Spending start prediction for delayed project: A Machine Learning Approach

Md Ismail Hossain

2025-07-01

Project Goal

The primary objective of this project is to develop predictive models to estimate *Expense Starting Days*—the number of days between a project's official start date and its first recorded expenditure. This metric is essential for financial planning and administrative decision-making. The analysis incorporates project-level predictors such as funding amount, duration, personnel count, and funding type.

Four regression models were evaluated: Ordinary Least Squares (OLS), Decision Tree (DT), Random Forest (RF), and XGBoost. These models were compared using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to assess predictive accuracy.

Overall Summary

```
##  Project Number     Project Fund Source                Award Type
##  Length:360         Federal:127         Others              :120
##  Class :character   Others :112         Research - Sponsored:240
##  Mode  :character   State  :121
##
##
##
##  Number_of_Project_as_Principal_Investigator Total_project_person
##  Min.   :  1.000                             Min.   : 3.000
##  1st Qu.:  2.000                             1st Qu.: 4.000
##  Median :  4.000                             Median : 5.000
##  Mean   :  9.758                             Mean   : 4.825
##  3rd Qu.: 11.000                             3rd Qu.: 6.000
##  Max.   :173.000                             Max.   :12.000
##  Project_duration Academic_Semester Project Funding Amount
##  Min.   :  61.0   Fall  :115        Min.   :   1000
##  1st Qu.: 407.2   Spring:102        1st Qu.:  25450
##  Median : 729.0   Summer:143        Median :  88271
##  Mean   : 808.8                     Mean   : 207568
##  3rd Qu.:1080.2                     3rd Qu.: 227015
##  Max.   :2556.0                     Max.   :8997490
##  Expense_starting_days                           Reduced_FA
##  Min.   :  0.0         No Reduction                    :156
##  1st Qu.: 28.0         Reduced Indirect Cost Off Campus: 24
##  Median : 62.0         Reduced Indirect Cost On Campus :180
##  Mean   :109.8
##  3rd Qu.:153.2
##  Max.   :936.0
##  subawards subcontracts Approval_Required Chemical_or_Hazard
##  No :337                Unknown:280       No :350
##  Yes: 23                Yes    : 80       Yes: 10
##
##
##
##
##  Foreign_Involvement Technology_or_IP_Involved
##  No :336             No :329
##  Yes: 24             Yes: 31
##
##
##
##
```

Bi-variate Analysis

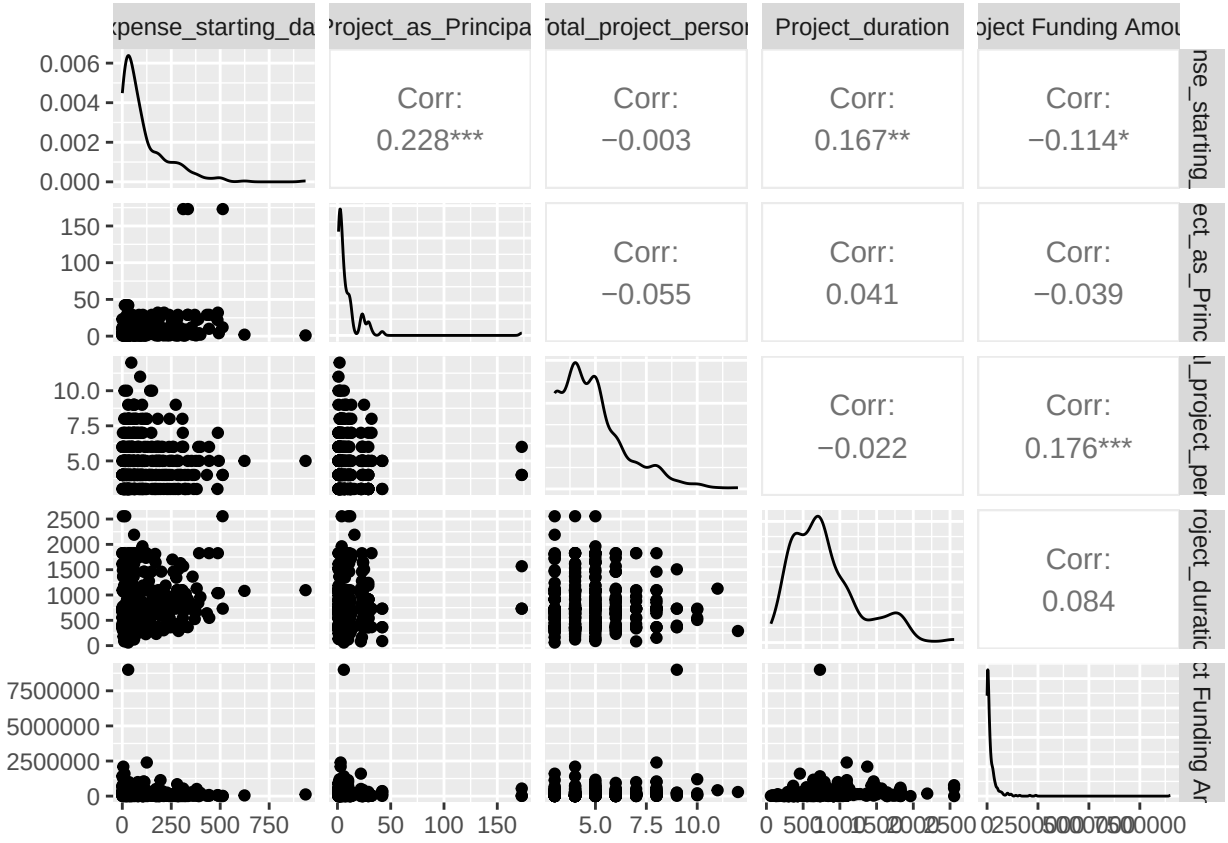
Categorical Variables

```
## # A tibble: 6 x 11
##   Variable           .y.   group1 group2    n1    n2 statistic    df     p p.adj
##   <chr>              <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <dbl>
```

```
## 1 Approval_Required  Expe~ Yes    Unkno~    80   280    -2.30  191.  0.022 0.132
## 2 Award Type         Expe~ Others Resea~   120   240    -1.30  272.  0.194 1
## 3 Chemical_or_Hazard Expe~ No     Yes      350    10     2.95   11.7 0.013 0.078
## 4 Foreign_Involveme~ Expe~ No     Yes      336    24     0.688  30.7 0.497 1
## 5 Technology_or_IP_~ Expe~ No     Yes      329    31     0.948  39.3 0.349 1
## 6 subawards subcont~ Expe~ No     Yes      337    23     3.45   34.1 0.002 0.012
## # i 1 more variable: p.adj.signif <chr>
```

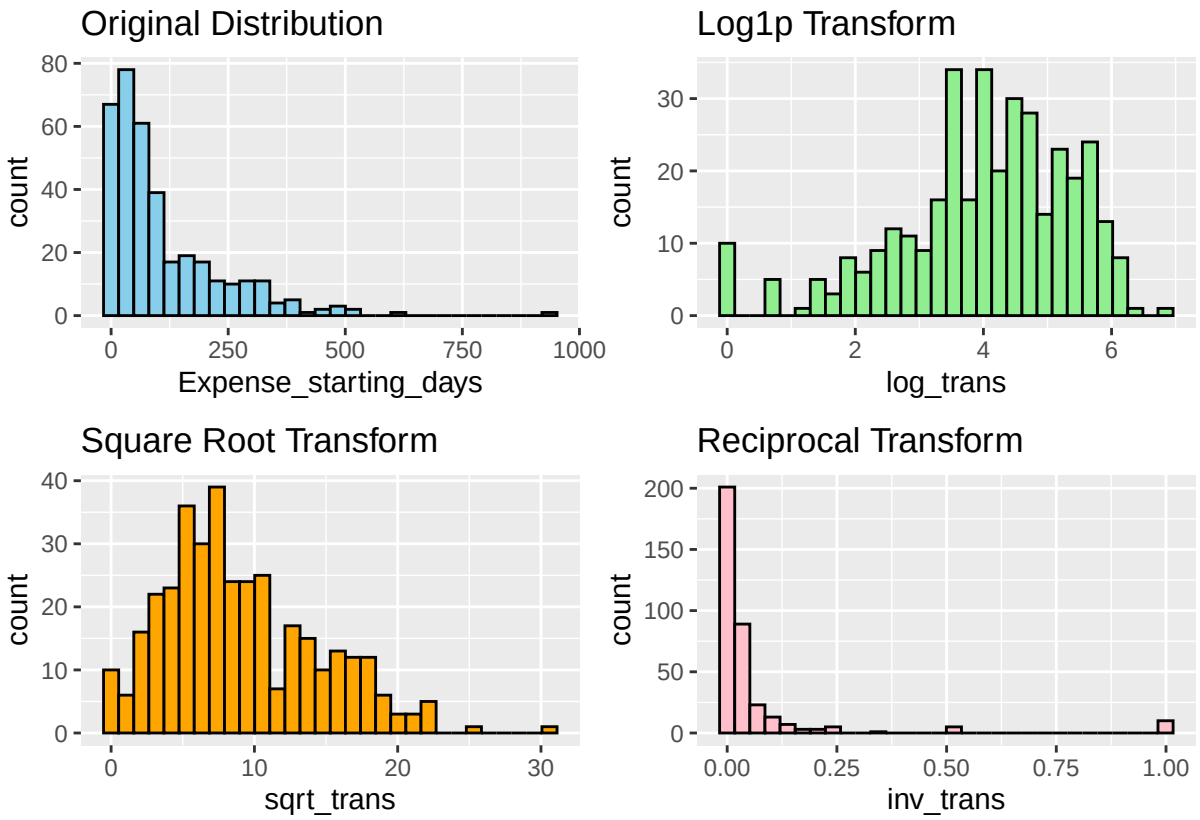
```
## # A tibble: 3 x 8
##   Variable            Effect   DFn   DFd     F     p ‘p<.05‘   ges
## * <chr>               <chr>  <dbl> <dbl> <dbl> <dbl> <chr>   <dbl>
## 1 Academic_Semester   Group      2   357 1.98  0.139 ""      0.011
## 2 Project Fund Source Group      2   357 0.612 0.543 ""      0.003
## 3 Reduced_FA          Group      2   357 2.76  0.065 ""      0.015
```

Contineous Variables



Dependent Variable heavily right-skewed:

As the dependent variable heavily right skewed, we should use transformation of the dependent variable.


Original Distribution
count
Expense_starting_days
0
250
500
750
1000
0
20
40
60
80
Log1p Transform
count
log_trans
0
2
4
6
0
10
20
30
Square Root Transform
count
sqrt_trans
0
10
20
30
0
10
20
30
40
Reciprocal Transform
count
inv_trans
0.00
0.25
0.50
0.75
1.00
0
50
100
150
200

Methodology

This analysis aimed to predict `Expense_starting_days`, defined as the number of days between a project's award date and the actual expense start date. Accurate forecasting of this interval is important for improving financial planning and administrative readiness. Multiple regression models were evaluated to identify the most accurate and interpretable approach.

Data Preparation

The response variable was log-transformed to reduce skewness, and all predictors were centered and scaled to ensure compatibility across algorithms.

Modeling Approach

We applied and compared several regression models, including **Ordinary Least Squares (OLS)**, **Decision Tree**, **Random Forest**, and **XGBoost**. Each model was trained using **10-fold cross-validation** to ensure robust and unbiased performance estimates.

Performance Metrics

Model performance was evaluated using:

- **Root Mean Squared Error (RMSE):** Measures the typical magnitude of prediction errors, penalizing larger deviations more heavily.
- **Mean Absolute Error (MAE):** Represents the average prediction error in days, offering direct interpretability.

Final Model Selection

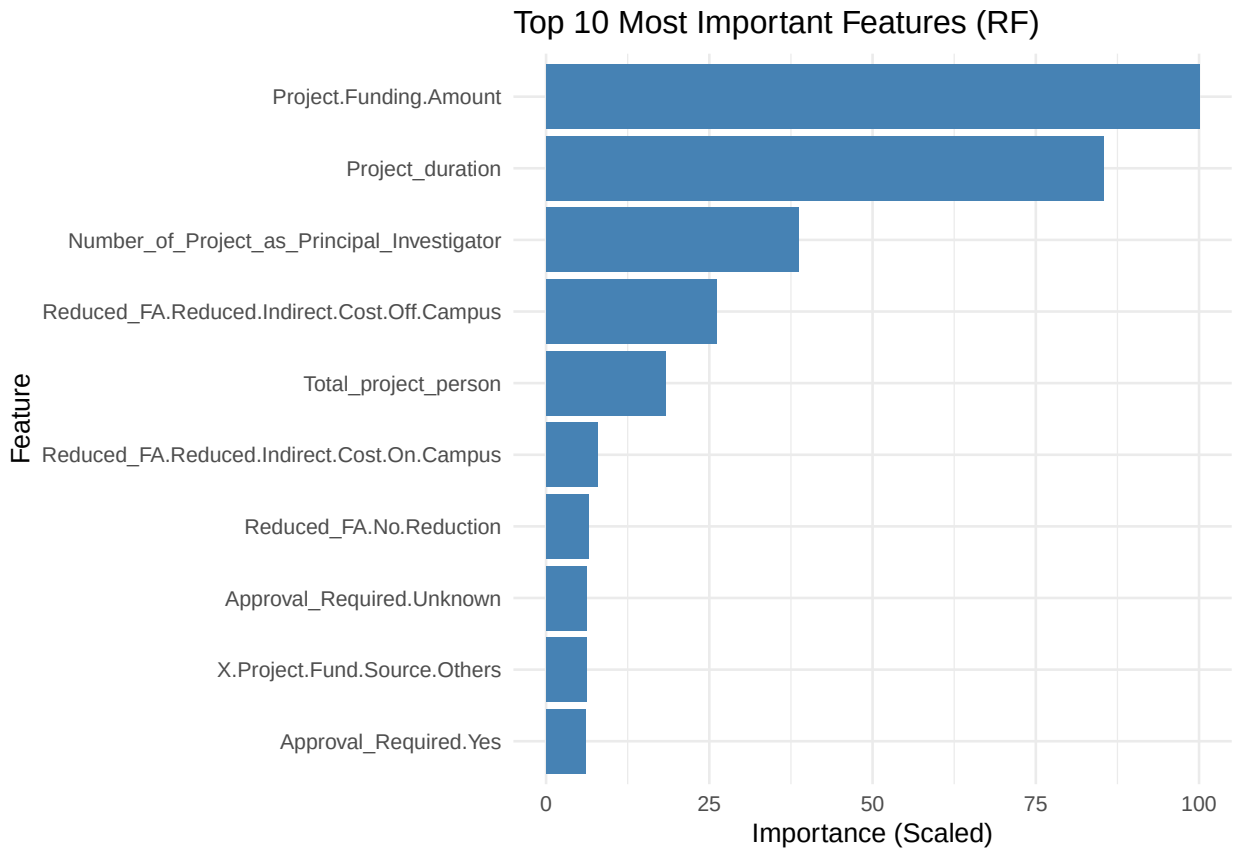
Among all models tested, **Random Forest** produced the lowest RMSE and MAE, indicating superior predictive performance in estimating expense start delays. It was therefore selected as the final model.

Model Comparison and Discussion

```
##               Model Train_minutes   RMSE   MAE
## 3     Random Forest          0.11 109.13 63.61
## 4           XGBoost         11.28 111.51 65.74
## 1 Linear Regression          0.00 124.59 77.22
## 2     Decision Tree          0.01 126.34 78.54
```

Model Comparison and Discussion

Feature Importance plot:



Conclusion:

Based on our analysis of 370 project records, predictive modeling for estimating expense start dates demonstrated limited reliability. Although the Random Forest model performed better than others, it still produced high error margins, making its predictions unsuitable for practical use. The overall lack of strong predictors, combined with a relatively small dataset and the absence of key operational variables (such as internal approvals or sponsor-related delays), limits the models' effectiveness. To improve forecasting accuracy, future efforts should focus on expanding the dataset and incorporating additional administrative and process-related features.