# Risk Management in Home Equity Loans: Data-Driven Approach to Forecast Loan Defaults

## Final Submission
## Loan Default Prediction


### Juan Campos


(**MIT**-PE ADSP March 23'C)

**June 13th, 2023**

Great Learning
POWER AHEAD

# Contents / Agenda

# Business Problem Overview

Credit default is a significant issue faced by banks and other financial institutions. Defaulting on loans affects the institutions' financial stability and increases loan interest rates for all customers, even those who consistently make their payments on time. It is estimated that a notable percentage of a retail bank's profits come from interest on home loans. Unfortunately, Non-Performing Assets (NPAs) or bad loans can substantially impact these profits, rendering some loans more of a liability than an asset.

Traditional loan approval processes are often manual, time-consuming, and prone to errors and biases. These processes can sometimes lead to the approval of loans for customers who are likely to default or, conversely, the rejection of customers who are capable of loan repayment.

By leveraging data science and machine learning techniques, the company can automate these processes, significantly reducing human error and bias and leading to more accurate predictions of creditworthiness. According to a 2019 report by the McKinsey Global Institute, it's estimated that the integration of AI technologies could yield up to $1 trillion in additional value each year for the global banking sector. This potential value stems from a broad range of applications, from risk management to customer experience enhancement. However, many banks have found it challenging to scale AI technologies across the entire organization, often getting stuck in the experimental phase for specific use cases[1].

JPMorgan Chase is a concrete example of a bank leveraging AI to drive value. Their ambitious tech strategy, which includes substantial investments in AI and data analytics, aims to deliver $1.5 billion in business impact by the close of 2023. These efforts are complemented by an infrastructure modernization program that is expected to yield an additional $1.5 billion in cost savings and efficiencies over the same period. Interestingly, the bank has already seen significant success with its AI-driven initiatives. In 2022 alone, using AI for personalized offerings in retail and for suggesting products and growth plans in commercial business generated $320 million. Currently, JPMorgan Chase has more than 300 AI use cases in production, demonstrating the broad scope of AI's potential impact on banking[2].

Therefore, addressing this problem and creating a reliable, unbiased, and interpretable classification model that can predict potential defaulters with high accuracy is crucial. By doing so, financial institutions can protect their profits and maintain lower interest rates, provide better customer service, and uphold their reputation in the market.

---

[1] McKinsey Global Institute. (2019). AI in banking: Can banks meet the challenge? Retrieved from AI in banking: Can banks meet the challenge? | McKinsey

[2] American Banker. (2023). JPMorgan Chase aims to create $1.5 billion in value with AI by yearend. Retrieved from JPMorgan Chase aims to create $1.5 billion in value with AI by yearend | American Banker

# Objectives

The main goal is to build an accurate and interpretable classification model to predict clients likely to default on their loans. This model will consider essential features such as the amount of loan approved, the amount due on an existing mortgage, the current value of the property, the reason for the loan request, job type, years at the present job, number of major derogatory reports, number of delinquent credit lines, age of the oldest credit line, number of recent credit inquiries, number of existing credit lines, and debt-to-income ratio.

By accomplishing this, we aim to achieve the following objectives:

- **Enhance the Credit Approval Process:** Improve the credit approval process's efficiency and accuracy.
- **Reduce the Non-Performing Assets (NPA):** Decrease the number of loan defaults, leading to higher profits for the bank.
- **Minimize Bias:** Ensure fairness in loan approvals by eliminating human bias and adhering to the Equal Credit Opportunity Act guidelines.
- **Improve Customer Experience:** By making accurate and quick decisions, enhance customer satisfaction and maintain the bank's reputation.
- **Ensure Compliance and Explainability:** Build an interpretable model to justify and explain loan approval/rejection decisions.

By meeting these objectives, the company can reduce financial risk, improve service, and strengthen its position in the financial market.

# Data Overview

The dataset provided has 5960 entries, each associated with 13 distinct attributes. These attributes provide crucial information about the applicant's loan and credit status and personal details like job type and reason for the loan request.

Our model aims to predict the "BAD" outcome variable, which indicates whether the lead turned into a default customer. This variable helps train the machine learning models to reduce financial risk.

It is important to note that several attributes have missing entries. Before further analysis, we'll need to handle these missing values.

We will also perform an exploratory data analysis to understand our data distribution better, identify outliers, and uncover potential correlations between the variables.

Despite the missing values, this dataset provides a robust base for building a machine-learning model to predict potential loan defaulters.

# Univariate Analysis

FIG 1

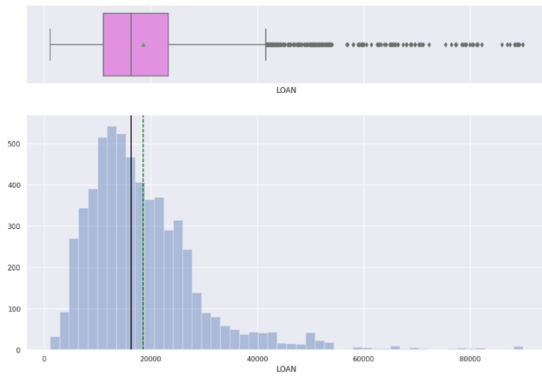



FIG 2

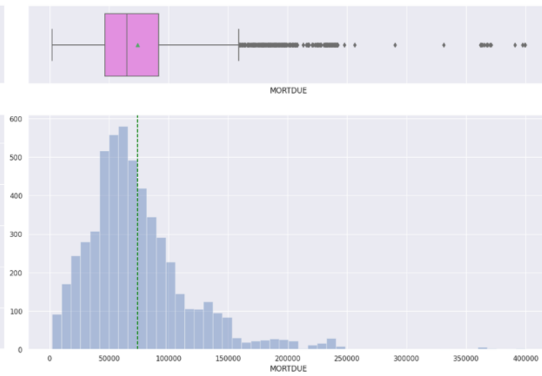



FIG 3

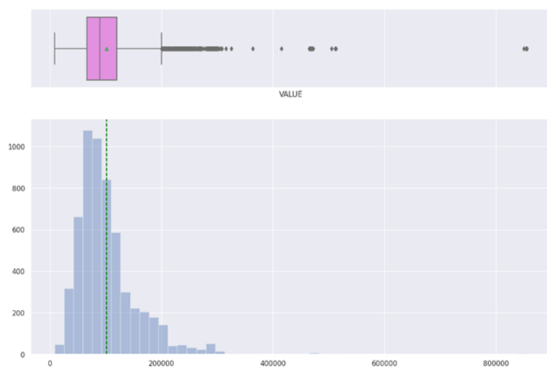
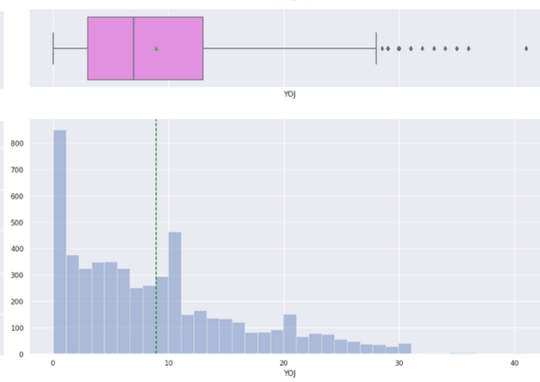



FIG 4




**LOAN (fig 1)**: The median loan amount is $16,300, implying that half of the loan amounts are less than this and half are more. The first quartile (Q1) at $11,100 and the third quartile (Q3) at $23,300 suggest that 50% of the loan amounts lie within this range. The minimum loan amount is $1,100, and the maximum is quite large at $89,900, indicating a wide range of loan values. The standard deviation (std) is quite significant, suggesting a high variation in loan amounts.

**MORTDUE (fig 2)**: The outstanding mortgage due shows a similar pattern to the loan amount. The median due is $65,019, and 50% of the dues range between $46,276 and $91,488. However, some mortgage dues are as high as $399,550, significantly skews the data.

**VALUE (fig 3)**: The median property value is about $89,235.5. Half of the properties in the dataset are valued between $66,075.5 and $119,824.25. The minimum property value is $8,000, and the maximum value is exceedingly large at $855,909, suggesting there might be some exceptionally valuable properties.

**YOJ (fig 4)**: Half of the borrowers have been in their current job for 3 to 13 years, with a median of 7 years. The range is from 0 to 41 years, which implies that some borrowers are either very new to their job or have a long tenure.

FIG 5

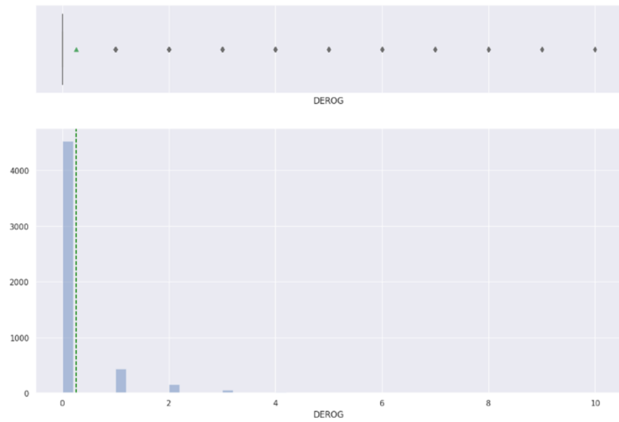
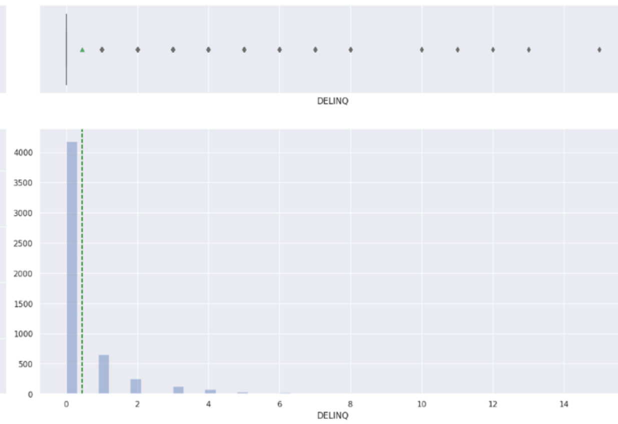



FIG 6




==========================================================================================

FIG 7

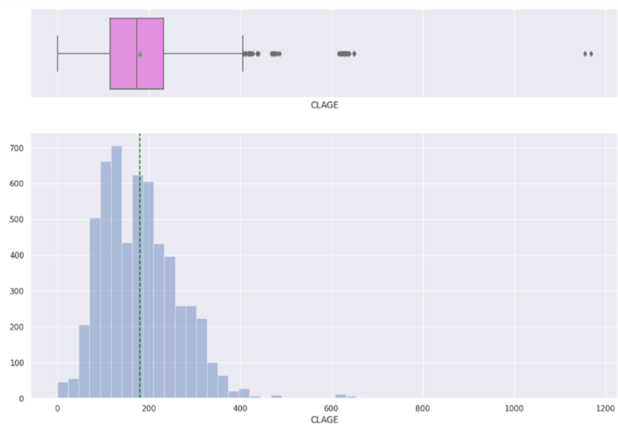
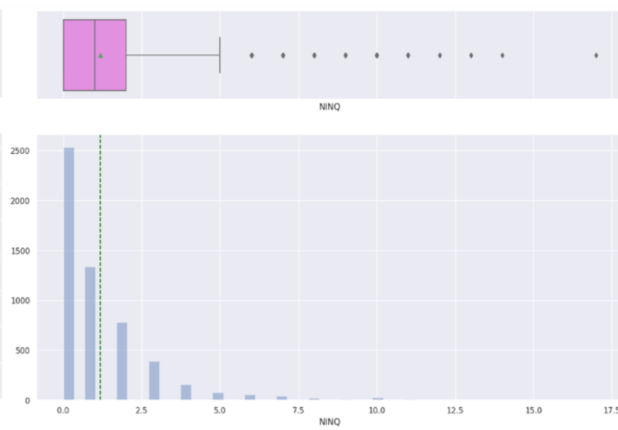



FIG 8




**DEROG (fig 5), DELINQ (fig 6)**: Most borrowers don't have derogatory remarks or delinquent credit lines, as the 25%, 50% (median), and 75% quartiles are all 0. The maximum number of derogatory remarks and delinquencies is 10 and 15, respectively, indicating a few borrowers with negative credit behaviours.

**CLAGE (fig 7)**: The age of the oldest credit line is measured in months. The median is around 173.47 months (roughly 14.5 years), and 50% of the credit lines fall within 115.12 and 231.56 months (approximately 9.6 to 19.3 years); this suggests a relatively old credit history for the borrowers.

**NINQ (fig 8)**: The number of recent credit inquiries shows a median of 1, meaning that half of the borrowers had at least one credit inquiry. The maximum goes up to 17 inquiries, which is significantly higher and could impact the credit scores of those borrowers.

**FIG 9**

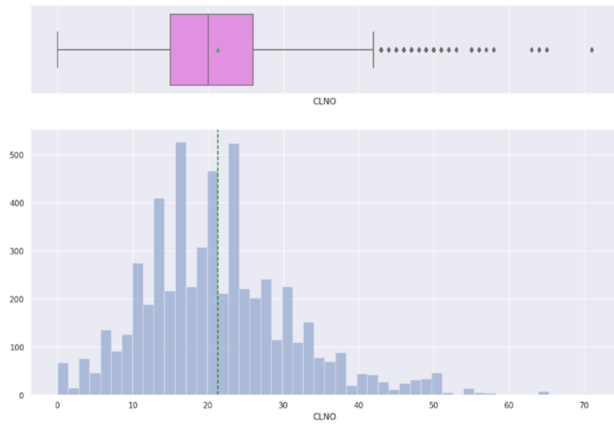
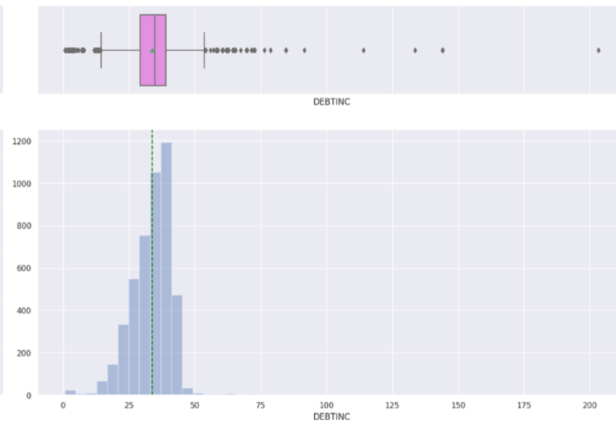



**FIG 10**




**CLNO (Fig 9)**: This variable represents the number of credit lines, which shows a median of 20, suggesting that half of the borrowers have 20 or more lines of credit. The maximum number of credit lines is 71, indicating some borrowers have access to extensive credit.

**DEBTINC (Fig 10)**: The debt-to-income ratio is a key risk metric. The median ratio is approximately 34.82, which means that debt represents around 34.82% of the income of a typical borrower. The maximum ratio is as high as 203.31, which is unusually high and suggests severe indebtedness.

In conclusion, the data varies and shows wide ranges for most variables. For some, like DEROG, and DELINQ, most values are zero, but outliers with high values suggest potential credit risk; Variables like LOAN, MORTDUE, and VALUE show wide distributions and high standard deviations, indicating high variability in these attributes among borrowers.

FIG 11

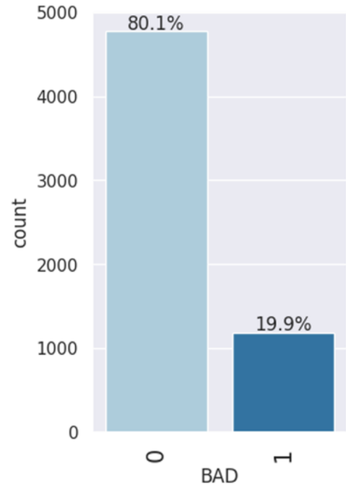



FIG 12

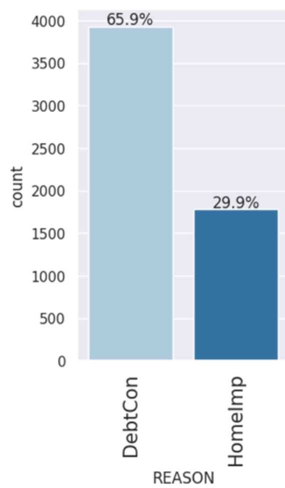
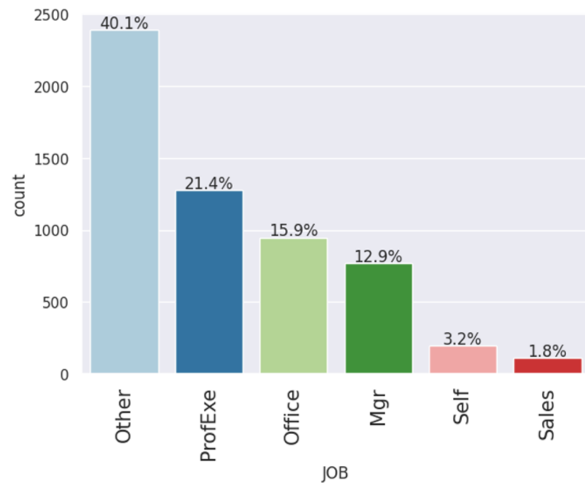



FIG 13




**BAD (fig 11):** This represents whether the applicant is a good or bad risk. In the dataset, 80.1% of the cases are '0' (not defaulters), and 19.9% are '1' (defaulters); this shows that most loan borrowers in this dataset are not defaulters.

**REASON (Fig 12):** This is the reason for the loan: debt consolidation (DebtCon) or home improvement (HomeImp). The bar plot indicates that 65.9% of loans are for debt consolidation, and 29.9% are for home improvement; this suggests that the most common reason for seeking a loan in this dataset is to consolidate other debts, with home improvement being a less common but still significant reason. These categories do not cover about 4.2% of data due to missing values.

**JOB (fig 13):** This represents the job category of the borrower, and the most common job category in this dataset is 'Other' (40.1%), followed by 'ProfExe' at 21.4%. Office workers represent 15.9%, and managers 12.9% of borrowers. Self-employed borrowers and salespeople are the least common, at 3.2% and 1.8%, respectively. This distribution can give insights into the demographic of borrowers; however, the 'Other' category is quite large and unspecified, which may limit the interpretability of this variable.

# Bivariate and Multivariate Analysis

FIG 14

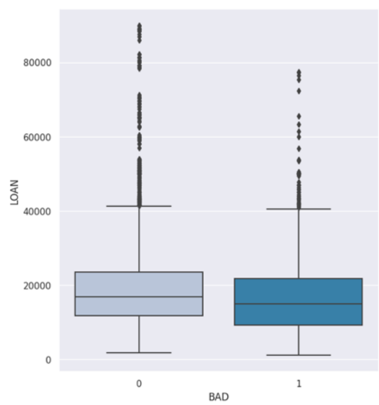



FIG 15

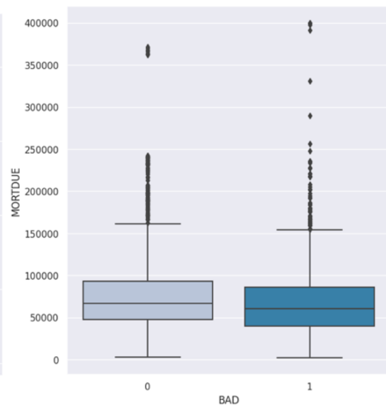



FIG 16

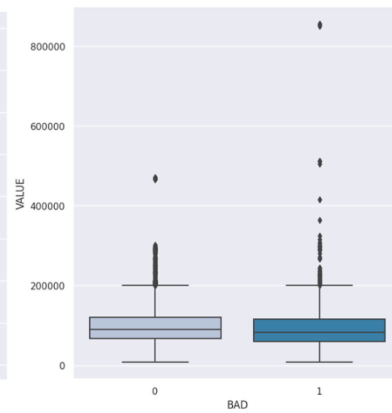



FIG 17

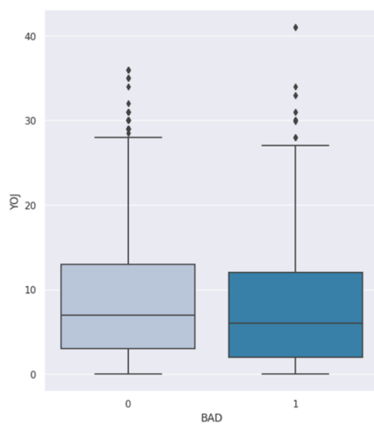
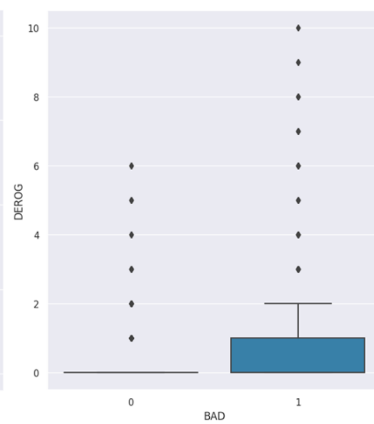



FIG 18




**Loan Amount (LOAN) (Fig 14):** For both categories 0 and 1, the median loan amount is lower for category 1 (14900.0) as compared to category 0 (16900.0). It's important to mention that there are several outliers in both categories with larger loan amounts, particularly in category 0; this could suggest that the risk represented by the "BAD" variable may be less associated with the loan amount itself, as there are loans of substantial size in both categories.

**Mortgage Due (MORTDUE) (Fig 15):** The median mortgage due for Category 0 is higher (66839.0) than Category 1 (60279.0). Additionally, the upper quartile (Q3) for category 0 is much higher, indicating that a higher mortgage due might be associated with a lower risk of falling into delinquency. However, we must consider many outliers for both categories before drawing conclusions.

**Property Value (VALUE) (Fig 16):** Similar to mortgage due, the median property value is higher for category 0 than category 1; there are numerous outliers in both categories on the higher end, more so in category 0, suggesting the possibility of less risk with higher-valued properties.

**Years at Job (YOJ) (Fig 17):** The median years at the job is higher for category 0 (7.0 years) compared to category 1 (6.0 years); although the difference isn't significant, we need to consider the presence of outliers in both categories that have a substantially high number of years on the job may deserve further investigation.

**Number of Derogatory Reports (DEROG) (Fig 18):** Interestingly, for category 0, the median, lower quartile (Q1), and upper quartile (Q3) are all 0. However, in Category 1, the upper quartile (Q3) is 1, and there are numerous outliers above this, suggesting that the presence of derogatory reports can be a significant risk factor for falling into delinquency.

FIG 19

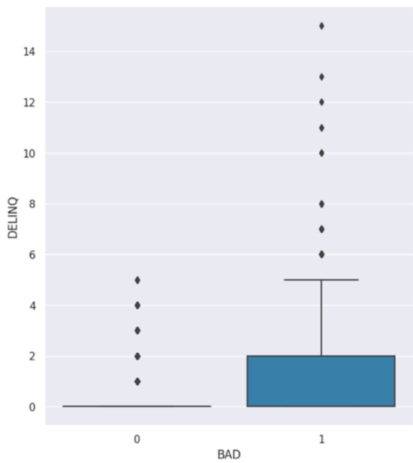



FIG 20

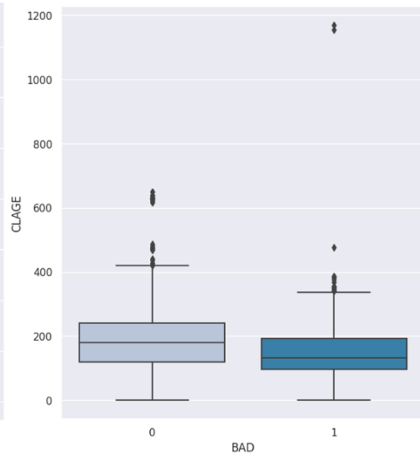



FIG 21

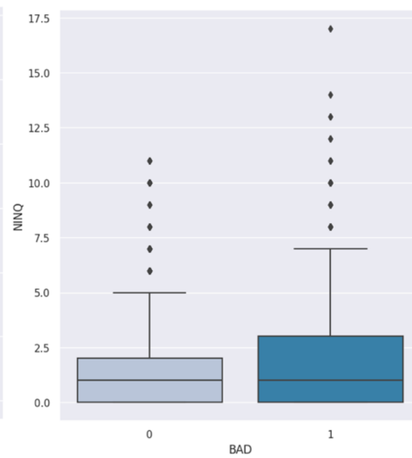



FIG 22

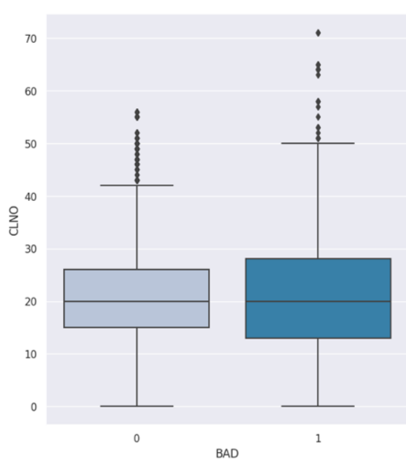
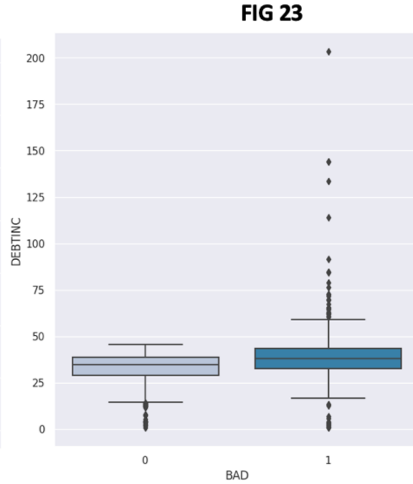



FIG 23




**Number of Delinquent Credit Lines (DELINQ) (Fig 19):** Similar to derogatory reports, delinquent credit lines are more common in category 1, as suggested by the higher median and Q3. Furthermore, outliers are more prominent and higher in Category 1, emphasizing the risk associated with delinquency.

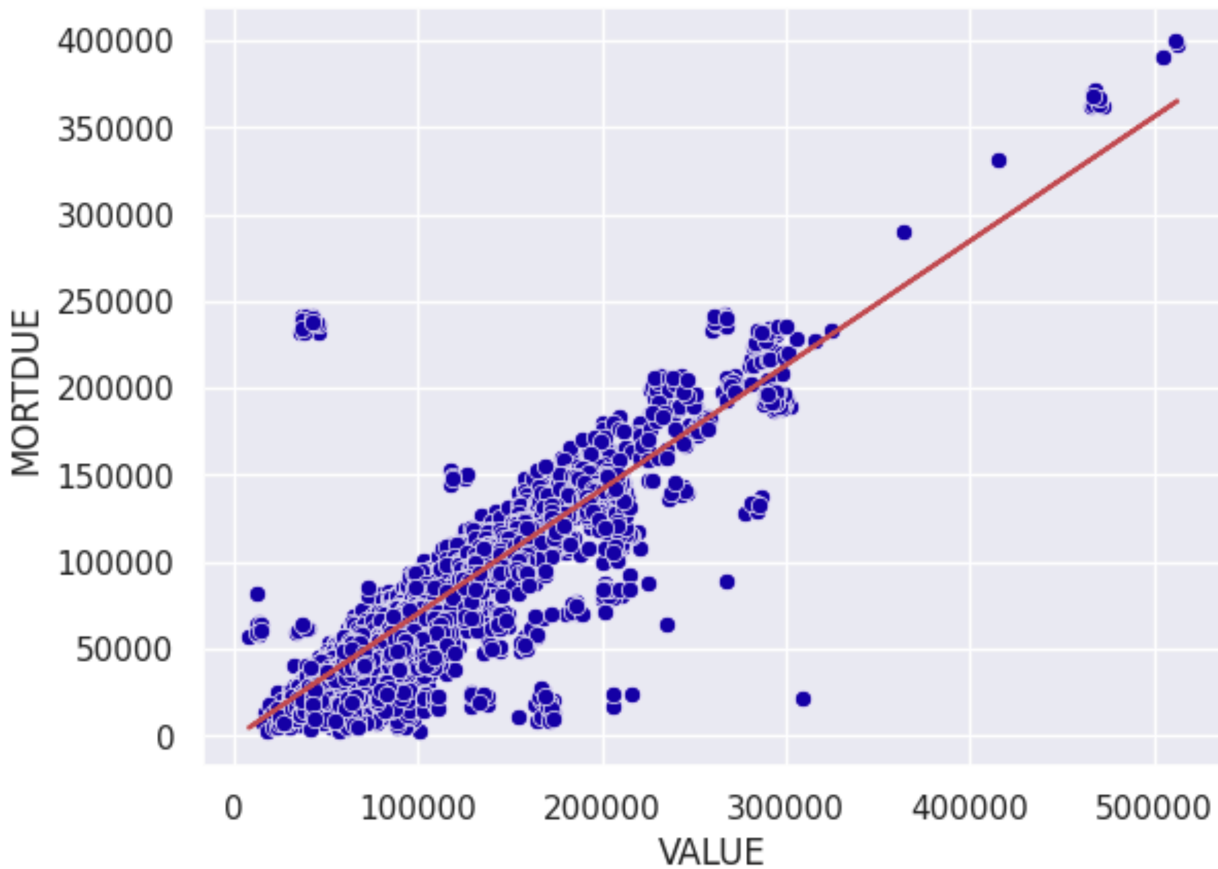
**Age of Oldest Credit Line (CLAGE) (Fig 20):** The median age of the oldest credit line is higher for category 0, indicating that a more extended credit history might be associated with a lower risk of delinquency. However, outliers in both categories warrant further investigation.

**Number of Recent Credit Inquiries (NINQ) (Fig 21):** While the median for both categories is 1, the upper quartile is higher for category 1 (Q3 = 3) compared to category 0 (Q3 = 2). There are also more outliers with more recent inquiries in Category 1, suggesting that more recent credit inquiries could be associated with increased risk.

**Number of Credit Lines (CLNO) (Fig 22):** The median number of credit lines is slightly higher for Category 0 than for Category 1. The overall spread of data and outliers doesn't clearly distinguish between the categories.

**Debt-to-Income Ratio (DEBTINC) (Fig 23):** The debt-to-income ratio contrasts categories 0 and 1. The median DEBTINC for category 1 (38.5) is significantly higher than for category 0 (30.8). This difference is also observed in the interquartile ranges, with category 1 showing a higher spread of values. The outliers in both categories fall on the higher end of the ratio. However, the considerable increase in DEBTINC values for Category 1 suggests that a high debt-to-income ratio can be a significant indicator for predicting the risk of defaulting on a loan. This is particularly notable as a high debt-to-income ratio can indicate that a borrower may have trouble repaying their loans.

FIG 24




**The scatterplot (Fig 24) of "VALUE"** (Home Value) against "MORTDUE" (Mortgage Due) exhibits a strong positive correlation, as shown by the high correlation coefficient of 0.876. This correlation implies that, as the Mortgage Due amount increases, the Home Value also tends to increase, and vice versa. The positive correlation between the two variables is expected in a typical real estate lending scenario, as the mortgage amount usually reflects a significant portion of the home's total value.

The slope of the regression line, 0.715, suggests that for every unit increase in the Mortgage Due amount, the Home Value increases by approximately 0.715 units, on average. This finding indicates that the Mortgage Due is a significant Home Value determinant but not a one-to-one relationship, as the slope is less than 1. It suggests that other factors contribute to the Home Value, and a portion of the value is independent of the Mortgage Due amount.

In conclusion, this analysis provides significant insight into how Mortgage Due affects Home Value. Since the Mortgage Due amount correlates strongly with Home Value, we can utilize these variables to decide on lending rates, loan approvals, and risk evaluation.

FIG 25

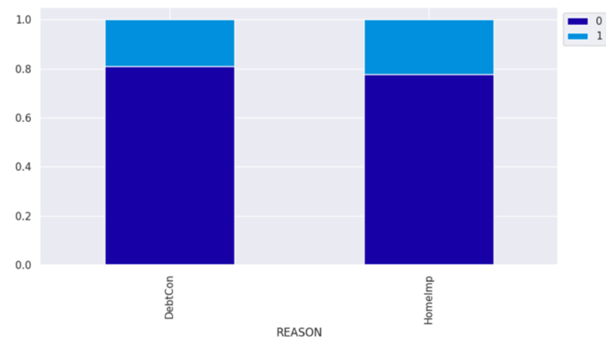
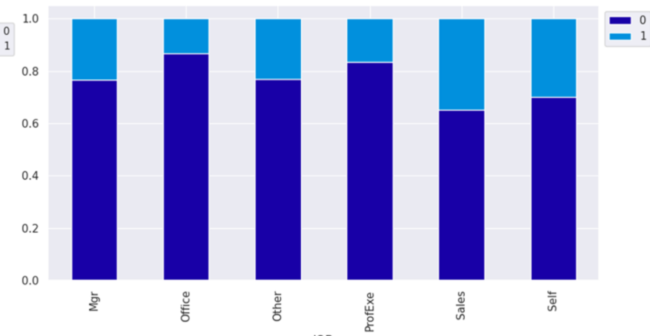



FIG 26

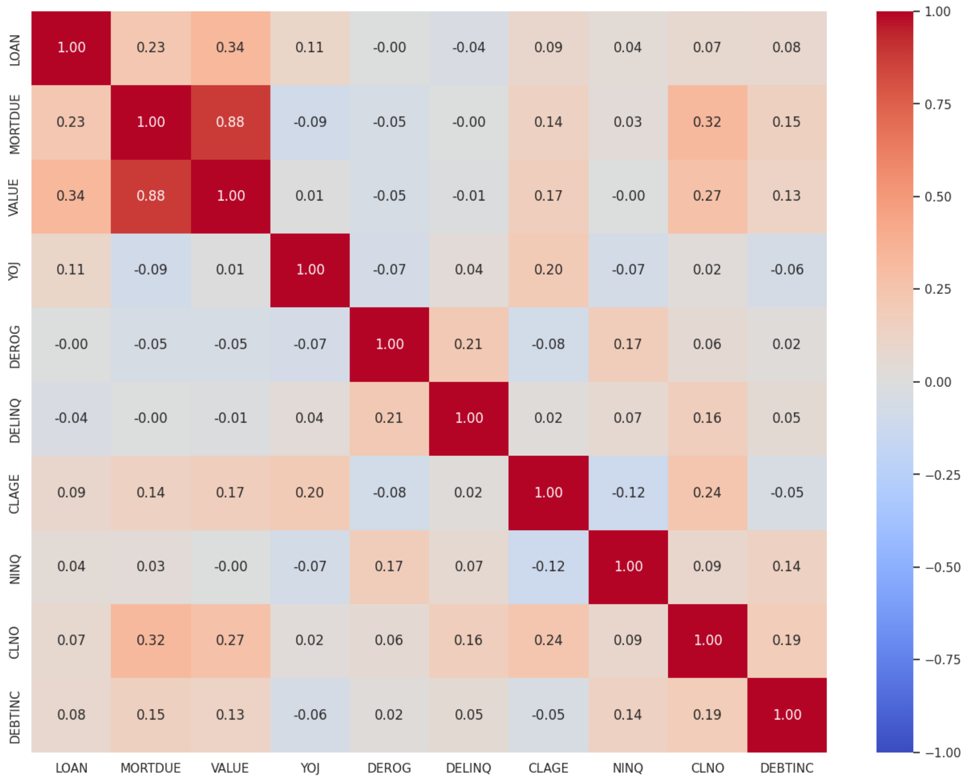



The stacked plots show the distributions of different categorical variables ('REASON' and 'JOB') against the 'BAD' variable, which indicates whether the customer is defaulted (1) or not (0).

**In the first stacked plot, the 'REASON' (Fig 25) variable** represents the reason for applying for the loan, either 'DebtCon' (debt consolidation) or 'HomeImp' (home improvement). It can be seen that there were 3,928 loans for debt consolidation, out of which 745 turned defaulted. This represents approximately 19% of the 'DebtCon' category. On the other hand, there were 1,780 loans for home improvement, out of which 396 were defaulted, representing around 22% of the 'HomeImp' category. Therefore, loans for home improvement are slightly riskier than debt consolidation loans.

**In the second stacked plot, the 'JOB' (Fig 26)** variable shows the job category of the loan applicants. Looking across different job categories, it appears that 'Other' has the highest number of loans at 2,388, with 554 being defaulted. This represents roughly 23% of the 'Other' category. Despite having the least total number of loans (109), the' Sales' category has the highest proportion of bad loans, with roughly 35% of sales job holders failing to pay their loans. In summary, these plots can be very helpful in profiling the risk associated with different categories.

**FIG 27**




**Mortgage Due and Property Value (Correlation = 0.88) (Fig 27)** Have a very strong positive correlation suggesting that, generally, as the property value increases, so does the Mortgage due. This makes sense as more valuable properties would likely result in higher mortgage amounts.

**Loan Amount and Property Value (Correlation = 0.34) (Fig 27) and Number of Credit Lines and Mortgage Due (Correlation = 0.32)** have a similar weak-moderate positive correlation.

**Credit Age and Years on Job (Correlation = 0.20) (Fig 27)** shows a weak positive correlation between the age of the oldest credit line and the years at the current job, suggesting that individuals who've worked longer also tend to have older credit lines.

**Number of Credit Lines and Debt to Income Ratio (Correlation = 0.19) (Fig 27)** shows a weak positive correlation indicating that the debt-to-income ratio also tends to increase as the number of credit lines increases.

**Delinquencies and Derogatories (Correlation = 0.21) (Fig 27)** show a weak-moderate positive correlation. This correlation suggests that customers with more delinquencies also tend to have derogatory remarks about their credit history.

From a business perspective, it's essential to understand these relationships to predict and manage credit risk. For example, customers with a higher number of credit lines and a high debt-to-income ratio could be riskier. Also, a high correlation between delinquencies and derogatories could indicate a pattern of poor credit behaviour.

# Data Preprocessing

To build our Classification Models to predict which leads are more likely to Default, we encoded categorical features, then split the data into train and test to be able to evaluate the models.

The training and test sets are consistent in their dimensions and class distributions, which is indispensable for constructing valid models.

To improve the accuracy of our model's prediction, we utilized the Interquartile Range (IQR) method, which is a reliable technique for handling outliers. This approach helps minimize the influence of extreme values on the data.

We used the median instead of the mean to handle missing numerical values because outliers less influence the median. For categorical variables, we replaced missing values with the mode.

We have created a binary flag for columns with missing values to retain potentially important information. The model will determine if the missing data has any predictive significance.

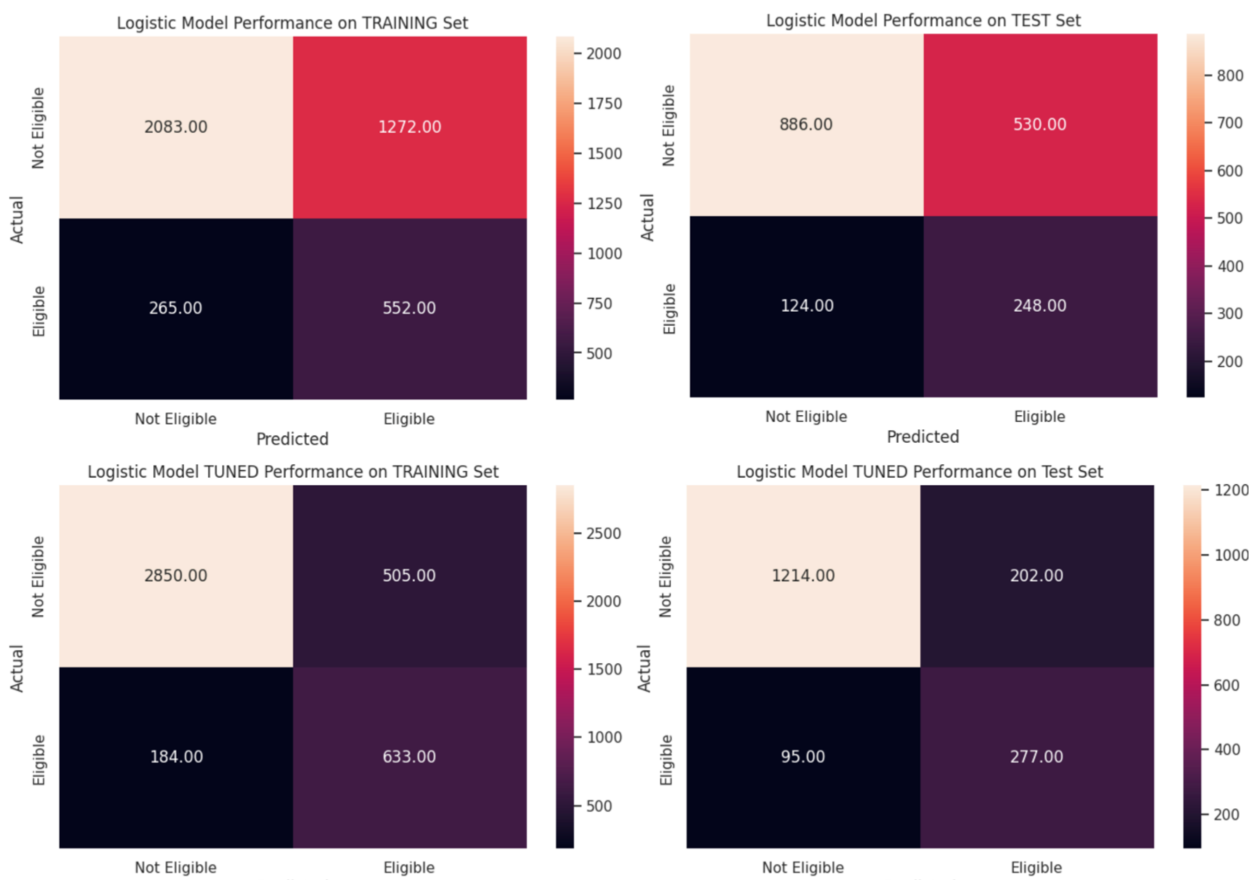
The dataset is slightly imbalanced, with about 80% for class 0 and 20% for class 1. This imbalance might affect the model's performance, potentially making it biased toward the majority class. To address this issue, we will adjust the class weight.

# Model Building and Performance Summary

We analyzed several models, such as Logistic Regression, Decision Tree, Random Forest, and XGBoost. We fine-tuned each model to improve performance and compared them to their original versions.

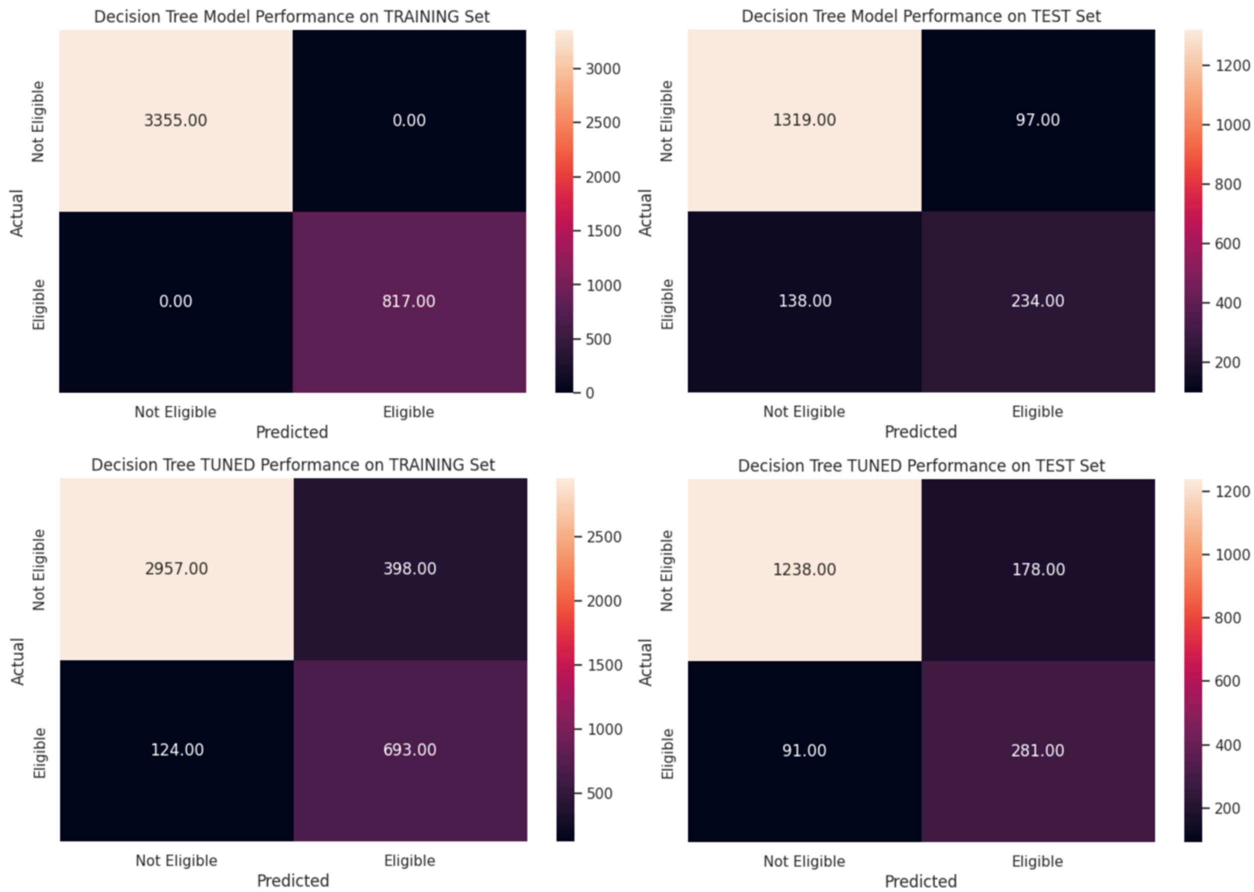
To establish a benchmark for measuring the impact of hyperparameter tuning, we built the logistic regression model, decision tree, random forest, and XGBoost model using default parameters. After the initial modelling, we fine-tuned each model to enhance its performance. This involved adjusting parameters such as maximum depth, minimum samples leaf, and learning rate.

***FIG 28***



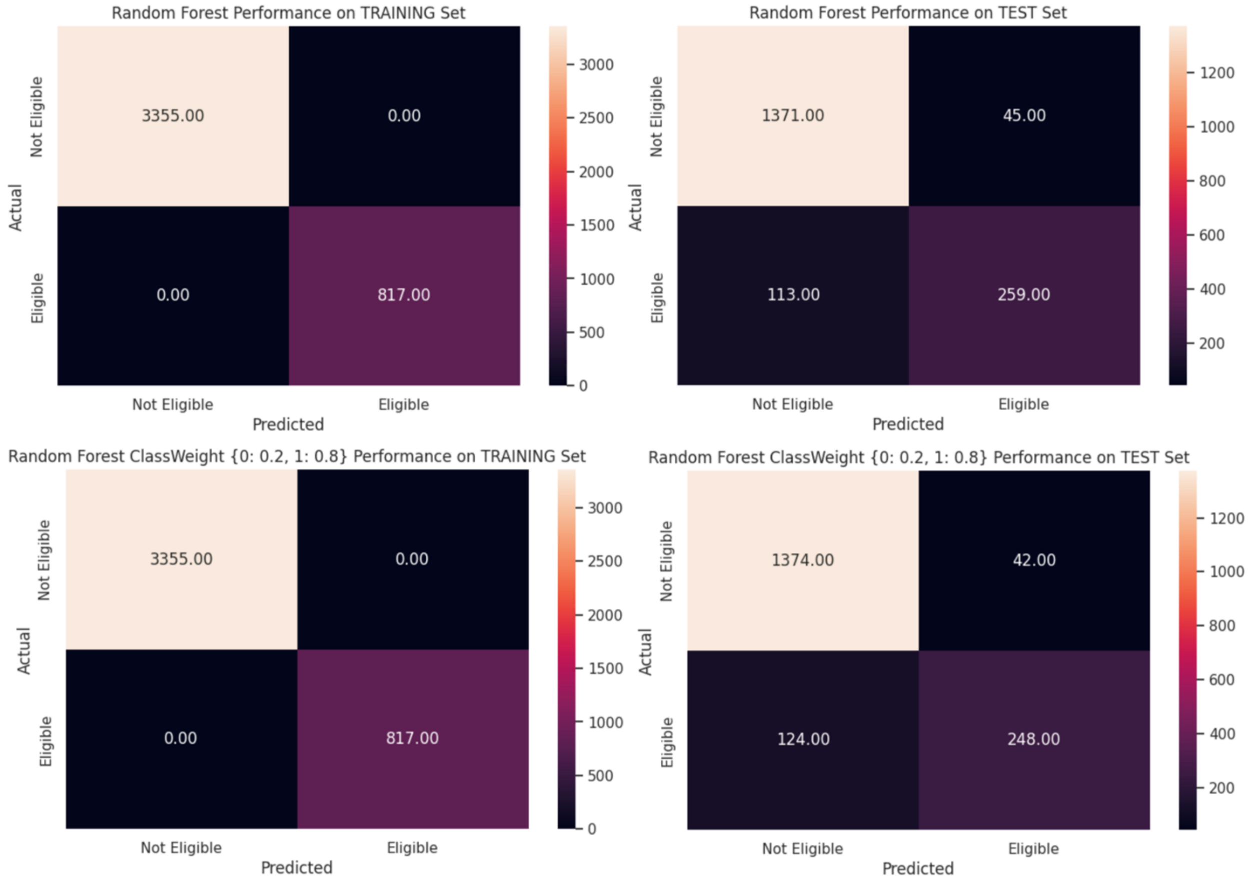

The **base logistic regression model (Fig 28)** performed modestly with an accuracy of 63% on both training and test datasets, with a recall of 67% on the training set and 65% on the test set. However, upon tuning the model parameters, there was a significant improvement in all metrics. The accuracy jumped to 83%, and the recall improved to 75% on the test set. This indicates that the tuned logistic regression model is much more robust in predicting credit risk.

***FIG 29***



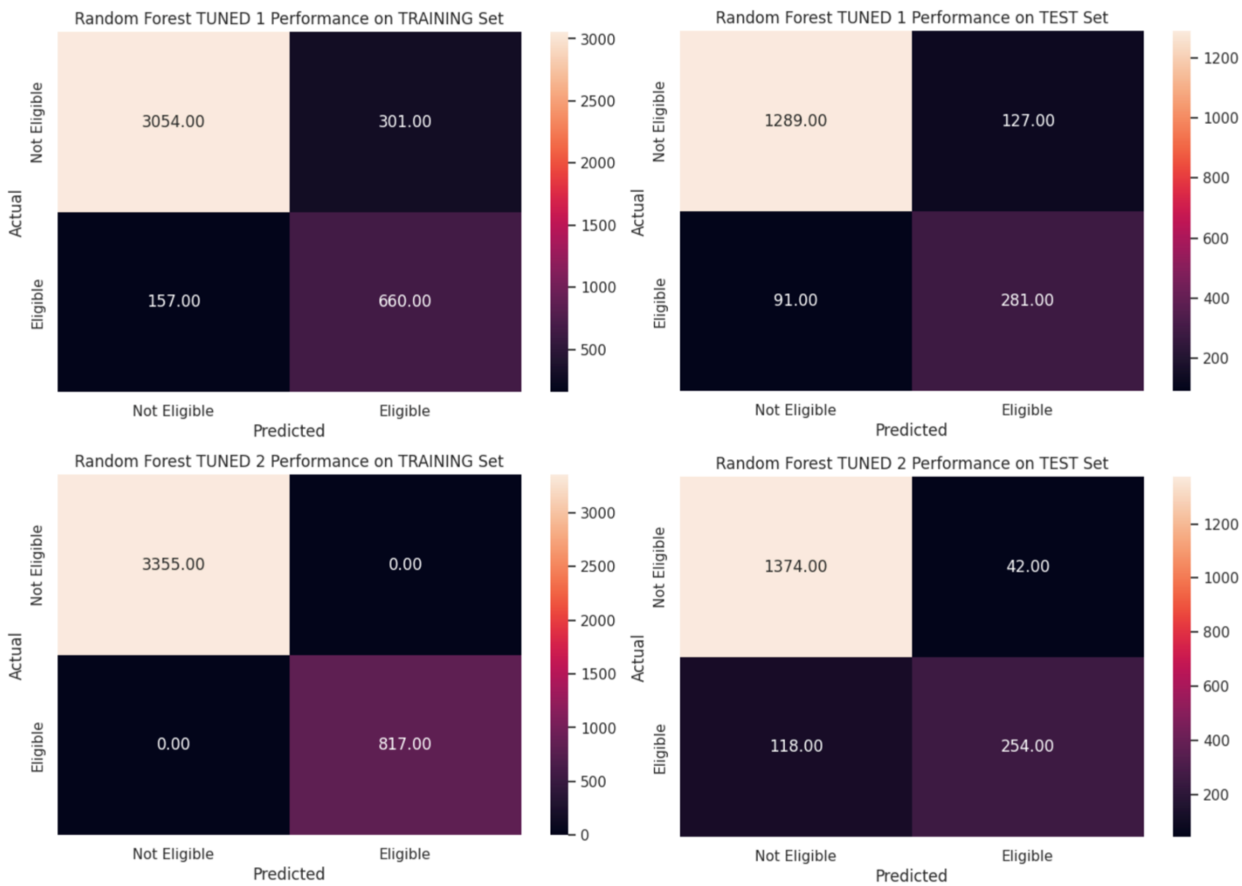

The **Decision Tree Model (Fig 29)** exhibited clear signs of overfitting in its base form, demonstrated by a perfect 100% accuracy on the training data but a drop to 87% on the test data. Overfitting suggests that the model is too complex and captures the noise in the training data, which shows a decrease in the performance on unseen data. After tuning, the model performance became more balanced, with an accuracy of 85% on the test set and a substantial improvement in recall to 76%.

***FIG 30***




We've explored four distinct models in our analysis of **Random Forest classifiers**. **The base Random Forest model (Fig 30)** performed exceptionally well on training data, demonstrating an accuracy of 100%. However, it showed a significant performance drop on the test data, reaching only a 91% accuracy, which indicates an overfitting issue. The base model's recall on the test data was 71%, suggesting it could correctly identify 71% of positive instances. Meanwhile, **the Class-Weight-Tuned Random Forest model (Fig 30)** maintained the overall accuracy but had a slightly lower recall rate of 67%. While this model proved adept at handling imbalanced data, it showed a slight decrease in correctly identifying positive instances.

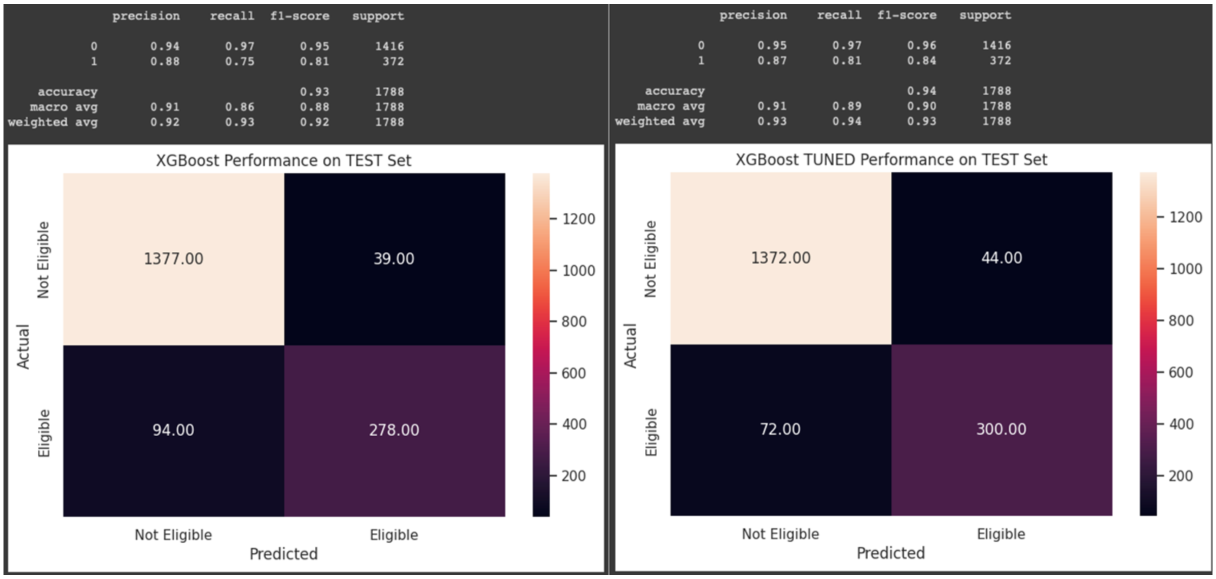
***FIG 31***




The third **Random Forest model, tuned with Grid Search (Fig 31),** demonstrated a more balanced performance. Despite its marginally lower accuracy of 88% compared to the base model, it showed an improved recall of 76%, making it more effective in identifying the positive class. **The final Random Forest model (Fig 31), which incorporated the best parameters**, matched the base model in accuracy (91%) but fell slightly short in recall (68%). While this model achieved commendable accuracy, its decreased recall indicates potential limitations in identifying positive instances correctly.

In conclusion, based on our evaluation, the **Random Forest model tuned with Grid Search (Fig 31)** provides the most balanced performance, with satisfactory accuracy and improved recall. Given the potentially high costs of false negatives in credit risk modelling, a model with higher recall might be more advantageous despite having a slightly lower accuracy.

***FIG 32-A***



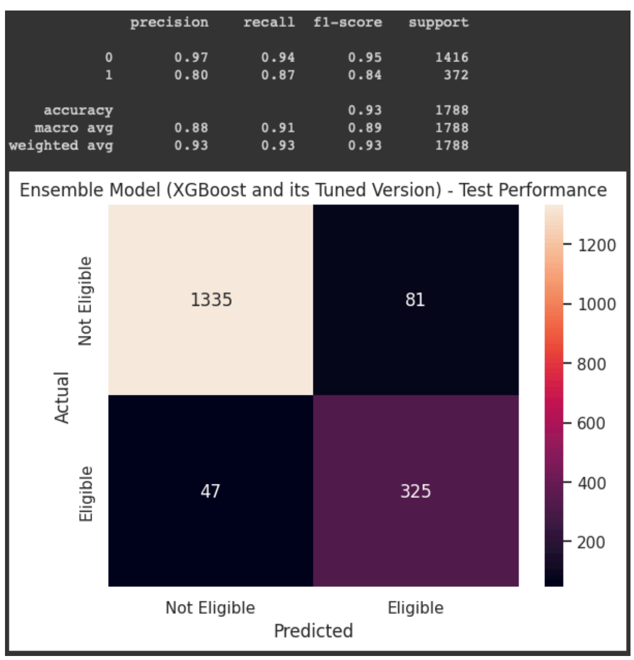
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.97 | 0.95 | 1416 |
| 1 | 0.88 | 0.75 | 0.81 | 372 |
| accuracy | | | 0.93 | 1788 |
| macro avg | 0.91 | 0.86 | 0.88 | 1788 |
| weighted avg | 0.92 | 0.93 | 0.92 | 1788 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.97 | 0.96 | 1416 |
| 1 | 0.87 | 0.81 | 0.84 | 372 |
| accuracy | | | 0.94 | 1788 |
| macro avg | 0.91 | 0.89 | 0.90 | 1788 |
| weighted avg | 0.93 | 0.94 | 0.93 | 1788 |



The **XGBoost Models (Fig 32-A)**, both the base and tuned version, demonstrated excellent performance, with the tuned model standing out with an impressive 94% accuracy on the test data. The recall of the tuned model on the test set was also high at 81%. One impressive attribute of the base version of the XGBoost model is its precision, which at 88% is the highest among all models. Precision is the ability of the model to correctly identify the positive cases out of all the predicted positive cases, which is crucial in credit risk modelling.

***FIG 32-B***



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.94 | 0.95 | 1416 |
| 1 | 0.80 | 0.87 | 0.84 | 372 |
| accuracy | | | 0.93 | 1788 |
| macro avg | 0.88 | 0.91 | 0.89 | 1788 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1788 |



The ensemble model **(Fig 32-B)** created based on both versions of the XGBoost Model has achieved impressive results with **93% accuracy** and excellent performance in both classes, particularly the minority class **(Recall 87%).** While the recall results are outstanding, we are still reviewing this to ensure a more balanced trade-off. As a result, it is not yet ready for deployment. However, a modified version of this may be deployed as our next model.

# Conclusion and Key Findings

An extensive exploratory data analysis (EDA) was conducted to identify significant correlations and trends among various loan attributes. In addition, we have gained valuable insights into predicting credit risk by exploring different machine-learning models. A few key observations emerged from this:

**Relationship Between Variables**: There is a strong positive correlation between Mortgage Due and Property Value, indicating that the property's value is a significant factor in determining the mortgage. The Loan Amount also shows a moderate positive correlation with Property Value.

**High-risk Factors**: Customers with more credit lines or higher Debt-to-Income ratios appear to pose a higher financial risk. The moderate positive correlation between Delinquencies and Derogatory Remarks also indicates a pattern of poor credit behaviour, suggesting a higher probability of loan default.

**Risk-based on Job and Reason**: In the categorical data analysis, clients who take out loans for debt consolidation have higher default rates. Also, 'Other' and 'Self' job categories have higher rates of loan defaults.

**Distribution of Key Variables**: The analysis of key variables indicates that certain values of these variables may be related to a higher probability of loan default. For instance, certain debt-to-income ratios and the number of credit lines are associated with higher instances of loan default (BAD = 1).

**Balancing Accuracy and Recall (Appendix, Fig 36, 37):** The classification problem at hand requires not just a high predictive accuracy but also a robust recall score. A model that predicts with high accuracy but fails to identify high-risk applicants (false negatives) can have serious financial implications for the business.

**Model Performance:** The XGBoost Tuned Version demonstrated the highest test accuracy of 94% and a strong recall rate of 81%, representing the best overall performance among the models tested. The model's scores were especially impressive when identifying potential loan defaults, providing a confidence level necessary for deployment in production.

**Overfitting in Random Forest Models**: While the Random Forest models showed perfect accuracy on the training data, the test data revealed potential overfitting, as demonstrated by the drop in accuracy.

**Advantage of Tuned Models:** Hyperparameter tuning improved the performance of most models, highlighting the importance of this process in model selection.

# Recommendations

In the Home Equity Lines of Credit industry, accurately predicting credit risk is crucial to financial success. A lack of precision in risk prediction could result in significant defaulted loans, leading to financial losses. Our current processes are based on traditional methods and need more sophistication to handle the complexity and scale of today's market demands.

To solve this problem, we have leveraged the power of machine learning and developed a predictive model using the XGBoost algorithm. XGBoost has been chosen for its superior performance in handling large datasets, its resilience to outliers, and its capability to model complex relationships between variables. Based on these insights, we recommend the following:

**Implementing Risk Mitigation Strategies**: Considering the high risk associated with customers with more credit lines and higher Debt-to-Income ratios, the company should plan risk mitigation strategies for these customers. This might include stricter underwriting, closer monitoring, or more strict repayment schedules.

**Deployment of the Tuned Version of the XGBoost model:** Considering accuracy and recall **(Appendix, Fig 36, 37)**, the XGBoost model provides the most balanced performance also is an excellent choice for production environments because it can handle complex and subtle datasets. Its speed, efficiency, and performance make it a top contender for deployment. Furthermore, XGBoost's robustness and ability to handle new data types and distributions could reduce costs in the long run by requiring less frequent model retraining and maintenance.

**Continued Model Tuning and Validation and Ensemble Learning (Appendix, Fig 38):** Given the dynamic nature of financial data, we recommend regular tuning and validation of the model to maintain its predictive power. Additionally, we are experimenting with **ensemble techniques** between both versions of the XGBoost Models, which has shown an outstanding **Recall Increase of 87%**; unfortunately, **in the trade-off, the precision was downgraded to 80% (Appendix, Fig 38).**

**Address Overfitting:** Overfitting can be mitigated by cross-validation, collecting more data, or feature selection. Regularly monitor models for signs of overfitting to ensure they maintain their generalization ability.

**Explore Other Models and Techniques:** The field of machine learning is vast and rapidly evolving. We should continue exploring other models or ensemble methods to improve prediction performance further.

**Continued Staff Training:** To maintain the effectiveness and relevance of our new model, we recommend ongoing training sessions for our staff. These should address any updates or improvements made to the model and ensure new team members are fully trained in its use.

# Expected Benefits and Costs

Assuming a **loan portfolio of $100M** with a high **default rate of 20%,** the **expected annual loss from defaults is $20M**. If we **reduce the default rate by 10%** through this implementation, the **savings would be $2M annually.**

The predictive model's development, integration, maintenance and associated programs are expected to cost about $200k annually. Adding to this, the cost of data acquisition and management could be $50k annually, regulatory compliance costs might be around $30k, model iteration and improvement could cost $70k, staff training might cost another $50k, and we should factor in a potential loss of $100k annually due to reduced customer satisfaction or business.

Thus, the **total annual cost would be $500k**. This means that the **net annual benefit would be $1.5M**, representing an **ROI of 300%**; from a financial standpoint, this is a valuable pursuit.

| Category | Amount |
|---|---|
| Savings from defaults | $ 2,000,000.00 |
| Development, integration, and maintenance | $ 200,000.00 |
| Data acquisition and management | $ 50,000.00 |
| Regulatory compliance | $ 30,000.00 |
| Model iteration and improvement | $ 70,000.00 |
| Staff training | $ 50,000.00 |
| Customer satisfaction impact | $ 100,000.00 |
| Total cost | $ 500,000.00 |
| **Net annual benefit** | **$ 1,500,000.00** |

# Challenges

While **XGBoost** is a powerful model known for its high performance and predictive accuracy, one of its main challenges is the interpretability of its results, which can make it difficult to justify rejections.

### Challenges:

- **The complexity of the Model:** XGBoost is a gradient-boosting model that uses numerous decision trees to make predictions. The complexity of these trees, especially when there are many, makes it difficult to understand the decision process intuitively.

- **Feature Interactions:** The model can capture high-order interactions between features, making the contribution of individual features to the final prediction more obscure.

- **Lack of Direct Coefficients:** Unlike linear models, where each variable has a clear coefficient that directly explains its impact on the response, XGBoost doesn't provide such a clear relationship.

### Overcoming the Challenges:

- **Feature Importance (Appendix, Fig 35):** One way to make XGBoost more interpretable is by using feature importance plots, which show the relative importance of each feature in making predictions giving a sense of which features are most influential in the model's decisions.

- **Partial Dependence Plots (PDPs):** PDPs offer a way to visualize the impact of certain features on the predicted outcome. They allow us to see how changes in a feature's value affect the model's predictions, holding all other features constant.

- **SHAP (Appendix, Fig 33):** SHAP values provide a suitable measure of feature importance and effects. They give an understanding of the contribution of each feature to the prediction for individual instances, thus offering local interpretability.

- **LIME (Local Interpretable Model-Agnostic Explanations) (Appendix, Fig 34):** The LIME technique can provide an understandable and accurate explanation of a classifier's predictions using a locally interpretable model approximation.

Applying these methods makes it possible to offer transparency and interpretability to the model decisions, making it easier to justify adverse behaviours. This interpretability is crucial for aligning the model with business understanding and meeting the regulatory requirements.

# Deployment Plan

In the rapidly evolving landscape of the financial sector, introducing Machine Learning (ML) models brings unprecedented opportunities for growth and efficiency. As our organization embarks on this transformative journey, it's critical to be mindful of potential challenges that may arise.

We have developed a robust rollout plan for a smooth transition, supplemented by a well-defined rollback strategy to address unforeseen issues; our two-pronged strategy aims to maximize the benefits of implementing our machine learning model while minimizing operational risk.

**Baseline Performance Metrics:** As a first step, we will establish a comprehensive understanding of our current operations. This includes recording detailed key performance indicators (KPIs) that define the success of our current loan approval process. These metrics will provide a crucial baseline against which we can measure the improvements brought on by the machine learning model.

**Phase-wise Rollout:** Rather than a complete system overhaul, we propose a gradual implementation of the ML model. The new model will Initially process only a small percentage of the total requests. As we gain confidence in the model's accuracy and stability, we will increase the proportion of requests handled by the ML model.

**Parallel Run:** To ensure continuous service and risk mitigation during the transition phase, we propose a parallel run of the current system and the new ML model. This approach allows the current system to serve as a backup, taking over the loan requests seamlessly if the ML model encounters any issues. It also provides a platform for a side-by-side comparison between the performance of the ML model and the existing system.

**Inclusion of a Manual Review System:** To boost our automated decision-making, we intend to integrate a system for manual review in instances where the model exhibits uncertainty. This condition is identified when the predicted default probabilities fall within a 'grey zone.' Initially, we propose that this grey zone be defined by a probability threshold between 0.4 and 0.6. However, this range can be adjusted based on the continuous performance assessment of the model. This strategic move will enhance our prediction accuracy, increase confidence in our risk assessment, and provide an additional layer of risk management.

**Staff Training:** To ensure a smooth transition to the new model, we must comprehensively train all our employees involved in the loan approval process. This training will cover the model's operations, outcomes, and new procedures. We'll also provide guidelines for interpreting ambiguous 'grey zone' cases to ensure consistency and efficiency in our loan approval process under the new system.

**Automatic Rollback:** A resilient system needs robust contingency plans. We will develop a mechanism that enables an automatic switch to the old method if it detects significant anomalies in the ML model's responses. Specific conditions, such as a sudden drop in key performance indicators or an unexplained increase in processing time, could trigger this.

**Continuous Monitoring and Alerting System:** As a part of our proactive approach to managing potential issues, we will implement an efficient monitoring system. This system will continuously assess the performance metrics and operational parameters. Any deviation or anomaly in model performance or system metrics will trigger alerts, enabling swift action. If a rollback is required, this system will ensure we can respond quickly and effectively, thereby minimizing any potential disruption or negative impact on our services.

By taking these steps, we aim to provide a successful transition to a more efficient credit risk assessment process. Our plan encapsulates the balance between embracing innovation and maintaining operational stability, setting us on a path for a brighter and more profitable business model.

# Appendix

*FIG 33*

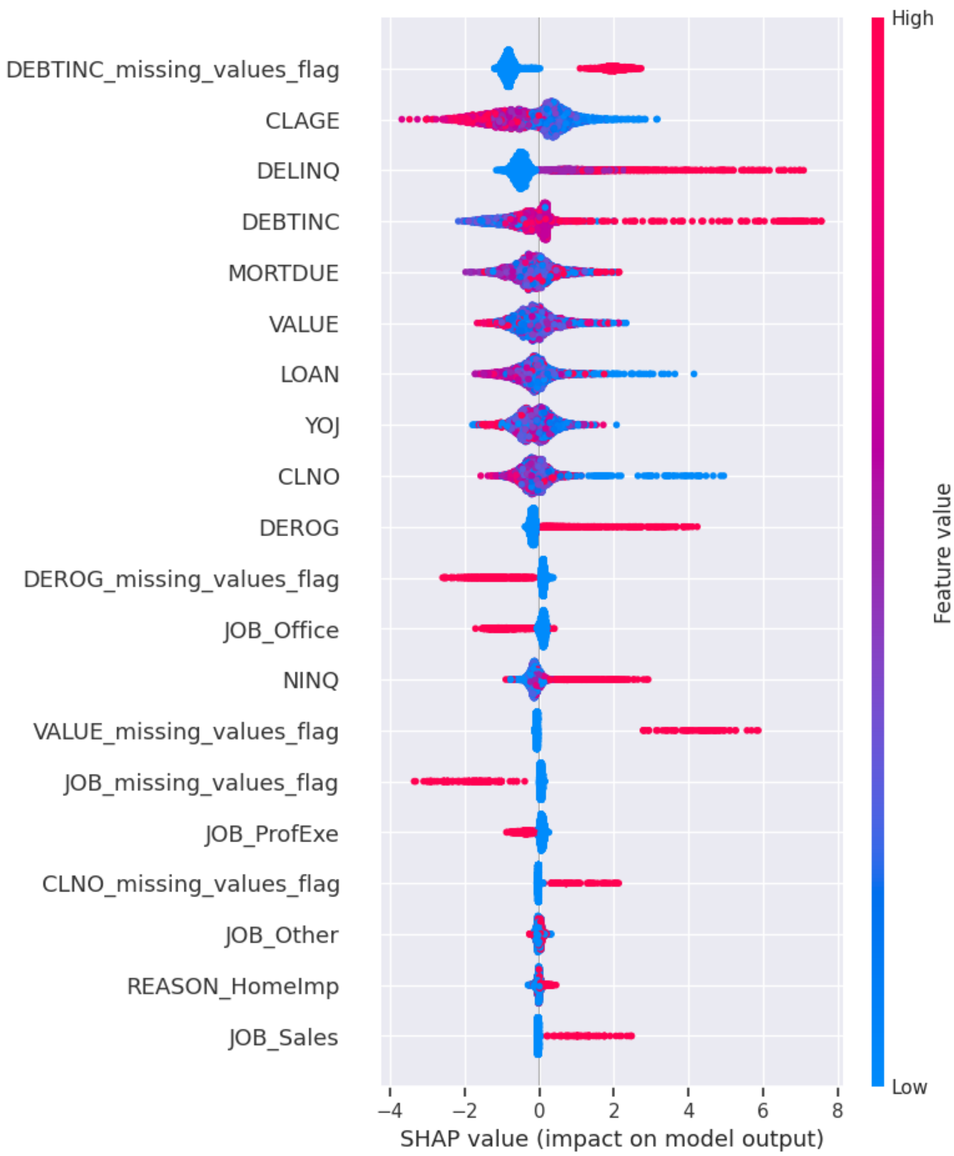



*FIG 34*


| Feature | Value |
|---|---|
| VALUE_missing_values_flag | 0.00 |
| DELINQ | 0.00 |
| DEBTINC_missing_values_flag | 0.00 |
| DEROG | 0.00 |
| DEBTINC | 34.34 |

***FIG 35***

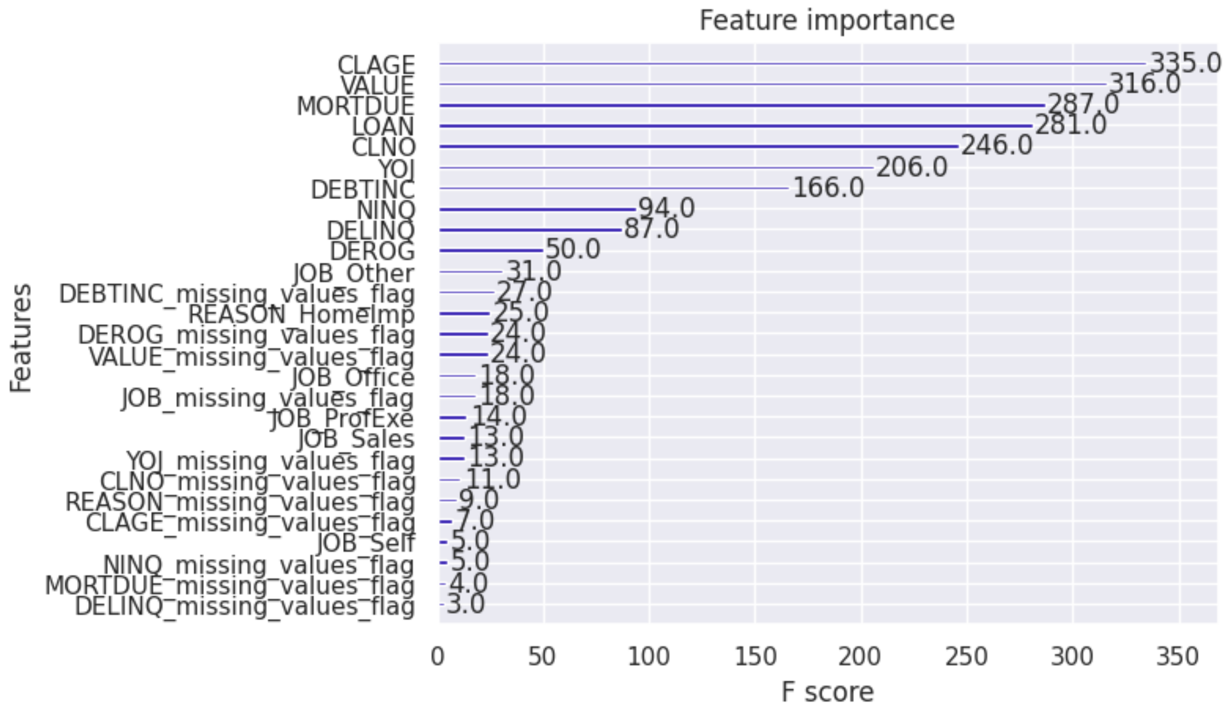

Feature importance
CLAGE
VALUE
MORTDUE
LOAN
CLNO
YOJ
DEBTINC
NINQ
DELINQ
DEROG
JOB_Other
DEBTINC_missing_values_flag
REASON_HomeImp
DEROG_missing_values_flag
VALUE_missing_values_flag
JOB_Office
JOB_missing_values_flag
JOB_ProfExe
JOB_Sales
YOJ_missing_values_flag
CLNO_missing_values_flag
REASON_missing_values_flag
CLAGE_missing_values_flag
JOB_Self
NINQ_missing_values_flag
MORTDUE_missing_values_flag
DELINQ_missing_values_flag
335.0
316.0
287.0
281.0
246.0
206.0
166.0
94.0
87.0
50.0
31.0
27.0
25.0
24.0
24.0
18.0
18.0
14.0
13.0
13.0
11.0
9.0
7.0
5.0
5.0
4.0
3.0
Features
0
50
100
150
200
250
300
350
F score

***FIG 36***

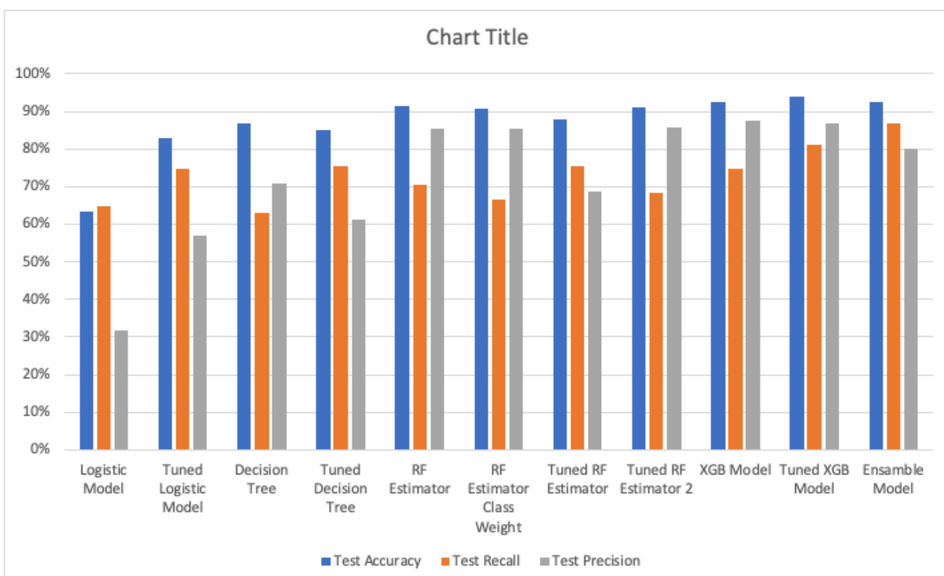


Chart Title
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Logistic Model
Tuned Logistic Model
Decision Tree
Tuned Decision Tree
RF Estimator
RF Estimator Class Weight
Tuned RF Estimator
Tuned RF Estimator 2
XGB Model
Tuned XGB Model
Ensamble Model
Test Accuracy
Test Recall
Test Precision

## *FIG 37*

| Model | Test Accuracy | Test Recall | Test Precision |
|---|---|---|---|
| Logistic Model | 63% | 65% | 32% |
| Tuned Logistic Model | 83% | 75% | 57% |
| Decision Tree | 87% | 63% | 71% |
| Tuned Decision Tree | 85% | 76% | 61% |
| RF Estimator | 91% | 70% | 85% |
| RF Estimator Class Weight | 91% | 67% | 86% |
| Tuned RF Estimator | 88% | 76% | 69% |
| Tuned RF Estimator 2 | 91% | 68% | 86% |
| XGB Model | 93% | 75% | 88% |
| Tuned XGB Model | 94% | 81% | 87% |
| Ensamble Model | 93% | 87% | 80% |

## *FIG 38*

```python
# Define and train the base model
xgb_model_class_0 = XGBClassifier()
xgb_model_class_0.fit(X_train, y_train)

# Define and train the Tuned model
xgb_model_class_1 = XGBClassifier(scale_pos_weight=4)
xgb_model_class_1.fit(X_train, y_train)

# Predict probabilities with both models
class_0_pred_prob = xgb_model_class_0.predict_proba(X_test)
class_1_pred_prob = xgb_model_class_1.predict_proba(X_test)

# Set a rule to make final predictions
threshold = 0.3  # you can adjust this threshold as necessary
final_pred = np.where(class_1_pred_prob[:,1] > threshold, 1, class_0_pred_prob.argmax(axis=1))


# Print the classification report
print(classification_report(y_test, final_pred))

# Compute the confusion matrix
cf_matrix = confusion_matrix(y_test, final_pred)

# Create a heatmap
sns.heatmap(cf_matrix, annot=True, fmt='d', xticklabels=['Not Eligible', 'Eligible'], yticklabels=['Not Eligible', 'Eligible'])

plt.title('Ensemble Model (XGBoost and its Tuned Version) - Test Performance')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

```
              precision    recall  f1-score   support

           0       0.97      0.94      0.95      1416
           1       0.80      0.87      0.84       372

    accuracy                           0.93      1788
   macro avg       0.88      0.91      0.89      1788
weighted avg       0.93      0.93      0.93      1788
```