Individual Project Report – Kickstarter (a crowdfunding platform)

Modelling Methodology: The data was initially explored and large number of missing data from *category* and *launch_to_state_change_days* was replaced with the constant value. For modelling, the *feature selection* process was followed and based on logic the variables such as *project_id* and *name* were removed from the training data since they were unique for every project with no predictive power. The variables such as *deadline, state_changed_at, created_at and launched_at* didn't have any importance as day,month,year and hour were separately mentioned in the data along with the number of days between creation to launch and launch to deadline. The *disable_communication* was a unary variable (consisted only 15685 False observations),so it was eliminated. The *currency* was removed as country would provide the same analysis as currency. All the variables related to *state_changed_at_**[1] has been removed since the project focussed on only success and failure states so before deadline there won't be no changes in the project status. The variables such as *backers_count, staff_pick, spotlight* were removed as these data elements would not be known before the launch of the project and also these directly determines the success/pledged value of the project which was clearly observed from the correlation matrix. The variables *name_len, blurb_len,created_at_yr and launched_at_yr* were removed to avoid collinearity issues due to high correlation with one other variable.

Regression Task: For this task, based on logic some features like *state* and *pledged* were removed as these data elements would not be known before prediction. In addition to this we needed to predict USD pledged which is just the conversion of pledged value into USD so if we already knew pledged value then there is no point of prediction, therefore *pledged* was eliminated. *Lasso method* was applied with optimal hyperparameters[2] and based on the feature importance; 14 useless predictors were eliminated. Models were then built on the optimal features and after the application of various models it was found that **Random Forest Model**

[1] This refers to all the variables with "state_changed" key word

[2,3,4,5] This refers to the optimal hyperparameters used in various models for tuning (coding file contains all details of hyperparameters for each model).

performed better as it generated the lowest **MSE** of ***$16.5 bn*** with optimal hyperparameters[3] and therefore this model was finalized for the regression task.

Classification Task: Similarly for this task, *pledged, usd_pledged and static_usd_rate* were eliminated as these data elements won't be available to us for projects whose state needs to be predicted and they directly determine the success of the project as observed from the correlation matrix. The Random Forest model was applied with optimal hyperparameters[4] to identify the feature importance of the remaining variables and about 60 irrelevant predictors were eliminated (assumed coefficient value < 0.01). Models were then built on the optimal features and after the application of various models, it was found that ***Gradient Boosting Model*** generated the highest **accuracy score of approx. 74%** with optimal hyperparameters[5] as compared to other models and therefore this model was finalized for the classification task.

Clustering: For the clustering task, I have taken the most relevant features(9) which are significant for the prediction of success/failure of the project and can give many insights. From the Elbow and Silhouette method, four optimal number of clusters were obtained which were then treated with K-Means algorithm to form clusters of different variable characteristics. K-Means algorithm was used as it is less computationally expensive as well as it works well and faster with large dataset.

Cluster Characteristics: On average, in cluster one, 38% of the successful projects had a goal of $20k and 62% of failed projects had a goal of $71k. Similarly, in cluster three 21% of successful projects had mean goal of only $18k and 79% of failed had a mean goal of $357k. Cluster two has only 64 successful projects with mean goal of $203k, whereas cluster four has $17k mean goal for 34% of successful projects and $60k mean goal for 66%failed projects. This implies that setting a right goal is very important to be successful and the goal set wisely under a reasonable range is more likely to be successful than the projects with huge goal

[1] This refers to all the variables with "state_changed" key word

[2,3,4,5] This refers to the optimal hyperparameters used in various models for tuning (coding file contains all details of hyperparameters for each model).

amounts. It was also noticed that for a project to be successful the number of backers and the pledged amount plays a key role and this was clearly illustrated by all the clusters and lesser the number of backers and pledged value, the projects have higher likelihood of failure. Most of the successful projects requires ~30 to 40 days on average to be completed except for projects in cluster three which requires 56 mean days. When it comes to the length of the name of project, there is not much difference between the successful and failed projects on an average in all clusters. On an average the name length of the successful projects lies between ~5 to 6 and without keywords is ~13. Similarly, on an average the successful projects are completed in nine months in cluster one, whereas in cluster two and three it gets completed in 7 months but in cluster four it gets completed in just 4 months. This implies that shorter the duration of the projects, more successful they are. In terms of year, 2015 was quite interesting because in cluster three and four, 40% and 46% of the projects failed respectively and 27 % succeeded in the same year in cluster two. In cluster one and three ~33% of projects were successful in 2014 and 2016 respectively.

Business Benefit: Kickstarter makes money by taking 5% of the total amount of money that is funded on the site. The funding only happens by the backers when the projects are successful on the deadline day. This model predicts the success/failure rate of the projects with 74% accuracy, 65% precision and 53% recall means that algorithm would identify projects as successful with the probability of 53%. This would ensure that company is able to invest in good creative projects by advertising and marketing them effectively so that they get more pledges from the backers and the chances of project becoming successful goes up. Similarly, the model with MSE of $16.5 bn, will be able to predict the pledged amount in USD with only 11.23% of the variation based on the r squared value. So overall, this predictive model will help Kickstarter to predict success/failure rate of projects with higher accuracy before having any information of backers and the pledged value on projects.

[1] This refers to all the variables with “state_changed” key word

[2,3,4,5] This refers to the optimal hyperparameters used in various models for tuning (coding file contains all details of hyperparameters for each model).