Data Mining and Visualization Individual Project Report

Kartikeya Lakhanpal – 260914945

Objective

The objective of the task assigned is to predict the performance of a startup featured on KickStarter by predicting the state(successful/failed) of the startup and the amount pledged to it. This report explains the process behind creating two models, one for classification of startups and one regression model for predicting the pledged amount.

Phase I – Data Exploration

The dataset provided was an excel file that had columns that explained features about a kickstarter project and the first task was to find patterns and features about the dataset. The following is a table discussing interesting features of the data columns.

Column Name	Feature
Project_id,Name	Not considered due to identity issues and to avoid extreme overfitting
goal	Inconsistent due to multiple currencies, therefore converted into **goal*static_usd_rate**
pledged	Inconsistent due to different currencies, column **usd_pledged** preferred.
State	Initially Consisted of 5 states, but the data has been processed now to only consider failed or successful projects. Failed projects are approximately twice the number of successful projects
Staff_pick	Not considered because the value of staff pick is realised after the project is launched

Backers_count	Not considered because the value of backers_count is realised after the project is launched
Cateogry	Included. Some values of category were found to be *null* and were replaced with "Others" value for easier dummification.
Spotlight	Not Considered because the value of spotlight is realised after the project is launched
Deadline_month/hour Day/year	Included, however to decrease the number of dummies the columns were split into buckets of specific range. This is also done, because whether the day is in the first week of a month is more relevant than the actual day itself. The assumption is based on business rules like the fact that salaries are disbursed during first two weeks of a month.
Launch to state change days	This column had more than 60% of its value as *null* and therefore it was removed so as the maintain the coherence of data.

Phase II – Identifying high collinearity

A correlation matrix was generated to identify for columns for high collinearity. It was observed that the values in countries and currency were highly correlated with pearson correlation value greater than 0.5 and therefore it was decided to remove the column **currency**.

Phase III – Feature Selection

Two feature selection models were used to compare and identify the set of features and model that gave us high accuracy for Classification Model.

i. Recursive Feature Selection
    a. Logistic Regression – This model gave us around 58 predictors with an accuracy of 69% approx.
    b. Random Forest Regression – This model gave us 128 predictors with an accuracy of 71%
    c. Support Vector Machine with a Linear Kernel – The result was inconclusive as the computation time was subjectively long.
    d. An assumption was made that SVM with a different kernel would not give us a higher result than random forest.

ii. Random Forest Feature Selection
    a. Logistic Regression – This model gave us around 30 predictors and an accuracy of 65%
    b. Support Vector Machine with Linear/Rbf/Sigmoid Kernel -The accuracy for these ranged between 67% to 73% with linear kernel showcasing the best performance.
    c. Random Forest Regression – Gave us an accuracy of 75% with around 107 predictors.

Comparing the performance of all models, it was observed that Random Forest Classifier under a Random Forest Feature Selection gave us a higher accuracy at an optimal number of predictors. The code for the above-mentioned models can be found commented in the appendix section of the attached code.

Two feature selection models were used for Regression Model.

i. Lasso – Lasso was run at varying levels of a penalty function(alpha). It was found that the feature selected at alpha 50 performed optimally with a Random Forest

Regressor. Higher Alphas gave us better performance, but they were not chosen to avoid the possibility of overfitting. A lower alpha was not chosen due to the obvious difference in performance.

ii. Random Forrest – Random forest feature selection gave us comparable performance to Lasso. The number of predictors was higher in random forest as compared to Lasso and therefore Lasso was chosen.

Please note that the code for the models mentioned above is available in the appendix section of the attached code.

Phase IV – Model Building

Random Forest Algorithm was used for both classification and regression problem statements as their performance was higher to other models as stated previously.

The measure of accuracy in a classification model is the accuracy score and the measure of accuracy in a regression model is the mean squared error value.

Therefore, while selecting the hyper-parameters for each model, the aim was to include that hyper parameter that increase the accuracy score for the classification model and decreases the MSE value for the regression model.

A loop was written to find the combination of the ideal set of hyper parameters. The code for the loops can be found in the attached code in the appendix section. The results were found to be as follows

i. Classification - random_state=0, max_features=70, max_depth =10, min_samples_split=79, min_samples_leaf=2, bootstrap=0, n_estimators=100

ii. Regression - max_features = 19,max_depth = 6, min_samples_split = 40, min_samples_leaf = 6, bootstrap = 1,n_estimators = 100

Phase V – Model Inference

The result of the model on the Kickstarter dataset were as follows:

i. Classification Model – The average accuracy score was 73.8%

ii. Regression Model – The MSE was around 12Bn.

This implies that when the model tries to classify a project it will, on average get 73% of the projects right. Another interesting metric to measure is the confusion matrix.

N = 4706	Predicted No	Predicted Yes
Actual No	2611	451
Actual Yes	820	824

Therefore, we can infer that the model is actually good at identifying unsuccessful projects.

An MSE of 12Bn implies that on average we will predict the amount pledged to a project within a boundary of 109K usd of the actual ground truth of the project. Therefore, on average we are within a 110K range(above or below) the actual amount the project raised. The performance of this model is subjective, because if every project aims at raising millions then our error is acceptable.

Conclusion

Overall the models developed are robust and do not tend to over fit the data. Higher diversity in the data would help increase the accuracy of the data and the overall performance of the model.