Report On

Taxi Trip Fare Prediction

Submitted in partial fulfillment of the requirements of the ML Course project in

Semester VII of Final Year Computer Engineering

by

Aman Sheikh (Roll No. 13)

Harshpratap Singh (Roll No.16)

Akhila Anilkumar (Roll No. 17)

Mentor

Prof. Megha Trivedi

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(A.Y. 2023-24)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the Mini Project entitled **" Taxi Trip Fare Prediction "** is a bonafide work of **Aman Sheikh(Roll No. 13) ,Harshpratap Singh(Roll No. 16) , Akhila Anilkumar (Roll No. 17)** submitted to the University of Mumbai in partial fulfillment of the requirement forthe award of the degree of **"Bachelor of Engineering"** in Semester VII of Final Year **"Computer Engineering" .**

Supervisor
Prof. Sneha Mhatre

Dr Megha Trivedi
Head of Department

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Course Project Approval

This ML Course Project entitled "Taxi Trip Fare Prediction" by **Aman Sheikh(Roll No. 13) ,Harshpratap Singh(Roll No. 16) ,Akhila Anilkimar(Roll No. 17)**is approved for the degree of **Bachelor of Engineering** in Semester V of Third Year **Computer Engineering .**

Examiners

1…….

(Internal Examiner Name & Sign)

2……

(External Examiner name & Sign)

Date:

Place:

Contents

Abstract

Data that is an archive is used in predictive analysis to estimate longer-term events. Using historical information, a mathematical model is utilised to capture relevant trends. The model then uses current data to make longer-term predictions or to determine activities that call for optical results. Predictive analytics has recently received a lot of praise, thanks to advancements in the machine learning and large data support technologies. Many sectors employ predictive analytics to create precise forecasts, such as providing the cost of a city ride. The prediction, for instance, makes it possible to more accurately predict cab fares, which facilitates resource planning. When starting a cab business, many aspects are taken into account. This project tries to know the patterns and use different methods for fare prediction. This project is developed for predicting the cab fare amount within a certain city. The project involves different steps like training, testing by using different variables like pickup, drop-off location for predicting cab fare.

Acknowledgments

We have immense pleasure in presenting the report for our project entitled “project name”.

We would like to take this opportunity to express our gratitude to a number of people who have been sources of help & encouragement during the course of this project.

We are very grateful and indebted to our project guide and respected HOD Dr Megha Trivedi for providing their enduring patience, guidance & precise suggestions. They were the one who never let our morale down & always supported us throughout the project.

Introduction:

1.1 Introduction:

Taxis play an important role as a transportation alternative in many cities. In developed countries, **taxis** tend to be used as a substitute for private vehicles. Going for a cheap taxi ride is so much more convenient than driving yourself to different places and going through the hassle of finding the right place to park your car. A stress-free and comfortable ride in the taxi not only saves the fuel of your transportbut also saves up your driving energy too.

For forecast utilizing frameworks the AI idea is generally utilized everywhere. There are various kinds of approaches utilizing AI for forecast some of them are directed, unaided learning. In issues related with business needs, AI can likewise be called prescient. AI is inferred into various. Supervised learning is one of the most commonly used learning in machine learning, to train our data here we need some factors like the dataset and we have to use algorithms to predict the output of the program which we are doing. If we do not have a data set to train our model then the output which we get is less reliable and the fare will not be accurate. Unsupervised learning is like learning without the help of any dataset or external factor, it automatically selects a suitable algorithm and tries to predict the output which will be accurate most of the time. On account of registering advances, the AI which we are seeing today isn't at all like AI which was utilized before some time. This explicitly demonstrates that frameworks can do the errands which were thought to be finishedby people as it were. AI can likewise be utilized in the area of Artificial insight. It can also be used in data mining. We can perform many things at an affordable price and also can do many complex things easily with the help of the advancement of machine learning. It is principally used to make precise forecasts by utilizing an alternate scope of strategies and calculations it gives to individuals. So it is broadly utilized in many organizations since we utilize the idea of AI day by day in our everyday lives and furthermore it will be utilized in the future likewise as its extent is exceptionally high.

1.2 Problem Statement :

Building a predictive model so as to suggest the taxi fare based on various factors like duration of trip, number of customers, tip amount, pay type etc using machine learning models like logistic regression, random forest or gradient boosting algorithm to acquire a model with better precision and accuracy.

1.3 Objectives :

This project is intended on predictive analysis of taxi fare which can help customer to get an accurate, correct and precise fare amount on the basis of all fact values that have been entered by the customer itself.

1.4 Scope:

Helps customer as well as taxi drivers to predict an accurate rate for the tour without even need of questionnaire to the customer, other taxi drivers or not even need to do random searches in google.

1.5 Course Project Contribution:

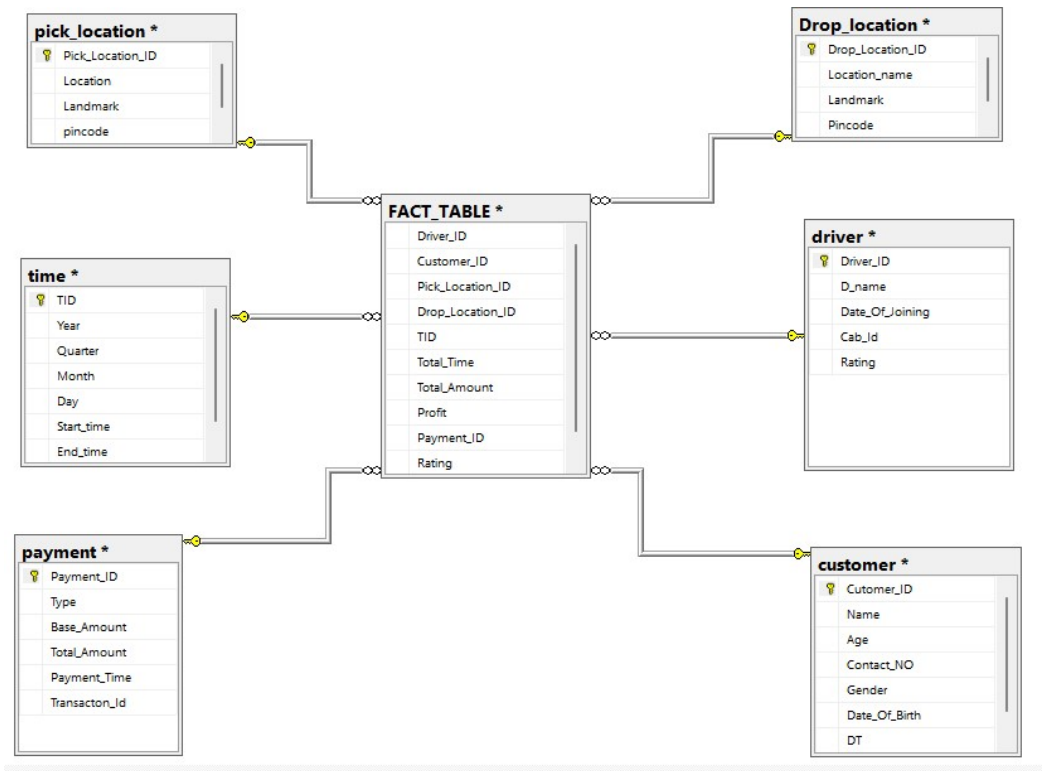
Each individual has their own strength and weakness. While learning and implementing this project, group was well structured and everyone contributed their 100 percent to it. We assigned each other a portion of project to accomplish, pulled together everyone inputs and reviewed each others work and updated weekly .We all conducted the research for the selection and continuation of the project at our own side. We all have equally developed the code by conducting google meet .

Proposed System:

2.1 Introduction:

For predicting the longer-term events predictive analysis use data which is an archive. For capturing the trends which are important mathematical model is used from past data. The model then uses present data to predict the longer-term or to derive actions to require optical outcomes tones of appreciation in recent time for predictive analytics thanks to development in support technology within areas of massive data inmachine learning. Many industries use predictive analytics for making an accurate forecast like giving the amount of fare for the ride within the city. These resource planning are enabled by the forecast as an example, cab fare can be predicted more accurately. A lot of factors are taken into consideration for a taxi start-up company. This project tries to know the patterns and use different methods for fare prediction. This project is developed for predicting the cab fare amount within a certain city. The project involves different steps like training, testing by using different variables like pickup, drop-off location for predicting cab fare.

2.2 Star Schema:



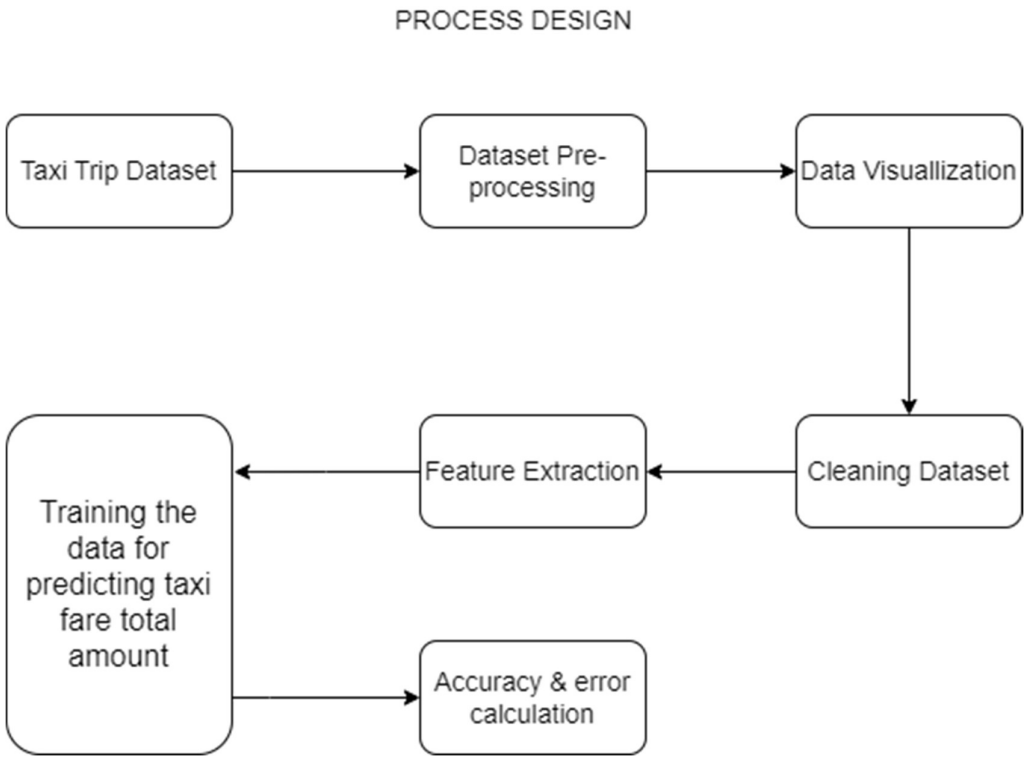
2.3 Algorithm & Process Design:

2.3.1 Algorithm:

Gradient Booster Regressor : Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. While you can build barebone gradient boosting trees using some popular libraries such as XGBoost or LightGBM without knowing any details of the algorithm, you still want to know how it works when you start tuning hyper-parameters, customizing the loss functions, etc., to get better quality on your model.

2.3.2 Process Design:

The below figure shows how the data flow through the entire process of prediction and analysis.



2.4 Data Dictionary:

The datasets that we have is used is of New York City. It contained 83,692 data entries and 17 features.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate 2=JFK** **3=Newark** **4=Nassau or**

	Westchester **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.

Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.
Congestion_Surcharge	Total amount collected in trip for NYS congestion surcharge.
Airport_fee	$1.25 for pick up only at LaGuardia and John F. Kennedy Airports

2.5 Experiment and Results for Validation and Verification

PREPROCESSING

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# % matplotlib inline
plt.style.use('seaborn-whitegrid')
df_train =  pd.read_csv('train.csv', nrows = 2_000_000)
df_train.dtypes
df_train.describe()
print(df_train.isnull().sum())
df_train.groupby('trip_distance').mean()
df_train=df_train.fillna(df_train.groupby('trip_distance').transform('mean'))
```

```
FEATURE EXTRACTION:
df_train=df_train.drop('store_and_fwd_flag',axis=1)
print('Old size: %d' % len(df_train))
```

```
df_train = df_train[df_train.fare_amount>=0]
print('New size: %d' % len(df_train))
df_train=df_train.drop('ehail_fee',axis=1)
df_train=df_train.drop('lpep_pickup_datetime',axis=1)
df_train=df_train.drop('lpep_dropoff_datetime',axis=1)
```

VISUALIZATION:

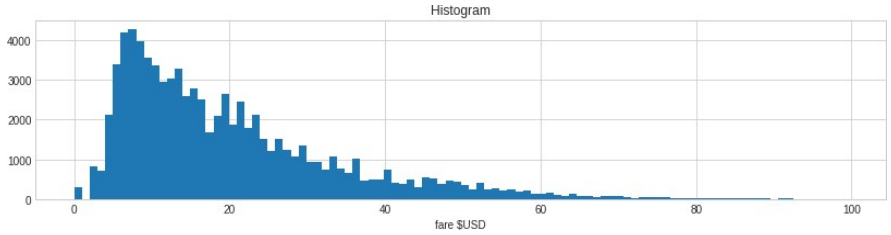
```
# plot histogram of fare
```

1)
```
df_train[df_train.fare_amount<100].fare_amount.hist(bins=100
, figsize=(14,3))
 plt.xlabel('fare $USD')
 plt.title('Histogram');
```

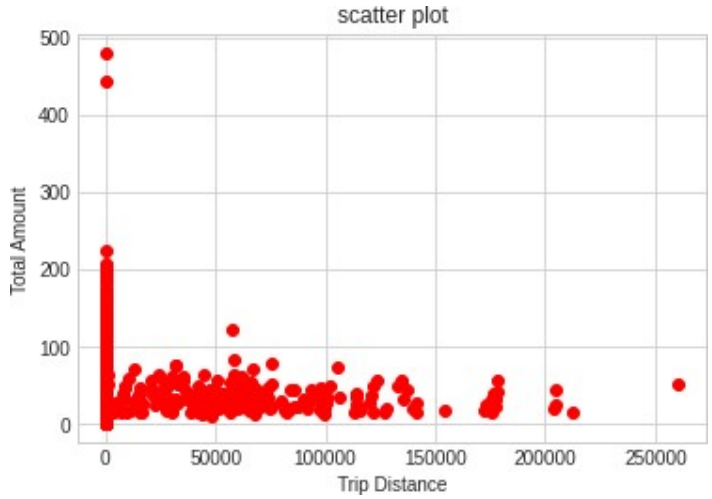
2)
```
plt.scatter(df_train['trip_distance'],df_train['total_amount
'], color='red')
plt.title("scatter plot")
plt.xlabel("Trip Distance")
plt.ylabel("Total Amount")
plt.show()
```

3)
```
plt.scatter(df_train['trip_distance'],df_train['fare_amount'
], color='red')
plt.title("scatter plot")
plt.xlabel("Trip Distance")
plt.ylabel("Fare Amount")
plt.show()
```

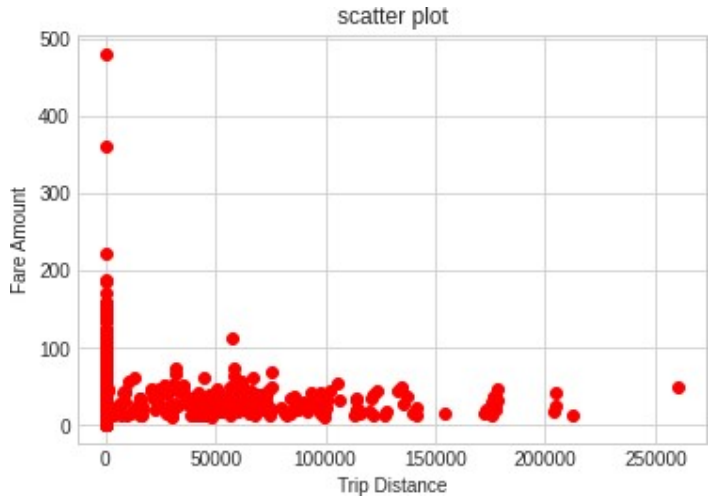
4)
```
df_train[df_train.trip_distance<30].trip_distance.hist(bins=
100, figsize=(14,3))
plt.xlabel('Trip Distance (in km)')
plt.title('Histogram');
```



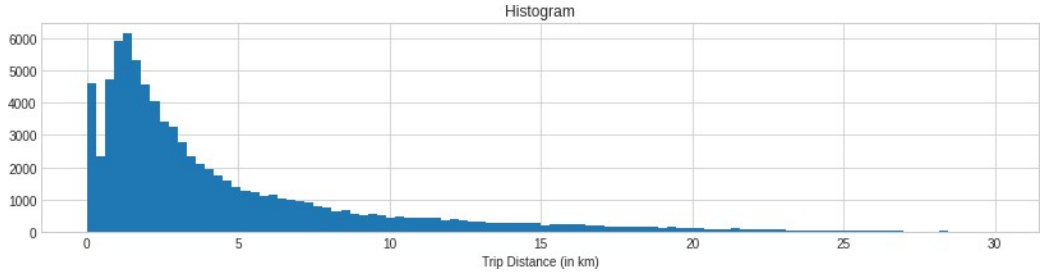
(1)



(2)



(3)



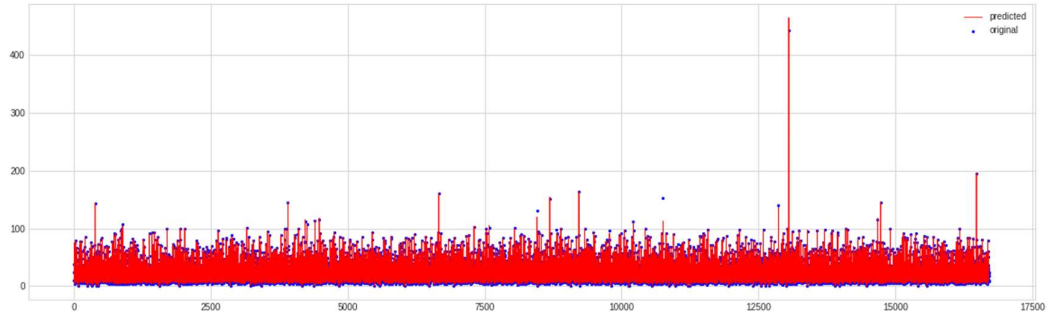
(4)

MODEL TRAINING:

```
# example of a normalization

from numpy import asarray

from sklearn.preprocessing import MinMaxScaler
```

```
scaler= MinMaxScaler()


x= scaler.fit_transform(x)

x

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

from sklearn.metrics import mean_squared_error

from scipy.stats import pearsonr

from sklearn.model_selection import cross_val_score,
cross_val_predict

from sklearn import metrics

warnings.filterwarnings('ignore')

%matplotlib inline

from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
#80% for Training and 20% for Testing

print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)

from sklearn import ensemble

df_predict = ensemble.GradientBoostingRegressor(n_estimators =
100, max_depth = 5, min_samples_split = 2,

          learning_rate = 0.1, loss = 'ls')

df_predict.fit(x_train, y_train)

df_predict_test=df_predict.predict(x_test)

df_predict_train=df_predict.predict(x_train)

pd.DataFrame({'actual unseen data':y_train,'predicted unseen
data':df_predict_train})

scores = cross_val_score(df_predict, x_test, y_test, cv=5)

scores

predictions = cross_val_predict(df_predict, x_test, y_test, cv=5)

accuracy = metrics.r2_score(y_test, predictions)

accuracy
```

```
x_ax = range(len(y_test))

plt.figure(figsize=(20,6))

plt.scatter(x_ax, y_test, s=5, color="blue", label="original")

plt.plot(x_ax, df_predict_test, lw=0.8, color="red",
label="predicted")

plt.legend()

plt.show()
```



ERROR CALCULATION:

```
print('MAE=
',metrics.mean_absolute_error(y_test,df_predict_test))

print('MSE= ',metrics.mean_squared_error(y_test,df_predict_test))

print('R2 value= ',df_predict.score(x_test,y_test))

print('Adjusted R2 value= ',1 - (1 -
(df_predict.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))

print('RMSE (train)=
',np.sqrt(mean_squared_error(y_train,df_predict_train)))

print('RMSE (test)=
',np.sqrt(mean_squared_error(y_test,df_predict_test)))
```

OUTPUT:

```
MAE=  0.2643833352910766

MSE=  0.471513715842156

R2 value=  0.9984364867629286

Adjusted R2 value=  0.9984155000080686

RMSE (train)=  0.3923853209502581

RMSE (test)=  0.6866685633128664
```

2.6 Conclusion:

Precise fare the taxi trip is predicted on the basis of data entries provided by the customer on ride. The accuracy of the model created is 98.4%

Future Work :

The future work may aim to create more efficient models using other data mining classification techniques such as support vector machine, principal component analysis, etc. We can also widen the scope of the project by focusing on other areas aswell.

References :

LINKS : https://www.kaggle.com/datasets/anandaramg/taxi-trip-data-nyc

Plagarism Report: