OLA BIKE RIDE REQUEST FORECAST USING ML

[1] Mrs.S.Yoga., M.Sc(CS&IT).,M.Sc(Maths).,M.Phil(CS)., M.Phil(Maths)., Assistant Professor

[2] J. Sangeetha, M.Sc (Computer Science),

[1] [2] Department Of Computer Science, Sakthi College of Arts and Science for Women, Oddanchatram.

ABSTRACT

The market shares and significance of ride hailing or Transportation Network Companies (TNCs) like Uber, Lyft, and Ola are growing in numerous transportation markets. Big data technology and algorithms should be utilized to analyse the huge amounts of information accessible to enhance service reliability, estimate the efficiency of these systems, and assist them in meeting the demands of passengers. In this study, a novel model is developed using data from Ola, the leading ride-hailing service in Bangalore, to estimate the gap between rider demands as well as driver supply in a particular time period and specified geographic location. The data set used in this study was a ride request dataset. This dataset would have the following attributes: ride booking time, pickup location, and drop point latitude longitude. The number of data points related to ride requests are, the columns of the data are Id of the customer, timestamp booking, pickup latitude, pickup longitude, drop latitude and drop longitude. On the phone-based Ola application, a passenger "calls a ride," (makes a request) by inputting the point of origin and destination and tapping the Request Pickup button. A driver answers the request by taking the order. Even while the training set is little in comparison to the whole Ola ride-hailing industry, it is large enough for patterns to be found and generalized.

I INTRODUCTION

The ride-hailing (Ola) service sector has been expanding for a few years, and it is anticipated to continue expanding in near future. Ola drivers must decide where to wait for passengers since they may arrive rapidly. Additionally, passengers like an immediate bike service whenever required. People who have issues with booking Ola bikes, which sometimes cannot be fulfilled or the wait time for the arrival of the trip is particularly lengthy owing to the lack of a nearby Ola bike. If you successfully reserve an Ola bike in one go, consider yourself fortunate.

Ola is acquiring a greater market share and significance in a variety of transportation markets. Big data technologies and algorithms should be employed to handle the enormous amounts of information that are available to enhance service efficiency[2]. This will allow for more accurate estimates of efficiency as well as assistance in meeting the needs of riders[3]. This work develops a model to forecast supply and demand mismatches using information from the leading ride-hailing company in Bangalore. The percentage of Indians who travel by taxi, bus, or rail is among the highest in the world and few of the Indians 1.4 million residents own automobiles [4]. The leading ride-hailing business in Bangalore, Ola, handles more than 1 lakh rides daily and gathers more than 5GB of data.

It has become important for Ola (and other e-haling) company to forecast the demand for their Ola bikes so that they may better understand that demand and maximize the efficiency of their fleet management.

A novel model based on users' ride request dataset is proposed to address these problems; it would include characteristics such as ride booking time, pickup place, and drop point latitude-longitude. This model would predict demand for a certain period in various city areas, assisting the business in maximizing the density of Ola bikes to meet consumer demand.

II LITERATURE SURVAY

Machine learning is a technique that allows computer to learn from past data and anticipate fresh samples. Machine Learning models may be used in any sector. Medical records are likewise not exempt from machine learning. For numerous years, the medical industry has used models in various settings. Many of the studies used machine learning approaches to forecast medical costs B. Nithya [1] et.al In Predictive Analytics in Health Care, machine learning models were used. For predictive analysis, they used a variety of supervised and unsupervised models. They also claimed that machine learning tools and techniques are crucial in health-care sectors, and that they are exclusively employed in the detection and prognosis of various malignancies. Ahuja Tike[2] et.al applied hierarchical decision trees for the medical price prediction systems. Their experiments showed that the price prediction system achieves high accuracy. Moran et al. [3] utilized linear regression techniques to anticipate Intensive Care Unit (ICU) expenses and utilize understanding socioeconomics, DRG (Diagnostic Related Group), length of stay in the clinic, and a couple of others as highlights. Gregory [4] et.al applied various regression models for analyzing medical costs in the health care system. They mainly concentrated on reducing the bias in the cost estimates to achieve good results. Dimitris Bertsimas[5] et.al applied different data mining techniques which provided an accurate prediction of medical costs and represent a powerful tool for the prediction of healthcare costs.

2.1 Medical Expense Prediction System using Machine learning Techniques and Intelligent Fuzzy Approach (2020, H. Chen Jonathan, M. Asch Steven) Prediction isn't a new concept in medicine. Clinical predictions based on data are becoming commonplace in medicine., ranging Risk categorization of patients in the critical care unit ranges from risk scores to anticoagulant treatment (CHADS2) and cholesterol medicine usage (ASCVD) (APACHE). You may easily develop prediction models for hundreds of similar clinical questions using clinical data sources and contemporary machine learning. These approaches might be used for everything from sepsis early warning systems to superhuman diagnostic imaging. The real data source, on the other hand, has an issue. Unlike traditional techniques, which rely on data from cohorts that have been thoroughly prepared to prevent bias, new data sources are sometimes unstructured due to the fact that they were developed for various purposes (clinical care, billing, etc.). Patient self-selection, indication misunderstanding, and inconsistent outcome data can all contribute to unintentional biases and even racist programming in machine prediction. As a result of this understanding, discussing the potential of data analysis to aid medical decision-making isn't just wishful thinking.

2.2 Predicting Days in hospital using health Insurance Claims (2020, Yang Xie, C.W. Chang, Sandra Neubauer) Identifying and managing patients most at risk within the health care system is vital for governments, hospitals, and health insurers but they use different metrics for identifying the patients they perceive to be at most risk[1]. Hospitals focus on readmission rate [2], [3] and cumulative risk of death during hospitalization [4]. Accurately predicting these indicators could assist in allocating limited resources and thus improve the hospital's operational efficiency. Health insurers are mostly concerned with insurance risk, because they agree to reimburse health-related services in exchange for a fixed monthly premium. Poor risk measure could result in exceeding a financial budget. Therefore, one of the most obvious goals for health insurers is to Various predictive models have been developed to identify high-risk customers by predicting health-care expenses [5]. Traditional prediction models used demographic information and prior costs to predict future costs [1]. More sophisticated models that incorporate diagnoses [6], [7], drug claims [6], [7] and self-reported health status data [8], have been shown to improve prediction performance. Zhao et al. [6] reported a coefficient of determination (R2) of 0.168 when both drug and diagnosis were used to estimate costs for the coming year. Bertsimas et al. proposed two models: a decision tree model and a clustering model [7], to improve the performance of an earlier model based on classical regression models [6], [9]. Since hospitalization is usually the largest component of health expenditure [10], a separate identification of subpopulations at higher hospitalization risk could improve current underwriting processes and pricing methodology. Moreover, insurance companies also manage insurance risk by using specific interventions in different sub-populations (especially high-risk groups) to minimize the resources they require [11]. Programs that use case and disease management have been developed, which target different sub-groups of customers, such as aged care programs, chronic disease programs, and more recently telehealth programs, all of which have been shown to improve health care outcomes. identify high risk customers by predicting their health care expenditures.

III THEORETICAL BACKGROUND

3.1 PROBLEM IDENTIFICATION

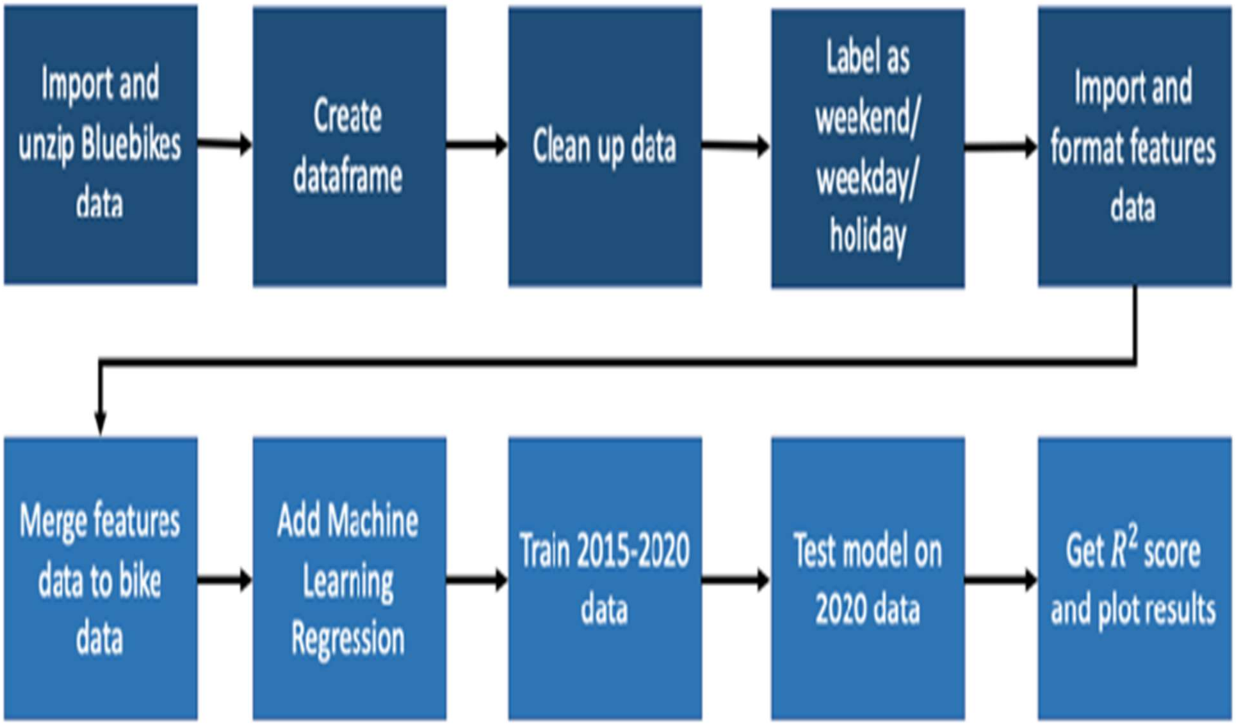
- In sample sizes ranging from small to large, statistical approaches (E.g., Lenear regression) suffer due to the zsero point spike and skewed distribution of health care expenses with a strong right-hand tail. To address this issue, advanced methods have been proposed, such as The data source, on the other hand, has a flaw. Unlike traditional techniques, which rely on data from cohorts that have been thoroughly prepared to eliminate bias, new data sources are sometimes unstructured since they were developed for various purposes (clinical care, billing, etc.). A range of issues, ranging from patient self-selection to indication uncertainty and inconsistent outcome data, can induce unintentional biases and even racist programming in computer prediction. As a result of this understanding, speculating on the potential of data analysis to aid medical decision-making isn't far-fetched.
- Having an online system that can book rides whenever necessary offers several advantages. Firstly, it provides convenience and saves time for passengers, as they can book a ride with just a few taps on their phone. No more waiting on hold or searching for a cab on the streets.

Secondly, it ensures reliability and availability, as the system can connect passengers with nearby ride quickly. Inaccurate fare estimates can lead to unexpected costs and inconvenience for passengers. By developing a machine learning model that accurately predicts cab fares, we can help passengers plan their budgets, avoid overpaying, and have a more seamless and transparent cab booking experience. This project aims to solve the problem of uncertainty and provide a reliable tool for estimating cab fares.

3.2 PROBLEM SOLVING

- The Medical information and expenditures billed by health insurance companies are included in the data. It has 1338 rows of information with the following columns: age, gender, BMI, illnesses, smokers, and insurance costs In these features, insurance charge is a dependent variable and the remaining features are called independent variables. In regression analysis, we need to predict the value of the dependent variable using independent variables. First, we collected the dataset and applied various data preprocessing methods. Data preprocessing is a technique in which we can remove missing values in the data. Because of these missing values, it is not possible to apply machine learning algorithms. After removal of missing values, we need to apply label encoding, one hot encoding data to the categorical features.
- the prediction accuracy of the ML model. As shown in Fig. 16, the prediction is very close to actual rides, but there is some variance between the predicted number of rides and actual number of rides per day. To determine how accurate the prediction is, the $R2$ value of the prediction was calculated, as discussed in Section 6. The $R2$ value was calculated to be $R2$= 0.85. This is a strong $R2$ score compared to the typical value of $R2$ score of 0.7. Thus, this score shows that the model is a fairly strong predictor of count of rides per day.

3.3 SYSTEM ARCHITECTURE


Import and unzip Bluebikes data
Create dataframe
Clean up data
Label as weekend/ weekday/ holiday
Import and format features data
Merge features data to bike data
Add Machine Learning Regression
Train 2015-2020 data
Test model on 2020 data
Get R^2 score and plot results

IV SYSTEM IMPLEMETATION

4.1. MODULE

- Data Preparation
- Data Features
- Multi-Step Time Series Forecasting

MODULE DESCRIPTION

- **Data Preparation:**
  In order to create a prediction model for the demand for rides in a certain area at a given time, the data must first be preprocessed to determine the actual estimated demand by consumers. I eliminated requests for rides that were very likely to be problematic in order to evaluate the genuine demand.

- **Data Features:**
  The resultant features are customer Id (unique Id given to each user), booking timestamp (booking timestamp of ride IST) and pick up cluster Id (this is computed by clustering over pickup latitude and longitude). To determine the number of demand/ride requests from a region: In order to collect the number of customer ids that booked trips from those areas during that timeframe, we split time into 30min intervals, creating a total of 24Hours * 2 = 48 (30min intervals).

- **Multi-Step Time Series Forecasting:**
  The challenge of anticipating a succession of values in a time series is known as multistep-ahead prediction. Applying a predictive model step-by-step and using the anticipated value of the current time step to calculate its value in the next time step is a common strategy known as multi-stage prediction.

V CONCLUSION & FUTURE WORK

5.1 CONCLUSION

In order to handle the issue of ride demand forecasting, a novel XGBoost regressor model is proposed in this work. The data preprocessing, geospatial engineering methods are utilized to convert latitude and longitude, to cluster Id using Mini-Batch Kmeans algorithm, and then multi-step forecasting is used to forecast the demand for ride requests coming from an area at a certain time. The proposed XGBoost Regressor model score is 0.916, and the RMSE values for train and test are 2.287 and 2.456.

REFERENCE

[1] Gunjan panda,Supriya p.panda.”Machine learning using exploratory analysis to predict cab fare”. International Journal for Research in Applied Science & Engineering Technology (IJRASET) Aug 2019.

[2] Pravin, A., Jacob, T. P., & Asha, P. (2018). Enhancement of plant monitoring using IoT. International Journal of Engineering and Technology (UAE), 7(3), 53-55.

[3] Kelareva, Elena. "Predicting the Future with Google Maps APIs." Web blog post. Geo Developers Blog, https://mapsapis.googleblog.com/ 2015/11/predicting-future-with-google-mapsapis.html Accessed 15 Dec. 2016.

[4] Jacob, T. P., Pravin, A., & Asha, P. (2018). Arduino object follower with augmented reality. Int. J. Eng. Technol, 7(3.27), 108-110.

[5] Wu, C. H., Ho, J. M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems, 5(4), 276-281.

[6] "A Deep Learning Framework for Cab Fare Prediction" by Xia et al. (2019) - This study proposes a deep learning framework for cab fare prediction, which uses historical fare data and traffic data to make predictions.

[7] "Cab Fare Prediction Using Time Series Analysis and Machine Learning" by Agrawal et al. (2018) - This study uses time series analysis and machine learning techniques such as ARIMA and Random Forest Regression to make predictions of cab fare prices.

[8] J. Ke et al., "Hexagon-Based Convolutional Neural Network for Supply-Demand Forecasting of RideSourcing Services," IEEE Trans. Intell. Transp. Syst., vol. 20, no. 11, pp. 4160–4173, 2019, doi: 10.1109/TITS.2018.2882861.

[9] Z. Ara and M. Hashemi, "Ride hailing service demand forecast by integrating convolutional and recurrent neural networks," Proc. Int. Conf. Softw. Eng. Knowl. Eng. SEKE, vol. 2021-July, no. Ml, pp. 441–446, 2021, doi: 10.18293/SEKE2021-009.

[10] I. Saadi, M. Wong, B. Farooq, J. Teller, and M. Cools, "An investigation into machine learning approaches for forecasting spatio-temporal demand in ridehailing service," 2017, [Online]. Available: http://arxiv.org/abs/1703.02433