Predicting Kickstarter Project Success

BY RASHMI WALAVALKAR

Data Science Intern at Technocolabs

Table of Contents

1. **Introduction:**

On April 28, 2009 an online platform, named Kickstarter, was founded that acts as centralized marketplace to connect “creators” with the capital to pursue their visions. It helps creators post their projects and support fundraising for the survival of their firm.

Kickstarter is a funding platform for creative projects where everything from films, games, and music to art, design, and technology is placed together. The only mission of Kickstarter is to help bring creative projects to life.

2. **Aim:**

The purpose to work on this website or Platform is to browse through the entire genre of available projects and predict the project to be getting successful or not based on the other parameters.

3. **Data:**

The data used for my analysis is taken from webrobots.io website. The dataset is scrapped from the mentioned website in csv format. It comprises of three years data on a monthly basis excel sheets.To be precise it ranges from April 2019 to April 2021.

Each month data consists of 39 columns and 3k+ rows, where there are several missing values, incorrect and irrelevant information.

Column named Blurb consist of statements which further can be used as tokens to find relation.

Column named Backers_count consist of numeric data ranging from 0-count of iterations taken place in respective month.

Based on the Funding position of each task our Target variable will be showcasing the result.

4. Import packages

Initial step to begin with analysis is to import required packages likely to be used while working.

Basic Packages I would import:

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt
import plotly.express as px

from sklearn import preprocessing
```

5. Data Preparation:
   i) Data Retrieving >> On loaded data in csv form. It was found that theres no need of using sep or decimal at that moment.
   ii) On checking data type, if found that data is in object type. We need to convert the categorical data into category and numeric data into int or float values.

iii) Replacing few white spaces with mean, model is another important step. To change the case of all text data, “str.upper()” can be used.

iv) Make use an appropriate function to replace blank spaces with one of the following values: - a fixed value - the column-wise median value - the column-wise mean value - or ignoring all observations containing missing values.

6. Data Manipulation and Transformation:

i) After cleaning of data, another important aspect is to go through Data wrangling methods. check the dataset format. Many times, it is in incorrect form and therefore causes problem in performing EDA.

ii) Data Transformation is performed on dataset to achieve normality. Using logarithmic transformation on the right skewed distribution is reduced. But from the square-root transformation we can remove the skewedness from the data and normalized it.

iii) There are few outliers present in numeric values of dataset are checked. Dataset has some outliers which are identified and can be treated using capping technique.

7. Exploration:

i) Dataset consist of column named “blurb” for each project present in each datasheet. We employ text modelling approaches to extract valuable information from this column.

ii) For each given categorical x value and target y, mean encodings are performed whereas one-hot encoding is performed on parent category

iii) Word embedding is a NLP based process which captures the similarity between words. Similar to in-mapper function, here two different methods are used where neighboring words in a sliding window is used to predict center word and reverse.

8. Modelling:

i) Since our dataset is labelled structured data, we can use supervised techniques for training and testing.

ii) We can use Classifiers like j48 which proves to be a decent model producing least errors and overfitting.
E.g. On reducing the ratio or percent of parameters from 70% to 44% though training error remains unchanged overfitting increases which is most undesirable.

iii) Classifier IBK, where K values are manipulated ranging from 2 to 50 over train - test split data, test error rate is higher with high k values and eventually reduces with decrease in K value resulting in least overfitting

iv) Random forest is one of the ensemble methods that bags decision trees combining their individual predictions. Likewise, Random tree is a proven classifier.
E.g. on increasing the K value from 0 to 500 keeping other parameters constant, the RMSE for training set increased similarly for testing set it increased but it remained saturated after the interval.

9. **Results**:

Using Various classifiers, and encoding the target variable is forecasted. Keeping all the instances and events, each of the project is forecasted to be succeeding or failing.

10. **Conclusion:**

The TPR, FNR acquired from the modelling operations provide us F1 score, Accuracy, Precision and Recall values. These values suggest us the efficiency and reliability to the models. Thus, better the results better the model.

11. References:

I) About — Kickstarter
II) GitHub document