UNIVERSITY of RWANDA

AFRICAN CENTER OF EXCELLENCE
IN DATA SCIENCE

ACE-DS

A machine learning approach towards micro, small and medium business profitability growth prediction.

By Student Name: Eva Mpagi

Registration Number: 219014051

A dissertation submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN DATA SCIENCE (DATA MINING)

In the College of Business and Economics

University of Rwanda

Supervisor Name: Dr. Niyizamwiyitira Christine

Month and Year of submission: September 2020

Declaration

I declare that this dissertation entitled A machine learning approach towards micro, small and medium business profitability growth prediction is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Names: Eva Mpagi

Signature:

Approval sheet

This dissertation entitled A machine learning approach towards micro, small and medium business profitability growth prediction written and submitted by Eva Mpagi in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Data Mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 18 % which is less than 20% accepted by the African Centre of Excellence in Data Science (ACE-DS).

Supervisor: Dr. Niyizamwiyitira Christine

Head of Training

Acknowledgements

This dissertation is the result of my work at the African Centre of Excellence in Data Science at the University of Rwanda from 2018 to 2020, where I was part of the first cohort of the Masters in Data Science programme. During this time, I received training and support from tutors, the school administration and my sponsors at the Inter University Council of East Africa.

I would therefore like to express my deepest gratitude, first and foremost to my supervisor Dr. Niyizamwiyitira Christine. Her patience and personal guidance helped me grow as a researcher and provided a suitable environment that allowed me to develop my own ideas and freely express myself. Further, I would like to thank the center administration for all their organizational support and persistent engagement particularly as we were conducting this research during the Covid pandemic.

My special gratitude goes to the Inter University Council of East Africa, who supported this research financially and funded my studies during my two years at the University of Rwanda. Finally, I want to express my deepest gratitude towards my family especially my mother, Dr. Jeniffer Mugisha, and my siblings, William and Christine. Thank you for always being there for me and for all your unconditional love.

Abstract

The importance of Micro, Small and Medium Enterprises for East Africa's economic endurance and innovation has been acknowledged by policy makers, credit institutions and business owners. The highly uncertain and volatile nature of the economic environment in which they operate is constantly affected by external factors such as politics, pandemics, and many other factors, hence making the process of analysing information to evaluate whether a Micro, Small and Medium Enterprise will be profitable a challenging task. This prediction problem shows that there is need for a model, which facilitates an unbiased approach to Micro, Small and Medium Enterprises profitability prediction. In this dissertation, the objective is to understand how the application of machine learning on survey data can be operationalized to study Micro, Small and Medium Enterprises profit growth prediction and develop a reproducible model via web-based deployment.

Predicting Micro, Small and Medium Enterprise profitability growth models currently almost exclusively relies on data collected by credit institutions based on their relationship with their customers. Additionally, data is collected by conducting physical business site visits. This data is used to assess the business. It is therefore very difficult to apply this method on a wide scale in a repeatable and automated way for growth prediction. This may also be a hindrance for Micro, Small and Medium Enterprises that are usually startups meaning they lack substantial credit history. In this dissertation, data from the national small business survey in Uganda is used and machine learning methods are applied as they are known to provide a better balance of speed and accuracy in the decision-making process than previously used methods.

Overall, three different models are applied to predict profit growth, which are random forest, extreme gradient boosting, and logistic regression. These models will allow us to compare performance of traditional models versus advanced models. Goodness-of-fit tests are applied to the models, and the best ones are extreme gradient boosting and random forest which are ensemble methods, with accuracies of 92.31% and 92.02%. The most relevant variables in the best performing models are 'sales made last year', 'operation time' and 'business owner education level'. Models generated in this dissertation can be used to predict Micro, Small and Medium Enterprise profit growth rate in a repeatable way, using annually available survey data.

Key words: Micro, Small and Medium Enterprises, Profit, Boosting, Bagging, Classification.

Table of Contents

List of figures

List of tables

List of Abbreviations

MSME	Micro, Small and Medium Enterprises
GEM	Global Entrepreneurship Mentor
UGX	Uganda Shillings
GDP	Gross Domestic Product
XGB	Extreme Gradient Boost
COBE	Census of Business Establishments
CAPI	Computer Assisted Personal Interviewing
SACCO	Savings and Credit Cooperative Organisations.
ANOVA	Analysis of Variance
ROC	Receive Operating Characteristic
AUC	Area under Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
WEKA	Waikato Environment for Knowledge Analysis

1 Introduction

This chapter provides the general motivation, objectives, and background of this dissertation.

1.1 Motivation

Micro, Small and Medium Enterprises (MSMEs) play an important role in the advancement of the Ugandan economy. In Uganda, an East African country, MSMEs make up about 90% of the private sector and employ more than 2.5 million people [1]. MSMEs in Uganda are comparatively young initiatives with many of them aged between one and ten years old. According to GEM Uganda, while 10 percent of Ugandans opened a business in 2016, a fifth of them closed those businesses the following year [2]. These statistics are disquieting and should be a key concern for policy makers. While there are multiple factors contributing to this situation for example lack of information, poor management, and inadequate investments in terms of skills and innovation, the key causatives to the failure of most MSMEs are financial. They center on and around inadequate access to and the cost of funding. It is clear from the 2015 national small and medium business survey that a major challenge is that credit or financial institutions have rigorous requests around collateral or financial security which MSMEs are not able to satisfy. This is mostly because most MSME’s are startups and are seen as risky ventures with high mortality rates.

In the recent years, there have been increased efforts from various stakeholders to counter the challenges faced by MSMEs for example a symposium on "Modalities for Financing Small and Medium Scale Enterprises" took place in 2002 in Uganda and was attended by hundreds of participants from SMEs, banks, government and the donor community. Some of the principal objectives were to identify the main obstacles to MSMEs accessing finance from formal sources and to suggest hands-on solutions and inventive ways to resolve the identified issues. Some of the recommendations from this symposium were for the government, the banking community and donors to set up a special loan guarantee fund for MSMEs. In addition, they resolved to provide assistance to MSMEs in terms of the best practices required for an MSME to become a profitable business [3].

Regardless of all these efforts by various stakeholders, it is still a challenge for various MSMEs to access these services quickly and efficiently as they are still seen as risky ventures. There is therefore need for a competent way to determine whether businesses are likely to grow in terms of profits made and hence determine their credit worthiness and provide a quantified avenue for MSMEs to access funding.

All businesses fail in their own distinctive way. This means it is important to study and analyse as many failed companies as possible to learn and identify key factors that led to losses in the first place. In addition, the time aspect of business growth and the fundamental nonfinancial factors will be recognized. The authors of failure processes and causes of company bankruptcy: a typology, highlight such non-financial elements, which not only include the management team, but also the relationship with different stakeholders [4].

This research will therefore mostly utilise supervised machine learning data mining approaches to create a solution that will allow all stakeholders for example business owners, investors, venture capitalists etc. to determine the viability of a business and hence ease access to funding.

1.2 Research questions

There are several research gaps identified in the domain of data mining using survey data and profitability prediction modeling in East Africa. Many models that exist especially in the MSME domain are mostly credit scoring models in the banking or credit industry. Most of these models are institution specific and can only be built on data that a credit institution has collected based on their relationship with a specific MSME. In addition, most of the current models, rely on extensive physical business site visits to collect information. It is therefore very difficult to apply this method on a wide scale in a repeatable and automated way for growth prediction.

Although numerous studies and surveys on MSMEs exist, data mining and analysis solutions from this data are rare, yet annual surveys produce information which can provide insight into MSMEs growth.

In order to address these gaps, we investigate the use of publicly available survey data for MSMEs profit growth prediction. In particular, we aim at understanding how the annual data surveys can be used to systematically generate business-relevant knowledge. Thus, the first research question is stated as follows:

Research Question 1. How can applying machine learning on survey data be operationalised to study MSME profit growth prediction?

Profitability of MSMEs is a multifaceted mechanism which is characterized by many internal and external factors. Therefore, it is very important that profitability models are developed for specific industries. Moreover, the fundamental factors for growth differ depending on the type of business i.e., sector definition must be a component of the modeling process. In this case, input variables will include definitions of sector type.

Research Question 2. To which extent can we develop a web data-based profitability growth prediction model for MSME businesses?

This dissertation is intended to achieve a balance between concepts and practice. Thus, the proposed research approach will combine literature review with practical studies. To understand the underlying mechanism of MSMEs growth, we will survey the key determinants of growth by applying several techniques of machine learning. The best performing model will be deployed using Flask framework.

1.3 Dissertation Outline

The dissertation is structured as follows: The study begins with the motivation and research questions. The next Chapter provides further information about MSMEs and growth models in East Africa specifically in Uganda via literature review. It provides a detailed definition of MSMEs, the challenges they face and a brief description of available business success or growth models available in the industry. Chapter 3 then focuses on the methodology used to undertake this dissertation. This is followed by Chapter 4, in which the described methodologies are applied in order to address the above-mentioned research questions. Chapter 5 then offers description for the selection of the best model. Then a comprehensive discussion of the data analysis and results follows and Chapter 6 concludes with a general discussion of the key findings of this research.

2 Literature Review

2.1 Description of an MSME

The conditions for defining MSMEs vary from nation to nation. The European Commission's description of MSMEs is based on the annual turnover and number of employees. The choice of MSME definition could be contingent on numerous factors, such as level of worldwide economic incorporation, business culture, the size of the country's populace and industry [5].

In Uganda, MSMEs comprise of all types of businesses notwithstanding of their legal status (such as sole proprietorships, family businesses or cooperatives) or whether they are formal or informal businesses [6]. MSMEs can be defined using any two of the conditions i.e., number of employees, capital investment and annual sales turnover that is:

- A micro enterprise is a business employing up to four people, with an annual sales/revenue turnover or total assets not exceeding Uganda shillings 10 million.
- Small enterprises employ 5 to 49 and have assets totaling from Uganda shillings 10 million but not exceeding 100 million.
- The medium enterprise employs between 50 to 100 individuals with total assets amounting to more than 100 million but not above 360 million Ugandan shillings.

MSMEs are the backbone of progression for economic advancement in Uganda and the world at large. They encompass multiple sectors in Uganda with 33% in commerce and trade, 10% in manufacturing, 49% in service sector and 8% in others. More than 2.2 million individuals are engaged in this industry, and they account for roughly 89% of the private sector, producing over 82% of mass-produced outputs that contribute to over 18% of the gross domestic product (GDP) [7].

2.2 Major challenges facing MSMEs.

The interconnection between the access to funds and challenges faced by MSMEs cannot be underestimated. In the subsequent discussion, it is revealed to be integrally embedded in every challenge that MSMEs encounter. In this section we shall dissect challenges such as access to affordable finance, inadequate information and skills, limited access to appropriate technology and competition.

Restrictions to Affordable Funding

A key limitation for MSMEs is inadequate access to equitable funding required to satisfy their needs. This usually affects initiatives started primarily by vulnerable groups, women and youth as well as agriculture centered initiatives which are perceived to be risky. Interest tariffs charged are usually extremely high and many MSMEs cannot meet those expenses [6].

Inadequate information

Very few businesses can withstand premature mortality without effective information and communication management. In order to grow into a highly successful business, information administration should be channeled over social networking channels and other relevant information channels in which social capital is derived [8].

Inadequate technical skills

Most MSMEs are usually categorized as informal businesses. Unlike the businesses in the formal sector that are endowed with individuals who are trained, skilled and gifted with knowledge and practical skills, MSMEs experience a gap especially in terms of technical entrepreneurship and management skills [6].

Suitable Technology Limitations

In order to enhance the value of the products and services, MSMEs require various extensive technologies. MSMEs use these technologies to guarantee that they stay ahead i.e., competitiveness in regards to manufacturing and production activities [6].

Competition from large enterprises

Ugandan MSMEs deal with a lot of unfair competition from larger corporations in the country and other foreign economies. In addition, because they lack the finances and appropriate technologies, they find great difficulty in obtaining credit ratings that fulfill the investors’ requirements. Unlike their larger counterparts that can easily access capital and utilize sophisticated technologies to perform their daily activities, MSMEs are usually unable to prove that they possess acceptable equity to contribute to their companies [8].

2.3 Brief Analysis of Business Growth Prediction Models

Business growth prediction models aim to predict whether companies will remain open or profitable depending on measurement of growth. Since the aim is to predict growth or decline in terms of profit growth, this is therefore a classification problem. Classification involves placing each micro, small and medium business into two groups i.e. either the profits increased group or the profits decreased group [9]. It is imperative to assess and evaluate as many businesses as possible to ascertain significant factors that led to losses or profit growth in the business in the first place.

McKenzie et al [10] in their research describe the relative and absolute performance of different methods in forecasting outcomes for participants in a business strategy competition in Nigeria. They included marks assigned to business plans by judges, simple ad-hoc prediction models utilised by researchers, and machine learning methodologies i.e., Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machines (SVM) and a Gradient Boosting algorithm. The study found that overall, all methods applied had a low predictive power, however the human discretion approach i.e., grading from judges to forecast outcomes was very unreliable as the scores assigned were uncorrelated with business survival, sales, or profits three years later. They however found that various demographic characteristics increased prediction accuracy e.g., age and gender. They also concluded that the machine learning methods performed slightly better than the other methods and that using more advanced algorithms and extensive grid searchers could possibly improve performance. This dissertation will disregard manual methods of forecasting business growth and will only compare traditional and advanced machine learning methods e.g., Extreme Gradient Boosting, as they are shown to perform better than the human generated scores. Characteristic importance will also be calculated during modeling to improve the overall model performance. Their research faced the challenge of inaccuracies from respondents misreporting certain indicators. Therefore, in order to avoid this, data used will come from interviews conducted by trained research officers at the various business locations in a country wide survey, instead of using business plans written by MSME owners or entrepreneurs.

On the other hand, Ibukun Afolabi et al [11] utilized both the Naïve Bayes algorithm and the J48 algorithm for the prediction of SME success. The Naïve Bayes algorithm is based on the naive bayes algorithm that assumes that the existence of a particular feature in a class is not related to the existence of any other feature while the J48 is a decision tree that is an implementation of the Iterative Dichotomiser 3 algorithm developed by the WEKA project team.

The selection criteria for their algorithm were based on the percentage accuracy generated by each algorithm on the data using the WEKA workbench. The researchers in an attempt to increase the accuracy of prediction used the stratified cross validation method, which in the case of a binary classification, means that each fold contains approximately equal proportions of the two categories, to train and test the data. They were able to obtain accuracy of above 50 % for all three of their target variables which were “to know the success of your entrepreneur skills”, “to predict the probability of the success of your business” and “to predict how long your business will last”. Their research implemented their model within a system, making it easy for non-technical individuals to apply the model. For this dissertation, we shall also deploy the model within a system with a graphical user interface containing a simple form for input.

Machine learning models have been found to be relevant in predicting start up success. Most startups are usually MSMEs and face similar challenges to their growth and success. Cemre Unal’s research empirically demonstrates the application of several machine learning procedures to forecast startup success. Overall, six different models were executed i.e. extreme gradient boosting(XGB), recursive partitioning tree, full logistic regression, conditional inference tree , reduced logistic regression and random forest. In order to tackle the over fitting problem in the tree models, they were extended to random forests. The XGB model performed the best among other models with an accuracy of 94.45%. The best 3 performance models were determined as recursive partitioning tree, random forest and XGB, grading the same number and type of variables as their most relevant features, which were company age, last funding to date and first funding lag [12].

In this dissertation, a similar methodology in terms of the models will be applied. We shall assess ensemble methods i.e., the random forest which utilizes bagging, the XGB model which utilizes boosting and conventional methods such as logistic regression. This will allow us to compare performance between advanced and traditional machine learning algorithms. Bagging is a method used in decision trees to reduce variance and eliminate the challenge of overfitting. Boosting on the other hand aims to reduce bias and focuses on the subsequent model attempting to correct the errors of the previous model [13].

Predicting MSME profit growth is a demanding undertaking and the related financial costs are relatively on the higher end. This study will attempt to provide a modeling process on financial and non-financial data to predict MSME profitability growth using various machine learning methods.

3 Methodology

The proposed methodology for this research is the standard Knowledge Discovery in Databases (KDD) approach. This refers to the comprehensive procedure of finding information in data and stresses the high-level application of specific data mining approaches. It is most commonly used by academic researchers in data visualization, pattern recognition, databases, machine learning, artificial intelligence and statistics [14].

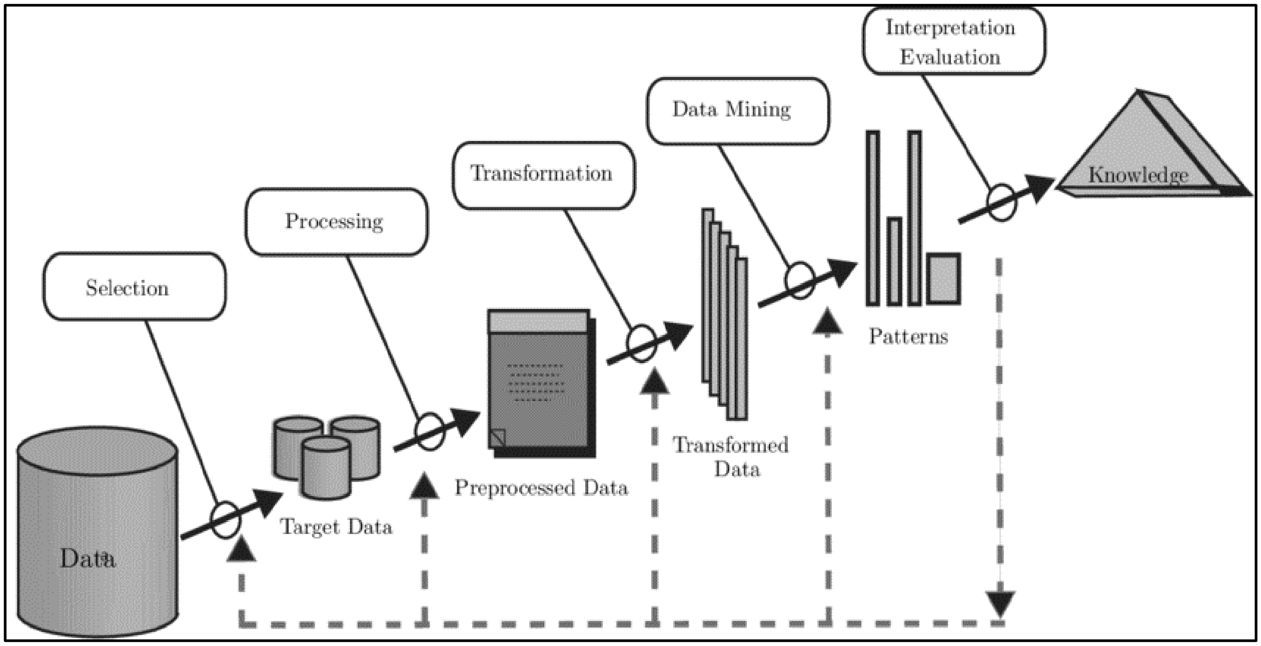


Figure 1: Knowledge Discovery in Databases Approach

3.1 Data gathering

This research intends to apply machine learning algorithms on survey data to predict profit growth. The data set is created using data from the small and medium business research 2015 in Uganda. A quantitative approach was utilized with a sampling frame of a 454,104 MSMEs that have 1-20 employees from the Census of Business Establishments (COBE). Out of that frame, a sample of 2000 MSMEs was drawn and only 1,839 face to face interviews were completed with these enterprises between March and August 2014. A structured questionnaire in Computer Assisted Personal Interviewing (CAPI) was used to facilitate the interviews such as tablet devices comprising of a pre-scripted questionnaire which each assessor had been fully skilled to administer [15]. For the purpose of this survey, the definition of MSMEs was originally agreed as follows:

- Micro: enterprises having four or less employees
- Small: enterprises having five to nine employees
- Medium: enterprises having ten to twenty employees.

3.2 Data selection and processing.

The dataset has 523 columns and 1839 rows in excel format. We shall not be using all these variables and hence we will explore the various columns, perform feature engineering and choose the ones that are most relevant.

The following steps were followed to have the final dataset:

1. Only the relevant columns were selected by gaining a simple understanding of the problem and dataset via data exploration.
2. All columns with more than 50 % null values were removed.
3. 1.5 % of the Age of the business field is missing values. We impute them with the age means of each sector. Imputation is a method of filling in missing values with an estimate. There are many types of imputation for example substitution, which includes using value from a new MSME not previously selected in the sample, hot deck imputation which involves finding other MSMEs with other similar values and randomly selecting one of their values to fill the age variable, cold deck imputation which is similar to hot deck but eliminates the random variation i.e., selecting the value always from the first MSME etc. [16]. This dissertation will utilize mean imputation for this variable. This means that we find the average of the business ages in the various sectors i.e., mining, financial and manufacturing etc., separately and replace the missing values with the means of each sector. We used mean imputation as it easy to understand and apply and will not require us to reduce the sample size further like other methods for treating missing values.
4. All duplicates columns were removed. For example, some columns like 'Legal status' and 'What is the current legal status of the business?' have nearly the same information. The latter contains the answers the users gave when they chose 'Other' on the former rather than one of the choices.

5. All rows with profit growth variable, having ‘Don’t know’ and ‘Stayed the same’ were also removed to make it easier for the model to predict success (increase in profits).

3.3 Feature transformation

3.3.1 Feature Engineering

Feature engineering is the process of creating new features that best represent the underlying problem from raw information or existing features. This process is usually intended to improve model accuracy[17]. There are limitless possibilities for this stage, therefore some of the techniques we shall apply in this dissertation include, combining similar features to create the investment and financial services variable, removing unused or redundant values which are those that are not clear or don’t make sense to pass into our machine learning algorithms from the target variable i.e., increased profits from the same month last year, and creating dummy variables to convert all the text/categorical features into a numeric format because most machine learning algorithms work better with numeric format data.

i. **Investments variable**

YES value assigned when answering yes to one or more of the following variables.

NO value assigned when answering no to all of the following variables.

- What investments have you made in the past year in your business? Machinery and equipment (including computers and software)
- What investments have you made in the past year in your business? Buildings/land
- What investments have you made in the past year in your business? Trainings/human capital for you or your employees
- What investments have you made in the past year in your business? Other specify
- From fiscal year 2012 through 2013 did this establishment provide formal training to any of its employees specifically for the development and/or introduction of innovative products or services and processes?

ii. Financial Services

YES value assigned when answering yes to one or more of the following variables.

NO value assigned when answering no to all of the following variables.

- Does your business use any products or services from Credit/finance dealer?
- Does your business use any products or services from Commercial Bank?
- Does your business use any products or services from Microfinance institution?
- Does your business use any products or services from SACCO?
- Does your business use any products or services from Mobile money?
- Does your business use any products or services from Money lender?
- Does your business use any products or services from Other financial institution?

After creating these features, the variables that are used to create them are deleted from the dataset.

iii. Target variables

As documented by MSMEs studies [18], most SMEs in East Africa, particularly indigenous ones, hardly survive beyond three years and the continued profit growth over a year is rare. With this hindsight, this study classified a profitable MSME as having increased profits from the same month last year. This variable describes whether the business increased, decreased or remained the same compared to last year in the same month in terms of profits made. This variable is filtered to remove the 'I don't know' and 'Stayed the same' variable. This is implemented to remove ambiguity that may affect the model prediction process.

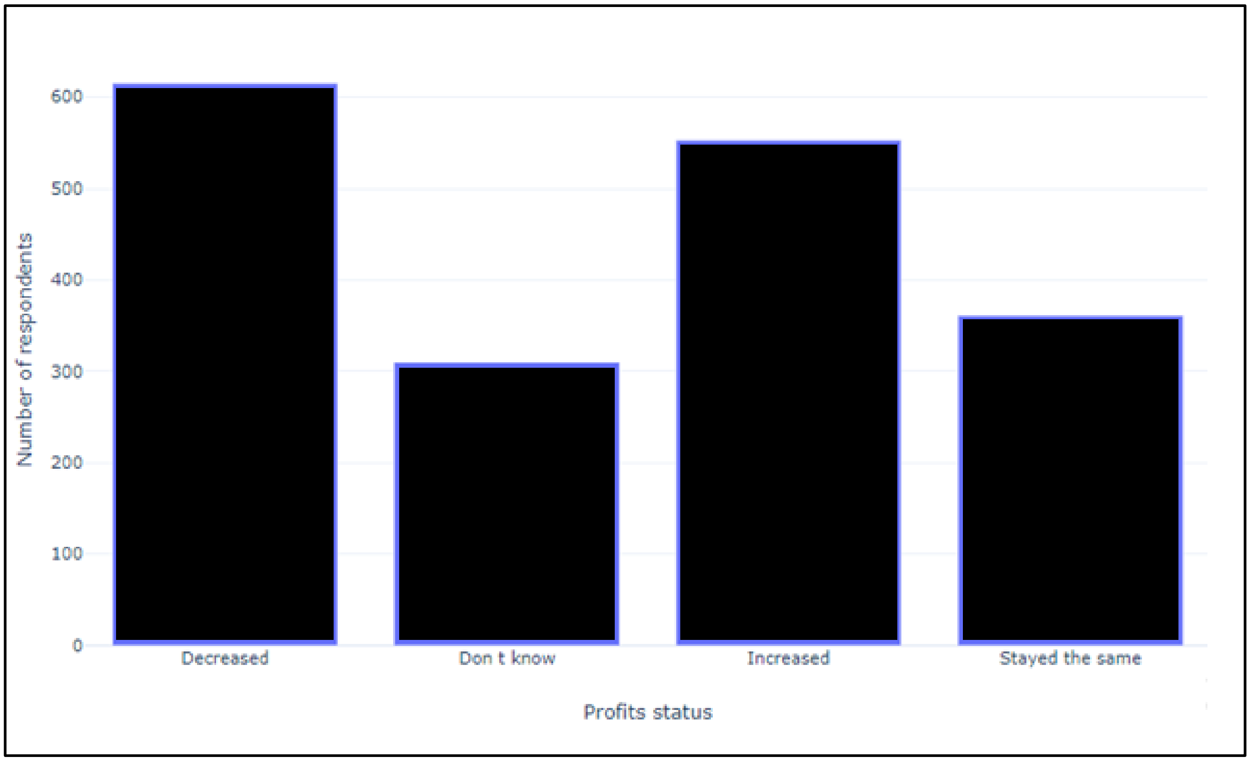


Figure 2: Target Profit variable (increased profits from the same month last year) before removing unused or redundant values from the target variable.

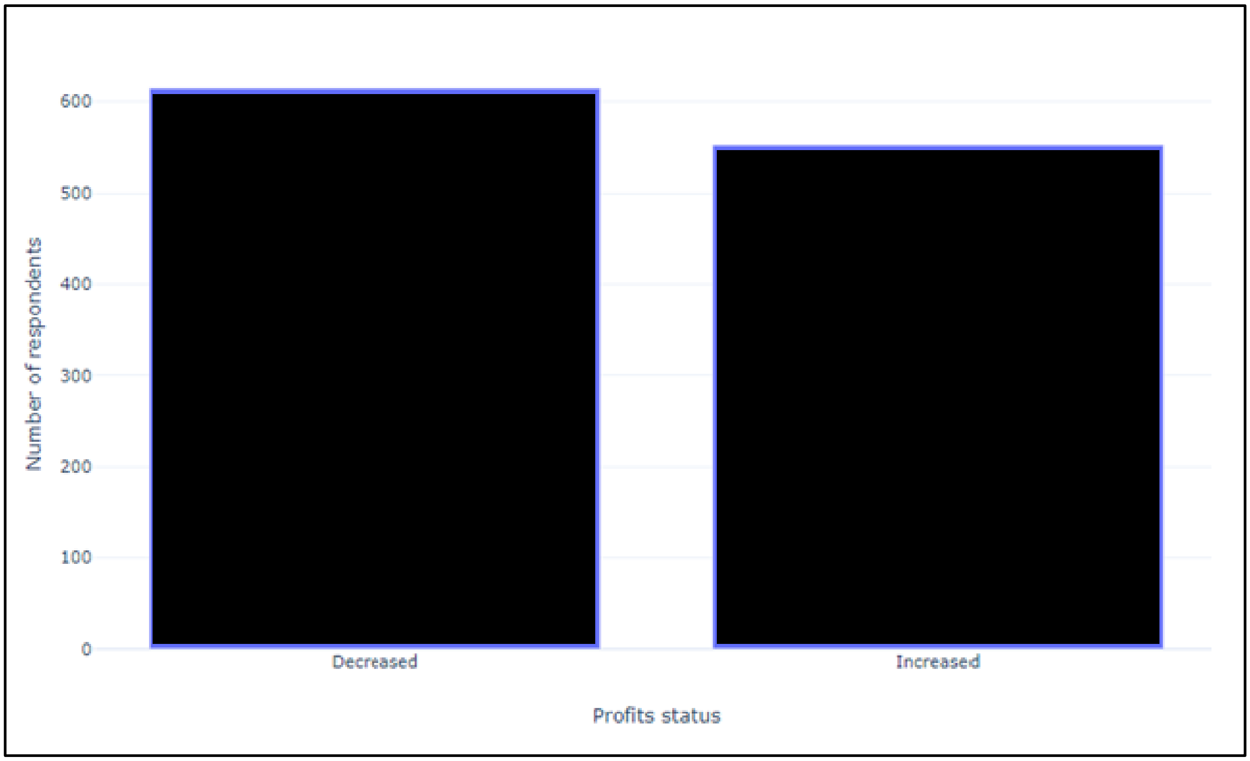


Figure 3: Target Profit variable (increased profits from the same month last year) after removing unused or redundant values from the target variable.

Fortunately, the target variable is balanced therefore, we shall not have to implement treatments for imbalance during the fitting process. A balanced dataset is one which the number of MSMEs are approximately the same number of instances in each class i.e., approximately the same number of decreased and increased in the target profit variable (increased profits from the same month last year).

After data cleaning, the lists of variables to be used throughout this dissertation are summarized in the table below:

Table 1: Description of variables data types

Variable	Type
Sector	Categorical
Region	Categorical
Role in business	Text
Age of owner	Numeric
Gender of owner	Categorical
Primary activity	Text
Operation time	Numeric
Legal status	Categorical
Operation premises	Categorical
Business foundation	Text
Startup capital funds	Text
Business influence start	Text
Physical receipts in an organized manner	Categorical
Bookkeeping for revenues (but not necessarily expenditures)	Categorical
Do full financials (revenue and expense excluding tax accounting)	Categorical
Do full financials (revenue and expense including tax accounting)	Categorical
Firm total sales status	Categorical
Profits status	Categorical
Investments	Categorical
Formally trained employees	Numeric
Experienced employees	Numeric
In charge	Text
Business owner education level	Categorical
Member of any association	Categorical

Preferred information source	Text
Use of expert advice from outside the business	Categorical
Financial services	Categorical
Rejected loan apps	Categorical
Customers come to you or do you take products to your customers	Categorical
Internet access	Categorical
Introduce any innovative product, service or process?	Categorical
Business size	Categorical

3.3.2 Encoding Variables

Encoding variables is the process of converting categorical variables into numeric formats. Generally encoding variables is mainly a limitation of the effective application of machine learning procedures rather than hard constraints on the algorithms themselves. While some algorithms can work with categorical and text data directly, many machine learning algorithms cannot operate on label data directly e.g., depending on the specific implementation, a decision tree can learn directly from categorical data with no data transformation required. Most algorithms implementations however require all input variables and output variables to be numeric [19].

This dataset contains mostly data in categorical and text format. Therefore, a combination of class mapping, label and one hot encoding is used in this dissertation to convert text/categorical data into numerical data and enable the algorithms to make sense out of it. Class mapping is used to encode variables that are ordinal. Ordinal variables are variables for which possible values are ordered. It will assign the weight in direct proportion of the values i.e., size of the MSME, 'Medium': 3,'Small': 2 and 'Micro': 1. We shall then apply label encoding for the nominal variables, which are variables that do not have order or rank i.e., gender ‘Female’: 1, ‘Male’: 0, ‘Yes’:1, ‘No’:0 etc. [20].

Using label encoding for nominal variables with more than two possible values may lead the model to assume that encoded integer values for each input variable have an ordinal relationship. For example, for the variable sector with value ‘utilities’ encoded as ‘16’ may be perceived as being better or lesser than value ‘accommodation’ encoded as ‘1’ which is untrue. Therefore, to prevent this, one hot encoding is applied to the values which are mapped into new binary variables, one new variable for each categorical value. An example of this would be the variable preferred information source with values government institutions, media etc. would create new variables ‘preferred information source government institutions’, ‘preferred information source media’ etc. After the encoding and modeling process is over, the predicted output variable will be converted back to categorical/text format by the model to present them in an easy-to-understand format i.e., from [0,1] to [decreased, increased].

Table 2: Class mapping of categorical/ text data

Business Size Before Encoding	Business Size After Encoding
Micro	1
Small	2
Medium	3

Table 3: Label encoding of categorical/ text data

Gender Before Encoding	Gender After Encoding
Female	1
Male	0

Table 4: One hot encoding of categorical/ text data

ID	Preferred Information Source		ID	Business Assoc.	Financial Inst.	Govern-ment	Media	Customers
1	Business Assoc.		1	1	0	0	0	0
2	Financial Inst.	→	2	0	1	0	0	0
3	Government		3	0	0	1	0	0
4	Media		4	0	0	0	1	0
5	Customers		5	0	0	0	0	1

3.4 Data Mining and feature selection

The subsequent step is to examine if there is an observable difference between profit and loss making MSMEs. In order to determine and compare the performance of the algorithms built in the next sections, we shall assess relationships between the predictor variables and independent variables to determine whether those input variables are significant to the results being predicted. The relationship between categorical variables will be assessed using the chi-square test while that between categorical and continuous variables will be tested using the ANOVA test.

A few strengths of chi-square are that it is easier to compute than some statistics. Also, it can be used with data that has been measured on a categorical scale, like the data used in this dissertation. It can also be used to see if there is a difference between two or more groups of participants. Another strength is that unlike other statistics, chi-square makes no assumptions about the distribution of the population [21]. There is a significant number of categorical variables that contain multiple groups in the dataset used for example the education of the owner variable has several classes for example secondary, primary, university and tertiary. ANOVA is therefore particularly valuable in this case when determining whether differences in mean values between three or more groups are significant.

3.4.1 Correlation between categorical variables

In the event of classification problems such as this one, where most of the input features have a data type of the categorical nature, pearson's chi-squared χ^2 statistical hypothesis tests are utilised to decide whether the target variable is reliant on or independent of the input variables. The χ^2 test of independence [22] involves forming a cross tabulation table between variables with *r* rows and *c* columns. Depending on the cell counts, presence of a relationship between variables and the strength of that relationship can be tested by testing the difference between the expected count, E, and the observed count, O. The subscript *i* is used to symbolise the row group, i.e., row *groupi*, while *j* will be used to symbolise the column group, i.e., column *groupj*, meaning the cell will be denoted by *celli,j*. If variables are found to be independent, then they might be irrelevant to the problem at hand and therefore could be considered for removal.

i. Hypotheses

- Null hypotheses: The two categorical variables are independent (no relationship between the two variables)
- Alternative hypotheses: The two categorical variables are dependent (there is a relationship between the two variables)

ii. Test statistic

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - \hat{E}_{i,j})^2}{\hat{E}_{i,j}}$$

One would reject the null hypothesis, H_O, if the calculated χ2 test statistic is > the critical χ2 value based on the degrees of freedom and α level. Degrees of freedom are calculated using (r−1)(c−1) where r is the number of rows and c is the number of columns.

In terms of a p-value and a chosen significance level (alpha), the test can be interpreted as follows:

- If p-value <= alpha: significant result, reject null hypothesis (H0) i.e., dependent.
- If p-value > alpha: not significant result, fail to reject null hypothesis (H0) i.e. independent.

3.4.2 Correlation between categorical and continuous variables.

Correlation between categorical and continuous variables, can be complicated. There are three possible methods to understand if a continuous and categorical are significantly correlated i.e. biserial correlation, logistic regression, and Kruskal Wallis H Test or parametric forms such as t-test or ANOVA. In this case, we shall use the ANOVA test. [23].

i. Hypothesis

$$H_0: \overline{x}_1 = \overline{x}_2 = \overline{x}_3 = \ldots = \overline{x}_k$$

H_A:At least one of the groups means differ

ii. Test statistic

This is the F- statistic and it compares the mean square between samples (MSB) to the mean square within sample (MSW). This F-statistic can be calculated using the following formula:

$$F = \frac{MS_B}{MS_W}$$

Where;

$$MS_B = \frac{\text{Sum of square between sample } (SS_B)}{(k-1)}$$

$$MS_W = \frac{\text{Sum of square within sample } (SS_W)}{(n_T - k)}$$

k is the number of groups

n_T is the total number of observations

and where,

Sum of square between sample (SS_B) = $\sum_k \; n_k(\overline{x}_k - \overline{x})^2$

Sum of square within sample (SS_W) = $\sum_{i,k} \; (x_{i,k} - \overline{x}_k)^2$ or can be calculated as $\sum_k \; (n_k - 1)s_k^2$

One rejects the null hypothesis, H0, if the computed F-static is greater than the critical F-statistic. The critical F-statistic is determined by the degrees of freedom and alpha, α, value. Reject H_0 if calculated F-statistic $>$ critical F-statistic.

4 Model development

Machine learning algorithms are arranged into a taxonomy based on the desired result of the algorithm. These include supervised, unsupervised and reinforcement learning. Supervised learning maps named information to known result, while unsupervised learning investigates patterns and forecasts without known labels. Reinforcement learning takes after a trial-and-error strategy [24].

This dissertation will utilize supervised learning algorithms which generate a function that maps inputs to desired outputs i.e., the function is inferred from labelled training data. Some of the problems that can be solved by supervised algorithms include regression and classification. Classification is involves approximating a class label while regression entails estimation of a continuous variable from a list of input factors [24]. Since we are addressing a supervised classification problem i.e., one in which the algorithm approximates class labels for a given set of input data, we shall compare the performance using the traditional and simple logistic regression algorithm and the more advanced extreme gradient boosting and random forest algorithms.

4.1 Extreme Gradient Boosting Algorithm

Extreme gradient boost is an enhanced distributed gradient boosting library developed to be highly proficient, flexible, and portable [25]. XGB provides a parallel tree boosting that resolves multiple data analytical and science problems in an efficient and precise way [25], while implementing machine learning under the Gradient Boosting framework.

Extreme gradient boosting (XGB) adds to gradient boosting by punishing trees for misclassifications, using additional randomization parameters to guarantee low variance, refining computing efficiency and reduction of the leaf nodes. It creates additive regression models by consecutively fitting a parameterized function to current pseudo-residuals by least squares at each iteration. It may be used to solve several machine learning problems i.e., regression, classification and user-defined prediction problems [26].

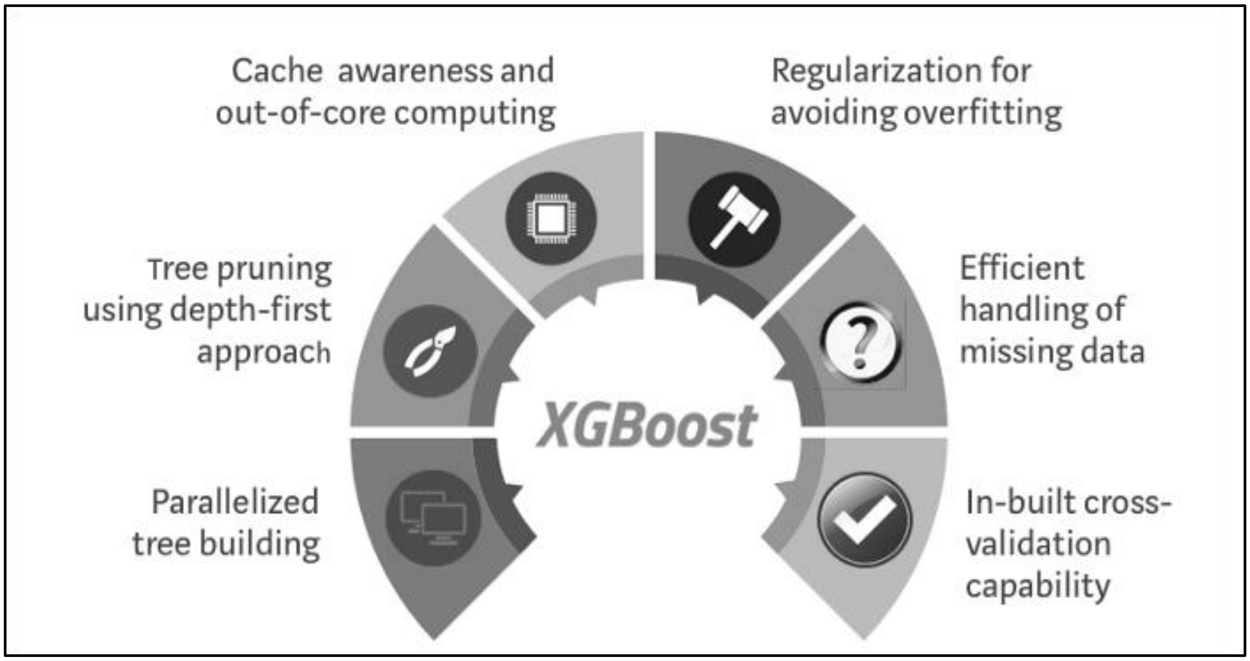

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Figure 4: How XGB improves upon gradient boosting [27]

The following steps are involved in gradient boosting [28]:

- ✓ F0(x) – with which we initialize the boosting algorithm – is defined as:

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

- ✓ The gradient of the loss function is computed iteratively:

$$r_{im} = -\alpha\left[\frac{\partial(L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \textit{ where } \alpha \textit{ is the learning rate}$$

- ✓ Each hm(x) is the fit on the gradient obtained at each step.
- ✓ The multiplicative factor γm for each terminal node is derived and the boosted model Fm(x) is defined as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

4.2 Logistic Regression

Logistic regression is a classification algorithm, used when the target variable is a binary variable. The algorithm builds a regression model to predict the probability that a given data entry belongs to a certain category. Logistic regression models the data using the sigmoid function that enables us to shrink real valued continuous inputs into a range of (0, 1) [29].

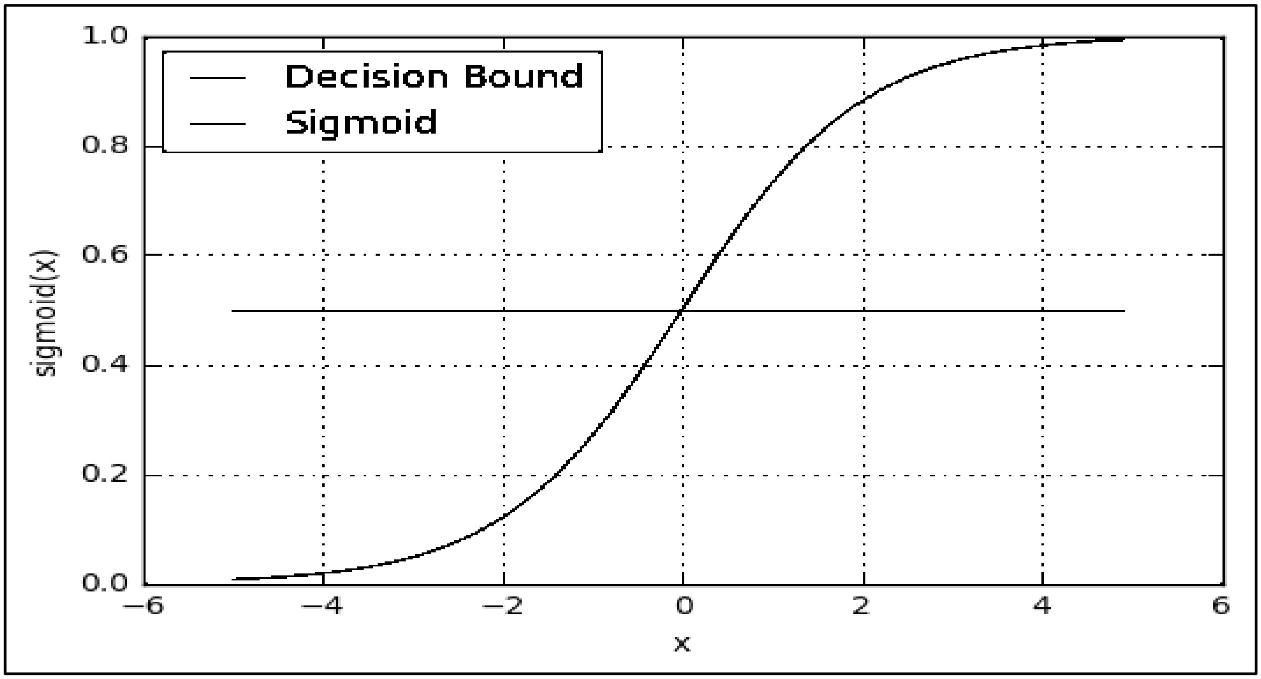


Figure 5: Logistic regression decision boundary

The logistic function:

$$g\left(z\right) = \frac{1}{1 + e^{-z}}$$

Logistic regression uses a decision boundary that acts as a threshold value which allows us to classify values into classes depending on whether they are above or below the boundary.

4.3 Random forest

Random forest is a supervised learning algorithm. It is an ensemble of decision trees, trained with the bagging method which is based on the principal that a combination of multiple decision trees increases the accuracy. It can be used for both classification and regression.

The model also overcomes the over fitting problem usually faced by decision trees as it consist of multiple trees from which a random choice is made [30]. Overfitting happens when the algorithm models the training data too well i.e., it learns the detail and noise in the training data as concepts of the model hence impacting the model's ability to generalize and model new data [31]. Random forests can also handle null values [30], which in our case is not a problem as we treated all missing values as shown in the previous chapters. Our data set consists of multiple categorical values which the model is well known for handling properly.

i. Modeling pseudo code [30]:

1. Randomly choose a number of variables 'a' from total variables 'b'. Where a << b.
2. From the chosen variables, calculate the node "d" using the best separation point.
3. Separate the node into daughter nodes using the best separation point.
4. Repeat 1 to 3 stages until only one node is reached.
5. Build forest model by recapping stages 1 to 4 for "x" amount of times to generate "x" amount of trees.

ii. Prediction pseudo code [30]:

1. Utilise the test variables and the rules of each randomly generated decision tree to predict the outcome and store the result.
2. Compute the votes for every predicted target.
3. Take the high voted predicted target as the final likelihood or prediction from the random forest model.

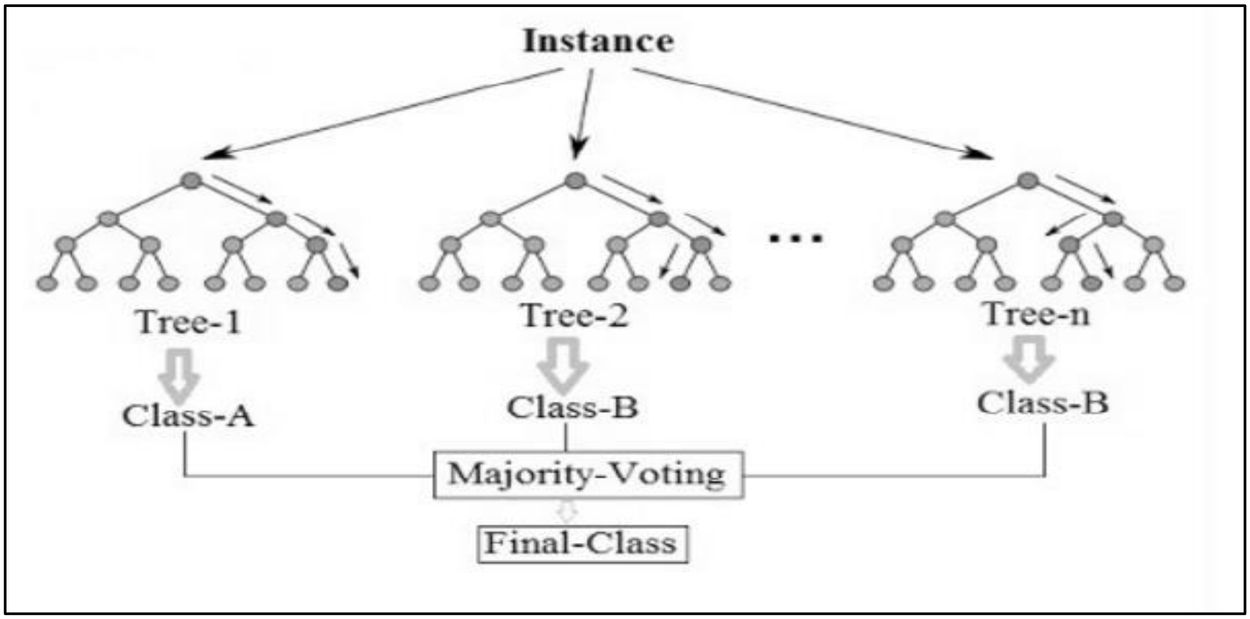


Figure 6: Random forestry majority voting

5 Model Selection

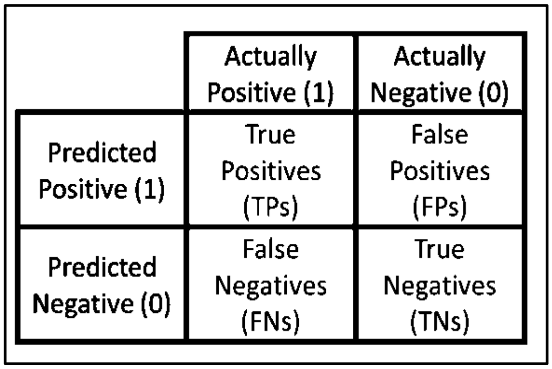
Before implementing models on a dataset and comparing them, it is not possible to say which method will be more superior in terms of performance as there is not a single algorithm that performs better in all problems. In the following section, estimations of three models and their outcomes will be explored and discussed. Tuning each model's parameters is vital to acquire the best performance possible. To adjust the models' parameters, grid search using the sklearns GridSearchCV is used to find the optimal value for the hyper-parameters. While there are other methods of selecting the best parameter such as random search, grid search was chosen as it allows for one to try every combination of the set of parameters defined unlike other techniques such as random search which may not test all the combinations. Grid search however does not scale well and uses more computation time [32]. The approach queries a given number of combinations of parameters in a sequence, where each hyper-parameter consists of a list of pre-defined lower and upper bounds created using the arrange method or function and formatted into a list. Furthermore, to validate the optimized classifiers to the training set, a k-fold cross-validation procedure is applied for each model, with the train dataset being split into 10 samples [33]. Cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample, such as the one used in this study. One subsample is reserved as the validation data for testing the algorithm, while the remaining k-1 (9) subsamples are utilized for training data. The procedure is repeated 10 times, with every 10 samples used only once. The 10 outcomes from the folds are averaged to yield a single performance approximation on the training set for model selection [34]. The performance of the final model is then reported on the test set.

5.1 Model Performance Optimization

This section elaborates the use of receiver operating characteristic (ROC) to optimize the performance measures for machine learning algorithms covered in the model development section. ROC is used as a scoring tool within the grid search fitting of the data to select the best hyper parameters. ROC can also be described as a probability curve and Area under Curve (AUC) represents degree or measure of separability. It conveys how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting false as false and true as true [35].

5.2 Performance Measures

The explanation of the models' performance represents the most important final step. Here, we discuss the performance measures used in this dissertation. When validating the performance of a classifier on the test set, the predicted outcome produced by the classifier are counts of the accurate and inaccurate classifications from each class. This information is usually displayed in a confusion matrix. A confusion matrix is a table indicating the differences between the true and predicted classes for a series of labeled data values [36], as shown in Figure 7 for a binary classification case. While the models we use will generate a general classification accuracy measure usually in percentage format e.g., 90%, this measure may hide the details we need to better understand performance. The confusion matrix on the other hand, is easy to implement and gives insight into the types of errors that are being made. It will allow us to assess not only how well the model predicted which MSME will make profits but also how well the model predicted which MSME will make losses i.e., sensitivity and specificity.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 7: Confusion matrix

Performance measures that can be derived from the confusion matrix include [36]:

i. **Accuracy** which refers to the overall percentage correctly classified.

$$(\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})$$

ii. **Sensitivity** which is the probability that a true positive sample is correctly classified into true positive group.

$$\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

iii. **Specificity** which is the probability that a true negative sample is correctly grouped into the true negative group.

$$TN/(FP + TN)$$

6 Results and Discussion

6.1 Correlation Analysis Results

Table 5: Chi Square test results between the independent variables and profit growth target variable

Independent variables	P
Sector	0.017
Region	0.000
Role in business	0.000
Age of owner	0.445
Gender of owner	0.467
Operation time	0.078
Firm_last_year_month	0.000
Nationality of owner	0.132
Primary activity	0.012
Legal status	0.000
Operation premises	0.133
Business foundation	0.071
Startup capital funds	0.641
Physical receipts	0.000
Bookkeeping	0.002
Do full financials (excluding tax)	0.000
Do full financials (including tax)	0.000
Investments	0.000
Trained employees	0.001
Experienced employees	0.051
In charge	0.000
Business owner education level	0.000
Member of any association	0.000
Preferred information source	0.012
Use of expert advice	0.000
Financial services	0.000
Rejected loan apps	0.116
Customers Come to You	0.073
Internet access	0.179
Introduce any innovation	0.000
Business size	0.043

P-value is the probability that the correlation between x and y in the sample data occurred by chance. A p-value of 0.05 means that there is only 5% chance that results from your sample occurred due to chance. A p-value of 0.01 means that there is only 1% chance. So lower p-values mean there is a strong evidence of correlation[37]. From this table we can tell that there are more variables that are correlated to the target variable than those that are not. While understanding the correlation of variables independently with the target variables is relevant, it is important to note that models usually perform based on a combination of variables rather than the strength of each individual variable. Therefore all variables are input into the model and feature importance is used to decide which variables are more relevant to the prediction process.

6.2 Extreme Gradient Boosting Implementation

The parameters affecting the model performance are initially set to learning rate: 0.01, n_estimators: 300 and max_depth: 5. In order to improve performance, and to determine the best parameters we use 10 fold grid search cross validation which allows for sequential looping through pre-defined parameters and chooses the best ones for fitting our dataset. The best parameters generated are as follows; depth is reduced to 3, iterations are set to 260 with a subsample of 0.8 and a learning rate of 0.01. Gamma, the loss reduction parameter to control the over fitting problem, is set to 1. The model achieved an accuracy of 92.31%.

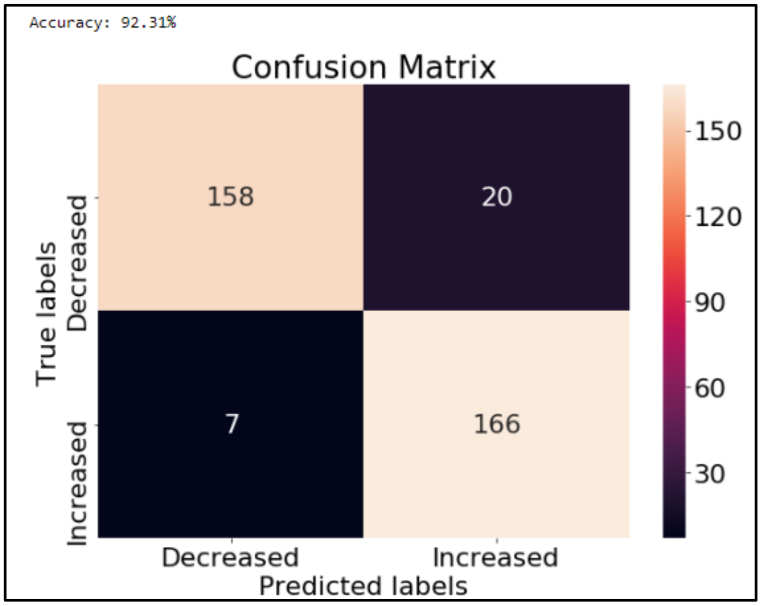


Figure 8: XG Boost Confusion matrix

Variable importance for the XGB model is measured using F score which is sometimes referred to as the weight. It refers to the number of times a variable is used to split the data across all trees [38]. Sales made last year, operation time and business owner education levels are seen to be the most important variables for the XGB model. Information source and business influence start variables are also among the top ten variables that contribute towards the model's decision.

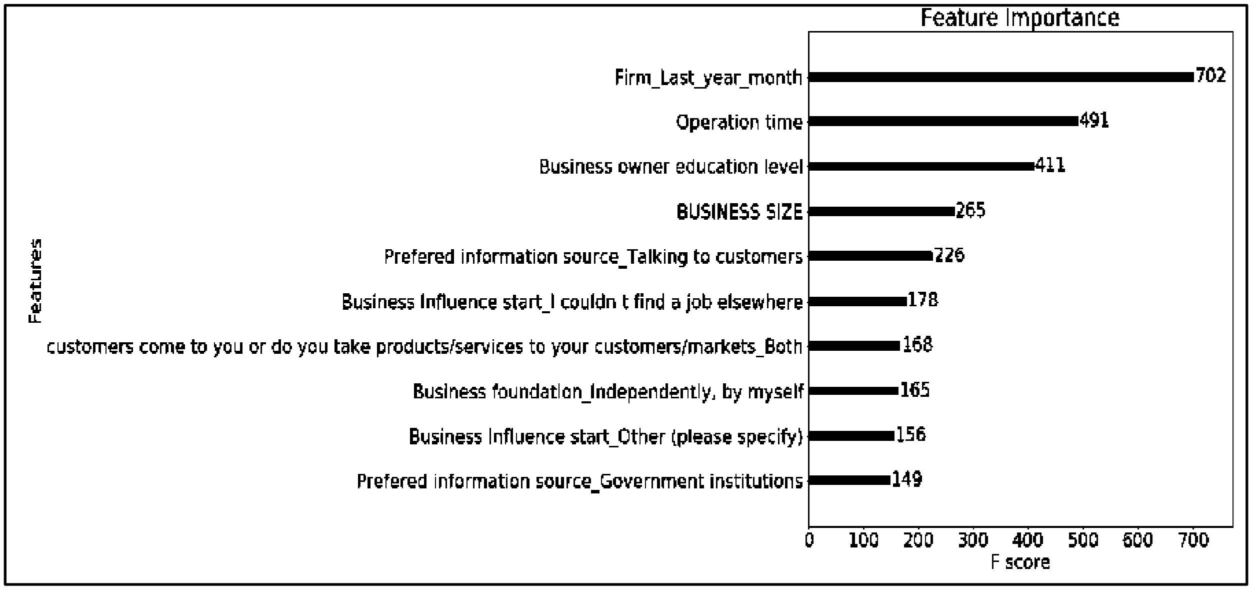


Figure 9: XG Boost Feature Importance

6.3 Random Forest Implementation

Decision trees are more inclined to over fitting, especially when a tree is deep. This is because of specificity due to the smaller sample of events available. We should be aware of this because small samples could lead to unsound conclusions [39]. Our data set contained 1839 interviews and was made smaller due to clean up process. In order to overcome these hurdles, a forest of decision trees is created. Using the grid search with cross validation, the best parameters are set to 200 estimators, with a max depth of 7. This model performs slightly worse than the XGB model, achieving an accuracy of 92.02%

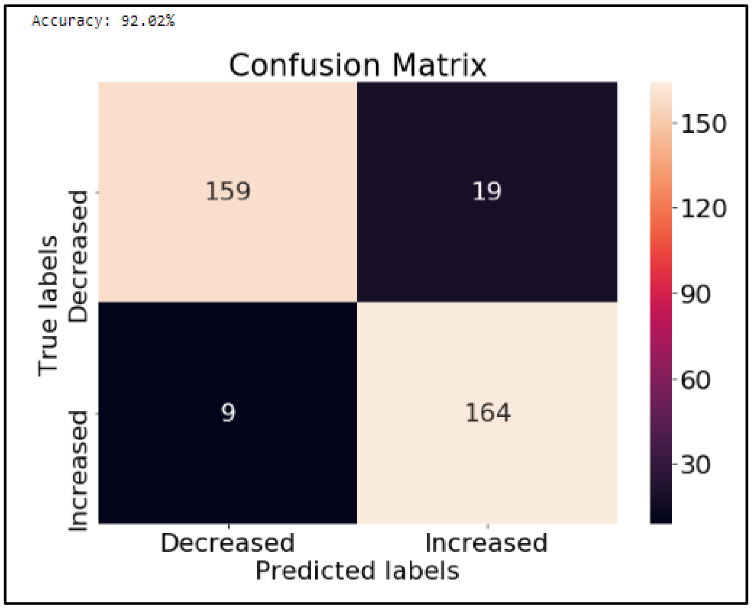


Figure 10: Random Forest Confusion matrix

Random forests consist of nodes in which a condition on a single feature splits the dataset into two. The measure based on which ideal condition is chosen is called impurity [40]. We can therefore compute how much each feature decreases the weighted impurity during the training phase. For a forest, the impurity decrease from each feature can be averaged and the features ranked according to this measure. Just like the XGB model, sales made last year, operation time and business owner education level are the most important variables.

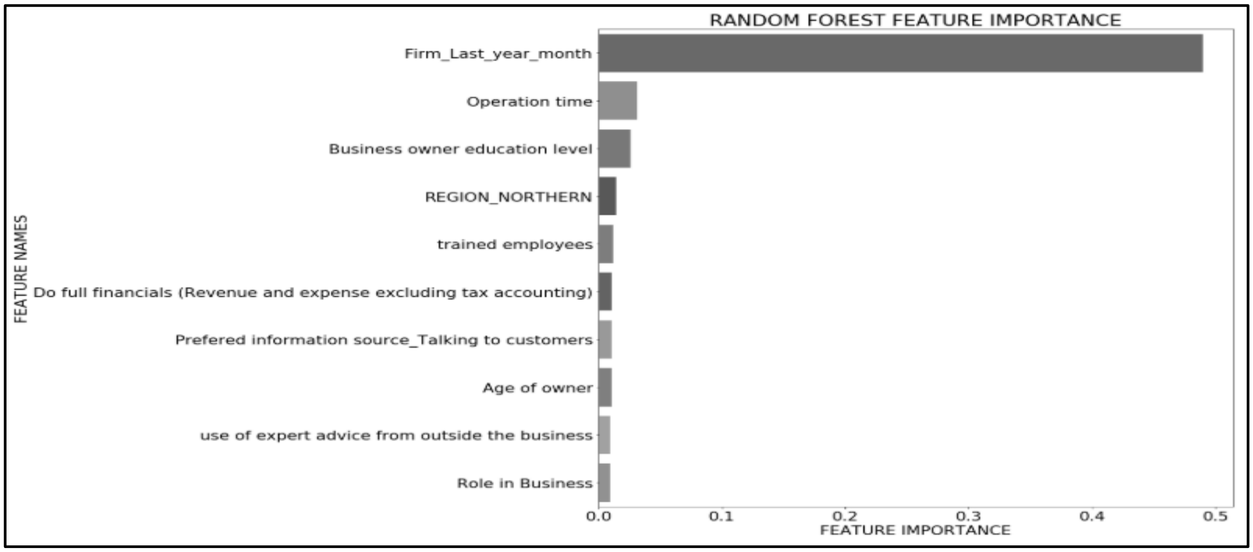


Figure 11: Random Forest Feature Importance

6.4 Logistic Regression Algorithm

Logistic Regression is one of the traditional algorithms used in machine learning. In this scenario the grid search is performed with cross validation. The regularization strength is set to 0.01 and the penalty to l2. This model achieves an accuracy of 90.88%. Although this is a good performance, it still does not measure up to the other models.

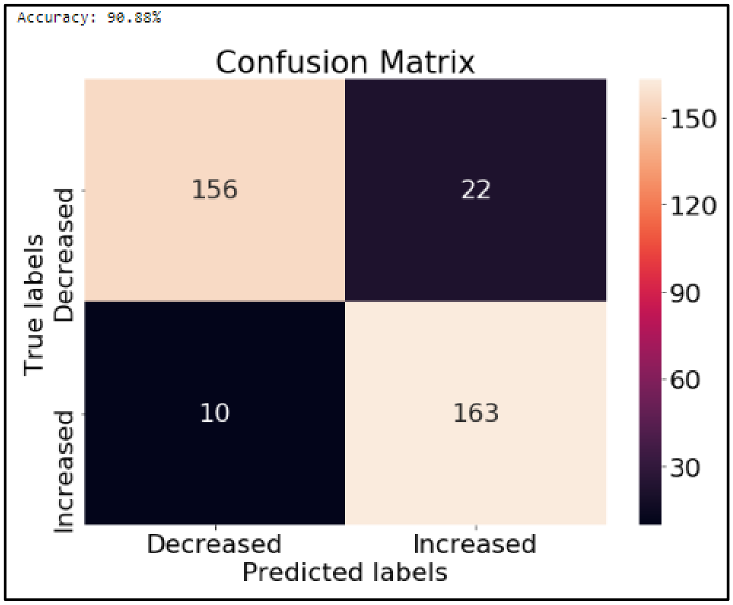


Figure 12: Logistic Regression Feature Importance

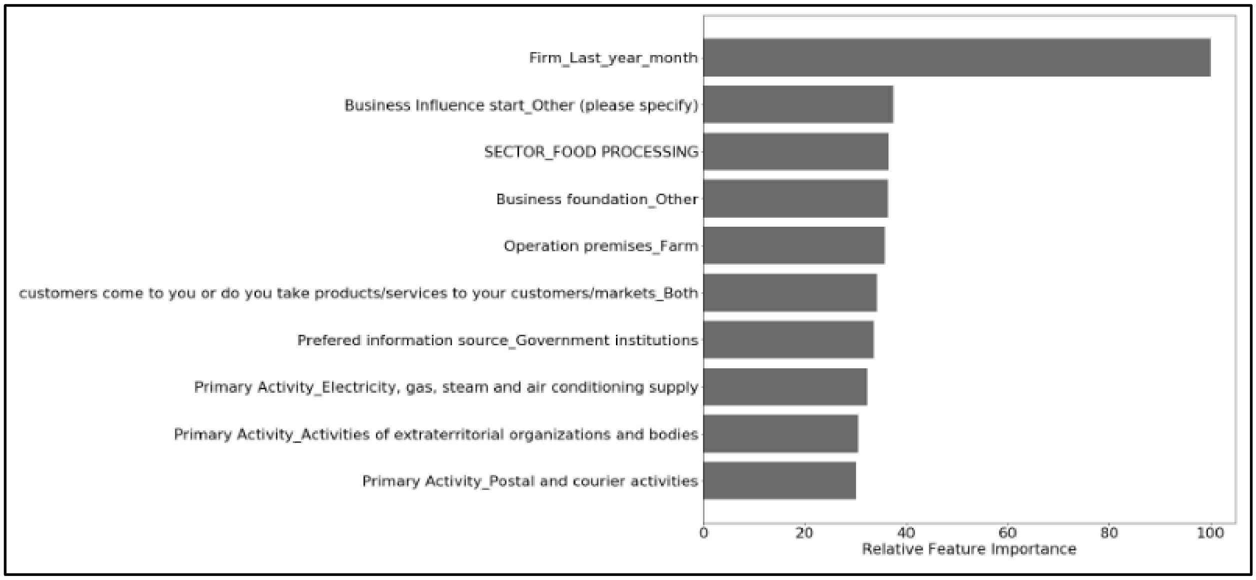


Figure 13: Logistic Regression Feature Importance

Observing the variable importance, which in this case is generated by the coefficients that provide a basis for a crude feature importance score, shows the most important variable is sales made last year. The other variables in the top ten however vary greatly from the other models.

6.5 Comparison of Models

There many metrics that can be used to determine the model performance. In this case we utilized the following, which can all be derived from the confusion matrix. Below is an overview of the various comparison metrics.

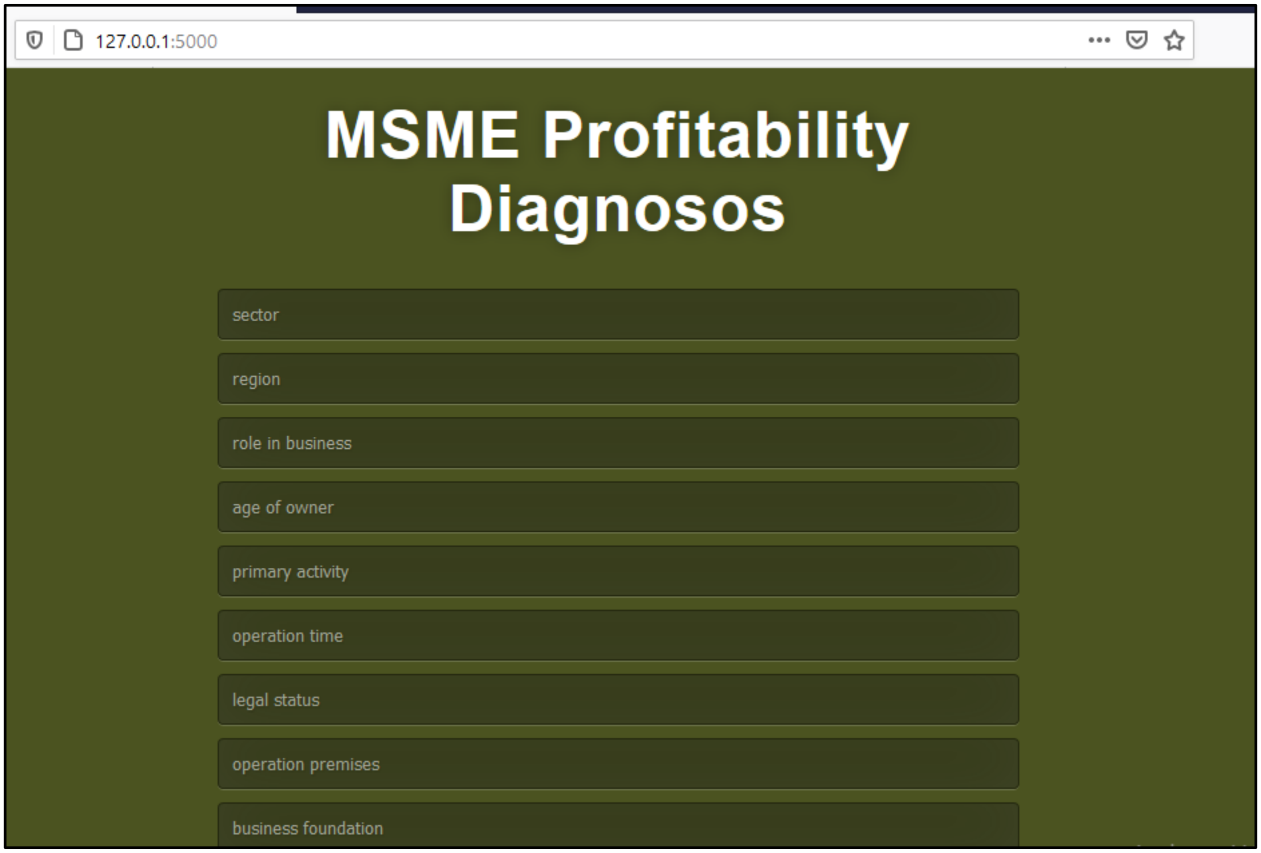
Table 6: Comparison of models

Model	Accuracy	Sensitivity	Specificity
XG Boost	92.31%	88.76%	95.95%
Random Forest	92.02%	89.33%	94.79%
Logistic Regression	90.88%	87.64%	94.22%

These metrics rank the XGB model as the best performing method. This means XGB is able to label decrease and increase in profits classes better, in comparison to the other methods. However we also note that the XGB model demonstrates a lower sensitivity than the Random Forest model meaning that Random forest is better at correctly predicting MSMEs that are making losses than the XGB model. However since the goal is to predict MSMEs making profits, we can see that the XGB model has the highest specificity, hence making it the best performer. Random forest is the close second after XGB. This shows that the general classification performance of the ensemble methods is better than methods such as logistic regression which is considered to be more traditional.

6.6 Deployment

In fulfillment of the second objective of this research, a web-based application was developed from which predictions of profitability can be made. The application is a simple form that allows the user to answer questions and generates a prediction of profitability growth. The best model i.e. XGB is stored in pickle serialized format and then is deployed within this application which is built using flask. Flask is a web framework that enables one to develop web applications easily and has a small and easy-to-extend core [41].



127.0.0.1:5000

MSME Profitability Diagnosos

sector

region

role in business

age of owner

primary activity

operation time

legal status

operation premises

business foundation

Figure 14: Diagnosis form

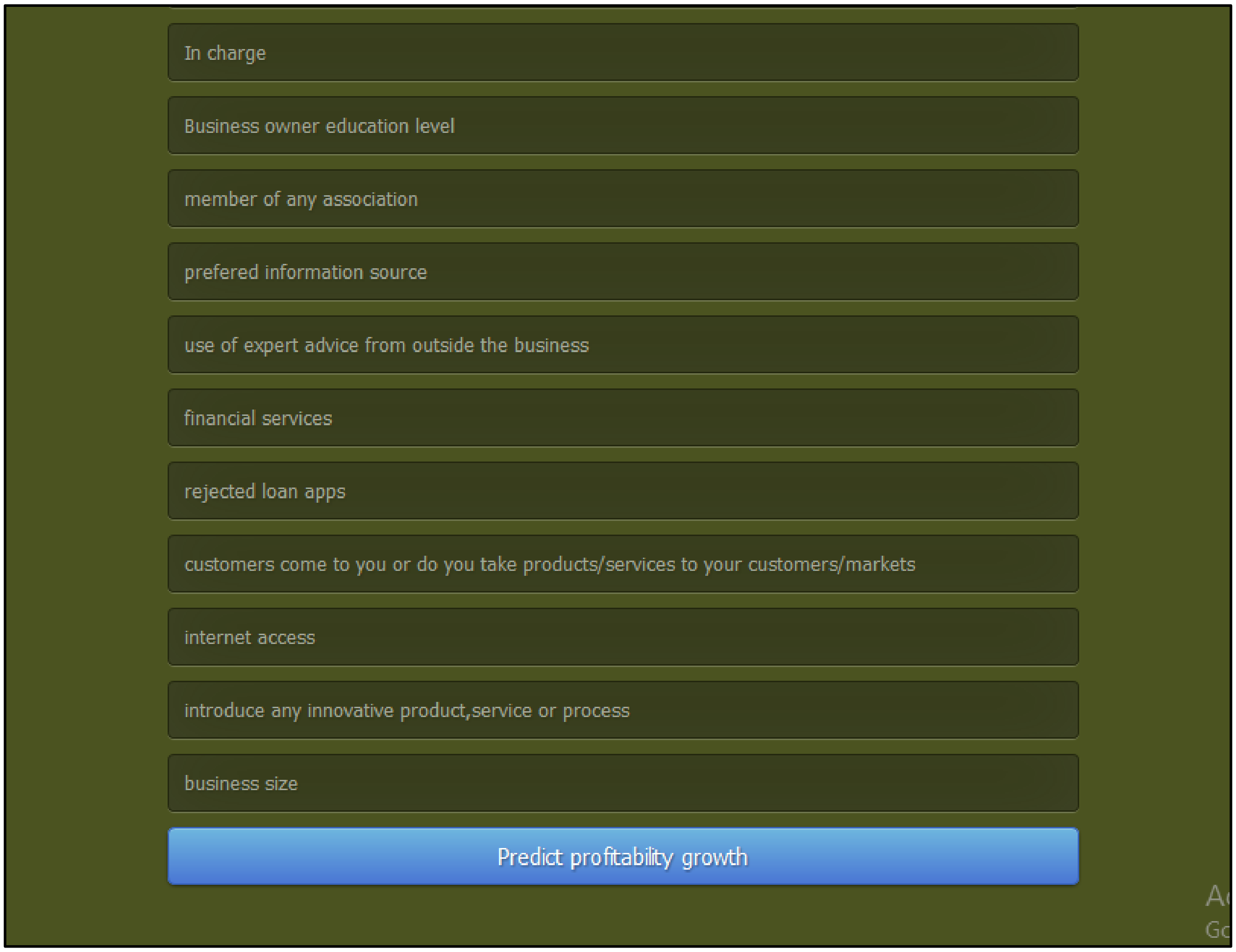

In charge
Business owner education level
member of any association
prefered information source
use of expert advice from outside the business
financial services
rejected loan apps
customers come to you or do you take products/services to your customers/markets
internet access
introduce any innovative product,service or process
business size
Predict profitability growth

Figure 15: Diagnosis form

6.7 Key Findings and discussion

The central objective of this research is to investigate the potential of utilising survey data for MSMEs profitability growth prediction. A resultant research question is formulated in the introduction, where we investigate the applicability of mining annual survey data for MSME growth prediction. Working with survey data can be a challenging task due to the moderately high quantity of vague responses for example ‘don’t know’ and ‘not sure’ responses in the data set. Therefore it is very important to deal with this data appropriately. I conducted a systematic literature search that allowed me to narrow down the number of fields to 32 useful and relevant factors. Regarding the target variable, to avoid inducing erroneous bias into the model it was filtered to remove vague variables such as ‘don’t know’ and ‘stayed the same’.

Correlation between individual independent variables and the dependent variable is important but could not be fully relied on to make feature importance decisions. We therefore use the Fscores and relative importance to determine how important a variable is in predicting growth. The features appearing in the top ten feature importance for both of the best performing models include ‘sales made last year’, ‘operation time’ , ‘business owner education level’ and ‘preferred information source talking to customers’. In figure 16 below, we can see that businesses whose sales were higher last year, are more likely to be profitable this year. We can also see that even though most MSMEs are owned by individuals with only a secondary school education, the education level group that has the highest percentage of profitable business is the university education level group. Secondary level has (160 increased: 170 decreased), primary level has (69 increased: 126 decreased), primary level has (4 increased: 12 decreased), while university level owners have a ratio of (151 increased: 131 decreased) and tertiary have (151 increased: 131 decreased). This shows that for owners with education equal to or lower than secondary education level, the likelihood that they will fail is higher than for those with a tertiary or higher level education. We can also see that majority of firms have been operating for more than 2 years and the older the business the more profitable it is likely to be. In addition it is clear that most businesses using talking to customers and business associations as their main source of business relevant information are more profitable than those who are not.

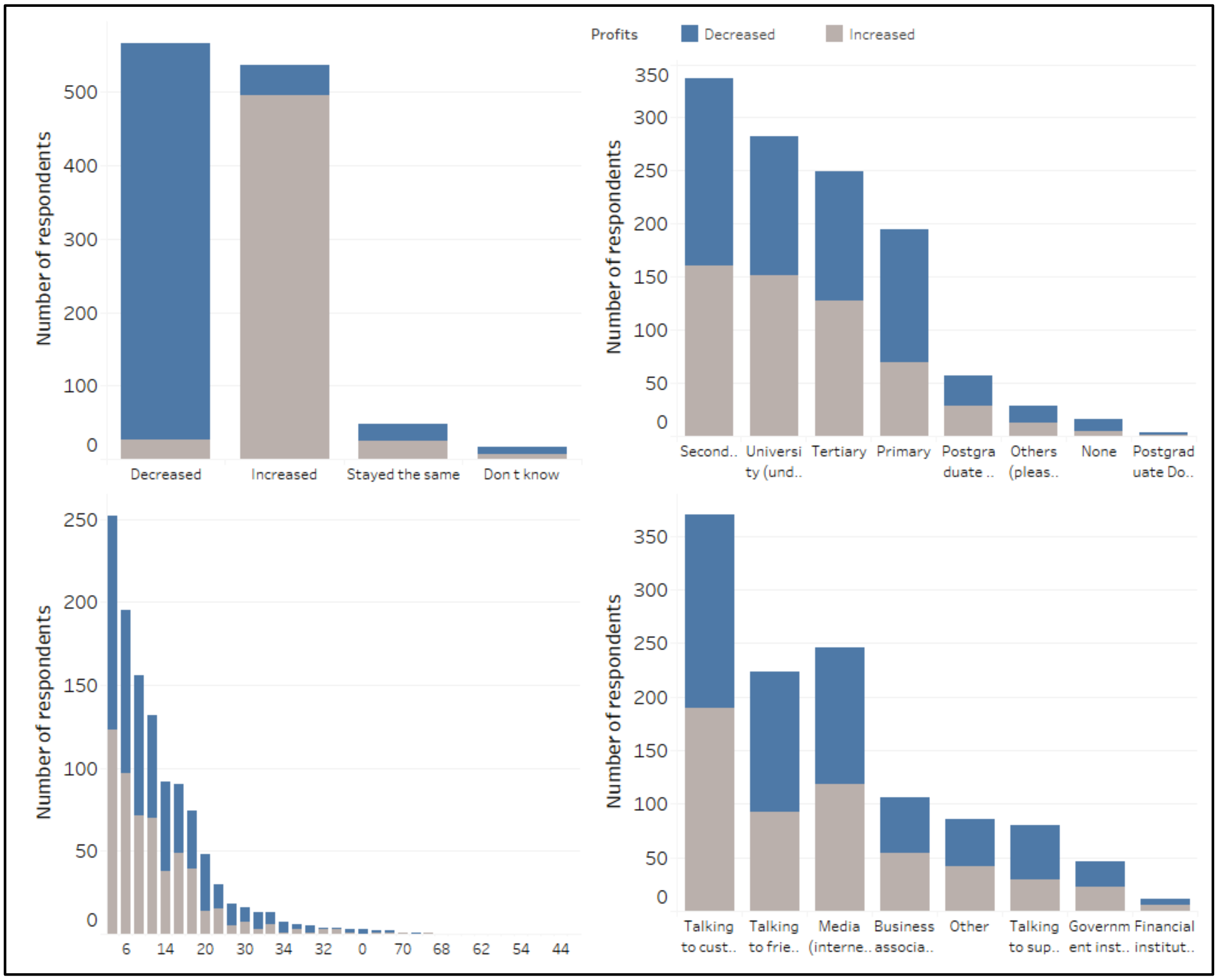


Figure 16: Analysis of the features in top ten feature importance of best models.

This study demonstrates that the application of annual survey data for MSMEs growth prediction is a very promising approach. Based on data from the annual small business survey, I was able to predict the profitability growth of MSMEs with an overall average accuracy of 91.7% i.e., the average of all the accuracies from the three implemented models. The traditional logistic regression model is the simplest model that was implemented on this data set with an accuracy of 90.88, therefore it provides a basis from which the performance of more advanced models can be evaluated. The results of the two more advanced models are considerably better with accuracies of 92.31% and 92.02 % for the XGB and Random forest algorithms, respectively. Given that the growth mechanism of MSMEs is highly multifaceted and that the constructed growth model is based on a single year of data, the results are encouraging both for further research and commercial implementation.

7 Conclusion

This research sets out to predict the profitability growth of MSMEs. MSMEs are more likely to be affected by political, economic or natural disasters such as the current Covid pandemic, than their larger counterparts hence stakeholders i.e. credit providers, entrepreneurs, venture capitalists, startups, policy makers and various small business owners can highly benefit from a quantified method, when it comes to making decisions in such high risk economic environments.

Given the importance of MSMEs for the economy and society, policy makers and researchers have made efforts to promote MSME growth and to improve overall economic performance. This has been more prevalent during the current Covid pandemic. Therefore, analyzing and predicting the growth of MSMEs is becoming an important area of research. There has been an increase in funding for surveys concerning MSMEs; however, the prospective of using the data collected from these surveys for the profit growth modeling of MSMEs in East Africa has not yet been thoroughly evaluated. The use of survey data offers many advantages as there is a huge and consistent amount of easily and publicly accessible survey data that can be obtained cost effectively and in large quantities.

The data used in this dissertation is based on a survey called the national small business survey. It is primarily conducted on an annual basis hence providing a consistent data input source for profitability growth models that can assess MSMEs on an industry wide scale and with up-to-date information allowing them to adjust to the current circumstances hence making them more relevant to stakeholders. The dataset is also fairly balanced as the businesses who made losses and those who made profits are comparable in number therefore there is no bias. This is eliminated the need for imbalanced data treatments. One disadvantage of using this data set in this dissertation is the sample size, which is small i.e., only 1839 respondents.

In total, three separate models are implemented i.e., logistic regression, random forest and extreme gradient boosting. Logistic regression is implemented for comparison reasons as it is one of the most common and traditional methods used in data science and research literature. It also enables us to set and to construct a benchmark for the subsequent models. Logistic regression compared to other two implemented models did not exhibit satisfactory predictive ability and had different relevant variables from the other models. Random forest exhibited better performance than the logistic regression model demonstrating that the random forest model has lower costs for misclassification of the profitable companies. Even though random forest is already an ensemble technique, the research has been expanded to utilizing extreme gradient boosting for its effectiveness and demonstrated performance in the latest research and competitions. As seen with its other applications in various projects, XGB attained the best outcomes among other models implemented. With an accuracy of 92.31%, a specificity of 95.95% and sensitivity of 88.76%, XGB is marginally superior to the random forest approach. The best performing models, XGB and random forest, graded the same variables as their relevant features, which are sales made last year, operation time and business owner education level.

There still a lot that needs to be done in this area as most surveys and research are centered on corporate businesses or credit risk models in the credit industry. In addition the data and record keeping in the MSME sector is very inadequate especially on the side of the business owners themselves. Most financial services are also unwilling to provide credit to these businesses as there is very little research in terms of financial feasibility and the costs of failure are usually felt in terms of high monetary and opportunity expenses. This research however provides a practical and quantified modeling process, to predict MSME profitability using machine learning methods and data from the annual business surveys, hence providing evidence of feasibility to various stakeholders.

8 References

[1] Uganda Investment Authority, International Monetary Fund, Imf, and Uganda Investment Authority, "SMEs Driving the Economy*," Uganda Investment Authority*. [Online]. Available: https://www.ugandainvest.go.ug/smes-driving-economy/. [Accessed: 05-May-2020].

[2] The Guardian. 2020. *Uganda Is A Land Of Entrepreneurs, But How Many Startups Survive?*. [online] Available at: <https://www.theguardian.com/global-development-professionals-network/2016/feb/16/uganda-is-a-land-of-entrepreneurs-but-how-many-startups-survive> [Accessed 06 May 2020].

[3] Nahamya, S., Ruffing, L., Ssendaula, G., Toure, D., Apire, R., Griffiths, F., Kisaame, J., Mbagut, H., Badagawa, G., Ruffing, L., Cannière, L., Osei-Yeboah, D., Thompson, J., Ocii, C. and Kasekende, L., "Proceedings Of The Symposium On Modalities For Financing Smes In Uganda". In: *United Nations Conference on Trade and Development*. New York And Geneva, United Nations, 2002. Accessed on: May. 05, 2020. [Online]. Available https://unctad.org/en/Docs/itetebmisc8_en.pdf

[4] H. Ooghe and S. D. Prijcker, "Failure processes and causes of company bankruptcy: a McKenzie, David & Sansone, Dario. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. Journal of Development Economics. 141. 10.1016/j.jdeveco.2019.07.002.

[5]"Micro-, Small and Medium-sized Enterprises Day 27 June", *Un.org*, 2020. [Online]. Available: https://www.un.org/en/events/smallbusinessday/background.shtml. [Accessed: 20-May- 2020].

[6] Ministry of Trade, Industry and Cooperatives, "Uganda Micro, Small and Medium Enterprise (MSME) Policy - Sustainable MSMEs for Wealth Creation and Socio-Economic Transformation", Ministry of Trade, Industry and Cooperatives, Kampala, 2020.

[7]"Small and Medium Enterprises", *Uganda Investment Authority*, 2020. [Online]. Available: https://www.ugandainvest.go.ug/sme/. [Accessed: 20- May- 2020].

[8] E. Turyahikayo, "Challenges Faced By Small And Medium Enterprises In Raising Finance In Uganda", Ph.D, Uganda Management Institute Kampala-Uganda, 2015.

[9] J. Brownlee, "4 Types of Classification Tasks in Machine Learning", *Machine Learning Mastery*, 2021. [Online]. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/. [Accessed: 18- March- 2021].

[10] McKenzie, David & Sansone, Dario. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. Journal of Development Economics. 141. 10.1016/j.jdeveco.2019.07.002.

[11] I. Afolabi T, C. Ifunaya, F. G Ojo and C. Moses, "A Model for Business Success Prediction using Machine Learning Algorithms", Department of Computer and Information sciences, Covenant University, 2019.

[12] C. ¨Unal, "Searching for a Unicorn: A Machine Learning Approach towards Startup Success Prediction", Masters, Humboldt-Universit¨at zu Berlin - School of Business and Economics - Institute for Statistics and Econometrics, 2019.

[13] A. Garrido, "What is the difference between Bagging and Boosting? ⋆ Quantdare", *Quantdare*, 2021. [Online]. Available: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/. [Accessed: 19- Mar- 2021].

[14] University of Regina DBD, “Overview of the KDD Process,” *KDD Process/Overview*. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html. [Accessed: 06-May-2020].

[15] Financial Sector Deepening Uganda, "NATIONAL SMALL BUSINESS SURVEY OF UGANDA REPORT", Financial Sector Deepening Uganda, Kampala, 2015.

[16] K. Grace-Martin, "Seven Ways to Make up Data: Common Methods to Imputing Missing Data - The Analysis Factor", *The Analysis Factor*, 2021. [Online]. Available: https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/. [Accessed: 19- Mar- 2021].

[17] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil and D. Turaga, "Learning Feature Engineering for Classification", in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017.

[18] R. Namatovu, S. Dawa, C. Katongole and F. Mulira, "Understanding Women Micro and Small Business Entrepreneurs in Uganda", Dakar, 2012.

[19] Y. Kinha, "Top 4 ways to encode categorical variables- Edvancer Eduventures", *Edvancer.in*, 2021. [Online]. Available: https://edvancer.in/encode-categorical-variables/. [Accessed: 19- Mar-2021].

[20] M. Guy, "Types of data measurement scales: nominal, ordinal, interval, and ratio", *My Market Research Methods*, 2020. [Online]. Available: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/#:~:text=Summary,the%20difference%20between%20each%20one. [Accessed: 21- Mar-2021].

[21] "When Chi-square Is Appropriate - Strengths/Weaknesses | Chi-Square Test for Goodness of Fit in a Plant Breeding Example - passel", *Passel2.unl.edu*, 2021. [Online]. Available: https://passel2.unl.edu/view/lesson/9beaa382bf7e/14. [Accessed: 21- Mar- 2021].

[22]"Chi-square Test of Independence", *Pythonfordatascience.org*, 2020. [Online]. Available: https://www.pythonfordatascience.org/chi-square-test-of-independence-python/. [Accessed: 01-Jul- 2020].

[23]"One-way ANOVA with Python", *Pythonfordatascience.org*, 2020. [Online]. Available: https://www.pythonfordatascience.org/anova-python/. [Accessed: 08- Jul- 2020].

[24] H. Hormozi, E. Hormozi and H. Nohooji, "The Classification of the Applicable Machine Learning Methods in Robot Manipulators", *International Journal of Machine Learning and Computing*, p. 560, 2012. Available: 10.7763/ijmlc.2012.v2.189 [Accessed 21 March 2021].

[25]"XGBoost Documentation – xgboost 1.2.0-SNAPSHOT documentation", Xgboost.readthedocs.io, 2020. [Online]. Available: https://xgboost.readthedocs.io/en/latest/ [Accessed: 10-Jul-2020].

[26] D. Elsinghorst, "Machine Learning Basics - Gradient Boosting & XGBoost", *Shirin's playgRound*, 2020. [Online]. Available: https://www.shirin-glander.de/2018/11/ml_basics_gbm/. [Accessed: 10- Jul- 2020].

[27]"XGBoost Algorithm: Long May She Reign!", *Morioh.com*, 2020. [Online]. Available: https://morioh.com/p/32c506939ad5. [Accessed: 10- Jul- 2020].

[28]"XGBoost Algorithm | XGBoost In Machine Learning", *Analytics Vidhya*, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/. [Accessed: 10- Jul- 2020].

[29] "Logistic Regression — ML Glossary documentation", *Ml-cheatsheet.readthedocs.io*, 2020. [Online]. Available: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html. [Accessed: 08- Aug- 2020].

[30] "How the random forest algorithm works in machine learning", *Dataaspirant*, 2020. [Online]. Available: https://dataaspirant.com/random-forest-algorithm-machine-learing/. [Accessed: 09- Aug- 2020].

[31] J. Brownlee, "Overfitting and Underfitting With Machine Learning Algorithms", *Machine Learning Mastery*, 2021. [Online]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/#:~:text=Overfitting%20refers%20to%20a%20model%20that%20models%20the%20training%20data%20too%20well.&text=This%20means%20that%20the%20noise,the%20models%20ability%20to%20generalize. [Accessed: 26- Mar- 2021].

[32] P. Worcester, "A Comparison of Grid Search and Randomized Search Using Scikit Learn", *Noteworthy - The journal blog*, 2019.

[33] "3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.24.2 documentation", *Scikit-learn.org*. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html. [Accessed: 16- May- 2021].

[34] J. Han, M. Kamber and J. Pei, *Data mining*, 3rd ed. Amsterdam: Elsevier/Morgan Kaufmann, 2012, pp. 370-371.

[35] A. Bhandari, "AUC-ROC Curve in Machine Learning Clearly Explained - Analytics Vidhya", *Analytics Vidhya*, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. [Accessed: 16- May- 2021].

[36] J. Brownlee, "What is a Confusion Matrix in Machine Learning", *Machine Learning Mastery*, 2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/. [Accessed: 30- Aug- 2020].

[37] Z. Jaadi, "Eveything you need to know about interpreting correlations", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8. [Accessed: 16- May- 2021].

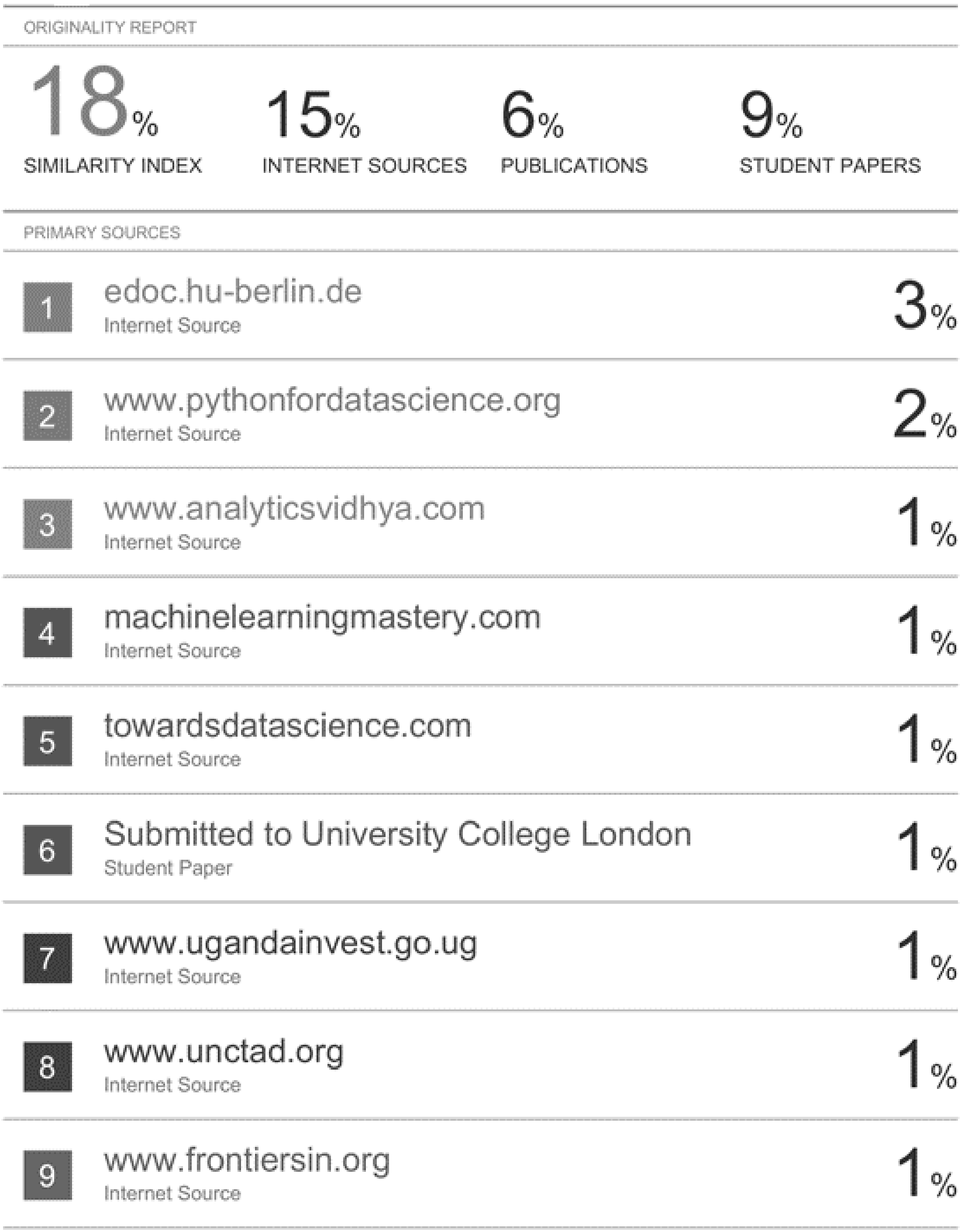
[38] J. Brownlee, "Feature Importance and Feature Selection With XGBoost in Python", *Machine Learning Mastery*, 2016. [Online]. Available: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/. [Accessed: 16- May- 2021].

[39] "Decision Trees and Random Forests", *Medium*, 2017. [Online]. Available: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991#:~:text=Decision%20trees%20are%20prone%20to,could%20lead%20to%20unsound%20conclusions. [Accessed: 16- May- 2021].

[40]"Selecting good features – Part III: random forests | Diving into data", *Blog.datadive.net*, 2020. [Online]. Available: https://blog.datadive.net/selecting-good-features-part-iii-random-forests/. [Accessed: 31- Aug- 2020].

[41] "What is Flask Python - Python Tutorial", *Pythonbasics.org*. [Online]. Available: https://pythonbasics.org/what-is-flask-python/#:~:text=Flask%20is%20a%20web%20framework,like%20url%20routing%2C%20template%20engine. [Accessed: 16- May- 2021].

9 Appendix



ORIGINALITY REPORT

18%	15%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	edoc.hu-berlin.de Internet Source	3%
2	www.pythonfordatascience.org Internet Source	2%
3	www.analyticsvidhya.com Internet Source	1%
4	machinelearningmastery.com Internet Source	1%
5	towardsdatascience.com Internet Source	1%
6	Submitted to University College London Student Paper	1%
7	www.ugandainvest.go.ug Internet Source	1%
8	www.unctad.org Internet Source	1%
9	www.frontiersin.org Internet Source	1%