International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

2022 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 27th Nov 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 11

DOI: 10.48047/IJIEMR/V11/ISSUE 11/23

Title **PREDICTIVE ANALYSIS OF CAB SERVICES USING PYTHON**

Volume 11, ISSUE 11, Pages: 145-148

Paper Authors

K.Akshay, Dr. G. Venkata Rami Reddy

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

PREDICTIVE ANALYSIS OF CAB SERVICES USING PYTHON

K.Akshay[1], Dr. G. Venkata Rami Reddy[2]

[*1]Student, M. Tech (computer Science), School of Information Technology, JNTU Hyderabad, Hyderabad, India
[*2]Professor of IT, School of Information Technology, JNTU Hyderabad, Hyderabad, India

ABSTRACT: This project is mainly about analysis on the data and predicting outcomes through analysis and drawing different patterns of the data and visualization of the data. Other main aim is to discuss about the price prediction of different Uber cabs that is generated by the machine learning algorithm. This problem belongs to the regression supervised learning category. To implement this two different machine learning algorithms are used such as Random Forest Regressor, and Gradient Boosting Regressor but finally, choose the one that proves best for the price prediction. Need to choose the algorithm which improves the accuracy and reduces overfitting.

Keywords *–Prediction; Gradient Boosting; Random forest; Cab Prediction; Predictive Analysis;*

1. INTRODUCTION

To ensure the demand and supply of cabs we need to analyze the data of cabs and predict the outcomes which can help organizations to ensure their services can meet up to the user's expectations. To implement this project we are using python programming language and different libraries .In this project we are going to predict the outcomes and visualize those outcomes using the data of cabs. through this project we can understand the user ideology towards transportation system .organizations can divide the company's target market into groups of potential customers with similar needs and behaviours. cabs price is predicted using the two models which are random forest and gradient boosting regression .these two algorithms applied on the pre-processed data. Algorithms are come under supervised learning In supervised learningRandom-forest and Gradient-boosting-regression come under supervised learning. In Supervised learning, a training-set and a test-set are there . Examples of input and output vectors are included in both the training and test sets. The objective of supervised learning algorithms is to produce a function that efficiently maps input vectors to output vectors. The Boston-Uber Dataset's price will be predicted using machine learning methods. Several features will be selected from columns. To find significant and practical-patterns in enormous amounts of data, predictive analysis uses computational approaches.

2.LITERATURE SURVEY

2.1 Large Scale Real-time Ridesharing with Service Guarantee On Road Networks

Urban traffic gridlock is a familiar scene. At the same time, the mean occupancy rate of personal vehicle trips in the United States is only 1.6 persons per vehicle mile. Ridesharing has the potential to solve many environmental, congestion, pollution, and energy problems. In this paper, we introduce the problem of large scale real-time ridesharing with service guarantee on road networks. Trip requests are dynamically matched to vehicles while trip waiting and service time constraints are satisfied. We first propose two scheduling algorithms: a branch-and-bound algorithm and an integer programming algorithm. However, these algorithms do not adapt well to the dynamic nature of the ridesharing problem. Thus, we propose kinetic tree algorithms which are better suited to efficient scheduling of dynamic requests and adjust routes on-the-fly. We perform experiments on a large Shanghai taxi dataset. Results show that the kinetic tree algorithms outperform other algorithms significantly.

2.2 Bus Pooling: A Large-Scale Bus Ridesharing Service

Ridesharing, a shared service that uses the information and knowledge matching, can efficiently utilize scattered social resources to reduce the demand for vehicles in urban road networks. However, carried sharing has the problems of low capacity and high cost, and it cannot satisfy demands for recurring, long-distance, and low-cost trips. In this paper, we formally define the bus ride sharing problem and propose a large-scale bus ridesharing service to resolve this problem. In our proposed model, the rider can use an online bus-hailing service to upload his or her trip demand and wait to be picked up when it gathers enough people. The provider assigns drivers to riders after integrating the matched ride requests. To maximize ridesharing success rate, we developed both exact algorithms and approximate algorithms to optimize the ride-matching service. A real-life dataset that contains 65,065-trip instances extracted from 10,585 Shanghai taxis from one day (Apr 1, 2018) is used to demonstrate that our proposed service can provide higher

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

cost performance and on-demand bus services for every ride request. Meanwhile, it reduces the number of vehicles used by 92%and 96% and the amount of oil used by 87% and 92% compared with car ridesharing and no ridesharing, respectively.

2.3 The Merits of Sharing a Ride

The culture of sharing instead of ownership is sharply increasing in individuals' behaviours. Particularly in transportation, concepts of sharing a ride in either carpool-ing or ridesharing have been recently adopted. An efficient optimization approach to match passengers in real-time is the core of any ride sharing system. In this paper, we model ridesharing as an online matching problem on general graphs such that passengers do not drive private cars and use shared taxis. We propose an optimization algorithm to solve it. The Outlined algorithm calculates the optimal waiting time when a passenger arrives. This leads to a matching with minimal overall overheads while maximizing the number of partnerships. To Evaluate the behaviour of our algorithm, we used a NYC taxi real-life data set. Results represent a substantial reduction in overall overheads.

3. IMPLEMENTATION

3.1 Project Architecture

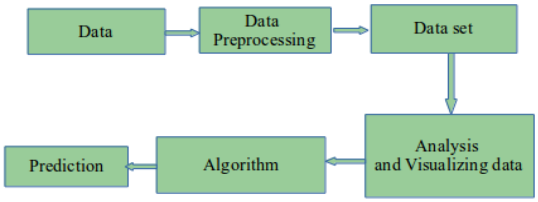

Fig 1: Project Architecture

3.2 PREPROCESSING

First we need to load and collect the data so that we can read the data and have some results at the end. We have to import some basic modules like pandas and numpy all the basic modules. These will help us to create some beautiful visualization to manipulate data. to analyze data in much more efficient way. We have multiple datasets so we need to concatenate or combine all the data sets into one .

Label Encoding

The dataset has a mix of categorical and continuous variables, which are difficult for most machine learning algorithms to comprehend or interpret. This indicates that when the data is represented as numbers rather than categories, machine learning algorithms will perform better. Label coding is the process of turning classifier values into machine-readable digital data. For example to perform analysis need to know data types of each variable .date and time should be timestamp data type if it is not changed to timestamp.

Filling NAN Values

We utilise the isnull() method to check for missing data in a pandas DataFrame. In dataset price column has many Nan values So, fillna() method used to fill the not assigned values in the dataset.

Drop Useless Columns

Many features which cannot be part of the model evaluation, unnecessary features can be deleted. To remove rows or columns based on certain column names and associated axes, we utilise the drop() function.

Feature Engineering

Feature engineering is the most important part of the data analytics process. It deals with, selecting the features that are used in training and making predictions. All machine learning algorithms use some input data to create outputs.

It has mainly two goals:

- Preparing a suitable input dataset that complies with the demands of the machine-learning algorithm.
- Enhancing machine learning model performance.

RFE (Recursive Feature Elimination)

Feature selection is an important task for any machine learning application. This is especially crucial when the data has many features. The optimal number of features also leads to improved model accuracy.

There are two important configuration options when using RFE:

- deciding how many features to choose from (k value)
- the selection of the feature-selection algorithm.

RFE works by selecting a subsets of features from the training-dataset, starting with every feature, and then successfully deleting each column one at a time until the required number remains. Recursive feature removal is being used through scikit-learn via sklearn. feature_selection.RFE class.

4. MODEL GENERATION

4.1 Random Forest

Random forest is a supervised learning algorithm which can be used for both classification and regression problems. It is a collection of Decision Trees. In general, Random Forest can be fast to train, but quite slow to create predictions once they are trained. This is due because it has to run predictions on each tree and then average their predictions to create the final prediction. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications the random forest algorithm is fast enough, but

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

there can certainly be situations. where run-time performance is important and other approaches would be preferred. A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees, with some helpful modifications. Random forest first splits the dataset into a number of samples and then applies a decision tree on each sample individually. After that, the final result is predicted accuracy whose majority is higher among all. Random Forest depends on the concept of ensemble learning. An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

4.2 Gradient Boosting

Gradient boosting is a technique which can be used for both classification and regression problems. This model combines the predictions from multiple decision trees to generate the final predictions. Also, each node in every other decision tree takes a different subset of features for selecting the best split. But there is a slight difference in gradient boosting in comparison to random forest : gradient boosting builds one tree at a time and combines the results along the way. Also, it gives better performance than random forest. Gradient Boosting trains many models in a gradual, additive, and sequential manner.

The modelling is done in the following steps:-

- First split the dataset into a training set and a testing set.
- Then train the model on the training set.
- And At last, test the model on the testing set and evaluate how well our model performs.

Price Prediction

The price prediction function takes the input parameters by taking parameters as input. The parameters are cab name, source, surge multiplier, and icon (weather).Create a manual for users which gives instructions about the input like what do you need to type for a specific thing and in which sequence.

Testing

Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to evaluate the regression problem's accuracy.

5. EXPERIMENTAL RESULTS

Serial No.	Models	Accuracy
1	Random Forest	0.962269474
2	Gradient Boosting Regressor	0.963187213

Fig 2 :Accuracy

No	Testing Model	Accuracy
1	Mean Absolute Error	3.40607721
2	Mean Squared Error	20.0334370
3	Root Mean Absolute Error	4.47587277

Fig 3 : Error table of random forest

No of rides in monthly

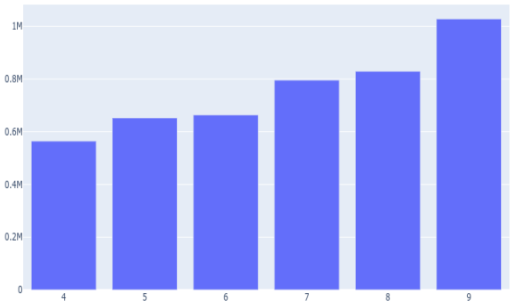

Fig 4: No of Rides

Analysis by each day in month

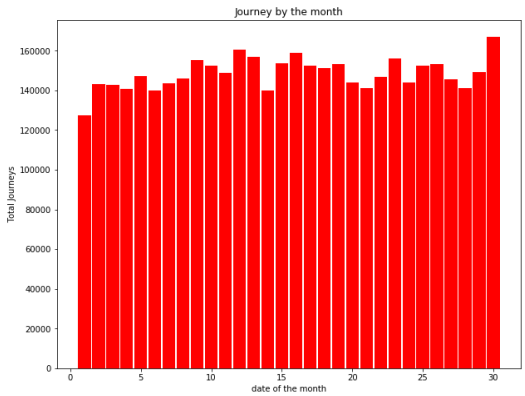

Fig 5: Analysis by each day

6. CONCLUSION

In this work we performed Analysis on the data and predicting outcomes through analysis and drawing different patterns of the data visualization and different graphical representation, also the price variations of other cabs and different types of weather. Two different models On remaining dataset both Random Forest, and Gradient

Boosting Regressor prove best with 96%+ accuracy on training for our model. This means the predictive power of all these two algorithms in this dataset with the chosen features is very high but in the end, we go with random forest because it does not prone to over fitting and design a function with the help of the same model to predict the price. Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to evaluate the regression problem's accuracy.

7. REFERENCES

1. [Agatz et al., 2012] Agatz, N., Erera, A., Savelsbergh, M., and Wang, X. (2012). Optimizationfor dynamic ride-sharing: A review. European Journal of Operational Research, 223(2):295–303.

2.Intelligent Shared Mobility Systems: A Survey on Whole System Design Requirements, Challenges and Future Direction FATEMEH GOLPAYEGANI , MAXIME GUÉRIAU 2 ,PIERRE-ANTOINE LAHAROTTE SAEEDEH GHANADBASHI 1 , JIAYING GUO 1 ,(Graduate Student Member, IEEE), JACK GERAGHTY 1 , AND SHEN WANG.

3. P. Ehsani and J. Y. Yu, ‘‘The merits of sharing a ride,’’ in Proc. Annu. Allerton Conf. Commun., Control, Comput., Oct. 2017, pp. 776–782.

4.L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, T.-M.-T. Nguyen, and J. Jakubowicz, ‘‘Dynamic cluster-based over- demand prediction in bike sharing systems,’’ in Proc. Int. Joint Conf. Pervasive Ubiquitous Comput., 2016, pp. 841–852.

5.J. Liu, L. Sun, W. Chen, and H. Xiong, ‘‘Rebalancing bike sharing systems: A multi-source data smart optimization,’’ in Proc. Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1005–1014.

6.[Ma et al., 2013] Ma, S., Zheng, Y., and Wolfson, O. (2013). T-share: A large-scale dynamic taxi ridesharing service. In Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pages 410–421.

7. C. Tian, Y. Huang, Z. Liu, F. Bastani, and R. Jin, ‘‘Noah: A dynamic ridesharing system,’’ in Proc. ACM SIGMOD Int. Conf. Manage. Data, Jun. 2013, pp. 985–988.

8. N. Masoud and R. Jayakrishnan, ‘‘A decomposition algorithm to solve the multi-hop peer-to-peer ride-matching problem,’’ Transp. Res. B, Methodol., vol. 99, pp. 1–29, May 2017.