OPTIMISED ROUTE PREDICTION ALGORITHM

CS 6111 CREATIVE AND INNOVATIVE PROJECT

DATED 06/05/2021

PRESENTED BY

BALAJI S (2018101014)
DHANANJEYAN AK (2018103523)
ASHWATH NARAYAN KS (2018103517)

OBJECTIVES

Urban transportation is going through a rapid and significant evolution. In the recent past, the emergence of the Internet and of the smart-phone technologies has made us increasingly connected, able to plan and optimize our daily commute while large amounts of data are gathered and used to improve the efficiency of transportation systems. Today, real-time ridesharing companies like Uber or Lyft are using these technologies to revolutionize the taxi industry, laying the ground for a more connected and centrally controlled transportation structure, and building innovative systems like car-pooling.

A field that can make such important contributions is vehicle routing, i.e., the optimization of each vehicle actions to maximize the system efficiency and throughput. In this paper, we have proposed

- A working model that predicts optimal route for taxi travel with a lesser time complexity.
- Optimal routes based on distance and weather constraints

INTRODUCTION TO DOMAIN

Travel optimization needs many criterias to look at. There are many attributes which contribute to travel path. The features should be looked at and look for optimal path considering the surroundings and physical and social conditions of the location.

The goal of the project is to optimize travel routes for a delivery vehicle by using **machine learning model predictions.** This is a two-component problem: first, we train a machine learning model on the data to predict how long it will take a delivery vehicle to go from point one point to another, and we feed these predictions into a **Genetic algorithm** which decides which is the most time efficient visit order for a given set of points.

"Genetic algorithm (GA) is a type of algorithm inspired by the process of natural selection to generate high-quality solutions to problems, which otherwise would be too difficult to solve."

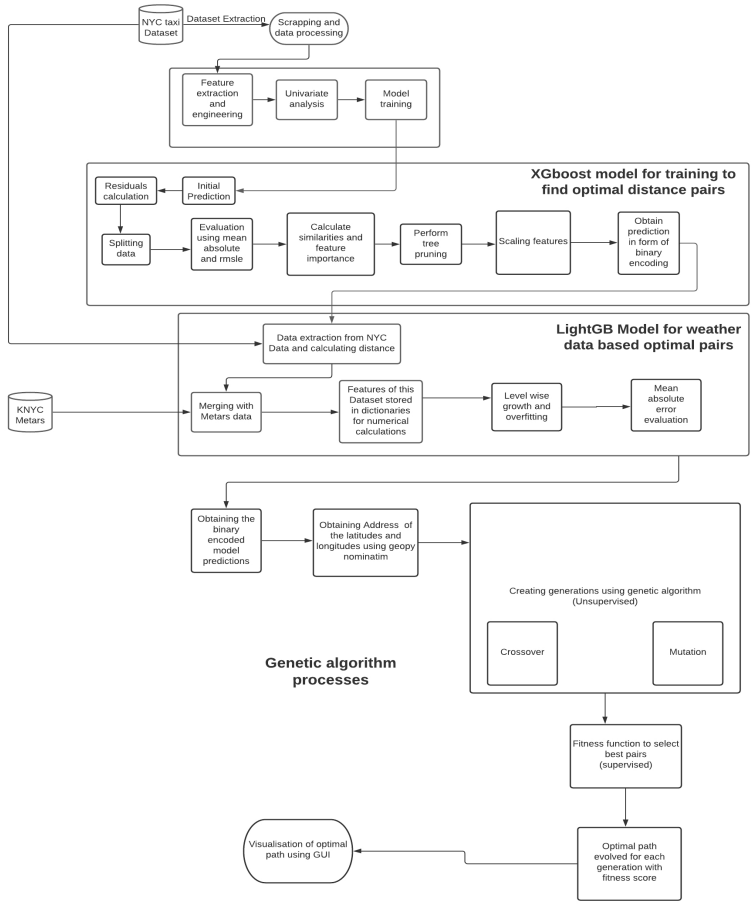
METHODOLOGY

Data preprocessing and feature scaling -> XGBoost Training -> LGBM Training -> Genetic algorithm -> Visualisation

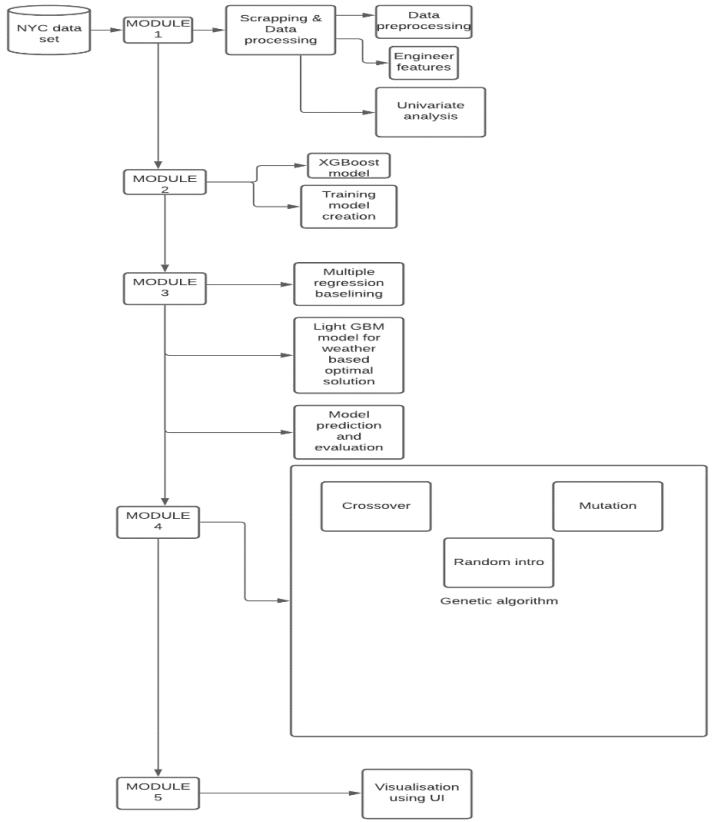
First we obtain taxi data (based on distance) to find optimal path pairs based on distance using a XGboost model .Then we work on weather data (in most cases optimal path metric can be weather and climatic conditions) from Metars dataset .Using Light gradient boost model, we try to obtain optimal path pairs based on weather . Then, we use another training model to predict number of pickups as accurately as possible for like 10 min interval. 10 min is choosen because for a metropolitan city,one can commute 1 mile with normal traffic.

We perform univariate analysis, regression and create training models in this case also. Finally optimal path pairs are given into genetic algorithm, populated as chromosomes, processes like crossover, mutation takes place to find optimal path among the pairs. Then a GUI is used to visualize the optimal travel path.

SYSTEM ARCHITECTURE



MODULAR FLOW



MODULE SPLITUP

MODULE 1.Data processing , Feature extraction ,univariate and bivariate analysis:

DESCRIPTION:

The first step is to obtain details of vehicle users and data of travel based on their destination routes. We will explore the data and modify dataset as per the our requirement for the further analysis of the problem. Data is obtained from dataset train.csv which contains details of taxi travels in new York city for an amount of period. Then we find unique id's in dataset.We alter date formats for convenience and use calculate function to find distance between pickup and dropoff coordinates.We import several packages like numpy,pandas for calculations . Then we perform univariate analysis before proceeding to training sets.

MODULE 1 CONTD.

Algorithm:

- Importing packages
- Import training set split from sklearn model
- Train.csv is read by a variable
- Sample[“dropoff time]=convert into (Y,m,d ,H,M,S) format.
- New construct of other variables

Input: Dataset of NYC taxi (2020) and KNYC_Metars (weather data).

Output: Data (Table) of vehicle id ,pickup,drop location,latitude,etc…

MODULE 2

MODULE 2.Training model creation and evaluation using weather and distance data:

DESCRIPTION:

After determining the objectives of delivery users, the travel time is analyzed as the main evaluation indicators of the optimal route. The travel time is the upper target. When the travelling distance of two paths are different, the optimal path will be updated to the path with lower distance considering the traffic and weather input from a training model.For any given set of locations, these location are fed to the machine learning model, which predicts how long it will take to travel between each two given points. Then ,other features are engineered and values according to calculation and modelling is done.

MODULE 2 CONTD.

The data is split into training set ,test and validation sets. We extract details of class labels from pandas .Then the XGBoost model is trained and contents saved in a (.sav) file for quick access for upcoming processes.(As dataset elements are huge ,conversion to byte stream helps for easy access).

Input : Instances of modified dataset.

Output : Training model

Evaluation metric for model:

Square(log(predicted+1) – (log(true+1)).mean()**0.5

MODULE 3

MODULE 3.Creating XGBOOST and LightGBM models and training of data:

DESCRIPTION:

We need a model to train on our dataset to serve our purpose of predicting the NYC taxi trip duration given the other features as training and test set. Since our dependent variable contains continuous values so we will use regression technique to predict our output.

We use multiple regression techniques to explain relationship between one continuous dependent variable and two or more independent variables.The values taken as training data from the input dataset is analysed(classified) using the Linear Regression model. In the LR model, the spots is in the X axis. For example , yy is distance in the X axis the corresponding cost of the Y axis is xxx and so on. Like this the values are plotted in the graph and analyzed using the LR model. Also, for model complexity, and XGBoost model is used to numeric and categoric features.

MODULE 3 CONTD.

The aim is also to use a LightGBM model when adding weather data because this package can handle categorical data and runs faster..

In addition to the functionality of training set,this LightGBM model adds weather data to predict optimal routes based on weather also.This is a experimentary try to find optimal route across a network of roads.Then ,Model prediction and evaluation takes place.

Output: Optimal Distance between two points(node to node)

MODULE 4

GENETIC ALGORITHM:

Traditional travelling salesman problem requires enormous amount of time to solve the np hard problem for large number of nodes. So using machine learning approach of chromosomal genetic connections ,we try to use the genetic algorithm to this area of optimal route prediction.

The distance between two points is taken as chromosomes. They're sorted according to their random number. Then crossover between two parents occur to form offsprings. (Parent is cities and produced offsprings are helpful finding optimal path). In Crossover, parents fuse to form fittest offsprings.

MODULE 4 CONTD.

Mutation: Original two point distance array is mutated into required style

Step1:

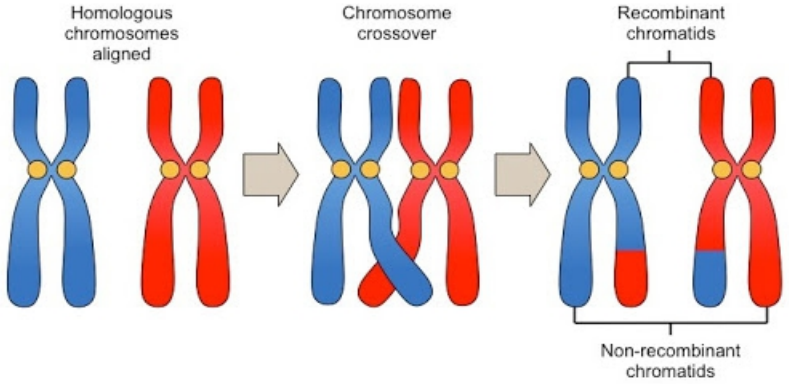
Generate a double array of size (n-1). First field is random number and second is sequence (numbers). Sort array based on first column , Sequence in second column gives chromosomes.For example:The first table is converted into second,

Random number	Sequence
0.23	2
0.65	3
0.49	4
0.58	5
0.75	6
0.34	7

Random number	Sequence
0.23	2
0.34	7
0.49	4
0.58	5
0.65	3
0.75	6

MODULE 4 CONTD.

Step2: (Crossover):



Parent 1:

1	3	4	2	5	7	6

Parent 2:

1	7	5	2	3	4	6

- Crossing over with 576

MODULE 4 CONTD.

Offspring 1:

1	2	3	4	6	7	5

Crossing over 346 (Offspring 2)

1	2	5	7	4	6	3

Mutation:

1	3	4	2	5	7	6

MODULE 4 CONTD.

Offspring: (Changing 425)

1	3	5	2	4	7	6

Pseudocode for genetic algorithm:

For 50 points(nodes),

For(iteration count < threshold){

Gen(50 chromosomes ie. distances);

Gen(32 chromosomes through cross over);

Gen(14 chromosomes over mutation);

Insert.chromosomes(4);

Select(50 best fit chromosomes ie. 32+14+4);

Iteration count++;

}

Present best solution obtained so far

Output: Optimal path

MODULE 5

PRESENT OPTIMAL ROUTE

Then , optimal path is visualized into a path in route map using interface as it allows users to navigate between their paths easily and reach their destination in the quickest route with minimum amount of time.Visualisation of datapaths would be performed by folium maps and creations of gpx HTML files which can be loaded onto MyMaps to see the optimization.

FINAL OUTPUT: Visual optimal path between source to destination.

EXPERIMENTAL RESULTS

DATASET DESCRIPTION:

Nyc Taxi Dataset

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

KNYC Metars (Weather Data)

METAR is a format for reporting weather information. A METAR weather report is predominantly used by pilots in fulfillment of a part of a pre-flight weather briefing, and by meteorologists, who use aggregated METAR information to assist in weather forecasting.

DIFFERENT INPUT TEST CASES

TEST CASE ID	NO.OF TEST LOCATIONS	PRODUCED OUTPUT
TC_01	5	['L4', 'L5', 'L1', 'L2', 'L3', 'L4']
TC_02	2	Unfortunately , two node routes are not predicted correctly , resulted in a lengthier path than expected.
TC_03	1	It produces a **travel time of 3.25 mins** by the model which is not possible.So this scenario **cannot be considered.**
TC_04	3	[‘L3’,’L1’,’L2’,’L3’]
TC_05	11	[‘L9’,’L10’,’L11’,’L3’,’L5’,’L8’,’L2’,’L1’,’L7’,’L6’,’L4’,’L9’]
TC_06	7	[‘L7’,’L5’,’L6’,’L1’,’L4’,’L3’,’L2’,’L7’]
TC_07	9	[‘L9’,’L1’,’L3’,’L6’,’L8’,’L5’,’L4’,’L7’,’L2’,’L9’]

PERFORMANCE METRICS

Distance data is processed , feature scaling is done and model used to predict is Random Forest Regressor which results in high error scores.

We use XGBoost Model to reduce RMS error to produce accurate results with high degree of ensembling . We combine distance and weather using ensembling technique of boosting the weaker features to produce strong learners. If we use Random Forest this specific focus will be missed.

At last Point we can see that our MSE is around 10000 and the MSE Error in Random Forest was two times greater.

We can observe that XGBoost model produces lesser RMSE error compared to Random Forest regressor and hence it is improved to produce predictions for travel.

Also ,the gap at end indicates the overfitting conditions.

Model : Regression Model with Ensembling

Hyper parameters:Learning Rate

Evaluation : Root mean square error

Epochs : Varying epochs

Testing: Mean absolute error

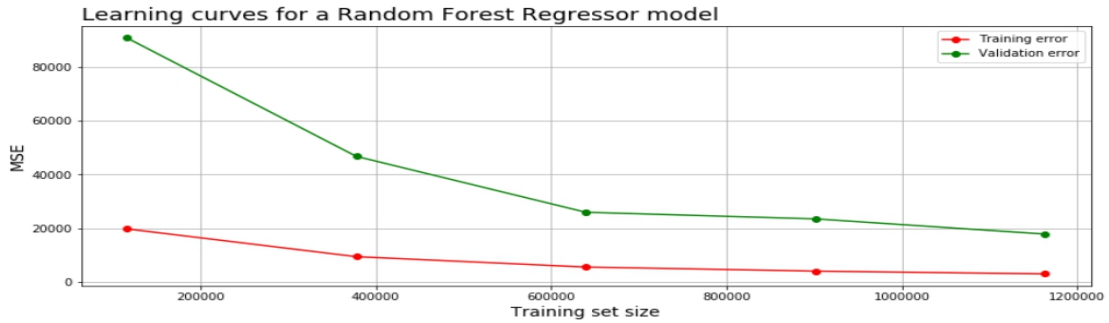
Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. Furthermore, the learning rate affects how quickly our model can converge to a local minima (aka arrive at the best accuracy).

INFERENCE TABLE

Test	Validation	Epochs	LearningRate	RMSE(Test)	RMSE (Eval)	Mean Error	Mean Error Final (All features considered)
30%	25%	500	0.2	0.07641	0.34768/0.34767 **(Underfitting)**	4.9699	4.906
40%	35%	700	0.1	0.07913	0.3417/0.3417	5.073	4.990
40%	35%	1000	0.1	0.0565	0.34188/0.34188	5.0449	4.991
40%	35%	1300	0.1	0.04191	0.34197/0.34197	5.027	4.9928
30%	25%	2000	0.05	0.0944	0.327426/0.327425/0.327426**(Indicates overfitting.Hence we stop the epoch)**	4.9564	4.8273

GRAPHICAL REPRESENTAITON OF RESULT

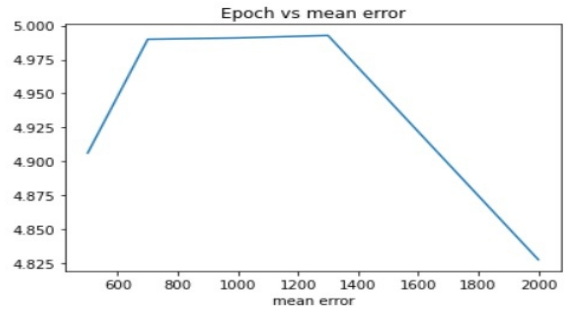
Learning Curve of Random Forest:



Learning curve of XGBoost (Used in our model):

EPOCHS VS MEAN ERROR:

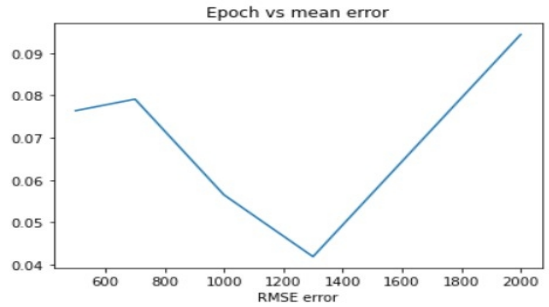
In [6]:
```
import matplotlib.pyplot as plt

#line graph
a = [500,700,1000,1300,2000]
b = [4.906,4.990,4.991,4.9928,4.8273]
plt.plot(a,b)
plt.title("Epoch vs mean error")
plt.xlabel("epoch")
plt.xlabel("mean error")
plt.show()
#scatter graph
x_values = [500,700,1000,1300,2000]
squares = [4.906,4.990,4.991,4.9928,4.8273]
plt.scatter(x_values,squares, s=10, color = "pink")
plt.show()
```



We can see that mean error gets reduced gradually with increase in epochs.

EPOCHS VS RMSE ERROR:

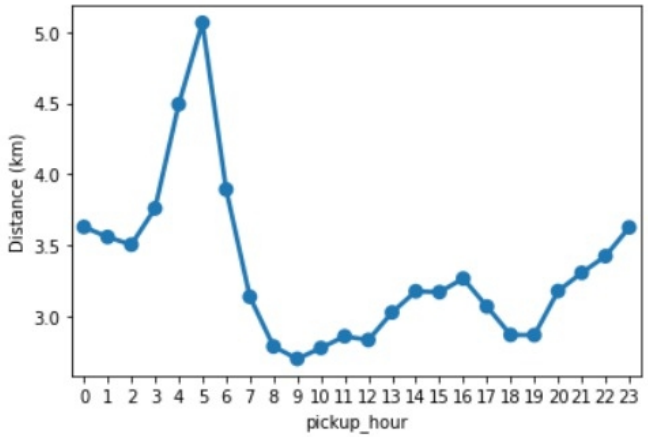
In [7]:
```
import matplotlib.pyplot as plt

#Line graph
a = [500,700,1000,1300,2000]
b = [0.07641,0.07913,0.0565,0.04191,0.0944]
plt.plot(a,b)
plt.title("Epoch vs mean error")
plt.xlabel("epoch")
plt.xlabel("RMSE error")
plt.show()
#scatter graph
x_values = [500,700,1000,1300,2000]
squares = [4.906,4.990,4.991,4.9928,4.8273]
plt.scatter(x_values,squares, s=10, color = "pink")
plt.show()
```



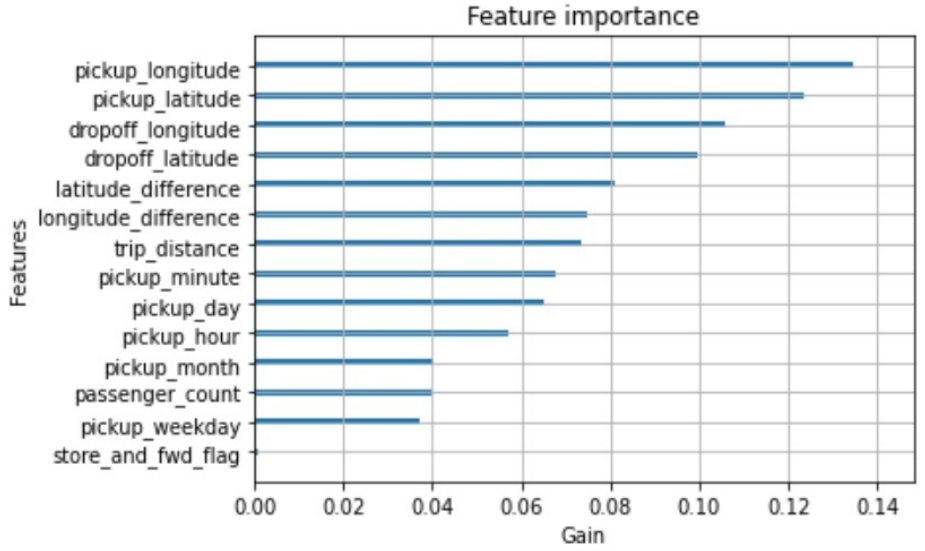
We can see an increase in RMSE error due to **overfitting.**

DISTANCE AND HOUR GRAPH:

```
[9]: group5 = sample_df.groupby('pickup_hour').trip_distance.mean()
     sns.pointplot(group5.index, group5.values)
     plt.ylabel('Distance (km)')
     plt.show()
```



FEATURES VS INFORMATION GAIN:



CONCLUSION

The XGB model had an error of 4.9 minutes in estimating a single trip's duration for a taxi. While this may seem acceptable for one trip, the error may get bigger the more locations are visited. The genetic algorithm itself is fairly straightforward, but it must be noted that every genetic algorithm gives an optimal approximation, but not the single best solution there is since it is a regression problem.

Since the algorithm had difficulties In finding the optimal paths for few of the test locations(For ex: 1 test location),it did better in generating an optimal path for other cases.

REFERENCES

- Kaas, K. Benefit of Traffic Incident Management, National Incident Management Coalition (NTIMC).
- http://www.transportation.org/sites/ntimc/docs/Benefits11-07-06.pdf
- Hagen, L.T. Best Practices for Traffic Incident Management in Florida; CUTR Report 2005, No. 21170543;U.S. Department of Transportation: Washington, DC, USA, 2000.
- Yin, Y. A scenario-based model for fleet allocation of freeway service patrols. Netw. Spat. Econ. 2008, 8,407–417
- https://www.analyticsvidhya.com/blog/2017/07/introduction-to-genetic-algorithm/
- cs.princeton.edu (The Boosting Approach to Machine Learning An Overview)
- https://www.hindawi.com/journals/mpe/2020/6348578/
- https://blog.goodaudience.com/taxi-demand-prediction-new-york-city-5e7b12305475 (Taxi demand prediction)
- Mitchell, M. An Introduction to Genetic Algorithms; MIT Press: Cambridge, MA, USA, 1996; ISBN 9780585030944
- Farradyne, P.B. Traffic Incident Management Handbook; Prepared for Federal Highway Administration, Office of Travel Management; U.S. Department of Transportation: Washington, DC, USA, 2000