Bike Sharing Prediction

Vishnu Chityala & Sarthak Mittal
e23cseu0049@bennett.edu.in & e23cseu0059@bennett.edu.in

Abstract

This study investigates how two predictive models Poisson regression and Random Forest regression–can be applied to a dataset that contains count data. The Poisson regression model is typically well–suited for count data predictions. Additionally, a Random Forest regressor was used due to the dataset's complexity and possible nonlinear relationships. The performance, interpretability, and predictive accuracy of both models were assessed.

Phases of Project

Model Selection and Designing

We selected Poisson Regression and Random Forest Regression as suitable models for bike-sharing demand prediction. Key evaluation metrics were defined to assess model performance.

Data Set Preparation

The dataset was cleaned, and relevant features were engineered to improve model performance. The dataset was then split into training and testing sets.

Dataset

Testing and Validation

Both models were trained and tuned. The model with the best performance, considering accuracy, interpretability, and computational cost, was selected.

Model Selection and Designing

Poisson Regression

Poisson Regression is a statistical method used to model count data. It's particularly suitable for predicting the number of occurrences of an event within a specific time interval. In the context of bike-sharing, it can be used to predict the number of bike rentals at a particular time and location.

Random Forest Regression

Random Forest Regression is a powerful ensemble learning technique that combines multiple decision trees to make accurate predictions. It's robust to overfitting and can handle complex relationships within the data. This makes it a strong candidate for predicting bike-sharing demand, as it can account for various factors like weather conditions, time of day, and day of the week.

Dataset Preparation

We began by cleaning the dataset to ensure data quality and relevance. We removed irrelevant columns like **instant**, **dteday**, **yr**, and **mth** that did not contribute to the predictive goal. To prevent data leakage, we excluded the casual and registered columns, as they are subcomponents of the target variable.

To capture seasonal patterns, we transformed the categorical season variable into numerical representations using one-hot encoding. This allowed the models to interpret seasonality without assuming an ordinal relationship.

Finally, we split the dataset into training and testing sets. The training set was used to train the models, while the testing set was used to evaluate their performance on unseen data.

Testing and Validation

We began by cleaning the dataset to ensure data quality and relevance. We removed irrelevant columns like **instant**, **dteday**, **yr**, and **mth** that did not contribute to the predictive goal. To prevent data leakage, we excluded the casual and registered columns, as they are subcomponents of the target variable.

Metric	Poisson Regression	Random Forest
Mean Squared Error (MSE)	1,852,055	1,775,758
Root Mean Squared Error (RMSE)	1,360.90	1,332.58
Mean Absolute Error (MAE)	1,360.90	1,091.74
R-squared (R^2)	0.506	0.557
Explained Variance	0.506	0.562

Results

The results indicate that **Random Forest Regression** outperformed **Poisson Regression** in predicting bike-sharing demand. Random Forest exhibited lower error metrics and higher R-squared and Explained Variance scores, suggesting better accuracy and ability to capture complex patterns in the data.

Thank You !

Project Website