Effective Fare Prediction using Machine Learning Algorithms

Dr. K Anuradha, Associate Professor, Department of MCA, Karpagam College of Engineering, Coimbatore. anu_kce@yahoo.com

Ashika Jayaraj, PG Scholar, Department of MCA, Karpagam College of Engineering, Coimbatore. ashikajayaraj11@gmail.com

Abstract : Predictive analytics makes use of historical data to forecast future events. Past data is typically used to create a mathematical model that captures significant trends. The predictive model is then used to current data in order to forecast the future or advise actions to take in order to achieve the best results. Due to advancements in supporting technology, particularly in the fields of big data and machine learning, predictive analytics has gotten a lot of attention in recent years. Predictive analytics is also used by businesses to make more accurate estimates, such as estimating the cost of a cab journey in the city. These estimates allow for more effective resource planning, such as the scheduling of multiple cab hires. When starting a cab rental start-up company, a lot of factor determines the fare amount. The goal of this study is to decipher all trends and use analytics to fare prediction.

The suggested system aims to create a system that estimates the cost of a cab journey in the city. The goal is to create regression models that can forecast the continuous fare amount for each cab journey and aid in prediction based on a variety of time-based, positional, and general parameters.

This analysis is going to benefit those who are looking to predict the fare for their future transactional cases. As, we can see there are number of cab companies like Uber, Ola, RedTaxi etc. And these cab companies deliver services to lakhs of customers daily. So it becomes really important to manage their data properly to come up with new business ideas to get best results. In this case, earn most revenues. Hence, it becomes really important estimate and gives fare details by using various machine learning algorithms like Decision Tree, Random Forest, Naïve Bayes and KNN algorithms.

Keywords - Decision Tree, Random Forest, Naïve Bayes, KNN Algorithms

I. INTRODUCTION

Taxi services is a transportation medium that will use travel information for the passengers, thus resulting in better way of providing best service to the needful . However, the problem with the system is that it is expensive, and not suitable for long distances. This actually takes location of taxis such as the geographical location and the source and destination, while selecting the optimum route for the travel and satisfying the customer ride bookings. Further this shows the best and cheapest travel amount in considerably short time.

Computational statistics, which focuses on making predictions with computers, is closely related to machine learning (ML). Data mining (DM) is a branch of machine learning that focuses on unsupervised learning for exploratory data analysis. Machine learning is also known as predictive analytics when it is used to solve business challenges. Several major categories exist for machine learning tasks.

The procedure develops a mathematical model from a set of data that includes both the inputs and the expected outputs in supervised learning. supervised learning techniques include classification and regression methods. The outputs of regression algorithms are continuous, meaning they can be any value within a range.

The approach develops a mathematical model from a set of data that comprises only inputs and no desired output labels in unsupervised learning.

Unsupervised learning techniques are used to discover structure in data, such as data point grouping or clustering. Unsupervised learning, like feature learning, can find patterns in data and organise inputs into categories. The practise of lowering the number of "features," or inputs, in a piece of data is known as dimensionality reduction.

Machine learning and data mining both use similar approaches and have a lot of overlap, however while ML focuses on prediction based on known qualities learnt from

 DOI : 10.35291/2454-9150.2022.0215

the training data, data mining focuses on finding previously undiscovered properties in the data. This is the knowledge discovery in databases (KDD) analytical stage [1]. DM employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy; on the other hand, ML employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.

1. Data

The goal is to create regression models that can estimate the continuous fare amount for each cab journey based on a variety of positional and time-based characteristics. This problem statement belongs to the area of forecasting, which is concerned with predicting future continuous values (the continuous value is the fare amount of the cab ride).

• pickup_datetime - timestamp value indicating when the cab ride started.

• pickup_longitude - float for longitude coordinate of where the cab ride started.

• pickup_latitude - float for latitude coordinate of where the cab ride started.

• dropoff_longitude - float for longitude coordinate of where the cab ride ended.

• dropoff_latitude - float for latitude coordinate of where the cab ride ended.

• passenger_count - an integer indicating the number of passengers in the cab ride.

All these factors together help in determining the fare amount.

The paper is divided as follows: Methodology is explained in Section II and the model of the system is explained in Section III. The developed system is evaluated in Section IV. And finally the paper is concluded in Section V.

II. METHODOLOGY

This section explains the methodology used in the paper. Many researchers used Machine Learning approaches for various applications [6-11]. Flight Pride prediction is carried out in [10] [11] using Machine Learning Algorithms.

A. Exploratory Data Analysis

Pre-processing data is the initial step in any endeavour. We gain a sense of the data at this point. This is done by examining plots of independent and target variables. If the data is disorganised, we try to clean it up by sorting it and removing superfluous rows and columns. This stage is referred as Exploratory data analysis. This stage usually entails data cleansing, merging, sorting, and searching, looking for missing values in the data and imputing missing values, if they are found approaches such as mean, median, mode, KNN imputation, and so on.

B. Removing values which are not within desired range(outlier) depending upon basic understanding of dataset.

Based on a fundamental understanding of all the variables, we will remove values from each variable that are not within the intended range and deem them outliers. You might wonder why I didn't just make those values NA instead of eliminating them. I did, but the dataset ended up with a lot of missing values (NA's). When the percentage of missing values rises too high, there is no benefit in using the imputed data.

C. Missing Value Analysis

We seek for missing values in the dataset in this step, such as empty row column cells that were left after special characters and punctuation marks were removed. Some missing values are represented by the letter NA. Missing values left over after outlier analysis; missing values can take any shape. Unfortunately, we discovered some missing values in this dataset. As a result, we'll conduct some missing value analysis. We chose random row no. 1000 and made it NA before imputed so that we could compare the original and imputed values and find the best method that would impute the value closest to the actual value. The missing value percentage of variables is shown in Fig 1.

Variables	Missing_percentage	
0	passenger_count	29.563702
1	pickup_latitude	1.966764
2	pickup_longitude	1.960540
3	dropoff_longitude	1.954316
4	dropoff_latitude	1.941868
5	fare_amount	0.186718
6	pickup_datetime	0.006224

Fig 1 : Missing value percentage of variables

D. Outlier Analysis

Boxplots are used to check for outliers in the dataset. Outliers can be found in the data.

These outliers have been eliminated. This is how we did it,

I. We replaced them with Nan values or, to put it another way, we generated missing values.

II. We then used the KNN approach to impute those missing values.

• For the time being, we will only perform outlier analysis on Fare amount, and we will perform outlier analysis after feature engineering latitudes and longitudes.

• Boxplots for the target variable (univariate).

E. Dataset

In machine learning, training data is the data you use to train a machine learning algorithm or model. Training data requires some human involvement to analyze or process the data for machine learning use.

 DOI : 10.35291/2454-9150.2022.0215

Model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set.

Training data is typically larger than testing data. This is because we want to feed the model with as much data as possible to find and learn meaningful patterns. Once data from our datasets are fed to a machine learning algorithm, it learns patterns from the data and makes decisions.

Here in this work, there was a total of 80 data out of which 70% of data are used for training and the remaining 30% are used for testing.

The dataset for this work was downloaded with reference to following :

https://data.world/datasets/uber [12]

III. MODELLING

Fare amount is understood to be dependent on numerous behaviours in the early stages of analysis during pre-processing.

As a result, it's critical to construct a model that incorporates all essential inputs and fits the model in such a way that it produces the most accurate result among all other models. The dependent variable might be one of four types: nominal, ordinal, interval, or ratio.

A. Decision Tree

A decision tree is a tree-like graph with nodes representing where an attribute is selected and queried, edges representing query replies, and leaves reflecting the actual output or class label. Nonlinear decision trees exist. Classification and Regression Trees are two terms for decision tree algorithms (CART).

B. Random Forest

Random forest is a tree-based approach that builds numerous trees (decision trees) and then combines their output to improve the model's generalisation capabilities. An ensemble approach is a technique for mixing trees. To create a strong learner, the ensemble combines weak learners (individual trees). Random Forest can be used to tackle problems involving regression and classification. The dependant variable in regression issues is a continuous variable. The dependent variable in classification issues is categorical.

C. K-NN

The K-Nearest Neighbors (K-NN) algorithm is a supervised machine learning technique that can address both classification and regression issues. The K-NN method believes that items that are similar are nearby. K-NN makes predictions straight from the training dataset. For a new instance (x), predictions are formed by scanning the whole training set for the K most similar instances (neighbours) and summing the output variable for those K instances. This could be the mean output variable in regression, or the modal (or most common) class value in classification. A distance metric is used to determine which of the K instances in the training dataset are most similar to a new input. The most typical distance for real-valued input variables is measured in Euclidean distance.

IV. MODEL EVALUATION

The quality of a regression model is determined by how well its predictions match actual values, and error metrics are used to assess model quality, allowing us to compare regressions with different parameters.

A. Root Mean Squared Error (RMSE)

is a type of error that occurs when the root mean square (RMSE)

When dealing with time series forecasting and continuous variables, the Root Mean Squared Error (RMSE) and R-Squared are utilised.[8]

The R-Squared is a relative measure of fit, whereas the RMSE represents the model's absolute fit to the data.

RMSE must be compared with the dependent variable as RMSE is in the same units as the dependent variable.

1) Smaller The Result, Better The Performance Of The Model: To understand how well the independent variables “explain” the variance in the model, the R-Squared formula is used.

2) For The R-Squared, The Closer The Value To 1, The Better The Performance Of The Model: According to the underlying model.

V. COMPARISON AND RESULTS

Among various algorithms used and compared, it is found that Random Forest has predicted more accurate result.
The accuracy can be calculated using the following formula:
Accuracy = (Predicted fare / Actual fare)* 100
Accuracy in result when compared with algorithms.

Method	Accuracy %
Decision Tree	72.06
Random Forest	76.98
Linear Regression	62.02
KNN	71.09

Table 4.1 : Accuracy Table

From the above table it is observed that Random Forest Algorithm gives the best result with 76.98%
The results are shown in Fig 4.1

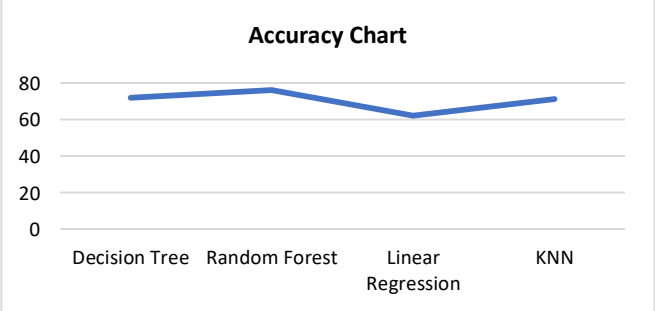

Fig 4.1 Comparison of Accuracy

 DOI : 10.35291/2454-9150.2022.0215

Here in the above graph we can see X axis represents the algorithms being used and Y axis represent accuracy of the algorithms. From this it is observed that Random Forest gives the good accuracy

VI. CONCLUSION

The paper titled "Effective fare prediction using Machine Learning Algorithms" has been developed to meet the requirement analysis phase which were fulfilled in development of this project. This project is validated for accuracy and the results were found to be satisfactory. Hence, the new system is more reliable, accurate, efficient and effective.

Validating each and every field in the forms ensure the accuracy of the data. The user can enter only the valid data to the system. It decreases the delay to make the reports according to the user's need.

We propose a representation of a trip sharing facility as a novel service, in contrast to the current ride services supplied by worldwide firms. Despite the fact that some of these characteristics have been tried before, the outcomes have not been sufficient. This model is appropriate for daily riders who want to save money, time and effort by choosing their route of travel. Our solution can reduce the number of private vehicles on the road, resulting in reduced traffic, shorter passenger journeys, a driver's bespoke schedule management, and lower ride costs. Essentially, it outperforms better use for passengers.

VII. FUTURE SCOPE

As is known, with an increase in the number of features; underlying equations become a higher-order polynomial equation, and it leads to overfitting of the data. Generally, it is seen that an overfitted model performs worse on the testing data set, and it is also observed that the overfitted model performs worse on additional new test data set as well. A kind of normalized regression type - Ridge Regression may be further considered.

VIII. REFERENCES

[1] Salman Salloum, Joshua Zhexue Huang and YulinHe. (2019). Random Sample Partition: A Distributed Data Model for Big Data. IEEE Transactions on Industrial Informatics.

[2] San Yeung (2017) PhD Forum: Toward a Human-in-the-Loop Smart Ridesharing System with Self Driving Technologies. IEEE International Conference on Smart Computing.

[3] Nicolae VlamidirBozdog, Marc X. Makkes, Aart van Halteren, Henri Bal (2018). RideMatcher: Peer to-peer Matching of Passengers for Efficient Ridesharing. IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.

[4] Jing Fan, Jinting Xu, Chenyu Hou, Bin Cao, Tianyang Dong, Shiwei Cheng (2018) URoad: An Efficient Algorithm for Large-scale Dynamic Ridesharing Service. IEEE International Conference on Web Services.

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

[6] S Shriram, K Anuradha, KP Uma (2021), Future Stock Price Prediction using Recurrent Neural Networks LSTM and Machine Learning, International Journal of Engineering Research & Technology, Vol.9, Issue. 5, 309 – 308.

[7] Anuradha K, Abinesh K, Barath S (2022). Diabetes Prediction using Machine Learning, Proc. of the International Conference on Emerging Trends in Science and Technology, ISBN: 978-93-92032-16-5, PP 78.

[8] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Computational Intelligence and Neuroscience, vol. 2021, Article ID 8387680, 11 pages, 2021. https://doi.org/10.1155/2021/8387680.

[9] Anuar Nayan, N., et al.: Cardiovascular Disease Prediction from Electrocardiogram by using Machine Learning Method: A Snapshot from the Subjects of the Malaysian Cohort (2020).

[10] Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar, Predicting The Price of A Flight Ticket With The Use Of Machine Learning Algorithms, International Journal of Scientific & Technology Research, 8(12), (2019), 3297 – 3300.

[11] Janhvi Mukane, Siddharth Pawar, Siddhi Pawar, Gaurav Muley. Aircraft Ticket Price prediction using Machine Learning, International Journal for Research in Applied Science and Engineering Technology, 10(11), 2022.

[12] https://data.world/datasets/uber

 DOI : 10.35291/2454-9150.2022.0215