Norwegian School of Economics
Mallorca, Fall 2022

Machine Learning for Predicting Voluntary Audits

How do loan and risk factors influence small private commercial U.S. banks' decision to get no audit?

Amalie Husby & Liv-Kari Bakka Solås

Supervisor: Ulf Mohrmann

Master thesis, Economics and Business Administration

Major: Business Analytics & Accounting and Auditing

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

This master's thesis was written in the fall of 2022 at the Norwegian School of Economics (NHH); for Amalie as a part of the master's study in economics and administration with a specialization in Business Analytics, and for Liv-Kari as part of the master's study in accounting and auditing.

We would like to thank our friends and family for their support in writing this thesis. In addition, we would like to thank NHHS and NHHI for years of valuable experiences, challenges, and friendships during our time at NHH.

This thesis would not have been possible without our supervisor Ulf Mohrmann. We are grateful for the valuable advice and support he has provided during the process.

Norwegian School of Economics

Mallorca, December 2022

Amalie Husby

Liv-Kari Bakka Solås

Abstract

Commercial banks and other financial institutions are essential to the modern economy, and government agencies and regulators strive to identify and counteract risks in banking institutions. With an emphasis on loan and risk based factors, this thesis explores what influences small private commercial U.S. banks' decision to get no voluntary external audit. Using bank regulatory data spanning 10 years from 2010 to 2020, we predict audit choice using four machine learning algorithms for classification; logistic regression, LASSO, random forest, and LightGBM. The models make use of 16 specially selected independent features. This thesis analyzes the machine learning algorithms based on various performance metrics (accuracy, specificity, precision, recall, and F1) and studies the feature importance measured by each model. To verify the results, the thesis uses two methods of feature selection; ANOVA and Mutual Information.

Our findings suggest that the proportion of agricultural loans to the total sum of loans is an important factor in predicting audit choice. Bank size and asset quality are also important factors in the banks' audit decisions. The best models are the tree-based models, with random forest being considered the best. Random forest predicts with a high level of accuracy and argues that the relationship between audit choice and the bank data is nonlinear.

Keywords – Banks, Audit, Classification, Machine Learning

Contents

List of Figures

List of Tables

1 Introduction

Economic and financial systems depend on financial institutions and commercial banks. Banks facilitate economic growth and development by allowing people and organizations to access the capital they need to start and grow businesses, buy homes, and invest in the stock market. The economy becomes more efficient through the allocation of funds from savers to borrowers, and banks play a vital role in today's society. At the end of the third quarter of 2022, commercial banks held over 23.6 trillion USD in accumulated assets in the U.S., which is 91.8% of the gross domestic product (Bureau of Economic Analysis, 2022; FDIC, 2022). Regulatory authorities monitor all banks, but unlike most other Western countries, the U.S. does not require an external audit of all banks (Nicoletti, 2018). This is in stark contrast to the European Union (EU) and the European Economic Area (EEA), where banks are considered Public Interest Entities and are subject to additional legal and regulatory requirements due to the nature of their business (Accountancy Europe, 2017). In order to ensure that their financial statements are accurate, reliable, and transparent, all banks in the EU and EEA are obliged to have statutory audits (Accountancy Europe, 2017).

The objective of the regulators supervising and regulating U.S. banks is similar to those of the auditors. The regulators' job is to ensure that banks and other financial institutions employ safe and sound business practices and comply with all applicable laws and regulations (Federal Reserve, 2021). Meanwhile, auditors are responsible for examining and verifying the accuracy and reliability of a company's financial statements (Tuovila, 2022). Banks differ from private firms in that they are reviewed by regulatory authorities regardless of whether they are audited (Federal Reserve, 2021). Private firms lack this type of control mechanism from an external party, and this interaction between regulators and auditors brings a new aspect to the choice of auditing for banks.

In most U.S. states, private banks are not obligated to have an external audit, but may choose to do so voluntarily. Research has shown that credit defaults are reduced by 38% in banks with voluntary external audits (Barton et al., 2015), yet 82% of the banks studied in this thesis have opted not to engage external audits for parts of or the whole 10-year

period. The question why so many banks opt not to have an external audit remains a conundrum to the field of accounting. To answer such a question, one must first identify which factors drive these banks not to employ an external audit when given the opportunity.

This thesis aims to fill a gap in the literature by identifying drivers of the banks' decisions. It has previously been researched as to why private firms choose to audit and on the value that an audit provides in general. However, very little research has been conducted on which factors influence banks in the decision of getting no voluntary external audits. Understanding why some banks decide not to procure a voluntary external audit contributes to better understand these banks and their behavior. Additionally, this research can help identify and decipher differences between private banks that operate in various sectors and activities, and facilitate further research into the value derived from an audit.

In this thesis, we investigate which factors influence the choice of not acquiring a voluntary audit for small[1] private commercial U.S. banks using bank regulatory data from 2010 to 2020. Vanstraelen & Schelleman (2017) found that there is much divergency in factors driving audit demand for private companies when studying what is currently known about the costs and benefits of auditing private companies (Vanstraelen & Schelleman, 2017). To the best of our knowledge, little research has been conducted on this regarding private banks. As such, we believe the study of voluntary audits profits from a more thorough analysis of which factors are the most important. For this analysis, we have narrowed down our focus to loans and risk factors as potential influences. It is our opinion that the most value can be found in identifying banks that do not acquire external audits, which will be explained in detail in the following paragraphs. Therefore, the thesis addresses the two-part research question:

How do loan and risk factors influence the decision to get no audit?

Banks' primary source of income is the interest spread from their loans, and it is clear that loans are one, if not *the*, main activity of a bank (Gadre et al., 2016). As such,

[1] banks with total assets below 500 million USD

loans may naturally influence most other aspects of a bank's decisions, including audit choices. We look at the types of loans banks issue, and specifically in what sectors, who the borrower is and how they are secured. These categories are defined as agricultural loans, commercial and industrial loans, real estate loans and loans to individuals. Lease financing receivables, which is net of unearned income on direct and leveraged financing lease and any allowance for losses on leases (Federal Reserve, n.d.), is included as well.

We assume that regulatory authorities are interested in high risk banks and their behavior. Bank failures may have enormous consequences, as brutally demonstrated in the financial crisis of 2007-2008, and government agencies and regulators strive to pinpoint and counteract risks in banking institutions. The relevance and importance of understanding banks is further highlighted by the 2022 Nobel Prize in Economics. The laureates' research has improved our understanding of banks' role in the economy and the findings show why avoiding bank collapses is vital (Nobel Prize Outreach AB 2022, 2022). As such, we assume there is a distinct interest in whether there is a connection between risk indicators and the decision to get no audit. This assumption led us to structure the research in a way that emphasizes the negative decision, i.e. to get no audit. The thesis examines risk factors, specifically cash, loan loss allowance, loan loss provision, return on assets, and ratios for Tier 1 and Tier 2 capital.

In order to study private U.S. banks' decisions to get an external audit or not, we have applied machine learning and predictive modeling techniques. Using machine learning, one is able to extract relevant information from a large quantity of available raw data, due to the fact that a machine learning algorithm searches for patterns and underlying structures it finds meaningful to its decision making (Alpaydin, 2020). Accordingly, this method allows us to investigate the behavior of banks, without implementing our own biases on the data. In this thesis we predict banks' choices of audit using a subset of the reported information in financial statements. We apply four machine learning models to the problem; logistic regression, LASSO, random forest and LightGBM. We use loan and risk variables as features in the models, in addition to variables derived from other necessary financial statement data.

The results of this thesis are twofold. First, the trained random forest and LightGBM models predict audits with considerable accuracy, random forest being the best. These predictive models offer a new approach to studying voluntary audits as it allows further research to predict decisions of banks' prior to the choice being taken. Second, it concludes that there are variables in both the loan and the risk category that affect the choice to get no audit. Intriguingly, the proportion of agricultural loans to total loans emerges as the most important variable. In terms of risk, loan loss allowance and loan loss provision, as proxies for asset quality, are large influences on the decision to get no audit. Lastly, in accordance with prior literature (Chan & Kogan, 2011), we find that bank size is important.

The thesis is divided into seven chapters. First, we review relevant literature about auditing private firms and banks in Chapter 2. This will provide an understanding of what we already know about auditing for private companies and banks and what research shows drives audit demand. The thesis will use prior literature as a foundation in further investigating what might drive audit demand amongst private U.S. banks.

In Chapter 3 we present the data we have collected, the data cleaning process and some discussion on the selected features. The following chapter, Chapter 4, will explain the multivariate analysis. Here, we introduce and describe the applied machine learning models, and show how these models are performing in terms of predictive power. Next, in Chapter 5, we explore the most important factors for the different models, and verify the results using two approaches of feature selection; ANOVA and Mutual Information. Further, in Chapter 6, we compare and interpret our results, and discuss the robustness of the models and their predictions. Additionally, we provide suggestions for further research. Ultimately, we summarize and conclude based the results of the analysis in Chapter 7.

2 Literature Review

Across the world, there are different financial disclosure requirements for private companies. In the major capital markets worldwide, mandatory audits are required of publicly traded companies, and by extension, of other public-interest entities (Vanstraelen & Schelleman, 2017). However, the requirements for audits of private companies differ considerably. Dedman et al. (2013) stated that the reasons for requiring a mandatory audit for private companies are less clear than for public companies because private companies exist in a different environment (Dedman et al., 2013). In private companies, the stakeholders tend to be closer to the company, which makes it easier to request information directly. Generally, private companies in the U.S. are neither required to disclose their financial results nor have their financial statements audited (Minnis & Shroff, 2017).

Auditing requirements for private companies vary across countries, and regulations change over time. For example, Sweden and Norway dropped the requirement that all private companies, including small limited liability companies, were subject to a statutory audit in 2011 (Vanstraelen & Schelleman, 2017). Between 1967 and 1994, all UK companies were obligated to undergo an external audit (Dedman et al., 2013). However, over time, the UK has significantly increased its exemption criteria for private companies, meaning fewer companies are required to have an external audit (Dedman et al., 2013). In comparison to research on public companies, little economic research has been conducted on privately held companies' external audits. However, changes in regulations and institutional settings, where certain private companies are not subject to a mandatory external audit, have allowed researchers to investigate the drivers of voluntary audit demand in a new way (Vanstraelen & Schelleman, 2017).

Using survey data on a sample of privately held U.S. companies, Abdel-Khalik (1993) concluded that private firms choose audits to compensate for the loss of control when there is a long chain of command, and observability of subordinates' actions decreases (Abdel-Khalik, 1993). Furthermore, Dedman et al., (2013) found that private companies

are more likely to voluntarily engage an external audit if they have greater agency costs (Dedman et al., 2013). These findings suggest that private companies value independent information verification when they have an asymmetric relationship between employees, management, and shareholders. According to research, private firms tend to acquire voluntary audits when they are larger, are growing, are more complex, have more contractual relationships with external parties such as creditors and suppliers, and when they require advice or consulting services that are provided by their external auditor (Vanstraelen & Schelleman, 2017). Audits are costly, and the cost-benefit analysis for private company audits are firm specific. There are variations in the different needs from users of private companies' financial statements and as a result of this, a one-size-fits-all audit may not be optimal for all types of companies (Dedman et al., 2013).

When it comes to the choice of auditing in banks, Chan & Kogan, (2011) used machine learning to predict the choice of getting a voluntary external audit. They researched FDIC-insured banks to evaluate if banks' choice of voluntary audits is systematic. In finding evidence for the choice being systematic, they also found that several of the factors indicating that a private firm is audited, are similar to the ones of banks. Their research suggests that banks' size, profitability, growth, leverage, complexity of operations and ownership structure significantly impact the decision to procure an audit. Furthermore, banks with lower levels of capitalization had lower likelihood of procuring an audit, while larger, more profitable, and growing banks had higher likelihood. (Chan & Kogan, 2011)

In terms of risk, Dedman et al. (2013) found that private UK companies choose to get an external voluntary audit when they are riskier (Dedman et al., 2013). Banks' lending activity is a substantial part of the their risk elements as their primary income is the interest rate they receive from loans (Gadre et al., 2016). This makes auditors' effect on loan loss provision an interesting study. Loan loss provision and loan loss allowance are created and maintained by a bank against potential losses on loans and certain risks (Nicoletti, 2018). A loan loss allowance reduces the reported value of issued loans as a contra-asset, while a loan loss provision increases the allowance and lowers reported earnings (Nicoletti, 2018). Both variables are used in research as proxies for asset quality

(Chan & Kogan, 2011). Nicoletti (2018) presents evidence that regulators and auditors have different influences on loan loss provision, with auditors being the dominant influence (Nicoletti, 2018). When banks are subject to greater regulatory scrutiny, audited banks are less timely, relative to unaudited banks (Nicoletti, 2018). In other words, when there is an increase in nonperforming loans, banks delay the recognition of expected losses when getting an external audit.

In the aftermath of the financial crisis of 2007-2008, regulators have increasingly focused on banks' capital ratios. As a result of the crisis, an internationally agreed upon set of standards for adequacy, stress testing, and liquidity are set as requirements for all banks (BIS, n.d.). Tier 1 and Tier 2 capital are banks' core capital and supplementary capital respectively, and banks' total capital can be calculated by adding these together (Nickolas, 2019). Ratios of these are seen as the most important measures for banks' financial health, and regulatory authorities monitor them as low values can be risky (Nickolas, 2019). As it is more difficult to liquidate Tier 2 capital, it may be considered less reliable than Tier 1 capital. Illiquid balance sheet values such as intangible assets can thus be a risk to banks if they make up a great percentage of their assets.

Having a voluntary external audit is shown to be valuable to banks. Barton et al. (2015) found that having a voluntary audit of banks' financial statements is associated with a 38% decrease in the likelihood of credit defaults (Barton et al., 2015). Interestingly, they found no differential effect between mandatory and voluntary audits (Barton et al., 2015). This suggests that the correlation between audit and reduced likelihood of failure is due to having an audit, rather than choosing to have an audit. As such, one can argue that research indicates that auditing of banks leads to real economic benefits.

3 Data

Regulatory authorities who oversee U.S. banks include the Federal Reserve (the Fed), Federal Deposit Insurance Corporation (FDIC), and Office of the Comptroller of the Currency (OCC) (Federal Reserve, 2021). The primary supervisor of a bank is decided by which type of institution the bank is, and who licensed the bank to operate (Federal Reserve, 2021). All commercial banks in the U.S. are obligated by regulators to fill out Reports of Condition and Income (Call Report) each quarter regardless of their size, independent audit status or trading status (WRDS, n.d.b). The Call Report follows Generally Accepted Accounting Principles (GAAP) and contains a balance sheet, income statement, and other supporting documents (WRDS, n.d.b). The Call Report is part of what we refer to as bank regulatory data. For this analysis, we focus mainly on the subsection of the Call Report named Report of Condition (Schedule RC-C) which provides detailed information on assets, liabilities, and capital accounts. The banks are required to fill out all information for domestic and consolidated bank operations (FFIEC, n.d.a). Submission of the Call Report can be done either digitally or by paper (FFIEC, n.d.a). The quality of the data is further discussed in Chapter 6.

To complete our analysis, we have used Rstudio which is an integrated development environment for R, supplemented by Python in Visual Studio and Power Query in Excel.

3.1 Presentation of Data

The data consists mainly of bank regulatory data for commercial banks collected from Wharton Research Data Services (WRDS). This data is retrieved from Call Reports filed by commercial U.S. banks, and we use the Schedule RC-C which includes balance sheet data. Data from the Center for Research in Security Prices, LLC, (CRSP), and the Federal Financial Institutions Examination Council (FFIEC) is used as a supplement. CRSP data is used for identifying publicly listed banks, while FFIEC data is used for filling out information on loans where the variables were not available throughout the 10-year period. As a foundation, the analysis has used variables from the quarterly Call Reports collected by the Central Data Repository (CDR). Specifically, we have investigated the variables

from the balance sheet, loans and lease financial receivables, and income statement. All banks and bank holding companies are identified by their RSSD ID, which is a unique number that the Federal Reserve assigns to each bank (FFIEC, n.d.c). The data spans ten years, beginning at the first quarter of 2010 and ending in the first quarter of 2020.

3.2 Data Cleaning

The raw data contains 271 553 observations covering 8 319 commercial banks. Domestic values are reported with the prefix RCON, and the consolidated bank values with RCFD (Federal Reserve, 2017). As a general rule for the data, where RCFD observations are missing, we have used the respective RCON data as a substitution. Furthermore, in the case of reporting variables expiring or having been edited throughout the period, we have filled them by calculating their value using other reporting variables, see Appendix A1 for detailed information. Duplicated data is excluded, removing 51 observations.

We have included data from 2020 due to the fact that information on whether a bank has received an external audit is reported in the March Call Report the following year (WRDS, n.d.a). After lagging the audit information so that it is present for the year in question, data from 2020 is removed. This process removes 5 382 observations.

The original data set includes some listed bank holding companies. These companies are identified using monthly stock security files from the CRSP database. In CRSP all companies use a PERMCO identifier instead of the RSSD ID. We link these using the current table of RSSD ID to PERMCO ID links from the Federal Reserve Bank of New York and exclude all bank holding companies that have been listed at some point during our time period. This removes 48 bank holding companies.

All observations where banks have total assets of over 500 million USD are removed as FDIC requires an external audit of all banks insured by them with total assets over this threshold (FDIC, n.d.a). Some discussion on these larger banks is presented in Chapter 6.

Moreover, we remove data from U.S. territories outside the 50 states and Washington D.C. These are banks that operate in American Samoa, Federated States of Micronesia (including PN for Pohnpei), Guam, Puerto Rico, and the U.S. Virgin Islands. Additionally, there were some spelling errors in reporting of state abbreviation, which were edited or removed as necessary[2].

To identify states that demand an external audit, we have computed the rate of audited banks per state. Appendix A2 presents the results. All states that surpass a ratio of 95% are assumed to have mandatory external audits. When employing a strict percentage limit, states with few operating banks may be incorrectly classified, as the ratio might be considerably affected by just one or two banks. Therefore, for states with less than 25 banks, we have investigated more comprehensively due to the small sample size[3]. This removes a total of 884 banks from the data.

All observations where total assets (RCFD2170) and total loans (RCFD2122) are missing, zero or negative are removed from the data set because this indicates that the bank is not currently operating. This removes 588 and 3 259 observations respectively.

Similarly, observations with missing values for loan loss allowance (RCFD3123) and lease financial receivables (RCFD2127/28) are removed from the set. For our objectives, we are not able to use observations without this information. This filters out 13 227 observations.

Ultimately, the data has been winsorized at the 1st and 99th percentile to reduce the effect of outliers. The final set contains 179 455 observations on 5 934 distinct banks from the first quarter of 2010 to the last quarter of 2019.

[2]IK is IL for Illinois, KA is KS for Kansas, 19 is Florida, and further removed banks reporting 0 as their state abbreviation, belonging to the Federated States of Micronesia.

[3]Alaska; Alaska Statutes Title 44. (2021), Hawaii; Hawaii Code of Financial Institutions Chapter 412 (n.d.), Idaho; Idaho Credit Union Act Chapter 21 (n.d.)

3.3 Discussion of Variables

3.3.1 Target Variable

To identify which banks are audited, the audit indicator RCON6724 is used to make a dummy variable that takes the value 1 if the bank has received an external audit that year, and 0 if not. The aim of the machine learning modeling in Chapter 4 is to predict this target variable. Banks report their audit indicator on a scale from 1 to 9, where each number indicates the most comprehensive level of auditing work received, with 1 being the most extensive (WRDS, n.d.a). If a bank reports either 1 or 2, the bank is classified as *audited* following prior literature (Chan & Kogan, 2011; Nicoletti, 2018). Three banks have reported a value of 0, and all associated observations are classified as *not audited*.

3.3.2 Predictors

Grouping	Name	Associated variable
Loans	Agricultural loans	AG_loans
	Commercial and Industrial loans	CI_loans
	Loans to Individuals	IN_loans
	Real Estate Loans	RE_loans
	Lease Financial Receivables	LFR
	Total sum of loans	total_loans
Risk	Liquidity	cash
	Loan Loss Allowance	LLA
	Loan Loss Provision	LLP
	Return on Assets	ROA
	Tier 1 capital	tier1_ratio
	Tier 2 capital	tier2_ratio
Other	Federal funds sold	fed_funds_sold
	Intangible Assets	intangible_assets
	Premises and Fixed Assets	prem_fixed_assets
	Size	size

Table 3.1: Grouping of Variables

As specified earlier, we focus on loan and risk related variables, and include some other necessary financial statement data, as illustrated in Table 3.1. These variables combined are the input used in order for the machine learning models to make a prediction. The next section explains this categorization and the reasonings behind selecting these features.

The complete and detailed variable definitions are presented in Appendix A3.

Loans

All loan variables are defined using the subsections of loans defined in the main asset side of the balance sheet of the Call Report, except for *total_loans* which is all loans accumulated. These categories are determined based on the loans' security, borrower, and purpose (FFIEC, n.d.b). Due to the vast difference in bank sizes in the data, it is more appropriate to include the ratios of these categories rather than an absolute value. We have calculated these ratios for each category scaled to the sum of all loans. The *total_loans* variable, on the other hand, is scaled by the sum of all assets to compute the ratio of loans to assets.

While *AG_loans, CI_loans, IN_loans*, and *RE_loans* are relatively straightforward to comprehend, *LFR* requires some additional explanation. *LFR* represents the balance of outstanding receivables which are related to direction and leveraged financing leasing on property that is acquired by the bank for leasing purposes (Federal Reserve, 2022). It can be argued that the variable therefore should be included in the risk category. Though, for all variables, it is important to consider that risk is a part of all activities and operations in a bank. This will be further discussed in the next part.

Risk

The risk related variables aim to identify signs of risk in the banks' financial statements. Risk variables are less intuitive than loan variables, as it is harder to evaluate to a specific value. As mentioned, risk is a broad concept, and most reported information may signal high or low risk in some way or another. A bank's overall risk is not measured solely from one or a few variables, but must be seen in relation to how the overall state of the bank appears. This is further a main reason as to why the loan and risk variables are supplemented by more financial statement data. Therefore, the grouping of the variables into a specific risk category is merely a way of organizing the most prolific high risk indicators. As such, we can identify if these specific variables are important to the decision not to audit. The following paragraphs explains the risk variables in further detail.

The *cash* feature is a proxy for liquidity as it indicates how much available reserve the bank has in proportion to the total assets. Low liquidity is undoubtedly a risk to banks as they may struggle to pay their commitments. Furthermore, extremely high portions of liquidity may come as a response to high levels of uncertainty (Breitenlechner et al., 2022).

We consider asset quality by the loan loss allowance (LLA) and loan loss provision (LLP). LLA is the reserve for losses on loans and leases, while LLP represents an income statement expense of what banks estimate to lose (Alpert, 2021). LLP shall be charged or credited to bring LLA to an appropriate level (OCC, n.d.), and if accurate, a large LLP should indicate larger losses.

Banks' overall financial health is addressed by Tier 1 and Tier 2 capital ratios, computed by Tier 1 and Tier 2 capital scaled by total risk-weighted assets. As discussed in the literature review, these ratios are a strong indicator for the condition of banks' assets. Accordingly, low levels are a sign of a bank being high risk.

The return on assets (ROA) represents companies' profitability in relation to their assets (Hargrave, 2022). It is computed by the total operating income divided by total assets (Hargrave, 2022). As banks' assets are mainly their loans, a low ROA may indicate low earnings on loans or that their interest spread is too low. A low ROA may be connected with low-quality loans (Britannica, n.d.) and asset inefficiency (Hargrave, 2022).

Other

Other variables from financial statements have been scaled by total assets, except for *size*. The variable *size* is represented by the natural logarithm of total assets, and has further been normalized to fit a scale from 0 to 1. We use this normalization as a means of not skewing the predictions. As all other values are ratios, i.e. between 0 and 1, the *size* variable has larger values before normalization The same process is repeated for LLP and ROA as both variables have negative values. This means that all variables are on a scale from 0 to 1 before modeling. For the two generalized linear models, this increases

interpretability when studying the β coefficients. This standardization does not have any significant effect on the tree models, but transforming the variables will not affect output, and it makes training faster (Charbuty & Abdulazeez, 2021).

We have also considered two other aspects of banks' assets, federal funds sold and premises and fixed assets. Federal funds sold are the excess reserves banks have after trading on their regional federal reserve (Chen, 2021). It represents how much more cash the commercial bank has than the standard reserve requirement, which is set by central banking authorities (Chen, 2021). As such, *fed_funds_sold* represents the ratio of excess reserves compared to total assets. A high level of *fed_funds_sold* indicates that banks have an additional buffer available, which is associated with lower risks. Premises and fixed assets, *prem_fixed_assets*, are often referred to as premises and equipment, and represent all the property and locale where the bank does its business as well as other fixed assets (FDIC, n.d.b). Among other fixed assets, vaults, fixed machinery, capitalized leases and real estate acquired for future expansion of business locales may be included (OCC, 2018). While larger amounts of premises and other fixed assets may indicate a larger, more reliable institution, this is no guarantee. In order to provide a broader picture of the banks, both features are included in the analysis.

Intangible assets include all assets that are not tangible, like human capital, technology and organization, and are therefore inherently difficult to measure (Kenton, 2021). Accordingly, intangible assets are highly illiquid. The ratio of intangible assets to total assets lets us know the portion of banks' assets which are tangible assets or material items they can sell for monetary value (Kenton, 2021). As such, if a large portion of total assets are intangible and illiquid, it can signal low liquidity or bad financial health.

4 Methodology

Before we dive deeper into the methodology of this thesis, it is beneficial to provide some background on machine learning. Machine learning is a scientific field or subject that studies how a machine can learn and improve with experience (Mitchell, 2006). The study of machine learning began 50 years ago and has grown into an essential part of everyday life (Mitchell, 2006). Machine learning as a field encompasses a wide range of models, algorithms, and applications, and is advancing at monumental speed (Mitchell, 2006). We will use one of these applications, specifically classification, in our analysis.

In classification, the problem has an outcome that is divided into two or more categories, also referred to as classes (Alpaydin, 2020). For this analysis, there are two classes; *audited* and *not audited*. This is referred to as a binary classification problem (Alpaydin, 2020). The machine learning model, called classifier, aims to assign the combined input to one of these classes (James et al., 2013). In this case, the input can be referred to as $x_1, x_2, \ldots .x_n$ where n is the total number of observations. Each observation, denoted x, contains p lines of data which is given by a set of features $X_1, X_2, \ldots X_p$ (James et al., 2013), which is also known as the model variables. Using this input, the classifier will predict whether this bank is *audited* or *not audited*. To perform effectively, i.e. to make accurate predictions, the model must first be trained (Soni, 2018). There are different methods within this training, typically divided into supervised and unsupervised learning (Soni, 2018). The first method trains the model using historical data with a correct class already assigned, and the latter only uses input data and searches for regularities and patterns (Soni, 2018). In this analysis, we conduct supervised training.

An important step in machine learning is feature selection (Cai et al., 2018). The aim of feature selection is to create a subset of features from the original features according to a particular evaluation criterion (Cai et al., 2018). Good feature selection can improve learning accuracy, reduce learning time and simplify learning results (Cai et al., 2018). We view feature selection as a valuable tool in this analysis as we search for the most important variables that explain the banks' choice to get no audit. However, in this

thesis, we are selecting potential variables based on our research question. Due to the thesis' focus on loan and risk factors, we have already identified features within these categories. Thus, we use feature selection as a means of validating the importance of features in the machine learning models rather than a preliminary step.

The most essential part of machine learning is testing. It is in this phase that the model can learn and may further improve (James et al., 2013). The iterative testing and training process is what makes machine learning such a powerful tool in predictive modeling (Brownlee, 2018). Testing involves verifying that the model functions and performs well, and therefore it must be applied to data that were not available during training (James et al., 2013). This testing will demonstrate whether the model overfits on the training data—that is, whether it produces perfect training data predictions using an overly complex model (James et al., 2013). This process of testing on unseen data is called cross-validation or model validation (James et al., 2013). The first step in cross-validation is to divide the data, which raises the question of how the data should be split. Many alternatives exist for dividing or distributing the data, but a simple and often applied method is to randomly divide it into k folds, and create training and tests sets (James et al., 2013). The easiest version is therefore to use $k = 2$ folds, i.e. divide the data into one training set and one test set based upon a selected percentage. Using this latter option saves a significant amount of computational time, which is primarily why it is used in this analysis.

Researchers diverge on an optimal training-test size ratio, but it is reasonable to use somewhere between 10-30% of the data for testing (Joseph, 2022; Gholamy et al., 2018). Recent research investigates sample size determination, but it is still a very new subject (Melvin, 2021). The split depends on the original sample size as well as the number of dimensions in the data. Even at 10%, a large sample size will yield a substantial test set, while a smaller sample size will require a larger percentage to obtain a sufficient test set. We have used 15% of the original data for the test set, and 85% for the training set. Due to the substantial sample size and the high dimensionality of the data — that is, the number of variables — we selected this slightly smaller test set size.

4.1 Exploratory Data Analysis

Investigating the distribution of the numerical variables and how they differ from one another, and from *audited* to *not audited* banks, helps us understand the data. A good comprehension of the data will provide us with a better basis for recognizing important features. Additionally, it is useful to be aware of the distributions, the existence and the location of outliers, and other anomalies. The Exploratory Data Analysis (EDA) enables us to determine whether the modeling results are likely to be accurate or whether we should take special precautions (Hartwig & Dearing, 2002). This section will compare descriptive statistics of all features and identify their correlation. Initially, we will examine differences within the features, while in the next part, we will examine correlations between the features.

4.1.1 Descriptive Statistics

The purpose of this part is to examine feature-wise differences in the data. In the data set, there are more observations with *audited* = 1 by roughly 40 000 observations. In total, there are 179 455 observations in the data. As such, we are examining differences in the distributions' shapes and means rather than the specific frequencies. This comparison is made between distributions of observations that have been audited and those that have not been audited. Descriptive statistics for all features are displayed in Appendix A4, along with histograms displaying each feature for each class in Appendix A5. All features, except *size*, are ratios without a unit, i.e., dimensionless, due to the scaling of the features. The *size* feature is measured in ln(USD).

Most features in the data set are skewed towards the left, and it is evident in most diagrams that there are many values being equal to zero. The majority of the banks are smaller banks. This is corroborated by histograms for *size*, which show that most banks are to the left of 0.5. Furthermore, the means for most features lie close to zero, with notable exceptions being *total_loans*, *size* and *RE_loans*. It is logical to assume that since many banks are smaller, they do not offer loans in every category, which helps explain these low values. Many values being equal to zero is an important issue when working with this data, and it may be especially worrying in *fed_funds_sold*, *intangible_assets*,

and *LFR* where there are few unique values in the distributions. This is further explored and problematized in Chapter 6.

4.1.2 Correlation

Correlation is a statistical measure of how variables covariate. The degree of correlation is given by the correlation coefficient denoted r. There are several different correlation coefficients, and the most commonly used is Pearson[4] correlation, which we use in this analysis. The correlation values are expressed as a value between -1 and 1, where -1 indicates a strong negative relationship and 1 a strong positive. A result of zero signals that there is no linear relationship between the variables. (Fernando, 2021)

A high correlation between variables can cause problems in a prediction model. All explanatory variables shall be independent of each other. As such, the model can estimate the relationship between each independent variable and the dependent variable separately. A too-high degree of correlation between the variables can make it challenging to fit the model and interpret the results. This is especially an issue with models based on a linear relationship between the dependent and independent variables. When there is a high level of correlation between the variables, the coefficients can swing wildly based on other independent variables in the model, and the coefficient may become sensitive to small changes. This phenomenon is called multicollinearity. Furthermore, multicollinearity can reduce the estimated coefficients' precision, which weakens the statistical power of a regression model. The result of multicollinearity is that one might not be able to trust the test's p-values to identify statistically significant independent variables. (Frost, 2017)

Several components of banks' financial statements are intuitively highly correlated. The total assets of the bank and the absolute value of loans issued are in all likelihood related. This is one of the reasons why all variables are normalized to a larger entity. Additionally, selecting the most relevant variables for the study was heavily influenced by the desire to reduce correlations between independent variables. For financial attributes which by nature represent the same aspect of the bank, we selected the most relevant based on

[4]Pearson's $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2.(\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

our research question. In the final selection, no variables have more than 74% absolute correlation. The correlations between the selected variables are displayed in Figure 4.1. The highest correlation is between *AG_loans* and *RE_loans*, at a negative 73.4%. Except for this, only three correlations are above 50%. Although it is generally accepted that correlations of 80% or above indicate a major issue (Senaviratna & Cooray, 2019), this does not directly exclude multicollinearity in our analysis.

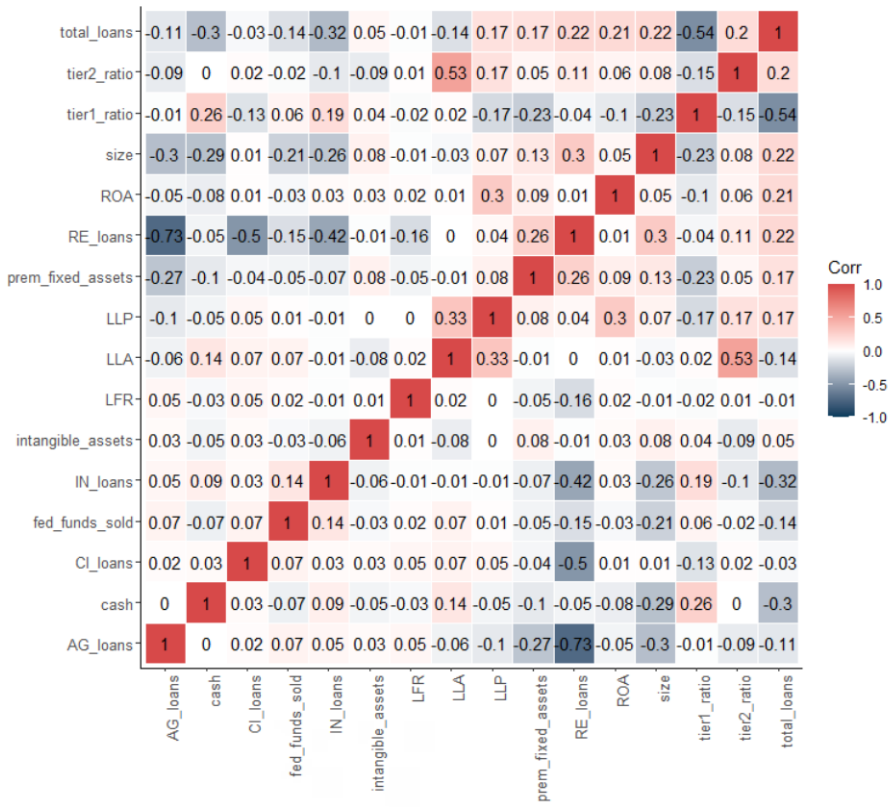

Figure 4.1: Correlation Matrix

4.2 Modeling

This section will outline the machine learning models we utilize to conduct our analysis in order to examine the factors influencing banks to not engage an external auditor. Prior to this, we give some background regarding hyperparameters and the evaluation criteria for prediction. We use two generalized linear models; logistic regression, and LASSO, as well as two tree-based models; random forest and LightGBM. These models offer different perspectives in terms of flexibility and interpretability, as illustrated in Figure 4.2. As

expressed in the introduction, the aim of the analysis is to identify the important features for the decision to get no audit, as well as produce a good prediction model. As such, we examine models from both extremes.

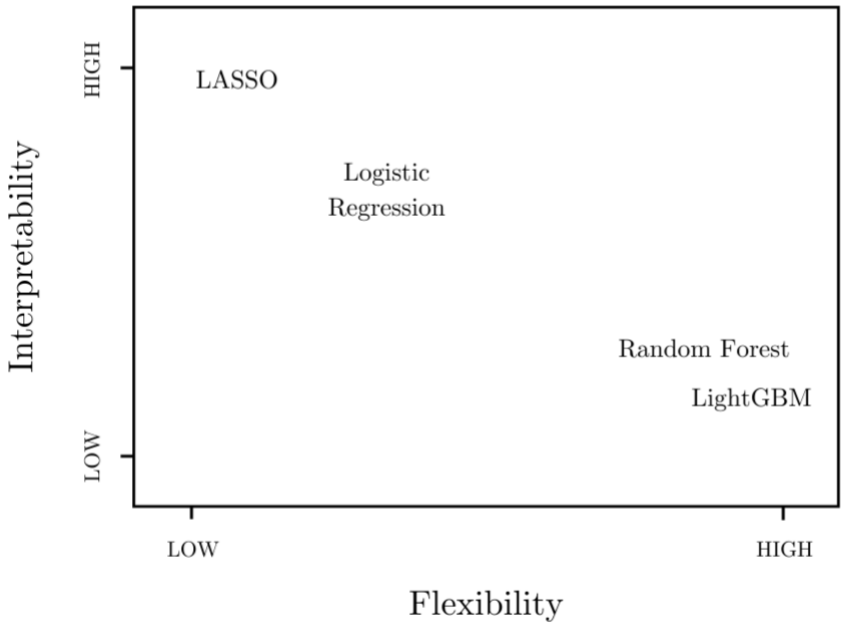

Figure 4.2: Flexibility and Interpretability for Models

(James et al., 2013)

In the following subsections, we introduce and briefly describe the classifiers and their application to our data. All models are trained and tested on the training and test sets previously defined. This continuity allows us to compare the accuracy of the models, which in turn helps validate the conclusions on feature importance and influence. Validation of the models' results are discussed in Chapter 5.

There are 16 features to consider as independent features for this analysis, see Appendix A3. The classifiers use these features to predict one of two outcomes; *audited* or *not audited*. The outcome is represented by the dependent variable *audited*;

$$audited_i = \begin{cases} 1, \textit{if } x_i \textit{ is audited} \\ 0, \textit{otherwise} \end{cases}$$

4.2.1 Hyperparameters and Tuning

Hyperparameters, or tuning parameters, regulate the complexity of different machine learning models (Gu et al., 2020). Tuning parameters for this analysis include, for example, the penalization parameter in LASSO and the number of trees in the random forest model. These parameters are usually tuned by iteratively improving the models on a validation set (Gu et al., 2020). As one must reestimate the model on the training set for each iteration, this process is very computationally heavy. This step is simplified for efficiency reasons and we look to former research when selecting hyperparameters.

4.2.2 Evaluation Criteria on Predictive Performance

In order to compare the different models, one must first consider what makes it an effective model in the applied context. We use the evaluation metrics; accuracy, precision, recall, specificity, and F1, in this analysis to evaluate the models' predictions. When computed using out-of-sample data, i.e. the test set, the aforementioned metrics will collectively provide a credible picture of the models' predictive powers. Figure 4.3 displays the formulas for the evaluation criteria and their components visualized in a confusion matrix.

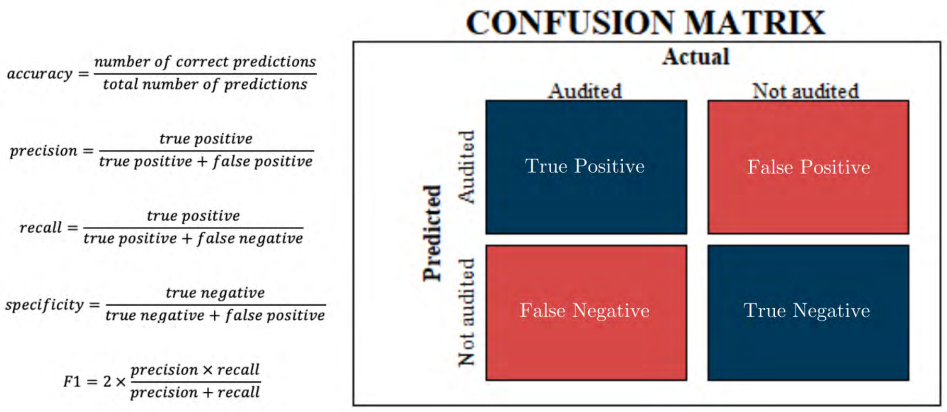

Figure 4.3: Evaluation Criteria and Confusion Matrix

(Luque et al., 2019)

The most commonly used evaluation criterion when comparing classification models is accuracy (Bratko, 1997), but this metric in isolation is not always sufficient to

evaluate a model. Looking at the test set where roughly 70% of the observations are *audited*, one should be careful using only one metric for measuring predictive performance. Accuracy measures a model's overall predictive power, i.e. all correct predictions compared to all predictions, and a model which predicts *audited* for all observations will accordingly have an accuracy of 70%. However, even if it is accurate to 70%, the model should not be considered reliable because all *not audited* observations are mispredicted. Thus, it is essential to consider multiple assessment metrics.

One important thing to consider is whether the highest cost for predictions is false positives or false negatives. Where there is a high cost associated with false positives, one should focus more on precision, whereas if false negatives have higher costs, one should focus on recall. Recall and precision generally contend with one another, as when one is increased, the other usually declines. To seek a balance between precision and recall, one can use the F1 score of the model, which is a function of recall and precision. The F1 score is especially useful when working with an uneven class distribution, which is present in our test set. While the accuracy can be disproportionately affected by many accurate predictions because of little weight added to false predictions, F1 score takes false predictions into account, providing a more balanced view. (Jordan, 2017)

In the following analysis, we have selected evaluation criteria based on our assumptions and research question. The aim of this thesis is to predict what drives banks' decision to get no audit. Thus, false negative predictions, i.e. the model predicting banks which are *not audited* as being *audited*, are the biggest issue. For this reason, we believe that false negatives have the highest associated cost. Therefore, we favor recall above precision, but balance this by including F1 scores. However, only using these criteria, it is possible that a model which only predicts *audited* could achieve great scores. Since there in such a case are no negatives, there would be no cost associated with false negatives. We therefore include the specificity metric into the evaluation criteria to prevent this from happening. Specificity will show how many of the unaudited banks were actually predicted accurately, which reveals if the model is favoring predicting *audited*. All five metrics displayed in Figure 4.1 are used in reviewing the predictions, but with a slight emphasis on recall.

4.2.3 Generalized Linear Models

The most common approach to regression analysis is the linear regression model, which uses the principle of ordinary least squares (OLS). However, standard linear regression is not fit for predicting a qualitative response, such as in classification problems (James et al., 2013). Generalized linear models is an umbrella term for models that linearly relate the response variable to the independent variables via a link function such as the logarithmic function (Jørgensen, 2012). The generalized linear models fit the model to the data using maximum log likelihood, which is iteratively fitting using reweighted ordinary least squares (Jørgensen, 2012).

Both generalized linear models used in this thesis predict the probability of engaging a voluntary external audit. The analyst can therefore determine a threshold value, which the model will predict according to. For this analysis, we have selected a threshold of 50%. In other words, if the probability surpasses 50%, the model predicts *audited*.

Logistic Regression

First, we use the logistic regression model which is an expansion of the linear regression model. The logistic regression model uses the natural logarithmic function to find a linear relationship (James et al., 2013). It follows from this that the response variable, called the logit[5], is linear in X. The aim of logistic regression is to estimate the β coefficients so that we maximize log likelihood (James et al., 2013). We predict the probability of X, defined:

$$p(X) = \frac{e^{\beta_0+\beta_1 X+...+\beta_p X}}{1 + e^{\beta_0+\beta_1 X...+\beta_p X}}$$

where β is the estimated coefficients (James et al., 2013).

We fit this logistic regression model on the training data, and use a threshold of 50%. Applying this model on the test data, we get the confusion matrix presented in Figure 4.4.

[5]logit: $log(\frac{p(x)}{1-p(X)}) = \beta_0 + \beta_1 X + ... + \beta_X$ (James et al., 2013)

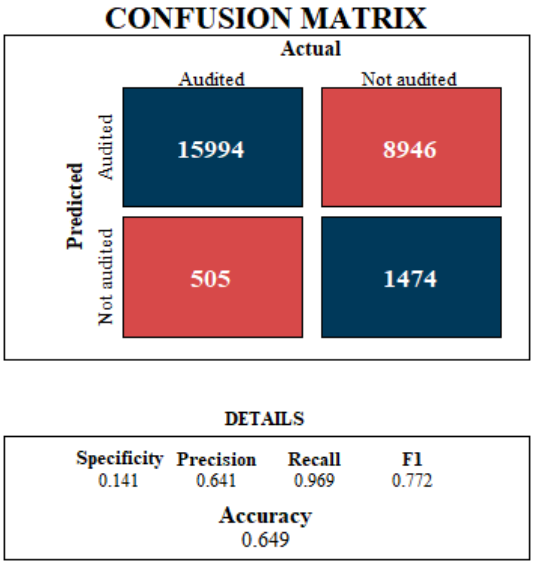

Figure 4.4: Confusion Matrix for Logistic Regression

The confusion matrix presents a high recall of 96.9%, signifying that the model predicts more than 96 out of 100 actually audited banks correctly. Accordingly, very few actually audited banks are labeled as *not audited*. Due to our focus on minimizing false negatives, one could therefore argue that this is a good model. However, other metrics are substantially worse. The F1 score, which we use to balance recall, is quite a bit lower, due to the low precision, and the specificity of the model is extremely low. This score signifies the issue presented in Section 4.2.2, where the model favors *audited*. As a result, few actually *not audited* observations are predicted correctly, and we should be cautious in using this model for our purposes.

LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is one of the most well known regression models, and is a type of shrinkage method (Tibshirani & Wasserman, 2017). The LASSO method is used for high-dimensional regression, and it builds on the concept of linear regression by adding a penalty parameter, λ, which will shrink all coefficients towards zero (James et al., 2013). LASSO problems are convex, which makes them efficient to solve because we can ensure that the identified solution is the global extrema, i.e., the most optimal solution (Tibshirani & Wasserman, 2017).

There are two models often considered when discussing shrinkage methods, LASSO and ridge regression (James et al., 2013). In contrast to its frequent competitor, LASSO has sparse solutions (Tibshirani & Wasserman, 2017). In other words, the solution using ridge regression will have all β coefficients unequal to zero, while LASSO sets p of β^j equal to zero, where j is $\in 1, ..., p$. In other words, it sets a subset of the coefficients equal to zero. As a result, LASSO is a popular feature selection method because it reduces the coefficients of less important features to zero (Tibshirani & Wasserman, 2017). As the aim of this thesis is to find the most important features, we use LASSO instead of ridge regression.

Figure 4.5 displays the binomial deviance for each log λ using 10-fold cross-validation. The upper horizontal axis shows the number of features with coefficients unequal to zero, and the leftmost dotted line represents the λ value that produces the lowest binomial deviance (λ_{min}). Using this λ_{min}, at 0.0003, no coefficients are zero, meaning that it includes all variables in the model. The dotted line to the right displays the last point at which the binomial deviance is within one standard deviation of the minimum error. This line represents a $\lambda^* = 0.0023$, which we will refer to as the optimal λ^*.

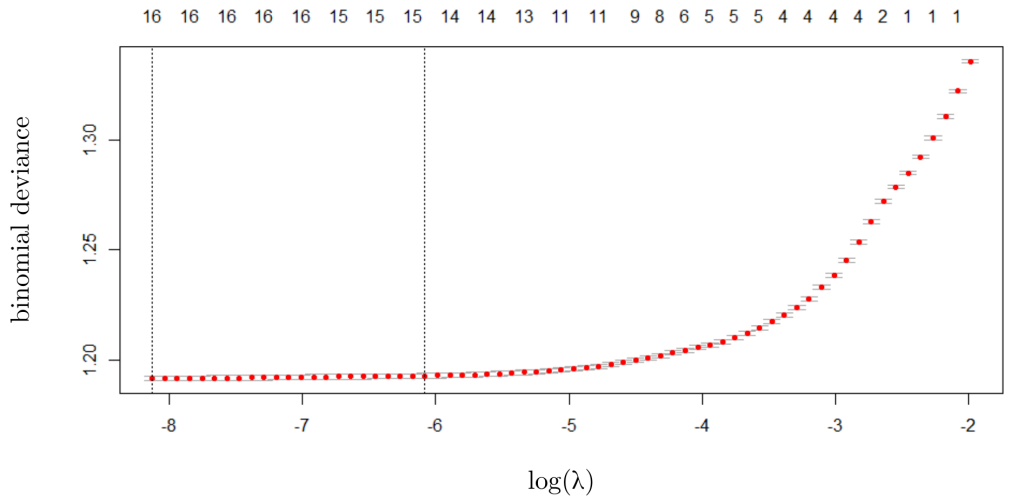

Figure 4.5: Plot for Binomial Deviance over log λ

We select the penalty parameter with the lowest binomial deviance while keeping the model as simple as possible. Therefore, we use the optimal λ^* as our penalty parameter in the prediction model. As such, we have a less complex model, in terms of the number of features, than we would have using λ_{min}. Using λ^* results in 14 nonzero coefficients.

The confusion matrix for this model is presented in Figure 4.6 below.

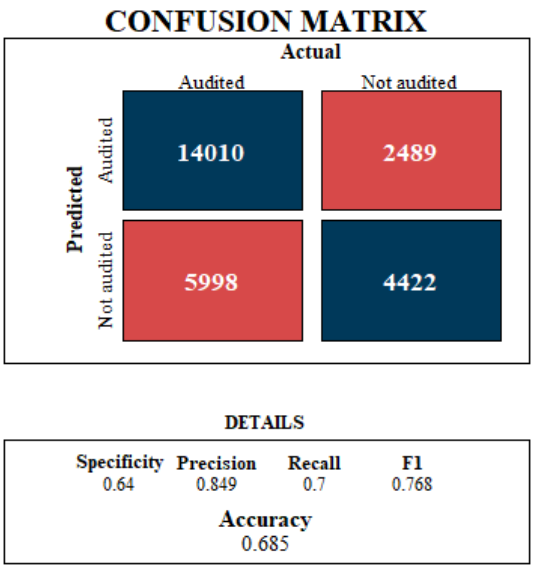

Figure 4.6: Confusion Matrix for LASSO

LASSO scores the highest for precision, meaning that there are few false positives. However, recall and F1 scores are both lower, which indicates that the model does not equally successfully reduce false negatives. Furthermore, the scores are not particularly high, in terms of accuracy and specificity. As a result, overall, the model should be considered as unsatisfactory in light of our objectives.

4.2.4 Tree Models

A popular machine learning method for incorporating multiple predictors is decision trees. Decision trees are inherently nonparametric and nonlinear which allows them to not impose any assumptions on the relationship between independent variables and the target variable (Almaça et al., 2013). As such, they are more flexible than the generalized linear models. An example of a simple decision tree is displayed in Figure 4.7.

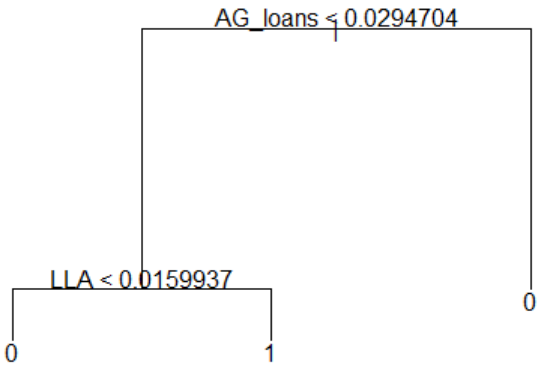

Figure 4.7: Example of Decision Tree

Decision trees have hierarchical levels where each level conducts tests on a numerical feature compared to a threshold (Abdulazeez & Jijo, 2021). A decision tree leaf node ends in a boolean outcome, either TRUE or FALSE (1 or 0) (Abdulazeez & Jijo, 2021), which, in this case, is the *audited* variable. One individual decision tree is a weak[6] predictor as it is prone to overfitting and often does not predict particularly well on its own (Brownlee, 2021; Natekin & Knoll, 2013). Therefore, adding more decision trees and comparing their results improves overall predictions. This approach is referred to as an ensemble method and includes both random forest and LightGBM (Sun et al., 2020).

Random Forest

Random forest modeling is an expansion of the simpler decision tree. It builds on the basis of bootstrap aggregation, or bagging as it is more often called (James et al., 2013). In this method, one draws a number, B, different bootstrap samples of the data (Gu et al., 2020). For each sample, the model creates a decision tree using a random sample of a specified number m features (James et al., 2013). In classification, the output of a random forest model is the popular vote of these classifiers (Gu et al., 2020).

Both the number of trees, B, and the maximum number of features, m, are hyperparameters of the model. Random forest uses a large set of trees as a basis for the prediction, which in turn will reduce the influence of each tree. Consequently, overfitting is less of an issue when using random forest compared to other machine learning models (James et al., 2013).

[6]only performs slightly better than random guessing or flipping a coin (50%) (El Fares, 2022)

Furthermore, increasing the number of trees will result in higher accuracy, as it computes a broader foundation for the popular vote (Breiman, 2001). As a result, we choose a large number while keeping in mind that each tree increases computational time. We have decided to use n =1000 trees. For the number of features considered per tree, it is useful to do some hyperparameter tuning. Ideally, one would want to use repeated cross-validation across k folds on a sequence of different m values, referred to as a k-fold grid search (Beheshti, 2022). As mentioned previously, we have simplified this tuning and followed former research. Generally, one can use $m = \sqrt{p}$, where p is the total number of features (James et al., 2013). We are therefore using $m = \sqrt{16} = 4$. The confusion matrix for random forest is displayed in Figure 4.8 below.

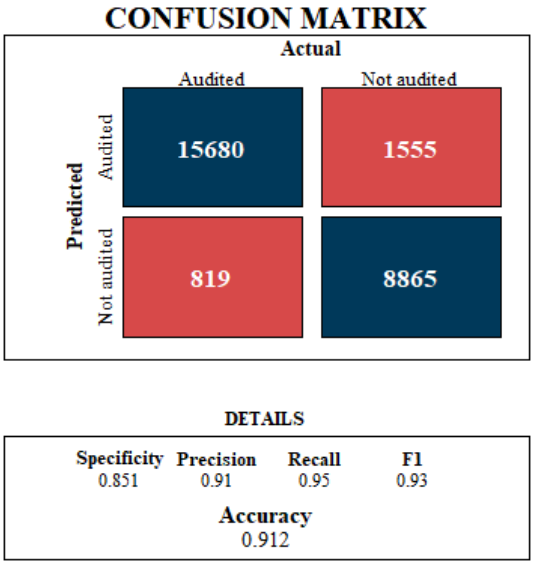

Figure 4.8: Confusion Matrix for Random Forest

The random forest model has consistently good scores for the evaluation metrics, with a high accuracy score of 91.2%. As the recall of the model is 95%, there are few false negative predictions, resulting in a model that is more suitable for our needs. The specificity, however, is marginally lower than all other values, indicating that the model, similarly to the earlier models, favors *audited* slightly. Overall, the model performs well on the test set in light of our research question.

LightGBM

Gradient boosting is another ensemble method using decision trees. The trees are trained in sequence rather than aggregate resampling (Ke et al., 2017). This process is called boosting (Ke et al., 2017). Boosting poses the question of whether one can "boost" a weaker model, such as the decision tree, into a more accurate learning algorithm (Freund & Schapire, 1999). A gradient boosted decision tree (GBDT) uses algorithms to incrementally fit new models to better estimate the response variable (James et al., 2013). With this algorithm, the base learners are constructed so that they are maximally correlated with the negative gradient of the loss function for the whole ensemble (Natekin & Knoll, 2013). The LightGBM model aims to minimize a loss-function (Sun et al., 2020). In this analysis, we use the binary log loss function.

GBDT is known for its efficiency, accuracy and interpretability (Ke et al., 2017). Two of the most popular gradient boosting machines are LightGBM and XGBoost (Ke et al., 2017). LightGBM uses leaf-wise growth as opposed to level-wise used in the XGBoost algorithm (Ke et al., 2017). This makes LightGBM faster as it only needs to fit the next level for the individual node, rather than an entire new level. A disadvantage with LightGBM is that the algorithm is prone to overfitting (Saha, 2022). To counteract this, one must select an appropriate maximum depth for each tree (Sun et al., 2020). Ke et al., (2017) found that in their experiments, LightGBM outperformed XGBoost (amongst others) by some margin in training time, while maintaining the same level of accuracy (Ke et al., 2017). Due to our limited computational resources, it is therefore more beneficial to use the LightGBM algorithm rather than the XGBoost.

Generally, in boosting, there are three tuning parameters; (i) the number of trees, B, (ii) number of splits in each tree, and (iii) the shrinkage parameter, λ (James et al., 2013). LightGBM uses these three, and more. According to Sun et al. (2020), the main parameters of the LightGBM function are the number of leaves in each tree, the learning rate (λ), the maximum depth of each tree, the minimum number of records a leaf may have, feature fraction and bagging fraction (Sun et al., 2020). For the most part, we use the default values included in the `lightgbm` package in R studio. However, to avoid

overfitting, we limit the maximum depth of each tree to 5 levels. We set the number of trees to 1000 as computational time for this model is relatively short, and since it is the same number of trees used in the random forest model.

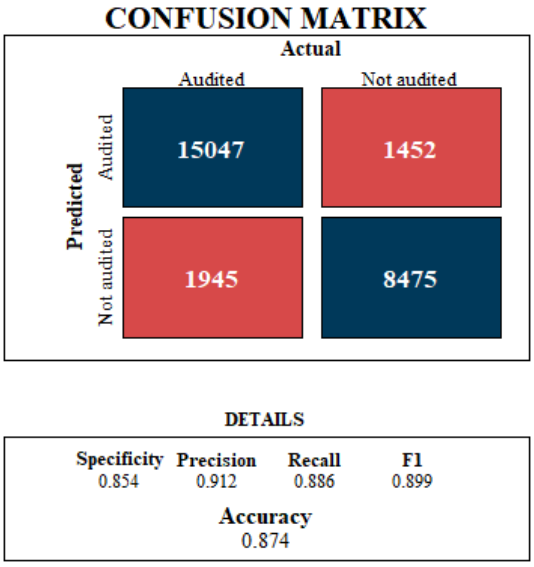

Figure 4.9: Confusion Matrix for LightGBM

Figure 4.9 presents the confusion matrix for the LightGBM model applied to the test set. The model achieves high scores for all evaluation criteria. In this model, precision is slightly higher than recall, which indicates a slight bias towards reducing false positive predictions. Similar to the random forest model, this model has lower specificity than the remaining metrics, suggesting that it may favor *audited* predictions. Overall, the model performs well.

4.2.5 Summary of Model Performances

All confusion matrices are displayed alongside each other in Appendix A6. The worst performing model is the logistic regression model, which predicts audits for approximately 96% of the observations of the test set. LASSO slightly outperforms the logistic regression, and is more consistent across the evaluation metrics. However, it is evident from the confusion matrices that the best performing models are the two tree-based models. These models are nonlinear, which could imply that the relationship between the independent variables and the response variable is better represented in a nonlinear function, at least without additional feature engineering. It is important

to note that these models depend on their respective hyperparameters. That being said, tuning hyperparameters to achieve a better test set result may not necessarily result in a better model outside of the data, as the model may be overfitted on the test data.

The predictive performance must be considered when analyzing the feature importance in the next section. A good model will give a more accurate representation of reality, while a bad model may be biased or provide false information. As a premise for this analysis, we have assumed that the cost of a false negative is higher than a false positive, which means that models with high values of recall are better. Logistic regression performs well in this regard, but when other performance metrics are considered, it is evident that we cannot rely on logistic regression alone. The great predictive performances of the two tree-based models argue that the results of these should be weighted more heavily in the next chapter of the thesis.

5 Results

We have analyzed two generalized linear models and two tree-based models. All models have been evaluated using the same evaluation criteria for prediction. However, the models differ in how they measure feature contribution and importance. As a result, in this section of the thesis, we will compare the results of the models to those of its similar counterparts. In other words, we will compare the generalized linear models and tree-based models separately. We use the predictive performance on the test set to either validate or invalidate the results. Additionally, we will conduct two statistical approaches for feature selection; ANOVA and Mutual Information.

5.1 Feature Importance

5.1.1 Generalized Linear Models

The models based on the generalized linear model employ a linear relationship between the independent variables and the response variable. As such, feature importance can be interpreted by the absolute size of the β coefficient of the variable X, given that the units are equal (Brownlee, 2020). As mentioned previously, all features but *size* are dimensionless. Hence, one should exercise caution when comparing *size* to the other variables. In general, a larger β represents a more important feature, as a change in this feature leads to a larger change in the probability (Brownlee, 2020). Additionally, the relative importance is easier to interpret as all variables are on the same scale, 0 to 1.

In this part, we will use LASSO for its feature selection purposes before presenting the β coefficients for both of the generalized linear models used for prediction in the previous chapter. First, we look at the applied LASSO model. Using the optimal λ^* we have the following β coefficients presented in Table 5.1.

Table 5.1: Coefficients for λ^*

	β
(Intercept)	-1.18
AG_loans	-4.82
cash	0.60
CI_loans	-0.18
fed_funds_sold	1.85
IN_loans	-0.89
intangible_assets	2.81
LFR	-7.55
LLA	22.50
LLP	1.60
prem_fixed_assets	6.89
RE_loans	0.00
ROA	0.00
size	1.23
tier1_ratio	-0.36
tier2_ratio	5.82
total_loans	-0.71

The β for *RE_loans* and *ROA* are set to zero, which means that the model removes these features from the model. In other words, the LASSO model determines that they are not important. *LLA* has the highest β by a considerable margin at 22.50, followed by *LFR* at -7.55. As previously explained, when increasing the λ, we are increasing the number of coefficients equal to zero. The variables that are near to zero will, intuitively, reach zero first when λ increases. Accordingly, by iteratively selecting a larger λ, we aim to identify the features that LASSO believes are the most important. We refer back to Figure 4.5.

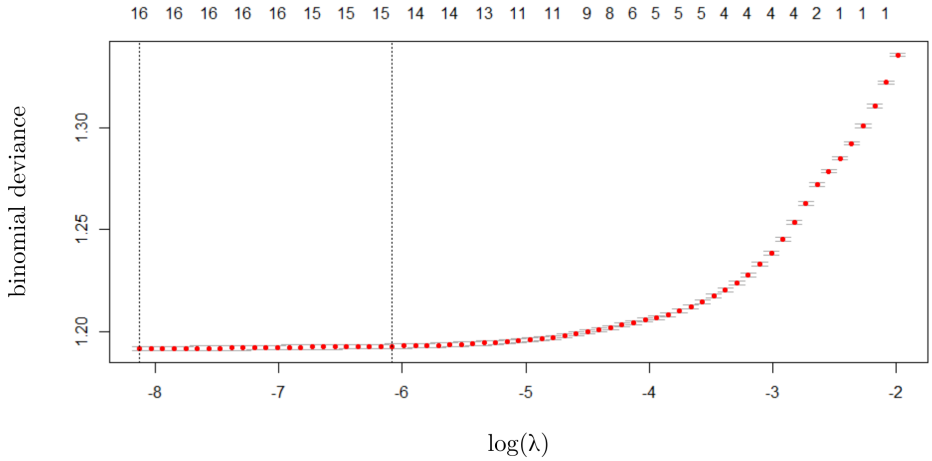

Figure 4.5: Plot for binomial deviance over log λ

Figure 4.5 illustrates that the binomial deviance of the model starts to increase more rapidly around log $\lambda = -4$. This is close to where the number of nonzero coefficients decreases from six to five, and we want to identify which features remain at this point. To investigate this further, we use the λ interval: 0.003 (just above λ^*) to 0.018 (log $\lambda = -4$), and fit new LASSO models. The β coefficients for these models are displayed in Appendix A7.

At $\lambda = 0.018$, we are left with *AG_ loans*, *LFR*, *LLA*, *LLP*, *prem_ fixed_ assets* and *size*. Of these, *AG_ loans* and *LFR* have negative coefficients which means that increases in these values will shift the predicted probability towards zero, i.e. *not audited*. The other features consequently contribute positively towards *audited*. *LLA* has the largest absolute β, as it did for λ^*. In order to verify that this feature reduction has not drastically affected the prediction accuracy, we have examined the accuracy of all models for all λ with the accuracy staying within 1% of the original model. The results can be seen in Appendix A8.

We use the features found for λ^* and for $\lambda = 0.018$, model (2) and (3) respectively in Table 5.2, and rerun the logistic regression model in order to check if there is any differences in significance for the β coefficients. The results are displayed along with the original logistic regression model with all features, model (1), in Table 5.2. In the original model used in the previous section, all features are significant at the 5% level, and further all except *ROA* are significant on the 1% level. Overall, there are no differences in the models' significance.

Table 5.2: Logistic Regression Models

	Dependent variable:		
	audited		
	(1)	(2)	(3)
AG_loans	−6.053***	−6.141***	−4.882***
	(0.128)	(0.127)	(0.061)
cash	0.777***		
	(0.073)		
CI_loans	−1.489***	−1.468***	
	(0.133)	(0.132)	
fed_funds_sold	2.270***	1.944***	
	(0.171)	(0.169)	
IN_loans	−2.336***	−2.428***	
	(0.160)	(0.160)	
intangible_assets	5.122***	4.952***	
	(0.801)	(0.801)	
LFR	−10.904***	−11.157***	−8.760***
	(0.638)	(0.638)	(0.591)
LLA	21.393***	22.265***	27.894***
	(0.850)	(0.845)	(0.680)
LLP	1.723***	1.671***	1.492***
	(0.050)	(0.048)	(0.046)
prem_fixed_assets	7.435***	7.172***	6.894***
	(0.451)	(0.450)	(0.431)
RE_loans	−1.096***	−1.114***	
	(0.113)	(0.113)	
ROA	−0.066**		
	(0.027)		
size	1.262***	1.179***	1.172***
	(0.029)	(0.028)	(0.026)
tier1_ratio	−0.813***	−0.744***	
	(0.085)	(0.084)	
tier2_ratio	15.849***	15.918***	
	(3.214)	(3.212)	
total_loans	−0.952***	−1.047***	
	(0.048)	(0.047)	
Constant	0.034	0.216*	−1.623***
	(0.130)	(0.128)	(0.024)
Observations	152,536	152,536	152,536
Log Likelihood	−90,857.030	−90,916.040	−91,329.490
Akaike Inf. Crit.	181,748.100	181,862.100	182,673.000
Note:		*p<0.1; **p<0.05; ***p<0.01	

5.1.2 Tree Models

As mentioned, both random forest and LightGBM are nonparametric, nonlinear models. Both models are high in flexibility, but low in interpretability, meaning that the feature importance is less intuitive than in the previous models. To determine the importance of a feature, one must look to other, more complex measurements, such as a decrease in accuracy or information gain per feature, rather than just the significance and absolute value of a coefficient in a linear expression. In this section, we will present the metrics for each model and compare the results. Both tree models predict with a very high degree of accuracy across all evaluation criteria, which suggests that the most important features really influence the decision to get no audit.

Before looking at the feature importance of the tree models it is beneficial to get some background on entropy. Entropy is an information measure that measures chaos in a system (Gray, 2013). This information measure is used in constructing decision trees to handle the unknown (Ye, 2022). The entropy averages the information gain for the variable, and a low entropy means that there is much information gain, while high entropy expresses the opposite (Ye, 2022). When constructing a decision tree, features with lower entropy are placed higher in order for the tree to be as effective as possible in directing inputs down a series of conditions to a correct outcome (Ye, 2022). The higher up a feature split is in the tree; the more information the variable gives to the model.

Random Forest

The plots in Figure 5.1 below display the importances for all features in the random forest model, using two different measures. The left-hand plot displays the mean decrease in accuracy when removing the variable from the model, and the right presents the mean decrease in node impurity (Anaraki & Haeri, 2022; R Core Team, 2022). Each feature's mean decrease in impurity, or Gini importance, is calculated by the sum of feature importance across all trees proportionally to the number of splits (Anaraki & Haeri, 2022). The scale for Gini importance is relative only to itself (Alfaro-Cortes et al., n.d.), meaning that, for example, a variable with a mean decrease of 4000 is twice as important as a variable with a mean decrease of 2000.

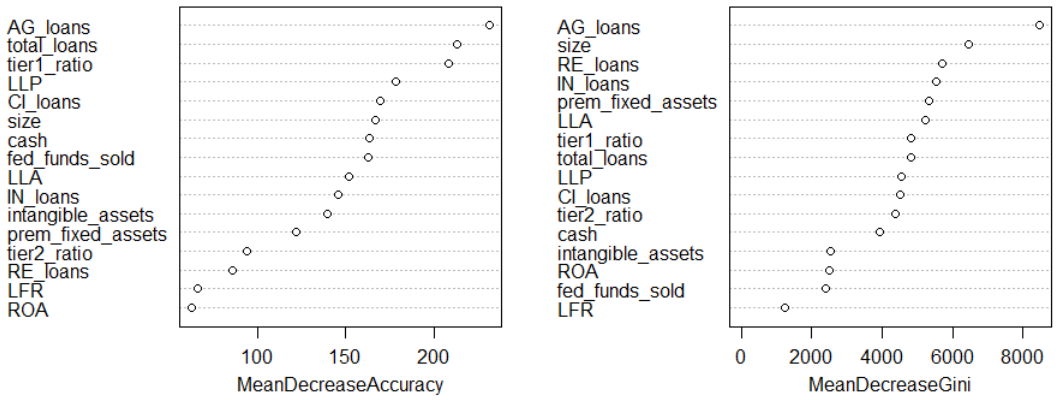

Figure 5.1: Plot for Random Forest Feature Importance

The main difference between the two measures is that the Gini measure is used when creating the best model (Menze et al., 2009), and mean decrease in accuracy is a product of running the model several times (R Core Team, 2022). It is necessary to repeat the model since the mean decrease in accuracy denotes how the removal of a particular feature affects accuracy on average, scaled by the standard deviation (R Core Team, 2022). A feature being high in both accuracy and Gini indicates that there are many trees and splits that include this feature. Additionally, the feature is present in many trees that produce the correct prediction.

There are some substiantial differences in the rankings between the two plots in Figure 5.1. An example is *RE_loans*, achieving a much lower score of accuracy than Gini. Furthermore, *fed_funds_sold* have a lower score for Gini importance, which might be a result of the data set having few unique values, meaning that splits by these features are not common. Given that *intangable_assets* and *LFR* also have few unique values and have very low Gini scores, this seems like a plausible bias for the Gini measure.

It is also possible to identify some more prominent features for the prediction model as *AG_loans*, *LLP* and *tier1_ratio* all score relatively high for both accuracy and Gini. Slightly lower, *LLA* is also fairly good. *AG_loans* outperforms all other features by a substantial margin in both metrics.

The mean decrease in accuracy is also presented per class (*audited* = 1, *not audited* = 0) in Table 5.3.

Feature	0	1
AG_loans	210.61	185.74
cash	154.94	142.01
CI_loans	149.74	150.02
fed_funds_sold	142.98	148.83
IN_loans	140.72	136.37
intangible_assets	126.23	138.44
LFR	60.63	70.21
LLA	145.85	134.82
LLP	177.18	156.80
prem_fixed_assets	125.84	112.47
RE_loans	81.80	82.06
ROA	60.78	21.20
size	166.66	154.30
tier1_ratio	202.85	165.72
tier2_ratio	98.05	77.43
total_loans	207.36	176.45

Table 5.3: Random Forest Feature Importance per Class

The findings are consistent with the left-hand plot of Figure 5.1, as the identified most important variables receive the highest scores for both classes, 0 and 1, in Table 5.3. Table 5.3 presents that *AG_loans*, *total_loans*, *tier1_ratio* and *LLP* have substantially higher scores of mean decrease in accuracy when predicting *not audited*. In terms of the overall class-specific mean decline in accuracy, the majority of features have similar scores, and only a few receive higher marks for *audited*. This gap between the classes is likely due to the test set's overweight of *not audited* observations. If one removes a feature from the set and reruns the model, it is likely to affect more *not audited* observations due to the fact there are more in the set.

LightGBM

In LightGBM, we measure the feature importance in information gain. Information gain is defined as the total gain for all splits by that feature divided by the total gain for all splits (Lee, 2020). Figure 5.2 illustrates the importance of the features in terms of the percentage of information gained per feature, presented numerically in Appendix A10.

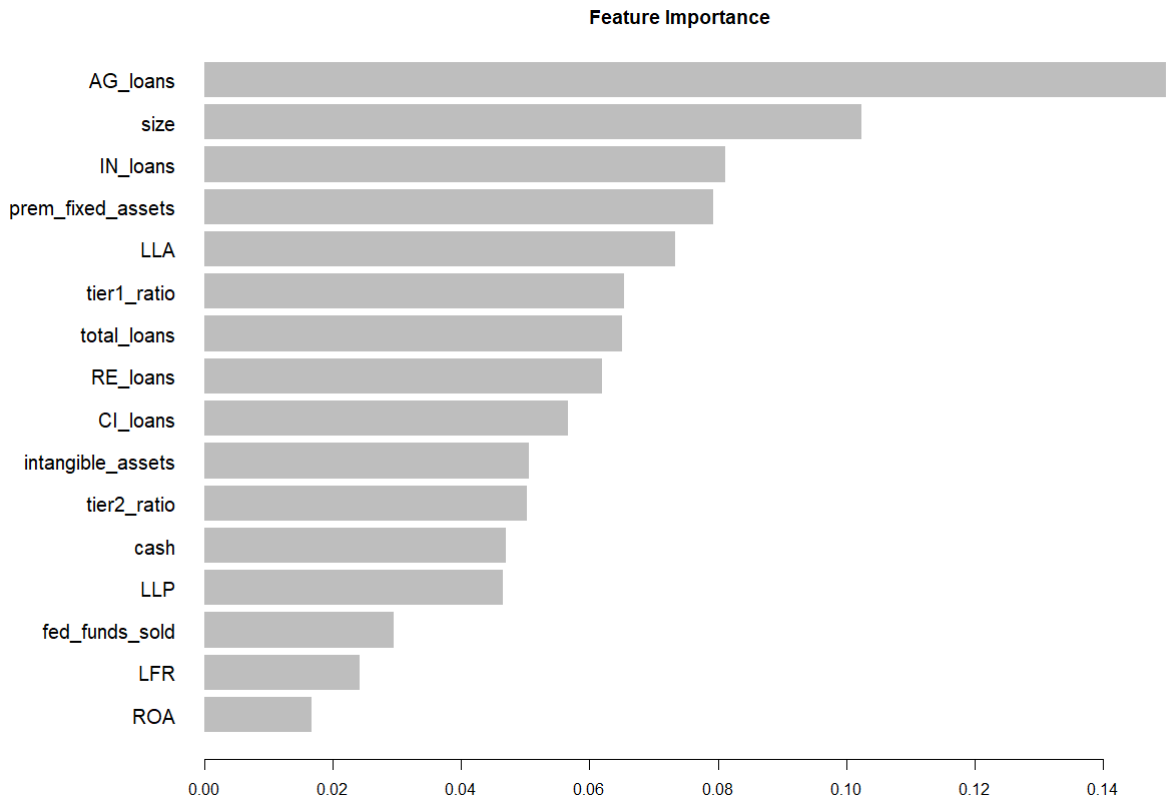

Figure 5.2: Plot for LightGBM Information Gain

AG_loans and *size* are the only ones to surpass 10%, followed by *IN_loans*, *prem_fixed_assets* and *LLA*. Moreover, *tier1_ratio* scores high in information gain, as it also did for mean decrease in accuracy in the random forest model. Furthermore, among the top five features, two features are categories of loans. This suggests that these types of loans contain much information on the decision to get no audit.

5.2 Statistical Analysis for Feature Selection

As explained previously, feature selection is typically carried out before the machine learning models are trained, in order to decrease the number of features. However, because the features are selected based on prior literature and our research question, we will use this statistical analysis to verify the results presented in the previous paragraphs.

Similarly to model training, one often separates feature selection into supervised and unsupervised feature selection (and sometimes semi-supervised) based on whether the training data is labeled, unlabeled, or partially labeled (Brownlee, 2019). The data in

our data set is labeled, meaning we have a y-value for every $x = (x_1, x_2, x_3, \ldots, x_n)$. We conduct the feature selection on the entire data set as opposed to just the training set. Generally in predictive modeling one disregards statistical learning as the aim is solely to make good predictions (Brownlee, 2018). However, statistical learning can still be useful in this analysis in order to increase interpretability of the relationship with the response variable (Brownlee, 2018).

It is essential to consider whether the input and output is numerical or categorical when choosing feature selection techniques. In our model, all input variables are continuous and numerical, and the output variable is categorical binary output. We have selected two approaches that fit these criteria; Mutual Information and ANOVA. Mutual Information studies statistical dependence of any kind (Vergara & Estévez, 2013), while ANOVA is interested in linear dependency (Asaithambi, 2018).

5.2.1 Mutual Information

Mutual Information measures how much information a variable has about another variable, and is explained through primarily two aspects. First, Mutual Information can assess any relation between variables, including nonlinear relationships. Second, Mutual Information is invariant under transformations in the feature space that are invertible and differentiable. The latter means that the Mutual Information is invariant with, for example, translations, rotations, and any transformation that preserves the order of the original elements. (Vergara & Estévez, 2013)

The information gained to predict y if we know x is defined as:

$$I(x, y) = H(x) - H(y|x)$$

where H is entropy. (Vergara & Estévez, 2013)

Mutual Information is zero where the variables x and y are statistically independent (Vergara & Estévez, 2013). We have identified the relationship between the discrete *audited* variable and the continuous independent variables using the `mutual_info_classif` function from the `sklearn` library in Python. The results for the top ten most important variables are displayed in Figure 5.3.

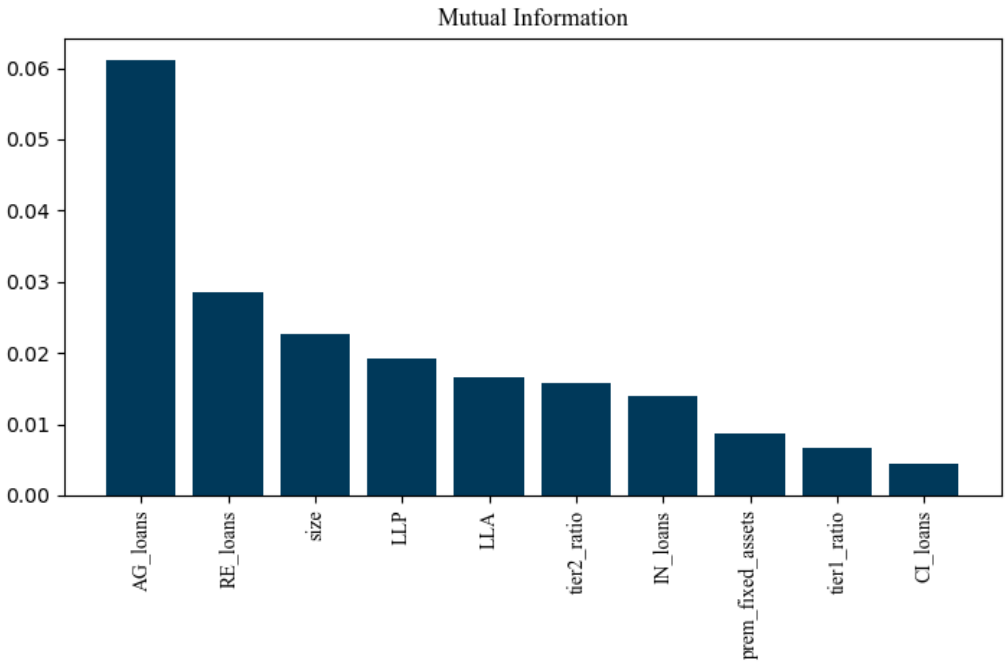

Figure 5.3: Plot for Mutual Information scores

The values for mutual information to the target value, *audited*, are low. There are, however, some higher values for the different loan categories. The highest is *AG_loans* which outperforms all other features, followed by *RE_loans*. *IN_loans* and *CI_loans* are also present, signifying that all four loan categories contribute information to the audit decision. Furthermore, *size*, *LLP* and *LLA* are all amongst the top five scoring features, though all below half of the score of *AG_loans*.

5.2.2 ANOVA

ANOVA stands for Analysis of Variance and is one of the most widely used statistical models (Soumare, 2020). Using ANOVA, one investigates whether there are statistical differences between the means of independent groups by analyzing the levels of variance within the groups through samples taken from each of them (Qualtrics, n.d.). One divides ANOVA into one-way and two-way ANOVA, depending on how many independent variables are in the model (Qualtrics, n.d.). In order to investigate two or more independent variables, one has to use two-way ANOVA (Qualtrics, n.d.).

In classification, ANOVA tests for a change in deviance instead of the standard variance (University of Pennsylvania, n.d.). This is often referred to as ANODev, Analysis of Deviance (Härdle & Huang, 2013), and was first introduced by Nelder and Wedderburn (Jørgensen, 2012). The null hypothesis for the ANODev is that the means for each class are equal. We verify this at a 1% level (Hu, 2022). The F-score for an ANODev is defined as:

$$F = \frac{\text{distance between classes}}{\text{compactness of classes}}$$

(Islam et al., 2019)

In order to conduct an ANOVA, or in this case, an ANODev test, one must ensure that the data is normal and homogeneous. First, the independent variables must be normally distributed. It follows from the central limit theorem that a sample containing $n>30$ observations approaches the normal distribution (Kwak & Kim, 2017). As such, one could argue that the sample size in itself is sufficient. It should still be noted that when investigating the histograms of the numerical variables, all of them do not seem to resemble the normal distribution. Soumare (2020) conducted multiple experiments and found that type 1 errors[7] were comfortably controlled even where data was closer to a Poisson or Negative Binomial distribution (Soumare, 2020). These distributions are closer to what we observe in the histograms. Despite the fact that Soumare's (2020) findings were based on count data, we conclude that we may conduct the test after taking both the their findings and the central limit theorem into consideration. Before conducting the test, the data is standardized such that we have a standard deviation equal to 1 and a mean equal to 0. This is prefaced by denormalizing *size*, *LLP* and *ROA* to ensure homogeneity in the data.

The analysis is conducted using the `f_classif` function from the `sklearn` library in Python. The results are visualized in Figure 5.4 below.

[7]type 1 error: probability of rejecting the null hypothesis when it is in fact true (Soumare, 2020)

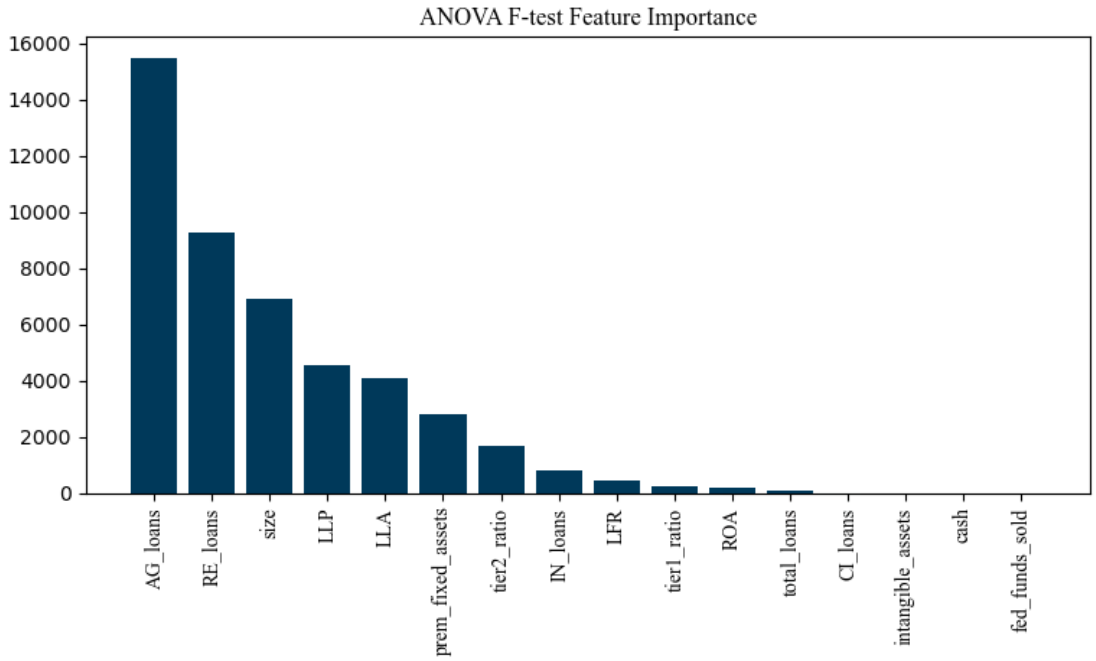

Figure 5.4: Plot for ANOVA F-scores

The ANOVA test shows higher importance for the loan types *AG_loans* and *RE_loans*, as well as *size*, *LLP*, and *LLA*. The high importance indicates that these features have a relatively high impact on the banks' audit choices. However, it does not indicate in which direction the features are moving. ANOVA discards the null hypothesis for most features, using a significance level of 1%. The exceptions are *fed_funds_sold*, *cash* and *intangible_assets*, implying no significant difference between *audited* and *not audited* for these features.

6 Discussion

6.1 Interpretation of Results

The logistic regression and the LASSO model both predict at roughly the same level of accuracy, scoring below the 70% threshold. This means that predicting all audits would outperform these models on the test set. Furthermore, the logistic regression model scores extremely low for specificity, which is a major drawback of this model. As a result, it is our opinion that the results from the logistic regression model should not be considered. Instead, we emphasize the tree models. Even if the model did not predict particularly well, the results from LASSO can be helpful when discussing the results from the tree models. Especially one should consider the features that remain for $\lambda = 0.018$ which were *AG_loans*, *LFR*, *LLA*, *LLP*, *prem_fixed_assets* and *size*.

The results from the tree models argue that the most important feature is *AG_loans*. This is further supported by the results from the statistical analysis, where *AG_loans* perform substantially better than all other variables. The statistical analysis identifies the most important variables as *size*, *LLA*, *RE_loans*, and *LLP*, in addition to *AG_loans*. These variables receive the highest scores for both the Mutual Information metrics and F-scores for ANOVA. In the Mutual Information model, the highest scoring variables also include *tier2_ratio*, which in ANOVA does not perform especially well. Furthermore, *prem_fixed_assets* and *IN_loans* are present in the top ten variables in both analyses. Comparing these results to the results of the tree models, it is evident that *size* and *LLA* are the most prominent features at the end of this analysis, in addition to the aforementioned *AG_loans*.

The differences in the two random forest importance measures provide insight into entropy reduction and how it differs from accuracy. *IN_loans* and *prem_fixed_assets* score relatively well in the LightGBM model and are both ranked in the top half of the statistical analysis. However, in the random forest model, while the scores are high in terms of Gini importance, they are low for mean decrease in accuracy. The same is

true of *RE_loans* which is one of the best in terms of the statistical analysis. These discrepancies, and the fact that high Gini scores coincide with high feature selection scores, are attributed to entropy. Both tree models try to reduce the amount of entropy. For random forest modeling, this is represented by the Gini importance measure, where the highest score represents the largest decrease in entropy (Menze et al., 2009). LightGBM's information gain metric is entropy measured before and after a split (Yıldırım, 2020). As Mutual Information is a measure of the differences in entropy, it is coherent that it will be very similar to these two other metrics. Moreover, it can be argued that one should give accuracy greater weight since entropy is measured using several metrics.

There are several variables where the results differ. For example, *total_loans* and *tier1_ratio* perform substantially better in the tree models than in the statistical analysis with the two being amongst the top three in the random forest mean decrease in accuracy. In light of the previous paragraph, these results suggest that although these variables may not be present in many trees, they may be important to increase accuracy.

Amongst the selection of high risk variables, *LLA* is found to be the most important. Although *LLP* performs well for the random forest model and statistical analysis, it ranks amongst the worst for the LightGBM information gain. In contrast, *LLA* is more consistent all around despite being somewhat less important in the mean decrease in accuracy. Since both variables are created and maintained by banks to account for loan losses, we do not consider the selection of one of the two variables as very important. As a result, we are concentrating on the more consistent *LLA*. A high *LLA* does not necessarily mean that a bank has bad financial health, but it indicates that the quality of the bank's assets are worse than those of banks with low *LLA*.

We must look at the importance per class to interpret how it affects the decision not to audit. It is however important to note that, as a result of using nonlinear, nonparametric models, it is not straightforward to gather what direction a feature is pulling, like it is in LASSO or logistic regression. Constructing a decision boundary is also difficult due to the number of features. As such, although *AG_loans* has a higher score for predicting

not audited, it is not necessarily true that higher AG_loans steers in the direction of *not audited*.

6.2 Robustness of Machine Learning Models

All models are validated using cross-validation, which uses data that was withheld from the model during training. This gives a more accurate representation of the models' performances. To further back up our conclusions, we have run all four models on a range of training and test sizes within the 10-30% test size interval. These results are in Appendix A13. The accuracies are consistent for all models.

Furthermore, credibility is built by including several prediction evaluation criteria as opposed to solely accuracy. As explained, a broader assessment allows us to identify irregularities in model performance, and discard models that favor a specific class. Additionally, we can disregard models that favors minimization of false positives at the expense of false negatives.

6.3 Limitations and Further Research

A central limitation of the analysis is the quality of the data. Many variables had missing values and had to be computed from other reporting variables in the original data set. Furthermore, as a result of many of the Call Reports being filed manually on paper before entering the database, there is risk of spelling errors (as we found for state abbreviations), false data, and missing data. These errors may also occur for entirely digital processes, but there is a higher risk of such mistakes happening by hand due to human error.

Additionally, many of the values in the data set are equal to zero. While many of these observations are probably correctly reported, it can cause an issue in the prediction (Delucchi & Bostrom, 2004). It is evident in the histograms, where many features have a pillar of zeros. A reason for these zero values is the focus on loans in different categories. Smaller banks do not typically have loans in all categories, thus many observations will have a value of 0. As we have investigated the data thoroughly, we are certain

that this is not due to missing values. Because observations with no loans issued in a category are also important for predictive modeling, removing these values would be problematic. Additionally, transforming the data is not an option because it would not solve the issue of numerous scores having the same value (i.e., zero) (Delucchi & Bostrom, 2004).

It should also be mentioned that the filtering of banks that operate in states where external audits are mandatory, is flawed. Navigating the American legal system is challenging, primarily due to the large differences between the states. Although being a convenient approach, using a threshold of audited banks to the total number of banks in each state may not be accurate. It is therefore possible that some states have voluntary audits even when having a percentage of audited banks above our threshold, and vice versa.

Additionally, it is fair to question if the financial statements of audited and unaudited banks can be analyzed on the same basis. Nicoletti (2018) showed that regulators and auditors influence banks' financial reporting differently. As regulators and auditors have different focuses during their work, set by their different objectives, can one question whether the accounting principles are employed in the same way across *audited* and *not audited* banks.

In terms of further research, the analysis could be replicated on different data, as this will test the model robustness and also the validity of the conclusions. Testing the models on more recent data after the Covid-19 crisis could be particularly relevant because it provides an opportunity to look at banks in the upcoming years. Other machine learning algorithms could also be used. As our results indicate that the relationship between the decision to get no audit and the independent variables is not linear, other nonlinear models should be considered. Additionally, we are not including a time aspect in our data, which assumes that the systematic choice not to audit does not change over time. In further research, one could introduce this time aspect in the data and look for trends. One should in such a case consider a dedicated time series classification algorithm as they are better suited than the ones applied in this analysis.

6.4 Remarks on Large Banks

While arguably a limitation of the analysis, we believe it is useful to provide some discussion on the larger banks removed in the data cleaning in a section of its own. To ensure that all banks used in the analysis are not obligated to have an external audit, we removed all banks over the FDIC threshold of 500 million USD. Removing these banks is customary in the little research we found on the topic, and is done in an effort to remove banks that are assumed to have mandatory audits. We find this to be inaccurate in our data.

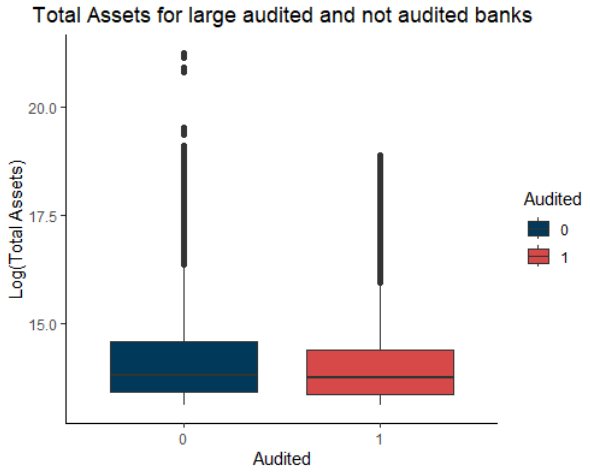

Figure 6.1: Box Plot for Large Banks

Figure 6.1 displays a box plot of total assets for class-specific observations that were removed from the data set during cleaning and whose total assets exceeded 500 million USD. The results for the two classes are visibly similar, although the maximum values for the unaudited observations are higher. The medians for each class are both close the threshold, but given that the scale is logarithmic, their assets are still substantially larger than the 500 million USD mark. Although close, the 75th percentile of unaudited observations is higher. As such, in addition to the fact that these observations have the most extreme values, one may conclude that the observations that were not audited have more overall assets than the audited observations.

It should further be noted that the majority of these observations are classified as *not audited*. As FDIC requires an external audit by all banks (over the threshold) insured by them, this indicates that FDIC does not insure these banks. Therefore, by excluding these banks, one may wonder if we are leaving out any effects that could be worth investigating.

7 Conclusion

This thesis uses bank regulatory data to investigate factors influencing small private U.S. banks' decision to get no audit. Previous research has found that this choice is systematic, but few have researched what the drivers are. We use machine learning to create four prediction models: logistic regression, LASSO, random forest, and LightGBM.

The analysis finds two well-performing predictive models. The two generalized linear models, logistic regression and LASSO, are discarded in favor of the tree models, random forest and LightGBM, due to the substantial difference in prediction accuracy across all evaluation criteria. As such, the results of the analysis suggest that the relationship between the choice of getting no voluntary external audit and the selected variables is nonlinear. The best performing model is the random forest model which predicts with over 90% accuracy on the test set. Either of the tree-based models may be beneficial for identifying banks that do not conduct audits and, as a result, facilitate further research on what characteristics these banks have in common.

Findings show that across all prediction models, the most important variable for the decision to get no audit is the proportion of agricultural loans to the total sum of loans. The agricultural loans' importance is further supported by Mutual Information and ANOVA feature selection. The thesis concludes that this feature is the most important in terms of reducing entropy and increasing accuracy. It follows from these findings that different loan categories, based on their security, borrower, and purpose, may have a large impact on the decision to get no audit. To the best of our knowledge, this phenomenon has not previously been discovered, which urges future research on the topic.

Bank size and loan loss allowance are found to be important features for all of the models used in this thesis. In terms of size, this is coherent with former research that has identified both bank size and the size of other private firms as drivers for audit choice. The fact that our results appear to be in agreement with previous research on both private firms and banks, suggests that the factors that lead to

procuring no voluntary external audit may be similar for banks and private firms. Loan loss allowance is also shown to be an important feature for our prediction models. As specified in the literature review, this factor is related to asset quality since it reveals what the bank loses on subprime loans. Accordingly, we conclude that asset quality and the size of total assets are important factors when deciding to get no audit.

This thesis provides a foundation for further research on banks and their behaviour. It is important to note that the performance evaluation of the models is based entirely on the test set. However, future research may be able to verify the accuracy of the models in the context of newer data, as well as fine-tune model hyperparameters. Although the analysis is conducted on data of imperfect quality, the results point toward an exciting new direction in research on voluntary audits. Furthermore, the knowledge of the relationship between features being nonlinear can be used in future analyses on banks as this thesis sets a foundation for using machine learning in investigating banks' behavior.

References

Abdel-Khalik, A. R. (1993). Why do private companies demand auditing? A case for organizational loss of control. *Journal of Accounting, Auditing & Finance*, 8(1):31–52. doi = 10.1177/0148558x9300800103.

Abdulazeez, A. M. & Jijo, B. T. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 02(01):20–28. doi = 10.38094/jastt20165.

Accountancy Europe (2017). Definition of Public Interest Entities in Europe. https://www.accountancyeurope.eu/wp-content/uploads/171130-Publication-Definition-of-Public-Interest-Entities-in-Europe_1.pdf.

Alaska Statutes Title 44. (2021). State Government § 44.81.270. Audits and examinations of bank . https://codes.findlaw.com/ak/title-44-state-government/ak-st-sect-44-81-270.html. Online; accessed 15-December-2022.

Alfaro-Cortes, E., Gamez-Martinez, M., & Garcia-Rubio, N. (n.d.). R: Plots the variables relative importance. https://search.r-project.org/CRAN/refmans/adabag/html/importanceplot.html. Online; accessed 13-December-2022].

Almaça, J. A. F., Quirós, J. T., & Justino, M. d. R. F. (2013). Nonparametric decision tree: The impact of ISO 9000 on certified and non certified companies. *Intangible Capital*, 9(3). doi = 10.3926/ic.303.

Alpaydin, E. (2020). *Introduction To Machine Learning.* Mit Press.

Alpert, G. (2021). Loan loss provision. https://www.investopedia.com/terms/l/loanlossprovision.asp#:~:text=A\%20loan\%20loss\%20provision\%20is.

Anaraki, S. A. M. & Haeri, A. (2022). Soft and hard hybrid balanced clustering with innovative qualitative balancing approach. *Information Sciences*, 613:786–805. doi = 10.1016/j.ins.2022.09.044.

Asaithambi, S. (2018). Why, How and When to apply Feature Selection. https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2.

Barton, J., Hodder, L. D., & Shepardson, M. L. (2015). Audits and Bank Failure: Do Financial Statement Audits Reduce Losses to Capital Providers? *SSRN Electronic Journal*. doi = 10.2139/ssrn.2719198.

Beheshti, N. (2022). Cross Validation and Grid Search. https://towardsdatascience.com/cross-validation-and-grid-search-efa64b127c1b.

BIS (n.d.). Definition of capital in Basel III - Executive Summary. https://www.bis.org/fsi/fsisummaries/defcap_b3.pdf. Online; accessed 9-December-2022.

Bratko, I. (1997). Machine learning: Between accuracy and interpretability. *Learning, Networks and Statistics*, 382:163–177. doi = 10.1007/978370912668-4_10.

Breiman, L. (2001). Random Forest. *Machine Learning*, 45(1):5–32. doi = 10.1023/a:1010933404324.

Breitenlechner, M., Geiger, M., & Scharler, J. (2022). Bank liquidity and the propagation of uncertainty in the U.S. *Finance Research Letters*, 46:102467. doi = 10.1016/j.frl.2021.102467.

Britannica (n.d.). Bank - the principles of central banking. https://www.britannica.com/topic/bank/The-principles-of-central-banking. Online; accessed 10-December-2022.

Brownlee, J. (2018). The Close Relationship Between Applied Statistics and Machine Learning. https://machinelearningmastery.com/relationship-between-applied-statistics-and-machine-learning/.

Brownlee, J. (2019). How to Choose a Feature Selection Method For Machine Learning. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/?fbclid=IwAR1DuS46BzQ775sljIVEsAeMsVkKe1m_M26Pw1_dLeZbZmSNebflXDzW8KI. [Online; accessed 15-December-2022].

Brownlee, J. (2020). How to Calculate Feature Importance With Python. https://machinelearningmastery.com/calculate-feature-importance-with-python/.

Brownlee, J. (2021). Strong Learners vs. Weak Learners in Ensemble Learning. https://machinelearningmastery.com/strong-learners-vs-weak-learners-for-ensemble-learning/.

Bureau of Economic Analysis (2022). Gross Domestic Product (Second Estimate) and Corporate Profits (Preliminary), Third Quarter 2022 | U.S. Bureau of Economic Analysis (BEA). https://www.bea.gov/news/2022/gross-domestic-product-second-estimate-and-corporate-profits-preliminary-third-quarter.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79. doi = 10.1016/j.neucom.2017.11.077.

Chan, D. Y. & Kogan, A. (2011). Machine Learning for Predicting the Procurement of an Audit at Small Private Banks: Is the Decision to Procure an Audit Systematic? *SSRN Electronic Journal*. doi = 10.2139/ssrn.1856367.

Charbuty, B. & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2:20–28. doi = 10.38094/jastt20165.

Chen, J. (2021). Federal funds definition. https://www.investopedia.com/terms/f/federalfunds.asp.

Dedman, E., Kausar, A., & Lennox, C. (2013). The Demand for Audit in Private Firms: Recent Large-Sample Evidence from the UK. *European Accounting Review*, 23(1):1–23. doi = 10.1080/09638180.2013.776298.

Delucchi, K. L. & Bostrom, A. (2004). Methods for Analysis of Skewed Data Distributions in Psychiatric Clinical Studies: Working With Many Zero Values. *American Journal of Psychiatry*, 161(7):1159–1168. doi = 10.1176/appi.ajp.161.7.1159.

El Fares, A. (2022). The power of ensembling in machine learning. https://medium.com/@El_Fares_Anass/the-power-of-ensembling-in-machine-learning-28806fbb8d06.

FDIC (2022). Quarterly Banking Profile (FDIC Third Quarter 2022). Federal Deposit Insurance Corporation.

FDIC (n.d.a). Banker Resource Center: Internal and External Audit Programs. https://www.fdic.gov/resources/bankers/corporate-governance-and-auditing-programs/internal-external-auditing-programs/part-363.html. Online; Accessed 2-December-2022.

FDIC (n.d.b). PREMISES AND EQUIPMENT. https://www.fdic.gov/regulations/safety/manual/section3-5.pdf. Online; accessed 13-December-2022.

Federal Reserve (2017). The Fed - About MDRM. https://www.federalreserve.gov/apps/mdrm/about_mdrm.htm. Online; accessed 15-December-2022.

Federal Reserve (2021). The Fed Explained: What the Central Bank Does. doi = 10.17016/0199−9729.11.

Federal Reserve (2022). The Fed - Micro Data Reference Manual. https://www.federalreserve.gov/apps/mdrm/data-dictionary/search/item?keyword=2165&show_short_title=False&show_conf=False&rep_status=All&rep_state=Opened&rep_period=Before&date_start=99991231&date_end=99991231.

Federal Reserve (n.d.). The Fed - Micro Data Reference Manual. https://www.federalreserve.gov/apps/mdrm/data-dictionary/search/item?keyword=2165&show_short_title=False&show_conf=False&rep_status=All&rep_state=Opened&rep_period=Before&date_start=99991231&date_end=99991231.

Fernando, J. (2021). Correlation Coefficient Definition. https://www.investopedia.com/terms/c/correlationcoefficient.asp.

FFIEC (n.d.a). GENERAL INSTRUCTIONS. https://www.fdic.gov/resources/bankers/call-reports/crinst-031-041/2020/2020-09-generalinstructions.pdf.

FFIEC (n.d.b). SCHEDULE RC-C - LOANS AND LEASE FINANCING RECEIVABLES. Online; accessed 1-December-2022.

FFIEC (n.d.c). Search Institutions. https://www.ffiec.gov/npw. Online; accessed 6-October-2022.

Freund, Y. & Schapire, R. E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780.

Frost, J. (2017). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.

Gadre, G., Bhargava, A., & Mehta, L. (2016). Smart things to know about sources of income for a bank. https://economictimes.indiatimes.com/wealth/save/smart-things-to-know-about-sources-of-income-for-a-bank/articleshow/54377370.cms.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*, 1290.

Gray, R. M. (2013). *Entropy and Information Theory*. Springer Science & Business Media.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33. doi = 10.1093/rfs/hhaa009.

Härdle, W. K. & Huang, L.-S. (2013). Analysis of Deviance in Generalized Partial Linear Models. *SSRN Electronic Journal*. doi = 10.2139/ssrn.2892644.

Hargrave, M. (2022). How to Use Return on Assets When Analyzing a Company. https://www.investopedia.com/terms/r/returnonassets.asp.

Hartwig, F. & Dearing, B. E. (2002). *Exploratory Data Analysis*. Sage Publ.

Hawaii Code of Financial Institutions Chapter 412 (n.d.). §412:3-112 Submissions to commissioner. https://files.hawaii.gov/dcca/dfi/Laws_html/HRS0412/HRS_0412-0003-0112.htm. Online; accessed 15-December-2022.

Hu, S. (2022). Complete Feature Selection Techniques 4 - 1 Statistical Test & Analysis. https://summer-hu-92978.medium.com/complete-feature-selection-techniques-4-1-statistical-test-analysis-611ede242fa0.

Idaho Credit Union Act Chapter 21 (n.d.). Title 26 Banks and Banking. https://www.finance.idaho.gov/legal/statutes-rules/documents/2022-idaho-credit-union-act.pdf. Online; accessed 15-December-2022.

Islam, M. R., Rahim, M. A., Islam, M. R., & Shin, J. (2019). Genetic Algorithm Based Optimal Feature Selection Extracted by Time-Frequency Analysis for Enhanced Sleep Disorder Diagnosis Using EEG Signal. *Advances in Intelligent Systems and Computing*, pages 881–894. doi = 10.1007/9783030295134_65.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. Springer.

Jordan, J. (2017). Evaluating a machine learning model. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/.

Jørgensen, B. (2012). Generalized Linear Models. *Encyclopedia of Environmetrics*. doi = 10.1002/9780470057339.vag010.pub2.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):531–538. doi = 10.1002/sam.11583.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Online; accessed 12-December-2022.

Kenton, W. (2021). Goodwill To Assets Ratio. https://www.investopedia.com/terms/g/goodwill-to-assets-ratio.asp.

Kwak, S. G. & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2):144–156. doi = 10.4097/kjae.2017.70.2.144.

Lee, C. (2020). Feature Importance Measures for Tree Models - Part I. https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3.

Luque, A., Carrasco, A., Martín, A., & 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. doi = 10.1016/j.patcog.2019.02.023.

Melvin, R. L. (2021). Sample Size in Machine Learning and Artificial Intelligence

,Äì Perioperative Data Science. https://sites.uab.edu/periop-datascience/2021/06/28/sample-size-in-machine-learning-and-artificial-intelligence/.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213. doi = 10.1186/1471210510213.

Minnis, M. & Shroff, N. (2017). Why regulate private firm disclosure and auditing? *Accounting and Business Research*, 47(5):473–502. doi = 10.1080/00014788.2017.1303962.

Mitchell, T. M. (2006). The Discipline of Machine Learning. http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf.

Natekin, A. & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(21). doi = 10.3389/fnbot.2013.00021.

Nickolas, S. (2019). Tier 1 Capital vs. Tier 2 Capital: What's the Difference? https://www.investopedia.com/ask/answers/043015/what-difference-between-tier-1-capital-and-tier-2-capital.asp.

Nicoletti, A. (2018). The effects of bank regulators and external auditors on loan loss provisions. *Journal of Accounting and Economics*, 66(1):244–265. doi = 10.1016/j.jacceco.2018.05.003.

Nobel Prize Outreach AB 2022 (2022). The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2022. https://www.nobelprize.org/prizes/economic-sciences/2022/press-release/.

OCC (2018). Bank premises and equipment. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/bank-premises-equipment/pub-ch-bank-premises.pdf.

OCC (n.d.). Interagency policy statement on the allowance for loan and lease losses. https://www.occ.gov/news-issuances/bulletins/2006/bulletin-2006-47a.pdf. Online; accessed 29-October-2022.

Qualtrics (n.d.). What is ANOVA (Analysis of Variance) and What It's Used For. https://www.qualtrics.com/experience-management/research/anova/.

R Core Team (2022). R: The r project for statistical computing. https://www.r-project.org/.

Saha, S. (2022). Xgboost vs lightgbm: How are they different. https://https://neptune.ai/blog/xgboost-vs-lightgbm?fbclid=IwAR1bxwW3id2yUOn5ecculKBuihPvQ0mEts30D2Pf7jb-GEBUdiFLMSqRpx0. Online; accessed 18-December-2022.

Senaviratna, N. A. M. R. & Cooray, T. M. J. A. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 5:1–9. doi = 10.9734/ajpas/2019/v5i230132.

Soni, D. (2018). Supervised vs. unsupervised learning. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d.

Soumare, I. (2020). Comparing performance of anova to poisson and negative binomial regression when applied to count data.

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on lightgbm. *Finance Research Letters*, 32. doi = 10.1016/j.frl.2018.12.032.

Tibshirani, R. & Wasserman, L. (2017). Sparsity, the Lasso, and Friends Statistical Machine Learning, Spring 2017. https://www.stat.cmu.edu/~ryantibs/statml/lectures/sparsity.pdf.

Tuovila, A. (2022). Audit. https://www.investopedia.com/terms/a/audit.asp.

University of Pennsylvania (n.d.). 6.3.4 - Analysis of Deviance and Model Selection. https://online.stat.psu.edu/stat504/lesson/6/6.3/6.3.4#fullScreen.

Vanstraelen, A. & Schelleman, C. (2017). Auditing private companies: what do we know? *Accounting and Business Research*, 47(5):565–584. doi = 10.1080/00014788.2017.1314104.

Vergara, J. R. & Estévez, P. A. (2013). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186. doi = 10.1007/s0052101313680.

WRDS (n.d.a). Wharton wrds - dictionary. https://wrds-www.wharton.upenn.edu/documents/566/6481_to_6933.txt. [Online; accessed 8-December-2022].

WRDS (n.d.b). Wrds. https://wrds-www.wharton.upenn.edu/documents/541/Who_Must_Report_on_What_Forms.pdf. [Online; accessed 5-October-2022].

Ye, A. (2022). Understanding Entropy: the Golden Measurement of Machine Learning. https://towardsdatascience.com/understanding-entropy-the-golden-measurement-of-machine-learning-4ea97c663dc3.

Yıldırım, S. (2020). Understanding the lightgbm. https://towardsdatascience.com/understanding-the-lightgbm-772ca08aabfa.

Appendix

A1 Data Transformation

Table A1.1: All Filled-Out Information

Identifer	Name	Comment
RCFD1010	Cash	Missing values equal $RCON0071 + RCON0081$
RCFD1600	Commercial and Industrial loans	Using RCFD1766
RCFD1350	Federal Funds Sold	Missing values equal $RCONB987 + RCONB989$
RCON2011	Loans to Individuals	Data from after 2010 equal to $RCFDB538 + RCFDB539 + RCFDK137 + RCFDK207$
RCFD2143	Intangible Assets	Missing values equal $RCON3163 + RCON3164 + RCON5507$
RIAD4000	Operating Income	Missing values equal $RIAD4107 + RIAD4169 + RIAD4020 + RIAD4079$

A2 Mandatory Audits in States

We consider an external audit as mandatory if more than 95% of the banks in the state have employed an external audit. States with less than 25 banks are investigated separately.

Table A2.1: Identified State Mandatory Audits

	Number of Banks	Audited Banks	Ratio of Audited Banks	State Mandatory Audit
Alabama	153	128	83.66	No
Alaska	8	7	87.50	Yes
Arizona	42	39	92.86	No
Arkansas	134	102	76.12	No
California	292	272	93.15	No
Colorado	144	104	72.22	No
Connecticut	55	53	96.36	Yes
Delaware	43	36	83.72	No
Florida	278	244	87.77	No
Georgia	299	273	91.30	No
Hawaii	9	6	66.67	Yes
Idaho	18	17	94.44	Yes
Illinois	627	387	61.72	No
Indiana	147	133	90.48	No
Iowa	368	138	37.50	No
Kansas	338	98	28.99	No
Kentucky	197	173	87.82	No
Louisiana	156	148	94.87	No
Maine	30	30	100.00	Yes
Maryland	87	83	95.40	Yes
Massachusetts	173	165	95.38	Yes
Michigan	143	121	84.62	No
Minnesota	412	121	29.37	No
Mississippi	92	72	78.26	No
Missouri	348	172	49.43	No
Montana	73	26	35.62	No
Nebraska	232	50	21.55	No
Nevada	33	31	93.94	No
New Hampshire	26	24	92.31	No
New Jersey	118	112	94.92	No
New Mexico	53	49	92.45	No
New York	199	158	79.40	No
North Carolina	102	95	93.14	No
North Dakota	92	24	26.09	No
Ohio	240	197	82.08	No
Oklahoma	256	85	33.20	No
Oregon	37	35	94.59	No
Pennsylvania	225	215	95.56	Yes
Rhode Island	13	12	92.31	No
South Carolina	89	81	91.01	No
South Dakota	85	30	35.29	No
Tennessee	196	187	95.41	Yes
Texas	630	468	74.29	No
Utah	60	51	85.00	No
Vermont	14	13	92.86	No
Virginia	119	109	91.60	No
Washington	87	74	85.06	No
West Virginia	66	65	98.48	Yes
Wisconsin	282	120	42.55	No
Wyoming	37	27	72.97	No
District of Columbia	6	6	100.00	Yes

A3 Variable Definitions

Table A3.1: Variable Definitions

Name	Explanation	Formula
audited	1, if bank is audited 0, otherwise	
AG_loans	Agricultural loans scaled by total assets.	$\frac{RCFD1590}{RCFD2122}$
cash	Proxy for liquidity. Cash scaled by total assets.	$\frac{RCFD0010}{RCFD2170}$
CI_loans	Commercial and Industrial loans scaled by total loans.	$\frac{RCFD1600}{RCFD2122}$
fed_funds_sold	Federal funds sold scaled by total assets.	$\frac{RCFD1350}{RCFD2170}$
IN_loans	Loans to Individuals scaled by total loans.	$\frac{RCON2011}{RCFD2122}$
intangible_assets	Intangible assets scaled by total assets.	$\frac{RCFD2143}{RCFD2170}$
LFR	Lease Financing Receivables scaled by total loans.	$\frac{RCFD2165}{RCFD2122}$
LLA	Loan Loss Allowance scaled by total loans.	$\frac{RCFD3123}{RCFD2122}$
LLP	Loan Loss Provision scaled by total assets.	$\frac{RIAD4230}{RCFD2170}$

Name	Explanation	Formula
prem_fixed_assets	Premises and Fixed Assets scaled by total assets.	$\frac{RCFD2145}{RCFD2170}$
RE_loans	Real estate loans scaled by total loans.	$\frac{RCFD1415}{RCFD2122}$
ROA	Return on assets. Total operating income scaled by total assets.	$\frac{RIAD4000}{RCFD2170}$
size	Proxy for size.	$ln(RCFD2170)$
tier1_ratio	Tier 1 capital ratio.	$\frac{RCFD8274}{RCFDA223}$
tier2_ratio	Tier 2 capital ratio.	$\frac{RCFD8275}{RCFDA223}$
total_loans	Total loans scaled by total assets.	$\frac{RCFD2122}{RCFD2170}$

A4 Descriptive Statistics

Table A4.1: Descriptive Statistics for Explanatory Variables

Statistic	N	Mean	St. Dev.	Min	Max
AG_loans	179,455	0.092	0.134	0.000	0.576
cash	179,455	0.102	0.087	0.011	0.455
CI_loans	179,455	0.128	0.088	0.000	0.455
fed_funds_sold	179,455	0.018	0.035	0.000	0.189
IN_loans	179,455	0.058	0.064	0.000	0.363
intangible_assets	179,455	0.003	0.007	0.000	0.043
LFR	179,455	0.002	0.010	0.000	0.078
LLA	179,455	0.016	0.009	0.003	0.056
LLP	179,455	0.172	0.138	0.000	1.000
prem_fixed_assets	179,455	0.017	0.013	0.0004	0.064
RE_loans	179,455	0.703	0.190	0.137	0.998
ROA	179,455	0.321	0.221	0.000	1.000
size	179,455	0.601	0.234	0.000	1.000
tier1_ratio	179,455	0.185	0.087	0.058	0.600
tier2_ratio	179,455	0.011	0.002	0.003	0.013
total_loans	179,455	0.614	0.162	0.171	0.903

A5 Histograms

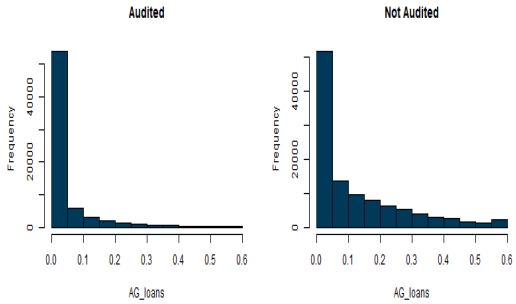

Figure A5.1: AG_loans

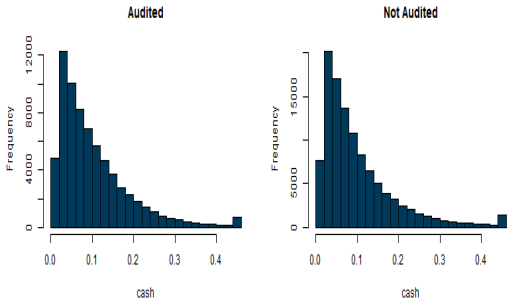

Figure A5.2: cash

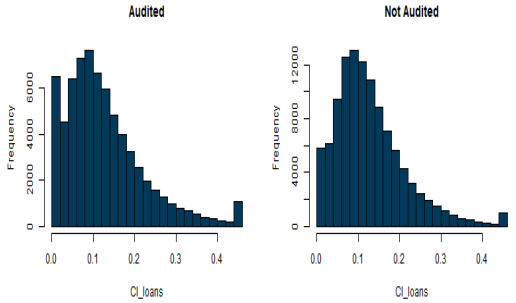

Figure A5.3: CI_loans

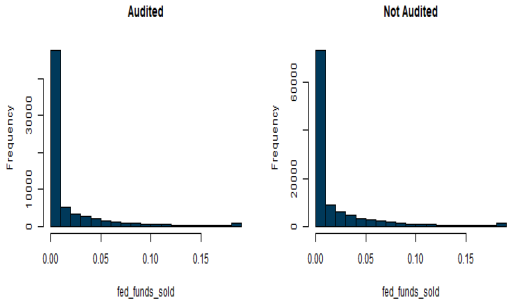

Figure A5.4: fed_funds_sold

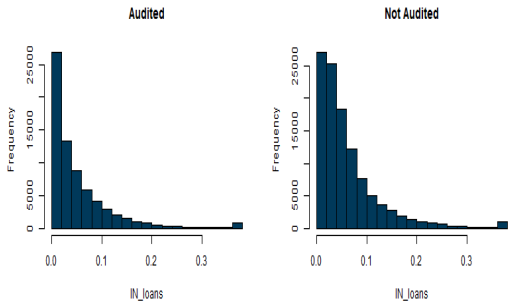

Figure A5.5: IN_loans

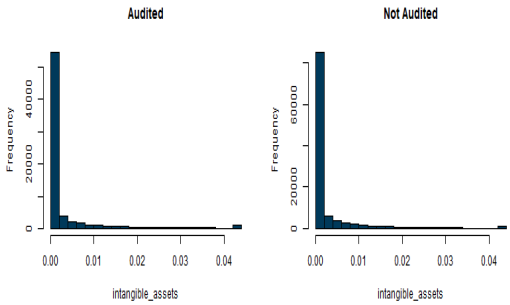

Figure A5.6: intangible_assets

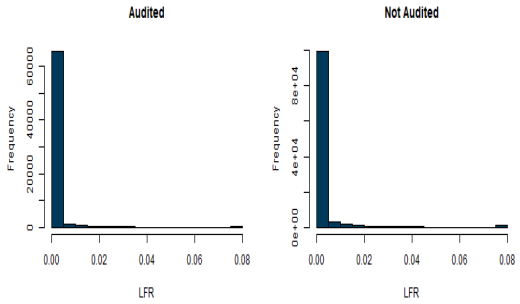

Figure A5.7: LFR

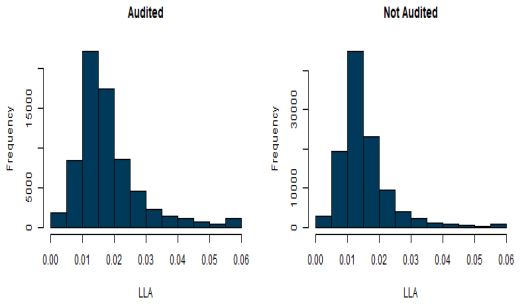

Figure A5.8: LLA

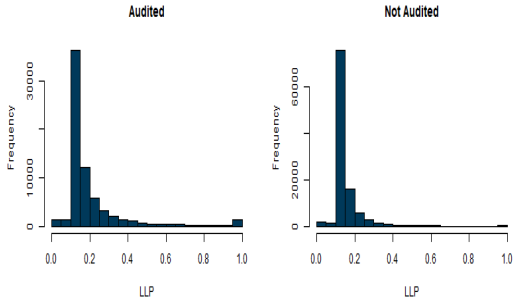

Figure A5.9: LLP

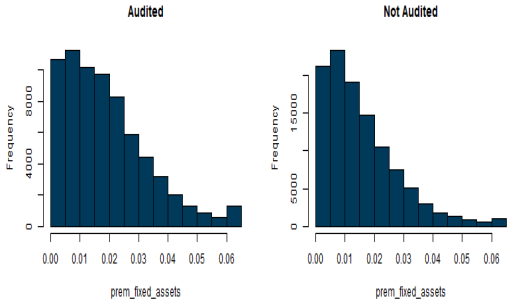

Figure A5.10: prem_fixed_assets

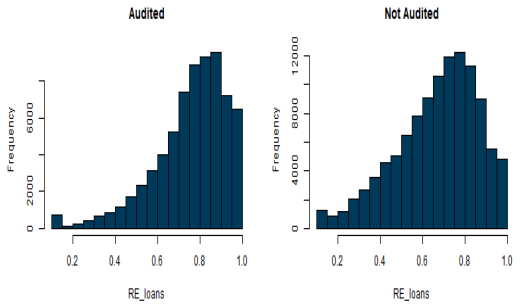

Figure A5.11: RE_loans

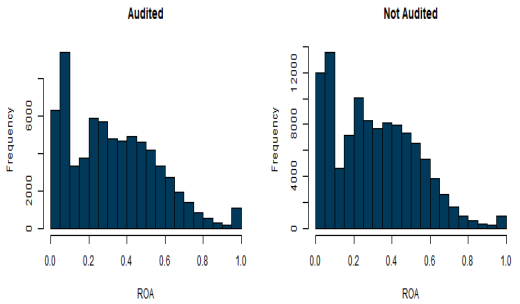

Figure A5.12: ROA

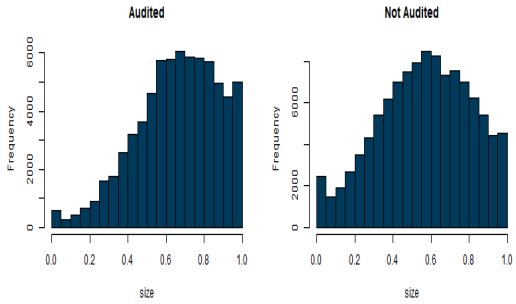

Figure A5.13: size

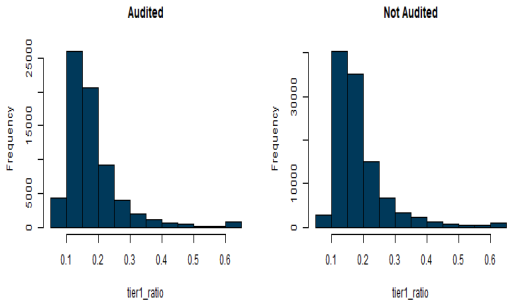

Figure A5.14: tier1_ratio

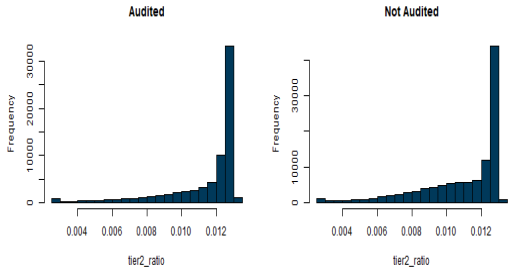

Figure A5.15: tier2_ratio

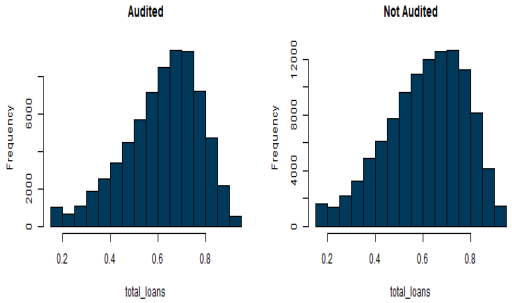

Figure A5.16: total_loans

A6 Performance of Machine Learning Models

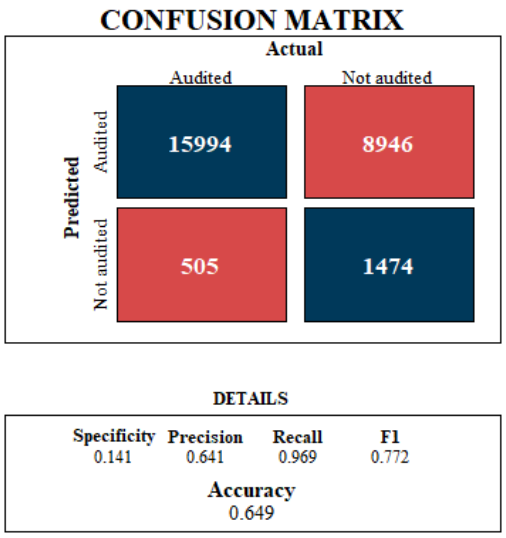

Figure A6.1: Logistic Regression

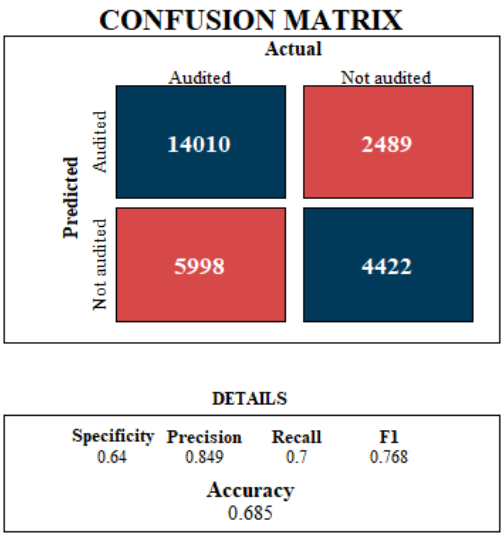

Figure A6.2: LASSO

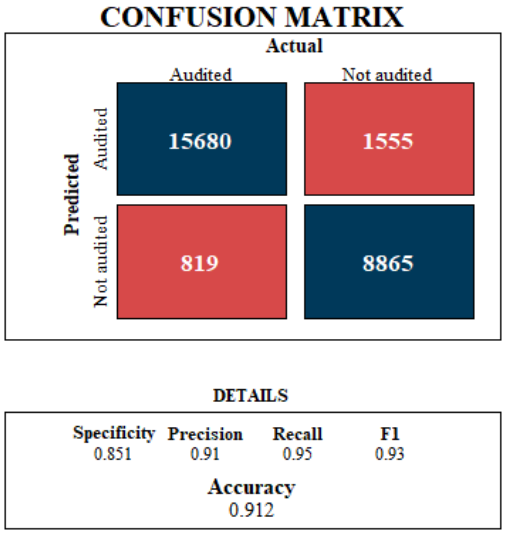

Figure A6.3: Random Forest

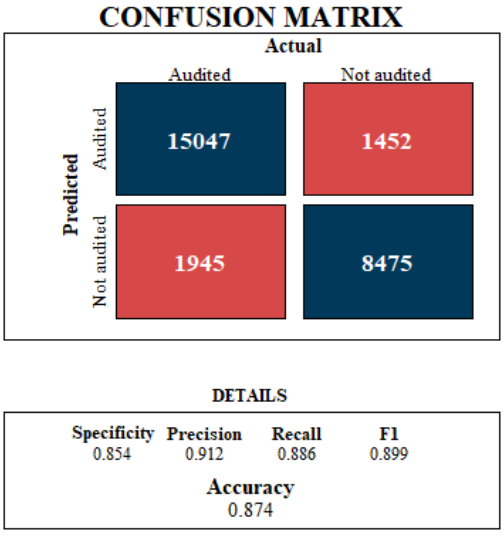

Figure A6.4: LightGBM

A7 LASSO Coefficients for $\lambda = [0.003 - 0.018]$

Table A7.1: Lasso Coefficients for λ Interval

	0.018	0.017	0.016	0.015	0.014	0.013	0.012	0.011	0.01	0.009	0.008	0.007	0.006	0.005	0.004	0.003
AG_loans	-4.083	-4.122	-4.161	-4.2	-4.23	-4.259	-4.29	-4.318	-4.366	-4.414	-4.463	-4.513	-4.564	-4.618	-4.702	-4.768
cash.	.	.	.	.	.	.	.	.	0.037	0.104	0.171	0.239	0.307	0.376	0.455	0.539
CI_loans.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-0.042	-0.123
fed_funds_sold.	.	.	.	.	.	.	.	0.095	0.278	0.475	0.672	0.871	1.071	1.272	1.474	1.692
IN_loans.	.	.	.	.	.	.	.	.	-0.09	-0.193	-0.296	-0.399	-0.504	-0.607	-0.725	-0.82
intangible_assets.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.422	1.25	2.158
LFR	-0.221	-0.645	-1.074	-1.505	-1.912	-2.317	-2.729	-3.147	-3.622	-4.103	-4.592	-5.087	-5.59	-6.101	-6.665	-7.176
LLA	19.086	19.541	19.999	20.461	20.79	21.101	21.417	21.717	21.913	22.074	22.24	22.411	22.586	22.795	22.769	22.61
LLP	1.03	1.052	1.075	1.098	1.132	1.168	1.204	1.241	1.28	1.321	1.362	1.404	1.447	1.489	1.53	1.573
prem_fixed_assets	2.627	2.856	3.086	3.316	3.596	3.885	4.175	4.468	4.789	5.125	5.464	5.804	6.147	6.476	6.655	6.789
RE_loans.	.	.	.	0.001	0.014	0.028	0.043	0.06	0.056	0.049	0.042	0.035	0.028	0.022	.	.
ROA.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
size	0.851	0.867	0.884	0.901	0.922	0.944	0.966	0.991	1.017	1.047	1.076	1.106	1.136	1.166	1.189	1.212
tier1_ratio.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-0.119	-0.26
tier2_ratio.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.449	3.993
total_loans.	.	.	.	.	-0.041	-0.088	-0.135	-0.181	-0.229	-0.275	-0.321	-0.368	-0.415	-0.462	-0.545	-0.639

A8 Accuracies for LASSO Models for Different λ

Table A8.1: Accuracy for λ Interval

λ	Accuracy
0.003	0.6782
0.004	0.6784
0.005	0.6787
0.006	0.6793
0.007	0.6799
0.008	0.6803
0.009	0.6808
0.010	0.6810
0.011	0.6817
0.012	0.6824
0.013	0.6829
0.014	0.6834
0.015	0.6842
0.016	0.6846
0.017	0.6848
0.018	0.6844

A9 Random Forest Importance

Table A9.1: Random Forest Importances per Class and Means

Feature	0	1	MeanDecAccuracy	MeanDecGini
AG_loans	210.61	185.74	231.83	8485.53
cash	154.94	142.01	163.70	3939.04
CI_loans	149.74	150.02	169.82	4525.27
fed_funds_sold	142.98	148.83	162.77	2388.01
IN_loans	140.72	136.37	145.76	5535.91
intangible_assets	126.23	138.44	139.69	2521.56
LFR	60.63	70.21	66.10	1237.59
LLA	145.85	134.82	151.94	5236.30
LLP	177.18	156.80	178.57	4550.30
prem_fixed_assets	125.84	112.47	122.29	5337.34
RE_loans	81.80	82.06	86.13	5716.95
ROA	60.78	21.20	62.57	2495.64
size	166.66	154.30	167.04	6468.02
tier1_ratio	202.85	165.72	208.69	4829.06
tier2_ratio	98.05	77.43	94.03	4371.23
total_loans	207.36	176.45	213.11	4808.08

A10 LightGBM Importance

Table A10.1: All Measures of LightGBM Importance

Feature	Gain	Cover	Frequency
AG_loans	0.15	0.07	0.07
size	0.10	0.09	0.10
IN_loans	0.08	0.08	0.09
prem_fixed_assets	0.08	0.08	0.09
LLA	0.07	0.06	0.07
tier1_ratio	0.07	0.07	0.08
total_loans	0.07	0.07	0.07
RE_loans	0.06	0.08	0.07
CI_loans	0.06	0.07	0.07
intangible_assets	0.05	0.06	0.05
tier2_ratio	0.05	0.06	0.06
cash	0.05	0.06	0.06
LLP	0.05	0.04	0.05
fed_funds_sold	0.03	0.04	0.04
LFR	0.02	0.03	0.02
ROA	0.02	0.04	0.03

A11 Mutual Information Results

Table A11.1: Mutual Information Scores

Feature	Score
AG_loans	0.06069
RE_loans	0.02906
size	0.02252
LLP	0.02174
tier2_ratio	0.01637
LLA	0.01634
IN_loans	0.01400
prem_fixed_assets	0.00961
tier1_ratio	0.00638
LFR	0.00630

A12 ANOVA Results

Table A12.1: ANOVA F-scores and p-values

Feature	Score	p-value
AG_loans	13124.4239	0.000
RE_loans	7825.5882	0.000
size	5836.2487	0.000
LLP	3889.0310	0.000
LLA	3465.6344	0.000
prem_fixed_assets	2418.0886	0.000
tier2_ratio	1441.0866	0.000
IN_loans	647.4538	0.000
LFR	378.4408	0.000
tier1_ratio	211.2016	0.000
ROA	184.4517	0.000
total_loans	56.3852	0.000
CI_loans	9.1123	0.003
intangible_assets	5.1808	0.023
cash	4.4280	0.035
fed_funds_sold	3.8097	0.051

A13 Test Size

Table A13.1: Accuracies for Different Test Sizes

Test size	0.10	0.15	0.20	0.25	0.30
Logistic regression	0.641	0.649	0.648	0.644	0.649
Lasso	0.679	0.685	0.686	0.679	0.684
Random Forest	0.916	0.912	0.908	0.901	0.900
LightGBM	0.876	0.874	0.866	0.850	0.849