INSY 662 Individual Project

Yichen Wang (260761601)

Q1.

For this task, a classification model is required to predict whether the variable "*state"* will take the value "*successful*" or "*failed*". To implement this, *DecisionTreeClassifier* is used to complete the task.

The first step is to clean the data. For this model, we need to first drop columns with many null values, which are "*category*" and "*launch_to_state_change_days*", with 1684 and 13182 null values respectively. There are also other 4 columns each having 5 null values, in this case we only drop the corresponding row. Then, *project_id* and the *names* of the project are dropped as they don't contribute to the prediction. All the dates are dropped as well since the date information can be hard to be dummified, and they have their own specific data columns as well. Furthermoe, "*usd_pledged*","*name_len*" and "*blurb_len*" are removed because they have similar columns (clean version), which may lead to collinearity problems. All the states that are not successful or failed will be dropped and object type predictors will be dummified, such as *"country","currency"* and so on. Finally, features that can only be considered after launch needed to be dropped, including *"spotlight" (perfectly correlated to state as well so not included), "staff_pick","backers_count"* and all the features related to state change.

For feature selection, I use the *RandomForestClassifier* to fit the current data and print out all the predictors with their corresponding feature importance. Finally, top 5 features with highest feature importance are selected to be the predictors: *'goal', 'create_to_launch_days', 'name_len_clean', 'create_at_day','launched_at_hr'*.

By splitting the data into training and test data by a ratio of 0.33, I trained the *DecisionTreeClassifier* model and the *accuracy_score* on the sample grading data is 0.6815, showing an acceptable accuracy.

According to the feature importance provided by the *RandomForestClassifier*, the goal amount set is the most important factor to the success of the project. It could be deduced that an unreasonable goal may scare away backers, while a good and realistic goal may give backers confidence. Other factors such as *“name_len_clean”* and *”create_to_launch_days”* also contribute a lot, meaning that a clear and intuitive name and a suitable preparation period for the project launch will likely indicates success.

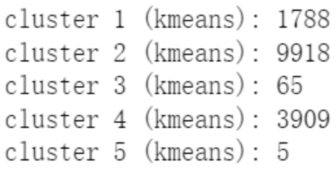
Q2.

The second task requires to develop a clustering model to group projects together. Since we have many different predictors, it would be wise to filter some before doing clustering. Otherwise, the clusters will have lower similarity within and a higher similarity to records outside, which is undesirable.

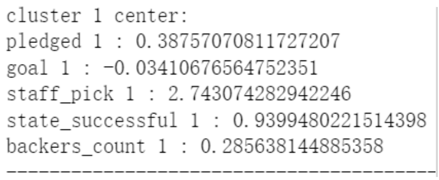
Similar to Q1, some columns and rows with null values are dropped. I also dropped all the rows where the *state* is not *successful* or *failed*. For faster and more accurate purposes, only five features are selected based on feature importance performed similarly as before as these will provide more meaningful insights: *'pledged','goal','staff_pick','state_successful','backers_count'*.

Before clustering, all the features needed to be standardized. The clustering method I used is KMeans so a pre-determined number of clusters is required. This could be determined by using the elbow method. Since adding another cluster almost always reduce the variance, a number k that improves little should be our target k. In this case, k=5 is a good choice. To test whether the cluster assignment’s performance, the silhouette method is used, and the silhouette score is equal to 0.856, much greater than 0.5, meaning it provides good evidence of the reality of the clusters in the data.

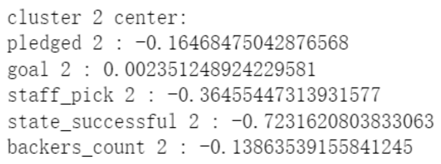
The results for 5 clusters are shown below, as well as the information about their cluster centers:



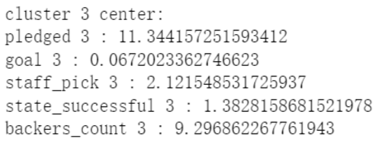
```
cluster 1 (kmeans): 1788
cluster 2 (kmeans): 9918
cluster 3 (kmeans): 65
cluster 4 (kmeans): 3909
cluster 5 (kmeans): 5
```



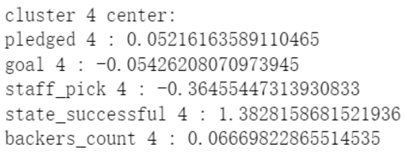
```
cluster 1 center:
pledged 1 : 0.38757070811727207
goal 1 : -0.03410676564752351
staff_pick 1 : 2.743074282942246
state_successful 1 : 0.9399480221514398
backers_count 1 : 0.285638144885358
----------------------------------------
```



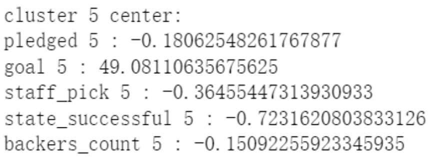
```
cluster 2 center:
pledged 2 : -0.16468475042876568
goal 2 : 0.002351248924229581
staff_pick 2 : -0.36455447313931577
state_successful 2 : -0.7231620803833063
backers_count 2 : -0.13863539155841245
```



```
cluster 3 center:
pledged 3 : 11.344157251593412
goal 3 : 0.0672023362746623
staff_pick 3 : 2.121548531725937
state_successful 3 : 1.3828158681521978
backers_count 3 : 9.296862267761943
```



```
cluster 4 center:
pledged 4 : 0.05216163589110465
goal 4 : -0.05426208070973945
staff_pick 4 : -0.36455447313930833
state_successful 4 : 1.3828158681521936
backers_count 4 : 0.06669822865514535
```



```
cluster 5 center:
pledged 5 : -0.18062548261767877
goal 5 : 49.08110635675625
staff_pick 5 : -0.36455447313930933
state_successful 5 : -0.7231620803833126
backers_count 5 : -0.15092255923345935
```

We can obtain a lot of insight from these results, especially about how the several features affect the rate of success. Since the data has been normalized, value equals to 0 meaning is among average.

- For cluster 1, it has an astonishing high value for *staff_pick*, meaning if the project has been promoted and highlighted on the site, there will be a great chance of achieving success.
- For cluster 2, almost all the features were below average, and the success rate is very low. This indicates that if there is nothing special about your project, there is little chance being successful.
- For cluster 3, the amount pledged and the number of backs are insanely high, and it is also being picked by the staff, meaning if there's a lot of capital and resources provided, there is a great possibility of getting a success.
- For cluster 4, it is very similar to cluster 1, except it has a lower percentage of being selected by the staff. However, if the goal amount is reasonable and the project has some amount pledged with some backers, there is still great chance being successful.
- For cluster 5, if a project wants to raise an insane amount of money but with below average amount of resource support, it is nearly doomed to be failed as investors will have no faith in such projects.