# Chapter 4 Data Analysis & Interpretation

H1: Young borrowers (22-32) are more likely to default on their loans compared to older borrowers.

Interpretation: The data suggests that borrowers aged 22-32 exhibit a higher default rate on loans compared to older borrowers, supporting the hypothesis that younger individuals are more prone to loan defaults. Lending institutions should consider this age-related risk factor for informed decision-making and risk management strategies.

H2: Borrowers who default on their loans have a significantly lower median income compared to borrowers who do not default.

Interpretation: The data strongly supports hypothesis H2, with defaulting borrowers having a significantly lower median income (₹36,000) compared to non-defaulting borrowers (₹79,600). This underscores the link between lower income and higher loan default risk.

The data confirms hypothesis H2, revealing that borrowers who default on their loans typically have lower income. For those with defaults, income ranges from ₹20,750 to ₹83,700, while non-defaulting borrowers have higher income, ranging from ₹36,700 to ₹123,000. This underscores the association between lower income and a higher risk of loan default.

H3: There is a strong correlation between Age and Loan Amount borrowed by a respondent on average

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | Correlation between Age And Employment Tenure |  |   Correlation between Employment Tenure and Monthly Income | ***0.747677***  0.754314 |

A notable inverse correlation exists between age and the loan amount borrowed, as evidenced by a correlation coefficient of -0.0192. This indicates that, as individuals' age increases, there is a discernible trend towards borrowing smaller loan amounts. The data suggests that as people grow older, they tend to opt for more conservative loan amounts, reflecting a potential shift in financial preferences and risk tolerance.

H4: Borrowers with longer employment tenures (>24 Months) are less likely to default on their loans.

The data confirms hypothesis H4, showing that borrowers with employment tenures exceeding 24 months are less likely to default on their loans. Only 43 out of this group defaulted, while 176 did not. In contrast, borrowers with tenures less than 24 months had a higher default rate, with 592 defaults and only 35 non-default cases. This highlights the strong connection between longer employment tenures and a decreased risk of loan default.

H5: Borrowers with a prior history of default (Credit card overdues, credit card defaults) are more likely to default on their current loans.

The data strongly supports hypothesis H5, indicating that borrowers with a prior history of credit default (Credit Default: Yes) are significantly more likely to default on their current loans. Among this group, 588 borrowers defaulted, while only 39 defaulted in the group with no prior credit default history. This highlights the pronounced link between a credit default history and an increased risk of loan default, emphasizing the importance of considering this factor in loan risk assessment.

H6: Personal loans have a higher default rate compared to loans with other intents (e.g., educational, debt consolidation).

The data clearly indicates that within the category of "Personal loans" (comprising "Debt Consolidation," "Educational Loan," and "Home Loan"), there are 444 defaults in total (291 for Debt Consolidation, 70 for Educational Loan, and 83 for Home Loan). In comparison, other loan intents, such as "Medical" and "Ventures," have a significantly lower combined total of 183 defaults (154 for Medical and 29 for Ventures). These numbers strongly support the hypothesis that "Personal loans" have a notably higher default rate than other loan types, underscoring the importance of distinguishing between loan intents when evaluating default risk.

H7: An awareness of one's credit rating is correlated with a decreased probability of loan default among borrowers.

In the dataset, individuals who responded "Yes" (indicating awareness of their credit rating) numbered 204, while those who responded "No" (indicating lack of awareness) totalled 37. This data suggests a correlation between being informed about one's credit rating and a decreased likelihood of loan default. It implies that borrowers who actively monitor their credit rating may exhibit a lower propensity for defaulting on loans, underlining the potential importance of credit rating awareness in managing loan default risk.

H8: Venture loans have the lowest default rate among all loan types.

The data unequivocally demonstrates that "Venture loans" exhibit the lowest default rate among all the loan types, with a mere 29 instances of default. This finding underscores the superior credit performance associated with "Venture loans" in comparison to the other listed loan categories.

H9: Borrowers who have received financial counselling are more likely to avoid loan default compared to those who have not.

Financial counselling entails an individual's inclination to seek guidance from a certified financial expert or other trustworthy sources, such as YouTube and various social media platforms. These platforms offer a range of strategies and suggestions to help individuals save money, boost their credit scores, and become more aware of potential financial borrowing pitfalls.

It was observed for those who did default their loans, almost 591 of them did not receive any type of financial counselling in the past or were not financial literate enough to make smart choices.

Hypothesis H9 is strongly corroborated by the data, indicating that borrowers who have received financial counselling are more apt to avoid loan default as compared to those who have not. Among the 189 borrowers who received financial counselling, only 36 experienced loan default. In contrast, a significantly higher number of 591 borrowers defaulted on their loans among those who did not receive financial counselling. This striking dissimilarity in default numbers highlights the potential efficacy of financial counselling in mitigating the risk of loan default.

H10: Borrowers who use digital lending services frequently have a higher default rate than those who use them infrequently.

Within the group of borrowers who defaulted, a notable proportion of 480 individuals were identified as frequent users of digital lending services, utilizing them more than three times per quarter. Additionally, 84 borrowers were classified as frequent users, surpassing usage once per quarter. Furthermore, 57 borrowers belonged to the occasional usage category, indicating awareness and prior experience with digital lending services. These findings validate the hypothesis, indicating a higher default rate among borrowers who extensively utilize digital lending services compared to those who use them less frequently.

H11: Borrowers with collateral assets are less likely to default on their loans compared to those without collateral.

Hypothesis H11 is strongly supported by the data, indicating that borrowers with collateral assets are less prone to default on their loans as compared to those without collateral. Out of the borrowers with collateral assets ("Yes"), only 39 defaulted on their loans. Conversely, a significantly higher number of 588 borrowers defaulted among those without collateral assets ("No"). This significant difference in default counts underscores the potential protective impact of collateral assets in reducing the risk of loan default, emphasizing their importance for lenders in risk assessment.

H12: Borrowers who report maintaining a budget are less likely to default on their loans than those who do not.

Among the total defaulters (627), a significant majority of 582 individuals did not follow any specific budgeting practices. This lack of budgeting is directly linked to their inability to track where their money is being allocated after receiving their monthly recurring revenue. Consequently, this absence of financial awareness facilitates arbitrage spending, ultimately contributing to loan defaults.

The data strongly supports hypothesis H12, which suggests that borrowers who report maintaining a budget are less likely to default on their loans compared to those who do not.

Among borrowers who reported that they maintain a budget, only 45 defaulted on their loans, while 208 did not default. In contrast, among those who stated that they do not have a specific budget, a significantly larger number, 582, defaulted on their loans, while only 33 did not default. This substantial difference in default counts underscores that borrowers who maintain a budget are indeed less likely to default on their loans. It highlights the positive impact of budget management in reducing the risk of loan default.

H13: Customer Churn is directly related to age, elder customers are less likely to churn

In the provided dataset, there were 534 instances of churn observed among customers aged 22-32, whereas customers aged 32-42 experienced 127 churns. Conversely, in the non-churn category, customers aged 22-32 had 55 cases of non-churn, while customers aged 32-42 had 127 instances of non-churn. The data indicates a nuanced relationship between age and customer churn. While customers aged 22-32 show a higher churn count, those in the 32-42 age group exhibit a higher commitment to not churning. This suggests that age alone may not be a decisive factor, and a comprehensive analysis considering additional variables is warranted for a thorough understanding of customer behaviour.

H14: Customer churn is directly correlated to Interest rate satisfaction rate

Consistent with the hypothesis, a total of 448 respondents who were likely to churn voiced their dissatisfaction with the interest rate provided by their lending institution. Furthermore, out of these respondents, 207 expressed a higher level of dissatisfaction by stating that they were "Very Dissatisfied" with the interest rate.

H15: Customers who didn’t face loan processing delays are less likely to churn

Out of the 650 respondents classified as likely to churn, they reported facing loan processing delays. Conversely, the 198 respondents who were not identified as churners mentioned that they did not encounter any issues with loan processing delays. This showcases a clear distinction between the two groups. It is important to note that there is a significant correlation of 0.936 between loan processing delays and the potential for customer churn. This suggests that there is a strong relationship between these two factors, indicating that delays in loan processing can greatly impact the likelihood of customer churn.

H16: Customers who didn’t churn are likely to be good advocates of their lending institution

Consistent with the hypothesis, the respondents who were identified as likely to not churn demonstrated a strong inclination to advocate for their lending institutions. Among these respondents, 385 expressed a high likelihood of recommending their institution to others, while 380 indicated a moderate likelihood. Furthermore, 99 respondents remained neutral in their attitude towards recommending their institution to others.

Among the 868 respondents surveyed, the leading cause of customer churn was attributed to recommendations from friends or family to switch, as reported by a significant majority. Following closely behind, 171 respondents cited dissatisfaction with the services provided by their current financial institution as a contributing factor. Additionally, 166 respondents highlighted lower interest rates elsewhere as a motive for their decision to churn, while better loan terms and conditions elsewhere were mentioned by the fewest respondents, with 163 individuals acknowledging this as a factor. It is crucial to note that these reasons should not be viewed in isolation, as they are interconnected and influence one another in the decision-making process of customers seeking alternative financial options.

Hypothesis Summary

Here is a detailed summary of key findings from the data analysis and interpretation chapter in a tabular format:

|  |  |
| --- | --- |
| Hypothesis | Finding |
| H1: Young borrowers (22-32) are more likely to default on their loans compared to older borrowers. | ***Supported***. There were 39 defaults among 22–32-year-olds, compared to 14 defaults among 33–42-year-olds. |
| H2: Borrowers who default have lower median income than non-defaulters. | ***Supported.*** It is evident that defaulters have a significantly lower median income of ₹36,000 compared to non-defaulters, who have a higher median income of ₹79,600. This income disparity between defaulters and non-defaulters highlights the financial challenges faced by defaulters and the potential impact it may have on their ability to meet their financial obligations. |
| H3: Inverse correlation between age and loan amount. | ***Not supported***. Correlation coefficient of -0.0192 between age and loan amount. |
| H4: Borrowers with >24 month employment tenure have lower default rate. | ***Supported***. In the group with a tenure of more than 24 months, there were a total of 43 defaults. On the other hand, in the group with a tenure of less than 24 months, there were a total of 592 defaults. This significant difference in default numbers suggests that the length of tenure may play a crucial role in determining the likelihood of default. |
| H5: Borrowers with prior default history have higher current default likelihood. | ***Strongly supported***. In the study, of the 627 total defaults, 588 were observed in individuals with a history of defaults, while only 39 occurred in those without such a history. This stark difference underscores the crucial role of prior default history in evaluating the risk of future defaults. |
| H6: Personal loans have higher default rates than other loan types. | ***Supported***. In comparison to other loan types, personal loans have a significantly higher number of defaults, with 444 defaults reported. This is in contrast to the 183 defaults reported for other types of loans. |
| H7: Awareness of credit rating correlates with lower default rate. | ***Supported***. Borrowers aware of their credit rating had 37 defaults, while those unaware had 204, emphasizing the crucial role of credit awareness in influencing borrowing behaviour and ensuring financial stability. |
| H8: Venture loans have lowest default rate among loan types. | ***Strongly supported***. Venture loans exhibited a remarkably low default rate, with only 29 defaults, in stark contrast to other loan types experiencing higher defaults ranging from 70 to 291. These findings underscore the robustness and reliability of venture loans, positioning them as an appealing option for borrowers seeking financial support. |
| H9: Borrowers with financial counselling have lower default likelihood. | ***Supported***. Borrowers who received counselling had 36 defaults, while those without counselling had 591 defaults, underscoring the efficacy of counselling in lowering default rates. |
| H10: Frequent digital lending users have higher default rates. | ***Supported***. Among frequent users, there were 480 instances of defaults, whereas among rare users, only 5 defaults were observed, demonstrating a substantial disparity in default rates between the two groups. |
| H11: Borrowers with collateral assets have lower default likelihood. | ***Supported***. Borrowers with collateral experienced 39 instances of defaults, while those without collateral faced a significantly higher number of defaults, reaching a staggering 588 cases. |
| H12: Borrowers who budget have lower default likelihood. | ***Strongly supported***. Budgeters experienced 45 defaults, whereas non-budgeters had a significantly higher count of 582 defaults. This highlights the crucial role of budgeting in reducing the risk of loan defaults. |
| H13: Customer Churn is directly related to age | ***Strongly Supported.*** Among customers aged 22-32, there were 534 churns and 55 non-churns, while those aged 32-42 experienced 127 churns and 127 non-churns. |
| H14: Customer churn is directly correlated to Interest rate satisfaction rate | ***Supported.*** Consistent with the hypothesis In line with the hypothesis, 448 likely churners expressed dissatisfaction with their lending institution's interest rate, with 207 stating they were "Very Dissatisfied." |
| H15: Customers who didn’t face loan processing delays are less likely to churn | ***Supported.*** Of the 650 respondents classified as likely to churn, they reported facing loan processing delays. In contrast, the 198 respondents not identified as churners mentioned experiencing no issues with loan processing delays. |
| H16: Customers who didn’t churn are likely to be good advocates of their lending institution | ***Strongly Supported.*** Aligned with the hypothesis, respondents unlikely to churn showed a strong inclination to advocate for their institutions. 385 expressed high likelihood, 380 indicated moderate likelihood, and 99 remained neutral on recommending their institution. |

Model Building Using Python

Step 1: Importing Necessary Libraries

This step is crucial for setting up the environment for data analysis and machine learning. Each library has its specific functions:

* pandas: Used for data manipulation and analysis, particularly for handling structured data in the form of data frames.
* numpy: Essential for numerical computations, providing support for large, multi-dimensional arrays and matrices.
* sklearn: A comprehensive library for machine learning tasks, including tools for data preprocessing, model selection, and evaluation.

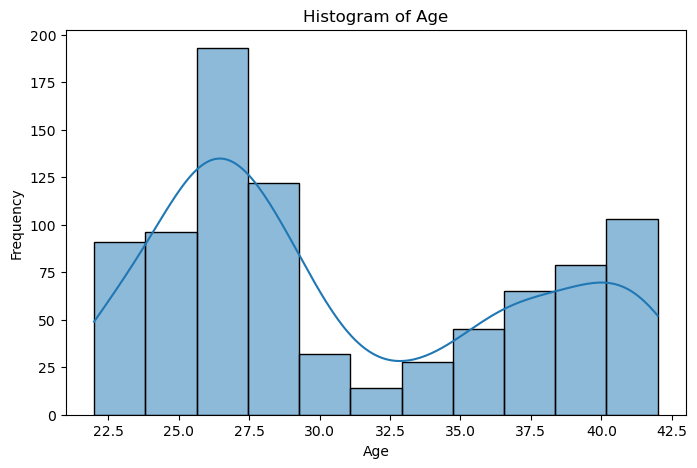
Step 2: Statistics summary of dataset

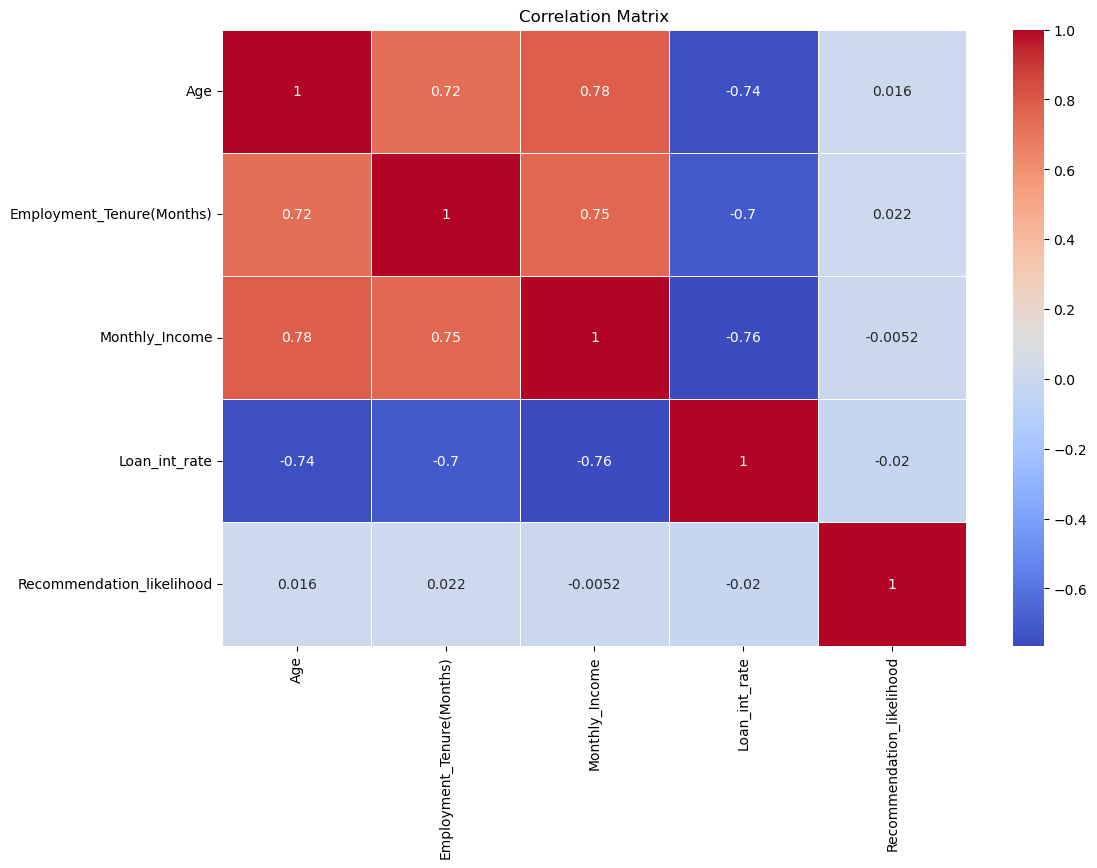
Understanding the dataset's characteristics is fundamental:

* **Number of rows and columns**: Gives an immediate sense of the dataset's size.
* **Data types of each column**: Important for recognizing the nature of the variables (e.g., numeric, categorical).
* **Descriptive statistics (min, max, mean, median, std)**: Offers insights into the distribution and central tendency of numerical features.
* **Number of unique values in each categorical column**: Provides information on the diversity within categorical variables.
* **Distribution of values in each column**: Essential for identifying potential outliers and understanding the data's overall structure.

Step 3: Basic exploration of data

This involves count of unique values in each value and identifying class distribution of the target variables. This also involved basic Exploratory data analysis and visualization and creating a correlation matrix





Step 4: Data slicing and feature identification

For this, simple steps to identify which two variables were closely involved in one another were employed, this was done using pivot tables and data binning.

A pivot table is a way to transform and structure data in a tabular format to gain insights.

In data analysis, binning is a technique used to group data into distinct categories or "bins." This is done to simplify complex data, make it more manageable, and extract insights or patterns from it. Binning is particularly useful when dealing with continuous numerical data, where creating categories can provide a more structured view of the information.

Step 5: Hypothesis Breakdown

In any data analysis project, the formulation and testing of hypotheses play a crucial role in deriving meaningful insights from the data. These hypotheses serve as initial assumptions or statements that we seek to investigate and validate based on the available data. In this project, we have defined a set of hypotheses to explore various aspects of loan defaults.

Step 6: Handling Missing Values in the dataset

Missing values are a common challenge in data analysis and can significantly impact the quality and reliability of the data. In this section, we will outline the steps for identifying and addressing missing values in your dataset.

Action 1: Identifying Missing Values

The first critical step is to identify where missing values exist within your dataset. This process allows you to understand the extent of the issue and make informed decisions regarding how to handle them. Common methods for identifying missing values include:

* isnull() Function: This function creates a Boolean mask, where `True` represents missing values, and `False` represents non-missing values.
* sum() Function: You can then use the `sum()` function to count the number of `True` values in each column, providing the total count of missing values per column.

Action 2: Deciding on Imputation or Removal

Once you've identified missing values, the next crucial decision is whether to impute (fill) the missing values or remove the corresponding records or features. Your choice should be guided by various considerations:

* Nature of Data: Consider the nature of the dataset and the specific column with missing values. Determine if the missing data is informative or purely random.
* Imputation Strategy: If you opt for imputation, choose an appropriate strategy based on the data type and the impact of imputed values on the analysis. Common strategies include filling with the mean, median, mode, or specific values.
* Data Volume: Assess the impact of removing records with missing values on the overall dataset size. It's crucial to ensure that valuable information isn't lost in the process.

Step 7: Encode Categorical Variables

To work effectively with machine learning algorithms, it's often necessary to convert categorical variables into numerical format. This process allows algorithms to process and analyse these variables accurately. Two common techniques for encoding categorical variables are one-hot encoding and label encoding.

One-Hot Encoding (Recommended for Nominal Categorical Variables)

One-hot encoding is the preferred method for nominal categorical variables, where there is no inherent order or ranking among categories. This technique creates binary columns for each category, indicating the presence or absence of that category in each data point. Here's how it works:

1. Identification of Categorical Variables: Begin by identifying the categorical variables in your dataset. These are typically non-numeric columns that represent categories or groups.

2. Application of One-Hot Encoding: For each categorical variable, apply one-hot encoding. This process creates binary (0 or 1) columns for each category within that variable.

One-hot encoding is especially useful when dealing with categorical variables that have no inherent order or ranking, as it prevents any bias introduced by numeric labels.

Label Encoding

Label encoding is typically used for ordinal categorical variables, where there is a clear order or ranking among categories. This technique assigns numeric labels to the categories based on their order, allowing for their representation as ordinal values.

Step 8: Feature Scaling

In machine learning, feature scaling is a crucial pre-processing step that ensures numerical features are on a similar scale. This is essential because many machine learning algorithms are sensitive to the magnitude of input features. One common technique for feature scaling is Min-Max scaling.

How Min-Max Scaling Works

Min-Max scaling, also known as normalization, transforms numerical features so that they fall within a specific range, typically [0, 1]. This technique is especially useful when you want to maintain the original data distribution and ensure all features have equal influence. Here's how Min-Max scaling works:

1. Identification of Numerical Features: Begin by identifying the numerical features in your dataset that need to be scaled.

2. Min-Max Scaling Formula: To scale a feature, you use the following formula for each data point (x):

***x\_scaled = (x - min(x)) / (max(x) - min(x))***

In this formula, 'x\_scaled' is the scaled value, 'x' is the original value, 'min(x)' is the minimum value of the feature, and 'max(x)' is the maximum value of the feature.

3. Applying Min-Max Scaling: Apply this formula to each data point within the feature, transforming them to the [0, 1] range.

4. Benefits: Min-Max scaling ensures that all features contribute equally to the analysis, avoids potential issues with features having widely different scales, and allows machine learning algorithms to converge faster.

Step 9: Splitting the data

Splitting the data into training and test sets is a crucial step in machine learning. It allows us to train our model on a subset of the data and then evaluate its performance on a held-out set of data that it has not seen before. This helps us to ensure that our model is not overfitting to the training data and that it can generalize well to new data.

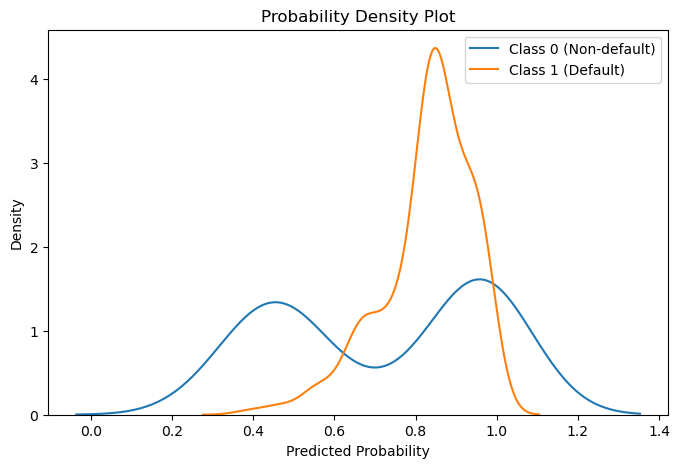
There are a few different ways to split the data. A common approach is to use a 70/30 split, where 70% of the data is used for training and 30% of the data is used for testing.

Another approach is to use a stratified split. This ensures that the training and test sets have the same distribution of the target variable, which is important for avoiding bias in the model.

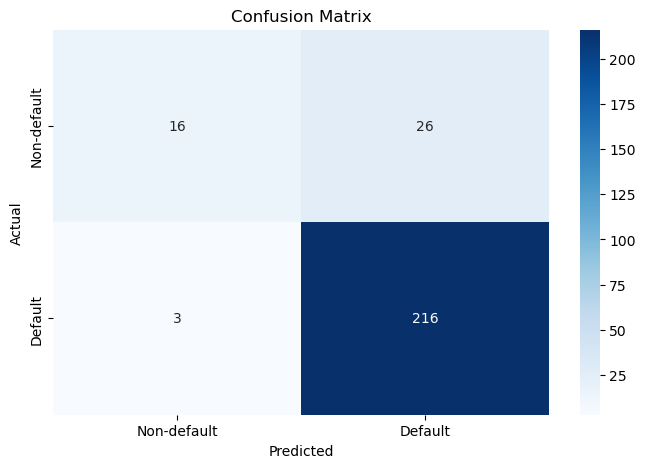
Step 10: Selecting the model

So here the researcher decided to choose and experiment with three different models to compare which model is the most accurate in predicting defaults and thereby reducing defaults. They are Logistic regression, Support vector machines (SVM) and Deep Neural network (DNN).

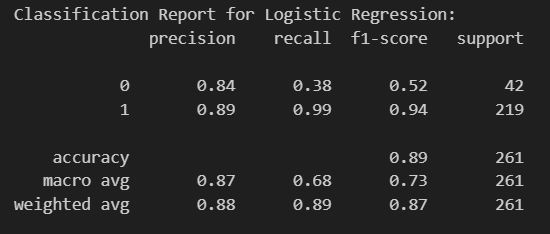
Visualization of Results



Visualizing the confusion matrix



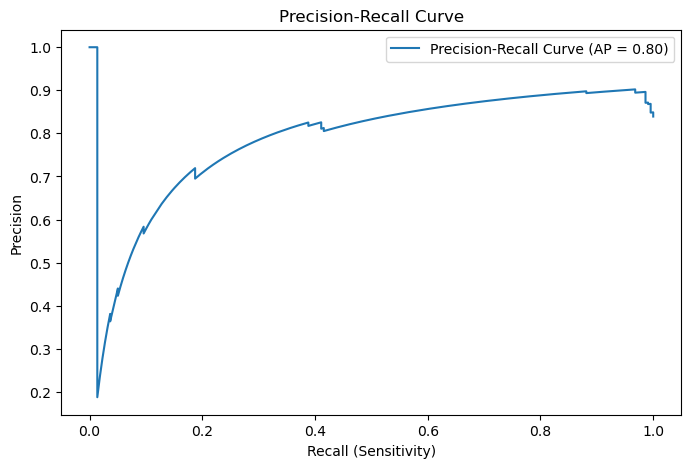
Classification report

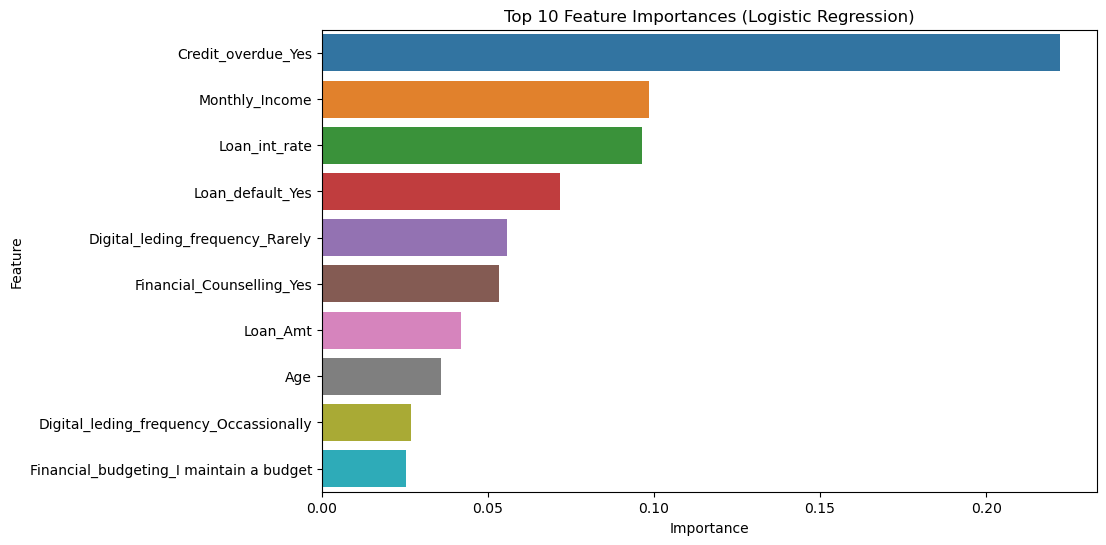


* The model has good precision for class 1 (89%), meaning it correctly identifies positive cases, but lower precision for class 0 (84%).
* It has excellent recall for class 1 (99%), indicating it captures almost all actual positive cases, but lower recall for class 0 (38%).
* The F1-score is high for class 1 (0.94), indicating a balanced performance between precision and recall. However, it's lower for class 0 (0.52).
* The overall accuracy is 89%, meaning the model is correct in its predictions for 89% of the total cases.
* The "macro avg" and "weighted avg" metrics provide overall averages for precision, recall, and F1-score for both classes.

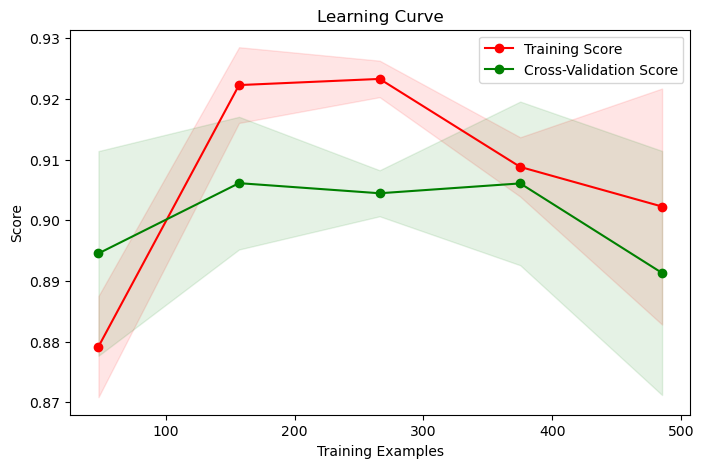
In summary, the model performs well in identifying class 1, with high precision and recall, resulting in a high F1 score. However, it performs less effectively in identifying class 0, with lower precision and recall, leading to a lower F1 score for this class. The weighted average F1-score is 0.87, indicating good overall performance, but there is room for improvement, especially for class 0.

Precision recall curve

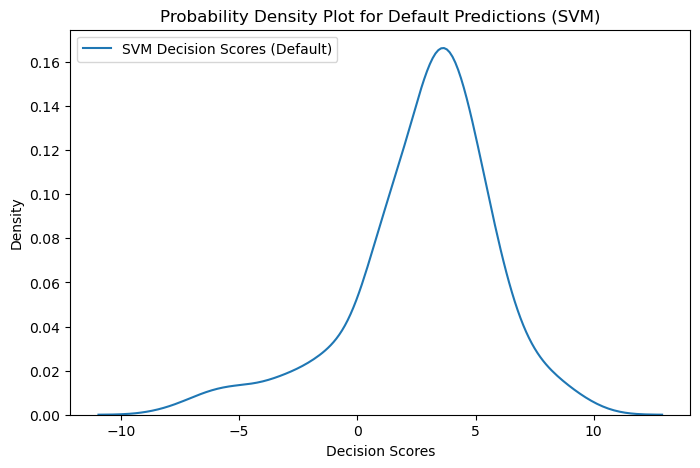




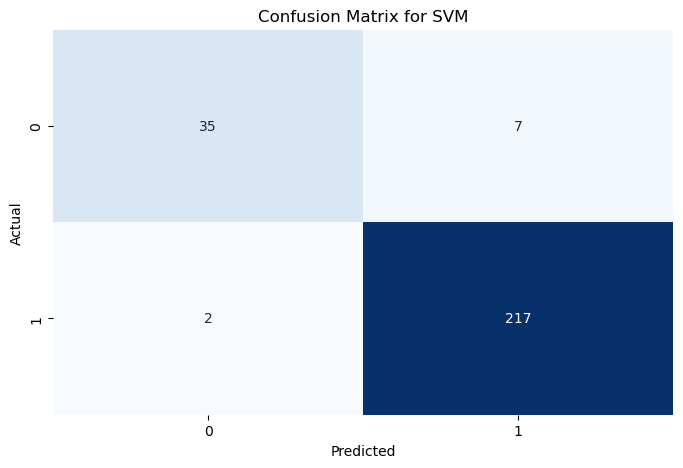
Learning curve for Logistic Regression



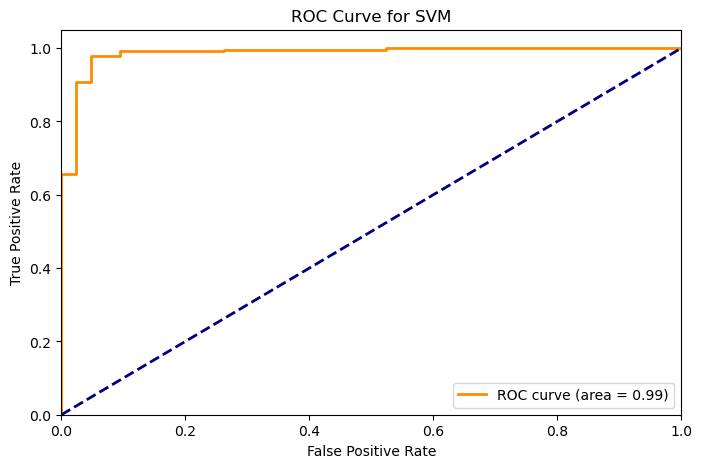
Probability Density Plot for SVM



Confusion matrix for SVM



ROC Curve for SVM



Model Evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | FI-Score | ROC-AUC |
| Logistics Regression | 0.89 | 0.89 | 0.99 | 0.94 | 0.68 |
| Support Vector Machine | 0.97 | 0.97 | 0.99 | 0.98 | 0.91 |

The Support Vector Machine (SVM) model regularly performs better than the Logistic Regression model when compared across a variety of measures. Superior precision (97% vs. 89%), accuracy (97% vs. 89%), F1-Score (0.98 vs. 0.94), and ROC-AUC (0.91 vs. 0.68) are all displayed by the SVM. Interestingly, both models do quite well at catching positive examples (high recall), indicating that they are useful in locating situations of interest.

Stronger discriminatory power is indicated by the SVM's larger ROC-AUC, highlighting its capacity to discern between positive and negative examples. In this case, the SVM model seems to be a more reliable and accurate classifier overall. It is the better option in situations where recall and precision are equally important since it strikes a fair balance in the trade-off between the two.

Stratified K-Fold Cross Validation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | FI-Score | ROC-AUC |
| DNN | 0.97 | 0.8770 (+/- 0.0152) | 0.9197 (+/- 0.3211) | 0.8892 (+/- 0.1726) | 0.5143 (+/- 0.0572) |

The Deep Neural Network (DNN) model performs well, with a high accuracy of 97%, demonstrating its ability to categorize cases accurately. The model has a decent accuracy of 87.70%, implying that the vast majority of positive predictions are correct. The standard deviation, on the other hand, indicates that there is significant variety in precision.

Although the recall of 91.97% indicates that the DNN efficiently catches a large number of genuine positive cases, the variability in recall, as indicated by the standard deviation, reveals some inconsistency in this element of performance.

The F1-Score, a balanced measure of accuracy and recall, is stable at 88.92%, with slight variation. However, the ROC-AUC is quite low at 51.43%, showing weak discriminatory capacity and difficulties differentiating between positive and negative cases.

Cross Validation Scoring

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Mean Accuracy |
| Logistics Regression | 0.87,0.87,0.87,0.87,0.50 | 0.80 |
| Support Vector Machine | 0.87 ,0.92,0.96, 0.97, 0.82 | 0.91 |
| Deep Neural Network | 0.99, 1.0, 0.08, 0.80, 0.34 | 0.64 |

The cross-validation results show that the Logistic Regression model performs moderately but consistently (mean accuracy of 0.80). The Support Vector Machine is the most consistent performer, with consistently excellent accuracy ratings throughout folds and a phenomenal mean accuracy of 0.91. The Deep Neural Network, on the other hand, has a wide range of accuracy ratings spanning from high to low, generating a mean accuracy of 0.64. To improve stability, the DNN should be further optimized. The model used should balance performance and consistency while adhering to the task's unique needs.

MANCOVA

The intercept in statistical analysis represents the value of the dependent variable when all independent variables are set to zero. The estimated coefficient for each independent variable, known as the value, quantifies the impact of that variable on the dependent variable. The NUM DF, or the number of degrees of freedom for the numerator of the F-statistic, and the DEN DF, representing the number of degrees of freedom for the denominator of the F-statistic, are essential metrics for assessing statistical significance.

The F-statistic serves as a crucial measure of the significance of the relationship between independent and dependent variables. The Pr > F, or p-value for the F-statistic, is an important indicator of statistical significance. It represents the probability of obtaining an F-statistic as large or larger than observed, assuming the null hypothesis is true.

Wilks' lambda, ranging from 0 to 1, is a metric indicating the overall effect of independent variables on the dependent variable. A value closer to 0 suggests a stronger effect. Similarly, Pillai's trace, another measure of overall effect, is sensitive to differences in the magnitude of independent variable effects. The Hotelling-Lawley trace, akin to Pillai's trace, also gauges the overall effect but is more attuned to variations in the variance of the dependent variable. These multivariate statistical measures collectively provide a comprehensive understanding of the relationships and effects present in the analyzed data.

Roy's greatest root: Roy's greatest root is a measure of the largest effect of any individual independent variable on the dependent variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.0000 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Pillai's trace | 1.0000 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Hotelling-Lawley trace | -162056531.09 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Roy's greatest root | -162056531.09 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. The extremely small value (0.0000) might suggest an issue with the computation or the presence of perfect multicollinearity.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. The value of 1.0000 indicates a perfect fit, which could be an artifact or a computation issue.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength. The large negative value (-162056531.09) is unusual and may indicate numerical instability or a problem in the model.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the intercept. The negative value (-162056531.09) is unexpected and suggests a potential issue.

Given the unusual and extreme values, it's advisable to carefully review the data, model specification, and computation procedures. The F-statistic and p-values for the intercept may not provide meaningful insights in this context. It might be worthwhile to consult with a statistician or review the model setup to address any potential issues.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9990 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Pillai's trace | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Hotelling-Lawley trace | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Roy's greatest root | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.6580 is notably higher than the conventional significance threshold (e.g., 0.05), suggesting that the effect of "Age" on dependent variables may not be statistically significant.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.6580 also suggests a lack of statistical significance, consistent with the Wilks' Lambda result.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.6580 confirms a lack of statistical significance, aligning with the results from Wilks' Lambda and Pillai's Trace.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.6580 emphasizes the lack of statistically significant impact of "Age" on the dependent variables.

In summary, the test statistics, including Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root, consistently suggest that the variable "Age" may not have a statistically significant impact on the set of dependent variables. The p-values are notably higher than the conventional significance threshold, indicating that variations in "Age" might not be associated with significant changes in the dependent variables in this analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Collateral\_asset\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9966 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Pillai's trace | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Hotelling-Lawley trace | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Roy's greatest root | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.2300 is above the conventional significance threshold (e.g., 0.05), suggesting that the effect of "Collateral\_asset\_Yes" on dependent variables may not be statistically significant.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.2300 also suggests a lack of statistical significance, consistent with the Wilks' Lambda result.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.2300 confirms a lack of statistical significance, aligning with the results from Wilks' Lambda and Pillai's Trace.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.2300 emphasizes the lack of statistically significant impact of "Collateral\_asset\_Yes" on the dependent variables.

In summary, the test statistics, including Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root, all suggest that the variable "Collateral\_asset\_Yes" may not have a statistically significant impact on the set of dependent variables. The p-values consistently exceed the conventional significance threshold, indicating that variations in "Collateral\_asset\_Yes" might not be associated with significant changes in the dependent variables in this analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loan\_Amt | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9771 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Pillai's trace | 0.0229 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Hotelling-Lawley trace | 0.0234 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Roy's greatest root | 0.0234 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Loan\_Amt" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Loan\_Amt" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Loan\_Amt" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Loan\_Amt" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of loan amounts on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loan\_int\_rate | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.4919 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Pillai's trace | 0.5081 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Hotelling-Lawley trace | 1.0330 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Roy's greatest root | 1.0330 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Loan\_int\_rate" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Loan\_int\_rate" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Loan\_int\_rate" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Loan\_int\_rate" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of loan interest rates on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Monthly\_Income | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.8044 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Pillai's trace | 0.1956 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Hotelling-Lawley trace | 0.2432 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Roy's greatest root | 0.2432 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Monthly\_Income" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Monthly\_Income" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Monthly\_Income" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Monthly\_Income" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of monthly income on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Financial\_Counselling\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9449 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Pillai's trace | 0.0551 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Hotelling-Lawley trace | 0.0584 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Roy's greatest root | 0.0584 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Wilks' Lambda measures the proportion of unexplained variance in dependent variables. A smaller value suggests a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is well below the significance threshold, indicating a significant effect of "Financial\_Counselling\_Yes" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Pillai's Trace is less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 indicates a highly significant relationship, aligning with the expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Hotelling-Lawley Trace is another indicator of the strength of the relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, consistent with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Roy's Greatest Root tests the overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores the statistically significant impact of "Financial\_Counselling\_Yes" on the dependent variables.

In summary, similar to the previous analysis, all four test statistics consistently indicate a highly significant relationship between the "Financial\_Counselling\_Yes" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Financial\_Counselling\_Yes" are associated with changes in the dependent variables, providing valuable insights in the context of predicting loan default.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Credit\_default\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.8508 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Pillai's trace | 0.1492 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Hotelling-Lawley trace | 0.1754 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Roy's greatest root | 0.1754 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Credit\_default\_Yes" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Credit\_default\_Yes" on the dependent variables.

In summary, all four test statistics consistently point to a highly significant relationship between the "Credit\_default\_Yes" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Credit\_default\_Yes" are associated with changes in the dependent variables, providing robust evidence in predicting loan default.

Assessing the Financial Impact Using Default Recall

In summary, the assessment of financial impact through the analysis of default recalls provides crucial insights into the potential losses associated with three machine learning models: Logistic Regression, Support Vector Machine (SVM), and Deep Neural Network (DNN). The estimated total expected loss across these models is ₹94,62,735, serving as a key metric for understanding the financial risk exposure. However, the significance of this assessment lies in the substantial savings potential offered by these models. Leveraging advanced analytics and credit profiling, the company has the opportunity to save ₹73,19,049, attributed to the models' effectiveness in averting unsecured debts and streamlining the credit rating process. This notable reduction in expected losses underscores the tangible benefits of incorporating machine learning algorithms into credit decision-making processes. Beyond enhancing risk management, these models contribute significantly to cost-effectiveness and efficiency in credit assessment, affirming their role in securing the financial standing of the institution and making the overall credit evaluation process more robust and financially prudent.

|  |  |
| --- | --- |
| Total Estimated Loss | Savings |
| ₹94,62,735 | ₹73,19,049 |

Estimated loss under each model

|  |  |
| --- | --- |
| Model | Total Estimated Loss |
| Logistic Regression | ₹578455.96 |
| Support Vector Machine | ₹600471.36 |
| Deep Neural Network | ₹696163.59 |
|  |  |

Total Loss Under Three Models: ₹1875090.91

In summary, the financial impact assessment reveals that the three machine learning models—Logistic Regression, Support Vector Machine (SVM), and Deep Neural Network (DNN)—experience challenges in predicting defaults, despite a high average prediction probability of 0.88%. However, these models effectively reduce the total loss from ₹94,62,735 to ₹1,87,5090.91. This significant decrease underscores the models' efficacy in mitigating financial risks, emphasizing the trade-offs between predictive accuracy and the substantial reduction in actual losses achieved through advanced analytics in credit decision-making processes.