McGill University

Desautels Faculty of Management

INSY662

Data Mining and Visualization

Individual Project:
Kickstarter

Sebastian Salazar 260983868

Regression Model

The first step in the regression task was to apply Data Cleaning, as was clear that the target variable distribution had very extreme outliers on its right tail. For such, all observations with a Pledged amount higher than US$ 438,805 (99^{th} percentile) were dropped. Additionally, some projects had no Category and were classified as "Others" to avoid dropping these observations.

Next, feature engineering was applied to the dataset, creating four new variables for the model with the following purposes:

- Creating variables closer to the target variable: 'Goal_USD' (as 'goal' was in the project's local currency) and 'Goal_USD_PerDay_LaunchtoDeadline' (to calculate how much the project was expecting to make per day, which can be related to project realism).
- Simplifying already existent predictors: 'Region' (the continent of the project's country) and 'Launch_Quarter' (grouping the month of launch to the quarter of the year).

Following that, 11 predictors are dropped because they do not add value to the regression task due to being repetitive or having no logical relationship to the target variable. Also, other 11 variables are dropped because they are invalid predictors as they are determined after the project in launched which beats their purpose as predictors. Then, the predictors were dummified to apply Isolation Forests and eliminate anomalies from the dataset. This resulted in a dataset with 85 predictors and 14,906 observations.

For feature selection, the filtered dataset was standardized, and the Random Forest Regressor algorithm was run creating a rank of all predictors based on their importance. To determine the optimal number of features a Gradient Boosting Regressor algorithm was run in a loop that calculated the MSE through cross validation with 5 folds. The model with the lowest MSE determined the number of predictors.

In conclusion, the **final regression model uses the 41st most important predictors resulting in a test MSE of 1,317,825,226.38 USD dollars.** From a business perspective, it would be recommended to drop more outliers as 75% of the dataset has pledged amount lower than $7,000.

Classification Model

Given that this model did not depend on continuous variable with high variance the dataset was not cleaned of outlier like in the Regression Model. However, for consistency this model's dataset underwent the same feature engineering process and 22 predictors were dropped (11 for repetitiveness and consistency, and 11 for being invalid). Then, the predictors were dummified resulting in a dataset with 85 predictors and 15,685 observations.

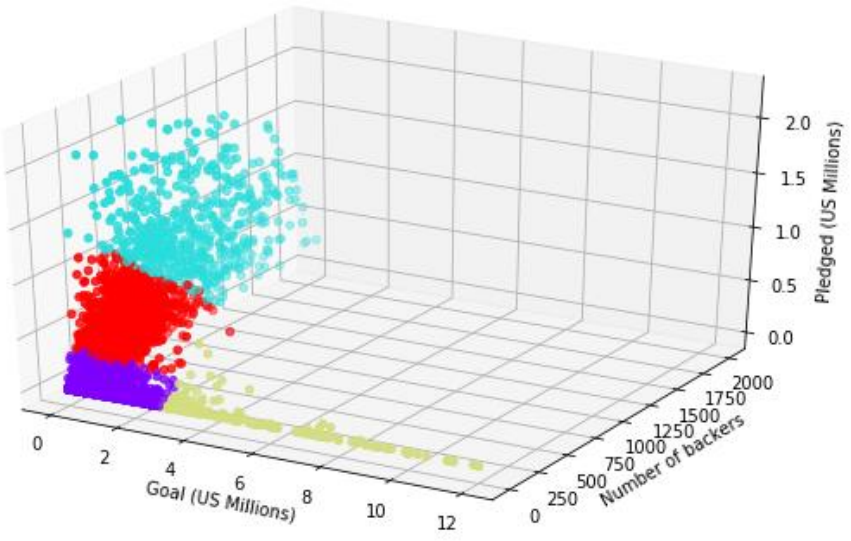
For feature selection, the filtered dataset was standardized, and the Random Forest Classifier algorithm was run creating a rank of all predictors based on their importance. To determine the optimal number of features a Gradient Boosting Classifier algorithm was run in a loop that calculated the accuracy through cross validation with 5 folds. The model with the highest accuracy determined the number of predictors.

In conclusion, the **final classification model uses the 65th most important predictors resulting in a test accuracy of 75.21%.** Business-wise, this model is very accurate in predicting the success of a project, additional hypermeter-tuning might increase its performance even higher.

Clustering Model

For the clustering model it was determined that project will be clustered by three predictors: 'goal_usd' (featured engineered), 'usd_pledged' and 'backers_count'. The dataset was cleaned of anomalies through Isolation Forest resulting in 3 predictors and 15,685 observations.

Clustering was done through K-Means algorithm, which required the dataset to be standardized. The optimal number of clusters of 4 with the Elbow Method, resulting in the following clusters:



Small/Medium Niche Projects: These projects have a small or medium funding goal and appeal to small number of backers which could lead to a successful project if the goal is small or a failed project if the goal is relatively high.

Kickstarter's Big Failures: These projects have a big funding goal and appeal to small number of backers which results in sure failure for the creator.

Kickstarter's Standard Projects: These projects have a small or medium funding goal and appeal to regular number of backers. These projects are somewhat innovative and appeal to backers needs leading to successful funding most of times.

Kickstarter's Golden Projects: These projects are one of a kind, innovate and appeal to a big number of backers. Even though they might initial have a small/medium goal, they are so well received that they break their goal multiple times and become and instant hit.