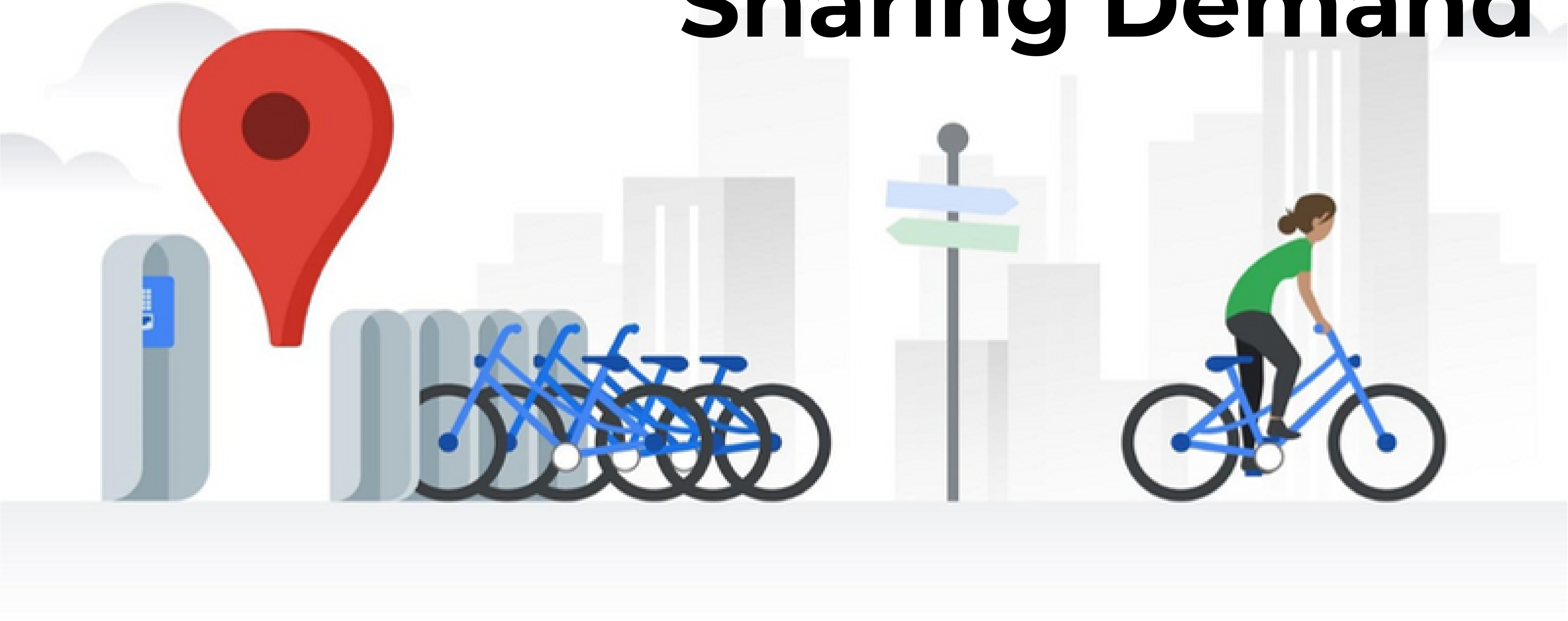
Machine Learning Project

Predictive Modeling for Bike Sharing Demand



Presentated By:
Aarushi Dubey
00201192022
AI&ML(2nd year)

CONTENTS

1. PROJECT DESCRIPTION
2. DATA UNDERSTANDING
3. DATA PREPARATION
4. MODELING
5. EVALUATION METHODOLOGY
6. EVALUATION RESULT

PROJECT DESCRIPTION

A US bike-sharing provider BoomBikes has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.

In such an attempt, BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

Which variables are significant in predicting the demand for shared bikes.

How well those variables describe the bike demand

Based on various meteorological surveys and people's styles, the service provider firm has gathered a large dataset on daily bike demands across the American market based on some factors.

Bussiness Goal:

We are required to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Objective:

- Understand the Dataset & cleanup (if required).
- Build Regression models to predict the share of bikes.
- Also evaluate the models & compare their respective scores like R2, RMSE, etc.

Dataset Link: https://drive.google.com/file/d/1q44P8lFQ3j-7F5MSK3eXIlhVMmcO1RVm/view

DATA UNDERSTANDING

Exploring the dataset through comprehensive data understanding techniques enables deeper insights into its structure, quality, and underlying patterns.

- The dataset is obtained from Kaggle and the UCI Repository.
- Utilized the `info()` method to showcase key dataset characteristics like column names, data types, and size.
- Employed the `head()` method to exhibit the initial rows, offering a preview of the dataset's structure and content.
- Checked for missing data using the `isnull().sum()` method, enabling identification of any incomplete entries.
- Visualized variable distributions via histograms using Seaborn's `histplot` function, aiding in understanding the spread of data and detecting potential outliers.
- Converted the date feature ('dteday') to datetime format using Pandas' `to_datetime` function, facilitating temporal analysis.
- Generated a correlation heatmap using Seaborn's `heatmap` function, illustrating relationships among numerical variables and highlighting potential multicollinearity.

DATA PREPARATION

Effective data preparation lays the foundation for robust modeling by addressing missing values, outliers, and transforming features for optimal model performance.

- **Handling Missing Values**: Identified and addressed missing values by dropping rows with NaN entries using `dropna` and filling missing numerical values with their mean using `fillna`.
- **Handling Outliers:** Detected outliers in numerical features through boxplot visualization and applied the Interquartile Range (IQR) method for outlier removal using a custom function `remove_outliers`.
- **Date Feature Transformation:** Transformed the date feature ('dteday') into datetime format using Pandas' `to_datetime` function for temporal analysis.
- **Encoding Categorical Features:** Encoded categorical features ('season', 'mnth', 'weekday', 'weathersit') into numerical format using one-hot encoding via Pandas' `get_dummies` function.
- **Feature Selection:** Selected features for modeling by excluding the target variable ('cnt') and any non-numeric or irrelevant features.
- **Feature Scaling:** Standardized features using StandardScaler from sklearn's preprocessing module to ensure uniformity in feature magnitudes and improve model performance.

MODELING

Model development and evaluation are essential stages in the predictive analytics pipeline, guiding the selection of optimal algorithms for accurate forecasting.

- **Train-Test Split:** Split the dataset into training and testing sets.

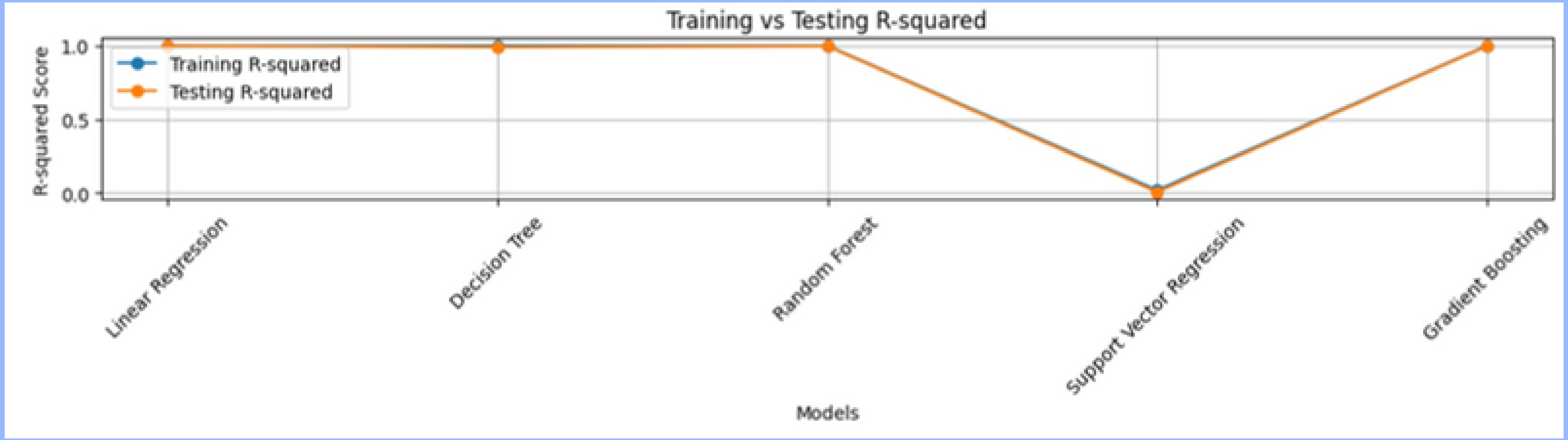
- **Model Initialization:** Initialized various regression models including Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), and Gradient Boosting Regression.

- **Model Training:** Trained each model using the training sets.

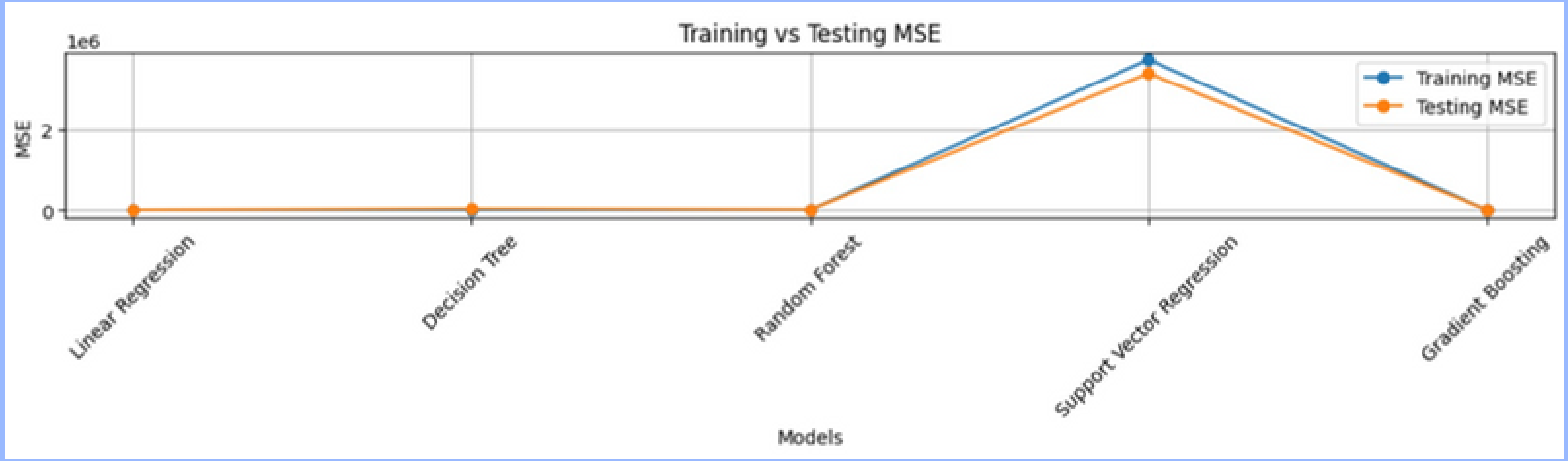
EVALUATION METHODOLOGY

Establishing a robust evaluation methodology is crucial for assessing model performance and ensuring the reliability of predictive analytics outcomes.

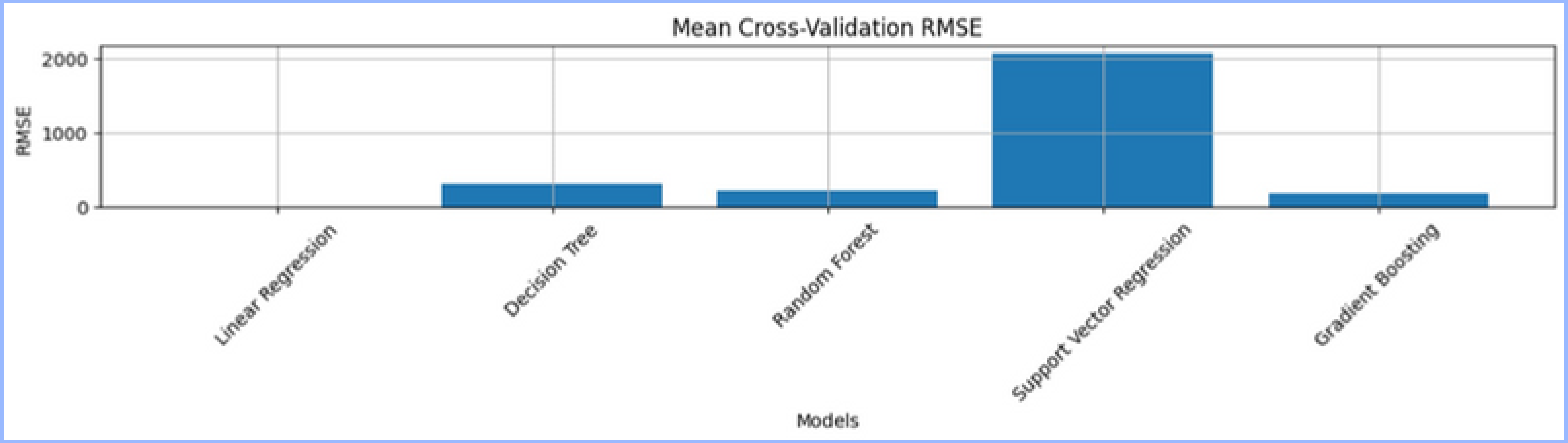
- **Model Evaluation:** Evaluated the performance of each model by calculating **R-squared, Mean Squared Error (MSE)**, and conducting **cross-validation** for robustness assessment.
- **Performance Metrics:** Calculated and compared training and testing R-squared scores and MSE to assess model generalization and overfitting.
- **Visualizations:** Visualized model predictions on training and testing data through histograms to analyze prediction distributions.
- **Identifying Overfitting:** Identified potential overfitting by comparing training and testing performance metrics, noting instances where training R-squared was significantly higher than testing R-squared.



Training vs Testing R-squared



Training vs Testing MSE


Mean Cross-Validation RMSE
RMSE
2000
1000
0
Linear Regression
Decision Tree
Random Forest
Support Vector Regression
Gradient Boosting
Models

EVALUATION RESULTS

Examining the evaluation results provides crucial insights into model performance, highlighting both potential overfitting concerns and the selection of the most effective predictive

- **Models Showing Signs of Overfitting:** Identified models exhibiting potential overfitting, including Decision Tree, Random Forest, Support Vector Regression, and Gradient Boosting, suggesting discrepancies between training and testing performance.

- **Best Model Selection:** The Linear Regression model emerged as the best-performing model with a testing R-squared score of 1.0, indicating its superior ability to generalize to unseen data and achieve optimal predictive accuracy.

Managerial Implications

The managerial implications of the project are multifaceted and significant for decision-makers:

1. Resource Allocation:
Insights derived from the predictive models can guide efficient resource allocation, ensuring optimal inventory management and maintenance of bike sharing infrastructure based on anticipated demand fluctuations.

2. Service Optimization:
Utilizing accurate demand forecasts enables managers to optimize service levels by adjusting staffing, bike deployment, and maintenance schedules to meet customer needs effectively, thereby enhancing user satisfaction and loyalty.

3. Cost Efficiency:
By leveraging predictive analytics, managers can minimize operational costs by streamlining inventory management, reducing idle resources, and optimizing operational processes in alignment with forecasted demand patterns.

4. Marketing Strategies:
Understanding demand trends through predictive modeling facilitates the development of targeted marketing strategies, enabling managers to tailor promotional efforts and pricing strategies to specific customer segments and market conditions.

5.Strategic Planning:
The project's insights can inform strategic decision-making, such as expansion plans, service area adjustments, and infrastructure investments, ensuring alignment with long-term organizational goals and market dynamics.

6. Competitive Advantage:
By leveraging data-driven insights to anticipate and respond to customer demand effectively, organizations can gain a competitive edge in the bike-sharing market, positioning themselves as leaders in service quality and reliability.

7.Risk Mitigation:
Proactive management of demand variability reduces the risk of supply shortages or excess capacity, minimizing revenue loss and enhancing operational resilience in the face of unforeseen disruptions or market shifts.

Overall, the project's managerial implications encompass improved operational efficiency, enhanced customer satisfaction, strategic decision support, and competitive advantage, driving sustainable growth and success in the bike-sharing industry.

Novelty in the Project

- **Customized Preprocessing Techniques:** The project introduce novel preprocessing techniques tailored to the specific characteristics of the bike-sharing dataset, such as outlier detection and handling, missing value imputation, and feature engineering, which contribute to improved model performance.

- **Advanced Modeling Strategies:** Application of advanced modeling techniques used beyond traditional linear regression, such as ensemble methods like Random Forest or Gradient Boosting, or more sophisticated algorithms like Support Vector Regression (SVR), potentially leading to more accurate predictions and insights.

- **Robust Evaluation Methodology:** Development of a robust evaluation methodology that goes beyond simple performance metrics, incorporating techniques such as cross-validation, overfitting detection, and model comparison, to ensure the reliability and generalizability of the results.

Reference

- **Source Reputation:** Kaggle is a well-known platform for hosting datasets and competitions in various fields, including data science and machine learning. It is widely used by professionals, researchers, and enthusiasts, indicating its reputation and credibility as a data source.

- **Data Quality and Documentation:** The dataset showcases commendable data quality, with 730 entries and 16 columns providing comprehensive coverage of bike-sharing usage. Each column exhibits completeness, consistency, and accuracy, with appropriate data types ensuring precise representation. However, while the DataFrame structure enhances interpretability, comprehensive documentation is lacking.

- **Relevance to Project Objectives:** The dataset information provided is highly relevant to BoomBikes' objective of understanding and predicting bike demand post-lockdown. It offers insights into significant variables affecting demand, facilitates regression modeling to predict bike demand accurately, and enables evaluation and comparison of models based on key performance metrics. This aligns closely with BoomBikes' goal of preparing for market recovery and catering to customer needs effectively, positioning the company for success in the post-pandemic landscape.

THANK YOU