Forecasting Bicycle Rentals

Project Report

Project Report on

Forecasting Bicycle Rentals
Course: Data Mining(CMPE255)
MS Software Engineering(Fall 2023)

Submitted by:
Team TrailBlazers
Ruchitha Reddy Koluguri (017434534)
Harshavardhan Valmiki (017442984)
Suresh Ravuri (017451824)

Guided by:
Prof. Vijay Eranti
Academic Year: 2023

ABSTRACT

Bike-sharing programs are becoming increasingly popular in cities worldwide, including in Washington D.C., USA. These programs allow people to borrow bikes for short trips, but predicting the exact demand for bikes in various locations and times is a significant challenge. Our study aimed to address this issue by developing methods to estimate the number of bikes needed across the city at different times.

The demand for bikes in a bike-sharing program can fluctuate widely due to factors such as weather and time of day. Accurately predicting this demand is crucial for user satisfaction and operational efficiency. To tackle this challenge, our study employed data analysis and machine learning techniques, which are computer methods that learn from data. We gathered and processed extensive data, including weather conditions and peak usage times, and applied various computer models to forecast bike usage.

Our methodology involved using advanced machine learning models like Random Forest, AdaBoost, and XGBoost, known for their ability to interpret complex data patterns. The study's findings were revealing; the AdaBoost model was the most accurate in predicting bike usage, though XGBoost proved advantageous in reducing prediction errors. These insights are valuable for bike-sharing program operators, aiding in better decision-making for bike allocation.

Our research has wider implications for urban planning, showing how data and technology can improve city transportation systems. This study is particularly relevant for enhancing bike-sharing programs and overall urban mobility in cities like Washington D.C. It highlights the significance of using technology and data analytics for smarter city management and sustainable urban living.

1. INTRODUCTION

In recent years, bike-sharing programs have surged in popularity across global cities, transforming urban mobility and presenting unique challenges and opportunities. This report delves into the intricate world of bike-sharing, a domain where technology and urban planning converge, creating a dynamic landscape that shapes the way people navigate cityscapes.

At the heart of our investigation is the compelling challenge of forecasting bike-sharing demand. This task is not just a statistical exercise but a crucial step towards optimizing resource allocation, enhancing user experience, and promoting sustainable urban living. By accurately predicting how many customers will rent bikes under varying conditions, we can ensure that bike-sharing remains a reliable, efficient, and environmentally friendly mode of transportation.

Our approach is grounded in the meticulous analysis of time-series data, encompassing variables like temperature, weather conditions, and wind speed. This dataset, a rich tapestry of daily rental activities, offers a window into the patterns and preferences of bike-sharing users. By dissecting this data, we aim to uncover the nuanced interplay between environmental factors and user behavior, providing actionable insights for bike-sharing operators.

Our findings paint a vivid picture of the factors influencing bike-sharing demand. We discovered a significant relationship between weather conditions and rental patterns, with customer numbers peaking during favorable weather. Our advanced predictive models, leveraging techniques like Random Forest and ensemble methods, offer a high degree of accuracy in forecasting demand. These models are more than statistical tools; they are a roadmap for efficient bike-sharing management.

In conclusion, this report is not just an academic exercise but a step towards a future where bike-sharing is an integral, well-orchestrated part of urban life. By harnessing the power of data, we aim to contribute to the development of smart, sustainable cities where mobility is a right, not a privilege.

2. RELATED WORK

For this project, We have used EDA and partial CRISP-DM Methodology.

In the field of demand forecasting for bike-sharing systems, established studies predominantly utilize time-series analysis, with a focus on linear regression models and basic statistical approaches. These methods effectively identify trends and seasonality in bike rental data, providing foundational insights into usage patterns influenced by environmental and temporal factors.

Our project builds upon this base by incorporating advanced machine learning techniques, notably ensemble methods like Random Forest, XGBoost, AdaBoost, and CatBoost. This selection of models, particularly the use of CatBoost for handling complex categorical data, represents a significant departure from the traditional time-series forecasting methods. Ensemble methods, by aggregating predictions from multiple models, offer enhanced accuracy and reliability in forecasts.

Additionally, our project places significant emphasis on rigorous data preprocessing, including techniques such as differencing, detrending, and transformations to address non-stationarity in time-series data. This level of detail in data preparation is not a standard practice in much of the existing literature.

In terms of model evaluation, our project aligns with common practices by using metrics such as MAE, MSE, RMSE, and R-squared. These metrics are critical in assessing the effectiveness of our predictive models, paralleling the methodologies used in similar studies.

Overall, our study adheres to the broader trend of employing time-series analysis in bike-sharing systems but distinguishes itself through a multifaceted approach that combines complex machine learning algorithms with detailed data preprocessing and a thoughtful selection of model evaluation metrics. This methodology aims to yield a more nuanced and precise forecast of bike rental demand.

3.DATA

3.1 Dataset

The dataset was sourced from OpenML, It is a continuous time-series data

➢ https://www.openml.org/search?type=data&status=active&id=43486&sort=runs

The dataset is a detailed time-series data collection with 2922 entries, each representing a single day from January 1st, 2011 to December 31st, 2018. It includes 23 features that capture various attributes relevant to each day, enabling a thorough day-to-day analysis over these eight years. This structure allows for an in-depth exploration of trends and patterns within the time-series framework.

Temperature is a key component of the dataset, represented through multiple attributes:

- **temp_avg**: the average temperature for the day.
- **temp_max**: the highest temperature recorded on that day.
- **temp_min**: the lowest temperature for the day.
- **temp_observ**: specific temperature observations made on that day.

Precipitation levels are captured in the **precip** column, providing insights into the amount of rainfall or snowfall. Similarly, wind records the wind speed, adding another layer of environmental context to each day's data.

Weather conditions are meticulously detailed through a series of **wt_** attributes. These columns record various weather phenomena, with binary indicators (0 or 1) signifying the absence or presence of a specific condition. For instance, **wt_rain** and **wt_fog** denote the occurrence of rain and fog, respectively. A day with **wt_rain** marked as 0 and **wt_fog** as 1 would indicate a foggy day without rain.

The dataset also distinguishes between two types of users:

- casual users, who aren't registered with the service.
- registered users, who have accounts.

The key variable in the study, total_cust, represents the combined total of casual and registered customers renting bicycles on a given day. This metric is central to understanding overall bicycle demand. The dataset also includes a holiday flag, highlighting days that are holidays, as these can notably affect rental patterns. The main goal is to forecast total_cust, aiming to accurately predict bicycle demand. This involves examining how factors such as temperature, weather conditions, wind speed, and holidays influence bike rental behavior among

different user groups. Accurate forecasts are essential for effective demand management and operational planning.

3.2 Data Pre-processing

3.2.1 Data Cleaning:

In the data cleaning phase, null values in the time-series data were handled carefully due to the continuous nature of the dataset. Missing values were filled using methods like forward filling for customer data, assuming closeness in time. For weather features, zeroes replaced missing values, appropriate given their binary nature. For average temperature, it was initially calculated as the mean of maximum and minimum temperatures. Plans were made to develop a more accurate method using a linear model, with temp_max and temp_min as inputs and temp_avg as the output, to fill in missing values.

```
df_v2
```

	date	temp_avg	temp_min	temp_max	temp_observ	precip	wind	casual	registered	total_cust	holiday	rain	fog	ice	datetime	year
0	2011-01-01	5.20	-1.57	11.97	2.77	0.07	2.58	330.0	629.0	959.0	0.0	1	1	0	2011-01-01	2011
1	2011-01-02	7.34	0.88	13.81	7.33	1.04	3.92	130.0	651.0	781.0	0.0	1	1	0	2011-01-02	2011
2	2011-01-03	2.01	-3.44	7.46	-3.06	1.88	3.62	120.0	1181.0	1301.0	0.0	0	0	0	2011-01-03	2011
3	2011-01-04	-0.66	-5.96	4.64	-3.10	0.00	1.80	107.0	1429.0	1536.0	0.0	0	0	0	2011-01-04	2011
4	2011-01-05	0.91	-4.29	6.11	-1.77	0.00	2.95	82.0	1489.0	1571.0	0.0	0	0	0	2011-01-05	2011
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2917	2018-12-27	3.50	-3.59	9.12	-1.06	0.02	2.10	1150.0	4280.0	5430.0	0.0	0	1	0	2018-12-27	2018
2918	2018-12-28	8.23	0.61	11.21	8.09	16.84	2.00	166.0	1959.0	2125.0	0.0	0	1	0	2018-12-28	2018

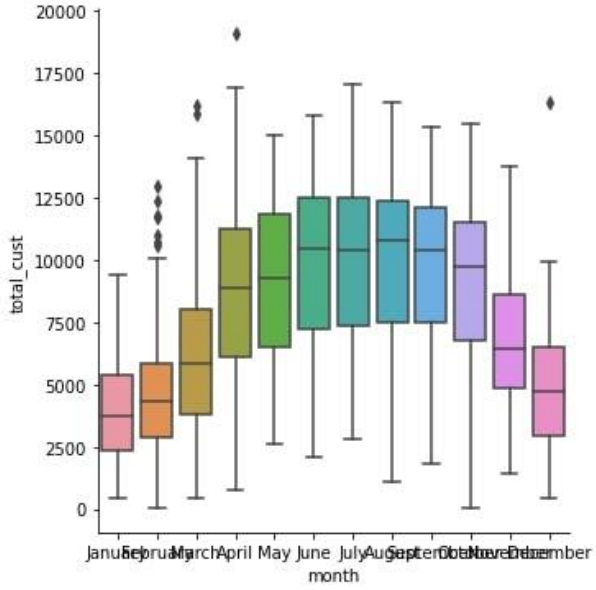
3.2.2 Create New Features:

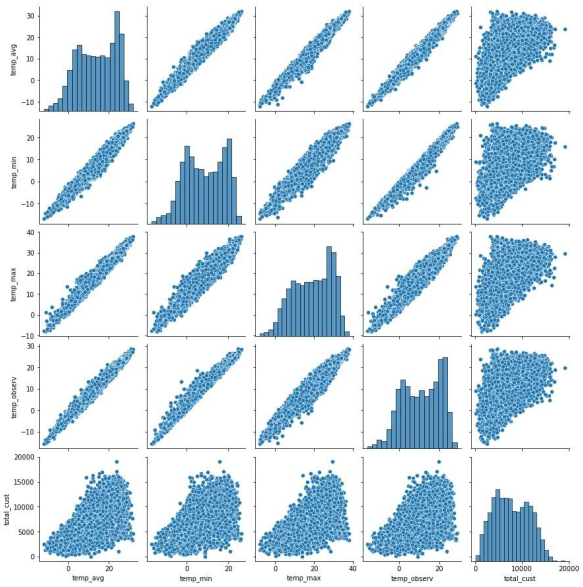
In the feature creation stage, new variables were formed to enhance the data analysis and modeling process. Weather attributes were consolidated into three categories: ice, rain, and fog, to simplify analysis. Additionally, from the datetime variable, features like year, month, and day of the week were extracted. This was done to better understand customer behavior patterns relative to these time-related factors.

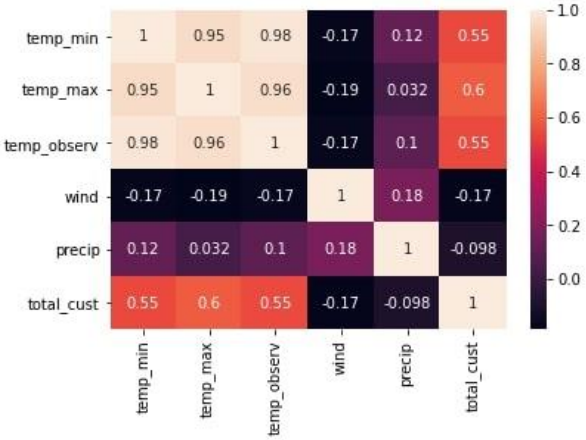
3.2.3 Data Analysis and Visualization:

In the data analysis and visualization phase, several plots were created to derive insights:

1. **Monthly Customer Trends:** A plot of total customers by month showed higher numbers from April to October, indicating a seasonal preference during warmer months. An outlier in February suggested a special event leading to increased registrations.
2. **Temperature vs. Customers:** Scatter plots revealed a linear relationship between temperature and customer numbers, indicating temperature as a significant factor affecting bike rentals.
3. **Heatmap for Correlation:** This visual showed the relationship between variables. While wind and precipitation showed no significant correlation with customer numbers, temperature variables statistically correlated with rental demand.





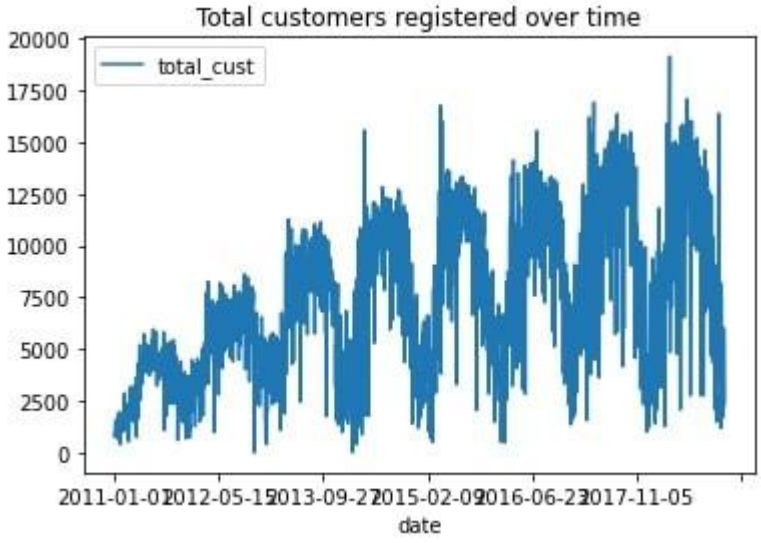


3.2.4 Time Series Specific Analysis:

We collectively focused on identifying whether our data is stationary or non-stationary, a crucial step in time series forecasting. Stationarity means that statistical properties like mean, variance, and covariance remain constant over time. Our first approach involves creating line plots to assess the data's stationarity visually. Our time series data is non-stationary, indicating changing statistical properties over time.

To further confirm the stationarity of our series, we employ two key statistical tests: the Augmented Dickey-Fuller (ADF) test, which checks for unit roots to ascertain difference stationarity, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, used to evaluate trend stationarity.

Upon determining the non-stationary nature of our series, we collectively decided to transform it into a stationary series before proceeding with modeling. We plan to use methods such as differencing to remove time dependencies, detrending to eliminate trend effects, and transformations like the log transfer method to stabilize variance and achieve stationarity. This preparatory work is essential for the effectiveness of our subsequent time series forecasting model.



4. MODELS

Ensemble is joining various models for predictions, not just using one model. This is a unique technique to combine the prediction power of multiple learning models, which can be termed as Base learners. In Ensemble, you have two methods:

Bagging
This method develops training subsets by replacements of data. The output is finalized by the number of votes attained, like in a Random Forest model.

Boosting
This method includes joining of weaker models with stronger learning models in a sequence which leads to better accuracy. Models like XG BOOST, ADA BOOST, and Gradient BOOST. Building trees one by one. Each new tree tries to fix the mistakes of the one before it.

4.1 Bagging meta-estimator

In Bagging met- estimator we split our data into random groups. Each group has all the features we need. Then, you stick a base estimator on each group. Finally, you take the guesses from each group. We add all those and we get the result. There are some special words in this process.

Base_estimator: It is the base estimator we apply to a random group from your data.

N_estimators: number of base estimators we need.

N_jobs: Number of many jobs we need to run at the same time.

Random_state: Process to split things randomly used to compare two models.

4.2 Understanding AdaBoost

AdaBoost is a simple decision tree. It just has one level and one split. AdaBoost's process starts by treating all data points equally. But those data points that make more mistakes get more attention. They get heavier weights. These 'heavy' data points are then prioritized in the next cycle of model-building. This keeps going until the model makes fewer errors for the regression issue. If our data set has some curvy patterns, AdaBoost comes in best. It grasps these curves, and that ends up improving the accuracy of our regression problem.

4.3 Gradient Boosting

This method centers on building models one after the other. It's all about correcting the errors from preceding models. The errors are slashed down by forming a fresh model from the flaws of the prior model. It cuts down the loss function by incorporating weak learners with the aid of gradient descent.

The hyperparameters we've selected include n_estimators, Learning rate, and max_depth. N_estimators refer to the number of trees (or weak learners) required in the model. As for the learning rate, a lesser value is usually more effective, given that enough trees are in place. Max_depth pertains to the decision's height. There exist multiple extensions of Gradient boosting.

4.4 XGBoost

It stands for eXtreme Gradient Boosting, an enhancement method. It has several unique features. XGBoost actively penalizes models using both L2 and L1 regularization, useful measures to prevent overfitting. In our data, there are one-hot encoded values, highlighting the data's sparsity. This reveals a need for a sparsity-sensitive split-finding algorithm, helping manage the range of sparsity patterns in the data. But XGBoost excels in more than just handling data - it also shines in computation. It utilizes multiple CPU cores, XGBoost performs highly. Its best usage revolves around large training sample sets, especially when a mix of numerical and categorical data comes into play.

4.5 CatBoost

Working with lots of categorical variables is difficult for more variables. If these variables have many labels, using one-hot-encoding bumps up the size of your model a lot. This can make preparing and training your dataset hard work, not to mention sifting out overfitting issues.

CatBoost can handle categorical variables on its own without the data prep we feed it into our machine learning model.

5. Experiments and Results

The most important step is checking the accuracy of the model. We rely on things like Mean Absolute Error, Mean Squared Error, R-Squared (also known as the Coefficient of Determination), and Root Mean Squared Error. These are our yardsticks in gauging the accuracy or performance of the model, especially in problems involving regression analysis.

MSE: This is the Mean Squared Error. It is a sort of average, but of the square of the difference. It helps us see how off the mark the errors are.

RMSE: The RMSE is just the square root of the Mean Squared Error.

R-squared, or Coefficient of Determination: It is change in one variable we can attribute to the change in another variable. Irrespective of how small or large these values are, this score is always less than one(<1).

Mean Squared Error(MSE) vs Root Mean Square Error: These methods are important for pinpointing big prediction errors compared to the Mean Absolute Error (MAE). Generally, RMSE is chosen over MSE when assessing the effectiveness of regression issues against other models.

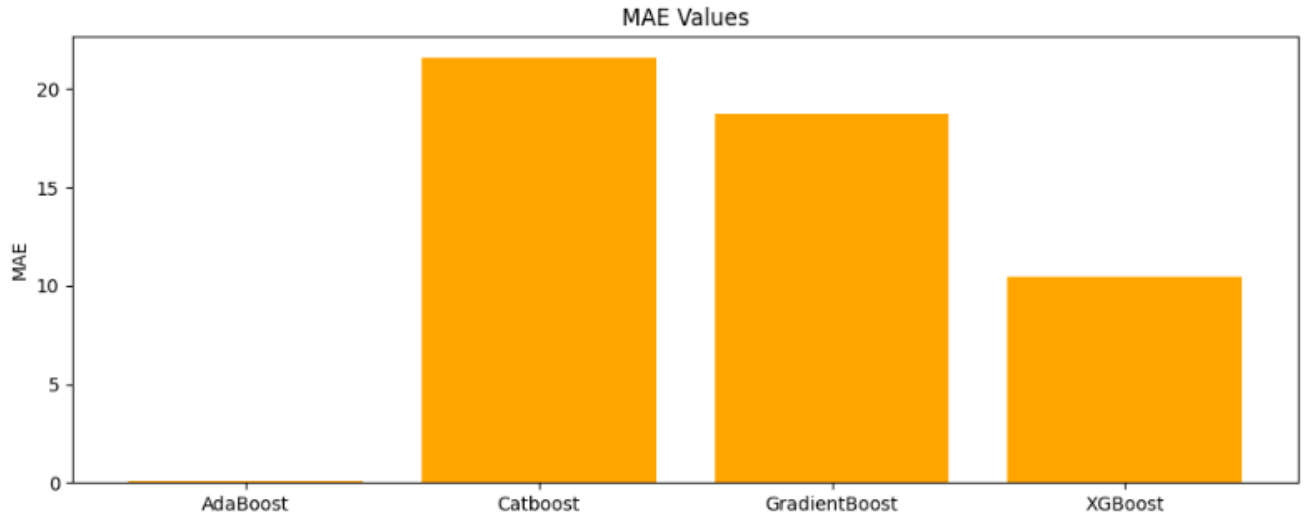
A smaller MAE, MSE, and RMSE value point to greater accuracy in a regression model. Also, calculating MSE won't take much time compared to MAE. This is due to its attribute of being differentiable, unlike MAE which is non-differentiable.

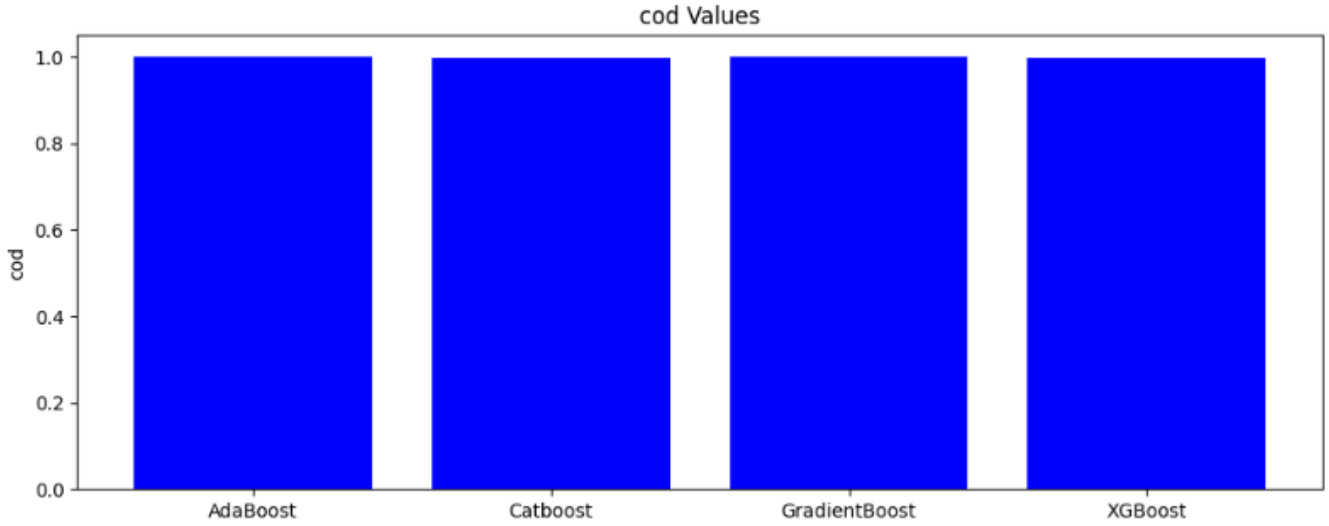
This is the reason why RMSE is frequently used to measure accuracy concerning a certain loss function. When wanting to compare different models' accuracies, most lean in favor of RMSE versus R-Squared.

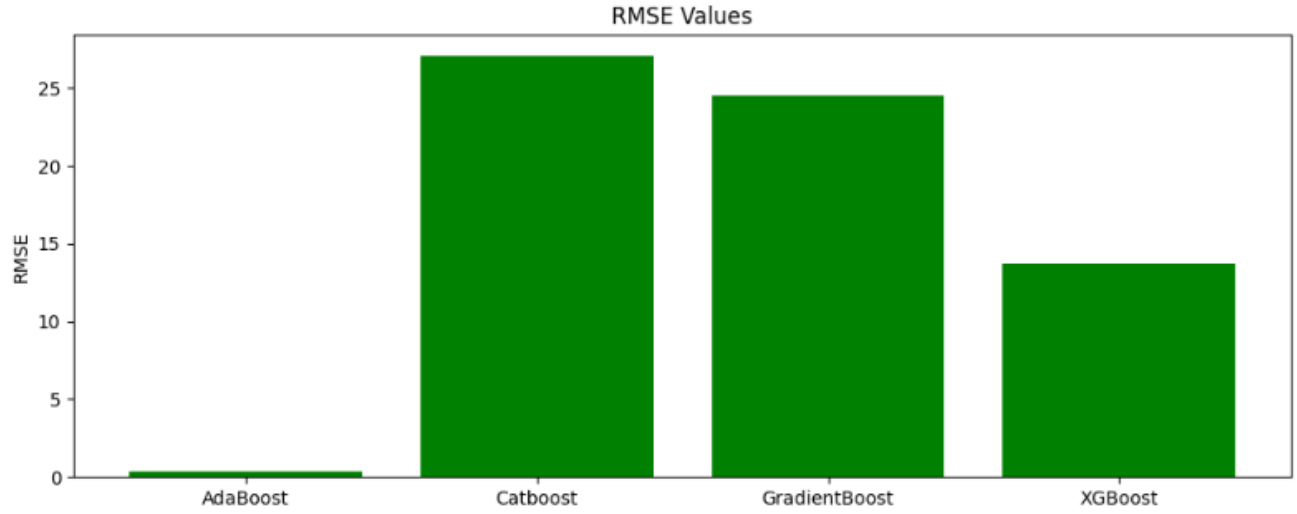
ML Models	RMSE	COD	MAE
Ada Boost	0.3791639239	0.9999999898	0.06063569682
Gradient Boost	24.53625508	0.999957226	18.7207903
XG Boost	11.89318835	0.9999899501	9.006094416
Bagging Estimator	3400.463494	0.178456845	2881.978636
Cat Boost	27.36333906	0.9999468013	21.77552653

As per the results, XG boost is the better model compared to Adaboost model although Adaboost gives the lowest error rate as XG boost tackles the overfitting issue of the model.

Visualization techniques of the model metrics:







6. CONCLUSION

Our goal of the project was to guess how many bikes would be rented from a bike-sharing program. Figuring out how many bikes will be needed at different places. However, we could help operators work out the best way to allocate bikes and run their service more efficiently.

During our study, we used machine learning models like Random Forest, AdaBoost and XGBoost. This allowed us to analyze a big dataset from OpenML. This dataset had daily records from 2011 through 2018. These records showed things like temperature, the weather, wind speed and number of people rented bikes. Using this, we wanted to figure out the total number of people who would rent bikes on any given day.

In short, in this work, we inspected various ways to look at a bike share dataset. We checked how every method is performing using several ways to calculate accuracy. We have taken GradientBoost, AdaBoost, CatBoost, XGBoost, Bagging Estimator among them the **AdaBoost** method outperformed other models according to our accuracy measures. But, we will use **XGBoost** to keep our model too fine-tuned.

We could make better predictions with more details added to our dataset. Making the model stronger. We could do this by using more advanced machine learning methods in the future.

7. FUTURE WORK

1. **Improving with User Feedback**: Users input is mandatory for improvement of model. By including their opinions, we can improve model quality and resolve any issues faced.
2. **Expand Features**: Include extras in our dataset to boost prediction power. Incorporating data such as city events or festivals could shed light on bike hire behaviors and predict their availability.
3. **Location Analysis**: Map the location of bikes across the city. Identify the patterns and locate hot spots in town so that bikes can be placed only on the high density areas.
4. **How Weather Affects Riding**: Weather influence on renting a bike . This includes sudden changes and extreme events. We can tweak the model based on the weather by taking the weather vs bike usage as two axes so that it gives an idea about the influence.
5. **Long-Term Prediction**: Extend the forecasting horizon for long-term predictions. Explore how the model performs when predicting bike rentals over extended periods, considering factors like seasonal changes and evolving user behaviors over the years.

8.REFERENCES

1. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squaredwhich-metric-is-better-cd0326a5697e
2. https://scikit-learn.org/stable/modules/ensemble.html
3. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/