Credit Rating Models for Central Banks

Najla Al Barrak
Hessa Al Sanousi
Central Bank of Kuwait, P. O. Box: 526, Safat 13006, Kuwait
Department of Computational Engineering Sciences,
Cranfield University, Cranfield, Bedfordshire
MK43 0AL, UK

Irene Moulitsas
Department of Computational Engineering Sciences
Cranfield University, Cranfield, Bedfordshire
MK43 0AL, UK

Salvatore Filippone
University of Rome Tor Vergata
Via Cracovia, 50, 00133 Roma RM
Italy

Abstract

Credit default is the hottest topic in the banking sector. Central banks are in pressure to conduct studies in order to have a suitable and robust regulations. Recent research on Kuwaiti banks provide evidence that internal rating models are more efficient than the standard approach for predicting the rate of credit default. Having a comprehensive model for the banking system, for Kuwait as an example, will help central banks to test and evaluate the internal models available for the banks. Having two type of banking, conventional and Islamic, will make the system accommodatable in all jurisdictions. Through machine learning solutions and long skewed data, our research is comparing different Ensemble models to reach the appropriate model structure. We have covered two type of loans, instalment and consumer loans to overcome the difference of the use of such loans.

Keywords: Machine learning, Central bank, Credit default model, Conventional banks, Islamic banks.

Introduction

The banking sector in Kuwait faces several risks (credit, market and operational). The major threat is the credit risk which is our focus here: risk of not being paid for loans granted.Banks manage this risk by following the regulations of setting aside part of the profits as a capital to cover those potential risks when occurred. In calculating the portion deducted from profits for credit risk, banks need to multiply the risk weight of each credit transaction by its exposure.A critical issue for banks is calculating the appropriate risk weight for each customer; this can be done either by

- standard approach (fixed rates) or
- using internal rating systems (calculating the default probability per customer),

The Bank for International Settlements in Basel, Switzerland, developed the Basel III regulation in 2010. This rule stipulates those banks should have enough regulatory capital to cover 8% of their Risk Weighted Assets (RWAs). Of all the RWAs at Kuwaiti banks, 60.7% of them are loans (Central Bank of Kuwait, 2020). Those loans are granted for corporations, small- and medium-sized businesses, and households. Each of these categories has its own risk rating weight. There are two types of risk rating: the standard model, in which risk weights are fixed, conservative, and very high; and the internal rating model that banks use to calculate the appropriate risk weight per customer. In Kuwait, such risk weights are calculated using the standard model. Loans are considered unrated, so their risk weight is the highest. Our aim in this study is to come up with an internal credit rating system that uses machine learning to calculate the risk weights for households. We aim to create models that will be approved by the Central Bank of Kuwait and be acceptable to the banks that use them.

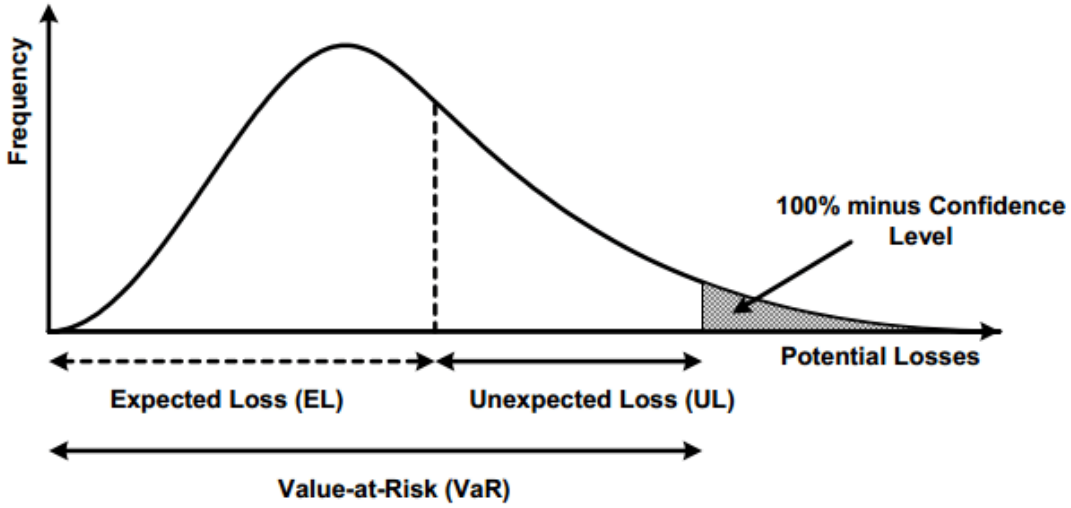
To calculate the final risk weight, banks can use the following equation:

$$Expectedloss = ProbabilityofDefault * ExposureAtDefault * LossGivenDefault)$$

From the probability of default, there will be categories of ratings that can be assigned specific risk weights. The main difficulty with applying Equation 1-1 is knowing what the probability of defaulting is. Banks can manipulate this to obtain low risk weights when estimating any unexpected losses.

From figure 1, we can calculate the percentage of unexpected loss, which is equal to the standard deviation of the expected loss. Expected losses are covered by the provision charges, whereas unexpected losses require capital charges to be anticipated (Basel Committe on Banking Supervision, 2005).

Figure **Error! No text of specified style in document.**: Expected and unexpected losses



Source: (Basel Committe on Banking Supervision, 2005: 3).

Our aim is to come up with a system that calculates the probability that a customer will default on a loan. It is hoped that the system will meet the approval of the Central Bank of Kuwait and be convenient for other banks to use with their existing data sets.

A critical problem for banks is how to calculate the appropriate risk weight for each customer. As mentioned, this can be done either by the standard approach (fixed rates) or by using an internal rating system (calculating the default probability per customer). Our research responds to the shortcomings of the standard approach by building a fruitful internal model and adding several new parameters to previous models. These new parameters are:

- The behaviour of consumers (net cash in and out), using sizes and types of data that have never been used for this purpose before.
- The number of transactions that customers make.
- The customers' credit card exposure.
- The length of the customers' relationship with the bank.

In any case, not all the work done in this context will be suitable for implementation in Kuwait unless it is properly customized for Kuwaiti residents. This is another major gap in the literature. Our research extends the current practice by collecting a significant amount of data that spans more than ten years and covers every local bank in Kuwait.

At this point, we should emphasise that 50% of the Kuwaiti banking sector is made up of Islamic banks, and 50% of it is made up of conventional banks. The credit rating models of Islamic banks have never been studied before. Conventional banks borrow money from depositors at a low interest rate and lend them to borrowers at a high interest rate. However, interest is forbidden in Islam, so Islamic banks enter into profit-sharing arrangements with both depositors and borrowers (Qian & Velayutham, 2017). Given the nature of the credit facility granted by Islamic banks, such financing is considered sales to be paid. The intention is that they will be fully repaid in the future. Hence, the existing models are best suited to deal with 'ordinary loans', rather than with Islamic products such as 'Murabaha' and 'Tawarruq'. Our model has been customized for these Islamic products. Finally, there has been a delay in research into Islamic banks as fewer people have adapted to this kind of banking. This is the first time that data for Islamic banking has been collected on such a broad scale, which will help to facilitate future studies.

Our research will tackle shortcomings of the standard approach by building fruitful internal model, we took some new parameters in addition to the previously used.

In any case, all work done in this context will not be suitable for implementation in Kuwait without proper customized testing for Kuwaiti residents; this is another major gap in the literature. Our research will extend the current practice by collecting a significant amount of data, spanning over ten years from all 8 local Kuwaiti banks. At this point we should emphasise that the Kuwaiti banking sector consists of 50% Islamic banks and for this type of banking has been studied before in(Albarrak, Alsanousi, Moulitsas, & Filippone, 2020).

There are several machine learning and deep learning options to inspect credit probability by default (Wang, Ma, Huang, & Xu, 2012). In a very recent International Monetary Fund working paper (Bazarbash, 2019), the advantages in financial technology are conferred. Specifically, the literature investigates how machine learning solutions could reduce the cost of credit and to provide much clearer solutions than the other templates for nontechnical audiences.

It is important to highlight that a credit rating model is fruitful for regulators and banks decision making evenly due to the advancement in the technology and the outperformance of the models developed and tested (Petropoulos, Siakoulis, Stavroulakis, & Klamargias, 2018). Baesens et al. (2003) compared 17 different methods for credit classes showing that it is suitable to conduct classification methods for credit rating. However, they did not use heavily imbalanced data in their case study. Imbalanced data in this context refers to data that is not well distributed between the two classes. The literature recommends that ensemble learning (gradient boosting decision trees) is a solution for solving the disadvantages of decision trees specially if the data is large and has long history, which is the case of our research (Bazarbash, 2019). The research done on similar work was on a period of three years (Petropoulos et al., 2018) while our aim is to expand it up to 11 years. Another important aspect is that most studies have relayed on same set of data gathered from the customer disregarding the data available with banks. Our study relied on data available with the banks currently (Nyathi, Ndlovu, Moyo, & Nyathi, 2014) and made use of it to estimate default customers.Therefore, this is the first time that such banking data is collected. The key performance indicator for my study is the Area Under Curve (AUC) for Receiver Operating Characteristic (ROC) curve. AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). We reached (results and literature review) that successful methods lay between Decision Tree and the three Ensembles.

It is important to highlight that for cross-validation we have used k-fold cross. It folds the dataset and takes various (random in default) portions of it to train the model, and then test it on the remaining instances, which is a pretty good way to check how good the model works for different train and test samples. We have picked a 70% training set and 30% test set. Given the nature of Ensemble methods that relay on subcategorizing the data in hand to several random groups to run the Random Forest Tree and reach for the almost accurate solution from each subgroup, the thing that could be considered as testing for overfitting.

The overfitting of boosting techniques is a topic that is not yet theoretically understood, but empirically results show that boosting seems to be very robust against overfitting (Breiman, 2001)(J. Friedman, Hastie, Rosset, Tibshirani, & Zhu, 2004). As a contribution to the used toolbox, we automated the analysis and set the comparison of the AUC as the key performance measure and the code will tell us what the best model is, given the 70% training sample and 30% holdout. The thing that will consider our work suitable for financial analysts and economists. In insuring that that our model can accommodate a large range of data, it is supported from the literature that in polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the selection is random. It has been found that even much smaller sample sizes will give very good results when tested against the confidence interval (Holmes, Illowsky, & Dean, 2017).

Materials and Methods:

Substantial data was recruited from all 10 Banks in Kuwait. Five of them are conventional and five are Islamic. We have requested and analysed an 11-year panel data from 2008 till 2018. Based on literature review, only some parameters like

- age,
- gender,
- nationality,
- income level,
- education and
- number of loans
- were tested and only for conventional banks. In this research, in addition to the previously mentioned parameters we have assessed, and for the first time for central banks as seen in (Albarrak et al., 2020),
- client-bank relationship duration,
- credit card exposure ‘outstanding balance’,
- monthly average net cashflow and
- monthly average number of transactions.

- We also included Islamic banks beside the conventional banks.

There are two type of loans in Kuwait, consumer and instalment. Consumer loans are defined as loans for the purpose of personal needs and durable goods with a limit of 25,000 KWD or 15 times the salary (whichever is less). Instalment loans are defined as loans for the purpose of maintenance or purchase of private residents with a limit of 70,000 KWD.

Our work is based on both type of loans. We chose to share the results of Instalment Loans first due to the higher amount granted and the economical factor of such loans since they are exposed to the real estate sector.

Default cases are cases in which the customer fails to meet their total monthly obligations 'loan payment and interest' for three consecutive months while there is a remaining outstanding balance. This means that the customer is defaulting (bad customers). Hence, the ongoing payments of monthly obligations are considered non-defaulting (good customers).

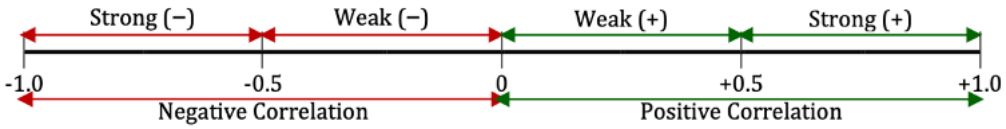
Several studies use correlation analysis to determine the level connection between research variables. Linear correlation analysis is a tool for representing the understanding of one associated variable to another. The linear correlation coefficient (r or R) is a measure offering evidence to how much two variables have association. The (Hauke & Kossowski, 2011) supports the Pearson's Product Moment Correlation Coefficient (R or r) as a measure of strength testing the linear connection between variables. Pearson's formula to quantify the degree of relationship (R) between variables can be given as

$$R = \frac{n(\sum XY) - (\sum X) \cdot (\sum Y)}{\sqrt{(n(\sum X^2) - (\sum X)^2)(n(\sum Y^2) - (\sum Y)^2)}}$$

Where
$n = Numberofobservations$
$X = Measuresofvariable\ 1$
$Y = Measuresofvariable\ 2$
$\sum XY = Sumofproductofrespectivevariablemeasures$
$\sum X = Sumofthemeasuresofvariable\ 1$
$\sum Y = Sumofthemeasuresofvariable\ 2$
$\sum X^2 = Sumofsquaredvaluesofthemeasuresofvariable\ 1$
$\sum Y^2 = Sumofsquaredvaluesofthemeasuresofvariable\ 2$

Based on the calculation result, the degree of correlative measure can be Positive, Zero or Negative correlation. If the sloping of a variable is positive and almost equal to another variable, there may be probability to have positive connection of each other and such association can provide positive correlation coefficient. And if the trend of a variable is almost negative to another variable such association can result in negative correlation coefficient. Basically, the coefficient of correlation R will range between -1 and +1. According to (Gogtay & Thatte, 2017), the correlation coefficient can be read based on its rate as in

Figure 2: **Error! Reference source not found.**Basic spectrum of interpreting correlation coefficient



Source (Gogtay and Thatte, 2017, p. 79)

An analysis of the correlation coefficient has been made for each of selected variables under consolidated data in regard to test the closeness of the variables for central bank model.

The aim is to develop the most appropriate credit rating model to predict households default rate in central banks. In order to achieve this, we made a comparison between different classification models; Logistic regression, Decision tree, Support vector machine, Bayesian network, Bag, AdaBoostM1 and RUSBoosted.

Logistic regression

The logistic regression model uses the logistic function to hold the output of a linear equation between 0 and 1

$$logistic(\eta) = \frac{1}{1 + exp(-\eta)}$$

The stage from linear regression to logistic regression is direct. In the linear regression model, it demonstrate the relationship between outcome and features with a linear equation

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}$$

For classification, we prefer odds to be between 0 and 1, so we replace the right side of the equation into the logistic function. Which forces the output to accept only values between 0 and 1

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}))}$$

The explanation of the weights in logistic regression differs from the explanation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1. In this study, we will differentiate between two classes of creditors, good and bad (Brown & Mues, 2012). The weights do not affect the probability linearly any longer. The weighted sum is distorted by the logistic function to a probability. Therefore, the equation is rewritten for the interpretation so that only the linear term is on the right side of the formula

$$log(\frac{P(y=1)}{1-P(y=1)}) = log(\frac{P(y=1)}{P(y=0)}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

The term called for the log function "odds" (probability of event divided by probability of no event) and called log odds.

This formula shows that the logistic regression model is a linear model for the log odds. However, we can work out how the prediction changes when one of the features x_j is changed by 1 unit. To reach that, we can first apply the exp() function to both sides of the equation

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Then we compare the chances when we increase one of the x's by 1. Except, instead of looking at the difference, we look at the ratio of the two predictions

$$\frac{odds_{x_j+1}}{odds} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_j (x_j + 1) + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p)}$$

While we apply the following rule

$$\frac{exp(a)}{exp(b)} = exp(a-b)$$

And we remove many terms

$$\frac{odds_{x_j+1}}{odds} = \exp\big(\beta_j (x_j + 1) - \beta_j x_j\big) = \exp(\beta_j)$$

In the end, we reach the fact that as exp() of a feature weight. A change in a feature by one unit changes the odds ratio (multiplicative) by a factor of $\exp(\beta_j)$. So, a change in x_j by one unit, increases the log odds ratio by the value of the corresponding weight.

Decision Tree

In classification decision trees, a single node is the preliminary point followed by binary differences (1,0). This results in the most indication about the class (Speybroeck, 2012). Then, the process is recurrent with the subsequent new node until we reach a place to stop. Usually the tree is too big, so it is back-tested through a cross-validation. The dependent variable Y is categorical in type, so, by using information theory in calculating how much we know about it from knowing the value of another separate variable A

$$\mathrm{I}[\mathrm{Y};\mathrm{A}] = \sum_a \Pr(A=a)\, I[Y; A=a]$$

I [Y ; A = a] is the value of the doubt about Y decreases from knowing that A = a given that we go from full population to sample where A = a. Therefore, I [Y ; A] is how much our uncertainty about Y reduces on average from knowing the value of A.

Support Vector Machines

By assuming a training set of N {(Xi, Yi)} Ni=1 with input data Xi ε Rn and consistent binary class labels Yi ε {-1+1}, the support vector machines classifier in Vapnik's theory satisfies the

$$\mathrm{I} yi[w^T \varphi(xi) + b] \geq 1, \quad i = 1, \ldots, N$$

The non-liner function of φ (.) designs the input space to a high dimensional feature space (Baesens et al., 2003). In this space created, the mentioned variations construct a hyperplane WT φ (X) + b = 0 differentiating between two classes. In the original weighted space, the following equation is used for the classifier

$$y(x) = sign[w^T \varphi(x) + b]$$

However, it is never evaluated in this form where the curved optimization problem could be defined as

$$\min_{w,b,\xi} j\,(w,b,\xi) = \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi i$$

subject to

$$\begin{cases} yi[w^T\varphi(x_i)+b] \geq 1-\xi_i, & i=1,\dots,N \\ \xi_i \geq 0, & i=1,\dots,N \end{cases}$$

The variables used in ξi are loose variables which are needed to allow the misclassifications to occur in the set of inequalities due to overlying distribution. The first section of the objective function is set to maximize the margin between two classes in the feature space. The second part is set to minimize the misclassification error.

Bayesian Network

Bayesian Network is a simple and high performance classifier (Baesens et al., 2003). This classification model works through learning the class condition probability p (Xi|Y) from each input variable Xi I = 1 … n given the class label Y. Then, a new observation is classified by Bayes' rule to calculate the following probability of each class of Y given the vector of observed feature values

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

To make things easier, an assumption behind the naïve Bayes classifier is that the features are in theory independent given the class label, therefore

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

The probabilities p (Xi|Y) are then predictable through using the frequency counts for the distinct features and a normal based method for the continues features.

Bag

Bootstrap aggregating, often abbreviated as bagging, requires that each model in the ensemble has equal weight of the random forest tree. In order to endorse model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set with bagging to achieve very high classification accuracy (Breiman, 1996).

In bagging, the samples are shaped in such a way that the samples are different from each other, however, replacement is allowed. In this case, an instance may appear in several samples, or it may not show in some of them. These samples are then given to several learners and then the results from each learner are combined in the form of voting.

AdaBoost

The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier (J. Friedman, Tibshirani, & Hastie, 2000). Originally, AdaBoost was designed in such a way that at every step the sample distribution was modified to put more weight on misclassified samples and less weight on correctly classified samples. The final prediction is a weighted average of all the weak learners, where more weight is placed on stronger learners

$$\sum_{n=1}^{N} w_n \ \exp(-y_n f(X_n))$$

where

- yn ∈ {–1,+1} is the true class label.
- wn are observation weights normalized to add up to 1.
- f(xn) ∈ (–∞,+∞) is the predicted classification score.

The gradient boosting method which is an ensemble algorithm proposed by (J. H. Friedman, 2002). It relies on incremental minimization of the error term which improves the accuracy of the prediction function (Brown & Mues, 2012). After setting the learner base, every tree calculated is fit to the 'pseudo residual', which is the deviation from the median and not from the expectation, from the earlier predictions in order to lower the error in general.

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots + \beta_n T_n(x)$$

G_0 is the initial value for the set. T_1 … T_n are the trees and β_1 … β_n are the coefficients for specific tree nodes calculated by the algorithm.

RUSBoost

Random under-sampling boosting (RUSBoost) is especially effective at classifying imbalanced data. It works by setting K classes, then, for each weak learner in the ensemble, RUSBoost takes a subgroup of the data with N observations from each of the K classes. The boosting process follows the technique in AdaBoostM1 for reweighting and constructing the ensemble (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2008).

Due to aggregation and the huge number of observations along with the skewed data sets, we ran the models and focused on almost all Ensemble methods (Bag, AdaBoostM1, RUSBoost, LogitBoost, GentelBoost and Robust Boost) (MATLAB, n.d.). The reason behind that is those later methods are a combination of several approved methods with more effective enhancement especially when we compare the AUC results from our data. As well, they are based on several Random Forest Trees that have been approved to be efficient for our data (Albarrak et al., 2020).

LogitBoost

LogitBoost works likewise to AdaBoostM1, except it minimizes binomial deviance (J. Friedman et al., 2000). Binomial deviance assigns less weight to badly misclassified items. LogitBoost can give better average accuracy than AdaBoostM1 for data with poorly divisible classes

$$\sum_{n=1}^{N} wnlog\left(1 + exp\left(-2ynf(xn)\right)\right)$$

where

- yn ∈ {–1,+1} is the true class label.
- wn are observation weights normalized to add up to 1
- f(xn) ∈ (–∞,+∞) is the predicted classification score

Binomial deviance assigns less weight to extremely misclassified observations (observations with large negative values $ofynf(xn)$).

Learner t in a LogitBoost ensemble fits a regression model to response values

$$\tilde{y}_n = \frac{y_n^* - p_t(x_n)}{p_t(x_n)(1 - p_t(x_n))}$$

where

- y_n^*∈ {0,+1} are re-labelled classes (0 instead of –1)
- $p_t(x_n)$ is the current ensemble estimate of the probability for observation x_n to be of class 1

GentleBoost

Combines features of AdaBoostM1 and LogitBoost. Like AdaBoostM1, GentleBoost minimizes the exponential loss, but its numeric optimization is set up differently (J. Friedman et al., 2000). Like LogitBoost, every weak learner fits a regression model to response values **Error! Reference source not found.**

$$\sum_{n=1}^{N} d_n^{(t)}(\tilde{y}_n - h_t(X_n))^2$$

where

- d(t)n are observation weights at step t (the weights add up to 1).
- ht(xn) are predictions of the regression model ht fitted to response values yn.

RobustBoost

Boosting algorithms such as AdaBoostM1 and LogitBoost increase weights for misclassified observations at every boosting phase. These weights can become very large. Therefore, boosting algorithm infrequently emphases on a few misclassified observations and indifferences the bulk of the training data. Accordingly, the average classification accuracy suffers. Unlike AdaBoostM1 and LogitBoost, RobustBoost does not minimize a loss function. Instead, it maximizes the number of observations with the classification margin above a certain threshold (Freund, 2009).

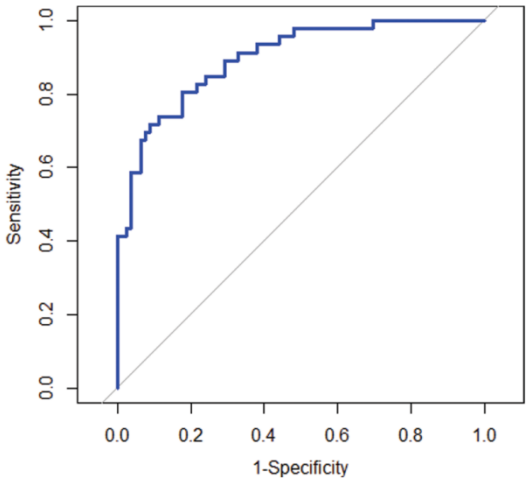
RobustBoost trains based on time evolution. The algorithm starts at t = 0; at every step, RobustBoost solves an optimization problem to find a positive step in time Δt and a steady positive change in the average margin for training data Δm. RobustBoost stops training and exits if at least one of these three conditions is true,

- time t ranges to 1,
- cannot find a solution to the optimization problem with positive updates Δt and Δm

- grows to as many learners as you requested
- Results can be usable for any expiry condition. It is advised to use cross validation to estimate the classification accuracy.

The selection of the appropriate model will be through evaluation of the Area Under Curve (AUC) for the Receiver Operating Characteristic (ROC) curve (Yang & Berdine, 2017). To better interpret these results, we could define the ROC as a curve connecting more predictive data points starting from less than the lowest value observed, (0,0), and ends at greater than the highest value observed (1,1).

Figure 3: Hypothetical ROC curve



Source: The receiver operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles, 5(19), 34.

After drawing the ROC curve, the AUC is the whole area underneath the ROC curve. The diagonal line between points (0,0) and (1,1) designates that the values on this line are not symbolic of a better estimate than a random guess (AUC = 0.50). The more the point in the ROC space above the diagonal line, the better the predictive value of the test.

Next, an investigation on true positive rates and false negative rates, from the confusion matrix, is conducted to improve our analysis. Through the information developed in the confusion matrix in, we will construct our analysis on calculating the accuracy rate, type I error and type II error.

Table 1: Confusion matrix for credit scoring

		Actual condition	
		Positive (non-risk)	Negative (risk)
Test result	Positive (non-risk)	True positive (TP)	False positive (FP)
	Negative (risk)	False negative (FN)	True negative (TN)

The three measures are defined as

$$Averageaccuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$TypeIError = \frac{FN}{TP + FN}$$

$$TypeIIError = \frac{FP}{TN + FP}$$

For cross-validation, we have used k-fold for it folds the dataset and takes various (random in default) portions of it to train the model, and then test it on the remaining instances, which is a pretty good way to check how good the model works for different train and test samples. We have picked a 70% training set and 30% test set.

Results

Form all 7 banks, the correlation table 2 for instalment loans indicate that, as expected, the strongest correlation between variables is between relationship duration and age which has a positive linear relation of 0.49. As well, there is a slightly negative linear relationship between credit card exposure and income with a rate of -0.30. The correlation data shows that there is no relation between average monthly net cashflow and the rest of the variables.

Table 2: Correlation coefficients between selected variables for instalment loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthlynumtx'	'mthlyavecf'	'count'
'exposure'	1.00	-0.30	-0.09	-0.09	-0.02	0.00	-0.01	0.03
'Income'	-0.30	1.00	0.13	0.15	0.10	0.15	0.01	0.04
'age'	-0.09	0.13	1.00	0.04	0.49	-0.09	0.00	-0.04
'gender'	-0.09	0.15	0.04	1.00	0.00	0.01	0.00	0.06
'Relationship duration'	-0.02	0.10	0.49	0.00	1.00	0.15	0.00	-0.10
'mthlynumtx'	0.00	0.15	-0.09	0.01	0.15	1.00	0.00	0.03
'mthlyavecf'	-0.01	0.01	0.00	0.00	0.00	0.00	1.00	0.00
'count'	0.03	0.04	-0.04	0.06	-0.10	0.03	0.00	1.00

The correlation

Table for consumer loans is inline as well where the strongest correlation between variables is between relationship duration and age with a positive linear relation of 0.50. As well, there is a slightly negative linear relationship between credit card exposure and income with a rate of -0.36. The correlation data supports the addition of the new to the research variable, average monthly net cashflow, which has 0 relation with the rest of the variables.

Table 3: Correlation coefficients between selected variables for consumer loans

	'exposure'	'Income'	'age'	'gender'	'Relationship duration'	'mthlynumtx'	'mthlyavecf'	'count'
'exposure'	1.00	-0.36	-0.10	-0.09	-0.09	-0.06	-0.01	0.06
'Income'	-0.36	1.00	0.09	0.13	0.15	0.23	0.00	-0.05
'age'	-0.10	0.09	1.00	0.11	0.50	-0.15	0.00	0.19
'gender'	-0.09	0.13	0.11	1.00	0.00	0.00	0.00	0.04
'Relationship duration'	-0.09	0.15	0.50	0.00	1.00	0.14	0.00	-0.01
'mthlynumtx'	-0.06	0.23	-0.15	0.00	0.14	1.00	0.00	-0.09
'mthlyavecf'	-0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
'count'	0.06	-0.05	0.19	0.04	-0.01	-0.09	0.00	1.00

In general, all the correlations conducted, based on previous studies and the referenced figure 1, indicate that all relations between variables are considered weak. Therefore, it is supported to include all listed variables as factors of predicting the probability of household loans default for both type of loans.

After combining 7 banks' data, between conventional and Islamic banks, bringing in 48,341 customers having instalment loans and 8,738 default customers with default rate of 18.07% as in table 4. Total number of observations reached 1,496,121.

Table 4: Number of observations and customers

All Banks	Instalment Loans
Observations	1,496,121
Default cases	28,530
Non-Default cases	1,467,591
Rate of default cases	1.907%
Number of customers	48,341
Number of default customers	8,738
Number of non- default customers	39,603
Rate of default customers	18.076%

After analysing the AUC for different methods it is as expected that Ensemble models, Bag, RUSBoost and AdaBoostM1, are the most effective models with close performance as in table 5.

Table 5: Area Under Curve results

All Banks	AUC Instalment loans
Logistic Regression	0.62
Decision Tree	0.74
Linear Support Vector Machines	0.54
Bayesian Network	0.67
Bagged	0.87
RUSBoosted	0.79
AdaBoostM1	0.79

To reach better conclusions, and after the result from the confusion matrix available in table 6, it is better to compare the performance of almost all Ensemble models where we included a combination of different models, LogitBoost and GentelBoost along with the existing models. A Robust Boost was also selected to add additional angel of different statistical performance to ensure that we have not missed its' abilities.

Table 6: Confusion Matrix results

All Banks	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	98%	N/A	100%	2%	N/A	2%	51%
Decision Tree	98%	83%	17%	2%	91%	2%	15%
Linear Support Vector Machines	98%	N/A	N/A	2%	N/A	2%	N/A
Bayesian Network	98%	6%	94%	2%	52%	2%	49%
Bagged	99%	87%	13%	1%	93%	1%	12%
RUSBoosted	99%	4%	96%	1%	52%	1%	49%
AdaBoostM1	98%	N/A	N/A	2%	N/A	2%	N/A

It is clear the ensemble models' running time are at acceptable range in table 7 given that this is the time execution for calculating the probability importance factors and training the mode. For prediction process, the enquiry does not exceed few seconds.

Table 7: Models running time

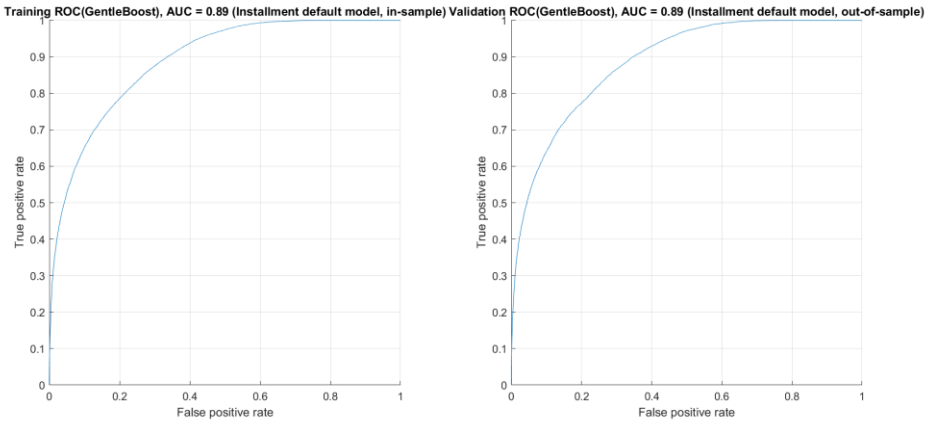
All Banks	Instalment Loans	
	Prediction Speed	Training Time
Logistic Reg.	~1800000obs/sec	72.58 sec
Tree	~1500000 obs/sec	72.159 sec
SVM	~5800 obs/sec	9112.6 sec
Bayesian	~940000 obs/sec	34.087 sec
RUSBoosted	~110000 obs/sec	269.86 sec
Bagged	~55000 obs/sec	2663.9 sec
AdaBoostM1	~120000 obs/sec	1102.2 sec

Therefore, we reached the conclusion that a deep comparison of almost all Ensemble models is more suitable for the Central Bank model. By comparing the AUC of the testing sets, GentleBoost is the most effective model for all banks data. We have picked a 70%, out of total data, to be our training in-sample to compute training accuracy that produced an AUC of 0.89.

While the remaining 30%, out of total data, is the testing (validation) out-of-sample to compute validation accuracy that produced an AUC of 0.89.

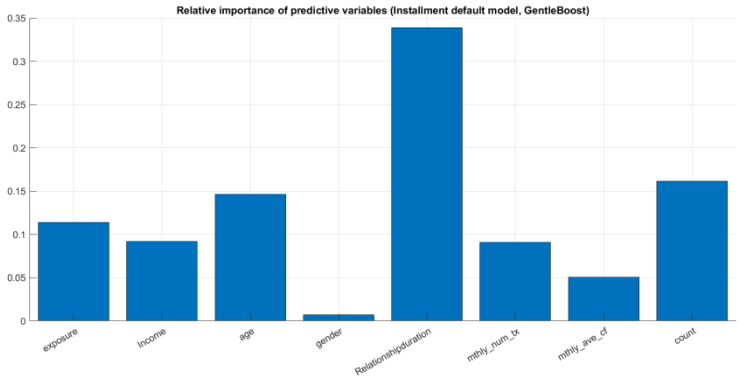
Figure 4: AUC results for 70% in-sample and 30% out-of-sample Figure 4 is an illustration from our in-sample and out-of-sample accuracy results.

Figure 4: AUC results for 70% in-sample and 30% out-of-sample



The following figure 5 shows the importance of input variables for our prediction model in regard to the central bank.

Figure 5: The relative importance of predictive variables



As expected, the relationship duration with the bank is the most influencer variable for predicting bad customers from good customers for both type of banking. Followed by number of loans that have been driven mostly from Islamic banks. Although monthly average number of transactions in the account was significant for bank-by-bank analysis, however, it lost momentum when we compared all banks data. Yet we would suggest individual banks to maintain this variable in their prediction model vis-à-vis central banks.

We have combined 7 banks data, between conventional and Islamic banks, bringing in 33,699 customers having consumer loans and 6,343 default customers with default rate of 18.82% and number of observations equal to 1,048,793 as in table 8.

Table 8: Number of observations and customers

All Banks	Consumer Loans
Observations	1,048,793
Default cases	21,756
Non-Default cases	1,027,037
Rate of default cases	2.074%
Number of customers	33,699
Number of default customers	6,343
Number of non- default customers	27,356
Rate of default customers	18.823%

After analysing the AUC for different methods, we can witness that Ensemble models, Bag, RUSBoosted and AdaBoost, are performing well with very close AUC results as in table 9.

Table 9: Area Under Curve results

All Banks	AUC Consumer loans
Logistic Regression	0.73
Decision Tree	0.82
Linear Support Vector Machines	0.5
Bayesian Network	0.71
Bagged	0.88
RUSBoosted	0.85
AdaBoostM1	0.85

From comparing the accuracy rate, type I error and type II error we can indicate that Bag has the most accuracy rate of 96% as in table 10.

Table 10: Confusion Matrix results

All Banks	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	98%	9%	91%	2%	54%	2%	48%
Decision Tree	98%	70%	30%	2%	84%	2%	23%
Linear Support Vector Machines	98%	N/A	N/A	2%	N/A	2%	N/A
Bayesian Network	98%	5%	95%	2%	52%	2%	49%
Bagged	99%	92%	8%	1%	96%	1%	7%
RUSBoosted	99%	9%	91%	1%	54%	1%	48%
AdaBoostM1	98%	17%	83%	2%	58%	2%	46%

From examining the efficiency of all models in general, and Bag in particular, it is stated that Bag model takes more than 25 minutes to operate as in table 1.

Table 11: Models running time

All Banks	Consumer Loans	
	Prediction Speed	Training Time
Logistic Reg.	~2300000 obs/sec	42.308 sec
Tree	~1900000 obs/sec	45.771 sec
SVM	~6700 obs/sec	5632.5 sec
Bayesian	~1000000 obs/sec	17.954 sec
RUSBoosted	~110000 obs/sec	184.68 sec
Bagged	~53000 obs/sec	1514 sec
AdaBoostM1	~120000 obs/sec	764.34 sec

Given that we are aiming to have robust model for central bank, it is advised to rerun the prediction model with focus on new Ensemble models which are the results of combining other models, namely LogitBoost and GentelBoost. Therefore, we reached a conclusion that a deep comparison of almost all Ensemble models is more suitable for the Central Bank model. In performing k-fold test as comparison of performance tool through comparing the AUC of the testing sets, GentleBoost is the most effective model for all banks data. We have picked a 70%, out of total data, to be our training in-sample to compute training accuracy that produced an AUC of 0.91. While the remaining 30%, out of total data, is the testing (validation) out-of-sample to compute validation accuracy that produced an AUC of 0.92. The figure 6 below is an illustration from our in-sample and out-of-sample accuracy results.

Figure6: AUC results for 70% in-sample and 30% out-of-sample

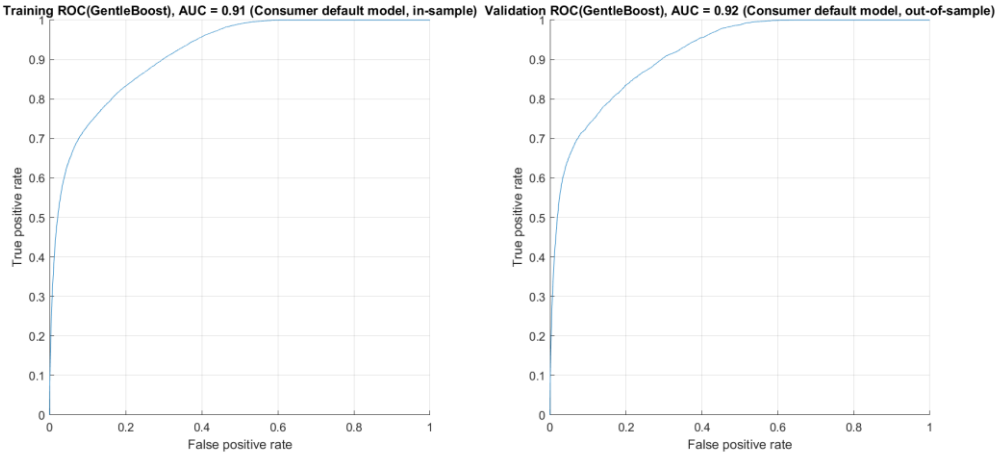
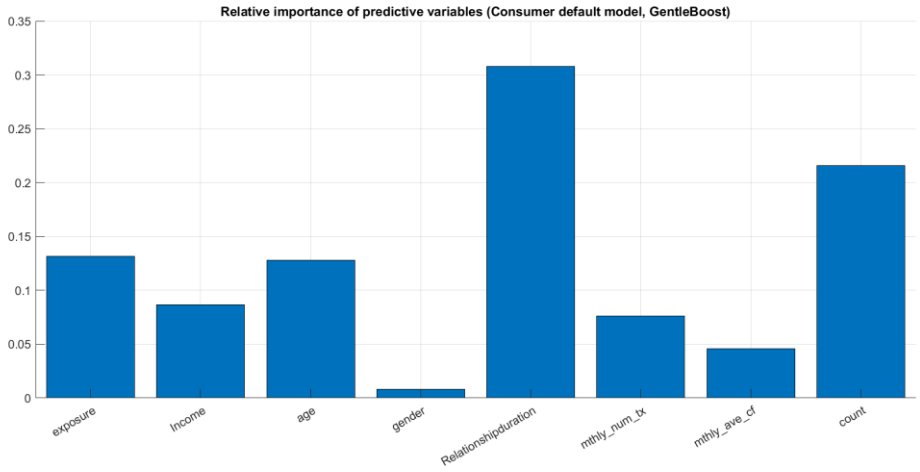


Figure7 shows the importance of input variables for our prediction model.

Figure7: The relative importance of predictive variables



As expected, the relationship duration with the bank is the most influencer variable for predicting bad customers from good customers for both type of banking. Followed by number of loans that have been driven from both type of baking.

Discussion

The results of our work have shown that the credit risk weight for each customer will be reduced for banks from the current standard approach in Kuwait for unrated customers of 75% of total loan amount to approximately 30% or less due to our internal based model. This information could help central banks to adjust its requirements and regulation on banks in calculating their capital adequacy.

This will also be used in assessing the level of liquidity needed for the banking system. Central banks can accommodate the lower capital adequacy requirements and allow banks to use the excess funds as a result in managing their liquidity, in which, reducing the reliance on central banks for liquidity supports. Nevertheless, the need for interbank activities could be reduced, hence, more efficient/reliable pricing for such activities.

The role of central banks will be enhanced given that there will be a robust system, with new undiscovered related variables, that calculates the credit default probability. When central banks have such thorough system, like the presented model, it will also be considered as risk assessment for the clients when applying for credit, adding more detailed system than what is currently used.

In evaluating and reviewing the relative importance of the independent variables selected for our model, central banks can consider those variables in their periodic stress testing by adding new impact factors for predicting default cases.

Nevertheless, the inspection of banks internal models would be more efficient when the inspection team have fair view of the selected variables and their importance weight. As well, a reasonable range of credit risk weight for specific clients with structured criteria could be approved and selected as a base for the banks to follow.

Conclusion

In general, classification methods are increasingly applied in fields other than computer science. The literature review is full of studies demonstrating the efficiency of such models in knowing the expected resulting different classes. Nevertheless, classification methods have been used to classify credit default classes, good or bad, in order to prepare regulators and bankers to better anticipate risks. We have compared different types of Ensemble models due to our 11 years, skewed, data. We have also tested for new variables than been used by central banks. after testing for the performance of the model through comparing the results of AUC, GentleBoost method was selected. The previous work done for the problem under study was conducted on a small range of data, the thing that, if enhanced, would provide more robust solutions and open-up for new models to perform better. We combined several methods by using the Ensemble models and testing for new Ensemble models for credit rating than what is available in the literature review. This is the first time such work is made for central bank of Kuwait, and due to the dual banking system available, our work could be adopted for both type of banking, conventional and Islamic.

References

Albarrak, N., Alsanousi, H., Moulitsas, I., & Filippone, S. (2020). USING BIG DATA TO COMPARE CLASSIFICATION MODELS FOR HOUSEHOLD CREDIT RATING IN KUWAIT. *International Journal of Soft Computing and Artificial Intelligence*, *8*(2), 1–7.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Basel Committe on Banking Supervision. (2005). *An Explanatory Note on the Basel II IRB Risk Weight Functions*. *Bank for International Settlements*.

Bazarbash, M. (2019). FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. *IMF Working Papers*, *19*(109), 1. https://doi.org/10.5089/9781498314428.001

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). *Random Forests*. Berkeley, California. Retrieved from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Central Bank of Kuwait. (2020). *Financial Stability Report*.

Freund, Y. (2009). A more robust boosting algorithm. Retrieved from http://arxiv.org/abs/0905.2138

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). Discussion of boosting papers. *Ann Stat*, *28*(2), 337–407. Retrieved from http://www.yaroslavvb.com/papers/friedman-discussion.pdf%5Cnpapers2://publication/uuid/C33BE497-39B4-4315-A405-3BBE029833BF

Friedman, J., Tibshirani, R., & Hastie, T. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016120463

Gogtay, N. J., & Thatte, U. M. (2017). Principles of correlation analysis. *Journal of Association of Physicians of India*, *65*(MARCH), 78–81.

Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Holmes, A., Illowsky, B., & Dean, S. (2017). *Introductory Business Statistics*. *openstax*.

MATLAB. (n.d.). Ensemble Algorithms - MATLAB & Simulink. Retrieved from https://www.mathworks.com/help/stats/ensemble-algorithms.html

Nyathi, K., Ndlovu, S., Moyo, S., & Nyathi, T. (2014). Optimisation of the Linear Probability Model for Credit Risk Management. *International Journal of Computer and Information Technology*, *03*(06), 1340–1345.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *The Use of Big Data Analytics and Artificial Intelligence in Central Banking*, *50*(August), 30–31. Retrieved from https://www.bis.org/ifc/publ/ifcb49_49.pdf

Qian, D. J., & Velayutham, S. (2017). Conventional Banking and Islamic Banking: Do the Different Philosophies Lead to Different Financial Outcomes? *Journal of Wealth Management & Financial Planning*, *4*(June), 3–14. Retrieved from https://www.islamicfinance.com/2014/12/

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. *Proceedings - International Conference on Pattern Recognition*, (December). https://doi.org/10.1109/icpr.2008.4761297

Speybroeck, N. (2012). Classification and regression trees. *International Journal of Public Health*, *57*(1), 243–246. https://doi.org/10.1007/s00038-011-0315-z

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61–68. https://doi.org/10.1016/j.knosys.2011.06.020

Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles*, *5*(19), 34. https://doi.org/10.12746/swrccc.v5i19.391