A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction with Explainability

Sefik Ilkin Serengil, Salih Imece, Ugur Gurkan Tosun, Ege Berk Buyukbas, and Bilge Koroglu

***Abstract*—Credit risk estimation and the risk evaluation of credit portfolios are crucial to financial institutions which provide loans to businesses and individuals. Non-performing loan (*NPL*) is a loan type in which the customer has a delinquency; because they have not made the scheduled payments for a time period. *NPL* prediction has been widely studied in both finance and data science. In addition, most banks and financial institutions are empowering their business models with the advancements of machine learning algorithms and analytical big data technologies. In this paper, we studied on several machine learning algorithms to solve this problem and we propose a comparative study of some of the mostly used non performing loan models on a customer portfolio dataset in a private bank in Turkey. We also deal with a class imbalance problem using class weights. A dataset, composed by 181.276 samples, has been used to perform the analysis considering different performance metrics (i.e. Precision, Recall, F1 Score, Imbalance Accuracy (IAM), Specificity). In addition to these, we evaluated the performance of the algorithms and compared the obtained results. Also, we studied on explainability of the benchmarked techniques with several eXplainable Artificial Intelligence tools. According to these performance metrics, LightGBM gave the best results among the logistic regression, support vector machines, random forest classifier, bagging classifier, XGBoost and LSTM for the dataset.**

***Index Terms*—Non performing loans, non performing loan prediction, big data, machine learning, supervised learning, explainable artificial intelligence.**

I. Introduction

A Non-performing loan (*NPL*) is a loan type in which the customer is in delay; because they have not made the scheduled payments for a clearly defined time period. Although the exact points of default situation may vary depending on the particular loan's terms, "no payment" is usually described as zero payments of either interest or principal. The specified period also varies according to industry and loan type. Mostly, this period is 90 or 180 days. In the banking industry, if the borrower does not pay interest or principal within 90 days, a commercial loan is considered as non performing [1].

Non performing loans can be the consequence of financial misfortune, however it is not just an indicator of a borrower's inability to pay [2]. In addition, the uncertain debt makes it harder to make investment and get new funding. On the side of the lender, the ratio of *NPLs* to total credits is related to the quality of bank assets and reflects the risk that the underlying cash flows from loans [3]. If the loan is non performing, the probability that the borrower will repay it in full are considerably lower. As of 2020, there are approximately one trillion euros worth of non performing loans in Eurozone banks. According to the IMF Euro Area Policies, *NPLs* have reached 1 trillion euros in July 2015 [4].

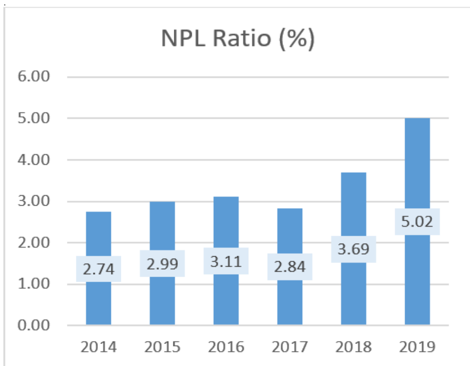

Fig. 1. Situation of non-performing loans in Turkey (2014- 2019).

Fig. 1 shows the ratio of bad loans in Turkey from 2014 to 2019. As can be seen from the table, the rate of non performing loans has been increasing over the years. To maintain the profitability and sustainability for the bank, it is very important that such loans can be determined in advance. According to Banking Regulation and Supervision Agency (BDDK), the gross amount of *NPLs* in Turkey is around 152 billion TRY, more than 7 times the amount in 2009, by March 2020 [5], [6]. Because of these indicators, regulators, such as BDDK, define rules on loans to prevent the increase in *NPLs*. According to the ratio of *NPLs* over total loans, banks pay fines. Furthermore, banks have to block a specific ratio of their assets, and this ratio increases by the *NPLs* ratio of the institution. Namely, *NPLs* not only decrease the banks' profit but also restrict their future moves and investments.

Today, all data stored in the banks' it infrastructure is a perfect example of big data source. Over the last two years, 14 trillion daily financial transactions have been processed for global payments [7]. The extensive use of banking services has concentrated on the non-performing loans management for developing methods desiring to decrease financial risks.

In the light of this big data source and banking services' targets, we suggest a comparative study of some non-performing loan prediction algorithms, the most preferred machine learning algorithms in the literature, to predict if a customer loan will be continue to be paid in a healthy way or not for the next months. For this purpose, we studied the algorithms on the data provided by a private bank in Turkey. The performance of trained models are evaluated based on

Manuscript received September 14, 2021; revised March 30, 2022.
The authors are with Yapi Kredi Technology, Turkey (e-mail: salih.imece@ykteknoloji.com.tr).

doi: 10.18178/ijmlc.2022.12.5.1102

traditional machine learning success metrics i.e. Specifity, AUC, F1 Score, Precision and Recall. In addition, Imbalance Accuracy Metric (IAM) is also investigated to gain insight about the the predictors. Imbalance Accuracy Metric (IAM) is developed for multi-class imbalance datasets [8]. The most promising approaches, also in terms of the explainability of the adopted models using different explainable Artificial Intelligence (XAI) techniques.

The paper is organized as follows. Section II analyzes machine learning algorithms about non-performing loans prediction. Section III is about our methodology for non-performing loan prediction. In Section IV, the evaluation of the results and explanation results has been given. Section V includes the obtained results and suggestions for future study areas.

II. Related Works

Predicting credit loans before they become non-performing is important for banks since the consequences are excruciating unless provisions are made. Many banks have credit risk management departments to perform *NPL* prediction. Traditional approaches to predicting bank loan delinquency mainly use financial probabilistic models, for example the credit scorecards, which utilize a shallow linear regression or classification model with the borrower's financial information. Statistical software tools, such as SAS, SPSS, and pre-defined financial rules may help analysts in this stage.

However, traditional approaches put human in the center and the credit risk managers manually perform the credit risk assessment procedure. Managing *NPLs* is extremely time consuming, because of the paper-intensive works. Moreover, expertise in the field is crucial, as the process are highly depend on the human decisions. This makes the system vulnerable to human mistakes and increases the cost because of the necessity of expertise.

Because of the data availability and computation power with the advanced technology, many researches have been conducted on the widespread issues of banking activities since *NPL* prediction is very important for banks to survive. In literature, the problem is mainly studied under different disciplines; *NPL* prediction, default prediction or credit risk assessment.

In 1960s, data analytics was used for bankruptcy prediction. Seminal univariate analysis of Beaver [9] and multiple discriminant analysis work of Altman [10] has initiated the usage of statistics and data analytics for prediction in finance. After the studies of Beaver [9] and Altman [10], it can be observed that there was a more focus for analysis of financial credit risk. In 1970s, methods such as ordinary least squares [11], discriminant analysis [12], and logistic regression [13] were deployed for prediction. In 1980s, factor analysis [14], logit analysis [15], and other similar techniques were introduced to the area of credit default risk. In 1990s, Altman [16] introduced the original Z-score method being extended to private firms. This study examined the applicability of the same machine learning algorithm to several areas, namely in finance.

In the last years, traditional approaches have been replaced by machine learning algorithms; because, it is hard to build up a model for non performing loans prediction because of the curse-of-dimensionality and class-imbalance problems. Machine learning algorithms are also able to extract non-linear relations on the datasets.

Bahnsen *et al.* [17] found that when several characteristic features have complex nonlinear relationships, traditional algorithms such as logistic regression are not as effective as before. Zhang [18] proposed a standard early warning risk model and BP neural network algorithm: They trained the sample data with BP Neural Network to evaluate the predetermined risk as a personal loan indicator. Ribeiro *et al.* [19] recommended to use SVM+ to improve the default risk model. They stated that, the use of SVM+ further improves versatility. Feki *et al.* [20] suggested a discrimination technique for finance sector based on non-performing loan indicators. It was studied with different kernels of SVM and Gaussian Naive Bayes algorithm. In addition, the strategies for variable selection were also suggested. Wang *et al.* [21] additionally observed that the tree based algorithms have the benefits of adaptability and robust interpretability in explaining the motives for credit scoring. Also, we studied on NPL prediction earlier and got remarkable results [22]. However, that study did not cover the explainability of the models.

Today, besides the applied machine learning techniques, the focus is also on the explainable artificial intelligence models, which means that the results of these models should be easily understood by the experts in risk monitoring division of the financial institutions. Explainability in banking regulations is one of the biggest concerns for complex high precision models. The regulations strictly forbid to deploy models which are not explainable even if they are highly robust. Machine learning interpretability is a newly emerged field, and the related studies in the banking sector is recent. In a recent study by Bussman *et al.* [23], the credit risk scores for medium enterprises are predicted and explainable AI outputs are examined. XGBoost has been used for the prediction model, and Shapley values and SHAP framework are used for explainability. In the paper, four companies' shapley values to feature importance plots are given where the two of them are default and the others are not. Another study, similar to the previous one, has been conducted by Misheva *et al.* [24], the credit risk scoring prediction has been applied using the Lending Club data from Kaggle; LIME and SHAP methods are examined for explainability. In the modelling part, Logistic Regression, XGBoost, Random Forest, Support Vector Machine and Neural Network models have been deployed. LIME has been preferred for local explanations and SHAP has been used for global explanations. They have implemented special frameworks of SHAP such as Kernel Explainer, Tree Explainer, Deep Explainer to compare the feature importance results. They found out LIME and SHAP has established consistent explanations in financial area.

In this study, we suggested a benchmark of different machine learning algorithms from the literature and gradient boosting algorithms such as XGBoost and LightGBM to deal with non-performing loans prediction for MEs (medium enterprises) on the data set obtaining from a private bank in Turkey. Addition to these different machine learning algorithms, we also studied on LSTM deep learning algorithm. Our analysis aims to increase F1 score and reduce false positives to decrease the misclassification costs. Moreover, to understand why the models give that prediction result we also focused on the explainability and interpretability of the selected models through the XAI tools.

III. Methodology

We studied the algorithms on data provided by private bank in Turkey. We focused customers' historical data such as: customer payment behavior history, balance sheets, previous credit card payments, risk and limit amounts of customers in other banks from consumer reporting agency, business sector of the customer and so on. Specifically: (i) we first performed analysis to specify non-informative features to exclude from our analysis by computing the missing values; (ii) after, we studied on features' correlation for dropping out additional non-informative features; (iii) we replaced the remaining missings of relevant features with the column-based computed average values. (iv) finally, we applied normalization for risk and limit amounts reported by consumer reporting agency.

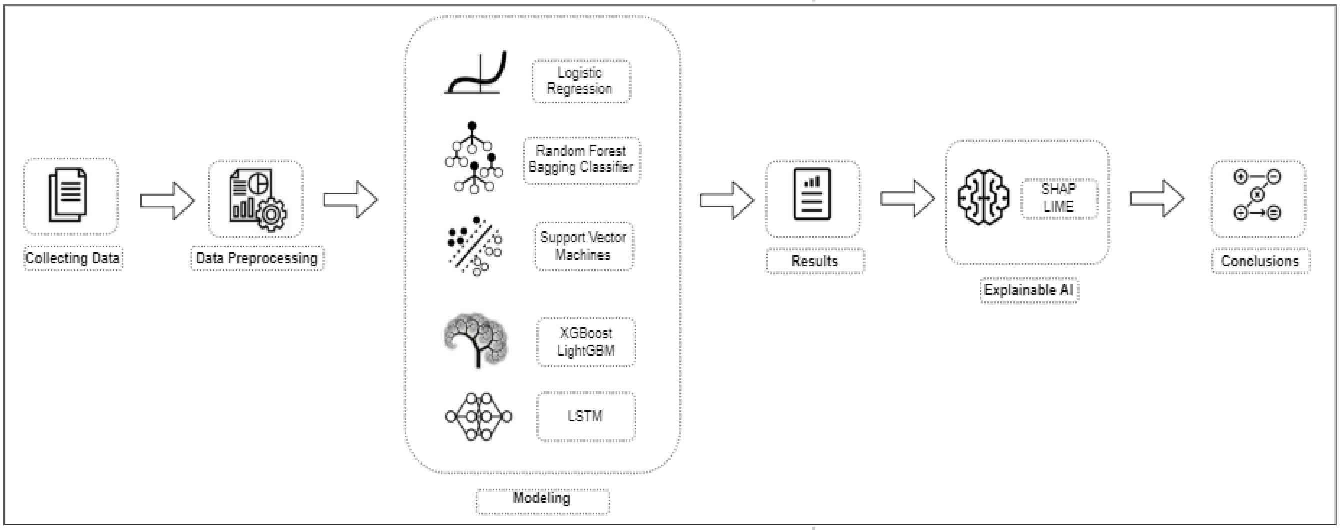

Fig. 2. General framework of proposed model.

To handle time series data in a classification model, we enlarged features of corresponding instance by adding historical values for 6 months period in columns with T and depth postfix. For example, customer business card limit t1 values refers to the business card limit values of a month before for the customer. The other step performs the non performing loans prediction with respect to a given customer that is affected by the imbalance data set problem, which is characteristic of the non performing loans, because of the large number of customer loans continue to be paid in a healthy way. The ratio of customers with *Healthy* status, *Late* status and *NPL* status in the dataset are 96 percent, 3 percent and 1 percent respectively. While working on the LSTM algorithm, we added historical features of the corresponding instance as dimensions instead of columns.

TABLE I: Dataset Characterization

Label	Total Case
Current	174.346
Late [15 – 90 days]	5.485
NPL [91+ days]	1.736
Total	**181.567**

In addition, we created 3 categorical variables with the payment status of customers in the next 3 or 6 months, which is the target variable of the algorithms. As seen in Table I, these are; customers who continue their payments with delinquency lower than 15 days, customers who have a delay of 15 to 90 days in their payments, and customers who have a delay of 91 days or more in their payment, that is, those who are in *NPL* status.

After applying the preprocessing techniques on the raw data, we compared the performance of the logistic regression, random forest, support vector machines, bagging classifier and gradient boosting algorithms (XGBoost tree algorithm [25] and LightGBM tree algorithm [26]) and LSTM. Table II shows the machine learning algorithms and performance metrics for the evaluation. According to literature, these algorithms are the most preferred ones on tabular data for the problem of non performing loan prediction. Addition to these machine learning algorithms, we also worked on the LSTM deep learning algorithm; because the data set we are working on has time series characteristics and LSTM algorithm gives good results on time series data prediction [27]. Lastly, we focused on obtaining human understandable reasons from the trained models with XAI frameworks, namely, SHAP and LIME.

TABLE II: Classifier Methods and Performance Metrics

Models	Performance Metirics
Logistic Regression	AUC
Random Forest Classifiers	Precision
Support Vector Machines	Recall
Bagging Classifier	F1 Score
LGBM	Imbalance Accuracy
XGBoost	
LSTM	

After training the algorithms, we studied on the comparison of several XAI tools for explaining the obtained model results. In particular, we compared two different XAI tools: SHAP and LIME. Both of these methods clarify the inner workings of black box models, which explains the reason for the prediction [28], [29]. Explanation and interpretability is indispensable in financial applications, it is needed to make sure that the resulting machine learning algorithm is able to capture financially correct heuristics from the training data.

To sum up, stages of the our machine learning project pipeline is shown in Fig. 1 namely, collecting data, preprocessing, modeling, evaluation, and reasoning of the trained models in a human understandable manner.

IV. Experimental Results and Evaluations

The aim of the evaluation is compare the different algorithms using different class weight strategies based on different evaluation metrics

For the proposed analysis, we have chosen a dataset provided by a private bank in Turkey, which includes financial credit portfolios of customers provided by the bank between 2017 and 2020, and consists of 181.567 samples and 705 features. We consider loan status as the target class of our study, and its values are listed in the Table 1. We divided the dataset into train, validation and test datasets with a ratio of 70:15:15 sequentially.

A. Evalutaion Metrics

In non performing loan prediction problems, several metrics can be identified to measure the performance of the models. Just using accuracy as a metric may not be a good choice for these types of problems [30]; as we cannot measure the loss which are caused by different type of errors, i.e. false negative and false positives. It is preferred to use these metrics such as Specificity, Precision, Recall, F1 Score in order to better understand the model results for this problem. Moreover, we used imbalance accuracy metric (IAM) for model evaluation. IAM can be considered as a newly defined metric compared to other metrics. This metric aims to eliminate the use of multiple metrics [8]. If the IAM value is less than 0, it shows that the model performance is bad, and if it is greater than 0, it shows that the model has a good fit on the dataset.

In addition to these metrics, the Area Under Curve metric defines the area under ROC curve, that shows the trade-off between the true positives and false positives for a model.

$$\text{Specifity} = \frac{\text{TrueNegative}}{\text{TrueNegative}+\text{FalsePositive}} \tag{1}$$

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive}+\text{FalsePositive}} \tag{2}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive}+\text{FalseNegative}} \tag{3}$$

$$\text{F1 Score} = 2\frac{\text{PrecisionRecall}}{\text{Precision + Recall}} \tag{4}$$

$$IAM = \frac{1}{k}\sum_{i=1}^{k}\frac{c_{ii}-\max\,(\sum_{j\neq i}^{k} c_{ij},\sum_{j\neq i}^{k} c_{ji})}{\max\,(c_{.i},c_{i.})} \tag{5}$$

The formulations of the metrics mentioned above are given in (1), (2), (3), (4) and (5). c_{ij} are the elements of the k by k confusion matrix where $i, j = 1,2,\ldots,$ k for a k- class dataset in eq. (5).

B. Experimental Results

We evaluated the performance of the algorithms based on the different parameters. Particularly, we have studied on six classification models (Logistic Regression, Random Forest, Support Vector Machines, Bagging Classifier, LightGBM and XGBoost) and LSTM which is one of the most preferred deep learning algorithm for time series problem. We used grid search to iterate the machine learning algorithms over a set of hyperparameters to determine the best set of hyperparameters.

TABLE III: Dataset Characterization

Model	Total Case
Linear Regression	penalty: ’l2’, solver: ’lbfgs’
Random Forest	n_estimators: 15, max_depth: 20
SVM	C: 1.0, kernel: ’rbf’
Bagging	n_estimators: 10
LGBM	obj: ’multiclass’, metric: 'multilogloss'
XGBoost	obj: ’multi:softmax’, n_estimators: 100
LSTM	activation: ’relu’, optimizer: ’adam’, architecture: ’3 ×127, 200, 3’, loss: 'categorical crossentropy'

All the information related to the model performance metrics and hyperparameter space on the test dataset is listed in Table III. We also compared the results with study [31] which also studies *NPL* detection problem. We also studied on the Bagging Classifier model, which is indicated as the best method in [30], to the data set. According to metrics results in Table IV, LightGBM emerged as the best method among all methods.

Besides, we evaluated the results in the business perspective as well. The confusion matrix is given in Table 5. Here, 1,347 *NPL* and 3,223 cases with *Late* status predicted correctly. Moreover, 194 cases with *Late* status predicted as *NPL* whereas 355 *NPL* cases predicted as *Late* status. Although these predictions are categorized as misclassified instances in the confusion matrix, it can be evaluated as early warning signals from the business perspective; because it gives us that the customer will have delinquency in payments even if it can not predict the magnitude of this delinquency correctly.

TABLE IV: Model Results on Test Data

Model	Specifity	Precision	Recall	IAM	F1
LR	0.79	0.37	0.53	- 0.59	0.38
Random Forest	0.91	0.75	0.74	0.45	0.74
SVM	0.81	0.70	0.71	- 0.37	0.71
Bagging	0.88	0.85	0.72	0.48	0.77
LGBM	**0.90**	**0.87**	**0.77**	**0.60**	**0.82**
XGBoost	0.88	0.88	0.76	0.58	0.81
LSTM	0.89	0.71	0.79	0.43	0.74

Furthermore, 538 *Healthy* cases predicted as unhealthy but this will not cause any business loss. Finally, 2,068 *Late* and 34 *NPL* cases predicted as *Healthy*. The misclassified 2,068 *Late* cases has a right-skewed distribution. In other words, they are very close to the upper boundary of *Healthy* cases. It will not cause any business loss because *Late* target is considered for just early alerts. Herein, 34 cases causes business loss but that's just 2 of 10,000 cases.

TABLE V: CONFUSION MATRIX

		Prediction		
		Healthy	Late	NPL
Actual	Healthy	173.808	531	7
	Late	2.068	3.223	194
	NPL	34	355	1.347

In addition to the success metrics results in Table V, the ROC curve and AUC scores of the models were also examined in Fig. 3. *Late* and *NPL* status (label 1 and label 2) were gathered under a single target variable and each target variable was converted to the binary class label. From the Fig. 3 it can be observed that, Random Forest Classifier model has the highest AUC score for binary classification.

However, for this problem, the target with three classes is more convenient to meet business needs. So, the algorithm of LightGBM model, which gives the best results with three target labels, were used as the input for the XAI frameworks.

C. Explanation Results

Decision makers of business teams tend to build logistic regression models. This habit comes from some regulations in the banking industry. No matter how accurate a model is, black box models are not allowed to be deployed in the production. Logistic regression models are highly explainable but they are not as strong as non-linear models such as deep neural networks, or gradient boosting. On the other hand, non-linear models are almost black boxes and they almost offer no explainability.

Feature importance is an important tool to explain built models. In generalized linear models such as linear regression or logistic regression, feature importance values are related to the coefficients of features in the regression equation. This explains what happens to output if the value of a feature changes one unit when others become same. Similarly, feature importance could be extracted from decision trees with the decision metric (e.g. entropy or gini) by multiplying number of instances on each branch. Besides, decisions of built trees can be read and understood by human clearly.

Nowadays, we have the state of the art XAI frameworks such as SHAP or LIME. We can now explain fully black box algorithms. LIME explains single predictions with human readable if-else statements similar to single decision trees whereas SHAP can analyze the importance of the each feature in terms of probability contribution of being which target class. By using these tools, the black box structure of the models is eliminated. Thus, a more efficient study is made in which the model results can be evaluated with their reasons.

We have 3 distinct classes in the target label: *Healthy* (0, 14 days), *Late* (15, 90 days) and *NPL* (91+ days) based on the past due days in the next 3 months. Previous payment patterns seem to be the most dominant feature among all classes. Having a few past due days becomes an important indicator to predict non performing loans for the short and mid-term delinquency estimations. Similarly cash debt and non-cash debt appear in the list as well. Payment debt ratio and debt limit ratio of the business cards in the customer portfolios of the dataset come after payment patterns.

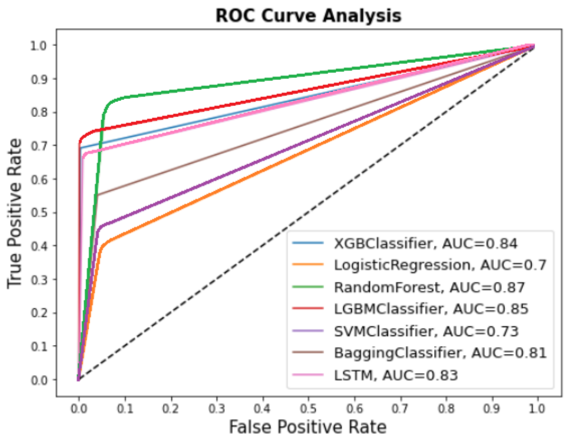

Fig. 3. ROC curve and AUC results for the models.

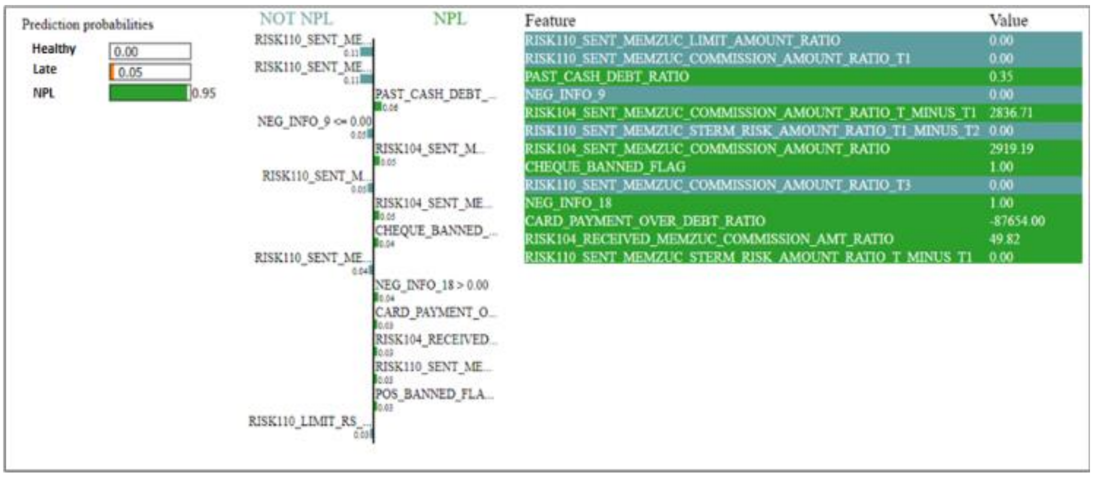

Fig. 4. Explaining an NPL Case with LIME.

Demographic information such as customer activeness tenure and digital banking customer tenure are also significant features. Interestingly, loan payments tend to go to *NPL* after the first COVID-19 case (2020 March). What's more, RISK104 states restructured loans and RISK105 states non-cash loans. A bank sends information to the central bank about its customers with sent key whereas it receives information from the other banks with received key. In other words, features with received key states intelligence about a customer in other banks. From the feature importance analysis, it can be observed that the higher the commission and re-discount amounts, the more probable to the customer portfolio will have a delinquency in the short and mid-term. Furthermore, credit class code 101 states a customer is in watch list because of some negative intelligence in the subsidiary companies such as factoring or leasing.

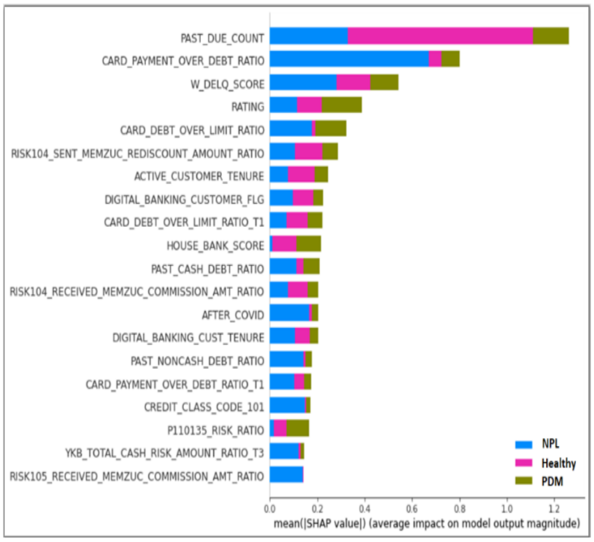

Fig. 5. Feature Importance of LightGBM Model with SHAP.

Even though, we can read the decisions of single trees clearly, this is not an easy task for gradient boosted trees. Because there are many trees. For example, 2160 trees were built in our experiment. Here, LIME can extract decision rules similar to single decision trees. In this way, we can read and understand how a single prediction was made. Its final form looks like if-else statements of single decision trees.

For example, Fig. 4 shows an explanation of a true positive prediction with LIME. For example, Also Fig. 5 shows the feature importance of the model with SHAP. That customer is in *Healthy* condition when prediction is made but the model predicts that it will be *NPL* in 3 months with 0.95 score. The most dominant features shown in Fig. 4 explain why it is assigned as *NPL* class. Negative intelligence with 18 code states that it has a dud cheque. It directly causes the customer portfolio to have a delinquency with 90 days or more. Supportively, it has a cheque ban and this causes it to *NPL* as well. Commission amount of restructured loans increase the probability of being *NPL* status. Even though, some of those features don't appear in the most important feature tables, they are crucial for this local prediction. This shows the significance of local explainability in machine learning pipelines. Moreover, SHAP feature importance results of the all models were also examined by experts in risk monitoring division. According to the examination, when experts looked at top features list, they stated that the most important features of LightGBM (clients historical payment behaviors, business cards payments and risks , types of loan products, customer tenures etc.) are also more meaningful in terms of financial and business.

V. Conclusion

Non-performing loans have a negative impact on the banking sector and, accordingly, macroeconomic balances. Also, non-performing loans prediction is one of the most challenging problems in finance sector. While using traditional techniques of machine learning, class imbalance and curse-of-dimensionality problems are encountered.

The aim of this study is to make a comparison for machine learning algorithms to predict the credit default prediction of that can also handle imbalanced data sets. According to the performance metrics (i.e. Precision, Recall, F1 Score, Imbalance Accuracy (IAM), Specificity), LightGBM gave the best results among the logistic regression, support vector machines, random forest classifier, bagging classifier, XGBoost and LSTM for the dataset. Due to the lack of algorithmic transparency for the wider adoption of AI-based solutions in credit default prediction, we also studied on explainable AI tools for the model that gave the best results.

Further works will be explored to study with different sampling methods on these benchmarked algorithms that in some scenarios could show better results.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The authors who have worked together on the publication and contributed equally.

References

[1] T. Segal. Nonperforming Loan – NPL. (Apr. 2021). [Online]. Available: https://www.investopedia.com/terms/n/nonperformingloan.asp
[2] M. Balgova, M. Nies, and A. Plekhanov, "The economic impact of reducing non-performing loans," Working paper, European Bank, 2016.
[3] M. M. Cornett and A. Saunders, *Financial Institutions Management: A Risk Management Approach*, McGraw-Hill/Irwin, 2003.
[4] International Monetary Fund, Euro Area Policies, July 10, 2015.
[5] Banking Regulation and Supervision Agency, Turkish Banking Sector Main Indicators, Dec. 2009.
[6] Banking Regulation and Supervision Agency, Turkish Banking Sector Main Indicators, March 2020.
[7] McKinsey, Global Payments Report 2019, July 15, 2019.
[8] E. Mortaz, "Imbalance accuracy metric for model selection in multi-class imbalance classification problems," *Knowledge-Based Systems*, vol. 210, p. 106490, 2020.
[9] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of Accounting Research*, pp. 71–111, 1966.
[10] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.
[11] P. A. Meyer andH. W. Pifer, "Prediction of bank failures," *The Journal of Finance*, vol. 25, no. 4, pp. 853–868, 1970.
[12] E. B. Deakin, "A discriminant analysis of predictors of business failure," *Journal of Accounting Research*, pp. 167–179, 1972.
[13] D. Martin, "Early warning of bank failure: A logit regression approach," *Journal of Banking & finance*, vol. 1, no. 3, pp. 249–276, 1977.
[14] R. C. West, "A factor-analytic approach to bank condition," *Journal of Banking & Finance*, vol. 9, no. 2, pp. 253–266, 1985.
[15] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, pp. 109–131, 1980.
[16] E. I. Altman, J. Hartzell, and M. Peck, "Emerging market corporate bonds — a scoring system," *Emerging Market Capital Flows*, pp. 391-400, 1998.
[17] A. Correa Bahnsen, "Feature engineering strategies for credit card fraud detection," *Expert Systems with Applications*, pp. 51, pp. 134-142, 2016.
[18] Z. Zhang, "Research of default risk of commercial bank's personal loan based on rough sets and neural network," in *Proc. 2011 3rd International Workshop on Intelligent Systems and Applications*, 2011, p. 1–4.
[19] B. Ribeiro, C. Silva, N. Chen, A. Vieira, and J. C. das Neves, "Enhanced default risk models with SVM+," *Expert Syst Appl.*, vol. 39, pp. 10140–10152, 2012.

[20] A. Feki, A. Ishak, and S. Feki, “Feature selection using bayesian and multiclass support vector machines approaches: Application to bank risk prediction,” *Expert Syst Appl.*, vol. 39, pp. 3087–3099, 2012.
[21] M. G. Wang, L. L. Ge, and J. P. Zhao, “Research on risk monitoring of micro-online loan platform based on C5.0 algorithm,” *Chinese Management Science*, vol. 24, s1, pp. 345-352, 2016.
[22] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, and B. Koroglu, “A comparative study of machine learning approaches for non performing loan prediction,” in *Proc. 2021 6th International Conference on Computer Science and Engineering (UBMK)*, September 2021, pp. 326-331.
[23] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable AI in Fintech risk management,” *Frontiers in Artificial Intelligence*, vol. 3, no. 26. 2020.
[24] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, “Explainable AI in credit risk management,” arXivpreprint arXiv:2103.00949, 2021.
[25] T. Q. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho *et al.*, “Xgboost: extreme gradient boosting,” *R Package Version 0.4-21*, vol. 4, pp. 1–4, 2015.
[26] G. L. Ke, Q. Meng, T. Finley, T. F. Wang, W. Chen, W. Dong *et al.*, “Lightgbm: A highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.
[27] D. Babaev, M. Savchenko, A. Tuzhilin, and D. Umerenkov, “Et-rnn: Applying deep learning to credit loan applications,” in *Proc. the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019, pp. 2183-2190.
[28] L. S. Shapley, “The ‘Value of the Game’ as a tool in theoretical economics,” Technical Report, RAND CORP SANTA MONICA CA, 1967.
[29] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in *Proc. the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
[30] J. Abellán and J. G. Castellano, “A comparative study on base classifiers in ensemble methods for credit scoring,” *Expert Systems with Applications*, vol. 73, pp. 1–10, 2017.
[31] M. A. Elnaggar, M. A. EL Azeem, and F. A. Maghraby, “Machine learning model for predicting non-performing agricultural loans,” in *Proc. the International Conference on Artificial Intelligence and Computer Vision*, 2020, pp. 395-404.

Sefik Ilkin Serengil received his MSc in computer science from Galatasaray University in 2011. He has been working as a software engineer for 10 years. Currently, he is a member of applied data science squad at Yapi Kredi Technology. His research interests are machine learning and crytopgraphy. Moreover, he contributed many open source projects.

Salih Imece graduated from Yildiz Technical University Industrial Engineering Department in 2016. Afer that, he graduated M.Sc. in industrial engineering at Istanbul Technical University in 2019. He is currently pursuing PhD program in industrial engineering at Istanbul Technical University. Since October 2020, he has been working as an R&D engineer at Applied Data Science Department of Yapi Kredi Technology.

Ugur Gurkan Tosun graduated from Koc University Electrical and Electronics Engineering Department in 2018. He is currently pursuing his M.Sc. in Software Engineering at Boğaziçi University. After completing his Bachelor's degree, he worked in several positions in the field of computer science. Since February 2021, he has been working as a Research and Development Engineer at Applied Data Science Department of Yapi Kredi Technology

Ege Berk Buyukbas graduated from Middle East Technical University Industrial Engineering Department in 2020. He is also currently pursuing his M.Sc. in computer engineering at Middle East Technical University. After completing his bachelor's degree, he have started to work in Yapi Kredi Technology as Part-Time Research and Development Engineer at Applied Data Science Department..

Bilge Koroglu has received her BSc and MSc degrees from Computer Eng. Department at Bilkent University in 2010 and 2012 respectively. She has been working in development teams of software products which extensively employ AI, ML algorithms in banking and finance industry . She has wide experience on applying algorithms and techniques on financial data to produce end-to-end software solutions. Ms. Koroglu is currently leading Applied Data Science team in Yapi Kredi Teknoloji.