JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TAXI FARE PREDICTION USING MACHINE LEARNING

[1]Neha Andhare, [2]Aishwarya Jadhav, [3]Vaishnavi Jadhav and [4]Savita Sangam
Department of Information Technology,
SSJCOE, Dombivli, India

Abstract: To predict longer-term events, predictive analysis uses data that is archived. A mathematical model from past data is used to capture trends that are important. The model then uses the current data to predict the longer term or to derive actions that recently require tones of appreciation of optical results for predictive analytics due to the development of supporting technology in the areas of massive data in machine learning. Many industries use predictive analytics to make accurate predictions, such as providing a discount for driving around town. This resource planning enables forecasting, for example, taxi fares can be predicted more accurately. There are many factors to consider when starting a taxi business. This project tries to know patterns and use different methods to predict fares. This project is developed to predict taxi fares in a certain city. The project involves different steps like training, testing using different variables like pick up, drop off location to predict taxi fare.

Index Terms - Machine Learning, Fare Prediction, Predictive analysis, Supervised Learning.

1. INTRODUCTION

Taxi fare prediction is a regression problem that uses machine learning techniques to estimate the fare for a given taxi trip based on various input functions. The process begins with the collection and pre-processing of a comprehensive data set of historical taxi ride records, which typically includes information such as pick-up and drop-off locations (latitude and longitude), travel distance, travel time, fare amount, and any other relevant contextual features such as .weather conditions, traffic patterns and time of day. Data preprocessing steps can include handling missing values, removing outliers, and ensuring consistency of feature representations.

Once the data is ready, the next step is feature engineering, where relevant features are extracted or inferred from the raw data to capture the underlying patterns and relationships that influence fares. These features could include straight-line distances, travel times based on historical traffic data, peak or weekend indicators, and any other domain-specific knowledge that could improve the model's predictive power.

With the designed features and the target variable (fare), various machine learning algorithms such as linear regression, decision trees, random forests, gradient boosting machines and neural networks can be used for the regression task. These algorithms learn complex relationships between input features and fares from training data, allowing them to make accurate predictions of new, unseen taxi ride instances.

Model evaluation is a crucial step where the performance of the trained model is assessed on the test dataset using appropriate metrics such as mean square error (MSE), root mean square error (RMSE) or mean absolute error (MAE). This evaluation helps identify the most accurate model and leads to further tuning or selection of different algorithms if performance is unsatisfactory.

Once a satisfactory model is obtained, it can be deployed as a service or integrated into ride-hailing applications to provide real-time fare estimates for potential taxi trips based on input features. Accurate price forecasting not only improves user experience by providing price transparency, but also enables taxi companies to optimize their pricing strategies, resource allocation and demand forecasting based on predicted patterns.

However, it is important to note that the accuracy of taxi price prediction models can be affected by various factors such as the quality and completeness of the training data, the efficiency of feature engineering, the complexity of the underlying patterns, and external factors such as traffic conditions, road closures, and weather events that they can introduce uncertainty into predictions.

1.1: OBJECTIVES

1. Accurate Fare Estimation: Develop a model that can accurately estimate the fare for a given taxi trip based on input features such as pick-up/drop-off locations, travel distance and time, providing transparency and building customer trust.
2. Demand forecasting: Analyze historical data and relevant features to predict demand patterns for different areas and times, helping taxi companies better resource planning and fleet management.
3. Improved user experience: Provide customers with accurate advance fare estimates, allowing them to plan their journeys more efficiently and avoid fare disputes, improving the overall user experience.

2. LITERATURE REVIEW

This research will be useful for those involved in fare forecasting. In the previous generation, the fare changed to be the most convenient depending on the distance, but with the improvement in technology, the cab fare depends on many things such as time, area, number of passengers, traffic, number of hours, base fare and so on. The view of is based on supervised mastering, one application of which is prediction in device recognition. This research aims to observe predictive evaluation, which is a method of analysis in Machine Learning. Many corporations like Ola, Uber and many others are using artificial intelligence and system learning technology to find the answer to fix the fare prediction problem. [1]

The survey suggests that flight and taxi fares vary depending on various factors such as location, time of day and so on. Also cabins where the fare depends on a wide range of passengers, visitors and so on. Seller has facts about all factors but buyers can access records which are limited and we cannot expect price lists. Uber and Ola use factors such as traffic in a particular neighbor hood and the motive behind these articles is to explore the factors that impact rate variances and how they relate to trade within pricing. . [2]

The patterns and functions of the transportation system, including the traditional mode of sightseeing that includes taxis and subways, as well as revolutionary devices such as ride-hailing (Uber, Lyft, etc.), are critical subjects of study in economics, transportation, and operations studies. field. By calculating and analyzing the effect of these elements on the amount of Uber drivers' fees, we will reach conclusions that can be instructive and useful in practice. [3]

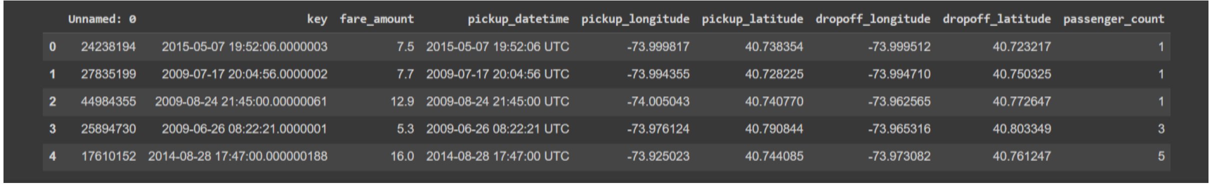
Using large-scale urban facts to expect taxis and Uber passenger demand in cities is valuable for designing higher taxi dispatch structures and improving taxi services. In this paper, we forecast taxi and Uber demand using real global datasets. Our technique involves two key steps. First, we use time-correlated entropy to measure the call for regularity and achieve maximum predictability. Second, we implement and validate 5 well-known representative predictors (Markov, LZW, ARIMA, MLP, and LSTM) to achieve as much predictability as possible. [4]

Using spatio-temporal modes of time series can help us to better recognize the demand for call services and anticipate it more accurately. This paper analyzes the overall prediction performance of 1 temporal model (vector autoregressive (VAR)) and spatiotemporal modes (Spatial-temporal autoregressive (STAR); least absolute shrinkage and selection operator applied to STAR (LASSO-ST AR)) and for characteristic scenarios (primarily based on the number of ohms and spatial delays) and performed for each peak and non-peak period. The implications show that the demand for taxi services does not need to be considered in spatial ways. [5]

This model is able to predict Uber surge multipliers, the overall average, and the historical average in all but 3 of the 49 Pittsburgh locations and outperforms the three nonlinear methods in 28 of the 49 locations. The cross-correlation of Uber and Lyft surge multipliers is also examined. [6]

3. METHODOLOGY

1. Data Collection:
   We Gather a dataset containing historical taxi rides from Kaggle.[8] Each record should include features such as pickup location, drop-off location. Each record should have the corresponding fare amount.
2. Data Understanding:
   To get the best results, to get the most effective model it is really important to our data very well. Here, the given train data is a CSV file that consists 9 variables and 200000 Observation. A snapshot of the data provided.



	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5

Figure 3.1: Train Data

The different variables of the data are:
- fare_amount : fare of the given cab ride.
- pickup_datetime : timestamp value explaining the time of ride start.
- pickup_longitude : a float value explaining longitude location of the ride start.
- pickup_latitude :a float value explaining latitude location of the12 ride start.
- dropoff_longitude : a float value explaining longitude location of the ride end.

- dropoff_latitude : a float value explaining latitude location of the ride end.
- Passenger_cont : an integer indicating number of passengers

3. Data Preparation :
The next step is, Data preprocessing. It is a data mining process that involves transformation of raw data into a format that helps us execute our model well. As, the data often we get are incomplete, inconsistent and also may contain many errors. Thus, Data preprocessing is a generic method to deal with such issues and get a data format that is easily understood by machine and that helps developing our model in best way. In this project also we have followed data pre-processing methods to rectify errors and issues in our data. And this is done by popular data preprocessing techniques, this are following below.

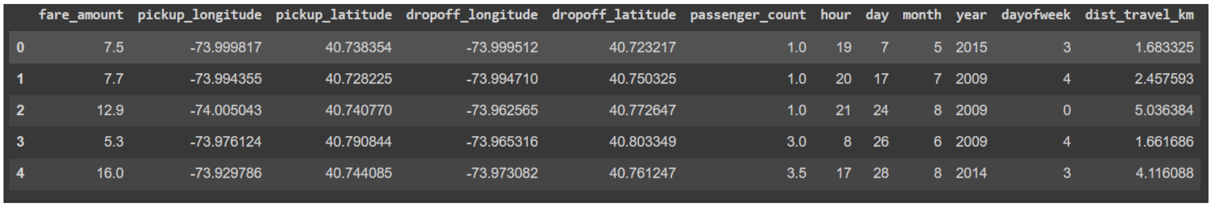
Note: I have removed the variable "pickup_datetime" as it is a timestamp value and it shows only the start time of pick up time , whereas there is no drop off time, so in this data set it seems, it will have no impact in the target variable, and also it lead to redundancy and model accuracy issues, so we preferred to drop it.

4. Feature Selection :
Sometimes it happens that, all the variables in our data may not be accurate enough to predict the target variable, in such cases we need to analyze our data, understand our data and select the dataset variables that can be most useful for our model. In such cases we follow feature selection. Feature selection helps by reducing time for computation of model and also reduces the complexity of the model.
After understanding the data, preprocessing and selecting specific features, there is a process to engineer new variables if required to improve the accuracy of the model.
In this project the data contains only the pick up and drop points in longitude and latitude. The fare_amount will mainly depend on the distance covered between these two points. Thus, we have to create a new variable prior further processing the data. And in this project the variable I have created is Distance variable (dist_travel_km), which is a numeric value and explains the distance covered between the pick up and drop of points. After researching I found a formula called The haversine formula, that determines the distance between two points on a sphere based on their given longitudes and latitudes. These formula calculates the shortest distance between two points in a sphere.
The function of haversine function is described, which helped me to engineer our new variable, Distance.

Used in Python :

```
import haversine as hs
travel_dist = []
for pos in range(len(df['pickup_longitude'])):

 long1,lati1,long2,lati2 = [df['pickup_longitude'][pos],df['pickup_latitude'][pos],df['dropoff_longitude'][pos],df['dropoff_latitude'][pos]]

 loc1 = (lati1,long1)

 loc2 = (lati2,long2)

 c = hs.haversine(loc1,loc2)

 travel_dist.append(c)

 print(travel_dist)

 df['dist_travel_km'] = travel_dist

 df.head()
```

After excecuting the haversine function in our project, I got new variable distance and some instances of data are mentioned below.



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	hour	day	month	year	dayofweek	dist_travel_km
0	7.5	-73.999817	40.738354	-73.999512	40.723217	1.0	19	7	5	2015	3	1.683325
1	7.7	-73.994355	40.728225	-73.994710	40.750325	1.0	20	17	7	2009	4	2.457593
2	12.9	-74.005043	40.740770	-73.962565	40.772647	1.0	21	24	8	2009	0	5.036384
3	5.3	-73.976124	40.790844	-73.965316	40.803349	3.0	8	26	6	2009	4	1.661686
4	16.0	-73.929786	40.744085	-73.973082	40.761247	3.5	17	28	8	2014	3	4.116088

Figure 3.2 : Engineering with new variable distance

5. Correlation Analysis:
   In some cases it is asked that models require independent variables free from collinearity issues. This can be checked by correlation analysis for the categorical variables and continuous variables. Correlation analysis is a process that is defined to identify the level of relation between two variables.
   In this project, our Predictor variable is continuous, so we will plot a correlation table that will predict the correlation strength between independent variables and the 'fare_amount' variable

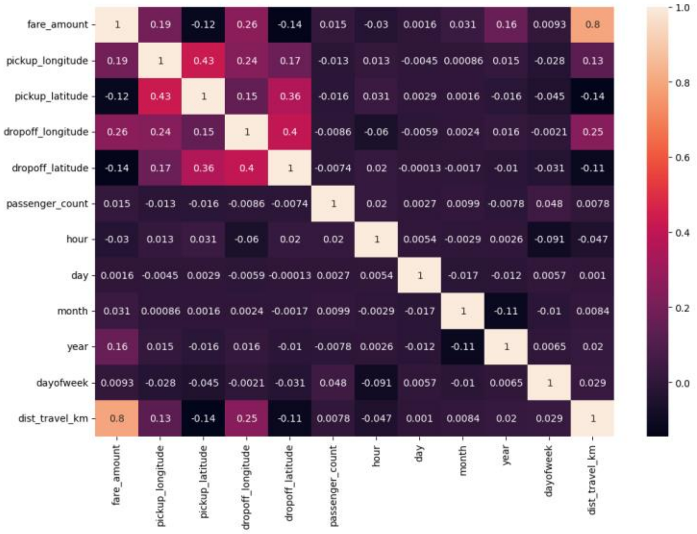


Figure 3.3 : Correlation Matrix

- From the above (fig. 3.3) plot it is found that most of the variables are highly correlated with each other, like fare amount is highly correlated with distance variable.

- Because all the variables are numeric the important features are extracted using the correlation matrix. All the variables are important for predicting the fare_amount since none of the variables have a high correlation factor, so all the variables for model building are kept.

6. Model Deployment:
   After all the above processes the next step is developing the model based on our prepared data.
   In this project we got our target variable as "fare_amount". The model has to predict a numeric value. Thus, it is identified that this is a Regression problem statement. And to develop a regression model, the various models that can be used are Random Forest and Linear Regression.

3.1 Proposed Architecture

The investigation found that although all the models' forecast error rates were below the industry standard of 5%, the regression tree model's mean error rate was higher than that of the multiple regression and lasso regression models. This indicates that while the regression tree model may have a lower error rate for some seeds, it has a higher error rate for the majority of seeds compared to the multiple regression and lasso regression models.This suggests that while the regression tree model may have some advantages over the other models in certain situations, it may not be the best choice for the given dataset.

The multiple regression and lasso regression models may have a more consistent performance and lower mean error rate, making them a better overall choice for the problem at hand. It's important to note that choosing the best model for a particular problem requires careful evaluation of the model's performance and consideration of the specific requirements and constraints of the problem.

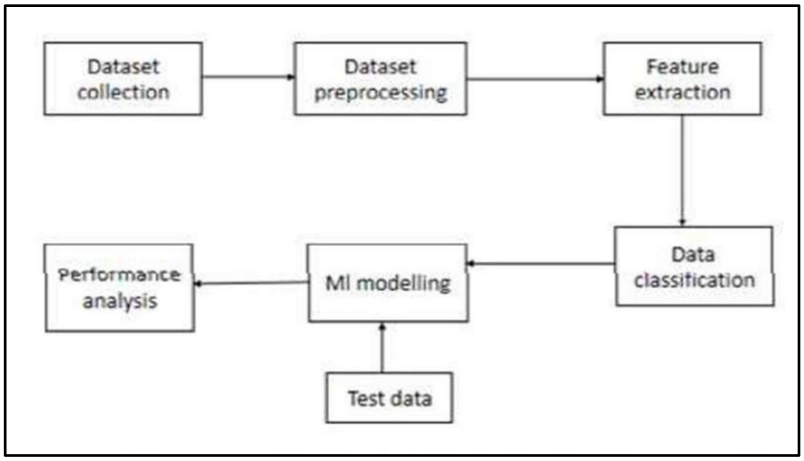


Figure 3.1.1: System Architecture Model

4. RESULTS

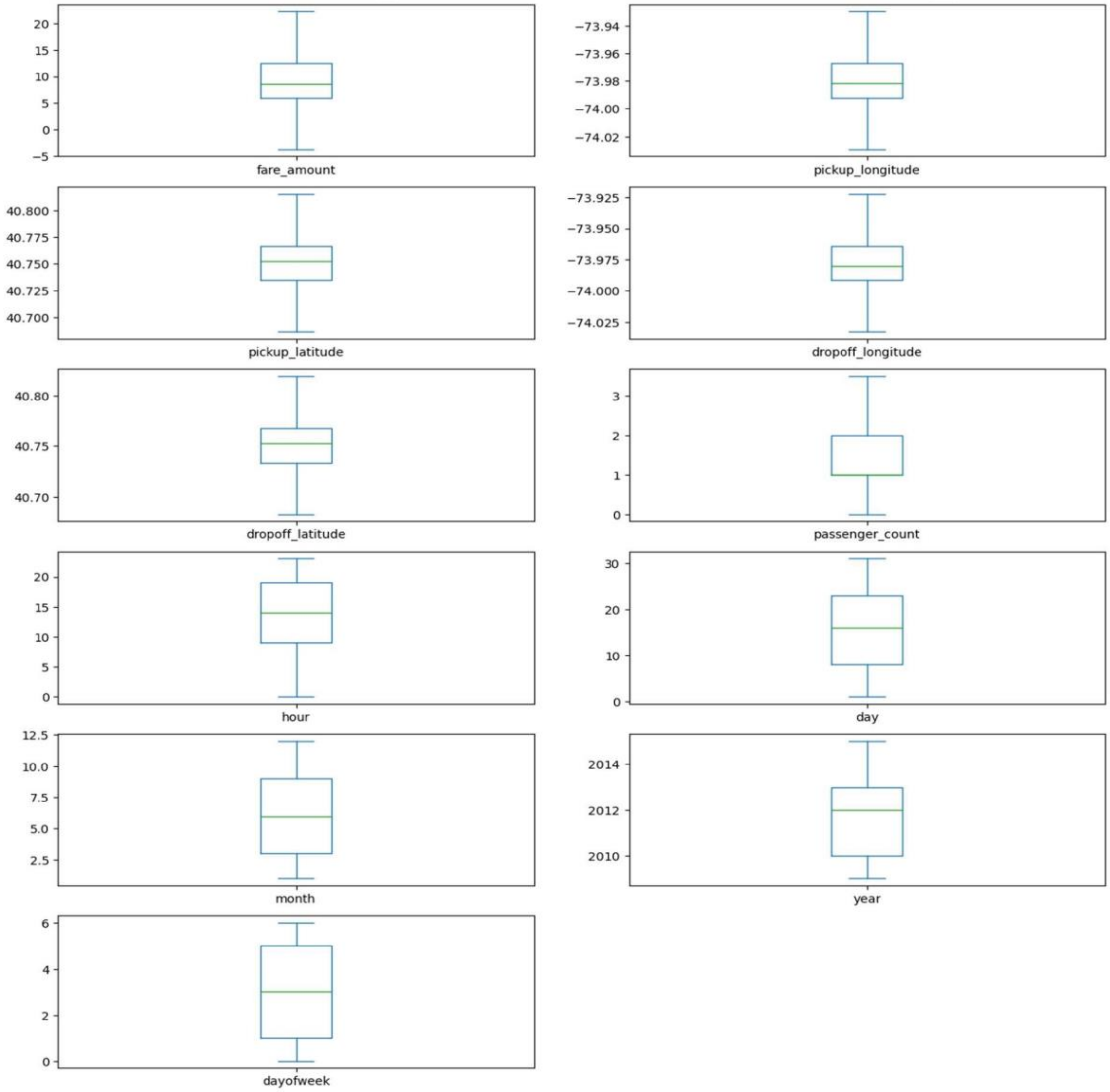


Figure 4.1: boxplot shows that dataset is free from outliers

Fig 4.1 indicates a box plot, or box-and-whisker plot, is a graphical representation of the distribution of a dataset that can help in identifying potential outliers. However, whether a dataset is "free from outliers" is a subjective determination that depends on the context and the specific criteria used to define an outlier.

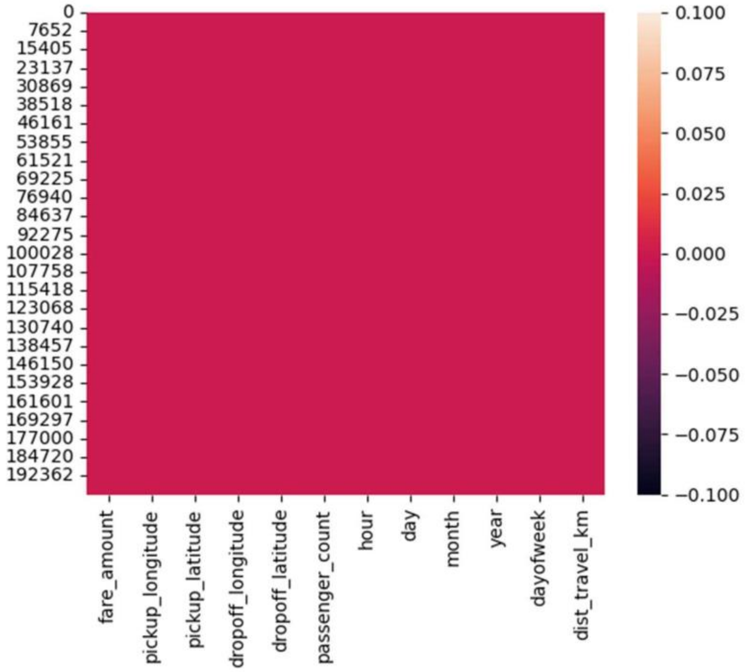


Figure 4.2: Heatmap

Fig 4.2 indicates a heatmap is a graphical representation of data where values in a matrix are represented as colors. It's a way to visualize data in a 2D format, where each cell in the matrix is assigned a color based on its value, allowing patterns and trends to be easily identified.

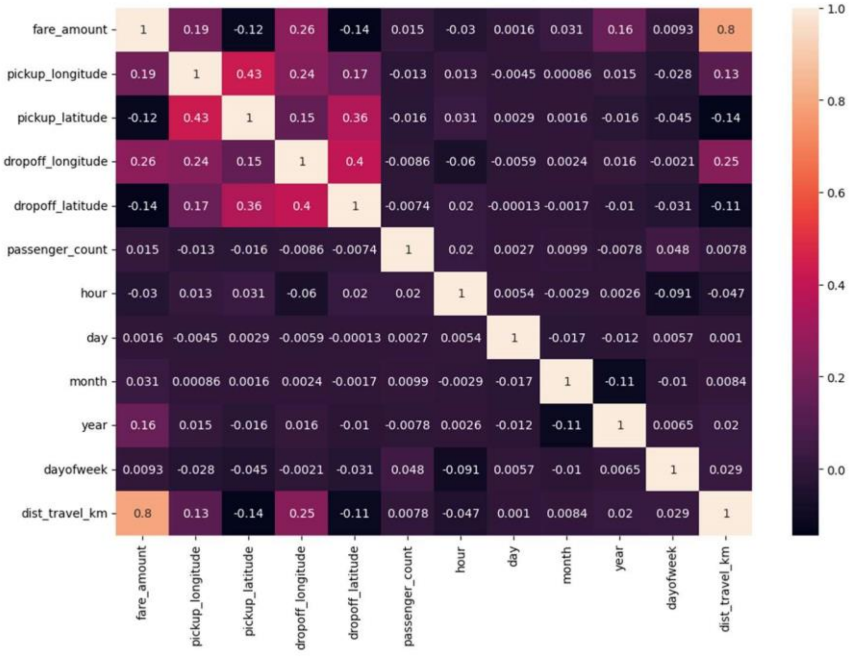


Figure 4.3: Heatmap subplot

Fig 4.3 indicates a heatmap subplot refers to a subplot within a larger figure or plot that specifically displays a heatmap

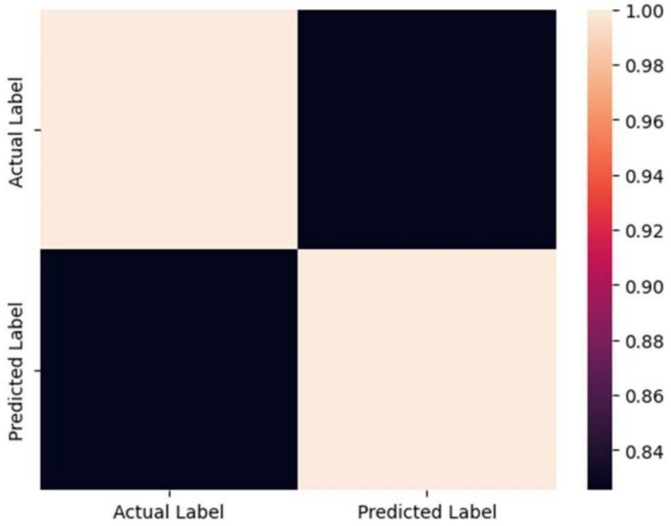


Figure 4.4: Heatmap comparison

Fig 4.4 indicates a heatmap comparison refers to the process of visually comparing two or more heatmaps to identify similarities, differences, patterns, or trends in the underlying data.

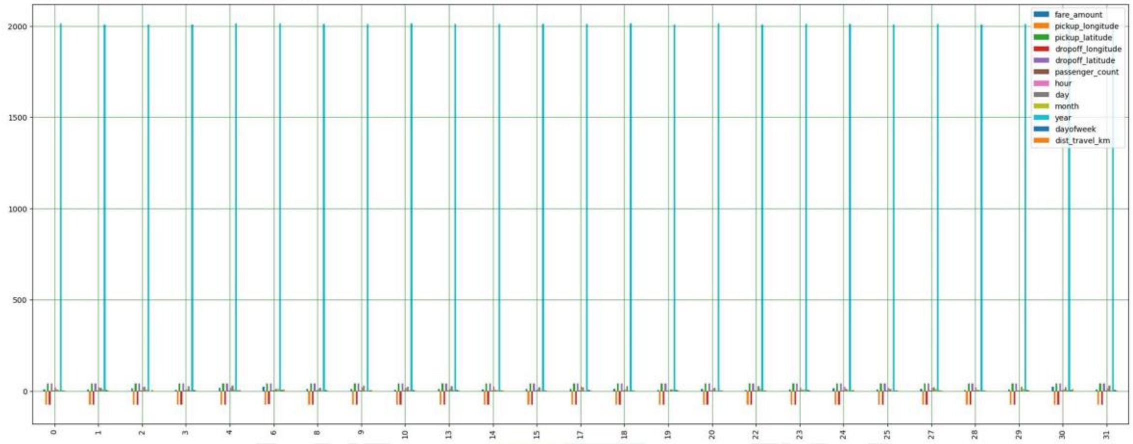


Figure 4.5: Cab bar

Fig 4.5 indicates a "cab bar" refers to a concept or technique in machine learning, it may be a recent development or a specialized term used within a specific community or domain.

CONCLUSION

The Project "TAXI FARE PREDICTION BY USING MACHINE LEARNING TECHNIQUES" can provide several benefits. The results of the analysis can provide valuable insights into the trends and patterns in the data, as well as provide accurate predictions of future cab fare prices. This information can be useful for cab companies to make informed business decisions and optimize pricing strategies.

After training and testing the results shown are fairly accurate. Random forest is useful in regression as well as classification whereas linear regression helps to find the linear relation among the variables. Hence we reached to the conclusion that Random forest is the best because it gives more accurate value as compared to linear regression model. That is why Random forest algorithm is the best fit for the model selection as it has the lowest RMSE value and the highest R square value.

FUTURE WORK

- 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 dataset, 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.

REFERENCES

[1] Banerjee, Pallab & Kumar, Biresh & Singh, Amamath & Ranjan, Priyeta & Soni, Kunal. (2020). Predictive Analysis of Taxi Fare using Machine Leaming. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 373-378. 10.32628/CSEIT2062108.

[2] Khandelwal, K., Sawarkar, A. ., & Hira, S. (2021). A Novel Approach for Fare Prediction Using Machine Leaming Techniques.

International Journal of Next- Generation Computing, 12(5). https://doi.org/10.47164/ijngc.v12i5.451 [3] Chao, Junzhi. (2019). Modeling and Analysis of Uber's Rider Pricing. 10.2991/aebmr.k.191217.127.

[4] Zhao, Kai & Khryashchev, Denis & Huy, Vo. (2019). Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability. IEEE Transactions on Knowledge and Data Engineering. PP. 1-1. 10.1109/TKDE.2019.2955686.

[5] Faghih, Sabiheh & Safikhani, Abolfazl & Moghimi, Bahman & Kamga, Camille. (2017). Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study.

[6] Predicting real-time surge pricing of ride-sourcing companies Matthew Battifaranoa , Zhen (Sean) Qiana,b, A Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States b Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213, United States