Applied Data Science
Teesside University

Middlesbrough, England, United Kingdom

PREDICTIVE MODELLING OF BIKE SHARING DEMAND

Grace Ebere Uzordinma
(Q2241217)

The Report is Submitted in Partial Fulfilment of the requirements
for a Master of Science Degree.

PREDICTING BIKE SHARING DEMAND: MACHINE LEARNING APPROACHES

Grace Ebere, Uzordinma
Q2241217@tees.ac.uk

The Department of Applied Data Science at
Teesside University's School of Computing, Engineering and Digital Technologies (SCEDT)
United Kingdom.

ABSTRACT

This study discusses models for predicting bike sharing counts using weather, hourly, weekly, and date data, focusing on the increasing popularity of rental bikes in big cities for improved mobility and comfort. The feature selection method that ranks the features according to how well they predict and removes non-predictive parameters. Out of the 7 statistical regression models that were trained and tested: Regression analysis includes linear regression, random forest, decision tree, support vector machine, KNN, Lasso regression, gradient boosting, ridge regression, and neural networks. The best model was Random Forest with the highest R-square value of 0.96 followed by Decision Tree with R-square value of 0.927, then Neural Network with R-square value of 0.926

1. INTRODUCTION

Bicycles are made accessible for shared use through bike-sharing system. The system allows individuals to rent bikes from one of its stations and return them to another station within the same network. The public bicycle system was first used in Amsterdam in 1965 (Shaheen et al., 2010). However, it wasn't until the 2000s that the system was expanded globally, due to advancements in IT technology. A typical bike-share has several distinctive features and characteristics, including membership and pass costs, per-hour usage prices, and station-based bikes and payment systems. Most programmes are sufficiently intuitive for even inexperienced users to grasp.

All citizens in South Korean cities can use the bikes that are part of the bike-sharing programme. The ability to borrow a bike from any system station and return it to any other station, increased mobility, and the ability to serve a larger user base are the advantages of bike sharing over renting. Becoming a member of a bike-sharing programme allows anyone to take advantage of the facility's advantages. For example, private users can utilise the city's bike fleet for little or no cost. (Sathishkumar V. E, et al. 2020)

Bike sharing can be quite helpful in mitigating the effects of greenhouse gases, such as carbon emissions, which are substantial contributors to climate change, given the state of the ecosystem today.

The main goal of this study is to develop an advanced statistical model that would predict the number of bicycles available for rental based on data availability and comprehend the patterns and variables influencing the number of leased bikes on a given day.

1.1. DATASET DESCRIPTION

The dataset used was gotten from Kaggle **https://www.kaggle.com/code/gauravduttakiit/bike-sharing-multiple-linear-regression.** it has 17379 records for the daily bike sharing count, with 16 attributes total. Out of the 15, 14 variables are independent with 1 dependent which form the basis of my regression analysis.

S/N	ATTRIBUTES	DESCRIPTION	DATATYPE
1	Dteday	Date	Year-Month-Day date
2	Season	1=Winter, 2=Spring, 3=Summer, 4=Autum	Categorical
3	Yr	Year (0=2011, 1=2012)	Categorical
4	Mnth	Month; from Jan to Dec.	Categorical
5	Hr	Hour of the day (0 to 23)	Categorical
6	Holiday	If the day was a holiday	Categorical
7	Weekday	Days of the week (Sun to Sat)	Categorical
8	Workingday	Neither a holiday nor weekend	Categorical
9	Weathersit	1= Sunny, 2=Rainy, 3=Windy, 4=Cloudy	Categorical
10	Temp	Temperature in degree celsius	Continuous
11	Atemp	Sentation of temperature in degree celsus	Continuous
12	Hum	Relative humidity	Continuous
13	Windspeed	Wind speed (km/hr)	Continuous
14	Casual	Total number of casual bike user	Continuous
15	Registered	Total number of Registered memberes	Continuous
16	count	Total numbers of rental I a day	Continuous

1.2. LITERATURE REVIEW

Jiang et al. (2019) combined LSTM and convolution neural network (CNN) to estimate reallocation sites. He took into account the user, the origin and destination, as well as meteorological elements like wind direction and temperature, when developing the forecast model. DPNst achieved an F1 score of 42.71% after outperforming the comparative baselines overall in a series of tests conducted using real station-less bike-sharing data.

Ali et al. (2012) conducted a comparative analysis of the categorization outcomes derived from Random Forest and Decision Tree techniques (J48). The classification parameters include F-Measure, Precision, Accuracy, Recall, and occurrences that are correctly and wrongly identified. According to the classification results, Random Forest performs better—96.13%—for the same number of characteristics and large data sets, or sets with more instances, whereas J48 performs better with small data sets.

The traditional multiple linear regression approach is inapplicable to the demand projection for bike sharing according to Feng Y et al. (2017). They suggest a random forest-based bike sharing demand forecasting model that uses a GBM package to enhance the decision tree's capacity during the random forest process. The random decision tree is integrated into the forest, and the forest has multiple random forest models with strong generalisation ability. During training, the trees were independent of one another without losing accuracy. The accuracy of the result increased significantly, to 82%.

1.3. ETHICAL CONCERNS

Data Privacy:

I got my dataset from Kaggle, and I ensured that they don't contain personally identifiable information (PII) unless user consent was given before training and evaluation, this is to respect user privacy.

Data Bias: I was cautious of potential biases in my datasets, such as skewed sampling methods, and ensure fairness and inclusivity to reduce model predictions.

Model Transparency: The model was transparent and interpretable, with detailed documentation of its architecture, training process, hyperparameters, and evaluation metrics to ensure reproducibility and accountability.

2. METHODOLOGY

Out of the 8 algorithms used for this assessment I will be discussing 4 algorithms.

Linear Regression:

Since this model implies a linear relationship between the input features and the goal variable, it was chosen because of its simplicity and interpretability. When a system is considered to be in linear regression, it typically indicates that the conditional mean of Y is an affine function of X given the value of X.

$Y = \beta 0 + \beta 1X + s$ (1)

Random Forest and Decision Tree:

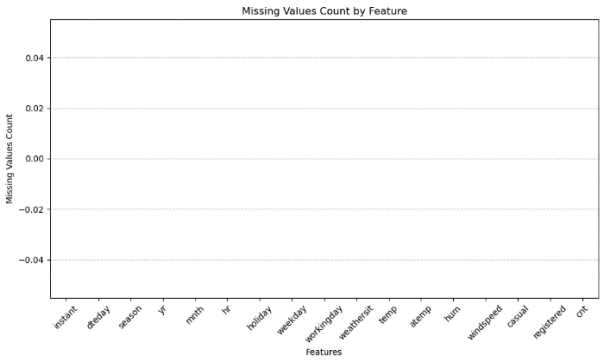
This model was used due to its ability to handle complex relationships and non-linear patterns in data. decision tree method, was created to divide the target dataset into subsets according to the preset parameters for each branch (Friedl and Brodley 1997). The models also reduce overfitting and improve the general performance. Random Forest can capture complex non-linear relationships between input features and target variables and provides a measure of feature importance, aiding in interpretability and decision-making. Its scalability makes it suitable for large datasets.

Neural Network:

I used this model because of its high scalability and handling of large data efficiently. Neural networks can generalize well to unseen data, making accurate predictions for new instances of bike sharing demand. They discover meaningful representations of input data at different levels of abstraction thereby improving the predictive performance.

3. DATA PREPROCESSING

3.1. Checking for missing values: The check indicates that neither the rows nor the columns of our data collection contain any missing or Null values.



3.2. Removing Unnecessary Columns: I dropped some of the columns as they were not important in the analysis and the prediction, columns such as instant, dteday, casual (it has the same total number as registered and it might cause bias in my model), atemp (almost like temp)

3.3. Renaming Columns: I renamed some columns to a readable name such as hr, hum, cnt, weathersit)

3.4. Encoding variables: I converted some of the categorical variable to numeric form and changed their names for readability.

```
# converting weather to categorical variable; 1:sunny,
bike.loc[(bike['weather'] ==1), 'weather'] = 'sunny'
bike.loc[(bike['weather'] ==2), 'weather'] = 'rainy'
bike.loc[(bike['weather'] ==3), 'weather'] = 'windy'
bike.loc[(bike['weather'] ==4), 'weather'] = 'cloudy'

bike.head()
```

	season	hour	holiday	weekday	workingday	weather	temp
0	winter	0	0	6	0	sunny	0.24
1	winter	1	0	6	0	sunny	0.22
2	winter	2	0	6	0	sunny	0.22
3	winter	3	0	6	0	sunny	0.24
4	winter	4	0	6	0	sunny	0.24

3.5. Creating Dummy Variables:

I created a dummy for the categorical variable by assigning a value of 1 to denote the existence of a specific category and a value of 0 to indicate its absence.

4. EXPLORATORY DATA ANALYSIS

4.1. Visualizing categorical variables:

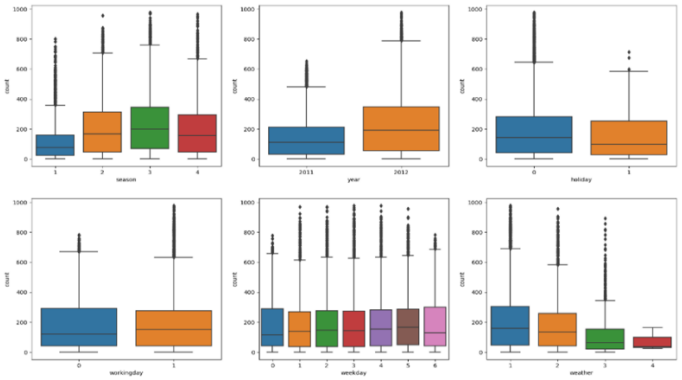


Fig.1. Boxplot showing all the categorical with the outliers.

4.2. Visualizing Numerical Variables:

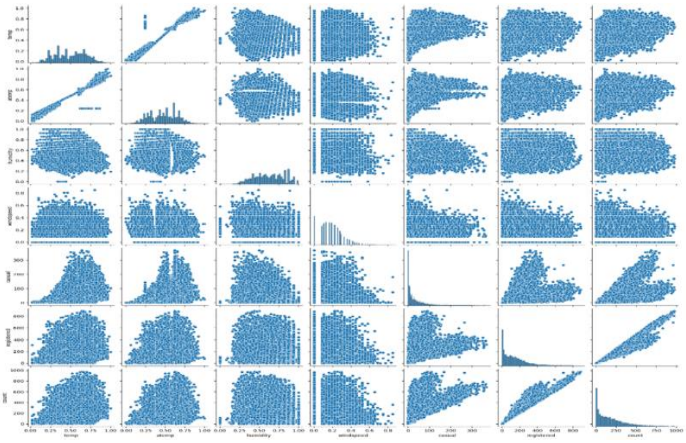


Fig.2. Pair plot showing all the numerical variables.

4.3. Visualizing Bi-variate analysis:

Bike sharing by Seasons: Fig1. Shows the seasons in a year where there is highest percentage of bike sharing. And we can see that people are more like to use the bike sharing during summer followed by spring then Autum.

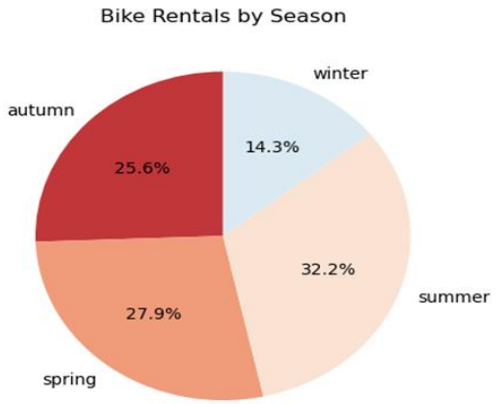


Fig.3. Pie chart showing the bike sharing by season.

Weather Distribution: Shows that people are more likely to rent the bike during a sunny day followed by rainy days.

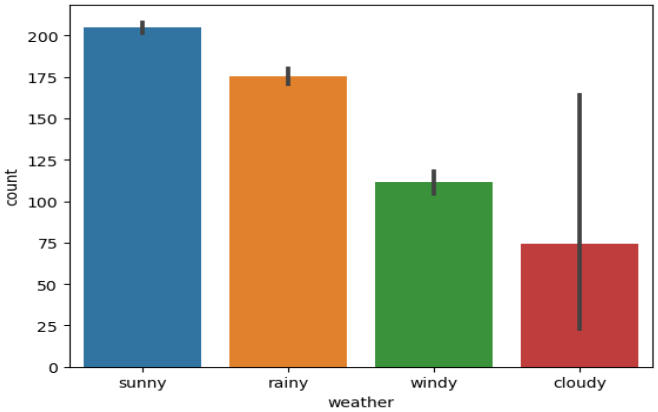


Fig.4. Bar chart showing the count of bike sharing by weather.

4.4. Multi-Variate Analysis:

Average User Count by Hour: The line chart shows that there is more count of rental between the hours of 7am-9am and 4pm-7pm and more rentals during the summer, spring and autumn season.

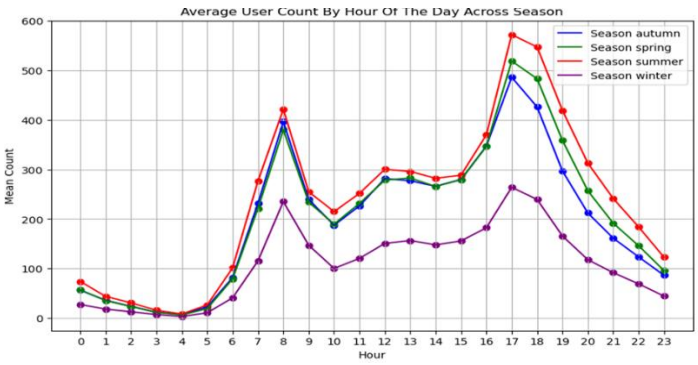


Fig.5. Multiline chart showing hourly count across the season.

Bike sharing by Month and Year: This shows that the rate of bike sharing was highest in 2012(Year 1) and that in September there was increased number of rentals.

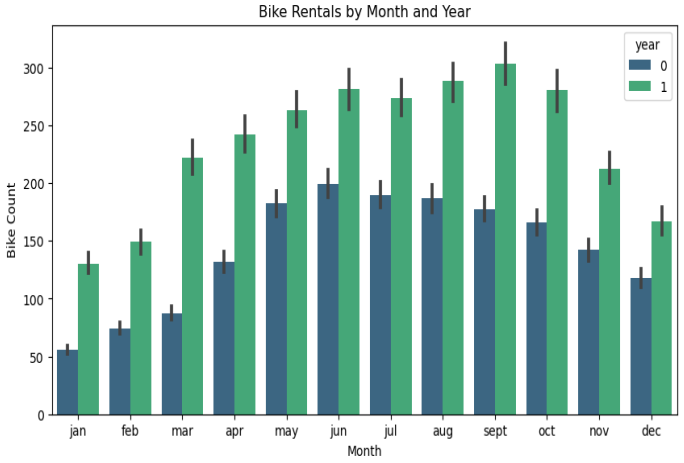


Fig.6. bar plot of bike sharing by month and year.

Registered Vs Casual Bikers Over year: This shows that there were more registered members for bike sharing in 2012 than causal rental.

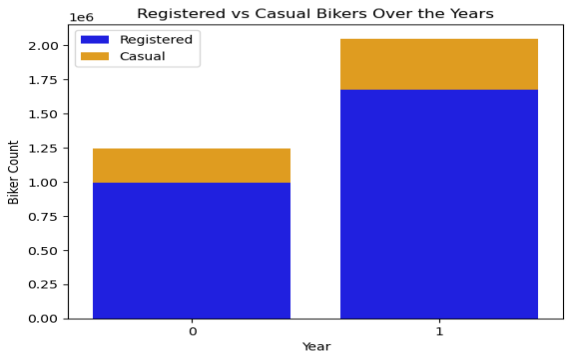


Fig.7. bar plot of registered and casual bikers over the year.

4.5. Correlation Analysis:

Fig 1 shows the correlation analysis before my dummy variables were created and we can see that we have some columns that are highly correlated (atemp, Registered) which I dropped and some feature which will not be important to my analysis (Humidity) which I also dropped.

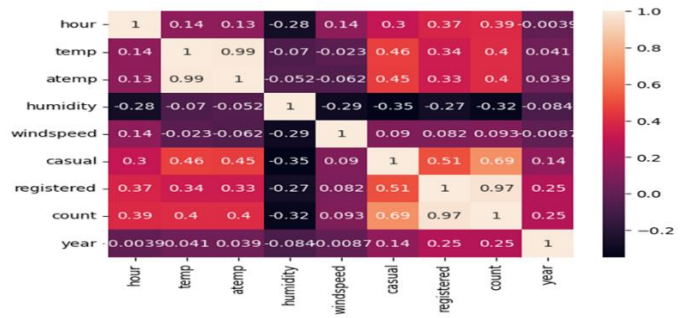


Fig.8. Correlation analysis.

4.6. Rescaling Features:

I rescaled the feature using the MinMaxScaler from sklearn. preprocessing, and then I created a heat map, as seen in fig1. Rescaling the features can increase the readability of the data, accelerate the convergence of machine learning algorithms, and reduce the possibility that a bias.

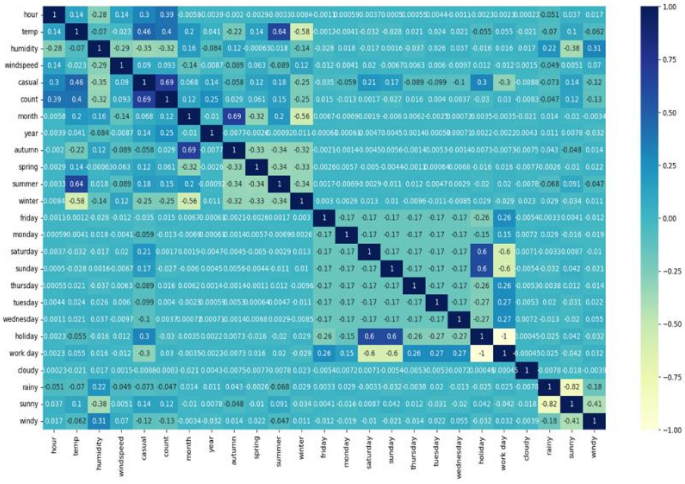


Fig.9. heat map shown the variables after scaling.

5. MACHINE LEARNING MODEL

5.1. Feature Selection:

This visualization shows the permutation importance of features from my KNN model, indicating the model performance when each feature's value is randomly permuted. My model shows that the hour, month, casual

and humidity feature performed better than year, autumn, rainy features.

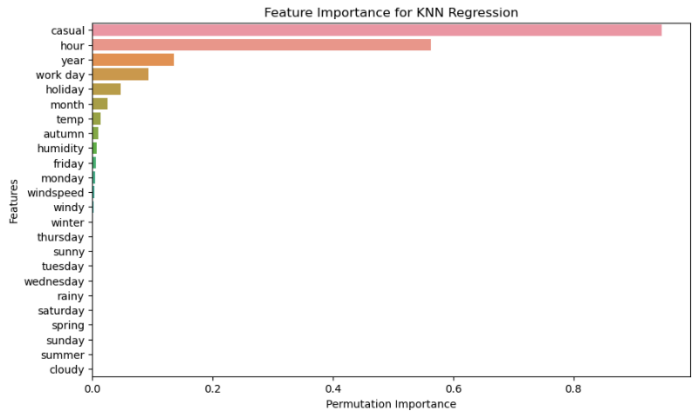


Fig.10. Bar chart showing the feature selection.

5.2. Proposed Algorithms:

Regression algorithms was used to predict the bike demand, which enhances efficiency and profitability in the bike sharing industry. The models used are, Linear regression, Random Forest regression, Decision Tree, SVM regressor, KNN regressor, Lassos regression, Ridge regression and Neural network.

6. EXPERIMENTS

Eight prediction models, including LR, RF, DT, SVM, Knn, Lasso Regression, Ridge Regression, and Neural Network, was trained on the count target variable in order to analyse the prediction performance of the algorithms used. As a result, the target variable is used to train and test these prediction models. A thorough examination of the variable's significance in relation to the most important variables across several models is carried out. The dataset is divided into two parts: 30% for testing and 70% for training. The regression algorithms are then trained using the generated training dataset. In the end, the test dataset assesses the prediction performance of each model using the different evaluation metrics. Afterwards, I ran the hyperparameter tuning to improve the performance of my machine learning model and to get better accuracy.

6.1. EVALUATION METRICS:

Regression models are trained using a repeated 10-fold cross-validation process in order to identify the best. I used 4 evaluation metrics in my model which are.

Root mean squared error (RMSE): This makes it possible to determine how the model response varies with regard to variance and to find substantial inaccuracies..

RMSE=n1∑i=1n(yi−y^i)2

Mean Squared Error (MSE): The average squared difference between the actual and anticipated numbers is what it calculates. More weight is assigned to larger errors.

MSE=n1∑i=1n(yi−y^i)2

Mean Absolute Error (MAE): This serves as a predictor of acuteness. By avoiding the offset between positive and negative errors, the MAE metric, like the RMSE metric, is a scale-dependent measure that efficiently depicts prediction error values.

MAE=n1∑i=1n|yi−y^i|

R-squared (R2) Score: The R2 score indicates the portion of the variation of the dependent variable that can be predicted from the independent variables. Better model fit is indicated by higher values, which range from 0 to 1.

R2=1−∑i=1n(yi−y¯)2∑i=1n(yi−y^i)2

6.2. HYPER-PARAMETER TUNNING:

Choosing the best hyper-parameter for models can be challenging, but there is a heuristic rule which can be followed and understanding the algorithm's workings are crucial. Once a good parameter is found, further refinement can occur within the range of the parameter. Also, after running a hyper parameter tunning to all my model, I noticed that all my model result improved most especially the R2 score.

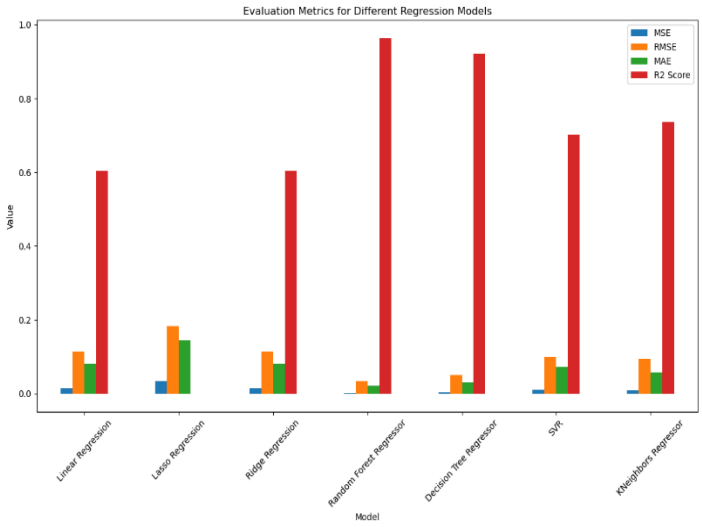


Fig.11. Optimization result for the model

7. RESULTS

7.1. Model Results

Model	Training				Testing			
	R2 Score	RMSE	MSE	MAE	R2 Score	RMSE	MSE	MAE
Linear Regression	0.613	0.117	0.014	0.082	0.605	0.115	0.013	0.082
Decision Tree	0.928	0.051	0.002	0.031	0.923	0.051	0.003	0.029
Random Forest	0.961	0.037	0.002	0.023	0.963	0.035	0.001	0.021
KNN Regression	0.666	0.108	0.012	0.071	0.715	0.097	0.009	0.062
SVM Regression	0.685	0.105	0.011	0.074	0.686	0.102	0.011	0.073
Lasso Regression	0.609	0.117	0.014	0.081	0.603	0.115	0.013	0.081
Ridge Regression	0.613	0.116	0.013	0.082	0.605	0.115	0.013	0.082
Neural Network					0.949	0.041	0.002	0.029

7.2. Result After Optimization

Model	R2 Score	RMSE	MSE	MAE
Linear Regression	0.605	0.115	0.013	0.082
Decision Tree	0.923	0.051	0.003	0.029
Random Forest	0.963	0.035	0.001	0.021
KNN Regression	0.715	0.097	0.009	0.062
SVM Regression	0.686	0.102	0.011	0.073
Lasso Regression	0.603	0.115	0.013	0.081
Ridge Regression	0.605	0.115	0.013	0.082
Neural Network	0.926	0.049	0.002	0.032

7.3. DISCUSSION ON INITIAL RESULT:

Linear Regression: The model achieved an R2 score of 60.5% of the target variable's variance, with moderate RMSE, MSE, and MAE values, indicating decent predictive performance.

Decision Tree: The model outperforms linear regression with R2 Score of 92.3% of the target variable's variance and showing good predictive performance.

Random Forest: The model outperformed the decision tree, achieving an R2 score of 0.963, demonstrating its robustness and ability to handle complex datasets.

KNN Regression: The model had an R2 score of 0.715, which showed a good performance but lags decision tree and random forest models in predictive accuracy.

SVM Regression: The model achieved an R2 score of 0.686 and showed moderate data fit, indicating its potential for optimal performance with careful hyperparameter tuning.

Lasso and Ridge Regression: showed similar results to linear regression, with comparable R2 scores and error metrics.

Neural Network: The model demonstrated strong data fit, achieving an R2 score of 0.926, demonstrating its ability to capture complex relationships in data.

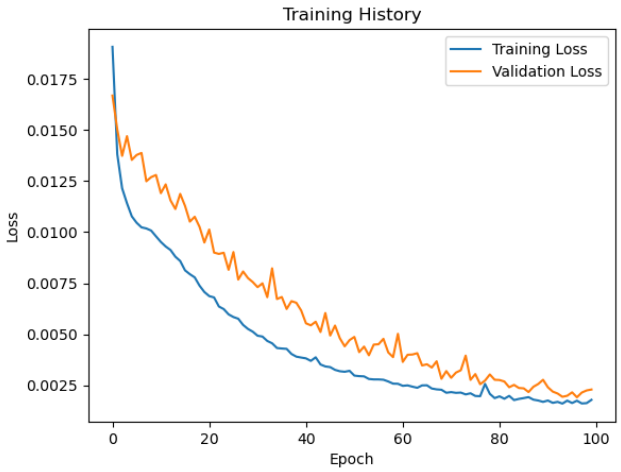


Fig.12. Chart showing validation loss for neural network.

7.4. DISCUSSION ON HYPERPARAMETER TUNNING RESULT:

Ridge, Lasso, and Linear Regression: The R2 score, and error metrics remained similar post-hyperparameter tuning, suggesting that these models may not have significantly improved.

Random Forest and Decision Tree: The models showed a high R2 score and slight improvements in RMSE, MSE, and MAE after tuning, indicating slight enhancement in predictive accuracy.

KNN and SVR Regression: The models showed enhanced predictive performance after tuning, with slight improvements in R2 score and error metrics.

8. CONCLUSION

In conclusion, my work used several regressors analysis to create a model for predicting the demand for shared bikes in a bike-sharing system. The study found that several important variables, including weather, temperature, season, month, and weekday, affect the demand for bikes.

The result of my study has benefits for planners of urban transportation and operators of bike-sharing programmes. Through a comprehensive understanding of the primary elements influencing bike demand, operators may boost customer satisfaction, optimise resource utilisation, and improve service quality.

To make sure that the results can be applied broadly, more investigation and validation are required. However, this study offers a strong foundation for further research on the topic of

forecasting demand for bike sharing and offers practical data that regulators and practitioners can utilise.

REFERENCES

[1]. Shaheen, S., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia. Transportation Research Record, 2143, 159-167. https://doi.org/10.3141/2143-20

[2]. S.V. V E, Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366. https://doi.org/10.1016/j.comcom.2020.02.007

[3]. Jiang, J., Lin, F., Fan, J., Lv, H., & Wu, J. (2019). A destination prediction network based on spatiotemporal data for bike-sharing. Complexity, 2019, Article e7643905. https://doi.org/10.1155/2019/7643905

[4]. Sathishkumar, V.E, & Cho, Y. (2020). Season wise bike sharing demand analysis using random forest algorithm. Computational Intelligence. https://doi.org/10.1111/coin.12287

[5]. Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. Retrieved from https://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf.

[6]. Feng, Y., & Wang, S. (2017). A forecast for bicycle rental demand based on random forests and multiple linear regression. In Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 101-105). IEEE.

[7]. Breiman, L. (2001) 'Random Forests. Machine Learning,' Machine Learning, 45(1), pp. 5–32. https://doi.org/10.1023/a:1010933404324.

[8]. Friedl, M. A., & Brodley, C. E. (1997). Decision Tree Classification of Land Cover from Remotely Sensed Data. Remote Sensing of Environment, 61(3), 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7.