Capitalization versus expensing of R&D costs under IAS 38: An empirical investigation using machine learning

Georgios A. Papanastasopoulos
Department of Business Administration
University of Piraeus
papanast@unipi.gr

John N. Sorros
Department of Business Administration
University of Piraeus
sorros@unipi.gr

Antonios M. Vasilatos*
Department of Business Administration
University of Piraeus
vasilatos@unipi.gr

Abstract: We study whether financial ratios or raw accounting data can achieve better out of sample predictions of R&D accounting treatment. We show that when the traditional logistic regression is used, financial ratios outperform raw accounting data. However, if machine learning algorithms are used, we provide evidence that raw accounting data may convey more information about future R&D accounting treatment. Finally, we introduce a Hybrid model, consisted of both ratios and raw data, which has achieved the best out of sample performance compared to all the examined models.

Keywords: R&D capitalization; Raw financial data; Financial ratios; Machine Learning

JEL Classifications: C53, M41

* Address for correspondence: Antonios M. Vasilatos, Department of Business Administration, University of Piraeus, Karaoli & Dimitriou 80, Piraeus, 18534, Greece. Email: vasilatos@unipi.gr

1. Introduction

The implementation of International Financial Reporting Standards (IFRS) in 2005 for all European public firms has led to many debates about the quality of the standards (Ahmed et al., 2013; Paananen & Lin, 2009; Soderstrom & Sun, 2007). In terms of R&D accounting treatment, debate focuses on IAS 38. IAS 38 dictates that research expenditures must be expensed while under circumstances, when certain criteria are met, development expenditures are required to be capitalized. On the contrary, in some countries, like United Kingdom or France, the National GAAPs leave to the discretion of the firm whether to capitalize or not.

There are two main theories about R&D capitalization. Those in favor of capitalization support that it allows the management to convey information about the R&D program future success and it acts as a signal to the market (Lev & Zarowin, 1999), while those against support that it is used for earnings management or as a way to slow down the amortization of failed R&D projects (Prencipe et al., 2008).

The focus of this study is twofold; Our first objective is to develop an R&D accounting choice prediction model out of sample by using financial statement data from European listed firms. The second objective is to discover if financial ratios or raw financial data from the financial statements can predict this accounting choice better. There is a large stream of accounting research on the determinants of R&D accounting treatment (Ball, 1980; Canace et al., 2022; Cazavan-Jeny et al., 2011; Dinh et al., 2016; Healy et al., 2002; Landry & Callimaci, 2003; Oswald, 2008; Wyatt, 2005).

In doing so, we extend the existing R&D accounting related literature which focuses on explaining accounting choice within sample, and thus emphasizing at causality. We approach the issue in a different way, by predicting the R&D accounting treatment out of sample. Thus, although the explanatory analysis is the mainstream approach in the literature,

we focus on the neglected importance of the prediction problems in business and economics (Kleinberg et al., 2015).

In the literature, models containing financial ratios as predictors, estimated with a logistic regression is the most common approach. These ratios are chosen by researchers or financial analysts based on economic theory. We use the model of Cazavan-Jeny et al. (2011) as a Benchmark and we obtain out of sample predictions using the logistic regression. Next, we use five models, all containing raw data items derived by the financial statements. We employ Cecchini et al. (2010), a replication of Cecchini et al. (2010) by Bao et al. (2020), the Dechow et al. (2011) model, a decomposition of the ratios included in Cazavan-Jeny et al. (2011) and finally a combination of all the raw data items from the financial statements we have available, as predictors of R&D accounting choice, using the logistic regression. It is impossible to know beforehand, whether ratios identified by experts and analysts are more powerful than raw data prepared by accountants. Existing theories about R&D determinants of accounting treatment may be incomplete, so ratios that have been constructed based on these theories may lack predictive power. On the other hand, raw data may convey useful information. There is no need to specify a specific structure, and raw data can be described by more complex algebraic forms of relationship (Bao et al., 2020).

To compare in depth these two approaches, ratios versus raw data, we use state of the art machine learning (ML) algorithms. We compare the benchmark with the best performing raw data set from the initial analysis (in which we have used the Logistic regression) and test if machine learning can extract any “hidden” information from the raw data. Our results indicate the following. Machine learning outperforms traditional econometric methods, and as well as raw data have greater predictive power than ratios. We also examine the performance of a Hybrid model, in which we mixed ratios with raw data. This model outperformed any other model in this research.

Our sample includes public firms from 15 countries from Europe, which share common economic characteristics and can be characterized as advanced economies. Our sample size is 18,957 firm-year observations of R&D engaged firms. We start our analysis in 2005, since that was the year IFRS became mandatory for all listed firms in most European countries.

Our research varies from the existing literature as we provide out of sample predictions of R&D accounting treatment. Moreover, we use machine learning algorithms, a methodology that yields better results than the traditionally used logistic regression. Additionally, following Bao et al. (2020) approach on using raw financial items in accounting research, we provide evidence that indeed raw data are informative and can outperform ratios.

This paper is structured in the following way. In Section 2, we make a brief introduction to machine learning. Section 3 describes the data. In Section 4, we discuss the way we evaluate the performance of the models. In Section 5, the empirical results are presented, while in Section 6, we conduct a sensitivity analysis. Finally, Section 7 offers concluding remarks.

2. An Introduction to Machine Learning

The dominant approach of creating econometric models is to specify an algebraic form of relationship between a dependent variable and its regressors. The relationship between the dependent variable and the regressors, thus the type of the functional form (e.g. simple linear, double log) is determined by the economic theory. The relationship is specified by a function of the form $y = f(b, x)$ where y is the dependent variable, x are the independent variables (regressors) and b their parameters (Anand et al., 2019). Econometrics focus to estimate parameters $\hat{\beta}$ by solving an optimization problem, either by minimizing a loss function (Mullainathan & Spiess, 2017) or by maximizing a likelihood function (MLE estimator).

On the other hand, a machine learning approach seeks to produce predictions by finding patterns in the data. ML produces a function, which is used to make predictions on y by using data x (Varian, 2014). Traditional econometric methods create predictions by relying on the estimated $\hat{\beta}$and use them in another sample, different than the one used for estimation, to produce out of sample predictions (Elliott & Timmermann, 2008). To sum up, as mentioned by Mullainathan and Spiess (2017), simply put, the difference between ML and econometrics is that ML focuses on $\hat{y}$ while the latter focuses on $\hat{\beta}$.

In the analyses that follow, we introduce the K-Nearest Neighbor (KNN) and Random Forest (RF) algorithms. The KNN is a non-parametric, supervised learning algorithm. The classification is made based on the majority vote of the object's neighbors, with the object being classified in the class of its nearest neighbors. On the other hand, RF is an ensemble learning method based on Decision Trees. The classification of an object is based on the majority vote of the individual trees in the forest.

3. The Sample and Data

Our sample size is 18,957 firm-year observations and includes publicly listed firms from 13 European Union countries (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain) plus Switzerland and the United Kingdom. These countries share some common characteristics. First, since 2005, they all have adopted IFRS, thus they treat R&D costs using the same accounting policies. In addition to that, most of these countries can be classified as advanced economies, yet they exhibit different levels of R&D activity.

We start the sample in 2005 since that was the year the adoption of IFRS became mandatory in Europe. Before the implementation of IFRS, R&D costs accounting treatment was different in each country. In general, there are two ways to treat R&D costs. Either they

are expensed as incurred, or they are capitalized as an intangible asset in the balance sheet. Both major accounting principles US GAAP (SFAS 72) and IFRS (IAS 38) dictate that R&D costs must be expensed. Their main difference is that under IAS 38, R&D activity is separated in two distinctive phases, that is the research phase and development phase. In the research phase, all occurred costs are expensed, like US GAAP. In the development phase, if the criteria for intangible asset recognition are met, then the costs must be capitalized. These criteria, in general, require that the firm proves that the asset will be completed and generate future economic benefit to the firm. In contrast, before the implementation of IFRS, local GAAP in countries like the UK, France, and Italy, allowed firm's management to decide whether costs in the development will be capitalized. The capitalization criteria were almost identical to current IAS 38 criteria for intangible assets recognition. The specific difference between IFRS and the local GAAP in force before IFRS adoption in 2005, is the limitation of management's discretion in the latter.

3.1. The R&D Sample

Our R&D sample has been retrieved by the WorldScope-Datastream database. We have followed the approach of Cazavan-Jeny et al. (2011) , and have removed from the sample firm-years that they did not develop any R&D activity, meaning that neither development costs (WC02505) nor expensed research costs (WC01201) were not reported at their financial statements. Firms classified as financials have been excluded from the sample as they follow different accounting principles. After having obtained the R&D firms sample, we have divided our firms into capitalizers and expensers. We classify a firm as a capitalizer if the firm reports capitalized development or as an expenser, if it only reports research expenditure (Cazavan-Jeny et al., 2011;Oswald & Zarowin, 2007; Oswald, 2008).

As shown in Figure 1, since the adoption of IFRS (IAS 38), there has been a constant upwards trend in the capitalization rate; That was expected since, in all cases where the standard's requirements were met, the capitalization became mandatory rather than optional. By 2013, capitalizing firms had surpassed the expensers. In general, we noticed a shift from expensing to capitalization. From 2014 to 2020, there is a relatively stable ratio between capitalizers and expensers (55%-45%).

[Insert Figure 1 here]

Our prediction models require a training and a test period. To ensure that our training sample takes under consideration all trends in the R&D accounting treatment, we use the period from 2005 to 2014 as the training sample, and the last six years of the sample as the test period. For the test year 2015, the training period is 2005-2014, for the test year 2016, the training period is 2005-2015, for the test year 2017 the training period is 2005-2016 and so on.

3.2. The Benchmark Model

Cazavan-Jeny et al. (2011) created their R&D choice determinants model, guided by theory. Based on White et al. (2002), they stated that capitalization improves leverage ratios and was used to smooth earnings. They stated that according to Aboody & Lev (1998) larger firms tended to expense more R&D costs compared to smaller firms, as well as that profitable firms avoided to capitalize R&D (Cazavan-Jeny & Jeanjean, 2006), and that managers used R&D capitalization to achieve smoother earnings (Lev et al., 2005;Penman, 1996). To examine these hypotheses, they used eight accounting ratios[1] (Size, ROA, CF_RD, DebtCap, CV _ROA, CV_CFRD and CAPEX). We have used their suggested model as a benchmark and compared it with other models containing raw data from the financial statements.

3.3. Raw Data

To obtain our initial raw data items, we have decomposed the seven accounting ratios from the Cazavan-Jeny et al. (2011) model. In this way we derived our initial nine raw items from the financial statements. In their groundbreaking and much discussed[2] article examining the detection of accounting fraud, Bao et al. (2020) used three sets of raw items derived from two detection fraud models. Initially, they used a list of raw financial data selected by Cecchini et al. (2010) who had reviewed the relative literature and selected 40 raw financial items used in fraud prediction. Cecchini et. al. included in their final sample only items that do not contain more than 25% missing values, by that way, they ended up to 23 raw items. Bao et al. (2020) followed the same procedure and in their replication of Cecchini et al. (2010), they created a sample containing 24 raw financial items.

In addition to that, Bao et al. (2020) chose 11 financial ratios used by Dechow et al. (2011) that could be calculated by 23 raw items derived from the financial statements. We

[1] Variable definitions are provided in Appendix A
[2] See: Bao et al. (2021);Walker (2021)

have followed their approach, and have created a list of raw items, which have been used to calculate the eight ratios of our benchmark model. Our initial list contained 10 raw items, all of which can be found in the financial statements. We have noticed that from these 10 items, three of them (Capitalized Development Costs, Amortized Development Costs and Research Expenses) have caused data leakage. This can be explained by the fact that our target variable (capitalizer or expenser) is coded (1 or 0) based on the values of these items. We have noticed that for Expensers (firms that do not report capitalized development costs), Development Costs and Amortized Development costs are reported as missing[3]. This creates a pattern which is leaked to the target variable and for these reasons, we decided to exclude these three items and use seven raw financial items instead.

In our approach, to obtain more sets of raw financial items, we use the raw items used by Bao et al. (2020) in their replication of Cecchini et al. and Dechow et al. models plus items from the original Cecchini et al. Our initial list of the available at the time raw financial items, contains 32 items, i.e., 17 items from the Statement of Financial Position, nine items from the Statement of Comprehensive Income, four items from the Cash-Flow Statement, two Market Value items and two items from the Disclosures. We use these items to replicate the Cecchini and the Dechow models. Even though the raw items used in Bao's et al. research are focused on the accounting fraud literature, they are important items used to calculate widely used ratios in the accounting literature.

[Insert Table 1 here]

[3] These values are not missing at random. There are firms that always expense and never capitalize their R&D costs. We attempted to fill the missing values of Capitalized Development Costs with zero, but this approach still creates a data leakage. For this reason, these items were excluded.

4. Evaluation of the Models

In the studies that use machine learning in accounting and finance, it is well established that financial data cannot be randomly split in train and test sets because they are intertemporal in nature (see: Bertomeu et al., 2021; Zhu et al., 2022). Therefore, as correctly highlighted by Bao et al. (2020), a k-fold cross validation is unsuitable[4] for our data, as in the random splits of the cross-validation, values from future data points may be used to predict values from past data points.

Anand et al. (2019), in their working paper, suggested the use of time-series cross validation. Data are split in k-folds, taking under consideration their intertemporal nature. We have followed their paradigm and have used the TimeSeriesSplit from the scikit-learn library in Python (Pedregosa et al., 2011), in order to split our data. In this type of cross-validation, the training data are partitioned in blocks of years. The algorithms are trained in the first block and are validated in the second block. Then the first and second block are trained, and they are validated in the third block. This is repeated for k-1 times, where k is the number of the folds. During the training phase, we use grid search to find the optimum set of hyperparameters for our algorithm. In the end of the training phase, we have obtained the hyperparameters that performed better. We train the algorithm containing with the best hyperparameters in the whole training subsample and we predict the testing fold we left aside, to obtain our out of sample performance. In the next step we repeat the process by adding to the training set the test fold we just used to make our predictions, and when we have obtained the new set of hyperparameters, we retrain the algorithm, and we predict the next test fold. In our case, this procedure repeats in the following way:

1. Training 1: 2005-2014; Test 1: 2015

[4] In the case of purely autoregressive time-series, the standard k-fold Cross-Validation is the preferred method (Bergmeir et al., 2018).

2. Training 2: 2005-2015; Test 2: 2016
3. Training 3: 2005-2016; Test 3: 2017

We repeat this process until we get out of sample predictions for the whole initial set we set aside for testing. In total, we obtain six out of sample predictions. The reported out of sample performance score is the average score for all the test years $OOSscore = \frac{1}{n}\sum_{1}^{n} Score_n$

4.1 Performance Evaluation Metrics

The R&D accounting choice issue can be converted to a binary classification task (capitalize vs expense) and therefore to evaluate the prediction performance we need to use the appropriate metrics used in classification. The most easy and straightforward way to measure performance is the Accuracy score which is defined as $ACC = \frac{TP+TN}{TP+TN+FN+FP}$, where TP (True Positive) are the instances in the positive class (capitalizer) which are classified correctly as capitalizers, TN (True Negative) are the instances in the negative class (expenser) which are classified correctly as expensers, FP (False Positive) are the instances which are expensers but they have been classified as capitalizers and FN (False Negative) are the instances which are capitalizers but they have been classified as expensers. From the definition of the Accuracy score it is obvious that in an unbalanced classification scenario, this metric is biased towards the class with the more instances.

Another scoring option would be the F1 score, which is defined $F1 = 2 * \frac{precision*recall}{precision+recall}$, or alternatively in terms of TP and TN, as $F1 = \frac{TP}{TP+\frac{1}{2}*(FP+FN)}$. Precision or positive rate is the number of positive class predictions that indeed belong to the positive class to the total number of predicted positives, both correctly and incorrectly classified ($\frac{TP}{TP+FP}$), while Recall is the number of positive class predictions correctly classified as

positives to the total number of positive instances ($\frac{TP}{TP+FN}$). However, F1 score requires us to choose a threshold[5].

To address this issue, we have used the ROC-AUC score, like many other researchers in the ML discipline (Bao et al., 2020; Bertomeu et al., 2021; Larcker & Zakolyukina, 2012). As Fawcett (2006) explained, the ROC curve depicts the classifiers performance at all possible classification thresholds. In fact, the F1 Score can be calculated for any point on the ROC curve, so the ROC curve "averages" the F1 score for all possible thresholds. To get a single number as a score, we calculate AUC (Area Under Curve). AUC value ranges from 0 to 1, with a value of 0.5 denoting that the classifier makes random guesses.

[Insert Figure 2 here]

5. The Out of Sample Performance of the Benchmark Model

In this section we use the most common and classic algorithm used in classification, the logistic regression (LR) to obtain out of sample predictions using the benchmark model which contains only financial ratios selected from the theory. In the next step, we use the raw financial data models and test if they can outperform the benchmark.

Before we train our models, it is necessary to preprocess our data. The ratios and the raw items used in our study exhibit scale differences. Data scaling, although it is uncommon in accounting literature, it is very common in ML and operations research. Even in the case of the benchmark model that consists of ratios which are in fact raw financial items divided by Total Assets, predictions can be improved by normalizing or standardizing the data (Shanker et al., 1996). By normalizing our data, we rescale the firm-year observations in the range [0,1]. This is also called Min-Max normalization, which is achieved by applying the formula $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, where x' is the scaled value, in every firm-year observation. In the end,

[5] Refer to Chicco & Jurman (2020) for a detailed study on the F1 Score's limitations and drawbacks

all features will have the same scale. We train the benchmark model two times, one with normalized features and one without, using the LR. The results are in line with the relative literature about the effect of normalization, as the normalized features model performs better. The Cazavan-Jeny et al. model achieves an AUC score of 0.73, which is our benchmark score.

[Insert Table 2 here]

5.1 Can Raw Financial Items beat the Benchmark Model?

In this step of our analysis, we use the raw financial items to predict accounting choice for R&D and compare their out of sample performance with the performance of our benchmark model. We use five models with raw financial items as described in Section 3.3.

[Insert Table 3 here]

We notice that all raw data models have an average AUC score of approximately 0.7, which is close to the performance of our benchmark. However, they cannot beat the Benchmark. To further investigate the issue, we pick the best performing set of raw financial items, that is the one with the 32 items from the financial statements and compare it with the benchmark model, by using more complex algorithms. We seek to investigate whether a more complex algorithm would capture more patterns in the raw data and if that could help to beat the Benchmark model.

5.1.1 K-Nearest Neighbors

K-Nearest Neighbors (KNN) algorithm (Fix and Hodges, 1951)[6] is a simple, non- parametric algorithm which can be used both for classification and regression problems. KNN can be used for linear and non-linear problems. While it can be easily and fast implemented

[6] There is not an official publication about the KNN method; Fix and Hodges introduced the KNN in an unpublished US Air Force School of Aviation Medicine report in 1951 (Silverman & Jones, 1989).

algorithm (relatively to the sample size), it is considered an effective algorithm with many applications (Triguero et al., 2019).

A KNN algorithm maps every class instance of the training sample to the real n-dimensional space. The number of the dimensions is equal to the number of model's features. The basic assumption of KNN is that similar data points (data points of the same class) are in proximity. This proximity in mathematics is called distance. To measure the distance, we rely on distance metrics. The most widely used metric is the Minkowski distance[7]. The distance is calculated in the following way. For two data points $A = (a_1, a_2, \dots, a_n)$ and $B = (\beta_1, \beta_2, \dots, \beta_n) \in \mathbb{R}^n$, the Minkowski distance of order p is defined as:

$$D(A,B) = \left(\sum_{i=1}^{n} (|a_i - \beta_i|)^p \right)^{\frac{1}{p}}$$

By changing the *p*-value to 1 we get the Manhattan distance:

$$D_{MNH}(A,B) = \sum_{i=1}^{n} |a_i - \beta_i|$$

And when *p* equals to 2, we get the Euclidean distance:

$$D_{EUC}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - \beta_i)^2}$$

When the distance between data points is calculated and stored in the training set (those are the Neighbors), the algorithm is "fed" with the test set. The algorithm will search for the K-Nearest Neighbors to the data point of the test set in order to classify it. The number of K ranges from 1 to the total number of instances. We have used a grid search to find the number of K that maximizes the AUC score. However, we must note that a K=1 will cause the algorithm to overfit the data and a K equal to the sample size is computationally expensive.

[7] Other distance metrics like Cosine, Jaccard or any other custom distance measure can be used.

[Insert Figure 3 here]

5.1.2. Random Forest

The Random Forest algorithm is based on the Decision Trees and was introduced by Breiman (2001). It is a versatile, general purpose, but rather complicated algorithm, which is suitable for all types of data and problems. Random Forest can be described as a collection of Decision Trees, whose predictions are averaged. As Anand et al. (2019) explained, this procedure can be parallelized to the Fama-Macbeth procedure which averages regression coefficients estimated in different years.

A unique Decision Tree is grown for a bootstrap sample of the training set by using a random sample of features. This is repeated as many times as the number of the trees that have been specified by the user of the algorithm, and a forest is grown. When the training has been completed, the test data are fed to the trees of the forest, and the prediction is the result of the majority vote of each unique tree. The complexity of the algorithm is the fact that the user has to hyper tune many parameters. First a decision must be made regarding the number of the trees in the Forest. Many researchers believe that the number of the trees is not a parameter to tune and that a large number of trees is sufficient. Adding more trees in the forest increases accuracy, but adding more trees requires more computational power and does not guarantee a perpetual improvement in accuracy (Probst & Boulesteix, 2017).

We have used grid search to find the optimal hyperparameters for our Random Forest. Among many possible hyperparameters, we have chosen to tune the maximum depth of the trees, the minimum number of samples required to split a node and the minimum number of samples required in each leaf node. It is also possible to determine the criterion based on which each split occurs. The Gini impurity or the entropy can be used. However, performing a grid search examining both criteria at the same time is computationally expensive and most

of the times performance is the same. For this reason, we have used the Gini impurity as the splitting criterion.

[Insert Figure 4 here]

5.1.3 The Out of Sample Performance of the Complex Algorithms

When we use more complex algorithms, the raw data model achieves better AUC score compared to the benchmark model. In the case of KNN, the difference between the two models is indisputably significant. Despite this, we have noticed that this difference is not caused by better performance of the raw data model but by the worse performance of the benchmark model (compared to the LR).

[Insert Table 4 here]

When we use the Random Forest, performance is marginally better for the raw data model, but only by 2%. Random Forest has boosted the performance of both models and clearly is a better classifier than KNN for the given data.

To sum up, we have an indication that raw data perform better than ratios or in the worst-case scenario, they perform equally well.

5.2. Introducing a Hybrid Model: Ratios and Raw Data Combined

In this approach, we take advantage of the data Normalization, and we combine the benchmark model which consists of ratios, with the 32 raw financial data. We use the best performing algorithm (Random Forest) to fit the data. By that, we have achieved a further 5% AUC score improvement compared to the best performing model (raw data only) fitted with Random Forest.

[Insert Table 5 here]

5.2.1 Does Data Preprocessing Affects Random Forest?

We further extend our analysis by fitting a Random Forest to the Hybrid Model without any data preprocessing. We do not winsorize the dataset, so outliers will be present, and we do not normalize it. We input ratios and raw data items, all with different scaling. In this approach, we followed the analysis of Anand et al. (2019), who showed that Random Forest is insensitive to outliers and data scaling. Our results are in line with theirs and we also concluded that when Random Forest is used, no data preprocessing is necessary at all.

[Insert Table 6 here]

5.2.2 Is Random Forest a Black Box?

One of the criticisms of the ML methods (and especially neural networks) is that they are considered "black box" methods. The user can just pick an algorithm, fit it to the data and obtain some results. This is also the case too with the Random Forest, especially if it is compared to a single Decision Tree. The single Decision Tree is interpretable while the Random Forest contains so many trees that no visual interpretation is possible.

[Insert Figures 5 and 6]

Even though visual interpretation is impossible with Random Forest, we can calculate feature importance. Feature importance provides us with evidence of how important each feature was in the classification (Tuv et al., 2009). Each individual tree in the Forest performs feature selection by selecting splitting points. In each split, there is a change in Gini impurity. The bigger the change, the greater the importance (Breiman, 1984). In Figure 7 we report the top 10 most important features based on their contribution to the prediction, starting with the most important, as ranked by their average importance throughout the test period 2015-2020.

[Insert Figure 7 here]

We notice that four features are ratios, and the rest are raw financial items. The importance of the four features (which are used in the Cazavan-Jeny et al. model) is adequately documented in their study.

What is of great interest, is the raw data items. “Other Intangible Assets” is the most important raw item. Other Intangible Assets contain intangibles like computer software, patents, copyrights. There is evidence that firms that disclose patents and citations, these are more value relevant to accounting choice than capitalized R&D (Ciftci & Zhou, 2016). We can think of patents as the output of successful R&D investments. So, there is a causal relationship between the two features. The second most important raw item is the “Audit Fees”. Audit fees are reported in the notes on the financial statements. It is well documented in the literature that there is a connection between capitalized R&D and audit fees. Auditors believe that some firms capitalize R&D in order to manage earnings and thus charge higher fees (Cheng et al., 2016). Furthermore, R&D intensive firms prefer Big 4 audit firms and auditors who specialize in auditing R&D and charge higher fees (Godfrey & Hamilton, 2005).

The importance of “Property, Plant and Equipment” (or PPE) can be explained by the relationship between R&D and tangible investments. R&D activities may require facilities, machinery or a tangible asset to introduce to the market new products generated by an R&D project (Carboni & Medda, 2019). The relationship between sales and R&D activity is well documented too. The more research the better the sales growth (Morbey & Reithner, 1990). However, there is not solid evidence which accounting choice for R&D affects more the sales growth. Cazavan-Jeny et al. (2011) suggest that not only the choice but also the amount of R&D affects sales growth. The impact of “Common Shares Outstanding” maybe is related to the ownership effect on R&D investments (see: Choi et al., 2015).

Despite the fact that feature importance is a useful tool and offers insights on how the algorithm makes predictions, we do not have an indication in which direction our features affect the R&D accounting choice. We can only make assumptions relying on theory and past research. In our effort to shed light in the "black box" of ML, we use Partial Dependence Plots (PDP)[8] to visualize the interaction of our features with the target variable.

[Insert Figure 8 here]

From the PDP plots we have a clear explanation on how our features affect the decision to capitalize R&D costs. We notice a positive relation between the volatility of R&D cashflows (CV_CFRD). The more volatile the cashflows are, the more probable it is that the management will capitalize the R&D costs. This is in line with Fudenberg & Tirole (1995) who support that managers prefer smooth earnings and cashflows, and in this case, they use R&D capitalization to achieve it. We notice a negative association between ROA and the decision to capitalize, which is supported by Aboody & Lev. (1998), who suggest that profitable companies do not capitalize in order not to harm the perception of the analysts on the quality of their earnings. The same negative relationship is observed with the cash flows from R&D (CF_RD) too. There is a positive relationship between Other Intangible Assets and the decision to capitalize, as it is supported by prior evidence. From the rest of the plots, all seem not to effect capitalization much, apart from PPE and Audit Fees. Although it is documented that R&D investments lead to tangible investments, there is a negative relationship between PPE and R&D capitalization. Firms with zero or small PPE assets tend to capitalize R&D, but as PPE value grows, firms seem to avoid capitalization. As the Audit Fees increase, the possibility of capitalization decreases. A possible explanation would be that more reputable and specialized auditors in R&D (who charge high audit fees) question

[8] For more info about the PDP, mathematical definition, and computation methods, refer to https://scikit-learn.org/stable/modules/partial_dependence.html

the management's decision to capitalize and they examine more thoroughly whether the firm meets the capitalization criteria.

In the final step of the Partial Dependence Plots analysis, we examine the interaction between the two most important features, CV_CFRD and ROA and how this interaction affects capitalization.

[Insert Figure 9 here]

According to the PDP, firms with ROA higher than 0.06 and CV_CFRD lower than 0.7 are more likely to capitalize. In other words, firms with less volatile cash flows from R&D and higher ROA are more likely to be classified as capitalizers.

6. Sensitivity Analysis

How we split our data for training and testing affects the classification performance. It is important to have a sensible balance between training and testing sample size; Very small or very large training sample may have a negative effect on performance (Xu & Goodacre, 2018). This is the reason we did not make an explicit train-test split, but we rather used cross-validation and made several splits, reporting the average performance across all test sub-samples.

Despite this, we made a design choice on where the first split should occur; Based on Figure 1, we made the split in the point captures most of the trends in the R&D accounting choice, as previously stated. To verify our design choice, we use an alternative test period, from 2010 to 2020. We fit the best performing model (Hybrid model) with a Random Forest and report the out of sample performance. The results indicate that the alternate test period does not change the performance, as well as that our approach is robust.

[Insert Table 7 here]

7. Conclusion

Accounting treatment for R&D, and more specifically the determinants of R&D capitalization is an important stream of research in the accounting literature. In this study we aim to develop a novel out of sample capitalization prediction model based on a sample of European listed firms over the period 2005-2020. Because of the intertemporal nature of our data, we use the last six years of our sample (2015-2020) as the out of sample test period and the years before that as the training period. To validate the robustness of our model, we have used an alternate test period from 2010 to 2020. Our results are valid in the alternate period, as well.

Our research differs from the existing literature in several ways. First, in the best of our knowledge, we are the first that predict out of sample accounting choice for R&D rather than trying to explain capitalization determinants within sample. Secondly, we use raw financial data from the financial statements rather than financial ratios to predict capitalization. With this approach we provide evidence that raw data are also informative, and, under certain circumstances, they may have better predictive ability compared to ratios. Thirdly, we use powerful Machine Learning algorithms rather than the commonly used logit regression used in similar research. Our results show that ML outperforms traditional econometric methods.

We used the Cazavan-Jeny et al. (2011) model, which uses financial ratios to explain the determinants of R&D capitalization, as a Benchmark. Afterwards, we compared the performance of the Benchmark against different sets of raw financial data using the Logistic regression. We found that none of the sets of raw data can beat the Benchmark model. In the next step of our analysis, we compared the best performing raw financial data model against the Benchmark by using more complex ML algorithms like KNN and Random Forest. We discovered that by using more advanced techniques the raw data model performs better than

the model which uses financial ratios. To find the optimal model, we combined the raw data items with the ratios, thus creating a Hybrid model. We used Random Forest, as the best performing algorithm, to fit the data. We conclude this study by providing evidence that the Hybrid model achieves the best performance compared to any other model used in this approach.

The empirical results presented in this study hold implications for the existing accounting literature. First, from an economic point of view, the utilization of raw data in our model has revealed new possible determinants of R&D accounting treatment. Our results indicate that apart from earnings management, income smoothing and signaling, which are the main theories behind R&D accounting choice, there are other determinants that should be considered, like the tangible assets of the firm and the reported other intangible assets.

Second, for future researchers, we have introduced a new approach to study the R&D accounting treatment phenomenon. We provide evidence that raw data and a combination of raw data and ratios can be used to create better models compared to the traditional ratio approach. Moreover, we suggest that machine learning is a useful alternative to standard econometric methods, and it is possible that it provides better performing models.

Our study is relevant to the R&D capitalization determinants literature (Cazavan-Jeny et al, 2011; Oswald, 2008) and research about the financial ratios versus raw financial data debate (Bao et al., 2020). One limitation of our study is that we were not able to obtain adequate enough R&D data prior to 2005 and make a comparison of R&D capitalization prior and after the implementation of IAS 38. For future research, we suggest a “brute force” approach, by comparing a model that contains all available fields in the financial statements versus our Hybrid model.

Acknowledgements

This work has been partly supported by the University of Piraeus Research Center.

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 4th Call for HFRI PhD Fellowships (Fellowship Number: 9116).

We would like to thank Professor Dimitrios Thomakos for his insightful comments and suggestions.

Appendix A: Variable Definition

Table A: WorldScope-Datastream Items

Variable	Definition	Measurement
RDCAP	Development Costs- Gross	WC02505
TA	Total Assets	WC02999
RDAM	Development Costs – Accumulated Amortization	WC02506
RDEXP	Research & Development Expense	WC01201
NETSAL	Net Sales or Revenues	WC01001
NETINC	Pre-tax Income (Income before tax, extraordinary items and preferred dividends)	WC01401
IIN	Interest Income	WC04149
IP	Interest Paid	WC04148
CAPEX	Capital Expenditures (Additions to Fixed Assets)	WC04601

Table B: Computed Variables

Variable	Defintition	Measurement
TAFR	Total Assets	TA-RDCAP-RDAM
SIZE	Natural Logarithm of TAFR	ln(TAFR)
NFE	Net Financial Expenditure	IP-IINC
ROA	Return on Assets	IBT+NFE+RDAM+RDEXP/0.5*(TAFR+$TAFR_{t-1}$)
RDS	R&D Intensity	RDEXP/NETSALES
CF_RD	Cash Flow of R&D (irrespective of its Acc. Treat.)	RDS+NETSALES+ΔRDCAP/0.5*(TAFR+$TAFR_{t-1}$)
CV_CFRD	Coefficient of Variation in CFRD	SD CFRD/Abs. Mean CFRD
CV_ROA	Coefficient of Variation in ROA	~
DEBTCAP	Gearing	TD/0.5*(TAFR+$TAFR_{t-1}$)
CAPEX	Natural Logarithm of Cap.Expen.	CAPEX/0.5*(TAFR+$TAFR_{t-1}$)

APPENDIX B: Algorithm Hyperparameters

1. Logistic regression

{"C":np.logspace(-3,3,7), "penalty":["l1","l2"], "solver":["lbfgs", "liblinear","saga}

2. K Nearest Neighbors

k_range = [2,5,10,15,20,25,30,35,40,45,50,55],{'n_neighbors': k_range}

3. Random Forest

n_estimators=700

Number of features to consider at every split

max_features = ['sqrt']

Maximum number of levels in tree

max_depth = [1,5,8,10]

Minimum number of samples required to split a node

min_samples_split = [2, 5]

Minimum number of samples required at each leaf node

min_samples_leaf = [10,15,20]

Method of selecting samples for training each tree

bootstrap = [True, False]

References

Aboody, D., & Lev, B. (1998). The value relevance of intangibles: The case of software capitalization. *Journal of Accounting Research*, *36*, 161–191.

Ahmed, A. S., Neel, M., & Wang, D. (2013). Does Mandatory Adoption of IFRS Improve Accounting Quality? Preliminary Evidence. *Contemporary Accounting Research*, *30*(4), 1344–1372. https://doi.org/10.1111/j.1911-3846.2012.01193.x

Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019a). Predicting profitability using machine learning. *Available at SSRN 3466478*.

Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019b). Predicting profitability using machine learning. *Available at SSRN 3466478*.

Ball, R. (1980). Discussion of Accounting for Research and Development Costs: The Impact on Research and Development Expenditures. *Journal of Accounting Research*, *18*, 27–37. JSTOR. https://doi.org/10.2307/2490325

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2021). A Response to" Critique of an Article on Machine Learning in the Detection of Accounting Fraud". *Econ Journal Watch*, *18*(1), 71–78.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, *26*(2), 468–519. https://doi.org/10.1007/s11142-020-09563-8

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Canace, T. G., Jackson, S. B., Ma, T., & Zimbelman, A. (2022). Accounting for R&D: Evidence and Implications*. *Contemporary Accounting Research*, *39*(3), 2212–2233. https://doi.org/10.1111/1911-3846.12780

Carboni, O. A., & Medda, G. (2019). Does R&D spending boost tangible investment? An analysis on European firms. *Applied Economics*, *51*(28), 3049–3065. https://doi.org/10.1080/00036846.2018.1564119

Cazavan-Jeny, A., & Jeanjean, T. (2006). The negative impact of R&D capitalization: A value relevance approach. *European Accounting Review*, *15*(1), 37–61.

Cazavan-Jeny, A., Jeanjean, T., & Joos, P. (2011). Accounting choice and future performance: The case of R&D accounting in France. *Journal of Accounting and Public Policy*, *30*(2), 145–165. https://doi.org/10.1016/j.jaccpubpol.2010.09.016

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting Management Fraud in Public Companies. *Management Science*, *56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

Cheng, J.-C., Lu, C.-C., & Kuo, N.-T. (2016). R&D capitalization and audit fees: Evidence from China. *Advances in Accounting*, *35*, 39–48. https://doi.org/10.1016/j.adiac.2016.05.003

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Choi, Y. R., Zahra, S. A., Yoshikawa, T., & Han, B. H. (2015). Family ownership and R&D investment: The role of growth opportunities and business group membership.

Journal of Business Research, *68*(5), 1053–1061. https://doi.org/10.1016/j.jbusres.2014.10.007

Ciftci, M., & Zhou, N. (2016). Capitalizing R&D expenses versus disclosing intangible information. *Review of Quantitative Finance and Accounting*, *46*(3), 661–689. https://doi.org/10.1007/s11156-014-0482-0

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements*. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Dinh, T., Kang, H., & Schultze, W. (2016). Capitalizing Research & Development: Signaling or Earnings Management? *European Accounting Review*, *25*(2), 373–401. https://doi.org/10.1080/09638180.2015.1031149

Elliott, G., & Timmermann, A. (2008). Economic Forecasting. *Journal of Economic Literature*, *46*(1), 3–56. https://doi.org/10.1257/jel.46.1.3

Fawcett, T. (2006). An introduction to ROC analysis. *ROC Analysis in Pattern Recognition*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fudenberg, D., & Tirole, J. (1995). A Theory of Income and Dividend Smoothing Based on Incumbency Rents. *Journal of Political Economy*, *103*(1), 75–93. https://doi.org/10.1086/261976

Godfrey, J. M., & Hamilton, J. (2005). The Impact of R&D Intensity on Demand for Specialist Auditor Services*. *Contemporary Accounting Research*, *22*(1), 55–93. https://doi.org/10.1506/P9FJ-EKAL-FPJQ-CM9N

Healy, P. M., Myers, S. C., & Howe, C. D. (2002). R&D Accounting and the Tradeoff Between Relevance and Objectivity. *Journal of Accounting Research*, *40*(3), 677–710. https://doi.org/10.1111/1475-679X.00067

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Landry, S., & Callimaci, A. (2003). The effect of management incentives and cross-listing status on the accounting treatment of R&D spending. *Journal of International Accounting, Auditing and Taxation*, *12*(2), 131–152. https://doi.org/10.1016/j.intaccaudtax.2003.08.003

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting Deceptive Discussions in Conference Calls. *Journal of Accounting Research*, *50*(2), 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

Lev, B., Sarath, B., & Sougiannis, T. (2005). R&D reporting biases and their consequences. *Contemporary Accounting Research*, *22*(4), 977–1026.

Lev, B., & Zarowin, P. (1999). The Boundaries of Financial Reporting and How to Extend Them. *Journal of Accounting Research*, *37*(2), 353–385. JSTOR. https://doi.org/10.2307/2491413

Morbey, G. K., & Reithner, R. M. (1990). How R&D Affects Sales Growth, Productivity and Profitability. *Research-Technology Management*, *33*(3), 11–14. https://doi.org/10.1080/08956308.1990.11670656

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Oswald, D. R. (2008). The Determinants and Value Relevance of the Choice of Accounting for Research and Development Expenditures in the United Kingdom. *Journal of Business Finance & Accounting*, *35*(1–2), 1–24. https://doi.org/10.1111/j.1468-5957.2007.02060.x

Oswald, D. R., & Zarowin, P. (2007). Capitalization of R&D and the Informativeness of Stock Prices. *European Accounting Review*, *16*(4), 703–726. https://doi.org/10.1080/09638180701706815

Paananen, M., & Lin, H. (2009). The development of accounting quality of IAS and IFRS over time: The case of Germany. *Journal of International Accounting Research*, *8*(1), 31–55.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Penman, S. H. (1996). The articulation of price-earnings ratios and market-to-book ratios and the evaluation of growth. *Journal of Accounting Research*, *34*(2), 235–259.

Prencipe, A., Markarian, G., & Pozza, L. (2008). Earnings management in family firms: Evidence from R&D cost capitalization in Italy. *Family Business Review*, *21*(1), 71–88.

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, *18*(1), 6673–6690.

Shanker, M., Hu, M. Y., & Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega*, *24*(4), 385–397. https://doi.org/10.1016/0305-0483(96)00010-2

Silverman, B. W., & Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 233–238. JSTOR. https://doi.org/10.2307/1403796

Soderstrom, N. S., & Sun, K. J. (2007). IFRS Adoption and Accounting Quality: A Review. *European Accounting Review*, *16*(4), 675–702. https://doi.org/10.1080/09638180701706732

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *WIREs Data Mining and Knowledge Discovery*, *9*(2), e1289. https://doi.org/10.1002/widm.1289

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, *10*, 1341–1366.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Walker, S. (2021). Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch*, *18*(2), 61.

White, G. I., Sondhi, A. C., & Fried, D. (2002). *The analysis and use of financial statements*. John Wiley & Sons.

Wyatt, A. (2005). Accounting recognition of intangible assets: Theory and evidence on economic determinants. *The Accounting Review*, *80*(3), 967–1003.

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Zhu, W., Zhang, T., Wu, Y., Li, S., & Li, Z. (2022). Research on optimization of an enterprise financial risk early warning method based on the DS-RF model.

International Review of Financial Analysis, *81*, 102140.

https://doi.org/10.1016/j.irfa.2022.102140

Figures & Tables

Table 1. List of Raw Financial Data Items

Available DataStream items (**1**)	Cecchini et al. (2010) raw items **(2)**	Dechow et al. (2011) raw items derived from financial ratios **(3)**	Replicated Cecchini et al. (2010) by Bao et al. (2020) **(4)**	Cazavan-Jeny et al. (2011) **(5)**
Statement of Financial Position items				
Cash & cash equivalents	YES	YES	YES	-
Accounts Receivables	YES	YES	YES	-
Inventories	YES	YES	YES	-
Short-term investments	YES	YES	YES	-
Current assets	YES	YES	-	-
Property, plant & equipment	YES	YES	YES	-
Total Assets	YES	YES	YES	YES
Current Liabilities	YES	YES	YES	-
Total Liabilities	YES	YES	YES	-
Equity Capital & Reserves	YES	YES	YES	-
Development Costs- gross	-	-	-	YES
Other intangible assets	-	-	-	-
Deferred taxes	-	-	-	-
Loans-net	-	-	-	-
Development costs- amortized				-
Total Debt	-	-	-	-
Long term liabilities	-	-	-	-
Income statement items				

Net sales	YES	YES	YES	YES
Cost of goods sold	YES	YES	YES	-
Depreciation & amortization	YES	-	YES	-
Interest income	YES	-	-	YES
Interest paid	YES	-	YES	YES
Research & Development expense	-	-	-	-
Income before tax, extraordinary items and preferred dividends	YES	YES	YES	YES
Net income	YES	-	YES	-
Income taxes	YES	-	YES	-
Cash Flow statement items				
Long Term Borrowings	-	-	-	-
Net proceeds from sale/issue Of common & pref. stocks	YES	YES	-	-
Capital Expenditure (CAPEX)	-	-	-	YES
Cash dividends paid	-	-	-	-
Extraordinary items	-	-	-	-
Market value items				
Common shares outstanding	YES	YES	YES	-
Market Price, Year End	YES	-	-	-
Other disclosure items				
Employees	YES	-	-	-
Audit fees	-	-	-	-

Column (1) lists all the available raw data items and column (2) our replication of Cecchini et al. (2010). Column (3) shows our replication of Dechow et al. (2011) derived from financial ratios while column (4) presents the replication of Cecchini et al. (2010) by Bao et al. (2020). Finally, column (5) contains the raw items from which the ratios of Cazavan-Jeny et al. (2011) were derived.

Table 2. The Out of Sample Performance Evaluation Metrics of the Benchmark Model for the Test Period 2015–20

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
8 financial ratios	1) Logistic Regression	0.71
	2) Logistic Regression (Normalized Data)	0.73

Table 3. The Out of Sample Performance Evaluation Metrics of the Raw Financial Items for the Test Period 2015–20

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
7 raw items from Cajavan-Jeny et al.	1) Logistic Regression (Normalized Data)	0.69
17 raw items from Bao et al. (Repl. Cecchini et al.)		0.69
15 raw items from Repl. Dechow et al.		0.69
22 raw items from Repl. Cecchini et al.		0.70
32 raw items from the financial statements		0.71

Table 4. The Out of Sample Performance Evaluation Metrics of the Raw Financial Items vs Benchmark for the Test Period 2015–20

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
8 financial ratios	1)KNN (Normalized Data)	0,63
32 raw items from the financial statements		0,71
	2) Random Forest (Normalized Data)	0,82
		0,84

Table 5. The Out of Sample Performance Evaluation Metrics of the Hybrid Model for the Test Period 2015–20

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
8 financial ratios plus 32 raw items from the financial statements	1) Random Forest (Normalized Data)	0,89

Table 6. Processed versus Unprocessed Data

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
8 financial ratios plus 32 raw items from the financial statements	1) Random Forest (Normalized Data)	0,89
	2) Random Forest (Raw Data)	0,9

Table 7. The Out of Sample Performance Evaluation Metrics of the Hybrid Model for the Test Period 2010–20

	Performance Metrics Averaged over the Test Period 2015–2020	
		Metric
Input variables	Method	AUC
8 financial ratios plus 32 raw items from the financial statements	1) Random Forest (Normalized Data)	0,89

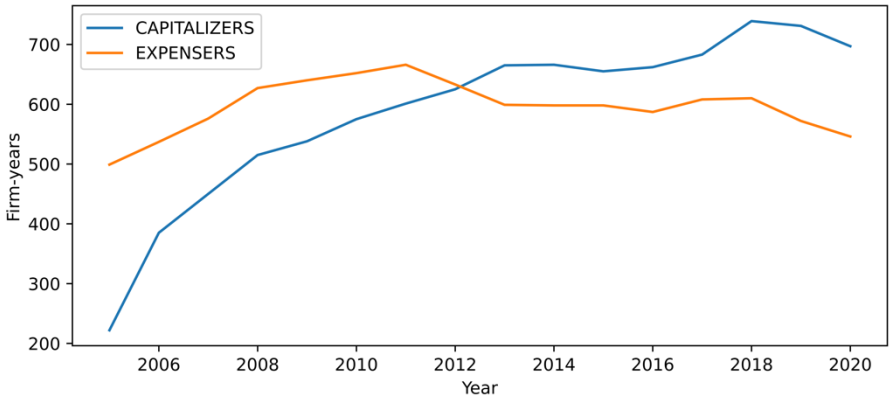

Figure 1. Accounting treatment of R&D per year

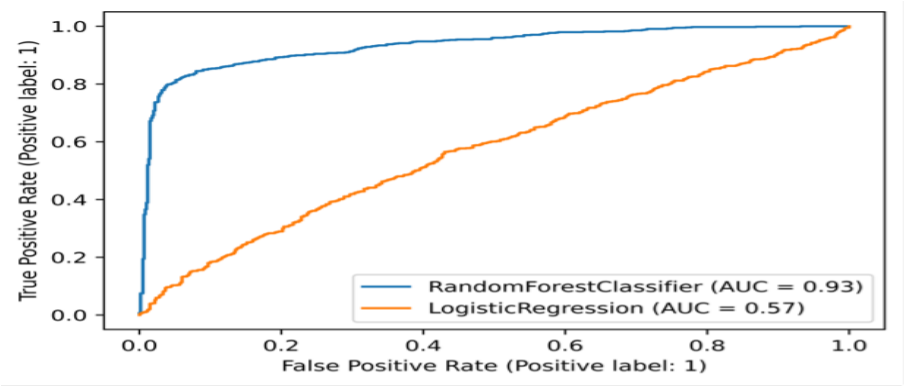

Figure 2. Example of ROC-AUC Curve

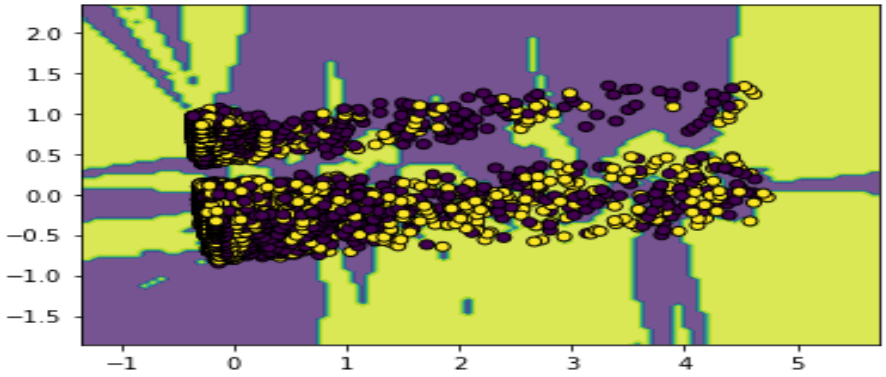

Figure 3. KNN decision boundary plot

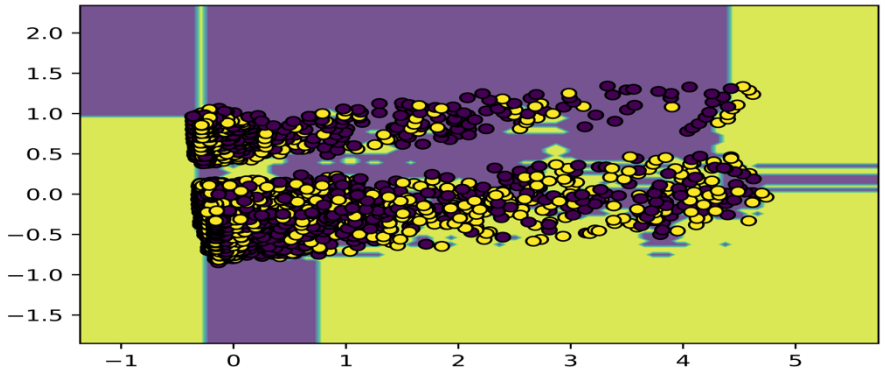

Figure 4. Random Forest decision boundary plot

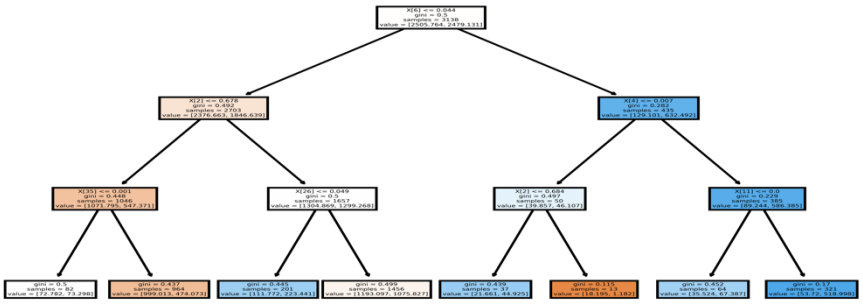

Figure 5. Pruned (four levels) Decision Tree

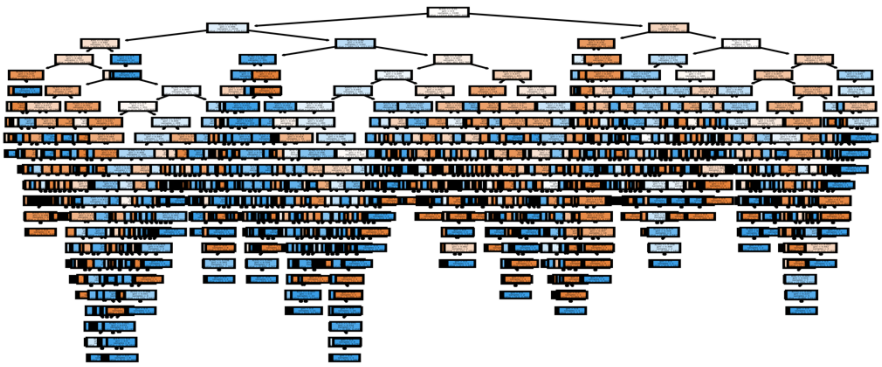

Figure 6. Fully grown single tree from the forest

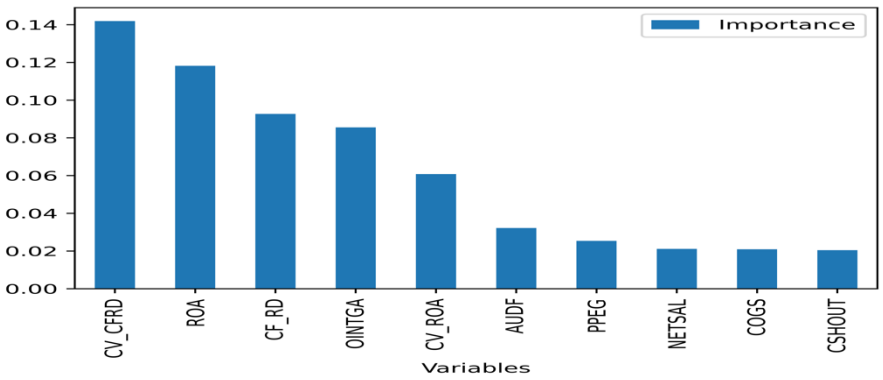

Figure 7. Top 10 feature importance

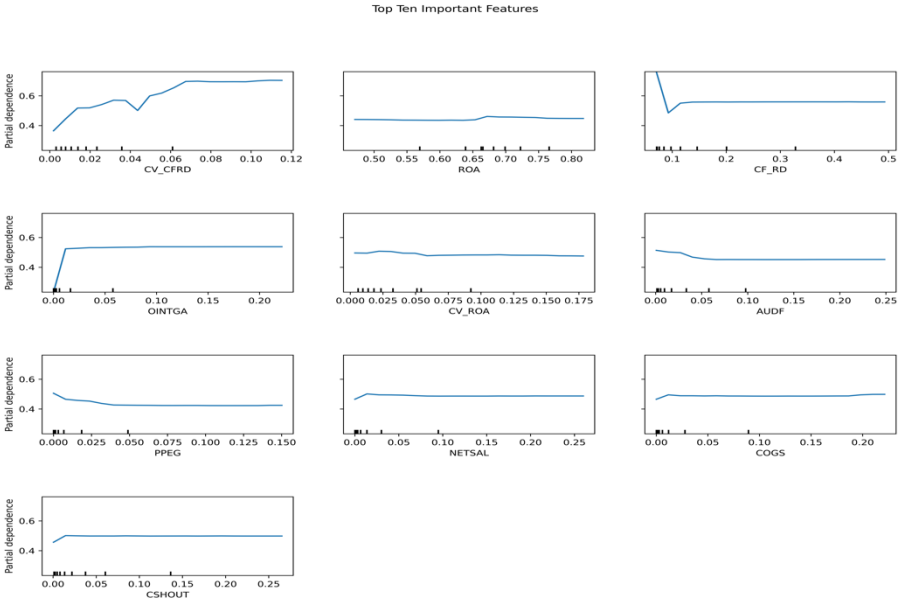

Figure 8. Partial Dependence Plot of the 10 most important features

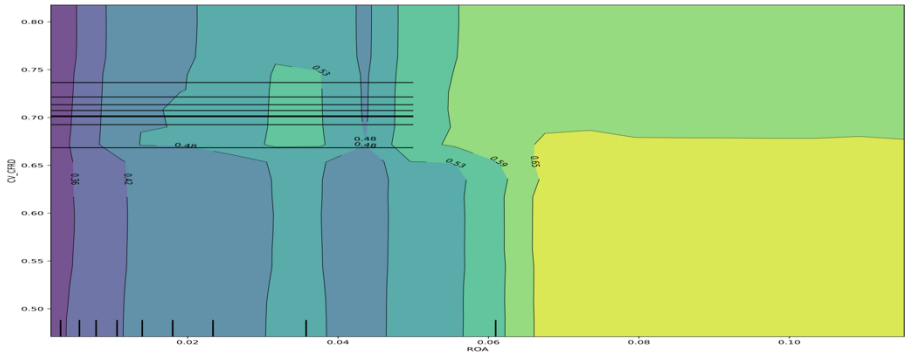

Figure 9. Two-way Partial Dependence Plot of the two most important features