Classification Model

I picked random forest as the classification model because random forest is one of the best performing models that is also very robust due to its ensemble nature. I picked a subset of predictors from the input data because only those subsets of data specified below are known at the launch time of the project.

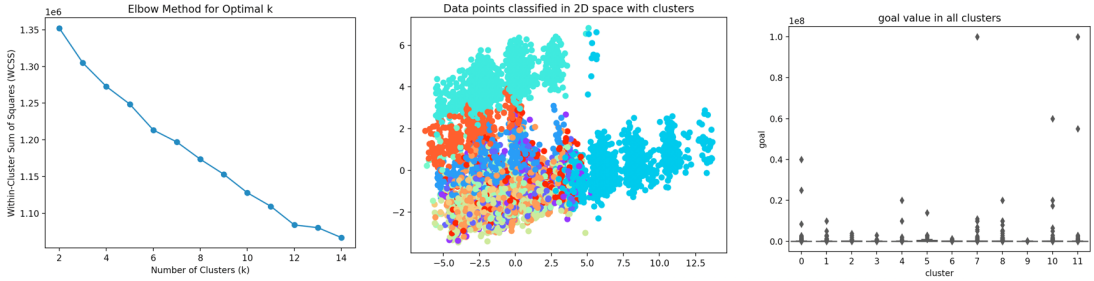
```
predictors = [
    'goal', 'country', 'currency', 'disable_communication', 'create_to_launch_days',
    'deadline', "created_at", 'static_usd_rate', 'category', 'name_len', 'name_len_clean',
    'blurb_len', 'blurb_len_clean',  "deadline_weekday", "created_at_weekday",
    'deadline_month', 'deadline_day', 'deadline_yr', 'deadline_hr',
    'created_at_month', 'created_at_day', 'created_at_yr', 'created_at_hr',
    'launched_at_month', 'launched_at_day', 'launched_at_yr', 'launched_at_hr',
    'create_to_launch_days',
    'launch_to_deadline_days',
]
```

In addition, I utilized feature engineering to extract useful features from datetime columns. I also introduced RFE feature selection, which choose a subset of features until the performance does not increase with more features for better model performance. Basically, the model takes the features that are known at the launch time of the project and use GridSearchCV for hyper parameter tunning to find the best set of hyper parameter base on 5-fold cross validation score. The best model is then obtained and used for prediction. The idea of random forest is that it combines multiple decision tree classifiers and pick the majority vote from the decision tree classifiers.

Clustering Model

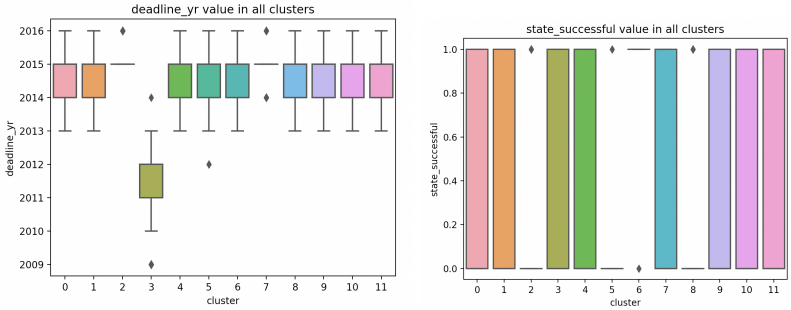
For Clustering, I utilized KMeans clustering algorithm, which is an unsupervised machine learning algorithm used for clustering data into groups or clusters base on distances.

To determine the optimal K and to balance the need for high accuracy and the ability to interpret and explain the results in a business context, I decided to pick k = 12, this is based on the Elbow Method. We can observe that the drop for Within-Cluster Sum of Squares flattens a bit after k = 12



From the PCA reduced dimension plot, we can see that we have a relatively well separated clusters but there are overlaps between clusters in the lower right corner, I then generated distribution and statistics about each cluster to investigate further, the 3rd image above is one example. 1) cluster 7, 11 seems to have some outlier values with very high goal, and the cluster with the highest success rate, which is cluster 6 have relatively lower goal values. 2) cluster 8, 9 seems to have the lowest pledged value and cluster 6 have highest. 3) cluster 3 and 6 seems to have the highest backers_count. 4) static_usd_rate seem to be the highest in cluster 4, lowest in cluster 9. 5) name_len are mostly the same across all clusters with some minor deviations and cluster 7, 8, 9 seems to have the lowest name_len value. 6) cluster 3 contains older projects that have deadline mostly in 2011-2012, others are newer, and correspoundinly cluster 3 also have older state_changes_at_yr. 7) cluster 7 is created later in the year comparing to to other clusters 8) cluster 2, 5, 8 have majority unsuccesful records while cluster 6 contains mostly succesful rows with a success rate of 0.9948, the average success rate is as follows: cluster 0 has a mean success rate of 0.327858, cluster 1 has a mean

success rate of 0.280751, cluster 2 has a mean success rate of 0.242553, cluster 3 has a mean success rate of 0.471493, cluster 4 has a mean success rate of 0.398673, cluster 5 has a mean success rate of 0.183333, cluster 6 has a mean success rate of 0.994801, cluster 7 has a mean success rate of 0.321598, cluster 8 has a mean success rate of 0.028114, cluster 9 has a mean success rate of 0.267606, cluster 10 has a mean success rate of 0.276306, and cluster 11 has a mean success rate of 0.368889.



I’ve also obtained the clustering centroids and write it into cluster_statistics.txt file and the size of the 12 clusters are as follows: 2312, 1658, 1583, 1439, 1397, 1154, 1105, 1051, 900, 705, 71, 60. We can see that the variance of cluster size is relatively large.