Abstract

Dataset

https://www.kaggle.com/kemical/kickstarter-projects

This dataset has more than 370,000 records of kick-starter projects. It's a collection of kick-starter data from 2009 to 2018, and a few data from 1970. There are 15 columns and columns are self-explanatory excepts following three columns :

usd_pledged: conversion in US dollars of the pledged column (conversion done by kickstarter)
usd pledge real: conversion in US dollars of the pledged column (conversion from Fixer.io API)
usd goal real: conversion in US dollars of the goal column (conversion from Fixer.io API)

There are massive amount of data, but the problem is we don't know what makes the projects successful yet.

Research Questions

What is the success rate of the entire project and each category ?
Is there any certain category more successful than others ?
Which project has highest success rate ? Which project raised the most funding ?
Is the funding period important ?
Is every project succeeded when the funding goal is achieved ?
What's the average number of backers of succeeded projects ?
Can we predict the success of projects beforehand?

Tools

R or Python

Techniques

Exploratory Analysis, Classification, and Regression.

First of all, I'm going to explore the dataset by visualization of research questions. And also find out the correlation between the columns. And then build a regression model to see if I can predict the project's success beforehand. And compare with other techniques to bring the best predictioin. After performing feature selection, I will also apply tree algorithm as well to see if I can classify the project's success and its accuracy for the entire projects. In this way, I hope to see the important factors contributing on the success of kick-starters projects and predict the likelihood of success of projects. I will use R for this project.