COLLEGE OF BUSINESS AND ECONOMICS

MSc Research Thesis

Tree-based and Logistic Regression Models for Business Success Prediction in Rwanda

Thesis submitted in partial fulfilment of the requirements for Master of Science in Data Science (Actuarial Science) at the African Center of Excellence in Data Science in University of Rwanda.

By**: Francis Kipkogei** Registration Number **219013854**
Supervisor: **Dr. Kabano Ignace**

September 2020.

DECLARATION

I declare that this dissertation entitled Tree-based and Logistic Regression Models for Business Success Prediction in Rwanda 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.

Name Francis Kipkogei

Signature … ………

Approval sheet

This dissertation entitled … **Tree-based and Logistic Regression Models for Business Success Prediction in Rwanda…**written and submitted by …**KIPKOGEI FRANCIS**…. in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in…**Actuarial Science**…. is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 17% which is less than 20% accepted by ACE-DS.

____

Supervisor

____

Head of Training

LIST OF SYSMBOLS AND ACRONYM

AUC ROC-Area Under Receiver Operating Characteristics Curve

TP- True Positive

FP- False Positive

VAT: Value Added Tax

PIT: Personal Income Tax

CIT: Corporate Income Tax

CITVAT: Corporate Income Tax and Value Added Tax

WHT: Withholding Tax

PAYE: Pay as you earn.

PITVAT: Personal Income Tax and Value Added Tax

XGB-Extreme gradient boosting

Log loss- Logarithmic Loss

ß -Beta

ϕ-Theta

1 Contents

Acknowledgements

Completing this research study was the most exciting as well as challenging experience in of my education at African Center of Excellence in data Science. Therefore, it is vital to acknowledge that my experience could not have been completed without the unswerving support and guidance of the kind persons around me who in one way or another contributed unreservedly with their precious succour in the groundwork and completion of this study.

First and foremost, very special gratitude goes to my supervisor who has been mentor, Dr. Ignace Kabano the Head of Training, African Centre of Excellence in Data Science, College of Business and Economics/University of Rwanda, who is also my academic supervisor for his stimulating suggestions that was a driving force in starting and completing my research. In the end, his sound advice and professional guidance ensured that come up with new ideas to enhance efficiency during my study.
My special gratitude also goes to Dr. Javier Zazo who is postdoctoral fellow at Harvard University who also played a key role in my research through his guidance, reviewing my manuscript, unfailing support, constant motivation contributed to completion of my study. He further supported ideas I contributed towards making work more efficient.

Nevertheless, I want to express my deepest gratitude prof. Ernest Fokue who is a Professor of the RIT College of Science for reviewing my manuscript and his great ideas which contributed towards making my research more efficient.
Finally, I wish to acknowledge my colleagues' students, lecturers and staff of, African Centre of Excellence in Data Science, College of Business and Economics/University of Rwanda, for their support, encouragement, advise and support. Thi contributed much to success of my study.

Abstract

Background: Businesses have been touted to contribute immensely to economic health of most countries. Many enterprises are started every year, among these, some tend to be successful while others are unsuccessful. Studies are instrumental in giving a glimpse of the geographical locations outlook on the factors affecting success of business, data that relates to the whole nation and focusing on more determinants may give more insights on the challenges and better give a prediction of success of an enterprise given the factors. This study used Rwanda Revenue Authority data to identify important variables that contribute to business success in Rwanda. Tree-based models were compared with logistic regression for prediction of business success. The most robust model was used for business success prediction.

Methods: Statistical learning models consisting of tree-based models and logistic regression were trained and evaluated using a dataset obtained from Rwanda Revenue Authority over a sample of 18,162 businesses in Rwanda. Metrics such as recall score, F1 score precision score and accuracy were used in evaluating the performance of each model in differentiating between successful and failed business. Further discriminant analysis such ROC AUC was used to compare and evaluate the discrimination power of machine learning models.

Results: Tree-based ensemble models such as gradient boosting, XGBoost, and random forest were among the top classifiers which showed high predicted sensitivity and specificity. Gradient Boosting particularly correctly identified over 93% of business success. On the other hand, the lowest performing model was logistic regression with a recall score of 90% and F1 score of 90.6% on average. Sector was found to be most important feature contributing to business success.

Conclusion: Evidence from this study suggests that tree-based models can be utilized within the current care model to essentially produce greater prediction accuracy in the prediction of business success. This study further suggested a need to segment sector to identify other classes within the sector of economy that could contribute to success of business.

Key words: *Business success, unsuccessful, tree-based, models, logistic regression*

2 INTRODUCTION

2.1 Background of the study

Businesses especially small and medium enterprises (SMEs) have been touted to contribute immensely to economic health that is stability of the economy, its growth, and development. This is because they are usually the greatest contributors to employment, add value to primary production including agricultural produce, and assist in building a resilient economic system (Ayandibu & Houghton, 2017). As a result of their contribution, it has elicited interest from economic planners, researchers, and policymakers who have sought to find the outgrowth and development of the businesses considering various programs, strategies, and economic policies (Nagaya, 2017). Despite their contribution to economic health, businesses face some challenges which may determine their success.

Enterprises have at times faced various challenges such as limited access to finance, taxation, poor infrastructure, low level of societal trust, challenges with contract enforcement, and a weak education system. Some in developing countries including those in Rwanda have inadequate abilities to develop their workers' skills and have limitations to explore local economies of scale in terms of raw materials (Bayisenge et al., 2020). It was estimated that 40% are likely to be unsuccessful during their first year, 60% their second year, 90% are likely to be unsuccessful in the first ten years of business existence (Ramukumba, 2014). Additionally, closure rates of new businesses are significantly higher than existing ones, and rates of failure of small businesses are also higher than large businesses (Bartoloni et al., 2020). It was found that determinants that contribute to business success are size, location, and age of business (Aqeel et al., 2011). Also, the businesses which survive for a longer period are more likely to be successful. Furthermore, businesspersons working in the agricultural sector would have a longer predicted duration in the business, because both competing perils are lower (Van, 2003). Survival rates of businesses were found to be similar to location, employment size, business type, economic sector, and distance from the mall (Van, 2003)

2.2 Problem Statement

Business success prediction of a venture has been a struggle for both researchers and practitioners. Nevertheless, some companies aggregate data about other firms thus making it possible to create predictive models and validate them based on an unprecedented amount of real-world examples (Żbikowski et al., 2021). This study sought to establish determinants of business success using data from Rwanda Revenue Authority using machine learning models, particularly logistic regression, and tree-based models. The study could be utilized by entrepreneurs, the government, Rwanda Revenue Authority, and even researchers to make more informed decisions in the everyday business of living as well as further research on the same area. Determinants identified could enable the government to plan and come up with better strategic policies that could promote enterprise activities hence reducing situations that may lead the business to be unsuccessful. The study examined the main challenges and successes faced by all kinds of businesses; small, medium, and large businesses in Rwanda. This study builds on the previous studies. While they shed light on this study, they had gaps that this study seeks to fill. Firstly, they focused on small and medium-scale businesses only. For instance, (Mutandwa et al., 2015) focused on determinants of enterprise performance of small and medium scale businesses in Rwanda. However, identifying variables that correlate with specific practices in successful businesses are also informative regardless of size and have an objective of growth in the future which authors sought to find out. This study will investigate determinants that can lead to all scales of business success. Moreover, machine learning has models that have been touted to enrich the insights hitherto foreseen or found with traditional models as machine learning and could uncover hidden patterns in data (Żbikowski et al., 2021). This study was motivated by (Gepp et al., 2010) who argued that accurate business failure (unsuccess) prediction models would be tremendously valuable to various industry sectors. (Gepp et al., 2010) also found out that decision trees which are part of tree-based models perform well in predicting business failure. However, boosted trees provide outstanding predictive performance for various tasks. Boosted trees have also been depicted to be among the superlative performing learning techniques based on public data evaluations (Ganjisaffar et al., 2011). (Zeng, 2017) suggested that using boosting to choose relevant predictors is a viable and competitive approach in predicting an aggregate. It was also found that all the top teams ranked by the Yahoo learning challenge all utilized tree-based ensemble methods. (Ganjisaffar et al., 2011). Therefore, this study takes advantage of boosted trees to have a better predictive performance for business success prediction.

This study also attempts to look at all lines of business such as startups and those which have been in the industry for a long time and taking advantage of the technological advances in machine learning and computational capabilities. This study also employs tree-based models which are gaining fame in areas such as artificial intelligence, medicine, and pattern recognition (Clark et al., 2017). Logistic regression which handles binary data in various fields is among the top in terms of computational speed and prediction accuracy (Zhu et al., 2003). Therefore, this study will compare the tree-based models and logistic regression which have not been harnessed to predict business success in Rwanda.

2.3 Objectives

2.3.1 General Objective

To find out the determinants of business success in Rwanda.

2.3.2 Specific Objectives

1. To find out the survival rates of businesses over time in Rwanda.
2. To compare various machine learning models and utilize the most robust model to predict business success in Rwanda.
3. To find the correlates associated with business success in Rwanda.
4. To offer strategic recommendations for business success in Rwanda.

2.3.3 Research Questions

1. What are the survival rates of businesses over time in Rwanda?
2. Which machine learning model has most robust in predicting business success in Rwanda?
3. Which are the correlates associated with business success in Rwanda?
4. What are strategic recommendations for business success in Rwanda?

2.3.4 Significance of the Study

This study will not only be of an imperative to the domain of academia but also to the investors and young entrepreneurs. On academic domain it will shape on what others have begun, using a new method. To the investors it will it will shade light on factors to critically investigate when determining factors to consider before venturing into business. To businesspersons in the industry it will give valuable determinants of business success in data thus higher return for cash outlay. To financial managers, it will be open on new insight into predicting performance of the firms they manage. Finally, to regulators it will give basis for determining the levels business risks by the various sectors in economy.

2.3.5 Justification

This study will aid in predicting the business success determinants.
Harnessing tree-based machine learning algorithms will shed light into the models that the various variables tend to follow.

2.3.6 Scope

This study will be within financial area; and will focus on the determining factors associated with the business success. The data will be obtained from RRA Rwanda. The methodology will be delimited to quantitative paradigm where modeling shall be done using tree-based and logistic regression machine learning algorithms. Table 1 below shows a summary of the research scope, in view of the objectives, areas of application, methods/techniques and algorithms which were applied.

Table 2: Scope of this study

Objective	Area	Method	Algorithm
Survival rates	Business success	Frequency distribution table	Crosstabulation
Survival rates	Survival of business over time	Survival curves	Kaplan Meier survival curves plot
Prediction	Business success	Supervised	logistic regression, decision tree, random forest, gradient boost and XGBoost.
Correlates	Determinants of business success	Feature importance	Most robust model amongst logistic regression, decision tree, random forest, gradient boost and XGBoost.

Table 3 depicts methods and all the algorithms that have been utilized in this study. These algorithms were used to carry on analysis and comparative evaluation between different models as well as predicting correlates that contribute to business success.

3 LITERATURE REVIEW

3.1 Introduction to the Review

This segment gives the state of earth through eyes of previous studies that this study builds on. It also reviews the various concepts under study and highlights on the gap in past studies that this study seeks to fill. It further lays the theoretical framework upon which this study is built on. Finally, there is a conclusion.

3.2 Previous studies in the area under study

Survival Rates of business success

The business which survives for a longer period are more likely to be successful. Furthermore, businesspersons working in the agricultural sector would have longer predicted duration in the business, because both competing perils are lower (Van Praag, 2003). Survival rates of businesses were found to be similar with respect to location, employment size, business type, economic sector and distance from the mall (Van Praag, 2003).

New businesses are alleged to have high rates of closure and these cessations are assumed to be failures, but then again two U.S. Census Bureau data bases elucidate that these suppositions may not be vindicated. The Business Information Tracking Series (BITS) depicted that around half of new employer businesses survive past four years and the Characteristics of Business Owners (CBO) depicted that around a third of unsuccessful businesses were successful at cessation (Headd, 2003)

Closure rates of new businesses are significantly higher than existing ones, and rates of small businesses being unsuccessful are also higher than those of large businesses (Miles, 2013). One of the major determinants of business failure or success rate is the age dissemination of the population of business and that failure rate has lesser counter-cyclical fluctuation (LANE & SCHARY, 1991).

Business failure rates (non- success rates) are higher for small business in wealthier countries, younger business, retail business, less fecund and less money-spinning business (McKenzie & Paffhausen, 2019).

Correlates associated with business success

The significant proportion of businesses that closed while successful calls into question the use of "business closure" as a meaningful measure of business outcome. It appears that many owners may have executed a planned exit strategy, closed a business without excess debt, sold a viable business, or retired from the work force. It is also worth noting that such inborn factors as race and gender played negligible roles in determining survivability and success at closure (Siow Song Teng, Singh Bhatia, & Anwar, 2011).

Assistance by the government plays an important contributor for the success of business especially the small-scale ones. On contrary, other scholars found that assistance by the government was not a prominent to success of the business. However, determinants that contribute to business success are size, location and age of business (Aqeel-Riaz et al., 2011).

Business failure or lack of success can be described as the business' discontinuance due to losses to creditors and shareholders .This infer that business success can be attributed to the perpetuation of business without losses (or making profits) to be factored in as a form of success s in business (Theng & Boon, 1996).

Business can be deliberated as successful if it has made an average profits for at least three preceding years in the industry and can be a failed business it has not been making profit in the preceding three years (Lussier & Pfeifer, 2001).

Assistance by the government plays an important contributor for the success of business especially the small-scale kinds of businesses. On contrary, other scholars found that assistance by the government was not a prominent to success of the business. However, determinants that contribute to business success are size, location and age of business(Aqeel et al., 2011).

Businesses having larger capitals with better funding and having workforces were found to have higher likelihoods of survival. Factors that could lead to failure of business include having little or no start-up capital and having a moderately young owner (Headd, 2003).

3.3 Gap in the past studies

This study builds on the previous studies. While they shed light on this study, they had gaps that this study seeks to fill. Firstly, they focused on small and medium scale businesses only. For instance, (Miles, 2013) explored factors needed for small business to succeed and (Mutandwa et al., 2015) focused on determinants of enterprise performance of small and medium scale businesses in Rwanda. Also (Feindt, Jeffcoate, & Chappell, 2002) investigate factors that lead to success and fast growth in SME E-commerce. Determinants contributing to success of all enterprises irrespective of size were left instead specific scale of business were explored. Identifying correlates that can lead to success of all scales of business is also informative as most businesses regardless of size have an objective of growth in future. This study will investigate determinants that can lead to all scales of business success.

Whereas various challenges have been looked at, mostly it has been surveys of individual business owners and may have focused on a specific geographical location within the country. For instance (Sibomana & Shukla, 2016) focused on two factors, in a single district. Whereas such studies are instrumental in giving a glimpse of the national outlook on the factors affecting success of business, data that relates to the whole nation and focusing on more determinants may give more insights on the challenges and better give a prediction of success of an enterprise given the factors. Moreover, machine learning has models that have been touted to enrich the insights hitherto foreseen or found with traditional models as machine learning could uncover hidden patterns in data (Gupta et al., 2016).

Moreover, they look at specific lines as for case (Saura, Palos-Sanchez, & Grilo, 2019) who detected main factors for startup business success may be less informative when generalized to other lines of business. This study attempts to sidestep this myopic view to look at all lines of business taking advantage of the technological advances in machine learning and computational capabilities. This study also employs tree-based models which are gaining fame in areas such as artificial intelligence, medicine and pattern recognition (Clark & Pregibon, 2017). However, the tree-based models and logistic regression have not been harnessed to predict business success.

3.4 Theoretical Framework

This study is based on (Asadi, Lin, & De Vries, 2014) who demonstrated that tree-based algorithms are efficient and effective in solving web ranking along with other problems in various spheres. Predictions can be made by augmenting the runtime performance of applying tree-based models given an already tuned model. Even though remarkably simple conceptually, various applications of tree-based machine learning models do not utilize current architectures of superscalar processor efficiently. The tree-based models have been proven to be more efficient in tackling problems in various fields such as computer vision, online advertising, genomic analysis and medical diagnosis.

According to (Ganjisaffar et al., 2011) boosted trees provides outstanding predictive performance for various tasks. Boosted trees have also depicted to be among the superlative performing learning techniques based on public data evaluations. (Ganjisaffar et al., 2011) further found out that all the top teams ranked by Yahoo learning challenge all utilized tree-based ensemble methods. (Ganjisaffar et al., 2011) applied LambdaMART which ranked algorithm by utilizing Gradient boosting to improve ranking cost function. Gradient boosting which is formed by combining many weak learning models together in order to form stronger predictive models was found to lessen bias by increasing the communicative power of the base learner and by compelling learning to attend to tuning cases that constantly are mis projected. Further, boosting aggregates the projections of manifold trees thus minimizing variance, however boosted trees are very powerful thus regularization usually is required to avert overfitting. (Ganjisaffar et al., 2011) combined bagging and boosting in order to boost learning-to-rank and also depicted that bagging boosted ensembles which have been slightly overfitted to their training set produces better results.

(Schapire, 2003) depicted that logistic regression and boosting can solve similar constrained optimization problem, apart from that in boosting, some normalization constrictions have been dropped. (Schapire, 2003) found that logistic regression could handle loss-minimization problem effectively and create a better understanding of boosting and donated to its extension in a more practical way. Thus, employing machine learning models to predict business success is vital since business relies on successful projection. Tree-based machine learning models which have depicted outstanding predictive performance were compared with logistic regression to predict determinants of business success.

3.5 Conceptual Framework on Determinants of Business Success

It is essential to have foundation for decision on variables to predict success of business. These will be the input features with much weight, that this study seeks to find. This study will infer a function Y = F(X) which utilized to map the inputs X to outputs Y (Shickel, Tighe, Bihorac, & Rashidi, 2018). That can be explained as supervised learners drawing insights and infer patterns from the labeled data so that they can learn from preceding examples to make reasonable predictions about new ones. In order to realize it, this study employed (Van Praag, 2003) approach in order to answer the vital research objective. This study was conceived to point out determinants of business success. According to (Van Praag, 2003), determinants of business success include; location of the business (urban and rural), size of the business(large, medium, small and micro), sector of the business (industry, service and agriculture), age of the business (duration) , income of the

business(high, medium and low), marital status of owner(married, single, widowed and divorced) and gender of the business owner(male, female and other).
(Van Praag, 2003)

3.6 Conceptual Model of the Study

The conceptual model of this research is principally based on (Van Praag, 2003), nonetheless some features were not analyzed since they were lacking from data obtained from RRA. The variables influence the survival and success of businesses in Rwanda. Some other variables were included while others were eliminated during feature selection.

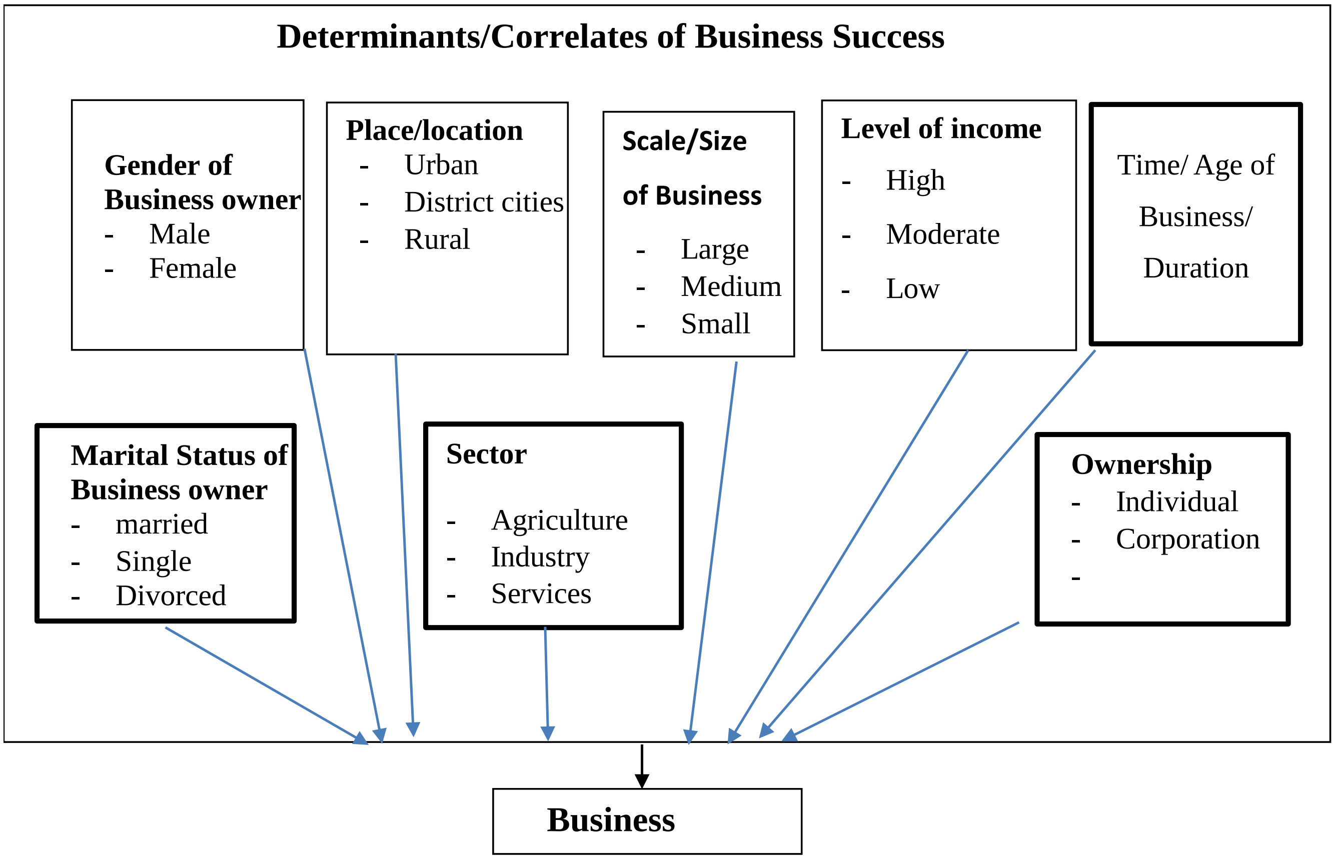


Figure 2 1: Conceptual Framework of Determinants of Business Success; Source (Van Praag, 2003)

4 METHODOLOGY

4.1 Introduction

Methodology utilized in this study include data description and its source, data description and evaluation metrics that would be used to compare tree-based machine learning algorithms. The general illustration of and process followed to predict business success is captured in Figure 1.1.

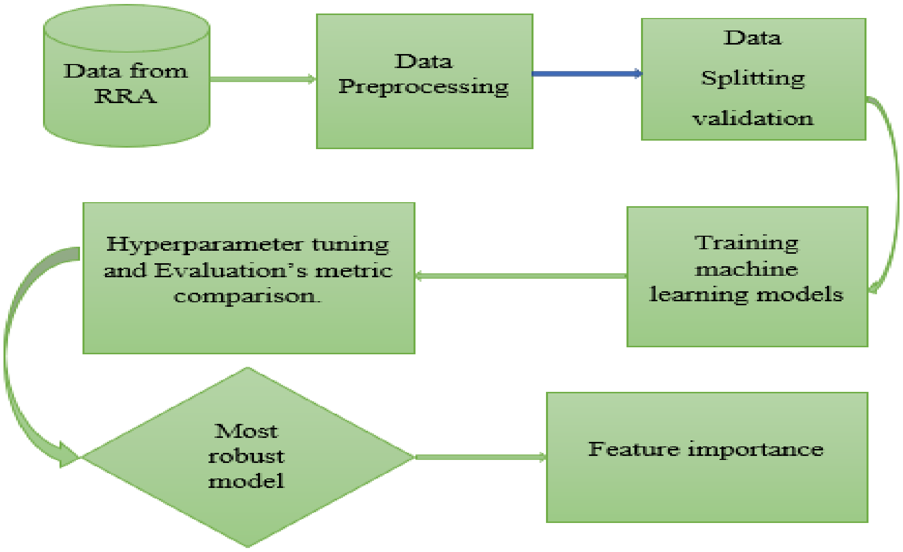

Data from RRA
Data Preprocessing
Data Splitting validation
Training machine learning models
Hyperparameter tuning and Evaluation's metric comparison.
Most robust model
Feature importance

Figure 1.1: General illustration of process followed to predict business success.

4.2 Data

The data for this study was obtained from Rwanda Revenue Authority, the government organization tasked with revenue collection. The original de-identified dataset consisted of 205,245 businesses between 1996 and 2020. Identifiable information such as business names, phone numbers, address and tin number details were not included. However, the authors established a few eligibility criteria that were in line with the prime objective of the study to enhance quality of data. The first exclusion criteria that ensured that the extracted dataset had the essential attributes required to accomplish the objective of the study was registration status. All businesses which were not deregistered due to continuous losses incurred by the owner were excluded from the dataset. During the merging process of different dataset, other reasons for exclusions were due to missingness of important features; further reducing the sample size to 18,162 businesses. Total postprocessed data consisted of 18,162 businesses, from which 13,565 were successful, and 4,597 were unsuccessful. The businesses which were deregistered or closed due to continuous losses incurred by the owner were extracted and categorized as unsuccessful businesses while registered businesses which are still operating and making profit were categorized as successful businesses. Each business owner had a unique TIN (taxpayer identification number). The response variable is registration status and it has two levels (Yes or No). The background characteristics are description, tax types, place, the scale, department, sector, level of income, duration, fraud status and origin. The registration status indicates that the business is still registered or deregistered due to losses. Description indicates whether the business is owned by an individual or a corporate business. Tax type indicates the types of taxes that business has been audited on, this include; value added tax, pay as you earn, withholding, custom taxes and others. Place indicates location where business is operating, this includes rural, urban and district cities. Scale indicates the size of the business, that is large, medium, small and micro. Department indicates whether the business is dealing with domestic products or imports, and exports. Sector indicates the sector of the business in the economy classified into three: agricultural sector, industrial sector and service sector. Level of income indicates whether the business is making high, moderate or low profit. Duration indicates the time difference between the time a business was registered and the time of this study for registered businesses while for deregistered (unsuccessful) businesses is time until deregistration. Fraud Status shows whether the business has committed tax fraud or not. Origin indicates whether the business owner is from Rwanda or from any other country.

4.3 Data Preprocessing

Data was obtained in excel format; however, messiness, duplicates, noise and outliers were prevalent in the data. Thus, data was cleaned by handling missing values, duplicates and noisy data, further data splitting was carried out in order to ease analysis, reduce misclassification and ensure improved model accuracy. Moreover, features that did not help improve results were removed. By removing them, it led to getting better results and made the data learning task less computationally expensive.

Missing data could have made training inconsistent. Imputation and reduced-feature models were used to solve the problems of missing data before training. Some data attributes could be redundant in the sense that their values can be obtained from other attributes. These were also reduced before training began. Data was split into 80% training set, 10% test set and 10% validation set after it was cleaned.

4.4 Models

After data cleaning, the data was split into training, test and validation sets before classification models were applied. The validation set helped in parameter tuning. This paper used supervised machine learning technique (decision trees classifier, random forest classifier, gradient boost classifier, XGBoost classifier and logistic regression classifier) to find out the determinants of business success in Rwanda.

4.4.1 Logistic classification algorithm

Logistic regression is a binary classification procedure in which a linear boundary is optimized to separate the input classes (Casella, Fienberg, & Olkin, 2013). However, it introduces nonlinearity logistic function over the linear classifier and the output is usually binary in nature. Logistic regression was suitable for the classification task at hand for the binary nature of success of business. Where the linear classifier is defined as:

$$\boldsymbol{g(x) = w^T x + b} \qquad 1$$

In case a straight line is fit to a binary label that is coded as 0 or 1, in this situation the predictions g(x)< 0 for some values of X and g(x)> 1 for others (except the range of x is limited) (Casella et al., 2013). If logistic regression is used to predict business success, logistic function is modelled such that predicted probabilities fall between 0 and 1.

The logistic classifier using a sigmoid function $\varphi()$ could be defined as:

$$\boldsymbol{\varphi\big(g(x_i)\big)\begin{cases}\geq 0.5 Y_i = 1\\ \;0.5 Y_i = 0\end{cases}} \qquad 2$$

The logistic function used in Type equation here.logistic regression can be written as;

$$\boldsymbol{\varphi\big(g(x)\big)} = \frac{\boldsymbol{e^{-g(x)}}}{\boldsymbol{1+e^{-g(x)}}} \qquad \textit{(Casella et al., 2013)}$$

4.5 Tree based Models.

Tree-based methods involve segmenting or stratifying the predictor space into several simple regions. To make projection for a given observation, the mode or the mean of the training observations are used in the region to which it belongs. Subsequently the set of splitting rules utilized in segmenting the predictor space can be abridged in a tree (Casella et al., 2013). Examples of tree-based machine learning models are decision trees, random forest, gradient boosting, and XGBoost (Dangeti, 2017).

4.5.1 Decision tree

Decision trees dispense data to predefined classification groups in the case of business success prediction, a decision tree usually dispenses each business to a successful or unsuccessful group. In general, decision trees are binary trees, which consist of a root node, non-leaf nodes and leaf nodes connected by branches (Gepp et al., 2010). Applying decision trees to classification problems like business success prediction, leaf nodes denote classification groups (successful or unsuccessful) and the non-leaf nodes each comprise a decision rule. Therefore, the tree is constructed through a recursive process where data is split when moving from a higher level of the tree to a lower level (Gepp et al., 2010). Using multiple input variables in a dataset the decision tree method enhances prediction of a value in the response variable (Dangeti, 2017).

4.5.2 Ensemble methods

Ensemble methods can be categorized into bagging and boosting techniques. In bagging, also referred to as bootstrap aggregating, multiple independent classifiers are trained and an aggregate result is reported (through a majority vote, for example (Dangeti, 2017). These multiple classifiers will be aggregated which aids to decrease variance. The models in bagging are built independently or rather in a parallel. Boosting on the other hand, trains simple classifiers on the input, and then improves the result by training subsequent models on the output. Subsequent models improve the model's performance (Dangeti, 2017). It conglomerates a set of weak learners and brings out better-quality prediction accuracy. It plays an important role in dealing with variance trade-off and biases. Some examples of boosting are gradient boosting and XGBoost.

4.5.3 Random forests

Random forest is formed by combining tree predictors such that each tree hinges on the values of the random vector sampled autonomously and with the similar distribution for entire trees in the forest (Breiman, 2001). Random forest utilizes an ensemble of decision trees to predict a response variable (for example business success) based on input features (input features, in our case sector, level of income, scale of business, place, tax type, ownership, duration and fraud status). The forecast is the result of successive, binary decisions that are orthogonal splitting in the multivariate space of features (Avanzi et al., 2019). Random forest handles classification problems, by determining individual tree forecasts and taking the response category which occur most frequently in the similar terminal node as the test case being forecasted (Sage, Genschel, & Nettleton, 2020)

4.5.4 Gradient Boosting

Gradient boosting is a machine learning technique containing optimized boosted decision trees which is formed by combining many weak learning models together in order to give stronger predictive models. Components in gradient boosting systems are summarized into two essential parts: that weak learner component and the additive component. The series of tweaks formed by weak learning algorithms boost the strength of the learner (Dangeti, 2017). Gradient boosting plays a significant role in minimizing loss or rather the variance between the value in the actual class of the training data set and the value of the predicted class. Minimizing errors in gradient boosting is achieved by taking calculated loss and carrying out gradient descent and thereafter the tree parameters are adjusted to minimize the residual loss. When adding a new weak learner to the model, the previous learners' weight is cemented in place or left unaffected on introduction of new layers. The ultimate ensemble model predictions are obtained from the averaging the predictions made by the prior tree models. The gradient boosting detects the faults of weak learners by means of gradients in the loss function (Dangeti, 2017). The loss function evaluates how good the coefficients of a model at fitting the given data are.

4.5.5 Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized distributed gradient boosted decision tree which is more efficient and portable due to its speed and performance that is dominative competitive machine learning. Machine learning algorithms under XGBoost are therefore implemented under the framework of gradient boosting (Dangeti, 2017). It has high scalability and is fast to execute this archetypally outperforming other algorithms. XGBoost model performance can be improved by hyper-parameter tuning which involves selection of data patterns and regularities by tuning thousands of what is known as "learnable" parameters automatically. It has regularization which aids in reducing overfitting (Dangeti, 2017).

4.6 Evaluation Metrics

Metric was used to evaluate on a test set where accuracy score, log loss, area under ROC curve F1_Score were utilized. Metrics play a significant role in optimizing the models, quantifying their performances as well as comparing them and improving their efficiency (Flach, 2003).

4.6.1 Binary Cross-Entropy (Log Loss)

This is used to quantify the performance of classification models by evaluating how good or bad are probabilities predicted from a given model. If the predictions are bad the log loss will return high values but when predictions are good, then log loss will return low values (Kull, Silva Filho, & Flach, 2017). As values returned by log loss tend to zero then it will result in low uncertainty and thus the better the model.

$\widetilde{Y_j}=\frac{1}{k}\sum_1^k Y_j \,.\, log\left(p(Y_j)\right)+(1-Y_j).\, log\left(1-p(Y_j)\right)$

$\widetilde{Y_j}$ Denotes Log Loss function

$p(Y_j)$ This is the predicted probability of the point being in the positive class for all k points.

Y_j Represents response variable.

4.6.2 Confusion matrix

Confusion matrix is a contingency table of actual class compared to model predictions.

True Positive (TP): Is when predicted values as positive and turns out to be true. For instance, the number of cases correctly identified that business will succeed.

False positive (FP) is when values predicted as positive and turns out to be false. For instance, the number of cases incorrectly identified that business will succeed.

False Negative (FN) is when values predicted as negative and turns out to be false. This is the number of cases incorrectly identified that business will be unsuccessful.

True Negative (TN): Predicted values as instance, the number of cases correctly identified that business will be unsuccessful.

4.6.3 Accuracy Score

Accuracy is the ratio of observations which are correctly predicted to the total observations.

Accuracy=$\frac{TP+TN}{TP+FP+FN+TN}$ (Dangeti, 2017)

4.6.4 Recall Score (Sensitivity)

Recall or sensitivity is the ratio of correctly predicted positive values to the all values in true class.

Recall =$\frac{TP}{TP+FN}$ (Dangeti, 2017)

4.6.5 Precision Score

Precision is the ratio of observations which are correctly predicted positive to the total number of observations predicted as positive.

Precision =$\frac{TP}{TP+FP}$ (Dangeti, 2017)

4.6.6 F1 Score

F1 Score is the weighted average of recall and precision score.

F1 Score=$\frac{2*Recall*Precision}{Recall+Precision}$ (Dangeti, 2017)

4.7 Discrimination Analysis

Evaluating the discriminative ability of any classification model is vital, to identify how cases with and without the outcome are separated (Steyerberg et al., 2010). An example is Area Under the Receiver Operating Characteristic Curve.

4.7.1 Area Under the Receiver Operating Characteristic Curve (ROC AUC)

The ROC AUC is used to check performance of classification problems at different thresholds settings. ROC AUC is used to compare and evaluate the discrimination power of machine learning models (Rodriguez & Rodriguez, 2006). ROC denotes probability curve on other hand AUC signifies degree of separability. It shows the extent of model capability to distinguish between classes. Models used were good since their AUC tends to 1 and so are their separability measure (Bowers & Zhou, 2019) .
The ROC AUC is used to check performance of classification problems at different thresholds settings. ROC denotes probability curve on other hand AUC signifies degree of separability. It shows the extent of model capability to distinguish between classes. Models used were good since their AUC tends to 1 and so are their separability measure (Bowers & Zhou, 2019). A model is good when AUC tends to 1 and so its separability measure is also good. When AUC tends to 0 then model is poorer than a random guess, and its separability measure is very poor. If AUC is 0.5, then the model has no class separation capability that is it   discrimination capacity to differentiate between the positive class and the negative class. Separability measure approximates the average number of cases in a dataset having nearby neighbor with the same response class. If AUC is almost 0, then model is said to reciprocate the classes by predicting positive class as a negative class and negative class as a positive class(Lenz, 2010). AUC ROC the plot of the specificity which is the proportion of the correctly classified negatives against the sensitivity which is the proportion of the correctly classified positives (Bhattarai, Shrestha, & Sapkota, 2019).

Figure 1:Area Under the Receiver Operating Characteristic Curve

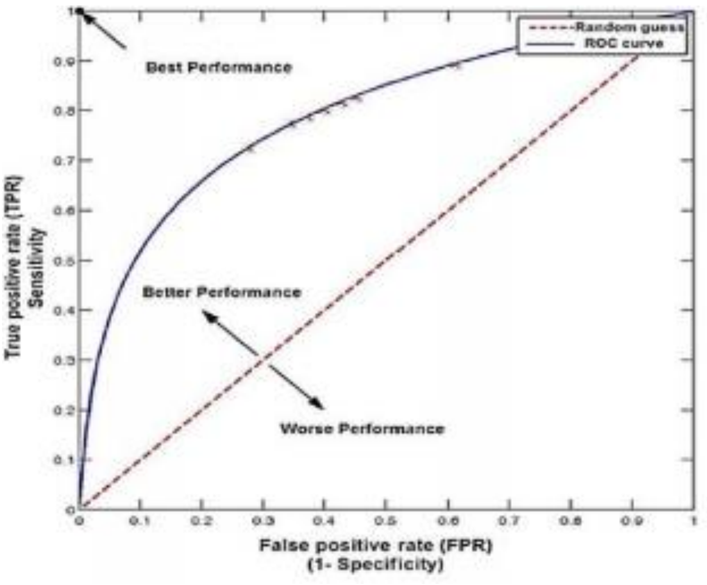


Figure 2 : ROC AUC (Bhattarai et al., 2019)

4.8 Calibration

(Recchioni, Tedeschi, & Gallegati, 2015) used calibration procedure to validate agent-based models . (Recchioni et al., 2015) also found out that an appropriate calibration aids the model in describing the predictor variables.

Hyperparameter Tuning

Selecting optimal hyperparameter values for models is a vital step in model design. This is done by lessening either generalization error estimate or some other related evaluation metric. (Duan, Keerthi, & Poo, 2003). Range of conceivable values for all hyperparameters was defined. Grid-search was utilized to find the best set of hyperparameters from subset which was chosen manually of the hyperparameter space of each model. Grid search was performed over the defined hyperparameters for logistic regression, decision tree, random forest, gradient boost and XGboost in order to yield optimal values. Optimized hyperparameters boosted the performance of each model. The hyperparameters with highest recall score, accuracy and F1 score was selected. However, in a situation where hyperparameters yields same results in terms of recall score, accuracy and F1 score then those with lowest runtime were chosen. Tables of hyperparameter tuning are shown in Appendix B.

5 DESCRIPTIVE ANALYSIS

Table 4 : Frequency distribution of all Input Features used in the Analysis and their Relationship with Registration Status of Businesses.

		Registration Status	
Classification of Business	**Frequency**	**Yes**	**No**

Sector			
1=Agriculture	180	168 (93.3%)	12 (6.7%)
2=Industry	4776	1068(22.4%)	3708 (77.6%)
3=Services	12915	12116 (93.8%)	799 (6.2%)
4=Others	291	213 (73.2%)	78 (26.8%)
Tax type			
1=VAT	2481	2371 (95.6%)	110 (4.4%)
2=CIT	456	379 (83.1%)	77 (16. 9%)
3=PIT	86	66 (76.7%)	20 (23.3%)
4=PAYE	60	57 (95%)	3 (5%)
5=WHT	1,222	1221 (99.9%)	1 (0.1%)
6=CITVAT	34	29(85.3%)	4 (14.7%)
7=PITVAT	34	28 (82.4%)	6 (17.6%)
8=OTHERS	1384	1371 (99.1%)	13 (0.9%)
9=CUSTOMS	12,405	8043 (64.8%)	4362 (35.2%)
Place			
1=Urban	12,928	9956 (77%)	2972 (23%)
2=District cities	2,730	1968 (72.1%)	762 (27.9%)
3=Rural	2,504	1641 (65.5%)	863 (34.5%)
Fraud Status			
1=Yes (Committed Fraud)	5347	4312 (80.6%)	1044 (19.4%)
0=No (Not Committed Fraud)	12815	9253 (72.2%)	3562 (27.8%)
Scale of Business			
1=Large	223	223 (100%)	0 (0%)
2=Medium	391	379 (96.9%)	12 (3.1%)
3=Small	14,493	10,980 (75.8%)	3513 (24.2%)
4=Micro	4,154	1983 (64.9%)	1072 (35.1%)
Description			
1=Individual	9437	5671 (60.1%)	3766 (39.9%)
2=Corporation	8725	7894 (90.5%)	831 (9.5%)
Origin			
1=National	18,114	13534 (74.72%)	4580 (25.28%)
2=International	48	31 (64.58%)	17 (35.42%)
Department			
1=Domestic	2108	1873(88.9%)	235 (11.1%)
2=Customs	16054	11692(72.8%)	4362 (27.2%)
Level of income			
1=High income	223	223 (100%)	0 (0%)

2=Moderate income	391	379(96.9%)	12 (3.1%)
3=Low income	18,162	12963 (74.7%)	4585 (25.3%)

From Table 5 above businesses under sector it is evident that most businesses operating in industry are more likely to fail with 77.6% chance of succeeding while those in the Services sector are most likely to succeed (96.8%), followed by those in the Agriculture sector (96,28%). Businesses classified by tax type depicts that those associated with customs are more likely to fail (35.2%) while businesses associated with WHT tax type had the highest chance of succeeding (99.9%). Businesses grouped by place shows that those operating in the urban areas had highest chance to succeed (77%), followed by those in district cities (72.1%) while those in the rural areas were least to succeed (65.5%). Businesses grouped by fraud status depicts that those which have committed fraud had higher chance (80.6%) to succeed followed by those which have not committed fraud (72.2%). This need to be investigated to find the nature and extent of fraud or may be fraud status may contribute less to success of the business. Classification based on business scale depicts those large-scale businesses had highest chance to succeed (100%) followed by medium scale (96.9%), small scale (75.8%) while micro scale had least chance to succeed (64.9%). Businesses grouped by description depicts that, businesses owned by corporation had higher chance of success (90.5%) compared to those owned by individuals (60.1%). Businesses grouped by region shows that, businesses owned by nationalists had higher chance to succeed (74.72%) compared to those owned by foreigners (international) (64.8%). Businesses grouped by level of income depicts that those earning high level of income had highest chance of succeeding (100%) followed by those earning moderate income (96.9%) and least to succeed were those earning low level of income (74.7%).

The most successful businesses are those in service sector since it comprise of many businesses and it is also the main sector of economy. Businesses in service sector include financial and insurance activities, real Estate Activities, Professional, Wholesale and Retail Trade, Scientific and Technical Activities, Information and Communication, Accommodation and Food Service, Human Health and Social Work, Education, Activities of Extraterritorial Organizations and Bodies, among others. It is followed by those in agricultural sector. The businesses in the industrial sector are least performing, these include manufacturing, Mining and Quarrying, Electricity, Gas and Air Conditioning Supply and construction industries.

Life table for Businesses

Life Table for the businesses Rwanda between 1996 and 2019 showing the number which are expected to be successful at the beginning of each time interval.

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Probability Density	Std. Error of Probability Density	Hazard Rate

0	18162	16	18154.000	377	0.02	0.98	0.98	0.021	0.001	0.02
1	17769	77	17730.500	931	0.05	0.95	0.93	0.051	0.002	0.05
2	16761	511	16505.500	478	0.03	0.97	0.90	0.027	0.001	0.03
3	15772	1223	15160.500	757	0.05	0.95	0.86	0.045	0.002	0.05
4	13792	1339	13122.500	617	0.05	0.95	0.82	0.040	0.002	0.05
5	11836	1679	10996.500	435	0.04	0.96	0.78	0.032	0.002	0.04
6	9722	1583	8930.500	423	0.05	0.95	0.75	0.037	0.002	0.05
7	7716	2476	6478.000	239	0.04	0.96	0.72	0.028	0.002	0.04
8	5001	1553	4224.500	126	0.03	0.97	0.70	0.021	0.002	0.03
9	3322	817	2913.500	57	0.02	0.98	0.68	0.014	0.002	0.02
10	2448	444	2226.000	38	0.02	0.98	0.67	0.012	0.002	0.02
11	1966	355	1788.500	40	0.02	0.98	0.66	0.015	0.002	0.02
12	1571	345	1398.500	18	0.01	0.99	0.65	0.008	0.002	0.01
13	1208	481	967.500	15	0.02	0.98	0.64	0.010	0.003	0.02
14	712	294	565.000	18	0.03	0.97	0.62	0.020	0.005	0.03
15	400	120	340.000	11	0.03	0.97	0.60	0.020	0.006	0.03
16	269	46	246.000	5	0.02	0.98	0.59	0.012	0.005	0.02
17	218	41	197.500	3	0.02	0.98	0.58	0.009	0.005	0.02
18	174	45	151.500	3	0.02	0.98	0.57	0.011	0.007	0.02
19	126	15	118.500	2	0.02	0.98	0.56	0.010	0.007	0.02
20	109	37	90.500	3	0.03	0.97	0.54	0.018	0.010	0.03
21	69	25	56.500	0	0.00	1.00	0.54	0.000	0.000	0.00
22	44	20	34.000	1	0.03	0.97	0.52	0.016	0.016	0.03
23	23	14	16.000	0	0.00	1.00	0.52	0.000	0.000	0.00
24	9	9	4.500	0	0.00	1.00	0.52	0.000	0.000	0.00

The median survival time is 24.00

.

From the life table, we can see when the businesses have the greatest risk of failing (closing). One high-risk period is between 1 and 2 years and between 3 and 4 years; this reflects that startups are highly vulnerable to failure. The other period where the failure rate is high is late in life, starting around sixth year.

The tables above show hazard rates of businesses (being unsuccessful) are high at the beginning of the interval and goes on decreasing with the time. This means that there are many startups failing during first two years and the chance of survival at this period is thus very low though it increases with the time.

Interval Start Time is the starting value for each interval. There is an extension of each interval from it's the time it starts until the starting time of the following interval.

Number Withdrawing during Interval is the number of businesses which were censored in this interval. These are businesses which are still active but thus far, they have stopped operating lengthier than the time period shown by this interval.

Number Exposed to Risk is the number of businesses surviving less one half the cases which have been censored. This is envisioned to account for the impact of the censored businesses.

Number of Terminal Events is the number of businesses that experience the terminal event in this interlude. These are unsuccessful businesses = 1.

Proportion Terminating is the ratio of terminal events to the number which are exposed to risk.

Proportion Surviving is one less the proportion terminating.

Cumulative Proportion Surviving at End of Interval is the proportion of businesses surviving from the beginning of the table to the culmination of the interval.

Probability Density is an estimate of the likelihood of the terminal event being experienced during the interval.

Hazard Rate is an estimate of the terminal event being experienced during the interval, conditional upon surviving to the beginning of the interval.

Survival curves disclose a huge amount of information about the businesses in Rwanda, such as if most businesses fail shortly after inception or if most survive (succeed) and likely to survive for many years

5.1 Survival Rates

Kaplan Meier survival curves was used to depict survival rates of various businesses over time in Rwanda. Survival curves also draw a clear picture on duration which was among most the important determinant of business success in Rwanda.

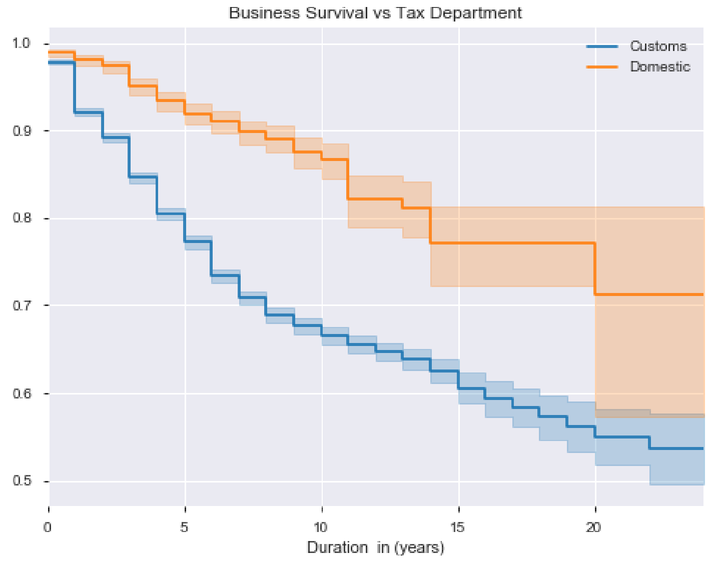


Figure 3: Survival curves of businesses' tax department

Figure 4 above depicts two survival curves, which explains cumulative survival pattern over time for businesses affiliated to tax department in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses affiliated to customs tax the first and second year of inception compared to the businesses affiliated to domestic. This curve shows that the failure rate for businesses affiliated to customs tax is higher compared to the those affiliated to domestic.

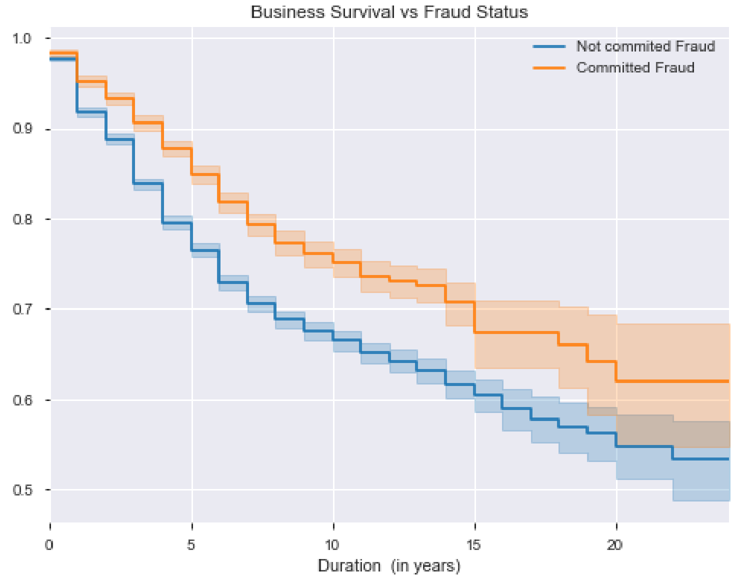


Figure 5:Survival curves of business' Fraud status

Figure 6 above shows two survival curves, which explains cumulative survival pattern over time for fraud status of businesses in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses which have not committed fraud along the first and second year of inception compared to the businesses which have committed fraud. Though there were businesses surviving at the end of the of the 22nd year, this curve shows that the failure rate for businesses which have not committed fraud is higher compared to the those which have not committed fraud.

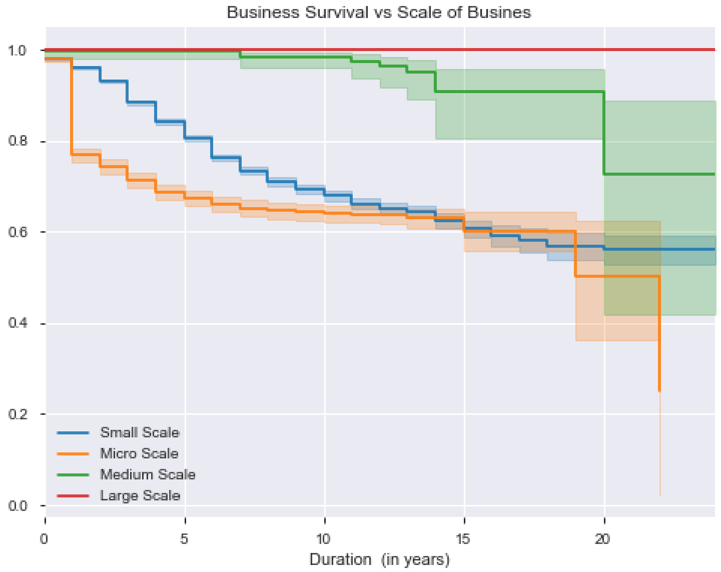


Figure 7: Survival curves of various scales of businesses

Figure 8 above shows four survival curves, which explains cumulative survival pattern over time for various scales of businesses in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the micro scale of businesses compared to the large and medium scale of businesses along the first and second year of inception. Though there were businesses surviving at the end of the of the 20th year, this curve shows that the failure rate for micro and small businesses are higher compared to large and medium scale businesses.

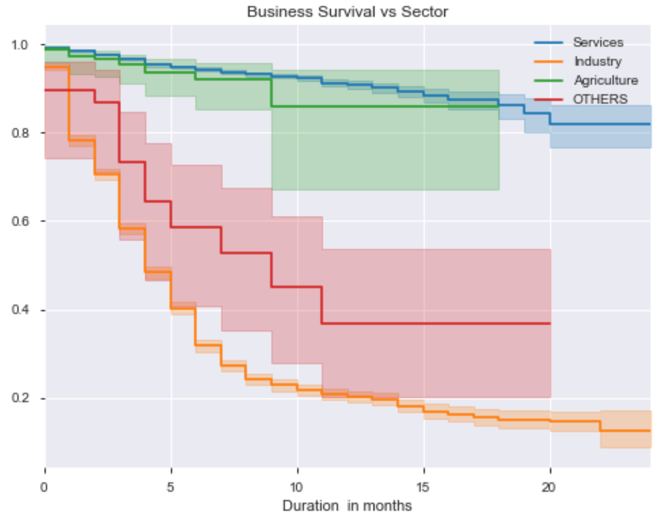


Figure 9: Survival curves of businesses in various sectors

Figure 10 above shows four survival curves, which explains cumulative survival pattern over time of businesses in various sectors in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses in industrial sector along the first and second year of inception compared to the businesses in the service and agricultural sector. Though there were businesses surviving at the end of the of the 22nd year, this curve shows that the failure rate for businesses in industrial sector is high compared to the those in service and agricultural sector.

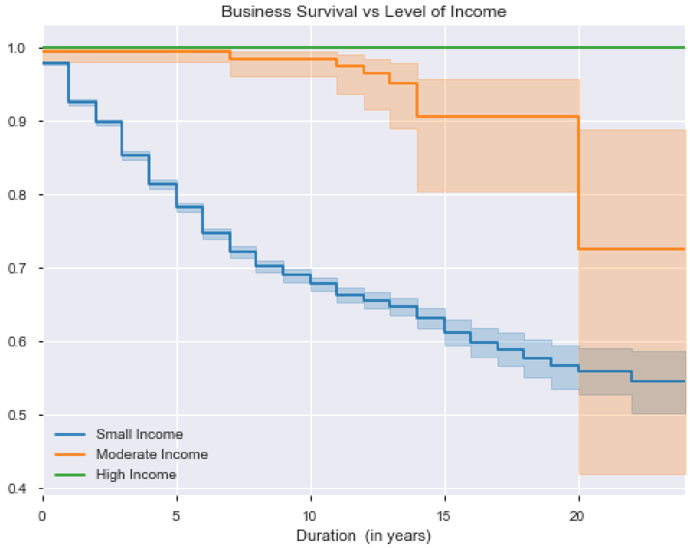


Figure 11: Survival curves of business' level of income

Figure 12 above depicts three survival curves, which explains cumulative survival pattern over time for businesses earning various levels of income in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses earning low level of income along the first and second year of inception compared to the businesses earning moderate and high level of income. Though there were businesses surviving at the end of the of the 22nd year, this curve shows that the failure rate for businesses earning low level of income is high compared to the those earning moderate and high level of income.

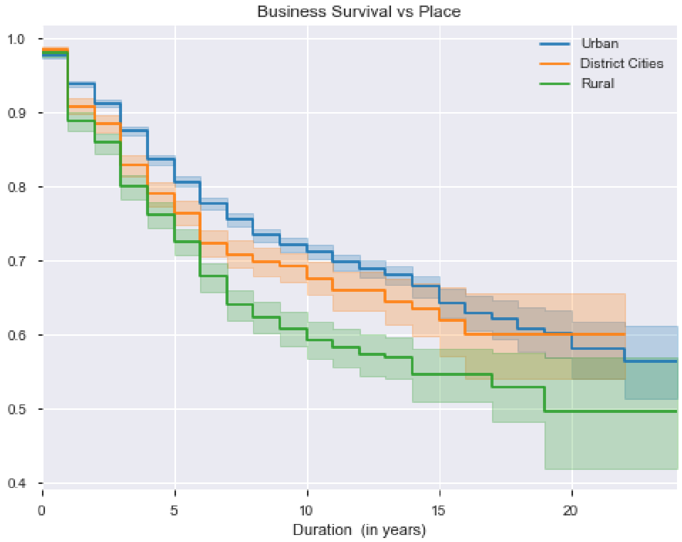


Figure 13:Survival curves of business in various places

Figure 14 above shows three survival curves, which explains cumulative survival pattern over time for businesses operating in different places in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses operating in rural areas along the first and second year of inception. Though there were businesses surviving at the end of the of the 22nd year, this curve shows that the survival and failure rate for businesses in rural areas is high compared to those district cities and urban.

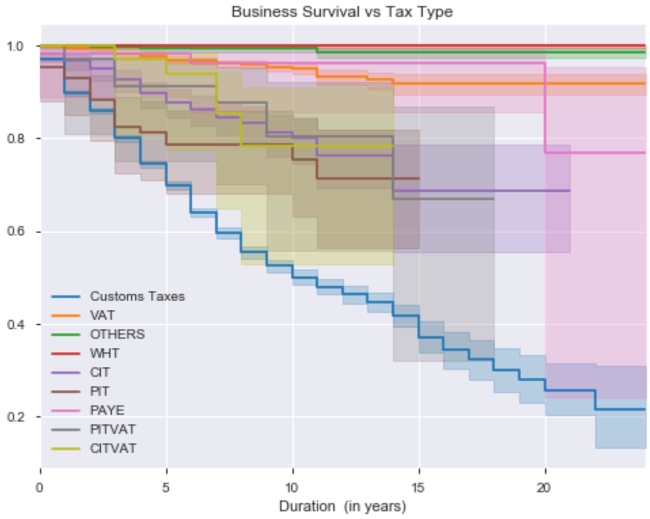


Figure 15: Survival curves of businesses' tax types

Figure 16 above shows nine survival curves, which explains cumulative survival pattern over time for businesses affiliated to various types of tax in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses affiliated to customs tax the first and second year of inception compared to the businesses affiliated to VAT, WHT, CIT, PIT, PAYE, PIVAT and CIVAT tax types. This curve shows that the failure rate for businesses affiliated to customs tax is higher compared to the those affiliated to VAT, WHT, CIT, PIT, PAYE, PIVAT and CIVAT tax types.

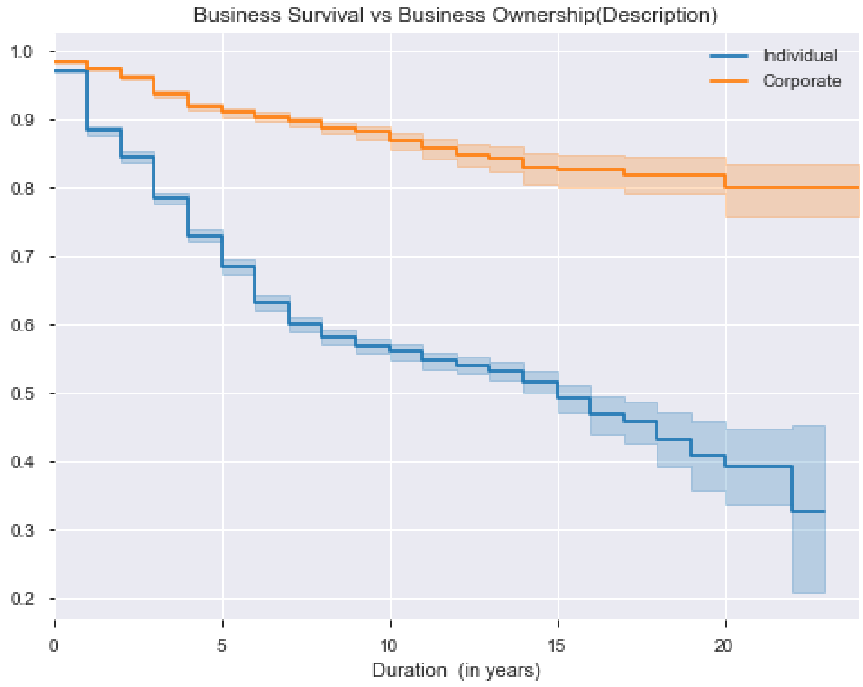


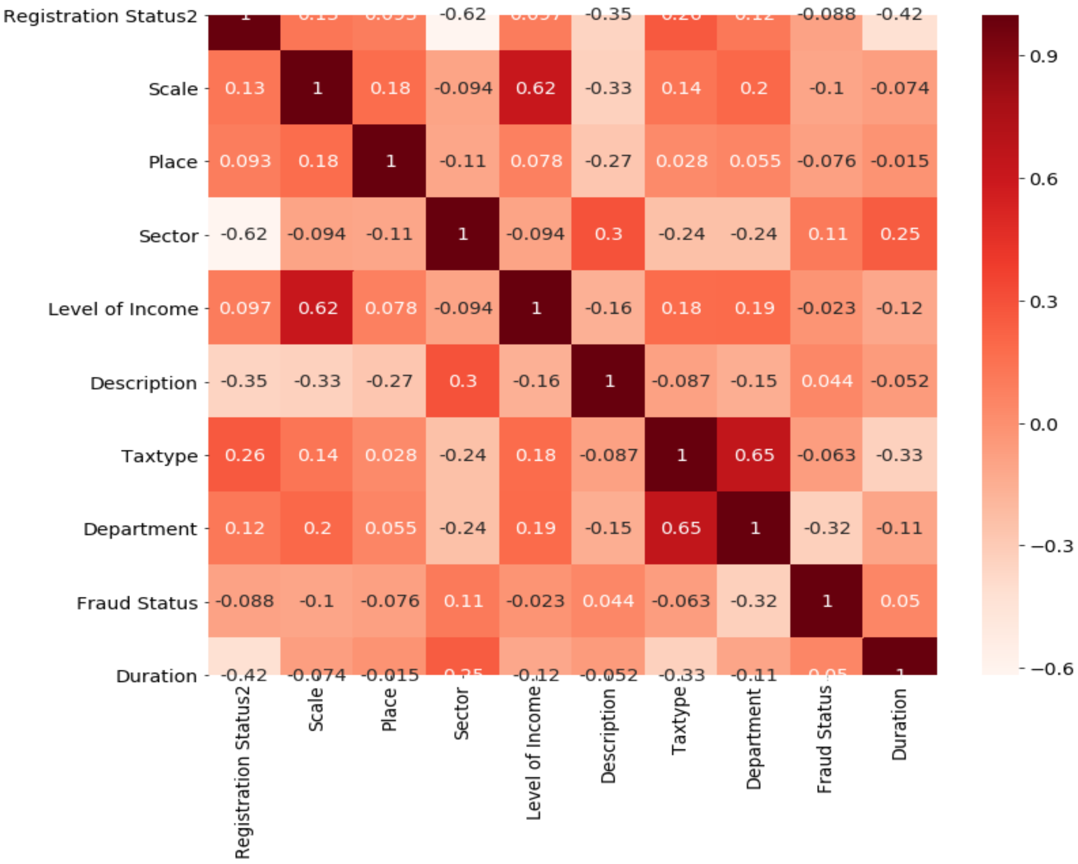
Figure 17: Survival curves of business' description

Figure 18 above depicts two survival curves, which explains cumulative survival pattern over time for businesses owned by individual or corporate in Rwanda. The sharp regression line at the beginning, meaning that there was high failure rate for the businesses owned by individual along the first and second year of inception. Though there were businesses surviving at the end of the of the 23rd year, this curve shows that the failure rate for businesses owned by individual is high compared to the those owned by corporate (Non-individual) during first three years.

5.2 Feature Selection.

5.2.1 Selecting Relevant Features Using Filter Method.

The first step of selecting correlates that contributes to business success was done using filter method. The variables were filtered in order to remain with relevant features. Pearson correlation was used to filter irrelevant and less relevant features. This was done by eliminating irrelevant features and redundant variables and setting a threshold to eliminate less relevant features.



1.

Figure 19: Pearson Correlation Heatmap

Figure 20 below depicts Pearson correlation heatmap which shows the correlation of input features with the response variable (registration status). Features which had correlation of above 0.06 with the response variable were selected. Features whose values were above 0.1 were viewed to be highly correlated with the response variable (registration status). However, features whose values were above 0.06 were considered in order to determine their importance using machine learning. All selected features were trained using five selected machine leaning models and most robust model was used to predict important correlates that contribute to business success according to their ranks.

Table 6: Correlation Of Input Features With The Response Variable (Registration Status)

Features	Coefficients
Sector	0.6180
Duration	0.4196
Description	0.3491
Tax type	0.2618
Scale	0.1334

Department	0.1180
Level of income	0.0967
Place	0.0929
Fraud Status	0.0885

Table 7 above shows correlation of input features with the response variable (registration status). Sector was the most determinant (0. 6180) of business success, followed by duration (0. 4196), description (0. 3491), tax type (0. 2618), scale (0.1334), department (0. 1180), level of income (0.0967), place (0. 0929) and the least is fraud status (0. 0885). However, five models were be used to predict business success and the most robust model was utilized to predict the most important features that contributes to businesses success as shown in the results below.

6 RESULTS

After cleaning the data and selecting relevant features, cross-validation was performed by splitting data into three sets, that is training set, validation set and test set. For training set 80% of dataset were used to fine-tune the algorithms. The training set provide a biased sense of model efficacy since actual samples were used to build the model. However, algorithms confidence was evaluated using training set. The quantity and quality of the training set contributes to success of model performance. For validation set 10% of dataset were hold back from training of the model and were used to give an unbiased sagacity of model efficacy. The validation set was used to evaluate performance on data which were unseen when test data was locked away. For the test set 10% of dataset were hold back from training of the model and were used to give an unbiased sagacity of a final model efficacy. The test set was locked away till fine-tuning of the model was complete thereafter an unbiased evaluation of the final hypothesis was obtained.

6.1 Classifiers Comparison Using Evaluation Metrics

Precision, recall, F1 scores, accuracy and log loss for the validation data

Table 8 shows the precision, recall, F1 scores, accuracy and log loss for the validation data for logistic regression, decision tree, random forest, gradient boost and XGBoost. From Table 9 the values of log loss for each of the models were: XGBoost (0.1688), gradient boosting (0.1706), logistic regression (0.2568), random forest (0.8294) and the one with the highest value of the log loss of the five models is decision tree (1.4497). From Table 10, the accuracy for each of the models were: for each of the models were gradient boosting (0.9450), XGBoost (0.9440), decision tree (0. 9440), random forest (0.9413) and logistic regression (0.9494). From Table 11, the F1 score for each of the models were: for each of the models were: gradient boosting (0.9267), decision tree (0.9261), XGBoost (0.9254), random forest (0.9218) and logistic regression (0.9041). The model which is most robust based on the above results is gradient boost because it has the highest precision score , accuracy, recall score and F1 score.

.

Table 12:Classifiers comparison using the validation data

	Model	Precision score	Recall score	F1_score	Accuracy	log loss
1	Logistic Regression	0.9139	0.8954	0.9041	0.9275	0.2568
2	Decision Tree	0.9352	0.9179	0.9261	0.9440	1.4497
3	Random forest	0.9351	0.9104	0.9218	0.9413	0.8294
4	Gradient Boosting	0.9402	0.9151	0.9267	0.9450	0.1706
5	XGBoost	0.9395	0.9134	0.9254	0.9440	0.1688

Precision, recall, F1 scores, accuracy and log loss for the for five classifiers before hyperparameter tuning for the test data

Table 13 shows the precision, recall, F1 scores, accuracy and log loss for the test data for logistic regression, decision tree, random forest, gradient boost and XGBoost before hyperparameter tuning, in this case default values of model hyperparameters were used . Table 14 the values of the log loss for each of the models were: gradient boosting (0.1495), XGBoost (0.1498), , logistic regression (0.2304), random forest (0.6985) and the one with the highest value of the log loss of the five models is decision tree (1.3793). From Table 15, the accuracy for each of the models were: for each of the models were gradient boosting (0.9523), XGBoost (0.9514), decision tree (0. 9468), random forest (0.9459), logistic regression (0.9321). From Table 16, the F1 score for each of the models were: for each of the models were: gradient boosting (0.9342), XGBoost (0.9326), decision tree (0. 9260), random forest (0.9244) and logistic regression (0.9063). The model which is most robust based on the above results is gradient boost because it has the lowest log loss, highest accuracy, recall score and F1 score.

Table 17: Evaluations metric comparison of the five classifiers before Hyperparameter Tuning on test data

	Model	Precision score	Recall score	F1_score	Accuracy	log loss
1	Logistic Regression	0.9096	0.9031	0.9063	0.9321	0.2304
2	Decision Tree	0.9332	0.9193	0.9260	0.9468	1.3793
3	Random Forest	0.9335	0.9161	0.9244	0.9459	0.6985
4	Gradient Boosting	0.9377	0.9308	0.9342	0.9523	0.1495
5	XGBoost	0.9380	0.9276	0.9326	0.9514	0.1498

Precision, recall, F1 scores, accuracy and log loss for five classifiers after hyperparameter tuning

After hyperparameter tuning was performed, the results for various tree-based models and logistic regression classifiers was obtained based on selected performance metrics. Table 18 shows the precision, recall, F1 scores, accuracy and log loss for the test data for the various learners with tuned hyperparameters, in this case optimal values of model hyperparameters were used. From Table 19 values of the log loss for each of the models were: gradient boosting (0.1239), XGBoost (0.1280), random forest (0.1615) , decision tree (0.1701) and the one with the highest value of the log loss of the five models is logistic regression (0.1718).From Table 20, the accuracy for each of the models were: for each of the models were gradient boosting (0.9626), XGBoost (0.9615), random forest (0.9593), decision tree (0. 9571), logistic regression (0.9494). From Table 21, the F1 score for each of the models were: for each of the models were: gradient boosting (0.9508), XGBoost (0.9491), random forest (0.9461), decision tree

(0.9437), logistic regression (0.9321). The model which is most robust based on the above results is gradient boost because it has the lowest log loss, highest accuracy, recall score and F1 score.

Table 22 below depicts evaluations metrics of tree-based and logistic regression model

Table 23: Evaluations metric comparison of the five classifiers after Hyperparameter Tuning on test data

	Model	Precision score	Recall score	F1_score	Accuracy	log loss
1	Logistic Regression	0.9475	0.9191	0.9321	0.9494	0.1718
2	Decision Tree	0.9482	0.9394	0.9437	0.9571	0.1701
3	Random Forest	0.9549	0.9381	0.9461	0.9593	0.1615
4	Gradient Boosting	0.9561	0.9459	0.9508	0.9626	0.1239
5	XGBoost	0.9565	0.9424	0.9491	0.9615	0.1280

6.2 Classifiers Comparison Based on Area Under Receiver Operating Characteristic Curve (ROC AUC)

The Figure 21: below depicts ROC AUC for logistic regression, decision tree, random forest, gradient boost and XGBoost.

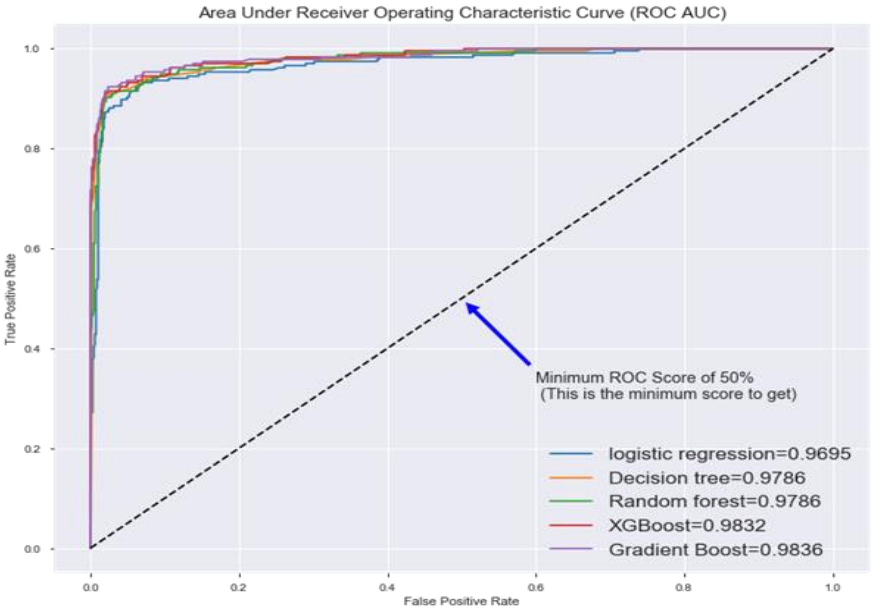


Figure 22 ROC AUC performance comparison of five classifiers

.

Figure 23 above depicts the Area Under Receiver Operating Characteristic Curves (ROC AUC) which shows how the predictive models used will be able to differentiate between the true positives (number of cases correctly identified that business will succeed) and true negatives (the number of cases correctly identified that business will be unsuccessful) for each of the trained models under test data. From Figure 24 gradient boosting has the high ROC AUC on the test data (0.9836), followed by XGBoost (0.9832), random forest (0.9786), decision tree (0.9786) and lastly logistic regression (0.9695)

6.3 Feature Importance

Figure 25 and Table 24 shows the feature importance using gradient boosting classifier. The classifier considers sector of the business as the most important factor in determining success of business (0.6579) followed by duration (0.2189), tax type (0.0523), scale (0.0227), description (0.0191), fraud status (0.0144), department (0.0114), place (0.0023) and lastly level of income of the business (0.0010).

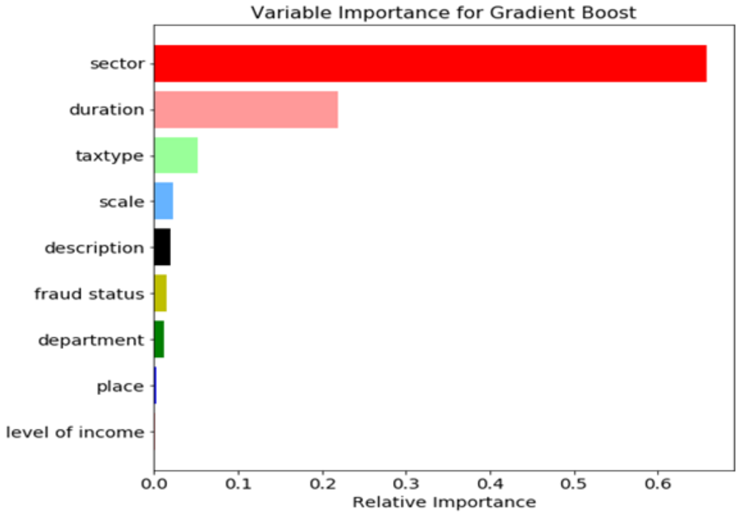


Figure 26: Feature importance using Gradient Boosting classifier

Table 25: Feature importance from Gradient Boosting

Feature	Importance
Sector	0.6579
Duration	0.2189
Tax type	0.0523
Scale	0.0227
Description	0.0191
Fraud status	0.0144
Department	0.0114
Place	0.0023
Level of income	0.0010

Figure 27 and Table 26 shows the feature importance with extracted dummy variables using gradient boosting model. The classifier considers Industry sector (0.5808) as most important feature, followed by customs tax type (0.1294), duration (0.1066), services sector(0.0437), micro scale (0.0258), committed tax fraud (0.02065), small scale (0.0204), not committed tax fraud (0.01334), international business (0.012), domestic department (0.0101), other tax types (0.0087), other sectors (0.0063), non-individual taxpayers (0.0039), individual taxpayers (0.0033), rural areas (0.0019), urban areas (0.0014), vat taxtype (0.0005), wht taxtype (0.0004), citvat taxtype (0.0004), cit taxtype (0.0004), pitvat tax type (0.0003), district cities (0.0003), pit taxpayers (0.0002), low income taxpayers (0.0001).

Figure 28: Feature importance from Gradient Boosting

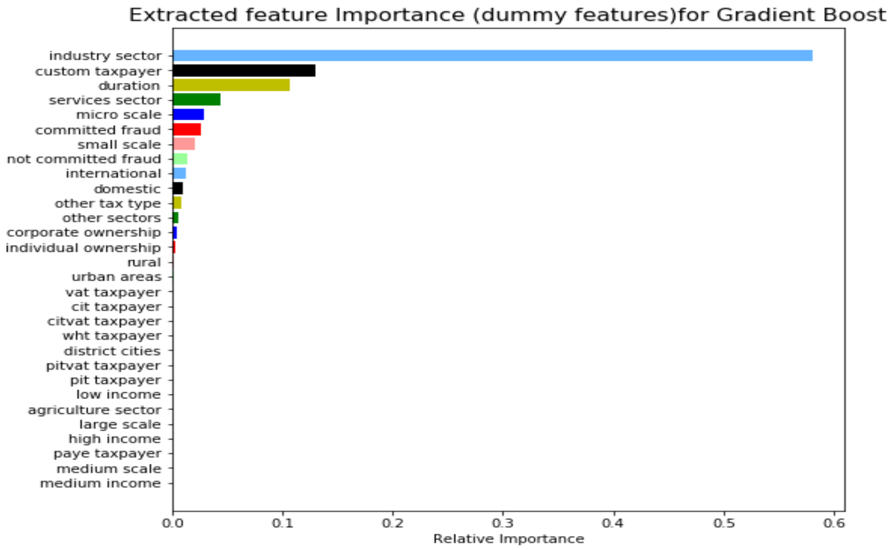


Table 27:Feature importance from Gradient Boosting

Boosting

Feature	Importance
Industry sector	0.6259
Duration	0.2289
Custom taxpayer	0.0461
Services sector	0.0346
Micro scale	0.0130
Individual ownership	0.0106
Corporate ownership	0.0076
Domestic	0.0067
Other sectors	0.0064
Not committed fraud	0.0038
WHT taxpayer	0.0031
Other tax type	0.0031
Committed fraud	0.0022
International	0.0022
Rural	0.0009

District cities	0.0008
Urban areas	0.0007
CIT taxpayer	0.0006
Agriculture sector	0.0006
Low income	0.0005
Small scale	0.0005
VAT taxpayer	0.0003
CITVAT taxpayer	0.0002
Large scale	0.0002
Medium scale	0.0001
PIT taxpayer	0.0001
PAYE taxpayer	0.0001
Medium income	0.0001
High income	0.0000
PITVAT taxpayer	0.0000

The feature importance using gradient boosting was compared with compared with that of logistic regression. The coefficients from logistic regression were both negative and positive. The positive coefficients depict variables that predicts class 1 (unsuccessful business), whereas the coefficients depict variables that predicts class 0 (business success). The higher value of this criterion when it is compared to another variable infers it is more imperative for generating a prediction. Large positive values signify higher importance in the prediction of positive class while large negative values signify higher importance in the prediction of negative class.

Logistic regression is an example of linear machine learning algorithms, the model is fit, and the prediction is the weighted sum of the input values. Larger coefficients are necessarily more informative because they contribute a greater weight to final prediction in most cases. Negative coefficient indicates a strong negative correlation we must rank features by absolute values of their coefficients.

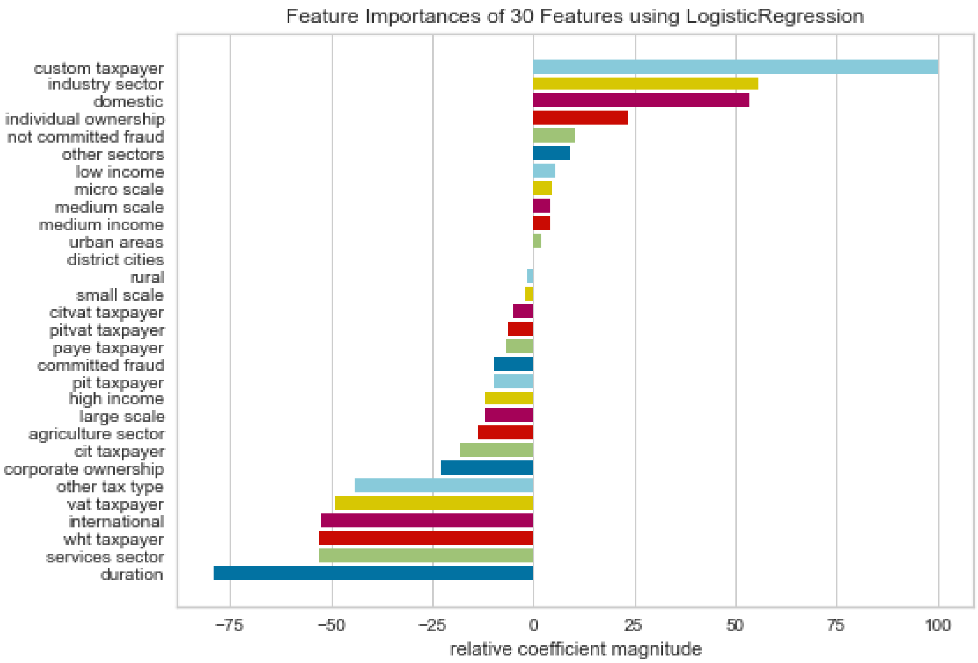


Figure 29:Feature importance using Logistic Regression

Table 28:Feature importance using Logistic Regression

Feature	Importance
custom taxpayer	1.6365
industry sector	0.9108
domestic	0.8781
individual ownership	0.3826
not committed fraud	0.1726
other sectors	0.1476
low income	0.0894
micro scale	0.0747
medium scale	0.0713
medium income	0.0713
urban areas	0.0342
district cities	-0.0028
rural	-0.0265
small scale	-0.0293
citvat taxpayer	-0.0805
pitvat taxpayer	-0.1014
paye taxpayer	-0.1110
committed fraud	-0.1608
pit taxpayer	-0.1621
high income	-0.1919

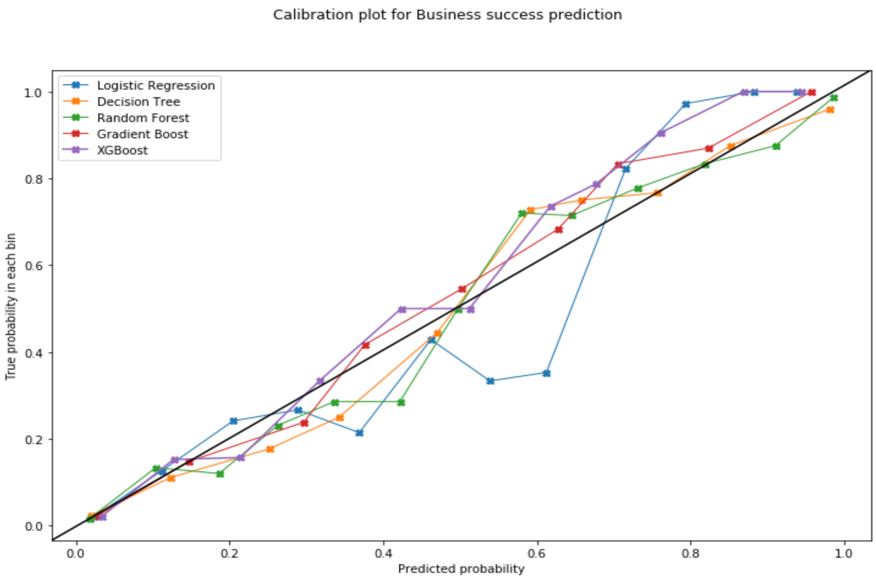
large scale	-0.1919
agriculture sector	-0.2225
cit taxpayer	-0.2977
corporate ownership	-0.3718
other tax type	-0.7235
vat taxpayer	-0.8034
international	-0.8613
wht taxpayer	-0.8632
services sector	-0.8650
duration	-1.2947

Figure 30 and Table 29 depicts the feature importance using logistic regression with dummy variables. For business which were likely to be unsuccessful classifier considers customs tax type (1.6365), as most important feature, followed by Industry sector (0.9108), domestic department (0.8781), individual enterprises (0.3826), not committed tax fraud (0.1726), other sector (0.1426), low income (0.0894), other tax types (0.0037), micro scale (0.0747), medium scale(0.0713), medium income (0.0713),urban areas (0.0342), customs department (0.0023) .

For business which were likely to be successful classifier considers duration (-1.2947), as most important feature, followed by Industry sector services sector (-0.8650), WHT tax type (-0.8632) , VAT tax type (-0.8034), corporate (non-individual) enterprises (-0.3718), CIT tax type (-0.2977), agriculture sector (-0.2225), large scale (-0.1919), high income (-0.1919), PIT taxpayers (-0.1621), committed tax fraud (-0.1608), PAYE tax type (-0.111), PITVAT tax type (-0.1014), CITVAT tax type (-0.0805),small scale(-0.00293), rural areas (-0.00265), district cities (-0.0028),

Calibration

Figure 31: Calibration curve for five classifiers



Miscalibration was evaluated using the calibration plot as shown in Figure 32 above. Tree-based models such as gradient boosting, XGBoost, random forest and decision tree were well calibrated. However, logistic regression exhibited some form of miscalibration. Such miscalibration can be biased and corrected by parameter tuning to ensure the model predict the true class event as a function of its un-calibrated predicted probabilities.

7 DISCUSSION

In this study, decision tree, random forest, gradient boost, XGBoost and logistic regression were compared as to their respective performance in predicting success of a business. A desirable classifier was taken as the one with high recall score, F1 score and accuracy. Finally, binary cross-entropy (log loss) was used to evaluate how good or bad are probabilities predicted from tree-based models and logistic regression models. The most desirable classifier was one with the lowest log loss.

Gradient boosting was most robust model based on results from Table 30 , Table 31 and Table 32 . However, Table 33 and Table 34 were considered since they were based on test data. From and Table 35 results it was evident that performance classifiers improved after hyperparameter tuning. Boosting algorithms gradient boosting, and XGBoost seemed to perform best when the hyperparameters of the decision tree did not let the tree to grow large. Since performance metrics of all classifiers improved after hyperparameter tuning, the results were reported based on tuned hyperparameters. Therefore, most robust model in Table 36 was used to predict business success in Rwanda

From the log loss results in Table 37, the most robust model is gradient boosting classifier since it has the lowest log loss (0.1239) while the model with the highest log loss is the logistic regression classifier (0.1718). The predictions from the models would be considered certain since the log loss returned low values from respective models. This implies that gradient boosting has the lowest uncertainty compared to other models.

From the accuracy results in **Error! Reference source not found.**, the most robust model is gradient boosting classifier since it has the highest accuracy (0. 9626) while the model with the lowest accuracy is the logistic regression (0.9494). The gradient boosting has high predictive accuracy since it is an optimized distributed gradient boosted decision tree which is more efficient and portable. Other tree-based models such as XGBoost random forest and decision tree had high accuracies too.

From the recall score results in Table 38, , the most robust model is gradient boosting classifier since it has the highest recall score (0.9459). It was followed by a XGBoost, random forest, decision tree and lastly logistic regression. Since the focus of the study was to predict success of the business, recall score was therefore an important metric since it gives the is ratio of correctly predicted successful businesses to the all values in true class. Recall score returns proportion of total relevant results classified by the algorithm. In this study therefore, evaluating a model based on recall score, gradient boosting is therefore the most recommended model to predict business success.

From Figure 33**Error! Reference source not found.**, gradient boosting has the highest ROC AUC (0.9836) on the test data. When the ROC AUC is closer to the upper left corner, the recall rate of the model is higher. The point on the ROC AUC closest to the upper left corner is the best threshold with the least classification errors, and the total number of false positive examples and false negative examples is the lowest. Based on the ROC AUC the authors could say that gradient boosting is the best classifier in business success prediction. It is followed in rank by XGBoost, random forest, decision tree and lastly logistic regression. Though all the classifiers have a ROC AUC of more than 0.96, which could be considered high, the boosted trees show higher ROC AUC than random forest, decision trees and logistic regression with each of the boosted trees being at least 0.98 as compared to logistic regression (0.9695).

The classification being binary, setting logistic regression as a baseline classifier was plausible since it would be expected to give high performance. Nevertheless, on ROC AUC, the tree-based models had higher performance than the logistic regression. Moreover, the boosted trees showed higher ROC AUC than bagged trees and standalone learners.

The relative importance of each input variable was measured by beholding the quantity of the tree nodes that utilize that feature, minimizing impurity on average. Figure 34, Figure 35 ,Table 39 and Table 40 Table 1 show the feature importance of gradient boost which was compared with that of logistic regression. Since gradient boost was the most robust model based on the log loss, recall score, ROC AUC and accuracy, thus it’s feature importance will be considered in predicting determinants contributing to business success in Rwanda, however it was an ideal to compare with feature importance logistic regression since it depicts variable contributing to success of business and unsuccessful ones. Feature importance gave a peep into the variables that contribute to the success of the business. Categorical variables were broken down into dummy variables to get separate feature importance per class in that variable. The most important feature in predicting business success was the sector of business. Businesses in the industrial sector were more likely to be unsuccessful while those in service and agricultural sectors were more likely to succeed. Thus, there is a need to put measures in place that will boost growth of businesses operating in the industrial sector. The second most important feature that contributes to business success was duration. Some startups were observed to be unsuccessful after one year. This therefore calls for the government to put measures in place to protect startups. Tax type was the third most important feature and businesses affiliated with custom taxes were observed to have higher rates of failure while those associated with WHT, VAT and PAYE had higher chance of success. Scale of business was the fourth most important feature. Small and micro scale businesses were least to succeed while large scale businesses had higher chances to succeed, therefore the government should put in place policies and mechanisms that will give a friendly environment for micro and small business such as tax reduction, training and coaching in order to improve their growth and chances of survival.

8 CONCLUSION and RECOMMENDATIONS

In this study it was found that sector was most important feature that contributes to business success in Rwanda. Therefore, this paper suggests further segmentation of sector to identify other classes within the sector of economy that could contribute to success of business. It was also found that predicting business success in Rwanda using a tree-based model was superior to logistic regression. Boosted trees depicted an outstanding predictive performance. Moreover, the results have demonstrated that boosted tree algorithms have lower training time than bagging, decision tree and logistic regression. Tree-based ensemble models require fewer hyperparameters tuning and in most situations default hyperparameters can lead in good performance. Nevertheless, gradient boosting was more robust than others including XGBoost, random forest, decision trees and logistic regression models in predicting business success in Rwanda. This successful application of gradient boosting in predicting business success could be a precursor for tackling a broad class of pattern detection of determinants that contributes to business growth. Ensembled tree-based models learn directly from high-level representations dataset, thus potentially evading traditional idiosyncratic thresholding-based criteria of business success prediction. The techniques employed in this study serve as an instance that may perhaps be automated, applied to other business success predictions. The developed gradient boosting model can achieve an accurate and hourly prediction which is very reliable and therefore could be utilized to identify determinants of business success (such as sector, duration, tax type , scale of business, place, level of income among others.), helping government, startups and investors or any business stakeholder to make informed decisions and operational optimization of business activities. Future research evaluates the capability of such approaches to prospectively detect determinants of business success missed by non-tree-based models and to translate into improved business success predictions. In future work, non-tree-based models such as artificial neural network, support vector machine, k- Nearest Neighbors, gaussian naïve among others, will need to be investigated for business success prediction.

9 APPENDICES

9.1 Appendix A

Hyperparameters

Table 41: Hyperparameter Tuning for Logistic Regression

Hyperparameters	Range	Optimal Value
solver	['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga’]	'liblinear',
penalty	['l1', 'l2', 'elasticnet]	'l2'
dual	[False, True]	FALSE

Table 42: Hyperparameter Tuning for Decision Tree

Hyperparameters	Range	Optimal Value
splitter	["random", "best"]	"best"
max_depth	[3 to 8]	4
min_samples_split	[2 to 8]	3

Table 43: Hyperparameter Tuning for Random Forest

Hyperparameters	Range	Optimal Value
n_estimators	[100 to 5000]	400
splitter	["random", "best"]	"best"
max_depth	[3 to 8]	4
min_samples_split	[2 to 8]	3

Table 44: Hyperparameter Tuning for Gradient Boost

Hyperparameters	Range	Optimal Value
learning_rate	[0.01 to 0.2	0.06
n_estimators	[100 to 5000]	500
max_depth	3 to 8	4
max_features	0 to 0.9	0.9
min_samples_leaf	1 to 4	2

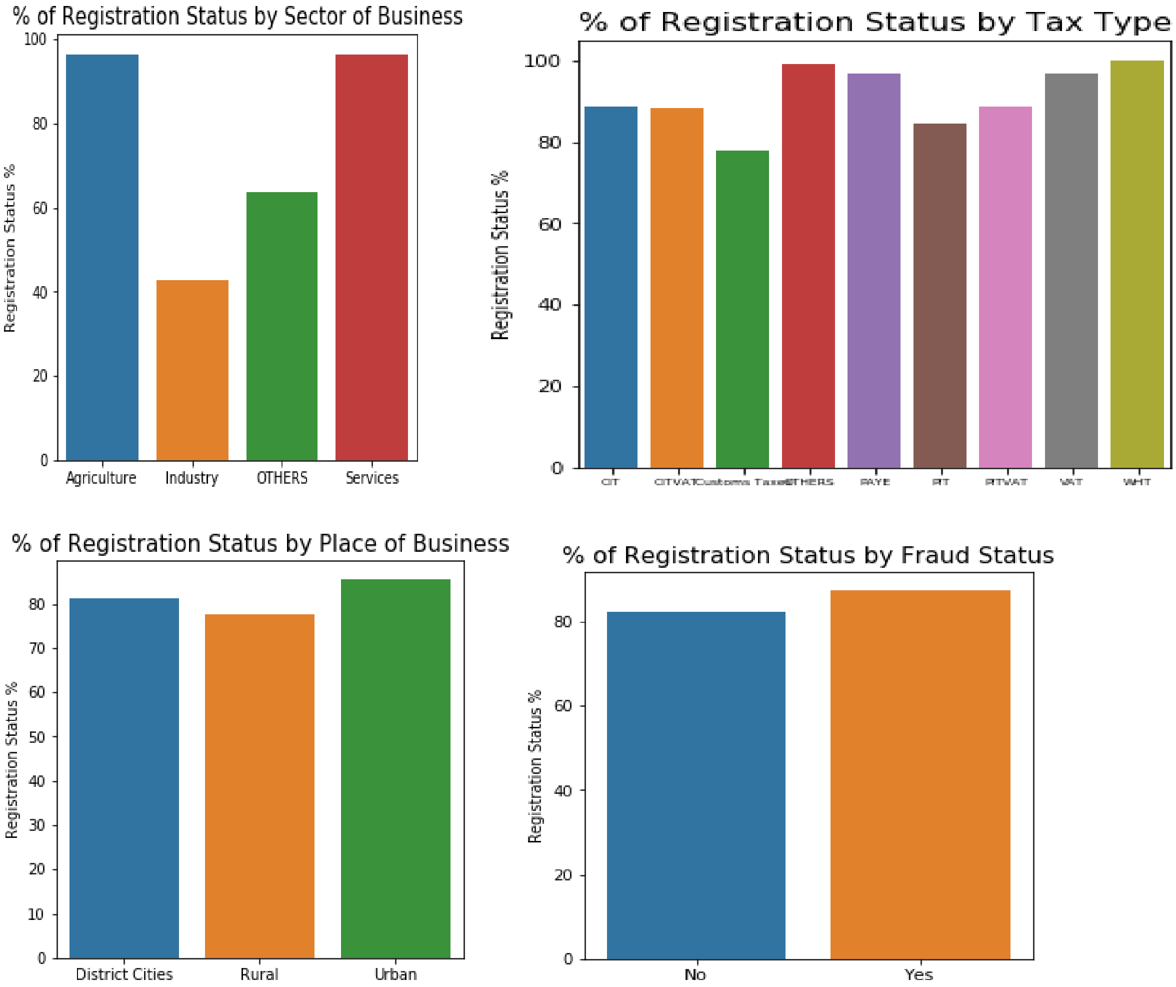
Table 45: Hyperparameter Tuning for XGBoost

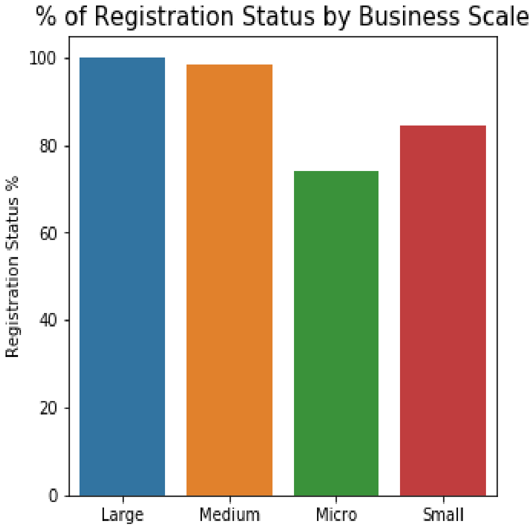
Hyperparameters	Range	Optimal Value

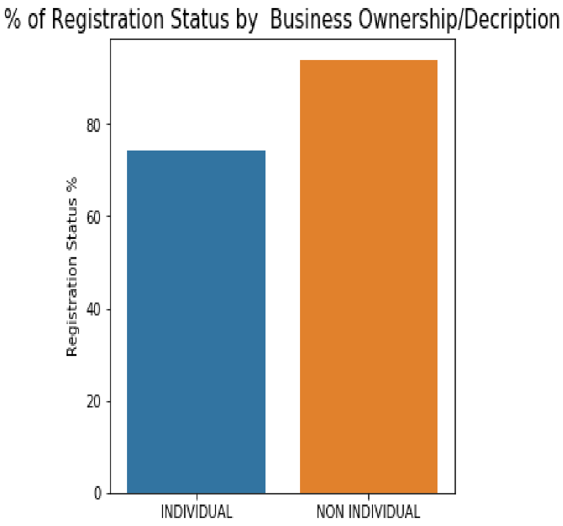
learning_rate	[0.01 to 0.5]	0.03
n_estimators	[100 to 5000]	300
max_depth	[3 to 8]	4
gamma	[0 to 0.9	0.9
min_child_weight	[1 to 4]	1

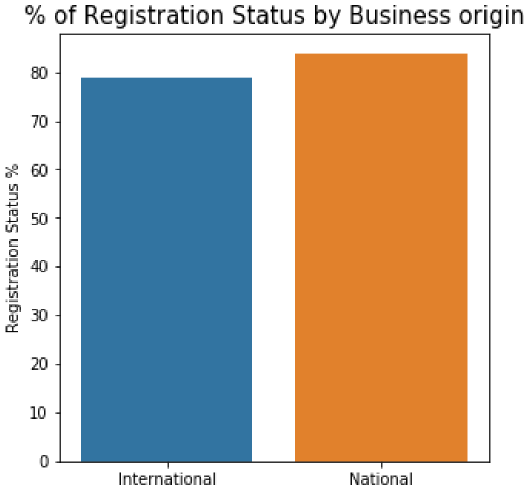
9.2 Appendix B

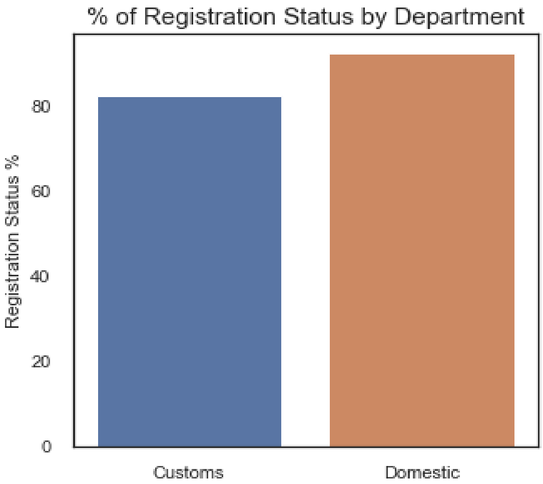
Bar graphs showing distribution of all Input Features used in the Analysis against Registration Status of Businesses .

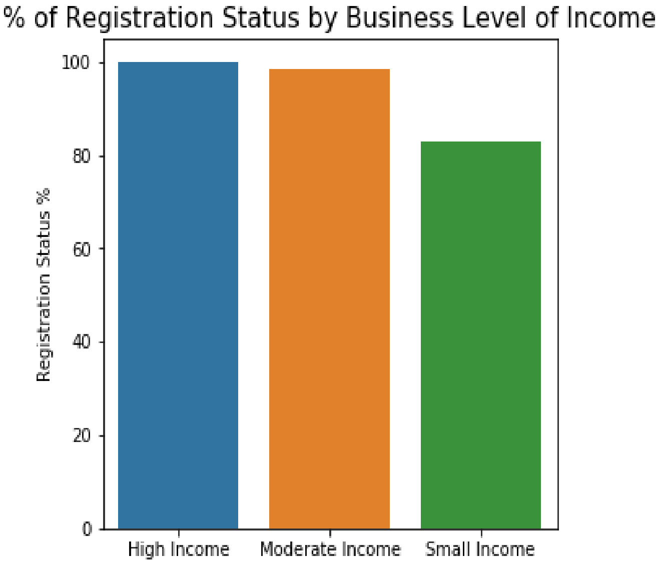



% of Registration Status by Business Scale


% of Registration Status by Business Ownership/Decription


% of Registration Status by Business origin


% of Registration Status by Department


% of Registration Status by Business Level of Income

10 REFERENCES

Aqeel, A. M. Bin, Awan, A. N., & Riaz, A. (2011). Determinants of Business Success (An Exploratory Study). *International Journal of Human Resource Studies*. https://doi.org/10.5296/ijhrs.v1i1.919

Asadi, N., Lin, J., & De Vries, A. P. (2014). Runtime Optimizations for Tree-Based Machine Learning Models. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2013.73

Avanzi, F., Johnson, R. C., Oroza, C. A., Hirashima, H., Maurer, T., & Yamaguchi, S. (2019). Insights Into Preferential Flow Snowpack Runoff Using Random Forest. *Water Resources Research*. https://doi.org/10.1029/2019WR024828

Ayandibu, A.O. and Houghton, J. (2017). The role of small and medium enterprise in economic development. *Journal of Business and Retail Management Research (JBRMR)*.

Bayisenge, R., Shengede, H., Harimana, Y., Bosco Karega, J., Lukileni, M., Nasrullah, M., … Emmerance Nteziyaremye, B. (2020). Contribution of Small and Medium Enterprises Run by Women in Generating Employment Opportunity in Rwanda. *International Journal of Business and Management*. https://doi.org/10.5539/ijbm.v15n3p14

Bhattarai, A., Shrestha, E., & Sapkota, R. P. (2019). Customer Churn Prediction for Imbalanced Class Distribution of Data in Business Sector. *Journal of Advanced College of Engineering and Management*. https://doi.org/10.3126/jacem.v5i0.26693

Bowers, A. J., & Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed at Risk*. https://doi.org/10.1080/10824669.2018.1523734

Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. https://doi.org/10.1016/j.peva.2007.06.006

Clark, L. A., & Pregibon, D. (2017). Tree-based models. In *Statistical Models in S*. https://doi.org/10.1201/9780203738535

Dangeti, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. In *Packt Publishing*.

Duan, K., Keerthi, S. S., & Poo, A. N. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*. https://doi.org/10.1016/S0925-2312(02)00601-X

Feindt, S., Jeffcoate, J., & Chappell, C. (2002). Identifying success factors for rapid growth in SME E-commerce. *Small Business Economics*. https://doi.org/10.1023/A:1016165825476

Flach, P. A. (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. *Proceedings, Twentieth International Conference on Machine Learning*.

Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/2009916.2009932

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting*. https://doi.org/10.1002/for.1153

Gupta, P., Sharma, A., & Jindal, R. (2016). Scalable machine-learning algorithms for big data analytics:

a comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1194

Headd, B. (2003). Redefining Business Success: Distinguishing between Closure and Failure. *Small Business Economics*. https://doi.org/10.1023/A:1024433630958

Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beyond Sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*. https://doi.org/10.1214/17-EJS1338SI

LANE, S. J., & SCHARY, M. (1991). UNDERSTANDING THE BUSINESS FAILURE RATE. *Contemporary Economic Policy*. https://doi.org/10.1111/j.1465-7287.1991.tb00353.x

Lenz, S. T. (2010). Gönen, Mithat (2007):Analyzing Receiver Operating Characteristic Curves with SAS. *Statistical Papers*. https://doi.org/10.1007/s00362-008-0156-x

Lussier, R. N., & Pfeifer, S. (2001). A Crossnational Prediction Model for Business Success. *Journal of Small Business Management*. https://doi.org/10.1111/0447-2778.00021

McKenzie, D., & Paffhausen, A. L. (2019). Small Firm Death in Developing Countries. *The Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_00798

Miles, K. J. (2013). Exploring Factors Required for Small Business Success in the 21st Century.

Mutandwa, E., Taremwa, N. K., & Tubanambazi, T. (2015). Determinants of business performance of small and medium size enterprises in Rwanda. *Journal of Developmental Entrepreneurship*. https://doi.org/10.1142/S1084946715500016

Nagaya, N. (2017). SME Impact on Output Growth, Case Study of India. *Palma Journal*.

Ramukumba, T. (2014). Overcoming SMEs challenges through critical success factors. In *ECONOMIC AND BUSINESS REVIEW*.

Recchioni, M. C., Tedeschi, G., & Gallegati, M. (2015). A calibration procedure for analyzing stock price dynamics in an agent-based framework. *Journal of Economic Dynamics and Control*. https://doi.org/10.1016/j.jedc.2015.08.003

Rodriguez, A., & Rodriguez, P. N. (2006). Understanding and predicting sovereign debt rescheduling: A comparison of the areas under receiver operating characteristic curves. *Journal of Forecasting*. https://doi.org/10.1002/for.998

Sage, A. J., Genschel, U., & Nettleton, D. (2020). Tree aggregation for random forest class probability estimation. *Statistical Analysis and Data Mining*. https://doi.org/10.1002/sam.11446

Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability (Switzerland)*. https://doi.org/10.3390/su11030917

Schapire, R. E. (2003). *The Boosting Approach to Machine Learning: An Overview*. https://doi.org/10.1007/978-0-387-21579-2_9

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2017.2767063

Sibomana, J. P., & Shukla, J. (2016). EFFECT OF VILLAGE SAVINGS AND LOAN ASSOCIATIONS ON SMALL AND MEDIUM ENTERPRISE (SME) GROWTH IN RWANDA: SURVEY OF KAYONZA DISTRICT. *International Journal of Business and Management ReviewOnline)International Journal of Business and Management Review*.

Siow Song Teng, H., Singh Bhatia, G., & Anwar, S. (2011). A success versus failure prediction model for small businesses in Singapore. *American Journal of Business*. https://doi.org/10.1108/19355181111124106

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Theng, L. G., & Boon, J. L. W. (1996). An exploratory study of factors affecting the failure of local small and medium enterprises. *Asia Pacific Journal of Management*. https://doi.org/10.1007/BF01733816

Van Praag, C. M. (2003). Business Survival and Success of Young Small Business Owners. *Small Business Economics*. https://doi.org/10.1023/A:1024453200297

Zeng, J. (2017). Forecasting Aggregates with Disaggregate Variables: Does Boosting Help to Select the Most Relevant Predictors? *Journal of Forecasting*. https://doi.org/10.1002/for.2415

Zhu, H., Yu, C. Y., & Zhang, H. (2003). Tree-based disease classification using protein data. *Proteomics*, *3*(9), 1673–1677. https://doi.org/10.1002/pmic.200300520