# **Introduction**

In recent years, the range of funding options for projects created by individuals and small companies has expanded considerably. In addition to savings, bank loans, friends & family funding and other traditional options, crowdfunding has become a popular and readily available alternative. Kickstarter, founded in 2009, is one particularly well-known and popular crowdfunding platform. It has an all-or-nothing funding model, whereby a project is only funded if it meets its goal amount; otherwise, no money is given by backers to a project.

A huge variety of factors contribute to the success or failure of a project — in general, and on Kickstarter. Some of these can be quantified or categorised, which allows for the construction of a model to attempt to predict whether a project will succeed or not. The aim of this project is to construct such a model and to analyse Kickstarter project data more generally, to help potential project creators, assess whether Kickstarter is a good funding option for them, and what their chances of success are.

## Business Problem

* Predict whether a Kickstarter Campaign is going to be Successful or Fail.
* What are the main factors that contribute to the success of a Kickstarter campaign?
* What is the best month/day to go live for a Kickstarter (is there one?)

## Project Approach

A picture containing graphical user interface

Description automatically generated

## Data Source

* The dataset used in this project was downloaded in .csv format from a web scrape conducted by a web scraping site called Web Robots.
* The dataset contains data on all projects hosted on Kickstarter between the company’s launch in April 2009 until the date of the web scrape on 24 March 2022.
* Merged data into one file (~ 2 GB), filtered for data>2018 and random sample of 0.3 %.
* Since target variable has 4 categories (successful, failed, live, cancelled). Kept rows for only 2 categories (successful, failed) in the target variable to convert this problem from multiclass classification to binary class classification.
* Final dataset has 65,270 rows.

## Data Cleaning

* Dropped duplicate rows.
* Flattened out JSON format columns like Category, Location to extract relevant features and merged with the original data frame.
* Dropped features with more than 70 % missing values (friends, is\_backing, is\_starred, permissions).
* Dropped features that won't add value to the model like ID, URLs etc.
* Converted Unix time date format features (encoded as number) to human readable date ( created\_at, launched\_at, deadline).

## Exploratory Data Analysis

* Number of Rows: 65,270
* Number of Columns: 15
* File Size: ~ 30 MB

## Variable Identification

|  |  |
| --- | --- |
| **Target Variable** | **Data Type** |
| state | Categorical |
|  |  |
| **Predictor Variable** | **Data Type** |
| spotlight | Categorical |
| ctgy\_name | Categorical |
| country | Categorical |
| is\_starrable | Categorical |
| staff\_pick | Categorical |
| lc\_state | Categorical |
| state\_changed\_at | Date |
| created\_at | Date |
| launched\_at | Date |
| deadline | Date |
| backers\_count | Numerical |
| pledged | Numerical |
| goal | Numerical |
| usd\_pledged | Numerical |

## Univariate Analysis

Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

**Continuous Variable**: We need to understand the central tendency and spread of the variable. **Categorical Variables:**For categorical variables, we’ll use frequency table to understand distribution of each category.

Chart, bar chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Table

Description automatically generated

Chart, histogram

Description automatically generated

Chart, pie chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Chart, treemap chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated