ISSN No:-2456-2165

Uber and Taxi Demand Prediction in Cities

Dr. G. Amudha[1] (Professor)
[1]Computer Science and Business Systems, RMD Engineering College

Balaji S[2] (Student)
[2]Computer Science and Business Systems, RMD Engineering College

Gowtham H[3] (Student)
[3]Computer Science and Business Systems, RMD Engineering College

Sudharshan Kumar M[4] (Student)
[4]Computer Science and Business Systems, RMD Engineering College

Abstract:- Traditional taxi systems in urban areas often face inefficiencies caused by uncoordinated actions as customer demand fluctuates. To forecast the upcoming number of taxis, we consider the taxis and uber demand in every region as a time-series data and simplify this prediction problem to a time series prediction. The varying temporal regularity of time series is addressed here. Furthermore, this lack of coordination leads to decreased passenger satisfaction due to long waiting times. The uber and taxi demand is predicted to avoid these inefficiencies. The Data is collected by using networked sensors and info like passenger count, demand rates in locations upon a date is stored as critical data in this system. This data presents opportunities to develop an intelligent transportation system that can efficiently control and coordinate taxis on a large scale. Taxi drivers can navigate to areas with high- demand, while ride-sharing companies like Uber can proactively reallocate the available resources to meet the rising demand.

I. INTRODUCTION

In this study, we aim to address two important questions regarding taxi demand prediction.

Firstly, we want to determine the prediction accuracy which is achieved by algorithms such as LSTM or ARIMA when they capture all the temporal patterns of the taxi demand time series. Secondly, we aim to identify the algorithm that can improve the prediction accuracy, given the maximum predictability.

We use the parameter (Пmax) which tells the max predictability of demand in a specific region. The maximum predictability is defined as the entropy of the taxi demand time series, considering both the randomness and the temporal correlation.

In our research, we define Пmax as the highest potential accuracy that any predictive algorithm can achieve in forecasting taxi demand. By analyzing Пmax, we can select the best predictor for taxi demand in the region under study.

II. LITERATURE REVIEW

Jun Xu, R Rahmatizade, and Ladislau B conducted a study on two major scenarios in the taxi industry, where case 1 represents a greater number of available vehicles and heavy competition among them, and case 2 represents more customer waiting duration and less taxi dependency. While they found many solutions for the second scenario, they focused mainly on the first scenario. To address this issue, they utilized time series forecasting techniques on a dataset from a large taxi network in Porto city, Portugal, which consisted of 63 taxi stands and 441 taxis. Their report indicates that they were able to achieve a 76% accuracy rate.

Sasu Tarkoma have found that cities have various functional regions, including markets, hospitals, malls, workplaces, holiday visitable places and homes. The movement motive of people in these regions differ, with individuals typically going to office during workdays and visiting places to spoil themselves like malls, shops, hotels etc.. during weekend. This study examines the movement motive of human in urban areas and infers the functions of regions. The evaluation carried out relies on the datasets of taxi Global positioning systems, which consist of 21M, 11M, and 17M GPS points. The specific areas in the cities are grouped into four different kinds such as office areas, public places, residential homes, and other areas. To determine the functional zones in the three cities based on temporal human activity, the study presents a novel quad-tree region division method that depends on the taxi visits and applies association rules. The functional zones that have been found have the potential to enhance the efficiency of data transmission in network applications, including urban Delay Tolerant Networks (DTNs). The new DTNs algorithm based on functional regions improves the delivery ratio by up to 183%.

Xing Xie have demonstrated that the growth of a city leads to the emergence of various functional areas, such as teaching areas and commercial districts. To find functional zones in a city, this study presents a framework called DRoF, which stands for Discovering zones of various Functions. It makes use of both motive movement of humans and places of interest (POIs). The larger region is divided into separate sub regions based on large roads, such as national highways, large flyovers, and city roads, that has peak traffic. A topic-based inference model that treats an

area as a document, a function as a topic, and categories of POIs as metadata is used to infer the functions of each region. Human mobility patterns are also considered as words to determine where customers would like to visit and for what reason and from where they would like to pick up. The resulting distribution of operations for each location and human movement motive for each operation are used to identify the intensity of each operation in different locations. Our framework has numerous applications, including urban planning, business location selection, and public suggestions. Using real-world datasets from Beijing, such as two point-of-interest datasets and two 3-month GPS trajectory datasets produced by more than 12,000 taxis, we assessed our methodology. Our findings show that our strategy outperforms baseline approaches that rely just on POIs or human mobility.

III. PROPOSED METHODOLOGY

The predictability of predictors can be assessed by examining their accuracy and computational costs. To determine the limits of predictability, an analysis of a location's taxi demand history can be conducted. In recent research, deep neural networks have been employed to forecast taxi demand in urban areas, as there is a correlation between the temporal pattern of human mobility and taxi pick-ups.

Our study reveals that the maximum predictability of taxi demand can reach 83%, indicating the presence of robust temporal patterns in New York City. Furthermore, our findings demonstrate that Uber demand is more predictable due to its demand-driven nature and higher level of temporal regularity.

We conducted a comparative analysis of five frequently employed predictors and observed that attaining maximum predictability is a feasible objective for actual prediction accuracy. The LSTM predictor exhibits superior accuracy when dealing with building blocks that possess low predictability, whereas the Markov predictor performs better in scenarios with high predictability. It is crucial to consider the predictability factor while selecting a predictor, as a less intricate predictor such as Markov might surpass a more intricate deep learning predictor like LSTM. Our discoveries hold general applicability and can be extended to other sets of time-series data.

A. Admin Login

The admin module enables a system administrator to configure the back end of a system and carry out fundamental system setup. It provides the functionality to define preconfigured drop-down fields and schedule classes based on time.

➢ *An Administration Module can allow a user to do the Following Things:*

- Managing the functioning of the Business Process Server.
- Facilitate the installation and removal of applications.
- Oversee the administration of all users and groups.

A login module refers to a portal module that enables users to input their username and password for the purpose of logging in. These login modules are integrated within applications to offer a specific form of authentication.

B. Importing Dataset

Importing datasets allows you to reuse datasets within the same tenancy, or merge and replace content, without the need to create a dataset from scratch. Additionally, it is possible to import data from external sources and merge it with the data gathered through Analytics. We have uploaded 2000 datasets in which half of the dataset is used for uber classification and the other half is used for taxi classification. Here are some methods for importing datasets:

- CSV files are the easiest way to import datasets from a CSV file.
- Authorized users can add data a row at a time or import data fromspreadsheets or other data files.
- To load a JSON file, you can load the Rison package and use from JSON to parse theJSON file.

C. Data Classification

Classification is a fundamental method utilized in the realms of data science and machine learning to anticipate the appropriate categories for data. It involves determining the specific category or sub- population to which a new observation belongs, relying on a training dataset that consists of known observations or instances with assigned category memberships.

➢ *Our Model Employs Two Classifications:*

- Uber Classification
- Taxi Classification. The stages are listed below:
- Scraping and Cleaning
- Modelling
- Feature Importance
- Conclusions and Trustworthiness

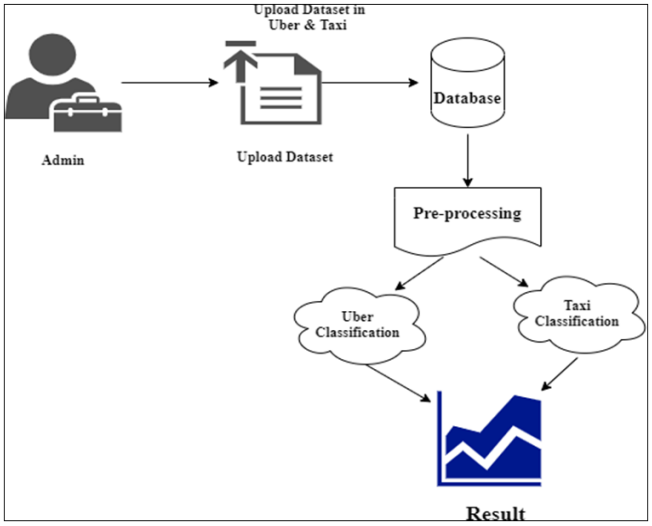


Fig 1 Architecture Diagram

ISSN No:-2456-2165

D. Data Visualization

Data visualization involves the depiction of data using various visual elements, including charts, plots, infographics, and even animations. These visual representations effectively convey intricate data relationships and insights derived from data analysis in a manner that is easily comprehensible. Typically employed for analytical purposes, data visualization presents the output in the form of visual aids like tables, graphs, and pie charts.

E. Evaluation of Predictors

The assessment of the predictors is conducted to determine the prediction algorithm that can achieve the maximum prediction upper bound (Πmax). comparison to other predictors across different predictability scenarios, we implemented our own LSTM predictor in a customized manner. Our LSTM model adheres to a standard design, wherein the inputs provided to LSTM cells encompass three gates: input - output - forget. Additionally, the model incorporates pointwise multiplication, addition, and the application of tanh, along with conventional neural network layers featuring sigmoid and hyperbolic tangent activation functions. To optimize the performance of our LSTM model, we conducted training on ten series extracted from the Yellow Taxi dataset, each with varying maximum predictability values. Ultimately, after careful evaluation, we identified a model configuration consisting of two layers and a lag size as two, as it exhibited the lowest errors in mean prediction.

➢ *ARIMA*

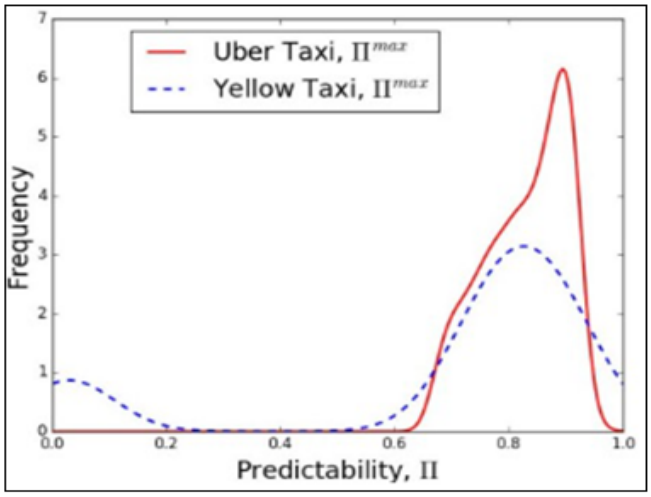


Fig 2 Distribution of Uber and Yellow Taxis

➢ *Predictors*

- *LSTM*

LSTM, a variant of Recurrent Neural Network (RNN), is specifically designed to effectively model sequential and time series data. Its distinguishing feature lies in its capability to capture long-term dependencies, making it a favored choice for predicting real-time taxi demand in urban areas. To assess LSTM's performance in ARIMA, a methodology developed by Box and Jenkins (1976), is a well-established approach for modelling and predicting univariate time-series data. It has undergone extensive research and has been demonstrated to possess robustness and adaptability. Nevertheless, its intricate nature and the prerequisite of expertise make ARIMA challenging to employ, necessitating substantial experience. Although it frequently yields satisfactory outcomes, the effectiveness of ARIMA is heavily reliant on the proficiency of the researchers involved.

ARIMA stands out among other algorithms due to its unique advantages. One of its key strengths lies in its ability to effectively represent diverse types of time-series data, such as autoregressive (AR), moving average (MA), and the combination of both (ARMA). Additionally, ARIMA harnesses the power of the latest samples from object sets to generate accurate predictions and seamlessly update its own models. With ARIMA, you can confidently navigate through complex time-series data and unlock valuable insights.

The ARIMA model assumes that the future value of a variable can be predicted by analyzing past observations and random errors. By doing so, we can better understand the process that generates the time series.

- *Markov Predictor*

A Markov Model is a powerful tool in statistical modelling that allows us to understand and predict the behaviour of systems that undergo random changes. It operates on the assumption that the future states of the system solely depend on its current state, disregarding any past events. This property, known as the Markov property or memory lessness, enables us to analyse and make accurate predictions about various phenomena. Markov Models have proven to be invaluable in the field of statistical modelling, offering insights into the dynamics of complex systems.

IV. RESULTS AND DISCUSSIONS

The taxi demand predictability in New York City is analysed. Our results indicates that taxi demand maximum predictability will highlighting a growing level of temporal regularity in mobility patterns. The entropy S and maximum predictability II of taxi demand for each building block using the yellow taxi data set is calculated,and the distribution of these values across all building blocks is obtained. Additionally, the difference in predictability between taxi and Uber is examined. Due to the sparsity of Uber taxis, the maximum predictability of Uber data sets at the neighbourhood's level is analysed. The hourly taxi demand in each neighbourhood and examine the maximum predictability of the hourly taxi demand sequence is grouped. The Uber taxi service's demand provisioning is significantly higher than that of regular taxis. This may be because yellow taxis typically use a random cruising strategy, while Uber taxis go directly to the passenger's location upon receiving a booking. As a result, the temporal data-correlation of taxi service demand in an area is better captured by taxis provisioned by Uber.

ISSN No:-2456-2165

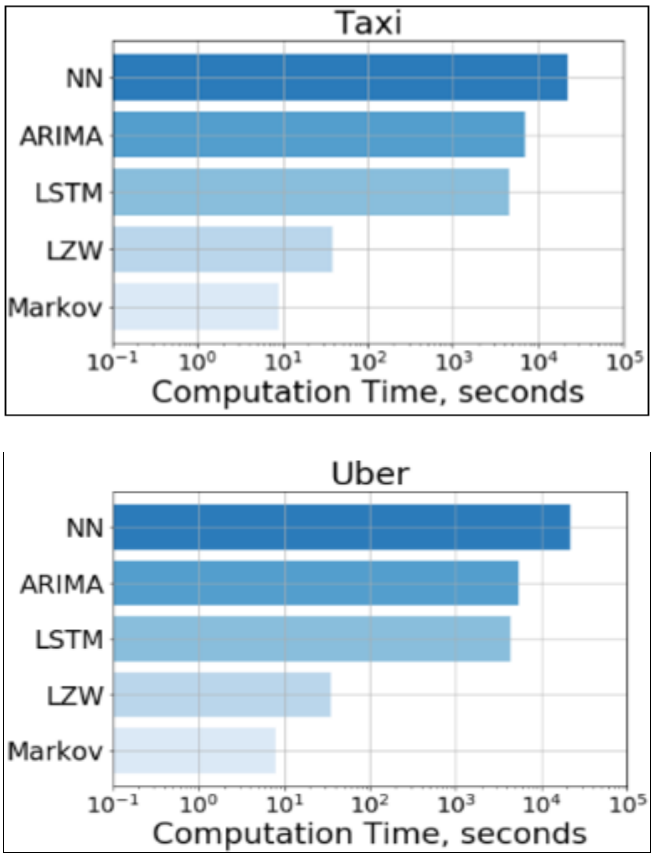
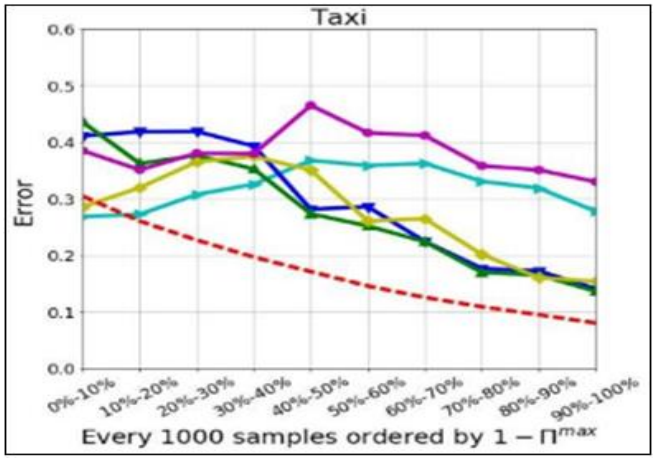


Fig 3 Computation Time taken by the Predictors in Seconds



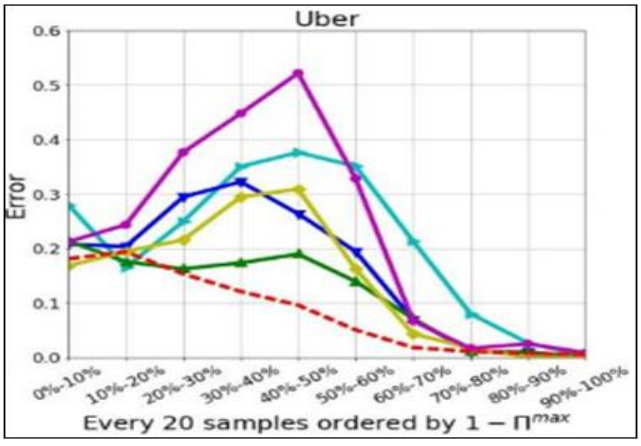


Fig 4 Error Probability Obtained with Respect to Prediction Percentage

V. SUMMARY AND FUTURISTIC DIRECTIONS

We have discovered a strong correlation between human mobility and taxi demand, with an average predictability of up to 83%. This indicates that there would be more likelihood of predicting the demand of taxi accurately based on temporal patterns. In our study, we compared different predictive algorithms to determine which one could achieve the highest predictability.

Interestingly, we found that the intensive computational deep learning algorithms, that has been widely used for prediction does not always outperform the Markov predictor algorithm in terms of the accuracy of the prediction obtained. The areas, which represents very less predictability factor ($\Pi max < 0.83$), the LSTM predictor was able to achieve high accuracy by capturing hidden long- term temporal patterns. On the other hand, in areas with high predictability, the Markov predictor showed great accuracy and minimal calculation time. This could be attributed to different cruising strategies observed with the data present in the taxi dataset.

When it comes to bike usage, the predictability is lower. However, both the Long Short-Term Memory and Neural Network predictors can achieve lower error rates by identifying complex non-linear patterns.

In higher predictability cases, the Markov predictor proved to be effective in achieving lower errors with minimal executional time. Based on our findings from testing these predictors on two different datasets representing uber demand and taxi demand, we can conclude that maximum predictability can guide the selection of predictors to achieve accurate predictions with minimal computational costs.

It's worth noting that the frequency of running prediction models mainly varies upon the various application. For instance, health care applications typically require predictions in seconds, while economic growth predictions are usually made over years.

REFERENCES

[1]. Y. Huang and J. W. Powell, "Detecting regions of disequilibrium in taxi services under uncertainty," in SIGSPATIAL'12, Redondo Beach, CA, USA, November 7-9, 2012, 2012, pp. 139–148.

[2]. B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset," in IEEE PerCom, Seattle, WA, USA, Workshop Proceedings, 2011, pp. 63–68.

[3]. J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases, ser. SSTD'11. Berlin, Heidelberg: SpringerVerlag, 2011, pp. 242–260.

[4]. R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing- theoretical perspective," The International Journal of Robotics Research, vol. 35, no. 1-3, pp. 186–203,2016.
[5]. F. Miao, S. Han, S. Lin, J. A. Stankovic, D. Zhang, S. Munir, H. Huang, T. He, and G. J. Pappas,"Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach," IEEE Trans. Automation Science and Engineering, vol. 13, no. 2, pp. 463–478, 2016.
[6]. F. Miao, S. Han, S. Lin, Q. Wang, J. A. Stankovic, A. Hendawi, D. Zhang, T. He, and G. J. Pappas,"Data-driven robust taxi dispatch under demand uncertainties," CoRR, vol. abs/1603.06263, 2016.
[7]. J. Xu, R. Rahmatizadeh, L. Blni, and D. Turgut, "Real-time pre-diction of taxi demand using recurrent neural networks," IEEETransactions on Intelligent Transportation Systems, vol. 19,no. 8, pp.2572–2581, Aug 2018.
[8]. C. Song, Z. Qu, N. Blumm, and A.-L. Barab´asi, "Limits of pre-dictability in human mobility,"Science, vol. 327, no. 5968, pp. 1018–1021, 2010.
[9]. C. Zhong, E. Manley, S. M. Arisona, M. Batty, and G. Schmitt,"Measuring variability of mobility patterns from multiday smart-card data,"J. Comput. Science, vol. 9, pp. 125–130,2015.
[10]. X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility," Scientific Reports, vol. 3, Oct. 2013.
[11]. C. T. Cheng, R. Jain, and E. van den Berg, "Mobile wireless systems: Location prediction algorithms," in Encyclopedia of Wirelessand Mobile Communications, 2008.
[12]. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time seriesanalysis: forecasting andcontrol. John Wiley & Sons, 2015.
[13]. S.Hochreiter and J. Schmidhuber, "Long short-term memory, " Neural computation, vol. 9, no.8, pp. 1735–1780, 1997.
[14]. T. M. Cover and J. A. Thomas, Elements of Information Theory.Wiley-Interscience, 2006.
[15]. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner,"Nonparametric entropy estimation for stationary processes andrandom fields, with applications to English text," InformationTheory, IEEE Transactions on, vol. 44, no. 3, pp. 1319–1327, 1998.
[16]. J. H. Wilkinson and J. H. Wilkinson, The algebraic eigenvalue problem.Clarendon Press Oxford, 1965, vol. 87.
[17]. "Matlab Documentation. Root of nonlinear function." https://www.mathworks.com/help/optim/ug/fzero. html,accessed: 2019-07-20.
[18]. R. P. Brent, "An algorithm with guaranteed convergence for find-ing a zero of a function," The Computer Journal, vol. 14, no. 4, pp.422–425, 1971.
[19]. L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location pre-dictors with extensive wi- fi mobility data," in Proceedings IEEEINFOCOM 2004, Hong Kong, China, March 7-11, 2004, 2004.
[20]. "Matlab Documentation. Root of nonlinear function." https://www.mathworks.com/help/optim/ug/fzero. html, accessed: 2019-07-20.
[21]. R. P. Brent, "An algorithm with guaranteed convergence for find- ing a zero of a function," The Computer Journal, vol. 14, no. 4, pp. 422-425,1971.
[22]. L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive wi- fi mobility data," in Proceedings IEEE INFOCOM 2004, Hong Kong, China, March 7-11, 2004, 2004. [27] E. Alpaydin, Introduction to Machine Learning, 2nd ed. The MIT Press, 2010.
[23]. H. Bozdogan, "Model selection and akaike's information criterion (aic): The general theory and its analytical extensions," Psychome- trika, vol. 52, no. 3, pp. 345-370, 1987.
[24]. "New York polygon data set," https://www1.nyc.gov/site/planning/data- maps/open-data.page.
[25]. A. Guttman, "R-trees: A dynamic index structure for spatial searching," in SIGMOD'84, 1984, pp. 47-57.
[26]. X. Xie, B. Mei, J. Chen, X. Du, and C. S. Jensen, "Elite: an elastic [infrastructure for big spatiotemporal trajectories," VLDB J., vol. 25, no. 4, pp. 473-493, 2016. [32] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit,"
[27]. S. Makridakis and M. Hibon, "The m3- competition: results, conclusions and implications," International Journal of Forecasting, vol. 16, no. 4, pp. 451-476, 00 2000.