Data Mining and Visualization – Individual Project

Submitted by: Arunima Agrawal, 261001915

Business Use Case:

Kickstarter enables creative projects to meet the right backers for funding. Imagine if Kickstarter could communicate the success/failure prediction of projects to the backers before the pledging occurs. If the backers know which projects have higher likelihood of being successful, they would be more easily attracted towards funding greater amounts of money for larger number of projects. We are attempting to understand the various characteristics of projects and the predictions about the success/failure of projects.

Data Pre-Processing:

The dataset consists of fields describing various characteristics of projects. The 'state' field indicating the success or failure of the project, was the target variable used for training the classification model. The first step was to prepare the data to be fed into different models:

- Removed rows where the state field not 'successful' or 'failure'
- Dropped columns: i) that are decided after the project is declared as successful or failure (like pledged, backers count), ii) that are insignificant in determining success (like project name, USD rate), iii) that have very high number of NULL values (like launch to state change days), iv) that are highly correlated with the predicted variable state (like spotlight), v) that have the same value for all the rows (like disable communication)
- Removed all the rows where any of the predictor values were NULL
- Created dummy variables for categorical fields and joined them in the main dataset.

Classification Model:

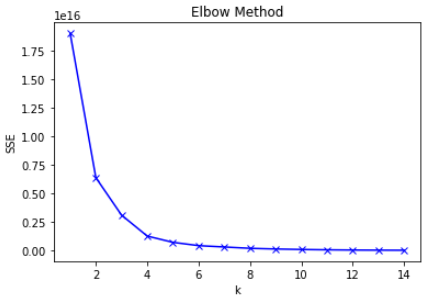
Ran various classification models including Gradient Boosting Technique (GBT), Logistic Regression, K-Nearest Neighbors, Random Forest Algorithm and CART Decision Trees. Received the highest accuracy (0.7800) and F1 Score (0.6280) from running the GBT algorithm.

Further, performed feature selection by finding their optimal parameters using loops. Used GBT for evaluating the performance since GBT accuracy was highest without feature selection:

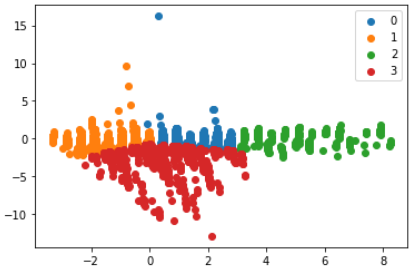
- Recursive Feature Elimination (RFE): Ranked all predictors, ran GBT on various combinations of predictors (like combination 1: rank 1; combination 2: rank 1,2, and so on). Highest accuracy (0.7803) and F1 Score (0.6286) achieved by using first 81 predictors.
- LASSO: Found optimal value of alpha by running loop from 0.01 to 0.05, with a step-size of 0.001, and obtained the same set of predictors as RFE, and therefore the highest accuracy (0.7803) and F1 Score (0.6286) remained same as RFE results.
- Random Forest: Obtained feature importance for each predictor. For the set of feature importance values, ran GBT by selecting predictors with a minimum feature importance from the set. By taking the minimum feature importance value of 0.001346, obtained the highest accuracy (0.7807) and F1 Score (0.6299).

Also tuned hyper-parameters using Grid Search Technique. However, the accuracy and F1 Score of the results hence obtained were slightly lower than the initial model. Therefore, finalized 62 predictors for running the GBT algorithm, where the feature importance is greater than 0.0012. By running GBT on these 62 predictors, obtained the highest accuracy and F1 Score. Also computed results using the 62 predictors on Logistic Regression, K-Nearest Neighbors, Random Forest Algorithm and CART Decision Trees. However, GBT continued to be the best-performing model.

Clustering Model:



The clustering analysis was performed only on the numerical fields that were identified as important predictors from random forest feature selection in the classification model. The analysis did not take the categorical fields into account since K-Means clustering works best with numeric data. Using the elbow method, identified 4 as the optimal cluster count. Also computed the silhouette score for these clusters (0.7759) and hence, deemed the clusters good. Since plotting 7 predictors on the graph is difficult, ran Principal Component Analysis to extract two principal components, which were plotted on the graph, colored based on the cluster number.



Next, centroids of each cluster were plotted for each of the 7 predictors used for clustering. Hence, we note that the first cluster has low number of staff picked projects, small number of days between project creation and public launch date, medium blurb length, and the projects launch, creation and deadline dates were recent. Likewise, projects in all the 4 clusters can be interpreted.