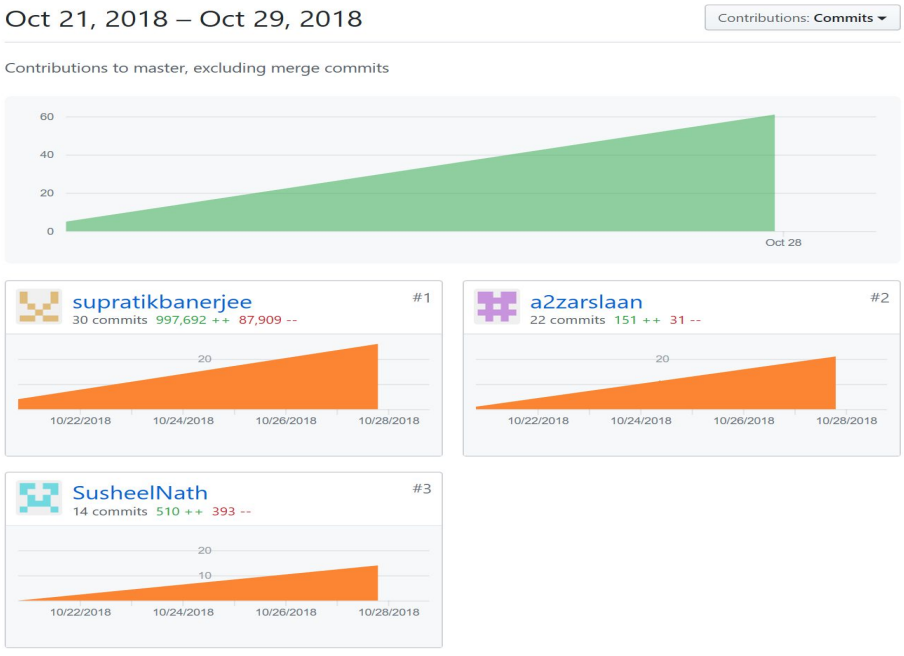
ML1819 Research Assignment 1

Team ID: 31
Task ID & Title: 102 – Dataset Pruning
Student Names: Supratik Banerjee, Susheel Nath, Amar Zia Arslaan
Student IDs: 18309796, 18300655, 18309976

Explanation for Individual Contribution:

Susheel Nath: Logistic Regression, Bagging Classifier and AdaBoost Classifier
Amar Zia Arslaan: Random Forest, Support Vector Cosine, MLP Classifier
Supratik Banerjee: Data Analysis, Preprocessor, Decision Tree Classifier and KNN

Word Count: 965
Source Code: Code
Source Code Activity: Activity
Screenshot Activity:



Classifier Reliability using Statistical Data Pruning

Supratik Banerjee
Team 31
sbanerje@tcd.ie

Susheel Nath
Team 31
naths@tcd.ie

Amar Zia Arslaan
Team 31
arslaana@tcd.ie

1 Introduction

For the past few decades, understanding the importance of data has played a key role; how it could potentially benefit an outcome when certain direction has been provided to it. Too much data risks overfitting and a data too small might not capture important information. Data pruning defines, elimination of certain unwanted data to observe an improvement in learning performance. In this paper, we use various classification algorithms, to observe the effect of data pruning on the accuracy of these classifiers. Certain statistical measures such as Z-Score and Grubbs' Test are used to identify outliers to validate our hypothesis and reduce the anomaly in the dataset. This reduces the complexity of the given dataset resulting in an increase in accuracy.

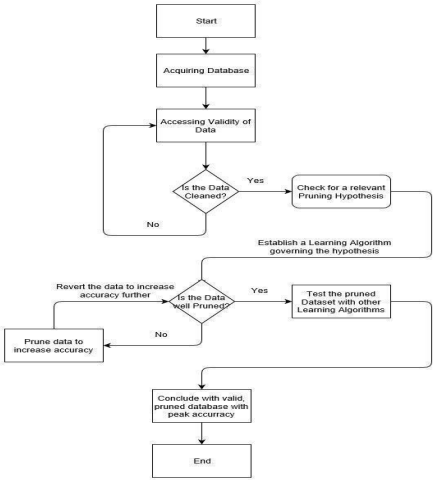


Figure 1: Flow Chart explaining the Data Pruning process.

2 Related work

Research by Anelia Angelova, DataPruning[1], dictates - "*The task is to identify if there are examples in the training set such that by eliminating them one may improve generalization performance*", which can be interpreted as, identifying outliers to prune the dataset. Removal of these "*Difficult Example*s", which are "*those which obstruct the learning process or mislead the learning algorithm*" as stated in the work would likely increase the accuracy of our models. To identify these outliers, we use the Grubbs' Test[2] to prune our dataset.

On analyzing Kaggle kernel[4] we find that amongst numerous learning algorithms used, LGBMClassifier results the highest accuracy of 68.12%, with pruning done in line In [21] and following with another Kaggle kernel[3], resulting in 68.46% accuracy using the LGBMClassifer.

3 METHOD

To solve our research problem, we use a dataset from Kaggle.com which provides information on various Kickstarter Projects[5] in '.csv' format containing 378661 observations and 15 variables. With the given data we are trying to predict the **success** or **failure** of a project. On analyzing the data, we explore the *'state'* variable, which describes the current state of a project. In Figure 2, the break up shows a failure rate of 52.2% and success rate of 35.4%. In Figure 3, we observe the distribution of the variable, *goal* using a linear plot, which yields an imprecise change. A logarithmic plot gives a clearer picture of the distribution, where approximately at 10^6 there is a sudden spike, which is also the variable's standard deviation.

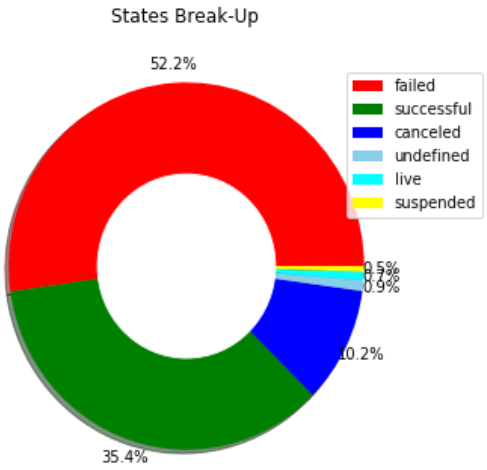


Figure 2: Break up chart explaining the various values of the variable '*State*'

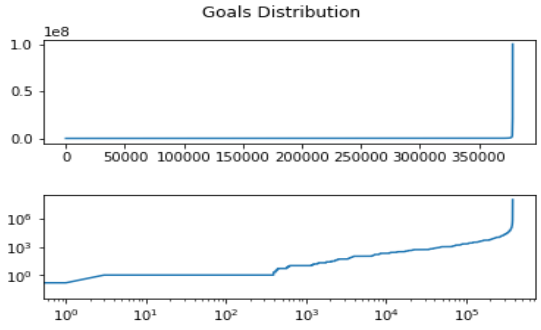


Figure 3: Distribution of the variable *'goal'*

Based on this observation, we further clean the *goal* variable, such that we only observe the *goal* values that are above the standard deviation and compare it with all the failed projects with goals above standard deviation. Figure 4 gives us a picture of goals above standard deviation which is just about 0.263% of the entire data out of which 98.62% of the projects have failed to achieve their goal.

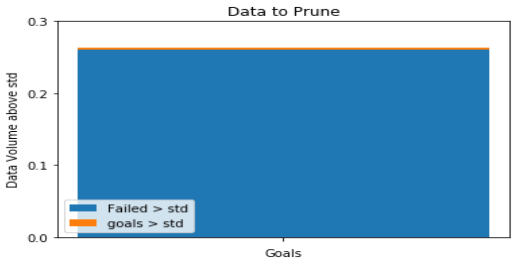


Figure 4: Logarithmic scale, failed to goals above standard deviation

Even though the number of failed projects above standard deviation is 0.259%, we can form our **hypothesis** out of this, stating that **pruning all projects, that have failed above standard deviation would yield a suboptimal dataset**. To verify this hypothesis, we run 2 tests wherein we look at the Z-Score of *goal* and then check whether it's above standard deviation. We get a result of 870 *goal* values above Z-Score 1. Followed by that, we perform Two-Sided Grubbs' Test, which yields the same result of 870 outliers. With this verification, we write two datasets, one based on our hypothesis, which yields an elimination of 862 *goal* values and second, where we eliminate all values detected as outliers by the Grubbs' Test. We then encode all unique string values in other variables numerically.

After pruning the data, we use eight classification algorithms. Most of the parameters are either set to default or are used as observed in LGBM 0.681[4].

4 RESULT

To answer our research question, we have trained Support-Vector-Cosine *SVC*, Multi-Layer-Perceptron *MLP*, Logistic-Regression *LGR*, K-Nearest-Neighbours *KNN*, Random-Forest *RF*, Decision-Tree *DT*, AdaBoost *AB* and Bagging-Classifier *BC*.

Pruning is performed on the data using our hypothesis mentioned in the previous section and Grubbs' Test, we obtain a statistically outlier-free dataset, using which we can train our models. Training is done by splitting the pruned dataset with a train size of 90% (test_size=0.1) as seen in Line In [49][4] and then followed by using 10% of the data from the unpruned dataset to test the models.

On comparing the graphs representing the sklearn accuracy_score as shown in Figure 5 - Unpruned, Hypothesis-Pruned and Grubbs-Test-Pruned respectively, we see that *SVC* underperforms in our hypothesis with 51.35% accuracy and in Grubbs-Test with 49.86% but does comparatively well in the unpruned dataset with 55.59% accuracy. An increment in accuracy is noticed in KNN algorithm in both the pruned datasets, 70.78% in the Hypothesis-Pruning and 70.92% in Grubbs-Test-Pruning and a significant rise is noticed in the accuracy of the pruned datasets in Decision-Tree Classifier with 80.99% accuracy in Hypothesis-Pruning and 81.26% in Grubbs-Test-Pruning as well as Bagging Classifier, with 78.24% accuracy in Hypothesis-Pruning and 78.25% in Grubbs-Test-Pruning as compared to 64.48% in the Unpruned Dataset. Other Machine Learning algorithms show no distinct change.

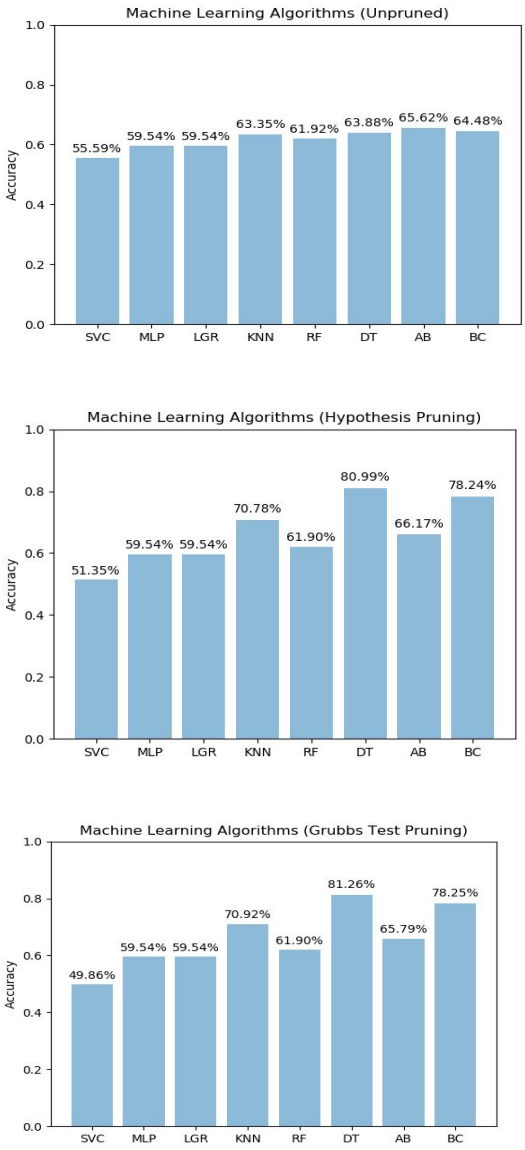


Figure 5: Comparison of accuracies of Classifiers in, a) Unpruned Dataset, b) Hypothesis Dataset and c) Grubbs' Test Dataset.

On removal of the outliers based on the variable *goal*, using Grubbs' Test and our hypothesis, a considerable increase in accuracy is noticed in most of the algorithms. Due to the noticeable increase in accuracy, we can conclude that our hypothesis to prune the dataset 'Kickstarter', holds true.

5 Limitations and Outcome

An outlier is a very subjective term, as it depends on the analyst to determine what the anomaly is in a dataset. Likewise, in the process of Data Pruning, it is very difficult to determine which is a difficult example which needs to be removed to enhance the performance of a learning algorithm. It will always depend on some assumption making it susceptible to an erroneous elimination.

We can further look into newer approaches of detecting outliers. This work has future scope in experimenting with learning algorithm parameters for better results. A concrete next step would also be to implement LGBMClassifier as implemented in the Kaggle kernels[3].

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin.

REFERENCES

[1] Angelova, Anelia. "Data Pruning." CaltechTHESIS, 2004, thesis.library.caltech.edu.

[2] Grubbs, Frank E. "Sample Criteria for Testing Outlying Observations." The Annals of Mathematical Statistics, Institute of Mathematical Statistics, 1950, projecteuclid.org/download/pdf_1/euclid.aoms/1177729885.

[3] Samuel, Dave. Kickstarter Success Classifier [0.685] | Kaggle, 2018, www.kaggle.com/majickdave/kickstarter-success-classifier-0-685.

[4] Kosovan, Oleksandr. "Kickstarter | LGBMClassifier [0.681]." Kickstarter | LGBMClassifier [0.681] | Kaggle, 2018, www.kaggle.com/kosovanolexandr/kickstarter-lgbmclassifier-0-681.

[5] Mouillé, Mickaël. "Kickstarter Projects." Kaggle: Your Home for Data Science, 8 Feb. 2018, www.kaggle.com/kemical/kickstarter-projects.

[6] LLC, NCSS. "Grubbs' Outlier Test ." Www.ncss.com, www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Grubbs_Outlier_Test.pdf.