129 w3

INCORPORATING EXPLAINABLE AI INTO THE AUDITING PRACTICES

Dr. Maria Mora-Rodriguez, Fujitsu Laboratories of Europe

Dr. Alicia Rodriguez-Carrión, Fujitsu Laboratories of Europe

Terunobu Kume, Fujitsu Limited

Akihiro Inomata, Fujitsu Limited

Toshimitsu Suzuki, Fujitsu Limited

ABSTRACT

The goal of this research is to evolve the auditing practices by the introduction of advanced Explainable AI (XAI) technologies for the prediction of auditor's opinion using financial statements and auditor's opinion reports from 12486 US-traded companies that publicly traded in Russell Global Index for the years 2005-2020. In this study, we evaluated sixteen models including traditional ones like linear regressions and demonstrated that CatBoost - a tree-based machine learning technique - offers the best results to predict the auditor's opinion. In addition, SHAP (Shapley Additive explanation) is utilized to interpret the results and analyze the importance of individual variables. The results show that CatBoost can predict auditor's opinion robustly with 92.75% of AUC and that the top ten variables that have a major influence in the probability to predict the auditor's opinions are: Working Capital to Total Assets, the size of the company, Operating cash flow ratio, Return on Equity (ROE), Accrued Expenses Turnover, Calculated Tax rate, Quick Ratio, Net Margin, Total Asset Turnover and Operating Cash flow to total assets.

Our results suggest that auditing practices may evolve through the adoption of AI tools that can be interpretable to support their judgment, and thus this type of research could open new approaches and opportunities not only for auditors, also sectors such as financial institutions or regulators in charge of auditing activities among others.

1. Introduction

With the 2008 global financial crisis, there has been an increased number of corporate bankruptcies at a global level, growing the spotlight on auditing practices and questioning the role of auditors. The financial crisis and crashes such as the Stock Market Crack of 1929, revealed deficiencies in audit procedures (Arnold, 2009). Since then, the quality of auditing practices on financial statements is not just a regulatory demand, but also a business requirement for companies to make themselves creditable and attractive to shareholders to invest money in their corporations (Leuz, 2010). The role of auditors became more difficult as businesses and economies grew, but became better supported after a major regulatory enhancement on internal controls, accounting standards, and auditing principles, as well as technology standards for financial reporting such as eXtensible Business Reporting Language (XBRL) (Cong *et al.*, 2019).

Still one of the biggest challenges in audit is the cost. The audit fees have been steadily increasing for the last 5 years, mostly associated with the fact that few audit firms control the major part of the market, and the human labor time required to provide enough certainty that the company's financial statements are not materially misstated (AccountancyAge, 2019). It is recognized in the literature that auditing firms have lagged on technology adoption in the past, and it is demanding partial automation due to its labor intensiveness and range of decision structures (Oldhouser, 2016). However, there is certain reticence for technology adoption because it threatens the auditor's core business: audits are sold as a service, where the auditors are selling their reputation, previous knowledge about the client, and the quality of their work (Pham *et al.*, 2017). It is the relationship between the client and auditing firm that makes the difference and convinces companies on who to hire. If all audits become completely automated, there will be no difference from one firm to another, and companies will hire the auditors that present the lowest price. Nevertheless, there is room for technology adoption, as long as it helps on clients' interaction and relationships meanwhile increases the certainty of the test, or provides the same certainty at a lesser cost than before (Byrnes *et al*., 2018).

Besides the audit business model barriers, there are more obstacles to facilitate the integration of technologies in audit practices. The accessibility of the data is one of them, in part due to the large complex datasets that companies are handling, and also to the data protection legal requirements that auditors will need to comply with and get approval. In many cases, companies are denied to give access to certain data arguing security concerns (Ernest and Young, 2015). Another challenge is the variety of accounting systems and other systems within the same company that auditors need to figure out how to handle (Oldhouser 2016). Barriers like these are making the auditing work time-consuming, less efficient, and more expensive.

Audits are performed throughout the year, but the most labor-intensive period occurs soon after the client's year-end. It is during this period when auditors analyze and test the year-end financial statements, which requires a deep analysis of a large number of transactions. If the auditor takes the company results at December 31st year-end, and performs the test for a month or two, it will potentially be able to issue an opinion by late March. Therefore, this means that an investor analyzing a company's financial statements to evaluate investment potential and the company's health cannot consider the auditor's opinion until the last quarter of the current year is already completed.

How to reduce the time and increase the effectiveness during the audit process, while finding the balance with the current audit core business, are key to facilitate successful market penetration of advanced technologies in auditing practices. In our research, we consider this balance a key goal, not only for contributing to previous literature, but also to the industry level itself by helping to reduce the cost and to increase competency in the audit market.

The auditing profession is standard-driven, requiring standard-setting boards to approve their methodologies and procedures. The development of the profession and its regulation is influenced by three types of organizations (Humphrey *et al.,* 2009): (1) The International Federation of Accountants (IFAC), (2) International regulators comprising the World Bank, the International Organization of Securities Commissions (IOSCO), the International Association of Insurance Supervisors (IAIS), the Basel Committee on Banking Supervision (BCBS) and the European Commission (EC); and (3) the large multinational audit firms. The profession is facing challenges adjusting the current auditing standards to adopt disruptive technologies, whilst companies and organizations are continuously generating and collecting a larger amount of data from exogenous sources. This fact is making clearer that Artificial Intelligence (AI) technologies will move the audit profession a step forward, since traditional audit procedures are becoming less effective and efficient with complex data ecosystems (Dai and Vasarhelyi, 2016).

The traditional auditing procedures are effective when the size of the database is small, but became ineffective in today's client's data ecosystems (Teeter and Vasarhely, 2015). These procedures need to evolve to encourage auditors to take advantage of AI to provide a higher level of assurance on a more frequent basis by improving audit effectiveness through the integration of new supporting evidence. Auditors can take advantage of AI technologies to examine complete transactions in a much shorter time, being able to put their professional skills to better use on high-level tasks and focus their efforts on the interpretation of the results produced by AI.

Taking into account the complex data ecosystems needs, and in order to adapt the auditing procedures to consider this new scenario, we envision the human-in-the-loop approach (Holzinger, 2016) as part of the solution, where humans and algorithms are configured to transform data into value-added output. What we envision with our study is a future where auditors work with artificial intelligence in the same way they have adopted excel spreadsheets and other tools to improve their day-to-day work, but not to replace their work. Innovations to accelerate planning and execution procedures within the auditing process is our goal, as well as increasing the speed and quality of the audit procedures. An approach like this will require time for adoption, as well as new auditor education and scientific testing regimes for monitoring artificial intelligence reliability.

In this study, we wanted to focus on the reliability side of the AI, through the interpretation of the results produced by the AI applied for the prediction of the auditor's opinion. This is known as AI algorithms that essentially can generate explanations understandable by humans, called Explainable Artificial Intelligence (XAI) with the main goal to help shed some much-needed light on AI algorithms' black box.

2. The role of interpretability within the auditing profession - Auditor's opinion prediction

In the literature, many techniques have been used to detect auditor's opinion from classical statistical modeling technique such as logistic regression models (Zdolšek *et al.*, 2015; Dopuch *et al.*,1987; Francis and Krishnan, 1999; Krishnan and Krishnan, 1996) and probit model (Francis and Krishnan, 1999), to the most recent studies using modern AI techniques such as decision trees (DT)(Saeedi, 2020), support vector machines (SVM) (Saif et al. 2012), neural networks (Sánchez-Serrano *et al.*, 2020, Gaganis *et al.*, 2007), K-nearest neighbors (k-NN) and rough sets (RS) (Saeedi, 2020).

To this end, annual financial reports and auditor's reports have been mostly considered to be the best suited datasets to detect and predict auditor's opinion, mainly because that is information publicly available, and well-structured and digitalized in some regulatory regimens.

Despite the benefits that these AI methods are offering within the auditing research, there is some clear weakness regarding the complexity of interpreting the results. AI model interpretation is becoming especially important in other domains such as accident detections (Parsa *et al.*, 2020) or train preference estimation (Hak Lee *et al.,* 2021), and thus several studies have started to take advantage of XAI models.

XAI is a research field that aims to make AI systems results more understandable to humans. The term was first coined in 2004 as explainable intelligence by Van Lent *et al*., 2004, to describe the ability of their system to explain the behavior of AI-controlled entities in simulation games. While the term is relatively new, the problem of explainability has existed since the mid-1970s in the area of expert systems (Xu et al., 2019). However, the academic interest in explainable technologies was reduced as the trend of AI research was shifted towards the application side in several industries. Nevertheless, the XAI topic has received attention again from academia and practitioners due to the current and future adoption across industries and its crucial impact in critical decision-making processes, in addition to the global investment forecast on AI field placed in 52.2 billion U.S. dollars by 2021 coming from 12 billion US in 2017 (International Data Corporation, 2018). Besides, the expected revenue growth from AI market worldwide from 480 billion U.S. dollars in 2017 to 2.59 trillion U.S. dollars by 2021 (Statista, 2018).

In the area of financial services, which is highly regulated, one of the important challenges of using AI is to provide well-argued reasons. For instance, in credit scoring, it is critical to provide the explanation of why a customer's loan application is denied, especially when the reasons behind that denial are the output from an opaque AI algorithm. Some credit agencies such as Equifax and Experian are working on novel research projects to generate automated reason and make AI-based credit scoring decisions more explainable and auditor-friendly (Equifax, 2018).

The XAI methods are helpful to answer questions like how a model behaves in general, or which variables drive the predictions and which ones are useless. Among the XAI methods most used for model interpretation are Local Interpretable Model-Agnostic Explanation (LIME) and SHapley Additive exPlanation (SHAP):

(i) Local Interpretable Model-Agnostic Explanation (LIME) by Ribeiro *et al.* (2016) is a technique that uses local surrogate models like linear regression or decision trees to explain individual predictions. LIME trains a surrogate model by generating a new data set from the data of interest. The way it generates the data set depends on the type of data. Currently LIME supports text, image, and tabular data. For example, for text and image data LIME generates the data set by randomly turning single words or pixels on or off. In the case of tabular data, LIME creates new samples by permuting each feature individually. The result of using LIME is what feature values influenced the prediction positively or negatively.

(ii) SHapley Additive exPlanation (SHAP) by Lundberg and Lee (2017) is a bit different. It bases the explanations on *shapely values,* which are a measure of the contribution of each feature in the model. This method does not only provide the contribution of each feature value as LIME, but in addition to that it provides the scale of these contributions.

Finding the right balance between accuracy and interpretability is key to choose the right method.

The following sections are organized as follows. In the methodology section, the sample data is detailed, as well as the list of variables utilized in the study. The prediction model (CatBoost) and the interpretation model (SHAP) algorithms are described in-depth. In the results section, the performance of several models is evaluated to predict the auditor's opinion, and it is demonstrated that CatBoost is the one that performs the best and finally, the interpretation of the results of the AI model are analyzed and discussed through a comprehensive feature interpretation and feature dependency study provided through SHAP.

3. Methodology

3.1. Sample selection

In the process of selecting and obtaining the sample of companies, the financial database Mergent Online was used to access both datasets we needed: the financial statements and also the auditor´s opinion. In this database, nearly 16000 companies are compiled. The sample of the study consists of US-traded companies that publicly traded in Russell Global Index for the years

2005-2020. We excluded the financial companies (i.e. banks, insurance, real estate companies) as done in the audit literature because their business model is highly different from other companies. Additionally, individual financial reports were also eliminated from the sample due to possible duplications, and only reports categorized from active companies (condition coinciding with the category 'active company' in the Mergent database) were selected.

After these exclusions, the sample size was reduced to 133478 company-year observations of 12486 firms that contained 4324 qualified opinions company-year observations and 129154 unqualified opinions. The auditor's opinion data is unbalanced as it contains only 3.2% of qualified opinions (Figure 1).

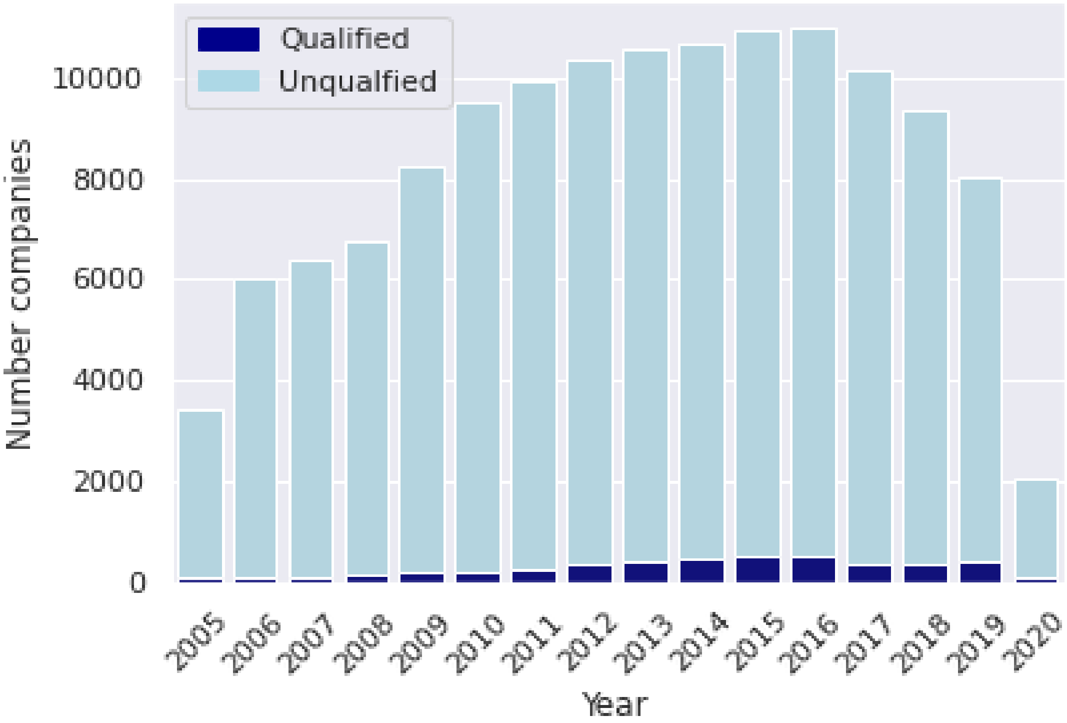

Figure 1-Sample distribution - qualified and unqualified

This acute unbalance of the dataset might hurt the model's learning capacity. In order to overcome this issue, two methods were considered. The first one is based on oversampling the minority class (unqualified opinions) using Synthetic Minority Oversampling Technique (SMOTE) (Nitesh *et al.,* 2002) and randomly undersampling the majority class. The combination of both oversampling and undersampling allows reducing the number of synthetic samples drawn from SMOTE that are needed to compensate for the high number of samples of the negative class, which is smaller after undersampling. The problem arising with this method is to find a good balance: not undersampling too much to keep as many real samples as possible (avoid losing much real information), whereas not oversampling too much to avoid creating too many synthetic samples that, although being close to the real samples, are just estimates that could lead to poor generalization of the model.

Therefore, we consider a second method referred as sample weights (De Prado, 2018), which keeps the data unbalanced as it is, and applies a weight to each sample when fitting the model.

These weights operate over the loss function to optimize when training the algorithms, giving more importance to the samples with higher weights, and lower importance to the samples with a lower weight. Using the balanced option, i.e., assigning weights inversely proportional to class frequencies to each sample, we get a higher weight for the minority class that compensates for its low frequency. This method prevents discarding real data and synthesizing new data.

The sectoral grouping has been carried out by categories according to the Global Industry Classification Standard (GICS) (2018) developed by MSCI and S&P Dow Jones Indices which allowed to distinguish 9 basic groups. The summary of the resulting sectoral distribution of the sample is presented in Table 1, where it can be seen the specific weight of each sectoral category.

GICS (Basic group)	ACTIVITY	COMPANIES	%
10	Energy	681	5.45%
15	Materials	1902	15.23%
20	Industrials	2535	20.30%
25	Consumer Discretionary	2402	19.24%
30	Consumer Staples Sector	648	5.19%
35	Health Care	1429	11.44%
45	Information Technology	1726	13.82%
50	Communication Services	768	6.15%
55	Utilities	395	3.16%
Total companies (qualified + unqualified)		12486	100.00%

Table 1 - Sample distribution per activity classification

To test the predictive capacity of the final model and its degree of generalization, 20% of the observations are selected using the stratified Random Sampling technique (Ding *et al.*, 1996) to respect the proportion of qualified and unqualified data, whereas the remaining 80% will be used to train and adjust the model.

3.2. Variables

One of the most relevant aspects in the development of prediction models for audit qualification is to determine the independent variables that will integrate it. In our case, they are mainly economic-financial ratios used in various research studies and usually statistically significant (Campillo *et al.*, 2019; Fernández-Gámez *et al.* 2016; Sánchez-Serrano *et al.*, 2020; Son *et al.*, 2019; Stanišić *et al.*, 2019; Zdolšek *et al.* 2015), and company profile information to facilitate the interpretation of the results. The economic-financial ratios are selected from the ratios of liquidity, solvency, profitability, and financial structure. All the items making up the ratios are derived from the balance sheet, cash flow, and income statements. The company's profile information was included through the size of the company and one categorical variable, the sector in which the company operates, considering nine categories corresponding to GICS classification. The dependent variable in the analysis is the auditor’s opinion, which is equal to “1” if the firm receives a qualified opinion and “0” unqualified. The list of variables is presented in Table 2.

	AUDITOR’S OPINION (Y)	
CODE	NAME	CONTENT
Y	Type of Audit opinions	1 Qualified , 0 Unqualified
	COMPANY PROFILE (X)	
CODE	NAME	CONTENT
C1	Sector	Global Industry Classification (GICS) 9 Sectors Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health care, Information Technology, Communication and Utilities.
C2	Size	Log(Total Assets)
	PROFITABILITY (X)	
CODE	NAME	CONTENT
P1	Gross Profit Margin	(Gross Profit / Revenue) * 100
P2	EBITDA Margin	(EBITDA / Revenue) * 100
P3	Operating Margin	(Operating Income / Revenue) * 100
P4	Net Margin	Net Margin: (Net Income / Revenue) * 100
P5	Cash Flow Margin	Cash Flow from Operations/Revenue
P6	Return on Assets (ROA)	(Net Income / Average Total Assets Over Period) * 100
P7	Return on Equity (ROE)	(Net Income / Average Stockholders' Equity Over Period) * 100
P8	Return on Investment (ROI)	(Operating Income / Average Invested Capital Over Period) * 100
P9	Retained earnings to total assets (RETA)	Retained Earnings/Total Assets Retained Earnings noted down as Stockholders’ or Shareholders’ Equity
P10	EBIT margin (EBITA)	EBIT/Total Assets
P11	Calculated Tax Rate	(Income Taxes / Earnings Before Taxes) * 100
P12	Property, Plant, & Equipment (PPE) Turnover	Annualized Revenue / Average PPE - net
P13	Cash & Equivalents Turnover	Annualized Revenue / Average Cash & Equivalents
P14	Accounts Payable Turnover	Annualized Revenue / Average Accounts Payable
P15	Accrued Expenses Turnover	Anrnualized Revenue / Average Accrued Expenses
	SOLVENCY / FINANCIAL STRUCTURE (X)	
CODE	NAME	CONTENT
S1	Total Debt to Equity	Total Debt & Leases / Shareholders' Equity
S2	Debt to assets	Total Liabilities / Total assets
S3	Interest Coverage	Operating Income / (0 - Non-Operating Net Interest Income)

LIQUIDITY (X)		
CODE	NAME	CONTENT
L1	Current Ratio	Current Assets / Current Liabilities
L2	Quick ratio	Quick Assets / Current Liabilities
L3	Operating cash flow ratio	Cash From Operations/Current liabilities
L4	Operating cash flow to total assets	Cash From Operations/Total assets
L5	Receivables Turnover	Annualized Revenue / Average Receivables - net
L6	Inventory Turnover	Annualized Direct Exp excl deprec / Average Inventories or Annualized Revenue / Average Inventories if Direct Expenses are not available
L7	Working capital to total assets	Working Capital/Total Assets Working Capital = Current Assets – Current Liabilities
L8	Total Asset Turnover	Annualized Revenue / Average Total Assets

Table 2-List of variables

3.3. Prediction Model - CatBoost

CatBoost (Prokhorenkova 2018) is a gradient boosting-based machine learning algorithm that outperforms existing state-of-the-art implementations of gradient boosted decision trees such as XGBoost (Chen 2016) and LightGBM (Ke 2017). Gradient boosting is, basically, a process of constructing an ensemble predictor by performing gradient descent optimization in a functional space. In other words, it iteratively builds a sequence of functions that gradually approximate better the function *F* that models the output variable *y* given a set of data *X* (*y* = *F(X)*), this is, minimize the difference between the functions constructed and the real one using gradient descent method.

CatBoost's good performance stems from the strong predictor it builds by iteratively combining weaker models, specifically, binary decision trees (Breiman 1984) in a greedy manner. Catboost addresses some shortcomings of other gradient boosting implementation through three novelties over the classical gradient boosting with decision trees:

- Using ordered boosting to overcome the overfitting caused by what the authors call prediction shift in gradient boosting. Prediction shift is a special case of target leakage, in which gradients used at each step of the boosting process are built using target values of already seen data. To avoid this, ordered boosting is proposed. This way, at each step of the boosting process, the current model is applied to new unseen training samples, thus avoiding overfitting.
- A new algorithm for processing categorical features (variables that are not numeric, but which discrete values correspond to different categories), especially effective for highly dimensional categorical features. This algorithm relies on dealing with each categorical feature by replacing each of its values for each sample with a numeric value equal to the expected target given the category to be replaced. To improve the method even further, the authors apply again an ordering approach similar to the ordered boosting mentioned above. In global terms, this algorithm allows increasing

the predictive power without increasing the overfitting many times caused by categorical features.

- Using oblivious trees (Feroy 2016) for faster execution. In oblivious trees, the same splitting criterion is used across each different level of the tree, thus providing balanced trees, which are less prone to overfitting and faster to execute.

With these novelties with respect to other gradient boosting algorithms, CatBoost is able to obtain better results than other state-of-the-art models in a wide variety of tasks (Prokhorenkova 2018). In Section 4 we demonstrate that CatBoost also performs best in the problem considered in this work.

3.4. Model interpretation – SHAP

SHAP, an important tool in Explainable AI, is a method proposed by Lundberg and Lee (2017) that helps users interpret predictions. It arose from the need to find the balance between the predictive performance and interpretability of model predictions, and it is a tool to interpret the results and analyze the importance of individual variables.

SHAP is based on calculating Shapley values (Shapley 1953), which reflect the importance of a variable by comparing what a model predicts with and without this variable. However, the order in which a model considers variables affect the predictions. Thus, this process is done in every possible order, so that the variables are fairly compared.

The way to calculate the Shapley value for a certain variable *i* (out of n total variables), given a prediction *p* is determined through:

$$Importance\ of\ i = p(with\ i) - p(without\ i)$$

The contribution of each variable *i* is the weighted average of all possible differences of the predictions with and without variable *i*, considering all possible subsets of variables to provide a fair comparison. This is, let *F* be the complete set of variables without variable *i*, *S* is each subset of *N* variables out of the *n* total variables (|*F*|=*n*) without variable *i*. The contribution of each variable *i* is expressed as:

$$\phi(i) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (p(S \cup i) - p(S))$$

SHAP also produces dependence plots, which are an alternative to partial dependence plots (PDP) with the benefit to overcome the limitations of PDP to produce unrealistic results when the variables are correlated (Molnar 2019).

SHAP is a model agnostic algorithm, which means that can be applied to any AI model. This fact makes SHAP a good candidate to use in our case to interpret the results of CatBoost model by analyzing the importance of individual variables comprehensively, as SHAP is capable to show how much each explanatory variable contributes, either positively or negatively, to the target variable.

3.5. Model evaluation

Among the many metrics available for classification problems, in this work we will use the Area Under the Receiver Operating Characteristic (ROC) curve, abbreviated as AUC. The ROC is the curve that depicts the percentage of samples from the negative class versus the percentage of samples from the positive class, which for the random case corresponds to a diagonal line from bottom left to upper right, meaning that all samples are unirformly distributed. This ROC for the random dummy model corresponds to an AUC of 50%. As the model performs better, the curve takes off from this diagonal to approach the upper left corner, to with corresponds an AUC value of 100%, which will be the target to aim for.

The main reason to select AUC is the highly unbalanced nature of the dataset that represents our problem. Therefore, typical metrics such as accuracy are not well suited. Accuracy, recall, false-positive rate, and all other metrics derived from the confusion matrix rely on the threshold used to convert the probability returned by the model to a final decision. The usual threshold is 50%, meaning that if the probability predicted by the model is above 50%, the sample is classified as the positive class, whereas if the probability is below 50%, the sample is classified as the negative class. However, this threshold is not appropriate for unbalanced datasets, in which the threshold must be carefully calibrated to detect as many positive samples as possible whilst misclassifying as few negative samples as possible. Also, since almost all samples come from the negative class, if the model always predicts the negative class, the accuracy (number of correct predictions/number of total predictions) will be as high as the percentage of the negative class samples (97% in our case), thus giving the impression of a good model when, in reality, it is a dummy classifier that never detects the positive class.

AUC removes both the dependency on the aforementioned threshold, as well as the effect of highly imbalanced datasets inaccuracy, thus providing a more robust evaluation metric.

4 Results

Initially, we provide a summary of the sample data as an initial description of the data through the analysis of the descriptive statistics of all the variables used as explanatory variables (Table 3). All the values are expressed in ratios and size is shown without logarithm transformation to facilitate the reading of the results. It is out of the scope of this statistical description the variable representing the sector since it is qualitative instead of quantitative.

In terms of median, which is a good reference because it is not influenced by extreme values, we notice negative values and considerable differences between Qualified (Q) and Unqualified(U) observations at several variables, namely EBITDA Margin (P2), Operating Margin (P3), Net Margin (P4), Cash Flow Margin (P5), Return on Assets (ROA) (P6), Return on Equity (ROE) (P7), Return on Investment (ROI) (P8), EBIT margin (EBITA) (P10), Operating cash flow ratio (L3) and Operating cash flow to total assets (L4). The negative values can be understood since they belong to qualified observations, and also the amount of qualified samples is much lower than unqualified ones. Nevertheless, these results offer some direction regarding what factors could influence the prediction of auditor's opinions, and also it is reflecting that mainly profitability and liabilities variables are the ones impacting the most. As will be seen later, the resulting model will consider the majority of these variables among the most important ones.

In terms of completeness of the data, only three variables present over 50% of missing data, namely Calculated Tax Rate (P11), Accrued Expenses Turnover (P15), Interest Coverage (S3). The way to handle missing values has been with "Arbitrary Value Imputation" technique (Song and Shepperd, 2007), which consists of replacing all the missing values within a variable with an arbitrary value that should differ from the median, mean, and mode, and not within the normal values of the variable. For example, we use -1 if the distribution is positive.

		Median	
Var	Name	Qualified	Unqualified
C2	Size	5.51e+07	6.68e+08
P1	Gross Profit Margin	0.25	0.30
P2	EBITDA Margin	-0.41	0.09
P3	Operating Margin	-0.43	0.07
P4	Net Margin	-0.52	0.04
P5	Cash Flow Margin	-0.12	0.10
P6	Return on Assets (ROA)	-0.20	0.04
P7	Return on Equity (ROE)	-0.38	0.09
P8	Return on Investment (ROI)	-0.17	0.09
P9	Retained earnings to total assets (RETA)	0.52	0.49
P10	EBIT margin (EBITA)	-0.14	0.06
P11	Calculated Tax Rate	0.22	0.28
P12	Property, Plant, & Equipment (PPE) Turnover	2.16	4.09
P13	Cash & Equivalents Turnover	3.64	8.10
P14	Accounts Payable Turnover	5.61	11.51
P15	Accrued Expenses Turnover	8.64	19.70
S1	Total Debt to Equity	0.58	0.47
S2	Debt to assets	0.59	0.51
S3	Interest Coverage	4.14	10.08
L1	Current Ratio	1.14	1.66
L2	Quick ratio	0.65	1.08
L3	Operating cash flow ratio	-0.26	0.33
L4	Operating cash flow to total assets	-0.06	0.08
L5	Receivables Turnover	4.66	6.08
L6	Inventory Turnover	6.35	6.61
L7	Working capital to total assets	0.03	0.17
L8	Total Asset Turnover	0.34	0.84

Table 3. Descriptive statistics of all the variables used as explanatory variables.

4.1 Prediction

4.1.1 Model selection

In order to select the model that performs best for the prediction of the auditor's opinion, the selection process was carried out in two phases. To evaluate each phase separately and avoid data leakage and overfitting of the final model, the total data sample was split into two different subsets: a training dataset comprised of 80% of the data, and a test dataset comprised of the remaining 20% of the data. Both subsets were selected in a random stratified manner to preserve the ratio of positive (qualified) and negative (unqualified) samples in all of them.

In the first phase of the selection process, sixteen models were trained using their default configuration parameters (also known as hyperparameters) and evaluated using the performance metric presented in section 3.5, the AUC. To perform the training and evaluation of each model, a cross-validation approach was followed. This way, the training dataset was split into 5 stratified folds, 4 of which were used for training and 1 for evaluation, repeating this process 5 times for the 5 different combinations of folds for training and evaluation. The resulting metric is the mean of the AUC, which provides a robust estimation of the predictive power of each model. Table 4 presents the list of the models evaluated together with their corresponding AUC results**.**

Model family	Candidate model	AUC
Tree-based	Extra Trees Classifier	92.53%
Tree-based	Light Gradient Boosting Machine	92.54%
Tree-based	CatBoost Classifier	92.09%
Tree-based	Random Forest Classifier	91.97%
Tree-based	eXtreme Gradient Boosting	91.73%
Tree-based	Gradient Boosting Classifier	91.53%
Tree-based	Ada Boost Classifier	90.44%
Linear model	Linear Discriminant Analysis	87.53%
Linear model	Logistic Regression	87.46%
Bayes	Naïve Bayes	82.97%
Neural networks	MultiLayer Perceptron Classifier	81.16%
Clustering	K Neighbors Classifier	77.87%
Non-linear model	Quadratic Discriminant Analysis	66.62%
Tree-based	Decision Tree Classifier	64.49%
SVM	Support Vector Machine – Linear Kernel	00.00%
Linear model	Ridge Classifier	00.00%

Table 4-16 models evaluation for auditor's opinion prediction

Looking at these results, it is important to note that in general, tree-based models present better results than traditional ones like logistic regression and complex ones like neural networks.

In the second phase, we evaluated the top three models from the previous iteration and adjusted their hyperparameters very precisely to fit with the observations, process known as "fine-tuning". For doing the hyperparameter tuning, a random search was performed among 100 sets of different values of hyperparameters for each candidate model. Again, the approach followed was a 5 stratified fold cross-validation, using the training dataset. Table 5 shows the resulting mean and standard deviation of the AUC for the optimal set of hyperparameters for each model.

Candidate model with hyperparameters tuned	AUC
CatBoost Classifier	92.98%
Light Gradient Boosting Machine	92.71%
Extra Trees Classifier	91.18%

Table 5- Best models results with hyperparameters tuned

Although the AUC results of these three models are quite similar, CatBoost presents the best one with 92.98% AUC, therefore we decide to select this model to proceed with our study,

4.1.2 Predictive performance analysis

To study the predictive performance of the selected model (CatBoost) in the previous section, it is needed to re-train the model with the whole training set (no cross-validation), and evaluated with the set of unseen data: the test dataset containing 20% of the original dataset. Figure 2 shows the ROC of the model, to which corresponds an AUC of 92.75%. This value reflects the high predictive power of the model, regardless of the future threshold selected to make the classification. For instance, selecting a typical threshold of 0.5, it can be seen that the recall (ability to detect as many qualified reports as possible) is 30.17%, whereas the precision (ability to select as few unqualified reports as possible) is 62.59%. By reducing this threshold, it would be possible to achieve a higher recall (detect more qualified reports) at the cost of also detecting more not-qualified reports as qualified ones, thus decreasing the precision.

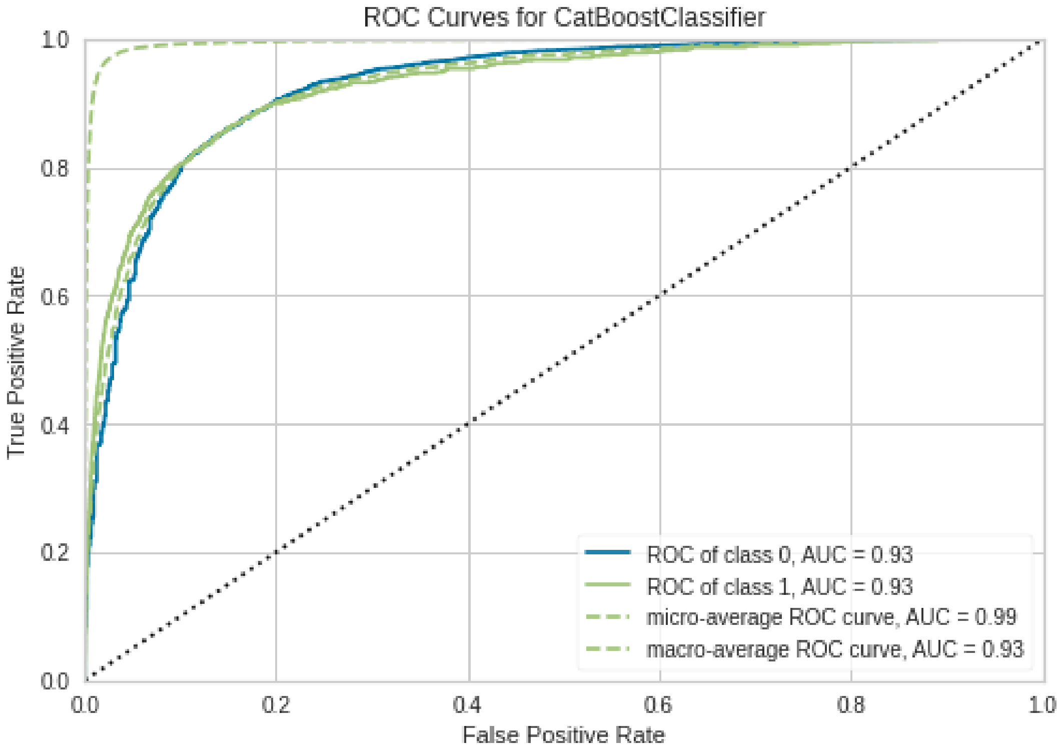

Figure 2- ROC results for CatBoost model to predict auditor's opinion

4.2 Interpretability

4.2.1 Feature importance analysis

For the interpretability of the predictive model (CatBoost), we analyze the results of the SHAP values plot (Figure 3), which express the contribution each variable has on the output of a model. In this plot, the horizontal axis presents the impact on qualified opinion prediction positively and negatively by increasing and decreasing the probability respectively. The horizontal axis shows the SHAP value that each observation has for each specific variable (y-axis), representing with colors the value of the variable, red for high values of the variable, and blue for low values. Therefore, we can see that the top ten variables that have a major influence on the probability to predict the auditor's opinions are: Working Capital to Total assets (L7) the size of the company (C2), Operating cash flow ratio (L3), Return on Equity (ROE) (P7), Accrued Expenses Turnover (P15), Calculated Tax rate (P11), Quick Ratio (L2), Net Margin (P4), Total Asset Turnover (L8) and Operating Cash flow to total assets (L4).

Looking at these top ten most influential variables, we can say that liquidity and profitability ratios have a major influential impact to predict the auditor's opinion, whereas solvency ratios are less influential.

The Working Capital to Total assets (L7) is the most important feature in the model, the SHAP values are pointing out that the lower this ratio the higher the probability that the auditor's opinion might be qualified. The second most influential variable is the Size, indicating that the smaller the company, the higher the probability that the auditor's opinion might be qualified and the larger the company this probability decreases. In general, it can be appreciated that at higher values within financial ratios values, the lower the probability to be qualified opinion, with exception of certain debt ratios such as Total Debt to Equity (S1) and Inventory and Debt to Assets (S2), which makes coherent that higher debt ratios the higher the probability of qualified opinion aligned with Cano-Rodriguez *et al.*, (2015).

Despite the results of the statistical description presented in the previous section and the results of other academic studies (Bradshaw *et al.*, 2001; Sánchez-Serrano *et al.*, 2020; Climent-Serrano *et al.*, 2018), the EBITDA Margin (P2), Gross profit Margin (P1), Operating Margin(P3), and Total Debt to Equity (S1) play a less important role in the prediction of auditor's opinion, at least not within the top ten.

Also, it is important to note that the Sector of the company is almost not influencing at all the prediction, as the majority of the sectors are ranked at the bottom.

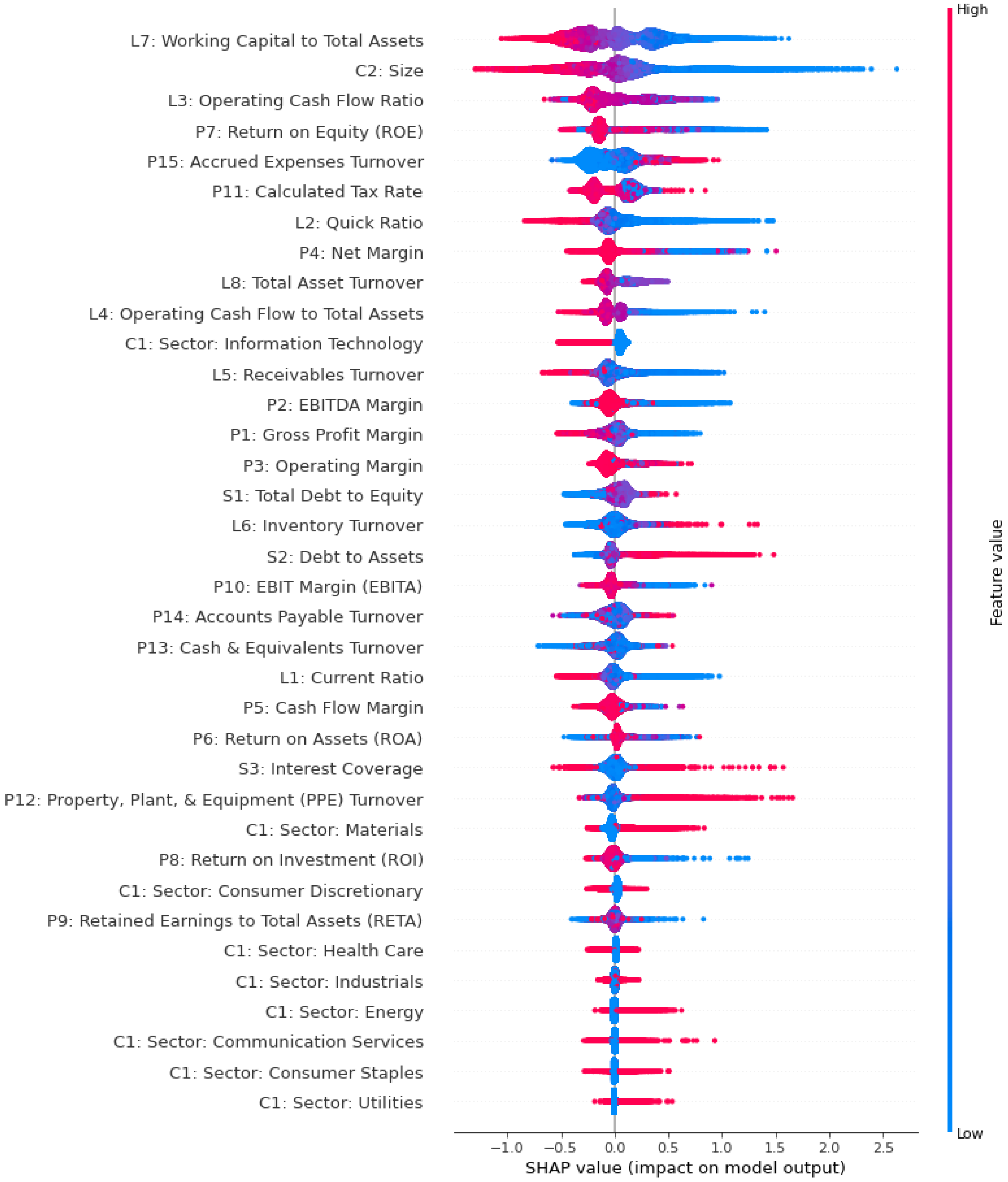

Figure 3- Results of SHAP values of all variables

4.2.2 Dependency analysis

To better comprehend what is behind the AI model, we analyze the relationship between certain main variables and their impact on the auditor's prediction model, this is called dependency analysis. Figure 4 presents the dependencies of three pairs of main variables that affect the probability of the auditor’s opinion. Each dot represents a row of the data, the x-axis is the actual value from the dataset, and the vertical location shows its impact on the probability of qualified opinion. In Figure 4(a), we select Operating Cash Flow to Total Assets (L4) as the variable to determine its impact when Net Margin (P4) increases. The red points represent higher values of P4 and the blue points represent the lower ones. It shows that the probability of qualified opinion (SHAP value) decreases when L4 and P4 increase. The probability of qualified opinion is positive until L4 increases to 0 then the probability to be qualified opinion becomes negative. In Figure 4(b), the Return of Equity (ROE) (P7) and Operating Cash Flow Ratio (L3) are selected as the variable to determine their impact on auditor’s opinion, which increase the probability of qualified when P7 are negative values and the majority of L3 values are also negative. Then when P7 becomes positive, L3 also increases over 0 in the majority of the cases, and the probability to be qualified opinion decreases. Finally, Figure 4(c), display the impact of the Size of the company (S2) and Calculated Tax Rate (P11) in the auditor’s opinion model. It shows a clear trend that bigger companies imply higher Calculated Tax Rate and a decrement in the probability of qualified opinion.

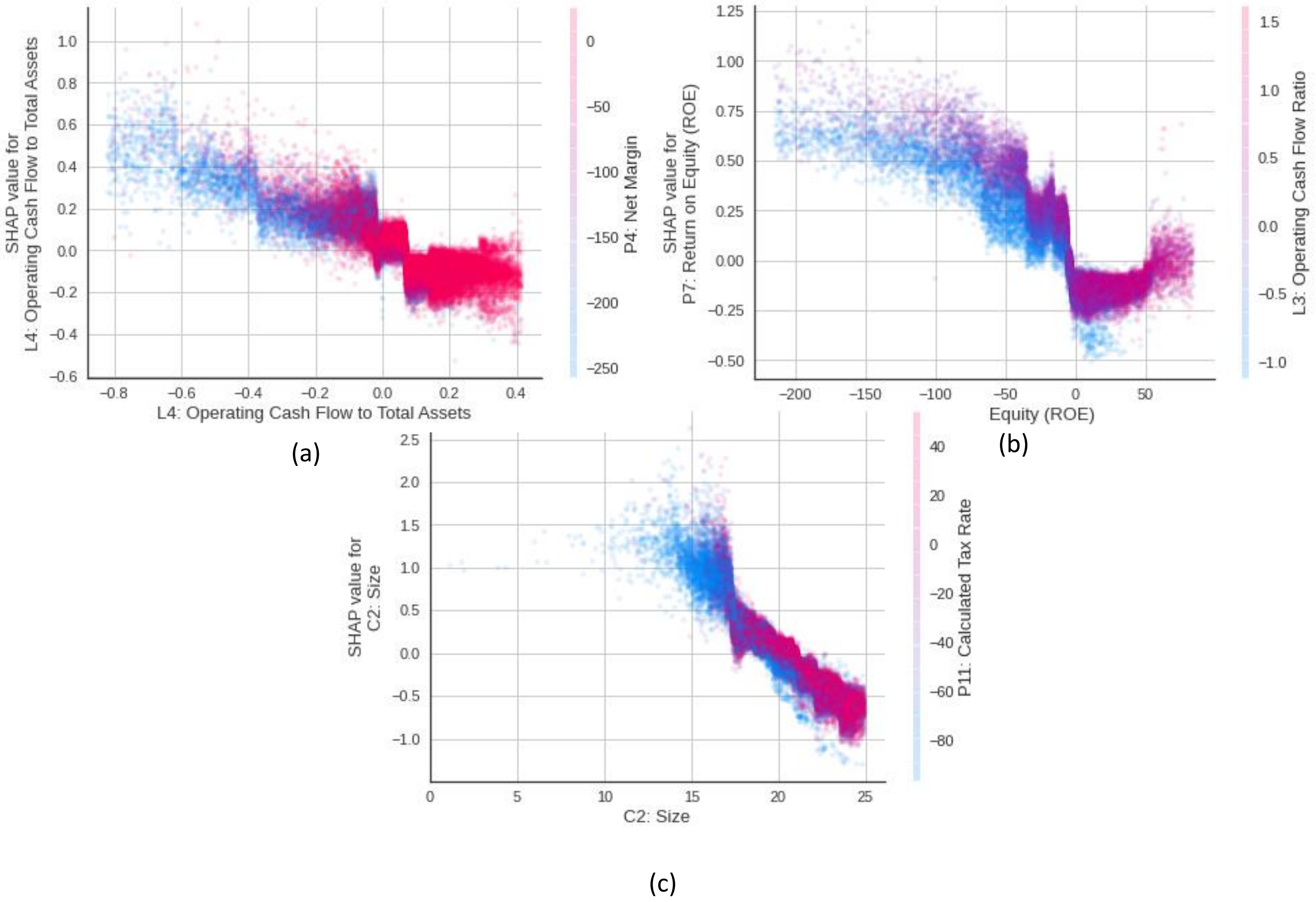

Figure 4 - Results of SHAPE dependency analysis:(a) Impact of Operating Cash Flow to Total Assets (L4) and Net Margin (P4) on model output, (b) Impact of Return on Equity (P7) and Operating Cash Flow Ratio (L3) on model output, (c) Impact of Size (C2) and Calculated Tax Rate (P11) on model output

5 Conclusions

Our study suggests that auditors may use advanced Explainable AI (XAI) technologies to take advantage of the interpretability potential of AI to plan specific auditing procedures to achieve an acceptable level of audit risk, as well as novel quality control tools for the auditing process.

During this study, sixteen algorithms were trained to predict the auditor’s opinion, being CatBoost the one that demonstrates better results. In total, 133478 firm-year observations of 12486 firms that contained 4324 qualified opinions firm-year observations and 129154 unqualified opinions from the years 2015-2020 are used to train the model.

In general, tree-based models perform better for the prediction of auditor’s opinion than traditional ones like linear regressions and complex ones like neural networks. CatBoost can predict auditor’s opinions robustly with 92.75% AUC, similar results were presented by Light Gradient Boosting Machine and Extra Trees Classifier with 92.71% and 91.18% respectively.

To facilitate the interpretability of the AI model results, feature importance analysis is applied to the selected model CatBoost, using SHAP. The results showed that especially liquidity and profitability ratios demonstrate a major influential power to predict the auditor’s opinion, whereas solvency ratios are less influential. Concretely these top ten ratios are Working Capital to Total assets (L7), the Size of the company (C2), Operating cash flow ratio (L3), Return on Equity (ROE) (P7), Accrued Expenses Turnover (P15), Calculated Tax rate (P11), Quick Ratio (L2), Net Margin (P4), Total Asset Turnover (L8) and Operating Cash flow to total assets (L4). In addition, to understand what is behind the AI model the dependency analysis was conducted using SHAP, and the impact of three pairs of variables of the model are captured and illustrated.

Future research could be directed towards various directions. First, in comparison to other academic studies, the EBITDA Margin (P2), Gross profit Margin (P1), Operating Margin(P3), and Total Debt to Equity (S1) play a less influential role in the prediction of auditor’s opinion, at least they are not placed within the top ten, which is an interesting point to continue this research. Second, we could not discriminate the reason for the qualified opinions due to the availability of the data that we were using. It would be interesting to develop AI models based on the reasons to be qualified, this could enrich the potential and usability of the auditor’s prediction. Finally, the inclusion of sustainability factors such as gender diversity, carbon emissions, or the number of board of directors could be also examined as a potential extension of the present research

The results obtained make it possible to think in the eminent practical application of AI tools within auditing practices that overcome the lack of transparency of AI models, commonly known as “black box AI”. We have demonstrated that complex AI models applied to auditor’s opinion prediction can be interpretable and therefore transparent to be applicable within the auditing practices and their decision tools. Sectors such as auditors, financial institutions, or regulators in charge of audit activities among others can apply these novels AI methods in their day-to-day to reduce errors and increase the efficiency in their production cycles. Knowing in advance the type of audit opinion would allow that, as well as, to design more specific work programs with a greater number of tests to expand the samples to be analyzed.

6 References

Arnold, P. J. (2009). Global financial crisis: The challenge to accounting research. Accounting, organizations and Society, 34(6-7), 803-809.

Accountancy Age (2019). What is the cost to do an audit? And how much time does it take to complete an audit? [Accessed: June 2021] https://www.accountancyage.com/2019/11/13/what-is-the-cost-to-do-an-audit-and-how-much-time-does-it-take-to-complete-an-audit/

Byrnes, P. E., Al-Awadhi, A., Gullvist, B., Brown-Liburd, H., Teeter, R., Warren, J. D., & Vasarhelyi, M. (2018). Evolution of Auditing: From the Traditional Approach to the Future Audit1. In Continuous Auditing. Emerald Publishing Limited.

Bradshaw, M. T., Richardson, S. A., & Sloan, R. G. (2001). Do analysts and auditors use information in accruals?. Journal of Accounting Research, 39(1), 45-74

Breiman, L., Friedman, J., Stone, C. J., and Olshen. R. A. (1984). Classification and regression trees. CRC press.

Campillo, J. P., Vargas, J. M., & Ibáñez, P. C. (2018). Análisis de la utilidad del algoritmo Gradient Boosting Machine (GBM) en la predicción del fracaso empresarial. Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad, 47(4), 507-532.

Cano-Rodríguez, M., Sánchez-Alegría, S., & Arenas-Torres, P. (2016). The influence of auditor's opinion and auditor's reputation on the cost of debt: evidence from private Spanish firmsLa influencia de la opinión de auditorãa y la reputación del auditor en el coste de la deuda: evidencia en las empresas españolas no cotizadas. Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad, 45(1), 32-62.

Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Climent-Serrano, S., Bustos-Contell, E., Labatut-Serer, G., & Rey-Martí, A. (2018). Low-cost trends in audit fees and their impact on service quality. Journal of Business Research, 89, 345-350

Cong, Y., Du, H., & Vasarhelyi, M. A. (2018). Technological disruption in accounting and auditing.

Dai, J., & Vasarhelyi, M. A. (2016). Imagineering Audit 4.0. Journal of Emerging Technologies in Accounting, 13(1), 1-15.

De Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.

Ding, C. S., Haieh, C. T., Wu, Q., & Pedram, M. (1996, November). Stratified random sampling for power estimation. In Proceedings of International Conference on Computer Aided Design (pp. 576-582). IEEE.

Equifax. (2018). Equifax Launches NeuroDecision Technology. [Accessed: June, 2021]. https://investor.equifax.com/ news-and-events/news/2018/03-26-2018-143044126

Ernest and Young, (2015). How big data and analytics are transforming the audit. EY Reporting. EY Reporting. [Accessed: June, 2021]. http://www.ey.com/gl/en/services/assurance/ey-reporting-issue-9-how-big-data-and- analytics-are-transforming-the-audit

Fernández-Gámez, M. A., García-Lagos, F., & Sánchez-Serrano, J. R. (2016). Integrating corporate governance and financial variables for the identification of qualified audit opinions with neural networks. Neural Computing and Applications, 27(5), 1427-1444.

Ferov, M. and Modry, M. (2016). Enhancing lambdaMART using oblivious trees. arXiv:1609.05610.

Global Industry Classification Standard (GICS) (2018) [Accessed June, 2021] Available at https://www.msci.com/gics

Hak Lee, E., Kim, K., Kho, S. Y., Kim, D. K., & Cho, S. H. (2021). Estimating Express Train Preference of Urban Railway Passengers Based on Extreme Gradient Boosting (XGBoost) using Smart Card Data. Transportation Research Record, 03611981211013349.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. Brain Informatics, 3(2), 119-131.

Humphrey, C., Loft, A., & Woods, M. (2009). The global audit profession and the international financial architecture: Understanding regulatory relationships at a time of financial crisis. Accounting, organizations and society, 34(6-7), 810-825.

International Data Corporation IDC. (2018). Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide. [Accessed: June, 2021]. https://www.idc.com/getdoc.jsp?containerId=prUS43662418

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 3149–3157.

Leuz, C. (2010). Different approaches to corporate reporting regulation: How jurisdictions differ and why. Accounting and business research, 40(3), 229-256.

Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1, 321–357.

Oldhouser, M. C. (2016). The effects of emerging technologies on data in auditing. Pham, N. K., Duong, H. N., Pham, T. Q., & Ho, N. T. T. (2017). Audit firm size, audit fee, audit reputation and audit quality: The case of listed companies in Vietnam. Asian Journal of Finance & Accounting, 9(1), 429-447.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 136, 105405.

Pham, N. K., Duong, H. N., Pham, T. Q., & Ho, N. T. T. (2017). Audit firm size, audit fee, audit reputation and audit quality: The case of listed companies in Vietnam. Asian Journal of Finance & Accounting, 9(1), 429-447.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), 6639-6649.

Ramlukan, R. (2015). How Big Data and analytics are transforming the audit. Ernst and Young Global Limited, April, 9, 9-12.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Saeedi, A. (2020). Audit Opinion Prediction: A Comparison of Data Mining Techniques. Journal of Emerging Technologies in Accounting.

Sánchez-Serrano, J. R., Alaminos, D., García-Lagos, F., & Callejón-Gil, A. M. (2020). Predicting Audit Opinion in Consolidated Financial Statements with Artificial Neural Networks. Mathematics, 8(8), 1288.

Shapley, L. S. (1953). A value for n-person games. In Contributions to the Theory of Games 2.28, 307–317

Son, H., Hyun, C., Phan, D. & Hwang, H.J. (2019). Data analytic approach for bankruptcy prediction. Expert Systems with Applications, 138, 112816

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. International journal of business intelligence and data mining, 2(3), 261-291.

Stanišić, N., Radojević, T., & Stanić, N. (2019). Predicting the type of auditor opinion: statistics, machine learning, or a combination of the two? *The European Journal of Applied Economics*, *16*(2), 1–58. https://doi.org/10.5937/EJAE16-21832

Statista. (2018). Revenues From the Artificial Intelligence (AI) Market Worldwide From 2016 to 2025. [Accessed: June, 2021]. https://www.statista.com/statistics/607716/worldwide-artificialintelligence-market-revenues/

Teeter, R. A., & Vasarhelyi, M. A. (2015). Audit analytics and continuous audit: Looking toward the future. New York, NY: AICPA Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the national conference on artificial intelligence (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press;

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the national conference on artificial intelligence (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing (pp. 563-574). Springer, Cham.

Zdolšek, D., Jagrič, T., & Odar, M. (2015). Identification of auditor's report qualifications: an empirical analysis for Slovenia. Economic research-Ekonomska istraživanja, 28(1), 994-1005.