Classification of Debt Vulnerability in Sub-Saharan African Countries Using Various Machine Learning Tree Based Algorithms

Danielle Shackley, Brendan Dao, Salem Othman*

Wentworth Institution of Technology 550 Huntington Ave Boston, Massachusetts 02115

*Corresponding author's email: othmans1@wit.edu

doi: https://doi.org/10.21467/proceedings.157.8

ABSTRACT

In this work, we compared the accuracy results of a classification problem with three different models i.e. Decision Tree, Random Forest and Gradient Boosted Tree. We took a small dataset from the Kaggle repository containing four hundred and thirty-five samples. We examined each model's choice of feature importance as well as their test and training accuracies. We found that the Gradient Boosted Tree produced the highest testing accuracy of 84%. Random forest was the second best accuracy of 83% and Decision Tree had the lowest with 82%. In addition to the accuracy, each model has a confusion matrix of the output of the testing data. Gradient Boosted Tree has the best true negative rate of 77.7% while Decision Tree has the worst true negative rate of 55.5%.

Keywords: Debt Vulnerability, Sub-Saharan Africa, Machine Learning

1 Introduction

The Heavily Indebted Poor Countries (HIPC) Initiative was created in 1996 with the intent to avoid poor countries from facing a debt burden they cannot manage. The International Monetary Fund (IMF) and Word Bank are the leaders in this initiative. In 2005, the HPIC was reinforced with the Multi- lateral Debt Relief Initiative (MDRI). This change allowed for another institution, African Development Fund (ADF) to assist with the debt relief of countries. Countries must meet a certain criteria, "commit to poverty reduction through policy changes, and demonstrate a good track record over time" in order to be considered for the relief package (Hakura 2020a). The debt of a country is considered sustainable as long as, "the government is able to meet all its current and future payment obligations without exceptional financial assistance or going into default" (Hakura 2020b). Analysts look at a number of factors that contribute to a country's debt risk. In this work we look at a data set that contains countries that received debt relief under the HIPC in order to classify them as still in debt risk or not. We used debt distress vulnerability data of Sub-Saharan African countries to classify whether a country would be classified as debt vulnerable. Using feature importance analysis and parameter tuning, we built different machine learning tree models to compare accuracy and performance to find an optimal model.

2 Research Questions

We aim to answer these two research questions:

- Can machine learning models predict the debt risk of a country, given a set of financial aspects spanning multiple years of that country?
- Using three tree models (decision tree, random forest and gradient boosting), which model performs the best on a classification problem?

Proceedings DOI: 10.21467/proceedings.157; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-970666-0-3

3 Methods

3.1 Dataset

The “Evolution of debt vulnerabilities in Africa” dataset is supplied by Kaggle’s data repository and used for classification tasks. It is a small dataset of four hundred thirty-five samples, making it an appropriate choice for Decision Tree models. It contains nineteen features that contribute to the classification of a country’s debt indication. The countries in this dataset have been granted debt relief from the Heavily Indebted Poor Countries (HIPC) initiative. This initiative granted cancellations of external debt owed to the World Bank, International Monetary Fund (IMF) and African Development Bank. Starting in 2005, the countries that received this relief package have been monitored for their debt vulnerabilities by the IMF and World Bank. The features include; ISO, year, inflation, Curr.acc.balance, Gen.gov,len.bor, Vol. Exp. Goods, GDP, GDP.per.cap, Gen.gov.revt, US.int.rates, Ext.Debt.Serv, Real.GDP.growth, Exch.Rate, Control of Corruption, Government Effectiveness, Pol. Stability Absence.of.Violence, Regulatory.Quality, Rule.of.Law, Voice.and.Accountability, Risk.ext.debt.distress. The features are a mix of categorical and integer values (Richter 2021).

3.2 Data Pre-Processing

Decision trees do not perform optimally on sparse or text data. Therefore, we converted the categorical feature values into numerical values with the Label Encoder method. This method applies to the dependent of the data. Its function is to “Encode target labels with values between 0 and n classes-1.” Ordinal encoder is the function used to convert the independent values (features) into integers (H 2020). This is used only on the independent variables because it returns a single column of integers per feature. Both of these methods automatically detect the range of values it needs to convert. For example, the “ISO” feature was con- verted into 29 values (0, 1, 2, . . .) because it has twenty nine unique options, the country’s abbreviation (Pedregosa *et al.* 2011). We tuned select hyperparameters within each model to provide the best accuracy. We also removed the column “Risk.ext.debt.distress” which had four values: low, medium, high, in-debt distress. After using this column in our model we found that the trees relied too heavily on this feature and to classify the debt indicator. It became the most important feature in each model by a large, as seen in Figure I. Though this feature helped with the accuracy, we wanted to explore the use of more features and see how the model performed. We explain the important features found in different models later in this report.

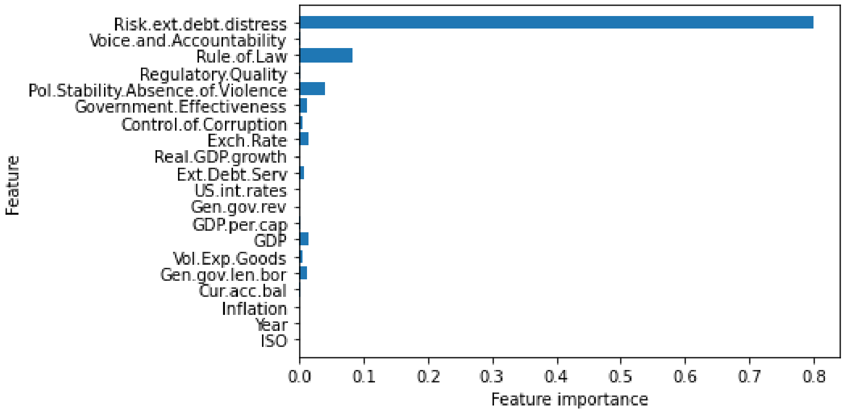

Figure 1: *Output of Decision Tree Feature Importance with Risk.ext.debt.distress Feature*

3.3 Related Work

A previous study, conducted by (Ali *et al.* 2012) compared Decision Trees and Random Forests using twenty datasets from the same data repository (Richter 2021). Their results found that Random Forests outperform

Decision Trees when the number of attributes are the same and the dataset is large. We follow a similar procedure with a large dataset and both models using the same nineteen features. We found similar results as this study in terms of classification performance.

3.4 Models

We used three different machine learning tree models to classify the debt indicator of low income African countries.

3.5 Decision Tree

The Decision Tree is a supervised learning algorithm, meaning we give our model labeled data to train with (Hoare 2020). This type of model is used in classification and regression problems. We state the goal, method and feature importance found of each of our models in the following sub-sections.

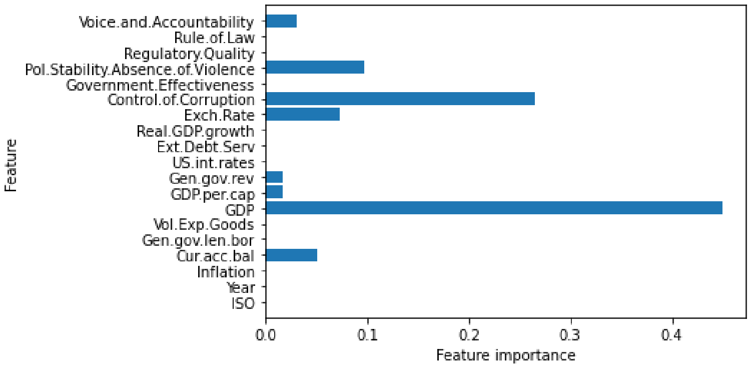

Figure 2: *Output of Decision Tree Feature Importance*

- Goal: To create a training model to be used for predicting class or value of the target by learning rules from previous data.
- Method: The design of the model can be described as "predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node" (Chauhan 2020).
- Feature Importance: Decision Trees can become overfitted if the max depth of the tree is too large. For our data we chose a depth of four. As seen above in Figure 2, the GDP feature is the most important. In terms of Decision Tree models, the most important feature is the one with the purest split.

3.6 Random Forest

Random forest tree models are also supervised learning algorithms. Random forest trees are multiple Decision Trees with split data which would be combined back into one main tree (Ali *et al.* 2012).

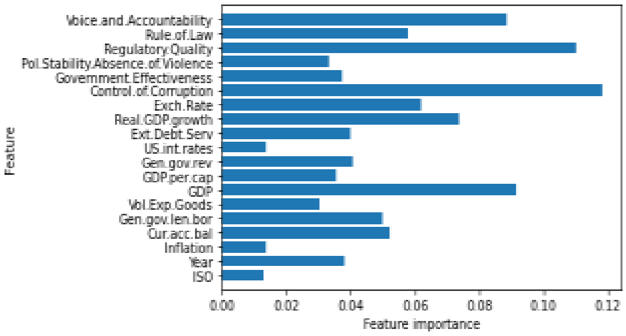

Figure 3: *Output of Random Forest feature importance*

- Goal: Random forest tree models built off of Decision Trees to create a collection of Decision Trees(for Geeks 2020).
- Method: To reduce overfitting of individual trees by building many different Decision Trees that do an acceptable job of predicting target class. Next it uses bootstrap sampling to assure randomness of trees. (Chauhan 2020)
- Feature Importance: The number of features that are ranked important increased as well as the features seen as most important changed from a Decision Tree.

3.7 Gradient Boosted Tree

The last model we examined is another supervised learning algorithm used for classification and regression problems (Maklin 2019).

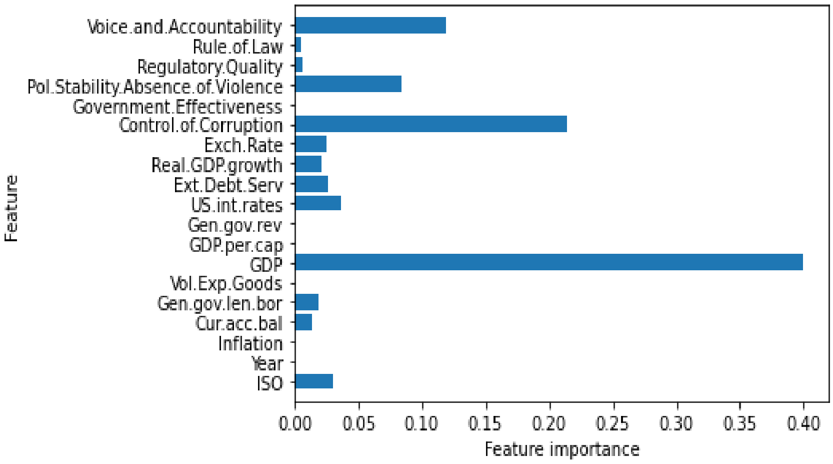

Figure 4: *Output of Gradient Boosted Tree feature importance*

- Goal: To combine multiple Decision Trees to create a more powerful model.
- Method: This method builds trees in a serial manner; each proceeding tree tries to correct the previous one's mistakes. There is no randomization default, instead it uses strong pre-pruning. Pre-pruning is stopping the tree before it has completed classifying the training set. Similar to Random Forests, it combines many simple models (known as shallow or weak trees).
- Feature Importance: The number of features that are ranked important decreased from the previous model. Similar to the Decision Tree model, GDP is found to be the most important.

There are three parameters that can be adjusted for the Gradient Boosted tree models. We built three different trees with different parameters, listed below. The results of this table show that the optimal parameters for this model are a max depth of 3, learning rate of 0.1 and the number of trees equaling 100.

Table I: *Parameter Adjustments during Gradient Boosted Tree Model*

Parameter Adjustments			
No. of Trees	**Max Depth**	**Learning Rate**	**Accuracy**
100	3	0.1	84.4%
100	1	0.1	81.7%
100	3	0.01	81.7%

4 Results

The following tables are data collected from the study on the tree models. We documented the top three features found from each model as well as accuracies from the training and test datasets.

As seen in Table II. there are two features that are in the top three for all three models, GDP and Control of Corruption. GDP is the total value of goods and services a country produces. Control of Corruption allows the model to take into account how corrupt the government is. The higher GDP means the country's economy is producing goods and services. Having a higher control of corruption allows for people to save their money without unfairly taxes to the government.

Table II: *Top three features for each model.*

Feature Rankings			
Feature Rank	Model		
	Decision	Random Forest	Gradient Boosted
1	GDP	Control of Corruption	GDP
2	Control of Corruption Political	Regulatory Quality GDP	Control of Corruption Voice
3	Stability Absence of Violence		and Accountability

Table III: *Accuracies of each model on testing and training set.*

Accuracy	Model		
	Decision	Random Forest	Gradient Boosted
Training	92.9%	99.1%	100.0%
Testing	81.7%	82.6%	84.4%

As seen in Table III, the model with the highest testing accuracy is the Gradient Boosted tree. This model was built to learn from the training set first and then from the mistakes of the preceding tree, giving it a "boosted" accuracy, which we can see from our results.

Table IV: *Decision Tree Confusion Matrix*

Decision Tree Confusion Matrix		
	No Debt	In Debt
No Debt	69	4
In Debt	16	20

Table V: *Random Forest Tree Confusion Matrix*

Random Forest Tree Confusion Matrix		
	No Debt	In Debt
No Debt	72	1
In Debt	11	25

Table VI: *Gradient Boosted Tree Confusion Matrix*

Gradient Boosted Tree Confusion Matrix		
	No Debt	In Debt
No Debt	64	8
In Debt	9	28

In addition to Table III. displaying the accuracy of each model, Tables IV-VI. contains the confusion matrix for each model. As shown in every confusion matrix, every model has a high likelihood of predicting correctly on concerning the no debt, however, we focused more on the improvement of each model on correctly predicting the people in debt.

To solve for how well each model predicts the people in debt, we would use the true negative rate. The true negative rate is a ratio of how many did the model correctly predict the negative over the total amount

of actual negatives. The Decision Tree's true negative rate is 55.5%. Although Random Forest improved to a 69.5% true negative rate, but Gradient Boosted Tree's true negative rate is the highest at 77.7%.
We ranked how well the model performed by comparing how well they were accurately predicted the people in debt because not knowing that they are in debt is more important than knowing that the people are not in debt. Although Random Forest performed better than Gradient Boosted Tree in identifying the people not in debt, it is worth less because Gradient Boosted identified the people in debt better than the Random Forest.

5 Conclusions

This project started with data selection. We chose a classification dataset that was compatible with our machine learning models we wanted to focus on. Data pre processing was performed to ensure the data was usable for our model. We built three different machine leaning tree models. Decision Tree, Random Forest and Gradient Boosted. An analysis on their accuracies and feature selection concluded with the Gradient Boosted tree having the highest accuracy at classifying the test set. After completing this project we compiled a few lessons learned in three different categories:

- Overfitting
  - Decision trees can overfit training data
  - Lowering the max depth of the tree can remediate this
- Learning Rate
  - A small learning rate is better for training
  - Lowering the learning rate can help reduce overfitting
- Performance
  - Gradient boosted trees produce shallow trees, which results in smaller models in terms of memory and speed
  - max features or max leaf nodes can drastically reduce space and time requirements for training and prediction

6 Future Work

This study contained an in depth analysis into three different machine learning tree models to compare accuracies. A possible expansion of this work includes adding another model to the study called extreme gradient boosting. Another direction we discussed was using a different dataset and running a regression model instead of classification and comparing results.

7 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in institutional affiliations.

How to Cite

Shackley *et al.* (2024). Classification of Debt Vulnerability in Sub-Saharan African Countries Using Various Machine Learning Tree Based Algorithms. *AIJR Proceedings*, 53-59. https://doi.org/10.21467/proceedings.157.8

References

Ali, J.; Khan, R.; Ahmad, N.; and Maqsood, I. 2012. Random Forests and Decision Trees. Technical report, International Journal of Computer Science.

Chauhan, N. 2020. Decision Tree Algorithm, Explained. for Geeks, G. 2020. Random Forest Regression in Python. H, B. 2020. How to Export Pandas DataFrame to CSV.

Hakura, D. 2020a. Debt Relief Under the Heavily Indebted Poor Countries (HIPC) Initiative. *International Monetary Fund*.

Hakura, D. 2020b. What Is Debt Sustainability? *International Monetary Fund*.

Hoare, J. 2020. Machine Learning: Pruning Decision Trees.*DISPLAYR*.

Maklin, C. 2019. Gradient Boosting Decision Tree Algorithm Explained.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit- learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Richter, E. 2021. Evolution of debt vulnerabilities in Africa. Technical report, Kaggle.

Stack, O. 2020. How do ’numpy.ndarray’ object do not ’numpy.ndarray’ object?