МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет «Києво-Могилянська академія»

Факультет економічних наук

Кафедра фінансів

Кваліфікаційна робота

освітній ступінь - бакалавр

на тему: "Scoring models in solvency evaluation" / "Скорингові моделі в оцінці кредитоспроможності"

Студента 4 курсу

галузь знань 07 «Управління та адміністрування»

спеціальність 072 «Фінанси, банківська справа та страхування»

Недождія М.А.

Науковий керівник: канд. економ. наук, PhD Литвин Антон Валерійович

Рецензент Шумська С.С.

Кваліфікаційна робота захищена

з оцінкою «_____»

Секретар ЕК _____

«____» __ _____ 2022 р.

Київ – 2022

Contents

Abstract

Ukrainian financial institutions and banks currently face a problem of a high number of non-performing loans. This study aims to propose a solution to that problem, by creating a machine learning-enabled credit scoring model, training it on an open source dataset, and applying it on to a set of potential Ukrainian borrowers.

The hypothesis is that the methods of credit scoring used by financial institutions in Ukraine are ineffective and could be vastly improved by implementing decision tree machine learning algorithms into day-to-day operations to increase the accuracy of default probability for individual borrowers. The results show that while these methods can be successfully applied for Ukrainian borrowers, the dataset to train the algorithm on has to be carefully picked to fit with the information that you can easily collect during a credit application process.

These results suggest that if the proposed prediction algorithms are trained on a diversified dataset, they can vastly reduce the amount of NPLs being given out by banking institutions in Ukraine.

Keywords: credit scoring, machine learning, banking, decision tree, gradient boosting

Анотація

Українські фінансові установи та банки зараз стикаються з проблемою великої кількості проблемних кредитів. Це дослідження має на меті запропонувати вирішення цієї проблеми шляхом створення моделі кредитного скорингу за допомогою машинного навчання, навчання її на загальнодоступному наборі даних та застосування на наборі даних потенційних українських позичальників.

Наша гіпотеза полягає в тому, що методи кредитного скорингу, які використовуються фінансовими установами в Україні, неефективні та можуть бути значно покращені шляхом впровадження алгоритмів машинного навчання (дерева рішень) для підвищення точності передбачення ймовірності дефолту для позичальників. Результати показують, що хоча ці методи можна успішно застосовувати для українських позичальників, набір даних для навчання алгоритму має бути ретельно відібраний, щоб він відповідав інформації, яку ви можете легко зібрати під час процесу отримання заявки на кредит.

Ці результати свідчать про те, що якщо запропоновані алгоритми прогнозування навчати на наборі даних, що рівномірно розподілений, вони можуть значно зменшити кількість проблемних кредитів, які видають банківські установи в Україні.

Ключові слова: кредитний скоринг, машинне навчання, банківська справа, дерево рішень, градієнтне підсилювання

Introduction

The main objective of credit scoring is to develop models, which split the loan applicants depending on how likely they are to experience financial distress/default on a scale of 0 to 1. Giving such a score to loan applicants allows the financial institution providing the loan to analyze the information provided by the applicant and accurately evaluate the probability of an individual returning the loan.

Throughout the years there have been a number of classification techniques adopted for credit scoring, such as logistic regression and discriminant analysis. While these models have existed since the 20th century, lately their development and application costs were able to be vastly reduced by using hardware-enabled analytical methods – machine learning. Many papers have been written on ML application for credit scoring in foreign financial institutions. In this paper, an attempt will be made to apply information collected by those foreign institutions and use it to predict financial distress amongst potential Ukrainian borrowers.

The goal of this paper is to attempt to use open-source US-based borrower data as a training dataset for a decision tree model and apply that model to a set of Ukrainian borrowers. High accuracy results in this use case would mean that such methods could be applied in banks and financial institution in Ukraine for default prediction of potential borrowers.

This paper contributes to the literature dedicated to ML-enabled credit scoring on the issue of applying such models to a real life dataset of potential borrowers in Ukraine to evaluate how effective it would be to deploy these methods of credit scoring in newly created or already existing financial institutions.

Relevance of the topic. Considering the current economic situation in Ukraine and the amount of non-performing loans in Ukrainian banks' credit portfolios (which will be discussed later), they are in a desperate need of maximizing the efficiency of their credit scoring procedures. Simultaneously, there are a lot of financial companies that provide loans outside of Ukraine's banking system, who could also greatly benefit from increasing the accuracy of their default predictions.

The purpose and main objectives of the paper. The main purpose of this paper is to propose an accessible alternative to the credit scoring methods used by Ukrainian financial institutions and evaluate the efficiency of those alternative methods. This paper has the following main objectives:

- To evaluate the problem of NPLs in Ukrainian banks;
- To review the history of credit scoring applications in banking;
- To review the recent developments in credit scoring applications;
- To build a model that can be easily applicable for banks, which performs on par with existing solutions or better;
- To find ways to further improve this model in case of its application in real-life financial institutions.

The paper starts with an overview of the literature written on the subject of credit scoring. Next, the methods used to create the decision tree scoring model are discussed, which is followed by a review of the results of the modeling process and suggestions on how the performance of the model can be improved.

Literature Review

Credit scoring is an essential part of day-to-day operations for most financial institutions throughout the world. Its importance lies in the necessity to separate loan applicants depending on how likely the probability of them paying back the loan is going to be [1]. The main goal of going through such a process is to estimate the probability of default of an individual or company, i.e., the event of a customer not paying back a loan in a given period [2].

Credit insolvency prediction is extremely important when applied to real-life debt. For example, the outstanding debt to nonfinancial businesses in the United States was about 17.7 trillion USD at the end of 2020, which means that an improvement in default prediction accuracy by just a couple of percentage points will potentially lead to tens of billions of dollars in savings [3, 4].

Credit insolvency in Ukraine is also a very prominent issue. The percentage of NPLs (short for Non-Performing loans, the loans that are over 90 days due) was at 30% as of January 1st 2022, however I would expect that this number is currently much higher due to the ongoing war [5]. This number, compared to NPL percentages amongst developed countries, is critically high, as the NPL ratio in EU countries sits at 2.06% as of Q4 2021, and at 1.07% and 0.53% as of Q4 2020 for the United States and Canada respectively [6, 7].

What also needs to be taken into consideration is that over 70% of NPLs in Ukraine are in Government-owned banks, where 47.1% of the loans were Non-Performing as of January 1st 2022 with the biggest Government-owned bank, PrivatBank, having 69.9% of their credit portfolio in NPLs [5]. I believe that such high numbers are caused by inefficient (and possibly severely outdated) credit scoring methods that were used when providing these loans. I believe that the methodology for providing loans in Ukraine (especially in

Government-owned banks) needs to be revised and improved in order to help bring the number of NPLs to less than 10%.

In the past decades a multitude of different classification techniques have been used for credit scoring. These techniques include traditional statistical methods (e.g. logistic regression), non-parametric statistical models (e.g. k-nearest neighbor and decision trees) and neural networks [8]. In this paper the focus will primarily be on machine learning-enabled decision trees, and how their low data requirements allow them to be used more effectively than traditional statistical methods when faced with a lack of historical data.

The idea of Machine Learning dates back to the mid-20th century, when the term was first coined in an IBM article by Artur Lee Samuel titled "Some Studies in Machine Learning Using the Game of Checkers" [9]. The early adoptions of this idea were incredibly hardware and software limited and I will touch on this topic later in this review. These early adoptions were mostly based on basic supervised learning algorithms, which, because of the hardware limitations, could only be implemented by major corporations, such as IBM.

One of the more public and significant applications of such algorithms were essentially hardware enabled bots that would learn how to play a certain board game. For example, in 1989 a supercomputer called Deep Blue was developed by IBM with a single function – playing chess.

It used evaluation functions to determine what moves to make, being fed a database of grandmaster chess matches and being able to evaluate 200 million positions per second [10, 11]. More information about this early adoption of AI can be found in Feng-Hsiung Hsu's book "Behind Deep Blue" and Steven Strogatz's article "One Giant Step for a Chess-Playing Machine". Moving on from general Machine Learning applications, let's dive in deeper into specific articles and publications regarding Scoring models.

Moving on to ML model's applications, while there is a lot of material written on the topic of Data Science/Machine Learning and it's general concepts, such as Foster Provost's and Tom Fawsett's "Data Science for Business" and research papers by the likes of Deloitte, literature regarding Scoring models applications is limited, and usually written from a technical perspective, not from a financial one [12, 13]. The former work briefly touches on Credit Scoring models, but mostly uses it as an example of applying data mining in finance, while the latter paper mentioned ("Business impacts of Machine Learning" by Deloitte) mostly describes generalist concepts within data science and how they could be applied in business, not touching on how to actually apply them.

Foster Provost and Tom Fawsett take a lot of their information regarding Credit Scoring models from a publication from Branko Soucek and The IRIS Group by the name of "Neural and Intelligent Systems Integration" published in 1991 [14]. While this work pre-dates the widespread application of Machine Learning, it gives a very important glimpse into how the history of scoring models evolved from a point where creating a basic prediction model required such hardware as evaluation boards.

Credit scoring models were one of the first fields of application when machine learning models became more widely accessible as there were less hardware limitations in the 80s and the 90s [15]. The work by multiple authors from HAL Open Science titled "Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds" goes deep into the history of credit scoring applications and how developments in fields of decision trees, k-nearest neighbors analysis, neural networks and support vector machines (SVMs) have helped make significant progress in getting credit scoring models to having a wide range of applications in the Financial Services industry.

The work, as evident from its title, also touches on a very important topic of machine learning applications compared to "standard" econometrics methods. The authors highlight that while ML methods (specifically, random forest models) largely outperform logistic

regressions and have become the standard choice for creating credit scoring models for banking, they have a major drawback, which is lack of explainability and interpretability in a sense that the credit approval/disproval process cannot be easily explained to customers and regulators.

The widespread application of Machine Learning models in banking has become evident as technology and financial services companies are absorbing 60% of AI talent worldwide, according to a paper by David Kelnar from MMC Ventures titled “The State of AI: Divergence” published in 2019 [16]. While “divergence” is a great term to describe the modern AI climate, within the Financial Services industry ML/AI applications mostly differ in methods and approach to execution of those methods, rather than the end goal, which is usually to measure credit risk when giving a loan to an individual or a company.

Delegating a part of the due diligence process to Machine Learning models has become a prominent trend in the Financial Services industry in recent years and one of the leading risk assessment analytics companies – FICO – have been utilizing such ML-enabled methods as decision trees to accurately determine a subject’s credit score. In a paper published by FICO in 2018 titled “Machine Learning and FICO Scores” they dive into their methods of determining credit risk using scorecards and how ML models perform compared to their own model of evaluating credit risk [17].

The results of their comparison suggested that ML-only credit risk models “are not equipped to counteract the significant selection biases due to truncation and cherry-picking that exist in unscorable populations”, and because of that possibility of biased predictions might still require human expertise. However, when making a direct comparison of how their model and an SGB-based and Neural Net-Based models perform side by side, the results are almost identical. As shown in Table 1, the ROC (Receiver operating characteristic) and KS Test scores for ML models and the FICO model are almost identical.

Table 1 FICO Score performance compared to ML-only methods

Score	Metric	Value
FICO® Score 9	ROC	.901
	KS	63.2
SGB-Based	ROC	.905
	KS	64.3
Neural Net-Based	ROC	.901
	KS	63.5

Reference: Built by the author based on [17]

While FICO themselves call these "measurable but modest differences", their scoring system uses a sample of millions of credit files, which in return should give their model a significant advantage, which, as evident from the data above, it doesn't have [14]. And considering that FICO Scores are used by 90% of the top US lending institutions for their risk assessment needs, I would say that even considering the need for human expertise post-risk assessment, Machine Learning models seem to be an a lot more accessible entryway to credit scoring, without the need of a large dataset [18].

Judging from FICO's evaluation of ML method application for credit scoring, SGB (short for Stochastic Gradient Boosting) performs better, when compared to Neural Net-Based models. Because the exact methods to calculate FICO scores are unknown, a direct performance comparison can't be made. However, considering that logistic regression is a widely used method for credit scoring, a direct comparison of pros and cons of Gradient Boosting methods and Logistic Regressions was made (Table 2).

Table 2 Pros and Cons of Logistic Regression and Gradient Boosting methods

	Logistic Regression	**Gradient Boosting**
Pros	Direct Parameters Interpretation	Fast Development
	Easy Deployment	Stress of parameters
	Greater stability over time	Flexibility (Target: Binary / Multinomial/ Interval)
	Business view	
	Market Confidence	
Cons	Development effort	Needs implementation environment
	Manual Fit (Iterations)	New Methodology for the Market
		Low variable interpretation

Reference: Built by the author based on [19]

When applying the information above to real-life use cases, logistic regression would be easier to deploy in an average financial institution, considering that it is a method that is easier to interpret and deploy without changing much in the company's pre-existing credit scoring algorithms. However, it could be argued that for newly created financial services companies (e.g. FinTech startups) creating the implementation environment and deploying gradient boosting credit scoring methods for their day-to-day operations would be relatively easy, since the methodology could be built from the ground-up without having to reorganize their data, since it hasn't been collected yet. Such models could be trained on publicly available datasets and provide a high prediction accuracy without having to collect data before deploying the model.

Overall, the above mentioned works provide a good understanding of general concepts within Machine Learning and, more specifically, in scoring models. I think there is a gap in the material written on the subject, which is that there aren’t any papers written to my knowledge about actual application of Decision Tree Credit Scoring models for the Ukrainian market. This gap is what I will try to fill with this paper by creating a Decision Tree Credit Scoring model and applying it to a dataset of my own.

Methods and Materials

To create the decision tree model, a dataset from the public domain was used, which features 150000 records of individual US borrowers. US-based data was used as an example of how the created credit scoring model would fare when applied to borrowers in developed economies, while also making it easier for future research since most of the publicly available borrower data is generated in the US.

It was applied to build a credit scoring model to predict the probability of those individuals experiencing financial distress. This dataset will be used to train a ML model using Gradient Boosting methods and applying the final model to a dataset of Ukrainian borrowers. The dataset features 11 variables.

The target variable is **SeriousDlqin2yrs** (represented in binary units, 0/1), which indicates whether the person experienced 90 days past due delinquency or worse, and the independent variables are the following:

- **RevolvingUtilizationOfUnsecuredLines –** Total balance on credit cards and personal lines of credit divided by the sum of credit limits;
- **Age** – Age of borrower in years;
- **NumberOfTime30-59DaysPastDueNotWorse –** Number of times borrower has been 30-59 days past due in the last 2 years;
- **DebtRatio –** Monthly debt payments, alimony, living costs divided by monthly gross income;
- **MonthlyIncome –** Monthly income of the borrower;
- **NumberOfOpenCreditLinesAndLoans –** Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards);

- **NumberOfTimes90DaysLate –** Number of times borrower has been 90 days or more past due;
- **NumberRealEstateLoansOrLines –** Number of mortgage and real estate loans including home equity lines of credit;
- **NumberOfTime60-89DaysPastDueNotWorse –** Number of times borrower has been 60-89 days past due but no worse in the last 2 years;
- **NumberOfDependents –** Number of dependents in a family excluding themselves (spouse, children etc.) [20].

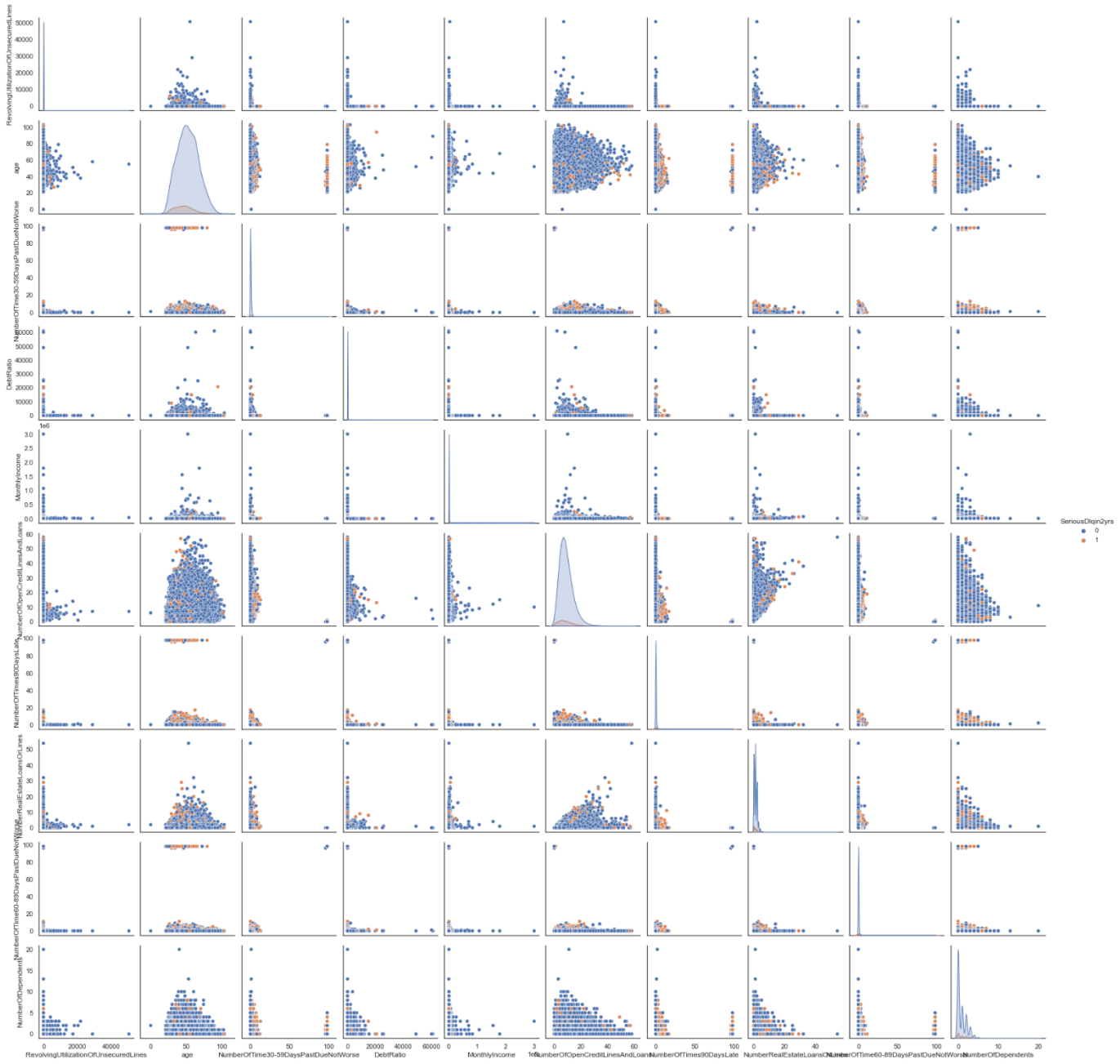
The model was created in Python using the Jupyter Navigator GUI. Out of 150000 individuals present in the training set, 29731 have no tracked monthly income and 3924 have no tracked dependents in the family, so those individuals were dropped from the dataset, which resulted in having 120269 individuals to work with.

To create the decision tree itself the scikit-learn machine learning library for Python was used [21]. Throughout the process two datasets were used: the *training* dataset, which features all of the above mentioned variables and the *test* dataset, which is a separate dataset of US borrowers to test the model on. The *test* dataset features 101503 records.

Overall, most of the dataset used is represented with individuals, for whom the *SeriousDlqin2yrs* variable is 0. Out of 120269 records, 8357 people will experience 90 days past due delinquency or worse, while 111912 will not. This might lead to an unwanted bias of the model towards predicting that most people will be able to pay back the loan in time.

To analyze the data collected before creating the decision tree a pair plot function provided by the seaborn package installed was used. This graph would allow us to see pairwise relationships in a dataset by creating a grid of Axes, such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column [22]. The pair plot for the dataset used is shown in Figure 1.

Figure 1 Pair Plot for the chosen dataset



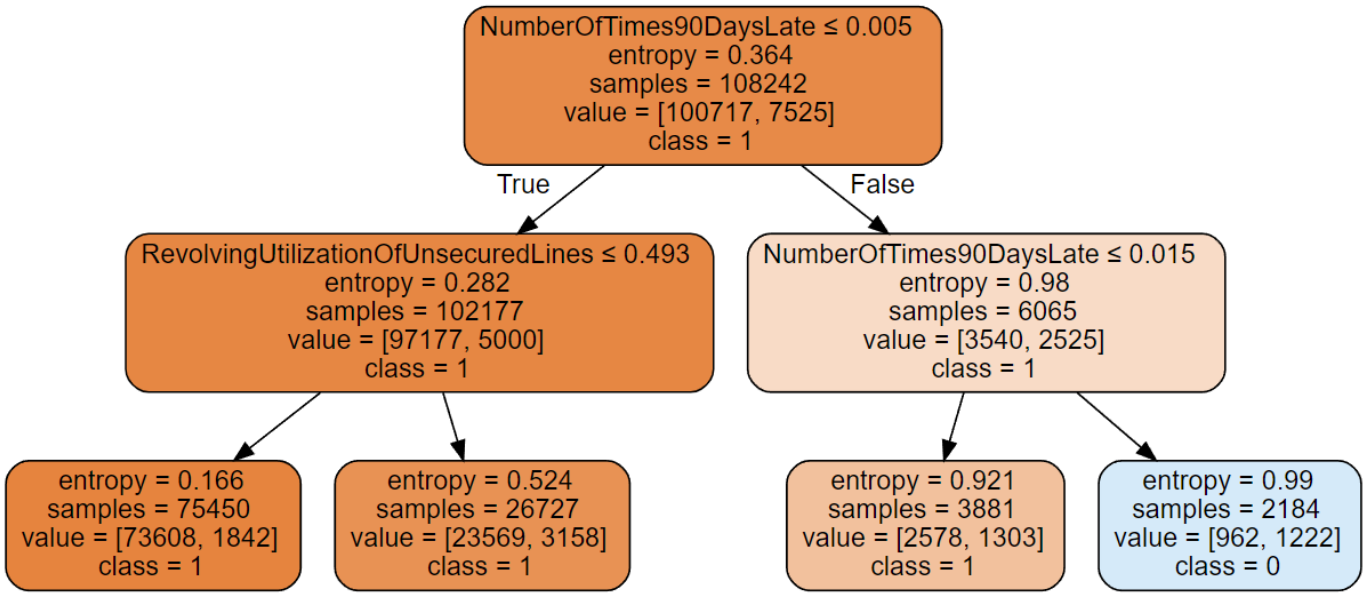
Reference: Author's developments

As can be observed from the visual representation of the data, most individuals have experienced 90 days past due delinquency or worse and there are no strong linear correlations between any of the variables in the dataset. In order to continue working on the dataset all

of the variables that can go above 1 to the 0-1 range were scaled to increase the model's accuracy.

Moving on to creating the decision tree itself using the entropy criterion (Figure 2). As can be seen from the graph, the model considers the number of times the borrower has been 90 days or more past due and the total balance on credit cards and personal lines of credit divided by the sum of credit limits as the two most important features in the dataset, with 0.67948 and 0.32 feature importance respectively.

Figure 2 Decision tree for the chosen dataset

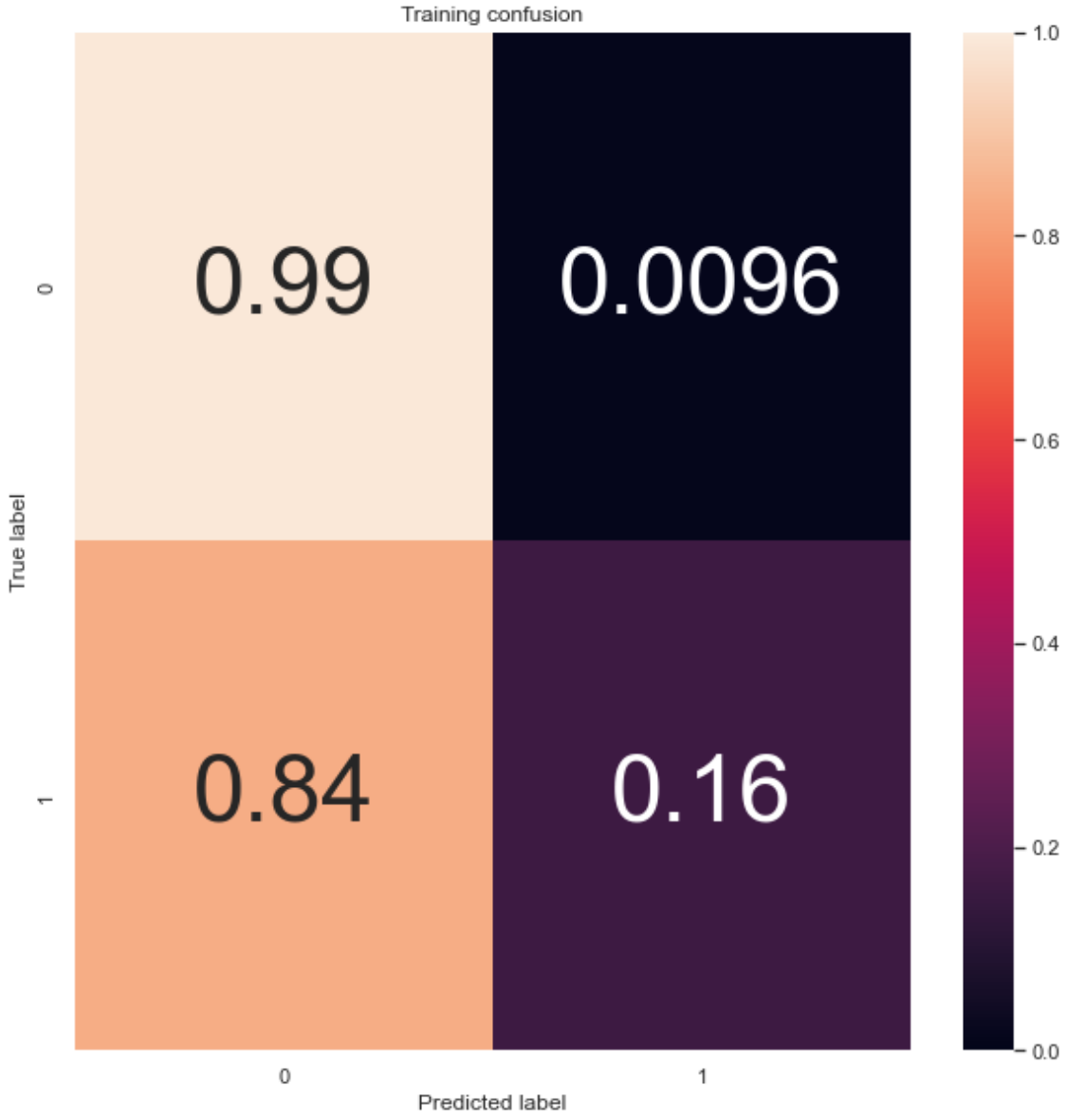


Reference: Author's developments

The accuracy score given to the model in this state is 0.93532 with the *training* dataset and 0.935 with the *test* dataset. To get a more accurate measurement of the model's accuracy confusion matrices with a Random Forest Classifier were used, which allowed us to compare actual and predicted classifications done by the model (Figure 3). The model was tuned before applying gradient boosting by testing different max depths of the tree (ranging from 2 to 5) and by using different *max_features* values, which is the number of features to consider when looking for the best split [23]. Unfortunately, all the tested combinations of

the two variables mentioned above resulted in an insignificant improvement, with the accuracy of predicting label "1" ranging from 0 to 19, compared to the 16% accuracy before tuning.

Figure 3 Recall confusion matrix of the model before tuning

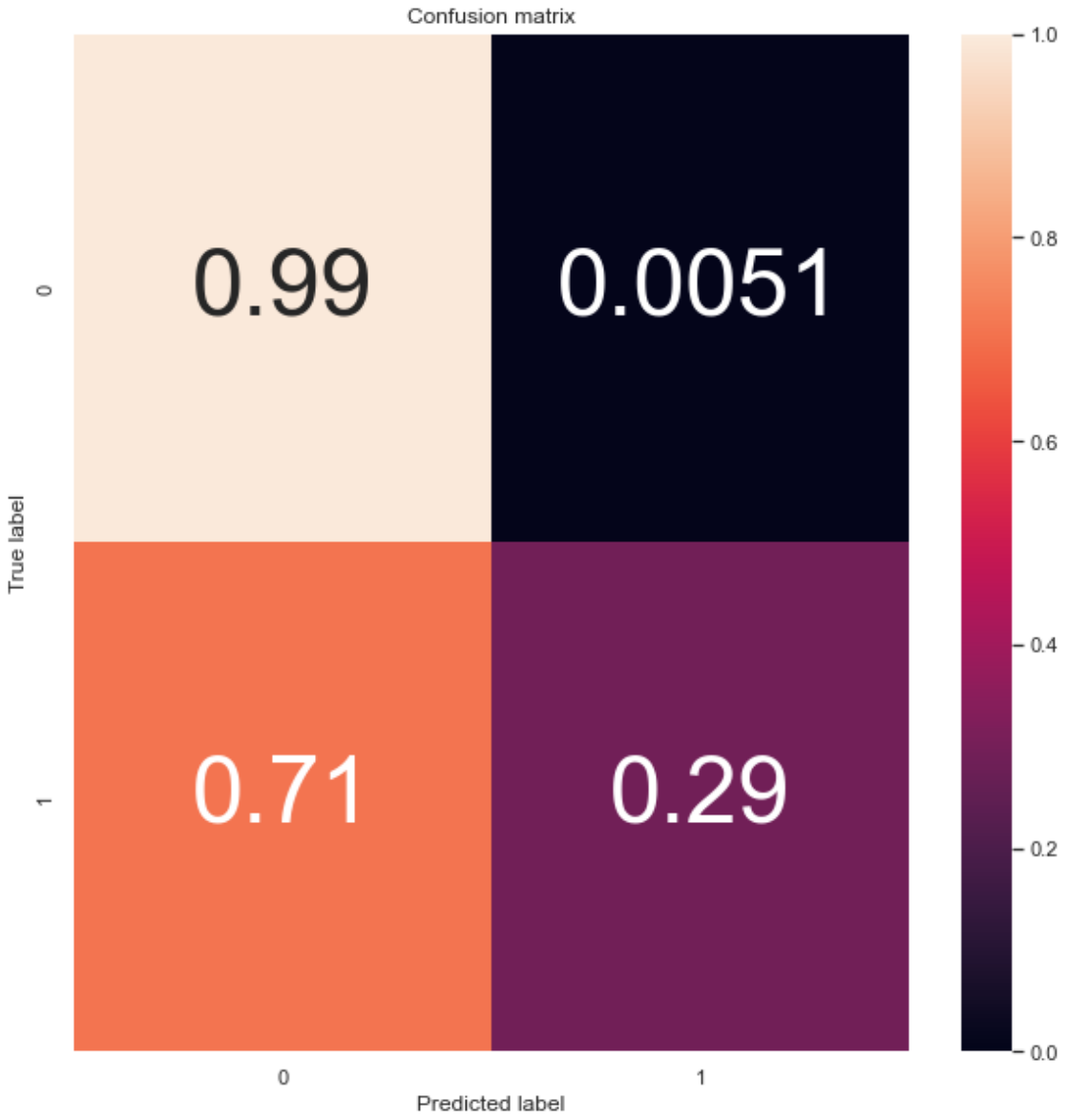


Reference: Author's developments

As evident from the confusion matrix, the label prediction only has high accuracy for the people that are least likely to default, in which case their credit score will be ≈0. With

the usage of gradient boosting methods provided by the XGBoost package installed, model's accuracy of predicting people, who are likely to default was improved by 0.05, leaving us with a 99% accuracy of predicting that an individual will not default and a 29% accuracy of predicting that an individual will default (Figure 4) [24].

Figure 4 Recall confusion matrix of the model using gradient boosting



Reference: Author's developments

While a 13% recall accuracy increase is a noticeable difference, it would be productive to attempt to tune the parameters provided by the XGBoost library in order to increase the accuracy further. There are two main performance metrics that need to be measure in order to pick the best parameters for the created model: *recall* and *precision*.

Recall is a measurement of completeness of the model, describing how well a process identifies items of specific interest compared with the total number of such items that exist in a dataset, *Precision* is a measurement of efficiency, describing how well a process identifies only those items of specific interest, by comparing the number if target items identified with the total number of pieces of data retrieved [25]. Both *recall* and *precision* can be calculated using absolute values in confusion matrices, the formulas for their calculation are listed below:

$$Recall\ (Sensitivity) = \frac{TP}{RP};\ \ Precision\ (Confidence) = \frac{TP}{PP}$$

where *TP – true positive elements in absolute values;*

RP – real positive elements in absolute values;

PP – predicted positive elements in absolute values.

These values can be defined directly from a contingency table with a systematic notation (Table 3).

Table 3 Systematic notation in a binary contingency table. Shading indicates correct (green) and incorrect (red) rates or counts in the contingency table.

	+R	**-R**	
+P	True Positive values	False Positive values	Predicted Positive
-P	False Negative values	True Negative values	Predicted Negative
	Real Positive	Real Negative	

Reference: [26]

To measure both *precision* and *recall* in a convenient format, the F1 measure (or F1 score) will be used, which is a harmonic mean of precision and recall [27]. For maximizing the F1 score the parameters provided in the XGBoost library were tuned as follows:

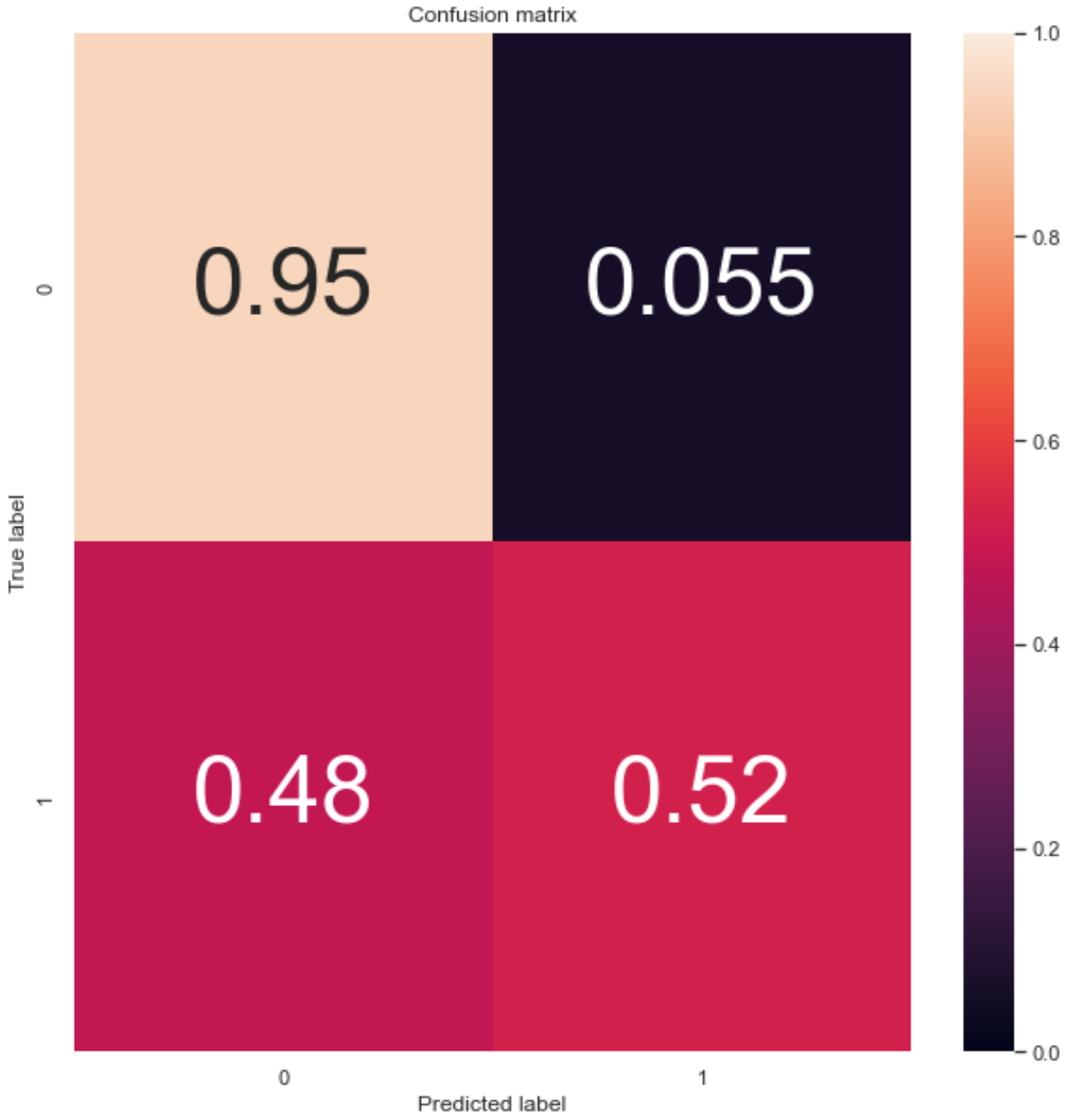
- *colsample_bytree* – subsample ratio of columns when constructing each tree, set to **0.8** (default=1);
- *gamma* – minimum loss reduction required to make a further partition on a leaf node of the tree, set to **0.2** (default=0);
- *learning_rate* – step size shrinkage used in update to prevent overfitting. After each boosting step, the weights of new features can be directly obtained, and learning_rate shrinks the feature weights to make the boosting process more conservative, set to **0.12** (default=0.3);
- *max_delta_step* – maximum delta step each leaf output is allowed to be, set to **1** (default=0);
- *max_depth* – maximum depth of a tree, set to **3** (default=6);

- *min_child_weight* – minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than *min_child_weight*, then the building process will give up further partitioning, set to **9** (default=1);
- *n_estimators* – number of boosting rounds, set to **375** (default=100);
- *random_state* – random number seed, set to **0** (default=random);
- *reg_alpha* – L1 regularization term on weights, set to **1** (default=0);
- *scale_pos_weight* – balance of positive and negative weights, useful for unbalanced classes; calculated as sum(negative instances) / sum(positive instances), set to **4** (default=1) [24,28].

The rest of the parameters were left at their default levels. Most of these parameters were tuned to make the model more conservative, due to the extreme class imbalance the used dataset is facing. Overall, the model post-tuning has the F1 score of **0.441**. The recall and precision confusion matrices for the end model can be seen below (Figure 5,6).

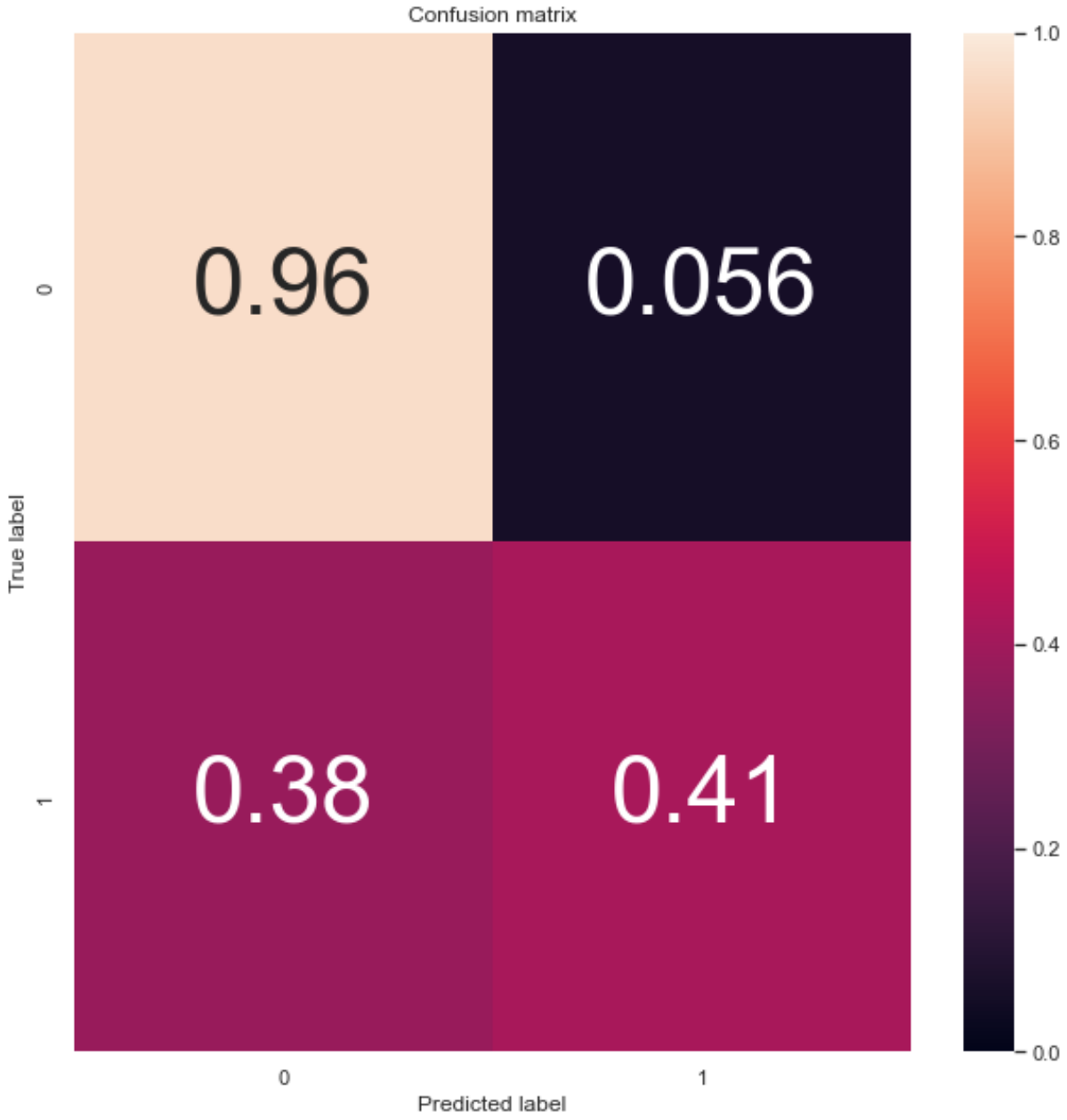
Overall, while the *recall* accuracy could be increased to up to 0.97 for both labels, it is important to maintain the balance between *precision* and *recall* scores when evaluating the accuracy of the model. For the model created, *precision* and *recall* accuracy scores are 0.395 and 0.5 respectively.

Figure 5 Recall confusion matrix of the model post-tuning



Reference: Author's developments

Figure 6 Precision confusion matrix of the model post-tuning



Reference: Author's developments

Results

The abovementioned model was applied to Ukrainian borrowers to see how effective the model will be when applied to credit markets outside of the US. The data used is a 15-person dataset, anonymously collected through a Google Form (Appendix A). Results of credit scoring for these individuals can be seen in Table 4.

Table 4 Credit scoring model application to Ukrainian borrowers

Individual code	Credit Score	Is likely to experience financial distress? (Yes/No)
1	0.16234079	No
2	0.27812165	No
3	0.38806403	No
4	0.36448362	No
5	0.17712267	No
6	0.36448362	No
7	0.18227558	No
8	0.09030385	No
9	0.93568563	Yes
10	0.8566867	Yes
11	0.17712267	No
12	0.16199848	No
13	0.24403746	No
14	0.16234079	No
15	0.17712267	No

Reference: Author's developments

The model created predicts that the vast majority of individuals in the dataset will not experience financial distress, therefore they would be eligible for a loan. While this information might be true considering that most individuals in the dataset do not have any history of past due loans, further analysis needs to be conducted on the dataset this model trained on and the two people in the dataset, who according to the credit score prediction will not pay back the loan.

Before gradient boosting was applied to the model the two most important variables in the dataset were identified as *NumberOfTimes90DaysLate* and *RevolvingUtilizationOfUnsecuredLines* with 0.679 and 0.320 feature importance scores respectively.

The former variable, which is the number of times borrower has been 90 days or more past due, could be interpreted in two ways: if the number is higher than 0 it means that the individual has a history of not paying back loans in time, but it also means that he has a credit history, so since the dataset doesn't have a variable which tells us whether or not the person has applied for loans before, the model deems this information as most valuable.

The 2nd variable, which is the total balance on credit cards and personal lines of credit divided by the sum of credit limits shows us what percentage of the total credit limit that an individual has access to they are using. Judging from the results of credit scoring, the higher this ratio is, the more likely the person is to experience financial distress (Appendix A).

As mentioned earlier in the paper, the model created tends to score individuals lower that their actual score due to large class imbalance present in the original dataset. The original dataset has only 8357 individuals with the *SeriousDlqin2yrs* variable of 1 and 111912 individuals with *SeriousDlqin2yrs* variable of 0, which leads to bias towards lower scores during training.

Another thing to note is how the original dataset was collected from US borrowers, therefore the variables used are the ones that are most applicable to that market. While calculating total balance on credit cards divided by the sum of credit limits would be relevant for the US market, it might be a lot less relevant for Ukraine. The average consumer debt per person in the US is $96371, which mostly consists of mortgages and borrowers with the highest FICO score (800-850) can go all the way up to $139280 in their consumer debt [29].

While this data is not tracked in Ukraine, it is fair to assume based on the available information that Ukrainian borrowers have much lower numbers of consumer debt than US borrowers. For example, the total household debt in the US reached $15,842bn in March 2022 [30]. By comparison, the total household debt in Ukraine and Poland are $9bn and 211.8b$ respectively, so it can be safely said that Ukrainian borrowers on average have much lower consumer debt, when compared to US borrowers [31, 32].

So, while the created decision tree machine learning model is applicable to a set of Ukrainian borrowers, there needs to be careful consideration when picking a training dataset to undertake such a task. A more fitting dataset from a country that is closer in total household debt to Ukraine (such as Poland) would certainly provide more accurate results.

Which is not to say that the model is unusable or unapplicable to Ukrainian financial institutions. While the dataset used is difficult to work with due to how the data is distributed, other machine learning-enabled classification methods could give better results. Due to the dataset being open source and listed on Kaggle as a competition, other credit scoring models have been created, some with comparable or higher accuracy scores than the one built for this paper. For example, a random tree forest model created by Max Fitzpatrick has the same high label prediction accuracy for “1” and has a 50% recall accuracy for the “0” label [33].

Different machine learning-enabled models can and will give varying results, depending on the complexity of the model itself and how well fit it is to the dataset available

to us. There is also potential room for experimentation with unsupervised machine learning models.

Most accuracy problems that occur when creating machine learning-enabled classification models can be linked to datasets that are imbalanced or are too small. The dataset used for the model created suffers from the former issue, which highly skews the prediction accuracy, even after fitting the parameters of the model. Using a different classification model could possibly result in higher accuracy, however, models that are tuned to unbalanced classes won't be very flexible when used on other datasets. The model created shows an example of how gradient boosted decision trees perform when applied to imbalanced classes and the best way to improve this model's accuracy would be to change the dataset itself.

To improve a gradient-boosted decision tree model on the chosen dataset to a point where the accuracy scores will be optimal, most likely there will be a need to resample the dataset used. Resampling methods can adjust the number of majority and minority classes, which is usually used to solve the imbalance in training data [34]. Such methods could help in maximizing accuracy of the model by resampling it to a, for example, dataset with 20000 positive and 20000 negative classes by oversampling the minority class with duplicates.

Which is not to say that the model built is not applicable to Ukrainian borrowers. As a fully automatic tool that can be deployed without having to hire a data science team and can be tuned depending on the needs of the financial institution, decision tree algorithms can have a very valid application in Ukrainian banks. And while the data that is publicly provided by those banks is scarce, it can be said with a high degree of certainty that internally such algorithms can and will be used to reduce the amount of NPLs in a bank's credit portfolio.

There's a lot of topics that can be identified for future research about the usage of ML-enabled classification models for Ukrainian financial institutions. One interesting topic would be to compare multiple different models and how they perform on the same dataset

(e.g., random tree forest, decision tree, k-nearest neighbors). Open-source datasets from Kaggle could be a good source of such information, as there are lots of already existing models created for those datasets.

Another promising avenue for future research would be to use a database provided by a Ukrainian financial institution with variables that represent data that is being collected during their regular credit scoring processes. That way a direct comparison can be made of how ML-enabled credit scoring methods fare in comparison to the methods they usually use.

Conclusions

In this paper the performance of decision tree algorithms trained on open-source data and application of those algorithms to the Ukrainian credit market were researched. While the accuracy of the model created is satisfactory, there is a lot of room for improvement before a model such as the one used could be applied in realistic scenarios.

The choice of a dataset when training a machine learning model is extremely important, and the dataset chosen is by no means perfect, considering the class imbalance between individuals, who will experience 90 days past due delinquency or worse, and individuals who will not. As was assumed in the methodology overview, this led to the model being skewed towards predicting that most people will pay back the loan, which might not be the case.

In general, there are several advantages to using ML-enabled classification methods for credit scoring, such as:

- Low entry data requirements;
- Low cost of deployment (depending on the difficulty of the model developed);
- Ease of use for newly created businesses due to the ability to fit the model to pre-existing data.

However, logistic regression remains the most widely used scoring model in the credit industry due to its stability and interpretability and considering how much results can vary between different ML-enabled models, it is not hard to see why. But there is room for such models in the credit market, and with ML algorithms already being used in the financial services industry, it is not long before these methods will be broadly applicable in commercial banking.

Appendices

Appendix A: Ukrainian borrower dataset used for testing the model

RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
0	21	0	0.5	1110.925	2	0	0	0	2
0.33	21	0	0.6	512.735	1	0	0	0	1
0.05	20	0	0.7	358.914	1	0	0	0	0
0	20	0	0.6	830.63	1	0	0	0	0
0	22	0	0.5	2768.767	0	0	0	0	0
0	21	0	0.7	1208.345	1	0	0	0	0
0.612	46	0	0.9	1025.469	2	0	0	0	1
0.1394	47	0	341.82	0	2	0	0	0	0
0.976	52	7	0.8	1709.115	5	7	1	7	2
0	21	1	0	512.735	0	0	0	1	0
0	21	0	0.476	717.828	0	0	0	0	0
0	21	0	0	820.375	0	0	0	0	0
0	21	0	0	153.8204	4	0	0	0	0
0	22	0	0.2875	1367.29	1	0	0	0	2
0	20	0	0.4444	615.2815	0	0	0	0	0

Bibliography

1. Thomas, L. C., Crook, J. N., Edelman, D. B., 2002. "Credit Scoring and its Applications".
2. Thomas Verbraken, Cristian Bravo, Richard Weber, Bart Baesens "Development and application of consumer credit scoring models using profit-based classification measures".
3. Deloitte "Is rising corporate debt a problem? Not necessarily.". URL: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-cfo-insights-august-2021.pdf
4. Amir F. Atiya "Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results". URL: https://authors.library.caltech.edu/75897/1/0922b4f35185c31d9b000000.pdf
5. National Bank of Ukraine "Loan Portfolio Quality (NPLs)". URL: https://bank.gov.ua/ua/stability/npl
6. European Central Bank "Non-performing loans ratio". URL: https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=420.SUP.Q.B01.W0._Z.I7000._T._Z._Z._Z._Z.PCT.C
7. The World Bank "Bank nonperforming loans to total gross loans (%), 2020". URL: https://data.worldbank.org/indicator/FB.AST.NPER.ZS?view=map&year=2020
8. Bart Baesens, Tony Van Gestel, Stijn Viaene, Johan A.K. Suykens, Jan Vanthienen, M. Stepanova "Benchmarking state-of-the-art classification algorithms for credit scoring".
9. Artur Lee Samuel "Some Studies in Machine Learning Using the Game of Checkers". URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560

10. Feng-Hsiung Hsu "Behind Deep Blue".
11. Steven Strogatz "One Giant Step for a Chess-Playing Machine". URL: https://www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html
12. Foster Provost, Tom Fawsett "Data Science for Business".
13. Deloitte Access Economics "Business Impacts of Machine Learning". URL: https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/process-and-operations/TG_Google%20Machine%20Learning%20report_Digital%20Final.pdf
14. Branko Soucek and the IRIS Group "Neural and Intelligent Systems Integration"
15. Elena Dumitrescu, Sullivan Hué, Christophe Hurlin, Sessi Tokpavi "Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds". URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3553781
16. David Kelnar (MMC Ventures) "The State of AI: Divergence".
17. Fair Isaac Corporation (FICO) "Machine Learning and FICO Scores". URL: https://www.fico.com/en/latest-thinking/white-paper/machine-learning-and-fico-scores
18. Fair Isaac Corporation (FICO) "FICO® Score". URL: https://www.fico.com/en/products/fico-score
19. Paulo Celio Di Cellio Dias, Serasa Experian; Melissa Forti, Bradesco;Marc Witarsa, Serasa Experian "A comparison of Gradient Boosting with Logistic Regression in Practical Cases". URL: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1857-2018.pdf
20. Kaggle "Give Me Some Credit". URL: https://www.kaggle.com/competitions/GiveMeSomeCredit/overview
21. scikit-learn Website. URL: https://scikit-learn.org/stable/
22. seaborn.pairplot Documentation. URL: https://seaborn.pydata.org/generated/seaborn.pairplot.html
23. sklearn.ensemble.RandomForestClassifier Documentation. URL:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

24. XGBoost Documentation. URL: https://xgboost.readthedocs.io/en/stable/index.html#
25. Powers, D.M.W. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation”. URL: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf
26. Nicholas M. Pace, Laura Zakaras “Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery”. URL: https://www.jstor.org/stable/10.7249/j.ctt3fh022.18?seq=1
27. Yutaka Sasaki “The truth of the F-measure”. URL: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf
28. XGBoost Python API Reference. URL: https://xgboost.readthedocs.io/en/stable/python/python_api.html
29. Chris Horymski "Consumer Debt Continued to Grow in 2021 Amid Economic Uncertainty". URL: https://www.experian.com/blogs/ask-experian/research/consumer-debt-study/
30. CEIC "United States Household Debt, 1999-2022, quarterly". URL: https://www.ceicdata.com/en/indicator/united-states/household-debt
31. CEIC "Poland Household Debt, 2003-2021, quarterly". URL: https://www.ceicdata.com/en/indicator/poland/household-debt
32. CEIC "Ukraine Household Debt, 2002-2022, quarterly". URL: https://www.ceicdata.com/en/indicator/ukraine/household-debt
33. Max Fitzpatrick "Probability of default: credit scoring model". URL: https://github.com/max-fitzpatrick/Credit-scoring-model
34. Taisho Sasada, Zhaoyu Liu, Tokiya Baba, Kenji Hatano, Yusuke Kimura "A Resampling Method for Imbalanced Datasets Considering Noise and Overlap". URL: https://www.sciencedirect.com/science/article/pii/S1877050920318676