Arnaud Guzman-Annes

ID: 260882529

I. Introduction

Kickstarter is an American crowdfunding company that gives Internet users the opportunity to finance projects that are still at the idea stage by reducing the burdens associated with traditional modes of investment. The goal of this project is to create 2 supervised models (regression and classification) and 1 unsupervised model (clustering) with the Kickstarter data. An analysis of the benefits that these models can bring in a business context is then presented.

II. Regression

To begin with, the data is cleaned in order to use the most reliable observations. For this, the "successful" and "failed" states are kept. For obvious reasons, the data that is collected after the launch of the project is removed. Finally, categorical variables are dummified. Then, Isolation Forest is used to remove anomalies from the dataset. For this, a contamination factor (C) of 0.05 is used in order to obtain and drop 700 anomalies (~ 5% of the total observations).

Once the anomalies are removed, Random Forest is again used to determine the most important predictors. With a Gini coefficient (GC) greater than 0.01, these predictors can be selected.

Now, it is possible to start modeling. For this, the model that performed the best in terms of MSE is GBM with cross validation. It should be noted that in order to improve this performance, the model has been hyper tuned. Table 1 shows the MSEs obtained with different models.

Table 1: MSE for regression

CART- Cross val.	11.5 Billion
RF- Cross val.	11.6 Billion
GBM-Cross val.	11.3 Billion
CART-Train test	16.7 Billion
RF-Train test	14.0 Billion
GBM-Train test	14.2 Billion

III. Classification

For this model, the data preprocessing steps are very similar. Indeed, the steps of cleaning the data, removing the anomalies and selecting the predictors are exactly the same (with C = 0.05 and GC = 0.005).

Therefore, the model which performed the best in terms of accuracy, recall and F1 score is GBM with test-train split. Table 2 to Table 4 show the different scores obtained with different models.

Table 2: Scores with CART classification model

Accuracy	71.2%
Precision	56.1%
Recall	50.0%
F1	52.9%

Table 3: Scores with RF classification model

Accuracy	71.2%
Precision	67.1%
Recall	26.2%
F1	37.7%

Table 4: Scores with GBM classification model

Accuracy	74.5%
Precision	62.5%
Recall	52.5%
F1	57.1%

IV. Clustering

Regarding the unsupervised model, the data preprocessing steps are still identical to the two previous models (C = 0.05). The choice of predictors is made with Random Forest and the top 5 numerical predictors used are: usd_pledged, backers_count, goal_usd, name_len_clean and create_to_launch_days. With the elbow method, 5 is identified to be the optimal number of clusters (Figure 1).

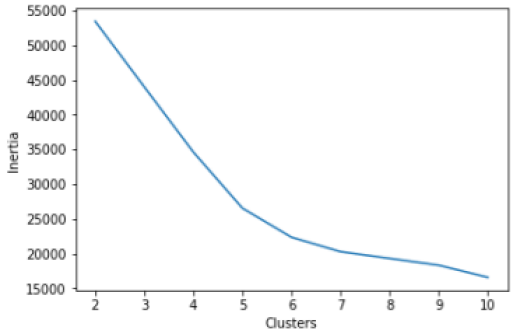

Figure 1: Elbow method. n = 5 cluster are selected

Moreover, Silhouette's method confirms that 5 of the 4 clusters are well placed and that one might be more difficult to define (Table 5).

Table 5: Silhouette score for each of the five clusters

Cluster 1	0.302
Cluster 2	0.386
Cluster 3	0.479
Cluster 4	0.330
Cluster 5	0.129

Cluster 1: This cluster groups together projects with the longest names and a large number of days between their creation and their official launch.

Cluster 2: This cluster looks a lot like the previous cluster except that projects have fewer number of days between their creation and their launch.

Cluster 3: This is the cluster with the most observations (6922) and it groups projects with a low pledge, few backers and a relatively small goal.

Cluster 4: This is the cluster that has the fewest observations (3) and it stands out from the others since it groups together projects with the highest goal.

Cluster 5: This cluster groups together projects with a high pledge, a large number of backers but a relatively small goal. Note that this is the cluster with the smallest Silhouette score (0.129) and that could provide scant evidence of its reality.

V. Conclusion

In this project, we created supervised and unsupervised models notably with GBM. This technique builds an ensemble of shallow trees in sequence with each tree learning and improving on the previous one. Furthermore, cross-validation is usually the preferred method to validate models because it gives it the opportunity to train on multiple train-test splits. This indeed gives a better indication of how well the model will perform on unseen data. However, in a business context, the model has its limitations. Indeed, the MSE is high, the scores (accuracy, precision recall and F1) are relatively low and, in terms of Silhouette scores, one cluster is not well defined. In this perspective, it would be relevant to obtain more data and predictors to make the model more accurate. In fact, the current model risks giving predictions that are not up to the demands of potential clients such as investors or managers of the platform.