INSY_662 Individual Project Report

Data Preprocessing

The variable launch_to_state_change_days was removed from the dataframe, given a high number of missing observations. Disable_communication was dropped, as it was a unary variable. The variables deadline, state_changed_at, created_at, and launched_at were dropped from the dataframe because they contained year, month, day, and time information grouped together for each row value and it was decided to analyze those features individually through other variables. Post-launch features, project_id, and name were removed, and country, category, currency, created_at_weekday, and launched_at_weekday were dummified. The target variable state was dummified as well, with state_successful = 1 representing a successful project and state_successful = 0 representing project failure.

Classification Model

A random forest model was built using gradient boosting. Using cross-validation, the optimal min_samples split for GBT was determined to be 3, and the optimal n_estimators was determined to be 400. These features yielded the highest model accuracy score of 0.712, thus making this the final model used for the classification task. The precision, recall, and F1 score of this model were 0.618, 0.471, and 0.535, respectively. According to the gradient boosting model, the variables that have the highest influence on a project's success (using a feature importance score threshold of 0.05) were the length of a project name, the year of project launch, the project category "Plays," the category "Software," and the category "Web." The variables with the greatest influence on a project being successful are the Web category, with a feature importance score of 0.173, followed by the Software category with a score of 0.136. All models used the same random_state of 5. Overall, the gradient boosting classification algorithm used is correct

71.2% percent of the time. If the model predicts a project as a success, this prediction is correct with 61.8% probability, and the model identifies successful projects 47.1% of the time.

Clustering Model

Data pre-processing for the clustering model was the same as in the classification task, with the exception that post-launch variables were included in the clustering analysis. A K-Means clustering model was built, with the elbow method graph indicating that the decrease in inertia became insignificant from 5 clusters onwards. Even though using four clusters yielded a higher total average silhouette score (0.953) than when five clusters where used (0.917), the author decided to use five clusters instead of four because five clusters reflected the distribution of the clusters in the scatter plot more accurately. The Pseudo-F statistic measure yielded the same p-value for models using 2,3,4, or 5 clusters (1.11*10^-16), indicating that the model performance for values of k ranging from 2 to 5 was not markedly different.

Clustering Model: Business Insights

In the clustering model, the variables static_usd_rate and backers_count for projects were plotted. We can observe five different clusters in the plot (see below). Projects that come from a country where the currency is worth significantly less than the US Dollar (red cluster) typically have fewer backers. Projects that come from a country where the currency is worth significantly more than the US Dollar, or who are from the US (green cluster) have greater numbers of backers. The cluster with the highest number of backers is observed to be from the United States (orange cluster), where the static_usd_rate is 1. Finally, the purple and blue clusters show an even distribution of project backers across countries where the currency is of similar value to that of the United States. Kickstarter projects can be crowdfunded from a number of different countries around the world. Furthermore, the US dollar is typically viewed as the benchmark for a currency's strength, with US dollar conversion rates = 1 or greater than 1 associated with a stronger currency, a stronger economy for that country, and a higher purchasing power (especially internationally) for that country's citizens. According to the Kickstarter website, the majority of backers for projects, especially in the initial stages, come from a creator's existing network and within that creator's country. International donors are observed less frequently on Kickstarter. It is possible that in countries where the currency is worth markedly less than the US dollar, the citizens of that country (from whom the majority of the pool of backers will be drawn from) will have lower purchasing power and less interest, or ability, to donate their money to Kickstarter projects, thus resulting in a lower number of backers for Kickstarter projects. In countries where the currency is worth more than the US dollar, consumers from projects' origin countries, who will make up the bulk of the backers of those projects, will have higher purchasing power and will be more willing to donate their money to Kickstarter projects, thus resulting in higher number of backers for projects where the USD conversion rate is significantly greater than 1. These factors could explain the variation in the numbers of backers, based on the static_usd_rate. Finally, the majority of Kickstarter projects are created in the United States, which could explain why the highest number of backers are observed in the United States, where static_usd_rate = 1 on the plot.