Title: Prediction of Success of a Project before release

Authors:

a. Name: Shubham Aggarwal, Net ID: sa4920

b. Name: Hardik Jivani, Net ID: haj272

Abstract: This project focus on predicting the success of a project before it is released in the market. This could save millions, if not billions, of lost dollars due to minor errors that could be debugged if carefully tested. The problem that we are solving is a prediction of the success of the project before it is released. If we can solve this problem efficiently, then people can run our model to test the success of the project and check for themselves on how to improve their project for greater success. We approached this problem by starting with data collection. We then performed data pre-processing and data visualization. We then performed data transformation and fed the data to the machine learning models like logistic regression and random forest classifier. Our results show that the choice of logistic regression gave very good prediction accuracy of the success of the project.

1. Introduction and Motivation

We believe that it might be too late to correct a mistake when everyone knows about it. Hence, we choose to predict the success of a problem before its released in the market. This could save millions, if not billions, of lost dollars due to minor errors that could be debugged if carefully tested. The problem that we are solving is a prediction of the success of the project before it is released. This could be a great use to the people as more and more people make projects and go-ahead for building a start-up. If we can solve this problem efficiently, then people can run our model to test the success of the project and check for themselves on how to improve their project for greater success. In order to achieve this, we applied most of the efforts in data cleaning, we then visualized the data to better understand it. We then transformed the data and applied machine learning prediction modelling techniques to get a 94% accuracy.

A huge variety of factors contribute to the success or failure of a project. Some of these can be quantified or categorized, 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 also to analyse Kickstarter projects data more generally, in order to help potential project creators, assess whether or not Kickstarter is a good funding option for them, and what their chances of success are.

There are many hurdles in solving this problem, some of them being working with the data collection of huge datasets from different sources. After obtaining the relevant data, cleaning the data is a big challenge here. Dropping irrelevant columns/values and performing feature engineering demands various operations to be executed in logical steps. After this step, understanding the data using proper visualization techniques is another hurdle since a data can have many different ways to be visualized but finding the key data points and designing appropriate visualization techniques for them make a huge difference in understanding what the data is trying to convey. Transforming the data and getting it ready to be fed to the Machine Learning Models is the next logical step. Finally, choosing an appropriate algorithm and getting a high value of prediction accuracy indicated how well the data was cleaned and prepared for the Machine learning models.

2. Methodology

We have divided the entire task of Data Analysis into manageable chunks of actions. We divided the task into these broad 5 categories as shown below and they further contain various small steps of data analysis:

Step 1: Web Scraping (Data collection)

We started with Data Gathering/Collection. This step allowed us to compile all the data that we need to work upon. The challenge here was to compile the relevant files to feed into our pre-processing step. We Started by collecting the URL's of all the files, then unzipped them and then found out the ones which were related to the projects on Kickstarter.

Step 2: Pre-processing the Data

Pre-Processing this huge dataset that we collected was a great challenge due to the size of the dataset. We performed various operations like dropping null columns, drop irrelevant columns from the dataset, getting column data into a suitable format, generated interesting features from the existing irrelevant features, generated multiple simple features from the complex column features and performed feature engineering on the short and long string types. This step was a huge step towards data cleaning. This step took a lot of time to prepare the data towards visualizing it.

Step 3: Exploratory Data Analysis

Now it’s time to understand the data by designing proper visualization techniques. This step is necessary to know what the data is trying to convey. Once we get the hang of the data, we can now talk to the data properly and perform transformations on it

Step 4: Transforming the Data

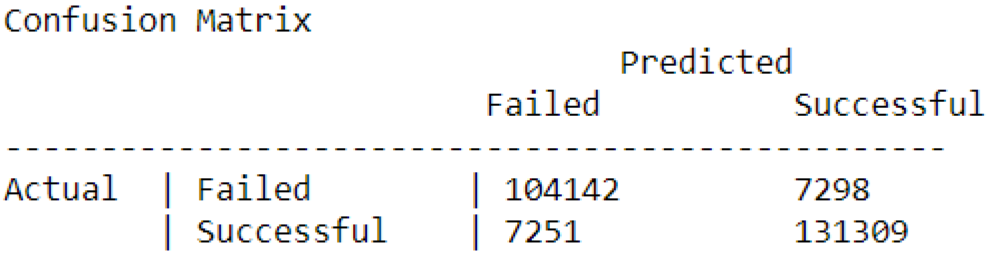
Now we prepare the data and transform it to make it ready to be fed to the Machine Learning models.

Step 5: Machine Learning Modelling of Data for prediction

Finally, we perform 2 models to check the accuracy of prediction of the success of a project, which was our initial data science problem that we trying to solve. We used Logistic Regression and Random Forest Classification to get the prediction accuracy.

3. Results

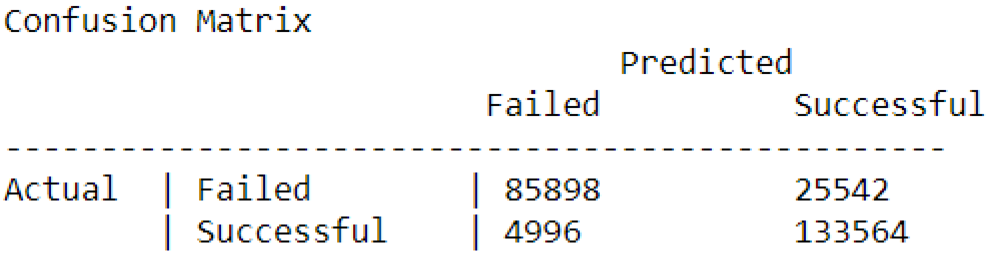
After applying the Logistic Regression and Random Forest Classifier for predicting the success of the project before its launched, we were able to achieve a high accuracy of over 94% using Logistic Regression. Logistic regression accuracy for training set is 0.94115 and Logistic regression accuracy for test set is 0.94164. Figure 1 shows the confusion matrix of Logistic Regression.



```
Confusion Matrix
                                Predicted
                         Failed          Successful
-------------------------------------------------
Actual  | Failed        | 104142          7298
        | Successful    | 7251            131309
```

Figure 1 Confusion Matrix of Logistic Regression

The results achieved by Random Forest Classifier were also comparably good. Random Forest a ccuracy for training set is 0.87677 and the Random Forest accuracy for test set is 0.87785. Figure 2 shows the confusion matrix of Random Forest Classifier.



```
Confusion Matrix
                                Predicted
                         Failed          Successful
-------------------------------------------------
Actual  | Failed        | 85898           25542
        | Successful    | 4996            133564
```

Figure 2 Confusion Matrix of Random Forest Classifier

4. Discussion and Conclusion:

Hence, we can see that after performing all the steps of Data Cleaning, Data Transformation, and Data Prediction modelling, we are able to get a 94.11% accuracy of predicting the result state of the project before its released using Logistic Regression technique. Logistic regression gave good results than random forest because, as the feature dimension goes large, the logistic regression goes on performing better than the random forest. Logistic regression optimizes the multinomial feature dimension dataset. Thus, our choice of selecting Logistic Regression turns out to be a good one. We hope to experiment with more ML prediction algorithms to try to test out the accuracies of every model. We can extend this model to techniques such as XG Boost classifier, Support Vector Machines Classifier to classify the new set of projects into success or failure.

REFERENCES:

[1] P. Yu, F. Huang, C. Yang, Y. Liu, Z. Li and C. Tsai, "Prediction of Crowdfunding Project Success with Deep Learning," *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, Xi'an, 2018, pp. 1-8.

[2] K. Chen, B. Jones, I. Kim, B. Schlamp, "KickPredict: Predicting Kickstarter Success", *Technical report California Institute of Technology*, 2013.