Taxi Demand Prediction using Machine Learning.

P. Sudheer Benarji[1], P. Sai Bharadwaj[2], B. Neeha[3], D. Srikanth[4], V. Ankitha[5]

[1]Professor, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad
[2345]Under Graduate Student, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad

---

Abstract - *Taxi demand prediction is the process of using historical data to forecast future taxi requests in a particular area. Managers may pre-allocate taxi resources in cities with the aid of accurate and real-time demand forecasting, which helps drivers find clients more quickly and cuts down on passenger waiting times. This project is aimed to choose the best model in predicting the taxi demand where we use various Machine learning techniques such as regression analysis and time series forecasting. Various baseline models, including moving averages (simple, weighted, and exponential), linear regression with grid search, random forest regressor with random search, and XGBoost regressor with random search are used. We find out which model is more suitable in predicting the output using the metrics we obtain.*

Key Words: Linear Regression with GridSearchCV, Random Forest Regressor with RandomSearchCV, XGBoost Regressor with RandomSearchCV.

1.INTRODUCTION

Taxi demand prediction is the process of using historical data to forecast future taxi requests in a particular area. Managers may pre-allocate taxi resources in cities with the aid of accurate and real-time demand forecasting, which helps drivers find clients more quickly and cuts down on passenger waiting times.

In our project, we've used data on taxi rides in New York city to to train and test the models using Linear regression, Random Forest regressor and XGBoost regressor.

Along with the Linear regression and Random Forest algorithms, we've also used the XGBoost algorithm. XGBoost is a machine learning algorithm that is commonly used in classification and regression problems. It is an ensemble learning method that combines the weak prediction models , such as decision trees to create a stronger overall prediction model. XGBoost has gained popularity due to its high accuracy, scalability, and ability to handle missing data.

With our project, we get an understanding of which model is best to predict the real time taxi demand and taxi companies be able to tailor strategies to allocate resources based on demand.

2. Literature Review

Multi-attention network-based graph prediction of taxi demand: In order to better address the taxi demand prediction problem, this study develops a Graph Multi-Attention Network (GMAN), which tries to forecast the taxi demands in every section of a road network.(Achieved a 72% Accuracy). Because only significant data needs to be learned by the models, applying attention increases model accuracy to extremely high levels. The Attention mechanism's drawback is that it requires a lot of time and is challenging to parallelize

Taxi demand forecast using the random forest model: Decision trees are employed in the random forest. The term "random" refers to our usage of a random bootstrap sampling, and the term "forest" refers to the collection of trees seen in decision trees. (Achieved 77% Accuracy). excellent forecasting abilities that improve application precision. Easy data preparation is made possible by not requiring normalization. Generally speaking, this algorithm is quick to train but takes a while to produce predictions after training

Demand projection for taxis XGBoost algorithm-based: The hot spot locations are identified and their boundaries are drawn using the density-based DBSCAN clustering technique, and the demand for the hot spot areas is predicted using the XGBoost algorithm. XGB provides various features, such as parallelization, cache optimization, and more. Like any other boosting method, XGBoost is sensitive to outliers.

Taxi Demand Forecast Based on Regional Heterogeneity Analysis and Multi-Level Deep Learning: With the aid of the taxi zone clustering technique and pairwise clustering theory, the Multi-Level Recurrent Neural Networks (MLRNN) model is put out.(83.33% Accuracy Attained). concentrates on how to exploit inter-zone heterogeneity to enhance prediction. The use of MLRNN results in high processing costs and greater complexity when fitting data.

Probabilistic Taxi Demand Prediction with Bayesian Deep Learning: Proposes a Bayesian deep learning approach for probabilistic taxi demand prediction. (Achieved Accuracy of 83%). Estimates the uncertainty of predictions and provides probabilistic forecasts. Greater technical complexity, defining a prior distribution can be hard using Bayesian statistics.

Prediction of Taxi Demand Using Ensemble Model: Utilizes a point of interest (POI) to match taxi demand with a location so that it can be studied using a different function.

This method is based on RNNs and XGBOOST. Achieved Accuracy of 72%). It increases the accuracy, improved resource allocation, effective data analysis. Limited coverage, incomplete data, limited generalizability.

BRIGHT,Drift-Aware Demand Predictions for Taxi Networks: Accurate forecasting of short-term taxi demand amounts using a novel combination of time series analysis techniques that can manage various sorts of concept drift.(Achieved Accuracy of 78%). Could offer a range of benefits, such as increased efficiency and revenue, and improved customer service. The cost of implementing the BRIGHT platform may be high, and ongoing maintenance and updates may also be required.

Taxi demand prediction using hybrid deep neural networks: Hybrid deep neural network approach to predict taxi demand based on a variety of factors. The authors compare their approach to other machine learning algorithms and find that their hybrid approach achieves the highest prediction accuracy.(Achieved Accuracy of 80%). Have shown to be effective at capturing complex patterns in data, and a hybrid approach that combines different types of networks can help improve prediction accuracy. The model is too complex and begins to memorize the training data rather than learning generalizable patterns.

Creating a Customised Transportation Model to Predict Online Taxi Demand: a personalized demand forecast model is suggested along with a broad demand prediction for online cab hailing. It is suggested that a model with two attentional blocks be used to account for both temporal and spatial viewpoints.(Achieved Accuracy of 75.7%). By using user-specific data, personalized transportation models can make more accurate predictions for demand and supply of rides. Personalized transportation models rely on user-specific data, which may not be available for all users. This can limit the effectiveness of the model and make it less accurate.

Convolutional Spatiotemporal Multi-Graph Network for Taxi Demand: They tested Deep TDP on two real-world traffic datasets, and the results showed that it was effective when compared to self- and other baseline variations.(Achieved Accuracy of 80.5%). STMGCNs can handle input data in various formats, such as graph-based data and time series data, making them versatile for different types of applications. The model does not have the ability to extract multi-scale correlations of non-adjacent frames.

A method for predicting taxi demand using an ensemble: The authors discuss their ensemble technique, which integrates many machine learning models, such as random forest regression, linear regression, and support vector regression.They made use of a variety of metrics, including root mean squared error, mean squared error, and mean absolute error. (Achieved Accuracy of 88%). Has several potential benefits, including improved prediction accuracy and robustness. It may also be more complex than other approaches.

3. Existing system

There are several existing systems for taxi demand prediction using machine learning. Here are a few examples:

Uber Movement: Uber Movement is a platform that uses machine learning to predict the demand for Uber rides in various cities. It provides historical and real-time data on traffic patterns, events, and weather conditions to help drivers optimize their routes and improve passenger wait times.

NYC Taxi Demand Prediction: The New York City Taxi and Limousine Commission (TLC) has developed a system for predicting taxi demand in New York City using machine learning algorithms. The system uses historical data on taxi trips, weather, and events to forecast the number of taxi requests in various parts of the city.

Didi Chuxing: Didi Chuxing is a Chinese ride-hailing company that has developed a machine learning-based system for predicting demand for its services. The system uses real-time data on traffic conditions, weather, and events to optimize ride allocation and reduce wait times for passengers.

GrabTaxi: GrabTaxi is a Southeast Asian ride-hailing company that uses machine learning to predict demand for its services. The system uses historical data on ride requests, traffic, and weather conditions to forecast demand and allocate drivers accordingly

4. Proposed system

The purpose of this project is to build a model to analyze data-patterns and predict the demand of taxis based on number of requests in a given time period to help the taxi companies to pre-allocate the resources optimally.

The objectives is to analyze taxi demand patterns, segment requests based on pickup and drop-off points and duration of ride, extract features for machine learning models, and build machine learning models using Linear regression, Random Forest algorithm and XGBoost regressor to predict taxi demand in the future

5. System Architecture

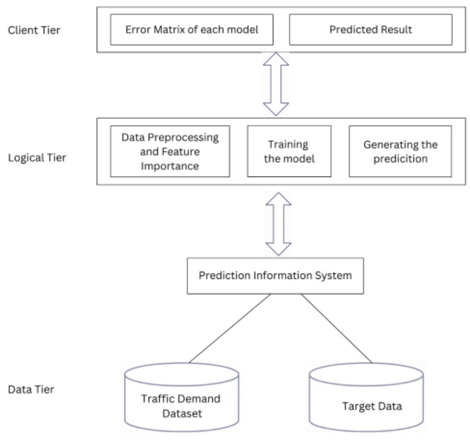

Client Tier
Error Matrix of each model
Predicted Result
Logical Tier
Data Preprocessing and Feature Importance
Training the model
Generating the predicition
Prediction Information System
Data Tier
Traffic Demand Dataset
Target Data

Fig 1: System Architecture

Client tier: This tier is responsible for presenting the user interface of the system. It includes web pages, mobile apps, or other interfaces that allow users to interact with the system. The client tier communicates with the logical tier to receive and display data.

Logical tier: This tier contains the business logic and application code of the system. It receives requests from the client tier, processes them, and sends responses back. The logical tier includes the data preprocessing, training models, and prediction algorithms that predict taxi demand. This tier also communicates with the data tier to access and retrieve data.

Data tier: This tier is used for storing and retrieving data from database. It includes the storage of data or data warehouse that stores historical taxi demand data, weather data, traffic data, and other relevant data sources used for prediction.

6. Functionalities

i) Functionality of User Input:

The user input is taken through the keyboard, then it analyses parameters and predicts the demand.

ii) Functionality of Pipeline Module:

Various machine learning algorithms are considered and stored in the pipeline. All the machine learning algorithms that are available in the pipeline are used to train the model with the dataset.

iii) Functionality of Model Selection Module:

RMSE and R2 score values for the above models are calculated. Then the model with the best figures of RMSE and R2 score is considered as the final model for the prediction.

iv) Functionality of Prediction:

Various data points are taken and the above selected model is being used to predict the taxi demand .

7. UML Diagrams

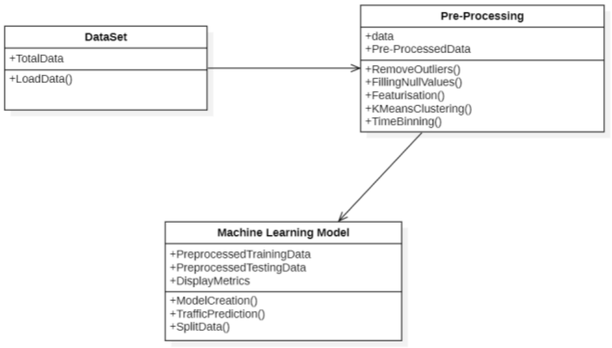


Fig 3: Class Diagram

A class diagram is another name for a static diagram. It shows the application's static view. A class diagram can be used to generate an executable code directly from the diagram for a software programmer as well as to visualize, describe, and document various components of a system. A class diagram describes the characteristics, actions, and limitations of a class.

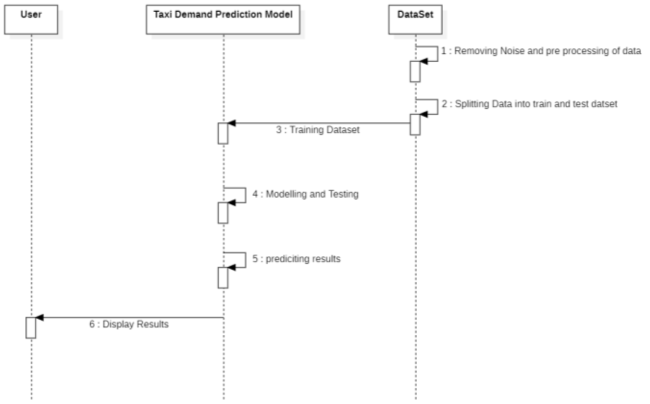

User
Taxi Demand Prediction Model
DataSet
1 : Removing Noise and pre processing of data
2 : Splitting Data into train and test datset
3 : Training Dataset
4 : Modelling and Testing
5 : prediciting results
6 : Display Results

Fig 4: Sequence diagram

The lifelines are:

- User
- Prediction (ML) Model
- Data Set

A sequence diagram is a type of interaction diagram because it shows how a group of actors communicate with one another. Software engineers and business professionals use these diagrams to understand the specifications for a new system

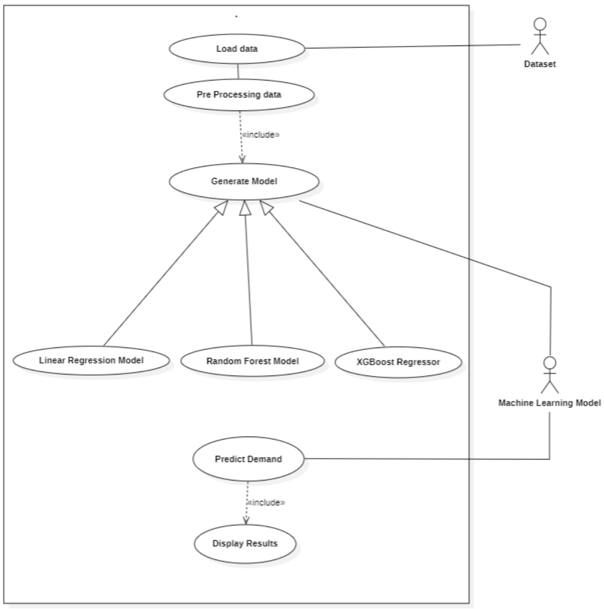


Fig5: Use Case Diagram

A written example of how users will carry out tasks on your website is known as a use case. It establishes how a system reacts to a request from the perspective of a user. A series of fundamental actions that begin with the user's aim and end when that goal is accomplished define each use case.

8. RESULTS

Models are trained using the dataset and accuracy of each model is analyzed to find out the best model for prediction that can best tell the demand.

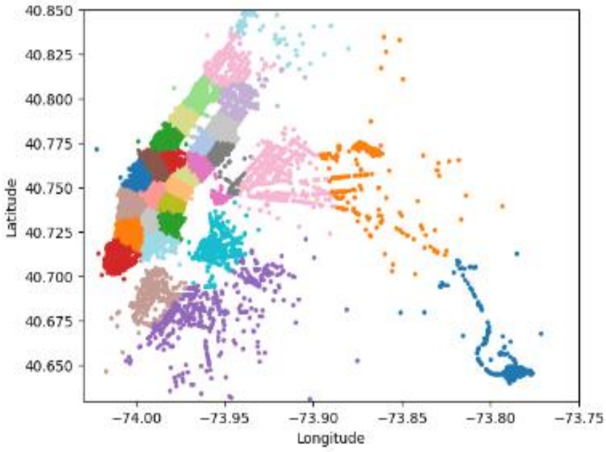
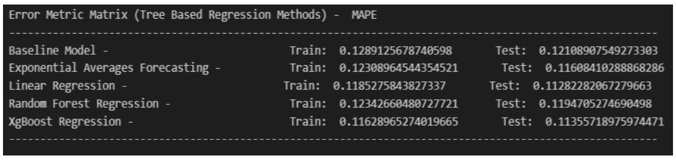
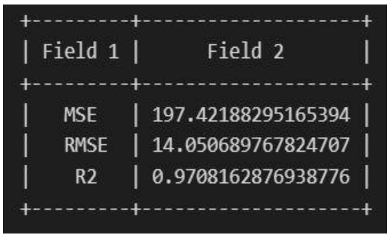


Fig – 6(a): Plot of Clusters

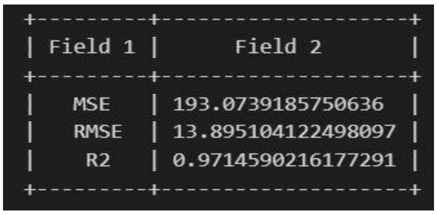


```
Error Metric Matrix (Tree Based Regression Methods) -  MAPE
---------------------------------------------------------------------------------------------------
Baseline Model -                              Train:  0.1289125678740598       Test:  0.12108907549273303
Exponential Averages Forecasting -            Train:  0.12308964544354521       Test:  0.11608410288868286
Linear Regression -                          Train:  0.1185275843827337       Test:  0.1128228206727966З
Random Forest Regression -                    Train:  0.12342660480727721      Test:  0.1194705274690498
XgBoost Regression -                          Train:  0.11628965274019665       Test:  0.11355718975974471
---------------------------------------------------------------------------------------------------
```



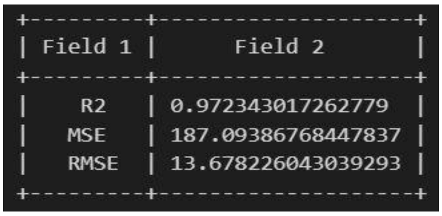
```
+---------+-------------------+
| Field 1 |      Field 2      |
+---------+-------------------+
|   MSE   | 197.42188295165394 |
|   RMSE  | 14.050689767824707 |
|    R2   | 0.9708162876938776 |
+---------+-------------------+
```

Fig – 6(b): Linear Regression Model Metrics



```
+---------+--------------------+
| Field 1 |      Field 2       |
+---------+--------------------+
|   MSE   | 193.0739185750636  |
|   RMSE  | 13.895104122498097 |
|    R2   | 0.9714590216177291 |
+---------+--------------------+
```

Fig – 6(c): RandomForest Regressor Metrics



```
+---------+--------------------+
| Field 1 |      Field 2       |
+---------+--------------------+
|    R2   | 0.972343017262779  |
|   MSE   | 187.09386768447837 |
|   RMSE  | 13.678226043039293 |
+---------+--------------------+
```

Fig – 6(d)XGBoost Regressor Metrics

9. CONCLUSION

In conclusion, taxi demand prediction using machine learning is a useful application that can help taxi companies optimize their operations and improve customer satisfaction. Use of machine learning provided many advantages in predicting Taxi Demand. The model saved time by preventing all the complex calculations and giving the demand of taxi in particular area. We used several essential attributes and regression techniques particularly linear Regression, Random Forest Regressor, XGBoost Regressor and K-Means for clustering of data points. We got even more accuracy than other models.

REFERENCES

[1] Mohamed Hanafy, Assiut University" Predict Health Insurance Cost by using Machine Learning and DNN Regression", Research gate, 348559741,2021.

[2] Shyamala Devi , Swathi Pillai , Vel Tech "Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning ",Research gate, 353231212,2021.

[3] Ch Anwar Ul Hassan, Jawaid Iqbal"A Computational Intelligence Approach for Predicting Medical Insurance Cost Hindawi journal,1162553,2021.

[4] Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, Manish Kumar" Health Insurance Cost Prediction using Machine Learning" IEEExplore,984201,2022.

[5] Chaparala Jyothsna, K. Srinivas, Bandi Bhargavi" Health Insurance Premium Prediction using XGboost Regressor" IEEExplore, 9793258,2022.

[6] Preet Jayendrakumar Modi, Vraj Jatin Naik "Insurance Management with Premium Prediction " IJRASET,2022

[7] Ghosh Madhumita "Health Insurance Premium Predictionusing Blockchain Technology and Random Forest Regression Algorithm" IJOEST article,346,2022.

[8] Keshav Kaushik , Akashdeep Bhardwaj "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums" NCBI,35805557,2022.

[9] Zhu, M., Li, Y., Li, L., & Li, Z. (2019). Time series prediction of taxi demand based on deep learning. IEEE Access, 7, 179647-179655.

[10] Yang, L., Zhang, B., Chen, Y., & Chen, L. (2018). A Deep Learning Method for Taxi Demand Prediction. IEEE Transactions on Intelligent Transportation Systems, 19(3), 782-791.

[11] Ma, T., Yang, Z., Wang, C., Hu, Y., Zhang, Y., & Liu, Y. (2019). Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction. IEEE Transactions on Intelligent Transportation Systems, 20(5), 1745-1756.

[12] Yu, X., & Jiang, Y. (2019). A hybrid deep learning model for taxi demand prediction. Transportation Research Part C: Emerging Technologies, 101, 206-218.

[13] Zhang, S., Chen, J., & Chen, C. (2021). Hierarchical Spatiotemporal Network for Taxi Demand Prediction. IEEE Transactions on Intelligent Transportation Systems, 22(5), 2918-2928.

[14] Zheng, X., Yang, X., & Sun, X. (2020). Deep Spatio-Temporal Residual Networks for Citywide Taxi Demand Prediction. IEEE Transactions on Intelligent Transportation Systems, 21(3), 1032-1042.

[15] Li, Y., Li, Z., Li, S., Li, W., & Li, G. (2020). A novel urban taxi demand prediction model using spatio-temporal deep learning method. Journal of Cleaner Production, 259, 120877.

[16] Zhang, K., Li, K., Lin, P., & Liu, K. (2021). A new attention-based neural network model for taxi demand prediction. Transportation Research Part C: Emerging Technologies, 126, 102988.