Bike Sharing Prediction using Poisson and Random Forest Regression

Vishnu Chityala, Sarthak Mittal
[1]Bennett University, Greater Noida, India
e23cseu0049@bennett.edu.in[1]
[2]Bennett University, Greater Noida, India
e23cseu0059@bennett.edu.in[2]

ABSTRACT

This study investigates how two predictive models—Poisson regression and Random Forest regression—can be applied to a dataset that contains count data. To capture the underlying trends and distributions related to count-based target variables, Poisson regression—which is typically well-suited for count data predictions—was employed. However, we also used a Random Forest regressor because we were aware of the dataset's complexity and possible nonlinear relationships. A comparative benchmark was the Random Forest model, which is renowned for managing high-dimensional data and complex feature interactions. The performance, interpretability, and predictive accuracy of both models were assessed.

Keywords: Poisson Regression, Random Forest Regression, Prediction Models

I. INTRODUCTION

Better comprehension and predictions of actual events have been made possible by predictive modelling, particularly in fields where goal variables stand in for counts. Because it assumes a Poisson distribution and works best when the variance equals the mean, the classic method known as Poisson regression is frequently the first option for count data.

Basic models find it difficult to capture the complicated, nonlinear relationships found in real-world datasets. In these situations, Random Forest regression is a versatile substitute that combines several decision trees to handle high-dimensional data and intricate feature interactions.

In this study, the prediction of count data using Random Forest and Poisson regressions is compared. While Poisson regression will work well for simpler count data, we anticipate Random Forest to perform better in datasets with intricate feature dependencies. The goal of the analysis is to give data scientists guidance on choosing models for count-based forecasts in challenging situations.

II. LITERATURE REVIEW

The application and comparison of Poisson and Random Forest regressions for count data prediction in this study are guided by a number of important sources.

Anderson, C. (n.d.) [1] provided essential insights into Poisson regression, particularly its suitability for count data where the mean and variance are related, informing its application in this study.

Chhablani, K. (2021) [2] offered valuable understanding of Random Forest regression, especially in handling complex, nonlinear relationships within data, which was crucial for implementing this model.

Click Reader. (2021) [3] explained the implementation of Random Forest regression in Python, serving as a practical guide for setting up the Random Forest model in our analysis.

Neptune.ai. (2021) [5] discussed the strengths and limitations of Random Forest, aiding in the evaluation of its performance compared to Poisson regression.

III. DATASET

The dataset includes several features relevant for predicting bike rental counts:

1. **Holiday**: A binary variable (0 or 1) indicating whether the day is a public holiday.
2. **Workingday**: A binary variable (0 or 1) representing weekdays (1) versus weekends/holidays (0).
3. **Weathersit**: A categorical variable (1-4) denoting weather conditions, where higher values indicate worse weather.
4. **Temp**: The normalized temperature in Celsius (scaled between 0 and 1).
5. **Atemp**: The "feels-like" temperature in Celsius, also normalized between 0 and 1.
6. **Hum**: The humidity level as a percentage (0-100%).
7. **Windspeed**: Wind speed in km/h.
8. **Cnt**: The target variable representing the number of bike rentals.
9. **Season (fall, spring, summer, winter)**: Binary variables indicating the season of the year.

These features capture key environmental and temporal factors influencing bike rentals.

IV. METHODS AND MATERIAL

The number of bike rentals was the target variable in this study, which was carried out using data from a bike-sharing service. To guarantee that the data was appropriate for efficiently training Poisson and Random Forest regression models, our data preparation and modelling procedure comprised several crucial steps. The steps taken in this study are listed below:

Data Cleaning and Preprocessing: Initial exploration of the dataset revealed columns that did not contribute directly to the predictive goal or were redundant. We dropped columns such as:

instant: An index column irrelevant to prediction.

dteday: A date column with no additional benefit for modelling in its current form.

yr and mnth: Variables representing year and month, which were deemed unnecessary given the dataset's seasonal structure.

casual and registered: Two columns representing sub-components of the target variable cnt (total count of rentals), which could introduce data leakage if included.

Feature Engineering: To make the data suitable for Poisson regression and capture potential seasonality, we performed the following transformation:

One-Hot Encoding: The season variable, which initially represented seasons as integer values (1: spring, 2: summer, 3: fall, 4: winter), was transformed into one-hot encoded columns: season_spring, season_summer, season_fall, and season_winter. This approach allowed the models to interpret seasonality without assuming an ordinal relationship.

Data Splitting: The processed dataset was split into training and testing subsets, with 80% of the data allocated for model training and 20% for model evaluation. This split allowed for a robust assessment of model performance on unseen data.

Model Training:

Poisson Regression: The Poisson regression model was trained on the processed dataset, assuming a Poisson distribution for the target variable, making it suitable for data where the mean and variance are related.

Random Forest Regression: A Random Forest regressor was trained to capture complex, nonlinear relationships using multiple decision trees, improving accuracy in datasets with interacting features.

Model Evaluation: Following training, each model's performance was evaluated on the test set using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), and Explained Variance Score. These metrics provided insights into the predictive accuracy and generalization capability of each model.

V. RESULTS AND DISCUSSION

Random Forest Regression outperformed Poisson Regression in predicting bike rentals. It showed slightly lower MSE, RMSE, and MAE, indicating better accuracy and smaller prediction errors. Additionally, Random Forest had higher R-squared and Explained Variance scores, suggesting it captured more complex patterns in the data than Poisson Regression.

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

While Poisson regression suits count data, Random Forest is better for this dataset due to its ability to handle complex, nonlinear relationships. Poisson regression may struggle with such complexities in more intricate datasets.

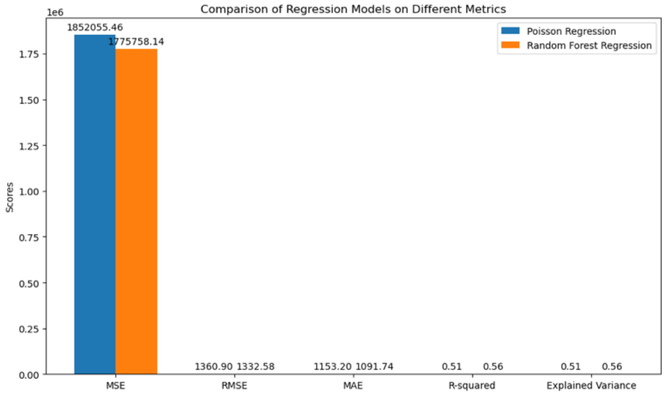

Figure 1: A sample line graph using colours which contrast well both on screen and on a black-and-white hardcopy

VI. CONCLUSION

In order to predict bike-sharing rentals, this study contrasted Poisson Regression versus Random Forest Regression. Random Forest Regression demonstrated superior accuracy and the capacity to identify intricate patterns, surpassing Poisson Regression in terms of MSE, RMSE, MAE, R-squared, and Explained Variance, even though Poisson Regression fared well for count data. Datasets having many characteristics and non-linear connections are more suited for Random Forest. For increased accuracy in further studies, the study recommends investigating additional machine learning methods and fine-tuning the model.

VII. REFERENCES

1. Anderson, C. (n.d.). *Introduction to Poisson Regression*. University of Illinois. Retrieved from https://education.illinois.edu/docs/default-source/carolyn-anderson/edpsy589/lectures/4_glm/4glm_3_beamer_post.pdf
2. Chhablani, K. (2021). *Predicting Stock Prices using Machine Learning*. Medium. Retrieved from https://medium.com/@kunalchhablani14/predicting-stock-prices-using-machine-learning-338ab2fe4e5b
3. Click Reader. (2021). *Random Forest Regression Explained with Implementation in Python*. Medium. Retrieved from https://medium.com/@theclickreader/random-forest-regression-explained-with-implementation-in-python-3dad88caf165
4. Data Overload. (2021). *Understanding Poisson Regression: A Powerful Tool for Count Data Analysis*. Medium. Retrieved from https://medium.com/@data-overload/understanding-poisson-regression-a-powerful-tool-for-count-data-analysis-b7184c61bfde
5. Neptune.ai. (2021). *Random Forest Regression: When Does It Fail and Why?*. Neptune.ai Blog. Retrieved from https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why
6. Roback, J. (n.d.). *Beyond Multiple Linear Regression: Poisson Regression*. Bookdown. Retrieved from https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html#introduction-to-poisson-regression
7. The Education Resources Information Center (ERIC). (2018). *Poisson Regression in Practice: An Applied Overview*. ERIC. Retrieved from https://files.eric.ed.gov/fulltext/EJ1173275.pdf