Divvy Bike-Share Analysis for Targeted Customer Marketing

Tamojit Roy1 • A. Sheik Abdullah2

Abstract

Although the age calls for motor-vehicles as the major shareholders of the commute industries, cyclists still continue to have impressive shares in some of the developing heavyweights and developed nations around the world. Some of the countries like Sweden, Denmark, Germany, UK, Japan and even China have a wonderful landscape for cycling. Although some of them are casual riders, quite a many are annual members for the major cycle manufacturing companies all over the world. Data analytics plays an important role in analyzing and boosting the sales of any company. It has an upper hand when it comes to implementing market plans to target the set of customers who are the most vulnerable by suggesting them various specialized schemes and membership benefits. Hence, it is no doubt one of the most potent tools helpful in boosting the sales of a product. Hence, it can play a pivotal role in increasing the annual membership of cycles of any company for good profits. This research work intensely focusses on analyzing all the major aspects and most of the if not all of the attributes of bike-share sales of a prominent bike-share company in Chicago named Divvy. This research work mainly revolves around understanding how subscribers and customers of Divvy bike-share service use bikes differently. The comparison along with other tasks have been used to design marketing strategies aimed at converting customers of the company to the subscribers of its services.

Keywords AR · ARIMA · VAR · RNN · LSTM · Linear Regression · Random Forest · FB Prophet · Darts · NaiveSeasonal · NaiveDrift · ExponentialSmoothing · AutoARIMA · RegressionEnsembleModel · RegressionModel · Theta · FFT · XGBoost

# 1 Introduction

In today's world, a significant lot of people beguiled by their work ethics and cultures need to travel a lot. This poses a serious challenge especially for those who rely on commute services to garner opportunities and those who seek major success in their careers. The competition becomes quite intense in regions with dense but competitive population. Congested spaces pose a murky hindrance to those who are struggling both socially as well as economically, as they have limited options when it comes to choosing commute services. In many cases, they are not aware of various bike-share services which can alleviate their suffering. The bike-share companies, on the other hand are not familiar with the needs of such sections and hence fail to fabricate plans to come up with customer-centric policies and subscriptions for boosting sales and for providing help as well. The first and foremost imperative move is to develop a research work which mainly revolves around understanding how annual members and casual riders use cyclistic bikes differently. The comparison along with other tasks can later be used to design marketing strategies aimed at converting casual riders to annual members.

This research work is focussed exactly to address the same problem by taking the instance of Divvy's bike share system. Divvy is a well-known and ubiquitous name and is well reputed to provide bike sharing facilities across Chicago and Evanston. Understanding how subscribers of Divvy bike-sharing trips and the bike-sharing service availing only-customers use the bikes differently for commute purposes,

recreational purposes, schooling, marketing, etc. and analyzing the multifarious shadow factors which influence

both the user categories in opting for the type of service which they show their inclination towards can prove to be of immense help. This comparison along with other tasks has been used to design marketing strategies aimed towards converting the customers of Divvy bike-sharing trips to the long-term subscribers of the bike-sharing service.

2 The literature review

This paper deals with analyzing and examining the various factors, which portray the bicycle-industry as a potent industry which can lead to the green recovery and sustainable development of economy and environment in Bangladesh. (Xames et al. 2022) performed a SWOT analysis after collecting and analyzing information with regards to the bicycle industry in the south-asian developing economy from many research publications, and by interviewing industry experts, goverment officials and university professors and finally a joint discussion by all the experts. Based on the findings in which both internal factors (strengths and weaknesses) as well as external factors (opportunities and threats) were taken into consideration, they came up with an internal factor evaluation matrix (IFEM) to find out that the bicycling industry in Bangladesh has enormous potential, given that some reforms be conducted upon the same with a view towards reducing carbon-emissions and decarbonizing the commute sector by facilitating green investment in bicycling as a non-motorized transport (NMT) option to reduce GHG emissions in the post-pandemic settings. Suggested potential strategies to facilitate the adoption of bicycles as a resilient and sustainable transport option such as by developing local manufacturing capability, dedicated infrastructure, reducing import duties, attracting FDIs, eliminating gender differences, etc., were also mentioned.

The analysis of the pattern of bike-sharing in a region plays the most significant role in predicting the bike usage in any particular region. (Jiang 2022) in this paper, have provided a comprehensive review of bike-sharing usage prediction with many invaluable approaches to predict the bike-sharing usage pattern with deep learning. Following a set of procedures of the following modules: data aggregation to build the prediction input-features and targets, defining 3 data formats (time-series format, grid format and graph format), addressing 3 types of prediction problems (time series-input prediction, graph-input prediction and grid-input prediction), quantifying the prediction error using different evaluation metrics, and finally some prediction challenges (complex spatial dependencies and complex temporal dependencies) and prediction models like FFNN, LSTM, RNN, GNN, GRU, MLP, SVR, etc., were illustrated with a different section for each type of model was mentioned for both docked and dock-less bike-sharing systems. Finally, application scenarios within the bike-sharing systems and beyond were brought up followed by the challenges and the development directions in the research paper. More open datasets, various applications based on bike usage prediction and potential research directions were summarized to encourage future research.

A vision of developing a sustainable bike-sharing system as viewed a PSS (product-service system) was expressed by (Moro et al. 2022) in this research paper. Although, it was initially developed considering a single focal-company, restricted to a particular region (southern region of developing country, Brazil) only, the authors emphasized the significance of their research-work claimimg that it was the first one to be done keeping in mind the scenario of a developing country. The design of the system-model organized in 4 stages: (i) value proposition, (ii) value configuration, (iii) value delivery and (iv) value capture was introduced. Although, the strategy doesn't hold the providers of the service as the manufacturers themselves, the business-model analyzed was a use-oriented in the context of shared mobility. By conducting face-to-face interviews with those involved in developing the business model, a research protocol was developed. The authors expressed that the PSS business model analyzed by them could represent significant contributions to improve micro-mobility.

This paper entirely deals with understanding, analyzing and illustrating the multiple modes and forms of relationships between build environment (i.e., land-use, transportation system and urban design) and bike-share usage. Quite a many variances between the build environment and bike-usage were stated and described with some outliers in notable cases. Variance in relationship in the build environment across different mobility patterns, docked and dockless bike-share patterns, w.r.t. trip purpose, between arrival and departure patterns, based upon the day of week, etc. and the bike-share usage were elaborated. The paper concluded with a brief summary of the major findings of (Guo et al. 2022) and them encouraging the recommendations for the future research works.

The study by (Hu et al. 2021) attempted to examine the associations of BnR (bike and ride) activities with metro area w.r.t. DBS (dockless bike sharing) systems, in the city of Shanghai, China. The study signalled that BnR behaviors were affected by features like station features, land use, socio-demographics, roadway designs, transportation facilities, etc. Mainly four metrics were employed in the entire study to understand BnR behaviors from the perspective of different participators viz. local govt., DBS users, etc. The metrics were BnR trip count, shared-bike utilization rate, metro catchment area and BnR rate, for the assessment of BnR performance. The generalized additive model (GAM) was utilized to build statistical inference. Several statistical issues such as over-dispersion, skewness and spatial autocorrelation were addressed while modelling DBS usage. The spatial distribution of the 4 metrics suggested that shared bikes were oversupplied in the city center while undersupplied in the suburb. Based on other things, various other conclusions were drawn for comprehensive analysis.

In this research paper, (Song et al. 2022) sought to investigate the correlation of the various factors of the perceived value upon the users willingness to pay for bike-sharing services in the first-tier and second-tier cities of China. A structural analysis was also conducted to validate the findings and visualize the significance of the different factors as variables. The paper analyzed the direct and indirect factors that affect bike-sharing users' willingness to pay. Based upon the findings, the authors concluded that, perceived usefullness and perceived ease-of-use have positive impact on perceived value; and perceived trust, perceived value, individual paying consciousness and environmental protection have positive impact on perceived value; the users' word-of-mouth and perceived entertainment have no significance; and finally perceived cost and perceived risk have negative impact on perceived value.

This paper investigates the various factors which make the bike-sharing services to be retained by the users, and not just be opted by them in the first place. For data collection purposes, questionnaires were collected through both online and offline survey. A total of 650 questionnaires were collected, including 500 field surveys and 150 online questionnaires, resulting in 622 valid questionnaires. (Zhang et al. 2021) introduced the related concepts of participation in purchasing decisions, customer engagement, and customer-perceived value, and uses a structural equation model to identify the interaction and influence mechanisms between the three variables and usage intent. It not only extends the scope of consumer behavior theory, but also provides management and marketing strategies for bike-sharing companies.

Through the paper, (Puneeth et al. 2021) tried to analyse and discuss the current bicycle market scenario in India and where the developing country presently makes its stand in the world when it comes to manufacturing, exporting and ranking in terms of bikes' usage and procurement of raw materials. The objectives which they proposed in the research papers include, gaining knowledge about India's cycle industry, learning abouth the industry's development, comparative analysis between sales and production, study of the future growth and analysis of the industry, the bicycle industry's contribution in international economic development, research on how latest gadgets can be added in bikes, understanding CORONA's impact on the Indian cycle industry, and finally, a SWOT analysis to aggregate the facts and figures for recommendation purposes favoring the success of the Indian cycle manufacturing industry in the future. After collecting the data from various sources, viz., journal, published papers, archived newspaper articles, official bicycle industry websites, and other ventures, the authors discussed various aspects: the growth of bicycle industry in India; future analysis of bicycle industry in India; major competitors of India in the industry; the contribution of bicycle industry in international economy development; the research and development centre for bicycle in India; the analysis highlighting the strengths, weaknesses, opportunities and threats of the Indian bicycle industry; CSR activities; COVID-19 impact on the Indian bicycle industry. Based upon the aforementioned, they provided their recommendations and concluded with areas for future development of the industry.

The paper provides a systematic literature review (SLR) of various research papers specializing in exploring and analyzing the multifarious machine learning approaches applies to bike-sharing systems (BSS). Based on a preferred reporting items for systematic reviews and meta-analyses (PRISMA) methodology, that consists of a checklist and a flow diagram, the systematic literature survey was performed. A four-phased flow diagram consisting of the following phases: identification, screening, eligibility, was aimed to describe and understand the items of the different sections. (Albuquerque et al. 2021) framed a process workflow to understand all the stages of the study, viz., keyword identification and search, repositories, bibliometric analysis, etc. The open source tool VOSviewer, was used as the bibliometric research tool for network analysis. The tool helped to to identify the main keywords, authors, co-authors and their respective relations, within the SLR data set, for quantitative analysis. Many graphs were created for each of the sections of interest, by the authors in order to bring out reasonable insights from the study. Based on the findings, the authors asserted that the two main problems addressed by the machine learning techniques in the context of bike-share systems are clustering (classification) and prediction. Three clustering algorithms, viz., hierarchical-clustering, community-detection-clustering and k-means-clustering are more commonly used. The authors additionally also discussed the various research and study limitations in their whole study. The study tried to raise some recommendations for the future work within the overarching theme of machine learning techniques applied to BSSs.

The research paper examines the various factors influencing the buying of two wheeler vehicles by the females in Palghar, distant suburb of Mumbai city. For collecting the data, (Jain et al. 2020) used structured questionnaires (primary data) and websites, journals, research articles and news reports (seconday data). For their research purposes, the authors collected data samples from a total of 150 respondents. Random sampling was used for the collection of primary data. The authors specified a set of hypothesis and performed a data analysis report, which put light upon the various aspects of the female buying behavior and more, viz., popularity of brands among female riders, use of internet in buying decision, awareness of celebrity endorsement, mode of purchasing the vehicle, etc. They concluded that necessity was the most influential factor affecting the buying decision of female two wheeler users. Based on the calculations, as the Chi square value was significantly high, it made them to accept the alternate hypothesis which they set. Other conclusion drawn were, celebrity endorsement had no significant imapact on the buying behavior and most of the women use the two-wheelers jointly with other family members.

Information from survey of local respondents and other sources were used for studying the factors which influences the purchase of two wheeler modes of transport in the city of Chennai. Both primary and seconday data were involved in the study by (Raj et al. 2020). The primary data was collected from the owners of two-wheelers using questionnaires. The secondary data was sourced from published reports, records, books, Journals, bulletins, magazines, internet and newspapers. About 100 respondents were selected for collecting the primary data with suitable sampling techniques. Some of the objectives put forward by the authors included: to analyze the age of the respondents, to identify the type of media created awareness among the consumers, to analyze the factors influencing the purchase of two-wheeler, and to know the expectation of consumers in the purchase of two-wheeler. This was followed by the analysis and implementation. The authors concluded the research paper with suggestions and conclusion.

(Yoon et al. 2020) scrutinized revenue management strategies for unlimited usage bike-share scenarios in their research paper. Citi Bike public data has been used by the authors for the overall analysis. Summarization of the basic data for understanding the behavioral patterns of the casual users and the number of trips that casual users take was estimated for relating between the two, because such data was not publicly available. The parameters of these distributions were derived from sample means and standard deviations using linear regression of daily short-term ticket sales and occasional passenger numbers. A path choice model was built using variables resampled from the fitted distribution by the bootstrap method. The multinomial logit model was used because it could represent not only individual vote probabilities but also aggregated market shares. As a result, the price combination of the two plans was optimized to maximize revenue based on estimated model output, and the impact on consumer surplus was quantified.

This paper deals with providing a set of illustrative examples and performing a series of sensitivity analysis tests to propose an original model to find out the number and layout of the bike stations as well as the number of bicycles and free-racks for modeling bike-sharing systems based upon spatial equity concept aimimg towards reducing implementation and operational cost of the existing such systems. In the research paper proposed by (Caggiani et al. 2020), a linear programming model was developed and tested to determine the number and layout of stations and the number of bikes and racks available at each station to set up and operate the BSS for each level of defined service, terms and conditions cost was minimized. The proposed linear mathematics problem minimized the setup and execution costs of a new BSS using a set of constraints that reflect the idea of ​​a balanced and fair level of service provided to all users of the system.

Cities all around the world are increasingly getting concerned about the detrimental effects of increasing number of vehicles on their roads, viz., greater pollution, emission rates, congestion, etc., which has led many of then to take active steps in order to prevent such things. One of the most effective and efficient alternative and life-saver which has come out from the horizon in this regards is the usage of BSSs or bike-sharing systems. Many cities have also been seen to adopt BSSs to tackle such growing problems and more. This increased use of bicycles has prompted many cities to either expand existing systems or introduce new ones. Due to the unbalanced spatio-temporal demands of cycle tours, many cycle stations are either empty or full during the day. This could have a significant impact on the reliability and usefulness of BSS and encourage drivers to return to using their vehicles or choose alternative modes of transport, resulting in increased congestion and emissions and pollution may increase. This reduces the number of BSS users and reduces system revenue. Operators recognized the imbalance and began building more bike stations closer to each other with the aim of keeping them within a five-minute walk. However, this solution is both financially and practically difficult to implement. This extensive and enormous report and aggregation of research paper methodologies proposes a new generation of BSSs, in which some can be portable, i.e, it takes into account both the types of BSSs (docked and dockless bike-sharing systems). Specifically, the framework consists of two levels: We use fast, online, and incremental learning approaches to predict the number of bikes at stations and balance the system with wearable stations. The goal is a framework that solves the dynamic bike-sharing relocation problem to minimize unmet demand and increase or decrease user satisfaction. The dissertation by (Almannaa 2019) contributes to the field in 5 ways. First, a multi-objective supervised clustering algorithm was developed to identify similarities in bicycle use with respect to time events. Second, a dynamic, interpretable and fast approach was developed for predicting the number of bikes at stations within a BSS. Third, a univariate inventory model was created using a Markov chain process that provides an optimal selection of bike levels at stations. The fourth was to explore the benefits of portable bike stations using an agent-based simulation approach as a proof of concept. Fifth, mathematical and heuristic approaches are proposed.

(Arsenio et al. 2018) published this research paper which is actually based on the Be4Schools R&D project implemented in the Portugal based city of Águeda. The intent of this study conducted was to analyze the preferences of the students aged between 15-21 in the context of using e-bikes while going daily to school. It also aimed at assessing their longings and preferences towards ICT related attributes.The information for examination of the results was collected in three parts. It comprised of a mobility survey and a stated-choice (SC) experiment. The first part was aimed at collecting the responses from the students of their travel preferences about what mode they use, thoughts regarding inclusion of ICT, perceptions barrier for not cycling and previous cycling experiences. This part was named as the "Simplifying Cycling Mobility". The second part dealt with the household budget and business perpectives and in this both the students and their parents were questioned in order to get insights on what equipments to be installed, ICT preferences, household budget constraits, etc. This part was named as the "Assessing students and their parents' preferences". In the third part a SC experiment was performed to understand the trade-offs information between car travel and e-bike relevant attributes by gathering 2232 observations in that regard. A comprehensive econometric analysis was conducted to assess the nature and degree of heterogeneity in student preferences, also considering gender issues. The authors of this research paper had also taken the final shot at studying the main determinants of both the traditional as well as electric bikes impact on school-to-home/home-to-school trips, using a detailed study of the design using exploratory data analysis and multivariable logistic regression model. Statistical analysis was performed using the IBM SPSS Statistics 21 software and R.

(Ricci 2015) of this research-paper worked on a very special and distinct idea and approach of reviewing the existing schemes and research-suggestions and findings of previous researches on bike-sharing: how well the schemes of implementation and operation suggested by the past research-approaches have actually come out victorious or to what extent such were actually guiding the working of the present bike-sharing schemes present in different parts of the world actually implementing such schemes. They provided a comprehensive review of such evidences. By following a two-fold measure: understanding and examining evidence-gaps and limitations that need further investigation; and by drawing on the evidence review and justifying whether the positive-sides of the approaches mentioned actually aid in transferability, beneficial impacts and operation-processes or not, they sought to put light on both the impacts as well as the processes (rather than just processes) as a target to enhance the current body of knowledge on bike-sharing and also contributing towards the ongoing academic and policy discourse on the increasing popular cycling measure. A sectional segmented approach was taken by the authors, as they sought to provide a critical overview of the increasing number of information sources and the growing body of knowledge about bike-sharing; discussing the evidence on users, usage and impacts of bike sharing's significance and limitations after summarizing them; providing an evidence on managing the business of bike-sharing from a process evaluation perspective; they concluded the paper by discussing how the evidence presented here can help strengthen and transfer positive results to other contexts in terms of impact and implementation processes, and identify key areas for further investigation. [16]

The long paper which comprised of multiple case study explored what various strategies the microenterprise owners in the artisan economy need to market using social media. Through the means of semistructured interviews and open-ended questions, data were collected from 5 bicycle framebuilding companies from a south-western US state. The diffusion of innovations theory was used by (Au 2015). A thematic analysis identified seven themes from the data, viz., technical proficiency, building a social media presence, effective use of social media platforms, effective communication skills, building a brand identity, time management, and obtaining external support. The overall study findings aimed to help artisan microenterprises learn to use social media effectively, which would lead to boost in sales and profits, further leading to good positive environment for growth and development. The findings also expect their way out towards helping the local economy as it helps to prevent the money from leaving local economies, thus building strong communities.

The study methodology proposed by (Frade et al. 2014) in this research paper aims to study the demands of the bicycle sharing systems, at the level of defining zones, with higher potential demand in urban areas. The literature survey conducted by the authors of this research paper presented some broad ideas, which included types of bicycle-sharing systems right from the antique to the modern systems, viz., free bike system, coin-deposit systems, information technology-based systems and multimodal systems; demand studies for cycling and bike-sharing: latent demand score method, 'revealed' or 'stated' preference surveys; etc. The methodology suggested by the authors to study the demand distributions for bike-sharing involves the two parts: quantifying the demand based upon other case studies; and defining the effect on demand caused by trip characteristics. The second part encompassed analyzing the factors such as purpose, distance, slopes, etc. and their effects on the the bicycle sharing demands thus making a cohesiveness study among the various factors proposed. After studying all these, the authors put the proposed methodology and their knowledge to exercise on the case study of Coimbra (a town located in the centre of Portugal). The conclusion came with pointing out the advantages being that the methodology proposed provides a quick-assessment technique, it can be adapted to other towns and cities, and that it is useful in the full design of the system. The paper ended with the authors pointing out the areas in which there is a scope for further studies in this area.

In this research paper, (Buck et al. 2013) investigated the travel behavior characteristics of bicycle system users. A comparative analysis was done to understand the differences between annual members and short-term user profiles on Capital Bikeshare (CaBi). The data used for the overall research was gathered from Washington, DC area regional household travel survey of 2007-2008, an intercept 10 survey of short-term CaBi users, and an online survey of annual CaBi members. This paper deals with a case study of short-term and annual bikeshare users of Capital 13 Bikeshare (CaBi) in Washington, DC and Arlington County, VA. The monthly and annual users were subjected to an online membership survey, while the users who has a 24-hour or 5-day membership were examined on the basis of 23 survey questions. The objectives for the long-term members were to collect info on demographic and use characteristics of CaBi members, 4 satisfaction with the system, and changes in travel patterns based on bikeshare availability. Short-term users who did not have time to complete the verbal 19 survey were provided with instructions for completing the survey online. The analysis part comprised of the discussion on various aspects: Demographics of Area Cyclists, Short-Term Users, and Annual Members; Income, Car Ownership, and Access to a Bicycle; Trip Purpose, Mode Shift, and Helmet Use. Finally before concluding the paper, the authors also touched upon some of the scopes for future research which included whether there could be significant differences or not between Area and CaBi users when controlling for other factors, admitting that the analysis cannot fully consider the spatial restrictions of the two CaBi surveys, presence of potential biases over the sample as CaBi annual member survey respondents self-selected from email and online solicitations, and 6 responded over the internet, Results for the Washington, DC area may not apply to other US cities, etc.

The approach of how the managing authorities of the bikesharing systems manage the problem of imbalance or bike rebalancing is surprisingly directly related to user-level satisfaction and thus towards profits. (Vallez et al. 2021) of this research paper sought to present a thorough review of the challenges and opportunities in rebalancing of bike-sharing systems (only for 4th generation of BSS). The objective of their research points out towards collecting research papers based on the repositioning-problem in dock-based bikesharing systems, classifying them and to suggest and divert to many novel research venues. A period-wise table containing the main research-topics in BSS research over the decades has been provided in the research paper. The four key-themes addressed by the research paper are: (a) overview of the state-of-the-art related to BSS research, (b) identification of available literature regarding reviews of bike-relocation problem, (c) list and classification of studies related to algorithms used to solve bike repositioning problem, (d) evaluation of further research to be done. The methods of data acquisition included performing keyword search in Google Scholar utilizing some selective constraints, and further steps, etc. The contextualization of research in bike-sharing field was done by performing an exploratory data analysis by making a graphic representation of the main key-words and topics that appear in titles and abstracts on the bike-sharing related papers, by making use of VOSviewer. The research-topics were clustered in 5 groups, and were also further analyzed. Following the BSS Rebalancing Problem Review Analysis, Literature Review Timeline, the authors provided the Summary of Results and Discussion. Finally, an exhaustive table that will assist researchers from different disciplines to address the open challenges in the field was also provided. [20]

3 The problem statement

The target has been to devise marketing strategies to increase the company sales and provide the customers with recommendations and specialized services by analyzing the data and summarizing the key insights from it. Data analytics and many machine learning approaches along with popular python libraries had been used to achieve this. The entire process has a multitude of steps, procedures, approaches, thoughts and implementation techniques. The data for performing all the analysis had been collected from the official site of the Chicago Data Portal. The procurement of the data is followed by cleaning the dataset and sampling operations. We also ran various statistical techniques, such as different machine learning techniques, statistical tests, visualization procedures to learn more about the different parameters hidden in the dataset, whatever we could do to find trends, to learn more about the data we were working with. We were able to get some key insights. We also trained and tested the data many times using different techniques and approaches each time to predict the outcome. We validated the accuracy of each statistical scheme, such as MSE, RMSE, etc., using appropriate metrics, and validated all that seemed reasonable depending on the reliability of the situation.

3 Proposed methodology using classical time-series forecasting algorithms, analysis using novel approach and data visualizations

3.1 Background of the proposed work

The methodology of the entire process of achieving the goal of devising target specific marketing strategies towards converting the customers of the Divvy's bike share system towards dedicated subscribers for enhancing and boosting the profits of the bike share company involves a multitude of steps, procedures, approaches, thoughts and implementation techniques. First of all, to begin with, the data for performing all the analysis is collected from the official site of the Chicago Data Portal. The Chicago Data Portal is an official vault of data which is dedicated towards promoting access to government data and also for encouraging the development of creative tools to engage and serve Chicago's diverse community. The official site managed by the government agency hosts over 600 datasets presented in easy-to-use formats about City departments, services, facilities and performance. The data of our interest is Divvy's bike sharing trips information. Divvy is a well-known and ubiquitous name and is well reputed to provide bike sharing facilities across Chicago and Evanston. This Chicagoland's bike share system provides residents and visitors with a convenient, fun and affordable option for getting around and exploring Chicago. Divvy, like any other bike-sharing system, consists of a fleet of purpose-built, robust and durable bikes locked into a network of docking stations throughout the region. Bikes can be unlocked at one station and returned to any other station in the system. People use bike sharing to explore Chicago, commute to work or school, run errands, and go to appointments and social events. Divvy is available 24 hours a day, 7 days a week, 365 days a year, giving riders access to every bike and station in the entire system. The dataset called Divvy Trips lists individual Divvy bike sharing trips, including the origin, destination, and timestamps for each of the trips. At the time of analysis, the dataset consists of 21,242,740 rows, containing valuable and extensive information about each of the bike trips from 2014 to 2019. The dataset is an example of a living dataset as regular updation adds more and more data to it, thus adding to the bulk of knowledge. The data contains info about the Trip Id, Start Time, Stop Time, Bike Id, Trip Duration, From Station Id, From Station Name, To Station Id, To Station Name, User Type, Gender, Birth Year, From Latitude, From Longitude, From Location, To Latitude, etc., of each of the bike sharing trips monitored by the company. The data has been made public by the company for public use for policy makers, transportation professionals, web developers, data analysts, and a lot of other people to use for maintaining their workflows. The data about each of the trips provided by Divvy is anonymous. The data is processed to remove trips made by staff during system maintenance and inspections. Rides under 60 seconds. This can result in false starts or the user attempting to redock the bike to ensure safety. The data is publicly available for download in a static format, options like CSV, CSV for Excel, JSON, RDF, RSS, TSV for Excel and XML are available.

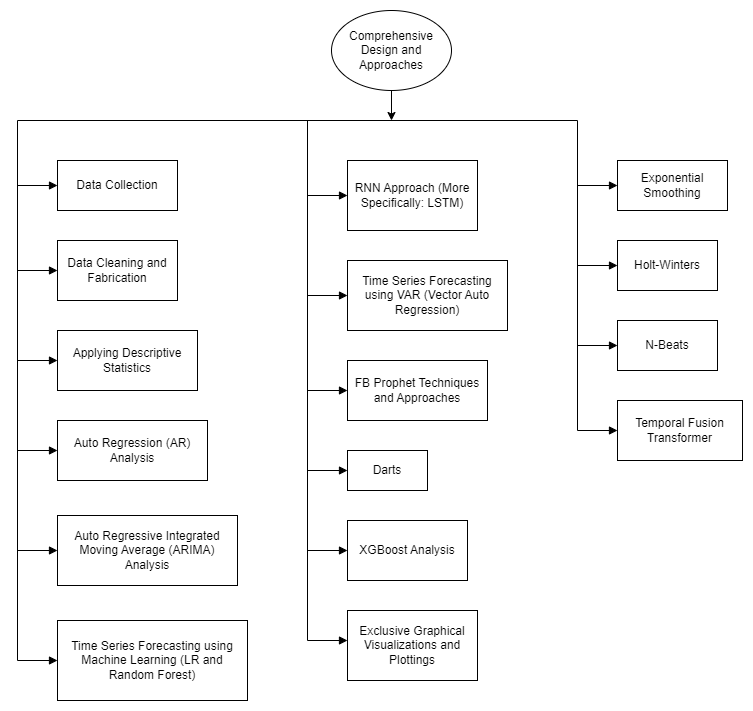


Fig. 1 Proposed methodology.

3.2 Data collection and procurement

As mentioned before, the process starts with searching for and downloading the dataset of Divvy Trips from the Chicago Data Portal, including it in our workflow if all the contraints comply. The ROCCC approach is used to determine the credibility of the data. It demands the data be: (i) Reliable - It is complete, accurate, and represents all bike rides made in the city of Chicago during the selected analysis period. (ii) Original - The data is provided by Divvy, which operates the city of Chicago's bike-sharing service Divvy. (iii) Comprehensive - The data includes all ride details such as start time, end time, station name, station ID and membership type. (iv) Current - It is up-to-date. (v) Cited - Data is quoted and available under a data license agreement. The coding and data analysis is done predominantly in R and Python. The first step involves the importing of data, cleaning the data, and sampling the data as we have limited computing power and our system cannot process over 21 Million data tuples all at once. Hence, to gain a basic understanding first, sampling of the data is necessary. For data cleaning we use R Language. We first loaded all the necessary packages for doing so, followed by sampling. We added new columns: Date, Year, Month, Day and Day of the Week. We then checked the data for errors. We cleaned the column names and checked for duplicate records in rows. The exporting of the cleaned data to a new file is followed by updating the working directory to the script path. After that, we started the descriptive analysis, in which saw the average ride time by each day for Subscribers vs Customers. We sorted the days of the week, analyzed the ridership data by type and weekday, and visualized the number of rides by rider type, and also visualized the average duration of ride by rider type.

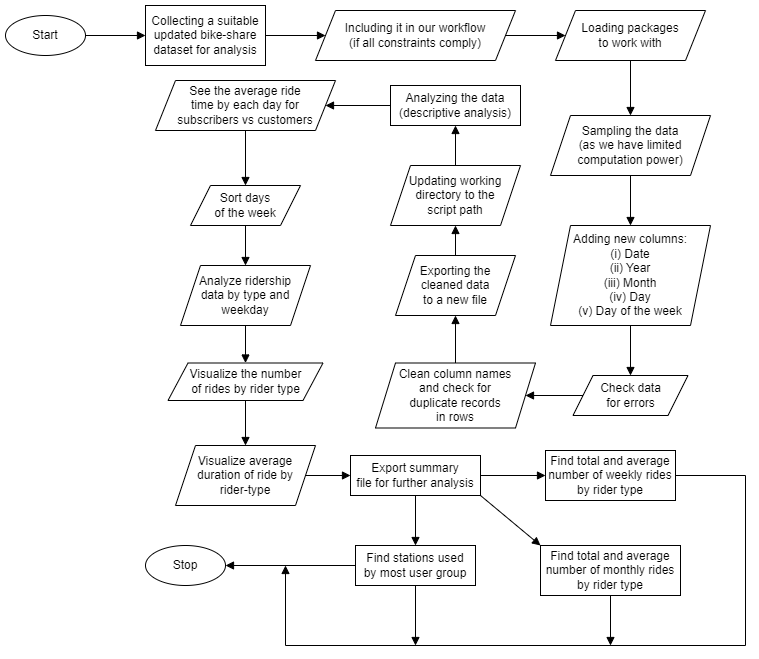


Fig. 2 Combined algorithm for data collection, data cleaning and descriptive analysis of the data.

We also performed a number of statistical techniques to know some more insights into the data that we were working on which included various machine learning techniques, statistical tests, visualization procedures and everything one could do to figure out the various parameters hidden in the dataset, the trends which it indicated and whatever it wanted to convey overall. We also trained and tested the data many times, each time with a different technique and approach and predicted the outcomes. Suitable metrics to validate the accuracy of each of the statistical schemes like MSE, RMSE, etc., were used, and whatever seemed justifiable according to the credibility of the situation.

3.3 Time-series forecasting techniques

3.3.1 Auto regression analysis

An Auto Regression (AR) analysis was performed on the dataset. In AR, we have some inputs and we multiply it with some weights and get the outputs as a continuous value. It is called Auto Regression because here the inputs are the previous values of the time series. In this analysis, we had tried to predict the future trip durations or the trip durations beyond the sample of data which we have actually taken into consideration. The inputs are the past trip durations, each of which came with the start time of the bike share trip. The trip durations corresponding to the actual date and original time stamps consisted of the inputs. So, the entire process starts with importing the necessary python modules which we used for analysis like numpy, pandas, etc. In addition, we also imported the AutoReg module from statsmodels for the Auto Regression part. After all these required inclusions, we read the previously cleansed dataset and extracted only the necessary parameter i.e. the trip duration along with the timestamps for our analysis, and also printed the descriptive info about the data for our convenience. We then checked our data for stationarity. Our data needed to be stationary for us to get a good time-series prediction. A stationary data means that the statistical properties should be constant, viz., mean should not change, there should be no change in variance, and there should be no seasonality or repeating patterns or trends in the data. After plotting the data, and making it certain that we were not sure whether the data is stationary or not via the method of visual inspection, we performed the Augmented Dickey Fuller (ADF) Test for checking for the stationarity. After performing the ADF Test, we got some values which included ADF value, P-Value, Number of Lags, Critical Values, etc. For making the conclusion whether the data is stationary or not we inspected the P-Value. The P-Value is a kind of probability value on whose value we can actually accept or reject the null hypothesis. As the P-Value for our case was less than 0.05 (statistically significant), we were able to reject the null hypothesis, by which we concluded that the data was stationary. Then, for the training and testing purpose, we reserved some percentage of the data for training on the model and the remaining small part for testing purpose. After splitting the dataset into training and testing set, we fed the training lot onto the AutoReg function specifying the no. of lags and then fitted the model. After training the model, we checked the parameters of the model and its summary. We got a lot of values after the training process like Log Likelihood, AIC, BIC, HQIC Scores, etc. We noticed that the P-Values were not less, which indicated that the time lags were not that significant. We then tried to make predictions on the test set by specifying the starting and ending index and do the comparison on the basis of our action of training the model. After we were done with making the predictions, we plotted the results in graphical format. We observed that the predictions were not that bad. After making the predictions and presenting our finding in the form of a graph, were calculted the error in our predictions by making use of RMSE (Root Means Squared Error) value. Finally, we also made future predictions on the trip data and observed our findings for the next timeframe.

3.3.2 Auto regressive integrated moving average

An Auto Regressive Integrated Moving Average (ARIMA) analysis was performed on the dataset. An ARIMA model is quite similar to ARMA (Auto Regressive Moving Average) model. In ARMA, we take into consideration the previous values of our data in addition to the previous errors which becomes our inputs for reaching the output or future values. The ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function) have a role to play in ARMA. The only difference in ARIMA is the order of differentiating to get stationary values (in addition to ACF and PACF). Hence, in ARIMA, we just convert non-stationary series to stationary series before proceeding. The inputs are the past trip durations, each of which came with the start time of the bike share trip. The trip durations corresponding to the actual date and original time stamps consisted of the inputs. So, the entire process starts with importing the necessary python modules which we used for analysis like numpy, pandas, etc. In addition, we also imported the pmdarima module for the ARIMA part. After all these required inclusions, we read the previously cleansed dataset and extracted only the necessary parameter i.e. the trip duration along with the timestamps for our analysis, and also printed the descriptive info about the data for our convenience. We then ran the ADF-Test as we had done in the previous case. After plotting the data, and making it certain that we were not sure whether the data is stationary or not via the method of visual inspection, we performed the Augmented Dickey Fuller (ADF) Test for checking for the stationarity. After performing the ADF Test, we got some values which included ADF value, P-Value, Number of Lags, Critical Values, etc. For making the conclusion whether the data is stationary or not we inspected the P-Value. The P-Value is a kind of probability value on whose value we can actually accept or reject the null hypothesis. As the P-Value for our case was less than 0.05 (statistically significant), we were able to reject the null hypothesis, by which we concluded that the data was stationary. Then, we tried to figure out the order for our ARIMA model (as ARIMA requires 3 parameters viz., PACF, ACF, and the order of differencing as mentioned before). So, from pmdarima we imported the auto\_arima function for figuring out the order. The auto\_arima function tries to use different combinations for figuring out the order for the model, and for every model it assigns a score, which is called the AIC (Akaike’s Information Criterion), and the goal is to minimize the AIC. After getting the best model i.e. the model with the minimum AIC, we splitted the dataset into training and testing as we did before. For fitting the model, we called the ARIMA function which we imported from statsmodels python package and fed the training part onto it after specifying the order. We then tried to make predictions on the test set by specifying the starting and ending index and do the comparison on the basis of our action of training the model. After we were done with making the predictions, we plotted the results in graphical format. After making the predictions and presenting our finding in the form of a graph, were calculated the error in our predictions by making use of RMSE (Root Means Squared Error) value. We observed that the RMSE value is very different and is actually quite more than the Mean value, which made us conclude the model to be bad one. After retraining the model on the entire dataset, we tried to make the future predictions.

3.3.3 Random forest and linear regression model building

Time series forecasting using machine learning was performed on the dataset. Two ML models namely Random Forest and Linear Regression were used to make predictions. We read the previously cleansed dataset and extracted only the necessary parameter i.e. the Trip Duration along with the timestamps for our analysis, and also printed the descriptive info about the data for our convenience. We sorted the data on the basis of timestamps. After that, we plotted the data in the form of a graph. In supervised learning, we need an input and an output for our model to learn the relationship between the input and output to implement our ML models. Hence, we created 3 additional columns in our dataset by shifting the values from the original trip\_duration column by one row, successively for each of the successive columns. By doing this, we ensured that we make the shifted values as the outputs for the original inputs, thus solving our problem of fetching the outputs for training our models. We imported functions like LinearRegression and RandomForestRegressor from sklearn package for our further work. After a slight data preprocessing, and splitting the dataset into training and testing sets, we fitted the 2 models onto the training data. After doing these operations, we predicted the results by using the test set for both Random Forest and Linear Regression as well and plotted the results in 2 separate graphs, one for each of the models. We observe that the models were capturing the trends quite well. To compare the accuracy of the predictions made by the two models, we calculated the MSE (Mean Squared Error) for each. The MSE for Random Forest Model came out to be around 1382.16, whereas that for Linear Regression was around 1328.63, which were not that different. Hence, we concluded that the Linear Regression model was slightly better that the Random Forest model, although the deviation was very insignificant.

3.3.4 Long short term memory approach

A RNN (Recurrent Neural Network) approach or more specifically LSTM (Long Short Term Memory) was also applied to the dataset. A RNN is a form of ANN that is commonly used to make predictions on data where there normally involves a sequence. LSTM is a form of RNN. So, the entire process starts with importing the necessary python modules which we used for analysis like numpy, pandas, matplotlib, etc. We read the previously cleansed dataset and extracted only the necessary parameter i.e. the Trip Duration along with the timestamps for our analysis, and also printed the descriptive info about the data for our convenience. We sorted the data on the basis of timestamps. After that, we plotted the data in the form of a graph for our convenience. To view the exact seasonal nature and how the seasonality in our data actually looked like, we imported a function called seasonal\_decompose from statsmodels package. We hadn't checked for the stationarity in our data, because technically RNN can work with non-stationary data as well and do not require the data to be stationary. After that, we splitted the dataset into training and testing sets. We preprocessed the data using a MinMaxScaler for converting the data into a scale of 0 to 1. For doing so, we first fitted the training set using the scaler object, and then we transformed both the training and testing sets using the transform function of the scaler object. We then formatted the data for feeding it to the neural network model. For doing so, we define the generator which consists of specifying the no. of inputs (assigning 3 to it), no. of features (assigning 1 to it). The no. of features would have been more if we were dealing with more than one time-series, but that is not the case with us here. Then we called the TimeSeriesGenerator function and give it the scaled trained input, the no. of inputs, etc. We finally created the model after calling the Sequential, Dense and LSTM classes from keras. We added an LSTM layer with a 100 neurons, and the activation function as the RELU activation function. We finally compiled the model using the Adam optimizer and MSE as the loss function. We fitted the model for 50 epochs. After training the model, we computed the loss for each epoch, and we also plotted it. We took the last 12 values in the training set to predict the first value in the test set. We observe that the original value was 0.01374463 whereas the model had predicted 0.01540671 which is pretty close. Then we made predictions on the testing set and converted it back into the original scale and performed the Inverse Transform upon the test predictions and appended those predictions to our original test set, and finally, plotted those two in the form of a graph to see how similar or different they are from each other. To put a number to see how good the predictions were we used RMSE which came out to be something around 1268.4964.

3.3.5 Forecasting using vector auto regression

Time series forecasting using VAR (Vector Auto Regression) was performed on the dataset. Normally, in Auto Regression we predict the future values based on the previous values or the previous time lags of a particular time series to make predictions into the future. In VAR we assume that there are two time series which have a correlation. We started by importing the cleansed dataset and this time in addition to the Trip Duration, we also included some other parameters like From Station Id, To Station Id, From Latitude, From Longitude, etc. We also imported various classes and functions like plot\_acf, plot\_pacf, VARMAX, VAR, grangercausalitytests, dfuller, etc., mainly from statsmodels package and others like tqdm, itertools, etc. And, as always we also imported python modules like numpy, pandas, matplotlib, etc. We plotted the 8 parameters into 8 different graphs. We checked for stationarity for all the 8 time-series using ADF Test. We inferenced after observing all the P-Values corresponding to each of the time series that they all are stationary as all of them give P-value less than 0.05. Before progressing, we checked whether there was any correlation between 2 suitable parameters or columns of the dataset, because as mentioned before it is the fundamental concept behind Vector Auto Regression. We did that through Granger Causality Test. According to that test if the p-value comes out to be less than 0.05, then that means the hypothesis is true. We concluded that from station id causes trip duration and to station id also causes trip duration. After all that, we splitted the dataset into training and testing set. To ascertain the number of lags, we fed the difference data of the training set into the VAR function to make the model. After specifying the maximum lags in the select\_order function of the model, we observed the summary of our analysis. The results came out in the form of 4 scores namely AIC, BIC, FPE and HQIC scores. For a model to be good, these scores must be as low as possible. For AIC the minimum score was observed to be 1.651 in the lag no. 81, minimum score of 1.929 for BIC in lag 29, 5.211 for FPE in same lag as that of AIC, and finally for HQIC, the minimum score of 1.773 was found to be in lag 45. To fit the model we used the VARMAX class in which we fed the training data. One advantage of using that class was that it is known to make forecasting very easy. After the computation, we got the AIC score to be 1473817.706, BIC as 1474584.888, and HQIC as 1474054.573, and we also got other informations such as results for equation trip\_duration, results for equation from\_station\_id, results for equation to\_station\_id, error covariance matrix, etc. For making the forecasts from the point where the training data ends, we specified the number of forecasts as 7600, and also calculated the mean of all the predictions. Finally, we calculated the mean values of each of the time series values and the RMSE for comparing the errors corresponding to each series. We observed that the RMSE of trip duration is more than its mean, which is statistically not good. Although the same is the case for from longitude and to longitude as well, but it is because of the negative natures of its values and hence including it makes no sense. For other parameters, the RMSE values are quite less than the respective mean values, which meant that it was statistically good.

3.3.6 Data forecasting and interactive visualizations using FB Prophet

Time series forecasting using FB Prophet was performed on the dataset. Prophet is an open-source software released by Facebook's Core Data Science Team available for free download on CRAN and PyPI. In order to forecast time series data, Prophet uses an additive model to fit the non-linear trends with yearly, weekly and daily seasonality, plus holiday effects. The underlying algorithm as mentioned before, the generalized additive model can be decomposed into three main components viz. trend, seasonality and holidays. Its worth mentioning that trend and seasonality are two important but very difficult to quantify components of a time series analysis but FB Prophet does a great job capturing them both. It is best suited for time series with strong seasonal effects and historical data for multiple seasons. Prophet is tolerant of missing data and changing trends, and usually handles outliers gracefully. We began by importing the necessary libraries such as pandas and prophet. Then we loaded the cleansed divvy tripdata dataset, dropped the NA values and resetted the index. Then we extracted the Start Time and the Trip Duration columns because we had to make predictions concerning these two columns only and discarded everything else. After that, we changed the columns names to 'ds' and 'y' respectively (The start\_time to ds and the trip\_duration to y). We had to do that because in order to use the functionalities its essential to do so, as the team at Facebook had hard coded it to make it work in that way. We parsed the start\_time to datetime format for future convenience. We splitted the dataset into training and testing sets and started making predictions. For making the predictions, we first invoked the Prophet class and fillted our model onto the training set, and to make the predictions we used the make\_future\_dataframe function and also specified the periods i.e. the number of time steps into the future for which we were going to make the predictions, and also the frequency. After all that, we made the predictions using the predict function and saved the predictions into the forecast dataframe. After making the predictions, we observed the values like yhat, yhat\_lower and yhat\_upper, which were actually our values of interest to make the plots. Then, using the built-in FB Prophet visualizations through the plot\_plotly function after feeding the model and the forecasted values we obtained the great visualization. We got an interactive plot where we had the option to set the plot to display as per 1-week data, 1-month data, 6-months data, 1-year data, and the whole trend as well. Also we could see a more detailed fraction of the entire plot by setting the starting and the ending timestamps manually on the figure window. Another visualization which the FB Prophet provides is the visualize components using the plot\_components\_plotly function. It gives the general trend throughout the dataframe, and also the yearly and the weekly trends. Finally after making the predictions, we had to evaluate how good the model was. In order to do so, we created a separate test set and calculated the RMSE between the actual test set and the predicted values. We also calculated the mean value of the test dataset. The RMSE was around 611.35 and the mean value was around 828.21 which indicated that the model was pretty good as the RMSE was less than the Mean.

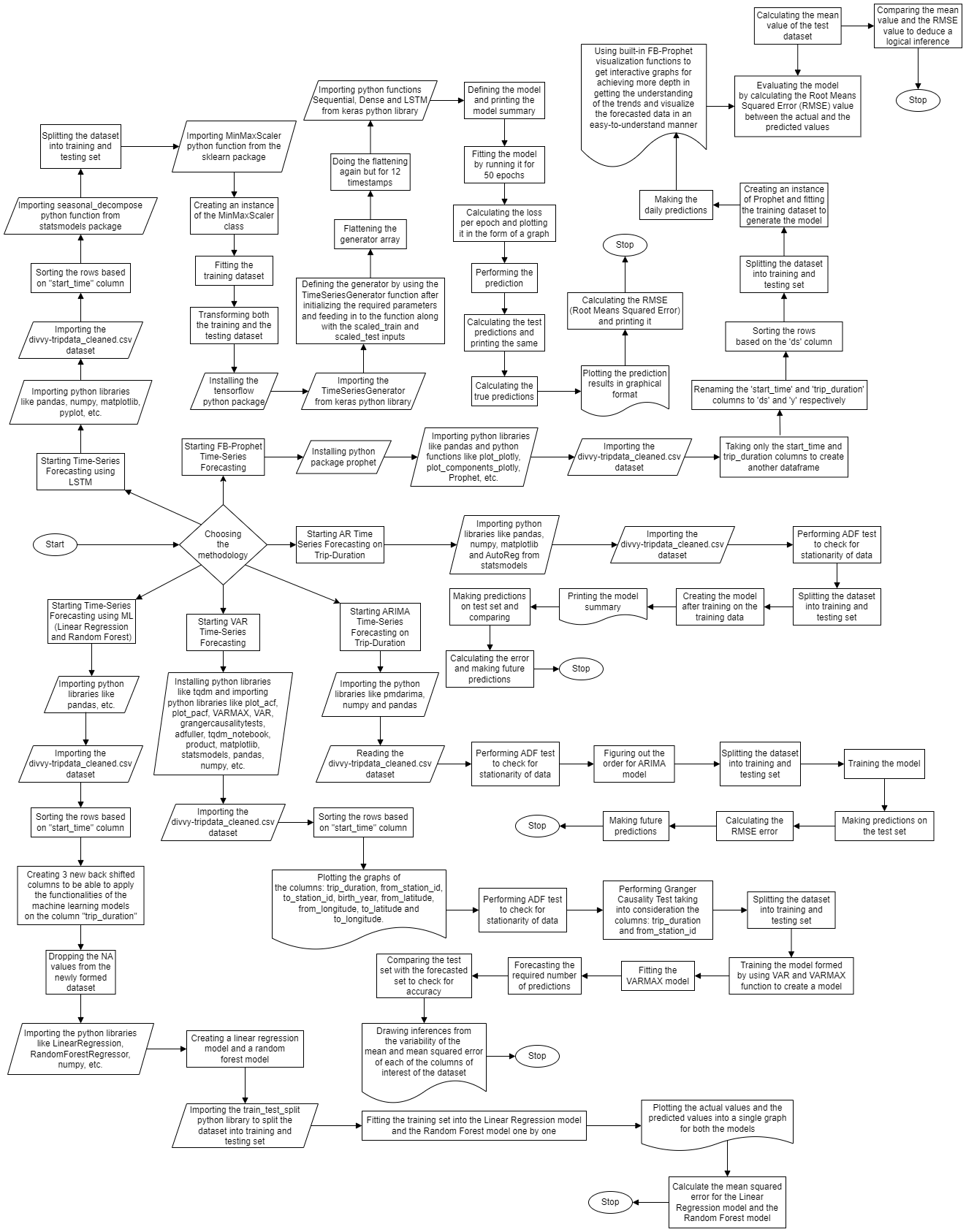


Fig. 3 Algorithm for various methodologies used for forecasting and analysis.

3.3.7 Data visualization techniques

Visual aid to get more insights about the sampled data was also leveraged to understand the underlying trends and as a preliminary step for attempting to formulate the final marketing strategies for converting more number of customers to the subscribers of the Divvy bike-share service. Some of the visualizations were achieved with the help of RStudio, whereas some others with basic MS Excel tips and tricks. Using RStudio and ggplot library, we obtained the graphs for the number of rides taken by each of the category of the service-users for each day of the week for the sampled data. According to the data, the customers were observed to take more number of rides than the subscribers on weekends. On the contrary, on the working week-days, the dedicated subscribers were noticed to take significantly greater number of trip rides than the customers. Another RStudio graph obtained showed the average trip duration of the customers and the subscribers on different days of the week. A straight observation yielded a clear-cut inference of the customers exceeding the subscribers in terms of the average trip duration taken on each day of the week. Based upon the original cleansed and sampled data, 3 datasets exported after performing the descriptive analysis, focussing upon the average trip duration and number of trips of each user type distributed across each week days, on the average trip duration and number of trips of each user type distributed across all the months, and the summary of the stations were generated separately. These datasets were used to generate graphs manually on MS Excel afterwards. The observations from month focussed data graphs revealed that the total number of rides taken by customers were lesser than the subscribers on every month of year, except in the month on August where surprisingly the total number of ride trips by the customers were actually noticed to be greater than the total number of ride trips taken by the subscribers although the difference was not that significant as compared to some of the extreme cases. On the other hand, the average trip duration was always higher for the customers in all the months, with the difference between the two parties remaining fairly consistent. In another effort, we tried to graphically show the most popular bike stations based upon the number of rides associated with each for both the subscribers and the customers separately. For the customers, the top 10 most popular bike docking stations came out to be Michigan Ave & Washington St, Indiana Ave & Roosevelt Rd, McClurg Ct & Illinois St, Lake Shore Dr & North Blvd, Museum Campus, Theater on the Lake, Millenium Park, Michigan Ave & Oak St, Lake Shore & Monroe St, and Streeter Dr & Illinois St in increasing order of their popularity. For the subscribers, the top 10 most popular bike docking stations came out to be Canal St & Jackson Blvd, Larrabee St & Kingsbury St, Columbus Dr & Randolph St, Clinton St & Madison St, Orleans St & Merchandise Mart Plaza, Franklin St & Arcade Pl, Dearborn St & Monroe St, Canal St & Madison St, Canal St & Adams St, and Clinton St & Washington Blvd in increasing order of their popularity.

3.3.8 Analysis with darts library

Darts is a very useful python library providing many useful tools to perform many data analytics operations and also time forecasting functionalities. We started by creating a new R-Script document. We updated the working directory to the script path, followed by loading the R packages like tidyverse and dplyr, aggregating the “trip\_duration” for each “year-month”, and writing the resultant data frame into a csv file. For doing the Darts time series forecasting, we began by installing the darts python package, importing python libraries like pandas and numpy, and importing the newly created dataset and reading the pandas dataframe. We then sorted the dataframe on the basis of “year-month/month” column, imported python libraries and functions like matplotlib, autoreload, sys, time, pyplot, datetime, reduce, functools, darts, TimeSeries, NaiveSeasonal, NaiveDrift, Prophet, ExponentialSmoothing, ARIMA, AutoARIMA, RegressionEnsembleModel, RegressionModel, Theta, FFT, mape, mse, check\_seasonality, plot\_acf, plot\_residuals\_analysis, warnings and logging, made a series from the pandas dataframe using TimeSeries function, plotted the series formed, splitted the series into training and validation sets and plotting them, created an instance of the NaiveSeasonal class (K=1) named naive\_model, fitted the naive\_model with the training set, plotted the actual data alongside the naive forecast (K=1), and used the plot\_acf function to plot the graph, taking the lags as 9 and alpha as .05 followed by checking for any seasonality in the data. We then created an instance of the NaiveSeasonal class (K=4) named seasonal\_model, fitted the seasonal\_model with the training data, created seasonal forecast for the next 120 values, plotted the actual data alongside the naive forecast (K=4), created an instance of the NaiveSeasonal class (K=13) named seasonal\_model, fitted the seasonal\_model with the training data, created seasonal forecast for the next 120 values, plotted the actual data alongside the naive forecast (K=13), created an instance of the NaiveSeasonal class (K=22) named seasonal\_model, fitted the seasonal\_model with the training data, created seasonal forecast for the next 120 values, plotted the actual data alongside the naive forecast (K=22), created an instance of the NaiveDrift class named drift\_model, fitted the drift\_model with the training data, created drift forecast for the next 120 values, created combined\_forecast with the drift\_forecast and the seasonal\_forecast eliminating the last value of the training set, plotted the combined graph of the trip\_duration, combined\_forecast and the drift\_forecast alongside each other in the same graph, followed by calculating the mean absolute percentage error (MAPE) for the combined naive drift and calculating the MAPE for each of the models: ExponentialSmoothing, Prophet, AutoARIMA, Theta. We then searched for the best theta parameter, by trying 50 different values, calculated the MAPE of the best theta model, plotted the combined graph of the best theta model prediction, calculated the Average error (MAPE) and the Median error (MAPE) over all historical forecasts and drawing the Industrial error scores (histogram) programatically, plotted the historical forecast theta data (backtest 3-months forecast (Theta)) with the original data, plotted residuals analysis over the best theta model residuals, created an instance of the ExponentialSmoothing class named model\_es, formed the backtest 3-months ahead forecast (Exponential Smoothing) plot, plotted residuals analysis over the exponential smoothing model residuals, created an instance of the ExponentialSmoothing class named model\_es and fitting the training data before doing the probabilistic forecast for 500 samples, followed by the plotting of the probabilistic forecast and the probabilistic forecasts for the 1-99th and 20-80th percentiles. We then formed an ensemble model using forecasting models like NaiveSeasonal(6), NaiveSeasonal(12) and NaiveDrift, fitted the ensemble model with training data and predicting for 120 values, plotted the ensemble forecast graph, calculated the MAPE for the ensemble prediction, drew the ensemble forecast (historical forecast) graph, and calculated the MAPE for the historical ensemble prediction.

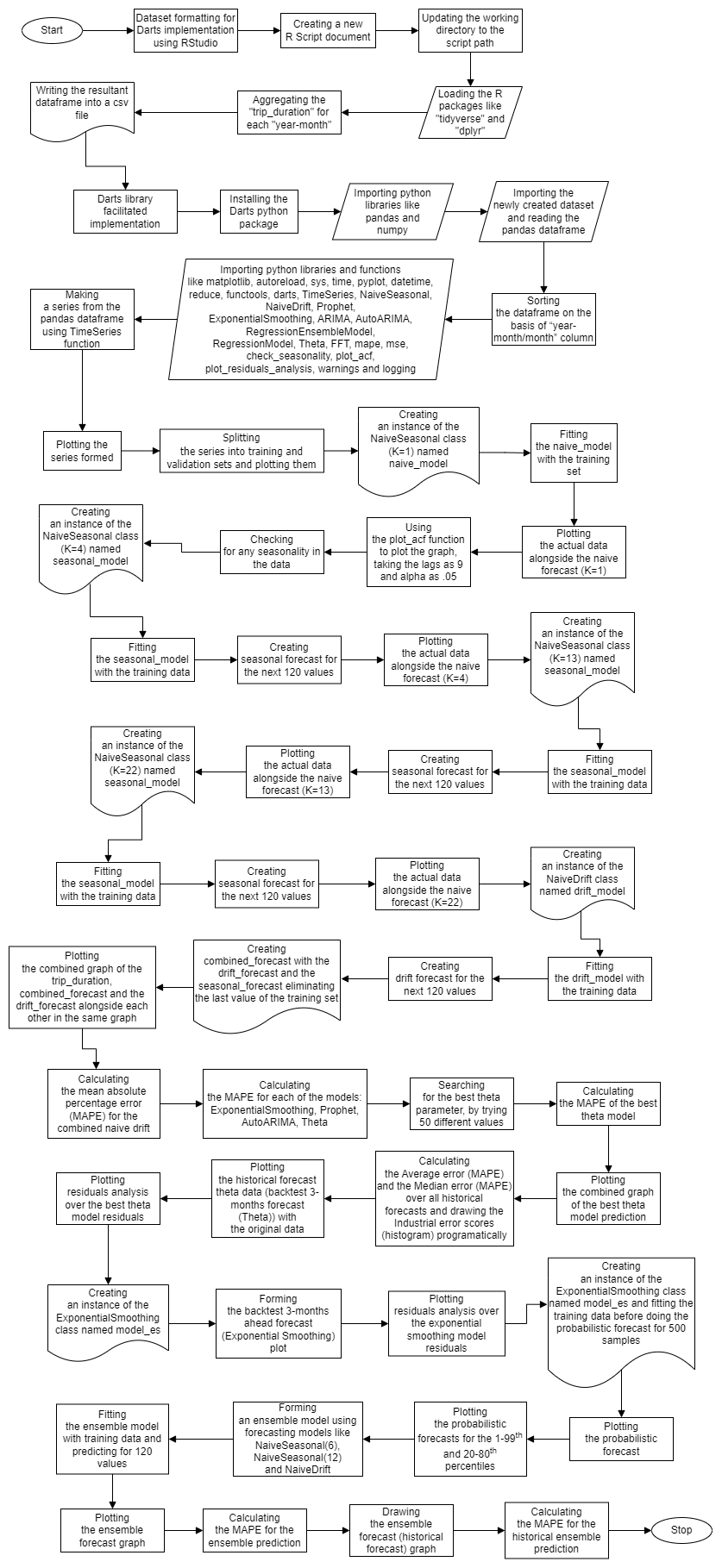


Fig. 4 Algorithm for Darts implementation.

3.3.9 XGBoost

XGBoost has also been used on the data for various analysis as it is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. First of all, we imported python libraries and functions such as numpy, pandas, matplotlib, pyplot, seaborn, xgboost, sklearn and mean\_squared\_error, followed by importing the dataset and converting it into a dataframe, sorting the dataframe on the basis of the “start\_time” column, plotting the data, splitting the dataset into training and testing sets with a suitable fragmentation mark/timestamp and plotting the resultant, plotting a week of data from the starting timestamp '2014-06-30 10:51:00' extending till the ending timestamp '2014-07-07 10:51:00', creating time series features based on time series index, visualizing the feature/target relationship, creating the model using the XGBoostRegressor function and fitting it, creating a graph for realizing the feature importance of the various different parameters, forecasting on the test data and plotting a graph for the same, plotting the Truth Data and the Prediction for the specific week of data, and then finally calculating the RMSE score and the error.

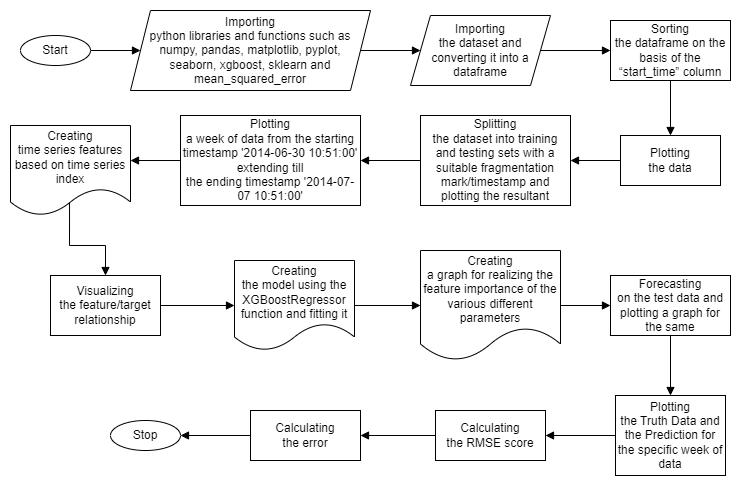


Fig. 5 Algorithm for XGBoost implementation.

3.3.10 Tableau data visualizations

Data visualizations have also been performed using Tableau. Tableau is a powerful and rapidly growing data visualization tool used in the business intelligence industry. It helps simplify raw data into a very understandable format. Tableau helps you create data that experts at all levels of your organization can understand. Users who are in the non-technical domain can also use it to create customized dashboards. Data blending, real time analysis and collaboration of data are some of the best features of Tableau. Data analysis using Tableau tools is very fast and the visualizations created are in the form of dashboards and spreadsheets. In order to utilize Tableau we had to do some operations first. We navigated to the official Tableau website on a web browser, moved to the Tableau Public download webpage for downloading the installation file of the version of the software for public usage, clicked on the link to download the Tableau Public 2022.4 execution file (TableauPublicDesktop-64bit-2022-4-0.exe), saved the execution file to the downloads folder, double-clicked on the executive file to install the software on the system, and finally agreed to the terms and conditions and then selecting suitable preferences to finally install the software on the computer. After following all the steps, we started the Tableau Public 2022.4 application. First of all, we imported the the divvy-tripdata\_cleaned.csv dataset into the workspace, then we created a new sheet for doing all the visualizations, dragged the ‘User Type’ to the Columns and the ‘Year’ onto the Rows to obtain the bar chart, followed by making other charts which included dragging ‘Gender’ to the Columns and the ‘Year’ onto the Rows, ‘Gender’ to the Columns and the ‘Day’ onto the Rows, ‘Birth Year’ to the Columns and the ‘Day’ onto the Rows, and ‘User Type’ to the Columns and the ‘Month’ onto the Rows, in order to obtain the respective charts. We also explored some other aspects of the data and familiarized ourselves with the Tableau interface.

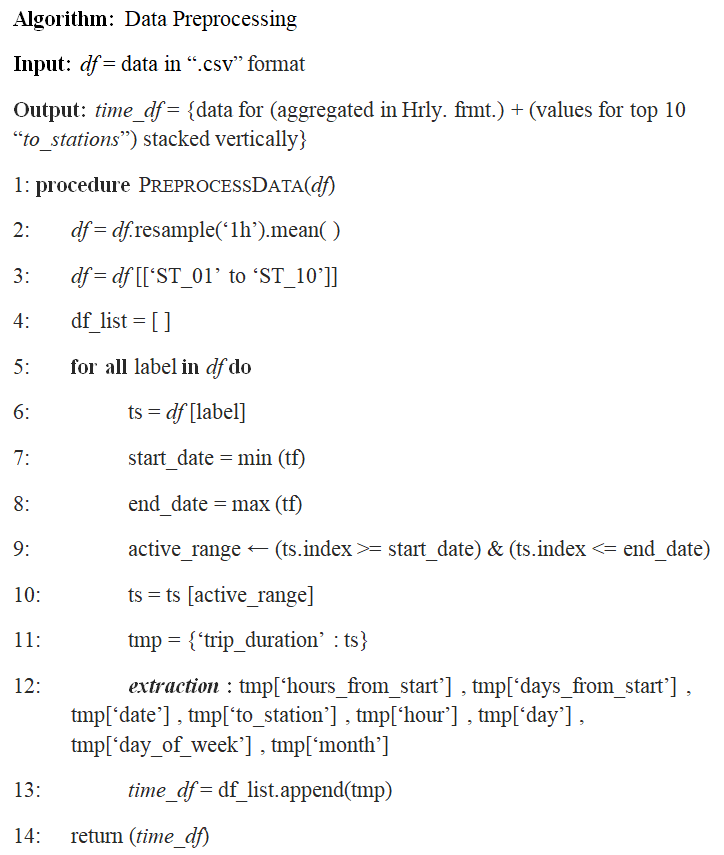


Fig. 6 Data preprocessing algorithm for novel approach.

3.3.11 Novel approach

Apart from all the data analysis and time-series forecasting methods and techniques used, we were able to come up with a novel analysis and prediction method built right from scratch using only some basic yet useful preliminary python libraries. The novel approach extends from data preprocessing all the way till modelling ways which further includes training the model, saving the best model, evaluating the best model that we got and finally the prediction based upon the former steps. Its an algorithm based upon medium complexity yet robust enough to come up with accurate predictions. The algorithm can be broadly divided into two, specifically the Data Preprocessing part and the Model Activities. For the data preprocessing, we first take the data.csv file as the input and then do the further operations on preprocessing for it to make it fit to be later fed into the TimeSeriesDataset function by aggregating it in hourly format and taking the corresponding top 10 bike-share stations centered data stacked vertically, thus creating new features in the process. We find the new data frame known as time\_df in the process. The next sub-algorithm we devised is for the modelling activities which itself is a combination of four major subsections: training, saving, evaluation and prediction. For this technique, we first of all input the time\_df dataframe which we got previously by the method we followed while data preprocessing. Before the beginning of the main approach, we did a little exploratory data analysis for seeing how our formatted or preprocessed data actually looked like then. For doing the exploratory data analysis, we took the most popular 10 stations and plotted the mean of the trip\_duration in a graphical format as each time-series had different properties. For simplicity, we plotted the first month of every time-series. We noticed that there was no noticeable trend but each time-series had slightly different seasonality and amplitude. We could have further experimented and checked for stationarity, signal decompositions, and so on, but in our case, we focussed on the model-building aspect only. After the small exploratory data analysis, we passed our time\_df dataframe to the TimeSeriesDataset format which was very useful since, we were exempted from writing our own Dataloader, we could specify how our algorithm would handle the dataset's features, and we could normalize our dataset with ease as we had the freedom to normalize each time-series individually and also normalization was mandatory because all time sequences differ in magnitude. The trip\_duration was the target variable in our case. After that, we found out the actuals and started building and then training our model. In training our model we used the EarlyStopping procedure so that it kicks in everytime the model was on the verge of getting over-trained. After that we fitted our model, and also loaded and saved the best model. For the evaluation of our model we calculated the average p50 loss by comparing the actuals and the predictions. We finally made the predictions and created one plot for each to\_station between the actual and the predicted so as to get a good picture of our study.

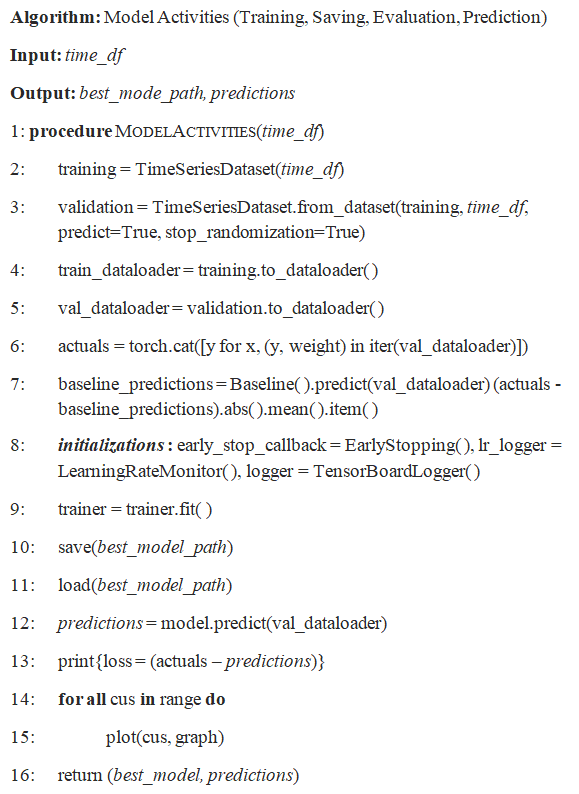


Fig. 7 Algorithm of modelling activities for the novel approach.

3.3.12 Other analysis techniques

Exponential Smoothing was effective in our study. After importing python packages like pandas, statsmodels, etc., python functions like ExponentialSmoothing, etc., importing the dataset and converting it into a dataframe, we decided which columns to include alongside the column acting as index, sorted the dataframe on the basis of the “start\_time” column, formed the dataframe on the basis of the frequency of “start\_time” column, created a model using the Exponential Smoothing function and also fitting it, specified the time series to model using the “endog” parameter, predicted the model forecast with taking steps of 100, and finally plotted the graph of the predictions.

Holt-Winters was another approach. First of all we imported the python packages like pandas, statsmodels, etc., python functions like ExponentialSmoothing, etc., imported the dataset and converted it into a dataframe, decided which columns to include alongside the column acting as index, sorted the dataframe on the basis of the “start\_time” column, formed the dataframe on the basis of the frequency of “start\_time” column, created a model using the ExponentialSmoothing function, specified the parameters: endog as trip\_duration, trend as add, seasonal as add, and seasonal\_periods as 7 and also fitted the model, predicted the model forecast with taking steps of 100, and plotted the graph of the predictions from timestamp of ‘2014-06-30 10:51:00’ onwards.

The N-Beats approch for time series forecasting involved importing python libraries like pandas, numpy, datetime, matplotlib, pyplot, darts, warnings, etc., python functions like TimeSeries, etc., reading the dataset, printing the head part of the dataset, forming a series using the TimeSeries function from the dataframe, plotting the series, importing the check\_seasonality function from the darts.utils.statistics package, checking for whether its daily seasonal or not and also finding out the period, checking for whether its weekly seasonal or not and also finding out the period, splitting out the series into training and testing sets, plotting the train as well as test parts, importing the NaiveSeasonal function from darts.models.forecasting.baselines package, creating a naive seasonal model by taking K as 6, fitting the naive seasonal model thus created with the train set, doing the prediction for the naive seasonal model using the items as 3, plotting the prediction results in graphical format, importing mae function from darts.metrics package, calculating the Mean Absolute Error for the naive model, printing the MAE, importing the NBEATSModel function from darts.models package, importing the Scaler class from darts.dataprocessing.transformers library, creating an instance of the scaler class and fitting it with the training set, creating the nbeats model by specifying the necessary parameters with suitable values and fitting the model, doing the prediction for N-Beats model and priting the MAE for it after calculation, importing the concatenate function from darts library, importing datetime\_attribute\_timeseries as dt\_attr from darts.utils.timeseries\_generation package, utilizing the concatenate function, creating the concatenated scaler instance and applying fit\_transform, doing the prediction for it and plotting the plot for the test set, and finally calculating the MAE and printing it.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **RMSE** | **MSE** | **MAE** | **Average error (MAPE)** | **Median error (MAPE)** |
| VAR | 1931.66 | 3731328.53 | - | - | - |
| ARIMA | 2546.03 | 6482289.68 | - | - | - |
| Linear Regression | 1328.63 | 1765265.62 | - | - | - |
| XGBoost | 1919.58 | 3684787.38 | - | - | - |
| Random Forest | 1382.16 | 1910356.73 | - | - | - |
| FBProphet | 611.35 | 373754.17 | - | - | - |
| AR | 567.40 | 321947.38 | - | - | - |
| Auto-ARIMA |  |  | - | 19.09% | |
| LSTM | 1268.50 | 1609083.24 | - | - | - |
| Naive Seasonal |  |  | 21.67 | - | - |
| Naive Drift + Seasonal | - | - | - | 15.67% | |
| Theta | - | - | - | 6.66 | 4.74 |
| TFT | - | - | - | - | - |

Table 1 Metrics table for all algorithms used in study.

4 Investigational results and discussions

4.1 Graphical interpretations

Based on our entire study, we were able to get many insights. Some of the major ones came from the graphical study itself. From the data sampled from between the aforementioned periods of time, its clear that the subscribers comprised of the majority of the bike services takers than the customers with the former exceeding the later by 3/5th of the total Divvy bike-share market whereas the later only has a 2/5th sharehold within the company services. The average ride duration of customers is nearly 3 times than that of subscribers. The number of rides taken by customers get doubled on weekends compared to weekdays, while the number of subscriber rides remains constant throughout the week. The subscribers average ride duration remains constant throughout the week, but the customers average ride duration fluctuates during the course of the entire week. This fluctuation might have been caused by the increase in number of customers on weekends. The average ride duration of customers does not change at the same rate as the number of rides. Still, there's a moderately positive correlation. It shows weekend customers tend to take longer rides than weekday customers. The average ride duration of customers increase and hits its peak on August but remains the same and nearly close to the peak in the months of March, April, May, June and July. The top five stations used by the customers are Lake Shore Dr & Monroe St, Michigan Ave & Oak St, Millenium Park, Theater on the Lake, and Museum Campus.

The logical conclusions which we deduced helped us to summarize our findings to a remarkable extent. Overall, customers take less number of rides but for longer durations. The customers of Divvy bike-share services mostly use bikes for recreational purposes. Unlike members who have consistent activity throughout the year, customers' use of bikes on weekends and holiday suggests they use them for recreational purposes. The customers are most active on weekends and tend to take longer rides. The customers take longest rides on the months of March, April, May, June, July and August. It peaks in the month of August.

4.2 Designing recommendations and schemes

After knowing how the customers of Divvy bike-sharing company use bikeshare differently, we were able to design recommendations and schemes to boost the sales of the company services. Designing riding packages by keeping recreational activities, weekend contests, and summer events in mind as more customers are inclined towards it can help to elevate purchase of such plans. If the customers are charged on duration basis, offered specialized discounts and coupons for regular and substantial users, this way users will be encouraged for more longer rides and thus it will result in high revenue. The designing of seasonal packages will allow flexibility and encourage customers to get membership for specific periods they want rather than paying for annual subscription. Effective and efficient promotions can be achieved by targeting customers at the busiest times and stations. The favorable days are weekends, the favorable months are June, July and August, and the most attractive stations are Lake Shore Dr & Monroe St, Michigan Ave & Oak St, and Millenium Park.

5 Conclusion and future work

By our study we were able to get a multi faceted view of the bike share infrastructure of the Divvy and also peep through the various parameters and statistical data of the company based in Chicago. We applied various pre-existing machine learning algorithms some of which are ubiquitous when it comes to doing time-series forecasting, besides using our own novel algorithm for achieving the feat. The various statistical figures we got by performing all the tests, the various types of visualizations we obtained by using some of the most dominant data science tools currently being used in the industry such as Tableau gave us a different view over the entire bike-share scenario in the region of our study. Based on our endeavoring efforts and never ending hardwork to scourge the data science community to learn about the latest and the best algorithms and scientific tools, we can confidently assert that our work is one of the most comprehensive and detailed work ever done especially in this field of research. Its important to point out that in some of the time-series forecasting algorithms we were able to achieve considerable accuracy and the predictions were also upto the mark. Nevertheless, its noteworthy that there is always a scope of improvement in any research study. We focussed on a single case study confined to a specific region of the globe but more generic ones can be done with proper planning and severe ambition in mind. Some of the pre-existing algorithms were used didn’t gave us satisfactory results and metrics, but still we included them in our study for future researchers to dive into and understand whether the problem actually was, be in the wrong way of implementation or even for the shortcomings of the specific algorithm itself to map the scenario. It would also be unjust to not speak of our novel algorithm, which despite performing quite well did show a lot of vices. It is upto the future researchers to pin-point those weaknesses and improve it and make it an even bigger grand success. Although we tried to cover the challenge of bike-rebalancing in this research work, we were not successful and our derived methodologies were not upto the mark and thus we decided to scrap that dimention of approach towards solving the problem but its not impossible and has great scope of implementation in the future studies.

Acknowledgements We would like to thank and express ingeniously our whole-hearted thanks to Dr. Nithyanandam P, Head of the Department, Project Coordinators, Dr. Abdul Quadir Md, Dr. Priyadarshini R and Dr. Padmavathy T V, School of Computer Science and Engineering, SCOPE, Vellore Institute of Technology, Chennai, for their valuable support and encouragement.

## Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

# References

Md Doulotuzzaman Xames, Jannatul Shefa, Ferdous Sarwar, "Bicycle industry as a post‑pandemic green recovery driver in an emerging economy: a SWOT analysis", Springer-Verlag GmbH Germany, part of Springer Nature 2022 (Environmental Science and Pollution Research), July 2022. <https://doi.org/10.1007/s11356-022-21985-2>

Weiwei Jiang, "Bike sharing usage prediction with deep learning: a survey", Springer-Verlag London Ltd., part of Springer Nature 2022, Neural Computing and Applications (2022) 34:15369–15385, June 2022. <https://doi.org/10.1007/s00521-022-07380-5>

Suzana Regina Moro, Paulo Augusto Cauchick-Miguel, "An Analysis of a Bike-Sharing System from a Business Model Perspective", Brazilian Journal of Operations & Production Management, Vol. 19, No. 2, e20221400, 2022, ISSN 2237-8960 (Online), June 2022. <https://doi.org/10.14488/BJOPM.2021.050>

Yuanyuan Guo, Linchuan Yang, Yang Chen, "Bike Share Usage and the Built Environment: A Review", Frontiers in Public Health (www.frontiersin.org), Volume 10, Article 848169, February-2022. <https://doi.org/10.3389/fpubh.2022.848169>

Songhua Hu, Mingyang Chen, Yuan Jiang, Wei Sun, Chenfeng Xiong, "Examining factors associated with bike-and-ride (BnR) activities around metro stations in large-scale dockless bikesharing systems", Journal of Transport Geography 98 (2022) 103271, Elsevier Ltd., December 2021. <https://doi.org/10.1016/j.jtrangeo.2021.103271>

Hanning Song, Gaofeng Yin, Xihong Wan, Min Guo, Zhancai Xie, Jiafeng Gu, "Increasing Bike-Sharing Users’ Willingness to Pay — A Study of China Based on Perceived Value Theory and Structural Equation Model", Frontiers in Psychology (www.frontiersin.org), Volume 12 | Article 747462, January 2022. <https://doi.org/10.3389/fpsyg.2021.747462>

Xiaonan Zhang, Jianjun Wang, Xueqin Long, Weijia Li, "Understanding the intention to use bike-sharing system: A case study in Xi’an, China", PLoS ONE 16(12): e0258790, December 2021. <https://doi.org/10.1371/journal.pone.0258790>

Puneeth B. R., Nethravathi P. S., "Bicycle Industry in India and its Challenges – A Case Study", International Journal of Case Studies in Business, IT, and Education (IJCSBE), 5(2), 62-74, ISSN: 2581-6942, Vol. 5, No. 2, August 2021. <https://doi.org/10.47992/IJCSBE.2581.6942.0120>

Vitória Albuquerque, Miguel Sales Dias, Fernando Bacao, "Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review", International Journal of Geo-Information, ISPRS Int. J. Geo-Inf. 2021, 10, 62, February 2021. <https://doi.org/10.3390/ijgi10020062>

Anil Jain, Nirmala Joshi, Anand J Mayee, "Factors motivating buying behavior of female two wheeler users in the district of Palghar", Journal of Management Research and Analysis, October-December, 2020;7(4):154-158, December 2020. <http://doi.org/10.18231/j.jmra.2020.037>

S. Diwakar Raj, Dr. N. Kannan, "Factors Influencing Purchase of Two Wheeler - A Study with Reference to Chennai City", International Journal of Management, 11(12), 2020, pp 2977-2982, ISSN Print: 0976-6502 and ISSN Online: 0976-6510, December 2020. <https://doi.org/10.34218/IJM.11.12.2020.278>

Gyugeun Yoon, Joseph Y.J. Chow, "Unlimited-ride bike-share pass pricing revenue management for casual riders using only public data", International Journal of Transportation Science and Technology 9 (2020) 159–169, January 2020. <https://doi.org/10.1016/j.ijtst.2020.01.003>

Leonardo Caggiani, Rosalia Camporeale, Branka Dimitrijević, Milorad Vidović, "An approach to modeling bike-sharing systems based on spatial equity concept", AIIT 2nd International Congress on Transport Infrastructure and Systems in a changing world (TIS ROMA 2019), 23rd-24th September 2019, Rome, Italy, Elsevier B.V., 2020. <https://doi.org/10.1016/j.trpro.2020.03.006>

Mohammed Hamad Almannaa, "Optimizing Bike Sharing Systems: Dynamic Prediction Using Machine Learning and Statistical Techniques and Rebalancing", DOI: 10.13140/RG.2.2.26034.43202, Thesis for: PhD, Advisor: Hesham Rakha, Project: Bike Research, April 2019. <https://doi.org/10.13140/RG.2.2.26034.43202>

Elisabete Arsenio, Elisabete Arsenio, Sofia Azeredo Lopes, Helena Iglésias Pereira, "Assessing the market potential of electric bicycles and ICT for low carbon school travel: a case study in the Smart City of ÁGUEDA", European Transport Research Review (2018) 10: 13, Springer, January 2018. <https://doi.org/10.1007/s12544-017-0279-z>

Miriam Ricci, "Bike sharing: A review of evidence on impacts and processes of implementation and operation", Research in Transportation Business & Management 15 (2015) 28–38, Elsevier Ltd., April 2015. <https://doi.org/10.1016/j.rtbm.2015.03.003>

Inês Frade, Anabela Ribeiro, "Bicycle sharing systems demand", EWGT2013 – 16th Meeting of the EURO Working Group on Transportation, Procedia - Social and Behavioral Sciences 111 ( 2014 ) 518 – 527, Elsevier Ltd., February 2014. <https://doi.org/10.1016/j.sbspro.2014.01.085>

Darren Buck, Ralph Buehler, Patricia Happ, Bradley Rawls, Payton Chung, Natalie Borecki, "Are Bikeshare Users Different from Regular Cyclists? A First Look at Short-Term Users, Annual Members, and Area Cyclists in the Washington, DC Region", Transportation Research Record Journal of the Transportation Research Board 2387(-1):112-119, DOI: 10.3141/2387-13, December 2013. <https://doi.org/10.3141/2387-13>

Carlos M. Vallez, Mario Castro, David Contreras, "Challenges and Opportunities in Dock-Based Bike-Sharing Rebalancing: A Systematic Review", Sustainability 2021, 13, 1829. https://doi.org/10.3390/su13041829, MDPI, February 2021. <https://doi.org/10.3390/su13041829>

Dissertations

Angela Au, "Social Media Strategies Used in Marketing Custom Bicycle Framebuilding Companies", Doctoral Study - Walden University ScholarWorks (Walden Dissertations and Doctoral Studies Collection), November 2015. <https://scholarworks.waldenu.edu/dissertations/1668>