ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 3| March 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

SUPERVISED LEARNING-BASED RIDE FARE PREDICTION

Akhilash Pennam[1], Pavan Kumar Potula[2*]

[1] Director of CRM Development,

[2]Assistant Professor, EEE Department

**Corresponding Author*

Article DOI: https://doi.org/10.36713/epra20759
DOI No: 10.36713/epra20759

ABSTRACT

One of the most well-known ride-sharing services worldwide is Uber. This Uber price prediction system will precisely forecast the cost of a ride by combining machine learning algorithms with past data in order to give clients the best service possible. Taxi services are currently the most popular means of transportation. Rapid change has occurred in corporations, and they are now moving toward digital innovation. Historically, software companies and product developers have developed a number of methods, but they haven't considered the necessity of a customer's mobility in a certain location. In order to develop a precise model for forecasting future pricing, the Uber price prediction system will examine historical trip data, traffic, weather, time of day, and other pertinent information. To produce its predictions, it will employ a range of methods including linear regression.

KEY WORDS: *Uber, Supervised Machine Learning, Business, Price Prediction.*

1. INTRODUCTION

This chapter provides a brief overview of machine learning, the algorithm used, and its application in predicting Uber ride fares. The prediction is based on multiple factors, including passenger count, trip distance, drop-off longitude, and total distance traveled. We employ a supervised machine learning method—linear regression—to train the system and achieve accurate pricing predictions. This trained model is then used to forecast the cost of an Uber journey.The fare of an Uber ride depends on several factors such as the distance between the source and destination, time of day, weather conditions, wind speed, dew point, visibility, latitude, and longitude. To enhance prediction accuracy, we also analyze correlations between different features in the dataset, identifying key patterns that influence ride costs.By leveraging machine learning, this system aims to provide precise, real-time fare estimations, assisting both passengers and drivers in making informed decisions about ride pricing.

1.1. Introduction to Machine learning

Binary and distributive issues are two subtypes of classification problems in machine learning. Machine learning, a branch of computer science and artificial intelligence (AI), focuses on utilizing data and algorithms to replicate how individuals learn, gradually improving the system's accuracy. The three primary types of machine learning approaches are supervised learning, autonomous learning, and reinforcement learning. We'll use the supervised learning approach. We employ supervised learning method, such as linear regression.

Linear Regression

Linear regression is one of the simplest and most widely used algorithms in machine learning. It's a supervised learning algorithm used to predict a continuous target variable based on one or more input features. The relationship between the input variables and the target variable is assumed to be linear. The objective is to find the best-fitting straight line (or hyperplane, in the case of multiple features) that minimizes the error between the predicted and actual values.

Mathematical Formulation:

In the case of simple linear regression, the relationship between the input feature XX and the target variable YY is expressed as:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

Where:

- Y is the dependent or target variable (what you're trying to predict),
- X is the independent variable or feature (input data),
- β0 is the intercept (the value of Y when X = 0),
- β1is the slope (how much Y changes when X changes by one unit),
- ϵ represents the error term (the difference between the observed and predicted values).

For multiple linear regression, where there are multiple features, the equation extends to:

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

Where X1,X2,…,Xn represent multiple input features.

Assumptions

Linear regression makes several assumptions that are important for its effectiveness:

1. Linearity: The relationship between the dependent and independent variables is assumed to be linear. If this assumption is violated, the predictions might not be accurate.
2. Independence of Errors: The errors (residuals) are assumed to be independent of each other. This is

important for the validity of statistical tests and confidence intervals.

3. Homoscedasticity: The variance of the residuals (errors) should be constant for all levels of the independent variables.
4. Normality of Errors: The errors should be approximately normally distributed. This assumption is important for hypothesis testing and constructing confidence intervals.

1.2. Benefits of Uber

Drivers can charge fees and receive payment through the Uber app, which also allows passengers to request rides. More precisely, independent contractors are employed as drivers by Uber, a ride-sharing business.

- **Convenient & Cashless**: With the e-hail app, users can hail a car from anywhere and have it arrive in a matter of minutes, eliminating the need to call and wait for a car service or chase down a taxi on the street. You are not even need to provide an address to Uber.
- **Professional Service**: Since Uber drivers drive their own vehicles, they are kept clean and in good condition.
- **Competitive Rates**: It is difficult for Uber to keep one and definite price for the ride. The price of the Uber ride changes according to surge multiplier.
- **Safety & Flexibility**: The Uber services are even available at night for the customers.

2. LITERATURE REVIEW AND RELATED WORK

A study report on a cab fare prediction system utilizing key feature extraction of artificial intelligence was published by Dr. Balika J. Chelia, Jai Singh, Devansh Chaturvedi, and Avinash Kumar Singh [1]. The "Journey Sharing" service reduces the overall demand for cars on the road by matching passengers. Yet, there is a significant drawback. Expensive and inappropriate for wide bandwidth or long distances. This article provides a scientific definition of the aforementioned difficulties. For example, the system provides information on travel demands to appropriately index taxi requests. She takes into account the taxi score, which is determined by a number of factors, including the location, point of origin, and final destination, while choosing the ideal building. A sizable amount of ratings show how accurate the suggested system is. Requests can be handled by the system in a matter of milliseconds. It makes the deployment simpler than with other providers.

Jeremy P Toner presented the research paper on the demand for taxis and the value of time [5]– welfare analysis. This study evaluated the value of taxi customers' walking, waiting, and in-vehicle times, as well as the elasticities of customer demand for taxis in terms of price and waiting time. The study's findings are presented in this document. Using expressed preference and transfer pricing approaches, the data were collected. Thus, of the three main sections of the report, the first deals with the conception and analysis of the expressed preference trial into the value of time, and the second employs transfer price information to recalibrate demand models. The social welfare study of the town's current pricing and quantity restrictions is presented in the third main part.

Ahmed.G.Kvazimpublishedastudytitled Thearticle"Taxi Fare PredictionUtilising Multilayer a perception system and Radial Basis" [7]. To forecast taxi fares, several prediction algorithms have been created. Some of these algorithms merely take into account the distance between the pickupand drop-off locations, while others also take the number of people and trip duration into account. The standard errors that resulted from the poor accuracy of these models' predictions varied from $2 to $4. Additionally, this problem has not been modelled using artificial neural networks. This study aims to develop a neural network-based system that takes into consideration the trip distance and other pertinent information in order to accurately predict the amount of fare for taxis in New York City. Both the radial based function network (RBFN) and the multi-layer perceptron are constructed; the former is done purely on the trip's journey distance, while the one that follows is accomplished by considering the trip's journey distance, the total number of passengers, the moment of day, the period of the week, and the month of the journey (temporal factors). The resultsshow that when the journey distance as well as additional influencing factors were included in the models, neural networks were effective in correctly modelling the taxi fare with a mean-square error < 0.005.

Ankit Kumar and Vishal Shah presented their research on A method using unsupervised machine learning to predict the cost of taxi rides [4]. In this research, a machine learning- based model that can predict the demand for taxis in various geographic locations of a city by optimising the within- cluster sum of squared distances has been suggested. In order to anticipate the cost of transportation from one fixed place to another fixed location depending on the time and location of booking, a reliable and accurate pricingprediction model has been built. Keywords K-means clustering for unsupervised learning Price prediction for a taxijourney. AlbertoRossi,GianniBarlacchi,M.Bianchini,B. Lepri of Computer Science has proposed the research paper on Modelling Taxi Drivers' Behavior for the Next Destination Prediction [3].A Recurrent Neural Network (RNN) approach is presented that models the taxi drivers' behavior and encodes the semantics of visited locations by using geographical information from Location – Based Social Networks (LBSNs).

3. SYSTEM ANALYSIS

We will cover the numerous trials conducted to identify the most precise model for estimating the cost of an Uber journey in this chapter. We will talk about the issue and the system we are developing to address it. We'll work with a supervised machine learning technique, such as linear regression.

Existing Problem

Uber is a transportation company with an app that allows passengers to hail a ride and drivers to charge fares and get paid. The existing problem of our project is how to calculate the most accurate price of a ride in Uber depending on the different parameters like destination etc.

The Suggested Repair

In order to develop a precise model for forecasting future pricing, the Uber price prediction system will examine historical trip data, congestion, the weather, time of day, and other pertinent information. To evaluate the model's accuracy ,a number of methods will be used, including linear regression. In the future, the price will be determined using the most precise algorithm.

4. METHODOLOGY

4.1. Defining the Problem

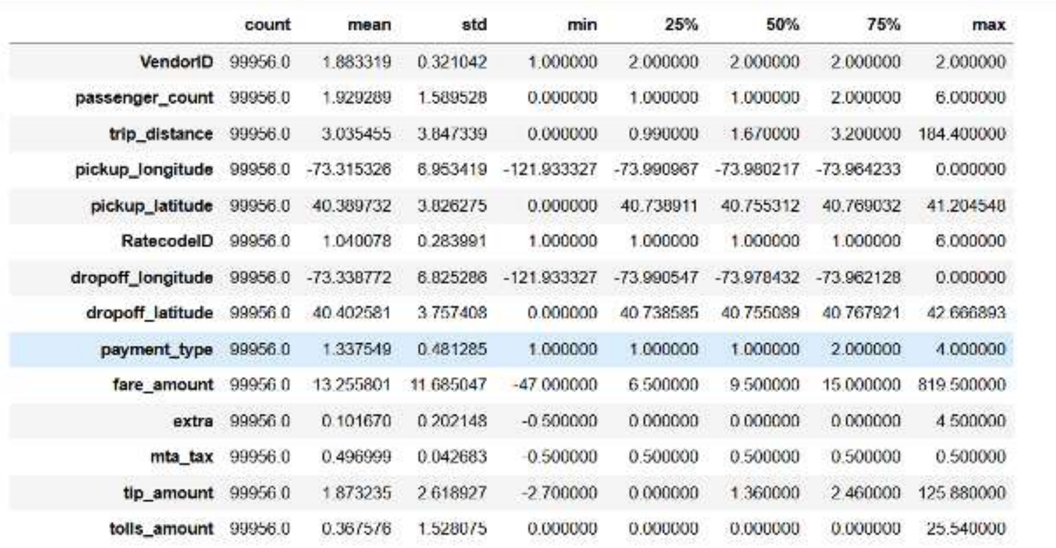
In order to develop a precise model for forecasting future pricing, the Uber price prediction system will examine historical trip data, traffic, weather, the moment in the day, and other pertinent information. To create its predictions, a supervised machine learning model such as linear regression is used.

4.2. Data Collection

With the aid of Kaggle, we are gathering the data for our research. Kaggle is a platform that is open-source that may be used for free to gather data for machine learning.

4.3. Exploratory Data Analysis

The figure contains a statistical summary of a dataset, showing various columns with their count, mean, standard deviation (std), minimum (min), 25th percentile (25%), median (50%), 75th percentile (75%), and maximum (max) values.



	count	mean	std	min	25%	50%	75%	max
VendorID	99956.0	1.883319	0.321042	1.000000	2.000000	2.000000	2.000000	2.000000
passenger_count	99956.0	1.929289	1.589528	0.000000	1.000000	1.000000	2.000000	6.000000
trip_distance	99956.0	3.035455	3.847339	0.000000	0.990000	1.670000	3.200000	184.400000
pickup_longitude	99956.0	-73.315326	6.953419	-121.933327	-73.990967	-73.980217	-73.964233	0.000000
pickup_latitude	99956.0	40.389732	3.826275	0.000000	40.738911	40.755312	40.769032	41.204548
RatecodeID	99956.0	1.040078	0.283991	1.000000	1.000000	1.000000	1.000000	6.000000
dropoff_longitude	99956.0	-73.338772	6.825286	-121.933327	-73.990547	-73.978432	-73.962128	0.000000
dropoff_latitude	99956.0	40.402581	3.757408	0.000000	40.738585	40.755089	40.767921	42.666893
payment_type	99956.0	1.337549	0.481285	1.000000	1.000000	1.000000	2.000000	4.000000
fare_amount	99956.0	13.255801	11.685047	-47.000000	6.500000	9.500000	15.000000	819.500000
extra	99956.0	0.101670	0.202148	-0.500000	0.000000	0.000000	0.000000	4.500000
mta_tax	99956.0	0.496999	0.042683	-0.500000	0.500000	0.500000	0.500000	0.500000
tip_amount	99956.0	1.873235	2.618927	-2.700000	0.000000	1.360000	2.460000	125.880000
tolls_amount	99956.0	0.367576	1.528075	0.000000	0.000000	0.000000	0.000000	25.540000

Fig1: Statistical Summary of a Dataset.

The correlation heat map visually represents the relationships between different numerical variables in the dataset, with colors indicating the strength and direction of correlations. Lighter shades represent strong positive correlations, while darker shades indicate negative correlations. One of the strongest positive correlations is observed between trip distance and fare amount, suggesting that longer trips result in higher fares. Similarly, tolls and extra charges show a positive relationship with fare amount, indicating that additional costs contribute significantly to the total fare. There is also a noticeable correlation between pickup and drop-off latitude/longitude, which reflects the geographic pattern of trips. However, variables such as Vendor ID and passenger count show weak or no correlation with most other variables, implying they have minimal influence on fare amount or trip distance. The heat map highlights key relationships that could be valuable for predictive modeling, particularly in forecasting fares based on trip distance and additional charges.

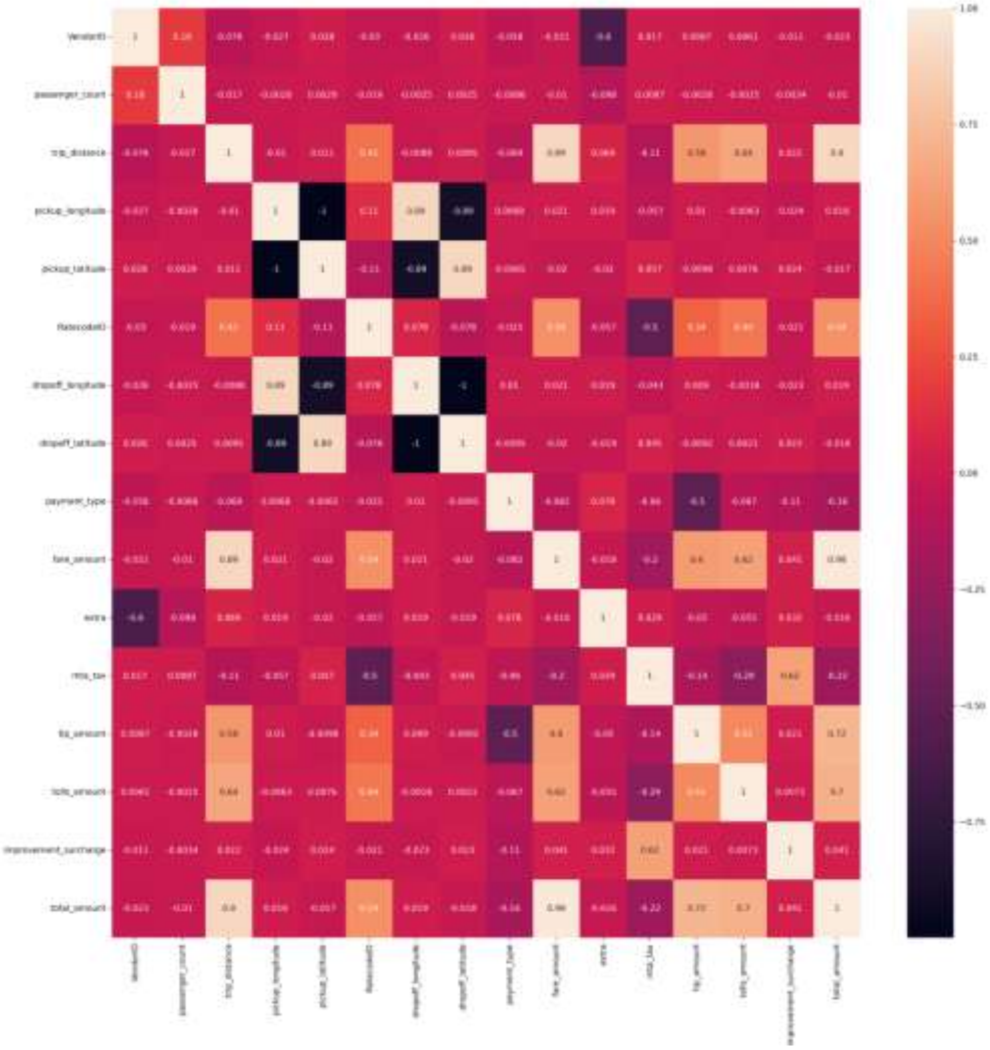
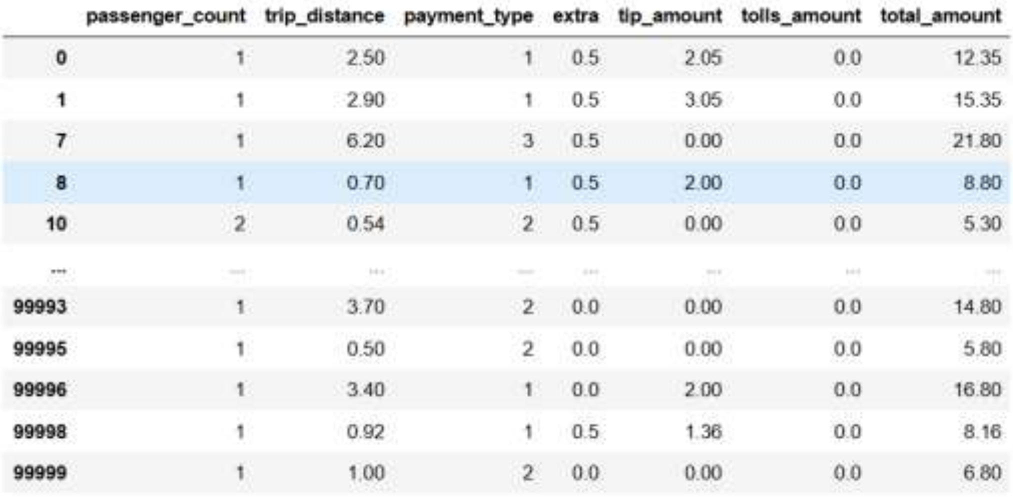


Fig2: Correlation Heat Map

4.4. Feature Engineering

Variance inflation factor (VIF) analysis is commonly used to eliminate the unwanted features from the given dataset. In this work, a VIF analysis is adapted to eliminate the unwanted features. After applying the VIF analysis, only few features the dataset has contained and which was shown in figure3.



	passenger_count	trip_distance	payment_type	extra	tip_amount	tolls_amount	total_amount
0	1	2.50	1	0.5	2.05	0.0	12.35
1	1	2.90	1	0.5	3.05	0.0	15.35
7	1	6.20	3	0.5	0.00	0.0	21.80
8	1	0.70	1	0.5	2.00	0.0	8.80
10	2	0.54	2	0.5	0.00	0.0	5.30
...	...	...	...	...	...	...	...
99993	1	3.70	2	0.0	0.00	0.0	14.80
99995	1	0.50	2	0.0	0.00	0.0	5.80
99996	1	3.40	1	0.0	2.00	0.0	16.80
99998	1	0.92	1	0.5	1.36	0.0	8.16
99999	1	1.00	2	0.0	0.00	0.0	6.80

Fig3: Statistical Summary of a dataset after VIF analysis

It is described as a process for transforming the initial data set into characteristics. It is used to enhance the performance and accuracy of machine learning models.

5. RESULTS AND DISCUSSION

The regression model's performance evaluation indicates strong predictive capability, as reflected in an R² score of 0.87, meaning 86.6% of the variance in the target variable (likely fare amount) is explained by the model. The Mean Absolute Error (MAE) of 1.33 suggests that, on average, the model's predictions deviate from the actual values by $1.33, while the Root Mean Squared Error (RMSE) of 1.80 further supports this by providing a typical error magnitude in the same unit as the target variable. The Mean Squared Error (MSE) of 3.24, though not as interpretable directly, emphasizes that larger errors have a stronger impact. Overall, the model demonstrates high accuracy with relatively low prediction errors, making it effective for fare estimation. However, further refinements, such as feature engineering or hyperparameter tuning, could enhance precision even more.

6. CONCLUSION AND FUTURE SCOPE

Uber is one of the most well-known ride-sharing services worldwide. In order to give clients the best experience possible, this Uber pricing prediction system will accurately estimate the cost of a ride by combining historical data with a machine learning algorithm. The Uber price forecasting algorithm will look at past trip data, traffic, the weather, the time of day, and other relevant factors to create an accurate model for predicting future prices. It uses a regression-based model to generate its predictions.

We are wrapping up the following features in this, such as the maximum number of trips by category, such as business. Maximum round trips are in which month whether the trip is meeting its requirement. Then booking of flight is least in which month like September these all the features we are concluding with the help of feature engineering. The future scope of this project is that we have used the supervised machine learning algorithms which we were not previously used with the help of which we can predict the price more accurately on the various factors. The additional feature has enhanced our project. We can tell where we require large number of ride at a particular time so that we can provide the customers with the better experience.

REFERENCES

1. *DevanshChaturvedi,JaiSingh,AvinashKumarSingh, and Dr. Balika J. Chelliah (2021) are the first. System for Predicting Taxi Fare Using Artificial Intelligence's Key Feature Extraction.*
2. *Gaurav Hajela, Pawel Pratyush, and AkshataGangrade (2021). using dynamic spatial and temporal analysis to anticipate demand for taxis. Alberto Rossi, Gianni Barlacchi, M. Bianchini, and B. Lepri (2018)*
3. *ModellingtheBehaviour ofTaxiDriverstoPredictthe Next Destination. Vishal Shah and Ankit Kumar(2022)*
4. *An unsupervised artificial intelligence method for tax price prediction. Jeremy P. Toner (1991)*
5. *TheValueOfTimeAndTheDemandForTaxis.MajidKhedmati and Mohammad Fili (2020).*
6. *Forecasting town visits using data mining methods. Ahmed. G. Quasim (2020)*
7. *Using Multilayer Perceptrons and Radial Basis, onecan predict taxi fares*

.

.