Indian Institute of Technology, Jodhpur

Applied Analytics for Predictive Lending: Classifying Loan Approvals and Predicting Risk Scores

MSL71440 Analytics Lab (Python) (Autumn, 2024)

Name:- Aman Kanshotia, Sahil, Abhas Malguri
Roll No.:- M21MA201, M21MA210, M23MA1001

M.Sc - M.Tech (Data and Computational Sciences)
M.Sc Mathematics

Department of Mathematics

Contents

List of Figures

List of Tables

1 Data Overview

A comprehensive overview of the demographic and financial details included in the dataset, along with credit information, loan application specifics, and risk indicators.

1.1 Demographic Details

- **Age:** Ranges from 18 to 80 years, with an average of around 40 years.
- **Gender:** Categories include Male, Female, and possibly others.
- **Marital Status:** Comprises categories such as Married, Single, Divorced, and Widowed.
- **Dependents:** Represents the number of family dependents, indicating financial responsibilities.
- **Education Level:** Applicant's education level, such as High School, Bachelor's, Master's, or PhD.
- **Employment Status:** Employment type, such as Employed, Unemployed, Self-employed, or Retired.

1.2 Financial Details

- **Annual Income:** Ranges widely, with a mean around 59,161.
- **Savings Account Balance and Checking Account Balance:** Indicators of financial stability.
- **Monthly Debt Payments:** The applicant's monthly debt obligations.
- **Debt to Income Ratio and Total Debt to Income Ratio:** Metrics for assessing financial health.

1.3 Credit Information

- **Credit Score:** Scores range between 343 and 712, with an average around 572.
- **Length of Credit History:** The duration for which the applicant has managed credit.
- **Number of Open Credit Lines:** Number of active credit accounts.
- **Payment History:** Reflects the applicant's past loan repayment behavior.
- **Number of Previous Loans:** Shows the applicant's prior borrowing experience.

1.4 Loan Application Details

- **Loan Amount:** Requested loan principal amount.
- **Loan Duration:** Loan term length in months, from 12 to 120 months.
- **Purpose of Loan:** Intended use of the loan, such as Education, Business, or Home Improvement.
- **Interest Rate and Monthly Loan Payment:** Loan costs, indicating the affordability for the applicant.

1.5 Approval and Risk

- **Loan Approved:** A binary variable (0 or 1) showing the outcome of the loan application.
- **Risk Score:** Ranges from 29 to 84, providing a risk assessment based on credit and financial data.

This overview captures the key demographic and financial characteristics that contribute to loan approval predictions and risk analysis.

2 Descriptive Analysis for the dataset

2.1 Dataset Structure Summary

Column Name	Data Type	Non-Null Count
Age	int64	20000
Annual Income	int64	20000
Credit Score	int64	20000
Employment Status	object	20000
Education Level	object	20000
Experience	int64	20000
Loan Amount	int64	20000
Loan Duration	int64	20000
Marital Status	object	20000
Number of Dependents	int64	20000
Home Ownership Status	object	20000
Monthly Debt Payments	int64	20000
Credit Card Utilization Rate	float64	20000
Number of Open Credit Lines	int64	20000
Number of Credit Inquiries	int64	20000
Debt to Income Ratio	float64	20000
Bankruptcy History	int64	20000
Loan Purpose	object	20000
Previous Loan Defaults	int64	20000
Payment History	int64	20000
Length of Credit History	int64	20000
Savings Account Balance	int64	20000
Checking Account Balance	int64	20000
Total Assets	int64	20000
Total Liabilities	int64	20000
Monthly Income	float64	20000
Utility Bills Payment History	float64	20000
Job Tenure	int64	20000
Net Worth	int64	20000
Base Interest Rate	float64	20000
Interest Rate	float64	20000
Monthly Loan Payment	float64	20000
Total Debt to Income Ratio	float64	20000
Loan Approved	int64	20000
Risk Score	float64	20000

Table 1: Summary of the Dataset Columns

2.2 Summary Statistics of the Dataset

Column Name	Mean	Std	Min	Max
Age	39.75	11.62	18.00	80.00
Annual Income	59161.47	40350.85	15000.00	485341.00
Credit Score	571.61	50.99	343.00	712.00
Experience	17.52	11.32	0.00	61.00
Loan Amount	24882.87	13427.42	3674.00	184732.00
Loan Duration	54.06	24.66	12.00	120.00
Number of Dependents	1.52	1.39	0.00	5.00
Monthly Debt Payments	454.29	240.51	50.00	2919.00
Credit Card Utilization Rate	0.29	0.16	0.00	0.92
Number of Open Credit Lines	3.02	1.74	0.00	13.00
Number of Credit Inquiries	0.99	0.99	0.00	7.00
Debt to Income Ratio	0.29	0.16	0.00	0.90
Bankruptcy History	0.05	0.22	0.00	1.00
Previous Loan Defaults	0.10	0.30	0.00	1.00
Payment History	23.99	4.95	8.00	45.00
Length of Credit History	14.96	8.37	1.00	29.00
Savings Account Balance	4946.05	6604.89	73.00	200089.00
Checking Account Balance	1782.56	2245.38	24.00	52572.00
Total Assets	96964.40	120799.94	2098.00	2619627.00
Total Liabilities	36252.41	47251.51	372.00	1417302.00
Monthly Income	4891.72	3296.77	1250.00	25000.00
Utility Bills Payment History	0.80	0.12	0.26	1.00
Job Tenure	5.00	2.24	0.00	16.00
Net Worth	72294.32	117920.02	1000.00	2603208.00
Base Interest Rate	0.24	0.04	0.13	0.41
Interest Rate	0.24	0.04	0.11	0.45
Monthly Loan Payment	911.61	674.58	97.03	10892.63
Total Debt to Income Ratio	0.40	0.34	0.02	4.65
Loan Approved	0.24	0.43	0.00	1.00
Risk Score	50.77	7.78	28.80	84.00

Table 2: Summary Statistics: Mean, Standard Deviation, Minimum, and Maximum

Column Name	25%	50%	75%
Age	32.00	40.00	48.00
Annual Income	31679.00	48566.00	74391.00
Credit Score	540.00	578.00	609.00
Experience	9.00	17.00	25.00
Loan Amount	15575.00	21914.50	30835.00
Loan Duration	36.00	48.00	72.00
Number of Dependents	0.00	1.00	2.00
Monthly Debt Payments	286.00	402.00	564.00
Credit Card Utilization Rate	0.16	0.27	0.39
Number of Open Credit Lines	2.00	3.00	4.00
Number of Credit Inquiries	0.00	1.00	2.00
Debt to Income Ratio	0.16	0.26	0.39
Bankruptcy History	0.00	0.00	0.00
Previous Loan Defaults	0.00	0.00	0.00
Payment History	21.00	24.00	27.00
Length of Credit History	8.00	15.00	22.00
Savings Account Balance	1541.75	2986.00	5873.25
Checking Account Balance	551.00	1116.00	2126.00
Total Assets	31180.25	60699.00	117405.25
Total Liabilities	11196.75	22203.00	43146.50
Monthly Income	2629.58	4034.75	6163.00
Utility Bills Payment History	0.73	0.82	0.89
Job Tenure	3.00	5.00	6.00
Net Worth	8734.75	32855.50	88825.50
Base Interest Rate	0.21	0.24	0.26
Interest Rate	0.21	0.24	0.27
Monthly Loan Payment	493.76	728.51	1112.77
Total Debt to Income Ratio	0.18	0.30	0.51
Loan Approved	0.00	0.00	0.00
Risk Score	46.00	52.00	56.00

Table 3: Summary Statistics: 25th, 50th, and 75th Percentiles

2.3 Summary of Categorical Variables

Column Name	Count	Unique	Top	Frequency
Employment Status	20000	3	Employed	17036
Education Level	20000	5	Bachelor	6054
Marital Status	20000	4	Married	10041
Home Ownership Status	20000	4	Mortgage	7939
Loan Purpose	20000	5	Home	5925

Table 4: Summary of Categorical Variables

The table above provides an overview of key categorical variables in the dataset. The count of each variable, the number of distinct values it contains, the most prevalent (or "top") category, and the frequency of this category are all combined to form a summary.

For example, among the employment statuses, the category "employed" is predominant, representing 17,036 entries. Similarly, the "Bachelor" level of education is the most common, occurring 6,054 times. The marital status distribution shows "married" as the most frequent category, with a count of 10,041. In terms of home ownership, "mortgage" is the leading status, appearing 7,939 times. Finally, the table indicates that "home" is the primary purpose for loans, representing 5,925 entries.

This summary helps to understand the categorical makeup of the data, shedding light on the dominant patterns in employment, education, marital status, home ownership, and loan purpose within the dataset.

3 Distribution Analysis

3.1 Analysis of Employment Status and Its Impact on Loan Approval Decisions

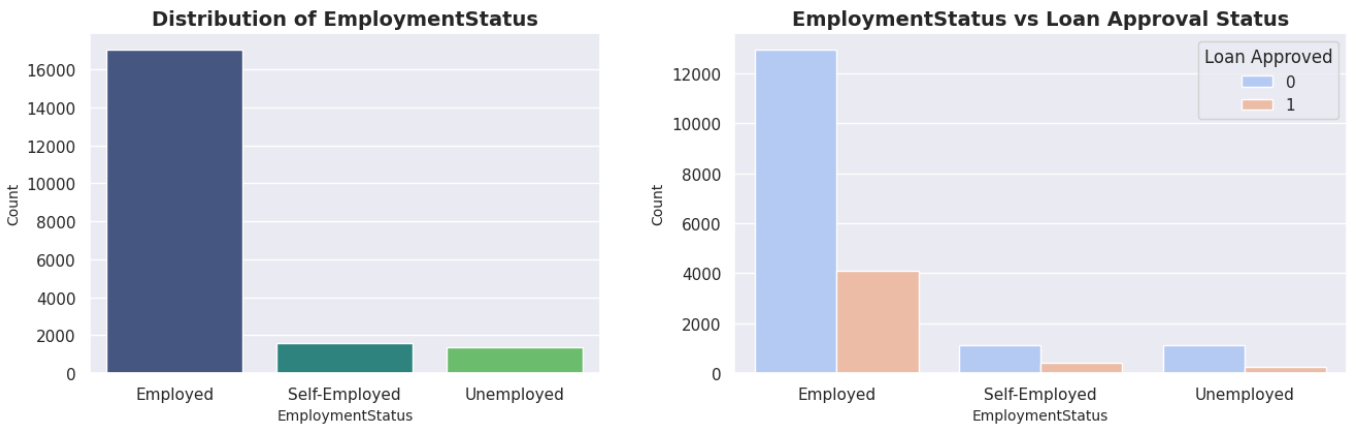

Figure 1: Employment Status and its impact on approval of loan

This visualization consists of two bar plots, each providing insights into the `EmploymentStatus` of individuals in relation to `LoanApprovalStatus`:

- **Left Plot - Distribution of Employment Status:**

  This bar plot displays the distribution of `EmploymentStatus` categories (*Employed*, *Self-Employed*, *Unemployed*) in the dataset. The vast majority of individuals are *Employed*, while much smaller groups are *Self-Employed* or *Unemployed*.

- **Right Plot - Employment Status vs Loan Approval Status:**

  This bar plot shows the breakdown of `LoanApprovalStatus` (approved or not approved) within each `EmploymentStatus` category. *Employed* individuals have a significantly higher count, both in terms of those approved and not approved for loans. However, a larger portion of employed applicants are not approved compared to those approved. For *Self-Employed* and *Unemployed* individuals, the count is much lower, and a similar pattern is observed, with more loans not approved.

3.2 Analysis of Educational Level and Its Impact on Loan Approval Decisions

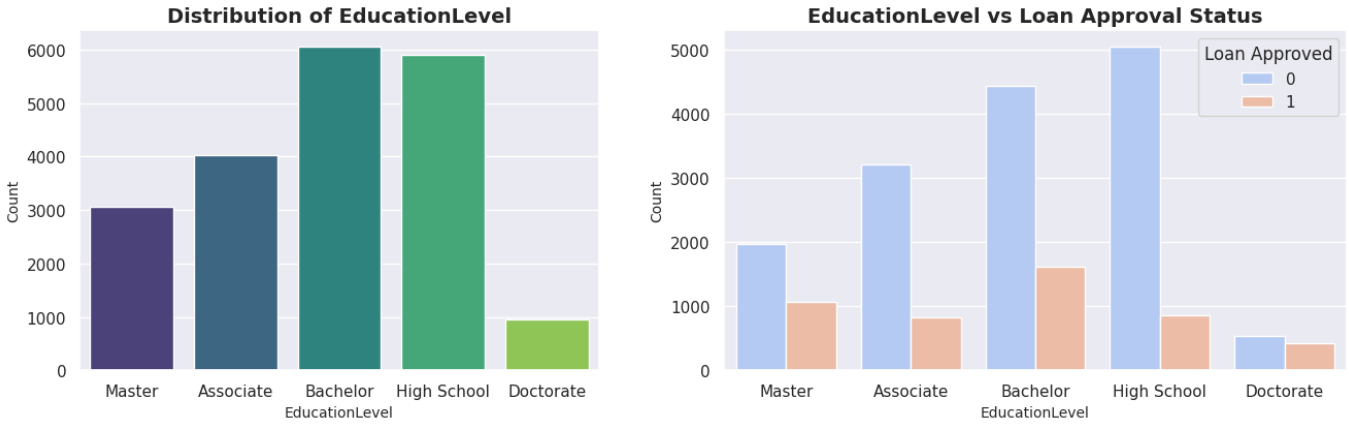

Figure 2: Educational Level and its impact on approval of loan

This visualization consists of two bar plots, each providing insights into the `EducationLevel` of individuals in relation to `LoanApprovalStatus`:

- **Left Plot - Distribution of Education Level:**

  This bar plot displays the distribution of `EducationLevel` categories (*Master*, *Associate*, *Bachelor*, *High School*, *Doctorate*) in the dataset. The largest groups are those with a *Bachelor's* and *High School* education, followed by *Associate* degree holders. Individuals with a *Doctorate* degree are the smallest group.

- **Right Plot - Education Level vs Loan Approval Status:**

  This bar plot shows the breakdown of `LoanApprovalStatus` (approved or not approved) within each `EducationLevel` category. Across all education levels, the count of individuals with loans not approved is higher than those approved. Individuals with a *Bachelor's* and *High School* education have the highest counts in terms of loan applications, with more applications in each category resulting in non-approval than approval.

3.3 Analysis of Marital Status and Its Impact on Loan Approval Decisions

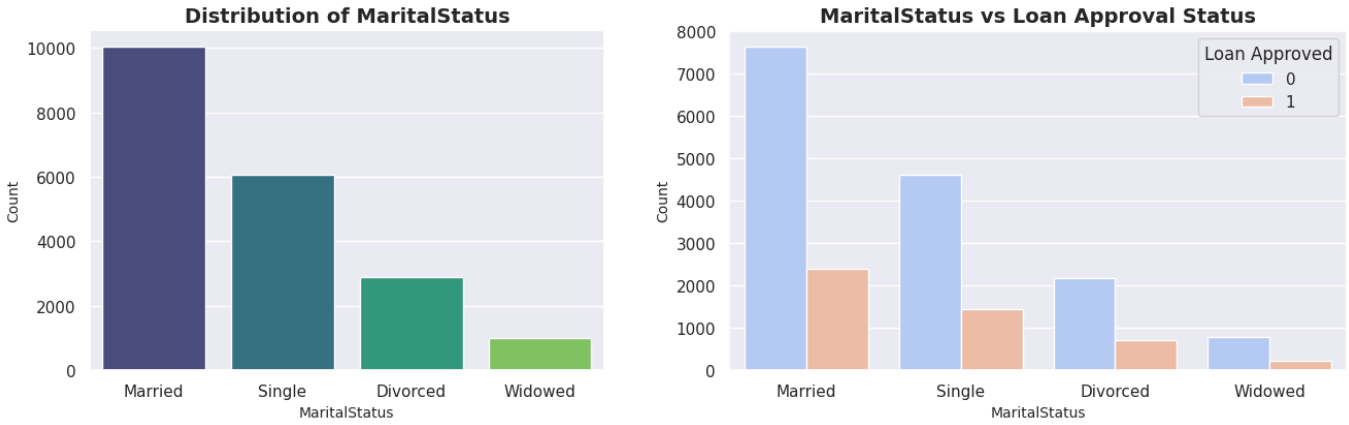

Figure 3: Marital Status and its impact on approval of loan

- **Left Plot: Distribution of Marital Status**
  1. The plot shows the count of individuals across four marital status categories: Married, Single, Divorced, and Widowed.
  2. The majority of individuals are married, followed by single, divorced, and widowed.
- **Right Plot: Marital Status vs Loan Approval Status**
  1. The plot shows the relationship between marital status and loan approval status.
  2. Loan approval status is indicated by two colors:
     (a) **Blue (0)**: Loan Not Approved
     (b) **Orange (1)**: Loan Approved
  3. Married individuals have the highest number of loan approvals compared to other groups.
  4. Single, divorced, and widowed individuals show fewer loan approvals, with a larger proportion of rejections.

3.4 Analysis of Home ownership Status and Its Impact on Loan Approval Decisions

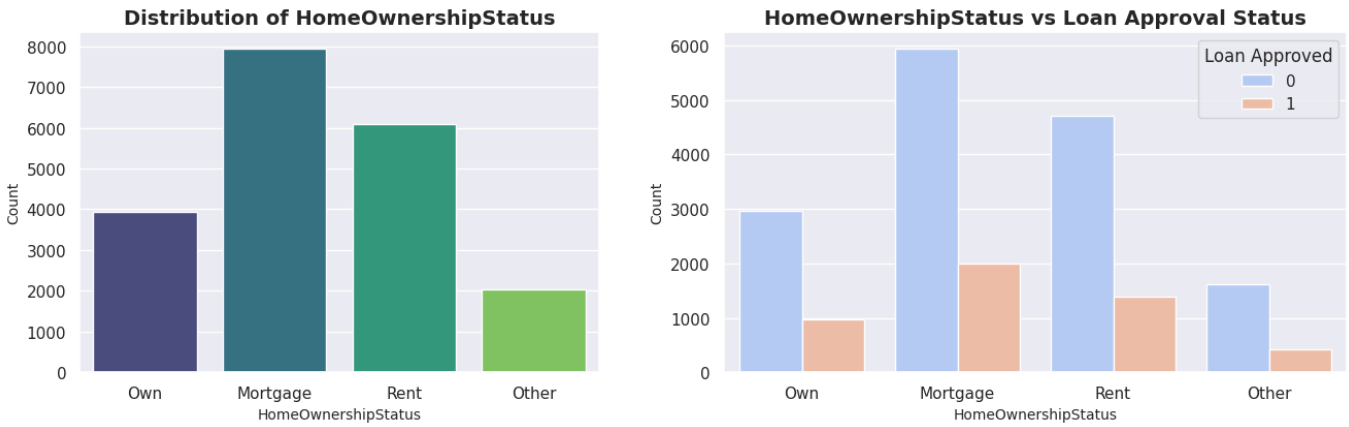

Figure 4: Home ownership Status and its impact on approval of loan

- **Left Plot: Distribution of Home Ownership Status**
  1. The plot shows the count of individuals across four home ownership categories: Own, Mortgage, Rent, and Other.
  2. The largest group is individuals with a mortgage, followed by those who rent.
  3. The smallest group falls under the ”Other” category.
- **Right Plot: Home Ownership Status vs Loan Approval Status**
  1. This plot illustrates the relationship between home ownership status and loan approval.
  2. Loan approval status is represented by two colors:
     (a) **Blue (0)**: Loan Not Approved
     (b) **Orange (1)**: Loan Approved
  3. Individuals with a mortgage or rent are more likely to apply for loans, though a significant proportion of these applications are not approved.
  4. Those who own their homes outright have fewer applications but show a higher proportion of approvals relative to rejections.
  5. The ”Other” category shows the fewest loan applications, with a very small number of approvals.

3.5 Analysis of Loan Purpose and Its Impact on Loan Approval Decisions

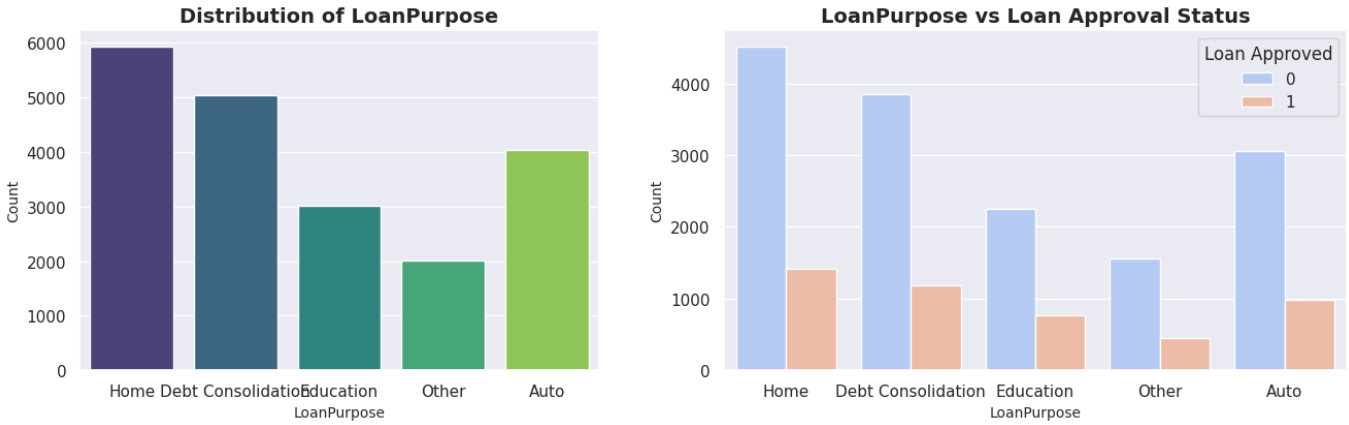

Figure 5: Loan Purpose and its impact on approval of loan

- **Left Plot: Distribution of Loan Purpose**
  1. The plot shows the count of loans requested for different purposes: Home, Debt Consolidation, Education, Other, and Auto.
  2. The most common purpose is for **Home**, followed closely by **Debt Consolidation**.
  3. Fewer loans are requested for **Education**, **Auto**, and **Other** purposes.
- **Right Plot: Loan Purpose vs Loan Approval Status**
  1. This plot shows how loan approval varies based on loan purpose.
  2. Loan approval status is represented by:
     (a) **Blue (0)**: Loan Not Approved
     (b) **Orange (1)**: Loan Approved
  3. Loans for **Home** and **Debt Consolidation** show high application counts, but with a relatively high number of rejections.
  4. **Auto** loans have a smaller number of applications, with more approvals compared to other purposes.
  5. Loans for **Education** and **Other** purposes have fewer approvals, with most applications getting rejected.

3.6 Loan Approval Distribution

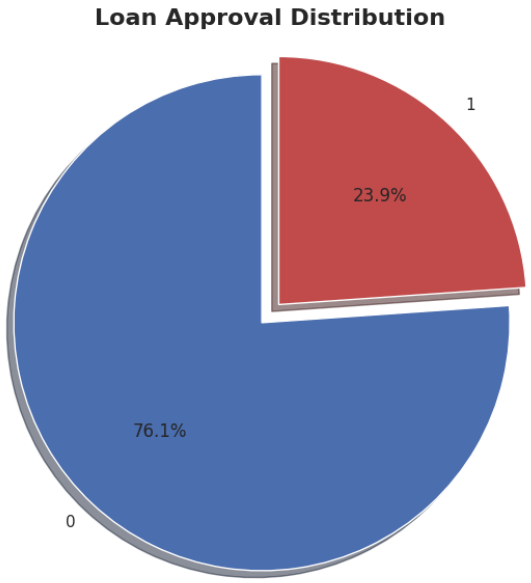

Figure 6: Loan Approval Distribution

Loan Approved	Count
0 (Not Approved)	15,220
1 (Approved)	4,780

Table 5: Summary of Loan Approval Counts

The total number of loans considered is $20,000$, indicating that a significant majority of the loans (approximately 76%) were not approved, while around 24% received approval. This distribution can provide insights into the lending policies and borrower qualifications within the dataset.

3.7 Analysis of Annual Income Distribution by Education Level

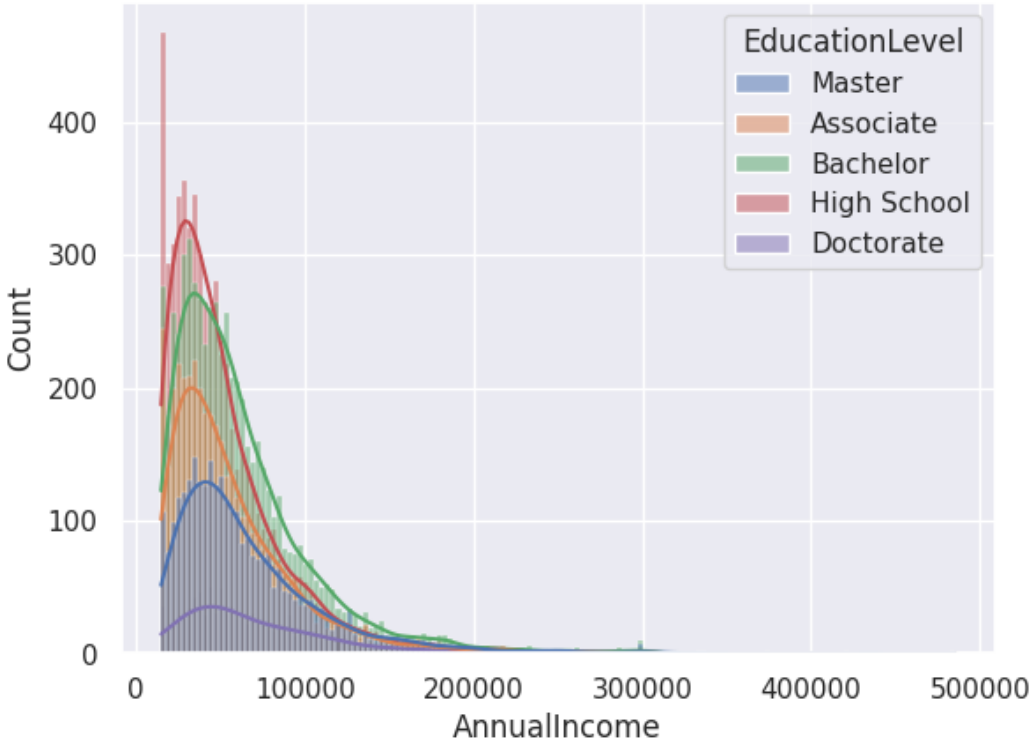

Figure 7: Annual Income Distribution

- The income distribution across all education levels is positively skewed, with most individuals earning less than $100,000 annually.
- Individuals with a Doctorate degree tend to have the highest incomes, although they represent a smaller portion of the sample.
- High school graduates exhibit the lowest income levels, with a large concentration in the lower-income range.
- Bachelor's and Master's degree holders show similar income distributions, with some overlap, indicating that a Master's degree does not always lead to significantly higher earnings.
- The income distribution for Associate degree holders aligns closely with that of Bachelor's degree holders but skews slightly toward the lower end.
- The KDE curves provide a smooth visualization of income trends, revealing that all education levels follow a similar pattern: a peak at lower incomes followed by a gradual decline.
- While higher education generally correlates with higher income, the overlap between adjacent education levels suggests that other factors beyond education may also influence earning potential.

3.8 Loan Amount Distribution by Education Level

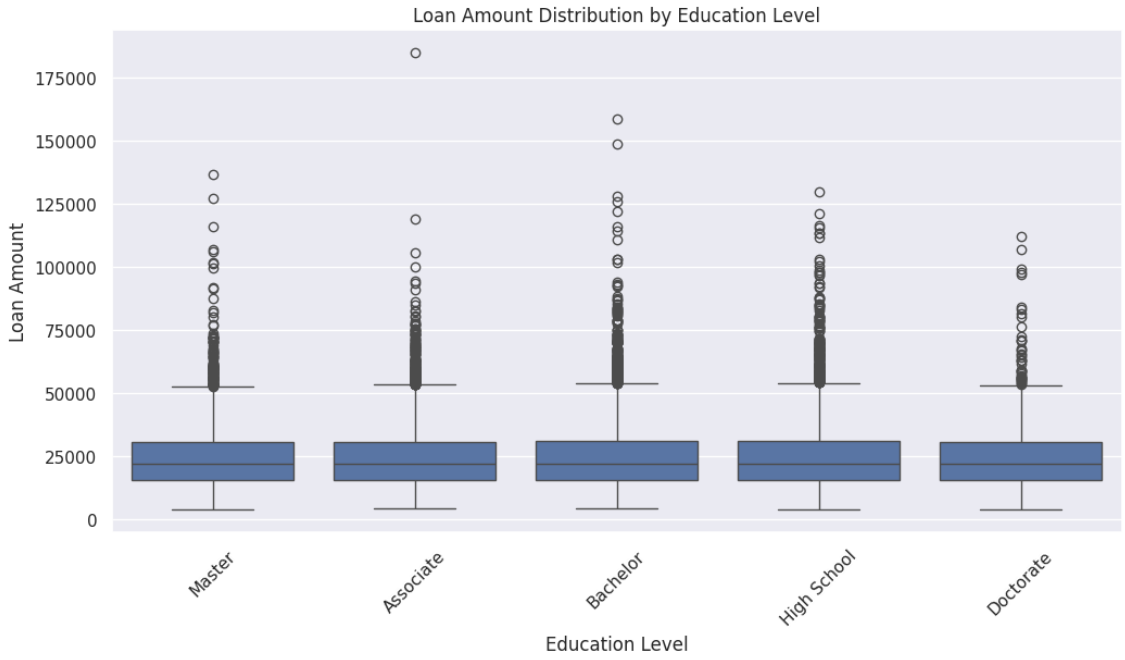

Figure 8: Loan Amount Distribution

- **Distribution of Loan Amounts**: The boxplot shows the distribution of loan amounts segmented by education level: Master, Associate, Bachelor, High School, and Doctorate.
- **Median Loan Amounts**: The median loan amounts are relatively similar across all education levels, suggesting that education level may not significantly impact the typical loan size.
- **Range of Loan Amounts**: The interquartile range (IQR) is consistent across all education levels, indicating that the majority of individuals, regardless of education, take loans of similar sizes.
- **Presence of Outliers**: Each education level has many outliers, with some loans exceeding 100,000. These outliers reflect a small subset of individuals taking exceptionally large loans.
- **Maximum Loan Values**: Doctorate holders and high school graduates show the largest outliers, with some loans exceeding $175,000$. However, such large loans are rare.
- **Upper Whisker Variation**: The upper whiskers (maximum values excluding outliers) extend slightly higher for individuals with higher education, especially for those with Master's and Doctorate degrees.
- **Lower Whiskers Consistency**: The lower whiskers (minimum loan values) are similar across all education levels, indicating that the smallest loans are generally consistent.

- **Conclusion**: While education level may influence the upper extremes of loan amounts, the typical loan size remains comparable across different educational backgrounds, suggesting that loan decisions are influenced more by individual financial needs or other factors beyond education.

3.9 Credit Score Distribution by Employment Status

Figure 9: Credit Score Distribution

- **Distribution of Credit Scores**: The violin plot shows the distribution of credit scores across three categories of employment status: Employed, Self-Employed, and Unemployed.
- **Central Tendency**: All three employment groups have similar median credit scores, indicated by the white dot in the center of each distribution.
- **Range of Credit Scores**: The spread of credit scores is similar across the groups, with most values lying between 450 and 650.
- **Peaks in Distribution**: Each distribution exhibits a density peak around the median score, suggesting that the majority of individuals in all employment categories have comparable credit scores.
- **Presence of Extreme Values**: Credit scores for each category show a few individuals with values extending towards the lower and upper extremes, around 300 and 700, respectively.

- **Comparing Employment Types**: There is no significant difference in the overall shape or distribution of credit scores between employed, self-employed, and unemployed individuals, indicating that employment status may not have a substantial impact on credit score distribution.
- **Symmetry of Distributions**: All three distributions appear symmetric, indicating that credit scores are fairly evenly distributed within each employment group without significant skewness.
- **Conclusion**: The similarity in distributions across employment statuses suggests that credit score is influenced by other factors beyond employment type, such as financial behavior or credit history.

3.10 Distribution of Loan Amount by Employment Status and Education Level

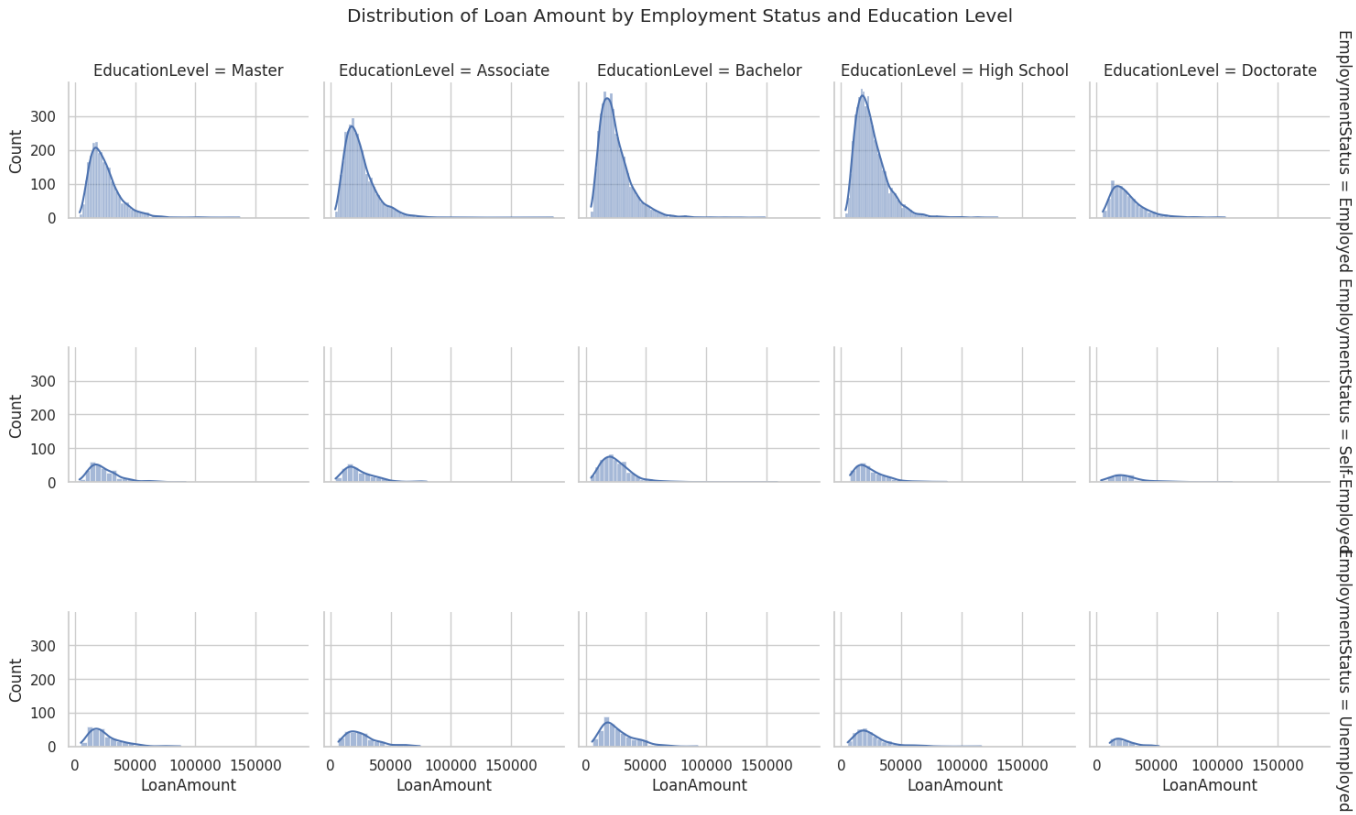

Figure 10: Loan Amount by Employment Status and Education Level

- **Overall Distribution:**
  1. Most loan amounts are concentrated at the lower end across all education levels and employment statuses.

2. The distributions show a long tail, indicating a few higher loan amounts.

- **Effect of Education Level:**

1. All education levels (Master, Bachelor, Associate, High School, and Doctorate) show a similar pattern with most loans under approximately 50,000.
2. The frequency of loans is highest for Bachelor's and High School graduates.
3. The distribution for individuals with a Doctorate is more spread out with fewer observations.

- **Effect of Employment Status:**

1. Employed individuals have the highest concentration of loans compared to self-employed and unemployed individuals.
2. Self-employed individuals also take smaller loans but with slightly broader variance than the employed group.
3. Unemployed individuals have a much lower frequency of loans, suggesting limited access or demand.

- **Joint Effect of Education and Employment:**

1. Across all employment statuses, Bachelor's and High School graduates exhibit the highest number of loans.
2. Doctorate holders, regardless of employment status, tend to take fewer loans, indicating possibly lower demand or better financial stability.
3. Loan amounts across employment types are similarly concentrated below 50,000, with occasional outliers reaching above 100,000.

4 Preprocessing and Splitting of the data using Numpy and Pandas and creation of Pipeline

The function `preprocessor` takes a DataFrame and processes it for machine learning tasks. The key steps involved in the function are outlined below.

```
def preprocessor(df, target_column='LoanApproved', test_size=0.2, random_state
    =42):
     ...
```

4.1 Parameters

The function accepts the following parameters:

- **df**: The input DataFrame containing features and the target variable.
- **Target column**: The name of the target variable, defaulting to `'LoanApproved'`.
- **Test size**: The proportion of the data to be allocated to the test set (default is 0.2).
- **Random State**: An integer used for reproducibility in data splitting (default is 42).

4.2 Data Preparation

The function begins by separating the features from the target variable:

```
X = df.drop(columns=[target_column], axis=1)
y = df[target_column]
```

4.3 Column Identification

It identifies numerical and categorical columns using:

```
num_column = X.select_dtypes(['int64', 'float64']).columns
cat_column = X.select_dtypes('object').columns
```

4.4 Train-Test Split

The dataset is split into training and test sets:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size,
    random_state=random_state)
```

4.5 Preprocessing Steps

The function defines a preprocessing pipeline using `ColumnTransformer` to apply different transformations:

```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), num_column),
        ('cat', OneHotEncoder(handle_unknown='ignore', drop='first'), cat_column
           )
    ]
)
```

In this step:

- `StandardScaler` standardizes numerical features.
- `OneHotEncoder` converts categorical variables into a format suitable for machine learning.

4.6 Pipeline Creation

The preprocessing steps are encapsulated in a pipeline:

```
preprocessing_pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor)
])
```

4.7 Data Transformation

Finally, the function fits the pipeline to the training data and transforms both the training and test sets:

```
X_train_transformed = preprocessing_pipeline.fit_transform(X_train)
X_test_transformed = preprocessing_pipeline.transform(X_test)
```

4.8 Conclusion

The `preprocessor` function efficiently prepares the dataset for machine learning by separating features and target variables, scaling numerical data, encoding categorical variables, and splitting the data into training and test sets. This structured approach facilitates effective modeling and ensures the data is in the correct format.

5 Analysis of Model Training Code

This section provides a detailed analysis of the Python code for training multiple classifiers to predict loan approval status. The analysis was broken down into manageable blocks, explaining each component's purpose.

5.1 Model Imports

The necessary libraries for building the models and calculating evaluation metrics are imported:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, precision_score, recall_score,
    f1_score
from sklearn.model_selection import cross_val_score
```

- This block imports classifiers from the `sklearn` library.
- Additionally, it imports functions to compute performance metrics and cross-validation scores.

5.2 Model Initialization

A dictionary is created to hold the different classifiers:

```
models = {
    'Naive Bayes': GaussianNB(),
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'KNN': KNeighborsClassifier(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier()
}
```

- Each classifier is instantiated and added to the `models` dictionary with a descriptive key.
- The models include Naive Bayes, Logistic Regression, KNN, Decision Tree, and Random Forest.

5.3 Results Storage Initialization

An empty dictionary to store the results of model evaluations is defined:

```
results = {
    'Model': [],
    'Accuracy': [],
    'AVG Cross Accuracy': [],
    'Precision': [],
    'Recall': [],
    'F1-Score': []
}
```

- This dictionary will later be populated with the evaluation metrics for each model.

5.4 Training Loop

The main loop for training the models begins:

```
for model in models.items():
    print(f'Training␣{model[0]}...')
    clf = model[1]
    clf.fit(X_train_transformed, y_train)
    y_pred = clf.predict(X_test_transformed)
```

- The loop iterates over the `models` dictionary, training each model on the transformed training data.
- The model's predictions on the test set are stored in `y_pred`.

5.5 Model Evaluation

Evaluation metrics are calculated for each model:

```
    accuracy = accuracy_score(y_test, y_pred)
    cross_accuracy = cross_val_score(clf, X_train_transformed, y_train, cv=5,
       scoring='accuracy')
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
```

- Accuracy: Computed using the `accuracy_score` function to evaluate overall performance.
- Cross-Validation Accuracy: Obtained from 5-fold cross-validation for a robust estimate of model performance.
- Precision: Calculated to assess the accuracy of positive predictions.
- Recall: Measures the ability of the model to identify actual positives.
- F1-Score: Provides a balance between precision and recall.

5.6 Storing Results

Metrics are appended to the results dictionary:

```
    results['Model'].append(model[0])
    results['Accuracy'].append(accuracy)
    results['AVG␣Cross␣Accuracy'].append(cross_accuracy.mean())
    results['Precision'].append(precision)
    results['Recall'].append(recall)
    results['F1-Score'].append(f1)
    print(f'Training␣{model[0]}␣COMPLETED')
```

- After each model's evaluation, the results are appended to the respective lists in the `results` dictionary.
- Progress messages are printed to the console for tracking purposes.

5.7 Creating Results Data frame

Finally, the results are compiled into a Data frame:

```
results_df = pd.DataFrame(results)
results_df
```

- The `results` dictionary is converted into a pandas Data frame for easier visualization and analysis.

6 Model Performance Summary

6.1 Classification Results

Model	Accuracy	AVG Cross Accuracy	Precision	Recall	F1-Score
Naive Bayes	95.35%	95.45%	87.67%	95.08%	91.23%
Logistic Regression	99.98%	99.96%	99.90%	100.00%	99.95%
KNN	95.98%	95.74%	97.98%	85.94%	91.57%
Decision Tree	99.00%	98.94%	98.13%	97.94%	98.03%
Random Forest	99.10%	98.82%	98.52%	97.94%	98.22%

Table 6: Performance metrics of various models

- Logistic Regression achieved the highest overall accuracy (99.98%) and perfect recall (100%), making it the most reliable model for loan approval predictions.
- Random Forest and Decision Tree models demonstrated high accuracy, indicating strong performance in handling loan approval classification tasks.
- Naive Bayes exhibited a lower precision (87.67%), suggesting it may produce more false positives compared to the other models.
- KNN had a decent accuracy of 95.98%, but its recall (85.94%) indicates it might miss some actual approvals, which could be critical in financial applications.
- The average cross-validation accuracy for all models suggests consistent performance across different subsets of the data.
- The F1-scores indicate that both Random Forest and Decision Tree maintain a good balance between precision and recall, making them robust choices.
- Overall, models with high precision and recall, such as Logistic Regression and Random Forest, are preferable for applications where false negatives are costly.

6.2 Classification Model Comparison

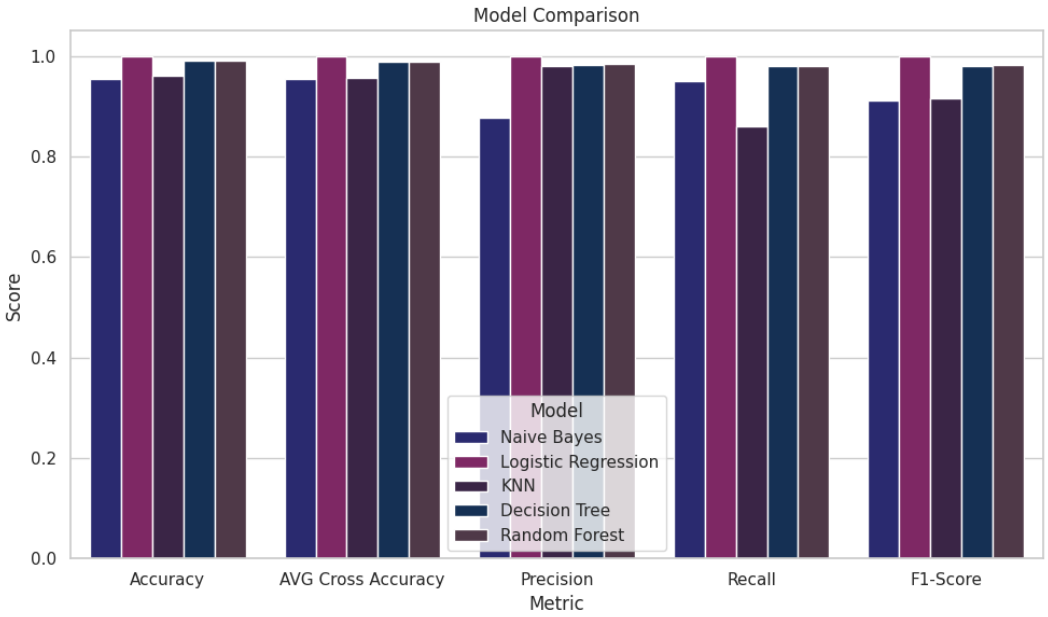

Figure 11: Model Comparison

6.3 AUC- ROC Curve for the models

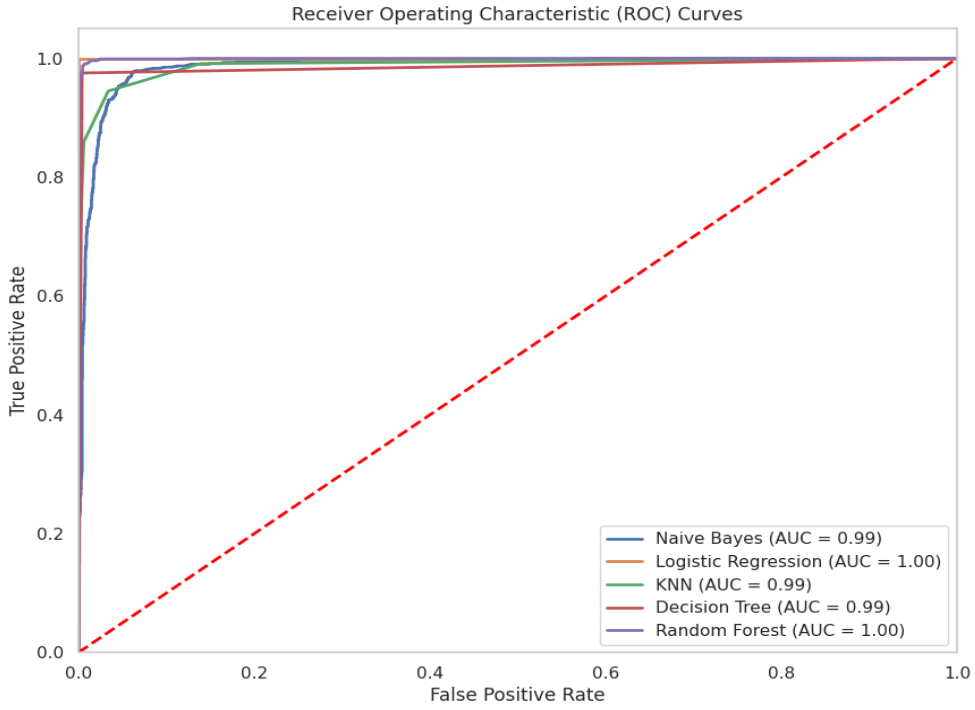

Figure 12: AUC- ROC Curves

6.4 Regression Results

Model	Train MAE	Validation MAE	Train MSE	Validation MSE	Train R2 Score	Validation R2 Score
Linear Regression	0.120340	0.120137	0.023143	0.022603	0.871535	0.873569
Ridge Regression	0.120347	0.120153	0.023143	0.022606	0.871533	0.873551
KNN Regressor	0.062734	0.079688	0.023791	0.034238	0.867938	0.808491
Decision Tree Regressor	0.000000	0.009375	0.000000	0.009375	1.000000	0.947561
Random Forest Regressor	0.005198	0.013772	0.000734	0.005089	0.995926	0.971535
SVR	0.060047	0.076004	0.005041	0.010635	0.972016	0.940510
XGBoost Regressor	0.002495	0.014341	0.000024	0.004595	0.999866	0.974299

Table 7: Performance metrics of various regression models

- **XGBoost Regressor** achieved the lowest validation MSE (0.0046) and highest validation R2 score (0.9743), indicating it is the most reliable model for predicting loan approval outcomes.
- **Random Forest Regressor** also performed well, with a high validation R2 score (0.9715) and low validation MSE (0.0051), suggesting strong capability in capturing complex patterns in the data.
- **Decision Tree Regressor** achieved a perfect train R2 score (1.0000) but had a slightly lower validation R2 score (0.9476), indicating it may be overfitting on the training data.
- **Linear Regression** and **Ridge Regression** exhibited similar performance, with validation R2 scores around 0.8735, indicating moderate predictive power without overfitting.
- **KNN Regressor** demonstrated lower performance, with a validation R2 score of 0.8085, suggesting it may struggle with generalizing across the dataset.
- **SVR** achieved a high validation R2 score (0.9405) with a relatively low validation MAE, indicating it maintains good balance in predictive accuracy while handling noise effectively.
- The comparison of validation R2 scores and MAEs across models highlights that **XGBoost Regressor** and **Random Forest Regressor** are preferable for this regression task, given their strong performance in both training and validation metrics.
- Overall, models like **XGBoost Regressor** and **Random Forest Regressor** are well-suited for financial applications where accurate predictions are crucial.

6.5 Regression Model Comparison

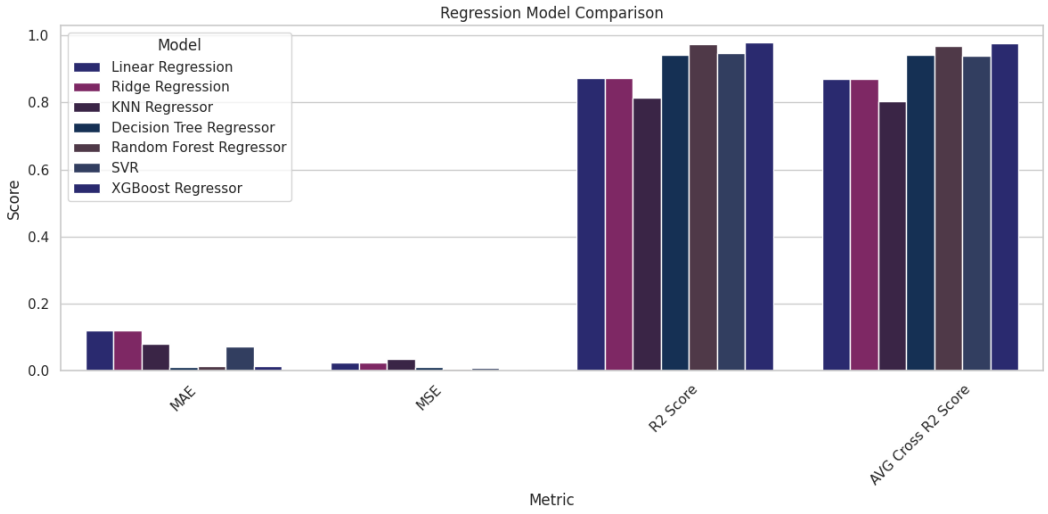

Figure 13: Model Comparison

********** THE END **********