SN Bus Econ (2021) 1:101
https://doi.org/10.1007/s43546-021-00104-2

REVIEW

Business success prediction in Rwanda: a comparison of tree-based models and logistic regression classifiers

Francis Kipkogei[1] · Ignace H. Kabano[1,2] · Belle Fille Murorunkwere[1,3] · Nzabanita Joseph[1,2]

Received: 1 February 2021 / Accepted: 17 June 2021 / Published online: 16 July 2021

Abstract

Businesses contribute immensely to economic growth. However, many enterprises started fail within a year of their operation. This study seeks to predict business success, elucidating on factors affecting the success of a business based on recent data for timely intervention. The study used Rwanda Revenue Authority data. Tree-based models were compared with logistic regression for the prediction of the business success. Log loss, Area under Receiver-Operating Characteristic Curve, accuracy, recall, and F1 score were used to evaluate the performance of each model in differentiating between successful and unsuccessful business. Tree-based ensemble models were more robust than other classifiers. However, gradient boosting was the most robust model. The results showed that the business industry (sector of the economy) is the most important factor determining business success. Other important factors are the nature of the business and type of ownership, duration of operation, and location of the business.

Keywords Business success · Tree-based models · Logistic regression

✉ Francis Kipkogei
francisyego4@gmail.com

✉ Ignace H. Kabano
kabanoignace@gmail.com

[1] African Centre of Excellence in Data Science, University of Rwanda, Kigali, Rwanda

[2] Department of Applied Statistics, College of Business and Economics, University of Rwanda, Kigali, Rwanda

[3] Rwanda Revenue Authority, Kigali, Rwanda

Introduction

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 and 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 10 years of business existence (Ramukumba 2014). In addition, 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). In addition, 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 Praag 2003). Survival rates of businesses were found to be similar to location, employment size, business type, economic sector, and distance from the mall (Van Praag 2003).

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 and Antosiuk 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. First, 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 and Antosiuk 2021; Gupta et al. 2016). 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 and Pregibon 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.

Materials and methods

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 Fig. 1.

Data

The data for this study were 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, addresses, and TIN details was not included. However, the authors established a few eligibility criteria that were in

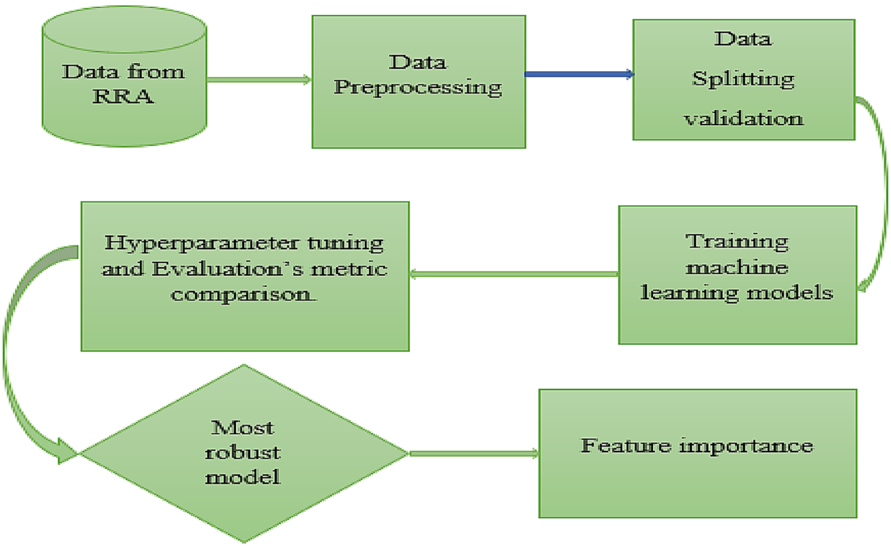
Data from RRA
Data Preprocessing
Data Splitting validation
Training machine learning models
Hyperparameter tuning and Evaluation's metric comparison
Most robust model
Feature importance

Fig. 1 General illustration of process followed to predict business success

line with the prime objective of the study to enhance the 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 the different datasets, other reasons for exclusions were due to the missing-ness of important features; further reducing the sample size to 18,162 businesses. Total post-processed data consisted of 18,162 businesses, from which 13,565 were successful, and 4597 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, scale, business exports, sector, level of income, duration, fraud status, and origin. The registration status indicates that the business is still registered or deregistered due to losses. The description indicates whether the business is owned by an individual or a corporate business. Tax type indicates the types of taxes that the business has been audited on, this includes value-added tax, pay as you earn, withholding, custom taxes, and others.

Place indicates the location where business is operating, this includes rural, urban, and district cities. The scale indicates the size of the business (large, medium, small, and micro). Business exports indicate 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: the agricultural sector, industrial sector, and service sector. The level of income indicates whether the business is making a 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 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 any other country. Below is a summary statistics of data description.

Summary of data description

Classification of business	Frequency	Successful businesses	Unsuccessful businesses
Sector			
1 = Agriculture	180	168 (93.3%)	12 (6.7%)
2 = Industry	4776	1068 (22.4%)	3708 (77.6%)
3 = Services	12,915	12,116 (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	1222	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	2730	1968 (72.1%)	762 (27.9%)
3 = Rural	2504	1641 (65.5%)	863 (34.5%)
Fraud status			
1 = Yes (committed fraud)	5347	4312 (80.6%)	1044 (19.4%)
0 = No (not committed fraud)	12,815	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	3055	1983 (64.9%)	1072 (35.1%)
Description			
1 = Individual	9437	5671 (60.1%)	3766 (39.9%)

Classification of business	Frequency	Successful businesses	Unsuccessful businesses
2 = Corporation	8725	7894 (90.5%)	831 (9.5%)
Origin			
1 = National	18114	13,534 (74.72%)	4580 (25.28%)
2 = International	48	31 (64.58%)	17 (35.42%)
Business exports			
1 = Domestic	2108	1873 (88.9%)	235 (11.1%)
2 = Customs	16,054	11,692 (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	17,548	12,963 (73.9%)	4585 (26.1%)

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.

Data preprocessing

Data were obtained in excel format; however, messiness, duplicates, noise, and outliers were prevalent in the data. Thus, data were cleaned by handling missing values, duplicates, and noisy data, further data splitting was carried out 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 were split into 80% training set, 10% test set and 10% validation set after it was cleaned.

Models

After data cleaning, the data were 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 algorithms (decision tree, random forest, gradient boost, XGBoost, and logistic regression) to find out the determinants of business success in Rwanda.

Logistic classification algorithm

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

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

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, the logistic function is modeled such that predicted probabilities fall between 0 and 1.

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

$$\varphi\left(g\left(x_i\right)\right)\{\geq 0.5Y_i = 1\ 0.5Y_i = 0 \tag{2}$$

The logistic function used in logistic regression can be written as (Casella et al. 2013):

$$\varphi(g(x)) = \frac{e^{-g(x)}}{1 + e^{-g(x)}}$$

Tree-based models

Tree-based methods involve segmenting or stratifying the predictor space into several simple regions. To project 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).

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 the prediction of a value in the response variable (Dangeti 2017).

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 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-offs and biases. Some examples of boosting are gradient boosting and XGBoost.

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 a 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, the 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 occurs most frequently in the similar terminal node as the test case being forecasted (Sage et al. 2020).

Gradient boosting

Gradient boosting is a machine-learning technique containing optimized boosted decision trees which are formed by combining many weak learning models to give stronger predictive models. Components in gradient boosting systems are summarized into two essential parts: the 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 dataset and the value of the predicted class. Minimizing errors in gradient boosting is achieved by taking a 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 the introduction of new layers. The ultimate ensemble model predictions are obtained from averaging the predictions made by the prior three models. The gradient boosting detects the faults of weak learners utilizing gradients in the loss function (Dangeti 2017). The loss function evaluates how good the coefficients of a model at fitting the given data are.

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 a 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).

Evaluation metrics

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

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 et al. 2017). As values returned by log loss tend to zero then it will result in low uncertainty and thus the better the model:

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

where $\overset{\sim}{Y_j}$ denotes Log Loss function, $p\left(Y_j\right)$ is the predicted probability of the point being in the positive class for all k points, and Y_j represents the response variable.

Confusion matrix

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

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

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

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

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

Accuracy is the ratio of observations that are correctly predicted to the total observations (Dangeti 2017):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}.$$

Recall or sensitivity is the ratio of correctly predicted positive values to all values in true class (Dangeti 2017):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Precision is the ratio of observations that are correctly predicted positive to the total number of observations predicted as positive (Dangeti 2017):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

F1 Score is the weighted average of recall and precision score (Dangeti 2017):

$$\text{F1 Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}.$$

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.

Area under the Receiver-Operating Characteristic Curve (ROC AUC)

The ROC AUC is used to check the performance of classification problems at different threshold settings. ROC AUC is used to compare and evaluate the discrimination power of machine-learning models (Rodriguez and Rodriguez 2006). ROC denotes the probability curve on the other hand AUC signifies the ability of the model to separate between the classes. It shows the extent of the model's capability to distinguish between classes. The models used were good since their AUC tends to 1 and so is their ability to differentiate between the classes (Bowers and Zhou 2019).

Results

Figure 2 depicts ROC AUC for logistic regression, decision tree, random forest, gradient boost, and XGBoost. The results in Fig. 2 were evaluated on the test data.

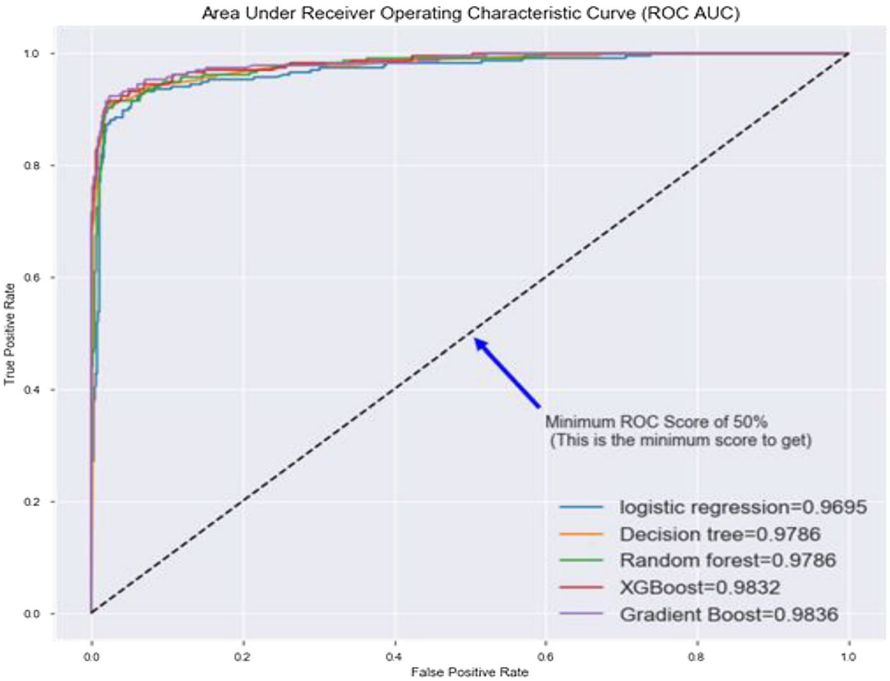

Fig. 2 ROC AUC performance comparison of five classifiers on the test data

Figure 2 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 would 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 Fig. 2, 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).

Precision, recall, F1 scores, accuracy, and log loss for test data

Table 1 depicts evaluation metrics for logistic regression, decision tree, random forest, gradient boost, and XGBoost. After training, test data were used to find the

Table 1 Evaluation's metric comparison of the five classifiers

	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

performance metrics of various tree-based models and logistic regression classifiers. The log loss for each of the models was: 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 1, the accuracy for each of the models was: 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 1, the F1 score for each of the models was: for each of the models were: gradient boosting (0.9508), XGBoost (0.9491), random forest (0.9461), decision tree (0.9437), and 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.

Feature importance

Table 2 shows the feature importance using a gradient boosting classifier. The classifier considers the sector of the business as the most important factor in determining the success of the business (0.6579) followed by duration (0.2189), tax type (0.0523), scale (0.0227), description (0.0191), fraud status (0.0144), business exports (0.0114), place (0.0023) and lastly level of income of the business (0.0010).

Table 3 depicts the feature importance using dummy variables. The classifier considers Industry sector (0.6259) as most important feature, followed by duration (0.2289), customs tax type (0.0461), services sector (0.0346), micro-scale (0.013), individual enterprises (0.00106), corporate (non-individual) enterprises (0.076), domestic (0.0067), other sector (0.0064), not committed tax fraud (0.038), WHT tax type (0.0031), other tax types (0.0031), committed tax fraud (0.0022), international (0.0022), rural areas (0.009), district cities (0.0008), urban areas (0.0007), CIT tax type (0.0006), agriculture sector (0.0006), low income (0.0005), small scale (0.0005), VAT tax type (0.0003), CITVAT tax type (0.0002), large scale (0.0002), medium scale (0.0001), pit taxpayers (0.0001), PAYE tax type (0.0002), medium income (0.0001), high income (0.0000), and PITVAT tax type (0.0000).

The feature importance of using gradient boosting was compared with that of logistic regression. The coefficients for gradient boosting were positive since it

Table 2 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
Business exports	0.0114
Place	0.0023
Level of income	0.0010

Table 3 Feature importance from gradient 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

returns the absolute values while those logistic regressions were both positive and negative. The positive scores indicate a feature that predicts class 1 (unsuccessful business), whereas the negative scores indicate a feature that predicts class 0 (business success). The higher value of this metric, when compared to another feature, implies it is more important 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.

Tables 3 and 4 depict the feature importance using logistic regression with dummy variables. The classifier considers customs tax type (1.6365), as most important feature, followed by duration (− 1.2947), Industry sector (0.9108), domestic (0.8781), services sector (− 0.8650), Withholding Tax type (WHT)

Table 4 Feature importance extracted from 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

(– 0.8632), international (– 0.8613) Value Added Tax type (VAT) (– 0.8034), individual enterprises (0.3826), corporate (non-individual) enterprises (– 0.3718), Corporate Income Tax type (CIT) (– 0.2977), agriculture sector (– 0.2225), large scale (– 0.1919), high income (– 0.1919), not committed tax fraud (0.1726), PIT taxpayers (– 0.1621), committed tax fraud (– 0.1608), other sector (0.1426), Pay as you earn tax type (PAYE) (– 0.111), Personal Income Tax and Value Added Tax type (PITVAT) (– 0.1014), low income (0.0894), other tax types (0.0037), Corporate Income Tax and Value Added Tax type (CITVAT) (– 0.0805), micro-scale (0.0747), medium scale (0.0713), medium income (0.0713), urban areas (0.0342), small scale (– 0.00293), rural areas (– 0.00265), and district cities (– 0.0028).

Discussion

In this study, decision tree, random forest, gradient boost, XGBoost, and logistic regression were compared as to their respective performance in predicting the success of a business. A desirable classifier was taken as the one with a high recall score, F1 score, accuracy, and ROC AUC. 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 the one with the lowest log loss.

From the log loss results in Table 1, the most robust model is the 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 Table 1, the most robust model is the 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 1, the most robust model is the gradient boosting classifier since it has the highest recall score (0.9459). It was followed by XGBoost, random forest, decision tree, and lastly logistic regression. Since the focus of the study was to predict the success of the business, recall score was, therefore, an important metric since it gives the ratio of correctly predicted successful businesses to all values in true class. Recall score returns proportion of total relevant results classified by the algorithm. In this study, evaluating a model based on recall score, gradient boosting is, therefore, the most recommended model to predict business success.

From Fig. 2, gradient boosting has the highest ROC AUC (0.9836) on the test data. 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. Figures 3, 4 and Tables 2 and 3 show the feature importance of gradient

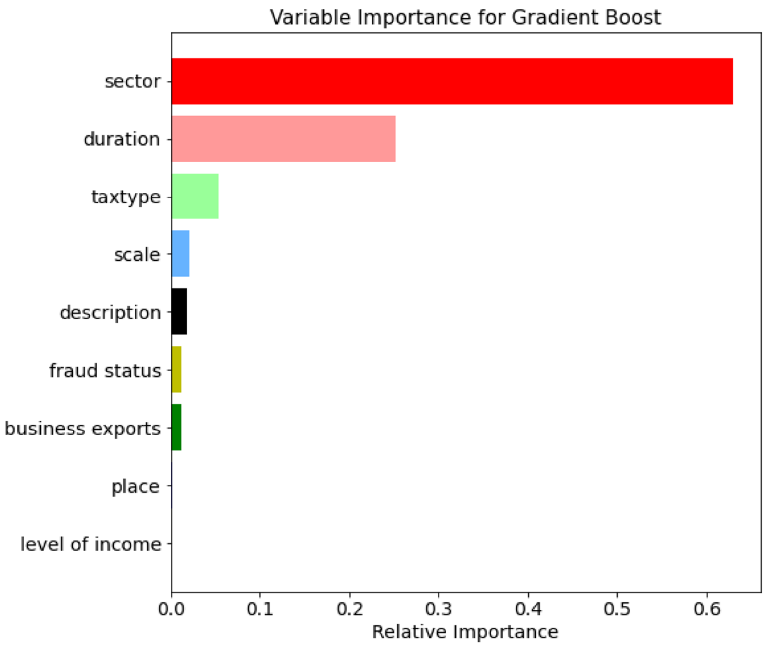

Fig. 3 Feature importance using gradient boosting

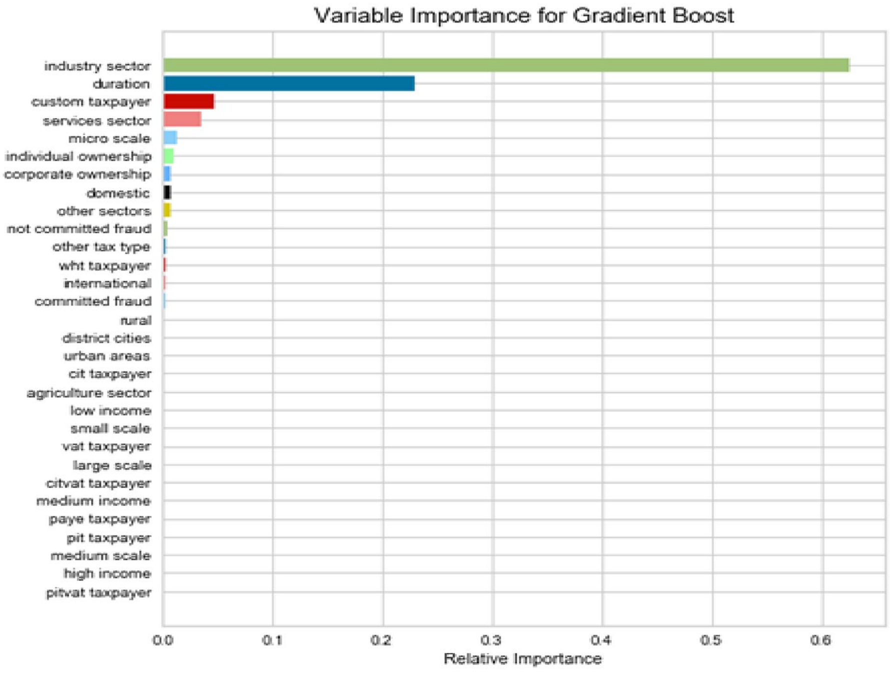

Fig. 4 Feature importance for dummy variables using gradient boosting

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 its feature importance will be considered in predicting determinants contributing to business success in Rwanda, however, it was ideal to compare with feature importance logistic regression since it depicts variable contributing to the success of the 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 the service were more successful since most businesses operate under this sector including international ones. Businesses in agricultural sectors were also more likely to succeed because almost the whole businesses in this sector are exempted from taxes. Thus, there is a need to put measures in place that will boost the growth of businesses operating in the industrial sector. The second most important feature that contributes to business success was the 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, Value Added Tax (VAT), and PAYE had a higher chance of success. The scale of business was the fourth most important

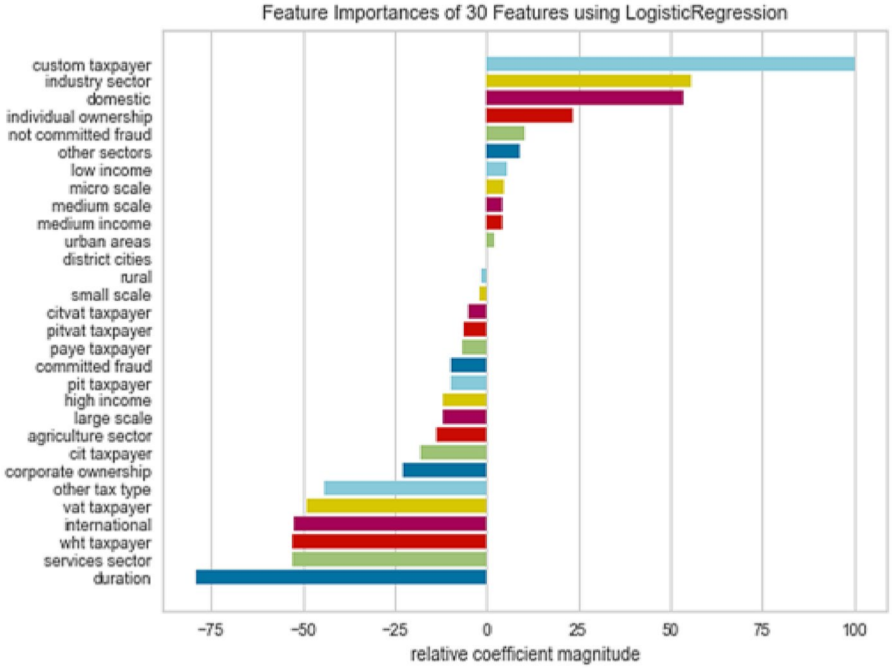

Fig. 5 Feature importance of dummy variables using logistic regression

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 to improve their growth and chances of survival (Fig. 5).

Conclusions

In this study, it was found that the sector was the most important feature that contributes to business success in Rwanda. Therefore, this study suggests further segmentation of sectors to identify other classes within the sector of the economy that could contribute to the success of the 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, 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. Ensemble tree-based models learn directly from high-level representations dataset, thus potentially evading traditional idiosyncratic threshold-based criteria of business success prediction. The techniques employed in this study serve as an example that could be automated, applied to other business success predictions. Future research should evaluate the ability of such methods to prospectively detect determinants of business success missed by non-tree-based models and to translate into improved business success predictions.

Author contributions FK conceptualized, analyzed, structured the methodology, did resource acquisition including downloading the needed software, did the validation, formal analysis, investigation, writing—original draft preparation, writing, visualization, visualization and prepared the draft manuscript. IK, as a principal advisor for FK, advised and supervised conceptualization, acquisition of the data, analysis, structuring the methodology, resource acquisition, writing - original draft preparation, writing, visualization, visualization and preparing of the draft manuscript. BF, and JN advised and supervised conceptualization, acquisition of the data, analysis, structuring the methodology, resource acquisition, writing, visualization, visualization and preparing of the draft manuscript. All authors have read and approved the final manuscript.

Funding Partial financial support was received from the African Center of Excellence in Data Science.

Data availability Data used was availed by the Rwanda Revenue Authority and may accessed upon reasonable request.

Code availability The python code used for this study is available on request. They shall be made publicly available on publication of this paper.

Declarations

Conflict of interest The authors declare that they have no conflict of interest or competing interest

References

Aqeel AMB, Awan AN, Riaz A (2011) Determinants of business success (an exploratory study). Int J Hum Resour Stud. https://doi.org/10.5296/ijhrs.v1i1.919

Avanzi F, Johnson RC, Oroza CA, Hirashima H, Maurer T, Yamaguchi S (2019) Insights into preferential flow snowpack runoff using random forest. Water Resour Res. https://doi.org/10.1029/2019WR024828

Ayandibu AO, Houghton J (2017) The role of Small and Medium Scale Enterprise in local economic development (LED). J Bus Retail Manage Res 11(2):133–139. https://doi.org/10.24052/JBRMR/262

Bartoloni E, Baussola M, Bagnato L (2020) Waiting for Godot? Success or failure of firms' growth in a panel of Italian manufacturing firms. Struct Chang Econ Dyn 55:259–275

Bayisenge R, Shengede H, Harimana Y, Bosco Karega J, Lukileni M, Nasrullah M, Xinrui H, Emmerance Nteziyaremye B (2020) Contribution of small and medium enterprises run by women in generating employment opportunity in Rwanda. Int J Bus Manag. https://doi.org/10.5539/ijbm.v15n3p14

Bowers AJ, Zhou X (2019) Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes. J Educ Stud Placed Risk. https://doi.org/10.1080/10824669.2018.1523734

Breiman L (2001) Random forests. Mach Learn. https://doi.org/10.1023/A:1010933404324

Casella G, Fienberg S, Olkin I (2013) An introduction to statistical learning. Springer Texts Statist. https://doi.org/10.1016/j.peva.2007.06.006

Clark LA, Pregibon D (2017) Tree-based models. Statistical Models 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. Packt Publishing

Flach PA (2003) The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: Proceedings, Twentieth International Conference on Machine Learning

Ganjisaffar Y, Caruana R, Lopes CV (2011) Bagging gradient-boosted trees for high precision, low variance ranking models. In: 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. J Forecast. 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

Kull M, Silva Filho TM, Flach P (2017) Beyond Sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electron J Stat. https://doi.org/10.1214/17-EJS1338SI

Mutandwa E, Taremwa NK, Tubanambazi T (2015) Determinants of business performance of small and medium size enterprises in Rwanda. J Dev Entrep. https://doi.org/10.1142/S1084946715500016

Nagaya N (2017) SME impact on output growth, case study of India. Palma J 16(13):11–170

Ramukumba T (2014) Overcoming SMEs challenges through critical success factors: A case of SMEs in the Western Cape Province, South Africa. Econ Bus Rev 16(1):19–38

Rodriguez A, Rodriguez PN (2006) Understanding and predicting sovereign debt rescheduling: a comparison of the areas under receiver operating characteristic curves. J Forecast. https://doi.org/10.1002/for.998

Sage AJ, Genschel U, Nettleton D (2020) Tree aggregation for random forest class probability estimation. Stat Anal Data Min. https://doi.org/10.1002/sam.11446

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Van Praag CM (2003) Business survival and success of young small business owners. Small Bus Econ. https://doi.org/10.1023/A:1024453200297

Żbikowski K, Antosiuk P (2021) A machine learning, bias-free approach for predicting business success using Crunchbase data. Inf Process Manag 58(4):102555

Zeng J (2017) Forecasting aggregates with disaggregate variables: does boosting help to select the most relevant predictors? J Forecast. https://doi.org/10.1002/for.2415

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