Kickstarter: Individual Project

For the first part of the project, I began by importing and pre-processing the data. I first dropped the column ‘name’ as it has a different value for each record, so it is not useful in our data mining algorithm. Then, I set the ID of the projects as an index so I can be able to refer each row to its project ID. I noticed that the variable ‘launch_to_state’ has 13182 missing values and that the ‘category’ variable has 1684 missing values. I first dropped both columns and proceeded with my analysis. However, after running several models, I noticed my accuracy scores were low so I figured that category could be an important variable to include in my analysis. Hence, I went back to my preprocessing and kept it in my dataset instead of dropping it. Next, I kept the failed and successful goals solely and converted the goal of each project to a goal in the USD currency (to have it on the same rate) using both ‘goal’ and ‘static_usd_rate’ variables. I also wanted to visualize the impact of days on the project. In other words, I wanted to see if a project that has more days to fund, would be more likely to succeed or fail. Thus, I performed feature engineering and created a new variable ‘create_to_deadline’ in the dataset from existing variables ‘create_to_launch’ and ‘launch_to_deadline’.

My first approach was to group some variables together to gather my dataset and perform well in my models. However, after some trial-and-error, I noticed it was better to dummify my variables as I would get better accuracy scores. After going through all the variables available in the dataset, I inserted in my X variable what I deemed to be relevant and observable at the time of prediction to predict the project’s failure or success. I then proceeded with dummifying my categorical variables ‘category’, ‘country’, ‘created_at_weekday’, ‘deadline_weekday’, and ‘created_at_year’ and adding them to my X variable in addition to ‘name_len_clean’, ‘blurb_len_clean’, ‘create_to_deadline’ and ‘goal_usd’. I replaced the values of my target variable being ‘failed’ and ‘successful’ with 0 and 1 respectively.

Then, I standardized my data and began performing different classification models such as logistic regression, KNN, Decision-tree, Random Forest, GBT and ANN. I varied the hyperparameters and random_state to test my scores and ran some loops to find optimal values, but also tested my models with and without standardization (except for the ones where standardization is necessary such as KNN and ANN models). The reason why I performed all classification models was to evaluate each performance and be able to see which model performs best.

Hence, my best accuracy score (0.7478149648262631) was found with the Random Forest model with a standardized dataset. However, my second-best accuracy score (0.7465359198465146) was found with my GBT model (non-standardized). Since the difference in my results between both models is minimal (roughly 0.1%), I decided to choose the GBT model over Random Forest because, in a business context, gradient boosting is more accurate than Random Forest given the fact that it builds one tree at a time and the next tree is a correction of the errors made by the previous tree. Hence, each tree learns and improves on the previous. It is also one of the most powerful techniques to build predictive models, especially in the case where we have a large dataset (such as Kickstarter) that could have a lot of missing values (as we've seen in the preprocessing) because it can deal with these missing values and outliers without the need to manually eliminate them. However, I kept my pre-processed data as it is to improve my model's performance. GBT is also best in a business context for its speed and accuracy improvement but also because it avoids overfitting (with cross-validation).

For the second part of the project, I pre-processed the data with the same steps that I took in the first part. However, since I will be getting insights from the clusters, I figured I could create new columns out of the dataset that I have, to make my clustering analysis more efficient. Hence, I grouped the countries by continents: America, Europe, Oceania and Asia. Then, I grouped the days of creation by weekdays and weekends (to see if projects created on weekends have more impact than projects created on weekdays) and did the same for the deadlines (weekdays and weekends). Finally, as 'category' is an important variable, I did some research on Kickstarter to visualize the different sectors of these categories and grouped my variables into four groups: Technology, Theater, Film and Photography_Music_Academic (it could also be considered as a category of "others"). Then, I divided the project's years of creation into two groups: those created before 2013 and those created after 2013 to make it visually better to interpret. I also dummified 'state' and added it to my X variable (that contains the same variables as in the first part except that my variables are now grouped and contain the variable state).

To get my optimal clustering number k, I first performed PCA on my X variable to reduce the data by finding a linear combination of my predictors such that the combination contains most variations and so we lose as little information as possible. I then performed the elbow method using my new X variable from the PCA analysis to find my optimal k. Since this method is subjective and commonly used in academic research, I decided to use it because with k=4, my graph almost doesn't decrease anymore, so I subjectively picked my optimal number of clusters to be 4 (run the plot to see the graph in Python). Then, I proceeded with clustering using KMeans, where I inserted the number of clusters being 4 (given the elbow method's result) and I printed the clustering results.

Given my results, projects in cluster 0 have the second lowest goal of projects, they are created in Oceania, their deadline is on a weekend, and they consist of technology projects. They have also been created after 2013 and these projects have failed. Projects in cluster 1 consist of projects with the lowest goal, their deadline were on a weekday, and they mainly consist of projects in Theater and are created before 2013 in Asia. Their state is successful. Projects in cluster 2 have a very high goal in USD, they are created in America, with both creation and deadline being on a weekday. They were created after 2013 and these projects have failed at achieving their goal. Projects in cluster 3 are created in Europe, they are technology projects that have a deadline on a weekday and were created on a weekday too, after the year 2013 but they have failed.

We can assume that among all clusters, those that have projects grouped with high goals have failed given the fact that it is harder to pledge a higher project cost than it is for a lower one.

(Interpretation is subject to change as levels change when running the data every time).