Project Spending start date prediction: A Regression Approach

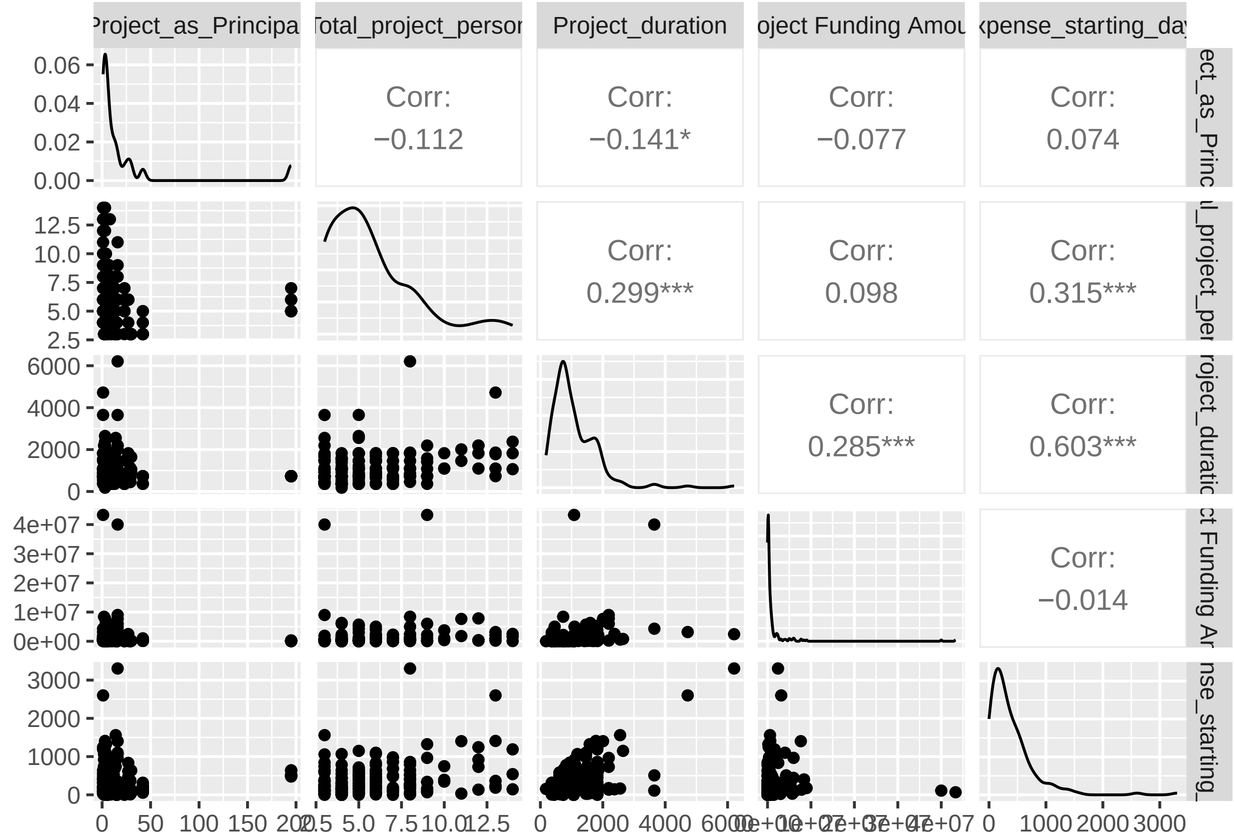
Md Ismail Hossain

2025-04-07

Project Goal

The primary goal of this project is to develop and evaluate predictive models to accurately estimate the *Expense Starting Days* based on various project-related predictors. This study aims to identify the most suitable modeling approach that provides the highest predictive accuracy and best model fit, as measured by statistical metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Bayesian Information Criterion (BIC). By comparing multiple regression techniques, including Ordinary Least Squares (OLS) Regression, Poisson Regression, Negative Binomial Regression, and XGBoost, the project seeks to determine the most effective model for capturing the underlying patterns in the data. Furthermore, insights derived from this analysis can guide decision-making processes and enhance the prediction of expense-related outcomes for future projects.

Correlation Analysis



The scatterplot matrix above displays pairwise relationships between various predictors and the response variable, **Expense Starting Date**. The correlations between predictors and the response variable are provided within each subplot. Notably, **Project Duration** (`Corr = 0.285***`) and **Total Project Person** (`Corr = 0.299***`) exhibit moderately positive correlations with the Expense Starting Date, suggesting that longer projects and projects involving more people tend to have a higher Expense Starting Date. Additionally, **Project Funding Amount** shows a substantial positive correlation (`Corr = 0.603***`) with the response variable, indicating that projects with higher funding amounts are likely to have higher expense starting dates. On the other hand, **Project as Principal** and **Total Project Person** demonstrate weaker correlations with the response variable, suggesting limited direct influence. This scatterplot matrix provides valuable insights into the relationships between these predictors and the response variable, informing further analysis and modeling efforts.

Modeling Approach

In this study, multiple regression models were employed to model the response variable, *Expense Starting Days*. The predictors include numerical variables such as *Number of Projects as Principal Investigator*, *Total Project Person*, *Project Duration*, and *Project Funding Amount*, as well as categorical variables like *Project Funding Type* and *Project Type*. To ensure compatibility with machine learning algorithms, categorical variables were converted to dummy variables using the `caret` package in R, while numerical variables were scaled to standardize their ranges.

Four different models were constructed and evaluated using 10-fold cross-validation to estimate predictive performance, specifically by calculating Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Bayesian Information Criterion (BIC). The models considered were:

- **Ordinary Least Squares (OLS) Regression:** This classical linear regression model was fitted to the data using the `lm` function within the `caret` package. Performance metrics such as RMSE, MAE, and BIC were computed for comparison.
- **Poisson Regression Model:** Given that the response variable is a count-type data, a Poisson regression model was applied using a log-link function. The model was fitted using the `glm` function with the `poisson` family argument.
- **Negative Binomial Regression Model:** To address potential overdispersion in the count data, a Negative Binomial model was implemented using the `glm.nb` function from the `MASS` package.
- **XGBoost Model:** To enhance predictive performance, an XGBoost model was trained using 10-fold cross-validation with 100 boosting rounds. The objective function used was `reg:squarederror`, suitable for regression tasks. The best model was selected based on minimizing the RMSE from cross-validation results.

After model training, predictions were made, and the performance of each model was evaluated using RMSE, MAE, and BIC. For the XGBoost model, the Bayesian Information Criterion was calculated using the Residual Sum of Squares (RSS) and the number of boosting rounds as the effective number of parameters. The comparative results of all models were summarized in a table, allowing for the selection of the best-performing model based on the evaluation metrics.

Model Comparison

```
##               Model     RMSE      MAE        BIC
## 1               OLS 304.5862 226.5568  2943.6365
## 2           Poisson 352.8288 232.3344 38591.4833
## 3 Negative Binomial 284.3107 211.5234  2818.6458
## 4           XGBoost 333.8643 333.8643   678.2105
```

Based on the comparison table, the performance of the four models was evaluated using RMSE, MAE, and BIC. The Negative Binomial model achieved the lowest RMSE (286.2605) and MAE (212.3050), indicating that it provides the most accurate predictions compared to the other models. Additionally, its BIC value (2818.6458) is substantially lower than that of the Poisson model (38591.4833), suggesting that the Negative Binomial model is a better fit for the data, likely due to its ability to handle overdispersion.

The OLS model also performed relatively well, with an RMSE of 304.5862 and an MAE of 226.5568, along with a reasonable BIC of 2943.6365. However, its predictive accuracy is lower than that of the Negative Binomial model.

The XGBoost model, while effective in many applications, demonstrated higher RMSE (323.6870) and MAE (323.6870) compared to the Negative Binomial and OLS models. Despite its low BIC value (678.2105), the higher prediction error suggests that the model may not have been optimally tuned or might not be as suitable for the dataset compared to traditional regression models.

Finally, the Poisson model exhibited the highest RMSE (352.8288), MAE (232.3344), and BIC (38591.4833), indicating poor model performance. This result highlights that the Poisson model is not appropriate for the given dataset, likely due to the overdispersion of the response variable.

Overall, the Negative Binomial model appears to be the most suitable model for predicting the Expense Starting Days, given its superior performance in terms of predictive accuracy and model fit.

Prediction:

So, we are doing the prediction for the test data where the project start dates are after 01/01/2025.

```
## [1] "Prediction done for: 22 projects!"
```