Spending start prediction for delayed project: A Machine Learning Approach

Md Ismail Hossain

2025-06-16

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.

Based on model performance, Random Forest was selected as the final model due to its superior accuracy in estimating expense start delays.

Overall Summary

```
##  Project Number                            Project Fund Source
##  Length:702         Federal Direct Sponsored Funds  :333
##  Class :character   Industry                        : 67
##  Mode  :character   Institutions of Higher Education: 69
##                     Non Profit/Foundation           : 53
##                     Others                          : 25
##                     State                           :132
##                     Unrestricted Operating          : 23
##   Project Funding Type         Project Designation              Project Type
##  Federal    :508       Award Project     :661      UW Grant           :661
##  Non-Federal:194       Cost Share Project: 41      UW Grant Cost Share: 41
##
##
##
##
##
##                       Award Type  Number_of_Project_as_Principal_Investigator
##  Instruction - Sponsored   :  8   Min.   : 1.00
##  Public Service - Sponsored:192   1st Qu.: 2.00
##  Research - Sponsored      :502   Median : 4.00
##                                   Mean   : 7.14
##                                   3rd Qu.: 9.00
##                                   Max.   :30.00
##
##  Total_project_person Project_duration Academic_Semester Project Funding Amount
##  Min.   : 2.000       Min.   :  61.0   Fall  :257        Min.   :       0
##  1st Qu.: 4.000       1st Qu.: 667.2   Spring:221        1st Qu.:   41887
##  Median : 5.000       Median : 994.0   Summer:224        Median :  121416
##  Mean   : 5.781       Mean   :1066.8                     Mean   :  441907
##  3rd Qu.: 7.000       3rd Qu.:1460.0                     3rd Qu.:  337740
##  Max.   :20.000       Max.   :6641.0                     Max.   :43298228
##
##  Expense_starting_days
##  Min.   :   0.0
##  1st Qu.:  30.0
##  Median :  77.0
##  Mean   : 130.4
##  3rd Qu.: 181.0
##  Max.   :1688.0
##
```

Bi-variate Analysis

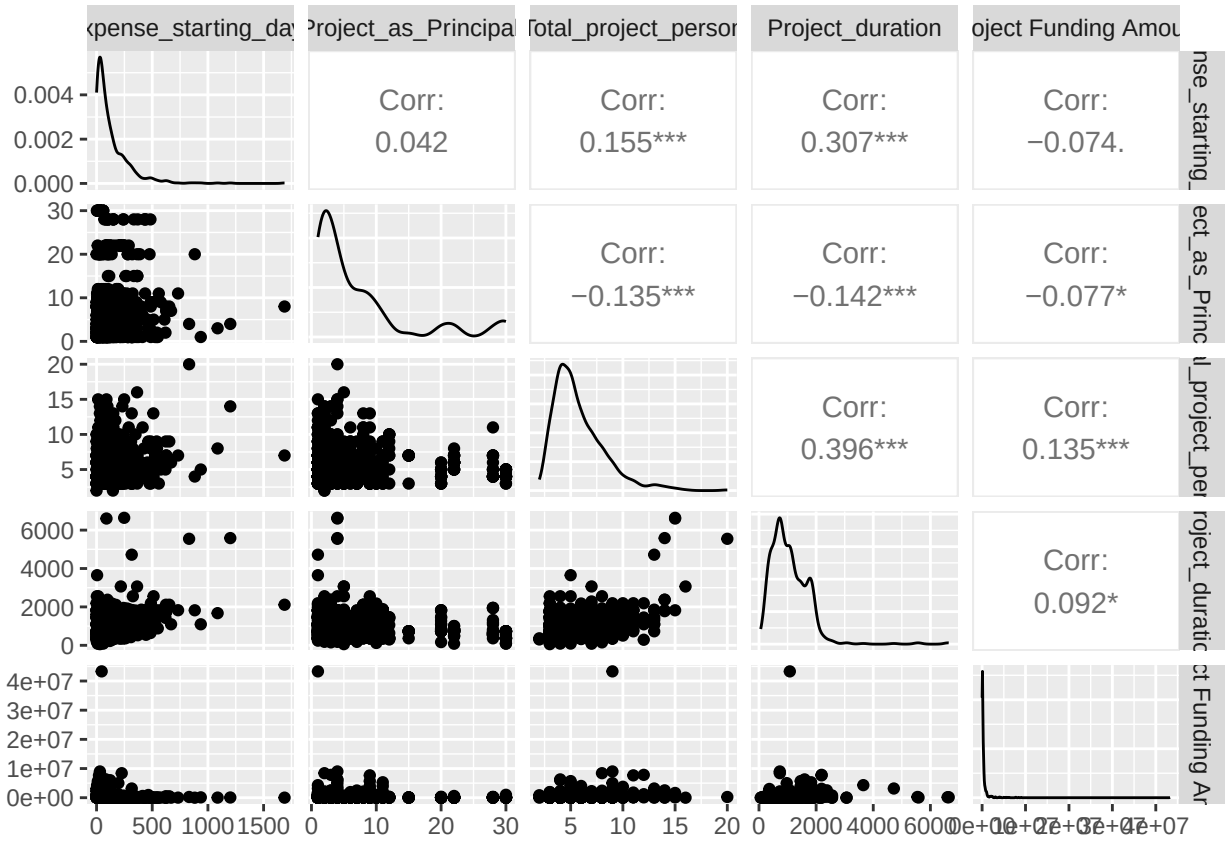
Categorical Variables

```
## # A tibble: 3 x 11
##   Variable       .y.   group1 group2    n1    n2 statistic    df        p   p.adj
##   <chr>          <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>   <dbl>   <dbl>
## 1 Project Desig~ Expe~ Award~ Cost ~   661    41     -4.12  44.7 1.64e-4 4.92e-4
## 2 Project Fundi~ Expe~ Feder~ Non-F~   508   194     -1.50 331.  1.34e-1 4.02e-1
```

```
## 3 Project Type   Expe~ UW Gr~ UW Gr~   661    41     -4.12  44.7 1.64e-4 4.92e-4
## # i 1 more variable: p.adj.signif <chr>
```

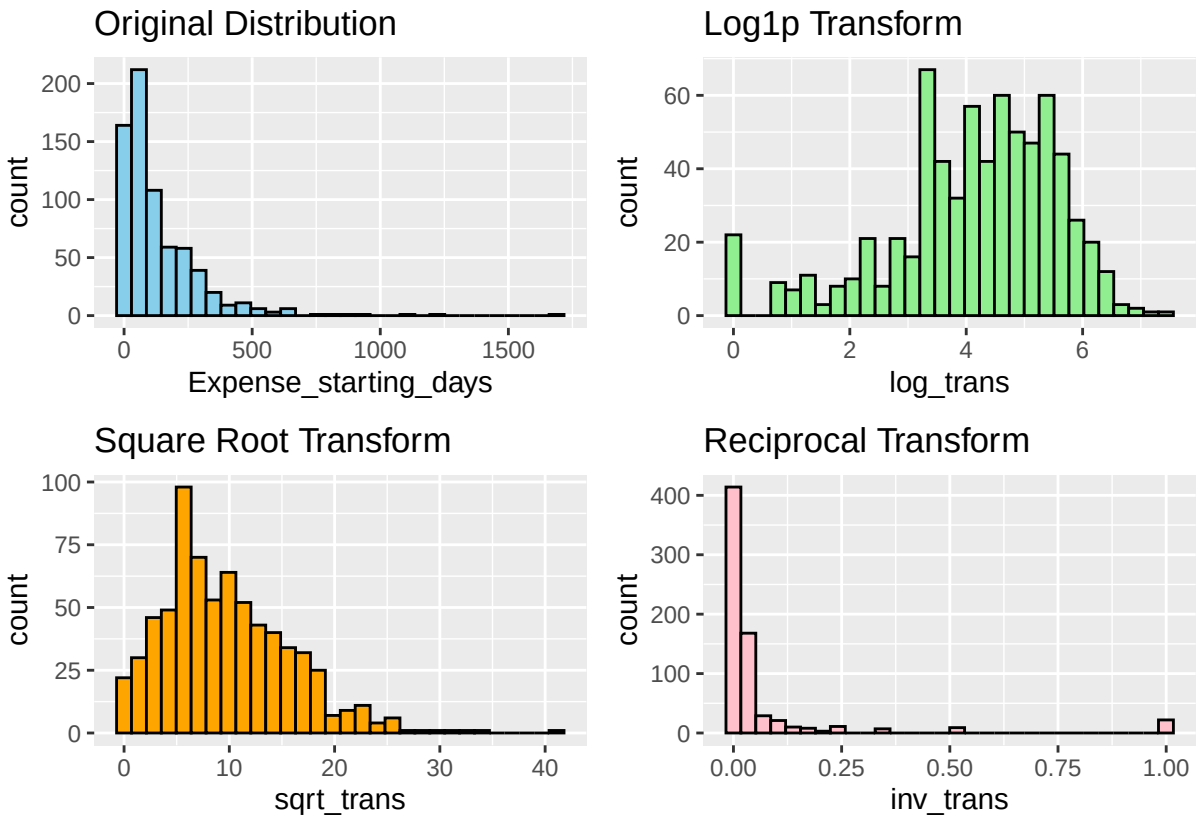
```
## # A tibble: 3 x 8
##   Variable            Effect   DFn   DFd     F            p 'p<.05'   ges
## * <chr>               <chr>  <dbl> <dbl> <dbl>        <dbl> <chr>   <dbl>
## 1 Academic_Semester   Group      2   699  6.78 0.001        *       0.019
## 2 Award Type          Group      2   699  5.62 0.004        *       0.016
## 3 Project Fund Source Group      6   695  7.71 0.0000000491 *       0.062
```

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
Log1p Transform
log_trans
Square Root Transform
sqrt_trans
Reciprocal Transform
inv_trans

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.

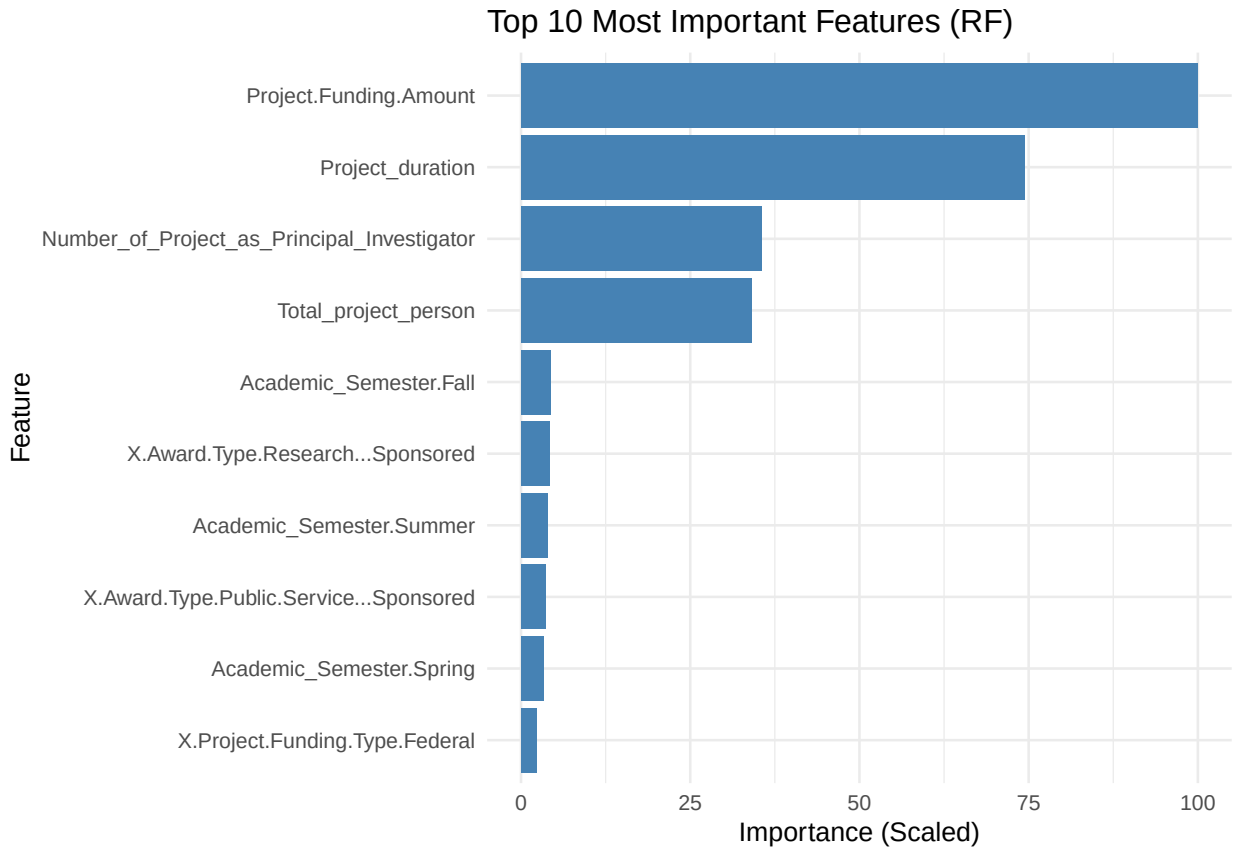
Model Comparison and Discussion

```
##               Model Train_minutes   RMSE   MAE
## 3     Random Forest          0.18 106.22 53.72
## 4           XGBoost          5.21 127.56 71.51
## 2     Decision Tree          0.01 156.58 87.52
## 1 Linear Regression          0.00 163.07 93.33
```

Among all models evaluated, Random Forest achieved the best predictive performance with the lowest RMSE (106.22 days) and MAE (53.72 days), indicating more accurate estimates of expense start delays. XGBoost followed closely but produced slightly higher errors (RMSE 127.56, MAE 71.51). In contrast, Decision Tree and Linear Regression showed substantially higher errors, suggesting limited predictive power for this task. Overall, Random Forest provided the most reliable predictions among the models tested.

Model Comparison and Discussion

Feature Importance plot:



The feature importance analysis from the Random Forest model highlights Project Funding Amount and Project Duration as the most influential predictors of expense start delays. These are followed by PI workload (number of projects) and team size, suggesting that project scale and complexity play key roles in determining when expenses begin. In contrast, categorical variables such as semester and award type contribute minimally to prediction accuracy, indicating they are less critical in explaining variation in expense start timing.

Save the best model: