Loan Grade Risk Assessment Using Machine Learning

Boran ASLAN  
*Middle East Technical University*Ankara, Turkey

*Abstract*— The central aim of this research is to compare the efficacy of various machine learning and statistical techniques in predicting loan credit grades. Based on a detailed dataset of loan applicants, the study begins with an exploratory data analysis to recognize the factors that influence creditworthiness. Multiple models were created after a thorough data cleaning and feature engineering phase that addressed missing data and data leakage. ANN, SVM, Random Forest, XGBoost, LightGBM, CatBoost, and logistic regression variants (multinomial and ordinal) were evaluated for performance. By optimizing hyperparameters with GridSearchCV and comparing key metrics like accuracy and F1-score, this project provides a comparative analysis of different approaches to solving this multi-class classification problem.

Keywords—Model Comparison, Credit Scoring, Classification, GridSearchCV, Machine Learning, Data Preprocessing, Loan Grade

# INTRODUCTION

For any financial institution, accurately classifying a loan applicant's credit risk is essential since it has a direct impact on risk management and profitability. A loan grade, which determines the loan's terms, usually highlights this risk. Building a predictive model to categorize these loan grades using a rich dataset which includes a borrower's credit history, financial demographics, and loan application data is the main goal of this project. The potential for this predictive modeling to produce a more effective, precise, and data-driven method of credit evaluation is what makes it significant.To do this, the study first identifies the primary factors influencing loan grades through a thorough exploratory data analysis (EDA).

To achieve this, a thorough data preprocessing step was followed to get the dataset ready for modeling. The predictive performance of algorithms including Random Forest, XGBoost, Support Vector Machines, and Artificial Neural Networks performed better against traditional methods like Multinomial and Ordinal Logistic Regression. The models are compared on key metrics such as accuracy and F1-score to determine the optimal solution for this classification challenge.

# LITERATURE REVIEW

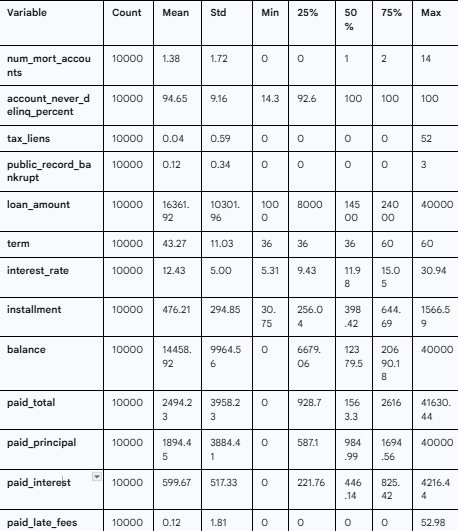
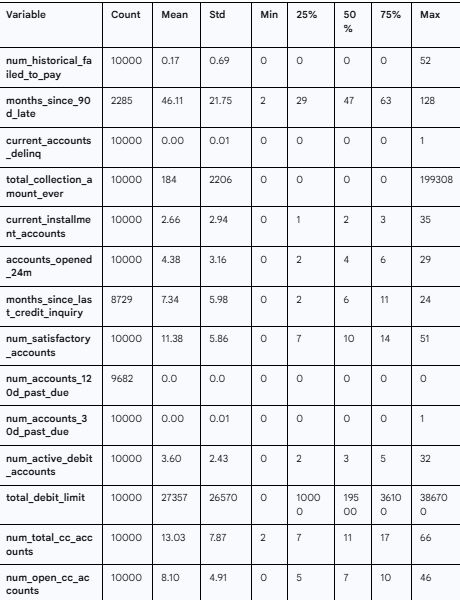
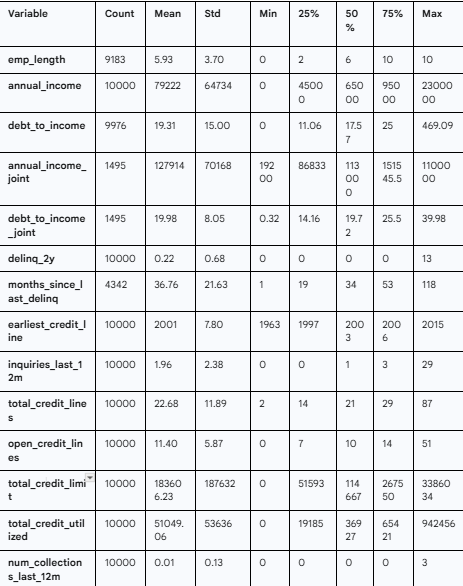
Credit scoring has been extensively studied, with traditional models such as logistic regression widely used for their simplicity and interpretability. However, as financial data became more complex, machine learning (ML) models have gained popularity for their superior predictive performance. Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) were among the 41 classifiers that Lessmann et al. (2015) benchmarked. They discovered that ensemble approaches, especially boosting algorithms like XGBoost, consistently outperformed classical models [1].

# METHODOLOGY

1. *Dataset*

The dataset used for this project is sourced from OpenIntro, which contains a sample of 10,000 loans made via the LendingClub platform. Predicting a borrower's loan grade is the main goal of this dataset, which is intended for credit risk analysis. This grade is a key indicator of the lender's evaluation of creditworthiness. The dataset includes 55 variables, such as the borrower's financial situation, employment details, credit history, and loan details, that give a comprehensive picture of each loan. Below is a list of some of the important variables that were used in this analysis.

1. *Descriptive Statistics*

The dataset's descriptive statistics offer crucial details for comprehending the variables' distribution, variability, and central tendency.

A right-skewed distribution is shown by the annual\_income distribution, which has a mean value of $79,222 and a median value of $65,000 when looking at the numerical variables. This suggests that a small number of high-income people are pulling the average upward. With a mean of $16,362 and a median of $14,500, loan\_amount falls between $1,000 and $40,000, indicating a more symmetrical distribution. Similarly, interest\_rate, which has a mean of 12.43% and a median of 11.98%, exhibits a minor right skew. Given that the mean (19.31) is greater than the median (17.57), the debt\_to\_income variable likewise exhibits a right-skewed distribution.

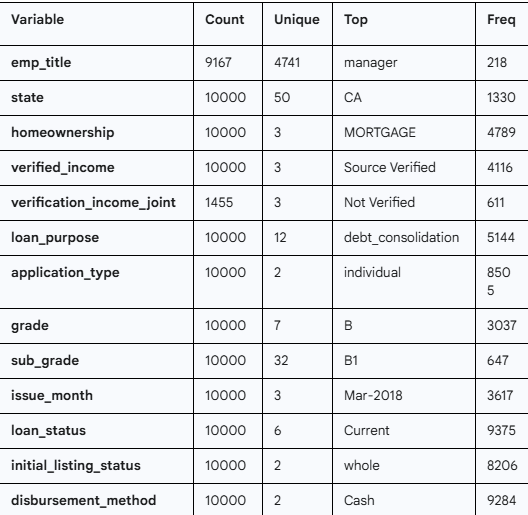


Table 2 Descriptive Statistical Summary of Categorical Data

The distributions within the dataset are visible when looking at the summary of categorical variables. Class 'B' is the most prevalent category in the grade variable (3037 occurrences). 'MORTGAGE' is the largest group in the homeownership category, with 4789 people. 'Debt\_consolidation' is the most frequent reason for the loan\_purpose variable, with 5144 occurrences. The bulk of the loans have a 'Current' status (9375 instances), and the great majority of applications (8505) were made on a 'individual' basis. This suggests that active loans make up the majority of the dataset.

1. *Exploratory Data Analysis*

As we can see from the density plots, box plots and summary statistics some variables have outliers. We’ve conducted Shapiro-Wilk Normality Test for each of the variables and with each variable having p-value<0.05, none of the variables are normally distributed. Outliers can be possible in the blood and can tell the diseases state, therefore we will not be conducting any outlier handling methods.

Therefore, in order to answer research questions, non-parametric tests such as Kruskal Wallis Test is conducted.

*C.1* *What is the overall distribution of loan grades?*

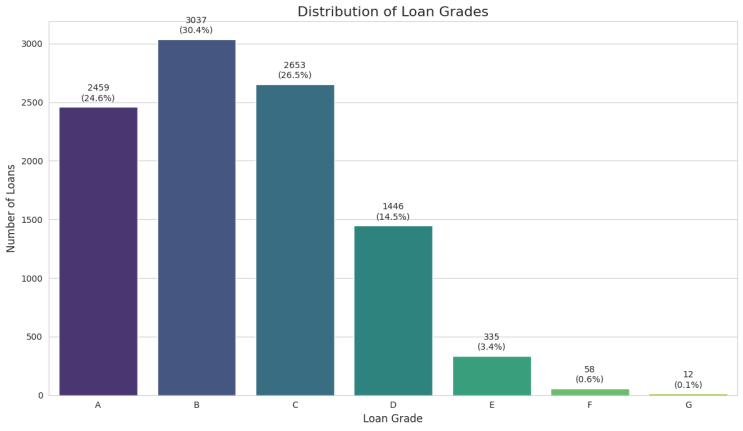
The distribution of loan grades is imbalanced. The majority of loans are concentrated in grades 'A', 'B', and 'C', which typically represent lower risk. Higher-risk grades ('D' through 'G') are progressively less common. Grade 'B' is the most frequent category, while 'G' is the rarest. This imbalance is critical! So afterwards we will apply some methods to solve it!

Figure 1 Box Plot x

The box plot shows that the expectation given above is almost satisfied. As it is observed from the plot, median changes as the Category variable does and Category 3 has the lowest albumin levels.

We have also conducted Kruskal-Wallis Test and we conclude that there is a significant difference in ALB levels of patients among the Category variable (p<0.05)

For each of the independent variable we have conducted a Kruskal-Wallis test to show whether they differ among different classes. Only Sex variable didn’t differ among classes. (p>0.05)

*C.2 Is there a connection between a borrower's annual income and the loangrade they receive?*

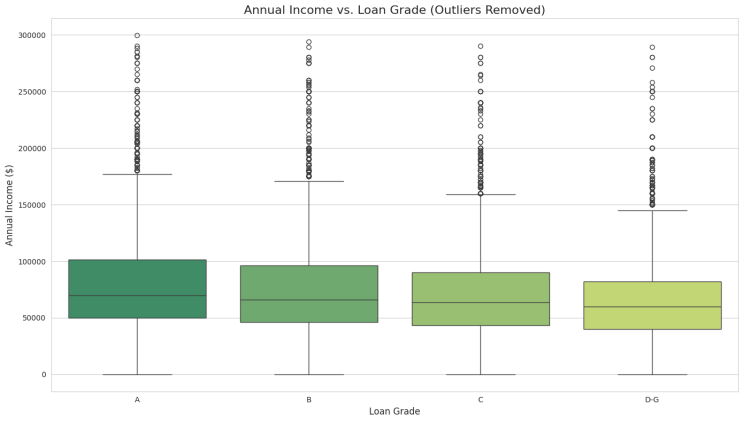
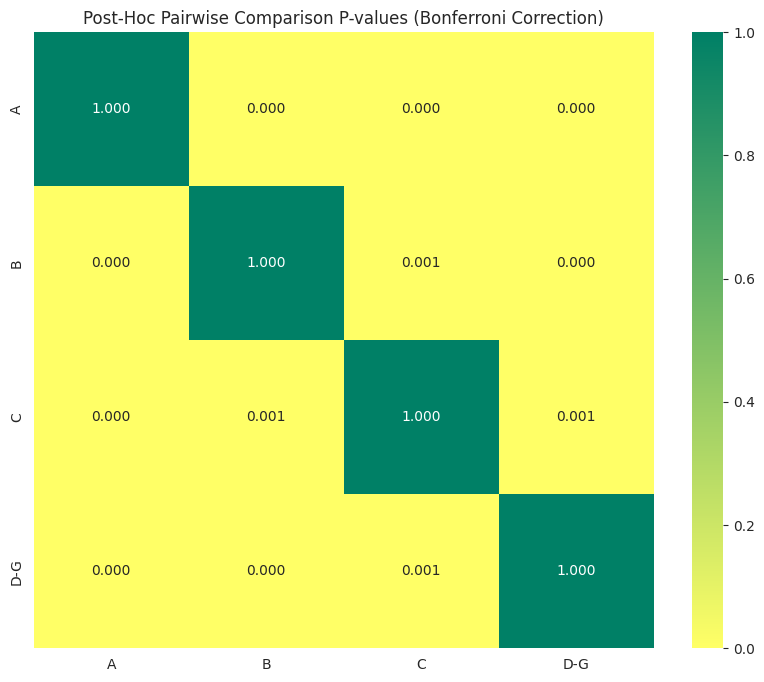
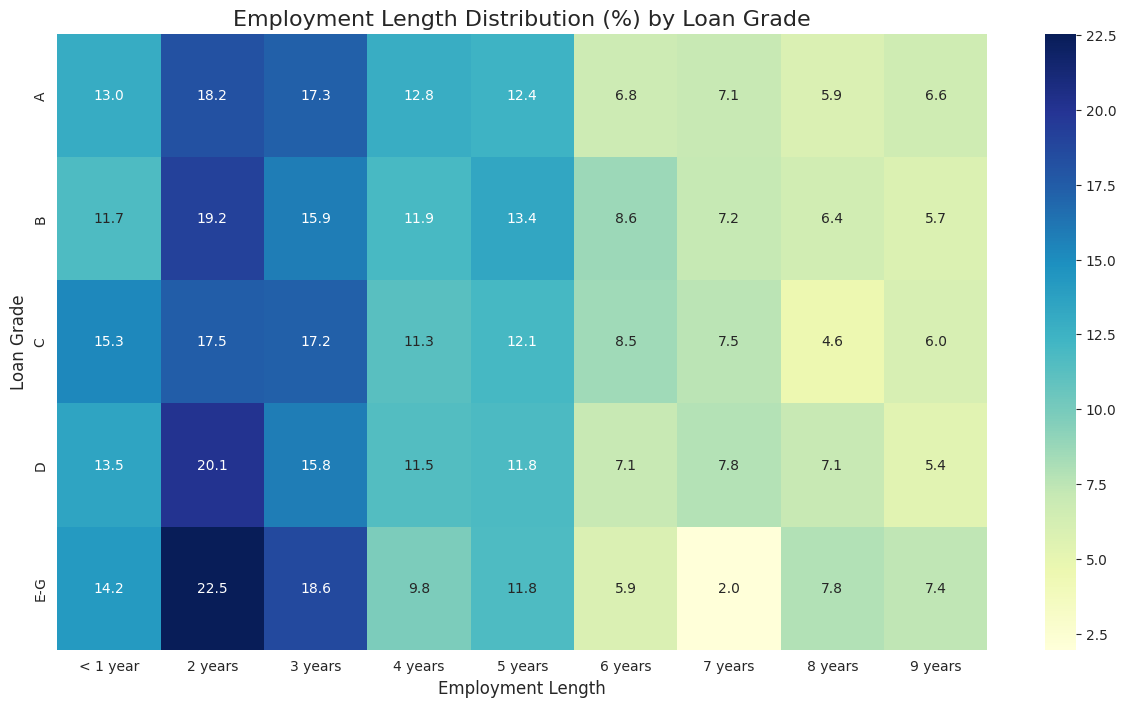
**

Figure 2 Correlation Plot



Kruskal-Wallis Test: H-statistic: 154.14: This large value indicates a substantial difference in the income distributions among the grade groups.P-value: 3.367e-33: This is an extremely small number (effectively zero).The null hypothesis stated that all loan grades share the same median annual income. This result confirms, with very high statistical confidence, that there is a significant difference in income among the grades.Post-Hoc Pairwise Comparisons: The Kruskal-Wallis test told us that a difference exists, and the post-hoc tests tell us exactly where those differences lie. The table and the heatmap show the p-values for each pair-by-pair comparison. The post-hoc analysis shows that the median annual income is statistically different between every single pair of loan grades.

*C.3 Does the length of a borrower's employment history correlate with their loan grade?*



A clear pattern indicates that higher loan grades are linked to longer work histories. Grade 'A' has the largest percentage of borrowers having '10+ years' of work experience, which progressively declines as the grade deteriorates. On the other hand, the riskier grades include a larger percentage of borrowers with shorter job histories. The idea that lenders see longer employment as an indication of stability and less risk is supported by this. However, with chi-sqr test The p-value (0.1506) is greater than 0.05. We fail to reject the null hypothesis.There is no statistically significant association between loan grade and employment length.

1. *Missingness*

There were missing values for ten predictor variables. A preliminary investigation was carried out in order to comprehend the underlying structure of the missing data. This demonstrated that the missingness showed clear, systematic patterns rather than being entirely random (MCAR). A targeted and rational imputation strategy is required instead of a uniform approach, as this structured absence of data indicates that the missingness mechanism is either Missing at Random (MAR) or Missing Not at Random (MNAR). 

Figure 3 Plot of Missing Values

>Group 1: Joint Application Variables (MAR): This group includes annual\_income\_joint, verification\_income\_joint, and debt\_to\_income\_joint. When the loan application is of the 'Individual' type, the data for these variables is consistently missing. This pattern is categorized as Missing at Random (MAR) since the lack of data is entirely explained by another observable variable (application\_type). These columns are removed due to the high percentage of missing values (>85%).

>Group 2: Employment and Credit Inquiry Variables (MAR): This group contains emp\_length, emp\_title, debt\_to\_income, and months\_since\_last\_credit\_inquiry. The missingness in these variables is also considered MAR. Missing employment data is conditional on the applicant being unemployed or self-employed. Which is issued in Lending Clubs forums[x].

Group 3: Delinquency History Variables (MNAR): This group, containing months\_since\_last\_delinq and months\_since\_90d\_late, is a classic example of Missing Not at Random (MNAR). The reason for the missing data is that these borrowers have never experienced a delinquency event.

According to the knowledge above, imputation techniques are divided like;  
Imputation with Mode:

This method fills missing values with the most frequently occurring value in the column. Variables imputed this way:

-emp\_length (employment length)

-num\_accounts\_120d\_past\_due (number of accounts 120+ days past due)

Predictive Imputation (using IterativeImputer):This is a more advanced method. It uses all other columns in the dataset to predict what the missing value should be, essentially running a regression model to estimate the best value.Variables imputed this way:

-debt\_to\_income (debt-to-income ratio)

-months\_since\_last\_credit\_inquiry

Imputation with Median:This method fills missing values with the median (the middle value) of the column.Variables imputed this way:

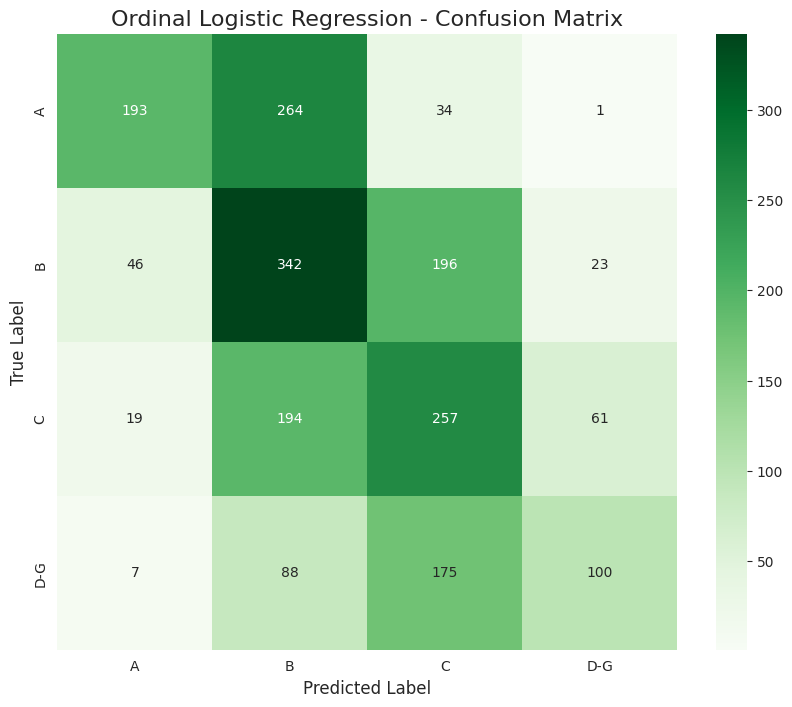
-months\_since\_last\_delinq



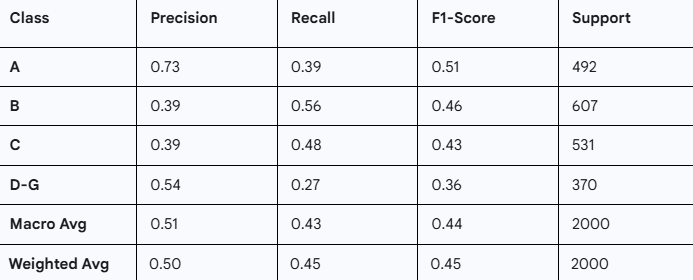
Also to make it better, last groups are combined.

## Modelling

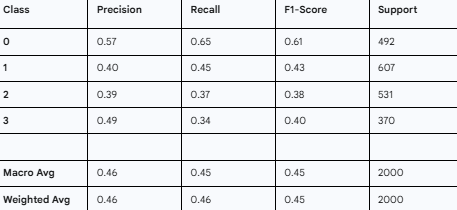
## Ordered Logistic Regression



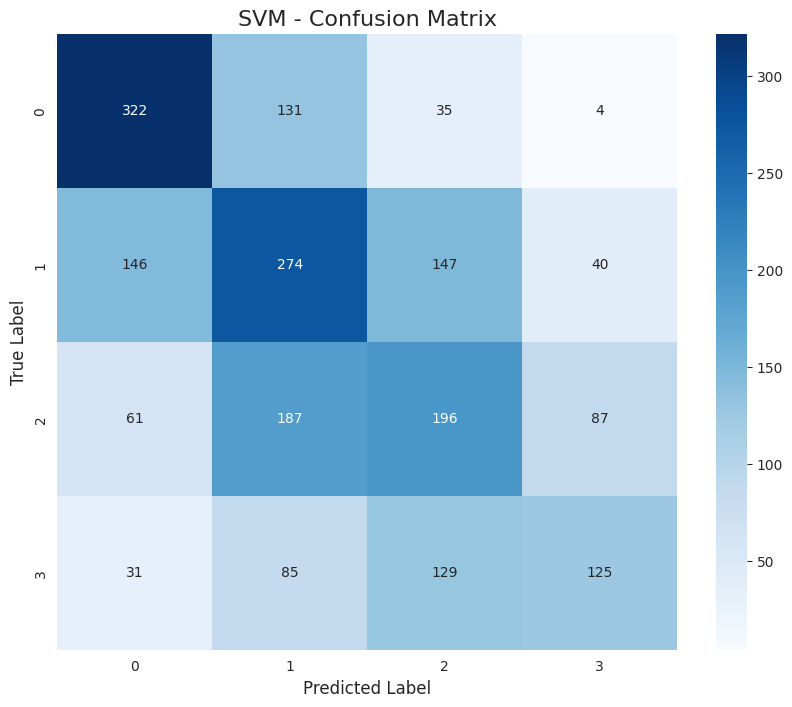
Contradictory Performance on Class 'A': For the best loan grade, the model shows very poor recall (39%) but very high accuracy (73%) instead. That implies the model is very likely to be right when it says a loan is 'Grade A'. It is much overly conservative, though, since it incorrectly classifies most (61%) of actual "A" grade loans as "B" or "C."



For also Likelihood Test;  
P-value (0.6599): The highest significant result. The performance of the entire model and the simplified model do not differ statistically significantly, as indicated by a p-value this high (far above the conventional cutoff of 0.05).According to the result, there is no statistically significant improvement over the reduced model from the extra characteristics in the complete model.

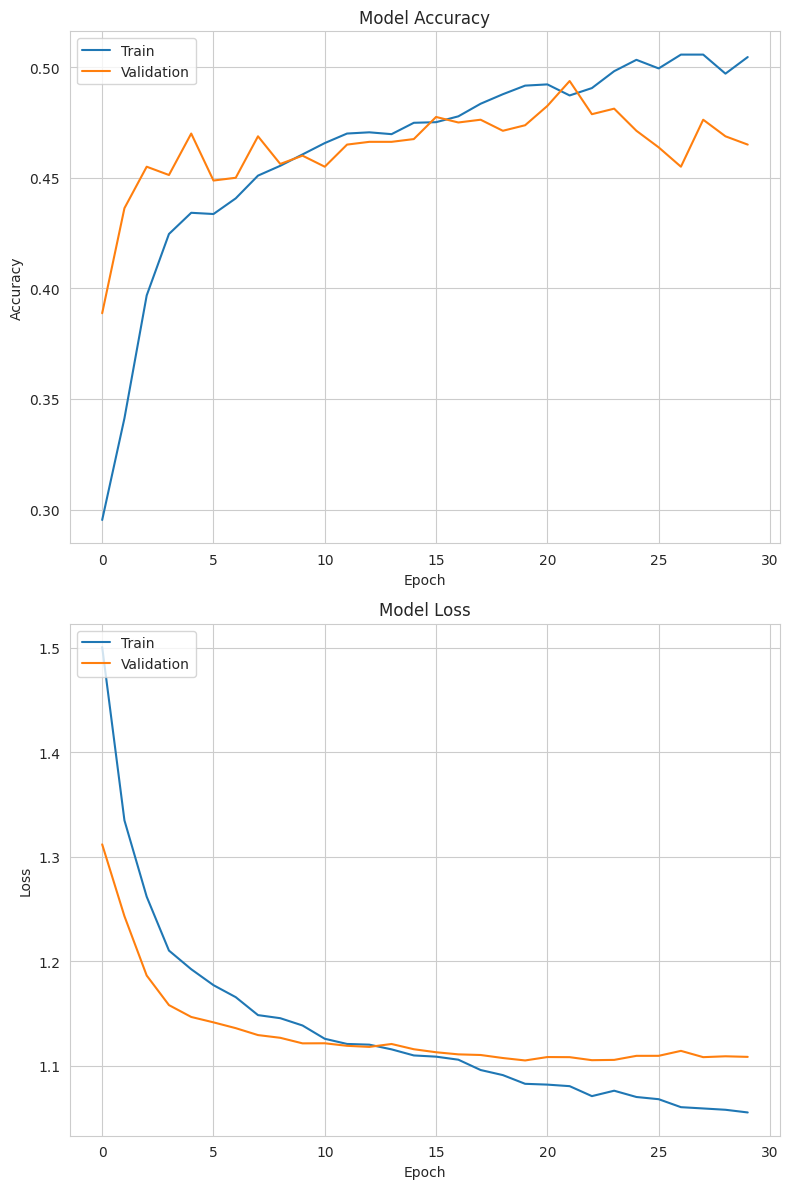
1. *Support Vector Machine* 

The accuracy (0.49) of the model is greater than the recall (0.34) for Class 3, which is the group at the highest risk. This suggests that the algorithm is hesitant to provide a high-risk rating to a loan. It has a reasonable level of confidence when it does predict Class 3. Its limited recall, however, indicates that it misclassifies most real high-risk loans as safer categories rather than identifying them.

**

The model identifies low risk better.

1. *A**rtificial Neural Networks*



Important information about how the Artificial Neural Network (ANN) learns throughout the training epochs is provided by the training history plots.Model Accuracy (Top Plot): The graph demonstrates that the validation accuracy (orange line) and training accuracy (blue line) both steadily improved and closely followed one another. This is encouraging since it shows that the model was able to identify patterns in the data without remembering them right away. As training comes to a close, the accuracy for both sets seems to plateau, indicating that the model has performed at its best.

Model Loss (Bottom Plot): For both the training and validation sets, the loss plot, which displays the model's error, steadily declines. The fact that the validation loss successfully reflects the training loss indicates that the Dropout layers were effective in avoiding notable overfitting. When the validation loss stopped getting better, the EarlyStopping callback probably ended the training, preserving the best version of the model.In summary, the training procedure was reliable and efficient. Although the model learnt effectively without overfitting, the classification task's intrinsic complexity was evident when its prediction power eventually stopped at an accuracy of about 50%.

1. *Random Forests*

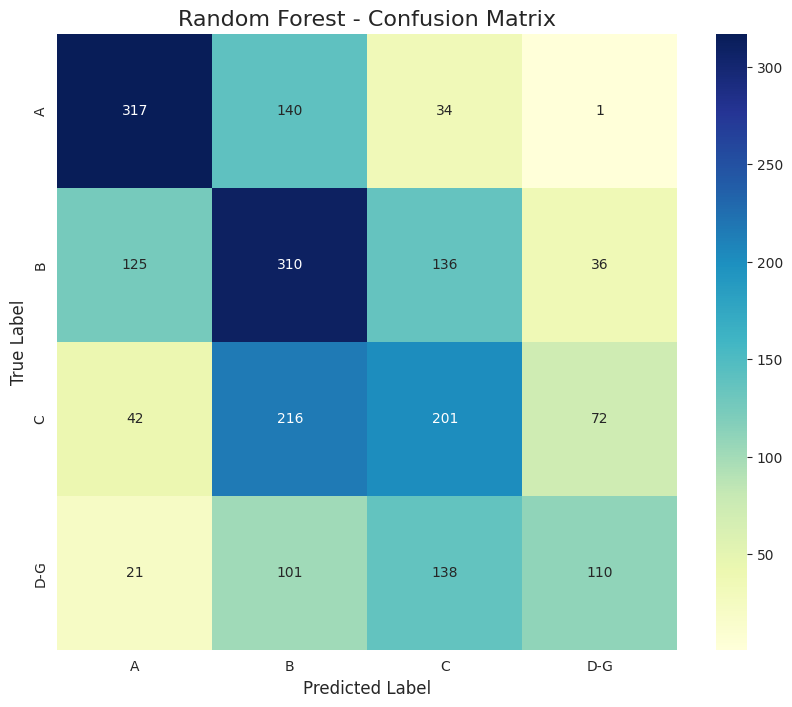
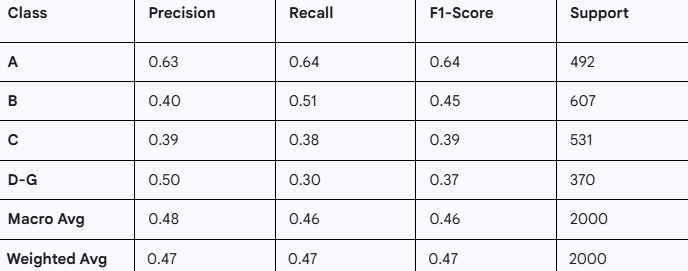
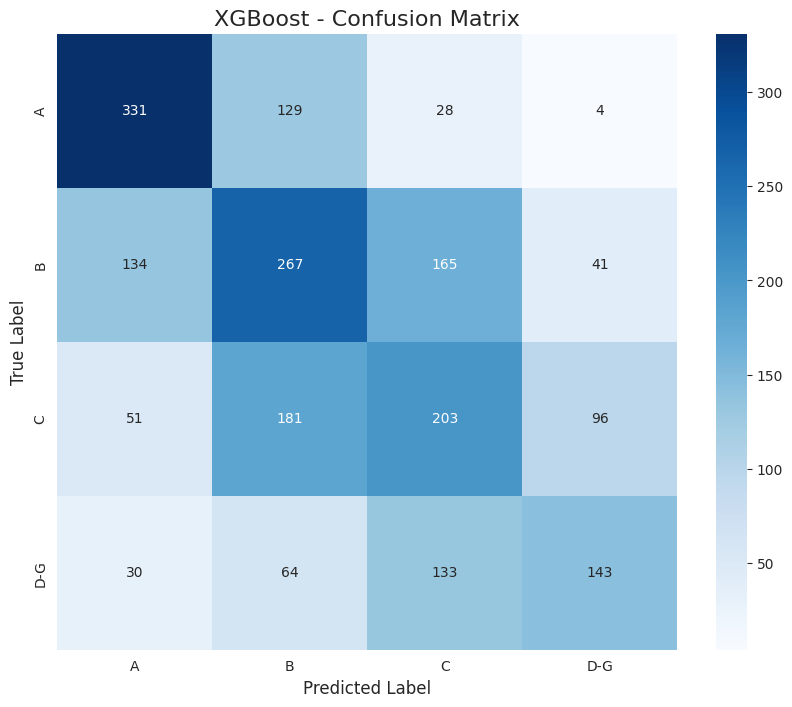


Figure 6 Variable Importance of Random Forest

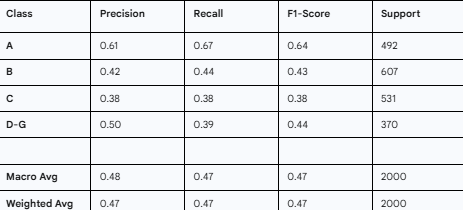
There is no apparent between the Random Forest model and the other approaches. The other models are a more practical choice because of its significant underperformance in recognizing the highest-risk 'D-G' category, even though it is better at predicting 'B' grade loans.



1. *XgBoost*

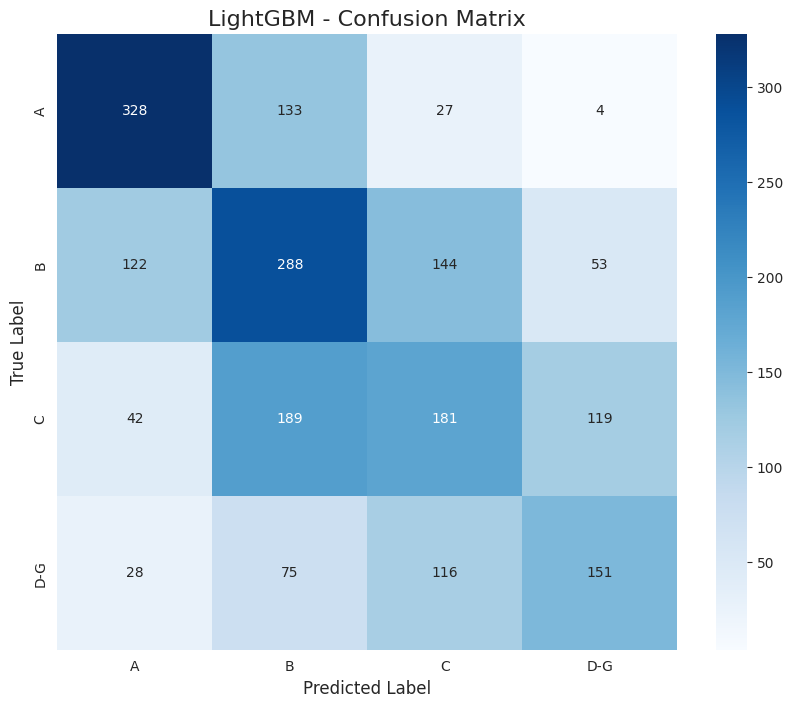


As the risk goes higher accuracy drops. The model struggles to distinguish the middle and low grades.

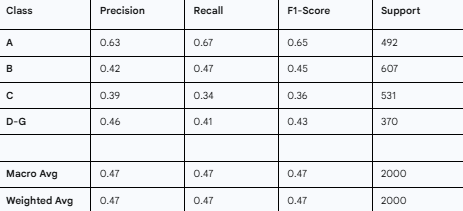


Accuracy of the model is 47.2%. Cohen's Kappa value is 0.2856.The model's predictions and the actual loan grades only show "fair" agreement, which is not great. Identifying the best loan grade ('A') is where the model excels, but it has severe problems with all other categories, particularly the high-risk 'D-G' category. 67 percent of all real "A" grade loans are appropriately identified.

1. *LightGBM*

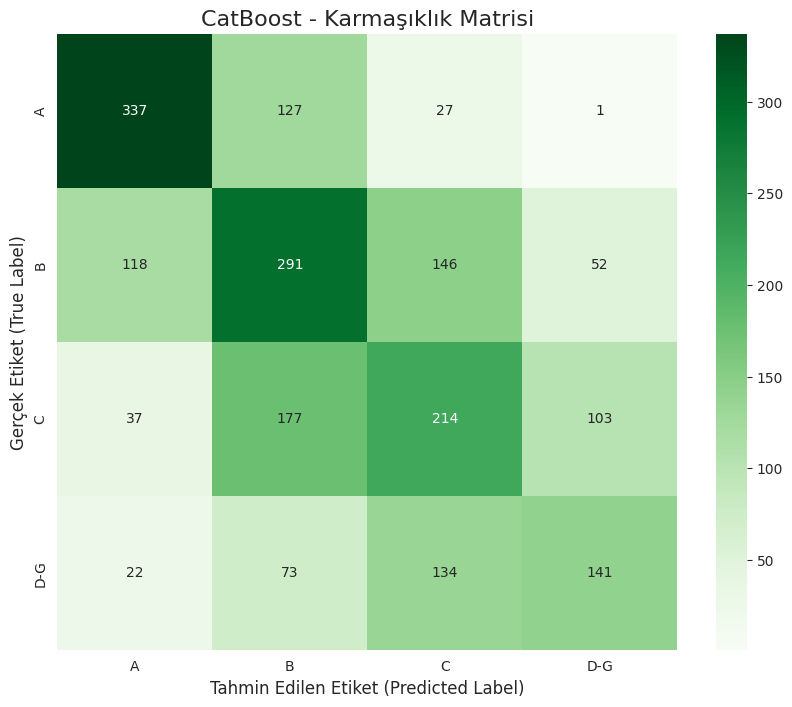


For LightGBM, also it predicts the higher grades better than lower.

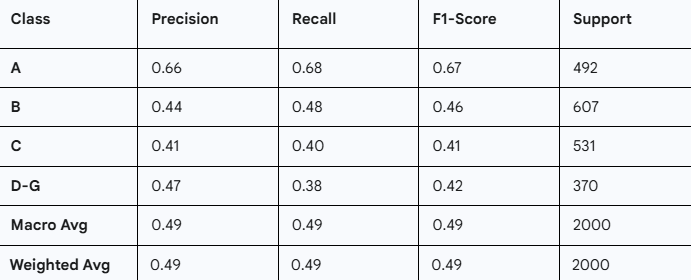


The model performs best on this category. With a recall of 67%. For 'B' grade loans, the model performs worse. The F1-score of 0.45 indicates considerable confusion with other courses and shows a poor balance between accuracy and recall.

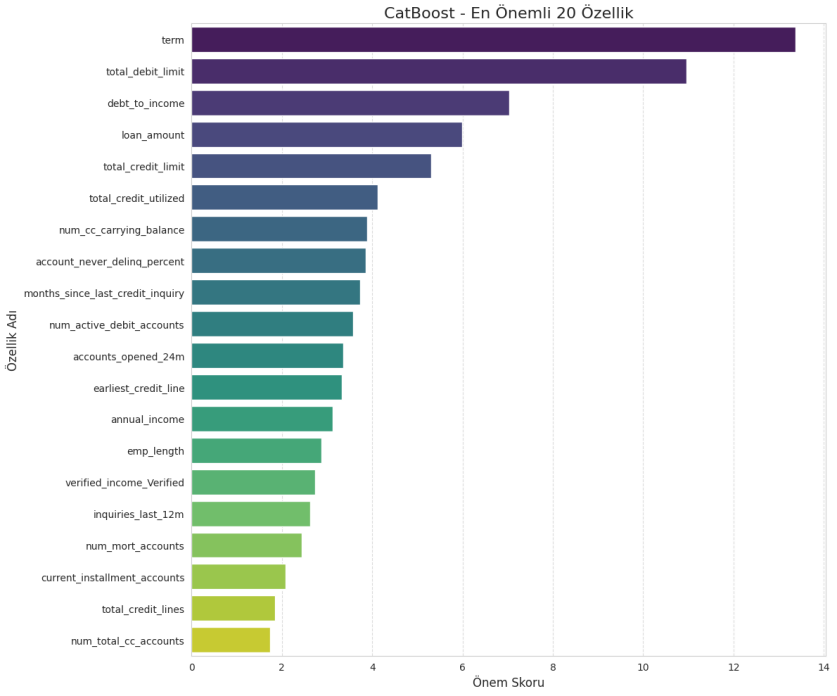
1. *Catboost*



Out of all the studied algorithms, the CatBoost model performs the best, with an overall accuracy of 49.15%. Its excellent macro and weighted average F1-scores demonstrate that it performs the most evenly and with the best accuracy across all classes.Outstanding Results on Class 'A': For 'Grade A' loans, the CatBoost model had the highest recall of any model, at 68%. This suggests that it is the best at accurately determining which applications provide the least danger.Its performance on classes 'B' and 'C' is competitive with the other top models, demonstrating a solid balance between not being unduly prejudiced (like the SVM) and not being overly reluctant (like the Ordinal model), even if it still finds the task challenging.



Feature Importance for catboost:



Most Significant Aspect: The term, or the loan's duration (such as 36 or 60 months), has the most impact. It has the greatest priority score, as indicated by its lengthy bar.Important Financial Data: After then, the model mostly depends on fundamental measures of financial health, such as debt-to-income and total\_debit\_limit.Decreasing Importance: Each feature's influence on the model's final decision reduces as you proceed down the list.

# RESULTS

These models' performances differed greatly. The total accuracy of the Support Vector Machine (SVM) model was 46.1%. Its inability to differentiate between neighboring classes and propensity to underestimate risk hindered its effectiveness; for example, it often incorrectly categorized high-risk 'D-G' loans into the safer 'C' group. With an accuracy of 47.5%, the Artificial Neural Network fared marginally better. The model's training process was consistent and prevented overfitting, but it had trouble properly classifying middle- and high-risk loan grades. For 'C' grade loans, its recall was very poor at 34%. Similar to other models, the XGBoost model failed to detect 61% of loans in the critical high-risk 'D-G' category, although achieving an accuracy of 47.2%. This caused by model avoids to categorize high-risk.

# CONCLUSION

This study used a huge financial dataset and a thorough analysis to predict loan risk grades. Significant correlations between the loan grade and important borrower characteristics, including yearly income, debt-to-income ratio, and length of work, were found in the preliminary exploratory data analysis. Several machine learning models were trained and assessed after data preparation, which included imputed missing values and data leak prevention. The CatBoost model's feature significance analysis supports the EDA's conclusions, emphasizing that the most important criteria influencing a borrower's loan grade are their overall credit history and debt management (term, total\_debit\_limit, and debt\_to\_income). This study concludes by showing that sophisticated machine learning models—in particular, gradient boosting algorithms like CatBoost—offer a reliable and efficient framework for automating and raising the threshold for loan risk categorization accuracy.

# REFERENCES

References

1. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

[2] Kim, J., & Ahn, J. (2012). Support vector machines for ordinal regression and its application to credit rating prediction. International Journal of Forecasting, 28(3), 524-534.

[3] Goldmann, A., Szimayer, A., & Alkan, S. (2024). A MIDAS approach to corporate credit rating prediction with ordinal logistic regression. Journal of Forecasting.