Yan Xu’s Springboard Data Science Career Track Capstone Project 1 Proposal

Background

Kickstarter is a crowdfunding website that allows entrepreneurs to obtain fundings for their projects, some of them eventually leads to a startup company that provide either an unique product or service. Currently, the success rate of the projects on Kickstarter is around 50%. As a for profit company, Kickstarter will charge a 5% fee if the project gets launched successfully. The main purpose of this project is to analyze the data of past and current projects to provide some actionable insight for improving the project success rate on Kickstarter as its competitor Indiegogo started to gain more market share. Kickstarter can use the results of this project to either provide better guidance to entrepreneurs or find a better promotion strategies to attract more entrepreneurs who are likely to be successful.

Dataset

The dataset will be used in this project is from Kaggle.com. It is called Kickstarter Projects. It contains data for more than 300K projects on Kickstarter. The date range is from May 2009 until December 2017. The latest version of the data included data up to December 2017. There are 15 variables included whether the project was success or fail. Below is the description of each variable:

ID - Internal Kickstarter ID
Name - Name of project. A project is a finite work with a clear goal that you’d like to bring to life.
Category - Project Category. There are total of 159 categories.
Main Category - Category of campaign. There are 15 main categories.
Currency - Currency used to support, such as USD. There are total 14 different currencies.
Deadline - Deadline for crowdfunding
Goal - Fundraising goal. The funding goal is the amount of money that a creator needs to complete their project.
Launched - Date launched
Pledged - Amount pledged by crowd
State - Current condition the project is in. There are 6 states: Failed, Cancelled, Successful, Live, Suspended, Undefined.
Backers - Number of backers.
Country - Country pledged from.
USD Pledged - Conversion in US dollars of pledged column(Done by Kickstarter)
USD Pledged Real - Conversion in US dollars of pledged column (from Fixer.io API)
USD Goal Real - Conversion in US dollars of goal column(from Fixer.io API)

Project Approach

Data Cleaning: Check any missing data, check if those missing data included some information. Check outliers to determine if those data will be included in the analysis. Find any abnormality in data set

Initial Analysis: Define which variables are continuous, which variables are categorical and which variables are ordinal. Perform some initial exploratory data analysis through data visualization to find any interesting factors. EDA will included a scatter plot or line graph for continuous variables and histogram for categorical or ordinal variables. Such as if the location of the project having an impact on the success rate or the month of the project was launched will impact success rate. Find possible correlations between two variables and determine if the correlation will impact prediction results.

Actual Analysis: Randomly divide data set into three groups: Training Set (40%), Validation Set (30%), and Testing Set (30%). Use Training and Validation set to develop a decision tree (boosted tree, bootstrap forest) to find out the variables and its corresponding values that contribute the most in determining whether the project will be successful or fail. For example, if main category is Design or Technology, it will have certain percentage of success rate. Then find out the best model that will predict the outcome of the project and make prediction. Finally, use Testing Set to test the model to see how successful the model is. Provide confusion matrix to determine the accuracy of our model. Then look at the statistics to draw conclusion and provide some recommendations or further analysis opportunities.

Writeup

The deliverable will include a PowerPoint presentation and Python code. PowerPoint presentation will include all related EDA graphs, any correlation analysis, final model(decision tree), confusion matrix, related statistics summary, and conclusion and recommendations. Python code will include all the active code and its corresponding comments.