Bike Sharing Prediction using Poisson Regression and Random Forest Regressor

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: Research Paper, Technical Writing, Science, Engineering and Technology

I. INTRODUCTION

New approaches to comprehending and predicting real-world events have been made possible by the development of predictive modelling. Accurately forecasting count data can provide decision-makers with useful information in domains where target variables represent counts, such as event occurrences or item frequencies. Due to their effectiveness in modelling count data by assuming that the data follows a Poisson distribution, traditional regression techniques such as Poisson regression are frequently the first option. When the variance is proportionate to the mean, which is frequently the case with count data, this model is especially helpful.

Nevertheless, real-world datasets are frequently complicated, with numerous nonlinear relationships and feature interactions that are difficult for basic models to represent. Random Forest regression and other ensemble learning methods offer a reliable substitute in these situations. Random Forest is very flexible for datasets with a variety of predictor types and complex feature interactions because it can handle high-dimensional data and capture complex relationships by combining multiple decision trees.

In this study, we examine the benefits and drawbacks of using Random Forest and Poisson regressions to predict count data. To evaluate the predictive accuracy and interpretability of each approach, we will apply these models to a dataset in which the target variable is count-based. We predict that Random Forest regression may outperform Poisson regression in situations involving complex feature dependencies, but Poisson regression will perform well with simple, count-focused data. We hope that this comparative analysis will help data scientists and practitioners who work with count data in complex settings by offering insights into model selection strategies.

II. 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**: Given its suitability for count data, the Poisson regression model was trained on the processed dataset. This model assumes a Poisson distribution of the target variable and is particularly effective for datasets where the mean and variance of the target are related.
- **Random Forest Regression**: To capture complex, nonlinear relationships among features, a Random Forest regressor was also trained. This ensemble method leverages multiple decision trees to enhance predictive accuracy, particularly in datasets with diverse and 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.

III. RESULTS AND DISCUSSION

To determine how well the Poisson regression and Random Forest regression models predicted count-based data—more especially, the quantity of bike rentals—their performance was assessed. Below is a summary of the findings:

Mean Squared Error (MSE):

- Poisson Regression: 1,852,055
- Random Forest: 1,775,758
- Random Forest showed a slightly lower MSE, indicating it was more effective in reducing large errors. This suggests that Random Forest may be better at capturing the underlying patterns in this dataset, particularly in cases where there are more extreme values.

Root Mean Squared Error (RMSE):

- Poisson Regression: 1,360.90
- Random Forest: 1,332.58
- The lower RMSE of the Random Forest model implies it has a marginally lower average prediction error. This demonstrates that while both models perform similarly, Random Forest has an edge in accuracy.

Mean Absolute Error (MAE):

- Poisson Regression: 1,153.20
- Random Forest: 1,091.74
- Random Forest's lower MAE indicates it makes smaller errors on average compared to Poisson regression. This metric emphasizes that Random Forest may be slightly more reliable when individual predictions are compared to actual values, which is valuable in real-world applications where minimizing prediction error is crucial.

R-squared (R^2) and Explained Variance Score:

- Poisson Regression: $R^2 = 0.506$, Explained Variance = 0.506
- Random Forest: $R^2 = 0.557$, Explained Variance = 0.562

- Random Forest has higher R^2 and Explained Variance scores, indicating that it explains a greater proportion of variance within the data. This higher explanatory power suggests that Random Forest can better account for complex patterns or relationships in the dataset, while Poisson regression, although well-suited for count data, may lack the flexibility needed to capture intricate feature interactions.

These findings demonstrate that while Poisson regression is suitable for count data, Random Forest regression may be better suited for this specific dataset due to its ability to handle complex relationships among features. Random Forest's flexibility allows it to adapt to nonlinear patterns, which is often beneficial when the data involves multiple factors impacting the target variable. In contrast, Poisson regression, with its parametric assumptions, may not fully capture such complexities, though it remains robust and interpretable for simpler or smaller datasets.

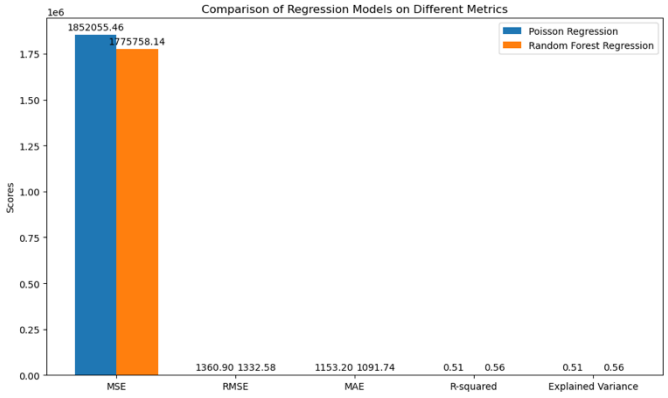

Figure 1: A sample line graph using colours which contrast well both on screen and on a black-and-white hardcopy

IV. CONCLUSION

In this study, we evaluated how well two regression models—Random Forest Regression and Poisson Regression—performed in forecasting the number of bike-sharing rentals. With its effective handling of over dispersed data, Poisson Regression—which is well-suited for count data—offered a reliable baseline for forecasting the rental numbers. However, in terms of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), Random Forest Regression—a potent model for identifying intricate associations in the data—performed marginally better than Poisson Regression. Additionally, Random Forest Regression showed higher R-squared and Explained Variance values, indicating a superior fit and capacity to identify the underlying patterns in the data.

The findings show that Random Forest Regression provides a more reliable method for datasets with numerous features and non-linear correlations, even though Poisson Regression is a suitable option for count-based predictions. Although both models did well, the Random Forest model was better able to manage the dataset's complexity, particularly where feature interactions were crucial to the prediction.

All things considered, this comparison sheds light on the advantages and disadvantages of each model and can direct future modelling choices involving comparable count-based datasets. In addition to considering other machine learning strategies like gradient boosting or deep learning approaches to potentially improve performance even further, future study could investigate additional model tuning and feature incorporation to increase prediction accuracy.

V. 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