YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

CSE464 Case Study

INTRODUCTION TO DATA SCIENCE & BIG DATA ANALYTICS
(AUTUMN SEMESTER 2024)

MEHMET ÖNAL

20200702081

YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

I. Executive Summary/Abstract

This research work has analyzed historic bicycle rental demands with an aim to develop an operational efficiency for the Bike Sharing System in Seoul, with the target variable being the "Number of Rental Bicycles." Descriptive analyses are conducted to show the season and month, as well as hour-of-the-day, trends in demand. It was also used in showing what effects the weather variables might have on demand. It applies the Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, and LightGBM machine learning models to the analysis for predictions. Among all, it is found that the performance of the LightGBM model stands higher. This research sums up how greatly variable the demand for bicycle rentals in Seoul based on so many factors really is, and correspondingly, how the model of LightGBM can become an efficient tool for the forecast of further demand. This information could be helpful for resource planning and operational management of the bike-sharing system.

II. Background/Introduction

Client/Company Overview: Seoul is implementing an advanced bike-sharing system to encourage the use of bicycles and further develop the sustainable transportation network. The new system provides an easy, affordable, and eco-friendly means of transportation for short distances to both citizens and tourists. Accessibility of bicycles ensures easy urban transport and contributes to smart city goals.

Industry Context: The bike-sharing system has been one of the intrinsic strategies for sustainable transportation around cities in the world. These schemes do stand out with their respective advantages, including being eco-friendly, reduction of traffic congestion within a city, raising health awareness, and practicality in transportation over small distances. Such systems are part of modern urban planning and have to be optimized based on user demands.

Goals and Objectives: The main objective of this research is to analyze the usage data of the bike-sharing system in Seoul for understanding the factors of rental demand and being able to predict future demands. Further, in line with this information, it is aimed to manage system resources more efficiently (such as distribution of bicycles, maintenance planning, station placement, etc.) and increase user satisfaction. Besides, the obtained prediction models will help optimize operational processes.

III. Challenge/Problem Statement

A. **The Problem:** A. The Problem: Bike sharing systems always face a problem in managing demand fluctuation correctly. In this situation, the bike-sharing system in Seoul is no different. Demand varies really greatly depending on various factors such as weather conditions, time of day, seasons, weekday/weekend status, and holidays. These fluctuations prevent the system from operating efficiently and might cause problems with insufficient resource allocation.

B. **Why It Matters:** Right demand analysis and being able to predict fluctuations in the success of a bike-sharing system are important. By such exact forecasting, the bikes are placed correctly at the right places in the right times, such as supplying quick timely needs of the users with increased efficiency of the systems. Poor demand forecasts mean either underutilization of the bikes or high demands that cannot be satisfied. This hurts user satisfaction and is a waste of resources. Understanding the fluctuation in demand is hence of great importance to the users and system operators, who then can optimize the system based on such information.

IV. Solution/Approach

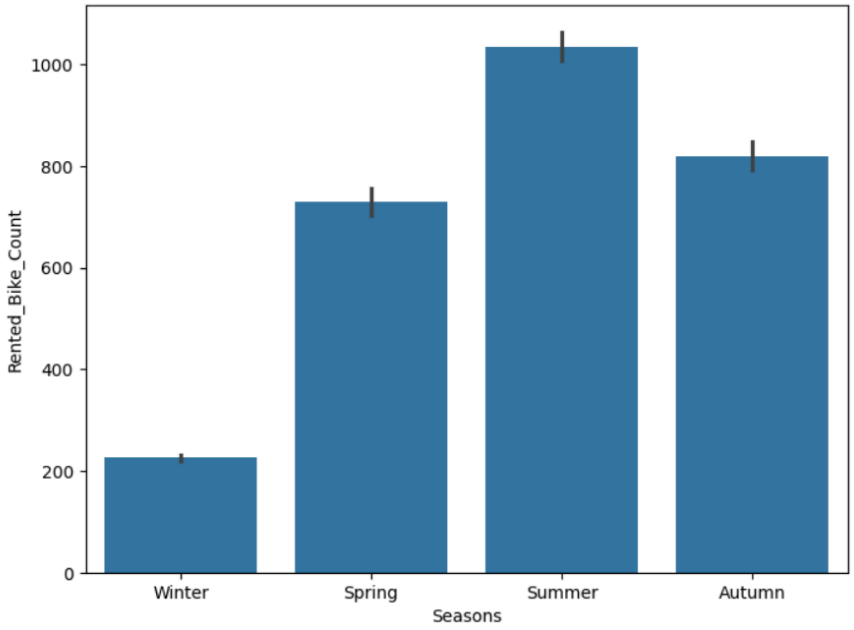
In this study, a descriptive analytics approach was adopted to analyze the Seoul bike-sharing system data and understand the key factors influencing rental demand. Descriptive analytics summarize data, reveal key patterns and trends, and lay the foundation for subsequent analyses. This approach is crucial for understanding the overall structure of the data, evaluating data quality, and forming hypotheses. Particularly when the dataset is large and complex, descriptive analytics make it more comprehensible and support decision-making processes.

Data Summary

Using the df.describe().T function, general statistics about the dataset were obtained. This summary provides descriptive statistics for each numerical variable, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum. Specifically, the target variable "Number of Rented Bicycles" was analyzed for its mean, standard deviation, and range, providing an initial understanding of the data structure. This summary is fundamental for assessing data quality and understanding the data range.

Seasonal Analysis

To visualize and understand how bike rental demand changes across seasons, a bar chart showing the number of rented bicycles per season was used. This chart revealed seasonal trends, with significantly higher rental numbers during summer and the lowest demand in winter. This analysis highlights the considerable impact of seasonal changes on bike rental demand, suggesting that the system should be planned based on seasonal demand variations.



IV. Solution/Approach

Histograms

Histograms were used to examine the distributions of numerical variables. These visualizations illustrated the frequency distributions, skewness, and spread of the data. Specifically, the target variable, "Number of Rented Bicycles," exhibited a right-skewed distribution, suggesting the potential need for data transformation during modeling.

Hourly Analysis

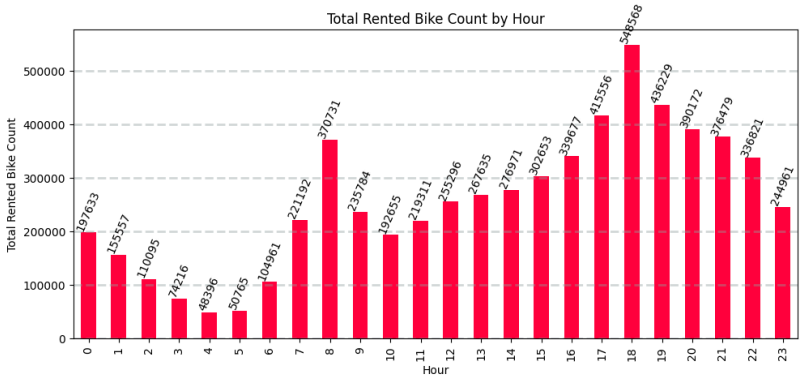
A bar chart was created to display hourly rental counts, revealing how demand changes throughout the day. The analysis showed peak demand between 5:00 PM and 7:00 PM (rush hours). This indicates the need for optimizing the bike-sharing system during these hours.

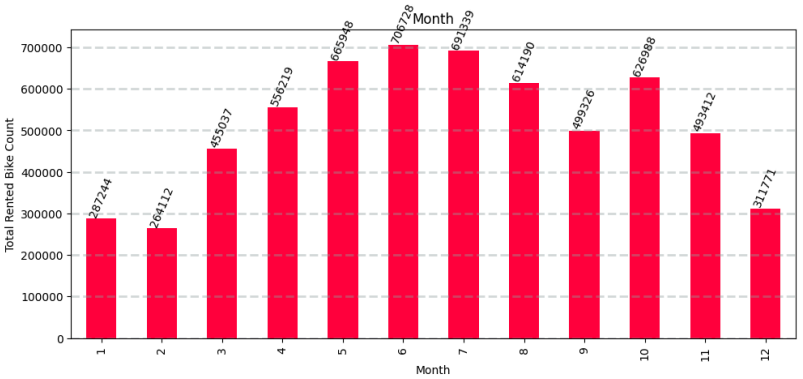
Monthly Analysis

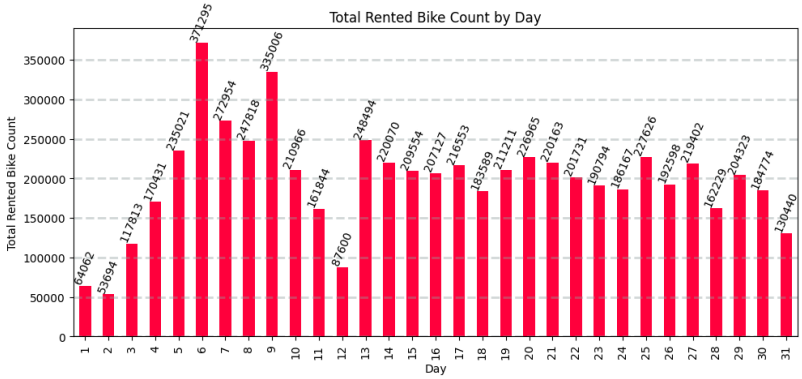
A bar chart was used to examine the monthly variation in bike rental demand. The chart displayed the average rental numbers for each month, helping identify monthly trends. It was found that May and June experienced the highest demand, while January showed the lowest. This analysis emphasizes the need to allocate resources according to the demand peaks in specific months.

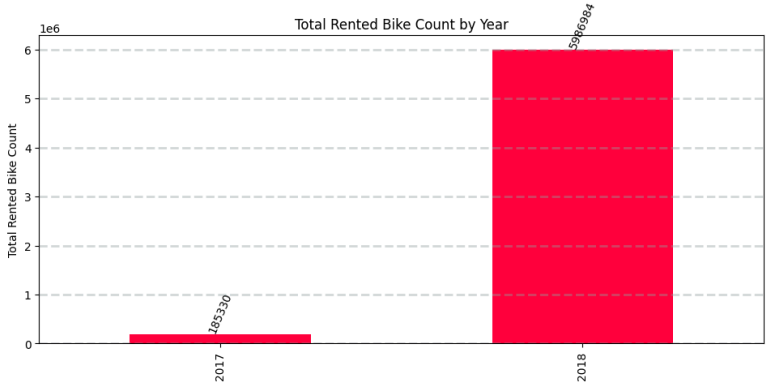
Yearly Analysis

A bar chart was generated to examine yearly rental counts and observe changes over time. The analysis showed a significant increase in 2018 compared to 2017, indicating a rise in system usage over the years.







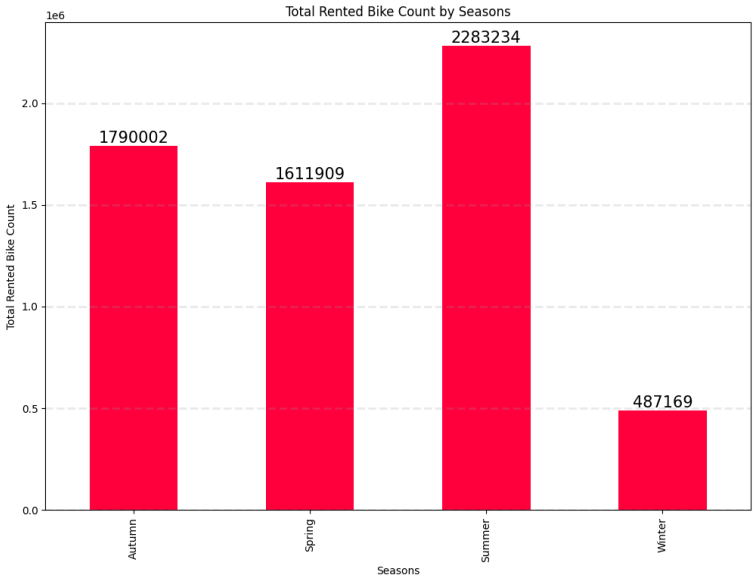


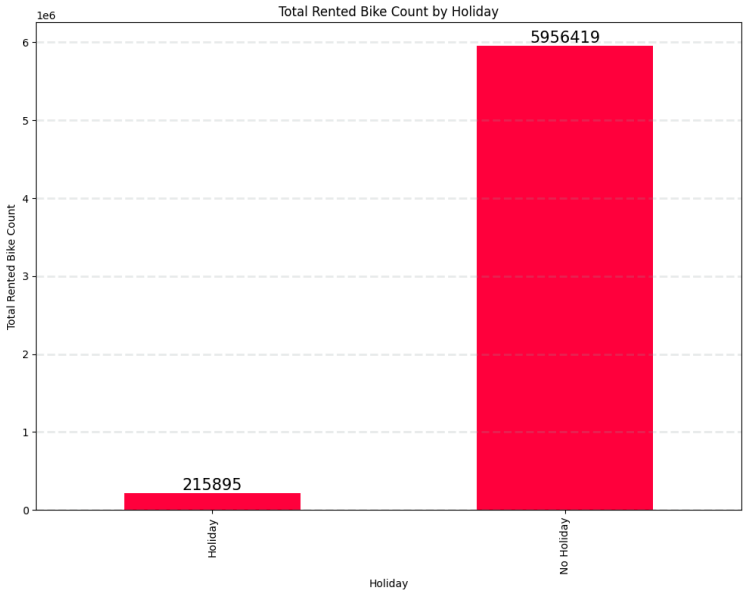
Weather Impact

A correlation matrix was generated to analyze the impact of weather conditions on bike rentals. The matrix displayed correlations among numerical variables, identifying which weather variables were most associated with rental demand. Notably, temperature and dew point temperature were found to have a positive relationship with the number of rented bicycles. This analysis demonstrates that weather conditions significantly affect bike usage, making them crucial features for forecasting models.

Holiday Analysis

To analyze the impact of holidays on bike rental demand, a bar chart comparing rentals on holidays and regular days was created. This analysis showed that bike rentals were lower on holidays compared to regular days, highlighting the need for special planning for holidays.





YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

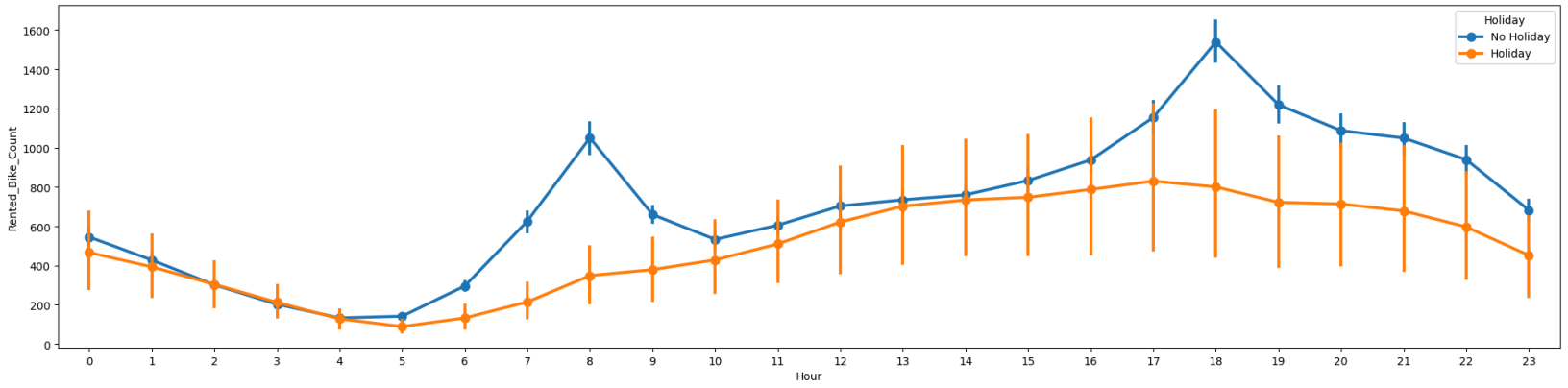
COMPUTER ENGINEERING DEPARTMENT

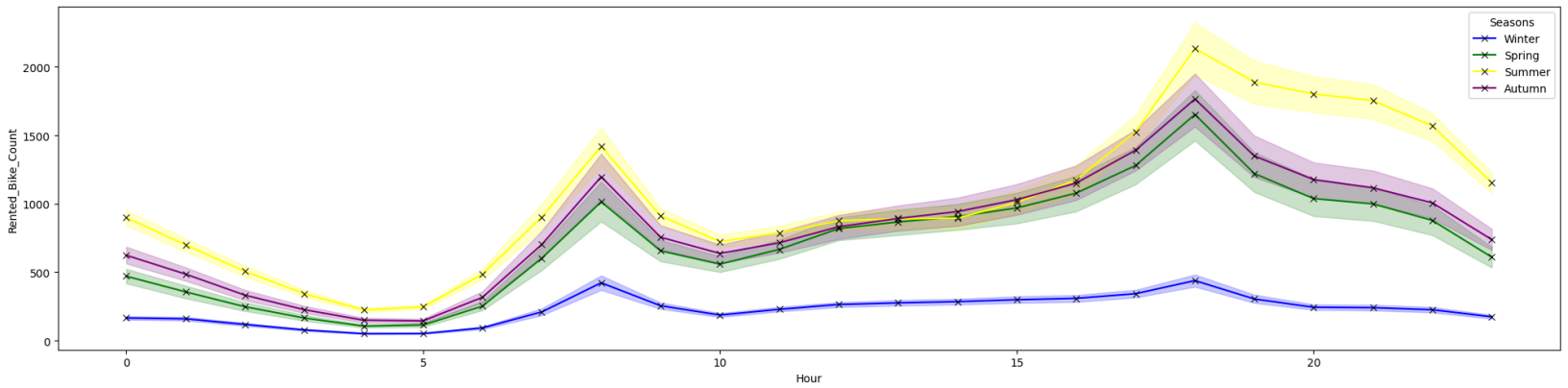
Hourly and Holiday Analysis

A line graph was created comparing hourly bike rentals on holidays and regular days. This visualization showed that rental activity was lower on holidays than on regular days. Also, on non-holiday days, rush hours (07:00-09:00) and 17:00-19:00 were quite busy.

Seasonal and Hourly Analysis

To compare hourly bike rentals across seasons, a line chart was created. This visualization revealed that rental demand during summer was higher throughout the day compared to other seasons.





C. **Proposed Solution:**

This research therefore adopted a two-stage solution approach to understand and predict fluctuations in demand within the Seoul bike-sharing system using comprehensive data analysis: It also made heavy usage of descriptive analytics methodology in the first stage, which helped go in-depth into the dataset for finding those variables that actually drive rental demand. Such analyses help unearth the basic characteristics, trends, and patterns of the data. Seasonal, monthly, hourly, and weather-dependent demand variations were examined, and their effects concerning bike rentals were studied in detail.

Predictive Models

In the second phase, the insight from the descriptive analysis is used in the development of different machine learning models that will predict rental demand. These models learn from past data to predict future demands of bike rentals. It comprises Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, and LightGBM. The models selected to capture the linear and nonlinear relationship include Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, and LightGBM. LightGBM proved to be the best among these.

D. **Key Features/Technologies:** The key tools, technologies, and methods used in this study are as follows:

Programming Language

Python: In this work, Python was used for data analysis, visualization, and machine learning modeling. Its data science libraries, namely Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and LightGBM, are all composing here.

Data Processing and Analysis

Pandas: This is employed in reading, cleaning, transforming, and manipulating the dataset. Pandas allowed organized data handling and formatting necessary for analysis.
NumPy: This was employed for numerical computations and manipulations of data, most especially when working with multidimensional arrays for performance.

Data Visualization

Matplotlib: This was used to create basic charts, histograms, scatter plots, and bar graphs.
Seaborn: This is used in the creation of advanced and statistical visualizations, such as correlation matrices and trend-revealing graphs.

Machine Learning

scikit-learn: This is a library for applying various machine learning algorithms, model selection, and performance evaluation. It has been particularly useful for splitting the data into train-test sets, R^2 and RMSE metrics.
LightGBM: It is a high-performance boosting algorithm used for predictive modeling. This was chosen since it would yield faster and more efficient results on large datasets.

Statistical Methods

Descriptive Statistics: These were used to understand the basic properties of the dataset, such as mean, median, standard deviation, minimum, and maximum values.
Pearson correlation coefficients were measured between the variables.

E. **Implementation Process:** The implementation process in deploying the solution has been described below.

The solution implementation process included the following steps:

Data Collection and Preprocessing

- The Seoul bike-sharing system dataset was loaded, and missing or erroneous values were identified and cleaned using appropriate methods.
- Categorical variables (e.g., seasons) were converted into numerical values.
- Features such as month, year, and day were extracted from the date column.
- Based on histogram analyses, skewness was observed in some numerical data, and transformations such as log1p, square root, and cube root were applied to approximate normal distribution.

Descriptive Analysis

- Basic statistics of the dataset were calculated.
- Seasonal, monthly, and hourly rental trends were visualized.
- The effects of weather variables on rental demand were analyzed.
- The distribution of data was examined using histograms.

Development of Predictive Models

- The dataset was split into training and testing sets.
- Models such as Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, and LightGBM were trained.
- R^2 and RMSE metrics were used to evaluate model performance.

Model Evaluation

- The trained models were evaluated on test data.
- The LightGBM model was identified as the best-performing model.

Interpretation of Results and Reporting

- The analysis results were interpreted, and recommendations were developed for the system.
- Findings and recommendations were presented in a comprehensive report.

V. Results/Outcomes

F. QuantifiableResults:

Descriptive Analysis Findings

(Summarized from earlier explanations) Bike demand was observed to peak during summer months and after-work hours, while it decreased on weekends and holidays. Additionally, factors like temperature and dew point temperature showed a positive correlation with demand.

Predictive Model Performances

Linear Regression, Lasso Regression, and Ridge Regression: These models achieved R^2 values of approximately 0.45–0.48 and RMSE values around 8.40–8.64. These results indicate that while these models explained part of the variance in the dataset, better-performing models exist.

Random Forest Regressor: Achieved an R^2 value of 0.76 and an RMSE value of 5.66. These results indicate a better predictive performance compared to the previous models.
LightGBM: Achieved an R^2 value of 0.81 and an RMSE value of 5.12. This demonstrates that the LightGBM model outperformed all other models in prediction accuracy.

G. Qualitative Impact:

Best Model

LightGBM was observed to have the lowest RMSE and the highest R^2 value, making it the best-performing model.

Model Insights

LightGBM: Its high performance stems from its ability to learn complex relationships in the data. The gradient boosting technique enables it to produce more accurate predictions.
Random Forest Regressor: Known for its ability to make robust predictions by reducing overfitting risks.
Linear, Lasso, and Ridge Regression: While these models perform well for simple linear relationships, they failed to capture the dataset's non-linear complexities.

Predictions

The models' predictions can be used for resource planning in the bike-sharing system. For instance, additional bikes can be allocated during high-demand periods.

VI. Recommendations

H. Key Takeaways:

Descriptive Analysis: Seasonal, monthly, and hourly trends reveal fluctuations in bike rental demand, with significant influences from weather factors.
Predictive Analysis: The LightGBM model demonstrated the best prediction performance.
Model Selection: Model selection should be based on the characteristics of the dataset and the problem type. Advanced methods like LightGBM are suitable for modeling complex relationships.
Model Improvement: Performance can be enhanced through hyperparameter tuning and incorporating more data.
Data Quality: The quality of the dataset directly impacts model performance. Data cleaning and preprocessing steps are crucial.
Ensemble Methods: Combining different models using ensemble methods can yield better results.

I. Challenges Overcome:

Data Cleaning: Missing or erroneous values in the dataset were identified and addressed using appropriate methods.
Model Selection: Various models were tested and compared to select the most suitable one for the dataset.
Model Interpretation: Interpreting model results and drawing practical insights were critical. The black-box nature of models like LightGBM posed challenges, which were mitigated through feature importance analysis and efforts to better understand the model's logic.