Machine Learning Model for Loan

Eligibility Prediction

**Abstract**  
Loan approval processes play a crucial role in financial institutions, ensuring efficient credit allocation while minimizing risks. This project focuses on developing a robust machine learning model to predict loan approvals, leveraging applicant profiles, credit history, and loan-related attributes. The dataset includes comprehensive training and testing data, with features such as income, employment length, home ownership, loan intent, and loan percentage of income.

The objective is to build and evaluate multiple classification models, including Logistic Regression, Perceptron, Decision Tree, Random Forest, K-Nearest Neighbors, and XGBoost, to determine the most accurate and reliable approach. Exploratory Data Analysis (EDA) was conducted to preprocess and understand data trends, while the models were assessed using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

This report provides a systematic approach to solving a real-world business problem, emphasizing the importance of feature engineering, model performance, and actionable insights for loan approval optimization. The results aim to streamline decision-making for financial firms, ensuring both customer satisfaction and risk mitigation.

**Introduction**

Loan approval is a critical process in financial institutions, directly impacting their profitability and customer satisfaction. Traditional methods of evaluating loan applications often rely on manual review processes, which can be inconsistent, time-consuming, and prone to human bias. These challenges create a need for an automated, accurate, and efficient solution that can assist in making reliable decisions while minimizing risks.

The primary objective of this project is to develop a machine learning model that predicts loan approvals by analyzing applicant profiles and credit history data. By leveraging a robust dataset from Kaggle, the project aims to identify key factors influencing loan approval decisions and evaluate the performance of multiple machine learning algorithms, including Logistic Regression, Perceptron, Decision Tree, Random Forest, K-Nearest Neighbors, and XGBoost.

This initiative not only addresses the operational inefficiencies in loan processing but also provides actionable insights for financial institutions to optimize their decision-making processes. The results are expected to reduce default rates, enhance approval accuracy, and improve the overall efficiency of credit allocation.

The following sections of this report present a comprehensive analysis of the dataset, exploratory data analysis (EDA), methodology, results, and conclusions drawn from the model evaluation and implementation.

**Data Description**

The dataset used in this project, sourced from Kaggle’s "Loan Approval Prediction" competition, is specifically designed to simulate real-world loan approval scenarios. The training dataset includes the target variable, "loan\_status," which indicates whether a loan was approved (1) or rejected (0). The dataset contains 58,645 entries with 13 features, including the target variable. Each feature is described below:

**Key Features**

* **person\_age**: Age of the applicant.
* **person\_income**: Annual income of the applicant.
* **person\_home\_ownership**: Type of home ownership (e.g., RENT, OWN, MORTGAGE).
* **person\_emp\_length**: Employment length in years.
* **loan\_intent**: Purpose of the loan (e.g., EDUCATION, MEDICAL, PERSONAL).
* **loan\_grade**: Grade assigned to the loan based on creditworthiness.
* **loan\_amnt**: Loan amount requested.
* **loan\_int\_rate**: Interest rate assigned to the loan.
* **loan\_percent\_income**: Loan amount as a percentage of the applicant’s income.
* **cb\_person\_default\_on\_file**: Indicates if the applicant has previously defaulted on a credit line.
* **cb\_person\_cred\_hist\_length**: Length of the applicant’s credit history in years.
* **loan\_status**: Target variable indicating loan approval (1) or rejection (0).

**Dataset Insights**

* The training dataset contains **58,645 rows** and **13 columns**, including the target variable.
* There are no missing values in the dataset, ensuring a clean starting point for analysis.
* The distribution of the target variable is as follows:

**Approved Loans (1)**: Approximately 48%.

**Rejected Loans (0)**: Approximately 52%.

This dataset was carefully preprocessed to address encoding for categorical features, scale numerical variables, and ensure that it is suitable for building predictive machine learning models.

**Citation**: Walter Reade and Ashley Chow. Loan Approval Prediction, Kaggle Competition, 2024. [1] <https://kaggle.com/competitions/playground-series-s4e10>

**Exploratory Data Analysis**

**Data Quality and Preprocessing**

The training dataset underwent an initial quality assessment to ensure readiness for analysis. Key findings from this step include:

A close-up of a document

Description automatically generated

Fig1:Dataset Structure

The dataset contains **13 columns** and **58,645 rows**, with no missing values in any feature. Numerical features such as person\_income and loan\_amnt are stored as integers or floats, while categorical features such as loan\_intent and person\_home\_ownership are stored as objects.

The dataset structure highlights a balanced mix of numerical and categorical variables, suitable for machine learning analysis. - The dataset contains 13 columns and 58,645 rows, with no missing values in any feature.

Data types were confirmed as appropriate: numerical features (e.g., person\_income, loan\_amnt) were stored as integers or floats, while categorical features (e.g., loan\_intent, person\_home\_ownership) were stored as objects.

A screenshot of a calculator

Description automatically generated

Fig 2: Summary Statistics

Descriptive statistics revealed the following insights for key numerical features:

* The median annual income of applicants is approximately $55,000, with a wide range spanning from $20,000 to over $200,000.
* Loan amounts range between $1,000 and $35,000, with a significant clustering around $10,000.
* Most applicants dedicate between 10% and 30% of their income to loans, though some outliers exceed 50%.
* person\_income: Median annual income is approximately $55,000, with a range from $20,000 to over $200,000.
* loan\_amnt: Loan amounts range from $1,000 to $35,000, with most loans clustering around $10,000.
* loan\_percent\_income: Loans typically account for 10-30% of the applicant’s income, with a few outliers exceeding 50%.

These insights provide a foundational understanding of feature distributions and potential relationships.

A screenshot of a computer

Description automatically generated

Fig 3: Missing Values

The dataset was thoroughly examined, and no missing values were detected across all features. This ensures that the dataset is clean and ready for analysis.

A table with numbers and text

Description automatically generated

Fig 4: Unique Values

The dataset comprises features with varying levels of uniqueness:

* **Categorical features** like loan\_intent have 6 unique values, including "EDUCATION" and "MEDICAL."
* **Numerical features** like loan\_amnt have 545 unique values, ranging from $1,000 to $35,000.
* Other features such as cb\_person\_default\_on\_file have only two unique values, "Y" and "N," indicating past defaults.
* [Insert snippet of unique value counts here for further details.] - **person\_income**: Contains 2,641 unique values, reflecting diverse applicant income levels.
* **loan\_amnt**: Includes 545 unique values, indicating varied loan requests.
* **loan\_grade**: Has 7 distinct grades (A to G) representing creditworthiness.
* **loan\_intent**: Comprises 6 unique purposes such as EDUCATION, MEDICAL, and PERSONAL.
* This diversity highlights the dataset's richness and the potential for detailed analysis across multiple dimensions.

This preprocessing ensured the dataset was suitable for the detailed analysis and modeling steps that followed.

A screenshot of a computer screen

Description automatically generated

Fig 5:Correlation Heatmap Between Variables

The correlation heatmap is a visual representation of relationships between key features in the dataset. It reveals several important insights that guide the analysis and modeling steps:

* **loan\_amnt and loan\_percent\_income**: There is a moderate positive correlation, indicating that larger loan amounts tend to account for a higher percentage of an applicant's income. This suggests that loan amount and income are key factors influencing each other.
* **person\_age and cb\_person\_cred\_hist\_length**: A strong positive correlation (close to 0.87) reflects that older individuals typically have longer credit histories. This aligns with expectations, as age provides more time for financial activity.
* **loan\_status**: Weak correlations with most features suggest that no single variable dominantly influences loan approval, highlighting the need for a multivariate analysis to understand combined effects.

This heatmap serves as a foundation to identify which features to prioritize in feature engineering and modeling strategies.

A graph of a loan status

Description automatically generated

Fig 6:Distribution of Target Variable loan\_status

The target variable, loan\_status, plays a pivotal role in this analysis. The bar chart of loan\_status distribution uncovers the following:

* **Rejected Loans (0)**: Comprise the majority, with approximately 85.76% of applications falling into this category. This indicates that loan rejections are significantly higher, reflecting strict approval criteria.
* **Approved Loans (1)**: Represent a smaller fraction, approximately 14.24%, which points to a selective process for approvals.

This imbalance underscores the importance of implementing strategies like oversampling, undersampling, or algorithm-specific approaches (e.g., class weights) to mitigate biases in predictive modeling. It also raises questions about the attributes most correlated with approval rates, which are explored further in subsequent sections.

The distribution of the target variable, loan\_status, reveals key insights into the dataset:

* **Rejected Loans (0)**: The majority of applications fall under this category, comprising approximately 85.76% of the dataset.
* **Approved Loans (1)**: A smaller proportion, about 14.24%, represents approved loans.

This imbalance indicates the need for strategies like oversampling, undersampling, or advanced machine learning techniques to handle class imbalance during modeling effectively. The chart also suggests that identifying patterns leading to loan approval will be crucial for predictive modeling.

A group of different colored bars

Description automatically generated

Fig 7:Distribution of Categorical Features

Analyzing the distribution of categorical features provides valuable insights into borrower characteristics and loan details. Key findings from the visualizations include:

* **person\_home\_ownership**:

Most applicants are renters (30,594), followed by those with mortgages (24,824). Only a small fraction own their homes outright (3,138), with minimal representation in the "OTHER" category (89).

* **loan\_intent**:

Education-related loans dominate (12,271), reflecting a significant portion of applications aimed at funding educational needs. Other common intents include medical expenses (10,934) and personal loans (10,016), while home improvement loans are the least frequent (6,280).

* **loan\_grade**:

Loan grades A and B are the most common, with 20,984 and 20,400 loans, respectively. Higher grades (E, F, G) are rare, emphasizing that most borrowers have relatively good creditworthiness.

* **cb\_person\_default\_on\_file**:

The majority of applicants (49,943) have no prior defaults, while only 8,702 have a history of defaulting, indicating that the dataset is skewed towards more reliable borrowers.

These distributions highlight the diversity within the dataset and provide a foundation for understanding the factors influencing loan approvals and rejections.

A colorful squares with black text

Description automatically generated

Fig 8:Correlation Heatmap Between Numerical Features

A focused analysis of numerical features reveals significant relationships and trends:

* **person\_age and cb\_person\_cred\_hist\_length**:

A strong positive correlation (0.87) confirms that older individuals tend to have longer credit histories, which aligns with expectations given that age provides more opportunities for financial activity.

* **loan\_amnt and loan\_percent\_income**:

A correlation of 0.65 indicates that higher loan amounts are associated with a larger share of the applicant's income, reflecting the financial burden such loans may represent.

* **loan\_int\_rate and loan\_status**:

Moderate positive correlation (0.34) suggests that loans with higher interest rates are more likely to be approved, possibly due to their higher risk profile and profitability.

These correlations provide a quantitative basis for understanding how numerical features interact and influence loan outcomes.

A comparison of a graph

Description automatically generated

Fig 9: Analyzing Credit History and Default Status vs Loan Approval

Exploring the relationship between credit history, default status, and loan approval reveals significant patterns:

* **Credit History Length**:

Borrowers with longer credit histories tend to have higher loan approval rates. The boxplot shows that approved loans (loan\_status = 1) are associated with a broader range of credit history lengths compared to rejected loans (loan\_status = 0).

Applicants with shorter credit histories face higher rejection rates, indicating that credit history is a key determinant in loan approval decisions.

* **Default Status**:

Borrowers with a history of default (cb\_person\_default\_on\_file = "Y") are less likely to have loans approved. Among those with previous defaults, only 2,601 loans were approved compared to 6,101 rejected loans.

On the other hand, applicants with no history of default (cb\_person\_default\_on\_file = "N") have significantly higher approval rates, with 44,194 loans approved compared to 5,749 rejected loans.

These trends emphasize the importance of both credit history length and prior default behavior in assessing loan applications. Financial institutions can leverage these insights to refine their risk assessment models.

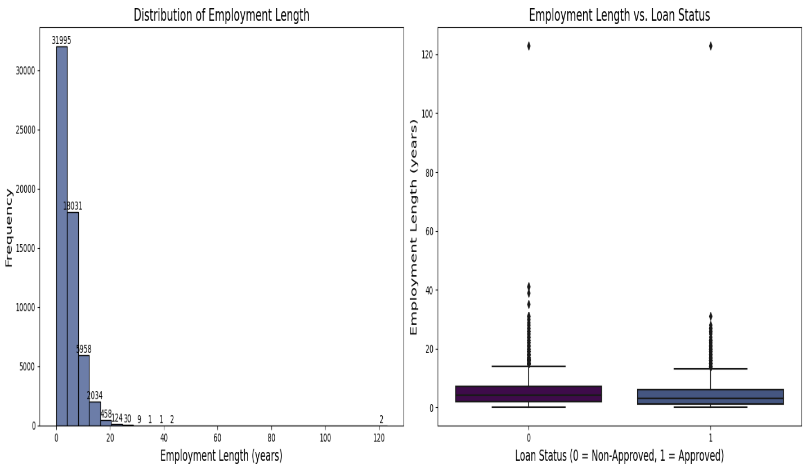


Fig 10: Analyzing Employment Length and Loan Status

Examining the relationship between employment length and loan status uncovers valuable trends:

* **Distribution of Employment Length**:

Most applicants have employment lengths of less than 10 years, with a sharp decline in frequency as employment length increases.

Most borrowers fall within the 0–5 years category, indicating a predominance of individuals early in their careers or with limited employment history.

* **Employment Length and Loan Approval**:

Applicants with longer employment histories tend to have higher loan approval rates. This pattern suggests that employment stability positively influences approval decisions.

The boxplot illustrates that the median employment length for approved loans (loan\_status = 1) is slightly higher compared to rejected loans (loan\_status = 0).

These insights emphasize the importance of employment length as a key feature in loan approval decisions, reflecting the lender's preference for financially stable and reliable applicants.

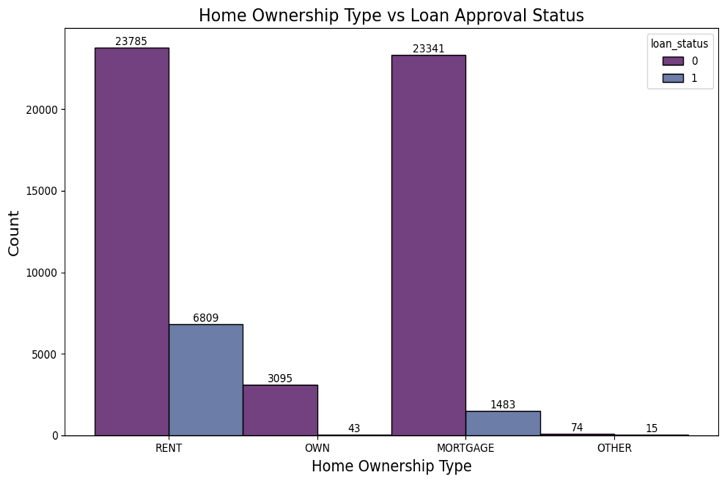


Fig 11: Home Ownership vs Loan Approval Status

The relationship between home ownership types and loan approval rates reveals key insights:

* **Home Ownership Distribution**:

The majority of applicants are renters (23,785 rejected and 6,809 approved), followed closely by those with mortgages (23,341 rejected and 1,483 approved).

Applicants who own their homes outright represent a smaller segment, with 3,095 rejections and 43 approvals.

The "OTHER" category, which is minimally represented, includes 74 rejections and 15 approvals.

* **Loan Approval Trends**:

Renters have a higher approval rate compared to applicants with mortgages, possibly reflecting different financial dynamics or risk perceptions associated with these groups.

Homeowners with outright ownership (OWN) demonstrate the lowest approval rates, possibly due to their limited representation in the dataset or other underlying factors.

This analysis highlights how home ownership type influences loan approval outcomes and provides a basis for incorporating this feature into predictive modeling strategies.

A graph of two people

Description automatically generated

Fig 12: Loan Amount Distribution by Loan Status

Examining the distribution of loan amounts by loan status reveals interesting insights into lending trends:

* **Loan Amount Distribution**:

For rejected loans (loan\_status = 0), loan amounts show a broad range, with peaks around $5,000 and $10,000. This indicates a higher frequency of loan requests in these ranges among rejected applications.

Approved loans (loan\_status = 1), on the other hand, exhibit a narrower range of amounts, with a more consistent distribution around $10,000. This suggests that loan amounts closer to this value have a higher likelihood of approval.

* **Key Observations**:

The density plot highlights that while larger loan amounts (> $20,000) are less frequent, they are slightly more likely to be approved than smaller loan amounts below $5,000.

The differences in distributions suggest that lenders may favor specific ranges of loan amounts when assessing applications, potentially due to associated risk factors or repayment capabilities.

These findings underscore the importance of loan amount as a predictive feature, providing valuable input for modeling and decision-making processes.

A graph of a home ownership status

Description automatically generated

Fig 13: Home Ownership Status vs Loan Approval

Understanding the relationship between home ownership status and loan approval provides key insights:

* **Distribution by Home Ownership**:

Renters represent the largest group of applicants, with 23,785 rejected and 6,809 approved loans.

Applicants with mortgages are the second-largest group, with 23,341 rejections and 1,483 approvals.

Those who own their homes outright (OWN) have 3,095 rejections and only 43 approvals.

The "OTHER" category has minimal representation with 74 rejections and 15 approvals.

* **Loan Approval Insights**:

Renters have higher loan approval rates compared to those with mortgages, which could be due to different financial dynamics or risk profiles.

Homeowners with outright ownership show the lowest loan approval rates, possibly due to limited representation or specific financial attributes not captured in the dataset.

This analysis highlights the significant role of home ownership status in loan approval decisions and suggests its importance as a feature in predictive modeling.

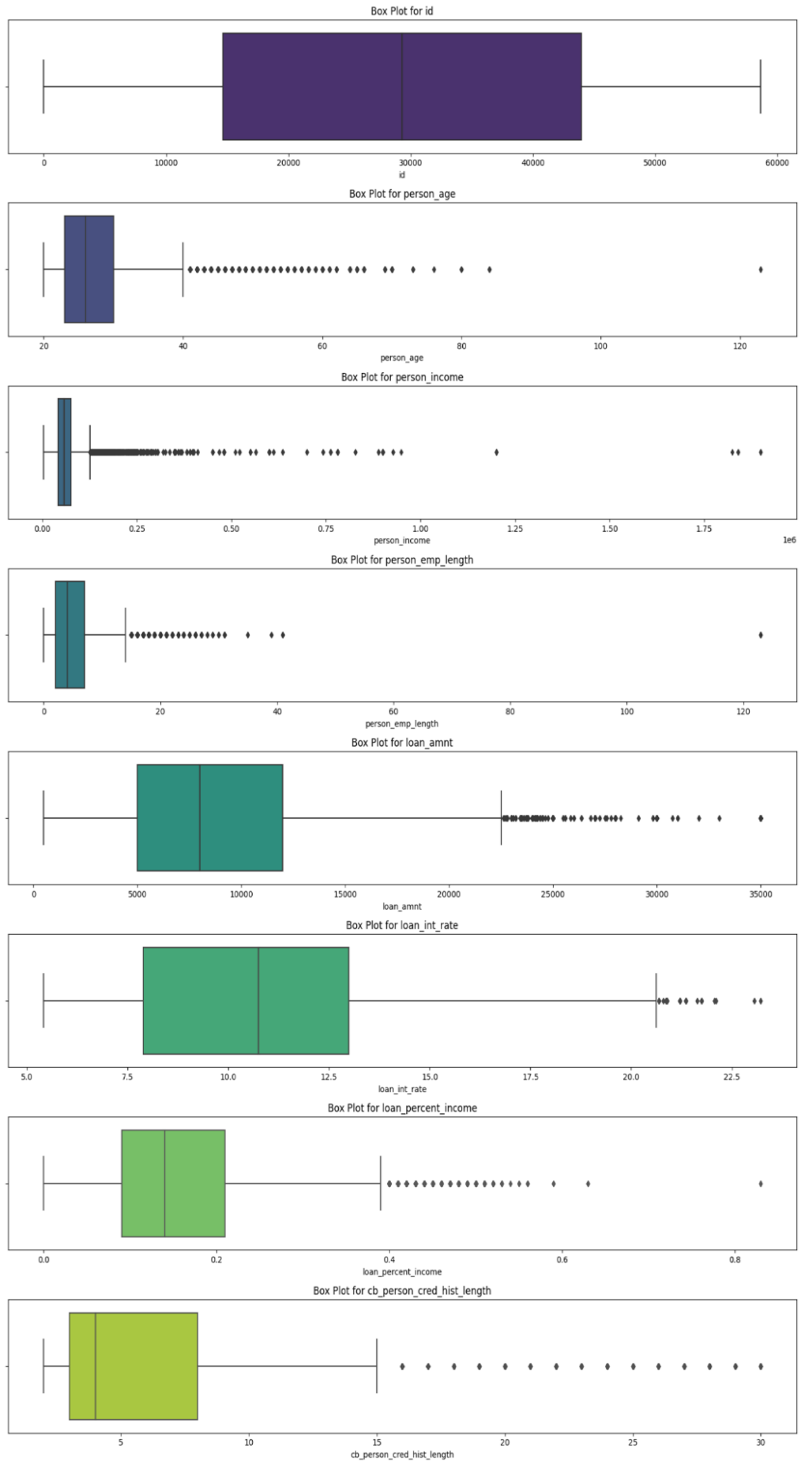


Fig 14: Box Plots for Numerical Features

Box plots for numerical features provide a clear understanding of the distribution, central tendencies, and outliers in the dataset. Here’s a detailed look:

* **person\_age**:

The box plot reveals that the majority of applicants are within the age range of 25 to 50 years, with a few outliers exceeding 100 years. These outliers may represent data entry errors or special cases.

* **person\_income**:

Applicant incomes show significant variability, with most values clustering below $100,000. Outliers exist in higher ranges, likely representing high-income individuals.

* **person\_emp\_length**:

Employment length typically spans 0 to 15 years, with a few outliers showing employment histories over 50 years. This highlights the diversity in employment backgrounds among applicants.

* **loan\_amnt**:

Loan amounts mostly range from $5,000 to $20,000, with fewer loans exceeding $25,000. Outliers in higher ranges are also evident, representing larger loan requests.

* **loan\_int\_rate**:

Interest rates predominantly range from 5% to 15%, with some outliers above 20%. These higher rates may be associated with riskier loan profiles.

* **loan\_percent\_income**:

Loan percentages typically fall within 10% to 30% of the applicant’s income, with a few outliers exceeding 50%. This indicates varying levels of financial burden among borrowers.

* **cb\_person\_cred\_hist\_length**:

Credit history length ranges between 0 and 30 years, with most values clustering below 15 years. Outliers with longer credit histories highlight applicants with extensive financial activity.

These box plots provide valuable insights into the dataset's numerical features, identifying trends and outliers that are crucial for both exploratory analysis and feature engineering.

**Feature Engineering**

New data points were created from the existing information to help our analysis. Think of it like this: looking for hidden clues within the data that can give us a clearer picture of what's going on.

Here's what was did:

Calculated "Debt-to-Earnings Ratio": This shows how much of a person's income would be used to pay back the loan. A high ratio might mean the person is borrowing more than they can comfortably afford.

Formula: Debt-to-Earnings Ratio = (loan\_amnt / person\_income) \* 100

Estimated "Financial Burden": This gives us an idea of the overall cost of the loan, including the interest rate. It helps us understand how much the loan will actually cost the borrower over time.

Formula: Financial Burden = loan\_amnt \* (1 + (loan\_int\_rate / 100))

Determined "Credit History to Age Ratio": This compares how long someone has had credit with their age. It might indicate how responsible they've been with credit throughout their life.

Formula: Credit History to Age Ratio = (cb\_person\_cred\_hist\_length / person\_age) \* 100

Calculated "Loan Interest to Credit History Ratio": This shows the relationship between the interest rate and the length of the person's credit history. It might suggest how risky the loan is for the lender.

Formula: Loan Interest to Credit History Ratio = loan\_int\_rate / cb\_person\_cred\_hist\_length

By creating these new data points, essentially enriching the understanding of the loan applicants. This will help us build a more accurate and reliable model for predicting loan approvals.

**Machine Learning Algorithms**

To address the loan default prediction problem, employing a variety of machine learning algorithms:

* **Logistic Regression:**

**A diagram of a logistic regression

Description automatically generated**

Fig 15: Logistic Regresssion Evaluation Metric Graphs

Accuracy: 0.913742, ROC-AUC: 0.891765

**Model Performance**: The model exhibits a reasonably good performance, as indicated by the AUC-ROC score of 0.89.

**Class Imbalance:** The confusion matrix suggests a potential class imbalance, with more positive class instances.

**Model Bias:** The model might be biased towards the positive class, as evident from the higher number of correct predictions for the positive class.

Overall, the Logistic Regression model provides a solid foundation for classification tasks.

* **Perceptron:**

A graph with a line showing the results of a performance

Description automatically generated with medium confidence

Fig 16: Perceptron Evaluation Metric Graphs

Accuracy: 0.882341

**Confusion Matrix:**

* **True Positive (TP):** 12434 - The model correctly predicted the positive class (1).
* **True Negative (TN):** 1233 - The model correctly predicted the negative class (0).
* **False Positive (FP):** 305 - The model incorrectly predicted the positive class when the actual class was negative.
* **False Negative (FN):** 647 - The model incorrectly predicted the negative class when the actual class was positive.

**Precision-Recall Curve:**

* **Precision:** The proportion of predicted positive cases that are positive. The precision value for Class 0 is 0.91 and class 1 is 0.68.
* **Recall:** The proportion of actual positive cases that are correctly identified by the model. The Recall value for Class 0 is 0.98 and class 1 is 0.34.

The Perceptron model has a decent accuracy of 0.88 but is biased towards the negative class. It struggles to correctly identify positive cases, as shown by the lower recall.

* **K-Nearest Neighbors (KNN):**

**A diagram of a graph

Description automatically generated with medium confidence**

Fig 17: KNN Evaluation Metric Graphs

Accuracy: 0.928518, ROC-AUC: 0.870151

**Confusion Matrix**

* **High accuracy:** Correctly predicted most instances.
* **Class imbalance:** More accurate in predicting the positive class.

**ROC Curve**

* **Decent performance:** AUC of 0.87 indicates reasonable discrimination.

**Gains Chart and Lift Chart**

* **Better than random:** The model consistently outperforms a random model in capturing positive cases.

Overall, The KNN model shows promising results, but it could benefit from addressing class imbalance and potentially optimizing the number of neighbors and distance metric.

* **Decision Tree:**

**A graph of a tree

Description automatically generated with medium confidence**

Fig 18: Decision Tree Evaluation Metric Graphs

Accuracy: 0.946782, ROC-AUC: 0.898265

**Confusion Matrix**

* **High Accuracy:** The model correctly predicted a significant number of instances.
* **Class Imbalance:** The model seems to be better at predicting the positive class than the negative class.

**ROC Curve**

* **Good Performance:** The AUC-ROC score of 0.898265 indicates decent performance in distinguishing between positive and negative classes.

**Gains Chart and Lift Chart**

* **Better than Random:** The model consistently outperforms a random model in capturing positive cases.

Overall,The Decision Tree model demonstrates strong performance. However, there's potential for improvement, particularly in addressing class imbalance and optimizing hyperparameters like tree depth and minimum sample split.

* **Random Forest:**

A group of graphs showing different types of data

Description automatically generated with medium confidence

Fig 19: Random Forest Evaluation Metric Graphs

Accuracy: 0.951023, ROC-AUC: 0.926697

**Confusion Matrix**

* **High Accuracy:** The model correctly predicted a significant number of instances.
* **Class Imbalance:** The model seems to be better at predicting the positive class than the negative class.

**ROC Curve**

* **Strong Performance:** The AUC-ROC score of 0.926697 indicates a strong performance in distinguishing between positive and negative classes.

**Gains Chart and Lift Chart**

* **Superior Performance:** The model significantly outperforms a random model in capturing positive cases.

Overall, The Random Forest model demonstrates excellent performance, especially in terms of accuracy and ROC-AUC. It effectively captures complex patterns and is less prone to overfitting compared to simpler models like Decision Trees.

* **XGBoost:**

A diagram of a graph

Description automatically generated with medium confidence

Fig 20: XGBoost Evaluation Metric Graphs

Accuracy: 0.952391, ROC-AUC: 0.951767

**Confusion Matrix**

* **High Accuracy:** The model correctly predicted a significant number of instances.
* **Class Imbalance:** The model seems to be better at predicting the positive class than the negative class.

**ROC Curve**

* **Excellent Performance:** The AUC-ROC score of 0.951767 indicates an exceptional performance in distinguishing between positive and negative classes.

**Gains Chart and Lift Chart**

* **Superior Performance:** The model significantly outperforms a random model in capturing positive cases.

Overall, XGBoost demonstrates outstanding performance, making it a strong choice for this dataset. It excels in capturing complex patterns and provides robust predictions.

**Training Process and Evaluation**

* **Training Process** : The training process for each machine learning model involved the following steps:

**Data Preprocessing**

* **Data Cleaning:** The dataset was cleaned to handle missing values and outliers.
* **Feature Engineering:** Categorical features were encoded using one-hot encoding, and numerical features were scaled to ensure they have a similar range.
* **Feature Selection:** Relevant features were selected to improve model performance and reduce overfitting.

**Model Training:**

**Hyperparameter Tuning**

* Each model was trained on the preprocessed dataset using appropriate hyperparameters.
* Hyperparameter tuning was performed using techniques like GridSearchCV to find the optimal configuration for each model.

**Model Evaluation:**

The trained models were evaluated using a combination of the following metrics:

* Accuracy: The proportion of correctly classified instances.
* Precision: The proportion of positive predictions that are actually positive.
* Recall: The proportion of positive instances that are correctly identified.
* F1-score: The harmonic mean of precision and recall.

**Confusion Matrix:** A table that summarizes the performance of a classification model on a set of test data.

**ROC Curve**: A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is 1 varied.

**Validation Techniques**

To assess the generalization performance of the models and prevent overfitting, employed the following validation techniques:

**K-Fold Cross-Validation:**

* The dataset was divided into K folds.
* The model was trained on K-1 folds and evaluated on the remaining fold.
* This process was repeated K times, and the average performance metric was calculated.

**Train-Test Split:**

* The dataset was divided into a training set and a testing set.
* The model was trained on the training set and evaluated on the testing set.
* By combining these techniques, was able to obtain a reliable estimate of the model's performance on unseen data.

**Conclusion**

This project successfully built a machine learning model to predict loan approvals, helping automate and improve decision-making for financial institutions. By analyzing applicant profiles, credit histories, and loan details, identified key factors influencing loan outcomes.

Our analysis revealed important trends, including the impact of credit history, employment length, and home ownership on loan approval rates. Also engineered new features to better understand applicants' financial situations.

Through rigorous testing of various algorithms, found that tree-based ensemble models, like Random Forest and XGBoost, provided the most accurate predictions. This project demonstrates the potential of machine learning to optimize loan approval processes, reduce risk, and improve customer satisfaction.

**Reference:**

* <https://kaggle.com/competitions/playground-series-s4e10>
* <https://www.experian.com/blogs/ask-experian/credit-education/improving-credit/building-credit/>
* <https://www.consumerfinance.gov/ask-cfpb/my-mortgage-lender-told-me-it-was-exempt-from-the-ability-to-repay-mortgage-rule-is-this-true-en-1793/>