A Bicycle Rental Demand Forecast Based on Random Forests and Multiple Linear Regression

A PROJECT REPORT

Submitted by

MANYA KARTHIK (2116210701152)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **"Predicting Used-Vehicle Resale Value in Developing Markets**" is the bonafide work of "**MANYA KARTHIK (2116210701152)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_____

Internal Examiner **External Examiner**

ABSTRACT

Bike sharing systems offer a convenient, automated way for users to rent bicycles, allowing them to pick up a bike from one location and return it to another. Initially, a multiple linear regression model was developed using SPSS software to predict rental demand. However, this model showed significant discrepancies when compared to actual usage data, indicating limited accuracy. The analysis revealed that the regression model did not adequately account for categorical variables such as time of day and season, which significantly affect rental patterns. To overcome these limitations and improve the accuracy of demand predictions, this study introduces the use of advanced machine learning techniques, specifically a random forest model and the Gradient Boosting Machine (GBM) within decision tree frameworks. These techniques are adept at managing complex interactions and non-linear relationships between variables. The random forest model demonstrated a notable improvement in predicting bike rental demand, significantly outperforming the initial multiple linear regression model. By employing random forest and GBM models, this study achieves more precise and reliable demand forecasts. These enhanced predictive models provide deeper insights into the factors influencing bike rentals, which can lead to better management and optimization of bike sharing systems. The results highlight the effectiveness of sophisticated machine learning approaches in capturing the complexities of real-world data, where traditional statistical methods may fall short.

Keywords: SPSS software, multiple linear regression, random forest model, GBM package.

I. INTRODUCTION

The global transition to low-carbon initiatives is gaining momentum, driven by the rising number of private vehicles that exacerbate urban traffic congestion and environmental pollution. In response, many cities are turning to public bicycle rental programs as an effective solution. Bicycles offer a flexible and efficient mode of transport for short trips, complementing existing public transportation systems and reducing reliance on private cars.

Current research on public bicycle rentals primarily focuses on predicting demand and optimizing station locations. However, the complexity of these systems, influenced by a wide range of factors, has limited comprehensive studies. One significant challenge is accurately forecasting demand, given the numerous variables at play, including socio-economic factors, infrastructure, and seasonal changes. Notably, the impact of weather conditions on bicycle rentals remains under-researched.

This study aims to address this gap by analyzing historical bicycle rental data to develop accurate demand forecasts, with a particular focus on weather conditions. We employ advanced analytical techniques, including random forest models and multiple linear regression, to harness the dataset's features.

Random forest models are well-suited for this task due to their ability to handle complex, non-linear interactions among variables and manage large datasets with many predictors. These models can identify the most significant factors affecting bicycle rentals, such as temperature, precipitation, and humidity. Multiple linear regression provides a traditional benchmark, highlighting the enhancements achieved through advanced machine learning techniques.

By comparing the performance of these models, we demonstrate the benefits of using sophisticated machine learning methods over traditional approaches. Our findings offer deeper insights into the influence of various factors, especially weather conditions, on bicycle rental demand. This information can assist city planners and transportation authorities in optimizing public bicycle rental programs, ultimately supporting more sustainable urban transport systems.

The results of this study improve the predictive accuracy of bicycle rental demand models, aiding in strategic planning and resource allocation. This contributes to reducing traffic congestion, lowering carbon emissions, and promoting healthier, more sustainable urban environments. By addressing gaps in current research, this study emphasizes the importance of innovative predictive models in enhancing the management of public bicycle rental systems, leading to smarter urban mobility solutions.

II. BASED ON MULTIPLE LINEAR REGRESSION MODEL

A. Data Introduction

A network of kiosk locations across a city facilitates the automated return of bicycles rented through a bike sharing system. With these systems, customers have the flexibility to rent a bike from any available pick-up point and return it to a location of their choosing. Globally, there are over 500 operational bike-sharing initiatives.

The data generated by these systems, including trip duration, starting and ending points, and elapsed time, presents a rich resource for academic research. This wealth of information positions bike sharing programs as valuable sensor networks, offering insights into urban mobility that are of great interest to researcher network.

The following elements are primarily included in the hourly rental statistics that has been provided for the past two years:

Date time – hourly date + timestamp

Season - 1 for spring, 2 for summer, 3 for fall, and 4 for winter.

Weather –

1: Clear, Few clouds, partly cloudy.

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

temp - temperature in Celsius.

feelstemp - "feels like" temperature in Celsius.

humidity - relative humidity

windspeed - wind speed

count - number of total rentals.

weather, temp, feelstemp, humidity, windspeed and count are numerical variable, Datetime, season and weather are discrete and discontinuous variable.

B. Multiple linear regression model.

This article focuses mostly on the demand forecast for bicycle rentals; multivariate linear regression analysis is the standard method for solving prediction problems. At first, the multiple linear regression model is employed [6–8]. The basic multiple linear regression analysis concept.

$$Y = \beta O + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \varepsilon. \quad (1)$$

The weather and season are qualitative variables whose values are discrete and continuous; they must be entered into the model as dummy variables, whereas the other six variables are common numeric variables. These are the seven factors that affected the rental bike market. The following can be used to construct the demanding forecast model for bicycle rentals:

$$\text{count} = \beta O + \beta 1X1 + \beta 2X2 + \cdots + \beta 7X7 + \varepsilon. \quad (2)$$

where ε is the random error term and βO to $\beta 7$ are the model's unknown parameters, also referred to as the regression coefficient. Count as the dependent variable, table bike rental demand forecast. X1 To X7 is the independent variable, which is the factors that influence.

The bike rental demand respectively called: weather, temp, attempt, humidity, wind speed, date time and the season. The regression model above has been established. Next, we will introduce the data to test required by the multiple regression analysis normality and linear relationship between the two-premise condition is satisfied. The SPSS work area will be imported, and the descriptive statistics of the continuous variables will be calculated as a result. Tables 1 and 2 are displayed below.

	quantity	*min*	*max*	*Standard Deviation*	*skewness*
temp	5737	0.82	41	8.16	0.113
feels temp	5737	0.76	45.46	8.87	0.005
humidity	5737	0	100	20.07	-0.105
windspeed	5737	0	57	8.27	0.576

TABLE I. DESCRIPTIVE STATISTICS RESULTS OF CONTINUOUS VARIABLE FACTORS.

	Quantity	*min*	*max*	*Standard Deviation*	*skewness*
count	5737	1	968	17.83	0.876

TABLE II. THE DEPENDENT VARIABLE OF CONTINUOUS VARIABLEDESCRIPTIVE STATISTICAL RESULTS.

We use the Matlab for continuous variables to the linear relationship between the drawings. According to the bicycle data collation. Shown as in Figures 1, 2, 3 and 4, we can find out that these factors have a strong linear relationship with the count.

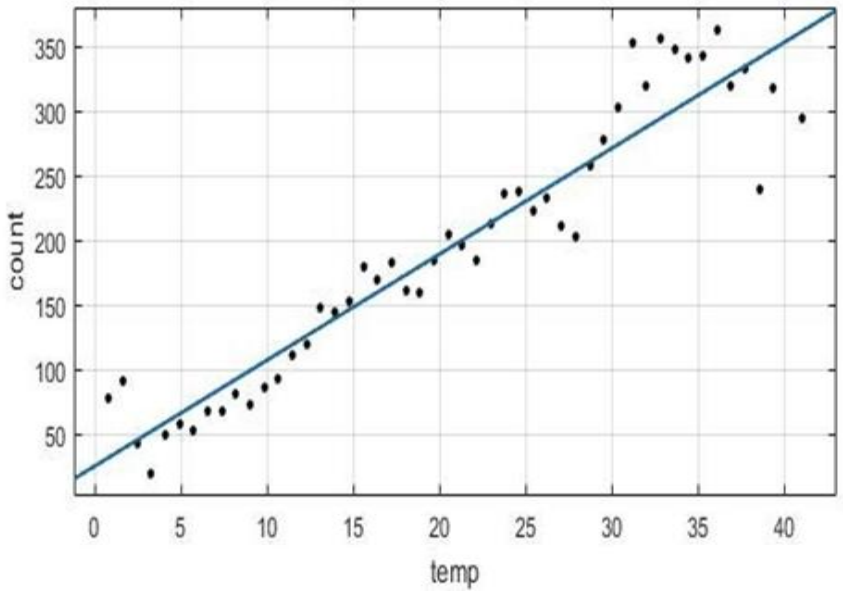


Fig.1 The relationship between the temp and the count

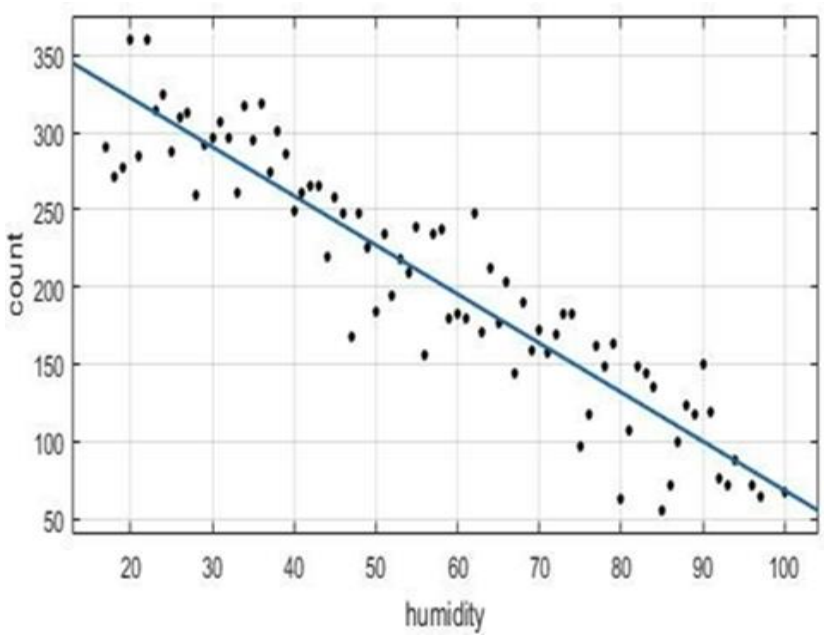


Fig.2 The relationship between the humidity and the count

According to the above analysis, using SPSS for multivariate linear regression analysis, multiple linear regression equation is obtained for, Linear regression equation, we see the Model Summary, in the table 3, there is the adjusted R square is 0.327, low value, show the fit of the equation worse, in the "one-way Anova, satisfy the F test, Sig. 0.00 is less than 0.005 with significant.

$$\text{count} = 16.143 + 7.504X_1 + 10.20X_2 - 5.934X_3 - 10.87X_4 + 17.018X_5 - 1.761X_6 + 1.008X_7 \quad (3)$$

Model	*R*	*R Square*	*Adjusted R Square*	*St Error of the Estimate*
1	.572[a]	.327	.327	146.32

TABLE III. THE MODEL SUMMARY

We forecast test set data using the multivariate linear regression equation, and the prediction accuracy is only 50%. The outcome showed that, while not terrible, the linear relationship between the different factors and the overall fit of the model is not good. The multivariate linear regression model's results for forecasting bike rental demand are not ideal, primarily due to the weather and season, which are included as dummy variables in the model and have an impact on the forecast's accuracy. Bike rental demand forecasting, therefore, is the conventional method of multivariate linear regression model but is not ideal.

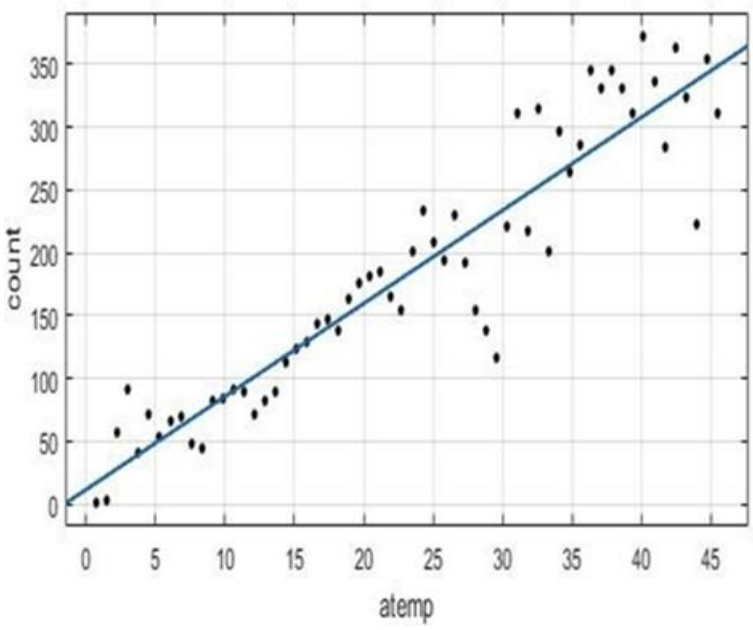


Fig.3 The relationship between the humidity and the count

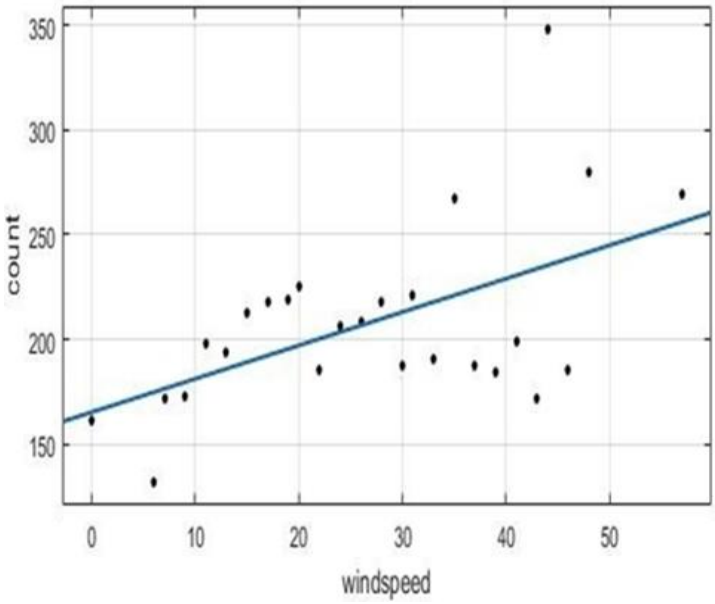


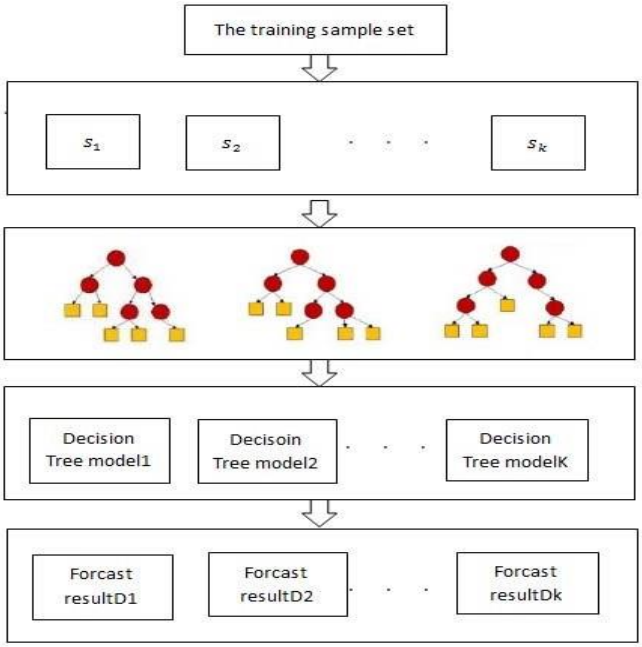
Fig.4 The relationship between the windspeed and the count

III. BASED ON RANDOM FOREST BICYCLE RENTAL DEMAND FORECASTING MODEL

We discovered that the traditional strategy is ineffective for forecasting bicycle rental demand when we used multivariate linear regression to the model used to anticipate demand for bicycle rentals. Upon examining the data once more, we discovered that the weather and the factors related to season are dummy variables. Based on the characteristics of this data, let me consider the method of random forests. As a result, linear regression analysis is not accurate. Thus, a random forest-based bicycle rental demand forecasting model is proposed in this study.

A. Random Forest method

Random forest classification by random vector growing into "tree", each tree growing without full pruning. Additionally, each node's variables at the time of the spanning tree are only a few, randomly chosen variables. Specifically, in the application of randomization to data (rows) and a variable (column). The term "random forest" refers to the process of randomly generating many trees for use in regression analysis and categorization. Each tree in the forest is dependent upon a different random vector, each of which is independently and uniformly distributed. The random forest categorization of the choice with the most votes serves as the basis for the final decision tree, which is based on the random vector potential tree "vote" that is formed**.** If the goal is to return, the mean value of the dependent variable will be calculated by averaging the outcomes of these trees.



Random forest classification by random vector growing into "tree", each tree growing without full pruning. Additionally, each node's variables at the time of the spanning tree are only a few, randomly chosen variables. Specifically, in the application of randomization to data (rows) and a variable (column). The term "random forest" refers to the process of randomly generating many trees for use in regression analysis and categorization. Each tree in the forest is dependent upon a different random vector, each of which is independently and uniformly distributed. The random forest categorization of the choice with the most votes serves as the basis for the final decision tree, which is based on the random vector potential tree "vote" that is formed**.** If the goal is to return, the mean value of the dependent variable will be calculated by averaging the outcomes of these trees.

B. The construction of random forest model

Recursive analysis is the initial step in training a set of data since random forests are not decision trees, which are a type of common single classifier; this results in the formation of shapes like inverted tree structures. A set of rules is produced by the second step analysis of the tree from the path of the root node to the leaf node; Ultimately, classification or projections for fresh data based on these guidelines.
The random forest model process is structured as follows: Random sampling was used to obtain n samples from the sample set.

Fig.5 The random forest construction and prediction process

k features were then randomly chosen from all the features to create the decision tree. m decision tree models are generated
by repeating the previous two steps m times, resulting in the random forest construction. For fresh data, the random forest construction and prediction procedure is as follows: forecasts are made at the end of each tree decision.

C.GBM improving the capacity of decision tree in the random forest.

When constructing decision trees with random forests, we utilize the GBM package to increase the decision tree's capacity. Each loss function model was created using the gradient descent direction of the prior model. The loss function indicates the degree of unreliability in the model; the bigger the loss function, the easier the model error (really, the problem of variance and deviation balance exists, but the assumption is that the more error prone the model, the greater the loss function).The ideal approach is to create the loss function in the gradient in the upward and downward direction if our model can be shrunk while maintaining the loss function, which demonstrates how our model is continuously improved.

In GBM package, important parameter Settings are as follows:

- distribution
- n.trees
- shrinkage
- bag.fraction
- interaction.depth

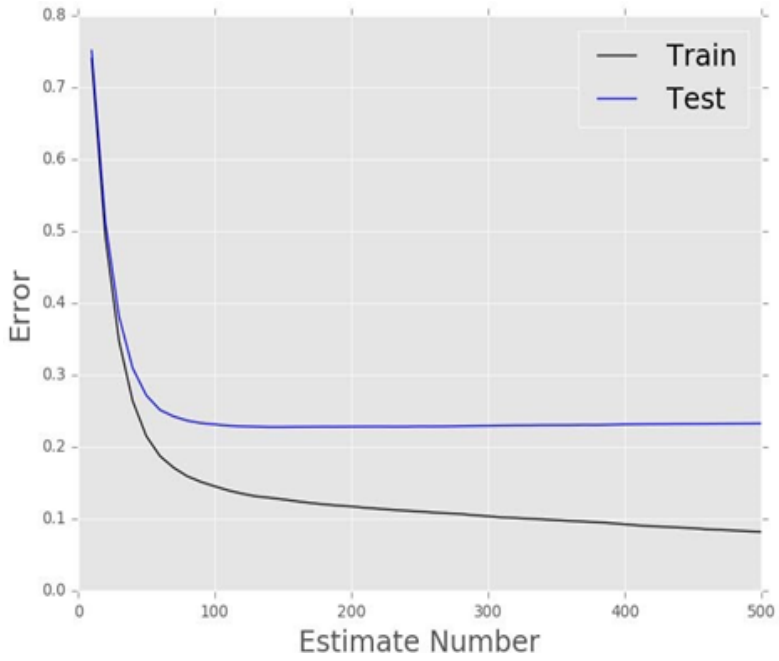


Fig.6 error vs number of estimator

The gaussian distribution is the one we chose because it allows us to forecast the problem and achieve the lowest possible shrinkage. However, if the shrinkage is too small, we will need to increase the number of iterations to reach the optimal model, which will increase the required time and computing resources. Thus, the shrinkage parameter is 0.005 and the number of trees is 5000.

As a result of our use of random forests for data modelling and the GBM package to enhance the decision tree's capacity, we were able to estimate the demand for bicycle rentals. As Figure 6 illustrates, the random forest forecast decreases error rates at 10% as the number of iterations increases.

We discovered that in this season, the weather, and time are factors that can effectively solve the random forests, forests that immediately increase the prediction accuracy. Through random forest prediction for the number of car rentals, with an 80% accuracy rate, the accuracy has been greatly improved when compared to multiple linear regression analysis. The random forest algorithm effectively solved the problems in the multiple linear regression.

V. CONCLUSION

First, the demand forecast for bicycle rentals above was created using a traditional multiple linear regression model, which has poor prediction accuracy despite a good linear relationship between the factors and a normal distribution of the factors. Some factors' characteristics, however, cause the result error to be extremely high. Therefore, the traditional multiple linear regression model is not appropriate to the bike rental demand projection in this article. Despite a decent linear relationship and a normal distribution of the components, the demand projection for bicycle rentals above was produced using a standard multiple linear regression model, which has a poor prediction accuracy. However, the peculiarities of some components lead to an exceptionally large outcome inaccuracy. Consequently, the demand estimate for bike rentals in this article is inappropriate for the conventional multiple linear regression model.

REFERENCES

[1] JIAO, Yuntai, LI, Wenquan, FENG, Peiyu, DING, Ran. A Scheduling Demand Model for Public Bicycle Rental Station [J]. Transportation and information security, 2014, 32(4): 8-13

[2] LU, FangQiang, Chen, XueWu, HuXiaoJian. Characteristic Research of Resident's Bus Trip Based on Bus OD Data[N]. Journal of Transportation Engineering and Information, 2010(2).

[3] LI, YanHong, Yuan, ZhenZhou. Analysion Trips Characteristic of Taxi in Suzhou Based on OD Data[N]. Journal of Transportation Engineering and Information, 2007(5).

[4] Camus R,Cantarella G E,Inaudi D,Real-time estimation and prediction of origin-destination matrices per time slince[J].International journal of Forecasting,1997,13(1):13-19

[5] Qian, Jin. Forecast and Analysis of the Demand for the Lease of City Public Bicycle’s Rental Station[D]. Chang’an University: , 2015.

[6] Wang, Huiwen, Meng, Jie. Multiple linear regression prediction modeling method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, (4):

[7] LinBin. Multiple linear regression analysis and its applicatio[N]. CHINA SCIENCE AND TECHNOLOGY INFORMATION May, 2010(9)