risks

MDPI

Article

Approaching European Supervisory Risk Assessment with SupTech: A Proposal of an Early Warning System

Pedro Guerra *, Mauro Castelli and Nadine Côrte-Real

Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal; mcastelli@novaims.unl.pt (M.C.); nreal@novaims.unl.pt (N.C.-R.)
* Correspondence: paguerra@bportugal.pt

Abstract: Risk analysis and scenario testing are two of the core activities carried out by economists at central banks. With the increasing adoption of machine learning to enhance decision-support systems, and the amount of collected data spiking, institutions provide countless use-cases for the application of these innovative technologies. Consequently, in recent years, the term sup-tech has entered the financial jargon and is here to stay. In this paper, we address risk assessment from a central bank's perspective. The uptrending number of involved banks and institutions raises the necessity of a standardised risk methodology. For that reason, we adopted the Risk Assessment Methodology (RAS), the quantitative pillar from the Supervisory Review and Evaluation Process (SREP). Based on real-world supervisory data from the Portuguese banking sector, from March 2014 until August 2021, we successfully model the supervisory risk assessment process, in its quantitative approach by the RAS. Our findings and the resulting model are proposed as an Early Warning System that can support supervisors in their day-to-day tasks, as well as within the SREP process.

Keywords: banking supervision; risk assessment; machine learning; sup-tech; EWS; scenario analysis; ECB risk assessment system

JEL Classification: C61; 681; O33; E58

Citation: Guerra, Pedro, Mauro Castelli, and Nadine Côrte-Real. 2022. Approaching European Supervisory Risk Assessment with SupTech: A Proposal of an Early Warning System. *Risks* 10: 71. https://doi.org/10.3390/risks10040071

Academic Editors: Mogens Steffensen and Anatoliy Swishchuk

Received: 9 January 2022
Accepted: 21 March 2022
Published: 24 March 2022

1. Introduction

In recent years, the use of decision-support systems has skyrocketed, with machine learning (ML) spearheading the change. The financial industry has always been one of the main drivers for that development (Zopounidis et al. 1997). As the amount of data collected soars and computing power rises to meet the challenge, the use of classical statistics such as linear and logistic regressions is gradually declining. Although they were once the mainstay of decision-support systems, nowadays they tend to be recalled sporadically, and mainly for their better comprehensibility in comparison to most ML models (Yang and Wu 2021). The current research problem is how to leverage ML to support risk assessment processes at central banks, using quantitative supervisory data.

Recent uses of machine learning have unveiled data patterns that were as of yet undiscovered (Huang et al. 2021). These applications have also expanded to the fields of regulation and supervision, as described by Hertig (2021). For supervisory purposes, there has been a huge increase of interest in developing sup-tech tools. As Beerman et al. (2021) reported, the number of ongoing ML projects in this field skyrocketed from 12 in 2019 to 71 as of December 2021. The pandemic forced an off-site approach to what was previously required to be done in person. In the past 2 years we have seen an increasingly higher number of production-ready systems applying ML to support central banks' tasks (Massaro et al. 2020). From the specific standpoint of supervision, the work from Filippopoulou et al. (2020) is a watershed in EWS development at central banks, using EBC Macro-prudential Database to address credit risk. This work, along with the EWS

proposed by Pompella and Dicanio (2017), supports the importance of these systems to support rating assignments and alert for distress signals.

The amount of data retrieved in the supervisory framework is overwhelmingly high (European Banking Authority 2013). Additionally, supervisors often ask for complementary information, either quantitative or qualitative. Even thought National Central Banks (NCBs) are equipped with business intelligence systems that allow them to organise most quantitative information, data analysis is mostly done in an ad hoc manner that is impractical for a prompt spotting of risky events (Broeders and Prenio 2018). Besides, this method only looks at past events, making it impossible to systematically test alternative economic scenarios. Furthermore, we must mention that risk methodologies might vary, making it difficult to compare not only the assessments, but also the evolution of the classifications.

Traditional approaches already set out an organised perspective of the reported data, through dashboards and reports that provide aggregated and specific views of key indicators (di Castri et al. 2019). However, these approaches only consider past events and they are constrained by the regulatory framework (not to mention, they lack what-if analysis and decision processes built on that data). The use of innovative technologies to support supervisory processes is defined by Broeders and Prenio (2018); Doerr et al. (2021) as sup-tech, and these authors summarise the barriers of adoption in three main items:

1. Frequent regulatory updates;
2. Conservative industry;
3. Lack of qualified human resources.

From a data perspective, Early Warning Systems for predicting banking crisis have also been in the spotlight. Casabianca et al. (2019); Consoli et al. (2021) are some of the many examples of landmark findings in that area, along with the previously mentioned Filippopoulou et al. (2020). However, none of these authors explore the information available in the European supervisory framework.

In a previous work, we have addressed the issues of using a single risk methodology, selecting literature-supported ML models to evaluate the risk level of banks, and using up-to-date real-world supervisory data from the Portuguese banking sector (Guerra et al. 2022). The previous work addressed the concept of liquidity risk since it is crucial for a bank's ability to operate (Vento and Ganga 2009) and it can render a bank nonoperational in a matter of days (Shah et al. 2018). In our paper we expand previous findings to the other risk perspectives comprised in the Supervisory Review and Evaluation Process (SREP).

In our current study, we have extended the sample from March 2014 until August 2021. This data is extensively validated by *Banco de Portugal* and European Central Bank (ECB) quality assurance processes. The quality of gathered information allows for accurate assessment, thus ensuring a positive correlation between risk prediction and the observed phenomena (Ng 2011).

Another key component of our approach is the way we set up the classification problem. Contrary to what is commonly found in the literature, we reiterate the importance of considering a multitier classification approach to this problem. Our data being provided by real world context, we feel highly confident in expanding from the *fail/no-fail* classes and adopting the four classes comprised in the RAS methodology, a European-wide risk assessment methodology:

1. Low risk;
2. Medium–low risk;
3. Medium risk;
4. High risk.

Our work also showcases literature-backed ML models for structured financial data that support the efficiency of supervisory processes.

Based on the findings of our study, we provide a comprehensive guidance for the development of valuable supervisory use-cases enhanced by innovative techniques.

The purpose of this work is to leverage the above-mentioned aspects, and expand the academic body of knowledge of quantitative risk assessment for prudential supervision. From a supervisor's standpoint, we aim to bring better insights into the data and attain higher efficiency—automating resource intensive tasks and freeing up analysts for more integrative analysis (Beerman et al. 2021). As pointed out, there is room for improvement in this field, since less than 25% of sup-tech systems are exclusively intended for quantitative purposes. Following this lead, this work develops a methodology to address each of the risk perspectives in the RAS methodology: credit, market, operational and profitability.

Related Work

The use of machine learning for risk assessment has been a highly debated topic, both from an academic and industry standpoint. Since the 2000s (Galindo and Tamayo 2000), risk assessment has been recurrently identified as a top-priority investment for developing the data literacy of financial institutions. As recently shown by Antunes (2021), risk assessment by central banks is paramount for accurate supervision and is less biased than the self-assessments carried out by the banks themselves.

Additionally, Galindo and Tamayo (2000) established that tree-based models perform consistently better than artificial neural networks (ANNs) considering structured financial data. This finding is one of the pillars of our approach and it has been confirmed by several other authors (Chen and Guestrin 2016; Climent et al. 2019; Xia et al. 2017).

In their literature review, Leo et al. (2019) highlighted the popularity of machine learning applications for risk management in banking industry, while also noting the experimental nature of most approaches. The authors also remark the discrepancy between the high level of academic research concerning this area versus the de facto industry applications.

This debate has focused around two main issues:

- Finding the right risk assessment measure.
- Finding an adequate machine learning algorithm to build a risk assessment model.

Guerra and Castelli (2021) studied both of these aspects appraising several methodologies for assessing distress signals. This review spans from 2004, when Hillegeist et al. (2004) turned the page on two landmark methods (the Z-score (Altman 1968) and the O-score (Ohlson 1980)) by proposing the use of the Black–Scholes–Merton option-pricing model, up until 2019, when Kou et al. (2019) listed the most common methodologies for assessing systemic risk in the financial system.

On the same topic, Climent et al. (2019) used XGBoost to identify the best predictors of bank failure and develop a classification model to label failed and nonfailed banks in the Eurozone. The data used in their study comprised 25 annual financial ratios for commercial banks.

The majority of the current literature converts the risk assessment problem into a binary classification task, where each bank is labelled as "failed or likely to fail" or "no fail" (Climent et al. 2019; Filippopoulou et al. 2020; Kolari et al. 2019; Leo et al. 2019; Wang et al. 2021). These studies usually rely on public datasets, where the target variable is derived from a set of financial ratios.

At central banks, as clearly pointed out by Stock and Watson (2001), economists are responsible for conducting risk analysis and performing scenario testing.

Since the appearance of the Single Supervisory Mechanism (European Commission 2015) we are bearing witness to a standardisation of reporting requirements and methodologies. The heterogeneous landscape of financial performance measures identified in the literature has been increasingly replaced by the use of the Supervisory Review and Evaluation Process (SREP) (European Central Bank 2022), leading us to leverage this risk assessment methodology. SREP is an ongoing work by the European Central Bank (ECB) and the National Central Banks (NCBs) that provides an integrated view on each bank according to five risk perspectives: liquidity, credit, market, operational and profitability.

The Risk Assessment System (RAS) is the quantitative pillar of the methodology, and it is the focal point of this work.

Selecting the adequate machine learning methods applied to central banking, we found that it recently became a hot topic from both an academic and NCBs standpoint (Alonso and Carbo 2020; Antunes 2021; Huang et al. 2021; Lee and Shin 2020; Wang et al. 2021). Beerman et al. (2021) report that the pandemic prompted NCBs to rely on sup-tech solutions in their everyday processes. Several of the surveyed authorities already have operational systems. For instance, the Central Bank of Brazil has a tool that examines the whole credit portfolio of a bank to detect exposures with unrecognised expected losses; the Bank of Spain is applying inference maps to model the relationships between borrower and evaluate the risk impact; and the Monetary Authority of Singapore is developing a tool to automate data analysis so that supervisors can rely on complete datasets, instead of sampling. For this reason, we expanded our research to applications of ML to risk assessment. By broadening this research, we can evaluate how ML has been used for financial structured data and then focus on the central bank case.

Stress testing is one of the many forms of risk assessment that is particularly used at central banks. Kolari et al. (2019) challenged the concept of a bank's resilience by suggesting that it mostly represents a bank's ability to deal with a specific risk supported by its own capacity to absorb it. In such a setting, applying a risk-focused methodology such as SREP allows supervisors to better assess the root causes of what might otherwise be perceived as a general business model issue.

Chakraborty and Joseph (2017) presented a series of ML applications for financial problems and they analysed the most frequently used algorithms, such as tree-based ensembles, artificial neural networks and clustering techniques. The authors also showcase three use-cases at central banks, that establish ML as a better solution than traditional statistics. The most relevant for our work is one that develops a series of alerts (EWS) based on the balance sheet structure of a bank, in a supervisory context. This shows not only how relevant supervisory data are for a proactive risk assessment, but also how this data can be used to sense the risk proclivity of supervised institutions.

Recent technological developments have allowed newer and more complex models to emerge (Strydom and Buckley 2019), such as deep learning (DL) and extreme gradient boosting (XGBoost) (Abellán and Castellano 2017). Evidence shows those analysis methods have a unique capacity of capturing the intricacies of financial phenomena (Huang et al. 2021; Ribeiro et al. 2012).

Iturriaga and Sanz (2015) showed that modelling time series is where DL excels. Moreover, Petropoulos et al. (2018) leveraged DL's precision and developed an Early Warning System (EWS) for predicting the failure of Greek banks (data in 2005–2015). This is a landmark report on the use of advanced ML in a daily supervisory context. Wang et al. (2021) proposed an add-on to the conventional logit-based EWS, which involves simulating expert voting through a Random-Forest-based system, and that showed valuable results in predicting systemic crises.

Broeders and Prenio (2018) organise supervisory innovation concepts and present a series of use-cases where early adopters are implementing innovative approaches (sup-tech), converting retrieved data into predictive indicators. These works are of great importance to systematise how to implement this technology. The increasing amount of available data is one of the main drivers for the development of ML-based systems, as Chakraborty and Joseph (2017) also have claimed. Banking supervision acknowledges the benefits of innovative technologies and the importance of keeping up with the variety of sup-tech initiatives being developed. These initiatives have the potential to dramatically change the supervisory process; anticipating the consequences of current behaviours instead of belatedly reacting to past events. The same authors explore several use-cases from the Central Bank of the Republic of Austria (OeNB), the Monetary Authority of Singapore (MAS), the Securities and Exchange Commission (SEC), among others. Business process effectiveness, cost reduction and increased analytical capability are noted as the main

drivers for the sup-tech endeavour. These supervisory agencies report several challenges in exploring and implementing these technologies, such as:

- The technical know-how and appropriate infrastructure to support these analytical solutions;
- The legal framework to support the use of the relevant information;
- The internal support from management to invest in these initiatives and from the end-users, to provide the expert knowledge and to use and promote the new analytical tools.

Financial Stability Board (2020) also shows how the balance of supply and demand ignited the development and use of sup-tech tools. From the demand side, these authors mention, among other aspects, enhanced supervisory and regulatory requirements and improved risk management capabilities, where the automation of data retrieval and summarisation can drastically improve supervisory processes. From the supply side, the increasing availability of data and new analytical methods are among the top supporters of the above-mentioned regulatory necessities. Listed benefits of implementing these ground-breaking tools include:

- Enhanced analytical capabilities;
- Enriched visuals, stemming from state-of-the-art data collection to sophisticated dashboards;
- Reduced costs, as a consequence of automation.

Nevertheless, adopting new analytical processes inevitably brings on fresh challenges. Recognising this aspect, Jagtiani et al. (2018) expand on the impacts of these new analytical solutions and possible risks of adopting them, such as:

- Third-party vendor risk, where banks give access to outside specialists—data scientists and business users involved in setting up the tool—that can lead to data breaches. Additionally, if the vendor is a dominant player in the market, that circumstance can create a single point of failure in the financial system.
- Cyber-security risk, which is related to the previous topic, as vendors might not comply with supervisors' security requirements. Additionally, by allowing for external sources of data, banks and central banks become exposed to that channel and the information therein contained.
- Model risk, where systems based on complex machine learning models or even black-boxes make decisions that might not make sense from a business perspective, hence providing wrong predictions.

Another factor with major impact in ML use is the comprehensibility of the models. Although ML models are seldom capable of explaining prediction, they consistently outperform the classic approaches. Dastile et al. (2020) published a systematic literature review contrasting these techniques for a credit scoring problem, and they stress the lack of interpretability of DL as the main barrier for adoption in supporting financial decisions.

It is worth bearing in mind that pointing out the direction of future research is as important as signalling risks associated with implementing ML models. Kou et al. (2019) present a thorough report on state-of-the-art applications and ML techniques to assess systemic risk. Based on the existing technology, they suggest several future work areas, such as big-data analysis, data-driven research and policy analysis with data science.

2. Methodology

Developments in the area of data science and machine learning usually fall into one of two categories: developing a new computational method to better solve an existing problem; or alternatively, using existing methods to address a new problem. In this work, we aim to address a problem that was yet to be solved using machine learning: supervisory risk modelling.

Figure 1 illustrates how we attained our objective in a step-by-step diagram, as a development of what was presented in Guerra et al. (2022). The first step comprises a

data-retrieval process from the *Banco de Portugal* supervisory data system, including a wide set of features and the target variables we want to model. Next, there is a transformation process that is responsible for cleaning the data, dealing with missing values and selecting the most significant features. In the following step, we compare the ML models for this task using train-test split, cross-validation and the TPOT AutoML framework (Olson et al. 2016). The f1-score and confusion matrices are used to compare the results and select the best model that can then support an Early Warning System for the RAS risk perspectives.

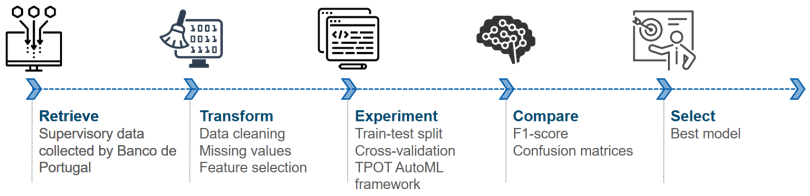


Figure 1. Methodology process overview.

In this section we present the steps carried out in this research, beginning with explaining how the data was retrieved, what transformations were required, which features were selected and its criteria, and finally, how the model's performance was evaluated.

2.1. The Data

One of the main pillars of this paper are the supervisory data collected by the *Banco de Portugal* (BdP) within the Capital Requirements Regulation (CRR) and Capital Requirements Directive IV (CRD IV) (European Parliament 2013). Our data ranges from March 2014 until August 2021 and most of the data used for the purposes of this paper are quarterly (European Banking Authority 2013). Due to confidentiality issues, the dataset used in this study cannot be made available for public consultation.

Data are extracted via an SQL query from BdP's production database (with no filters regarding reference date, banks or their consolidation level) into a *comma-separated-values* (csv) file. The result set is imported using a Python script within Jupyter notebooks. An extraction routine was implemented to assure consistency and automation in data gathering.

To account for all possible predictors, we have selected our feature space from the four main reporting frameworks for banking supervision: Financial Reporting, Common Reporting, Asset Encumbrance and Funding Plans.

The data resides in a relational database where each row represents a reported value. This means that in the data source, several rows represent a single observation. During extraction, data are anonymised using MD5 algorithm within an SQL's hashing function. This step assures the same identifier for every row in the same observation. The extracted dataset follows this column schema:

1. ID—a hash code representing each observation's identifier;
2. variable—a code with business meaning that represents each reported value;
3. val—the actual numeric value of the variable.

2.2. Transformations

Preparing the data for machine learning algorithms is the single most critical stage in such studies and projects. The first step in our study is to pivot the data with the aim of having each row corresponding to one observation. This transformation uncovers the sparsity of our feature space, requiring null columns to be dropped.

Another important step is to focus the dataset on the risk perspective to be evaluated. In our study we are addressing credit, market, operational and profitability risks. When investigating one risk perspective—one specific target variable—we drop all the others. This might lead to invalid observations, that is, observations that only made sense for a certain risk. As a consequence, we discard the rows for which the selected target value is null.

Dealing with missing values is the final step of the transformations phase. Our dataset is exclusively numeric and each column/feature has its own distribution. Therefore, we opted to input the missing values of each feature with the median, since this provides a more accurate perspective of the data's distribution when dealing with up to 20% missing values (Acuna and Rodriguez 2004).

By the end of these steps, our dataset consisted of 9262 rows and 82,576 columns.

2.3. Feature Selection

As we saw in the previous subsection, this dataset is extremely sparse—here, the inaccuracy of the term "extremely" endeavours to capture the fact that this is a wide dataset (more features than observations). Although we have considered using Principal Component Analysis (PCA), this method compromises model comprehensibility; since it projects the original features into a lower dimensionality feature space, there is always information loss from discarding the components with less variance/information. The selection criteria is based on the covariance matrix, and does not account for the target variable to be studied. As this dataset comprises five different target variables—one per risk—PCA might exclude features regardless of their contribution to a specific target.

To address the above-mentioned issues, we have used the Random Forest feature-selection algorithm, with an 85% threshold for feature importance. Tree-based models are best to perform this task since they not only take into account the target variable to be explained, but they also rank features according to how well they improve the purity of nodes (gini impurity) a priori. The closer a node is to the root, the greater impurity decrease occurs (i.e., the "cleaner" data becomes). Contrarily, leaf nodes have smaller impurity decrease. Hence, pruning below a certain node results in a subset of the most important features.

This method allowed us to technically assess the list of features that explain at least 85% of our target variable. From the original total of 82,576 features we selected 2608 features—for credit risk. This number varied for different target variables.

As a final check, we have computed the correlation matrix for each target variable to assure features and target were not highly correlated, with a Pearson's correlation coefficient less that 0.3.

2.4. Experiments

In the following subsections we lay out the three approaches followed to assess the best machine learning model:

- Train-test split: simply splitting the dataset in train and test sets.
- Cross-validation: using different partitions of the data to test and train the model on every observation, iteratively.
- TPOT Auto ML: an auto ML framework by Olson et al. (2016), for comparison purposes.

These approaches provide a performance measure that summarises the generalisation capability of every model and allows for a reliable and fast comparison among models. F1-score was used as a performance measure since it keeps a balance between precision and recall. Furthermore, since we observe uneven class distribution in the dataset, F1-score is more appropriate than the Area Under the Curve (AUC) (f1-score gives a score for a specific thresholds, whereas AUC averages over all possible thresholds). For a full detail of each evaluation, the confusion matrices are also provided.

For the purposes of this study we selected and evaluated the performance of each of the following models:

- Logistic Regression (LG), used only for benchmarking;
- k-Nearest Neighbours Classifier (kNN);
- Random Forest Classifier (RFC);
- Extreme Gradient Boosting Classifier (XGBC).

The TPOT framework is an AutoML framework that makes use of genetic programming to optimise the process of finding the best model to the problem at hand. This is a rising trend in the usage of machine learning and we have included it in order to to evaluate its adequacy to this problem.

All three approaches comprise an optimisation phase, where we experiment with a range of values for the hyper-parameters of each of the considered models. For both the train-test split and cross-validation we carried out a 5-fold cross validated grid search for the specific parameters of each model. The TPOT framework has an optimisation step within its pipeline that is fully documented.

Just before feeding the data to the ML algorithms, we used the MinMaxScaler to fit the features in the same scale. This approach preserves outliers and the original distribution of each feature, hence conserving the information embedded in the data.

All the experiments were executed at the *Banco de Portugal* using its computing infrastructure. The specifications of the node assigned to these experiments were the following:

- 4 Intel(R) Xeon(R) CPUs E7-8891 v4 @ 2.80 GHz, 32 GB of RAM, 1 TB SSD;
- Ubuntu 20.04.3 LTS;
- Python 3.8.10;
- Pandas 1.2.0;
- scikit-learn 0.24.0;
- TPOT 0.11.7.

2.4.1. Train-Test Split

Train-test split is the standard approach to model evaluation and the one we have used to begin with. The initial three-fold split was 60-20-20 for train, validation and test sets, respectively, and we organised the experiment in the following steps:

1. Prepare the data as specified in the previous subsection;
2. Iterate through the machine learning models considered previously;
3. Fit each model to the training data;
4. Use the validation set to run an hyperparameter optimisation process;
5. Compute the relevant scoring measures for train and test phases, along with the confusion matrices;
6. Persistently store the results.

2.4.2. Cross-Validation

Train-test split provides a good approximation of a model's potential performance on a specific dataset. However, for small-to-medium datasets, splitting the data might prove inaccurate, since the training set will probably misrepresent our universe of events, and overestimate the overall performance of the model.

In order to avoid this pitfall, we have used StatifiedKFold, a specific implementation of cross-validation within *scikit-learn* that preserves the proportion of samples among classes.

Cross validation splits the dataset in a specified number of folds and provides models of each of the folds as train and tests sets. This strategy balances the scores of the several splits, providing a more accurate view of how the model will perform on unseen data.

Despite its numerous advantages, the use of cross validation will concurrently entail difficulties, the most common being data leakage. This happens when the model trains from both training and test sets. The authors avoid this problem by providing the cross validation function the complete dataset and performing the necessary data transformations within each iteration. Arguably, this approach comes at a cost, but the benefit of assuring that no data leakage will happen largely compensates for the performance deterioration.

After training the models on the data, choosing the proper performance measures is key to correctly evaluating and comparing each resulting model. For this experiment we have chosen the f1-score as an overall performance measure and the confusion matrix for a detailed view on each model's classification decisions:

- f1-score represents the harmonic mean of precision and recall. It is most suited for uneven class distributions, as it is the case of our dataset. It is calculated as

$$\text{f1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

where

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

- The confusion matrix is an N $\times$ N matrix with each of the rows representing each class prediction, and the columns each actual value provided. In our case, classes 1, 2, 3 and 4 represent the risk tier of a given bank.

To achieve these measurements, we have evaluated each model through the following steps:

1. Defining a pipeline for scaling and training—MinMaxScaler;
2. Establishing a 5-fold cross validation—StratifiedKFold;
3. Using *cross-val-predict* to assess each model's generalisation capability;
4. Performing a nested cross-validated loop to tune the hyperparameters of each model;
5. Computing the performance measures—f1-score and confusion matrices;
6. Storing the results.

2.4.3. Optimisation Process

In order to improve the robustness of the two previous approaches, we carried out an optimisation step, as mentioned in steps listed above. This hyperparameter tuning phase is in everything similar for train-test and cross-validation, except for the data used to optimise those parameters. For the latter, we used 20% of the data corresponding to the validation set described in the respective subsection. For the former, we used a double, or nested, cross-validation. This is the preferred way for avoiding the bias created by selecting and tuning a model in the same data (Cawley and Talbot 2010).

In both approaches we used GridSearchCV with five-fold cross-validation to tune the hyper-parameters of each model. The parameters tuned for the purposes of this work were the following:

- Logistic Regression
  - **C**—100, 10, 1.0, 0.1, 0.01.
  - **penalty**—none, l1, l2, elasticnet.
  - **solver**—newton-cg, lbfgs, saga.
- k-Nearest Neighbours Classifier
  - **n_neighbors**—{1,2,3,...,21}.
  - **metric**—euclidean, manhattan, minkowski.
  - **weights**—uniform, distance.
- Random Forest Classifier
  - **n_estimators**—10, 100, 500.
  - **max_features**—sqrt, log2.
  - **criterion**—gini, entropy.
- Extreme Gradient Boosting Classifier
  - **max_depth**—3, 7, 9.
  - **n_estimators**—10, 100, 500.
  - **learning_rate**—0.001, 0.01, 0.1.

2.4.4. TPOT—An AutoML Approach

The first automation initiatives in model selection through grid search and similar methods were described in the 90s. The term *AutoML* has been used ever since, and a commercial version made its debut in 2018 (Zöller and Huber 2019).

In this work we opted for TPOT AutoML framework to accompany with this rising trend and compare it using our real-word problem. This framework employs genetic programming techniques to hone the model selection pipeline, by providing state-of-the-art optimisation methods to a broader audience:

1. **generations**—selected value was 5—represents the number of iterations given to the optimisation pipeline. By increasing the number of iterations, we increase the chances of finding a better (or even optimal) solution, always at the cost of time and computational resources.
2. **population_size**—selected value was 50—is the number of solutions in each generation, as a subset of the total population of solutions.
3. **cv**—selected value was 5—is the number of folds considered in the cross validation function—StratifiedKFold.
4. **verbosity**—selected value was 3—determines the amount of information TPOT shows to the user during run-time.
5. **scoring**—selected value was f1—determines the scoring method for the models.
6. **n_jobs**—selected value was 16—determines the number of processes to be used in parallel.
7. **random_state**—selected value was 42—is the random generator seed used to assure the same results across executions, given the same inputs.

By following these structured steps—feature preprocessing, feature selection, model training, optimisation and scoring—we assure the comparability of the three approaches.

3. Results and Discussion

In this section we present the results and their discussion, structured by risk perspective. We analyse how our observations span through the risk classes. Each risk perspective is described in terms of performance of the models assessed using the forementioned approaches: train-test split, cross-validation and TPOT.

Figure 2 shows the distribution of observations in our dataset by risk class. Most risks show a balanced distribution of classes, except for credit risk, which has significantly more observations on class 2. Using oversampling and undersampling techniques to deal with this issue was pondered. However, this distribution reflects the frequency of each class in the real world, so as a consequence we decided not to balance the dataset.

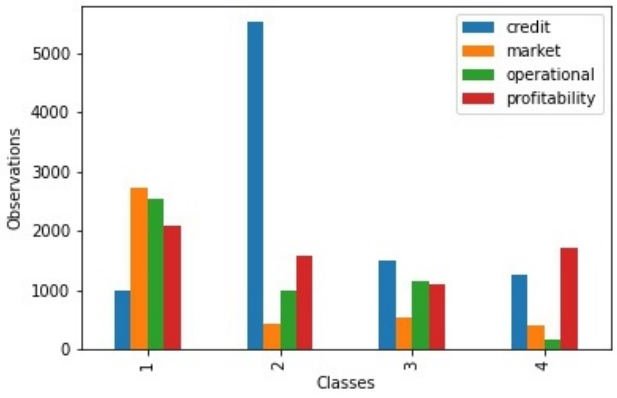


Figure 2. Number of observations in the dataset per target class, per target variable.

3.1. Credit Risk

The evaluation of credit risk was performed in a sample with 9262 observations, and 2608 features after the feature selection process. The processing wall time used to evaluate the performance of the listed models for credit risk were:

1. Train-test split: 28 min and 48 s;
2. Cross-validation: 1 h, 57 min and 44 s;
3. TPOT framework: 2 days, 15 h, 20 min and 34 s.

Figure 3 shows the results of each model comparing its training and test scores.

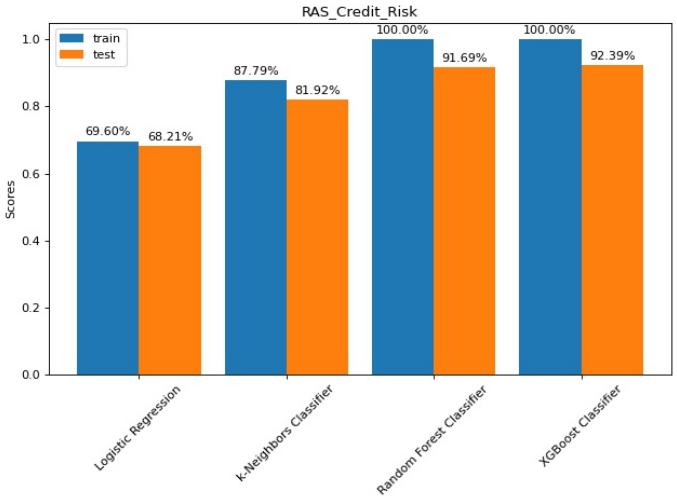


Figure 3. F1-scores of each model, using train-test split approach.

The Logistic Regression presents average results, slightly below 70% in both train and test sets. This could suggest that we are dealing with more complex decision boundaries. k-Nearest Neighbours shows better results, suggesting that the classes in our dataset are not linearly separated and they might not be independent. This is often the case with financial reporting data: variables are not completely independent from each other and the heterogeneity of the data creates more complex boundaries between classes.

When applying tree-based models with ensembles, such as Random Forest and Extreme Gradient Boosting (XGBoost), we see a 10 to 20% increase in test performance. The work by Chang et al. (2018) shows how these techniques capture the heterogeneous structure of financial data, making these models the most adequate choice.

Figure A1 in the Annexes, shows the detailed classifications of each model through their confusion matrices, supporting the above-mentioned findings.

For a more precise view on how the intricacies of the data affect the performance of these models, we applied cross-validation with with a hyperparameter optimisation process to assess the f1-score of each model (Figure 4). As already mentioned, this measure represents the harmonic mean of precision and recall, ideal for multilabel classifiers and imbalanced datasets. The figure also includes the results of TPOT, the autoML framework considered in this study. By using the whole dataset for the several train-test splits in cross-validation, we ensure that the final score is not biased to any specific split, missing some particular event that compromises the models' performance on unseen data.

In terms of relative performance, the models performed similarly when compared to each other. XGBoost presents the best performance, even when compared to TPOT—this last framework being resource-intensive and needing almost three days to optimise its pipeline. Furthermore, in terms of time and performance gains, it does not outperform XGBoost.

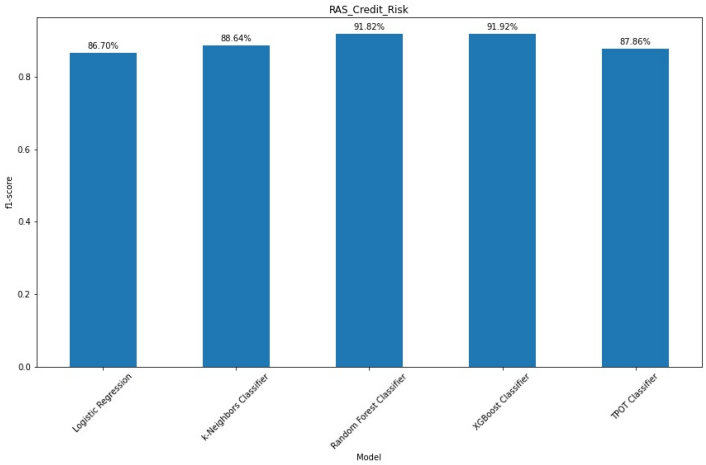


Figure 4. F1-scores of each model, using cross-validation approach.

The confusion matrices in Figure A2 offer a detailed perspective on each model's classification decision. As with the train-test split, we see k-Nearest Neighbours being penalised by class 2 imbalance for credit risk and XGBoost missing the least classifications.

3.2. Market Risk

This subsection presents the results of evaluating the market risk perspective. This sample is composed by 4080 observations and 3539 features, selected through Random Forest feature selection process. The wall time of each of the approaches was:

1. Train-test split: 4 min and 14 s;
2. Cross-validation: 16 h, 40 min and 8 s;
3. TPOT framework: 18 h, 44 min and 16 s.

The results of the train-test split evaluation are shown in Figure 5. Here the results show a distribution similar to what we observed with credit risk; however, the scores are slightly better.

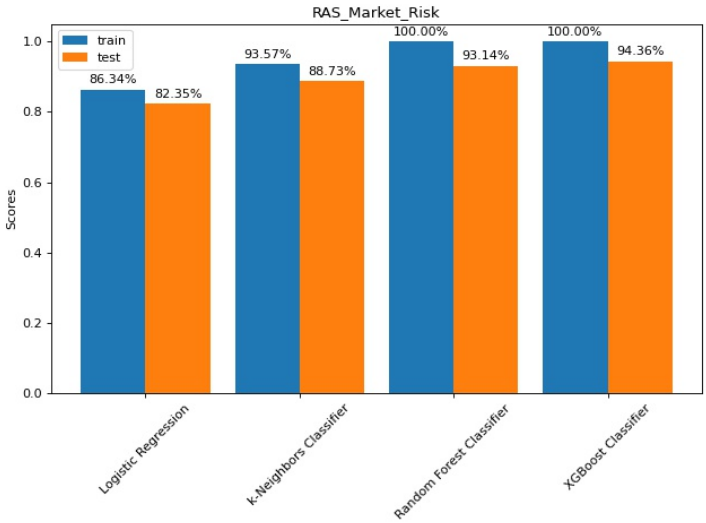


Figure 5. F1-scores of each model, using train-test split approach.

The Logistic Regression results suggest that we face linear (or close to linear) boundaries between classes. This reading is also supported by the fact that its score is closer to k-Nearest Neighbours'.

Still, the use of ensemble tree-based models show a significant increase in performance. The spike is not as prominent as with credit risk, and Random Forest has again a similar,

but lower, score than XGBoost—on the order of the decimal percentage points. Figure A3 shows the confusion matrices for the train-test split evaluation.

However, a random train-test split might give an undervalued or overvalued perspective of a model's performance. To validate these findings we applied cross-validation with f1-score to the whole dataset. The results of this process are shown in Figure 6 along with the evaluation of the TPOT framework.

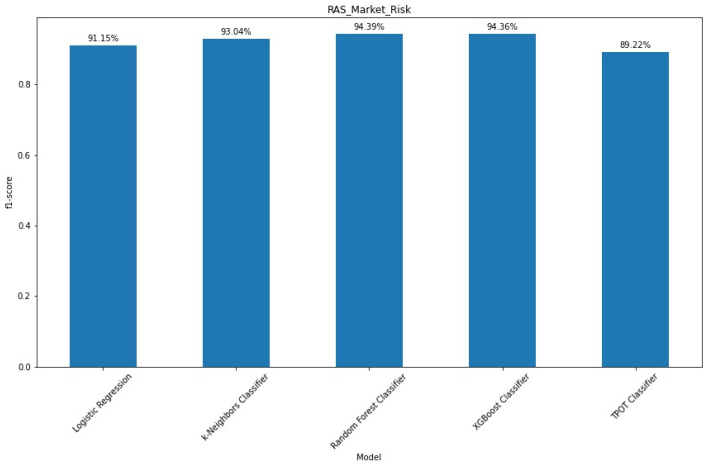


Figure 6. F1-scores of each model, using cross-validation approach.

The models show similar scores when compared to each other, with Logistic Regression faring close to the k-Nearest Neighbours. Contrarily to what we observed in the train-test split, a more discerning look at the results shows that Random Forest classifier slightly outperforms XGBoost. TPOT comes in third place in terms of performance, and it becomes even less appealing if we consider its wall time. Figure A4 presents the confusion matrices for this classification process.

3.3. Operational Risk

The sample provided to evaluate operational risk has 4819 observations and 3447 features. The wall time needed to evaluate the models on this sample was:

1. Train-test split: 5 min and 19 s;
2. Cross-validation: 18 h, 13 min and 52 s;
3. TPOT framework: 2 days, 15 h, 31 min and 41 s.

The train-test split results shown in Figure 7 paint a different picture than the other perspectives. Although we can observe a similar distribution of results, the Logistic Regression presents below-average results on unseen data. Furthermore, the k-Nearest Neighbours classifier exhibit a slight improvement to the previous model.

Random Forest and XGBoost classifiers again come into the spotlight, with the latter showing a modest advantage of less than two percentage points. Figure A5 shows the confusion matrices for the train-test split, for a detailed view of each classification.

Applying cross-validation to this sample reveals several performance adjustments (Figure 8). Our non-tree-based models—the Logistic Regression, and k-Nearest Neighbours classifier—expressed an increase in their score, due to the optimisation process.

For Random Forest and XGBoost we see minor adjustments in the f1-score, however, their performance difference is consistent with the train-test split approach. This finding confirms the ability to grasp the heterogeneity of regulatory financial data. The TPOT framework is again in third place, revealing to be a poor choice due to the more than 2.5 days of processing. Figure A6 shows the confusion matrices of the cross-validation process, for a detailed view of the classifications of each model.

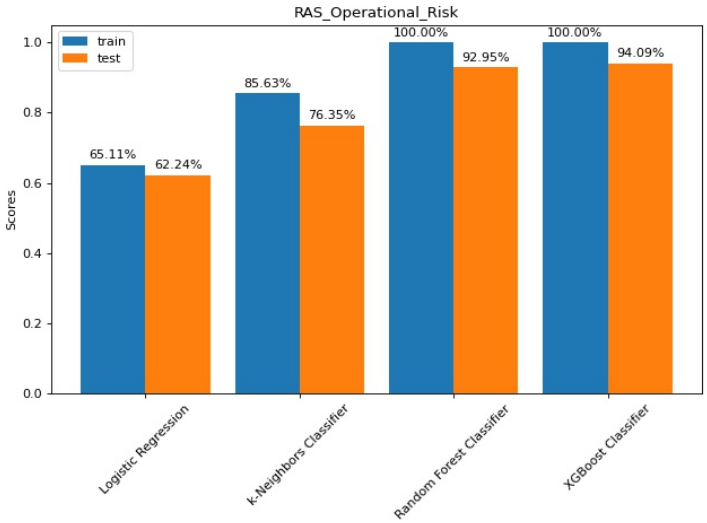


Figure 7. F1-scores of each model, using train-test split approach.

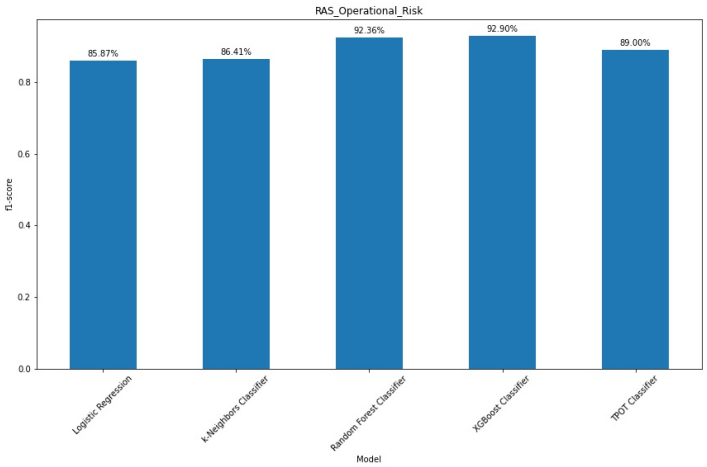


Figure 8. F1-scores of each model, using cross-validation approach.

3.4. Profitability Risk

As for our final risk perspective—profitability—we used a sample of 6448 observations and 3177 features. The processing and evaluation times for each of the approaches were:

1. Train-test split: 9 min and 14 s;
2. Cross-validation: 1 day, 2hours, 25 min and 58 s;
3. TPOT framework: 1 day, 11 h, 56 min and 42 s.

This is the risk perspective with the worse overall results. Figure 9 shows the train and tests scores for each model. Logistic Regression, k-Nearest Neighbours present a paltry performance. Even Random Forest and XGBoost show some decrease in performance, although still presenting good results. Figure A7 pictures the detailed classifications of these models through the confusion matrices.

The cross-validation process corrects for any misclassification resulting from a unfavourable train-test split. In Figure 10 we show the f1-scores for each model, including the TPOT framework.

Just as with train-test split, the Logistic Regression, and k-Nearest Neighbours present a low score, when compared to the other algorithms and their performance in other risk perspectives. Although this seems not related to class imbalance (see Figure 2), the complexity of the decision boundaries and the dependence of some of the features might be the root cause for these foundering results.

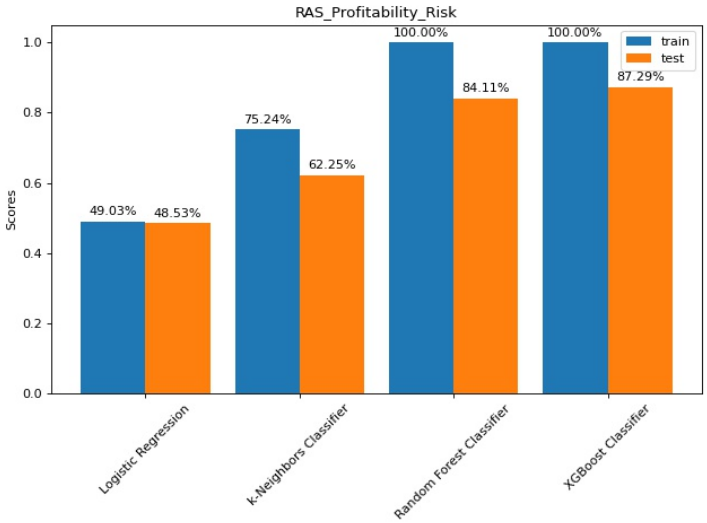


Figure 9. F1-scores of each model, using train-test split approach.

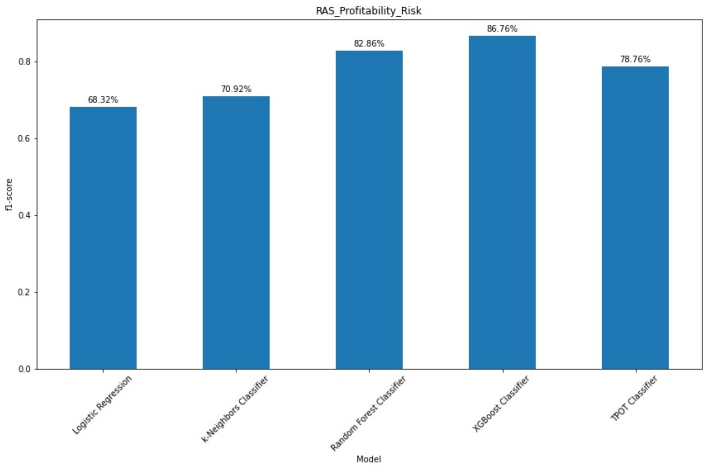


Figure 10. F1-scores of each model, using cross-validation approach.

Even so, the Random Forest and XGBoost show good results, with the latter again outperforming the former. The TPOT framework, comes in third with average results and one and a half day of processing, again making it an unsatisfactory alternative for this task. See Figure A8 for the confusion matrices of the cross-validation process.

3.5. Final Remarks

Following our previous work (Guerra and Castelli 2021), we clearly defined the required elements for modelling the supervisory risk assessment process comprised in RAS. First, we suggested the use of SREP's quantitative pillar—Risk Assessment System—as a standard methodology to compare the banks at European level. This methodology is already established across the Euro-area, hence making it the ideal choice for the task. Moreover, most works in this area adopt a binary classification of the risk level of the banks, limiting the classification to "failure" or "no failure". As mentioned before, this approach lacks the flexibility required for central banks to detect the effect of distress events gradually and earlier in time. This is accomplished through the progressive multiclass scale provided in the RAS. Additionally, we identified a research gap specifically addressing the supervisory use-case. Using real-world supervisory data, designed and retrieved for regulatory purposes, it has been proven to provide the most accurate outlook (Broeders and Prenio 2018; di Castri et al. 2019; Filippopoulou et al. 2020; Massaro et al. 2020).

We tested the above-mentioned elements and successfully modelled the liquidity risk of a bank (Guerra et al. 2022). Based on those findings, we set out to generalise the methodology and model the remaining risk perspectives comprised in the RAS: credit, market, operational and profitability.

From a technical standpoint, we confirmed that an optimised XGBoost outperformed the other considered models. This is accordance with previous literature results suggesting XGBoost performs best with structured financial data. In addition to that, we have tested it against the auto ML framework TPOT, a rising trend in the field. The results showed that due to the characteristics of the dataset—large number of features and sparse dataset—computing time was extremely taxing, even with low parameters for the GP algorithm. It might be interesting to reduce the number of features to fewer than 10, and see how TPOT performs.

From a business perspective, the novelty within the presented results is the fact that we are modelling a multi-class decision process with real-world supervisory data. Whereas other works have not explored supervisory data, we rely on the European regulatory framework and the data collected within it. This data is the pillar of supervisory processes and brings the structure and context to our models. By relying on these models, we can develop early warning systems capable of anticipating distress events, considering the risk measures above, and also give supervisors a tool to test alternative economic scenarios to prevent pitfalls.

4. Conclusions

Streamlining an effective supervisory methodology requires an integrated view of the risks a credit institution is subject to. In our previous work we have successfully modelled liquidity risk according to SREP methodology. Once that pillar was set, we were able to apply the same modelling techniques to the other risk perspectives comprised in the Supervisory Review and Evaluation Process (SREP) and its Risk Assessment System (RAS): credit, market, operational and profitability.

Based on the quantifiable mainstay of ECB's Risk Assessment Methodology, we classified credit institutions from the Portuguese banking sector according to their risk level on each of the perspectives encompassed in the methodology. We used real-life supervisory data and modelled this decision process by comparing several machine learning techniques, benchmarked against a widely used statistical method.

We have reached significant results clearly establishing that this decision process can be modelled and that the ML techniques used outperform the classic statistical approaches.

Regulatory supervisory data is highly correlated and heterogeneous, making the decision boundaries of this exercise a challenging task. Additionally, real-world events are seldom represented by balanced data. All risk levels are observed but with occurrences that are subject to events in a specific point in time. The complexities of such reality were best represented by ensemble tree-based models, such as Random Forest and XGBoost classifiers. These models can capture the heterogeneous nature of financial data and establish clear decision boundaries with little error—f1-score between 87% and 94%. These results were obtained after applying an optimisation process within the cross-validation cycle.

Given the computational resources available and the cutting-edge genetic programming optimisation pipeline available through TPOT, we expected it to outperform XGBoost. However, TPOT consistently came in third regarding f1-score, being outperformed by Random Forest and XGBoost. Its long processing times can be explained by the dedicated optimisation process, and the fact that our dataset is sparse (82,576 features). The feature selection process is costly in computational sense and it might account for a significant share of the wall time.

We firmly believe this work is a meaningful contribution to a set of stakeholders involved in risk assessment in the banking sector:

- National Central Banks (NCBs) can leverage the findings of this work and use these models to develop early warning systems. These sup-tech initiatives are currently in

the limelight, with many projects being developed in this area by the ECB, the Bank of International Settlements (BIS) and worldwide NCBs. A decision-support system such as this would provide an enhanced risk assessment perspective to supervisors.
- Banks and the consulting industry can convey these principles into their own systems. Consultancy companies can further support their clients in implementing their decision support systems using the data owned by the banks themselves. A bank can then proactively monitor and adjust their risk profile and strategy according to the regulatory requirements.
- Academia can use this work to extend and apply these types of ML methodologies to expand its usage on a regulatory perspective. Furthermore, we stress the postulates of using high-quality, highly validated relevant data, and adopting a universal methodology for risk assessment, one that standardises how to appraise a bank.

Through this paper, we aim to contribute to the technical understanding of ML that can be applied to sup-tech use cases according to the business needs. Grounded in historical supervisory data, we propose a sup-tech tool that improves the European supervisory risk assessment by providing early warnings on several risks.

Limitations and Future Work

There are several aspects we have identified over the course of this study that would merit revision and improvement.

The dataset we used in this work reflects the Portuguese banking sector. Ideally, expanding to the European level and using data from all central banks in the Euro-area would provide a complete supervisory perspective. Additionally, more diverse data, with more business models would strengthen the ML models presented here.

Each of the risks would also benefit from context-specific data, in order to enhance the generalisation of each model. This would also allow the supervisors to access more timely decisions. Supervisory data are mostly quarterly, which prevents quick reactions to adverse events. By combining this with daily data sources, such as market data, payments systems and credit responsibilities data, we might be able to obtain a daily signal for each aspect of a bank's risk. Confirming this decision path will strengthen the aforementioned models and provide a running risk assessment on which supervisors can rely.

The ability to explain the reasoning behind each model is of utmost importance, in particular for critical systems such as for crisis detection. Explainable AI benefits hold true not only for experts to validate the decision process carried out by the ML models, but also as common ground language to report any issue to banks. As such, investing in explainable models will deliver a better understanding of the technology, bringing supervisors closer to sup-tech, and will also set forth a clearer communication between institutions and central banks.

Combining our quantitative data with qualitative expert judgement, using Natural Language Processing (NLP), will allow for automated score adjustments based on internal supervisory notes and risk assessment reports.

As a final remark, consolidating the results of each risk model with the relevant qualitative data could provide a single integrated bank score as an additional measure for the SREP exercise.

Author Contributions: Conceptualization, P.G., M.C., N.C.-R.; Data curation, P.G.; Investigation, P.G.; Methodology, P.G., M.C.; Software, P.G.; Validation, P.G., M.C., N.C.-R.; Visualization, P.G.; Writing original draft, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

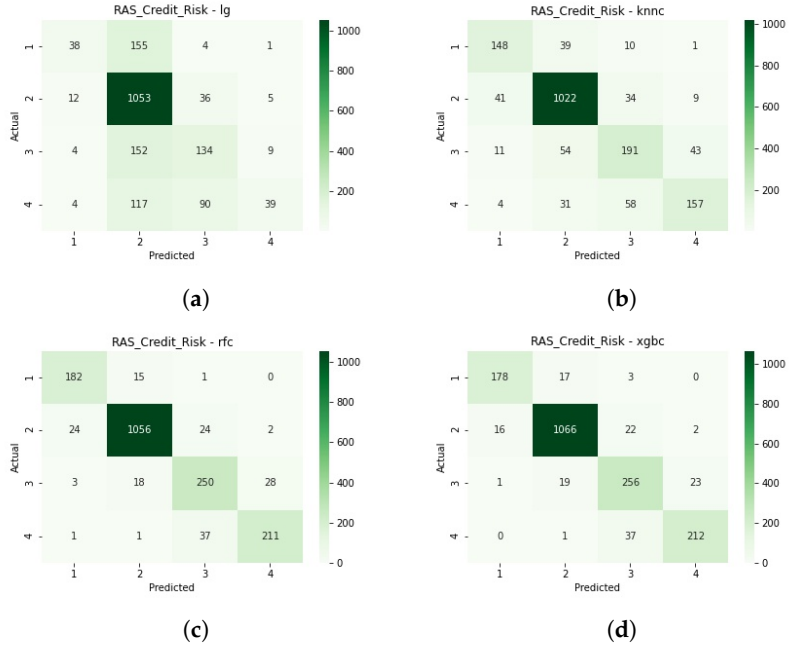
Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Disclaimer: The views, thoughts, and opinions expressed in the text belong solely to the authors.

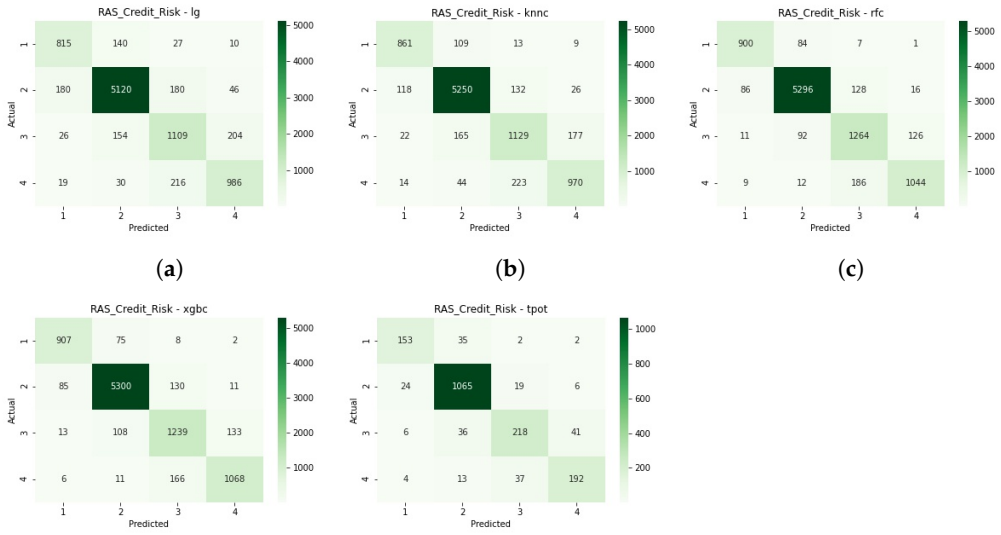
Appendix A. Confusion Matrices



(**a**) (**b**)

(**c**) (**d**)

Figure A1. Credit risk: confusion matrices generated when evaluating the above-mentioned models, using train-test split approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier.



(**a**) (**b**) (**c**)

(**d**) (**e**)

Figure A2. Credit risk: confusion matrices generated when evaluating the above-mentioned models, using cross-validation approach.(**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier. (**e**) TPOT classifier autoML framework.

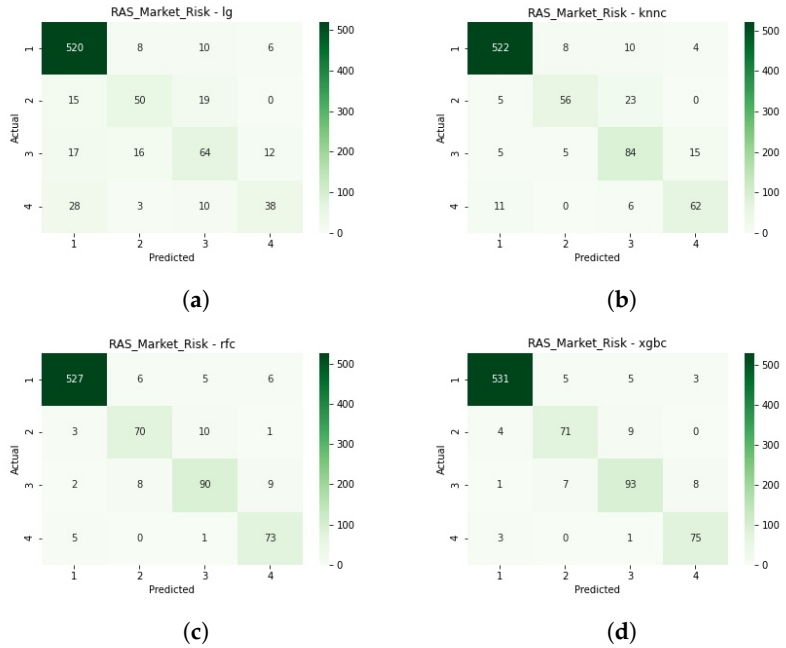


Figure A3. Market risk: confusion matrices generated when evaluating the above-mentioned models, using train-test split approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier.

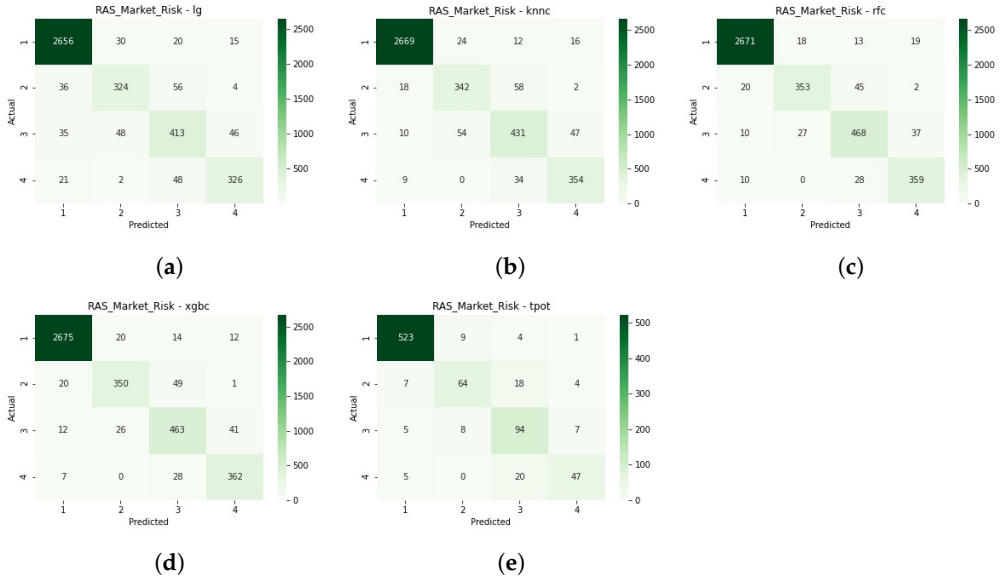
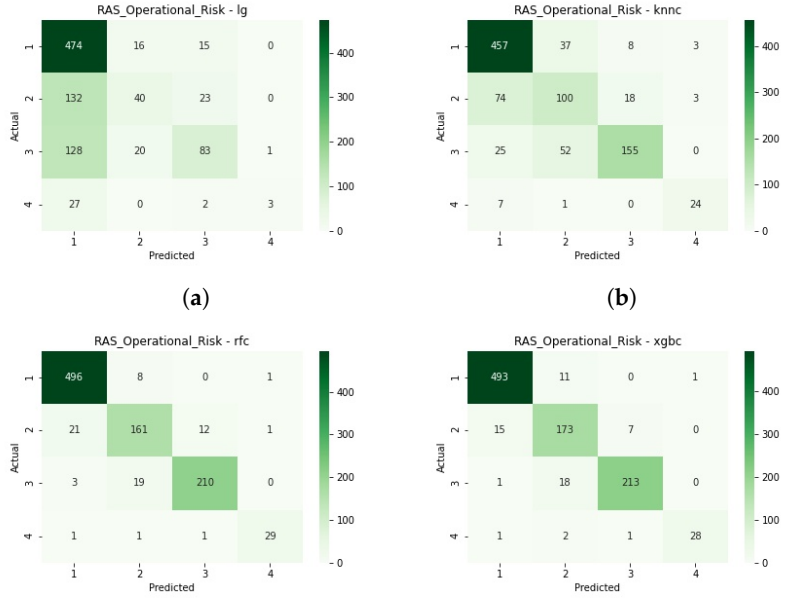


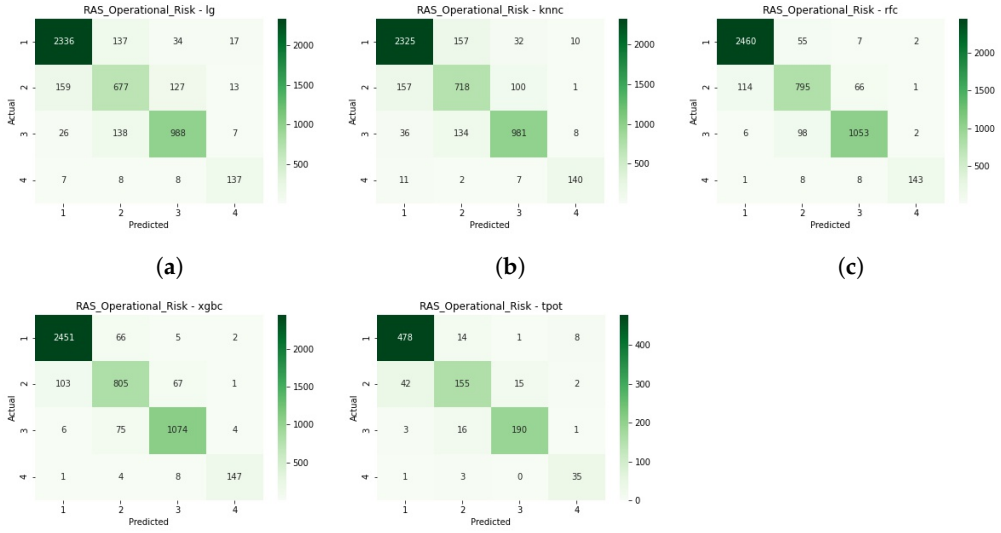
Figure A4. Market risk: confusion matrices generated when evaluating the above-mentioned models, using cross-validation approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier. (**e**) TPOT classifier autoML framework.



(**a**) (**b**)

(**c**) (**d**)

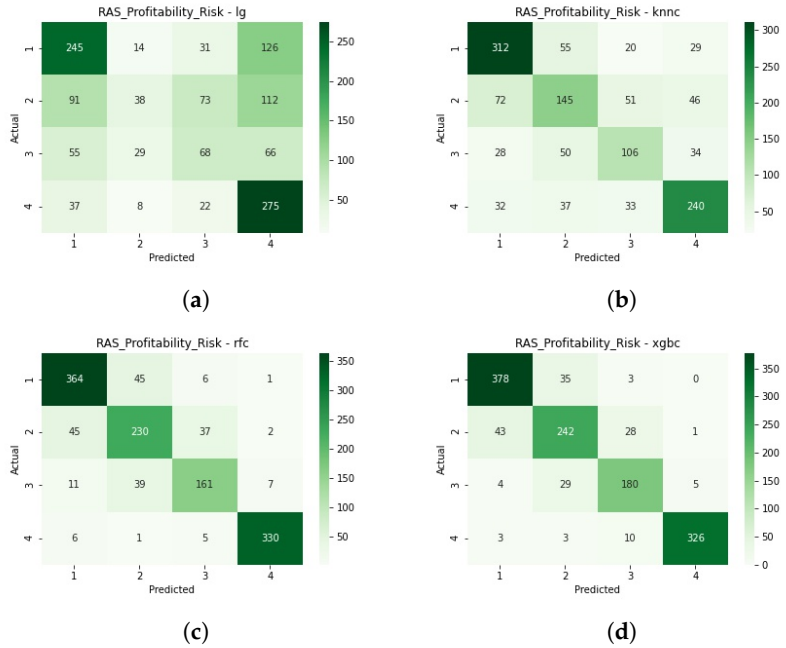
Figure A5. Operational risk: confusion matrices generated when evaluating the above-mentioned models, using train-test split approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier.



(**a**) (**b**) (**c**)

(**d**) (**e**)

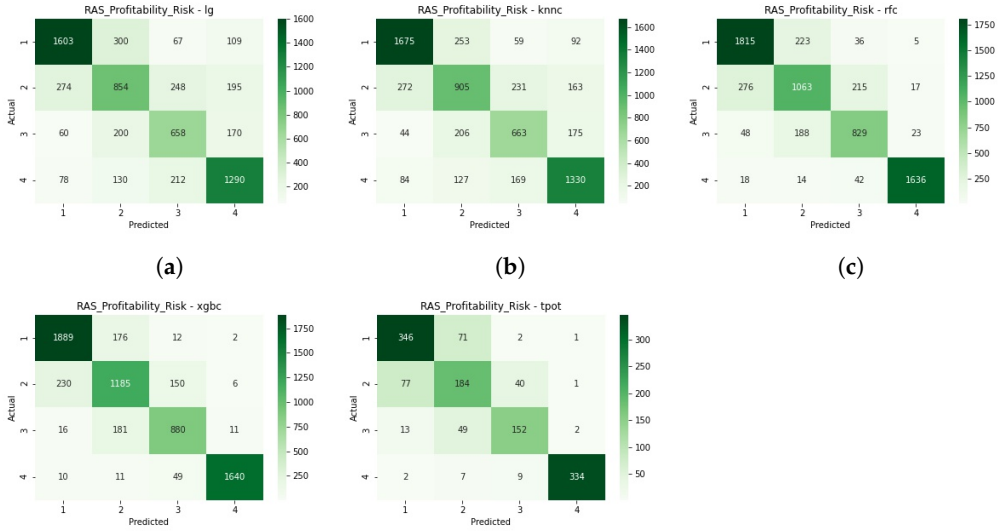
Figure A6. Operational risk: confusion matrices generated when evaluating the above-mentioned models, using cross-validation approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier. (**e**) TPOT classifier autoML framework.



(**a**) (**b**)

(**c**) (**d**)

Figure A7. Profitability risk: confusion matrices generated when evaluating the above-mentioned models, using train-test split approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier.



(**a**) (**b**) (**c**)

(**d**) (**e**)

Figure A8. Profitability risk: confusion matrices generated when evaluating the above-mentioned models, using cross-validation approach. (**a**) Logistic Regression. (**b**) k-Nearest Neighbours classifier. (**c**) Random Forest classifier. (**d**) Extreme Gradient Boosting classifier. (**e**) TPOT classifier autoML framework.

References

Abellán, Joaquín, and Javier G. Castellano. 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications* 73: 1–10. [CrossRef]

Acuna, Edgar, and Caroline Rodriguez. 2004. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*. Berlin and Heidelberg: Springer.

Alonso, Andrés, and Jose Manuel Carbo. 2020. Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost. *SSRN Electronic Journal*. [CrossRef]

Altman, Edward. 1968. Financial ratios, discriminant analysis and the prediction of corpporate bankruptcy. *The Journal of Finance* XXIII: 589–609.

Antunes, José Américo Pereira. 2021. To supervise or to self-supervise: A machine learning based comparison on credit supervision. *Financial Innovation* 7: 26. [CrossRef]
Beerman, Kenton, Jermy Prenio, and Raihan Zamil. 2021. Fsi insights no 37: Suptech tools for prudential supervision and their use during the pandemic. In *FSI Insights on Policy Implementation*. Basel: Bank for International Settlements.
Broeders, Dirk, and Jeremy Prenio. 2018. FSI insights innovative technology in financial supervision. In *FSI Insights on Policy Implementation*. Basel: Bank for International Settlements, vol. 29.
Casabianca, Elizabeth, Michele Catalano, Lorenzo Forni, Elena Giarda, and Simone Passeri. 2019. *An Early Warning System for Banking Crises: From Regression-Based Analysis to Machine Learning Techniques*. EconPapers. Orebro: Orebro University.
Cawley, Gavin C., and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11: 2079–107.
Chakraborty, Chiranjit, and Andreas Joseph. 2017. Machine learning at central banks. *SSRN Electronic Journal*. [CrossRef]
Chang, Yung Chia, Kuei Hu Chang, and Guan Jhih Wu. 2018. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing Journal* 73: 914–20. [CrossRef]
Chen, Tianqi, and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. Paper presented at the International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, pp. 785–94. [CrossRef]
Climent, Francisco, Alexandre Momparler, and Pedro Carmona. 2019. Anticipating bank distress in the eurozone: An extreme gradient boosting approach. *Journal of Business Research* 101: 885–96. [CrossRef]
Consoli, Sergio, Diego Recupero, and Michaela Saisana. 2021. *Data Science for Economics and Finance*. Cham: Springer International Publishing. [CrossRef]
Dastile, Xolani, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal* 91: 106263. [CrossRef]
di Castri, Simone, Stefan Hohl, and Arend Kulenkampff. 2019. Fsi insights on policy implementation no. 19: The suptech generations. *Financial Stability Institute* 19: 19.
Doerr, By Sebastian, Leonardo Gambacorta, and Jose Maria Serena. 2021. How do central banks use big data and machine learning? *The European Money and Finance Forum* 37: 1–6.
European Banking Authority. 2013. EBA Implementing Technical Standards (ITS). Available online: https://www.eba.europa.eu/regulation-and-policy/supervisory-reporting/implementing-technical-standard-on-supervisory-reporting (accessed on 22 January 2022).
European Central Bank. 2022. Supervisory Review and Evaluation Process. Available online: https://www.bankingsupervision.europa.eu/banking/srep/html/index.en.html (accessed on 22 January 2022).
European Commission. 2015. Single Supervisory Mechanism. Available online: https://www.bankingsupervision.europa.eu/about/thessm/html/index.en.html (accessed on 22 January 2022).
European Parliament. 2013. Directive 2013/36/eu. Available online: https://www.europex.org/eulegislation/crd-iv-and-crr/ (accessed on 22 January 2022).
Filippopoulou, Chryssanthi, Emilios Galariotis, and Spyros Spyrou. 2020. An early warning system for predicting systemic banking crises in the eurozone: A logit regression approach. *Journal of Economic Behavior and Organization* 172: 344–63. [CrossRef]
Financial Stability Board. 2020. The Use of Supervisory and Regulatory Technology by Authorities and Regulated Institutions: Market Developments and Financial Stability Implications. Available online: https://www.fsb.org/wp-content/uploads/P091020.pdf (accessed on 20 January 2022).
Galindo, Jorge, and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics* 15: 107–43. [CrossRef]
Guerra, Pedro, and Mauro Castelli. 2021. Machine learning applied to banking supervision a literature review. *Risks* 9: 136. [CrossRef]
Guerra, Pedro, Mauro Castelli, and Nadine Côrte-Real. 2022. Machine learning for liquidity risk modelling: A supervisory perspective. *Economic Analysis and Policy* 74: 175–87. [CrossRef]
Hertig, Gérard. 2021. Using artificial intelligence for financial supervision purposes. *Bank of England* 1–29. Available online: https://ethz.ch/content/dam/ethz/special-interest/dual/frs-dam/documents/Hertig%20WP%20AI%20and%20Financial%20Supverision%20(Feb-1-2021).pdf (accessed on 22 January 2022).
Hillegeist, Stephen A., Elizabeth K. Keating, Donald P. Cram, and Kyle G. Lundstedt. 2004. Assessing the probability of bankruptcy. *Review of Accounting Studies* 9: 5–34. [CrossRef]
Huang, Shian Chang, Cheng Feng Wu, Chei Chang Chiou, and Meng Chen Lin. 2021. Intelligent fintech data mining by advanced deep learning approaches. *Computational Economics* 1–16. [CrossRef]
Iturriaga, Félix J. López, and Iván Pastor Sanz. 2015. Bankruptcy visualization and prediction using neural networks: A study of u.s. commercial banks. *Expert Systems with Applications* 42: 2857–69. [CrossRef]
Jagtiani, Julapa, Larry Wall, and Todd Vermilyea. 2018. The Roles of Big Data and Machine Learning in Bank Supervision. pp. 1–11. Available online: https://ssrn.com/abstract=3221309 (accessed on 22 January 2022).
Kolari, James W., Félix J. López-Iturriaga, and Ivan Pastor Sanz. 2019. Predicting european bank stress tests: Survival of the fittest. *Global Finance Journal* 39: 44–57. [CrossRef]
Kou, Gang, Xiangrui Chao, Yi Peng, Fawaz E. Alsaadi, and Enrique Herrera-Viedma. 2019. Machine learning methods for systemic risk analysis in financial sectors. *Technological and Economic Development of Economy* 25: 716–42. [CrossRef]

Lee, In, and Yong Jae Shin. 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons* 63: 157–70. [CrossRef]
Leo, Martin, Suneel Sharma, and Koilakuntla Maddulety. 2019. Machine learning in banking risk management: A literature review. *Risks* 7: 29. [CrossRef]
Massaro, Paolo, Ilaria Vannini, and Oliver Giudice. 2020. Institutional sector cassifier, a machine learning approach. *SSRN Electronic Journal* 548. [CrossRef]
Ng, Jeffrey. 2011. The effect of information quality on liquidity risk. *Journal of Accounting and Economics* 52: 126–43. [CrossRef]
Ohlson, James A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18: 109. [CrossRef]
Olson, Randal S., Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. 2016. Evaluation of a tree-based pipeline optimization tool for automating data science. Paper presented at the Genetic and Evolutionary Computation Conference 2016, Denver, CO, USA, July 20–24, pp. 485–92. [CrossRef]
Petropoulos, Anastasios, Vasilis Siakoulis, Evaggelos Stavroulakis, and Aristotelis Klamargias. 2018. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *The Use of Big Data Analytics and Artificial Intelligence in Central Banking* 50: 30–31.
Pompella, Maurizio, and Antonio Dicanio. 2017. Ratings based inference and credit risk: Detecting likely-to-fail banks with the pc-mahalanobis method. *Economic Modelling* 67: 34–44. [CrossRef]
Ribeiro, Bernardete, Catarina Silva, Ning Chen, Armando Vieira, and João Carvalho Das Neves. 2012. Enhanced default risk models with svm+. *Expert Systems with Applications* 39: 10140–52. [CrossRef]
Shah, Syed Quaid Ali, Imran Khan, Syed Sadaqat Ali Shah, and Muhmmad Tahir. 2018. Factors affecting liquidity of banks: Empirical evidence from the banking sector of pakistan. *Colombo Business Journal* 9: 1. [CrossRef]
Stock, James, and Mark Watson. 2001. Vector autoregressions. *Journal of Economic Perspectives* 15: 101–15. [CrossRef]
Strydom, Moses, and Sheryl Buckley. 2019. *AI and Big Data's Potential for Disruptive Innovation*, 1st ed. Hershey: IGI Global. [CrossRef]
Vento, Gianfranco A., and Pasquale La Ganga. 2009. Bank liquidity risk management and supervision: Which lessons from recent market turmoil? *Journal of Money, Investment and Banking* 10: 79–126.
Wang, Tongyu, Shangmei Zhao, Guangxiang Zhu, and Haitao Zheng. 2021. A machine learning-based early warning system for systemic banking crises. *Applied Economics* 53: 2974–92. [CrossRef]
Xia, Yufei, Chuanzhe Liu, Yu Ying Li, and Nana Liu. 2017. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* 78: 225–41. [CrossRef]
Yang, Yimin, and Min Wu. 2021. Explainable machine learning for improving logistic regression models. Paper presented at the 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Palma de Mallorca, Spain, July 21–23, pp. 1–6. [CrossRef]
Zopounidis, Constantin, Michael Doumpos, and Nikolaos F. Matsatsinis. 1997. On the use of knowledge-based decision support systems in financial management: A survey. *Decision Support Systems* 20: 259–77. [CrossRef]
Zöller, Marc André, and Marco F. Huber. 2019. Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research* 70: 411–74. [CrossRef]