**SCALE A DATA-DRIVEN STARTUP IN THE SHARED MOBILITY BUSINESS**

Table of Content

1. Executive summary…………………………………………………………………………………………………. 3
2. Introduction to shared bike mobility…………………………………………………………………………5
   1. Market analysis of shared vehicle……………………………………………………………………….5
   2. Regional insight…………………………………………………………………………………………………..6
   3. Need for research analysis on shared mobility…………………………………………………….6
3. Literature review. …………………………………………………………………………………………………….7
   1. Prior studies on rebalancing………………………………………………………………………………..7
   2. Prior studies on demand analysis………………………………………………………………………..8
   3. Addressing the gap in prior studies……………………………………………………………………..9
4. Research methodologies…………………………………………………………………………………………11
   1. Consumer usage analysis………………………………………………………………………………… 11
   2. Daily demand analysis……………………………………………………………………………………….13
   3. Overview of Machine learning models………………………………………………………………13
      1. Random forest……………………………………………………………………………………….14
      2. XGBoost…………………………………………………………………………………………………14
      3. Mixed Integer Linear Programming……………………………………………………….14
   4. Assumptions and Tradeoffs in regression model……………………………………………….15
   5. Ordinary least square………………………………………………………………………………………..15
   6. R-Squared………………………………………………………………………………………………………….16
   7. Adjusted R-Squared…………………………………………………………………………………………..16
   8. RFE and VIF……………………………………………………………………………………………………….17
   9. Influence on P-value and F-Statistics…………………………………………………………………17
   10. Ethical Issues, Commercial sensitivity, and Intellectual property…………………18
5. Data analysis and results…………………………………………………………………………………………19
   1. Shared bikes consumer usage analysis………………………………………………………………19
   2. Daily demand analysis…………………………………………………………………………………….…20
      1. Understanding the dataset…………………………………………………………………….20
      2. Data cleaning and preprocessing……………………………………………………………21
      3. Finding trend and patterns using visualization……………………………………….23
      4. Architecture and RFE outputs: Random Forest………………………………………26
      5. Architecture and RFE outputs: XGBoost…………………………………………………26
      6. Architecture and RFE outputs: MILP………………………………………………………27
      7. Observation. …………………………………………………………………………………………27
      8. Hypothesis testing…………………………………………………………………………………28
      9. Testcase 1………………………………………………………………………………………………30
      10. Testcase 2………………………………………………………………………………………………31
      11. Testcase 3………………………………………………………………………………………………32
      12. Testcase 4………………………………………………………………………………………………33
      13. Result and Validation…………………………………………………………………………….34
6. Discussion and conclusion……………………………………………………………………………………….35
   1. Critical evaluation and future works………………………………………………………………….36
   2. Business recommendations………………………………………………………………………………37
   3. Future scope and ideas……………………………………………………………………………………..38
7. References………………………………………………………………………………………………………………39
8. **Executive Summary**

This executive summary provides a comprehensive overview of the findings and implications derived from an in-depth business analysis aimed at enhancing Lanterne's shared mobility business. The analysis covers critical facets of expanding the shared mobility business, including market segments, customer engagement, messaging strategies, and shared bike usage performance evaluation. The objective of this analysis is to establish a solid foundation for the success of Lanterne not only in its initial phases but also in the long term.

Initially, an extensive review of prior research is conducted, concentrating on two fundamental aspects. The study encompasses the utilization of the most recent rebalancing strategies, ensuring the availability of an appropriate fleet of shared vehicles, encompassing bikes, cars, and scooters, at each docking station. Diverse rebalancing methodologies, including the travelling salesman approach, MILP heuristic approach, employment of genetic algorithms, and other heuristic techniques, are comprehensively explored to provide a profound understanding of the rebalancing strategy landscape. Each study is meticulously categorized based on its strategy model used, approach, implementation process, and ultimately the achieved outcomes.

* Inventory Rebalancing and Vehicle Routing in Bike Sharing Systems frames bike-sharing rebalancing as a scheduling problem, focusing on vehicle routing and inventory management to achieve clustered MIP heuristic outperforms MIP model within 1 minute, yielding 5-15% improvement in solution quality across various vehicle families. (Jasper Schuijbroek, 2017)
* Another research - Barcelona and Seville Region: An Analysis of Bike-Share Rebalancing Strategies employs binary choice and linear regression models to explore rebalancing actions in bike-sharing systems and demonstrates that redistributing capacity to smaller stations enhances usage, indicating positive effects on system dynamics. (Ahmadreza Faghih-Imani, 2017)

Subsequently, the focus of this paper transitions towards an exhaustive exploration of predicting the daily demand for shared vehicles. This report presents an analysis of the shared mobility business, centring on pivotal domains such as customer behaviour, bike usage patterns, and the deployment of predictive models for daily usage demand projection. The primary objective is to furnish insights that serve as navigational beacons for strategic, tactical, and operational determinations, fostering the optimization of business expansion and acknowledging competitiveness within the ever-evolving shared mobility milieu. The latter part of daily demand prediction can be organised into two categories of analysis: (i) Shared mobility consumer usage analysis and (ii) Daily Demand predictive analysis.

**Shared mobile consumer usage analysis.**

The initial phase of analysis delves into consumer behaviour and preferences. Through a comprehensive survey, the report captures vital information regarding the user base's age, purpose of use, preferences, frequency of usage, and valuable feedback for improvement. This analysis provides Lanterne with insights into understanding the diverse needs of their customers, which is crucial for tailoring their services effectively. The results of this analysis showed below results:

**Daily Demand predictive analysis**

The report then shifts its focus to analysing daily bike demand. The dataset, sourced from a trusted open repository, undergoes meticulous cleaning and preprocessing to ensure data quality. Visualization techniques are employed to unearth patterns and trends, enabling a deep understanding of demand dynamics. The data is divided into training, validation, and test sets, paving the way for predictive modelling. Machine learning techniques, including Random Forests, XGBoost, and Mixed Integer Linear Programming (MILP), are harnessed to construct predictive regression models. The performance of each model is recorded and analysed using R-squared and Adjusted R-squared values. These models leverage weather, temporal, and historical data inputs to forecast daily bike demand accurately. The analysis includes a comparison of these models and the identification of the most effective one through recursive feature extraction.

**Model Validation and Optimization**  
The report bridges the gap between theoretical analysis and practical implementation using advanced machine learning techniques. By integrating sophisticated methodologies like Recursive Feature Elimination and Variance Inflation Factor, the report enhances model testing and optimization efficiency. This is particularly valuable in a rapidly evolving real-world scenario. The predictive model's significance is determined through F-Statistics, where a higher value denotes greater stability of the model. The lower VIF and p-value reinforce the model's significance. Coefficients of predictor variables shed light on their impact on bike demand. For instance, features like weather conditions, temperature, and weekdays play pivotal roles in predicting bike demand.

**Recommendations and Implications.**

The final section consists of sets of tiered recommendations: Strategic, Tactical, and Operational. These recommendations are tailored to the distinct decision-making levels, facilitating informed choices for business expansion and optimization.

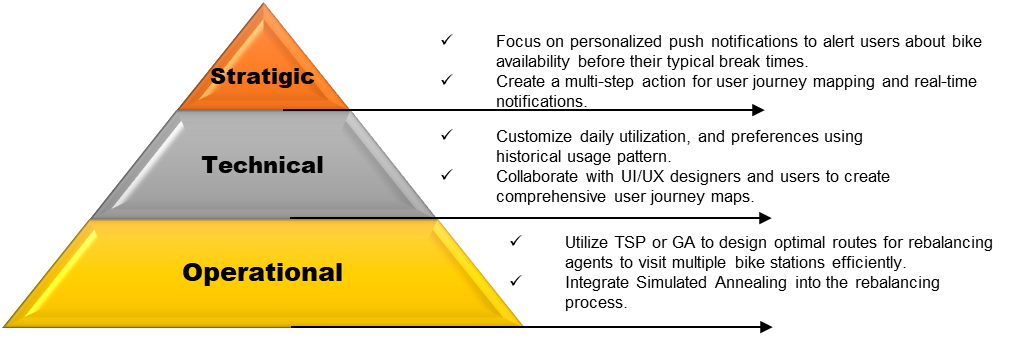


Figure 1: Strategic, Technical and Operational recommendation.

In summary, this comprehensive analysis aims to provide Lanterne with a profound understanding of shared mobility rebalancing methods, customer base, demand dynamics, and predictive modelling strategies.

**2. Introduction to shared bike mobility.**Shared mobility is transforming transportation systems by promoting collaborative use and challenging traditional private ownership of vehicle models. Bike-sharing systems are revolutionizing urban transportation, allowing users to borrow public bikes from designated docking stations and return them to any station after use. This innovative technology uses sensors to capture real-time data on users and their routes, supplying up-to-the-minute information on bike availability through smartphone applications. This data enables dynamic adjustments to the service to meet immediate user needs, while also helping long-term system augmentation by introducing bikes or stations in areas of high demand. Additionally, the data aids in the strategic planning of new bike lanes to cover often-travelled routes, making bike-sharing systems an essential part of modern urban transportation.

A shared mobility system usually consists of lightweight vehicles such as bikes, e-bikes, e-scooters, and ore-mopeds. It provides an environmentally beneficial mode of short-distance travel, such as last-mile transit, and aids in the reduction of city traffic congestion, resulting in a sustainable urban transportation system. (Shaheen S. A., 2019)

Bike sharing typically includes one of three common service models:

1. Station-based bike-sharing system where users can access bicycles through unattended docking stations for one-way service, allowing returns to any station (docking station only).
2. Dockless bike sharing systems where users may check out a bicycle and return it to any location within a predefined geographic region. Dockless bike sharing can include business-to-consumer or peer-to-peer systems enabled through third-party hardware.
3. Hybrid bike sharing systems where users can check out a bicycle from a station and end their trip either returning it to a station or a non-station location or the other way around where users can take any free-floating bike and either return it to a docking station or any free location.

**2.1 Market analysis of shared vehicles**

According to research done by Yahoo Finance, The Shared Vehicles market is expected to generate US$ 4.49 billion by 2027. The number of consumers in the Shared Vehicles market is estimated to reach 5.09 billion by 2027. The predicted user penetration is 61.9% in 2023 and 64.1% by 2027. (Global Bike Sharing Market (2022 to 2027) - Industry Trends, Share, Size, Growth, Opportunity and Forecasts, n.d.) The market is driven by a growing preference for shared vehicles due to their flexibility, convenience, and cost savings. Moreover, shared vehicles offer high-quality, personalized travel options to individuals who do not own a vehicle, which explains the increased usage of this service and drives market expansion. (Statista, 2023) During the COVID-19 pandemic, the shared vehicles market has experienced a significant decrease in demand and usage due to the regulatory norms imposed by the government. This adverse impact has been seen because of the outbreak.

The COVID-19 pandemic had a severe negative impact on the shared vehicles market, as the lockdown was imposed globally to stop the spread of SARS-CoV-2. This led to restrictions on travelling, as well as delays in decision-making by the corporate sector, local partners, and authorities on new agreements. The break and hold on supply chains also further affected the shared vehicles market. (Grand view research, 2020)

**2.2 Regional Insights**

The bike-sharing industry is predicted to experience the highest CAGR of 15.0% from 2022 to 2028. Bike-sharing services involve partnerships between companies and cities, supplying bicycles for public use. The increasing use of eco-friendly and fuel-efficient vehicles, along with government initiatives supporting bike-sharing systems, is contributing to the growth of this sector. During the forecast period, this segment of shared vehicles is expected to experience the fastest growth rate.

(Stellar, 2023)

According to the Ministry of Transport (Mot), China has over 70 bike-sharing firms, with 23 million bicycles and over 400 million riders around the country. Expanding urbanization and a fast-rising population in India and China are assisting industry growth. These countries have implemented a variety of laws to regulate bike-sharing maintenance, operation, and manufacture, allowing service providers to remove defective bikes from the fleet. After the Asia-Pacific region, Europe accounts for over 20% of the global market. Germany, France, and Italy are the leading countries. Other nations, such as the Netherlands, Denmark, and the Nordic countries, exhibit similar characteristics.

**2.3 Need for a research analysis on shared mobility.**

To scale the business, the availability of bikes on daily usage must be improved. However, Bikes cannot naturally attain a relatively balanced distribution, as traffic flow is asymmetrical and user needs are uncertain. The shared mobility system is facing persistent challenges such as the unavailability of bikes or docks for returning them, particularly in stations that have a non-scientific and slow rebalancing process. Consequently, many citizens are reluctant to use the bike-sharing service, which reduces the likelihood of its usage and hinders its growth potential. Improving the bike-sharing service and satisfying the public cannot be achieved solely by expanding facilities. To effectively address the challenge at hand, prioritizing this essential solution is crucial.  
  
The primary objective of this research is to provide a comprehensive analysis of the shared mobility ecosystem, focusing on consumer usage patterns and the prediction of daily demand. By understanding the factors driving shared bike utilization and developing accurate predictive models for demand, this research aims to optimize business expansion strategies, improve user experiences, and enhance operational efficiency.

This research is categorised into 2 categories of analysis:

1. Shared mobility consumer usage analysis.

* What pivotal elements contribute to shared bike utilization, and how does bike usage behaviour evolve across different age cohorts?
* In what manner do customer behaviour trends transform in response to the accessibility of bikes, and how does this influence strategic decision-making?
* How do service providers strategically address insights gleaned from customer analyses to enhance user experiences and operational efficiency?

1. Daily Demand predictive analysis.

* How this dataset propels the formulation of an optimal strategy, unravelling intricate usage nuances and patterns within the bike-sharing ecosystem.
* What is the impact of critical variables such as temperature, day of the week, weather conditions, and humidity on the day-to-day demand for bikes?
* Why is the integration of predictive models imperative, and how do these models react to the change in features?

The possible outcomes from the analysis of consumer usage patterns and predictive demand models can guide the expansion of shared mobility services to areas with higher potential demand. This optimization could lead to increased revenue and market share. Enhanced User Experiences can be attained by tailoring services to match consumer preferences and by ensuring better bike availability, user experiences could be significantly improved, attracting more users, and increasing retention rates. A study on prior research with its practical implementation using statistics and machine learning predictive analytics will help to attain these outcomes.

**3. Literature Review**

**3.1 Prior studies on rebalancing**

The study of bike-sharing mobility has been popular among researchers. (DeMaio, 2009) and (Shaheen, Guzman, & Zhang, 2010) describe the development of bike sharing and trace its beginnings to Amsterdam's first generation of "white bikes" in 1965. Third-generation IT-based systems that appeared after 1995 and were distinguished by the incorporation of cutting-edge technology for bicycle reservations, pickup, drop-off, and information tracking continued the development and identified 4 types of shared mobility analysis: Strategic design, Demand analysis, Service level analysis and Rebalancing operations.

This report however focuses on Demand analysis in two folds.

First, forecasting future demand to meet consumer demand. Secondly, it delves into elucidating factors that contribute to managerial decision-making (Andreas Kaltenbrunner, 2010) predicts the system inventory state and suggests making such information available to users. (Jon Froehlich, 2008) and (Neal Lathia, 2012) identify a temporal demand pattern and forecast the number of rentals. These studies could provide insights that help improve the service level requirements developed in this paper.

(Jenn-Rong Lin a, 2013) present a hub location inventory model for a strategic design challenge for bicycle-sharing systems that incorporates bicycle stock considerations. The research investigates the design work for many components of the bicycle-sharing system, such as the number and position of bicycle stations, the installation of bicycle lanes, the selection of pathways, and so on.

The analysis model incorporates various optimization techniques, including integer linear programming and mixed-integer linear programming, to find the best solutions for these strategic decisions. Demand modelling is addressed using simulation techniques and machine learning algorithms to forecast carsharing demand accurately. Data analytics and statistical methods are utilized to analyse user behaviour, identify factors affecting users' decisions, and assess the social and environmental impacts of car-sharing services. (Masoud Golalikhani, 2021)

The approach involved formulating the e-scooter rebalancing problem as a static rebalancing problem similar to the Traveling Salesman Problem (TSP). The goal was to find the optimal path for a relocation vehicle to efficiently redistribute e-scooters between locations. The genetic algorithm was chosen due to its ability to handle complex optimization problems with multiple constraints and uncertainties.

(Sujae Kim, 2022)

**3.2 Prior studies on the demand analysis**

To seamlessly integrate effective rebalancing strategies into the context of shared mobility systems, a crucial prerequisite involves accurate prediction of the bike requirement for each day. This predictive insight plays a pivotal role in orchestrating optimal rebalancing efforts. This endeavour furnishes a comprehensive tableau of the methodologies employed, spanning from traditional techniques to more contemporary approaches.

|  |  |  |  |
| --- | --- | --- | --- |
| **Research Paper Name** | **Objective / Model** | **Approach** | **Implementation** |
| The economic contribution of public bike-share to the sustainability and efficient functioning of cities.  (Craig Bullock, 2016) | It aims to provide a convenient and efficient mode of transport for residents and commuters by offering a network of bicycles available for short-term rentals at various docking stations across the city. | The approach adopted by the Irish government and Dublin City Council is to create a public-private partnership for the implementation and management of the Dublin bikes scheme. The advertising company J.C. Decaux provided the initial investment for the project, and in return, the company received valuable advertising space within Dublin City. | The scheme's strategic station placements and integration with public transport hubs have made it popular for commuting and sustainable transport. To support the expansion and ensure long-term sustainability, the scheme entered a commercial partnership with Coca-Cola. Furthermore, strategic station placements were made, including at key destinations and public transport hubs, to encourage bike commuting and facilitate seamless integration with other modes of transportation. |
| Prediction of bike sharing system for casual and registered users. (Tae You Kim, 2022) | Continuous model prediction using SVM and SoftMax regression | The research paper takes 2 different approaches to predict the bike-sharing demand. The first approach predicts the daily demand using all features and the second approach classifies them into multiple levels. | Initially, linear regression is used to fit the data and then RMSE is calculated for casual and registered over multiple features such as year, month, and day. Later the SVM is used to predict the bikes only for registered and casual users. |
| Prediction of Bike share demand by machine learning  (Kumari, 2023) | XGBoost formulations for the bike prediction analysis. Autoregressive moving average (ARIMA) to forecast the bike availability at the docking station. Sensitivity analysis is performed to measure the variation in parameters. | The research focuses on enhancing bike-sharing demand forecasting using machine learning techniques, specifically XGBoost and LightGBM. The study leverages the strengths of these algorithms, designed for efficient gradient boosting and tree-based learning, to predict bike demand accurately. Feature selection plays a crucial role, with various features like datetime, weather conditions, and temperature chosen for analysis. However, a few attributes like minute and second are excluded due to a lack of data. Redundant variables, like 'casual' and 'registered,' are also omitted. | The investigation extends to LightGBM, which introduces innovative techniques for overseeing extensive data instances and features. To evaluate model performance, common error metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Root Mean Squared Logarithmic Error (RMSLE) are employed. RMSLE, chosen for its robustness against outliers, is considered the most insightful measure. The chosen evaluation metric enhances the model's ability to manage varying demand levels and outlier scenarios. |
| The spatial-temporal pattern of public bicycle sharing program: the effect of weather and calendar events.  (Corcoran, 2014) | Description of bicycle trips, calendar events and weather. Case study region, Poisson and statical analysis | The outcomes are divided into three parts. To begin, investigate the impact of calendar events and weather conditions to determine the presence of system-wide correlations. The flow map is then used to assess the degree to which various calendar events and weather conditions change the spatial and temporal dynamics of City Cycle utilization at the sub-system level (suburb-to-suburb and station-to-station). | The implementation employs mathematical optimization models for both local and line-haul operations.  These models incorporate distance, charging, and operational costs. The aim is to maximize total benefit while minimizing total cost.  Optimization algorithms are used to solve the models and obtain approximate solutions. |
| Vehicle Rebalancing in A Shared Mobility System with Rider Crowdsourcing  (Ziliang Jin, 2023) | Time-space network with multiple service regions + Flow Balance. | 3PL plus crowdsourcing. The 3PL is more efficient for bulk relocation than rider crowdsourcing, which is better for irregular relocation demands. To service an area, the 3PL often relocates cars in batches from distant, low-demand locations during peak hours, whereas rider crowdsourcing relocates a few vehicles at a time from neighbouring regions throughout the day. | Solution Algorithm: First, we reduce the number of binary variables in the relocation problem (Q). Second, as the entire time-space network G is too large, we develop a temporal decomposition approach by iteratively solving a series of subproblems. Each subproblem considers only a part of the time-space network covering consecutive periods. Third, based on initial vehicle allocations obtained in the second step, we design a heuristic to further improve the solution quality. |

Table 1: Prior studies on Shared mobility demand analysis

**3.3 Addressing gaps in prior studies**

These studies employed standard metrics and mathematical operations to forecast demand for public bike-sharing systems. (Kumari, 2023) employs XG boost and ARIMA on a specific set of memberships-based customers but does not consider optimizing the model through any method. The data set used to develop the model has a significant problem with multi-collinearity, which might lead to model bias. Similarly, (Tae You Kim, 2022) makes use of all features. There are still gaps in the integration of environmental variables, temporal elements, and feature extraction to capture subtle patterns. To address these gaps, hypothesis testing, test cases, recursive feature extraction, and predictive machine learning models for regression analysis are being employed. (Tae You Kim, 2022) While the research effectively employs SVM and SoftMax regression for prediction, it may encounter limitations in terms of model complexity and interpretability.   
Furthermore, it is worth noting that the current research paper primarily focuses on the predictive aspect, often bypassing a comprehensive exploration of potential causative relationships that could underpin bike-sharing demand. This is a significant aspect, as understanding the causal drivers can provide valuable insights into the dynamics that contribute to demand fluctuations.

Addressing this, ROW4 and ROW5 papers present a statistical implementation of demand analysis by incorporating weather-related features. This intricate approach also involves mathematical optimization techniques to derive optimal solutions. While undoubtedly thorough and insightful, it is undeniable that such an approach can be intricate and time-consuming. The formulation of models and optimization processes demands a significant investment of effort and resources.

Moreover, the element of seasonality, encompassing variations in demand across different months, introduces another layer of complexity to overall demand patterns. Different weather conditions, holidays, and cultural events can contribute to distinct patterns of bike-sharing demand throughout the year. This factor necessitates a more nuanced analysis that can capture the interplay between various variables.

Despite the meticulousness of (Corcoran, 2014) and (Ziliang Jin, 2023) 's analytical approaches, it's important to acknowledge that the integration of different types of machine learning predictive models demands considerable computational resources and expertise. This implementation involves data preprocessing, model training, hyperparameter tuning, and thorough testing to optimize the models effectively.

Previous research has predominantly centred around implementation using rigid statical and single models to predict the demand. However, such an approach might not adequately capture the diverse range of factors that influence demand. To effectively address the gaps highlighted in the preceding section and accurately forecast daily demand, it becomes essential to identify the multifaceted drivers of demand. In this context, this report acts as a bridge between earlier isolated analyses and the pragmatic application of various machine-learning models.

By harnessing the power of machine learning techniques such as Random Forest, XGBoost, and Mixed Integer Linear Programming, this paper takes a step further. Additionally, the report introduces a more streamlined approach to testing models, incorporating methods like Recursive Feature Elimination, Variance Inflation Factor, and Ordinary Least Squares. These techniques facilitate efficient testing and optimization of the model within practical time constraints. This pragmatic approach expedites the evaluation of the model's efficacy, particularly in real-world situations where the dynamics of bike-sharing demand are in constant flux.

Additionally, seasonality, including variations in demand between months, can affect overall demand patterns. Thus, this approach provides a more comprehensive and accurate method for predicting demand in a new region on a particular day. Each of these sections must be considered individually as explained below.

1. Weather Data Integration:

Including historical weather data such as season and temperature is crucial in developing accurate weather-based demand models for bikes. Test cases can validate the relationships between temperature, precipitation, wind speed and bike demand, ensuring model robustness.

1. Temporal Factors:

Temporal factors such as weekdays, weekends, and business cycles strongly affect demand patterns. The statistical validation of their influence may be accomplished through hypothesis testing. The model's capacity to capture differences across diverse temporal contexts may be validated using test scenarios.

1. Feature Engineering:

Creating composite features that combine weather and temporal elements, like "rainy weekday" or "sunny weekend", is an effective way to capture nuanced demand. Hypothesis testing can assess the significance of these new features.

1. Machine Learning Models:

Harnessing machine learning potential, Random Forests, MILP, and Gradient Boosting are used to build predictive regression models. These models anticipate daily demand by synthesizing multifaceted inputs such as weather, temporal, and historical data. Hypothesis testing validates the suitability of each model for the task. Test cases rigorously examine their efficacy under diverse scenarios.

1. Evaluation, Optimization and Continuous Learning:

Training and testing sets are created from historical data to evaluate machine learning models for predictive accuracy. Continuous learning involves updating models with Influence on F-statistic, Influence on P-values, and Variance Inflation Factor (VIF) to remain attuned to evolving demand patterns and seasonal shifts.

**4.Research methods.**

The report performs 2 analyses in this research. (Wintjen, 2020)

* 1. **Shared bikes consumer usage analysis.**

At first, the consumer data is collected via a survey. The primary aim of conducting of first analysis is to capture a holistic understanding of the user base's preferences, requirements, and expectations. The survey aims to provide insight into elements like age, purpose of use, reasons for choosing shared mobility, frequency of weekly usage, and insightful comments or feedback for development.

Figure 2: Methodology of Shared bike consumer usage.

**4.2 Daily usage demand analysis.**

1. The data is collected from a trusted open-source repository, and subsequently subjected to rigorous cleaning and preprocessing procedures to ensure data quality and consistency. Before delving into statistical analysis, the data undergoes meticulous preparation.
2. Patterns and trends are uncovered using a range of visualization techniques, elucidating insights through various graphical representations. This visual exploration enables a comprehensive understanding of the data's underlying dynamics.
3. After the data has been pre-processed and formatted correctly, it is divided into three sets: Train, Validate, and Test datasets.
4. Transitioning seamlessly to predictive modelling, a variety of state-of-the-art machine learning algorithms are employed to forecast daily demand, facilitating demand analysis. This section covers three models: Random Forest, XgBoost, and Mixed integer linear programming. The model that performs the best is chosen and optimized using recursive feature extraction to predict the daily count of shared vehicles.
5. In pursuit of model optimization, an experimental methodology is employed. The intricacies of Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) are explored with precision, meticulously evaluating and refining models for enhanced performance. This optimization process is effectively communicated through a comprehensive OLS, encompassing P-value and F-statistics, which collectively illustrate the intricate relationships between these critical components.

The following diagram illustrates the step-by-step flow from data collection to the final predictive model.

A diagram of a data processing process

Description automatically generated

Figure3: Methodology of demand analysis.

**4.3 Overview of machine learning models**

**4.3.1 Random Forest.**

The Random Forest is a type of classifier in supervised machine learning. It constructs numerous decision trees with identical nodes, using data and partitioning conditions to create distinct leaves. These decision trees are built top-down, using node splits that minimize entropy in each sub-branch. As the bike-sharing problem requires regression, Mean Squared Error (MSE) is used to solve regression problems to see how data branches from each node. The MSE calculates the distance of each node from the predicted actual value to decide which branch is the better decision for the forest. (UNIVERSITY OF WISCONSIN–MADISON)

The algorithm of Random Forest works as follows –  
A screenshot of a computer program

Description automatically generated

Figure 4: Random forest algorithm

For each decision tree in the forest, a bootstrap sample is selected from S where S(I) denotes the ith bootstrap. The algorithm is then modified as follows: At each node of the decision tree, some subset of features (f) is selected from the set of features F where f ⊆ F and maximises the reduction in entropy. The node then splits based on the best feature in f rather than F (f is much smaller than F). Select the features (f) with maximum reduction in entropy in each resulting new node. The learning speed of the tree can be increased by narrowing the set of features. The above steps are repeated for each sub-node until all data is labelled at the leaf node or until the entropy of the leaf node is below some threshold.

Random forest follows bootstrap-aggregating (bagging and pasting) to create multiple different subsets of data for training the decision tree.

Bagging: The features are randomly sampled with a different subset of data and the decision tree is trained on each subset. Additionally, when sampling is performed without replacement, it is called Pasting. The resulting decision trees are then combined to form the final random forest model.

Thus, ensembling multiple models and then combining their predictions to make a final prediction helps to avoid overfitting and increases the accuracy of the classifier.  
This feature when applied on weather datasets such as Temperature, Humidity, windspeed etc., the bike requirement pattern can be understood in a more detailed way.  
To perform demand analysis, the model introduces a meticulously crafted grid of hyperparameters (param\_grid). Hyperparameters, including n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf, and max\_features, each hold a unique role in shaping the model's behaviour.

**4.3.2 XGBoost**

XGBoost is a distributed gradient boosting library that is built for optimal efficiency, flexibility, and portability. It utilizes machine learning algorithms within the Gradient Boosting framework, specifically offering a parallel tree boosting method (known as GBDT or GBM). This approach resolves a wide range of data science issues rapidly and accurately. It is compatible with various distributed environments (e.g., Hadoop, SGE, MPI), and can oversee problems with billions of examples.Boosting is a technique used in machine learning where multiple weak learners are combined to create a strong learner. Weak learners are simple models that perform slightly better than random chance, and in XGBoost, they are usually shallow decision trees to avoid overfitting.

XGBoost builds trees sequentially through gradient boosting, where each tree tries to correct the errors made by the previous ones. It calculates the gradient of the loss function concerning the predictions of the previous trees and then fits a new tree to the negative gradient of the loss, effectively minimizing the loss function.

To control the complexity of the individual trees and the entire ensemble and prevent overfitting, XGBoost introduces regularization terms. The two main types of regularization in XGBoost are L1 regularization (Lasso) and L2 regularization (Ridge). The objective function is a crucial part of XGBoost as it represents the loss function that the algorithm tries to minimize during training. For regression tasks, the most common objective function is the mean squared error (MSE), which measures the average squared difference between predicted and actual values. (Geeksforgeeks, 2023)

While building the model, RandomizedSearchCV is a helpful guide that conducts a randomized search across various configurations in the hyperparameter grid. The number of configurations evaluated is determined by the n\_iter parameter, ensuring a comprehensive exploration of the parameter space. The chosen metric, 'neg\_mean\_squared\_error', guides the search towards configurations that minimize the squared error. The use of cross-validation (cv=5), verbose logging, and parallel processing (n\_jobs=-1) further enhance the thoroughness of this exploration.

**4.3.3 Mixed Linear Integer Programming (MLIP).**

Linear regression is a crucial supervised ML algorithm used to model the relationship between a dependent variable (target variable) and one or more independent variables (or features). The objective of linear regression is to discover the best-fitting linear relationship that can predict the dependent variable based on the values of the independent variables.

Below is a step-by-step breakdown of the linear regression algorithm:

Model Representation: Linear regression assumes that the relationship between the independent variables (X) and the dependent variable (y) can be represented by a linear equation of the form:

Y = b0 + b1 \* x1 + b2 \* x2 + ... + bn \* xn

Y : Dépendent variable

x1, x2, ..., xn: Independent variables

b0, b1, b2, ..., bn: Coefficients (parameters) that need to be estimated.

Cost Function: The objective is to find the coefficients (b0, b1, b2, ..., bn) that minimize the difference between the predicted values (obtained from the linear equation) and the actual values of the dependent variable. The Mean Squared Error (MSE) is the typical cost function for linear regression. The calculation determines the average squared difference between the values that were predicted and the actual values.

Model Evaluation: After obtaining the optimal coefficients, you can use the linear equation to make predictions for new data points. The quality of the model is assessed using evaluation metrics like the coefficient of determination (R-squared), which measures how well the model fits the data. (Jolly, 2018)

**4.4 Assumptions/ Trade-offs of Linear Regression:**

1. Linearity: The relationship between the independent and dependent variables is assumed to be linear.
2. Independence of Errors: The errors (residuals) between predicted and actual values are assumed to be independent of each other.
3. Homoscedasticity: This refers to the condition where the variance of errors is constant at all independent variable levels.
4. Normality of Errors: This model assumes that the errors are normally distributed.

**4.5 Ordinary least squares**:

OLS regression analysis is a highly popular statistical technique for modelling the correlation between one or more independent variables and a dependent variable. The primary objective of OLS is to identify the line that reduces the sum of squared differences between predicted and observed values to a minimum. OLS regression is based on the principle of fitting a linear equation to the observed data points. (Ivezić, 2019) The general form of a linear regression equation is:

If a model has p explanatory variables, the OLS regression model is expressed as:

Y = β0 + Σj=1...p βjXj + ε

In this equation, Y is the variable that depends on other factors, β0 represents the starting point, and Xj is the factor that explains the change in Y (with j ranging from 1 to p). The unpredictable variability is represented by " ε " and has an average value of 0 and a standard deviation of σ.

(Lumivero, 2023) Concerning the analysis of daily demand for shared bikes,   
Y represents the predicted count of bikes that would be required on a particular day.  
β0 represents the constant or starting point.

βjXj represents multiple feature coefficients and associated values.

Assessing the goodness of fit of OLS regression models requires fundamental measures such as R-squared and Adjusted R-squared.

**4.6 Coefficient of determination (R-squared)**

R-squared measures the accuracy of a model by calculating how much of the dependent variable's variability can be explained by the independent variables. The value of R-squared ranges between 0 and 1 and the value closer to 1 indicates a better fit. (Tuffery, 2011) While a higher R-squared may suggest a better explanation for the dependent variable's variance, it does not necessarily ensure model validity and overfitting must be avoided.

A math equations on a white background

Description automatically generated

Figure 5: R-Squared

**4.7 Adjusted R-squared**

The Adjusted R-squared formula is an improved version of R-squared that accounts for the complexity of the model and penalizes the use of unnecessary predictors. This measure provides a more accurate evaluation of the goodness of fit for a model.

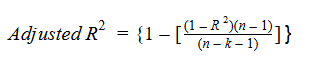


Figure 6: Adjusted R- Squared

Here,

* The variable 'n' represents the total number of data points.
* The variable 'k' represents the number of independent variables,
* 'R' represents the R-squared values determined by the model.

If the R-squared value does not significantly increase upon the addition of a new independent variable, then the Adjusted R-squared value will decrease.

When assessing how well a model fits the data, the Adjusted R-squared metric is useful in determining if the addition of a new predictor has significantly improved the fit beyond what could be due to chance. It also helps prevent overfitting by penalizing the inclusion of irrelevant variables. This metric is particularly useful for comparing models with different numbers of predictors, as it provides a fairer evaluation by balancing model fit with complexity. Regression analysis is an extremely powerful statistical tool for modelling relationships between variables.

To achieve optimal results, possessing a strong balance between Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) is crucial. These techniques play a critical role in selecting the appropriate variables and conducting accurate model assessments, which directly affect vital statistical metrics such as p-values and F-statistics (Rutecki, 2022).

**4.8 Recursive Feature Elimination (RFE)**

RFE aims to identify the key independent variables that greatly enhance the model's predictive ability. The aim is to choose features by repeatedly considering progressively smaller sets of features. technique works by iteratively fitting the model and eliminating less significant variables based on criteria such as p-values or coefficients. By progressively removing fewer features, RFE simplifies the model, minimizes overfitting, and often leads to better interpretability.

**Variance Inflation Factor (VIF)**

The VIF is essential for identifying multicollinearity in a regression model. Multicollinearity can lead to unpredictable coefficient estimates and inflated standard errors when predictor variables are highly correlated. VIF measures the extent of multicollinearity, making it an invaluable tool. Higher VIF values show more significant multicollinearity. Correlation between the variable and the other predictors potentially leads to issues in interpreting the effect of that variable.

**4.9 Influence on P-values**:

To effectively determine the statistical significance of individual predictors in a regression model, one must look at p-values. However, these values can be influenced by RFE, which removes variables that do not contribute to the model's predictive power. By eliminating less important variables, the remaining predictors can have more noticeable effects, which can lead to lower p-values. Therefore, RFE can enhance the identification of statistically significant predictors by removing unnecessary variables and noise.

The precision of coefficient estimates is affected by VIF, which in turn impacts p-values. High VIF values indicate the presence of multicollinearity, resulting in larger standard errors of coefficient estimates. This leads to wider confidence intervals and lower t-values, resulting in less significant p-values. By identifying variables with high multicollinearity, VIF can aid researchers in addressing this issue and improving the accuracy of p-values.

**Influence on F-statistic**

The F-statistic is used to evaluate the significance of a regression model by comparing it to a null model with no predictors. RFE can affect the F-statistic by eliminating less important variables, resulting in a model with fewer predictors. Although a more concise model could have a lower F-statistic, the F-statistic may remain significant or even increase if the decrease in predictors enhances model fit and explanatory power.

The VIF detects multicollinearity, affecting the F-statistic. High VIF values mean more multicollinearity, which can make interpreting coefficient estimates difficult and reduce the significance of the model. To improve the F-statistics stability, address multicollinearity by transforming or removing variables.

RFE helps select relevant predictors, improving model interpretability and positively influencing p-values and F-statistics. VIF identifies multicollinearity issues, which negatively impact p-values and F-statistics. Understanding these concepts is crucial for reliable regression analyses.

**4.10 Ethical Issues, Commercial Sensitivity, and Intellectual Property**

The intricately braided ethical, business, and intellectual property strands in the data analysis endeavour reflect the complex landscape of contemporary data investigation. In the age of technology, ethical data exploration is supported by a constant vigilance for privacy, a steadfast determination to eliminate biases, and a commitment to openness. It takes judgment to strike the correct balance between information sharing and corporate opacity. Nevertheless, maintaining an environment that values intellectual property is essential for sustaining innovation. The research in issue, distinguished by its strict methodology and ethical awareness, is a perfect example of the complex interplay between the obligations and problems of contemporary data analysis.

Ethical Considerations

This first step avoids the minefield of contaminated or biased data and emphasizes a commitment to honesty and openness. The collection of personal data such as age, education, occupation, and feedback necessitate strict adherence to data protection regulations and guidelines. It's crucial to ensure that all data collected is anonymized and aggregated, preventing any possibility of re-identification of individuals. Secondly, even though the data for demand analysis is from trustworthy sources, privacy and consent issues for users still raise ethical concerns. The analysis using these data cannot be perfect. Furthermore, as the research focuses on user behaviour prediction, the prevention of unintentional biases that can support inequity becomes crucial. (Cote, 2021)

Commercial Sensitivity

The project's findings have consequences that go beyond academia and may change the way that businesses operate. Shared mobility businesses may benefit from identifying user activity trends, but doing so also raises issues of commercial sensitivity. Selecting which discoveries to present becomes a complex dance. While sharing insights encourages the spread of information, it also has the potential to provide rivals with a strategic advantage.

Intellectual property

The creation of prediction models unfolds intellectual property issues inside the complex tapestry of data analysis. The use of cutting-edge machine learning techniques like Mixed Integer Linear Programming, Random Forest, and XGBoost is a prime example of innovation. In addition, the strategies used, such as recursive feature extraction and optimization algorithms, represent original intellectual work. These approaches' originality is protected by documentation and possible patenting.

**5. Data Analysis and Result:**

**5.1 Shared Bikes consumer usage analysis.**

A thorough survey of a varied group of people who actively use shared bikes and scooters was conducted as a part of this project. These users have been using shared bike and scooter services for 0 to 12 months. Our survey gathered data from a wide range of individuals, including students and working professionals aged 14 to 48. The survey aimed to gain profound insights into user behaviour patterns, exploring their purposes and intentions for using shared bikes and scooters. Additionally, it sought to delve into the intricate demographics of the customer base associated with shared bikes, unravelling a deeper understanding of the underlying characteristics of this user group. The aim is to find a pattern to enhance user experience, improve services, and align with user preferences and expectations by focusing on key points (Claire Chung, 2018). Below are the key takeaways from the survey:

A graph showing different colored bars

Description automatically generatedFigure 7: Factors behind using shared vehicles.

Few major factors that influence people when choosing shared bikes as a mode of transport are cost efficiency, convenience, and ease of exploring the city, while also reducing the carbon footprint.

A graph of different colored bars

Description automatically generatedFigure 8: Number of users taking shared bikes per week vs. age group.

Shared vehicles are popular among teenagers and young workers due to their affordability and faster mode for small distances. However, this trend is not common among those over 30, who prefer a more comfortable mode of transportation.

A pie chart with different colored numbers

Description automatically generated

As per feedback, it has been observed that a significant percentage of users, 23.9% to be precise, strongly feel that the availability of bikes should improve. This is closely followed by the importance of customer service and pricing.

Figure 9: Feedback received for improvement.

A graph with green dots

Description automatically generated

To effectively improve bike availability, it is important to consider feedback from customers who have been using shared bikes for 3-12 months per week as they have been using more frequently than the new users.

Figure 10: Frequency of users who advised improving the availability of bikes

The survey results unequivocally establish that access to shared vehicles is an essential determinant of the utilization of shared bike services. The mere existence of accessible bikes significantly influences prospective user’s adoption of these systems. Furthermore, the survey unequivocally demonstrates that availability is not the only crucial factor to contemplate. The availability of bikes must efficiently complete the daily demand of users so use the resources optimally.

**5.2 Daily demand analysis.**

**5.2.1 Understanding the Dataset.**

The dataset is collected from the UCI machine learning repository. (Fanaee-T, 2013) The increasing popularity of bike-sharing systems has sparked interest in understanding the dynamics of bike rentals. The dataset, comprising various features, presents an opportunity to gain insights into bike rental behaviour, usage patterns, and environmental factors' impacts.

The dataset at hand encompasses various attributes, each contributing to the understanding of bike rental trends. Here's an overview of the significant features:

|  |  |
| --- | --- |
| **Feature** | **Explanation** |
| Instant | A record index indicating the position of the entry within the dataset. |
| Date | The date of the data recording. |
| Season | Represents the season (1: winter, 2: spring, 3: summer, 4: fall) |
| Year | The year in which the data was recorded, with values 0 and 1 corresponding to 2011 and 2012, respectively. |
| Month | Ranges from 1 to 12, representing the month of the year. |
| Hour | Represents the hour of the day, ranging from 0 to 23. |
| Holiday | Indicates whether the day is a holiday or not. |
| Weekday | Represents the day of the week. |
| Working day | A binary variable indicates whether the day is neither a weekend nor a holiday. |
| Weather sits | 1: Clear, Few clouds, partly cloudy, partly cloudy  2: Few clouds+ Mist, Mist + Cloudy, Broken clouds+ Mist  3: Light Rain + Scattered clouds +Thunderstorm +, Light Snow, Scattered clouds + Light Rain.  4: Heavy Rain + Mist, Snow + Fog + Ice Pallets + Thunderstorm |
| Temperature | The normalized temperature is Celsius, considering a range from -8 to +39 degrees. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) |
| Feels-like Temperature | Normalized feeling temperature, considering the perceived temperature. The values are derived (only in hourly scale) via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 |
| Humidity | Normalized humidity, is expressed as a fraction of the maximum value. |
| Windspeed | Normalized wind speed, scaled to a fraction of the maximum value. |
| Casual Users | The count of casual users renting bikes. |
| Registered Users | The count of registered users renting bikes. |
| Total Count | The sum of casual and registered users' counts, represents the total number of bike rentals. |

Table 2: Dependent and independent variables present in the dataset.

The value of data points comprehends the contents of the dataset by giving succinct and unambiguous descriptions of each variable. Through data visualization and analysis, a plethora of information is available in this dataset, which may be examined.

Temporal Trends: Identifying the peaks and the falls of rental services on a day and various seasons by graphing rental counts against hours, days, months, and seasons.

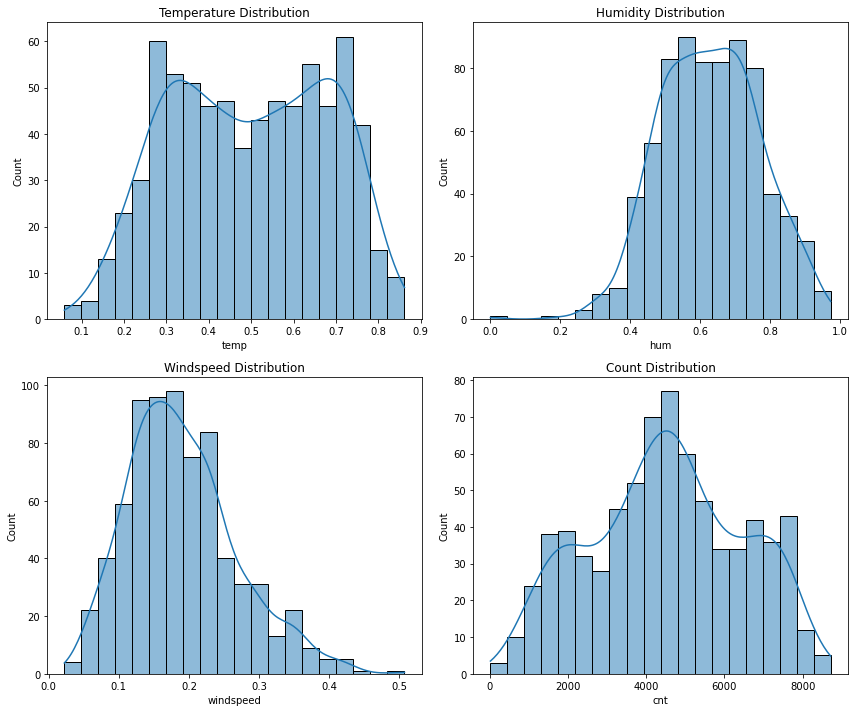
Weather Impact: Performing analysis and understanding how the weather affects the consumer’s choice renting of bikes. Do more people rent bikes on sunny days than on days when it is raining?

**5.2.2 Data cleaning and pre-processing followed by Data analysis.**

Data cleaning and preparation are essential for any data analysis or machine learning models. The accuracy and reliability of subsequent analysis and modelling are heavily reliant on proper cleaning and preparation of the dataset. This section presents a comprehensive code that addresses various aspects of data cleaning, including managing outliers, filling in missing values, converting data types, and scaling features. (Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython., 2017)

1. The process begins by understanding an overview of the dataset, including column names, data types, and non-null counts.
2. To ensure accurate analysis and modelling, it’s important to address missing values in a dataset. In this case, the columns ‘temp’, ‘atemp’, ‘hum’, and ‘windspeed’ had missing values that were handled by replacing them with the mean values of their respective columns. This imputation technique preserves the integrity of the data and prevents the loss of entire rows due to missing values.

The below figure contains a 2x2 grid of subplots, each containing a histogram and KDE plot for variables ‘temp’ (temperature), ‘hum’ (humidity) and ‘windspeed’. The histograms shows that usage of bike might be overly dependent on temperature, higher humidity and low windspeed might increase the bike usage.



A graph of a graph

Description automatically generated with medium confidence

Figure 11: Temperature, humidity and Windspeed distribution in the dataset

1. After imputing missing values, the dataset’s information again shows that the missing value issue has been resolved. This demonstrates transparency and ensures that the data is ready for further analysis.
2. When conducting statistical analyses and creating models, it’s important to be aware of outliers as they can skew results. To identify potential outliers in the numeric columns (‘temp’, ‘atemp’, ‘hum’, ‘windspeed’, ‘casual’, ‘registered’, and ‘cnt’), the code utilizes box plots. These plots provide a visual representation of the data’s distribution and make it easier to identify any extreme values. Outliers are effectively managed by the code, which employs a standard method of capping values that are more than 3 standard deviations away from the mean.   
   This approach ensures that extreme values are brought within a reasonable range without entirely removing them. The process involves iterating through the numeric columns, calculating the mean and standard deviation, and defining the upper and lower bounds for capping. After addressing the outliers, the code confidently displays the dataset’s updated information, demonstrating its ability to handle extreme values effectively. (Python data analysis perform data collection, data processing, wrangling, visualization, and more using python, third edition, 2021)
3. Categorical variables are efficiently transformed into binary columns using one-hot encoding in the code. The get\_dummies function is employed with drop\_first=True to create binary columns for each category while eliminating the original categorical columns. This approach effectively prepares categorical variables for use in machine learning algorithms.
4. Rescaling Min-Max Scaling: In datasets where features have varying scales, the differences in magnitude among variables can lead to biased results and inefficient model performance. Rescaling aims to mitigate these issues.   
   Also known as normalization, this method scales features to a specific range, often between 0 and 1. The formula for min-max scaling is:

X(new) = x-min(x) / max(x)-min(x)

**5.2.3 Finding Trends and patterns using Data visualization.**

Data in a structured format provides crucial insights and facts that can help make informed decisions. However, the complexity of raw data can make it difficult to understand. That’s where data visualization comes in – transforming complicated data into visually appealing representations that are easy to comprehend and analyse. In the realm of shared bike datasets, data visualization is particularly crucial as it provides valuable information about bike rentals over time. This section provides a few trends and patterns of shared bike services into the significance of data visualization. (Michael Waskom., 2022)

**A pie chart with numbers and a black background

Description automatically generated**

This visual representation in the form of a pie chart illustrates bike rentals according to the different seasons. The chart displays how bike rentals were distributed across four seasons - Winter, Spring, Summer, and Fall. In both 2011 and 2012, the highest percentage of bikes were rented during Summer (32.2%) and the lowest during Winter (14.3%). This outcome is logical because summer is a popular time for tourism, while in winter, people may opt for cabs, buses, or trains due to the cold weather.

Figure 12: Shared bike usage across seasons.

**A screen shot of a computer screen

Description automatically generated**

Understanding the correlation between temperature and bike rental counts is of great importance. This analysis can reveal patterns in bike rental behaviour based on temperature changes. For instance, it indicates that individuals are more likely to rent bikes during warmer temperatures, particularly in the spring.

Figure 13: Split of temperature V/s count of bikes.

**A graph of red and blue bars

Description automatically generated**

In 2012, there were more rentals than in 2011. By studying the correlation between bike rentals, months, and years, we can identify patterns that can assist in allocating resources and designing effective marketing strategies. September had the highest number of rentals, closely followed by August.

Figure 14: Comparison between usage in 2011 and 2012

**A colorful rectangles and rectangles

Description automatically generated**

In the second visual, the focus shifts to the impact of weather situations on bike rentals. This visualization aids in understanding how varying weather conditions influence bike rental demand, helping bike-sharing services prepare for changes in weather. The most favourable weather condition is Clear, Few clouds and partly cloudy.

Figure 15: Bike counts in various weather situations

**A screenshot of a computer screen

Description automatically generated**

Figure 16: Correlation matrix of features the in dataset.

This heat map helps analysts pinpoint which variables have significant positive or negative correlations, allowing them to identify potential factors that influence bike rentals. The scale on the right indicates the strength of the relationship. Factors such as year, month, temperature, feel-temperature, and registered subscribers play a major role in bike rentals. Conversely, wind speed, humidity, weather, and holidays decrease the number of bike rentals. (Yim, 2018)

**The architecture of predictive model and RFE outputs**

After the data has been pre-processed and formatted correctly, it is divided into three sets: Train, Validate, and Test datasets. These datasets are used to create a regression model that can be used for recursive feature extraction. This section covers three models: Random Forest, XgBoost, and Linear Regression. The model that performs the best is chosen and optimized to predict the daily count of shared vehicles.

**5.2.4 Model 1: Random Forest**

As mentioned in the previous section, Random Forest's significance comes from the combination of many decision trees that were all trained on different data subsets.

The behaviour of the model is influenced by a variety of hyperparameters, including n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf, and max\_features.  
param\_grid = {

'n\_estimators': [50, 100, 150, 200],

'max\_depth': [None, 10, 20, 30],

'Min\_samples\_split': [2, 5, 10],

'Min\_samples\_leaf': [1, 2, 4],

'Max\_features': ['auto', 'sqrt']

With the creation of Randomized Search CV object comes into existence. This method's elegance lies in its randomized search of the hyperparameter space, exploring various configurations and assessing them through cross-validation.

As the RandomizedSearchCV object undergoes training on the provided dataset (X\_train and y\_train), its quest is to uncover the golden combination of hyperparameters that best minimize the negative mean squared error. (scikit-learn developers, 2023)

Recursive Feature Extraction Output using Random Forest:  
All features in the model are crucial in making accurate predictions, even seemingly insignificant ones. However, "temp," "yr," and "atemp" reign supreme with importance scores above 0.2. "Temp" is the most influential feature with an importance score of 0.3706, showing its close correlation with the target variable and the model's confidence in making accurate predictions.

**5.2.5 Model 2: XGBoost**

The goal of regression-based prediction is to minimize the squared error, sometimes known as the "reg: squared error." An ensemble of decision trees is used by the well-known method XGBoost to augment its gradient-boosting abilities and improve prediction accuracy.

RandomizedSearchCV is a handy tool that does a randomized search across multiple configurations in the hyperparameter grid throughout the model development process. The n\_iter parameter controls the number of configurations that are assessed, ensuring a thorough examination of the parameter space. The 'neg\_mean\_squared\_error' measure is used to direct the search to configurations that minimize the squared error. This investigation's depth is further increased using cross-validation (cv=5), verbose logging, and parallel processing (n\_jobs=-1). (scikit-learn developers, 2023)

Recursive Feature Extraction Output using XgBoost.

Certain features contribute less to the model's predictions, including "weekday\_6," "holiday," "windspeed," "mnth\_4," and "mnth\_9." Others, like "weathersit\_2," "hum," "mnth\_5," and "mnth\_11," have a more significant impact. The top features are "weathersit\_3," "atemp," "season\_3," "season\_4," "temp," and "yr." "Yr" and "temp" are especially important. High VIF values for "atemp," "temp," "yr," and "hum" suggest possible multicollinearity issues.

**5.2.6 Model 3: MILP**

The architecture of this Linear Regression model epitomizes the fusion of data analysis and algorithmic prowess. Each step is a building block, contributing to a model that does more than predict; it interprets, empowers, and elucidates the intricate narratives concealed within data. The model delves into the complexities of multicollinearity through Variance Inflation Factors (VIF). These reveal relationships between features, identifying potential redundancies or dependencies. (scikit-learn developers, 2023)

Recursive Feature Extraction Output using Linear regression.

The magnitudes of the coefficients serve as a yardstick for gauging a feature's importance. Features with higher absolute coefficients wield more influence on predictions. In this context, "temp" emerges as a powerful predictor with a coefficient of 0.2768, followed closely by "yr" with 0.2302. These variables appear to have significant sway over the model's output. Meanwhile, "weathersit\_3" and "hum," with coefficients of -0.2358 and -0.1466 respectively, appear as influential factors with negative impacts.

**5.2.7 Observation:**

In all three models, the Variance Inflation Factors for a few features such as temp, atemp and the year have values higher than 5. Multicollinearity is a problem in statistical analysis and regression modelling because it can lead to several significant issues that affect the reliability, interpretation, and stability of the model's results.

🡪 Multicollinearity causes the standard errors of the regression coefficients to become inflated.  
  
🡪 Multicollinearity makes the coefficient estimates sensitive to small changes in the data, leading to instability. This can result in unstable model performance when the same model is applied to different datasets or slightly different versions of the same dataset.

🡪 The interpretation of each coefficient becomes less clear, as it is difficult to determine the unique contribution of each predictor.  
  
🡪 Multicollinearity can lead to overfitting, where the model captures noise in the data instead of true underlying patterns. This results in poor generalization to new, unseen data.

To eliminate the process of multicollinearity this section performs Feature Selection, Variance Inflation Factor (VIF) Threshold and maintains a low P-value by passing multiple test cases. These test cases will perform Feature importance and remove the features that produce noise in the model.

**5.2.8 Hypothesis Testing**

In regression modelling, the null hypothesis states that one or more features have VIF values above a certain threshold to avoid unreliable regression estimates caused by multicollinearity.

The provided null hypothesis (Ho) and alternative hypothesis (HA) pertain to conditions related to the Variance Inflation Factor (VIF) and p-values for feature variables. The hypothesis testing aims to optimize the F-statistics. Let's frame the context of these hypotheses and the aim:

**Null Hypothesis (Ho):**

Ho: At least one feature with VIF > 5 or at least one feature with a p-value > 0.5.

**Alternative Hypothesis (HA):**

HA: The VIF value of all the features is less than 5, and the p-value of all the features is less than 0.5.

**Aim:** Optimize F – Statistics value.

The OLS regression technique to determine the correlation between independent and dependent variables. Its goal is to establish the most suitable linear relationship between the independent variables and dependent variables while minimizing the total squared differences between the observed values and the values predicted by the regression model. Below is the status of VIF and P the value of all the features.  
Upon obtaining the initial Ordinary Least Squares (OLS) results, the Variance Inflation Factor (VIF) values for all the features have been assessed. The observations from this analysis are as follows:

(i) Notably, all features highlighted in red have VIF values exceeding the established threshold. The VIF values for 'temp' and 'atemp' are strikingly elevated. This anomaly suggests a potential bias towards these variables. Additionally, several other features display VIF values surpassing the threshold (VIF > 5).

(ii) After scrutiny, it has been observed that the P-values for 9 features exceed the significance level of 0.005. This outcome implies that the current model struggles to adequately capture the variability inherent within these features.

(iii) The F-statistics value is 105.7 is deemed relatively low. As per established standards, this value signifies that the model is not optimally fitting the data. Consequently, it becomes imperative to embark on a process of model optimization.



Figure 17: VIF values of all features.

A screenshot of a computer

Description automatically generated

Figure 18: OLS Regression results

**5.2.9 Test Case 1:**

* Removing all the features that have high P-Value (>0.5).
* "mnth\_4","mnth\_6","mnth\_7","mnth\_11","weekday\_1","weekday\_2","weekday\_3","weekday\_4".

Test case 1 Observations:

Figure 19: VIF values and Test Case 1 Observations

(i) Post the initial feature extraction, it is noteworthy that the VIF values for 'temp' (55.39), 'atemp' (51.37), and 'season\_3' (6.79) continue to remain elevated. This observation hints at the potential presence of bias associated with these variables.

(ii) The P-values for the 'mnth\_2' and 'mnth\_12' features surpass the conventional significance level of 0.05.

(iii) Having successfully cleared the first test case, the F-statistics value has shown improvement to 142.6. Although comparatively higher, this value does not attain the desired level of statistical robustness.

By established benchmarks, this finding points to the model's suboptimal fitting of the dataset. These results necessitate the need for a more refined model optimization approach to achieve an enhanced fit to the data.

**5.2.10 Test Case 2:**

* Removing all the features that have high VIF (>5). - **"temp"**

The terms "temperature" and "feels like temperature" are synonymous in this context. Given that the variable "tempt" exhibits a higher Variance Inflation Factor (VIF), this test case aims to alleviate multicollinearity by excluding the "temperature" feature. However, it's worth noting that in real-world scenarios, temperature holds a significant influence over consumers' decisions regarding shared mobility. As a result, the decision has been made to retain the "feels like temperature" ('atemp') variable for consideration, while removing the "temperature" ('temp') variable from the analysis.

Test case 2 Observations:



Figure 20: VIF values and Test Case 2 Observations

(i) The VIF of 'season\_3' is still higher than the threshold and the rest of the features VIF is less than 5.

(ii) P-values for "holiday," "mnth\_2," "mnth\_12," and "weekday\_5" continue to be greater than the predetermined threshold, indicating ongoing difficulties with the significance of these variables.

(iii) The F-statistics value showed a 5-point improvement to 147.6.

More test cases must be used since test case 2's results are unsatisfactory.

**5.2.11 Test Case 3:**

* Removing all the features that have high P-Value (>0.5)
* **"mnth\_12","mnth\_2","holiday","weekday\_5"**

Test case 3 Observations:



Figure 21: VIF values and Test Case 3 Observations

(i)The outcomes of test case 2 exhibit a substantial reduction in VIF values, signifying effective mitigation of multicollinearity concerns. While ‘season\_3’ retains a relatively higher value, this matter is presently not a primary focus of attention.

(ii) Notably, the P-values for ‘mnth\_3’, ‘mnth\_5’, and ‘mnth\_8’ continue to exceed the established threshold, reflecting persistent challenges in terms of these variables’ significance.

(iii) A remarkable enhancement is observed in the F-statistics value, surging from 147.6 to 185.6. This substantial improvement underscores the model’s previous reliance on ‘temperature’ (‘temp’), and the subsequent elimination of this feature has remarkably bolstered the overall stability of the regression model.

**5.2.12 Test Case 4:**

Removing all the features that have high P-Value (>0.5).

**“Mnth\_3”,” mnth\_5”,” mnth\_8”**

Test case 4 Observations:



Figure 22: VIF values and Test Case 4 Observations

From the output of the T4 model summary, it is evident that:

1. P values of all the feature is <0.5.
2. VIF of all the features is <5.
3. All our coefficients are not equal to zero.

Therefore, We REJECT the NULL Hypothesis and ACCEPT the ALTERNATE Hypothesis.

**T4 model coefficient values with constant = 0.8815**

|  |  |
| --- | --- |
| **Features** | **Coefficient** |
| yr | 0.2307 |
| Working day | 0.0535 |
| atemp | 0.475 |
| hum | (-0.1442) |
| windspeed | (-0.1078) |
| season\_2 | 0.1501 |
| season\_3 | 0.1033 |
| season\_4 | 0.1629 |
| mnth\_9 | 0.1148 |
| mnth\_10 | 0.0666 |
| weekday\_6 | 0.0662 |
| weathersit\_2 | (-0.0623) |
| weathersit\_3 | (-0.2353) |

Table 3: Feature and its final coefficient value

**5.2.13 Result and its validation**

To determine the significance of the model, F-Statistics is used. A higher F-Statistics value indicates a more significant model. In this case, the F-Statistics value is 223, which is greater than 1, the VIF is less than 5 and the P- value is less than 0.005, indicating that the overall model is significant and stable.

1. Const (Constant): The constant term (0.8815) serves as the baseline prediction for the count of shared bikes when all predictor variables are zero. This means that on a day with no influence from other variables, the model predicts an initial count of 0.8815 shared bikes.
2. yr (Year): With a coefficient of 0.2307, the "yr" variable (representing the year) positively impacts the predicted count of shared bikes. This suggests that as the years progress, the model predicts an increase in the daily count of shared bikes.
3. Workingday: The coefficient of 0.0535 indicates that the "workingday" variable has a positive effect on the predicted count. This suggests that on working days, the model predicts a higher daily count of shared bikes compared to non-working days.
4. atemp (Adjusted Temperature): A coefficient of 0.4750 implies that an increase in the adjusted temperature ("atemp") results in a higher predicted count of shared bikes. This suggests that warmer temperatures are associated with more shared bike rentals.
5. hum (Humidity): The negative coefficient of -0.1442 indicates that an increase in humidity ("hum") is associated with a decrease in the predicted count of shared bikes. Higher humidity levels are linked to a lower predicted daily bike count.
6. windspeed: Similarly, the negative coefficient of -0.1078 suggests that higher windspeeds lead to a lower predicted count of shared bikes.
7. season\_2, season\_3, season\_4: These dummy variables represent different seasons. The positive coefficients (0.1501, 0.1033, 0.1629) indicate that the model predicts a higher count of shared bikes in seasons 2, 3, and 4 compared to the reference season.
8. mnth\_9, mnth\_10: These dummy variables represent specific months. The positive coefficients (0.1148, 0.0666) suggest that the model predicts a higher count of shared bikes in September and October compared to the reference month.
9. weekday\_6: This dummy variable represents a specific weekday (likely Saturday). The positive coefficient of 0.0662 implies that the model predicts a higher count of shared bikes on this day.
10. weathersit\_2, weathersit\_3: These dummy variables represent different weather situations. The negative coefficients (-0.0623, -0.2353) indicate that the model predicts a lower count of shared bikes in weather situations 2 and 3 compared to the reference weather situation.

The equation of best-fitted surface based on the results of Test case 4 is:

**cnt** = 0.8815 + (yr × 0.2307) + (working day × 0.0535) + (atemp × 0.4750) -(hum × -0.1442) − (windspeed × 0.1078) + (season\_2 × 0.1501) + (season\_3 ×0.1033) +(season\_4 × 0.1629) + (mnth9 × 0.0.1148) + (mnth\_10 × 0.0666) +(weekday\_6 × 0.0662) − (weathersit\_2 × 0.0623) − (weathersit3 × 0.2353)

**6. Discussion and Conclusion**

The report embarked on a study of the shared mobility business, encompassing various dimensions crucial for Lanterne's success. From analysing bike rebalancing strategies and customer usage patterns to employing predictive models for daily demand projections. It delved into a wide array of research papers, research methods, statistical analyses, and business relations to unearth insights that could serve as strategic drivers for shared mobility systems.

The optimization of bike rebalancing is crucial for operational efficiency. Prior research papers use advanced optimization techniques and algorithms like TSP, HC, SA, and GA, orchestrating the bike redistribution. The operational aspect ensures maximum bike availability and creates a seamless bridge between supply and demand.

The focus of this study then shifts to a thorough investigation of forecasting the daily demand for shared automobiles. This research provides an examination of the shared mobility industry with a focus on key areas such as customer behaviour, bike utilization trends, weather forecast, and month and season to forecast daily usage demand. The goal is to provide a data-driven approach that acts as a compass for strategic, tactical, and operational decisions, promoting the optimization of corporate growth and recognizing competitiveness within the dynamic shared mobility environment. Two areas of analysis can be used to organize the final stages of daily demand forecasting.

1. Analysis of consumer utilization of shared mobility.

2. Predictive study of daily demand.

Addressing the core research questions, the analysis shed light on the determinants of shared bike utilization, the trend of user behaviour, and the intricate patterns in daily demand. The analysis deciphered the pivotal role of accessibility, convenience, and demographic factors in influencing shared mobility choices. Furthermore, the investigation into daily demand patterns brought to the fore the significance of weather conditions and various temporal aspects in shaping bike rental demand Advanced analytics and intuitive design allows for optimized bike availability and streamlined booking processes, resulting in a holistic user-centric experience.

Throughout the analysis, a plethora of results emerged that shed light on various facets of shared mobility:

* Factors such as cost-efficiency, convenience, and were key drivers for users opting for shared bikes. Further age emerged as a critical determinant, with younger individuals being more inclined towards this mode of transportation.
* Additionally, the correlation between weather conditions, temperature, and bike rental demand was established, unveiling seasonal variations in user behaviour.
* The predictive modelling phase further contributed by offering a holistic understanding of the factors impacting daily bike demand. Features such as year, temperature, working day status, and weather conditions exhibited notable influences on the bike rental counts. The developed predictive models, including Random Forest, XGBoost, and Mixed Integer Linear Programming, presented varying degrees of accuracy in forecasting bike demand.

**6.1 Critical Evaluation and Future Work:**

While the research analysis succeeded in its primary objective of providing insights into shared mobility and bike demand prediction, there exist certain limitations and avenues for future exploration. The predictive models, although effective, are subject to the inherent unpredictability of real-world dynamics. The report could delve deeper into user behaviour analysis by considering factors such as the incorporation of real-time data and advanced deep learning techniques could offer more dynamic predictions and enable adaptive strategies. Fine-tuning and continuous recalibration of these models could enhance their accuracy. Moreover, the report and analysis primarily focused on shared bikes, leaving room for incorporating other shared mobility options like cars and scooters to offer a more comprehensive overview.

* The findings on the influence of weather factors on the count of bikes are for a specific geographical locality, therefore they could not accurately reflect the diversity of urban settings and transportation practices throughout the world. A more thorough comprehension of the global dynamics of shared mobility could result from expanding the research to include a wider variety of cities and regions.
* The gathering procedure for the consumer data included in this research may have introduced biases. The outcomes might be impacted by sampling biases and non-representative data sources. Such biases might be reduced, and the legitimacy of the analysis would be improved by using representative and varied data-gathering sources.
* The data utilized in the analysis contains user information. Ensuring compliance with privacy regulations and ethical considerations is crucial. Future projects could be more sensitive towards privacy and ethical considerations to explore techniques for anonymizing and aggregating data to address privacy concerns while still extracting valuable insights.
* The research predominantly focused on shared bikes, but the shared mobility ecosystem often involves multiple modes of transportation, including scooters and cars. Exploring the synergies and competition among different modes and their implications for users' choices could yield valuable insights.

**6.2 Business Recommendation:**

The strategic recommendations for Lanterne can be organised across three key domains: user experience improvement and optimal bike rebalancing at three levels: Strategic, Tactical, and Operational.  
  
Recommendation 1: Enhancing User Experience and Maximizing Booking Success by Leveraging Advanced Analytics and User Journey Tracking

Scenario: Personalized Push Notifications  
The user habitually opens the app during lunch breaks but discovers that all the bikes are usually reserved.

Create thorough user journey maps that show the complete procedure, from the app's launch through the rental or abandonment of a bike. Introduce push notifications to let consumers know if there are bikes available in their preferred locations, hence resolving the "no-bike-found" annoyance.

Potential Benefits:

* Increased customer satisfaction because of improved app use and bike availability notifications.
* Excellent word-of-mouth promotion as delighted customers tell their friends and colleagues about their excellent experiences.

Recommendation 2: Optimal Bike Rebalancing - Employing Advanced Optimization Techniques for Enhanced Availability

Efficient bike rebalancing is critical for ensuring that bikes are accessible where and when consumers want them. To address this difficulty, Lanterne can use complex optimization techniques such as the Traveling Salesman Problem (TSP), Hill Climbing (HC), Simulated Annealing (SA), and Genetic Algorithm (GA). TSP can efficiently design paths for rebalancing agents, HC can iteratively improve routes, SA can successfully explore solution spaces, and GA can identify near-optimal solutions in complicated settings. Lanterne may reduce operational expenses while increasing bike availability in high-demand locations by utilizing these approaches.  
  
Scenario: Route Planning Using TSP

TSP is used by Lanterne's rebalancing team to determine the most effective path for a rebalancing agent to cover several bike stations. The agent thus visits more stations in less time, ensuring that bikes are appropriately allocated and easily accessible to users. Genetic Algorithms rebalance tactics over time, learning from prior patterns and optimizing for long-term success.

Potential Benefits:

* Optimization of rebalancing routes and resource distribution to save operating expenses. Better bike availability leads to better consumer satisfaction and utilization.
* Capability to adjust to variations in dynamic demand, providing quick bike redistribution during peak hours.

**6.3 Future Scope and Visionary Ideas for Advancing Shared Bike Mobility:**

By embracing visionary ideas encompassing technological innovation, environmental responsibility, user engagement, and societal impact, the company can transcend conventional boundaries and lead the industry forward. For sustained success in the evolving shared mobility landscape, embracing innovation and forward-thinking strategies is essential. This section explores visionary ideas that incorporate artificial intelligence (AI) innovations to further elevate shared bike mobility services.

1. AI Predictive Maintenance: Using AI for predictive maintenance of bike fleets provides a proactive approach. AI algorithms can forecast maintenance needs, such as worn-out parts or possible failures, by outfitting bikes with IoT sensors. This reduces operational disturbances, increases bike longevity, and increases user safety. Furthermore, AI can optimize maintenance plans to decrease downtime and maximize resource allocation.
2. Shared autonomous vehicles: The integration of autonomous bikes or cars with artificial intelligence and robotics interface with shared mobility. These self-driving bikes can autonomously redistribute to high-demand regions, assuring continuous availability. AI algorithms powered by real-time data analytics can optimize bike deployment, improve operational efficiency, and contribute to reduced traffic congestion. (Andreas Cornet, Matthias Kässer,Thibaut Müller,Andreas Tschiesner, 2017)
3. Sustainability and environmentally friendly initiatives: As environmental issues gain significance, introducing an electric or solar-powered bike fleet is in line with sustainability aims and appeals to environmentally aware riders. Collaborating with local communities to plant trees or offset carbon emissions from bike use may also be a bold move toward corporate social responsibility. (Shaopeng Zhong, 2023)
4. Augmented Reality Navigation: Integrating augmented reality (AR) navigation into the app may significantly improve the user experience. Through AR glasses or smartphone screens, users may see real-time bike availability, station locations, and navigation signals projected onto their surroundings. This user-friendly approach makes bike rentals even more convenient and accessible.

In conclusion, while the research report has achieved its objectives within its defined scope, it's crucial to recognize and address its limitations. The evolving nature of shared mobility, coupled with technological advancements, demands a continuous effort to refine methodologies, gather more comprehensive data, and embrace emerging techniques. By addressing the limitations outlined here, future research and experiments can contribute to a more holistic understanding of shared mobility and pave the way for more effective strategies and solutions in dynamic urban transportation.

# References

1. (2017). In W. Mckinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* Sebastopol: O’Reilly Media, Incorporated.
2. (2021). In A. F. Navlani, *Python data analysis perform data collection, data processing, wrangling, visualization, and more using python, third edition.* Packt Publishing.
3. Ahmadreza Faghih-Imani, R. H. (2017). An Empirical Analysis of Bike Sharing Usage and Rebalancing: Evidence from Barcelona and Seville . *Transportation Research Part A Policy and Practice*, 29.
4. Andreas Cornet, Matthias Kässer,Thibaut Müller,Andreas Tschiesner. (2017, September). *The road to artificial intelligence in mobility—smart moves required.* Retrieved from McKinsey & Company: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-road-to-artificial-intelligence-in-mobility-smart-moves-required
5. Andreas Kaltenbrunner, R. M. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Barcelona Media - Innovation Centre*.
6. Claire Chung, A. Y. (2018). *Matplotlib for Python Developers.* Packt Publishing.
7. Corcoran, J. (2014). *Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events.* Elsevier.
8. Cote, C. (2021). *5 PRINCIPLES OF DATA ETHICS FOR BUSINESS*. Retrieved from Harvard Business School- Business Insights: https://online.hbs.edu/blog/post/data-ethics
9. Craig Bullock (Dr), F. B. (2016). The economic contribution of public bike-share to the sustainability. *Sustainable Cities and Society (SCS)*, 12.
10. Craig Bullock, F. B. (2016). The economic contribution of public bike-share to the sustainability. *Elsevier*.
11. DeMaio, P. (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. *Digital commons*.
12. Fanaee-T, H. (2013). *Bike Sharing Dataset. UCI Machine Learning Repository.* Retrieved from UCI Machine Learning Repository.: http://archive.ics.uci.edu/dataset/275/bike+sharing+dataset
13. Geeksforgeeks. (2023). *XGBoost for Regression*. Retrieved from Geeksforgeeks: https://www.geeksforgeeks.org/xgboost-for-regression/
14. Géron, A. (2019). *Hands-on machine learning with Scikit-learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.
15. *Global Bike Sharing Market (2022 to 2027) - Industry Trends, Share, Size, Growth, Opportunity and Forecasts*. (n.d.). Retrieved from Yahoo finance.
16. Grand view research. (2020). *hared Vehicles Market Size, Share & Trends Analysis Report By Service (Car Rental, Bike Sharing, Car Sharing), By Region, And Segment Forecasts, 2022 - 2028*. Retrieved from Grand view research: https://www.grandviewresearch.com/industry-analysis/shared-vehicles-market-report
17. Ivezić, Ž. e. (2019). *tatistics, Data Mining, and Machine Learning in Astronomy : A Practical Python Guide for the Analysis of Survey Data, Updated Edition.* Princeton: Princeton University Press.
18. Jasper Schuijbroek, R. H.-J. (2017). Inventory Rebalancing and Vehicle Routing in bike sharing system. *European Journal of Operational Research*, 27.
19. Jenn-Rong Lin a, T.-H. Y.-C. (2013). A hub location inventory model for bicycle sharing system design: Formulation and solution. *Google Scholar*.
20. Jolly, K. (2018). *Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python.* Packt Publishing, Limited.
21. Jon Froehlich, J. N. (2008). Measuring the Pulse of the City through Shared Bicycle Programs. *Proc. of UrbanSense08*.
22. Kumari, K. (2023). *End-to-End Case Study: Bike Sharing Demand Prediction*. Retrieved from Analytics vidhya: https://www.analyticsvidhya.com/blog/2023/05/end-to-end-case-study-bike-sharing-demand-prediction/
23. Lumivero. (2023). *ORDINARY LEAST SQUARES REGRESSION (OLS)*. Retrieved from XLSTAT: https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols
24. Masoud Golalikhani, B. B. (2021). Carsharing: A review of academic literature and business practices toward an integrated decision-support framework. *Transportation Research*.
25. Michael Waskom. (2022). *seaborn: statistical data visualization*. Retrieved from seaborn: https://seaborn.pydata.org/
26. Neal Lathia, S. A. (2012). Measuring the impact of opening the London shared bicycle scheme. *Elsevier*.
27. Rutecki, M. (2022). *Multicollinearity - detection and remedies*. Retrieved from kaggle: https://www.kaggle.com/code/marcinrutecki/multicollinearity-detection-and-remedies
28. Schott, M. (2019, Apr 25). *Random Forest Algorithm for Machine Learning*. Retrieved from https://medium.com/: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb
29. scikit-learn developers. (2023). *Gradient Boosting regression¶*. Retrieved from Scikit learn 1.3.0: https://scikit-learn.org/stable/auto\_examples/ensemble/plot\_gradient\_boosting\_regression.html
30. scikit-learn developers. (2023). *RandomForestRegressor*. Retrieved from Scikit learn 1.3.0: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
31. scikit-learn developers. (2023). *sklearn.linear\_model.LinearRegression*. Retrieved from scikit learn 1.3.0: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html
32. Shaheen, S. A. (2019). Shared Micromoblity Policy Toolkit: Docked and Dockless Bike and Scooter Sharing. *UC Berkeley Transportation Sustainability Research Center*.
33. Shaheen, S., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future. *UC Davis Institute of Transportation Studies*.
34. Shaopeng Zhong, A. L.-J. (2023). Energy and environmental impacts of shared autonomous vehicles under different pricing strategies. *npj Urban Sustainability*.
35. Statista. (2023). *Bike-sharing*. Retrieved from Statista: https://www.statista.com/outlook/mmo/shared-mobility/shared-rides/bike-sharing/worldwide
36. Stellar. (2023). *Bike Sharing Market: Global Industry Analysis and Forecast (2023-2029) by Bike Type, Bike Model, Sharing System, and Region.* Retrieved from Stellar: https://www.stellarmr.com/report/Bike-Sharing-Market/44
37. Sujae Kim, G. L. (2022). Optimal Rebalancing Strategy for Shared e-Scooter Using Genetic Algorithm. *Wiley*.
38. Tae You Kim, J. S. (2022). Prediction of Bike Share Demand by Machine Learning: Role of Vehicle Accident as the New Feature. *International Journal of Business Analytics*.
39. Tuffery, S. (2011). *Data mining and statistics for decision making. Chichester, West Sussex.* Wiley.
40. UNIVERSITY OF WISCONSIN–MADISON. (2017, March). *pages.cs.wisc.edu.* Retrieved from UNIVERSITY OF WISCONSIN–MADISON: https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf
41. UNIVERSITY OF WISCONSIN–MADISON. (n.d.). Random Forests. *UNIVERSITY OF WISCONSIN–MADISON*, 2.
42. Wintjen, M. (2020). *Practical data analysis using jupyter notebook learn how to speak the language of data by extracting useful and actionable insights using python.* Packt Publishing.
43. Yim, A. C. (2018). In *Matplotlib for Python Developers: Effective Techniques for Data Visualization with Python, 2nd Edition.* Birmingham: Packt Publishing, Limited.
44. Ziliang Jin, Y. W. (2023, March). *Vehicle Rebalancing in A Shared Mobility System with Rider Crowdsourcing.* Retrieved from Informs Pubs online: https://pubsonline.informs.org/doi/abs/10.1287/msom.2023.1199#:~:text=To%20serve%20a%20region%2C%20the,neighboring%20regions%20throughout%20the%20day.

**THE END.**