A Supervised Learning Algorithm for Cab Fare Prediction

S. Sushma Priyanka[1], G.V.L. Sai Roshini[2], SD.Sabeera Begum[3], S.N. Tirumala Rao[4]

[1,2,3]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[4]Head of the Department, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

priyankasushma13@gmail.com[1], roshinigrandhe@gmail.com[2], sabeerasayyad07@gmail.com[3], nagatirumalarao@gmail.com[4]

1. ABSTRACT- This is a predictive machine learning model. In this project, my objective was to predict the fare of a cab based on different factors mainly distance and time. I applied the process of data cleaning, data merging, and data visualization. Calculating the distance between the source and destination by using the haversine formula is the main key to predicting the cab fare. I implemented some ML regression algorithms and tested the score of the model, with that score we can finalize the machine learning model. Major apps like Ola, Uber, etc are introduced to cab rides in the major cities, but this research helps those who live in suburban areas. The study is based on supervised learning.

KEYWORDS: Supervised Learning, Outlier Analysis, Machine Learning, Price Prediction, Predictive Analysis.

2. INTRODUCTION

Machine learning is the process through which a computer may automatically learn from data, improve performance based on previous experiences, and generate predictions. Machine learning uses a variety of algorithms that work with enormous volumes of data. Data is used to training these algorithms, and after training, they create a model and perform a certain task. There are 3 different categories:

- Supervised learning: Throughout the machine learning phase, supervision is necessary. It contains both the input and the desired output, and the model is set up to forecast the desired outcome.
- Unsupervised learning: The model learns on its own by seeing patterns in the dataset without the need for supervision. Just input is provided, the model self-trains, and output results. Clustering and association are two examples.
- Reinforcement learning: The model is created utilizing the hit-and-miss approach in this type of learning. Its nature is reliant. Its input is the result of the previous operation. For instance, puzzle chess.

We used the supervised learning approach in this work because it best meets the requirements of predictive analysis.

Predictions are produced based on past data acquired, and when the model is taught to deal with novel input and anticipate the predicted result, predictions are generated based on fresh data.

Today's taxi services are growing at a multiplication pace. In the current study, there was a lot of focus on cab fare prediction. Customers get a terrific experience with affordable costs thanks to how simple and flexible the services are to use. the steps this strategy involves.

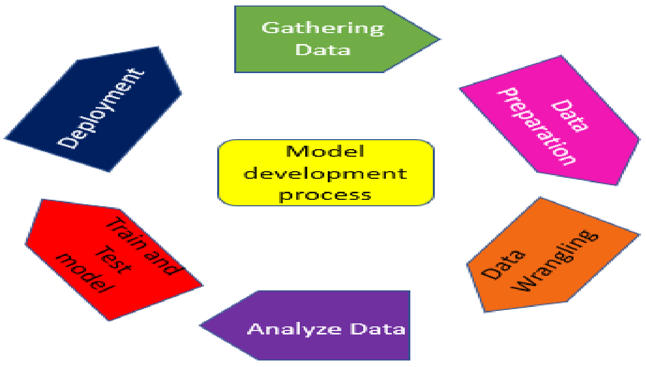


Fig.1 6 steps in model building

Many data mining algorithms have been developed as a result of data mining research. These algorithms may be applied directly to a dataset to build models or to derive important conclusions and inferences from it. Decision tree regression, linear regression, and random forest regression are some common data mining methods.

3.LITERATURE SURVEY

Understanding how the organization plans to use and benefit from this model is crucial before examining the data. This initial round of brainstorming aids in selecting the best algorithms, framing the challenge, and evaluating their effectiveness.

Although they frequently use the same techniques and have a lot in common, data mining and machine learning (ML) focus on different things. The goal of data mining is to discover previously undiscovered qualities in the data, whereas machine learning (ML) focuses on prediction utilizing prior attributes learned from training data. This is referred to as knowledge discovery (KDD) in databases [1].

Using various graphs, charts, plots, etc., data visualization makes it easier to understand the data. So that the data may be adequately examined, it is here depicted using a scatter plot, bar graph, and histogram. These analyses were all completed in JupyterLab [2].

The process of preparing the data comes after choosing a method. Data exploration, data rectification, and data visualization using graphs and charts are all terms used to describe the process of looking at data. A common name for this is Exploratory Data Analysis (EDA) [3].

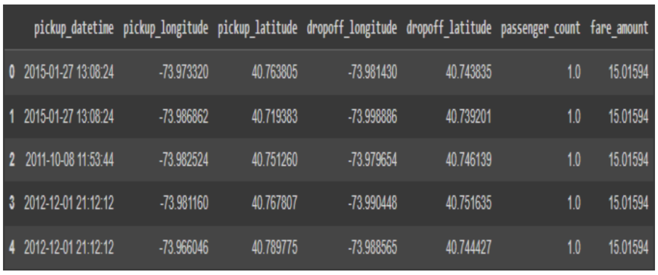
4. PROPOSED SYSTEM

4.1 Data preprocessing

Typical noises, missing numbers, and sometimes an inappropriate arrangement make significant-world data unsuitable for machine-learning algorithms.

4.1.1 Dataset Description

The website of Kaggle, which has around 7 columns, was used for data gathering. These columns are the fee, the pickup and drop-off locations, the pickup and drop-off longitudes and latitudes, and the number of passengers. There were around 5 million rows of data in there as well. We used about 80,000 rows from that data, spanning the years 2009 to 2016. The following data view is provided:



	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	fare_amount
0	2015-01-27 13:08:24	-73.973320	40.763805	-73.981430	40.743835	1.0	15.01594
1	2015-01-27 13:08:24	-73.986862	40.719383	-73.998886	40.739201	1.0	15.01594
2	2011-10-08 11:53:44	-73.982524	40.751260	-73.979654	40.746139	1.0	15.01594
3	2012-12-01 21:12:12	-73.981160	40.767807	-73.990448	40.751635	1.0	15.01594
4	2012-12-01 21:12:12	-73.966046	40.789775	-73.988565	40.744427	1.0	15.01594

Fig 4.1.1. A sample of the dataset

The following list includes one target variable and six predictor variables:

Predictors:

1) Pickup Datetime: A timestamp value describing the commencement of the cab journey.
2) Pickup Longitude: A floating point number that represents the longitude coordinate at which the cab ride began.
3) Pickup Latitude: A floating point number that represents the latitude coordinate at which the cab ride began.
4) Dropoff Longitude: A floating point value indicating the longitude coordinate of the final location of the cab journey.
5) Dropoff Latitude: A floating point value indicating the latitude coordinate of the final location of the cab journey.
6) Number of Passengers: An integer value that represents the total number of passengers in the cab.

Goal: Price

4.1.2 Handling Missing Data

Missing Data can be handled in one of two ways after the dataset has been imported into the model:

A) By removing those specific rows

B) By substituting mean or median for those values

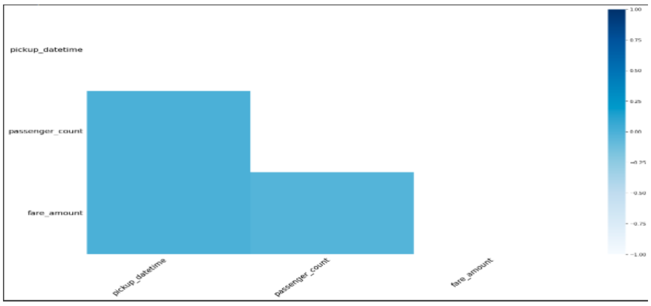
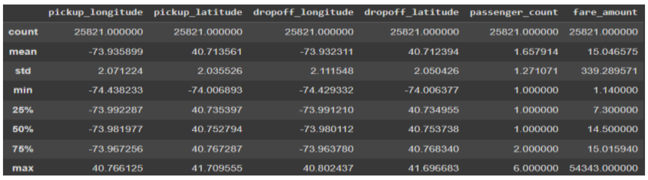


Fig 4.1.2(a): Visualization of missing values

We replaced missing values in our model with the mean of that particular column.

4.1.3 Outlier analysis

As there are many of noisy data, cleaning the data is crucial for improved model performance. In this instance, Turkey's technique is employed as a traditional method of eliminating outliers. We use box plots to show the outliers. Several insightful conclusions may be drawn from these plots, including the fact that each of the data sets has a significant number of outliers and extreme values.



	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	fare_amount
count	25821.000000	25821.000000	25821.000000	25821.000000	25821.000000	25821.000000
mean	-73.935899	40.713561	-73.932311	40.712394	1.657914	15.046575
std	2.071224	2.035526	2.111548	2.050426	1.271071	339.289571
min	-74.438233	-74.006893	-74.429332	-74.006377	1.000000	1.140000
25%	-73.992287	40.735397	-73.991210	40.734955	1.000000	7.300000
50%	-73.981977	40.752794	-73.980112	40.753738	1.000000	14.500000
75%	-73.967256	40.767287	-73.963780	40.768340	2.000000	15.015940
max	40.766125	41.709555	40.802437	41.696683	6.000000	54343.000000

Fig 4.1.3(a): refined data

The data is now refined as seen in Fig.4.1.3(a) after the outliers have been removed.

4.2 Correlation

The term "variable or attribute selection" is commonly used in machine learning. Choosing the traits that have the most influence on the aim variable is essentially what this procedure entails. The fare amount in this situation is the target or prediction variable, whilst the other factors are independent.

.

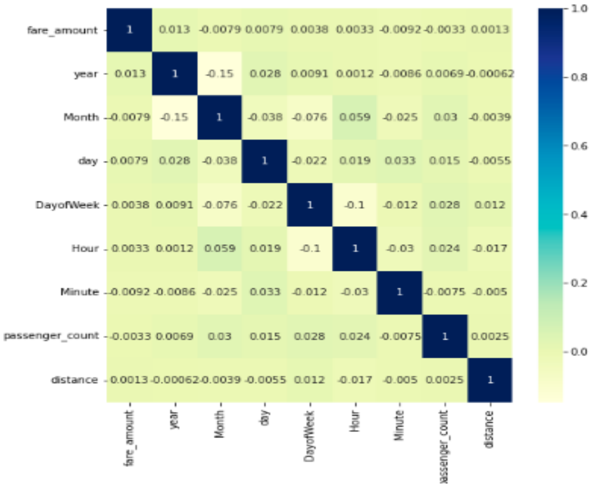


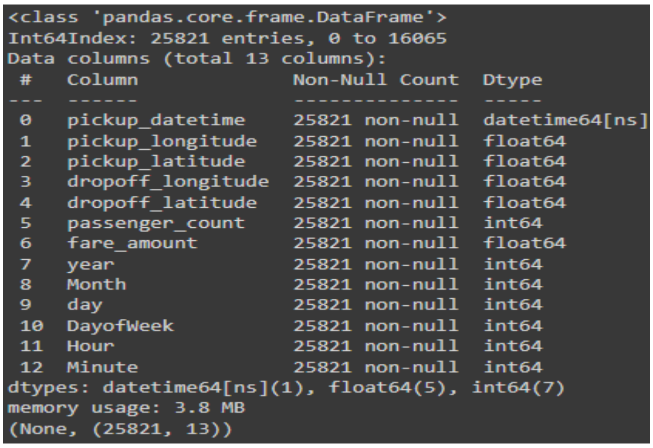
Fig 4.2(a): Correlation Heat Map

4.3 PCA

Principal component analysis (PCA) is a dimensionality reduction method that facilitates the identification of correlations and patterns in data collection.

4.4. Feature Engineering

To generate more useful information, it is crucial to draw certain conclusions from the available data. The distance covered in each ride may be simply estimated to establish a correlation between the fee amount and the distance since the longitude and latitude markers are present, determining the distance. It is crucial for navigational purposes.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25821 entries, 0 to 16065
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   pickup_datetime    25821 non-null  datetime64[ns]
 1   pickup_longitude   25821 non-null  float64
 2   pickup_latitude    25821 non-null  float64
 3   dropoff_longitude  25821 non-null  float64
 4   dropoff_latitude   25821 non-null  float64
 5   passenger_count    25821 non-null  int64
 6   fare_amount        25821 non-null  float64
 7   year               25821 non-null  int64
 8   Month              25821 non-null  int64
 9   day                25821 non-null  int64
 10  DayofWeek          25821 non-null  int64
 11  Hour               25821 non-null  int64
 12  Minute             25821 non-null  int64
dtypes: datetime64[ns](1), float64(5), int64(7)
memory usage: 3.8 MB
(None, (25821, 13))
```

Fig 4.4(a): Feature engineering

4.5 Model Selection

Modeling usually referred to as model selection is a crucial step that follows data pretreatment. The process of selecting a model from a wide pool of models to solve a prediction problem is known as model selection. Predicting the fare amount is our challenge. Regression is the issue here. The dependent variable in classification issues is categorical. Regression models will thus be used to analyze training data and make predictions about test data.

In this study, we employ a tree-based technique called Random Forest to address classification and regression issues. In regression models, the usage of a linear regression model is also common. The entire set of data was divided into two categories: train data (70%), and test data (30%).

4.5.1 Random Forest

Based on supervised learning, which supports both classification and regression, this model. Since it is composed of several decision trees that work together to forecast the goal value. A group of trees provides a value that is more accurate than one single tree.

4.5.2 Linear Regression

It is a supervised learning-based method. In light of independent factors, it provides the desired prediction value. It is mostly used to determine how dependent and independent variables are related.

Using the use of an accessible independent variable, To predict the value of a dependent variable, linear regression is performed. This strategy creates a linear relationship between the two groups of data a result. The following equation serves as the algorithm's foundation:

$$y=a+bx$$

Finding the appropriate values for a and b is the objective, with x acting as an independent variable and y acting as a dependent variable.

4.5.3 Decision tree

A decision tree is a graph that resembles a tree with nodes for the attributes that are selected and for which questions are asked, edges for the responses to those questions, and leaves for the actual output or class label. Nonlinearity describes decision trees. Classification and Regression Trees are two terms used to describe decision tree techniques.

4.6 Data Visualization

Using various graphs, charts, plots, etc., data visualization makes it easier to understand the data. So that the data may be adequately examined, it is here depicted using a scatter plot, bar graph, and histogram.

4.6.1 Histogram

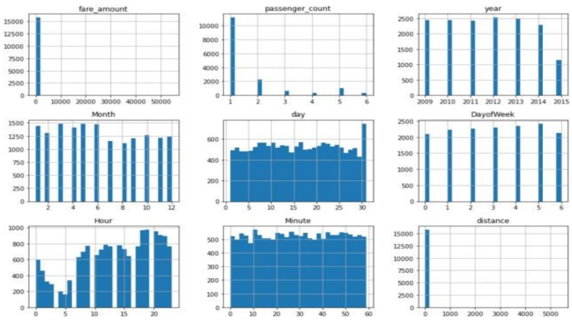


Fig 4.6.1 Histogram of all attributes of dataset

4.7 Proposed model

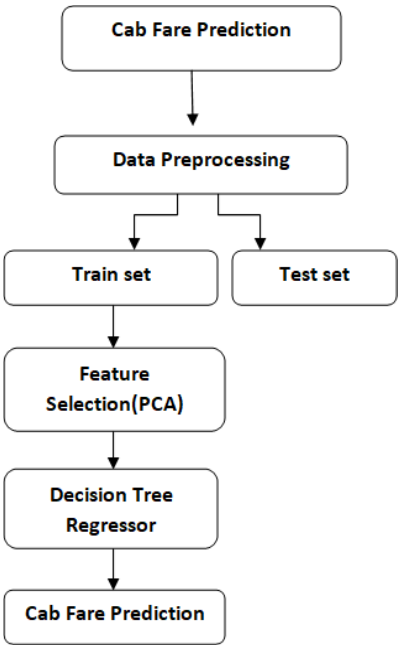


Fig 4.7(a): Design model of Cab fare prediction

5. RESULT ANALYSIS

The degree to which a regression model's predictions agree with observed values is a measure of its quality. Regressions may be compared to other regressions with various parameter values thanks to error metrics, which are used to evaluate a model's quality. We'll discuss particular regression error measures like -

- R square: The model is more accurate the higher the value. The values are converted to percentages ranging from 0 to 1.
- RMSE (Root Mean Square Error): The square root of the MSE error rate. Although the RMSE of 0.6 is modest, it is no longer that little. But the better, the lower the RMSE.

The model that is deemed to be the best and most accurate is also the one that has the highest R square and lowest RMSE values.

Algorithm	Accuracy
Linear regression	0%
Random forest	80%
Decision tree	100%

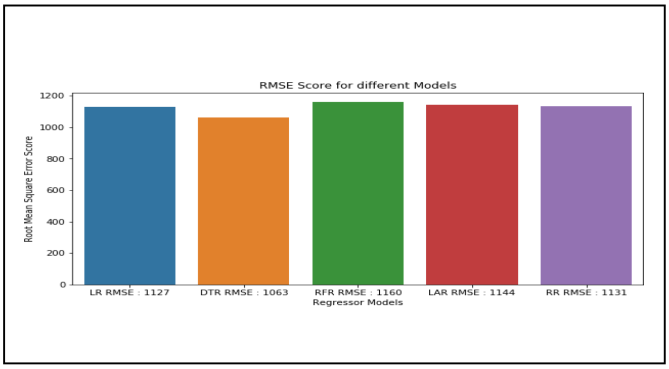
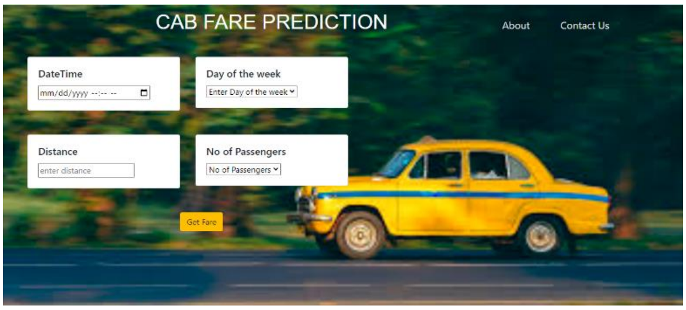


Fig 5(a): Comparison of algorithms with correlation

By considering the above Fig. 5(a) it is observed clearly that the RMSE of the Decision Tree Algorithm is low when compared to all the algorithms.

We are utilizing a decision tree method to forecast the cab fare using machine learning. This method can predict the result with high accuracy and less rmse value.

6. OUTPUT SCREENS



DateTime: mm/dd/yyyy --:-- --

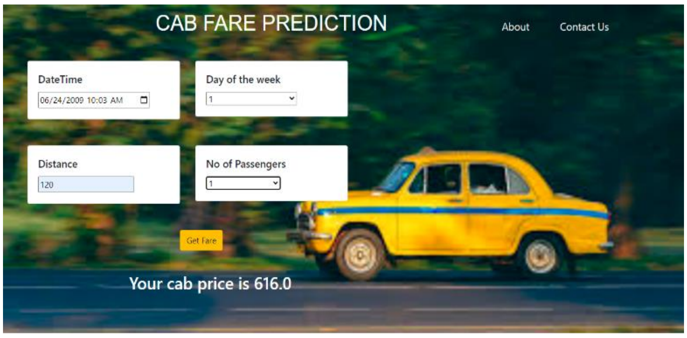
Day of the week: Enter Day of the week

Distance: enter distance

No of Passengers: No of Passengers

Fig 6(a): Cab fare prediction page

Using the above cab fare predictor, we predict the fareof the cab ride.



DateTime: 06/24/2009 10:03 AM

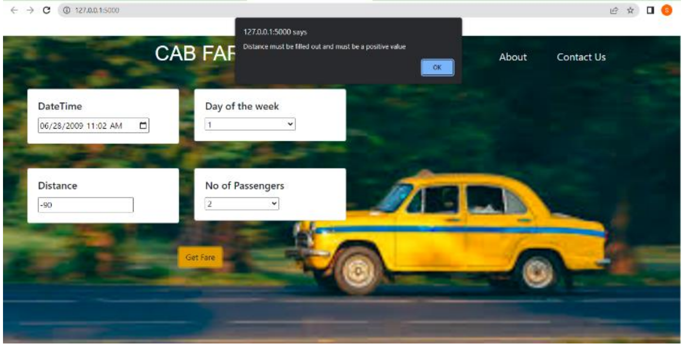
Day of the week: 1

Distance: 120

No of Passengers: 1

Your cab price is 616.0

Fig 6(b): Output of cab fare prediction



127.0.0.1:5000 says

Distance must be filled out and must be a positive value

DateTime: 06/28/2009 11:02 AM

Day of the week: 1

Distance: -90

No of Passengers: 2

Fig 6(c): Output screen when distance is negative

7. CONCLUSION

The degree to which forecasts and observed values agree determines whether a regression model is successful. The dependent variable is continuous in issues involving regression. The dependent variable is categorical in classification issues. Problems with classification and regression may both be solved with Random Forest. In contrast to linear regression, decision trees do not use an equation to describe the connection between the independent and dependent variables. The decision tree is the most effective model out of the remaining three since it has the best R-Squared and RMSE scores, which measure how well the model fits the data and how much variability it explains.

8. REFERENCES

[1] John D. Kelleher, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (The MIT Press), 1st Edn.

[2] Data science work with public datasets, [Online] Available: https://www.kaggle.com,

[3] Chong Ho Yu, Exploratory data analysis in the context of data mining and re-sampling, International Journal of psychological research, 2010, vol.3, No.1

[4] Ruben Oliva Ramos, Jen Stirrup, Advanced Analytics with R and Tableau, Packt Publishing, 2017, ISBN: 9781786460110

[5] Machine Learning in Python, [Online] Available https://scikit-learn.org/stable/

[6] T. Divya and A. Sonali, “A survey on Data Mining approaches for Healthcare”, International Journalof Biosciences and Bio-Technology, vol. 5, no. 5, (2013), pp. 241-266.

[7] Sebastian Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, [Online] Available https://arxiv.org/abs/1811.12808,2016 .