Machine Learning Engineer Nanodegree

Capstone Proposal
Predicting Kickstarter Campaign Success (and Success Factors)

Robert Martinez
July 21st, 2018

Domain Background

(approx. 1-2 paragraphs)

Crowdfunding – namely, the practice of raising money for a venture from a large base of small donors via the internet - has become a popular form of funding for small, creative projects. Kickstarter was one of the earliest crowdfunding platforms to become a household name in this space, and as a result has a large dataset of projects with which we can explore the factors that contribute to the success or otherwise of fundraising efforts.

Gaining a better understanding of the success/failure factors for crowdfunding projects could benefit all sides of the market, through a reduction in the rate of failed (and cancelled) projects. Given a sufficiently accurate model, Kickstarter could predictively suggest adjustments in project specifications to project proposers (goal amount conditional on project category, or campaign length, for example). My own motivation for this project comes from some of my recent professional work, exploring funding support for the animation industry in a developing-country context, where traditional sources of financing are frequently hard to come by.

Problem Statement

This project aims to discover a classification model that can accurately predict the success or failure of Kickstarter projects given a number of project characteristics. In other words, for a given project, given a set of input features/project characteristics x, I aim to train a classifier $f(x)$ such that it accurately predicts the target variable y, which indicates whether the project was successfully funded or not.

Datasets and Inputs

The dataset to be used for this project comes was retrieved from Kickstarter by crowdfunding enthusiast Mickaël Mouillé, and reviewed by data science platform Kaggle. It can be obtained by clicking here. Two datasets are included, one featuring roughly 324,000 projects up to February 2016, and the other featuring

379,000 projects up to January 2018. The latter set was used for this capstone project.

The dataset gives information about projects along 15 dimensions, including:

- Project Name
- Project Category (Specific Category and Main Category)
- Project Launch Date and Deadline
- Country
- Currency
- Goal (in specified currency and USD if they differ)
- Number of Backers

Given the low dimensionality of the dataset, some new and hopefully illuminating features will also be constructed from the given features.

Solution Statement

Given the problem statement, I intend to train at least three models to try to discover an effective solution to the classification problem: a Support Vector Classifier (or SVM), Adaptive Boosting (AdaBoost) with a decision tree base learner, and XGBoost. The accuracy of these models will be compared to determine the best performer, and the best performer will be optimised using Grid Search cross validation, with F1 scoring.

Benchmark Model

The benchmark model for this project will be a simple logistic regression model (see here) which uses only the given features in the dataset (i.e. no feature engineering was conducted). This model achieved an accuracy score of 63.19%. Given that the problem to be solved is one of binary classification, one could have compared to a 'coin-toss' benchmark of 50% accuracy, but this is perhaps not the most challenging starting point. The solution I propose aims to improve upon the simple logistic regression, through a more robust model selection process, as well as feature engineering.

Evaluation Metrics

Given that the task of predicting successful or unsuccessful Kickstarter projects is a binary classification task, an appropriate evaluation metric to be used is the **accuracy** score, i.e. the ratio of correctly predicted observations to total observations. This simple metric allows comparison to other models, namely the benchmark model specified above, and is easy to interpret.

However, in the model selection process, I will also make use of the **F1 score** to judge the appropriateness of different models and hyperparameters. The F1 score is a useful metric for evaluating classifiers especially when the distribution of

classes is somewhat uneven. F-scores are generally calculated as the harmonic mean of the precision and recall scores. Precision refers to the ratio of true positives to all classified positives, and Recall refers to the ratio of true positives to all actual positives. A high precision score means a low rate of false positives, and a high recall score means a low rate of false negatives.

The F1 score is calculated as:

- 2 x Precision x Recall / (Precision + Recall)

Project Design

I propose a multi-stage process for conducting this project. First, I will import the appropriate dataset, then perform necessary data pre-processing. Namely, I will explore the dataset for missing values and choose appropriate steps if they are encountered - either choosing a strategy for filling them in, or omitting rows containing missing value - depending on how frequently they occur and the circumstances in which they occur. Next I will encode categorical data as numerical data where necessary (e.g. through label encoding or one-hot encoding). This is a necessary step, since machine learning models require numerical data to be computable.

Next I intend to explore the data visually (and through summary statistics), looking for interesting relationships that may help in building accurate, interpretable classifiers. I will also be on the lookout for new and potentially interesting features (e.g. length of project name, length of campaign) to create that may (or may not) influence the success of trained classifiers. If some important input features contain major outliers I may apply scaling/normalisation so that the outliers do not affect the training process negatively.

In the next stage, I will build and specify some models, namely SVM, AdaBoost with a decision tree base learner, and an XGBoost classifier, however if another model provides significantly greater performance I will include this model in the analysis. Once the best model is selected – based on accuracy score, and (if there are significant disparities) training time – I will optimise the hyperparameters of the model using the Grid Search cross-validation technique and evaluate based on accuracy and F1 scores, then display the output of the final optimised model.

Finally, if the best classifier contains a *feature importance* attribute, I will extract the most important features of the model to explain which factors possibly contribute most to the success or failure of Kickstarter projects.

Proposed Project Workflow:

1. Import Data
2. Data Pre-Processing
    i. Explore/Clean Missing Values
    ii. Encode Categorical Data as Numerical Data
3. Exploratory Data Analysis

    i. Data Visualisation
    ii. Feature Engineering
4. Model Selection
    i. Support Vector Classifier
    ii. AdaBoost Classifier
    iii. XGBoost
    iv. Evaluation Metrics: Accuracy and F1 Score
    v. Model Tuning: Grid Search
5. (Conditional) Feature Selection
    i. Extracting Feature Importances