Machine Learning Engineer Nanodegree

Capstone Proposal

Syed Umar Farooq October 31, 2018

Proposal

Domain Background

The internet age has given rise to a new mechanism of financing for creators and entrepreneurs: crowdfunding. Traditionally, investors and entrepreneurs would be expected to be located in close proximity, reducing the costs associated with distance. Yet today, the average distance between investors and entrepreneurs on crowdufunding platforms is approximately 3,000 miles. By directly appealing to consumers, crowdfunding campaigns, a successful crowfunding campaign can create enormous media attention, establish a customer base, and provide a large influx of capital. For instance, Oculus, a virtual reality company, launched a prominent campaign in 2012 for its headset and raised more than 2.5 million dollars. Three years later, the company was bought by Facebook for two billion dollars.

Hundreds of crowdfunding websites have been established, but an early entrant into the market, Kickstarter, is the largest crowdfunding platform in the world. More than fifteen million people have backed a Kickstarter project, with more than four billion dollars of funding being pledged since the site launched almost ten years ago.

Problem Statement

Like all forms of investing, crowdfunding has risks. Ambitious projects sometimes do not meet their funding target, which means they must refund their backers in full. A failed project can damage the company's credibility and drive backers away from that platform or crowdfunding generally. Therefore, it is important for both backers and and creators alike to understand how to run a successful crowdfunding campaign.

For more entrepreneurial projects, like some consumer technology projects, recent legal changes have made finding a successful project to support as an early adopter more lucrative. In 2017, Regulation A+, a section of the JOBS Act, came into effect and allowed companies to have mini-IPOs. This process can cost up to 90% less than a full IPO and be completed in a few months rather than a few years, giving early backers of major projects the chance to become investors and giving projects access to individuals that can serve as shareholders and brand advocates.

Still, 63% of projects on Kickstarter fail to meet their funding goals, and 13% never receive a single pledge. This rate is even higher for some types of projects. Technology projects, for instance, fail to meet their funding goals nearly 80% of the time. It is apparent that many project leaders are not able to successfully run a crowdfunding campaign, and many backers are not able to successfully back projects with a strong chance of being fully funded. **What factors, then, determine a project's likelihood of being successfully funded?** This is a binary classification problem, where the dependent variable is whether or not the project was successfully funded. I believe the likelihood of a project being successfully funded will depend largely on the size of the fundraising goal, with projects with smaller goals being more likely to meet their funding goals.

Datasets and Inputs

To answer the above question we will explore a dataset of more than 200,000 Kickstarter campaigns. The data, hosted by Kaggle (downloadable here), includes the project's name, ID, main category (e.g. technology), subcategory (e.g. hardware). the number of backers, the amount of money raised, the projects goal, the project's country of origin, and if the project was successful in meeting its funding goal (also known as project status). The project status will be used as the dependent variable, with ongoing or canceled projects being removed from the data set. This will make project status a binary dependent variable.

To determine the value of project status, a combination of categorical and continuous variables will be used. Continuous variables will include project goal and the number of backers. If projects with a large goal are more successful, it could suggest those campaigns have the most resources. If not, a large goal could indicate a project overestimating its capacity to raise funds. Number of backers can indicate the popular support for a project, and might be associated with a larger chance of project success.

Categorical variables will include subcategory and country of origin. Category will be restricted to a single value due to the sheer size of the data set and because there are dozens of potential subcategories. Restricting the category allows for a clearer analysis of the relevant factors. Because many high profile Kickstarter campaigns and acquisitions have been of technology campaigns, and because technology campaigns have the highest rate of failure, technology will be the category examined in this project.

Specifications:

Size of Full Data Set: 378,661 **Size of Data Set after converting dependent variable to a binary variable:** 331,675 **Categorical Features:** Subcategory, Country **Continuous Features:** Goal, Number of Backers **Features present in the dataset but that will not be used:** Currency, Date Launched, Deadline, USD pledged

Solution Statement

I propose predicting the likelihood of a project's success using bagged and boosted random forests. This can be measured through an F-score or by comparing the accuracy on the training set to that of the cross validation set.

Benchmark Model

The standard approach to problems with binary dependent variables is to use a logistic regression model. Therefore, the logistic regression model. This can be measured through an F-score or by comparing the accuracy on the training set to that of the cross validation set.

Evaluation Metrics

Fortunately, both the benchmark model and the solution statement can be measured using standard evaluation metrics for supervised learning techniques. I intend to compare the F-scores and accuracies for the benchmark and solution models on training and cross validation sets.

Accuracy measures the percentage of data points our model predicts that are the same as the data's actual value. Its values can have a complicated meaning, though. If the data fits the training set very well but performs poorly on the testing set, that suggests the model is overfitting on the training data.

An F-score is a metric that measures how many of our true positives were correct. It takes into account both recall and precision. In our example, recall would be the percentage of funded projects which were correctly identified by our model, whereas precision would be the percentage of projects that our project predicted were fully funded which were actually fully funded. F-score takes in a parameter referred to as beta. This parameter specifies the tradeoff between recall and precision. This problem in this proposal is broad enough to care about both recall and precision, but the main interest is in avoiding projects that are failures, so we will set beta to try and maximize precision.

Project Design

Exploratory Data Analysis

Before formally analyzing the data, I'd like to visualize much of the data. What do rates of success look like across different countries? Were early kickstarter campaigns more likely to fail? What percentage of projects succeed across various technology subcategories? Using bar charts I can visualize the number of successes and failures and hope to compare those numbers across various segments.

Data Preprocessing

The data needs to be cleaned up a bit before the work can begin. The current dataset contains data that is not needed for binary classification, such as the starting and ending dates for each crowdfunding campaign. I do not plan to use them because of the difficulty of incorporating dates into the models I have outlined. Other variables, such as currency, will be dropped due to the presence of a similar enough proxy (country).

I also need to convert my categorical variables into a series of binary explanatory variables through one-hot encoding.

My initial foray into the data has not suggested that there are any missing values. Should I determine that there are missing values but that the number is small (e.g. a few hundred) then I will implement listwise deletion and not use those rows in my analysis. Fortunately, my dataset is large and removing a small percentage of samples, while sub-optimal, is an option if absolutely needed. If missing data is present among categorical variables, I may attempt to classify those projects as "Miscellaneous" (meaning the country or category of that project will be "Miscellaneous" instead of, say "US" or "Hardware"). This will allow me to keep my data and get a better understanding of what patterns exist in the data points missing their categorical designations. If missing values are present in continuous variables, I may implement listwise deletion, as I do not think there is a good way to infer what the correct value of that data is/should be.

Finally, I will have to normalize my continuous variables. number of backers for Kickstarter projects can range from zero to the thousands, and the goals can range from five hundred dollars to hundreds of thousands of dollars. Normalizing this data through a log-transformation should reduce how skewed the data is.

Data Analysis

The fun part! Here, I will use both the benchmark and solution model and run them on the training data set. Then, I will attempt to tune both models using the training and cross validation sets to find better parameters (e.g., I will see if changing the bagging or boosting parameters will increase the accuracy scores without leading to overfitting). If time allows, I would also like to try combining bagging and boosting with unsupervised learning methods like k-nearest neighbors, to see if certain patterns emerge in the data (e.g., patterns among projects in the same subcategory or originating from the same country).

Results and Conclusion

Lastly, I will regroup and summarize what the data analysis reveals about what percentage of the funding success of Kickstarter technology projects can be determined by the explanatory variables we analyzed. I will also discuss room for improvement and what additional variables might have proven useful.