer_id	timestamp	pick_lat	pick_lng	drop_lat	drop_lng	travel_distance	travel_time	trip_fare	datetime
0	ID001	CUST_001	1546709270211	17.442705	78.387878	17.457829	78.399056	2.806	12.609667	37	2019-01-05 22:57:50.211
1	ID002	CUST_002	1546709309524	17.490189	78.415512	17.450548	78.367294	11.991	24.075200	119	2019-01-05 22:58:29.524
2	ID003	CUST_003	1546709331857	17.370108	78.515045	17.377041	78.517921	1.322	8.708300	27	2019-01-05 22:58:51.857
3	ID004	CUST_004	1546709358403	17.439314	78.443001	17.397131	78.516586	11.822	24.037550	121	2019-01-05 22:59:18.403
4	ID005	CUST_005	1546709386884	17.432325	78.381966	17.401625	78.400032	6.978	16.120867	58	2019-01-05 22:59:46.884

Figure 7: Dataset with new column 'datetime'

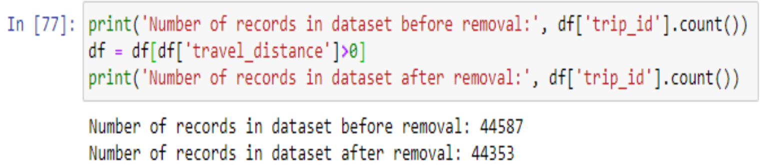
Also, we add a new column namely 'weekday' by extracting week days from 'timestamp' variable. Figure 8 shows the overridden dataset.



	trip_id	customer_id	timestamp	pick_lat	pick_lng	drop_lat	drop_lng	travel_distance	travel_time	trip_fare	datetime	weekday
0	ID001	CUST_001	1546709270211	17.442705	78.387878	17.457829	78.399056	2.806	12.609667	37	2019-01-05 22:57:50.211	Saturday
1	ID002	CUST_002	1546709309524	17.490189	78.415512	17.450548	78.367294	11.991	24.075200	119	2019-01-05 22:58:29.524	Saturday
2	ID003	CUST_003	1546709331857	17.370108	78.515045	17.377041	78.517921	1.322	8.708300	27	2019-01-05 22:58:51.857	Saturday
3	ID004	CUST_004	1546709358403	17.439314	78.443001	17.397131	78.516586	11.822	24.037550	121	2019-01-05 22:59:18.403	Saturday
4	ID005	CUST_005	1546709386884	17.432325	78.381966	17.401625	78.400032	6.978	16.120867	58	2019-01-05 22:59:46.884	Saturday

Figure 8: Dataset with new column 'weekday'

Another finding comes into existence while analyzing 'travel_distance' variable. There are 231 customers whose 'travel_distance' is 0 which might correspond to cancelled trips, for which minimum fare of INR 20 is charged. Also, there are 3 customers whose 'travel_distance' is -1 which might correspond to some error. Both of these observations seem to be absurd, so its better to remove them from the dataset. Figure 9 depicts the number of records before and after removal.

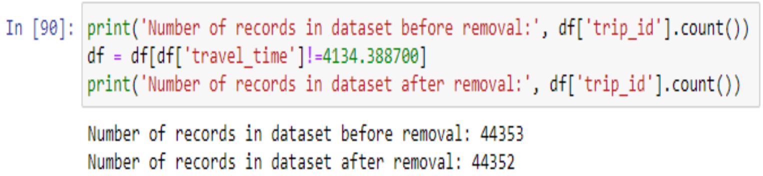


```
In [77]: print('Number of records in dataset before removal:', df['trip_id'].count())
         df = df[df['travel_distance']>0]
         print('Number of records in dataset after removal:', df['trip_id'].count())

         Number of records in dataset before removal: 44587
         Number of records in dataset after removal: 44353
```

Figure 9: No. of records before and after removal

While performing EDA on 'travel_time' variable, we observe an absurd value for 'travel_time', i.e., 4134 minutes having 'trip_fare' of INR 60 and 'travel_distance' of 6.8 KMs. Figure 10 can be seen for the number of rows before and after removal.

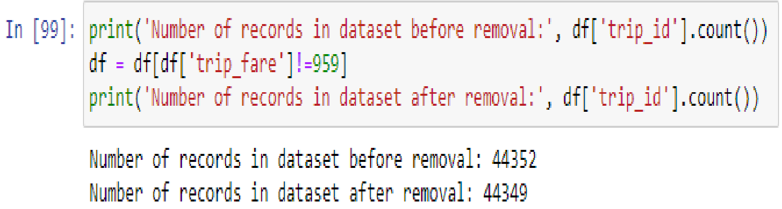


```
In [90]: print('Number of records in dataset before removal:', df['trip_id'].count())
         df = df[df['travel_time']!=4134.388700]
         print('Number of records in dataset after removal:', df['trip_id'].count())

         Number of records in dataset before removal: 44353
         Number of records in dataset after removal: 44352
```

Figure 10: No. of records before and after removal

Also, there are some absurd values for 'trip_fare' if we look at the corresponding values of 'travel_distance' and 'travel_time'. For Example, 'trip_fare' INR 959 seems to be absurd in some cases. The number of rows before and after removal can be seen in Figure 11.



```
In [99]: print('Number of records in dataset before removal:', df['trip_id'].count())
         df = df[df['trip_fare']!=959]
         print('Number of records in dataset after removal:', df['trip_id'].count())

         Number of records in dataset before removal: 44352
         Number of records in dataset after removal: 44349
```

Figure 11: No. of records before and after removal

We also check for null values and the results can be seen in Figure 12. It is clear that *t*here are no missing values in the dataset. So further we need not to clean the data for model building as it is already having no null values and no categorical column of object type.

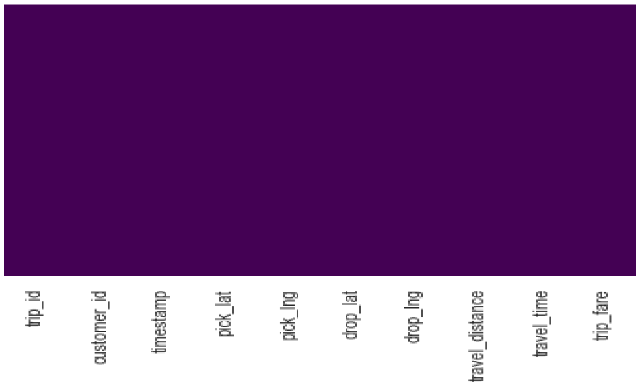


Figure 12: Heatmap for null values

Based on the analysis, 'travel_distance' and 'travel_time' are the most suitable variables to predict 'trip_fare'. First reason of selecting these two is that both are numeric as required for linear regression model, and second can be seen from the heatmap showing the strong correlations in Figure 13.

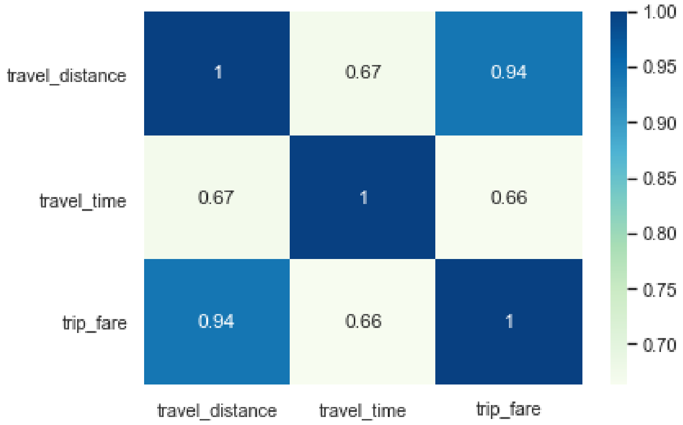
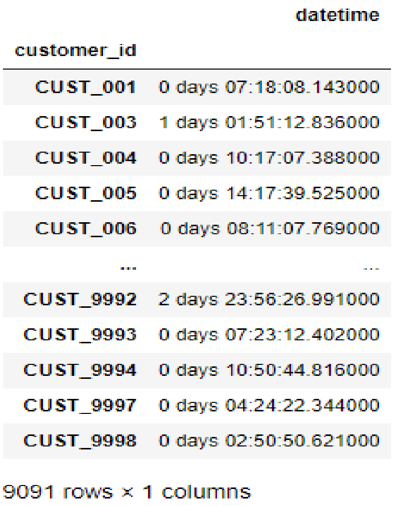


Figure 13: Heatmap for correlations with target variable

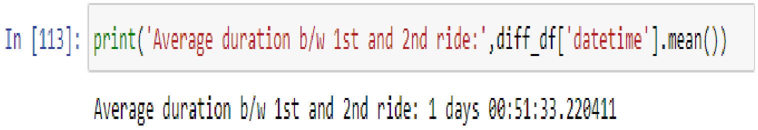
In the metric calculation for average duration between 1st and 2nd trip of customers, we first calculate the duration between 1stand 2nd trip of each customer (who has travelled more than once) individually as shown in Figure 14. Then we calculate the average of these durations as shown in Figure 15.



	datetime
customer_id	
CUST_001	0 days 07:18:08.143000
CUST_003	1 days 01:51:12.836000
CUST_004	0 days 10:17:07.388000
CUST_005	0 days 14:17:39.525000
CUST_006	0 days 08:11:07.769000
...	...
CUST_9992	2 days 23:56:26.991000
CUST_9993	0 days 07:23:12.402000
CUST_9994	0 days 10:50:44.816000
CUST_9997	0 days 04:24:22.344000
CUST_9998	0 days 02:50:50.621000

9091 rows × 1 columns

Figure 14: Duration between 1st and 2nd trip for each customer



```
In [113]: print('Average duration b/w 1st and 2nd ride:',diff_df['datetime'].mean())

          Average duration b/w 1st and 2nd ride: 1 days 00:51:33.220411
```

Figure 15: Average duration between 1st and 2nd trip

Now in the model building phase of the process, we build a linear regression model taking 'travel_distance' and 'travel_time' as independent variables to predict 'trip_fare'. Figure 16 shows the scatter plot of true values v/s predicted values (predicted by built model) of trip fare. It is clear that plot of true and predicted values comes out to be almost linear, which is a good sign.

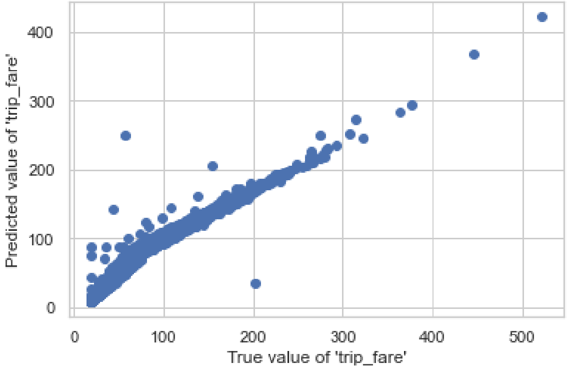
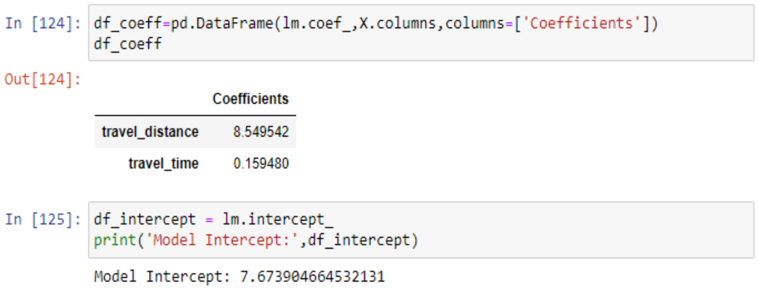


Figure 16: Scatter plot of true values v/s predicted values

The intercept and coefficients of independent variables can be seen in Figure 17.



```
In [124]: df_coeff=pd.DataFrame(lm.coef_,X.columns,columns=['Coefficients'])
          df_coeff

Out[124]:
```

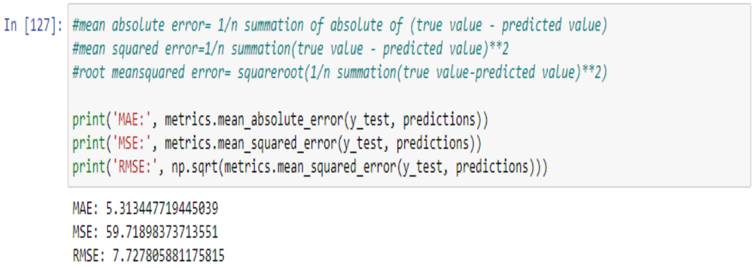
	Coefficients
travel_distance	8.549542
travel_time	0.159480

```
In [125]: df_intercept = lm.intercept_
          print('Model Intercept:',df_intercept)

          Model Intercept: 7.673904664532131
```

Figure 17: Intercept and coefficients of model

Figure 18 is showing various kinds of errors calculated for the model we built.



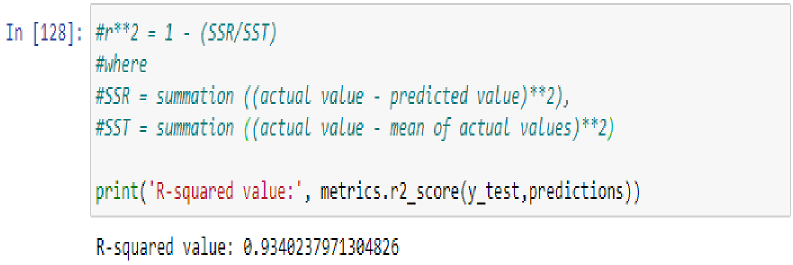
```
In [127]: #mean absolute error= 1/n summation of absolute of (true value - predicted value)
          #mean squared error=1/n summation(true value - predicted value)**2
          #root meansquared error= squareroot(1/n summation(true value-predicted value)**2)

          print('MAE:', metrics.mean_absolute_error(y_test, predictions))
          print('MSE:', metrics.mean_squared_error(y_test, predictions))
          print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))

          MAE: 5.313447719445039
          MSE: 59.71898373713551
          RMSE: 7.727805881175815
```

Figure 18: Calculated errors

The R- squared value for the model is shown in Figure 19. It is clear that the efficiency of our model comes out to be 93.40%.



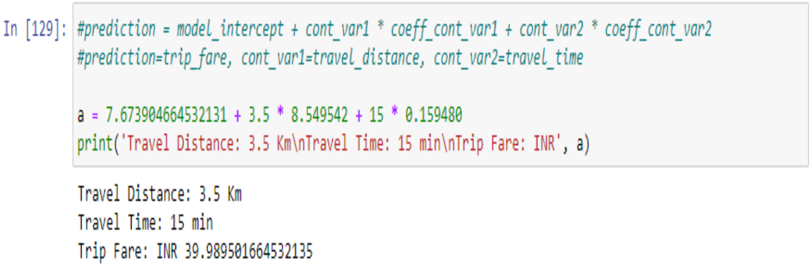
```
In [128]: #r**2 = 1 - (SSR/SST)
          #where
          #SSR = summation ((actual value - predicted value)**2),
          #SST = summation ((actual value - mean of actual values)**2)

          print('R-squared value:', metrics.r2_score(y_test,predictions))

          R-squared value: 0.9340237971304826
```

Figure 19: R- squared value of the model

In the last part of process, trip fare for distance of 3.5 KMs and duration of 15 minutes has been predicted by the model as shown in Figure 20.



```
In [129]: #prediction = model_intercept + cont_var1 * coeff_cont_var1 + cont_var2 * coeff_cont_var2
          #prediction=trip_fare, cont_var1=travel_distance, cont_var2=travel_time

          a = 7.673904664532131 + 3.5 * 8.549542 + 15 * 0.159480
          print('Travel Distance: 3.5 Km\nTravel Time: 15 min\nTrip Fare: INR', a)

          Travel Distance: 3.5 Km
          Travel Time: 15 min
          Trip Fare: INR 39.989501664532135
```

Figure 20: Final prediction

CONCLUSION AND FUTURE WORK

Rapido is a Ride-On-Demand (ROD) service that uses dynamic pricing to balance supply and demand in an attempt to increase service quality. The research focuses primarily on exploratory data analysis (EDA), metric calculation, and development of model. The dependent algorithm variable turned out to be 'ride fare,' while 'travel distance' and 'travel time' are the independent variables. The Linear Regression Model for prediction of dynamic price of trips is providing an efficiency of 93.40%. It is better suited for the prediction of target variable which is trip fare, and it performs very well. Further this work can be carried out using different machine learning algorithms and techniques in order to get higher efficiency and lower errors.

REFERENCES

[1] Mahajan, Ginika, and Bhavna Saini. "Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM." *International Journal of Computer Applications & Information Technology* 12, no. 1 (2020): 310-316.

[2] Sharma, Manik, Samriti Sharma, and Gurvinder Singh. "Performance analysis of statistical and supervised learning techniques in stock data mining." *Data* 3, no. 4 (2018): 54.

[3] Gautam, Ritu, and Manik Sharma. "Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis." *Journal of medical systems* 44, no. 2 (2020): 1-24.

[4] Zekić-Sušac, Marijana, Saša Mitrović, and Adela Has. "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities." *International journal of information management* (2020): 102074.

[5] Malik, Monica, Sameena Naaz, and Iffat Rehman Ansari. "Sentiment Analysis of Twitter Data Using Big Data Tools and Hadoop Ecosystem." In *International Conference on ISMAC in Computational Vision and Bio-Engineering*, pp. 857-863. Springer, Cham, 2018.

[6] Biradar, Abhilash, and S. G. Totad. "Detecting Depression in Social Media Posts Using Machine Learning." In International Conference on Recent Trends in Image Processing and Pattern Recognition, pp. 716-725. Springer, Singapore, 2018.

[7] Nguyen, Dong, Noah A. Smith, and Carolyn Rose. "Author age prediction from text using linear regression." In Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities, pp. 115-123. 2011.

[8] Ureyen, Mustafa E., and HueseyinKadoglu. "The prediction of cotton ring yarn properties from AFIS fibre properties by using linear regression models." Fibres and Textiles in Eastern Europe 15, no. 4 (2007): 63.

[9] Rath, Smita, AlakanandaTripathy, and AlokRanjanTripathy. "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, no. 5 (2020): 1467-1474.

[10] Ghosalkar, Nehal N., and Sudhir N. Dhage. "Real estate value prediction using linear regression." In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-5. IEEE, 2018.

[11] Kiran, S. "Prediction of Resale Value of the Car Using Linear Regression Algorithm."

[12] Menon, Sindhu P., RamithBharadwaj, Pooja Shetty, Prajwal Sanu, and Sai Nagendra. "Prediction of temperature using linear regression." In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pp. 1-6. IEEE, 2017.

[13] Cakra, YahyaEru, and BayuDistiawanTrisedya. "Stock price prediction using linear regression based on sentiment analysis." In *2015 international conference on advanced computer science and information systems (ICACSIS)*, pp. 147-154. IEEE, 2015.

[14] Brubaker, S. C., A. J. Jones, K. Frank, and D. T. Lewis. "Regression models for estimating soil properties by landscape position." *Soil Science Society of America Journal* 58, no. 6 (1994): 1763-1767.

[15] Engle, Robert F., Chowdhury Mustafa, and John Rice. "Modelling peak electricity demand." *Journal of forecasting* 11, no. 3 (1992): 241-251.

[16] Brissette, Ian, Ann Lowenfels, Corina Noble, and Deborah Spicer. "Predictors of total calories purchased at fast-food restaurants: restaurant characteristics, calorie awareness, and use of calorie information." *Journal of nutrition education and behavior* 45, no. 5 (2013): 404-411.

[17] Wei, Yuanwei, YaliXue, Jiongyu Yin, and Weidou Ni. "Prediction of municipal solid waste generation in China by multiple linear regression method." *International Journal of Computers and Applications* 35, no. 3 (2013): 136-140.

[18] Liu, Kiang. "Measurement error and its impact on partial correlation and multiple linear regression analyses." *American Journal of Epidemiology* 127, no. 4 (1988): 864-874.

[19] Ge, Yan, and Haixia Wu. "Prediction of corn price fluctuation based on multiple linear regression analysis model under big data." *Neural Computing and Applications* (2019): 1-13.

[20] Kani, SA Pourmousavi, S. M. Mousavi, A. KashefiKaviani, and G. H. Riahy. "A new integrated approach for very short-term wind speed prediction using linear regression among ANN and Markov Chain." In *Proceeding on International Conference on Power System Analysis, Control and Optimization*. 2008.

[21] Das, Sayan, RupashriBarik, and Ayush Mukherjee. "Salary Prediction Using Regression Techniques." *Available at SSRN 3526707* (2020).

[22] Meyfroidt, Geert, Fabian Güiza, Jan Ramon, and Maurice Bruynooghe. "Machine learning techniques to examine large patient databases." *Best Practice & Research Clinical Anaesthesiology* 23, no. 1 (2009): 127-143.

[23] Austin, Peter C., and Ewout W. Steyerberg. "The number of subjects per variable required in linear regression analyses." *Journal of clinical epidemiology* 68, no. 6 (2015): 627-636.

[24] Yao, Weixin, and Longhai Li. "A new regression model: modal linear regression." *Scandinavian Journal of Statistics* 41, no. 3 (2014): 656-671.

[25] Uyanık, Gülden Kaya, and Neşe Güler. "A study on multiple linear regression analysis." *Procedia-Social and Behavioral Sciences* 106 (2013): 234-240.

[26] Tanaka, Hideo, Isao Hayashi, and Junzo Watada. "Possibilistic linear regression analysis for fuzzy data." *European Journal of Operational Research* 40, no. 3 (1989): 389-396.