A REPORT ON

DATA SCIENCE Mini Project

FARE AMOUNT PREDICTION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE IN THE PARTIAL FULFILLMENT FOR THE AWARD OF

THE DEGREE OF BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

BY

- MISS. SANYOGITA DATTATRAY LONDHE. ROLL NO -358
- MISS.PRATIKSHA VIJAY RAUT. ROLL NO-374
- MISS.PAYAL PRAMOD BHOSALE ROLL NO-309
- MISS.POOJA KAILAS MEHER.ROLL NO-361

UNDER THE GUIDANCE OF PROF. S.N.NIMBALKAR

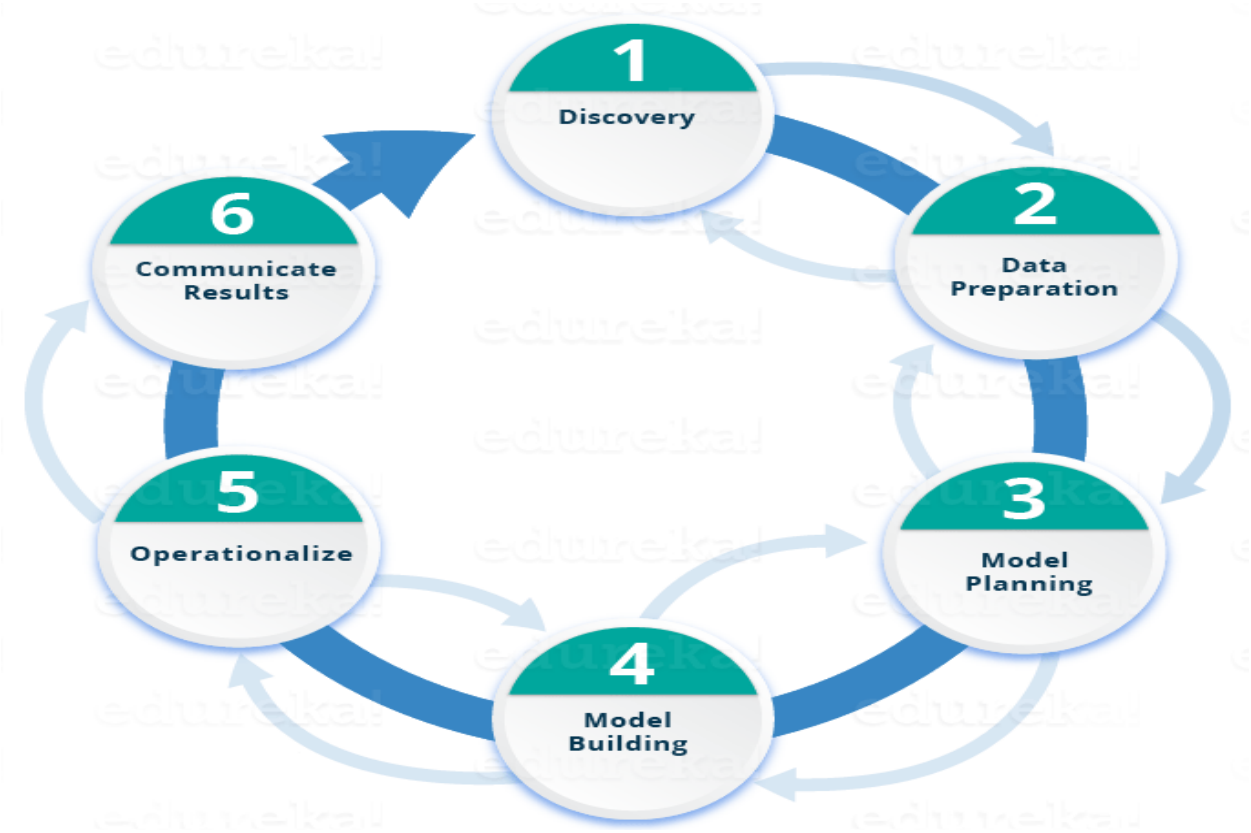
DEPARTMENT Of

COMPUTER ENGINEERING

Contents:

INTRODUCTION:-

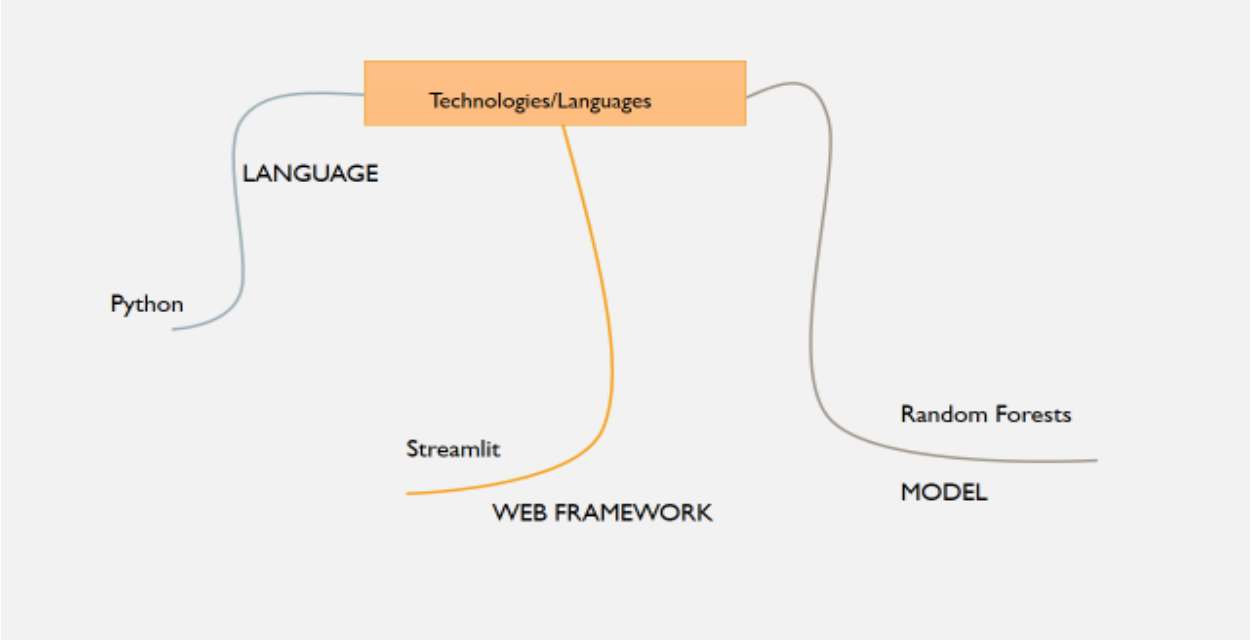
The taxi fare problem is one of several real-world problems that are used as case studies in the series of courses.Random Forest is far more flexible than a Linear Regression model. This means lower bias, and it can fit the data better. Complex models can often memorize the underlying data and hence will not generalize well. Parameter tuning is used to avoid this problem.



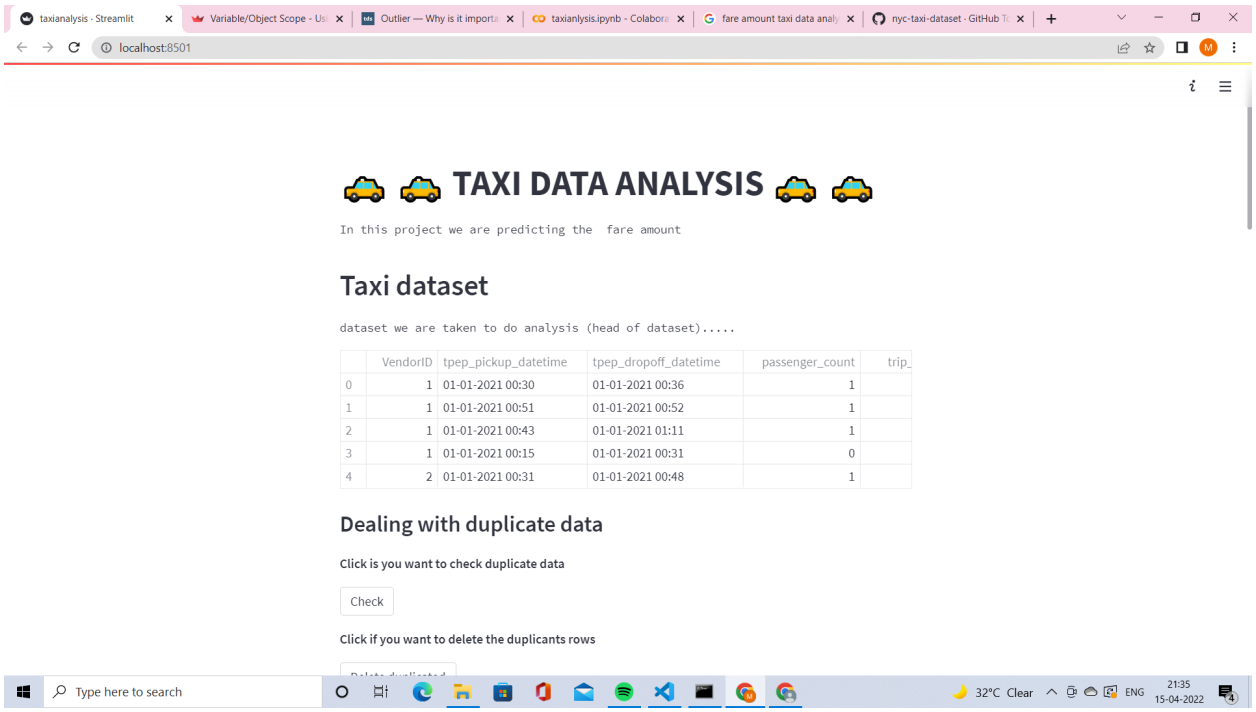
PROBLEM STATEMENT: -

We take cab rides on a regular basis (sometimes even daily!), and yet when we're hitting that 'Book now' button, we rely on manual on-the-fly calculations rather than hardcore ML ones. And that's what I aim to demonstrate here.

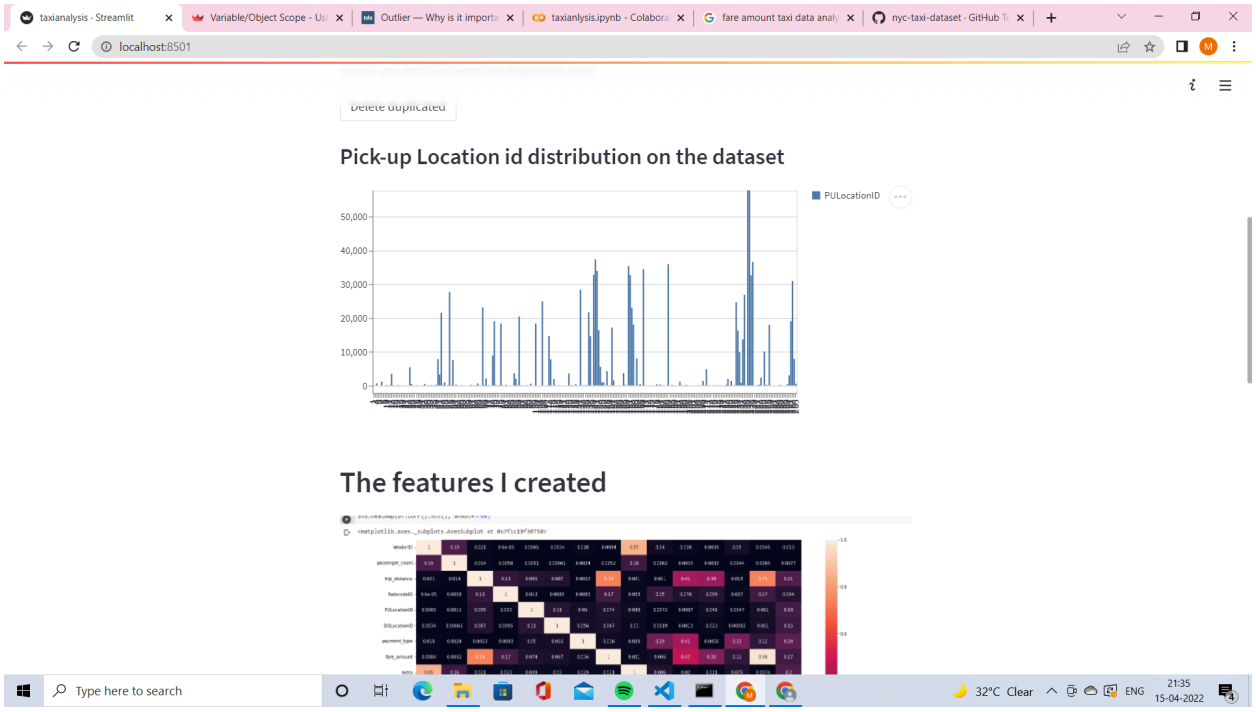
SOFTWARE REQUIREMENT SPECIFICATION:



GRAPHICAL USER INTERFACE

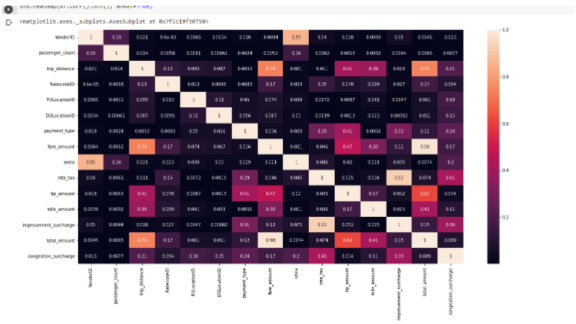


	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_
0	1	01-01-2021 00:30	01-01-2021 00:36	1	
1	1	01-01-2021 00:51	01-01-2021 00:52	1	
2	1	01-01-2021 00:43	01-01-2021 01:11	1	
3	1	01-01-2021 00:15	01-01-2021 00:31	0	
4	2	01-01-2021 00:31	01-01-2021 00:48	1	



taxianalysis - Streamlit | Variable/Object Scope - Us | Outlier — Why is it importa | taxianlysis.ipynb - Colabora | fare amount taxi data analy | nyc-taxi-dataset - GitHub T

localhost:8501



Correlational matrix to check relations between variables

- **first feature:** From above correlational matrix we clearly know that our target variable is fare_amount
- **second feature:** fare_amount is depend on DOLocationID ,payment_type,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,congestion_surcharge

Time to train model

Here you get to choose the hyperparameter of model and see how the performance chang

what sholud be the max_depth of the model?

Type here to search | 32°C Clear | ENG | 21:35 15-04-2022

taxianalysis - Streamlit | Variable/Object Scope - Us | Outlier — Why is it importa | taxianlysis.ipynb - Colabora | fare amount taxi data analy | nyc-taxi-dataset - GitHub T

localhost:8501

Time to train model

Here you get to choose the hyperparameter of model and see how the performance chang

what sholud be the max_depth of the model?

20

10 100

how many trees should there be?

100

here is a list of features in my data

	0
0	VendorID
1	tpep_pickup_datetime
2	tpep_dropoff_datetime
3	passenger_count
4	trip_distance
5	RatecodeID
6	store_and_fwd_flag
7	PULocationID
8	DOLocationID
9	payment_type

which feature should be used as input feature?

PULocationID

mean absolute error of model is

27.36624393992797

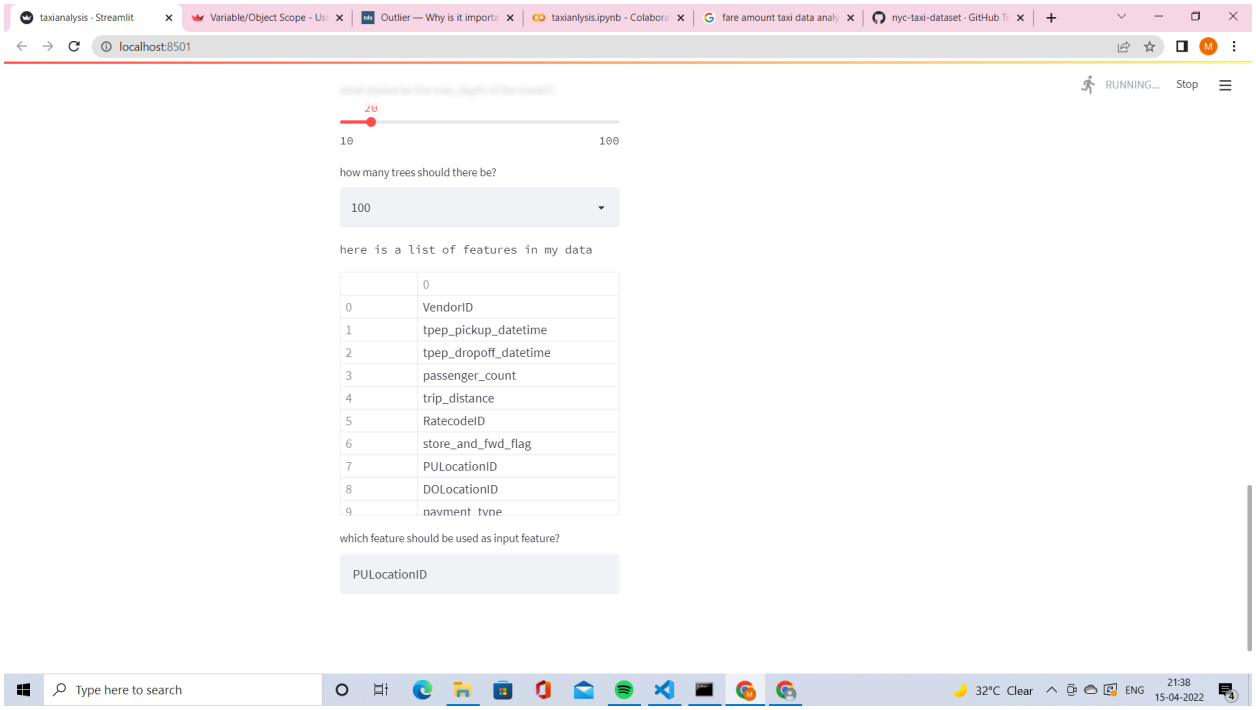
mean squared error of model is

1622.1387772067344

r2 score of model is

-5.136961691562553

Type here to search | 32°C Clear | ENG | 21:35 15-04-2022


taxianalysis - Streamlit
localhost:8501
RUNNING...
Stop
20
10
100
how many trees should there be?
100
here is a list of features in my data
0
0 VendorID
1 tpep_pickup_datetime
2 tpep_dropoff_datetime
3 passenger_count
4 trip_distance
5 RatecodeID
6 store_and_fwd_flag
7 PULocationID
8 DOLocationID
9 payment_type
which feature should be used as input feature?
PULocationID
Type here to search
32°C Clear
ENG
21:38
15-04-2022

CONCLUSION:

Random Forests significantly improved the predictive ability of our Machine Learning model. Another way to improve the model's accuracy is to increase the amount of training data, and/or building ensemble models. If we have a lot of dimensions (features) in the data, dimensionality reduction techniques can also help improve the model's accuracy.