Scientax
Jurnal Kajian Ilmiah Perpajakan Indonesia

djp

VOLUME 6 NO. 2 | APRIL 2025 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Assessing Taxpayers' Ability to Pay: A Machine Learning Approach[1]

Sukaryo [a], Adi Marhadi [b]
[a] Directorate General of Taxes, Jakarta, Indonesia. Email: sukaryo1@gmail.com
[b] Directorate General of Taxes, Manokwari, Indonesia. Email: a.marhadi@gmail.com

* Corresponding author: sukaryo1@gmail.com

ABSTRACT

Tax revenue remains one of the challenging fiscal issues in Indonesia. Improving tax collection performance through comprehensive reform has been an influential agenda, especially for the Directorate General of Taxes. One of the critical improvement areas is the utilization of information technology in tax assessment and audit functions. This study explores the taxpayers' ability concept as a complementary measure to the existing taxpayer monitoring module, particularly in case selection and targeting functions under the Compliance Risk Management (CRM) framework. The 5Cs of credit analysis (Character, Capacity, Capital, Condition, and Collateral) are employed as proxies for the taxpayers' ability to pay. This research aims to identify the most effective machine learning algorithm for classifying taxpayers' ability to pay to enhance the CRM's effectiveness for corporate taxpayers, limited to those administered in large and medium tax offices. Several machine learning algorithms were tested, including logistic regression as a baseline comparison, based on the quantitative and qualitative performance comparison. The findings reveal that the Light Gradient Boosting Machine algorithm provides the most effective results in terms of both accuracy and computational efficiency. However, several challenges need to be addressed to improve the model implementation.

Keywords: machine learning, ability to pay, taxpayers, compliance risk management, tax compliance

1. INTRODUCTION

Tax revenue is a daunting challenge in Indonesia, with a stagnated revenue-to-GDP ratio of roughly 11% or considerably lower than the regional and income-group average (de Mooij et al., 2018). In general, the low revenue ratio suggests an inefficiency in tax revenue collection (IMF, 2019). Consequently, there has been mounting pressure to improve tax collection through various policy and administrative measures. Furthermore, from 2017 to 2020, the government through the Directorate General of Taxes (DGT, Indonesia's tax authority) carried out the latest comprehensive tax reform, which encompasses policy and administrative pillars (DGT, 2020).

However, even though the reform generally aligned with the Medium-Term Revenue Strategy (MTRS) framework, there is still some potential for improvement, for example, the

[1] The Ability-to-pay module has been officially launched and deployed on 14 July 2021. This paper only explores the background study and initial development of the module.

DOI: 10.52869/st.v6i2.530
Received: August 14, 2022; Revised: February 24, 2025; Accepted: March 13, 2025; Published: April 30, 2025

Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia is Sinta 3 Journal (https://sinta.kemdikbud.go.id/journals/profile/9121)
How to Cite:
Sukaryo, & Marhadi, A. (2025). Assessing taxpayers' ability to pay: A machine learning approach. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia*, *6*(2), 95–106. https://doi.org/10.52869/st.v6i2.530

administrative reform (de Mooij et al., 2018). Arguably, the tax administration is one of the ever-growing challenges as the increasing taxpayers' baseline is followed by limited filing compliance and suboptimum administrative capability (IMF, 2019).In 2008, the DGT administered around 10.68 million taxpayers. It quadrupled in 2019, reaching approximately 45.95 million taxpayers. This surge translated into considerable improvement of the filing rate, rising from 30.96% in 2008 to 73.06% in 2019 (DGT, 2011; DGT, 2020). However, the filing rate is lower than the 85% on-time filing rates of OECD and selected economies (OECD, 2019), signaling persistent compliance challenges.

Arguably, these compliance challenges are aggravated by constrained human resources and administrative bottlenecks. Between 2009 and 2019, DGT human resources grew modestly, from 31.8 thousand in 2009 to 46.6 thousand in 2019. Compared to regional peers, Indonesia's labor force to tax officer ratio of 2952 is higher than China (1979), Singapore (1739), Malaysia (1142), and Australia (721) (OECD, 2019). At the same time, the DGT suffers from productivity complications: higher staff allocation for routine and supporting activities and audit resources misallocation (de Mooij et al., 2018).

To address administrative limitations and enhance compliance, DGT could harness the Information Technology (IT) potential to improve risk management and case selection functions for targeting high non-compliance risks (de Mooij et al., 2018). Building on this, IT offers a transformative approach to tackling compliance risks through enhanced audit assessment capabilities. The use of technology to promote tax compliance has earned a more pivotal role, especially in tax audit and assessment functions (OECD, 2016). In particular, the emerging implementation of the advanced analytics approaches improves audit case selection (OECD, 2016). By large, an analytics-driven approach could promote a more efficient process and improve efficiency (Davenport & Harris, 2007).

In recent years, DGT has introduced and incorporated technology initiatives, including data analytics, as one of its leading strategies (DGT, 2016; DGT, 2017; DGT, 2020). One of the highlights of data analytics implementation in the DGT is the Compliance Risk Management (CRM) program (de Mooij et al., 2018; DGT, 2017). In short, DGT (2019) reported that the CRM clusters the taxpayers based on their non-compliance risk and then maps them according to probability and consequence level. As a result, the tax verification staff could prioritize assessment activity for the highest risk group (DGT, 2019; OECD, 2004).

As a relatively new case selection tool, this study observes that during its initial implementation in 2019, CRM did not yet account for the taxpayers' future financial capability—specifically, their ability to pay outstanding tax liabilities had they been audited[2]. One commonly used methodology for assessing audit performance is the outcome approach, which is based on the cost-benefit principle (OECD, 2006). In other words, an audit is generally considered favorable if it results in considerable future revenue. Additionally, OECD (2006) highlights that case selection is a critical component of a successful audit process. Therefore, improving case selection by incorporating a model that assesses taxpayers' collectability or future ability to pay could significantly benefit DGT.

Based on micro-level tax return data, this paper will explore the ability to pay (ATP) concept in Indonesia. Specifically, this research seeks to develop a machine-learning model capable of predicting taxpayers' future ATP. To effectively measure ATP, this study employs key financial proxies such as current assets, net revenue, operating cash flow, and outstanding liabilities, which serve as indicators of a taxpayer's financial capacity. This paper adopts the machine-learning approach to utilize the DGT's data analytics infrastructures and address the pattern-recognition challenges with the micro-level dataset

[2] The initial implementation of CRM, as outlined in the DGT's Circular Letter, integrates the taxpayers' Ability to Pay concept specifically within the tax collection function, which defined as the map of taxpayer's compliance risk in fulfilling their tax liabilities. In contrast, the CRM audit and assessment function does not incorporate ATP, focusing instead on compliance risks tied to the likelihood of noncompliance and taxpayers' contributions to tax revenue (DGT, 2019).

(Khandani et al., 2010). Furthermore, several studies showcased that the machine-learning techniques offer a powerful prediction power when used in financial sectors, e.g. to model credit risk (van Liebergen, 2017; Khandani et al., 2010; Provenzano et al., 2020; Bellotti and Crook, 2009). Thus, we seek to reproduce the findings in tax administration settings.

This paper has three primary objectives: (1) to provide an introductory study on the ATP concept; (2) to provide the most effective and efficient machine learning algorithm to classifying ATP; and (3) to enhance the CRM framework for tax verification in Indonesia. The ATP module could provide valuable information for account officers and auditors, enabling them to better assess taxpayers before initiating tax assessments or prioritizing CRM execution based on taxpayers' projected financial capacity. The remainder of this paper discusses the ATP concept, study limitations, methodology, results, and conclusions.

2. CONCEPTUAL FRAMEWORK

Equity has been the heart of taxation principle, and one of the primary examples is Adam Smith's argument from 1776 that citizens should contribute to government support in proportion to their respective abilities (Smith, 1776)[3]. In particular, Musgrave (1996) argued that the interpretation of fairness in taxation is primarily aligned with the ability to pay concept. Furthermore, Musgrave (1996) added that the equity principle in taxation consists of two concepts, horizontal equity (where individuals with equal abilities should have equivalent tax burdens) and vertical equity (where those with greater abilities should bear higher tax obligations).

The quantification of the ability to pay notion is arguably somewhat limited, especially when applying it to the tax administration domain. From a policy perspective, we generally could determine the taxpayers' ability to pay based on their realized income or wealth, e.g., one of the bases for progressive income tax[4] (Slemrod, 1996; Gruber, 2011). However, under the administrative perspective (e.g., tax audit), the tax authority should assess both present and future taxpayers' taxable income and profit as the audit examines past compliance resulting in future tax liability[5] (OECD, 2006). Therefore, this study introduces a novel approach to adopt the ability to pay concept as taxpayers' financial capability based on an accounting perspective and credit analysis.

From an accounting perspective, assets are a relevant proxy of one's financial capability, reflecting an entity's economic resources (Alibhai et al., 2020). Traditionally, the ability-to-pay concept relies on income, which is defined by the International Accounting Standard (IAS) as increases in assets or decreases in liabilities that boost equity, excluding contributions from equity holders (Alibhai et al., 2020). This suggests that income is inherently tied to asset growth, positioning assets as a comprehensive measure of financial capacity.

Within this framework, IAS describes an asset as a present economic resource controlled by an entity due to past events, capable of generating economic benefits (Alibhai et al., 2020). Among these assets, IAS classifies current assets, such as cash, cash equivalents and items expected to be converted into cash, sold, or consumed within a normal business operating cycle, as a distinct subcategory based on their liquidity and immediate availability (Alibhai et al., 2020). Therefore, this study leverages current assets to capture taxpayers' financial capability, emphasizing their direct relevance to tax obligations.

[3] Musgrave (1996) argued that Smith's assertion combined two principles of equity, benefit and ability to pay.

[4] For example, Slemrod (1996) suggested that we could interpret the ability to pay principle as "equalizing the sacrifice due to tax"; thus, "tax should rise with income" even though the implementation might not as straightforward as the concept proposed. Similarly, Gruber (2011) use the Haig-Simons comprehensive income definition that define "taxable resources as an individual's ability to pay taxes" or "individual's potential annual consumption, potential annual consumption, plus any increase in his or her stock of wealth".

[5] One of the relevant concept in tax audit is the collectibility or a taxpayers' ability to pay the future tax assessment as a basis to conduct tax audit (OECD, 2006)

To assess taxpayers' Ability to Pay (ATP), this study uses current assets as key financial indicators, borrowing accounting perspectives commonly applied in credit assessments by credit rating agencies. Tangible current assets, such as cash, bank balances, accounts receivable, and inventories, represent liquid resources, which are arguably readily convertible to cash to meet tax obligations (Alibhai et al., 2020). Similarly, intangible current assets, including short-term investments and prepaid expenses, highlight the taxpayer's near-term financial stability by capturing resources that support liquidity and operational continuity despite their non-physical nature (Alibhai et al., 2020). This approach aligns with standard accounting principles, where assets reflect a firm's economic capacity to fulfil its financial commitments.

Liquidity measures, such as working capital and cash flow status, complement these asset-based indicators by providing a broader view of financial flexibility. Working capital (calculated as current assets minus current liabilities) highlights a taxpayer's ability to cover short-term debts, while cash flow status indicates the cash available to sustain operations and compliance (Kieso et al., 2016). By combining these elements, this study constructs a practical framework to evaluate taxpayers' ATP, tailored to strengthen CRM function, in particular case selection and prioritization, within DGT.

Building on this accounting foundation, credit analysis enhances the ATP assessment by offering a structured method to evaluate financial capability. Credit analysis, described by Ganguin & Bilardello (2005) as a systematic and thorough evaluation of a firm's capacity and willingness to meet its financial obligations, is a method widely employed by rating agencies to assess creditworthiness. This study adapts the "5 Cs of credit" framework—Character, Capacity, Capital, Conditions, and Collateral—to provide a comprehensive lens for understanding taxpayers' ATP (Ganguin & Bilardello, 2005; Golin & Delhaise, 2013). While no prior studies have directly applied this combined approach to tax compliance in Indonesia, it mirrors practices in credit scoring by rating agencies, ensuring a practical adaptation for the CRM context.

In this framework, Character is related to a firm's reputation, management quality, and financial policies, where aggressive growth strategies may signal higher risk (Ganguin & Bilardello, 2005). The second C, Capacity, described by Golin & Delhaise (2013) as a measure of firm's ability to generate cash or income, serving as the basis of financial capability. Capital is related to the owner's investment in the firm, indicating long-term stability, whereas Conditions portray the competitive environment and market position, influenced by factors like cost competitiveness (Ganguin & Bilardello, 2005). Lastly, Collateral highlights alternative funding sources, such as assets available to repay the obligations if needed (Golin & Delhaise, 2013). Among the five aspects, Golin & Dehlaise (2013) argued that Capacity, rooted in cash flow and financial risk analysis, remains the "core of credit analysis", as it directly ties to a firm's ability to meet obligations based on historical and current financial data. However, applying the 5Cs can be challenging due to variations in business types and data availability (Golin & Delhaise, 2013). Therefore, Golin and Delhaise (2013) illustrated some credit rating practices, such as sectoral-based credit analysis and different analytics approaches for various business scales.

3. METHODOLOGY

3.1 Methodology and Data

This study employs a supervised learning classification method as its core approach, with taxpayers' Ability to Pay (ATP) as the target variable. Specifically, this study explores multiple machine learning algorithms to predict and classify taxpayers' ATP for the upcoming year (t+1) using predictors from the current year (t+0). Current assets serve as the proxy for ATP, a choice based on accounting principles, where they represent liquid resources (Alibhai et al., 2020), and credit analysis practices, where they reflect financial capacity (Ganguin & Bilardello, 2005).

This study evaluates several machine learning classification models to identify an optimal

approach balancing accuracy and resource efficiency: Support Vector Matrix (SVM), Naïve Bayesian, tree-based Gradient Boosting, and Random Forest classifier methodology. The tree-based approach, such as Gradient Boosting, is highly relevant in this study as it could capture the non-linearity trend and multiple categoric variables, optimized for a relatively smaller dataset compared to the neural network algorithm (Friedman, 2001; Ke et al., 2017). The selection for the SVM algorithm is based on its capability in generalization and pattern recognition (Burges, 1998). Meanwhile, the Naïve Bayes approach is one of the most competitive classifications of machine learning techniques, especially in independent and dependent features (Rish, 2001). The tree-based Gradient Boosting algorithms are generally more favorable due to their efficiency, accuracy, and interpretability (Ke et al., 2017). Lastly, this paper also compares the Random Forest classification result as one of the most versatile algorithms in machine learning (Biau & Scornet, 2016). As a benchmark, this study compares these machine learning techniques against logistic regression, a widely recognized statistical method, to evaluate their performance relative to a traditional, non-machine learning approach (Hosmer et al., 2013).

When evaluating and comparing the models' output, this study uses 75% of the sample dataset as the "train data" to build the machine learning model (Hastie et al., 2009). The remaining 25% of the dataset will serve as "test data" for model evaluation (Hastie et al., 2009). The dataset was subjected to standard preprocessing techniques to handle missing values, remove outliers, and normalize numerical features. Missing values were treated using mean or median imputation, depending on the distribution of the data. Outliers were identified using the interquartile range (IQR) method and subsequently removed or adjusted to prevent model distortion. Categorical variables were encoded using one-hot encoding or label encoding where appropriate. Finally, numerical features were scaled using min-max normalization to standardize the dataset across different attributes.

This paper focuses on five parameters to evaluate the model's performance: accuracy, the area under the curve (AUC), recall, precision, and execution time. The accuracy measures the "correct" predictions based on the total number of samples (Fawcett, 2006). Meanwhile, precision and recall emphasize the proportion of the "correct" identifications, with precision focusing on "false positive" and recall focusing on "false negative" (Davis & Goadrich, 2006). The AUC parameter measures the performance of the classification model, focusing on the true positive rate (TPR) and false positive rate (FPR) (Bradley, 1997). The four parameters will take a value of 0 to 1, where 1 represents a model with 100% "correct" prediction. Lastly, this study incorporates execution time, measured in seconds, as a proxy for model efficiency, i.e., an efficient model will require a somewhat lower run time (Ke et al., 2017). Arguably, the lower run time will be one of the significant parameters during the scaling-up and deployment stage, especially applying the model with an extensive, high dimensionality dataset.

In order to forecast the ability to pay, there are two possible outputs: continuous or categorical. Under the supervised learning methodology, both are plausible under the "regression problem" for continuous values and the "classification problem" for discrete values (Hastie et al., 2009). However, to make the model easier to understand, explain, and possibly faster, this study adopts the latter approach by constructing the taxpayers' ability to pay as discrete value (Liu et al., 2002). This study categorized the taxpayers' ability to pay (i.e. current assets) into five classes: "very low", "low", "moderate", "high", and "very high". This was determined using a quartile-based approach on current assets, ensuring a balanced distribution across categories. The division are: (i) very low: Taxpayers in the first quartile (Q1), representing the bottom 25th percentile of current asset distribution; (ii) low: Taxpayers in the second quartile (Q2), between the 25th and 50th percentile; (iii) moderate: Taxpayers in the third quartile (Q3), between the 50th and 75th percentile; (iv) high: Taxpayers in the fourth quartile (Q4), above the 75th percentile but below extreme outliers; and (v) very high: Taxpayers beyond the

Table 1
Data Distribution, Train and Test Set

Tax Office	Total Data	Train Data (75%)	Test Data (25%)
Madya	54,878	41,158	13,720
Khusus	7,097	5,322	1,775
LTO	962	721	241
Total	62,937	47,201	15,736

Note. Authors' Calculation

90th percentile, representing firms with significantly high current assets.

This research utilizes administrative and external data for around 60 thousand corporate taxpayers, as summarized in Table 1. However, the data distribution shows significant stretching at both tails, potentially skewing the representativeness of these groups. To address this, this study further refines the classification by segmenting corporate taxpayers according to their administering tax offices, allowing cutoff points for the five ATP classes to vary across these groups. This adjustment accounts for regional or administrative differences in financial profiles, improving the model's applicability.

This paper finds that the "Madya" and "Khusus" tax offices group are comparable based on the data distribution. Thus, the corporate taxpayers will further be classified into two tax office groups: Large (for taxpayers from LTO) and Medium (from Madya and Khusus offices). Accordingly, the modelling will be built separately for the two classes.

3.2 Limitation

This study is subject to data availability limitations, impacting both the scope and depth of the analysis. Primarily, the data limitation arises from asymmetric information from tax return forms and third-party data exchange, especially for individual taxpayers who submit the 1770S or 1770SS forms. At the same time, the attempt to construct taxpayers' features under the 5Cs of credit analysis approach faces–at least–data representativeness challenges, e.g. relevant information and data matching problems. In other words, increasing the sample size would result in reduced dimensionality, as additional records may lack critical variables. Consequently, due to data availability issues, this study limits the observations for corporate taxpayers administered in the medium and large taxpayer offices.

Another data availability challenge is regarding the external data. While it is possible to construct tax return data using several fiscal years' information, the case of third-party external data is contrasting: the data is limited to recent years. Therefore, to balance these constraints and ensure recency, this study incorporates data from the 2019 and 2020 fiscal years. Also, aligning with the initial implementation of the CRM system.

The last identified challenge is obtaining a representative dataset to measure the 5Cs. Based on the administrative and external data, this study managed to proxy the 4Cs of credit – except Collateral. Arguably, the Capital and Collateral principles are highly correlated: the total asset figure represents the amount of capital investment and part of the firm's collateral ability (Ganguin & Bilardello, 2005; Golin & Delhaise, 2013). Therefore, this study employs the 4Cs principle: Capacity, Capital, Character, and Condition.

Table 2
Independent Variables Mapping and Classification

Variable	4 Cs	Type	Remarks
Turnover	Capacity	Numeric	
Commercial Net Profit	Capacity	Numeric	
Operating Cashflow	Capacity	Numeric	
Working Capital	Capacity	Numeric	
Net Revenue	Capacity	Numeric	
Gross Revenue	Capacity	Numeric	
Cash Equivalents	Capital	Numeric	
Net Assets	Capital	Numeric	
Total of Assets	Capital	Numeric	
Current Liabilities	Capital	Numeric	
Total Liabilities	Capital	Numeric	
Domestic account balance/Exchange of Information Data	Capital	Numeric	To complement the Cash Equivalents variable
Outstanding Credit	Capital	Numeric	To complement the Current Liabilities and Total Liabilities
Credit Collectability	Character	Ordinal	1. Current; 2. Special Mention; 3. Substandard; 4. Doubtful; 5. Bad.
Status of PKP (Value-added tax subject)	Character	Categorical	1. Non-PKP; 2. PKP.
Tax return filing status	Character	Categorical	1. Non-filers; 2. Normal; 3. Filer, correction.
Firm's maturity	Character	Ordinal	1. 0 – 5 Year; 2. 5 – 10 Year; 3. 10 – 15 Year; 4. more than 15 Years.
Operating Cashflow Status	Condition	Ordinal	1. Negative; 2. Stagnant; 3. Positive
Status of receiving a tax refund in the last six months	Condition	Categorical	1. No; 2. Yes
Conglomerate group	Condition	Categorical	1. No; 2. Yes

Note. Authors' Calculation

4. RESULTS AND ANALYSIS

Table 3 compares the classification result based on the prediction performance for every algorithm on the test set. The model output suggests that the gradient boosting algorithm generally outperforms logistic regression in predictive capability. The machine learning models average around 0.7 of accuracy, precision, and recall, while the run time varies across algorithms. Among the machine learning algorithms, the Gradient Boosting groups ("gbc", "lightgbm", "xgboost", and "catboost") achieve the highest performance and efficient running time – for both Large and Medium taxpayer models. For example, the non-

Table 3
Model Performance Comparison

Model	Accuracy	AUC	Recall	Precision	F-1 Score	Time (secs.)
Large Offices						
- Logistic regression	0.2500	0.5508	0.2169	0.1307	0.1631	0.1180
- SVM	0.1718	0.0000	0.2286	0.1893	0.2071	0.0560
- Naïve Bayes ("nb")	0.5261	0.8403	0.5204	0.5240	0.5222	0.0160
- Gradient Boosting Classifier ("gbc")	0.6954	0.9104	0.6880	0.7072	0.6975	0.5360
- Light Gradient Boosting Machine ("lightgbm")	0.6926	0.9152	0.6880	0.7072	0.6975	0.2020
- Extreme Gradient Boosting ("xgboost")	0.6770	0.9141	0.6674	0.6915	0.6792	24.2280
- Catboost Classifier ("catboost")	0.7082	0.9194	0.7053	0.7185	0.7118	8.2400
- Random Forest Classifier ("rf")	0.6718	0.9158	0.6649	0.6822	0.6734	0.4140
- Decision Tree Classifier ("dt")	0.6224	0.7600	0.6212	0.6257	0.6234	0.0200
Medium Offices						
- Logistic regression	0.1184	0.7836	0.2076	0.2671	0.2336	9.5620
- SVM	0.1357	0.0000	0.2196	0.3107	0.2573	0.6180
- Naïve Bayes	0.3904	0.7114	0.3807	0.3748	0.3777	0.4420
- Gradient Boosting Classifier	0.7557	0.9322	0.7672	0.7576	0.7624	40.6560
- Light Gradient Boosting Machine	0.7607	0.9342	0.7690	0.7623	0.7656	2.5680
- Extreme Gradient Boosting	0.7572	0.9324	0.7655	0.7586	0.7620	132.3740
- Catboost Classifier	0.7612	0.9344	0.7688	0.7629	0.7658	33.2160
- Random Forest Classifier	0.7552	0.9318	0.7662	0.7573	0.7617	4.7200
- Decision Tree Classifier	0.6565	0.7793	0.6663	0.6568	0.6615	0.7940

Note. Authors' Calculation

Gradient Boosting model performance for accuracy is around 0.49, while the Gradient Boosting approach reaches 0.73. However, Gradient Boosting models require more processing time, averaging 30 seconds per run (across the Large Offices and Medium Offices groups) compared to just 0.88 seconds for non-Gradient Boosting models, a difference of roughly 30 times.

Selecting the model: Performance and Efficiency. Following the result from Table 3, this study compares the six performance parameters to select the optimum model for predicting the taxpayers' ability to pay. Since these parameters (accuracy, AUC, precision, recall, F1-score, and run time) are measured on different scales and in different units, we first standardize them by calculating the z-score for each metric (Moore et al., 2017). This transformation centers all metrics around a mean of zero and scales the variability uniformly. Allowing for a direct and fair comparison across all performance and efficiency measures.

By converting each parameter into a standardized score, we can objectively assess each model's relative strengths and weaknesses. Also, model performance across traditional metrics often converges closely, with differences typically in the third decimal place, e.g., the accuracy result for "xgboost" is about 0.6770 or slightly higher than the "rf" with around 0.6718 of accuracy. In such cases, execution time emerges as a critical tiebreaker, especially scaling to the DGT's full datasets. For instance, while the "catboost" (Catboost Classifier) shows strong predictive performance, its considerably higher run time is also reflected in its z-score, indicating that its computational cost might outweigh its marginal

gains in performance for large-scale deployment: 8.24 seconds for Large Offices and 33.22 seconds for Medium Offices. In contrast, both "lightgbm" (Light Gradient Boosting Machine) and "rf" (Random Forest Classifier) not only deliver balanced performance (as indicated by their z-scores on traditional metrics) but also maintain a much lower run time: the "lightgbm" requires 0.20 (Large Offices) and 2.57 seconds (Medium Offices), whereas the "rf" doubles the "lightgbm" run time to 0.41 seconds and 4.72 seconds, respectively. Figure 1 compares the performance score for all algorithms.

The efficiency aspect, or running time, is highly significant during the scaling up and deployment stage, especially when there are around 2 million corporate taxpayers in the tax administration data. To illustrate, increasing the sample size from 241 (test set, Large Offices) to 15495 (test set, Medium Offices), or around 64 times, requires more than ten times the initial run time (11.4 times for "lightgbm" and 12.7 times for "rf"). Using the back-of-the-envelope approach, scaling up the dataset to around 2 million taxpayers would result in more than 300 times the execution time for "lightgbm" and "rf" models.

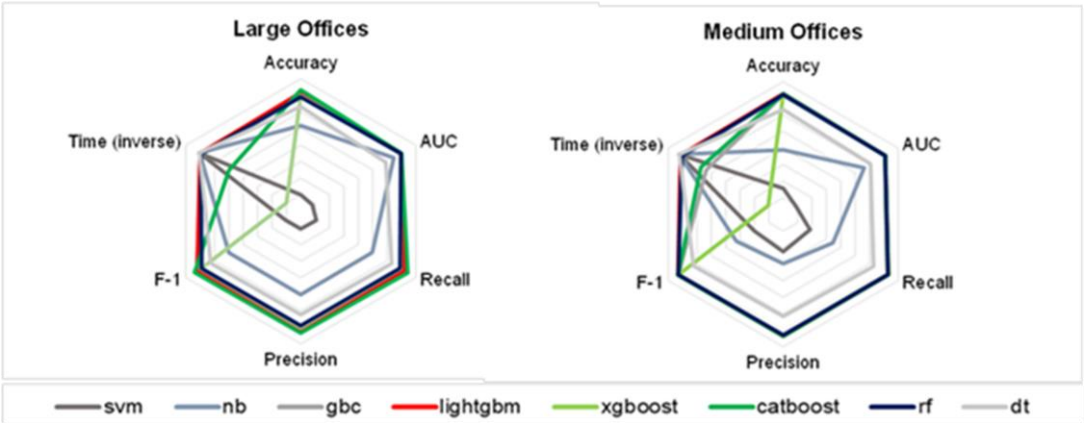
This study retains and presents the z-score result in spider charts (Figure 1) to enhance visual interpretability for stakeholders, e.g., DGT's officials, beyond the technical precision already provided in Table 3. The chart highlights trade-offs (e.g., "catboost"'s performance vs. runtime) without duplicating raw results, aligning with the study's aim of delivering transparent, actionable insights for CRM function.

To improve model transparency, feature importance analysis was also conducted for each machine learning model, particularly LightGBM and RF. The analysis revealed that Cash and Cash Equivalents are the most critical factor, as liquid assets directly indicate a taxpayer's ability to settle tax obligations. Additionally, Operating Cash Flow serves as a key measure of financial health, reflecting the taxpayer's capacity to generate cash from core business activities. Furthermore, Net Revenue acts as a strong indicator of a company's earning capability, which influences its ability to meet tax liabilities. Lastly, Current Liabilities are essential in assessing short-term financial obligations that may impact liquidity, providing further insights into a taxpayer's financial position.

Beyond these factors, Total Assets also play a significant role, capturing a firm's overall financial strength and long-term viability (Ganguin & Bilardello, 2005). While the framework prioritizes Current Assets for their immediate relevance to taxpayers' liquidity, Total Assets were tested as a robustness check to ensure model reliability across a broader financial context. This inclusion confirms that while liquidity-driven features dominate ATP prediction, broader asset measures contribute to a comprehensive evaluation, aligning with standard practices in model validation (Hastie et al., 2009). These findings highlight the relationship between

Figure 1

Machine Learning-based Model Performance Comparison



Note. Authors' Calculation

short-term liquidity and long-term stability in shaping taxpayers' financial capacity, offering insights for the DGT's CRM framework.

Aside from performance measurement using the test and train dataset, this research includes a practical validation phase with the operational unit (i.e. the Account Representatives) who directly interact with taxpayers. This approach is intentional because these officers are on the front lines of tax administration and can provide immediate, practical insights into the model's effectiveness in day-to-day operations.

For this validation, a survey was conducted on a sample of 120 taxpayers. The results indicate that, on average, about 65% of the cases were correctly classified by the model: 61.2% for Large Offices and 70.6% for Medium Offices. Moreover, Account Representatives reported that around 38.8% of the corporate taxpayers in Large Offices were classified as having a lower ability to pay than they perceived, while the figures were relatively lower for Medium Offices. The survey further showed under-classification rates of 17.6% for Medium Offices and over-classification of about 11.8%. Overall, the survey-based performance, with approximately 70% correct classification, aligns well with the quantitative metrics obtained from the datasets. Furthermore, follow-up inquiries with several Account Representatives revealed that some respondents factor in taxpayers' willingness to pay when assessing the model's fit with real-world cases. In some instances, even if the model accurately classifies a taxpayer's ability to pay, a low willingness to pay leads them to judge the taxpayer's actual ability as lower than the model suggests.

5. CONCLUSION AND RECOMMENDATION

Using several machine learning models, this study applies the 5Cs principle of credit analysis (Character, Capacity, Capital, Condition, and Collateral) to assess taxpayers' ATP. By using current assets as proxies for ATP, our results show that models such as the LightGBM and RF not only achieve robust predictive performance but also offer efficient computation compared to traditional methods like logistic regression.

In practical application, the ATP mapping developed in this research could become an integral aspect of the existing CRM engine. Incorporating ATP scores into the CRM framework enables Account Representatives to target taxpayers who pose a higher compliance risk and have a higher ability to pay. Subsequently, the combined insights can improve the quality of case selection and enhance the overall success of taxpayer assessments.

Nevertheless, as an initial study assessing the taxpayers' ability to pay, there are several challenges in the model development and output. First, expanding the range of variables, including qualitative indicators, such as willingness to pay and historical delinquency records. It may refine ATP estimates further. For example, the survey result introduced the qualitative variable of taxpayers' willingness to pay, which will be a relevant variable for taxpayers' Character. One possible proxy for the variable is the taxpayers' historical payment for tax assessments or taxpayers' delinquency records. Also, it is critical to include the Collateral principle in assessing the taxpayers' ability to pay, mainly using external data sources rather than the tax return information.

Lastly, improved dimensionality and sample are instrumental in improving the model's performance. With around 0.7 accuracies, it would be plausible to fine-tune the models to achieve higher classification performance. On the one hand, the 70% accuracy serves as a robust baseline for day-to-day taxpayers' assessments, for example, case selection and prioritization, especially given the inherent complexity and variability of taxpayers' behaviour. Although some cases may be misclassified, the model's output provides meaningful assistance to enhance the decision-making process, especially when combined with the Account Representatives' professional judgment and further qualitative insights. On the other hand, the 0.7 accuracies suggest that there remains room for improvement, for example, using more dimensions, ensemble methods, or more datasets.

It is crucial to test the model with taxpayers administered in small tax offices, which will bring more challenges of more observations and limited data availability. Future research on taxpayers' ability to pay needs to incorporate more comprehensive observation and historical data, allowing machine algorithms to "learn" distinct features related to the taxpayers' ability to pay. Also, if data and information are available, it is relevant to introduce the ATP for individual taxpayers. Moreover, exploring the possibility of employing an "ensemble" approach, combining several machine learning algorithms, is essential to creating a more robust prediction model while accounting for efficiency. Arguably, the taxpayers' ability to pay could be a critical complementary feature in the DGT taxpayers' assessment and supervisory functions. However, extensive model development and training are required to unlock its optimum capability.

REFERENCES

Alibhai, S., Bakker, E., Balasubramanian, T. V., Bharadva, K., Chaudhry, A., Coetsee, D., Dougherty, J., Johnstone, C., Kuria, P., Christopher, N., Ramanarayanan, J., Shah, D., & van der Merwe, M. (2020). *Wiley 2020 interpretation and application of IFRS® standards*. John Wiley & Sons.

Bellotti, A., & Crook, J. (2009). Support vector machines for credit scoring and discovery for significant features. *Expert System with Applications*, 36, 3302-3308.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159.

Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 121–167. https://doi.org/10.1023/A:1009715923555

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, (pp. 233–240). https://doi.org/10.1145/1143844.1143874

de Mooij, R., Nazara, S., & Toro, J. (2018). Implementing a Medium-Term Revenue Strategy. In L. E. Breuer, J. Guajardo, & T. Kinda (Eds.), *Realizing Indonesia's economic potential* (pp. 109-140). International Monetary Fund.

Directorate General of Taxes. (2011). *Annual report 2010*.

Directorate General of Taxes. (2016). *Annual report 2015*.

Directorate General of Taxes. (2017). *Annual report 2016*.

Directorate General of Taxes. (2019). *Surat edaran Direktur Jenderal Pajak nomor SE-24/PJ/2019 tentang implementasi compliance risk management dalam kegiatan ekstensifikasi, pengawasan, pemeriksaan, dan penagihan di Direktorat Jenderal Pajak*.

Directorate General of Taxes. (2020). *Annual report 2019*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Ganguin, B., & Bilardello, J. (2005). *Fundamentals of corporate credit analysis*. McGraw-Hill.

Golin, J., & Delhaise, P. (2013). *The bank credit analysis handbook: A guide for analysts, bankers and investors*. John Wiley & Sons Singapore.

Gruber, J. (2011). *Public finance and public policy (*3rd ed.) Worth Publishers.

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

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.) Wiley.

International Monetary Fund. (2019). *Indonesia selected issues: IMF country report no. 19/251*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787.

Kieso, D. E., Weygandt, Jerry J., & Warfield, T. D. (2016). *Intermediate accounting* (16th ed.) Wiley.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*, 393–423.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the practice of statistics* (9th ed.) W.H. Freeman.

Musgrave, R. A. (1996). Progressive taxation, equity, and tax design. In J. Slemrod (Ed.), *Tax progressivity and income inequality* (pp. 341-356). Cambridge University Press.

OECD. (2004). *Compliance risk management: Managing and improving tax Compliance*. Centre for Tax Policy and Administration. https://www.oecd.org/content/dam/oecd/en/topics/policy-issues/tax-administration/compliance-risk-management-managing-and-improving-tax-compliance.pdf

OECD. (2006). *Strengthening tax audit capabilities: General principles and approaches*. OECD Publishing.

OECD. (2016). *Advanced analytics for better tax administration: Putting data to work*. OECD Publishing.

OECD. (2016). *Technologies for better tax administration: A practical guide for revenue bodies*. OECD Publishing.

OECD. (2019). *Tax administration 2019: Comparative information on OECD and other advanced and emerging economies*. OECD Publishing.

Provenzano, A. R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., ... & Nordio, C. (2020). *Machine learning approach for credit scoring*. arXiv. https://doi.org/10.48550/arXiv.2008.01687

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41-46.

Slemrod, J. (1996). *Tax progressivity and income inequality*. Cambridge University Press.

Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations* (2008 ed.) Oxford University Press.

van Liebergen, B. (2017). Machine learning: a revolution in risk management and compliance? *Journal of Financial Transformation*, *45*, 60-67.