ABHIJITH (U00928982)

 JAYASHREE RAMESH REDDY (U00928079)

**Municipal Debt Risk Analysis using Machine Learning**

Contents

**[Abstract](#_Toc165208243)** [2](#_Toc165208243)

**[1.](#_Toc165208244)****[Introduction](#_Toc165208244)** [2](#_Toc165208244)

[i. Problem Definition 2](#_Toc165208245)

[ii. Background 3](#_Toc165208246)

[iii. Related Works 3](#_Toc165208247)

[ Research Gap 4](#_Toc165208248)

[iv. Objectives and Contributions 4](#_Toc165208249)

**[2.](#_Toc165208250)****[Methodology](#_Toc165208250)** [4](#_Toc165208250)

[i. Data Collection and Preprocessing 4](#_Toc165208251)

[ Data Collection 4](#_Toc165208252)

[ Preprocessing 5](#_Toc165208253)

[ii. Model Building 5](#_Toc165208254)

[iii. Model Evaluation 5](#_Toc165208255)

[iv. Model Comparison and Selection 5](#_Toc165208256)

**[3.](#_Toc165208257)****[Model Descriptions](#_Toc165208257)** [5](#_Toc165208257)

[i. CatBoost 5](#_Toc165208258)

[ii. MLP Classifier (Multilayer Perceptron) 6](#_Toc165208259)

[iii. Random Forest 6](#_Toc165208260)

**[4.](#_Toc165208261)****[Results and Experiments](#_Toc165208261)** [7](#_Toc165208261)

[i. Database 7](#_Toc165208262)

[ii. Training and testing logs 8](#_Toc165208263)

[iii. Discussion and comparison 8](#_Toc165208264)

[CatBoost model performance 8](#_Toc165208265)

[MLP model performance 9](#_Toc165208266)

[Comparison 10](#_Toc165208267)

**[5.](#_Toc165208268)****[Conclusion](#_Toc165208268)** [10](#_Toc165208268)

[i. Limitations 10](#_Toc165208269)

**[6.](#_Toc165208270)****[References](#_Toc165208270)** [11](#_Toc165208270)

# **Abstract**

The research in this study carries out a detection process for bad debt in municipal finance systems using advanced machine learning algorithms. The dataset used in the study was obtained from Kaggle. The dataset contains invoices and payment records for eight South African municipalities over two years. This dataset contains many signs and flags that tag accounts likely to cause concern. For the ML models used in the research study, three models have been developed: CatBoost, Multilayer Perceptron (MLP) and Random Forest. CatBoost is a better model than the rest, since it attains an AUC of 1.0 with low misclassifications, showing good precision and dependability. In contrast, the MLP model showed a higher rate of misclassification with 847 false negatives and 2 false positives, and hence, a slightly reduced AUC of 0.99. The Random Forest model was more or less the same as CatBoost in bad debt detection but showed a higher rate of misclassification of non-bad debt. This work elaborates the potential of machine learning in assessment of financial risk, providing a method for unpaid debt prediction, and in a way improving the financial stability of municipalities. It also revealed major flaws in data variability, quality, and imbalanced data issues, indicating that future research should be performed to strengthen such predictive models.

**Key Words** – MLP, Random Forest, CatBoost, AUC, etc.

# **Introduction**

This project will apply advanced machine learning techniques in the prediction and management of risks related to municipal debt. There is an expectation that, in the future, municipalities will find it increasingly difficult to manage their finances in the provision of key services to their citizens. The project's goal is to develop an advanced analytics tool using machine learning models in assessing and predicting the risks associated with municipal debt. By taking this approach, it introduces a novel way of running the finances of municipalities, giving way for a proactive approach in place of a reactive one for countering economic challenges.

## Problem Definition

The focus of this project is on forecasting and controlling the potential for bad debts in the municipal financial system, which comprises a major role in keeping the financial stability of the system and delivering needed public services. To develop this application, which can predict the likelihood of accounts turning into bad debts, a next-stage analytical tool based on machine learning techniques will be created. With this tool, the ability for proactive risk management will be enabled.

The major challenges being addressed include:

**• Data Complexity:** Municipal finances capture a broad spectrum of data types and formats, making very difficult to predict account performance in the usual ways.

**• Importance of Complex Models:** There is a huge need for sophisticated data-driven models that would capture and forecast the ever-changing financial patterns of municipalities.

• **Operational Efficiency:** The tool is designed to bring actionable information that would help the municipalities manage risks in the most effective manner, without compromising on the quality of services.

This study would compare many advanced machine-learning models such as CatBoost, MLP, and Random Forest in order to come up with the most efficient way of predicting and controlling risks associated with municipal debt, which is achieved via improvement in the ability of municipalities to take decisions and encouraging financial accountability.

## Background

A greater number of municipal finances have become increasingly complex, with severe consequences of mismanagement, such as bankruptcy or a reduction of services, leading to the need for this analysis of municipal debt risk (Rich, Roberts and Zhang, 2021). Traditional models of risk estimation overlook many important elements related to the complex and dynamically changing financial condition of a municipality, leading to poor techniques of risk management. The growing influence of modern, data-driven methodologies seeks to overcome these limitations. Recent advancements in machine learning and data analytics provide new opportunities to improve the accuracy and effectiveness of risk management systems in the municipal finance industry.

## Related Works

The authors (Altman, 2020) used multiple discriminant analysis (MDA) to examine the variables that influence municipal credit rating. Four main input variables were used to create a discriminant function. While the obtained coefficients were significant, only 50% of the testing sample was identified accurately.

The same database was used by the authors (Ioannou, 2022) in an effort to enhance the outcomes. Regression analysis was used in place of MDA. A total of 150 municipalities from around the country were selected for the study, with 75 of them having investment grade ratings. The study outperformed the previous study by achieving an 80% accuracy rate in classifying the testing data.

The authors (Omstedt, 2019) used conventional machine learning methods with the help of multiple discriminant analysis method, linear regression as well as ordinal logistic regression as well as improved neural network techniques, like feed-forward neural networks technique, probabilistic neural network approach as well as cascade correlation neural networks to analyze the State of Connecticut's municipal bond dataset. They achieved performance of around 98.8% with the help of PNN approach on a four-class data set and 96.3% on a nine-class data set. Their findings validate that neural network models, including both publicly accessible non-financial and financial data have the potential to generate very accurate predictions of credit ratings. This is in spite of rating organizations' emphasis on the crucial role performed by the subjective judgment of experts. Also, they believe that a substantial quantity of observations is essential while constructing credit rating models.

The researchers (Lin et al., 2022) provided evidence that it is feasible and reasonable to forecast Moody's municipal credit ratings with a relatively high level of accuracy. By analyzing data from 25 prominent cities in the United States, they achieved an accuracy of 83% across five credit rating categories and 95% accuracy across two combined categories. Also, they noted that their sample size was rather limited. A larger dataset would provide even more exceptional outcomes.

The authors (Antulov-Fantulin, Lagravinese and Resce, 2021) employed the MDA technique which is based on discriminant approach and the ANN technique to analyze bond ratings across four industries which are electric utility domain, gas industries, manufacturing as well as telephone sectors. The researchers discovered that the ANN technique is capable of categorizing bond ratings into either three or four groups. Also, they observed that the ANN method achieves greater accuracy in predicting ratings compared to the MDA method. The MDA technique demonstrated superior accuracy of around 98 percent to nearly 100 percent accuracy for the training samples compared to the ANN method that gave an accuracy of 88 percent to 98 percent. During the testing phase, the Artificial Neural Network method demonstrated superior accuracy compared to the MDA method in three out of four industry samples such as gas, manufacturing as well as telephone sector.

### Research Gap

A study of the existing literature on municipal debt risk analysis reveals significant advances in applying machine learning to control financial risks. However, gaps remain in dealing with the complexity and high dimensionality of municipal financial data, which frequently affects the accuracy and application of predictive models. Previous investigations, such as those by Altman (2020) and Ioannou (2022), used methods such as multiple discriminant analysis and regression analysis but had limited success in terms of predictive accuracy and generalizability due to these limits.

Also, while neural network models have demonstrated great accuracy in studies such as Omstedt (2019), they necessitate large datasets and are frequently too complex for practical use in municipal settings. This study seeks to address these restrictions through using advanced machine learning models like CatBoost, MLP and Random Forest, which are capable of handling large and complex datasets. This project aims to improve the accuracy and resilience of predictive models while also incorporating them into a realistic analytical tool for extensive municipal debt management. This approach is intended to increase operational efficiency and deliver practical insights, thus filling a significant gap in the existing body of research.

## Objectives and Contributions

* **Identify Key Risk components:**
* **Objective:** This involves identifying the main components that contribute to the risk connected with municipal debt, in order to have a better knowledge of the reasons that can lead to financial challenges.
* **Contribution:** Create an advanced risk assessment methodology that employs predictive analytics to define and analyze risk elements such as economic instability and revenue collection efficiency. This methodology will help in the identification of areas prone to financial instability, allowing for more proactive risk management.
* **Provide actionable insights:**
* **Objective:** To provide municipalities with practical information that will assist them in reducing risks and improving their overall financial well-being.
* **Contribution:** Use machine learning analysis of financial data to generate precise and practical strategies that aim to optimize revenue streams and effectively manage expenses.
* **Improve Decision-Making:**
* **Objective:** Enhance the decision-making processes related to municipal debt management and financial planning by offering a dependable basis of data-driven insights.
* **Contribution:** Share insights for implementing an analytical decision-support system that uses the results generated by advanced machine learning models.

# **Methodology**

## Data Collection and Preprocessing

### Data Collection

This research used a dataset sourced from Kaggle, which consists of extensive billing and payment data from eight South African municipalities over a duration of two years. The dataset contains a range of fiscal and socioeconomic indicators that are important for understanding municipal finances.

### Preprocessing

The preprocessing stage includes multiple essential processes to ensure the data is appropriate for analysis:

* **Removal of Column:** To edit the DataFrame in place, the column **'accountcategoryid'** was dropped from the DataFrame 'df' using the drop() function with inplace=True.
* **Label Encoding:** Scikit-Learn's LabelEncoder was used to transform the DataFrame's categorical variables to numerical values. To do this, a LabelEncoder was fitted to each column, the columns were iterated over using object dtype, and the columns were transformed in place.
* **Feature Scaling:** StandardScaler from scikit-learn was only used to scale the numerical features. To do this, the numeric columns were chosen, the scaler was fitted to them, and they were then transformed to have a zero mean and unit variance.

## Model Building

During the model building stage of the Municipal Debt Risk Analysis project, the training set will be used to train three distinct types of machine learning models. Each model will process the same data in order to forecast the probability of an account turning into a bad debt. The chosen models for this project are:

* **CatBoost**
* **MLP Classifier (Multilayer Perceptron)**
* **Random Forest**

## Model Evaluation

The models are evaluated using various metrics to evaluate their predicted performance and practical applicability:

* **Accuracy, Precision, Recall and F1 Score:** These metrics provide a complete picture of the models' performance, particularly their ability to correctly identify accounts as bad debts or not.

## Model Comparison and Selection

In the last stage, the performance of the CatBoost, MLP Classifier and Random Forest models is compared using evaluation criteria. The model that shows an ideal balance between accuracy, reliability and practical usability is chosen for deployment.

# **Model Descriptions**

## CatBoost

CatBoost is a type of supervised machine learning technique that uses the algorithm of gradient boosting on the decision trees. It excels in effortlessly managing categorical data without the need for significant preprocessing, which is usually necessary for other machine learning techniques (Liu et al., 2020). CatBoost is specifically engineered to deliver a combination of exceptional performance and speed, making it exceptionally effective for addressing both regression and classification tasks. The primary features of this system include the efficient management of large data sets, the inclusion of built-in techniques to minimize overfitting and the ability to automatically handle categorical variables.

**Advantages:**

* **Handling Categorical Variables:** Unlike many other machine learning algorithms, CatBoost handles categorical variables naturally and does not require considerable pre-processing.
* **Robust to Overfitting:** It is more stable when working with complicated datasets since it includes built-in techniques to prevent overfitting.
* **Efficiency:** Due to simplified implementation and GPU support it can train and predict at high speeds even on big datasets.

**Disadvantages:**

* **Parameter adjustment:** While CatBoost is easier to use out of the box than other gradient boosting models, reaching peak performance may still necessitate precise parameter adjustment.
* **Complexity:** The algorithm is complex and beginners may struggle to understand the entire range of its functionality and parameters.

## MLP Classifier (Multilayer Perceptron)

The multilayer perceptron is a type of feedforward ANN model based on neurons. It has at least three layers of nodes which consists of an input layer, a hidden layer as well as an output layer. Except for the input nodes, every node is a neuron with a nonlinear activation function (Li et al., 2019). MLP trains using the supervised learning approach which is also termed as backpropagation. MLP is different from linear perceptrons in that it has numerous layers and uses the approach of non-linear activation. It can identify data that is not linearly separable and provides a considerable benefit over simpler models.

**Advantages:**

* **Flexibility:** Its layered structure and use of non-linear activation functions enable it to model complex and non-linear relationships.
* **Scalability:** With more data and deeper structures, performance can be improved.
* **Generalization:** The ability to generalize well from situations if properly trained with appropriate data and regularization.

**Disadvantages:**

* **Prone to overfitting:** Particularly when working with small datasets or when appropriate regularization methods are not used.
* **Computational demands:** The process of training might need significant computational resources, especially when the network architecture grows bigger and more complex.

## Random Forest

Random Forest is supervised approach that depends on ensemble learning technique which can handle both discrete and continuous outputs. It works by generating several decision trees during training (Sheykhmousa et al., 2020). The Random Forest algorithm outputs the class that is chosen by the majority of the trees for classification tasks. The model is highly efficient in machine learning and offers strong prediction capabilities, minimal overfitting and straightforward interpretability. Despite the existence of highly correlated features, this technique remains valuable. Random Forest is capable of handling both continuous and categorical data. It offers a reliable approach for predicting missing data even when a significant amount of the data is missing and without compromising accuracy.

**Advantages:**

* **Performance:** Typically provides high performance for a wide range of problems right out of the box and without the need for substantial hyperparameter adjustment.
* **Interpretability:** Because each tree in the forest can be visualized and understood, this model is more interpretable than other black-box models.
* **Handling Different Data Types:** Effectively handles numerical as well as categorical data.

**Disadvantages:**

* **Prediction Delay:** The model's prediction speed may be slow, particularly when the forest's tree count grows.
* **Complexity in Large Forests:** As a forest gets very big, it can be challenging to manage, visualize and understand the model.

# **Results and Experiments**

## Database

Over a two-year period, this data was taken from the billing systems of eight South African municipalities and summarized based on the total amount invoiced as opposed to the total amount collected. There is a flag indicating if a certain account led to a bad debt for each of the accounts.

The purpose of this classification exercise is to ascertain whether it is possible to estimate the likelihood that an account will turn into a bad debt in order to anticipate the quantity (and value) of accounts that are susceptible to becoming bad debts.

The description of the variables is given as follows

|  |  |
| --- | --- |
| **AccCategoryID** | (Account Category ID) The numeric link in the database to the Account Category |
| **AccCategory** | (Account Category) A classification of the type of account |
| **AccCategoryAbbr** | (Account Category Abbreviation) An abbreviation of the classification of the type of account - to be used for One-hot encoding |
| **PropertyValue** | (Property Value) The market value of the property |
| **PropertySize** | (Property Size) The size of the property in square metres |
| **TotalBilling** | (Total Billing) The total amount billed to the account for all services |
| **AverageBilling** | (Average Billing) The average amount billed to the account for all services |
| **TotalReceipting** | (Total Receipting) The total amount receipted to the account for all services |
| **AverageReceipting** | (Average Receipting) The average amount receipted to the account for all services |
| **TotalDebt** | (Total Debt) The Total Debt that is at 90 days or more |
| **TotalWriteOff** | (Total Write Off) The Total amount of debt that has been written off |
| **CollectionRatio** | (Collection Ratio) The ratio between the Total Receipting and Total Billing (ie. Total Receipting/Total Billing) |
| **DebtBillingRatio** | (Billing Debt Ratio) The ratio between the Total Debt and Total Billing (ie. (Total Debt + Total Write Off)/Total Billing) |
| **TotalElectricityBill** | (Total Electricity Bill) The total amount billed for electricity. This field was put in place because it is used as a means to recover debt - ie. If an amount is outstanding for any service the municipality has the right to cut a consumer's electricity connection. |
| **HasIDNo** | (Has ID No.) The consumer has an ID number. This is similar to a Social Security number in the US and can be useful in legal proceedings. A consumer without any ID No. details is a lot harder to collect debt from. In addition, this field denotes that the account is held by a person and not a business. However, it is not very reliable as it's often not captured properly or at all. |
| **BadDebtIndic** | (Bad Debt Indicator) 1 = Is considered to be a Bad Debt, 0 = Not considered to be a Bad Debt |

Table 1: Description of the data

## Training and testing logs

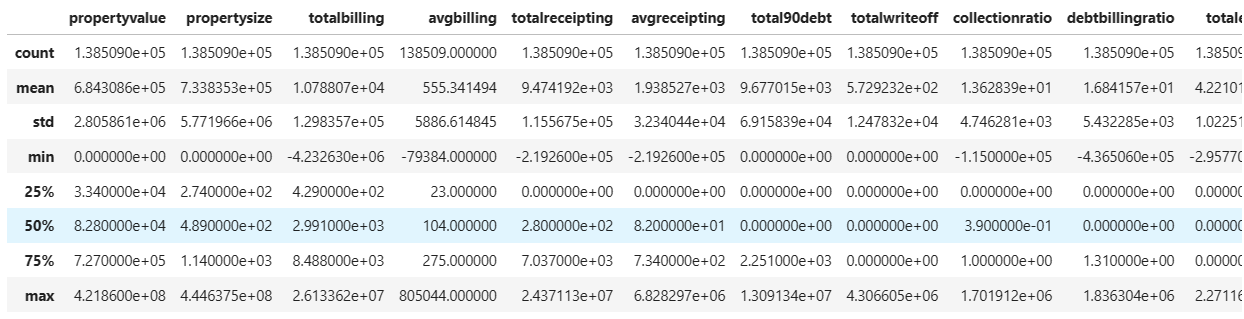


Figure 1: Descriptive statistics

The 90-day debt has an average total debt of around $9,677, with a $69,158 standard deviation. The total amount of debt varies quite a bit between accounts. Around $573 is the average total write-off, with a standard deviation of $12,478. This implies that the accounts have written off different amounts of debt. The percentage of billed amounts that are successfully received is shown by the collection ratio which is calculated by dividing total receipting by total billing. On average, this ratio is 13.63%. With an average of 16.84%, the debt billing ratio shows what percentage of billing amounts end up as debt.

A red and blue squares

Description automatically generated A red and blue rectangular shapes

Description automatically generated

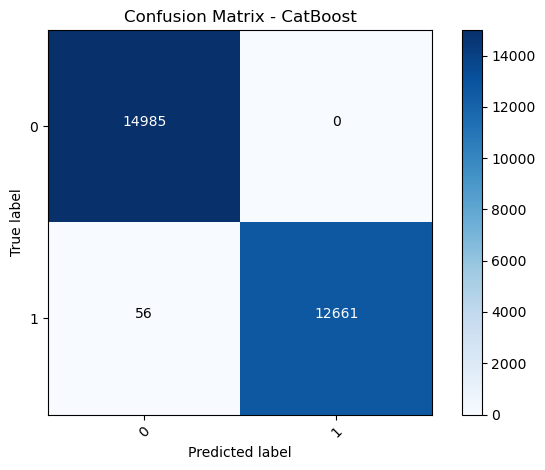
Figure 2: Average property value and size based on debt

The plot indicates that properties with bad debt have significantly lower average values compared to those without bad debt. Also, the property with debt have significantly lower size compared to those with no bad debt. These two variables can be significant predictors for determining bad debt.

## Discussion and comparison

The models are compared based on confusion matrix which is derived from prediction of bade debt by each models in test data. Other metrics such as weighted precision, recall and F1-score is also used to measure the amount of errors produced by each model.

### CatBoost model performance

A graph of a catboost

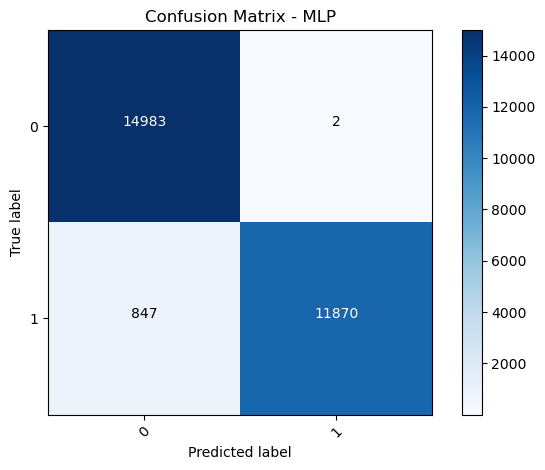
Description automatically generated

Figure 3: ROC curve and confusion matrix of CatBoost model

Correctly, the model detected 14985 samples that had no bad debt. No samples were incorrectly identified by the model as having bad debt when in fact they did not. 12661 samples with bad debt were accurately recognized by the model. When 56 samples had bad debt, the model incorrectly identified them as having none. The model appears to be highly effective at recognizing bad debt, with minimal misclassification, based on the low number of false negatives.

The ROC curve is a straight line positioned horizontally at a value of 1.0, which signifies flawless differentiation between the two categories. The Area Under the Curve (AUC) is 1.0 indicating optimal performance without any false positives.

### MLP model performance

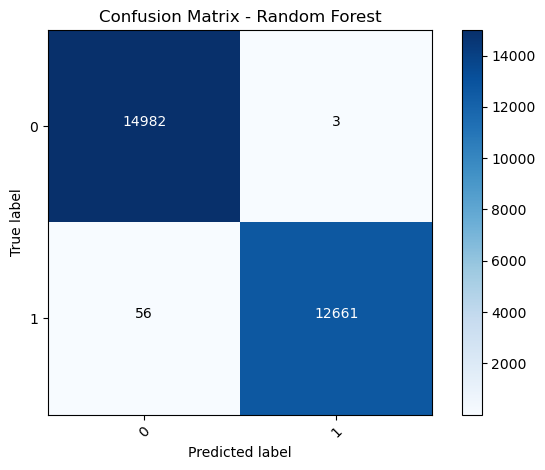
A graph of a curve

Description automatically generated

Figure 4: ROC curve and confusion matrix of MLP model

The ROC curve and the confusion matrix suggests that MLP misclassifies more bad debt samples compared to CatBoost model where 2 non bad debt samples had been misclassified as bad debt and 847 bad debt samples had been misclassified as non bad debt. The AUC score is 0.99 which is also indicating excellent distinguishing performance but is weaker compared to CatBoost model.

### Random Forest performance

A graph of a curve

Description automatically generated

Figure 5: ROC curve and confusion matrix of Random Forest

For classification of bad debts, random forest performed similar to CatBoost with similar amount of misclassification but considering non bad debts, the model produced misclassification compared to the CatBoost model.

### Comparison

This study assessed the predictive performance of three models in estimating the occurrence of bad debts using a dataset obtained from the billing systems of eight South African municipalities over a span of two years. The models included in this set are CatBoost, MLP, and Random Forest.

The performance of CatBoost in forecasting bad loans was outstanding, as evidenced by its confusion matrix which had just 56 false negatives and zero false positives. This demonstrates a significant level of precision and dependability. The ROC curve exhibited impeccable discrimination between the two groups, achieving an AUC of 1.0, indicating perfect performance.

The MLP (Multi-Layer Perceptron) exhibited worse performance in comparison to CatBoost, with a higher number of false negatives and false positives. The confusion matrix revealed that the MLP model incorrectly identified 2 samples as having bad debt when they were actually free of it and misclassified 847 samples without bad debt as having bad debt. While the AUC score of 0.99 is still rather good, it is comparatively lower than that of CatBoost.

Random Forest and CatBoost demonstrated comparable performance, with equal rates of misclassification in recognizing bad loans. But it indicated a higher percentage of misclassification when evaluating non-bad loans, suggesting a somewhat diminished level of accuracy in correctly classifying all instances.

# **Conclusion**

The analysis findings demonstrate that the CatBoost model has superior performance in accuracy compared to both the MLP and Random Forest models, particularly in discriminating between bad debts and non-bad debts. This is apparent from its excellent Receiver Operating Characteristic (ROC) curve and the minimal amount of misclassifications. The exceptional accuracy of CatBoost in identifying delinquent bills makes it the optimal choice for this particular dataset and problem space.

## Limitations

Although the research showcases the performance of various models, it is important to acknowledge a few limitations:

* **Data Variability:** The dataset shows a broad spectrum of property values and other factors suggesting significant variability. This variability may have implications for the model's capacity to be applied to other settings or datasets with accuracy and reliability.
* **Data quality:** It refers to the precision and accuracy of information. In this case, if there are missing or incorrectly recorded data, it might result in inaccurate predictions or biases in the model.
* **Imbalanced Data:** Imbalanced dataset distribution, specifically with regards to bad debts and non-bad debts might have a negative influence on the model's performance especially in accurately forecasting the minority class.
* **Overfitting:** Models that achieve high AUC values may be susceptible to overfitting indicating the importance of thorough validation to guarantee the models perform well on new and unexplored data.

Given these limitations, future research might focus on investigating approaches to tackle data quality concerns, maintaining a balanced dataset and evaluating models using other data sources to assure reliable generalization. Also, the use of cross-validation and other methodologies to prevent overfitting could improve the dependability and efficiency of the models.

# **References**

Altman, E.I. (2020). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance*, 23(4), pp.589–609. doi:https://doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Antulov-Fantulin, N., Lagravinese, R. and Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization*, 183, pp.681–699. doi:https://doi.org/10.1016/j.jebo.2021.01.014.

Ioannou, S. (2022). Regional and municipal debt in the Eurozone: a cross-country analysis. *Regional Studies*, pp.1–15. doi:https://doi.org/10.1080/00343404.2022.2047914.

Li, D., Huang, F., Yan, L., Cao, Z., Chen, J. and Ye, Z. (2019). Landslide Susceptibility Prediction Using Particle-Swarm-Optimized Multilayer Perceptron: Comparisons with Multilayer-Perceptron-Only, BP Neural Network, and Information Value Models. *Applied Sciences*, 9(18), p.3664. doi:https://doi.org/10.3390/app9183664.

Lin, X., Chen, S., Cheng, X. and Wang, J. (2022). Local government audit and municipal debt risk: Evidence from audit reform in China. *Finance Research Letters*, 50, p.103198. doi:https://doi.org/10.1016/j.frl.2022.103198.

Liu, W., Deng, K., Zhang, X., Cheng, Y., Zheng, Z., Jiang, F. and Peng, J. (2020). A Semi-Supervised Tri-CatBoost Method for Driving Style Recognition. *Symmetry*, 12(3), p.336. doi:https://doi.org/10.3390/sym12030336.

Omstedt, M. (2019). Reading risk: The practices, limits and politics of municipal bond rating. *Environment and Planning A: Economy and Space*, 52(3), pp.611–631. doi:https://doi.org/10.1177/0308518x19880903.

Rich, K.T., Roberts, B.L. and Zhang, J.X. (2021). Linguistic tone of management discussion and analysis disclosures and the municipal debt market. *Journal of Public Budgeting, Accounting & Financial Management*, 33(4), pp.427–446. doi:https://doi.org/10.1108/jpbafm-11-2019-0168.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P. and Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, pp.6308–6325. doi:https://doi.org/10.1109/jstars.2020.3026724.