Predicting Whether Kickstarter Project will be Successful

Sandy Weng

Abstract

The purpose of this project was to use classification models to predict whether Kickstarter campaigns be successful. I worked with data scraped by web robots and used numeric and categorial features to achieve this goal.

Design

The dataset contains 171,251 projects with 20 features, of which 3 are numerical. Some of the features include the funding goal, category, and campaign duration. I used f-beta score (beta=0.5) as the metric because I want high precision but still consider the recall score.

Algorithms

Feature Engineering

1. Converting categorical features to binary dummy variables.
2. Extracting the month and day of the week when the projects launched.
3. Subtracted the deadline from launched date to obtain the project duration.

Models

Logistic regression, KNN, decision tree, and random forest classifiers were used before settling on random forest as the model with best f-beta score.

Model Evaluation and Selection

The data was split into 80/20 train vs test sets, and all the scores were calculated with 10-fold cross validation on the training portion only. Predictions were done on the 20% holdout at the end after picking the model with the best f-beta score, so the scores were only seen once.

Final random forest 10-fold CV scores:

- Precision: 0.7259,
- Recall: 0.8038,
- F-beta score: 0.7402

Test Set

- Precision: 0. 0.7355
- Recall 0. 8168
- F-beta score: 0.7504

Tools

- Numpy and Pandas for EDA
- Scikit-learn for modeling
- Matplotlib, Seaborn, Plotly, and Tableau for visualizations

Communication

- Presentation and slides