National College of Ireland

Assessment Methodology for Credit Models to Meet European Regulatory Expectations

MSc Research Project
Data Analytics

Albert Winston
Student ID: X20136331

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Albert Winston
Student ID:	X20136331
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	17/08/2022
Project Title:	Assessment Methodology for Credit Models to Meet European Regulatory Expectations
Word Count:	6724
Page Count:	30

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Albert Winston
Date:	18th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Assessment Methodology for Credit Models to Meet European Regulatory Expectations

Albert Winston
X20136331

Abstract

This research considers the assessment of machine learning models for banking in the context of European regulatory expectations. This includes requirements beyond predictive performance, is not well addressed in existing literature and prevents the wider industry adoption. Research has shown that such models may lead to capital savings for banks, but introduce complexity and cost to risk management.

The research developed a range of models for predicting credit default risk. These were subjected to an assessment approach that considered predictive performance, explainability at a local and global level, complexity and capital impacts. Under these additional headings we see that the preferred model may change depending on the regulatory focus. In addition while models may display good predictive ability it may be difficult to promote them under other criteria. The distribution of probability may lead to unexpected capital effects due to the non-linearity of that relationship.

Further work is required to consider how to optimally weight each aspect of the assessment criteria within a model risk management framework, however the research provides an important starting point for this.

1 Introduction

Credit default models within the banking industry have remained relatively unchanged in decades, with Logistic Regression classifiers being the dominant approach, as detailed by for example Siddiqi (2012). In recent years considerable research has been undertaken into the potential of more sophisticated approaches to improve the accuracy of credit default prediction, demonstrating the potential of such models (Yu (2020)). However regulatory challenges exist in terms of transparency, explainability and governance. As a result uptake has been slow. Recent discussions and guidance documents from regulators have identified expectations around more sophisticated modelling approaches

1.1 Background and Motivation

Credit default models are used by banks assess default risk [1] and for regulatory Capital and Impairment calculations. They are subject to strict regulatory requirements, including that they are part of business processes, transparent, understood by management and

[1] default risk is the risk that an obligor will fail to repay a loan in line with contractual terms potentially exposing the lender to loss

subject to validation and monitoring. These present challenges to the use of advanced machine learning models. While industry is aware of this problem Kurshan et al. (2020), research to date has been much more limited as noted by Dastile et al. (2020). As a result Alonso and Carbo (2021) highlights interpretability and governance as two frequent concerns of supervisors when reviewing Machine Learning models.

The Irish Central Bank notes that average risk weight density [2] is 33% for European Banks, and 49% for Irish Banks as at June 2021 [3]. Given Basel requires 8% of RWA to be held as capital under Pillar 1 there are benefits to managing the components of RWA calculations [4]. This has benefits both for the industry and wider society, with Das and Deb (2017) demonstrating that both the quantity and quality of bank lending is positively impacted by regulatory capital levels.

This research presents an assessment approach for Machine Learning models for Credit Scoring that addresses key European regulatory expectations. The motivation is to facilitate the adoption of complex modelling approaches for credit modelling within industry. This has significant financial implications given such models may be a key part of both credit and capital management. The approach presented offers an end-to-end view of the model lifecycle, with scope for further commercial development in the future.

1.2 Research Question, Objectives and Contributions

The research question identified the gaps in current research and demonstrated that additional criteria may be utilised in selecting the preferred modelling approach. The scope of the research presents broader assessment criteria for a range of models commonly applied to Credit default data

Research Question: *"How can complex Machine Learning Models be assessed in a way that allows a clear demonstration of the preferred model in terms of both understandability of the model, pure model performance, capital impacts and relative model complexity?"*

Sub-Research Question: *"For a selection of models that frequently demonstrate superior predictive performance to classical approaches, how can we demonstrate if these are still the preferred model when additional regulatory concerns are taken into account?"*

Sub-Research Question: *"Can this information be presented in a manner that enables senior management of an organisation to satisfy their requirements around understanding of the chosen model?"*

The research problem was addressed through research objectives detailed in Table 1. The models fitted reflects the most promising classifiers based on existing research. The simpler approaches represent the baseline which any challenger must outperform. As per regulatory guidance models must be explainable and increased complexity should be justified by performance. A simple Logistic and a Weight of Evidence based scorecard were developed representing two common classical approaches, along with Naive Bayes. Random forest and Boosted models represent ensemble methods and a neural network

[2] Risk weight density is defined as risk-weighted assets (RWA) expressed as a percentage of the banks total loan exposure

[3] https://www.centralbank.ie/docs/default-source/publications/financial-stability-notes/risk-weights-on-irish-mortgages.pdf

[4] the calculation for Risk Weighted Assets under the IRB approach is detailed in the Capital Requirements Regulation: Regulation EU 575/2013 (CRR), Chapter 2, section 2. For details see https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/504

was also developed. While Logistic Boost hasn't featured heavily in research it is included as an interesting variation.

Table 1: Description of Research Objectives

Detailed Research Objectives		
Ref	Objective	Description
1	Literature Review	Critical review of key literature related to the research
2	Data extraction, preprocessing and exploratory analysis	Data was extracted, subjected to exploratory analysis and cleaned. Feature extraction and generation was be completed. Class Imbalance was addressed.
3	Development of Credit Models	All models for assessment is produced at this stage
3.0	Logistic Regression	
3.1	Logistic Regression	
3.2	Random Forest	
3.3	Neural Network	
3.4	XGBoost	
3.5	AdaBoost	
3.6	LogitBoost	
3.7	Naive Bayes	
4	Model Evaluations and comparisons	Model performance assessed. Graphical outputs produced for model performance.
5	Global Explanations	Global explainability of the model assessed using SHAP. Relevant summary and visual information produced, including factor importance plots
6	Local Explainability	Local explainability assessed using SHAP for a selection of datapoints in the test sample.
7	Complexity Assessment	This is based on the predictive latency of the testing dataset for the fitted model.
8	Capital Assessment	Compares RWA for each model.
9	Database of results	Generate a SQL Server repository containing the key datasets produced
10	Dashboard of results	A Power BI dashboard produced as a MI tool to ensure senior management understanding.Technical dashboard with a wider range of metrics produced to support developers and validators

The remainder of the report is structured as follows. Chapter 2 considers peer-reviewed literature aligned to the research objectives. Chapter 3 lays out the scientific methodology applied in carrying out the research. Chapter 4 describes the Implementation approach that was followed as well as detailing preparation undertaken, while Chapter 5 evaluates the results obtained during the research. Chapter 6 provides an overall conclusion, as well as identifying possible areas for future research.

2 Related Work

2.1 Introduction

In reviewing the current literature, several subsections will be considered that look at the development and assessment of credit default models by financial institutions and the associated regulatory obligations. These are: (i) The application machine learning to credit models, (ii) regulatory expectations, including the challenges and opportunities this provides (iii) the existing research into model explainability and assessment approaches.

2.2 Application of Machine Learning to Credit Default

A number of different approaches to applying Machine Learning approaches have been considered. Work by Yu (2020) indicated that for Credit card data random forest produced superior results due to the inability of classical approaches to address the non-linearity inherent in the problem. Wang et al. (2020) also favours Random Forest, albeit their comparison doesn't include many of the more powerful ensemble and Deep Learning methods. The use of Deep Neural Networks to model Credit Card delinquency was explored by Sun and Vasarhalyi (2021). This demonstrated that DNN's could provide an effective approach, emphasising the importance of the hyperparameter choice, a topic regulators have also focused on in guidance documents. Further work on applying DNN is provided by Chishti and Awan (2019), who attempts to apply a range of deep learning models to Credit Card data. This research also highlights the ability of this modelling class to produce a probability of default output, necessary to utilise these models for capital calculations.

The use of other machine learning approaches has also been explored, with a lack of consensus as to the 'best' modelling approach to apply. Research by Chen et al. (2021) finds no clearly preferred approach. This might be expected given credit risk covers a range of risk appetites and Butaru et al. (2016) has demonstrated this diversity with models of Credit Card default. Even within a single dataset, competing research has championed a range of approaches. A summary of research is provided by Dastile et al. (2020), considering 74 studies. Based on this ensemble approaches appear to have the edge, with some support for deep learning. However even classical approaches have been championed in some studies. Notably research focuses on model performance in terms of predictive performance, with AUC and accuracy frequently used to select the best model. Typically the complexity and transparency of the models is not considered. Considerations such as regulatory and privacy concerns feature less prominently in research, and only a limited number of papers Addo et al. (2018) attempt to address this, highlighting the gap between the research approaches and the regulatory focus on transparency.

2.3 Regulatory Expectations for Credit Default Models

The early view of models risk management is captured in the Comptroller of the Currency's SR 11-07 document [5], which heavily influenced US and European thinking. Since Basel II the requirements for internal credit models have steadily grown and binding

[5]Board of Governors of the Federal Reserve System Supervision and Regulation Letters (2011) 'SR 11-7: Guidance on model risk management'

guidance and regulation has developed. European Regulatory authorities have significant expectations for the use of Advanced Analytics approaches. In their Report on Big Data and Advanced Analytics [6] the EBA detailed key requirements around ethics, explainability, auditability, fairness, data and consumer protection.

As noted by Kurshan et al. (2020), current governance approaches struggle to be effective, timely and cost -effective here. Existing manual approaches are insufficient to address model governance for AI with 67% of financial institutions perceiving regulatory complexity as a topic of concern due to the balance between model performance and compliance. The authors propose building regulatory verification into the modelling framework. Within industry this is echoed by key players, with McKinsey [7] noting that Model Risk management faces challenges and changes in the near future. Kokaly et al. (2016) has highlighted the cost of compliance, proposing an integration of the model management and software compliance aspects. Our research aims to apply a practical approach to this in the context of European regulations for credit default models.

In November 2021 the EBA issued their initial discussion paper on Machine Learning for IRB Models [8], outlining the expectations for Machine Learning models used for Internal Capital purposes. An emphasis is placed on strong internal understanding and a balance between performance and explainability - complexity should be justified by a predictive improvement and validation and monitoring of model changes is expected. This is in line with the view of national central banks [9]. Additional European Commission regulation and guidance, as detailed by Smuha (2019) defines key concepts such as explainability.

Research into meeting regulatory expectations has been slow to evolve. In their review Onay and Öztürk (2018) note that regulatory aspects of Big Data challenges represent the trajectory of future research. In considering 248 distinct research works they find that while 41% of research was concerned with statistical techniques, only 5% considered the theme of regulation. Data privacy and fairness considerations have led to the latter topic regaining popularity in more recent research. However recent research by Dastile et al. (2020) still finds a figure of only 8%. Alonso and Carbó (2020) has considered the problems faced from a supervisory perspective, proposing a cost function to capture supervisory risk tolerance. This focuses on the use of the IRB approach, generalising through the requirements of the Basel 'Use-test' [10]. However this approach, and the supervisory cost function employed, does not leverage how explainability for example, should be used to bridge the acceptability gap between classic and complex approaches. Further work in the area carried out by Alonso and Carbo (2021) indicates that savings in regulatory capital of up to 17% may be possible. The sensitivity of RWA to model changes is also considered here as required under regulatory materiality guidelines [11]. However once again there are gaps versus the expectations of the regulations.

[6] EBA Report on Big Data and Advanced Analytics, January 2020 EBA/REP/2020/01

[7] https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/banking-models-after-covid-19-taking-model-risk-management-to-the-next-level

[8] EBA Discussion Paper on Machine Learning for IRB Models, November 2021 EBA/DP/2021/04

[9] Machine learning in risk models – Characteristics and supervisory priorities July 2021 Deutsche Bundesbank

[10] Use test is a CRR requirement whereby capital models must also be embedded in the wider business activities of the institution

[11] EU 529/2014: regulatory technical standards for assessing the materiality of extensions and changes of the Internal Ratings Based Approach and the Advanced Measurement Approach

2.4 Approaches to Explainability and Model Assessment

Explainability of Machine Learning models represents a growing field of research interest. Current work by Phillips et al. (2020) for NIST has introduced key principles that would be expected of explainable AI which align well with the view of the EBA on big data, while Guidotti et al. (2018) has provided a summary of many of the explainability methods currently employed, albeit Lipton (2018) highlights the underspecified nature of the problem. A local methodology is LIME (Locally Inerpretable Model Explanation), as demonstrated by Magesh et al. (2020) and Dieber and Kirrane (2022), which is often favoured in the explanation of specific (local) outcomes.Lundberg and Lee (2017) presents an approach, based on the use of SHAP (Shapley Additive ExPlanations) which extends the idea of an explanation model to explain the outputs of the target model. This approach presents a unifying framework for the various approaches based on properties of the common approaches, including Shapley values and LIME.

The need for explainability is further extended to credit scoring models by Demajo et al. (2020) who highlight regulatory requirement, including local and global elements, and identifies the limited scope of application in existing research. The authors suggest a number of enhancements for future work, including the generation of an overall risk rating and greater ability to manage parameters of the explanations.

Research on model selection often focuses on standard performance metrics, e.g. Aleksandrova et al. (2021), Hurlin et al. (2018) rather than incorporating other criteria due the difficulty in quantifying aspects such as interpretatability islam2020towards. While Carrington et al. (2018) have made some progress here their work focuses on SVM models. There is limited consensus on how to assess the validity of the explainability outcome as noted by Vilone and Longo (2021). Research to incorporate other aspects are relatively recent, for example Ouedraogo (2021). This is necessary to to satisfy regulatory expectation.

2.5 Comparison and Critique of Approaches and Findings for Developing and Assessing Regulatory Credit Models

While there is an extensive range of research on modelling credit risk using machine learning approaches, there is no consensus as to the best approach as noted by Chen et al. (2021). A range of metrics are also used to select the preferred model, in spite of the class imbalance that regularly appears with default datasets. Research into the explainability and assessment of these models is far less prevalent. In Table 2 below the findings of the literature review are summarised.

Table 2: Summary of Key Findings from Literature Review.

Subsection	Sources	Key Findings
Application of Machine Learning to Credit Default	Yu (2020) Sun and Vasarhalyi (2021) Chishti and Awan (2019) Chen et al. (2021) Butaru et al. (2016) Dastile et al. (2020) Addo et al. (2018)	• Machine learning shows promise • No consensus on best approach • Focus on Model performance • limited work to incorporate regulatory and other aspects
Regulatory Expectations for Credit Default Models	Kurshan et al. (2020) Kokaly et al. (2016) Smuha (2019) Kumar and Gunjan (2020) Onay and Öztürk (2018) Dastile et al. (2020) Alonso and Carbó (2020) Alonso and Carbo (2021)	• Regulators have significant expectations for explainability and governance • Limited focus in existing research • Significant benefits identified from a governance perspective to ML models • This is a frequent regulatory concern
Approaches to Explainability and Model Assessment	Phillips et al. (2020) Guidotti et al. (2018) Lipton (2018) Lundberg and Lee (2017) Cornacchia et al. (2021) Magesh et al. (2020) Dieber and Kirrane (2022) Chen et al. (2018) Demajo et al. (2020) Bracke et al. (2019) Aleksandrova et al. (2021) Hurlin et al. (2018) Islam et al. (2020) Carrington et al. (2018) Vilone and Longo (2021) Ouedraogo (2021)	• A number of approaches to Explainability are available • A model agnostic approach at local and global level is required • Limited work so far to incorporate this or other criteria into model assessments

2.6 Conclusion

While model risk management is developing to consider regulatory changes, research still tends to consider only model performance metrics. Limited work exists to develop a risk assessment approach that incorporates other aspects into model selection. However

regulators have expectations that model assessment involve more than just performance. The research has identified potential approaches to deliver additional assessment criteria to inform model selection. However the development of such an assessment approach remains a gap. Given the increasing cost and complexity of model risk management this would represent a significant benefit to the industry.

3 Research Methodology for Assessment of Credit Models

This section details the process employed in carrying out the research. This has been divided into steps to cover Data details, data exploration and processing, feature engineering, candidate model development, model assessment and model Explanation.

3.1 Description of Methodology for Approach to Assessment of Credit Models

The process flow applied for the methodology is illustrated in Figure 1. As noted by Plotnikova et al. (2020) adaptions of project methodology to enable integration with business processes are common and reflect the motivation for the adaptions employed here. The approach builds on the Crisp-DM process developed for industry Shafique and Qaiser (2014)

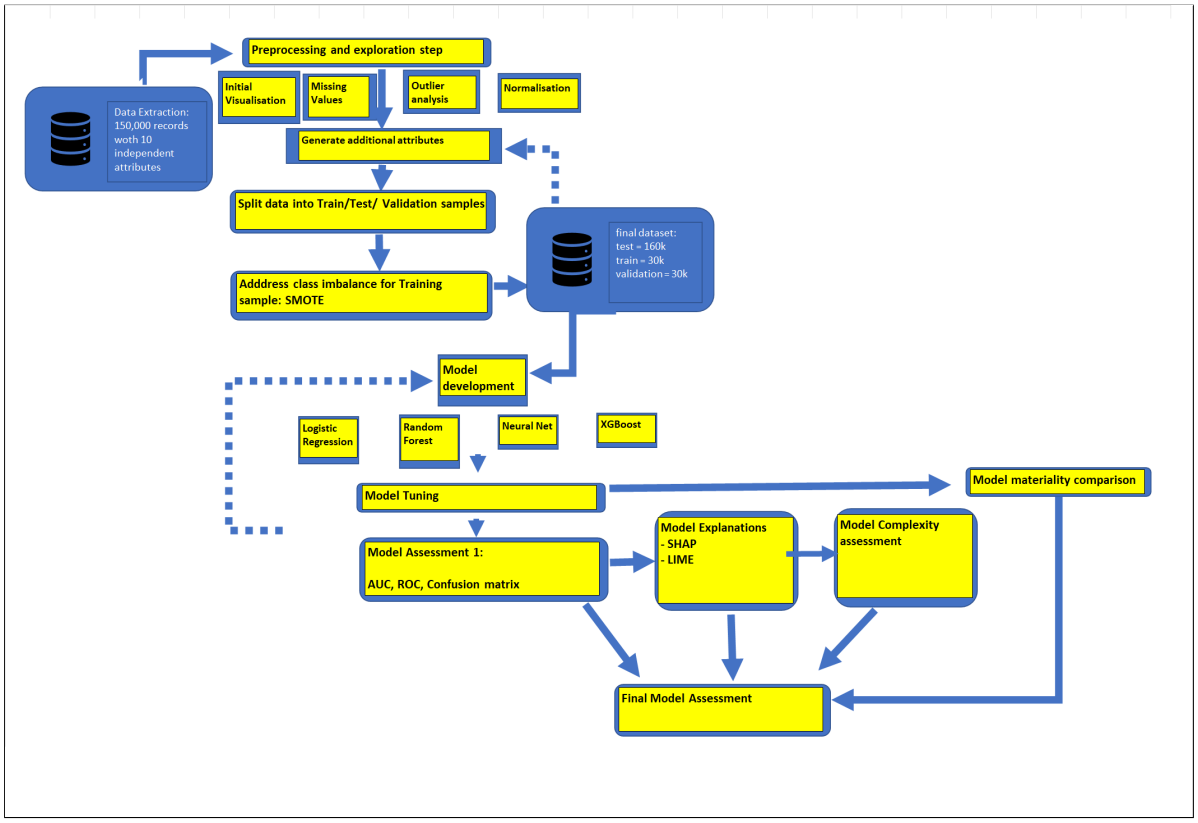

Figure 1: Process flow applied to research

3.2 Design Flow

For the research a two tier design process flow is applied. Tier (i) is a Presentation Tier representing the outputs from the Classification models, the model diagnostics and the exploratory analysis supporting the analysis. In practice this is technical reporting supporting model development or validation, and a management reporting. Tier (ii) is a business logic tier, representing the Feature extraction, data cleaning and the model development and assessment. This is illustrated in Figure 2

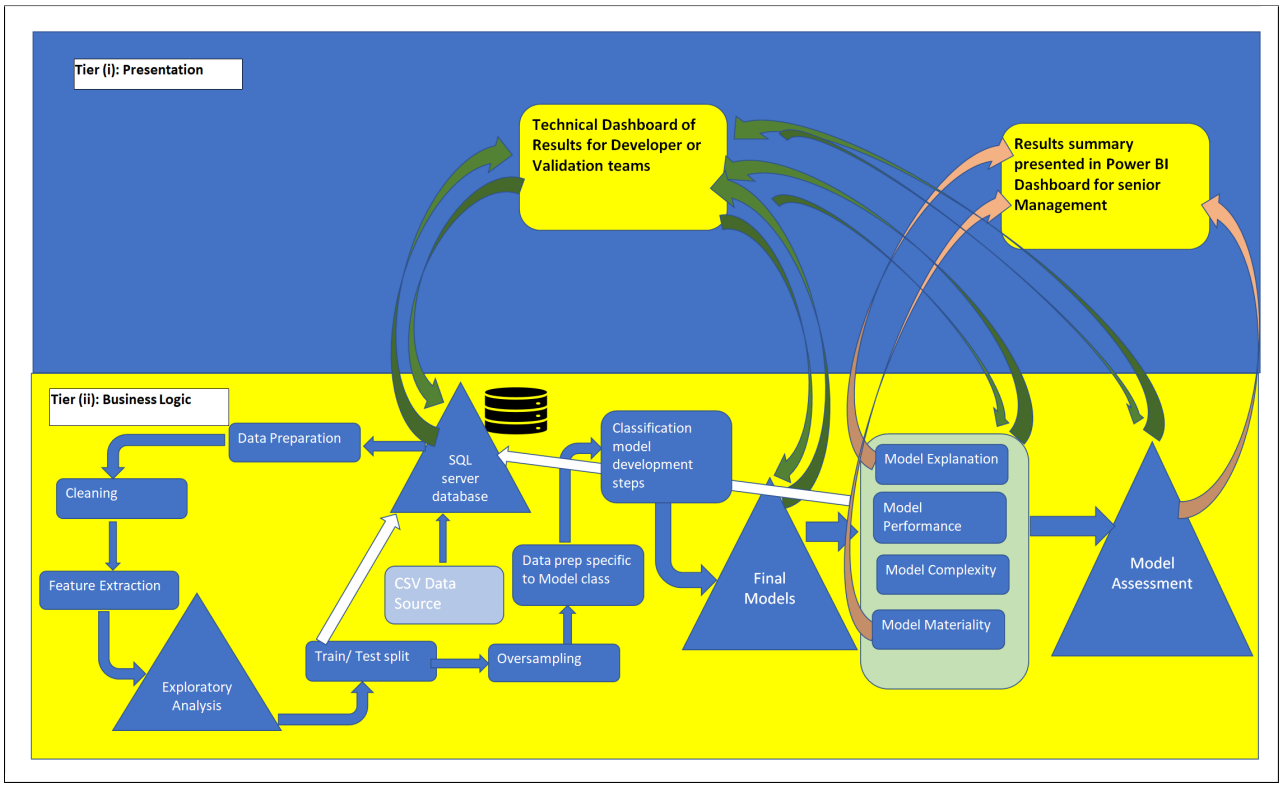

Figure 2: Design flow used in research

3.3 Data Collection and Description

Credit default modelling requires a binary classification dataset. For this research we have used a publicly available dataset, sourced from the Kaggle website. This was originally used in a competition in 2011, and can be found at: `https://www.kaggle.com/competitions/GiveMeSomeCredit/data`. The objective of this competition was to maximise model performance and as such transparency, explainability and complexity were not key considerations for the competitors. The data consists of a response variable representing cases that default in the two years post observation and 10 independent attributes. The total dataset size is 150,000 records and this is stored in a SQL server database. Summary informastion of the data is detailed Table 3.

Table 3: Summary description of attributes.

Attttribute	mean	std	min	max
RevolvingUtilizationOfUnsecuredLines	6.05	249.76	0	50,708
age	52.30	14.77	0	109
NumberOfTime30-59DaysPastDueNotWorse	0.42	4.19	0	98
DebtRatio	353.01	2037.82	0	329,664
MonthlyIncome	5348.14	13152.06	0	3,008,750
NumberOfOpenCreditLinesAndLoans	8.45	5.15	0	58
NumberOfTimes90DaysLate	0.27	4.17	0	98
NumberRealEstateLoansOrLines	1.02	1.13	0	54
NumberOfTime60-89DaysPastDueNotWorse	0.24	4.16	0	98
NumberOfDependents	0.74	1.11	0	20

3.3.1 Data Preparation and Cleaning

A number of variables displayed missing data. These were imputed using the MICE algorithm as originally proposed by Dempster et al. (1977). Work by for example Li et al. (2022) demonstrates this as an effective imputation method.

The dataset was split into Training, Test and Validation samples in a 60:20:20 rate. While the hyperparameter tuning applied cross validation, the Validation dataset is maintained separately for use as an additional dataset in the implemented solution.

Significant class imbalance was identified within the dataset. This could lead to biased models with poor accuracy over the minority class relative to the majority class As a result SMOTE oversampling Luque et al. (2019) was applied in the training sample.

3.3.2 Feature Selection

A number of variables displayed outliers and were cleaned as decribed in Table 4. However while extreme values were removed, plausible but high values were generally retained.

Table 4: Description of Data Cleaning Completed

Variable	Cleaning completed
Age	variable ranged between 0 and 109 years. Credit cannot be legally extended to customers ¡ 18 while extreme ages may be considered questionable. Data constrained in the range 18-80
Monthly income	Range from 0 to 3 million. Data constrained to an upper value of 100,000.
NumberOfTime60-89DaysPastDueNotWorse	Ranges from 0 to 98. Given this is over a 24 month period the upper value was limited at 25
NumberOfTimes90DaysLate	Similar approach to above
RevolvingUtilizationOfUnsecuredLines	This is a a ratio of credit used to available credit and as such should constrain at 1. To allow for overdraw on revolving credit lines the constraint was set to 1.5.
Debt ratio	Ranges between 0 and 329,664. Given this represents the level of debt to assets a high value indicates poor capacity to repay. This variable is constrained at 25,000

Following these steps the data was reviewed once again. While we still see skew within the distributions, the removal of implausible values clearly improved the attribute distributions.

A number of Ratio variables were also created to expand the feature set. Ratios are a common financial tool to represent creditwortiness Saygili et al. (2019). The following ratios were used:

Ratios looking at propensity to move to higher Arrears.

- (NumberOfTimes90DaysLate)/(NumberOfTime60- 89DaysPastDueNotWorse)
- (NumberOfTimes90DaysLate)/(NumberOfTime30- 59DaysPastDueNotWorse)
- (NumberOfTime60- 89DaysPastDueNotWorse)/(NumberOfTime30- 59DaysPastDueNotWorse)

Ratios looking at potential drags on available income:

- MonthlyIncome/ NumberOfDependents
- MonthlyIncome/ NumberOfOpenCreditLinesAndLoans
- MonthlyIncome/ NumberRealEstateLoansOrLines

In addition to Assess the capital impact associated with each model values for the Loss Given Default (LGD) and Exposure at Default (EAD) parameters used in capital calculation were simulated for each observation. The approach applied was as follows: For LGD a Beta Distribution was simulated, with parameters $\alpha = 1.5$ and $\beta = 10$. The Beta distribution is a distribution commonly used to model loss distributions De Servigny et al. (2004) and has form:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_{i=0}^{1}(t)^{\alpha-1}(1-t)^{\beta-1}dt}$$

with

$$E(X) = \frac{\alpha}{\alpha+\beta}, V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

For the distribution simulated this gives an average LGD of 15% and variance of 0.9%

For EAD a Gamma distribution was simulated Jiménez and Mencía (2009), with parameters k = 2 and $\theta = 200$. This distribution has form:

$$f(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} \; for x > 0 \alpha, \beta > 0$$

this will give an average loss given default of €400.

3.4 Modelling Approaches

the following Modelling approaches were applied to develop credit models. :Logistic regressions and Naive Bayes represent classical, explainable approaches while ensemble and Neural Net approaches represent more sophisticated approches. Any model must, as per EBA regulations, (i) be locally and globally transparent, (ii) have superior performance to the others and (iii) this improvement must be sufficient to overcome any increase in complexity seen. The EBA guidelines do not specify how this final point should be interpreted.

3.4.1 Logistic Regression

Logistic Regression represents an industry standard approach. Logistic regression represents a model from the Generalized Linear model class.

$$f(x) = \alpha + \sum \beta_i x_i$$

where $f(x)$ represents a logit link function

$$ln(\frac{p}{1-p})$$

Models in this form are most commonly seen in non-retail exposure classes where large scale scoring is not required. For this implementation the Python Package Scikit-Learn is used.

3.4.2 Logistic Regression using Binning based on weight of Evidence

This approach is often taken with Retail Exposure classes/. Each factor is binned into several bins

$$x_1, ...x_i$$

. For each level of an attribute a weight of evidence (WOE) is calculated and these are combined to produce information value (IV) for the attribute. The WOE values then form the inputs for the model.

$$WOE_i = ln(\frac{\%Goods_i}{\%Defaulted_i})$$

$$IV = \sum_{i=1}^{n} (\%Goods_i - \%Defaulted_i) * WOE_i$$

For this project the Python packages OptBinning and SciKit- Learn were used to generate an optimally binned set of attributes and generate a scorecard for these.

3.4.3 Naive Bayes

Naive Bayes in a relatively simple conditional probability model that leverages Bayes theorem and assumes that feature values are independent of the values of any other feature. Sckikit-Learn is used to implement.

3.4.4 Random Forest

An ensemble Learning method based on constructing a large number of Decision Tree classifiers. A bagging approach is typically applied to construct a large number of trees with controlled variance. As such Random Forest is a parallel learner.

The model was implemented in Scikit learn. Hyperparameter tuning was implemented using the Gridsearch and RandomSearch modules and the results of each estimation were saved to be used in assessing model performance versus complexity

3.4.5 Neural Network

Neural Networks were built using Tensorflow. We used a network with a single hidden layers with relu activation function and allowed the number of Epochs to reach up to 200. Early stopping was enabled where performance fails to increase across 10 successive runs, and a custom Tensorflow callback allows the capture of changes in performance and fit times as the number of epochs is increased.

3.4.6 XGBoost

XGBoost is a modelling approach that uses gradient boosted decision trees. It is deigned to be a fast and computationally efficient approach, which has produced strong results as a classifier approach. The model was implemented using the XGBoost library in Python. Hyperparameter tuning was implemented using cross-validation with the Gridsearch and RandomSearch modules and the results of each estimation were saved to provide information on the model evolution.

3.4.7 AdaBoost

Adaboost is an ensemble learning method based on iteratively improving the performance of weak learners. AdaBoost is a sequential learner using a boosting approach where the weak learners are decision tree stumps. It is equivalent to an additive tree regression that minimises an exponential loss function Hastie et al. (2009)

The model was implemented in Scikit learns AdaBoost Classifier package. Hyperparameter tuning was implemented with cross-validation using the Gridsearch and RandomSearch modules and the results were saved to provide information on the model evolution

3.4.8 LogitBoost

LogitBoost is another boosted ensemble learning approach. The approach was proposed by Friedman et al. (2000) and represents an additive tree regression that minimises a Logistic loss function. The model was implemented using the ogitBoost Package in Python, with hyperparameter tuning implemented using cross-validation with the Gridsearch and RandomSearch modules. Results from this were again saved.

3.5 Model Assessment

3.5.1 Model Performance

For each model fitted AUC was used as the primary performance assessment metric. This is consistent with industry practice Izzi et al. (2011), Engelmann (2006), and is typically employed ahead of accuracy due to class imbalance in default data. However a number of other metrics were also produced including Precision, recall and accuracy

3.5.2 Model Complexity

In general there is no agreement in the literature on a model agnostic approach to assessing model complexity. For this research complexity is considered in terms of the Prediction Latency - the time required to score a case using a given model. This also aligns to the results saved from the hyperparameter tuning. These results were also be collected into the final modelling Database to ensure the EBA expectation to consider the effect of model hyperparameters could be addressed

3.5.3 Model Explainability

Model explainability will be considered using SHAP (SHapley Additive exPlanations) values generated for each model. This will allow a model agnostic approach to comparing the contributors for each model, as well as explanation of the models at both a local and Global level. this in turn satisfies regulatory expectations to understand key model drivers as well as GDPR requirements to explain individual model predictions.

Shapley values present a game theoretic view of the global contribution of each factor, with Shapley values representing the fair allocations to each players, of gains in a cooperative game. LIME generates a locally weighted linear model on perturbed instances of the observation of interest. SHAP brings together these concepts through taking the original model and calculating Shapley values for a conditional expectation function.

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \backslash i)]$$

where $|z'|$ is the number of non-zero entries in z', $z' \subseteq x'$ is all vectors z' with non-zero entries as a subset of non zero entries in x'

3.5.4 Model Materiality

The RWA percentage associated with each record will also be calculated, using the formula contained within the Capital Requirements regulation, for each mode produced. For the purposes of the research the portfolio will be assumed to follow the Retail exposure class and the capital floors stated in the regulations will not be applied

4 Implementation of Process for Development and Assessment of Credit Models

4.1 Introduction

This section will describe the implementation of each stage of this research. It is divided into sections describing the implementation for each stage of the process, including data assessment, model generation and model assessment steps.

4.1.1 Exploratory Analysis of Dataset

Following initial data adjustments the correlation matrix was generated to assess relationships in the data. This is shown in Figure 3. Based on this we see that a number of variables display reasonable levels of correlation. However, given the small number of independent variables all variables were retained at this stage.

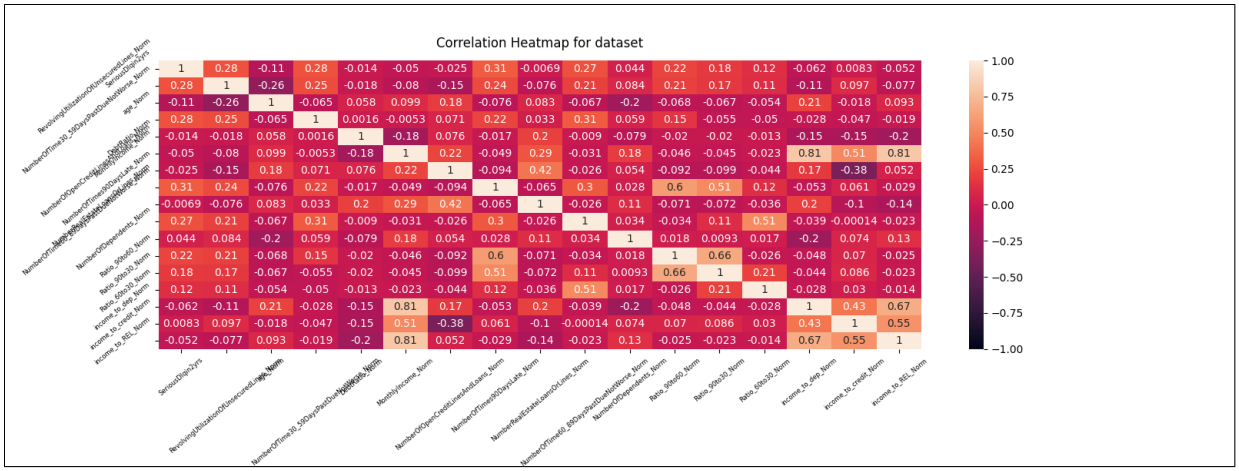

Figure 3: Correlation Matrix for Dataset

The predictive ability of each factor in the model was considered. Given that factors are all continuous this was visualised by applying a decile binning approach and calculating the event rate for each decile. if a factor is predictive this should be evidenced by differing event rates across the decile. Box plots of the distribution of the attribute for each level of the target variable were also generated,several variables appear predictive.

The data were standardised around a mean of zero and a sigma of one. A split into Train test and Validation datasets in a 60:20:20 ratio was completed.

4.2 Class Imbalance and Feature Engineering

Initial analysis of the data indicated that class imbalance was a concern as the default target represented only 6.7% of cases in the sample. This is an issue frequently seen in credit default. As Luque et al. (2019) has noted, class imbalance can impact the accuracy of many performance metrics, typically leading to low predictive ability for the infrequent class Ling and Sheng (2010). To address this imbalance in the development dataset SMOTE oversampling was applied as detailed in Chawla et al. (2002). This oversampling approach was applied only to the development dataset.

The Risk Weighted Assets (RWA) were calculated using simulated EAD. LGD and the predicted probability generated by the developed models, in accordance with the regulatory approach [12]. Additional ratio variables were also created as per Section 3.

4.3 Model Development

A range of Python libraries were used to develop the credit scoring models considered in the research. For the Logistic Regression, Random Forest, AdaBoost and Naive Bayes approaches SciKit learn was used. XGBoost was fitted with the xgbost library, Logit Boosting was fitted with the logitboost library, and Neural Net was fitted with TensorFlow. Hyperparameter tuning for the ensemble approaches was completed using RandomizedSearchCv and GridsearchCV. This addresses a regulatory expectation that hyperparameter evaluation be considered.

4.4 Performance Evaluation Approach

To evaluate the models the AUC metric is used, as well as considering the overall confusion matrix and precision-recall curves to understand the capabilities of each model tested.

Model Explainabiluty is considered through the SHAP values obtained. This will be applied to both the local explanation of selected observations and the Global feature importance for the model.

Model complexity is assessed using Prediction Latency as applied to the test dataset. As noted in the documentation for sci-kit learn (`https://scikit-learn.org/0.15/modules/computational_performance.html`), the main features influencing prediction latency are Number of features, Input data representation and sparsity, Model complexity and Feature extraction

The final step involves presenting a view of the model that considers performance, explainability, capital and complexity. Based on this a model selection can be proposed.

5 Model Evaluation and Assessment

5.1 Model Development and Performance Evaluation

5.1.1 Logistic Regression

The logistic regression model displays an AUC of 86.2% as as shown in Figure 4. The confusion matrix illustrates how the model makes a trade off between a strong ability to detect true defaults versus a tendency to incorrectly classify good customers as possible defaults. Considering additional statistics from the confusion matrix we see that the models precision, at 23.56% is relatively low. Precision represents the number of true positives as a ratio of total predicted positives, and a low precision indicates the models tendency to mis-classify negatives as positives. Conversely the high recall indicates that the model is strong at not mis-classifying defaults as goods. This illustrates a sterength of the Logistic regression, given the loss associated with having a default is much higher than the profit foregone by rejecting a good obligor.

[12]Regulation (EU) 575 2013, section 2

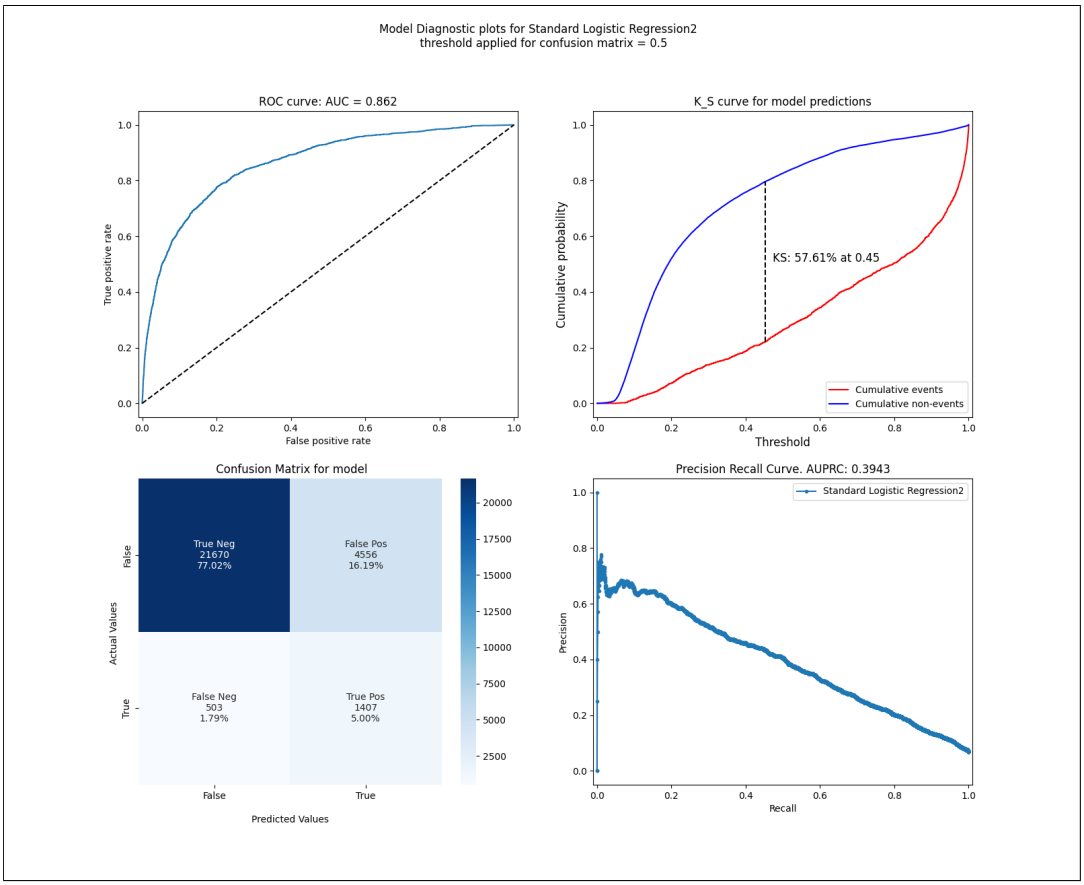

Figure 4: Diagnostic Plots for Logistic Regression Model

5.1.2 Binned Scorecard Logistic Regression Model

The Binned Scorecard logistic regression model displays a similar performance to the Logistic regression model with an AUC of 83% A small reduction in AUC is not unexpected due to the combining of continuous variables into discrete bins. The confusion matrix also displays similar behaviour in terms of strong ability to detect true defaults but a tendency to incorrectly classify good customers. This model has somewhat poorer recall at 56.84%.

5.1.3 Random Forest Model

Performance using Random Forest was disappointing given what has been seen in other research, with an AUC, at 85.6%, lower than the Logistic Regression model. However accuracy is quite high at 93.5%. When we consider Precision (54.9%) and recall (26.6%) we see that the model has reversed the trend in previous models, displaying a lower rate of both false and true positives - e.g. the model tends to more aggressively classify as non-default. The confusion matrices also show this, with a much lower tendency to declare a default but a better chance of correctly classifying non-defaulters.

5.1.4 XGBoost Model

For XGBoost we see performance with an AUC of 83%, that is again lower than the Logistic Regression model. Similarly to the Random Forest the model displays high

accuracy at 93.36%. Considering Precision (52%) and recall (28.7%) the model again tends to have higher true and false negatives.

5.1.5 LogitBoost Model

For LogitBoost, model performance is similar to other Boosted ensemble models, with an AUC of 85.3%, and lower than the Logistic Regression model. Accuracy is again high at 93.57%, with precision (57.7%) and recall (20.1%). The model has a much lower tendency to declare a default than the logistic regression. This is based on a threshold of 0.5 being applied for the outcome.

5.1.6 AdaBoost Model

In Figure 5 we display the diagnostic plots for the AdaBoost model. The model displays an AUC of 86.2%, equivalent to the Logistic Regression model. Accuracy is also quite high at 93.6%. When we consider Precision (57.3%) and recall (24.6%) we see behaviour similar to the other ensemble models. The KS Plot indicates a compression of the distribution for the two classes which we will see later.

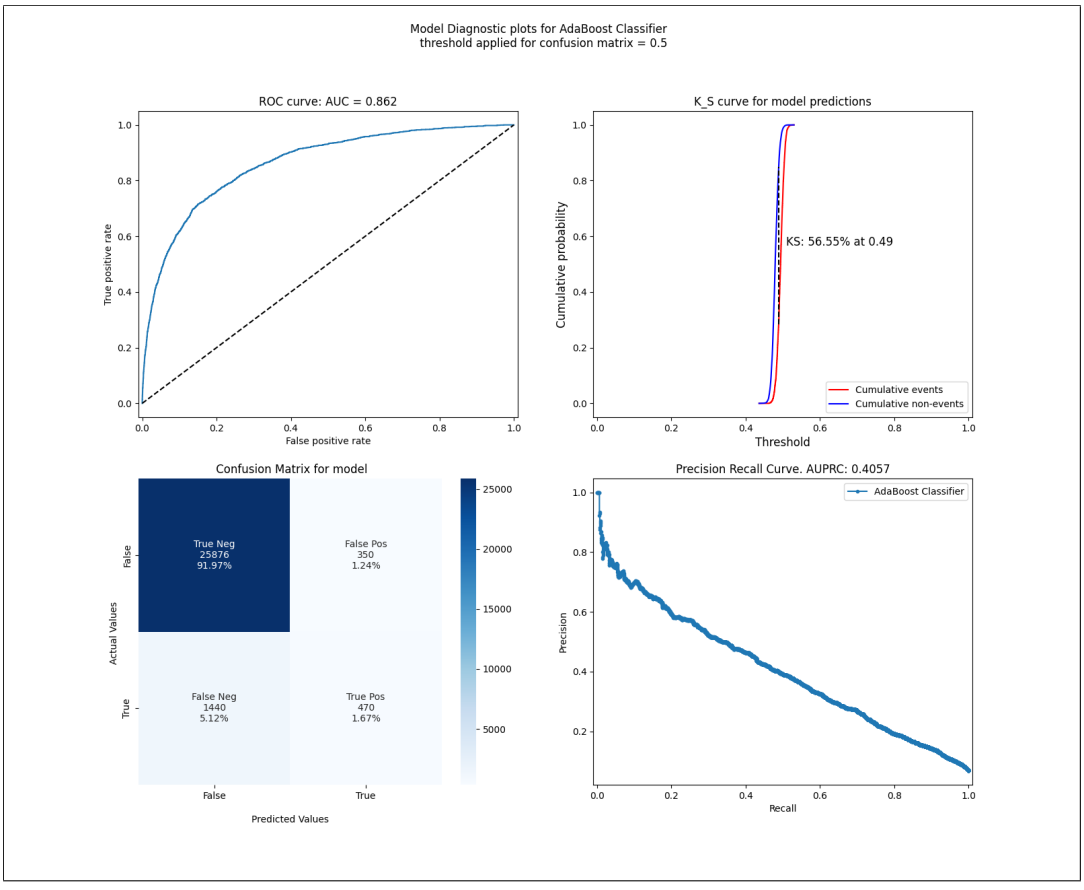

Figure 5: Diagnostic Plots for AdaBoost Model

5.1.7 Neural Network Model

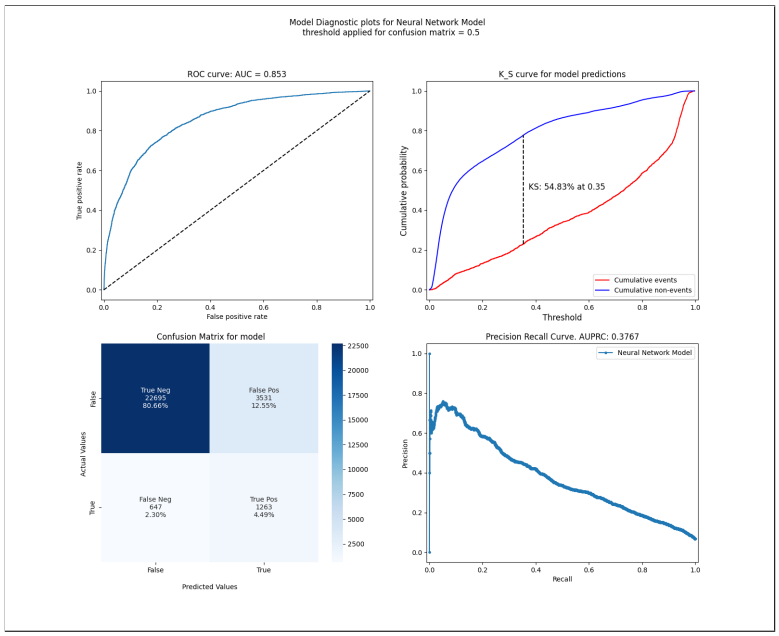

Figure 6: Diagnostic Plots for Neural Network Model

In Figure 6 we display the diagnostic plots for the Neural Network model fitted to the data. Performance here, with an AUC of 85.3%, is between the two Logistic models. Accuracy, at 85.1%, is also inferior to other candidate models.

5.1.8 NaiveBayes Model

Naive Bayes is the simplest model we have fitted, and as might be expected, performance is lower than other candidates, An AUC of 82.8% is obtained. Accuracy of 90.1%, precision (35%) and recall (53.6%). Overall the model behaves similarly to the logistic regression, with a greater tendency to assign a case as a default than other classifiers being evidenced in the precision and recall.

5.2 Model Explainability

5.2.1 Logistic Regression

As a standard Linear model Logistic Regression can be explained in terms of the parameter weights associated with each factor. However this approach is model specific and cannot be used to compare different types of models, for example to enable management understanding. Therefore we consider feature importance calculated using SHAP, using a Linear Explainer. This approach instead considers the marginal contribution that each feature makes to the model. When we apply a SHAP approach to the Logistic model we see the feature importance shown in Figure 7. There are some difference between the two methods Saarela and Jauhiainen (2021), however the most prominent drivers are consistent across the two approaches, with Utilization of Unsecured Lines still the most important factor .

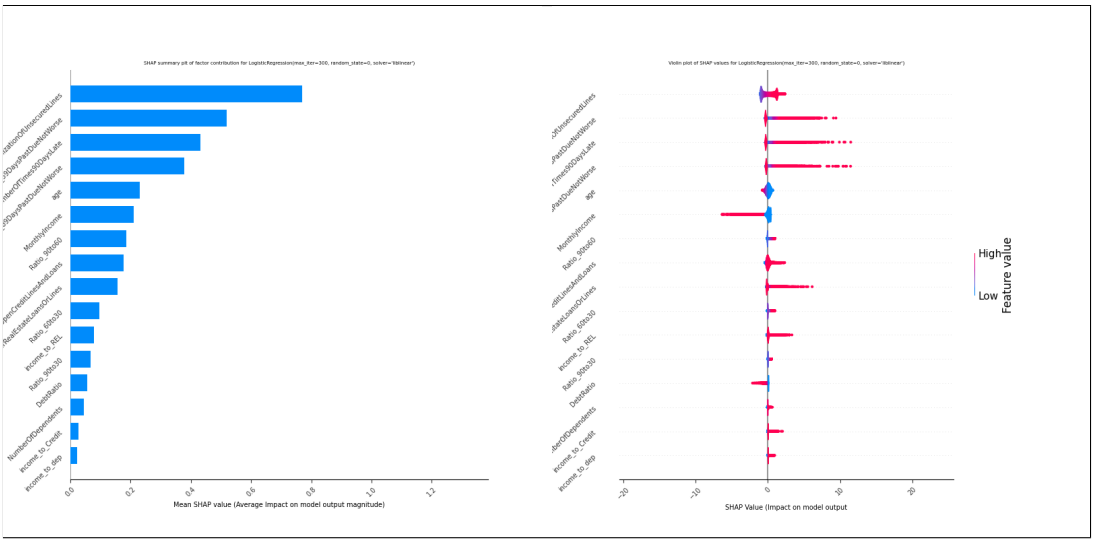

Figure 7: Feature Importance and Violin Plot for Logistic Regression Model

We can also see how individual observations combine to contribute to overall importance, with the effect being illustrated in the violin plot. This illustrates both the impact of the driver but also directionally how the variable influences the outcome.

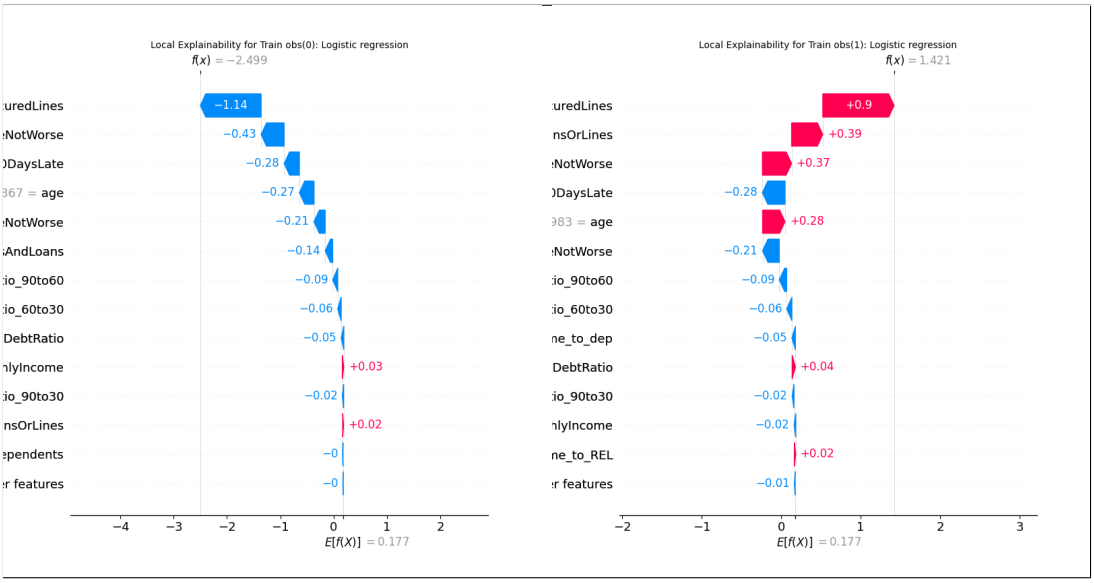

Figure 8: Local SHAP explanation for first two cases in Test dataset

Given we have this information it is also possible to explain the outcomes for individual observations. In Figure 8 we can see the waterfall plots for the first two datapoints of the Test dataset. From these we can see the contribution of drivers to the individual outcomes, providing local explainability. This is due to SHAP's unified approach to explainability, utilising local interpretability equivalent to LIME.

5.2.2 XGBoost

A similar approach is applied to the XGBoost model using SHAPs TreeExplainer algorithm. While different explainers reduce generality slightly they will improve the speed and accuracy of the overall results. Based on this analysis, illustrated in Figure 9 we can see that XGBoost is selecting a different range of key features, with the factors NumberofRealEstateLoansandLines and NumberofDependants being the top two model driver for this model. This flexibility to compare explanations enable business logic to be brought to bear in model selection as well as supporting algorithmic fairness by enabling models that over weight more controversial factors to potentially be identified and rejected. Once again this can also be extended to local explainability, with Figure 10 illustrating the explanations for the first two cases of the test dataset. Unsurprisingly these are different to what we see for the Logistic regression model

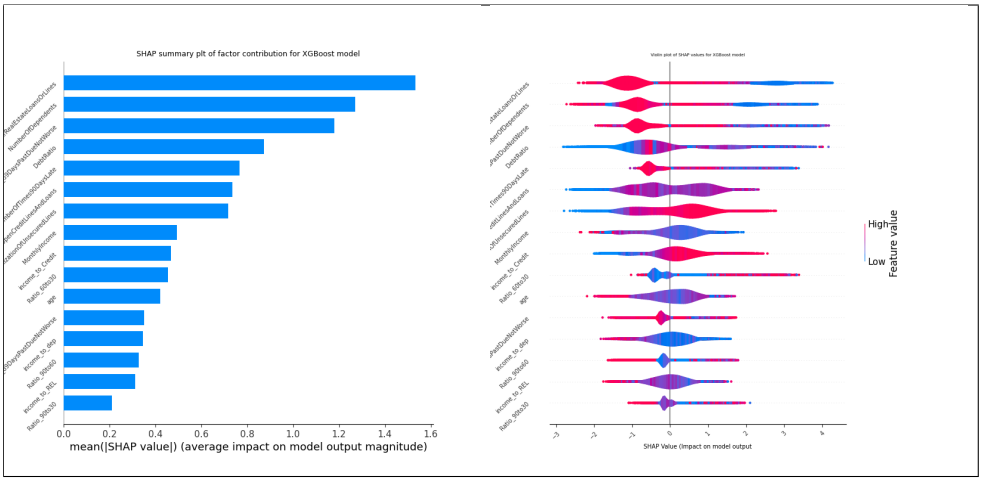

Figure 9: SHAP Model explanation for XGBoost model

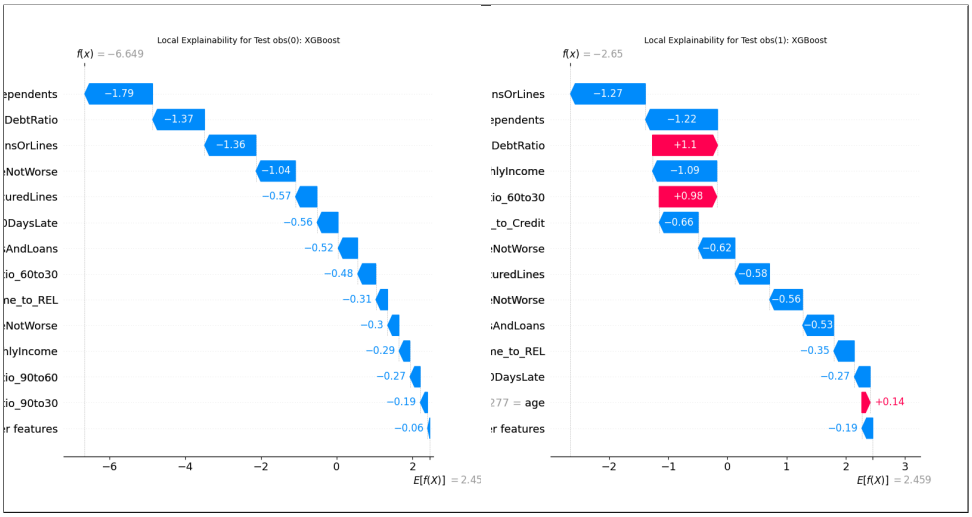

Figure 10: Local SHAP explanation for first two cases in Test dataset using XGBoost

5.2.3 Random Forest

The same tree explainer within SHAP can be applied within Random Forest. Note that due to the extremely long computation time encountered we restricted the fit to the first 1,000 values of the dataset. However this is sufficient to demonstrate the viability of the technique. Local explainability is also again possible, utilising a similar approach to that taken for XGBoost.

5.2.4 AdaBoost

Adaboost requires the use of SHAPs kernel explainer to produce explainability estimates. Given the slower speed of this approach, which fits a regression estimator locally to the data to produce its estimates, only a sample of the dataset was used.

Based on this analysis, illustrated in Figure 11 we can see that AdaBoost again selects a different range of key features, with the factors DebtRatio and NumberofUnsecuredLines being the top two model driver for this model. Based on this insight a business may wish to preferentially consider models that for example rely on long term stable macroeconomic factors.

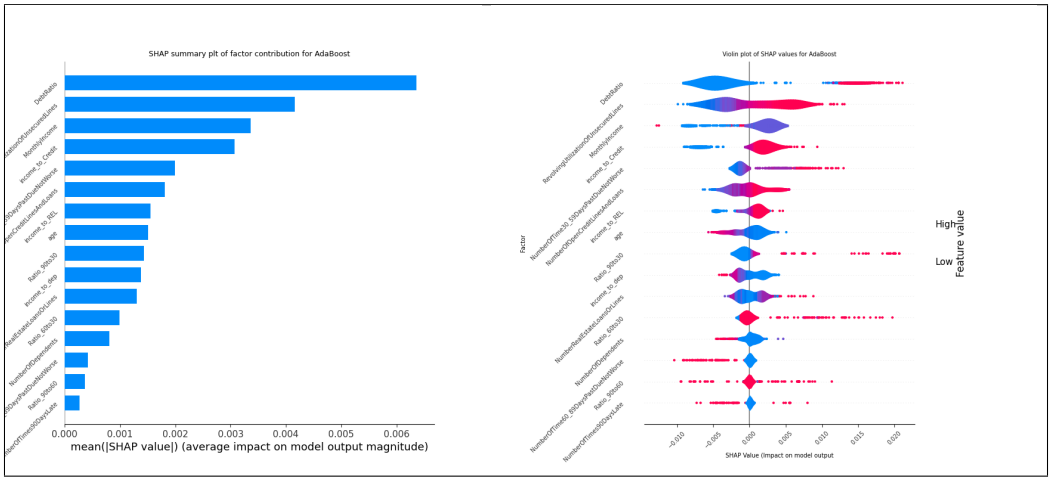

Figure 11: SHAP Model explanation for AdaBoost model

The same approach was extended to all models. Naive Bayes and Neural Net were fitted similarly, uing the Linear and DeepExplainers respectively. We encountered difficulties with producing explainability for the Scorecard model. This is related to the way the model is created by the OptBinning library. and While it may be surmountable we were unable to resolve within this research. However we note that by definition the scorecard is trivially explainable on a standalone basis.

5.3 Model Complexity

The complexity of the model is considered through the prediction latency - the time taken to fit the test dataset. For logistic regression we obtain a Bulk prediction latency of $4.3x10^{-}8$ seconds based on fitting 30 bulk repeats. An average atomic prediction latency of $6.1x10^{-}4$ seconds is obtained. The prediction latency of the Binned scorecard model is considerably higher than that of the Logistic regression model. Atomic latency

is $2.9x10^{-3}$ and Batch predictive latency is $6.6x10^{-7}$. This is potentially due to the extra processing required to assign the appropriate binning to the data. In Figure 12 the Atomic Latency is illustrated. It is clear that the Random Forest and logitBoost models have both the greatest Predictive Latency and the highest volatility for the Predictive Latency. When we remove the most extreme performers we can see that the XGBoost model performs relatively well in terms of Predictive Latency.

The Batch Predictive Latency is also assessed. This consists of fitting the full dataset and assessing the average fit time per observation. We expect some efficiencies due to fitting the data in bulk. The process is repeated 30 times to allow for variance in the process. The outcomes are illustrated in Figure 13. The results are somewhat similar. Once the most extreme performers are removed we see that AdaBoost again displays surprisingly good results.

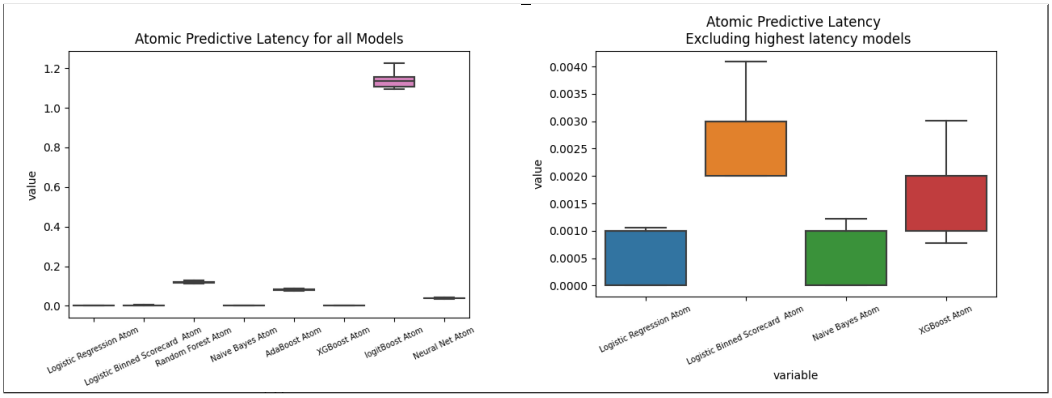

Figure 12: Atomic Latency for all models

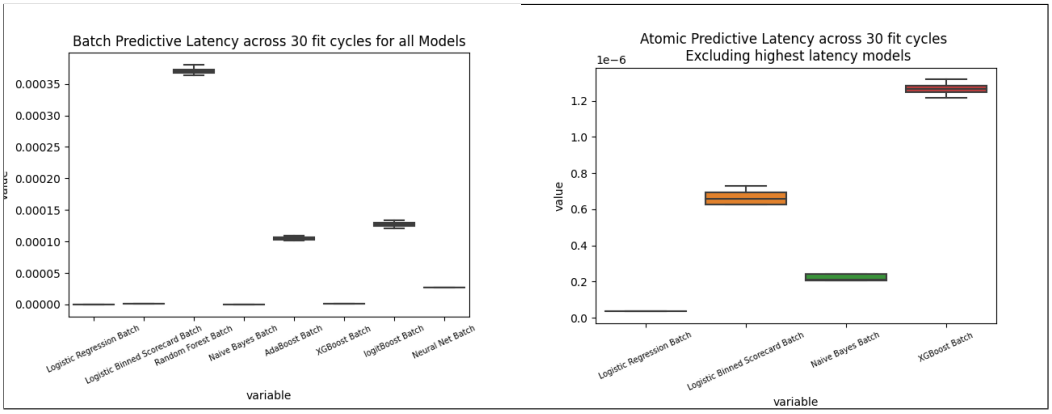

Figure 13: Batch Latency for models based on 30 fits

Predictive Latency appears to produce intuitive results in terms of the rank order of the models, with the more 'complex' machine learning approaches having a higher latency that would reflect both the larger model structure and the complexity to arrive at a decision. However further work is required to quantify the relationship between Predictive Latency for different models..

5.4 Model Risk Weighted Assets

Figure 14 displays the distribution of the predicted PD for the different models tested. A wide range of different distributions are generated by the models. While Logistic regression and Random Forest produce long tailed distributions the AdaBoost classifier produces a more symmetrical form. Heavily bimodal models also feature here. The shape of the distributions heavily influences the effect of any adjustment to the thresholds as a model adjustment strategy. Adaboost would display the largest initial movement under such a strategy, however the relationship between PD and RWA would imply limited changes in capital. On the other hand Naive Bayes would see little benefit. The shape here also reflects the confusion matrix, with some models being much more conservative in terms of assigning defaults

Figure 14 also illustrates the impact of this in terms of the RWA percentage associated with each model. Of particular interest is that the most predictive model doesn't inherently lead to the lowest average RWA percentage. In particular models with higher false negative rates will often lead to lower capital as they misclassify cases to a lower risk. However given RWA is a non-linear function this is not guaranteed, and we see AdaBoost actually leads to the highest capital requirements. Models with extreme bimodal classification should lead to the lowest capital requirements, and this is reflected in the models here. However this feature would likely be difficult justify to regulators.

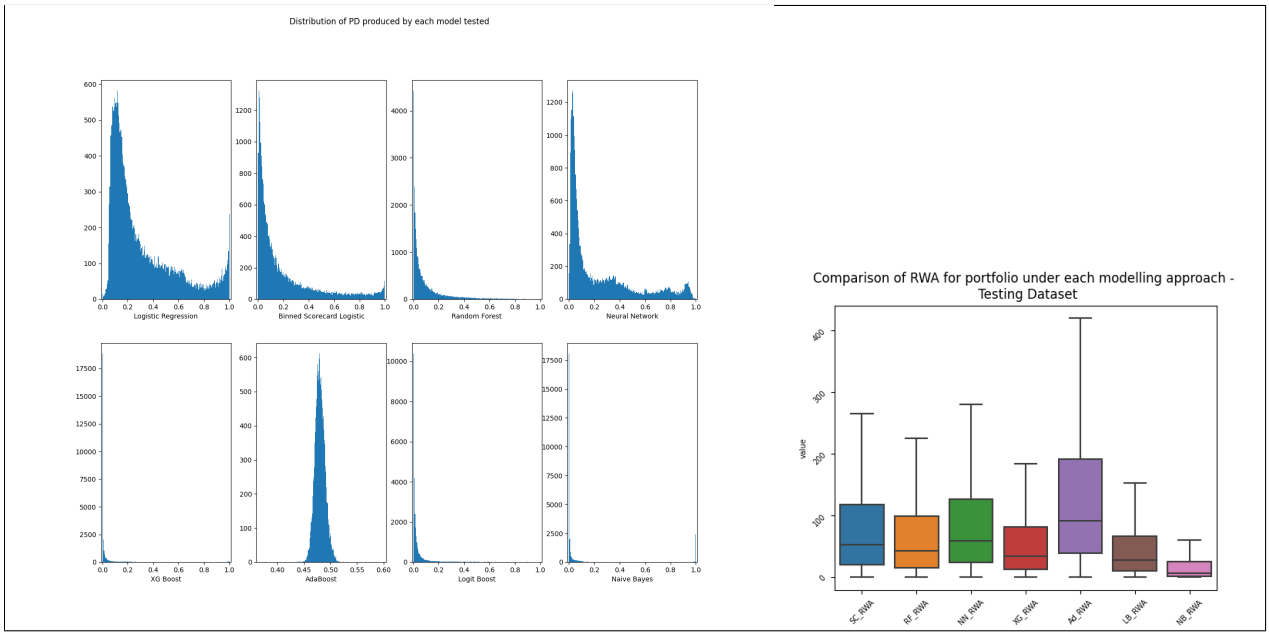

Figure 14: Distribution of PD values and RWA percentages for each model Model

5.5 Overall Model Assessment

Statistic	Logistic	Scorecard	AdaBoost	Random Forest	XGBoost	LogitBoost	Naive Bayes	Neural Net
AUC	0.862	0.83	0.862	0.856	0.83	0.853	0.828	0.853
Accuracy	0.82	0.87	0.936	0.935	0.934	0.936	0.901	0.852
Precision	0.236	0.277	0..573	0.549	0..520	0.577	0.35	0.263
Recall	0.763	0.568	0.246	0.266	0.287	0.201	0.536	0.661
F1	0.357	0.372	0.344	0.358	0.370	0.298	0.424	0.377
RWA [13]	0.272	0.230	0.361	0.170	0.161	0.135	0.0611	0.245
Atomic predictive Latency								
mean	$6.1x10^{-4}$	$2.9x10^{-3}$	$8.1x10^{-2}$	$1.3x10^{-1}$	$1.6x10^{-3}$	$1.1x10^{0}$	$6.0x10^{-4}$	$3.8x10^{-2}$
stedv	$4.9x10^{-4}$	$4.1x10^{-4}$	$1.8x10^{-3}$	$1.9x10^{-2}$	$5.0x10^{-4}$	$2.2x10^{-2}$	$4.9x10^{-4}$	$6.1x10^{-3}$
Batch predictive Latency								
mean	$4.3x10^{-8}$	$6.6x10^{-7}$	$1.1x10^{-4}$	$3.7x10^{-4}$	$1.3x10^{-6}$	$1.3x10^{-4}$	$2.3x10^{-7}$	$2.7x10^{-5}$
stedv	$1.5x10^{-8}$	$3.5x10^{-8}$	$3.1x10^{-6}$	$5.2x10^{-6}$	$1.0x10^{-7}$	$3.0x10^{-6}$	$1.8x10^{-8}$	$4.9x10^{-7}$

Table 5: Compiled Model Assessment Statistics

Based on the details in the sections above, the choice of model would depend on which of the aspects an bank would wish to prioritise and the weighting they would apply for each aspect. The EBA guidelines, while calling out the need to consider these aspects, do not explicitly state a threshold or priority to be applied. In Table 5 we provide a summary of the various elements to be considered in assessing a preferred model choice.

Table 6 ranks the models under a number of criteria. It is clear than no model is best under every criteria.

Model	AUC	Accuracy	Explainability	RWA	Atomic Prediction Latency	Batch Prediction Latency
Logistic	1	8	y	7	1	1
Scorecard	6	6	y*	5	4	3
AdaBoost	1	1	y	8	6	6
Random Forest	3	3	y	4	7	8
XGBoost	6	4	y	3	3	4
Logit Boost	4	2	y	2	8	7
Naive Bayes	8	5	y	1	2	2
Neural Net	4	7	y	6	5	5

Table 6: Model ranking against Criteria

A simple average of the ranks would suggest Logistic regression is the preferred model in this case, however from a banks perspective this would come at the cost of very high RWA cost. By adjusting the weighting it is possible to select a model based on for example, additional weight being given to capital intensive or more complex models. In practice, in the absence of a clear statement from regulators it would be up to an individual institution to justify the weighting to be applied to each criteria, while being mindful of a likely regulatory push back against models with either excessively low capital

requirements or excessively high complexity. For example AdaBoost demonstrates good predictive performance however it is unlikely to be a preferred model due to the high RWA driven by the clustering of predictions around a centre point. This is also likely to require significant justification to regulators. In addition its predictive performance comes at the cost of high latency. As such a bank may down-weight this model.

5.6 Enabling Reporting and Monitoring

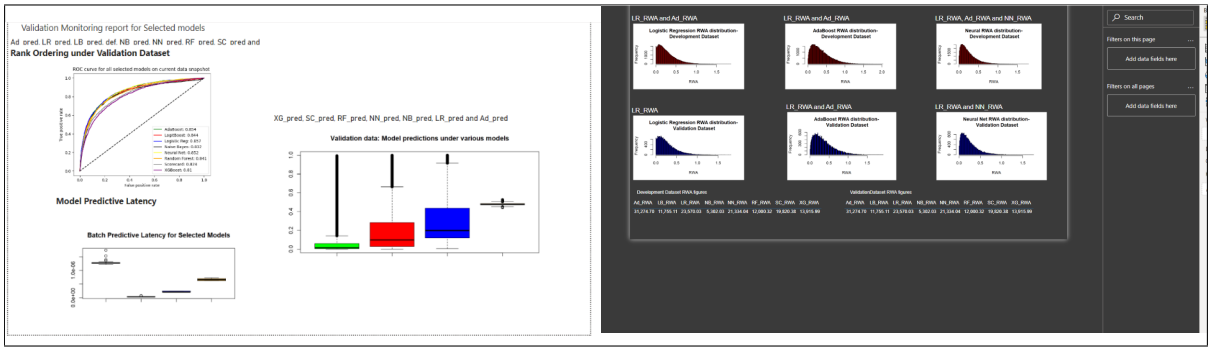

Figure 15: Example of Power BI visualizations

The results of the analysis are transferred into SQL server database tables. Distinct Tables are generated for each aspect of the research. From here a simple Power BI dashboard can be generated using Python and R scripts to illustrate how this can be monitored and reported to management in line with regulations. Examples of this are shown in Figure 15

5.7 Comparison to Other Research Findings

The research indicates that Logistic Regression is the preferred model in this case. This is consistent with the findings of Juneja et al. (2020), with AUC performance for this model 85.8% lying within the range of 82%1 - 89% found in that research. It is perhaps surprising to bee other methods performing less weel, however the Random Forest, when considered in terms of Accuracy of 93% compares well to the findings of Yu (2020), which achieved 95%. The Neural Netowrk performance falls short of that achieved by Sun and Vasarhalyi (2021), hoever this may be reflective of the greater depth applied in the latter model.

In addition the research aligns with the work of Kurshan et al. (2020) in that it extends the evaluation to a wider rangwe of risk dimensions, addressing many of the challenges raised in that paper. In line with the work ofAlonso and Carbo (2021) we also consider capital implications as a key aspect of regulatory review.

6 Conclusion and Future Work

This research project aimed to assess the following research question:

"How can complex Machine Learning Models be assessed in a way that allows a clear demonstration of the preferred model in terms of both understandability of the model, pure

model performance, capital impacts and relative model complexity?"

This has been addressed by reviewing and building on the existing research in the field [R.O 1] to explore how cutting edge machine learning models can be assessed in a model agnostic fashion that considers explainbility, complexity and wider regulatory concerns. The research has prepared a suitable dataset [R.O 2] and developed a range of models to perform credit default modelling, reflective of high performers in existing research [R.O 3-3.7]. The research has shown that In addition to the range of model performance metrics available [R.O 4] it is possible to compare the models in terms of their contributing features using a model agnostic approach [R.O 5]. Individual cases can also be explained using a local explainability approach [R.O 6]. In additrion we have incorporated a measure of the model complexity into the assessment [R.O 7], as well as considering the implications of model choice on the underlying RWA position of the bank [R.O 8]. A database of the generated results from the research has also been created [R.O 9], both the satisfy regulatory expectations around documentation and to enable further visualisation and assessment and monotoring by both developers and management [R.O 10].

The research provides for an assessment of credit models that considers a full range of regulatory aspects. Such an evaluation is required to enable these models to be widely accepted in industry. However limitations include (i) that regulatory requirements continue to evolve and may place future additional burdens and (ii) the approach does not consider the calibration of the model against the long-run PD.

Future work could include the assessment of other complexity measures, the incorporation of the model calibration into the assessment, and the development of an overall risk metric to unify the different dimensions considered here.

Acknowledgement

I'd like to thank my wife Emer and children Ava, Ethan and Adam for their patience and support throughout my studies. I would alsolike to extend thanks to my research superviser, Dr Catherine Mulwa and all the staff of NCI for their guidance throughout this research and my time at the college.

References

Addo, P. M., Guegan, D. and Hassani, B. (2018). Credit risk analysis using machine and deep learning models, *Risks* **6**(2): 38.

Aleksandrova, C. A. P. D. Y., Parusheva, S. et al. (2021). Performance evaluation of machine learning models for credit risk prediction, *Izvestia Journal of the Union of Scientists-Varna. Economic Sciences Series* **10**(2): 89–98.

Alonso, A. and Carbó, J. M. (2020). Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost.

Alonso, A. M. and Carbo, J. (2021). Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation, *Banco de Espana: Working Papers (Topic)* .

Bracke, P., Datta, A., Jung, C. and Sen, S. (2019). Machine learning explainability in finance: an application to default risk analysis.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W. and Siddique, A. (2016). Risk and risk management in the credit card industry, *Journal of Banking & Finance* **72**: 218–239.

Carrington, A., Fieguth, P. and Chen, H. (2018). Measures of model interpretability for model selection, *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, pp. 329–349.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S. and Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk, *arXiv preprint arXiv:1811.12615* .

Chen, S., Guo, Z. and Zhao, X. (2021). Predicting mortgage early delinquency with machine learning methods, *European Journal of Operational Research* **290**(1): 358–372.

Chishti, W. A. and Awan, S. M. (2019). Deep neural network a step by step approach to classify credit card default customer, *2019 International Conference on Innovative Computing (ICIC)*, IEEE, pp. 1–8.

Cornacchia, G., Narducci, F. and Ragone, A. (2021). A general model for fair and explainable recommendation in the loan domain.

Das, N. M. and Deb, J. (2017). Regulatory capital and its impact on credit risk: The case of indian commercial banks., *IUP Journal of Bank Management* **16**(4).

Dastile, X., Çelik, T. and Potsane, M. M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey, *Appl. Soft Comput.* **91**: 106263.

De Servigny, A., Renault, O. and de Servigny, A. (2004). Measuring and managing credit risk.

Demajo, L. M., Vella, V. and Dingli, A. (2020). Explainable ai for interpretable credit scoring, *arXiv preprint arXiv:2012.03749* .

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–22.

Dieber, J. and Kirrane, S. (2022). A novel model usability evaluation framework (muse) for explainable artificial intelligence, *Information Fusion* **81**: 143–153.

Engelmann, B. (2006). Measures of a rating's discriminative power—applications and limitations, *The Basel II Risk Parameters*, Springer, pp. 263–287.

Friedman, J., Hastie, T. and Tibshirani, R. (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.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018). A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* **51**(5): 1–42.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Hurlin, C., Leymarie, J. and Patin, A. (2018). Loss functions for loss given default model comparison, *European Journal of Operational Research* **268**(1): 348–360.

Islam, S. R., Eberle, W. and Ghafoor, S. K. (2020). Towards quantification of explainability in explainable artificial intelligence methods, *The thirty-third international flairs conference*.

Izzi, L., Oricchio, G. and Vitale, L. (2011). *Basel III credit rating systems: An applied guide to quantitative and qualitative models*, Springer.

Jiménez, G. and Mencía, J. (2009). Modelling the distribution of credit losses with observable and latent factors, *Journal of Empirical Finance* **16**(2): 235–253.

Juneja, S. et al. (2020). Defaulter prediction for assessment of credit risks using machine learning algorithms, *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, pp. 1139–1144.

Kokaly, S., Salay, R., Sabetzadeh, M., Chechik, M. and Maibaum, T. (2016). Model management for regulatory compliance: a position paper, *2016 IEEE/ACM 8th International Workshop on Modeling in Software Engineering (MiSE)*, IEEE, pp. 74–80.

Kumar, M. R. and Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis, *Ingeniería Solidaria* **16**(1).

Kurshan, E., Shen, H. and Chen, J. (2020). Towards self-regulating ai: Challenges and opportunities of ai model governance in financial services, *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8.

Li, G., Weng, J., Wu, B. and Hou, Z. (2022). Incorporating multi-scenario underreporting rates into mice for underreported maritime accident record analysis, *Ocean Engineering* **246**: 110620.

Ling, C. X. and Sheng, V. S. (2010). *Class Imbalance Problem*, Springer US, Boston, MA, pp. 171–171.
URL: *https://doi.org/10.1007/978-0-387-30164-8*$_1$10

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., *Queue* **16**(3): 31–57.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions, *Advances in neural information processing systems* **30**.

Luque, A., Carrasco, A., Martín, A. and de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix, *Pattern Recognition* **91**: 216–231.

Magesh, P. R., Myloth, R. D. and Tom, R. J. (2020). An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery, *Computers in Biology and Medicine* **126**: 104041.

Onay, C. and Öztürk, E. (2018). A review of credit scoring research in the age of big data, *Journal of Financial Regulation and Compliance* .

Ouedraogo, D. N. (2021). Interpretable machine learning model selection for breast cancer diagnosis based on k-means clustering, *Applied Medical Informatics.* **43**(3): 91–102.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A. and Przybocki, M. A. (2020). Four principles of explainable artificial intelligence, *Gaithersburg, Maryland* .

Plotnikova, V., Dumas, M. and Milani, F. (2020). Adaptations of data mining methodologies: a systematic literature review, *PeerJ Computer Science* **6**: e267.

Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models, *SN Applied Sciences* **3**(2): 1–12.

Saygili, E., Saygili, A. T. and Gokhan, I. (2019). An analysis of factors affecting credit scoring performance in smes, *Ege Academic Review* **19**(2): 159–171.

Shafique, U. and Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma), *International Journal of Innovation and Scientific Research* **12**(1): 217–222.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*, Vol. 3, John Wiley & Sons.

Smuha, N. (2019). Ethics guidelines for trustworthy ai, *AI & Ethics, Date: 2019/05/28-2019/05/28, Location: Brussels (Digityser), Belgium*.

Sun, T. and Vasarhalyi, M. A. (2021). Predicting credit card delinquencies: An application of deep neural networks, *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning*, World Scientific, pp. 4349–4381.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* **76**: 89–106.

Wang, Y., Zhang, Y., Lu, Y. and Yu, X. (2020). A comparative assessment of credit risk model based on machine learning——a case study of bank loan data, *Procedia Computer Science* **174**: 141–149.

Yu, Y. (2020). The application of machine learning algorithms in credit card default prediction, *2020 International Conference on Computing and Data Science (CDS)*, IEEE, pp. 212–218.