

The background of the slide is a blurred photograph. On the left, a coiled metal spring, likely from a bicycle seat, is visible. On the right, a red brick wall is partially seen. A large white rectangle is centered on the left side, and a pink rectangle is centered below it.

BIKE SHARING INSIGHTS IN SEOUL

A project by Uzma Naeem

INTRODUCTION

- The dataset provides a detailed view of urban transportation patterns by combining rental bike counts with weather and holiday data.
- This multifaceted dataset reveals how environmental factors influence bike rental demand, offering valuable insights for researchers and urban planners.
- By analyzing this data, they can improve bike-sharing systems, make mobility more efficient, and promote sustainable urban development.





Problem Statement

The purpose of our project is to build a predictive model that performs accurately to forecast bike-sharing optimization based on analysis data and provide feature importance to improve overall efficiency of the business. Seasonal trends and feature importance can be used for building marketing strategies to increase usage.

RELEVANCE TO DS-861

- Application of predictive analysis:
 - Practical application of course concepts.
- Connecting Data Mining Techniques with Real-world Problem Resolution:
 - Address urban mobility challenges.
 - Optimize bike-sharing systems.



Course Relevance

EDITABLE STROKE

RESEARCH & LITERATURE REVIEW

Background & Importance:

- Efficient bike allocation critical for customer satisfaction.
- Prediction challenges due to various factors like time, events, & weather.

Methodology:

- Deep LSTM RNN chosen for sequential data processing.
- Addresses long-term dependency learning without gradient issues.

Data Processing:

- Citi Bike System data used, split into training and test sets.
- Includes historical weather data & time-related features.

Experimentation & Results:

- Compares deep learning models, with LSTM outperforming.
- Lower RMSE values indicate precise demand prediction.

Conclusion & Future Applications:

- Accurate demand prediction aids in efficient bike allocation.
- Potential for optimizing distribution & reducing operational costs.
- Model's adaptability extends to other bike-sharing systems.




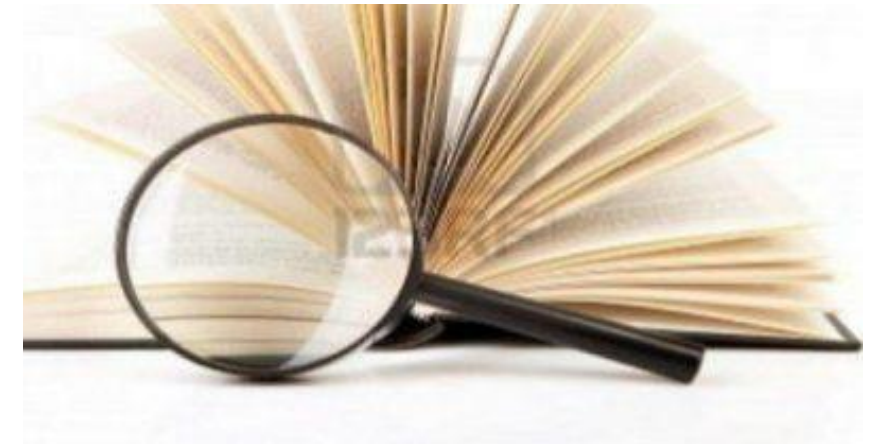
Procedia Computer Science

Volume 147, 2019, Pages 562-566



Predicting bike sharing demand using recurrent neural networks

[Yan Pan](#)^a , [Ray Chen Zheng](#)^a, [Jiaxi Zhang](#)^a, [Xin Yao](#)^b



Literature Review

METHODS

Seasonal Trends Analysis:

- Plotting the **average** across all seasons to discern patterns.

Data Pre-processing:

- Preparing the dataset for analysis through **cleaning** & organization.

Coefficient Identification:

- Employing **OLS model** to pinpoint significant coefficients.

Train-Test Data Split:

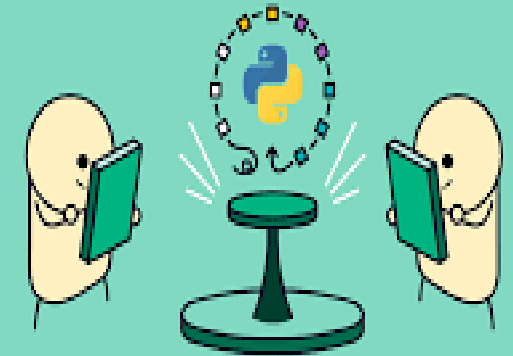
- **Dividing** data for effective model evaluation.

Predictive Modelling Techniques:

- Utilizing **Linear Regression, Decision Tree, Random Forest, LASSO, & Ridge Regression.**

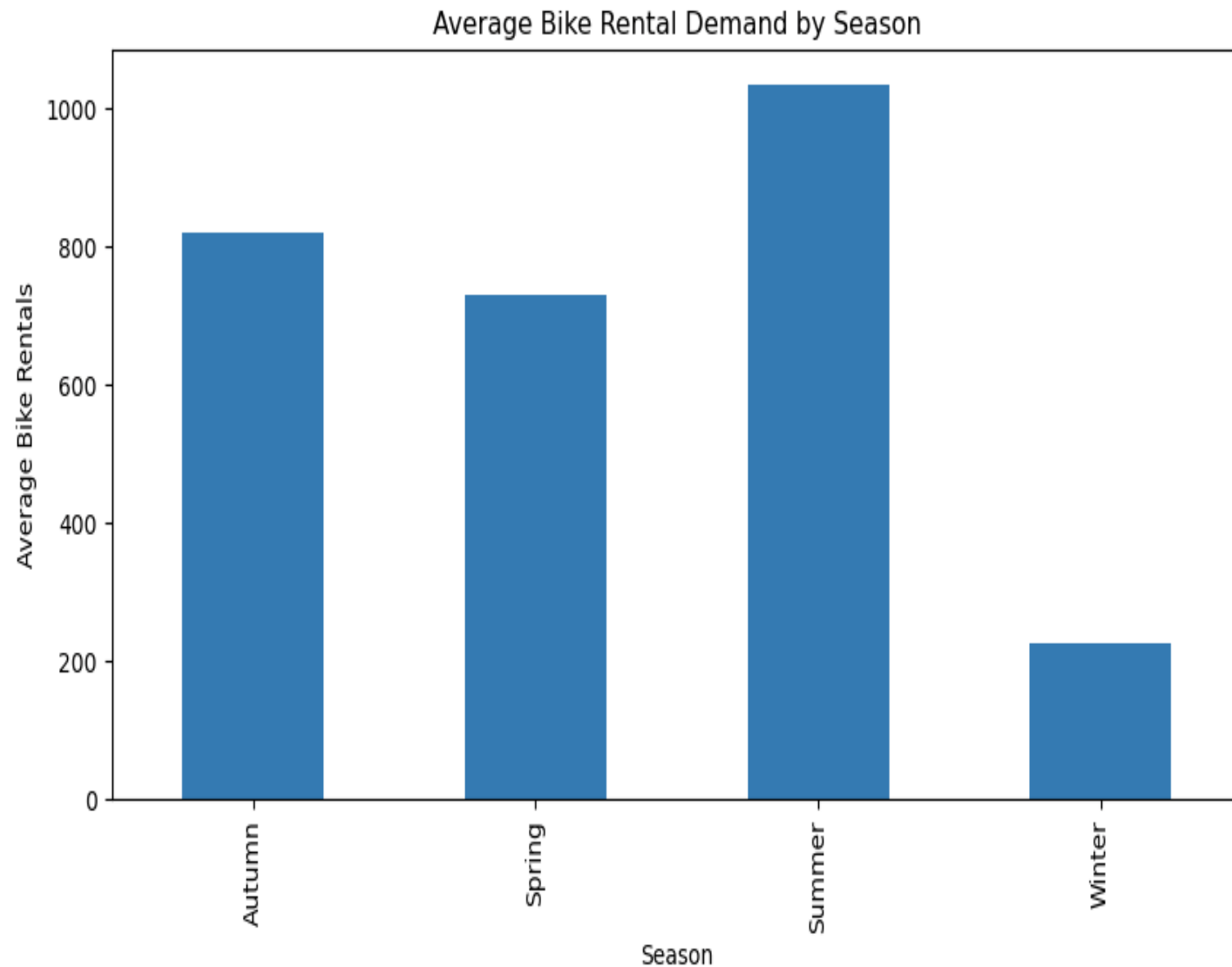
Performance Comparison:

- Assessing model performance with **RMSE & R-squared** metrics.



AVERAGE BIKE RENTAL DEMAND BY SEASON

The graph shows the average bike rental demand for each season from highest to lowest.



OLS REGRESSION RESULTS

- Used Log transformation
- R-squared: Indicates 77.3% of the variance in rented bike count explained by independent variables.
- Coefficients: Represents estimated impact of each independent variable on bike count.
- P-value: < 0.05 are statistically significant, except for:
 - Radiation
 - Snowfall (cm).
- The model shows good explanatory power.

OLS Regression Results

```

=====
Dep. Variable:    Rented Bike Count    R-squared:            0.773
Model:            OLS                  Adj. R-squared:       0.772
Method:           Least Squares        F-statistic:          1982.
Date:             Mon, 20 May 2024      Prob (F-statistic):    0.00
Time:             16:39:34              Log-Likelihood:       -7979.5
No. Observations: 7008                 AIC:                  1.598e+04
Df Residuals:     6995                 BIC:                  1.607e+04
Df Model:         12
Covariance Type:  nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	5.8849	0.009	651.433	0.000	5.867	5.903
x1	0.2933	0.010	29.720	0.000	0.274	0.313
x2	-0.0837	0.084	-0.997	0.319	-0.248	0.081
x3	-0.6225	0.041	-15.329	0.000	-0.702	-0.543
x4	-0.0425	0.010	-4.171	0.000	-0.062	-0.023
x5	0.0375	0.011	3.281	0.001	0.015	0.060
x6	0.8245	0.096	8.589	0.000	0.636	1.013
x7	-0.0032	0.013	-0.251	0.802	-0.028	0.022
x8	-0.2471	0.009	-26.262	0.000	-0.266	-0.229
x9	-0.0163	0.010	-1.702	0.089	-0.035	0.002
x10	-0.0524	0.011	-4.690	0.000	-0.074	-0.031
x11	-0.0901	0.009	-9.908	0.000	-0.108	-0.072
x12	1.1307	0.009	124.483	0.000	1.113	1.148

```

=====
Omnibus:            840.567    Durbin-Watson:           1.994
Prob(Omnibus):      0.000     Jarque-Bera (JB):        3284.089
Skew:               -0.556     Prob(JB):                 0.00
Kurtosis:           6.164     Cond. No.                 24.2
=====

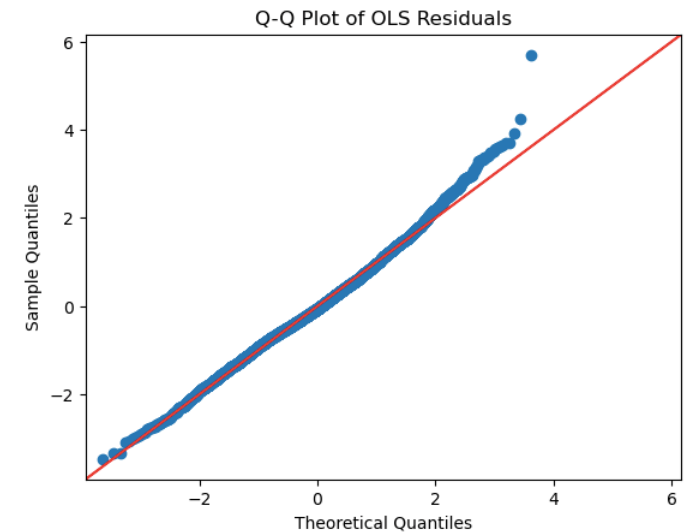
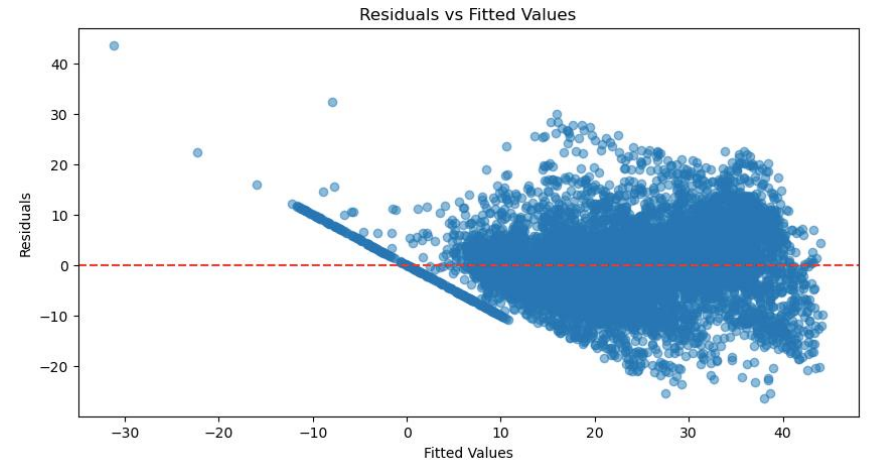
```

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

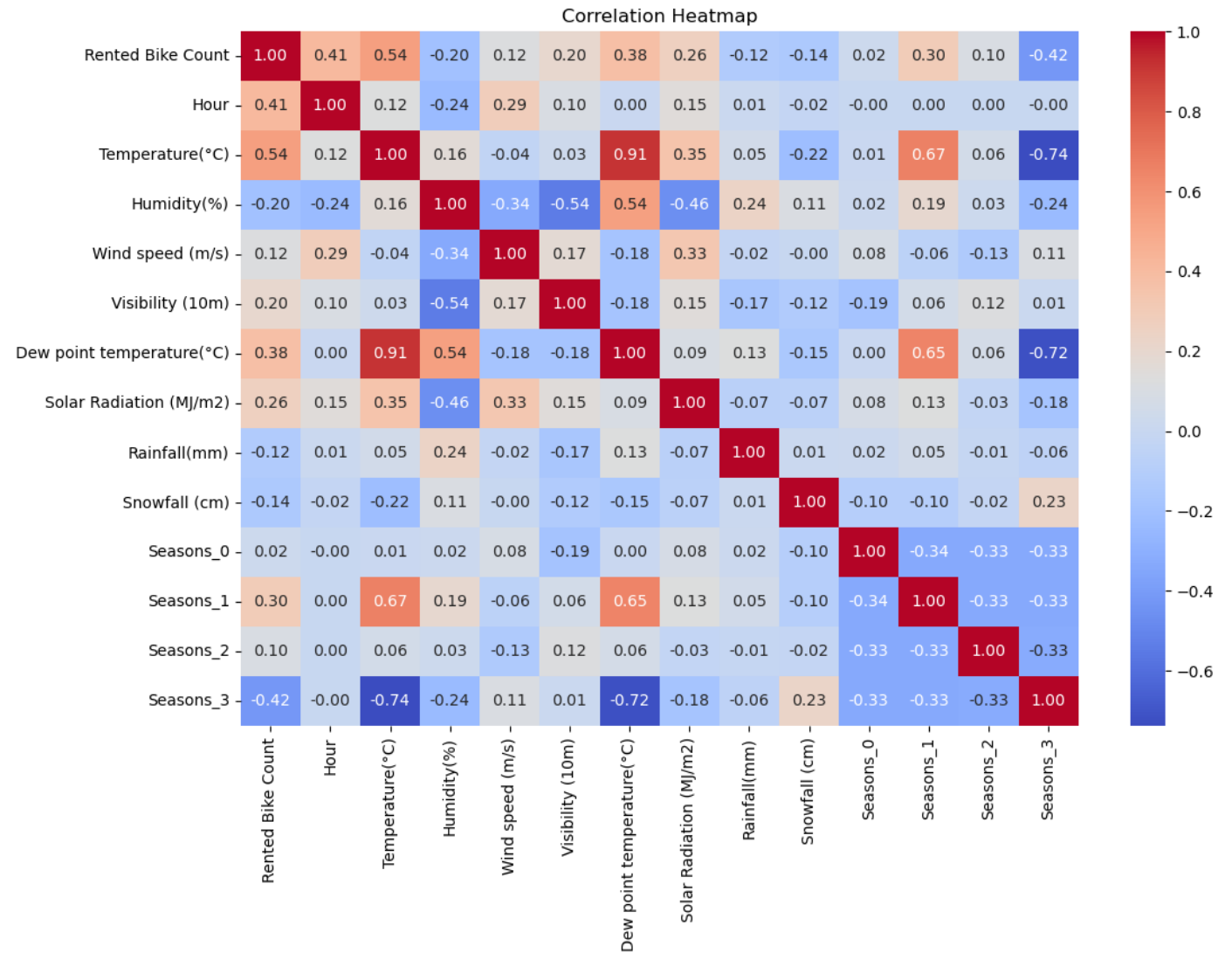
RESIDUALS & QQ PLOT AFTER BACKWARD ELIMINATION

- After removing coefficients with higher p-values, OLS model still did not achieve better results.
- Therefore, we will apply regularization models, such as Lasso and Ridge Regression to improve model performance.



CORRELATION MATRIX & HEAT MAP(HM)

- **Positive Correlation:** HM shows positive correlation with Temperature (0.54) and Hour (0.41) .
- **Negative Correlation:** HM shows a negative correlation with Humidity (-0.20) and Seasons_3 i.e. Winter (-0.42).
- **Weak Correlation:** Some variables show weak correlations with bike rental counts, like Rainfall (-0.12) and Snowfall (-0.14), indicating minimal impact on bike rentals.

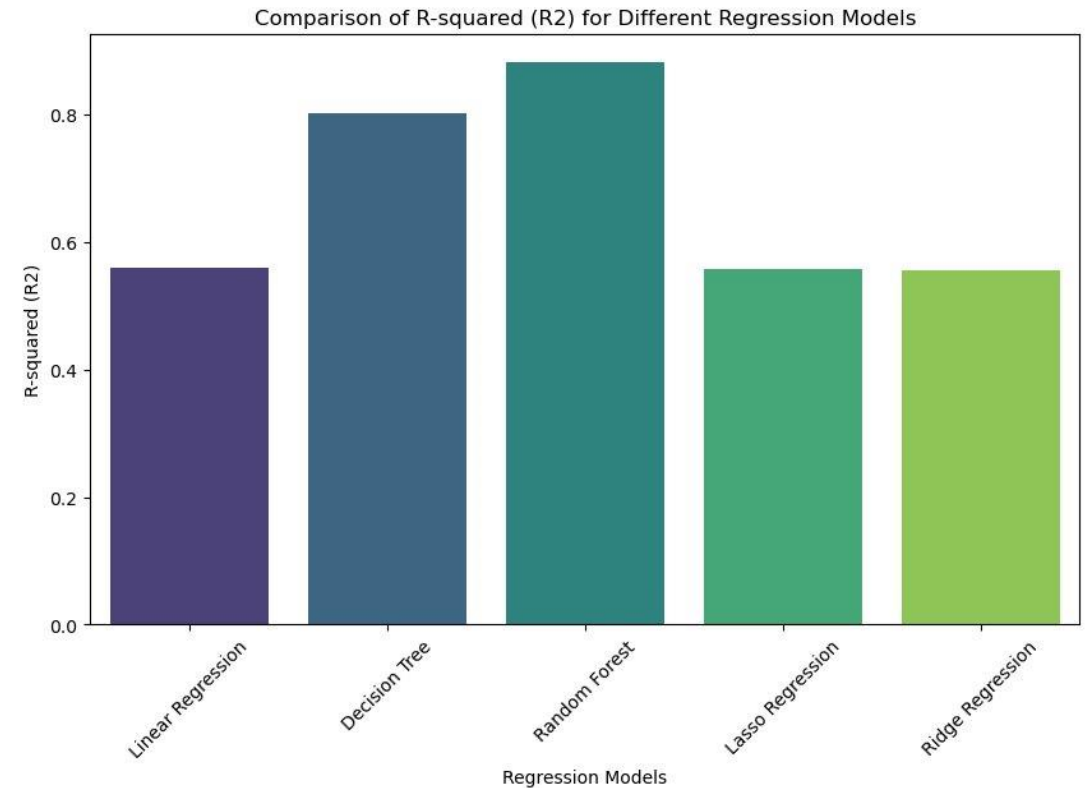
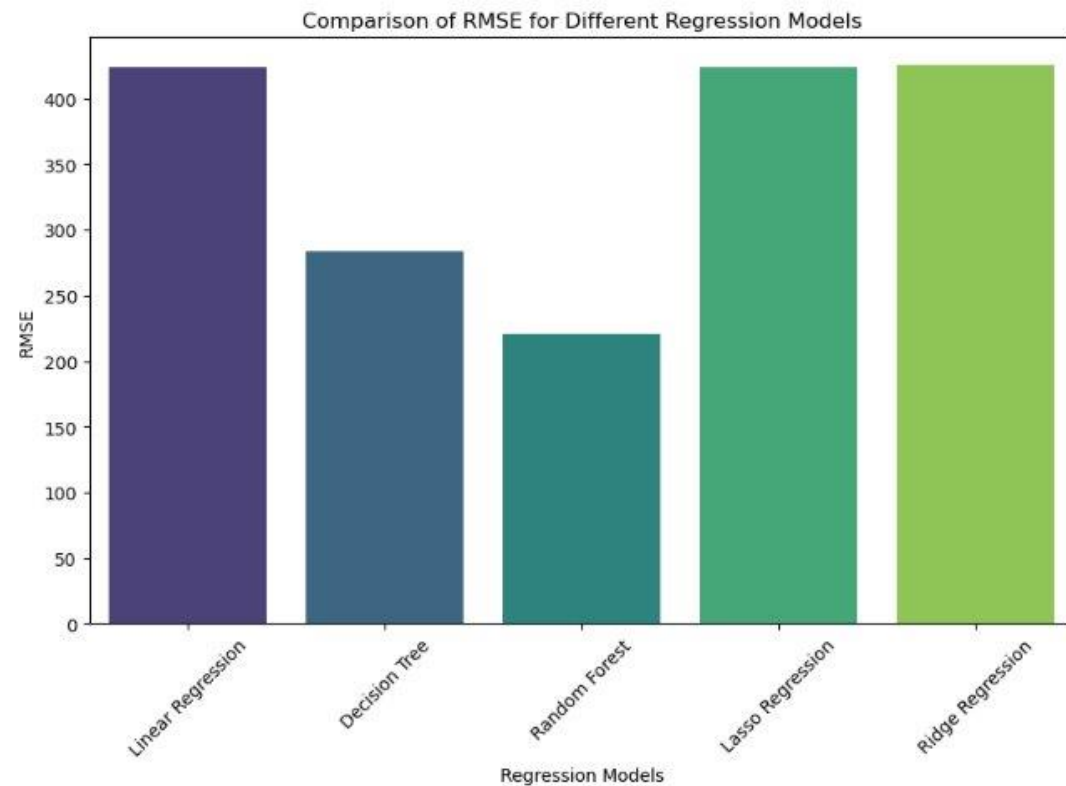


LET'S DIVE INTO NUMBERS!

- **Random Forest** demonstrates the **best** performance, exhibiting the **lowest RMSE & highest R-squared** value, providing the most accurate predictions for bike rental counts.
- **Decision Tree** follows closely with strong performance.
- **Linear** Regression, **Lasso**, & **Ridge** models show comparatively **lower accuracy** in predicting bike rental counts.

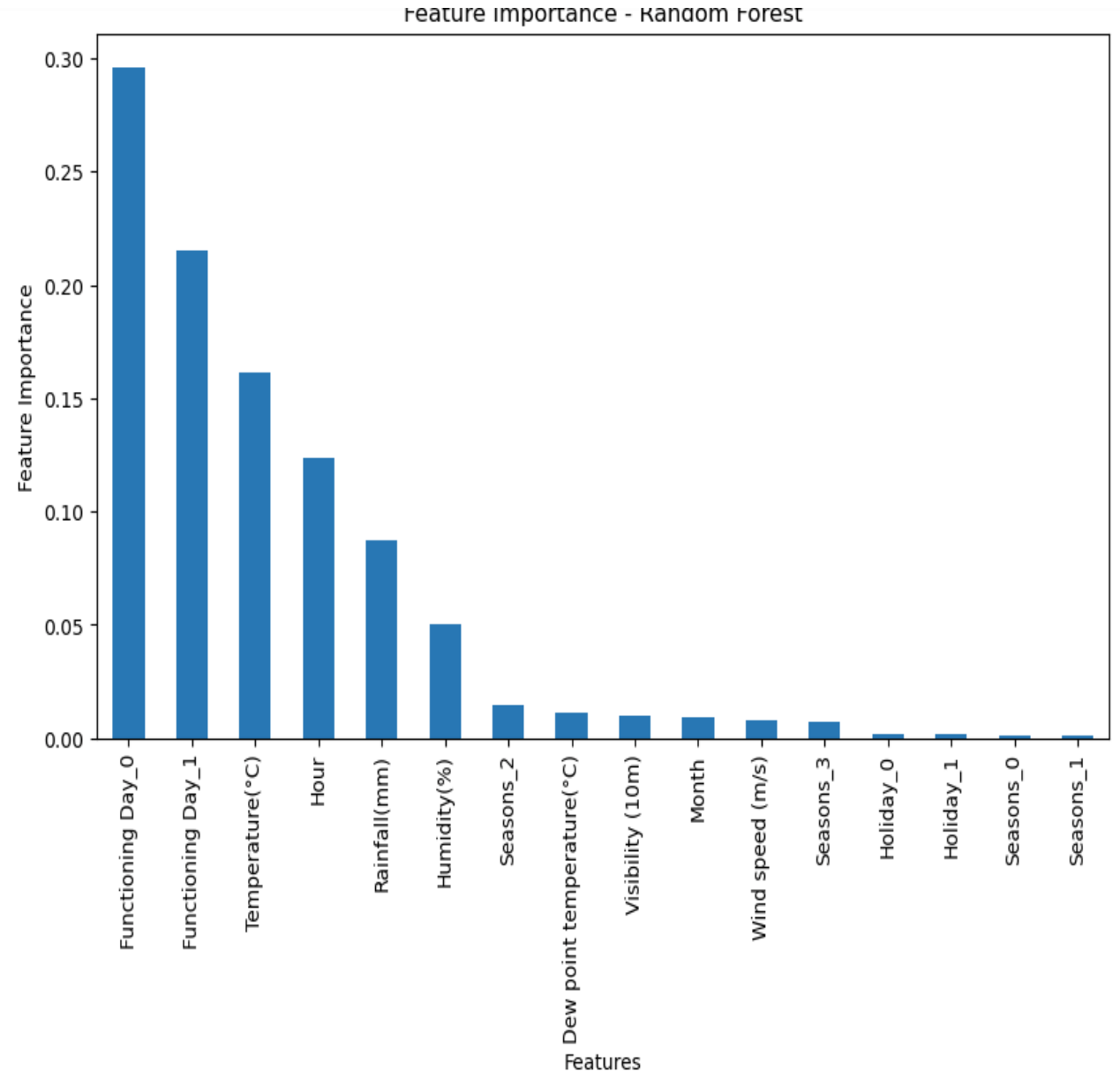
Model	RMSE	R-Square
Linear	7.056	0.67
Decision Tree	4.793	0.85
Random Forest	3.804	0.90
Lasso	7.066	0.67
Ridge	7.027	0.67

COMPARISON BETWEEN MODELS



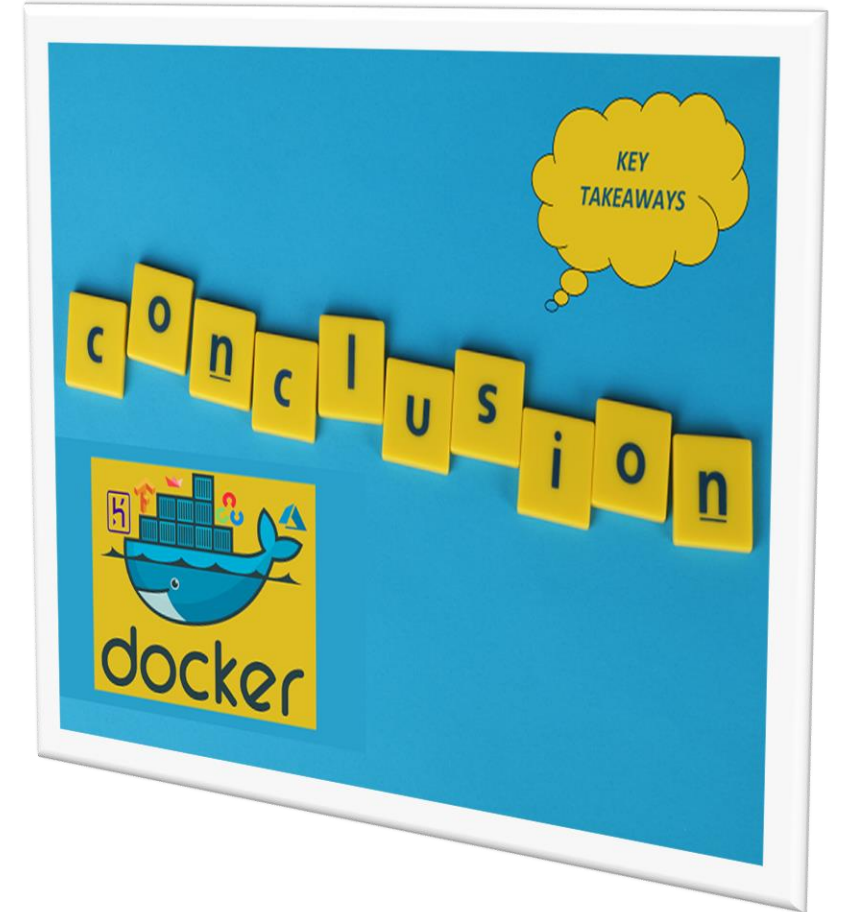
FEATURE IMPORTANCE - RANDOM FOREST

- **Key Influential Factors:-**
- **Functioning Day, Temperature and Hour** are the **top 3** features with the highest importance scores, indicating their significant impact on bike rental demand.
- Rainfall & Season_2(Autumn) also play crucial roles in predicting bike rental counts.
- **Secondary Influences:-**
- **Visibility & Wind Speed** have moderate importance, suggesting their secondary influence on bike rental patterns.
- Seasonal variables, particularly Seasons_3 (winter), contribute to rental predictions to a lesser extent.
- Holiday & **certain seasonal** categories exhibit **minimal** importance in predicting bike rental counts.



KEY TAKEAWAYS

- **Predictive Power:**
 - Machine learning models were employed to forecast bike rental demand.
- **Best Performer:**
 - Random Forest emerged as the most effective model.
- **Influential Factors:**
 - Functioning day, Temperature and Hour revealed as key drivers of bike rental demand.
- **Strategic Insights:**
 - Informed operational decisions, marketing strategies, & resource allocation for bike-sharing services, enhancing customer satisfaction & operational efficiency.





THANK YOU