Karan Goda | 460496371 3/10/18

INFO3406 Stage 2 Report (Individual Project)

Section 1: Setup

The dataset derived is a reduced dimension and cleaned Kickstarter dataset from the analysis and data manipulation that was conducted in the first stage of this individual project. The models for the research questions laid out below will be carried out in this stage of the project. Multiple models for each research question have been created in order to find the best scores for the data.

Research Questions

1. Do certain categories from certain countries of projects have a higher success funding rate?
   H0: Baseline values are better than the classifiers.
   H1: There is at least one classifier with better results than baseline, this will be judged using McNemar's P test.
2. Are number of backers a good indicator to predict the USD pledged.
   H0: There is a no correlation between backers and USD pledged.
   H1: There is a positive correlation between backers and USD pledged.
3. Are we able to use the project's features such as country or category to predict if the project reaches funding or not (This is a classification problem).
   H0: Baseline values are better than the classifiers.
   H1: There is at least one classifier with better results than baseline, this will be judged using McNemar's P test.

We will use machine learning modelling in Jupyter to create prediction models for our research questions so that it can be used in helping us find ways of improving how to predict potential successful Kickstarter campaigns before they are launched, **however only our third research question's models will be evaluated in this report.** This way focus is given, and the report is effective for the user.

We must be certain that our dataset is reliable in order to create statistical models in Jupyter so that we are able to obtain answers for our research question, hence we will use confidence interval on the quantitative columns of the dataset and significance testing to test the hypothesis of our research question.

A confidence interval is a range of values above and below a point estimate that enables us to capture the true population parameter at a fixed confidence level. For example, if we want the chance to obtain 95% of the true population parameter with a point estimate and corresponding confidence interval, our confidence level would be set to 95%. Higher confidence level means a wider confidence interval. We can check confidence interval of the total USD pledged per project, total USD goal per project, and number of backers of the dataset which will help us to have 95% of the encompassing values inside the range in the dataset. Significance testing challenges a hypothesis (H0) in order to determine if the alternative (H1) hypothesis is more acceptable. We can reject H0 if p value is less than 0.05.

We can measure effectiveness of the models of the research question by obtaining the **precision, recall, and F-measure** scores of the different models that we create using the columns from our Kickstarter dataset. We will then find which model(s) are effective in each of these categories by comparing the values derived from these models.
The formulae for these measures of effectiveness are:

Precision = True Positives / (True Positives + False Positives)
Recall = True Positives / (True Positives + False Negatives)
F-measure = (2 * True Positives) / ((2 * True Positives) + False Positives + False Negatives)

Section 2: Approach

We have used an extensive variety of models for the modelling of our data. We shall be using supervised methods since the data is labelled and their datatypes are known.

The proposed model and benchmark models we will be using are listed below.

Model (Learning techniques)	Assumptions (Features of the model)
Logistic Regression: Logistic regression is a model for regression used in categorical prediction of the dependent variable based on the independent values. Our dependent variable falls under the Bernoulli distribution where the result is either success or failure. The regression equation fitted for the training data can be used to classify the test data.	Homoscedasticity is not required. No multicollinearity among the independent variables. Linearity of independent variables and log odds Large sample size Error terms (residuals) do not need to be normally distributed
Linear SVC (Support Vector Clustering): SVC is a model that can deal with labelled and unlabelled data. An SVC training algorithm builds a model that assigns new examples to one category or another, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVCs can efficiently perform a non-linear classification using what is called the kernel trick.	Linear relationship between dependent variable and independent variables Multivariate Normality No Multicollinearity Homoscedasticity
Decision Tree: A decision tree is a model that created a structure with a root, leaves, and nodes. The leaf nodes are a final prediction for a particular observation which is the goal of a decision tree. A decision tree can be constructed in multiple ways because of different ways it can be split. Gini impurity can help us optimise our model.	Data is describable by features Class label is predicted by using set of decisions summarized by decision tree.
Random Forest: Random forest is an extension of decision tree which is another method of creating a model. Multiple decision trees are created through the process of taking a bootstrap sample from the training set. Each decision is forced to select a random subspace. This allows for better performance predictions which doesn't overfit as easily as the decision tree model.	Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. The tree is built using bootstrap sample than whole dataset. Aggregate tree outputs by averaging.
KNN (K Nearest Neighbours): KNN does not make any assumptions on the underlying data distribution. KNN is a lazy algorithm that does not use the training set to do any generalization for modelling. Most if not all the testing is conducted during the testing phase. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).	Non-parametric lazy learning algorithm. Data is in a feature space. Training data consists of a set of vectors with class labels We are also given a single number "k"

Model Construction

All the models contain hyperparameters that we are able to tune:

Model	Hyperparameters
Logistic Regression	C value
Linear SVC	C value, kernel, and gamma.
Decision Trees	Maximum depth, maximum features, minimum sample to split an internal node, and criteria used.
Random Forest	Maximum depth, maximum features, minimum sample to split an internal node, minimum sample required to be a leaf node, and number of estimators.
KNN	N neighbours (K value), and metric

Parameter Selection

In the parameter selection for the different machine learning models that we have created, we let the Python Jupyter functions decide on the default values because we will be trusting the methods to create the best models for us. However, the only hyperparameter that will be revised and decided upon is the number of estimators in the random forest method. We have set this to be 100 because this parameter change allows for maximum effectiveness and least overfitting. However, in the future, we should try to combine models to further improve accuracy and avoid overfitting if it ever occurs in any of the models.

Section 3: Results

The p value resulting from the significance test for all the models is lower than 0.05. This suggests that all the models outperform the baseline with Linear SVC having the lowest p value.

F1 Score	Hypothesis testing vs Baseline	Mean Accuracy Score	Model	Precision	Recall
0.579020	8.150972e-03	**0.632787**	Random Forest	0.593102	**0.632787**
0.533707	4.550026e-02	0.586885	Decision Tree	0.539872	0.586885
0.530620	3.114910e-04	0.603279	KNN	0.548402	0.603279
0.395869	9.799074e-26	0.632787	Logistic Regression	**0.815789**	0.632787
0.299676	**3.552964e-33**	0.360656	Linear SVC	0.394696	0.360656

From the individual models, we can see that **random forest has the highest accuracy score** (Recall) out of all the models. This suggests that random forest is the superior model in predicting which features of a project will fulfil funding. Since each feature is weighed equally, we can rely on the accuracy of this model. Note: Accuracy/Recall in machine learning is the proportion of actual positives identified correctly.

Another value that the random forest model is superior compared to the other models, is the F1 score. An F1 score is a measure of a test's accuracy, however one thing to note is that this score does not take true negatives into account. The best value is close to 1 while the worst is closer to 0. We had a sufficiently high F1 Score, so we are satisfied with the result.

Random forest interestingly did not have the highest precision, logistic regression takes the prize here. Precision is the proportion of true positives obtained correctly over total positives.

Overall, we believe all models had high/above average results except Linear SVC in predicting which Kickstarter projects will succeed based on their features. We will call these models as the "successful prediction models". The accuracy of the successful prediction models was above 0.5 which indicates that the models are correct more than half the time. The precision of these models was also above 0.5 which means that we had a reasonable classification threshold, and the models had a good proportion of positive identifications.

The dataset had no limitations, it was of reasonable size, the data was complete, there were no missing values, and no sections in the data were missing. No anomalies in the models were detected, so we know that our dataset is of sufficient quality and size. We observed limitations in the setup because we trusted the default values in majority of the models, however with further trial and error, we could have obtained better F1 scores, Precision and accuracy values had we decided to customize and tune the hyperparameters. This project is able be improved further by using the best machine learning models here in unification with each other to create a new custom model that specializes in Kickstart project success for research question three.

Section 4: Conclusion

Evaluating the process in our initial data mining project and the modelling stage allows us to appreciate the completeness and cleanness of the data as a lot of time consuming processes were not necessary in this CRISP-DM project.

In conclusion, I learned how important the process of gathering data or obtaining valid data is because without reliable and useful data, we are not able to generate valid models let alone conduct any data analysis to interpret what the data stands for, and if data is not interpreted and useful results are not generated, a great risk is posed to organizations because they are unable to improve their business processes. These organizations are then incapacitated, unable to take customer feedback into account.

So I talked about how this study improved my understanding of why we want data analytics, now I want to mention how greatly this project helped me gain the knowledge and skill to select reliable datasets, interpret the information that these datasets possess, and create my own machine learning models to predict future outcomes just from data. This truly is one of the greatest skills an IT engineer can possess and I am grateful for everyone involved in the creation of this study.

I would recommend my random forests model as a solution to the problem of Kickstarter project success prediction, because my project takes words of the project title into account. Usually this title is what attracts users in determining if they want to invest in the project and our ML model is following the same pattern in predicting the project's success. Further testing and tuning of the hyperparameters of the random forest model can be done in the future to improve the prediction capabilities of the model.