Individual Project Report

Kickstarter is a crowdfunding platform with the aim to help bring creative projects to life through investments from the public. Since the launch of Kickstarter in April 2009, the company has successfully funded around 190,000 projects with a grand total of $5.5 billion pledged. Each project would have a deadline and a minimum funding goal. If the project does not attain the required amount of money upon deadline, the funds collected would be refunded, and the project would be deemed to be failed. The company has two major stakeholders: Creators and the Backers. With the main revenue stream coming through transaction cost, the company strives to attract more backers and creators to use their platform crowdfunding. With over 6 million repeat backers, Kickstarter has revolutionized the crowdfunding space by rewarding the backers with amazing experiences and limited editions of the creative work in return for their support.

The key to attracting more customers to invest through their platform is if they could know before hand if a project launched would be successful or not, which can then be used for marketing purposes. With data on millions of successful and failed projects, we would be able to analyse the information and come up with machine learning models that could efficiently predict the potential of a newly created project based on their characteristics like the time of launch or the length of their name. In this project, I have used a sample of data on projects launched between the years 2009 and 2016 and used certain characteristics to predict how much a project would have pledged and the probability of success. For this, a regression and classification model has been created, respectively. I have also developed a clustering model which can be used to find similarity between different projects based on certain features. These models could be used to determine what characteristics make a successful project, target backers based on location, and analyse the performance of different project categories.

Data Pre-Processing:

Since the supervised models are intended for testing live projects, some of the characteristics are not considered for training the models since they cannot be realized during the live state. These include the number of backers a project has, if the project was featured on the Kickstarter page and whether the project was picked by a staff. I have also removed the time and date variables associated with current state change since these would be identical to the time of launch as we are testing projects upon their launch. The timeline of a project from its creation to deadline is an important aspect and has been used in the model. I have also computed the day of the week for launch, creation, and deadline. This can be compensated for the weekdays in the dataset and can be treated as continuous variables. I also found out the earliest dates of launch, creation and deadline for projects in my sample dataset, and used it as a base date. This was then used to compute how recent each project was related to the base dates. Country of origin and category of projects were also taken into consideration for the models. Here, countries like USA, Canada and Great Britain, and categories associated with technology were significant and used as standalone predictors. Finally, the goal set for each project was calculated in USD and added as a predictor. Once the above steps have been executed, the Isolation forest has been used to identify and remove the top 2% of anomalies from the dataset. The subsequent dataset split into training and test set and used for training the regression and classification models

Regression Model:

The Random Forest algorithm was used predict the amount of money that a project could pledge upon deadline. The algorithm was tuned to have 100 decision trees in the forest with each tree taking a maximum of 7 features from the original set used for training. Random states have been set for the Random Forest to replicate the results obtained. Here, the Random Forest yields an MSE of 7,274,184,052.74.

Classification Model:

The Gradient Boosting Classifier was used to predict whether a project would be success or failure upon deadline. The algorithm was set to create 100 decision trees to aid in classification with a min_samples_split of 3. Upon training the model, it was able to accurately classify around 75% of the test set with an F1 score of 59.2%.

Clustering Algorithm:

The K-Means clustering algorithm was used to group observations based on their outcome, desired goal, the number of days since earliest project launch date in the sample dataset, number of days from creation to launch to deadline, the length of name and blurb, and categories like Sound, Hardware, Wearables, Gadgets and other categories. The algorithm was able to yield 8 distinct clusters with a mean Silhouette score of around 0.63. From the clusters, I was able to infer that recent projects categorized as Gadgets with long name and blurb, tend to succeed, if they have low goals, an average create to launch time and high launch to deadline time. Similarly, the older projects categorized as Hardware, that also have long names and blurb, tend to succeed, if they have low goals and high create to deadline time. On the other hand, older projects categorized as Hardware tend to fail if they have very high goals. Projects like these tend to have very high create to deadline time and comparatively low name and blurb length. Likewise, projects that are more recent and categorized as Gadgets are more probable to fail if they have high goals and a small project name. Finally, recent projects categorized as Sound and Wearables have an even chance of success or failure, if they have an average goal compared to other projects, average time between create to launch and a comparatively low length of name and blurb. Last but not least, projects categorized as other, tend to fail if they have very high goals and very high create to deadline time, and tend to succeed if they have very low goals and create to deadline time.