Overview

Crowdfunding is the practice of collecting money from multiple individuals or sources to finance a new project. Often, crowd funders turn to social media to share their platform or idea to inspire others to contribute to the crowdfunding campaign. Kickstarter is one such platform helping individuals fund their creative projects.

The project helps out Kickstarter identify the future state of a project being a success or a failure at the time of project launch. This will help attract backers to fund projects with a higher probability of success.

Data Pre-Processing

The Kickstarter dataset explains the characteristics of every project concerning dates, amounts, country, etc. Upon exploring the available dataset, it was recognized that there were a few steps to be performed for cleaning the data before the model could be run on it.

- Since the model had to be developed for the instance where the project was getting launched, only “success” and “failure” observations were filtered on the state column.
- There were a few columns that weren’t relevant for the launch day predictions, like pledged, state_changed_at, backers_count, etc which were dropped.
- Variables insignificant to the analysis, like project_name were dropped. Also, the variables with a large number of NULL values were dropped. Also dropped date columns as the split for the day, date & year is given in separate columns
- Few observations ~8%, were removed that had NA values.
- Converted “state” values to binary for simplification in modeling.
- Dummified the categorical columns in the remaining dataset, joined the dummy columns.

Classification Modeling

Multiple classification models were used to predict the success/failure of a project. Firstly, five base models were used with a train-test split of 67% & 33%. Below are the accuracy scores for all five base models.

Model	Logistic Regression	KNN	CART	Random Forest	Gradient Boosting
Accuracy	0.6685	0.7401	0.7650	0.7870	**0.7900**
F1 Score	0.0477	0.4572	0.5434	0.6431	0.6414

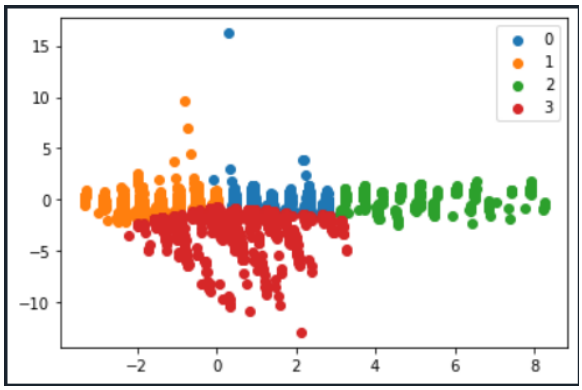
Post the base classification models, Gradient Boosting was chosen for further enhancements as it achieved the highest accuracy. Feature selection was performed over all the predictors using Recursive Feature Elimination (RFE), LASSO, and Random Forest selection. The random forest selection chose the best 22 predictors, although, the accuracy did not improve for the GBT model. The number of predictors was chosen in RFE dynamically as well. Again, the performance of the model was best when taking all the predictors for the training.

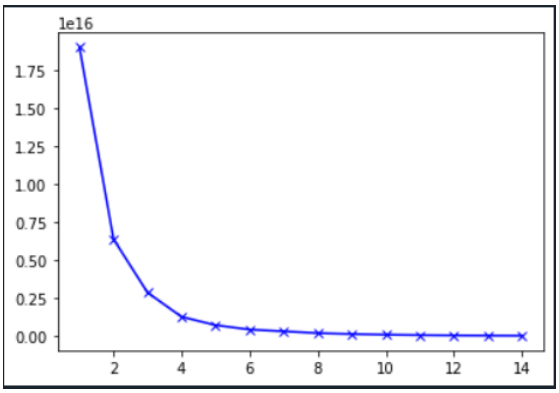
For further deep-diving into the tuning, hyperparameters were tuned using the Grid Search algorithm to obtain a more accurate model. The max depth was dynamic from range 1 to 20, n_estimators kept 50 or 100, and learning rates provided were 0.01, 0.1, and 1. Once again, the model did not improve basis the hyperparameter tuning.

Tuning the predictors, selecting the best possible predictors through feature selection, and comparing with different kinds of classification models, it was concluded that the Gradient Boosting method for classification with n_estimators as 100 and min_sample_splits as 3 performed the best. Hence, Gradient Boosting Model was chosen for classifying the project as successful or failure.

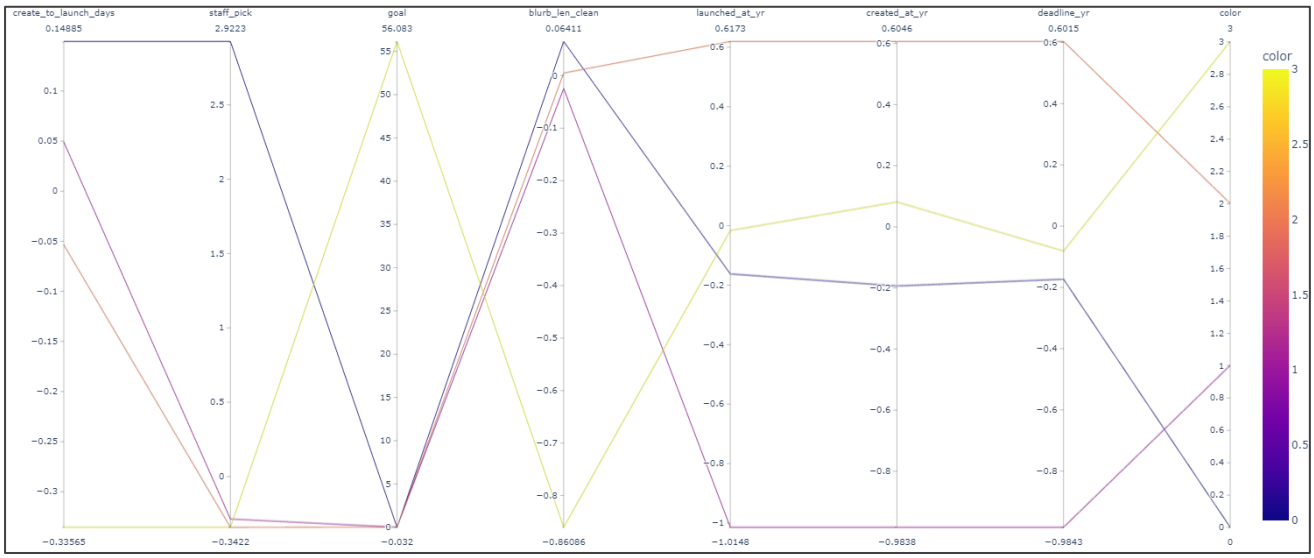
Clustering

For clustering, the preprocessed dataset (similar to classification) was used. To identify the optimal number of clusters, the elbow method was run and it was concluded that the optimal clusters would be k=4. Also, two-component PCA was used for interpreting the 7 selected predictors more accurately. The cluster formation for the 4 clusters is as below:





Next, a parallel coordinates plot was plotted using the Plotly library to interpret how the chosen variables were distributed across the four clusters. (Silhouette Score: 0.7736)



Cluster 0 has high values for create_to_launch_days, staff_pick, blurb_len_clean. Similarly, cluster 1 had only high blurb_len_clean, cluster 2 had launched_at_yr, created_at_yr, deadline_yr, while cluster 3 had highest values for goal.