**Personal Bank Loan Prediction ML Report**

Mukta Pathak and Athira John

School of Business, St Lawrence College

ADMN5016: Applied Artificial Intelligence and Machine Learning

Prof. Sujoy Paul

December 15th, 2023

**Personal Bank Loan Prediction Model Report**

**Introduction:**

In the ever-evolving landscape of financial services, the ability to make informed and precise decisions is paramount. Financial institutions continually seek methods to optimize their services, and one crucial aspect is the prudent evaluation of customers' eligibility for various financial products. This analysis centers around predicting the likelihood of a customer accepting a personal loan.

**Business Problem:**

The primary challenge faced by our financial institution is the need to identify customers who are more likely to accept a personal loan offer. This predictive capability is essential for resource allocation, risk assessment, and targeted marketing strategies. By accurately determining which customers are predisposed to accept personal loans, the institution can tailor its approach, leading to more effective marketing campaigns and improved customer satisfaction.

**Dataset Overview:**

The dataset at our disposal contains several variables that provide valuable insights into customers' attributes and behaviors. Here is an overview of the key variables in the dataset:

* ID: A unique identifier for each customer.
* Age: The age of the customer.
* Experience: The number of years of professional experience.
* Income: The customer's annual income.
* ZIP Code: The ZIP code of the customer's residence.
* Family: The size of the customer's family.
* CCAvg: The average spending on credit cards per month.
* Education: The customer's level of education.
* Mortgage: The amount of mortgage owed by the customer.
* Personal Loan: Binary variable indicating whether the customer accepted a personal loan (1) or not (0).
* Securities Account: Binary variable indicating whether the customer has a securities account (1) or not (0).
* CD Account: Binary variable indicating whether the customer has a certificate of deposit account (1) or not (0).
* Online: Binary variable indicating whether the customer uses online banking services (1) or not (0).
* Credit Card: Binary variable indicating whether the customer has a credit card (1) or not (0).

This dataset serves as the foundation for developing a predictive model that can assist in anticipating customers' responses to personal loan offers. By leveraging machine learning techniques, we aim to enhance the institution's decision-making process, enabling a more targeted and efficient approach to personal loan marketing.

**Exploratory Data Analysis (EDA):**

**Class Distribution:** The distribution of the target variable, 'Personal Loan,' is a critical aspect of understanding the dataset. The class distribution is as follows:

Class 0 (No Personal Loan): 4520 instances

Class 1 (Personal Loan): 480 instances

A clear class imbalance is evident, with a significantly higher number of instances in the "No Personal Loan" class. This imbalance should be considered during model development to avoid biases towards the majority class.

**Class Distribution Visualization:** A pie chart visualization reinforces the imbalance, highlighting that the majority (approximately 90%) of instances correspond to customers who did not accept a personal loan.

**A blue and orange pie chart

Description automatically generated**

**Descriptive Statistics:** Descriptive statistics provide insights into the central tendency and dispersion of numerical features. Key observations include:

A table of numbers and letters

Description automatically generated

The average age of customers is approximately 45. Average income is around 74, with a wide range indicating potential income disparity. The average credit card spending (CCAvg) is about 1.94, reflecting variability in spending habits. The dataset includes customers with diverse financial profiles, as indicated by the broad range of values across features.

**Handling Outliers:** Outliers in numerical features, specifically 'Income,' 'CCAvg,' and 'Mortgage,' were identified using boxplots. Outliers can significantly impact model performance and addressing them is crucial. The threshold for identifying outliers was set at 3 standard deviations.

A diagram of a graph

Description automatically generated with medium confidence

**Correlation Analysis:** Understanding the relationships between variables is essential for feature selection and model interpretability. The correlation heatmap provides insights into pairwise correlations among features.

A screenshot of a graph

Description automatically generated

The correlation matrix provides insights into the relationships between various features in the dataset. Some key observations include a strong positive correlation of approximately 0.99 between 'Age' and 'Experience,' indicating a natural association between age and professional experience. 'Income' exhibits a moderate positive correlation of around 0.65 with 'CCAvg' and 'Mortgage,' suggesting that individuals with higher incomes tend to have higher average credit card spending and mortgage amounts.

Furthermore, there is a noteworthy positive correlation of about 0.50 between 'Personal Loan' and 'Income,' indicating that individuals with higher incomes are more likely to opt for personal loans. The matrix also reveals a moderate positive correlation of approximately 0.32 between 'CD.Account' and 'Personal Loan,' suggesting that individuals with Certificate of Deposit accounts are more inclined to take out personal loans.

**Data Processing**

**Handling Outliers:** Outliers were identified and analyzed in specific columns ('Income,' 'CCAvg,' 'Mortgage'). The process involved calculating mean and standard deviation, setting a threshold, and locating instances beyond the threshold. Outliers were then visualized using histograms.

**A graph of income

Description automatically generated**

The outlier’s data consists of 228 rows with 14 columns, where each row represents an outlier identified in the dataset.

Upon examining the outliers, it is evident that specific instances deviate significantly from the general patterns observed in the dataset. For example, in row 3896, the individual is aged 48, with an income of 224 and a ZIP code of 93940, exhibiting characteristics that distinguish it as an outlier.

Understanding and addressing outliers is crucial for ensuring the robustness of predictive models. These extreme values can disproportionately influence model training, potentially leading to skewed results. Therefore, careful consideration and appropriate treatment of outliers are essential in the data preprocessing stage to enhance the accuracy and reliability of subsequent analyses.

**Correlation Analysis on Outliers:** A correlation heatmap was created specifically for outliers to explore relationships among outlier instances. This helps in understanding if outliers in one variable coincide with outliers in another.

A graph of a financial graph

Description automatically generated with medium confidence

**Missing Values**

The dataset exhibits a reassuring absence of missing values across all columns. Each variable, including 'ID,' 'Age,' 'Experience,' 'Income,' 'ZIP Code,' 'Family,' 'CCAvg,' 'Education,' 'Mortgage,' 'Personal Loan,' 'Securities Account,' 'CD Account,' 'Online,' and 'CreditCard,' has a count of 0 missing values. This completeness ensures a robust foundation for subsequent data analysis and modeling efforts, as the absence of missing data eliminates the need for extensive imputation strategies. The dataset's completeness contributes to the reliability of insights derived from exploratory data analysis and supports the development of accurate predictive models.

**Conclusion from EDA and Data Processing:**

1. Class Imbalance Awareness: The class distribution in the 'Personal Loan' target variable is imbalanced, with most instances belonging to the "No Personal Loan" class. Addressing this imbalance is critical to avoid biased model outcomes.
2. Outlier Impact: Outliers, particularly in 'Income,' 'CCAvg,' and 'Mortgage,' can significantly affect model performance. Robust preprocessing techniques, such as outlier handling, are necessary to ensure the model's resilience to extreme values.
3. Correlation Insights: Understanding the correlation structure among features is crucial for informed feature selection. Identifying potential multicollinearity helps in optimizing the model's interpretability and generalization.
4. Missing Values Consideration: The presence of missing values in certain columns requires strategic imputation methods. Imputing missing values accurately is essential to maintain data completeness and avoid introducing biases during model training.
5. Next Steps Preparation: The insights gained from EDA and data processing lay a strong foundation for the next steps in model development. Prioritizing class balance, outlier handling, thoughtful feature selection, and robust imputation strategies will be essential for building a reliable and accurate predictive model.

**Model Development**

**Data Splitting:** The dataset was initially split into features (X) and the target variable (y), with 'Personal.Loan' representing the target. This is a fundamental step in machine learning, allowing us to train and test models on different datasets. An 80-20 split was employed, allocating 80% for training and 20% for testing.

**A white background with black text

Description automatically generated**

**Balancing (Random Undersampling):** A careful examination of the target variable distribution revealed a significant class imbalance. The instances of '0' (no personal loan) greatly outnumbered '1' (personal loan). To address this, Random Undersampling was applied to the training set. This technique aims to reduce instances of the majority class to achieve a more balanced representation. It helps prevent the model from being biased toward the majority class during training.

A blue and orange circle with text

Description automatically generated

**Feature Engineering: Standardization**

Feature engineering is crucial to ensure all variables are on the same scale. Standardization using StandardScaler was applied to the features. This step is essential, especially for algorithms sensitive to the magnitude of features. It prevents variables with larger scales from dominating the model training process.

A white rectangular sign with black text

Description automatically generated

**Random Forest Model:** The Random Forest Classifier, known for its robustness and ability to capture complex relationships in data, was chosen for its versatility. The model was trained on the balanced training set and evaluated on the test set. The choice of Random Forest was motivated by its capability to handle non-linear relationships and capture feature importance effectively.

**A screenshot of a computer

Description automatically generated**

A close up of a sign

Description automatically generated

The precision of the model indicates its ability to correctly identify positive instances (personal loan acceptances). With a precision of 1.00 for class 0 and 0.30 for class 1, the model excels at accurately predicting instances where a personal loan is not accepted. However, its precision for personal loan acceptances is lower, indicating a moderate level of false positives.

The recall and F1-score metrics highlight the model's robust performance, especially in correctly identifying instances of no personal loan acceptance. The model achieves a commendable mean accuracy of 96.4% across different folds in 10-fold cross-validation, underscoring its consistent and reliable predictive performance.

**Hyperparameter Tuning:** Hyperparameter tuning is a critical step in optimizing model performance. RandomizedSearchCV was employed to explore various combinations efficiently. The best hyperparameters were identified, resulting in a more fine-tuned Random Forest model. This process aims to enhance the model's predictive capabilities by selecting the most effective configuration of hyperparameters.

**A screenshot of a computer program

Description automatically generated**

The hyper-tuned Random Forest model exhibits a notable improvement, with an enhanced accuracy of 91.33% after 10-fold cross-validation. The refined model demonstrates improved precision and recall, particularly in correctly identifying instances of "1" (indicating acceptance of a personal loan), achieving a more balanced and reliable predictive performance.

**A screenshot of a computer

Description automatically generated**

**Actual versus Predicted Comparison (Random Forest):** A visual comparison between the distribution of actual values and predicted values was conducted. This step provides insights into where the model performs well and where it may have challenges. It aids in understanding the alignment between predicted and actual outcomes, facilitating an in-depth assessment of the model's predictive power.

A graph of a graph of a diagram

Description automatically generated with medium confidence

**Logistic Regression Model:** In addition to the Random Forest model, a Logistic Regression model was implemented and evaluated. Logistic Regression is suitable for binary classification problems and provides a baseline for comparison. The classification report and accuracy metrics offer a detailed breakdown of the model's performance.

**A screenshot of a computer

Description automatically generated**

The logistic regression model demonstrates good predictive performance with an overall accuracy of 66%. It excels in correctly identifying instances of "1" (personal loan acceptance) with a high recall of 100%, indicating its effectiveness in capturing positive cases. However, the precision for "1" is lower at 23%, suggesting that while the model identifies most actual positive instances, it also has a higher rate of false positives. The confusion matrix further illustrates this, showing a substantial number of false positives (343) in comparison to true positives (105). The mean accuracy with 10-fold cross-validation is 88.67%, indicating consistent performance across different subsets of the dataset.

A graph of a diagram

Description automatically generated with medium confidence

**Model Comparison - ROC Curve:** To compare the models comprehensively, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were employed. The ROC curve visualizes the trade-off between true positive and false positive rates, and AUC provides a single metric for comparison. The Random Forest model exhibited a higher AUC, indicating superior discriminatory power. The ROC AUC values indicate that the logistic regression model achieved a higher area under the curve (0.97) compared to the random forest model (0.94), suggesting better overall discrimination and balance between true positive rate and false positive rate for the logistic regression model.

**A graph of a function

Description automatically generated**

**Conclusion:** In conclusion, our data-driven model development process culminated in the selection of a finely tuned Random Forest model for predicting personal loan acceptance. The model showcased superior accuracy, bolstered by a meticulous approach to handle class imbalance and outliers. Through hyperparameter tuning, we optimized model performance, outperforming the baseline Logistic Regression. The inclusion of visualizations like ROC curves enriched our understanding of the model's discriminatory power. This refined predictive tool not only offers immediate business insights but also lays the groundwork for ongoing improvements and strategic decision-making.

**Model- Without Outliers**

* Data Cleaning: Removed outliers from the dataset using the outlier detection methods, resulting in a dataset denoted as df\_no\_outliers.
* Data Splitting: Split the cleaned dataset into training and testing sets (80% training, 20% testing) for model evaluation.
* Balancing Data: Applied random undersampling to balance the class distribution in the training set, as shown by the pie chart.
* Standardization: Standardized the features of the dataset to ensure uniformity and improve model performance.
* Logistic Regression Model: Trained and tested a logistic regression model on the cleaned and balanced dataset, achieving an accuracy of 89.74% and providing a detailed classification report.
* Random Forest Model: Trained and tested a random forest model on the same dataset, resulting in an accuracy of 96.64%. The classification report illustrates precision, recall, and F1-score for both classes.
* Visualization: Utilized distribution plots to visualize the actual versus fitted values for both logistic regression and random forest models.
* Comparison: Compared the performance of both models, emphasizing accuracy, precision, recall, and F1-score metrics.

**Why we developed a Model without Outliers:**

The decision to create a model without outliers in the bank loan dataset was driven by the need for a more accurate and reliable predictive model. Outliers, being extreme values, can significantly impact model training and distort predictions, particularly in financial scenarios like personal loan approval. By eliminating these extreme data points, we aimed to enhance the model's ability to generalize patterns and make more accurate predictions on unseen data. This preprocessing step is crucial in aligning the model with real-world scenarios, where extreme cases might not represent the typical behavior of the majority. Thus, the model without outliers is better suited for providing actionable insights in the context of bank loan approval.

**Model with Outliers**

In building the model with outliers included in the bank loan dataset, we aimed to assess how extreme data points influence predictive performance. This model considers the entire spectrum of data, including outliers, to understand the impact of such extreme cases on model generalization. The under-sampling technique was applied to balance the class distribution and prevent the model from being biased towards the majority class. Both logistic regression and random forest classifiers were employed for comparison.

**Logistic Regression:**

* Accuracy: The model achieved an accuracy of 93%, indicating its ability to correctly classify instances.
* Precision, Recall, F1-score: With precision and recall both exceeding 90%, the model demonstrates a good balance between identifying positive instances and avoiding false positives.
* Cross-validation: The 88.33% mean accuracy across 10-fold cross-validation ensures robustness.

**Random Forest Classifier:**

* Accuracy: The random forest model exhibited superior performance with a 96% accuracy, showcasing its effectiveness in handling outliers.
* Precision, Recall, F1-score: Precision and recall scores above 95% emphasize the model's proficiency in correctly classifying instances.
* Cross-validation: The 96.03% mean accuracy in cross-validation indicates consistent performance.

**Conclusion for Model with Outliers:** The model with outliers provides insights into the impact of extreme cases on predictive performance. While logistic regression shows robustness, the random forest model outperforms, emphasizing its resilience to outliers. The decision to include outliers is contextual, based on the importance of extreme cases in the business scenario. In this case, understanding the model's behavior with outliers’ aids in making informed decisions regarding risk assessment and loan approval.

**Key Learnings:**

* Data Understanding: Thoroughly understanding the dataset, its features, and the context of the problem is crucial for effective model development. In this project, a comprehensive grasp of the bank loan dataset laid the foundation for subsequent decisions.
* Handling Imbalanced Data: The presence of imbalanced classes, where loan approvals significantly outnumber rejections, requires careful handling. Employing techniques like random undersampling ensured fair representation of both classes in the training data.
* Model Evaluation Metrics: Choosing appropriate evaluation metrics is essential. Precision, recall, and F1-score are particularly relevant for a bank loan scenario, where false positives (approving a risky loan) and false negatives (rejecting a worthy applicant) have different consequences.
* Balancing Interpretability and Performance: Striking a balance between model interpretability and performance is crucial. Logistic regression provides transparency, while Random Forest, though less interpretable, offers higher accuracy and robustness against outliers.
* Hyperparameter Tuning: Utilizing hyperparameter tuning techniques like random search enhances model performance. Fine-tuning parameters using cross-validated search ensures optimal settings for better generalization.
* Continuous Model Improvement: Models should be treated as evolving entities. Continuous monitoring, periodic reevaluation, and updates based on changing data patterns contribute to sustained performance.
* Ethical Considerations: Acknowledging and addressing ethical considerations, such as bias in lending decisions, is essential. Implementing strategies to identify and mitigate biases ensures fairness and responsible use of the model.