Predicting Status of Kickstarter Projects

MOSTAFA KHALID RAIHAN(0417052065)

MD. RIFAT HOSSAIN(0419052108)

1 INTRODUCTION

Kickstarter is a crowd funding site that maintains a global crowd funding platform focused on creativity. It tries to bring creative project to life. Kickstarter generates time series data containing different attributes of different projects and status of the projects. We used the March 2021 version of the dataset. Predicting the status of the project can be formulated as a data mining task. In this project, we tried to predict the status of the project by applying different algorithms and compared their performance.

2 DATASET

For this project we used the Kickstarter dataset for https://webrobots.io/kickstarter-datasets/. It contains 38 attributes of project and 213583 entries. It has four (04) status of the project- success, fail, cancel and live. In our dataset, there are 37830 successful project, 21721 failed project, 2671 cancelled project and 1853 live project. So it is clear said that our dataset is not a balance dataset.

3 DATA PREPROCESSING

Data pre-processing is a crucial part in data mining task. Performance of any data mining algorithm mostly depends on how one can successfully perform data pre-processing step. In this project we perform numerous data pre-processing tasks. As we described, we have 213583 entries and 38 attributes in the dataset, we discarded 4 attributes since it has only 96 non-null values and had little contribution to determine the status of the project. They were $friends$, $is\_backing$, $is\_starred$ and $permission$. Then we created some new features from the existing features. Some of them are category-name, project-duration-day, $per\_day\_usd\_req$, $per\_day\_usd\_pledged$. Also, it is possible to have redundant attributes in the dataset. So, in order to determine the redundant attributes, we used correlation matrix. If one attribute more than 90% correlated with another attribute then we remove one of them from the dataset since both attribute reveal almost same information in the project classification task. We used heatmap to show the correlation of different attributes.

4 ALGORITHM USED

The list of algorithm we have used for this project are -

(1) Decision Tree
(2) Random Forest
(3) K-nearest Neighbors
(4) XGBoost
(5) Logistic Regression
(6) ANN
(7) Naive Bayes

We measure the performance of these algorithm and compared there results.

Table 1. Experimental Result

Algorithm	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score	Roc_AUC
Decision Tree	0.999	0.929	0.933	0.929	0.931	0.876
Random Forest	0.874	0.957	0.937	0.957	0.940	0.965
KNN	0.905	0.905	0.879	0.905	0.891	0.838
XGBoost	0.958	0.958	0.962	0.958	0.938	0.969
Logistic Regression	0.927	0.927	0.889	0.927	0.904	0.831
ANN	0.874	0.874	0.876	0.874	0.858	0.912
Naive Bayes	0.815	0.815	0.874	0.815	0.838	0.799

5 RESULT ANALYSIS

In this section, we experimentally compared the performance of different algorithms on our dataset. Firstly we split the dataset into training and testing data(70-30 split). For the comparison of the performance of the algorithms, we used different matrices such as *Accuracy*, *Precision*, *Recall*, *F1Score*, *Roc_Auc*. We used the default settings of all the algorithm and Experimental results of different algorithms are shown in table 1. From table 1 we can see that Random Forest and XGBoost algorithm outperformed the rest and Naive Bayes algorithm performed the worst.

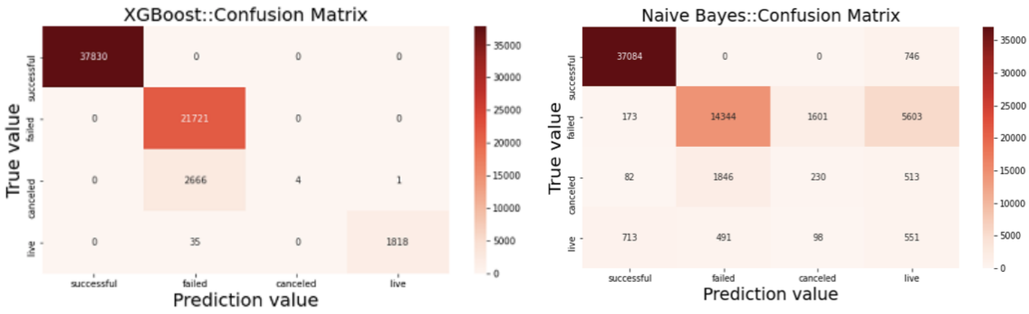

Fig. 1. Confusion matrix of XGBoost and Naïve Bayes algorithm

we visualized the confusion matrix for all the algorithms. As a reference to our result, here we show confusion matrix of XGBoost and Naïve Bayes algorithms in fig 1 as the best and the worst algorithm respectively for our dataset. From the figure, we see that XGBoost algorithm can successfully classify all successful project and Naïve Bayes can classify 37084 successful project out of 37830. For failed project, XGBoost classify all successfully whereas naïve bayes can classify 14344 out of 21721, for cancelled, XGBoost can classify only 4 and Naïve Bayes can classify 230 out of 2671. Most of the cancelled project classified as failed project. The main reason for this poor performance is that attributes of the cancelled and failed projects are similar and our model recognized cancelled project as a failed project. Finally, for live project, XGBoost classified 1818 and Naïve Bayes classified 551 out of 1853. In Naïve Bayes, most of the live project classified as successful and failed project since live project can be successful or failed in future.

Finally we calculated the precision, recall and F1 score for successful, failed, cancelled and live projects individually. We found that all the algorithms perform well for successful and failed projects but perform very poor for cancelled and live projects. The reason for this is the same as we describe above. As a finding of our work, we found the best classification techniques for our Kickstarter project dataset and best algorithm for our work is XGBoost algorithm.

6 CONCLUSION

In this project, we studied about Kickstarter projects and its attributes. We tried to classify the project successfully. We applied various algorithms and compared their results to find out best algorithm for this dataset. We also performed many pre- processing task like feature creation, feature selection and performed many measurement and visualize there results.

7 PROJECT LINK

https://github.com/mostafa-K-raihan/Kickstarter-Status-Predictor