Name: Krishnachander Govindarajan
Email ID: gkrishnachander1@gmail.com
Date: 03 November 2023
Phone: +49 175 6853865

Introduction

The dataset contains hourly rental data from a bike rental service (2011 & 2012), detailing counts of rentals of casual renters, registered renters and total renters, alongside seasonal attributes such as weather, humidity & season. Insights from this dataset can be used to optimize the inventory of the rental store and the workforce required to service the customers. Using these insights, we can balance supply and demand, ensure low customer waiting times and maximize profitability of the store.

Data Insights

1. By observing the monthly trends in Figure 1, a significant insight emerges regarding the seasonality of bike rentals. The average counts of bike rentals for both casual and registered users seem to peak around the middle of the year, particularly from May to September, which corresponds with warmer seasons.
2. This is further corroborated when we analyse the average count of bike rentals against temperature and seasons in Figure 2 & Figure 3. As temperatures rise, the average bike rentals increase as well, highlighting a positive correlation between warm weather and bike rental popularity.
3. Interestingly, an exception is observed, a drop in rentals at high temperatures (around 40 Degrees C) suggests extreme heat discourages biking.
4. Figure 4 illustrates distinct hourly bicycle rental patterns: a bimodal distribution on working days with peaks during typical commute times, and a more uniform midday peak on non-working days, suggesting leisure use. Late-night and early-morning hours show the lowest rental activity regardless of the day type.
5. Figure 5 shows box plots for total, casual, and registered bike rentals. There are many points far from the rest, which we call outliers. These unusual points could be mistakes or days with special deals. Removing these unusual points might help make our future predictions of bike rentals better.

Predictive Model & Explainability

1. **Outlier Removal:** A two-step outlier elimination was applied using the IQR method, first for 'casual' counts and then for 'registered' counts, to address discrepancies in hourly trendlines. This ensured a cleaner and more consistent dataset for modelling.
2. **Feature Selection and Transformation:** The feature 'yr' was excluded to avoid overfitting. ‘atemp’ was also excluded as it was highly correlated with ‘temp’. Log transformation was applied to the 'cnt' feature for normalization, with inverse applied post-prediction. Categorical features were directly used without one-hot encoding due to the capabilities of the XGBoost algorithm.
3. **Data Splitting:** The dataset was partitioned into an 80-20 split for training and testing, respectively.
4. **Model Selection:** The choice of XGBoost Regressor was driven by its superior performance compared to a spectrum of models. After establishing a baseline with Linear Regression, both Random Forest and a Feedforward Neural Network were evaluated with tuned hyperparameters. XGB Regressor outperformed these alternatives.
5. **XG Boost Model Evaluation:** The model underwent comprehensive training with hyperparameter optimization, achieving a **Mean Absolute Error (MAE) of 28.54** and an R-squared value of 87.20%, indicating high predictive accuracy and variance explanation.
6. **Feature Importance Plot (Figure 6)**: The 'hr' (hour of the day) feature has the highest F score, indicating its significant role in predicting bike rentals, followed by 'hum' (humidity) and 'temp' (temperature). Features like 'holiday' have the least importance, suggesting a smaller impact on the rental counts.
7. **SHAP Summary Plot (Figure 7):** High 'hr' and 'temp' influence predictions positively, reflecting the preference for comfortable weather conditions. 'Workingday' shows dual impact.

Name: Krishnachander Govindarajan
Date: 03 November 2023

Email ID: gkrishnachander1@gmail.com
Phone: +49 175 6853865

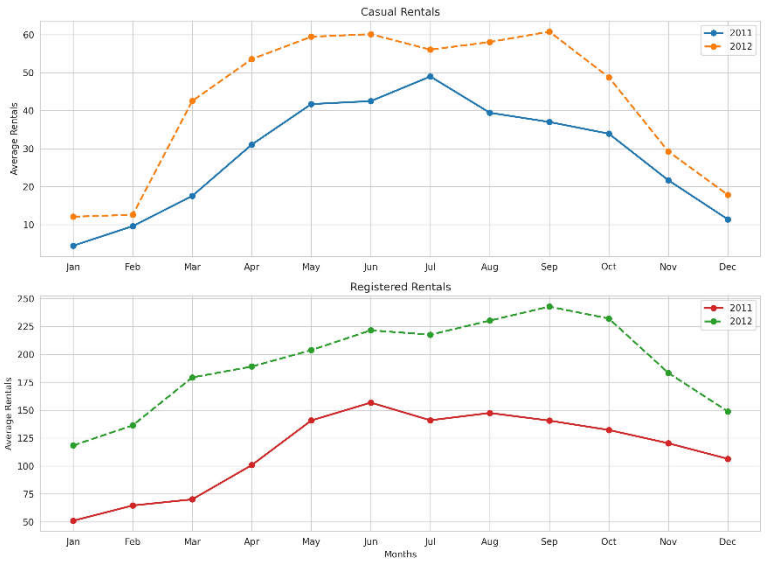

Figure 1 Year wise Average Monthly Bike Rentals for Casual & Registered Users

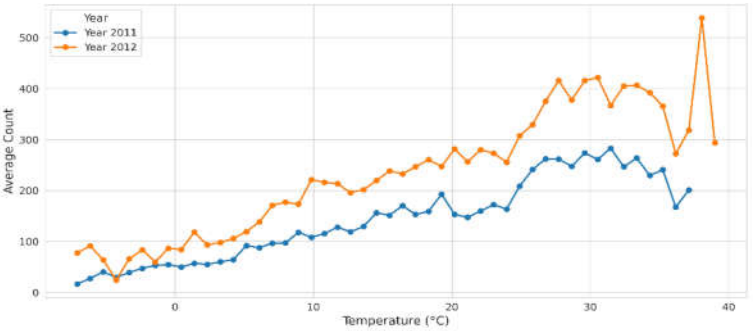

Figure 2 Average Bike Rentals vs Temperature & Year

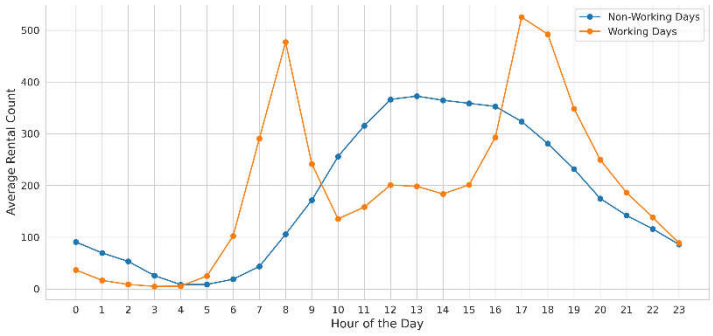

Figure 4 Average Bike Rental Counts by Hour for Working & Non-Working Days

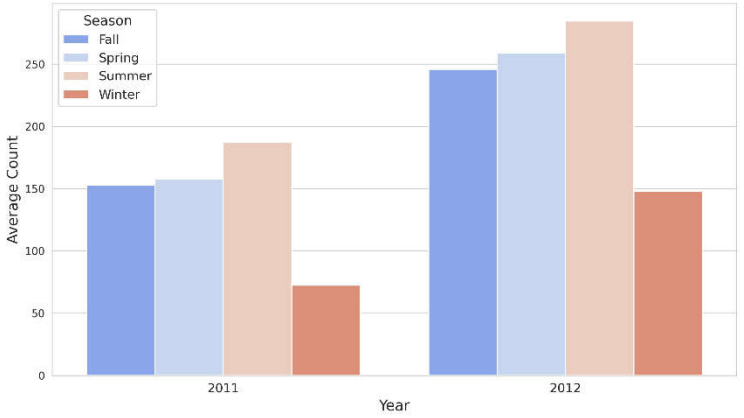

Figure 3 Average Bike Rentals per Season & Year

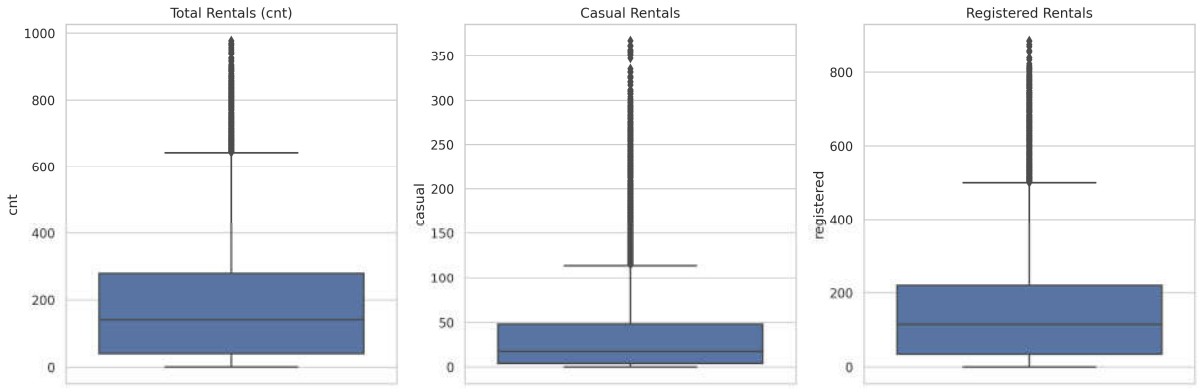

Figure 5 Box-Plots for 'cnt, 'casual' and 'registered'

Name: Krishnachander Govindarajan
Date: 03 November 2023
Email ID: gkrishnachander1@gmail.com
Phone: +49 175 6853865

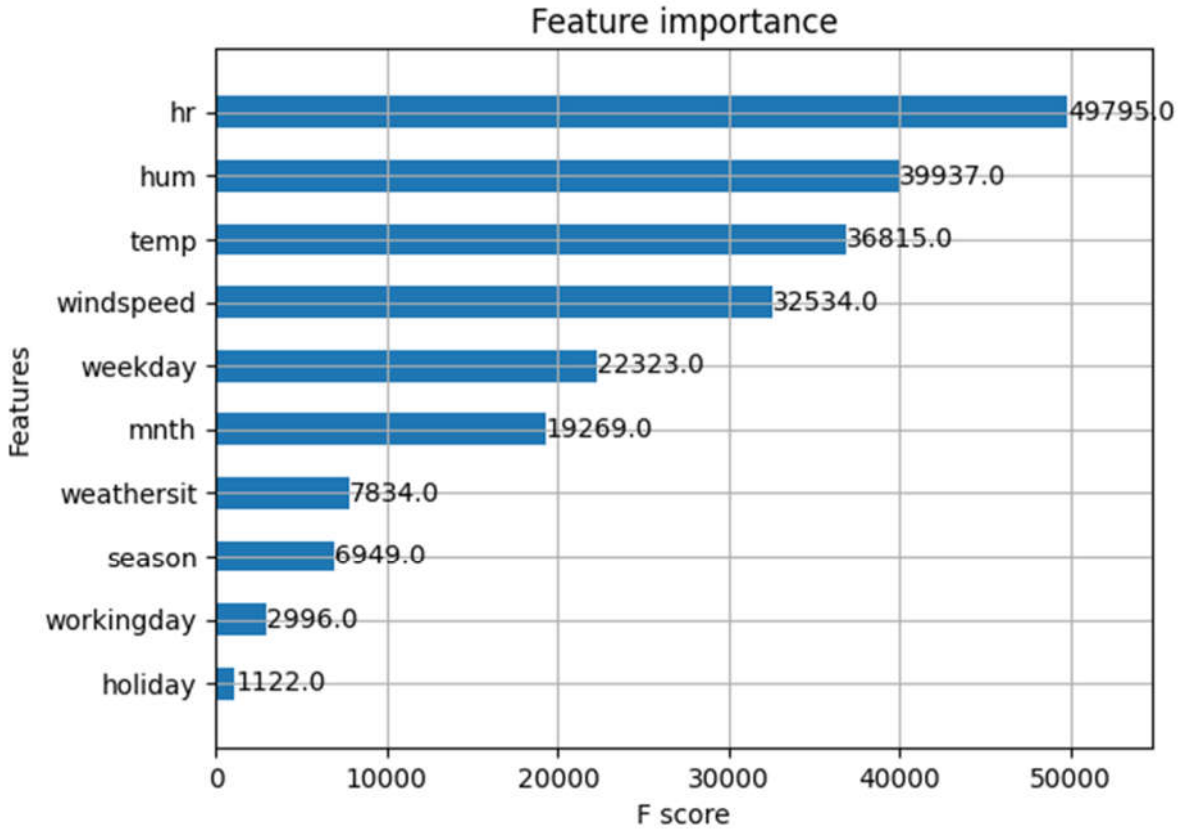

Figure 6 XGB Model - Feature Importance

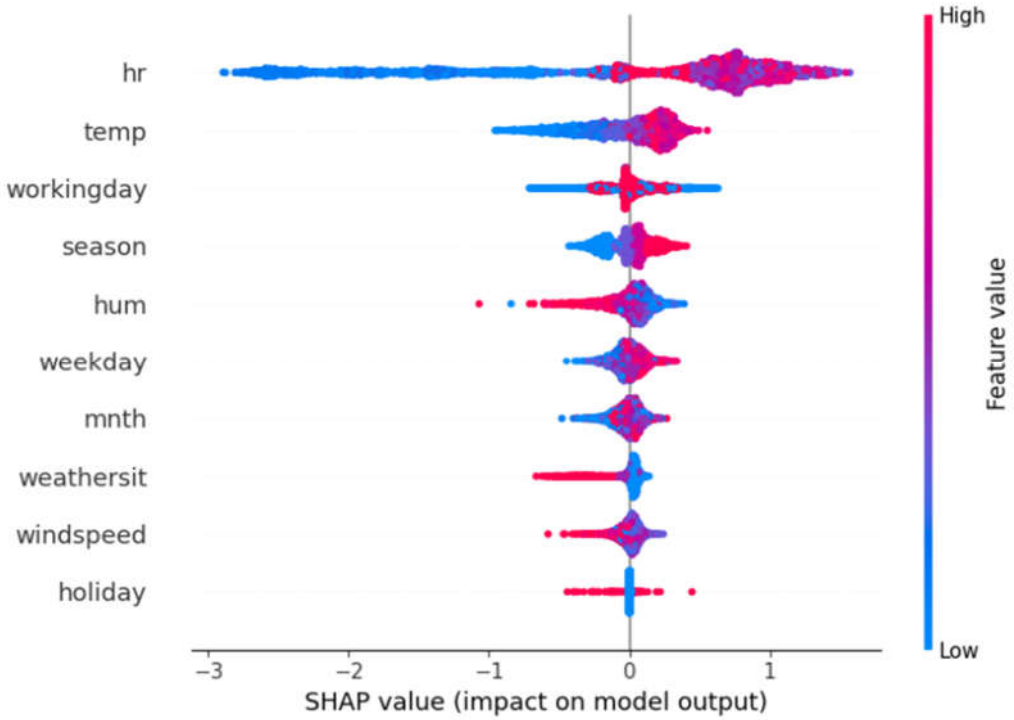

Figure 7 SHAP Values from Test Dataset

Name: Krishnachander Govindarajan
Email ID: gkrishnachander1@gmail.com
Date: 03 November 2023
Phone: +49 175 6853865

Part 2: Deploying models on a scaled dataset

Considerations for Deploying Model on a scaled dataset

1. **Data storage**: Traditional data processing libraries like Pandas are not equipped to handle datasets in the terabyte range due to memory limitations.
2. **Data Preprocessing Pipeline**: Executing a preprocessing pipeline for large datasets with Python can be complex and might require substantial computational resources.
3. **CPU Resource**: Although XGBoost can run in parallel across multiple CPU cores, providing scalability, datasets of terabyte magnitude need advanced, high-performance CPUs with an extensive multi-core architecture.
4. **Training and Hyperparameter Tuning**: The feasibility of executing comprehensive cross-validation on large datasets is not possible, given the considerable time and resource commitments required.
5. **Updating Model**: Once deployed, incorporating new data for model retraining can be a challenge since XGBoost does not natively support incremental learning.

Strategies for Addressing Problems

1. **Big Data Platforms**: Utilizing platforms such as Apache Spark and Hadoop can help manage and process large volumes of data. From my experience at Sapiens, we've successfully employed Spark for streamlining data processing tasks and Hadoop for efficient storage management.
2. **ETL tools & Workflow Automation**: Tools like Airflow and Talend can be used to create preprocessing pipelines and machine learning workflows for continuous and large datasets. My background includes developing preprocessing pipelines using Talend and Airflow for various data warehousing and analytical projects.
3. **Distributed Computing with XGBoost:** Distributed versions of XGBoost facilitate running the algorithm on Spark clusters, enhancing its ability to handle sizable datasets.
4. **Hyperparameter Optimization**: Optimal hyperparameters can be determined on a smaller sample of the dataset and then extrapolated to the larger dataset to streamline the process.
5. **Deployment and Scaling with Containers**: The combination of Docker and Kubernetes offers a scalable solution for deploying machine learning models, abstracting away much of the infrastructure management, thus allowing a more focused approach on model development. I have previously worked with Docker for containerization and Kubernetes for orchestration.
6. **Real-Time Data Streaming**: Apache Kafka can be implemented for real-time data streaming, enabling the model to ingest and process data instantaneously. Currently, at Sapiens, we are leveraging Kafka to funnel data in real-time from operational systems into data warehouses, ensuring immediate availability for analytical processing.

Limitations of Proposed Approach

1. **Training Duration**: Training this model on large datasets can be extremely time-consuming, which can impact the iterative process of model development and deployment.
2. **Cost & Complexity**: The financial and logistical complexities of setting up and maintaining a large-scale machine learning infrastructure can be very high. Performing a cost-benefit analysis of different available solutions, such as linear regression, compared with their performance metrics, is crucial for the bike rental company to make an informed decision that aligns with its budgetary constraints and performance thresholds.