# Loan Application Dataset Analysis and Model Performance Report

## Dataset Explanation

The dataset appears to be related to loan applications. Here’s a brief explanation of each column:

* **person\_age**: Age of the individual applying for the loan.
* **person\_income**: Annual income of the individual.
* **person\_home\_ownership**: Type of home ownership of the individual.
  + **rent**: The individual is currently renting a property.
  + **mortgage**: The individual has a mortgage on the property they own.
  + **own**: The individual owns their home outright.
  + **other**: Other categories of home ownership that may be specific to the dataset.
* **person\_emp\_length**: Employment length of the individual in years.
* **loan\_intent**: The intent behind the loan application.
* **loan\_grade**: The grade assigned to the loan based on the creditworthiness of the borrower.
  + **A**: The borrower has a high creditworthiness, indicating low risk.
  + **B**: The borrower is relatively low-risk, but not as creditworthy as Grade A.
  + **C**: The borrower’s creditworthiness is moderate.
  + **D**: The borrower is considered to have higher risk compared to previous grades.
  + **E**: The borrower’s creditworthiness is lower, indicating a higher risk.
  + **F**: The borrower poses a significant credit risk.
  + **G**: The borrower’s creditworthiness is the lowest, signifying the highest risk.
* **loan\_amnt**: The loan amount requested by the individual.
* **loan\_int\_rate**: The interest rate associated with the loan.
* **loan\_status**: Loan status, where 0 indicates non-default and 1 indicates default.
  + **0**: Non-default - The borrower successfully repaid the loan as agreed, and there was no default.
  + **1**: Default - The borrower failed to repay the loan according to the agreed-upon terms and defaulted on the loan.
* **loan\_percent\_income**: The percentage of income represented by the loan amount.
* **cb\_person\_default\_on\_file**: Historical default of the individual as per credit bureau records.
  + **Y**: The individual has a history of defaults on their credit file.
  + **N**: The individual does not have any history of defaults.
* **cb\_preson\_cred\_hist\_length**: The length of credit history for the individual.

## Detailed Analysis Report

1. **Loan Intent and Average Person Age**: This table shows the average age of individuals applying for different types of loans. The average age varies based on the loan intent, with ‘Home Improvement’ loans being taken by the oldest individuals (average age of 29 years), and ‘Education’ loans being taken by the youngest (average age of 26.6 years).
2. **Loan Intent and Repayment Status**: This table provides a breakdown of the total number of loans, defaults, and repayments based on the loan intent. ‘Venture’ loans have the highest total number of loans (5719), with 847 defaults and 4872 repaid. ‘Debt Consolidation’ loans have the lowest total number of loans (5212), with 1490 defaults and 3722 repaid.
3. **Interest Rate and Repayment Status**: This table shows the percentage of defaults and repayments across different interest rate ranges. The default percentage increases with the interest rate. Loans with an interest rate of 21-25% have the highest default percentage (84%), while loans with an interest rate of 5-10% have the lowest default percentage (16.82%).
4. **Home Ownership and Repayment Status**: This table shows the percentage of defaults and repayments based on the type of home ownership. Individuals who rent their homes have the highest default percentage (31.57%), while those who own their homes outright have the lowest default percentage (7.47%).
5. **Loan Amount and Repayment Status**: This table shows the percentage of total repaid loans across different loan amount ranges. The majority of repaid loans fall within the loan amount range of 5001-10000 (38.77% of total repaid), while the least repaid loans fall within the loan amount range of 30001-35000 (0.57% of total repaid).
6. **Credit History and Repayment Status**: This table shows the percentage of total repaid loans across different credit history lengths. Individuals with a credit history length of 2-6 years have the highest percentage of total repaid (56.53% for those with no default history and 9.21% for those with a default history).
7. **Income Range, Loan Intent, and Repayment Status**: This table shows the percentage of total defaults and repayments for different loan intents across various income ranges. For individuals with an income range of 4,000 - 99,999 and no default history, ‘Education’ loans have the highest percentage of total repaid (15.17%), while ‘Debt Consolidation’ loans have the lowest percentage of total repaid (10.83%).

This analysis provides insights into various factors that influence loan repayment. It’s important to note that correlation does not imply causation, and these factors should be considered as part of a broader context when assessing loan repayment probabilities.

## Model Performance Analysis

Here’s a detailed analysis of the performance of the three models:

* **Linear Regression Model**: This model has an accuracy of 86.05%, meaning it correctly predicted the loan status for 86.05% of the cases in your dataset. However, its sensitivity (true positive rate) is 94.91%, and its specificity (true negative rate) is 54.32%. This means it’s more reliable in predicting non-defaulters than defaulters.
* **Decision Tree Model**: This model has an accuracy of 85.63%. It has a sensitivity of 100%, meaning it correctly identified all the non-defaulters. However, its specificity is 34.15%, indicating it’s less reliable in correctly identifying defaulters.
* **SVM Model**: The SVM model has the highest accuracy of 90.3%. It also has a high sensitivity of 98.19% and a reasonable specificity of 62.01%. This makes it the most balanced model among the three.

### Tabular Comparison

**Table**

| **Model** | **Accuracy** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- |
| Linear Regression | 86.05% | 94.91% | 54.32% |
| Decision Tree | 85.63% | 100% | 34.15% |
| SVM | 90.3% | 98.19% | 62.01% |

## Predicted Loan Status for New Record

The predicted loan status for the new record using the three models is as follows:

**R**

new\_record <- data.frame(

person\_age = 24,

person\_income = 159000,

person\_home\_ownership = "RENT",

person\_emp\_length = 123,

loan\_intent = "EDUCATION",

loan\_grade = "B",

loan\_amnt = 65000,

loan\_int\_rate = 14.02,

loan\_percent\_income = 0.40,

cb\_person\_default\_on\_file = "Y",

cb\_person\_cred\_hist\_length = 5

)

* **Linear Regression**: 0 (No Default)
* **Decision Tree**: 1 (Default)
* **SVM**: 0 (No Default)

## Conclusion

Based on the performance metrics and the balance between sensitivity and specificity, the SVM model appears to be the best model to use for loan prediction in this case. It not only has the highest accuracy but also maintains a good balance in correctly identifying both defaulters and non-defaulters.