1. **Introduction**

*1.1 The Global Financial Crisis and Risk Management Practices*

The Global Financial Crisis (GFC) was the most severe international economic crisis since the Great Depression. The GFC had far-reaching consequences for the financial system, the housing market, the economy and wider society in general. It resulted in a recession leading to extremely high levels of unemployment in the United States and most of Europe, homelessness crises and general economic turmoil (Consequences of the Global Financial Crisis, Grant, Wyn (editor)). The crisis consisted of wide-ranging issues in credit markets centred in the United States of America, Europe and the United Kingdom. Housing bubbles in the United States, the United Kingdom, Spain and Ireland were caused by banks providing households with loans at very low interest rates. Households were able to borrow at levels that they could not realistically afford and purchase larger properties. Turmoil began to unfold when property prices began to drop and households found themselves in crippling levels of debt, in some cases their debt was larger than the value of the property. It later emerged that complex financial instruments had been built by large banks on these underlying household loans that were ultimately extinct (McKinsey). A quintessential component of the financial system is the prudential measures applied to combat against malpractice and avoid devastating impacts as observed by the fall out of the GFC. Following the GFC greater efforts were made by regulatory bodies to provide frameworks by which financial institutions must comply by to ensure sufficient and adequate risk management practices in lending money. The Basel III framework was published in June 2011 as part of the Basel Committee’s response to the GFC. It addresses several shortcomings in the pre-crisis regulatory framework and provides a foundation for a resilient banking system that will help avoid the build-up of systemic vulnerabilities. Basel III scrutinised the pre-crisis risk management practices of financial institutions. A particularly important component of the Basel III framework was the increase in the level of capital requirements to ensure that banks are sufficiently resilient to withstand losses in the times of stress. This increased scrutiny on the capital requirements for banks provided an impetus for institutions to improve risk management practices, one of which being the credit risk framework.

* 1. *Risk Weighted Assets and Credit Risk Modelling*

According to the CFA Institute, credit risk is the risk of loss resulting from a borrower failing to make full and timely payments of interest and/or principal (CFA). Credit risk is the biggest risk for banks. Other types of risks include market risk, operational and concentration risk. However, poor credit risk management can result in bank failure, and this was accentuated by failures of banks such as Northern Rock and Anglo-Irish bank during the GFC. Therefore, it is worthwhile for banks to implement sound and effective credit risk management systems/frameworks to remain sufficiently capitalised and resilient to economic shocks. Banks have developed sophisticated systems to model the credit risk arising from important aspects of their business lines. These sophisticated models are intended to aid banks in quantifying, aggregating, and managing risk across different segments of their disparate business services. The output produced by these models play a significantly important role in the risk management processes. The Basel Committee on Banking Supervision found that credit risk modelling may prove to result in better internal risk management (Basel paper 1999). The output that is produced by credit risk models are used as parameters to compute a bank’s level of capital requirements. Credit risk modelling practices assist in the computation of regulatory capital. Regulatory capital is a component of ‘Pillar One’ of the Basel capital framework, which sets out an institution’s minimum capital requirements. The quantitative regulatory capital under Pillar One has a core simple equation, to maintain financial resources equal to or greater than a percentage of its risk weighted assets (RWAs). Risk weighted assets (RWAs) are a measure of a bank’s assets adjusted for their perceived risks (Irish RWA paper). Capital requirements are in line with RWAs, for example the minimum risk-based capital requirements as established by the Basel framework is that total capital held by an institution must be at least 8% of RWA. The Basel Framework describes how to calculate RWA for credit risk. RWA for credit risk can be calculated either using the standardised approach or the internal ratings-based (IRB) approach. The IRB approaches are complex and require considerable resources as well as robust systems. Banks aspire to use IRB models as they are a mark of greater sophistication and usually, they produce lower risk calculation than the standardised approach (Irish RWA). There are three parameters in developing IRB models: Probability of Default (PD), Loss given default (LGD) and Exposure at default (EAD). This paper places a greater emphasis on the PD parameter.

* 1. *Probability of Default and Current Methods*

The probability of default quantifies the likelihood of a borrower that he will not be able to meet its contractual obligations and will default. Banks need to estimate rating-wise PD for the calculation of regulatory capital. The prime objective in modelling default risk is to measure credit risk in terms of default probabilities rather than ordinal rankings. PD can be estimated at a borrower, or an overall portfolio level and it depends on borrower-specific characteristics ranging from financial, behavioural, geographical or external market factors (Managing Portfolio Credit Risks in banks). Traditional statistical methods such as Logistic Regression or Decision Tree based methods are utilised by financial institutions for likelihood of default prediction. This paper compares advanced machine learning algorithms versus the typical statistical methods.

1. **Literature Review**

*2.1 Introduction*

There is a plethora of research outlining the use of machine learning concepts for credit risk models. The literature review included as part of this research proposal focusses on the development the Internal Ratings-Based approach to credit risk models, the early beginnings of various machine learning models and the relevant testing measures to compare the economic impact of each predictive model on capital requirements for financial institutions. Inspiration is drawn from various sources in particular to Alonso and Carbó 2020 in which they outline their methodology to understand the performance of machine learning models to predict credit default.

*2.2 Credit Risk & Probability of Default Definition*

Credit is the process of borrowing and lending capital. In order to compete with investment banking activities, commercial banks allow customers to acquire debt (Gande, 2008). There are numerous major risks faced by banks including credit, operational, market and liquidity risks. The most challenging risk associated with banking practices is Credit Risk. Loan performance affects profitability and the stability of bank growth and development. Credit risk is defined as the risk associated with a loan given by a bank, which will not be repaid-either partially or fully-on time (Campbell, 2007). One of the principle reasons of failure in the banking system, as accentuated by the Global Financial Crisis 2007 – 2008, is the persistent occurrence of non-performing loans on a financial institution’s balance sheet. Effective management of credit risk is a critical component of a comprehensive approach to risk management and essential to the long-term success of any banking organization (Basel, 2000). The Basel Committee on Banking Supervision issued its ‘Principles for the Management of Credit Risk’ in October 2000 encouraging banking supervisors to promote the sound practices for managing credit risk. Effective management of credit risk involves sound identification, measuring, monitoring and controlling of credit risk in order to determine the sufficient amount of capital to hold. This paper focusses on measuring credit risk. In January 2001, the Basel Committee on Banking Supervision published a consultative document on the ‘Internal Ratings-Based (IRB) approach’ for measuring credit risk. The paper stressed that this approach relies heavily upon a bank’s internal assessment of its counterparties and exposures. The two key objectives to be secured by the approach was additional risk sensitivity and incentive compatibility. The hope was that the IRB approach will provide a single framework by which a given set of risk components or ‘inputs’ are translated into minimum capital requirements. According to the Basel Committee on Banking Supervision, banks’ internal measures of credit risk are based on assessments of risk characteristics of both the borrower and the specific type of transaction. One such component to provide a measure of potential losses and, as such, credit risk is the Probability of Default of a borrower. The Probability of Default is one of the components that form the basic input to the IRB approach. According to (Grigutis, 2023) the Probability of Default (PD) is one of the most the important metrics in credit risk management. PD is a financial term that describes the likelihood of a borrower not being able to repay the obligations of a loan within a predefined timeframe. A higher PD indicates a higher risk of not repaying loan obligations, while a lower PD indicates a lower risk of not repaying loan obligations (Kenton, 2023). A bank’s estimate of PD to a certain exposure helps to map into a schedule of regulatory capital risk weights. Risk weighted assets (RWA) link the minimum amount of capital that a bank must have, to the risk of the profile of the bank’s asset book. The key role of RWA is to act as the measure of unexpected loss for banks portfolios (Lyons & Rice, 2022). Under the IRB framework, the risk inputs calculated using internal estimates, including PD, could map to different regulatory risk weights and produce a low level of risk allowing less capital held.

*2.3 Probability of Default Calculation Methods*

A classification is a method of predicting similar information from the value of a categorical target or categorical class variable (Reddy & Babu, 2018). Predicting borrowers default is a type of classification problem and, as such, classification methods are incorporated to predict a binary outcome. Logistic regression, linear-probability modelling and discriminant analysis are the techniques that are most commonly used in credit scoring. The considerations to take into account when selecting the techniques that is most appropriate for credit risk modelling are; ease of calculation, transparency, and whether or not it is statistically suitable for the problem. When credit scoring was first introduced in the 1950s and 1960s linear-probability modelling and linear discriminant analysis were the primary tools to measure risk. As time progressed, in tandem with the progression of computing power, maximum likelihood estimation (MLE) techniques became more feasible. Techniques such as logistic and Gaussian regression were adopted, both with less demanding statistical assumptions required compared to linear techniques. Logistic regression is used by between 80 and 90 per cent of scorecard (probability of default) developers (Anderson, 2007). Logistic regression is a well-known econometrics model in the banking industry (Alonso & Carbó, 2020). Research has been extended to model PD using alternative measures. Supriya et al. (2019) used Decision Trees as a machine learning tool to implement their model. Shoumo et al. (2019) used Support Vector Machines. Hamid and Ahmed (2016) used Data Mining techniques to build a model for classifying loan risk using three algorithms, J48, Bayes Net and Naïve Bayes. This paper will draw comparisons between traditional statistical modelling techniques to predict default (Logistic Regression) and more advanced machine learning methods to calculate PD.

*2.4 Logistic Regression*

Linear regression and linear probability modelling dominated early credit scoring techniques despite an obvious shortcoming, in that the target variable is binary. Logistic regression is more appropriate for credit scoring due to the prediction of a binary outcome. Logistic regression has its origins to the early 19th century for the description of the growth of populations and the course of autocatalytic chemical reactions (Cramer, 2003). Today, logistic regression uses a process called maximum likelihood estimation (MLE) to, transform the target variable into a log function, predict the coefficient size for independent variables, and determine the changes to coefficients, in order to maximize the likelihood of the log. It is a largely iterative and calculation-intensive approach, requiring almost six attempts to reach convergence, using one of several convergence criteria. The formulae for logistic regression can be found below:

A math equations with numbers and symbols

Description automatically generated with medium confidence

Logistic regression holds the following assumptions:

1. Categorical target variable;
2. Linear relationship with the log odds function;
3. Independent error terms;
4. Uncorrelated predictors;
5. Use of relevant variables.

Logistic regression is particularly effective due to its inherent ability to limit the predicted outcome between 0 and 1, and with the development of computer processing power has been adopted by major financial institutions in predicting loan outcome (Anderson, 2007). The emergence and development of Big Data, Machine Learning and Advanced Data Analytics methods has allowed for the possibility of more sophisticated tools to be used in the area of credit risk management (Alonso & Carbó, 2020).

*2.5 Machine Learning*

According to IBM, Machine Learning is a branch of artificial intelligence (AI) and computer science which focusses on the use of data and algorithms to imitate the way humans learn, gradually improving its accuracy. Machine learning is an important component of the growing field of data science. Through use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects (IBM, 2023). The origin of machine learning is associated with Frank Rosenblatt from Cornell University, who created a machine that was built for the purpose of recognizing letters of the alphabet (Rosenblatt, 1958). The machine, labelled ‘perceptron’, used both analog and discrete signals and became the prototype of modern artificial neural networks (ANN) (Fradkov, 2020). UC Berkely outline the three basic components of a typical supervised machine learning algorithm:

1. Decision process: calculations that attempt to predict the data pattern the algorithm is to find;
2. Error function: a method of measuring the prediction attempt by comparing to known examples (if available);
3. Optimization process: a method in which the algorithm can analyse the errors made and update how the decision process eventually selects a final decision.

Ayodele, 2010 outlines the main types of algorithms used in machine learning. These include; supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, transduction and learning to learn. Supervised learning entails learning a mapping between a set of input variables and an output variable and applying this mapping to predict the outputs for unseen data. Supervised learning is the most important methodology in machine learning (Cunningham et al. 2008). Unsupervised learning is where the machine simply receives inputs but obtains neither a supervised target, nor rewards from its environment. The goal of unsupervised learning is to build representations of the input variable that can be used for decision making, predicting future inputs or efficiently communicating the inputs to another machine (Ghahramani, 2003). Reinforcement learning is defined as the aim of building a machine that can learn from interacting with the environment (Sutton & Barto, 2018). The algorithm’s behavior is shaped through a sequence of rewards and penalties, which depend on whether its decisions towards a defined goal are correct or incorrect. This paper places emphasis on supervised machine learning algorithms mapping to the outcome of loan payment.

*2.6 Decision Tree Method*

A decision tree is a machine learning algorithm used for supervised learning. A decision tree simulates human thinking in tree structure for making a decision easily. Firstly, a decision tree picks an attribute to be the root node. Secondly, it creates a branch for each possible attribute value and separates the instance into multiple subsets. Thirdly, every subset links to a branch of the root node. Finally, the algorithm repeats the process iteratively on each branch (Kohavi, 2002). Decision trees are widely used in the banking industry due to their high accuracy and their ease of interpretability. Frydman et al. (1985) showed that results obtained through Recursive Partitioning Algorithm outperform discriminant analysis when using financial variables for credit risk measurement. Simha and Satchidananda (2006) applied both decision tree learning and logistic regression to predict default risk for agricultural loan information from India. The main conclusions were that decision trees generated good results and have the benefit of being easy to understand and implement. The figure below outlines a simple decision tree structure used for granting loan decision.

A diagram of a divorce

Description automatically generated

The Classification and Regression Trees (CART) algorithm is utilized as the learning algorithm in order to construct binary trees using the feature and threshold that yield the largest information gain at each node. Lewis (2000) introduces the concept of CART in order to create clinical decision rules found to perform as well or better than rules developed using methods that are more traditional. CART analysis consists of four basic steps. The first step involves building trees where each built tree is assigned a predicted class. The class is assigned irrespective of whether the node has been split into further child nodes. The second step involves stopping the tree building process, where a maximal tree is produced. The third step involves, ‘pruning’ a sequence of simpler trees by cutting off increasingly important nodes. The final step optimally selects trees, where the tree which fits information from the learning dataset, without overfitting, is selected among the new sequence of pruned trees. Luo, Hsu and Xu (2020) utilize this algorithm in the credit in an SME default prediction framework.

*2.7 Support Vector Machine*

Vapnik and coworkers first introduced a Support Vector Machine (SVM) for the purpose of classification. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVM is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. SVMs can be defined as systems which use hypothesis space of a linear function in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory (Jakkula, 2006). Chen et al. (2011) propose using a Support Vector Machine (SVM) to predict the default risk of German firms. In the paper, the performance of the SVM is compared to more traditional methods such as logit and discriminatory analysis. It is found that the nonlinear model classified by the SVM outperforms to logit model using the same predictors. The aim of the SVM algorithm is to find an optimal hyperplane that separates data points belonging to different classes. A kernel function maps input data point into a higher-dimensional space in order to find a hyperplane that can effectively separate non-linearly separable data in the original feature space. The optimal hyperplane is then described as the one that maximizes the distance between the hyperplane and the nearest data point from either class. The support vectors are defined as the data points that lie closest to the hyperplane and largely influence its placement. The figurebelow displays the output of SVM classification.



*2.8 Random Forest*

Random forests are a scheme proposed by Breiman (2001) to build a predictor ensemble using a set of decision trees that grows in randomly-selected subspaces of data. A random forest is a classifier consisting of a collection of tree-structured classifiers where the inputs are independent identically distributed random variables used to determine how the successive cuts are performed when building individual trees. The accuracy of a random forest depends on the strength of the individual tree classifiers and the measure of the dependence between them. For each input, the random forest procedure will be able to predict, with an optimal number of trees, the binary output value. A stopping criteria is used to minimize the prediction error (Addo et al., 2018). According to IBM, the key benefits of random forests are; a reduced risk of overfitting due to the robust number of decision trees in the forest lowers variance and error, providing flexibility as it can handle both classification and regression problems, and, easy to determine feature importance through metrics such as gini importance and mean decrease in impurity. Random forest have been applied in many disparate industries. Madaan et al. (2021) found that the random forest algorithm outperformed the decision tree model when prediction loan outcome for a peer-to-peer financial institution dataset. The figurebelow displays the structure of a random forest algorithm.

A diagram of a tree

Description automatically generated

*2.9 Artificial Neural Network*

Deep learning (DL) is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-driven decisions, particularly suited to contexts where the data available is complex and large (Kelleher, 2019). There have been three waves of developments in DL: cybernetics in the 1940s-1960s, connectionism in the 1980s-1990s and the current resurgence beginning in 2006 (Goodfellow et al., 2016). Haykin (2016) defines an Artificial Neural Network (ANN) as a massively parallel combination of simple processing units, which can acquire knowledge from an environment through a learning process and store the knowledge in its connections. Artificial/Feed-forward Neural Networks (ANNs) consist of an input layer, an output layer and a specific number of hidden layers. Neurons are contained in the input and output layers related to the number of input and output parameters. It is then an objective to determine the number of hidden layers and the number of neurons in each hidden layer. Activation functions are applied by each neuron in the hidden and output layers to the weighted sum of inputs. A characteristic of Artificial Neural Networks is the theory of Back Propagation. Back Propagation is a learning algorithm that is based on the Deepest-Descent technique. If the network is provided with an appropriate number of Hidden units, the network will also be able to minimize the error of nonlinear functions with high complexity. Practically, it is often necessary to provide ANNs with at least two layers of hidden units, when the function to compute is particularly complex, or when the chosen data, in order to train the back propagation are not reliable (Buscema, 1998). Artificial neural networks have been employed in various industries such as weather forecasting, medical diagnosis, aerospace, facial recognition, stock market and social media.

A diagram of a machine learning

Description automatically generated

Although modern software products enable relatively easy handling with artificial neural networks, their creation, optimization, and usage in real-life situations requires an understanding of the supporting theory and output required (Suzuki, 2011). Selection of the appropriate neural network architecture, hyperparameter tuning tuning, selection of the loss function, optimization algorithm and computational resources are a few of the considerations to take into account before developing a neural network. Bayraci and Susuz (2019) applied a Deep Neural Network to Turkish commercial bank data in order to predict loan default. It was found that the Deep Neural Network model significantly improved the performance of the credit scoring system when compared to logistic regression and support vector machine models for the loan application dataset. The performance metrics of choice included balanced accuracy and Type I and Type II error. However, the deep neural network did not outperform its counterparts in the loan performance dataset. It was found that traditional classification models may be preferred as opposed to more sophisticated deep learning models due to ease of implementation and interpretability. Byanjakar et al. (2015) found that a neural network-based credit scoring model performs effectively in screening default applications when applying such a model to peer-to-peer lending data.

*2.10 XGBoost*

Chen and Guestrin (2016) describe eXtreme Gradient Boosting (XGBoost) as a scalable machine learning system for tree boosting. It is a parallel tree boosting, also known as Gradient Boosted Decision Trees (GBDT) and Gradient Boosting Machine (GMB), which solve many data science problems in a fast and accurate way. In XGBoost individual trees are created using multiple cores and data is organized in order to minimize the lookup times. Schapire (2003) defines boosting as the machine learning method based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule. XGBoost has gradient boosting at its core, however, the difference between the simple gradient boosting algorithm and XGBoost algorithm is that, the process of adding weak learners does not happen successively. It takes a multi-thread approach whereby proper utilization of the CPU core of the machine are utilized, leading to greater speed and performance. XGBoost has been applied to credit risk in many academic papers. Wang (2022) applies an XGBoost model to build a personal credit risk evaluation model. It was found that the XGBoost model performs better than a decision tree and K-nearest neighbor model, based on AUC, accuracy and Type II error. Petropoulos et al. produced a paper where they used an XGBoost and deep learning neural network model to quantify the probability of default for a large dataset of corporate credit loans in the Greek banking system. The study’s conclusion found that the XGBoost model seemed to marginally outperform the deep neural network however there was greater flexibility offered by the neural networks, which may optimize the bias variance trade-off. XGBoost was selected as it can perform very well on any dataset size even with little features.

*2.11 Performance Measures*

Multiple models have been proposed as part of this research. Consequently, the performance of the proposed models will have to be compared against each other to formalize conclusions. There are many forms of performance measure in machine learning which depend on the type of problem to be solved. In our case the problem to be solved is loan default outcome which is, by definition, a binary classification problem. Standard performance measures of binary classifications problems include Type I Error, Type II Error, accuracy, precision, recall, AUC, confusion matrix etc. All these performance measures will be explored and reported. The performance metrics will be placed in sub categories based on their purpose. The sections will be accuracy-related metrics, discriminatory metrics and stability metrics.

*2.12 Accuracy-Related Performance Metrics*

Accuracy related metrics in machine learning measure the extent to which your model makes correct predictions of the target variable. The measure explored in this literally review is the Receiver Operating Characteristic (ROC) curve and consequently the Area under the Curve (AUC) metric.

*2.13 Receiver Operating Characteristic (ROC) Curve and Area under the Curve (AUC)*

The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are effective measures of the accuracy of model predictions. ROC curves are used to judge the discrimination ability of various statistical methods (Hanley & McNeill, 1982). Conventionally, a standard way of describing binary classification outcomes is by using a confusion matrix. A confusion matrix visualizes and summarizes the performance of a classification algorithm (Tiwari, 2022).



From the above confusion matrix the sensitivity of specificity of a model’s predictions can be calculated. The sensitivity of a model’s predictions is equal to the true positive fraction (TP/(TP+FN)). The specificity of a model’s predictions is equal to the true negative fraction (TN/(TN+FP)). These are performance measures in their own right but also are factored into the ROC curve measurement. The formal definition of the ROC curve is the plot of sensitivity versus 1-Specificity (Metz, 1978). One way to look at the ROC curve is to view it as a function built using two cumulative distribution functions (CDFs). With regards to credit risk modelling, the two CDFs used are credit scores relating to good and bad customers. Gönen and Haller (2010) summarizes the general formula for a ROC curve as below:



Where FB is a CDF of the scores of the bad customers, FG is a CDF of good customers, and FG-1 is its inverse.

A ROC curve can be summarized with its area under the curve (AUC). The AUC measures the discrimination power of the binary classifier. Generally, the higher the AUC, the better at predicting the outcomes the model is. The ROC curve is considered standard practice by credit-risk professionals (Anderson, 2007). The AUC is used to assess, improve, and compare credit risk models (Xia et al. 2020). With the ROC curve model, the path of the curve can be estimated based on sample results and the confidence intervals for AUC can be computed (Lahiri & Yang, 2018). However, computing the AUC in isolation might not be informative of the discriminatory power of the model. Řezáč and Koláček (2012) showed that some models excel at distinguishing good customers form bad ones dependent on the shape of the ROC curve. Tang and Chi (2005) show that a model with a much lower AUC outperformed a model with a higher AUC.

A diagram of a positive and negative test results

Description automatically generated

*2.14 Discriminatory Metrics*

The discriminatory power of a classification model is the ability to discriminate between the different classes, which serve as the output of the model.

*2.15 Gini Coefficient*

Gini coefficient is one of the methods that have been used to measure inequality in a population. It can be defined as the mean of the absolute differences between all pairs of individuals for some measure. In the context of credit risk, the gini coefficient can be computed by comparing the concentration of ‘bad’ borrowers (defaults) in the higher probabilities and ‘good’ borrowers in the lower probabilities. The gini coefficient is a rescaling of the area under the curve (AUC) metric from the receiver operating characteristic (ROC) curve.

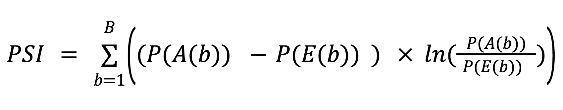
  
The gini coefficient is equivalent to a special case of Somers’ D statistic where one of the variables is binary and the other one is ordinal (Somers, 1962). The gini coefficient is a metric that is used by professionals to validate the credit risk models developed. It is recommended by the Basel Committee to be used. A gini coefficient of 50% indicates that the model outputs each value evenly and does not provide a better discriminatory model than a random guess. A gini coefficient of above 50% is considered useful and a better discriminator than random chance.

*2.16 Stability Metrics*

A crucial component during model development and testing is to assess whether the population experiences major changes.

*2.17 Population Stability Index (PSI)*

The population stability index (PSI) is a statistic that measures how much a variable has shifted over time, and is used to monitor applicability of a statistical model to the current population (Yurdakul & Naranjo, 2020). The formula for PSI can be found below:



The rules of thumb for PSI are; PSI < 0.1 means ‘little shift’, 0.1 < PSI < 0.25 means ‘moderate shift’, and PSI > 0.25 ‘significant shift’. The PSI is typically calculated for each independent variable in the model, however in this research paper it is proposed to apply PSI to the credit grades that are calculated as part of the probability of default model. The PSI is informative of a shift in the population of a model, however it does not inform model developers of the cause of the shift and further investigation is required to obtain this information.

*2.18 Feature Importance Measures*

A useful comparison between models will be to assess which features are important in each and perceive if certain features are better predictors for default than others. Numerous feature importance metrics will be used, dependent on model type. Coefficient size will be used for logistic regression, feature importance for random forest and decision and SHAP values for neural networks and XGBoost. SHAP values will be extended on in this literature review.

*2.19 Shapley Additive explanation values*

A recurring motif with machine learning models is the problems with interpretability of the models. It can be difficult to interpret the results of the model, in particular the key features that contribute to a model’s output. Lundberg and Lee (2017) introduced SHapley Additive exPlanations to address the problem of interpreting the predictions of complex models. With SHAP, an explanation model is used and represented by a linear model to determine an additive feature attribution method. According to C3.ai,the SHapley vlue is the average marginal contribution of a feature considering all possible combinations. SHapley values are computed for each individual prediction to assign which features are the most important in determining a particular output. These values can also be aggregated to a global level for each model and analyse which features are the most prominent. SHapley values are becoming more prominent in the banking industry as accentuated by their presence in the EBA’s response to their machine learning paper where Intesa Sanpaolo and the French Banking Federation mention their use of SHAP methodology (ref).

*2.20 Regulatory Capital Calculation*

For each type of exposure a bank has under its operations, risk weights are derived under a specific function. A risk-weighted asset is defined as the risk weight of a transaction multiplied by a measure f exposure for that transaction. Total risk weighted assets (RWA) are the sum of individual RWA across all transaction (BIS, 2001). As part of this research it is proposed to attach a tangible economic impact for the dataset under examination by calculating the capital requirement under each model. Baena et al. (2005) explains that theoretically statistical models with better predictive power have the potential to yield a better outcome in term of regulatory capital requirements. It is shown that the risk weighted function using the Internal Ratings-Based (IRB) approach is concave in PD. This implies that the capital requirement increases as PD increase. Therefore, if the predictive power of PD improved with more sophisticated models a lower capital requirement could be the outcome. Therefore, the capital requirements of advanced machine learning techniques will be assessed in terms of savings to regulatory capital holdings by banks. The formulas that are used to calculate capital requirements are included below (these may be subject to change dependent on the type of exposures included in the dataset, i.e mortgages vs corporate loans).

A math equation with black text

Description automatically generated with medium confidence

A mathematical equation with numbers and symbols

Description automatically generated



Various components of the above formulae will not be computed as part of this research (EAD, LGD). Loss-given default (LGD) can be benchmarked to a standardized figure published by the EBA dependent on the exposure class. Exposure at default (EAD) cannot be benchmarked and therefore the RWA figure cannot be calculated. As such, the capital requirement function will be used to compare different models.

*2.21 Conclusion*

There is a significant amount of research carried out on the potential applications of machine leaning techniques on predicting the outcome of borrower loan behavior for financial institutions. The above literature review explores multiple sources and methodologies upon which the realm of advanced data analytics has been extended into credit risk modelling. Upon investigation, a particular concept with little research are the actual economic impacts associated with these applications. Little research has been carried out to assess the financial impacts/trade-offs/consequences for banks in advancing their machine learning capabilities. This research paper intends to extend the outcome of machine learning applications from scientific measurement to a tangible measurement, namely capital requirement. In doing so, this paper may help answer the question as to whether it would be in the greater interest for financial institutions to invest time and resources in more sophisticated applications of artificial intelligence.