Machine Learning Engineer Nanodegree

Capstone Proposal

Vinícius Albuquerque
July, 2018

Proposal

Predicting if a project will succeed on Kickstarter platform.

Domain Background

Crowdfunding is a method of raising funds for a product/business/project through collective effort. That means that instead of the traditional way (which is having a small amount of investors that have large amount of the proposed service), people now ask for the community to help funding their products.

Kickstarter is a crowdfunding platform, thus, it is a intermediary where people can propose projects and ask for a specific amount of capital and in exchange offer something in return for those whom have helped achieve the goal, sometimes a first try on the product or a demo or a prototype.

Problem Statement

The goal of this project is to be able to predict if a project would achieve its goal of raising the asked amount of funding or not so people would know how to adjust their demands in a way that it would be probable that they would get the capital they are asking, thus having their projects possible to get done.

In stats collected on July 23rd, Kickstarter was used to successfully fund 147,688 projects and had 14,977,750 total funders (https://www.kickstarter.com/help/stats?ref=global-footer). As we can see, it is a platform accessed and used by millions of people so having a model that could successfully represent the results of the proposals would be rellevant to help those people know better how to get their needed funding.

Datasets and Inputs

The used dataset is a public dataset that can be found on the Kaggle platform (https://www.kaggle.com/kemical/kickstarter-projects).

Two different datasets can be found on this source but in this project we are only going to use the 2018's dataset so we keep the data as current as possible.

This dataset is decribed by having over 370k datapoints and 15 features. Some of the features are: category, amount pledged and date. These are some of the features that are going to be used as inputs to the model to predict the final status of the proposed project.

Solution Statement

The solution will be predict the final result of the proposed project having as parameters the information given by the person to the Kickstarter platform. First, we will have to have some kind of preprocessing to make the data workable to the chosen algorithms. Then, since it is a classification problem (either the project succeeds, or fails or it is canceled) we will choose a model that focuses on that type of issue, such as logistic regression.

Benchmark Model

I don't really have a benchmark to compare but my goal in this project is to be able to classify at least 3 out of 4 projects correctly, that means 75%. I'm considering that if my model does not achieve at least this percentage of accuracy, it is going to be considered as not good enough.

Evaluation Metrics

To define the accuracy of the model, I intent to use a grid search using the R2 score as my scoring function which will validate the predict data with the real actual data.

Project Design

The first steps of the projects would be to define which features would be relevant to our problems. Some of them seem to be giving the same information so maybe they can be reduced, others seem to be giving information that maybe are not relevant to the problem.

After that, a preprocessing will be made so the data type is represented in the same universe (numbers). I'm thinking of using panda's dummies transformation to convert categorical variable into indicator variables.

Now the data will be splitted and tested by various ways with some classification algorithms. I'll try to check if the data can be used as it is at this point or if it is better to have it normalized or maybe apply a PCA onto that. Then, after checking the results, I intent to use a GridSearch to try to improve even more the results that I already achieved at this point.

References

- Kaggle
- Kickstarter