SMART LOAN CLASSIFIER

Group 12

LIST OF CONTENTS

01 PROJECT SELECTION & PROBLEM DEFINITION

02 DATA SOURCE

03 DATA DESCRIPTION

04 DATA EXPLORATION, VISUALIZATION, AND PROCESSING

05 DIMENSION REDUCTION AND VARIABLE SELECTION

06 MODEL EXPLORATION AND MODEL SELECTION

07 CONCLUSION & RECOMMENDATIONS

PROBLEM DEFINITION

- **Loan approval** is a **critical process** for both financial institutions and borrowers
- Traditionally done through **manual evaluation** of:
- **Demographics**
- **Financial records**
- **Credit information**
- This approach is often **time-consuming, inconsistent,** and **prone to errors**
- Project Goal: **Automate** and **optimize loan** decisions using **machine learning**
- Approach: Build a **binary classification** model to predict **approval** or **denial**
- Model is trained using **historical loan data**
- Benefits:
- **Streamlined processes**
- **Consistent and accurate decisions**
- **Better risk management**

OBJECTIVE AND KEY RESULTS

- **Objective 1: Develop a High-Performing, Fair, and Compliant Loan Classification Model**
- KR1: Achieve a model accuracy of ≥ 90% while maintaining fairness across demographic groups.
- KR2: Reduce false approvals (risky loans approved) to ≤ 4% and false rejections to ≤ 6%.
- KR3: Achieve balanced precision and recall (≥ 85%) across all demographic groups to ensure fairness.
- **Objective 2: Enhance Model Interpretability and Trustworthiness**
- KR1: Ensure 100% of model predictions include a confidence score, with at least 85% confidence level.
- KR2: Conduct bias analysis to ensure that no demographic group is disadvantaged by more than 5% discrepancy in approval rates.
- **Objective 3: Optimize Model Deployment for Efficiency, Scalability, and Cost**
- KR1: Automate ≥ 75% of loan decisions while keeping manual interventions for edge cases.
- KR2: Reduce loan processing time per application from 2 hours to under 5 minutes.
- KR3: Reduce operational cloud costs by 15%, optimizing computing resources without impacting performance.

KEY PERFORMANCE INDICATORS

1. Model Performance & Fairness Metrics

- Accuracy: Target ≥ 90%.
- Precision & Recall: ≥ 87% for both approved and denied loans.
- False Positive Rate (FPR): Keep ≤ 4% (minimizing approval of risky loans).
- False Negative Rate (FNR): Keep ≤ 6% (minimizing rejection of trustworthy applicants).
- F1-score: Maintain ≥ 0.88.
- Fairness Score: Ensure a ≤ 5% difference in approval rates across demographic groups.

2. Loan Processing Efficiency & Automation

- Average loan processing time: Reduce from 2 hours to under 5 minutes.
- Percentage of fully automated approvals: ≥ 80%.
- Reduction in manual loan evaluations: ≥ 60% decrease.
- Average API response time: ≤ 1 second for loan eligibility requests.

3. Model Interpretability & Trustworthiness

- Percentage of predictions with confidence scores: 100%.
- Confidence level in predictions: Maintain at least 85% confidence.
- Top 5 influential factors displayed in loan approval explanations.
- Bias mitigation metric: Approval rates across demographic groups should have ≤ 5% deviation.

4. System Performance & Scalability

- Number of loan applications processed per second: Target ≥ 4 applications/sec.
- System uptime: ≥ 99.99%.
- Peak load handling capability: Ensure system stability for ≥ 15,000 applications per hour.
- Data processing efficiency: Reduce memory and CPU consumption by at least 20% through optimization.

DATA SOURCE

- 📂 **Source:** Kaggle – Loan Approval Classification Data
- 📊 **Records:** 45,000 entries
- 📌 **Features:** 14 variables
- 🏷️ **Target**: Loan Approval (Approved / Denied)

DATA DESCRIPTION

Feature Name	Description	Type
`person_age`	Age of the applicant	Numerical
`person_gender`	Gender (male/female)	Categorical
`person_education`	Education level (High School, Bachelor, Master, etc.)	Categorical
`person_income`	Annual income in USD	Numerical
`person_emp_exp`	Employment experience in years	Numerical
`person_home_ownership`	Type of home ownership: RENT / OWN / MORTGAGE	Categorical
`loan_amnt`	Loan amount requested	Numerical
`loan_intent`	Purpose of the loan (e.g., EDUCATION, PERSONAL, etc.)	Categorical
`loan_int_rate`	Interest rate on the loan (%)	Numerical
`loan_percent_income`	Loan as a percentage of the applicant’s income	Numerical
`cb_person_cred_hist_length`	Length of applicant's credit history in years	Numerical
`credit_score`	Applicant's credit score	Numerical
`previous_loan_defaults_on_file`	Whether applicant has defaulted on previous loans (Yes/No)	Categorical
`loan_status`	Target variable: 1 = Approved, 0 = Rejected	Numerical (Binary)

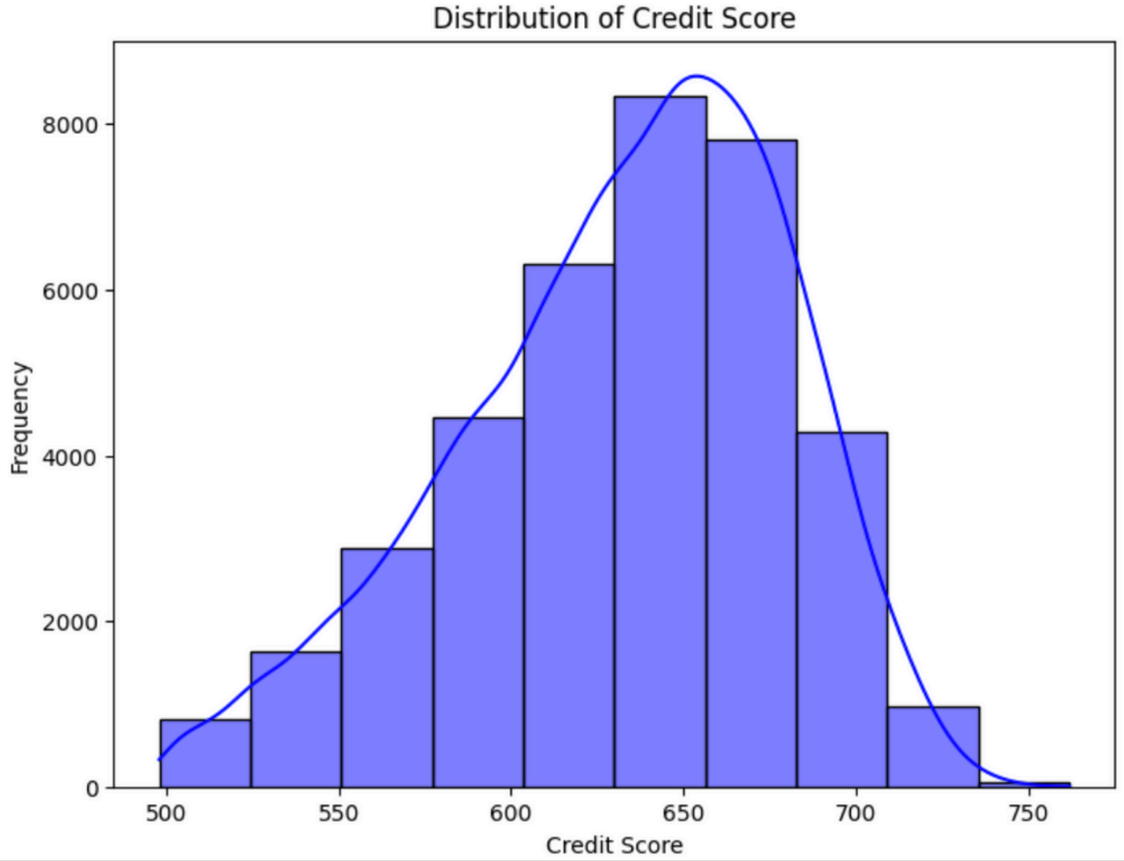
DATA DESCRIPTION

	person_age	person_income	person_emp_exp	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length	credit_score	loan_status
count	45000.000000	4.500000e+04	45000.000000	45000.000000	45000.000000	45000.000000	45000.000000	45000.000000	45000.000000
mean	27.764178	8.031905e+04	5.410333	9583.157556	11.006606	0.139725	5.867489	632.608756	0.222222
std	6.045108	8.042250e+04	6.063532	6314.886691	2.978808	0.087212	3.879702	50.435865	0.415744
min	20.000000	8.000000e+03	0.000000	500.000000	5.420000	0.000000	2.000000	390.000000	0.000000
25%	24.000000	4.720400e+04	1.000000	5000.000000	8.590000	0.070000	3.000000	601.000000	0.000000
50%	26.000000	6.704800e+04	4.000000	8000.000000	11.010000	0.120000	4.000000	640.000000	0.000000
75%	30.000000	9.578925e+04	8.000000	12237.250000	12.990000	0.190000	8.000000	670.000000	0.000000
max	144.000000	7.200766e+06	125.000000	35000.000000	20.000000	0.660000	30.000000	850.000000	1.000000

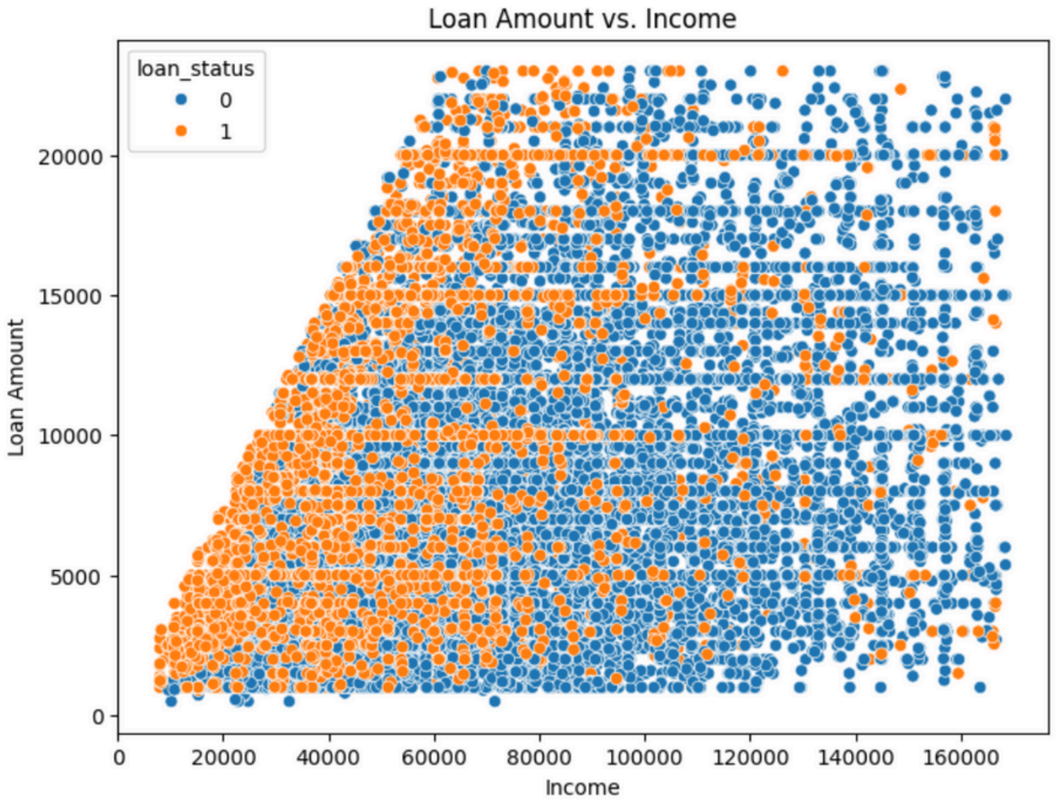
- **Outliers Detected:**
- **person_age:** Max value **144** is unrealistic → likely an outlier
- **person_income:** Max value **$7.2 million** also an outlier
- 💳 **Credit Scores:** Range from **390 to 850**, with a mean around **632**
- 📊 **Loan Amounts:** Mostly between **$5,000–$12,000 (25%–75% range)**
- 🧠 **Loan Percent Income:** Median applicant borrows about **12% of income**
- ✅ **Dataset is clean:** No nulls, no missing records

EXPLORATORY DATA ANALYSIS

- 📈 Most applicants have **credit scores between 600 and 680**, indicating a moderately good credit profile.
- 🔍 The distribution is slightly left-skewed, with fewer applicants having low scores (<550) or very high scores (>720).
- ✅ This range suggests that credit score is a **key factor** in determining loan eligibility and can effectively help separate risk levels.

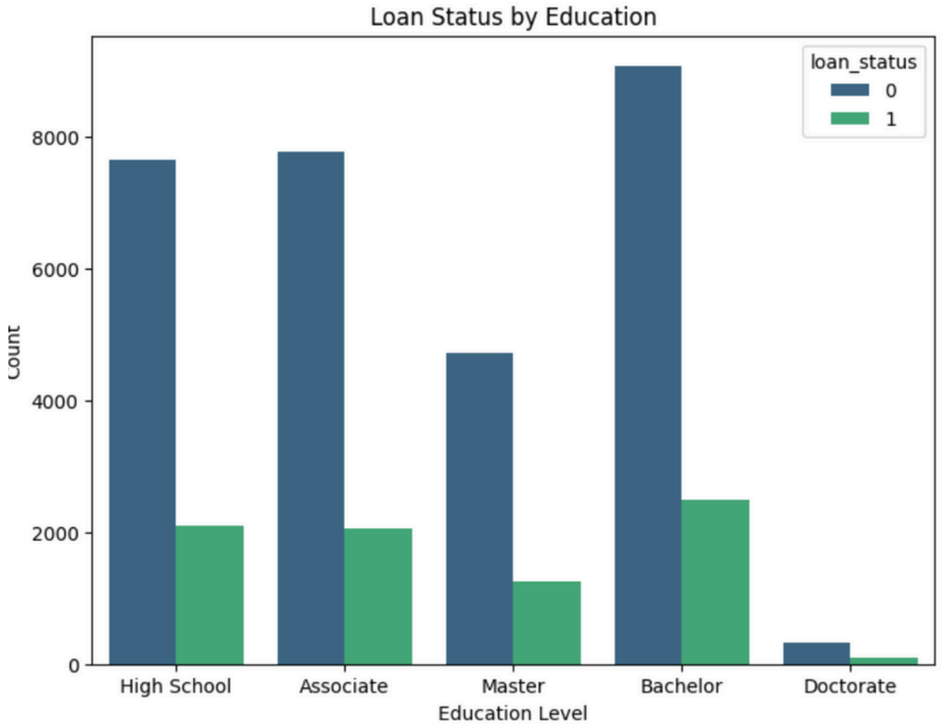


EXPLORATORY DATA ANALYSIS



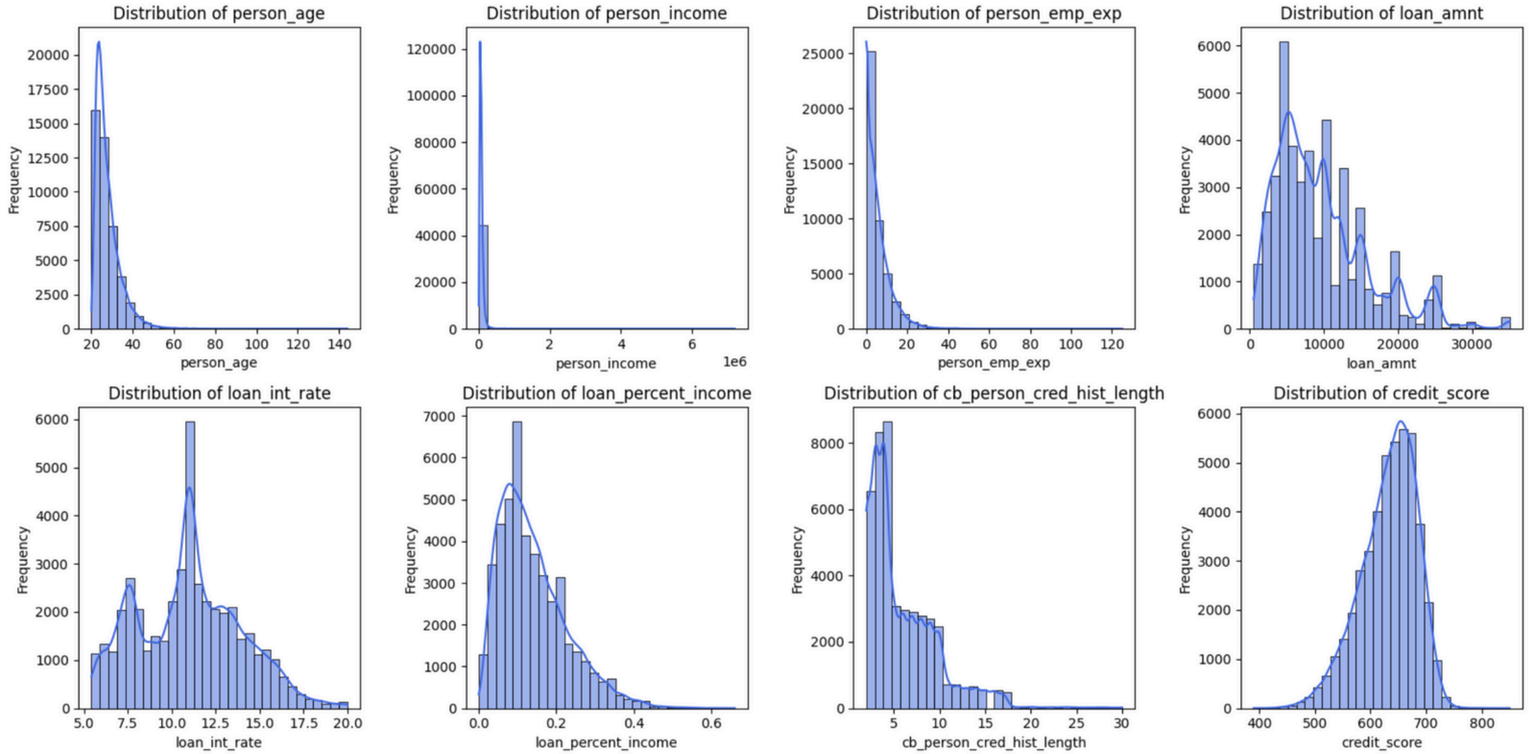
- 💡 **Loan approvals (orange)** are more frequent for **low to mid-income ranges** — especially when the requested loan amount is relatively moderate.
- ⚠️ **Higher income** does not guarantee approval — many **high-income applicants are still denied (blue)**, suggesting other factors influence decisions.
- 📊 The triangular shape shows a natural boundary: **loan amount increases with income**, but rarely exceeds a certain percentage of income — indicating income-to-loan ratio is a key eligibility factor.

EXPLORATORY DATA ANALYSIS



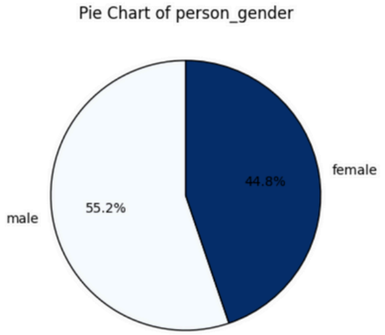
- 🎓 **Bachelor's degree holders have the highest number of applicants,** but also a high rate of rejections — education alone doesn't guarantee approval.
- 📉 **Doctorate holders are few in number,** with very low approval and rejection counts — possibly due to fewer applicants or other disqualifying factors.
- ⚖️ Across all education levels, the number of **rejected loans (status = 0) i**s significantly higher than approvals — indicating stricter criteria beyond education.

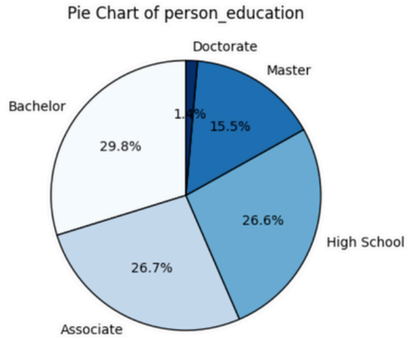
FEATURE DISTRIBUTIONS (NUMERICAL VARIABLES)

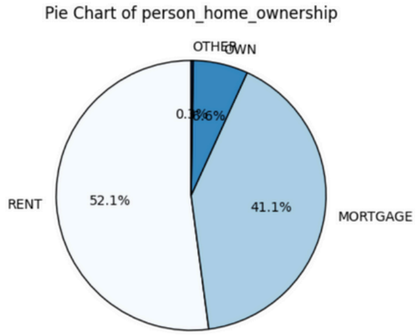


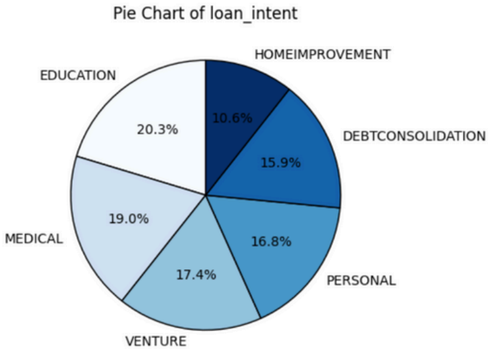
- 📊 Most numeric features like person_age, person_income, person_emp_exp show **right-skewed distributions**
- 🧮 credit_score is **nearly normal,** making it well-suited for modeling
- 📈 Features such as loan_amnt, loan_int_rate, and loan_percent_income display **peaks and outliers,** later cleaned during preprocessing

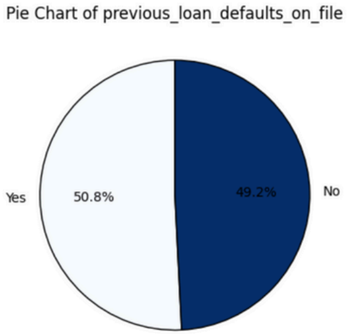
CATEGORICAL FEATURE OVERVIEW (PIE CHARTS)





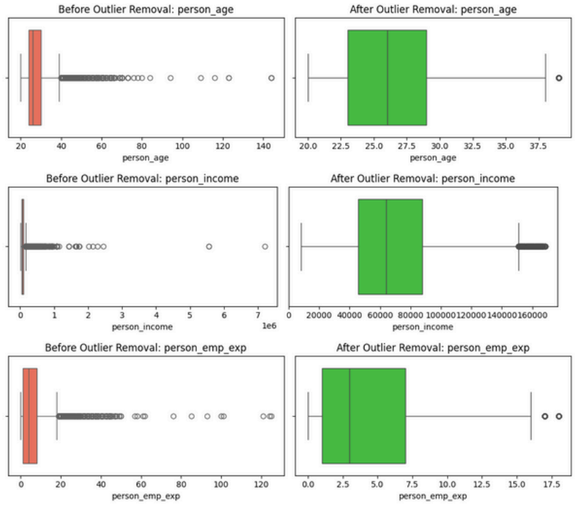


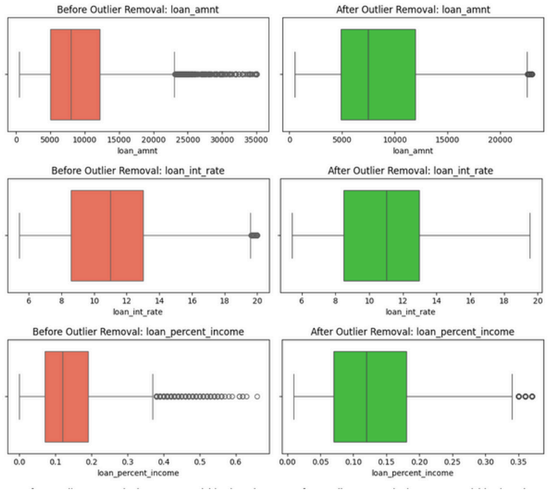


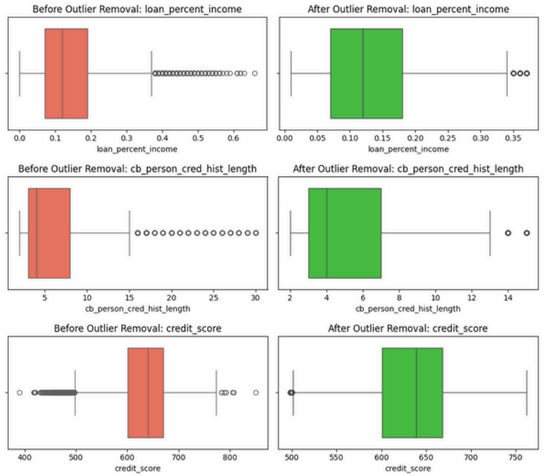


- 👤 **Gender Distribution**: ~55% male, ~45% female — relatively balanced
- 🎓 **Education Levels:** Bachelor's is the most common, followed by Associate and High School
- 🏠 **Home Ownership:** Most applicants **rent (52%)** or have a **mortgage (41%)**

DATA PREPROCESSING







- ✅ No missing values or duplicates detected in the dataset
- ➖ No imputation or deletion required for nulls

Outlier Detection & Removal

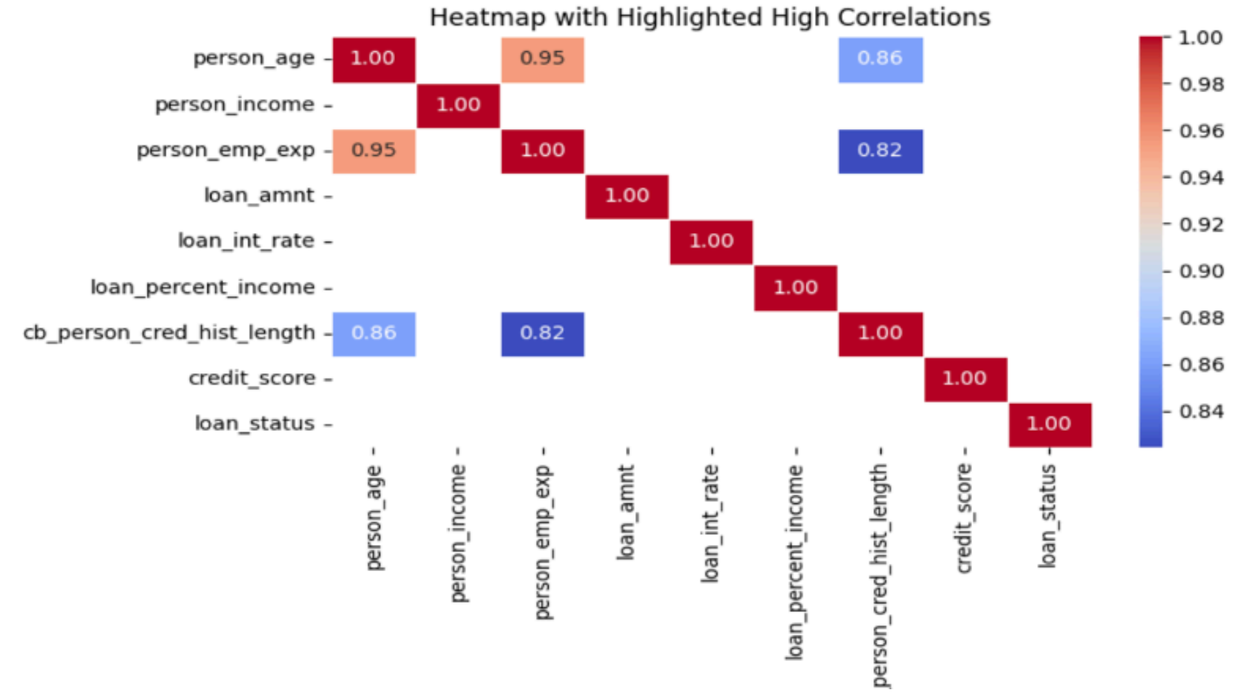
- 📦 Outliers identified using **box plots** across key numerical features
- ✂️ **7,451 outliers removed** from multiple columns
- 🧮 **Data shape after removal:** (37,549 rows × 14 columns)
- Features affected:
- **person_age, person_income, person_emp_exp, loan_amnt, loan_int_rate, loan_percent_income, cb_person_cred_hist_length, credit_score**

Effect of Outlier Removal

- 📊 Distributions are now more **normalized and centered**
- ⚖️ Reduced skewness improves **feature scaling** and **model training** stability
- 🚀 Preprocessed data is now optimized for **better accuracy** and **generalization**

CORRELATION HEATMAP ANALYSIS

	Feature 1	Feature 2	Correlation
0	person_age	person_emp_exp	0.954412
2	person_emp_exp	person_age	0.954412
1	person_age	cb_person_cred_hist_length	0.861985
4	cb_person_cred_hist_length	person_age	0.861985
3	person_emp_exp	cb_person_cred_hist_length	0.824272
5	cb_person_cred_hist_length	person_emp_exp	0.824272



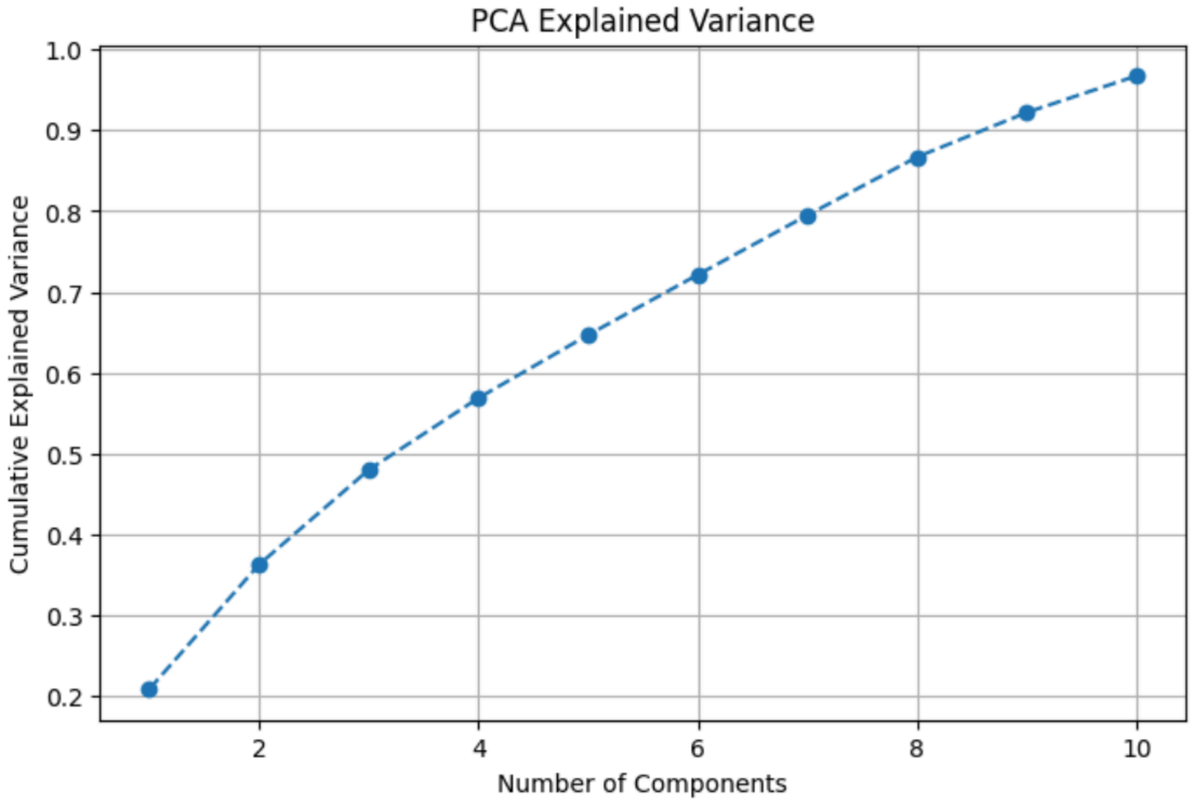
- Visualizes relationships between **numerical features** using **Pearson correlation coefficients.**
- Strong positive correlations observed between:
- person_age and person_emp_exp → **r = 0.95**
- person_age and credit_hist_length → **r = 0.86**
- emp_exp and credit_hist_length → **r = 0.82**
- These correlations suggest natural dependencies:
- Older applicants usually have **more work experience** and **longer credit history.**

DIMENSIONALITY REDUCTION (PCA)

```
PCA Results:
            PCA1      PCA2      PCA3      PCA4      PCA5      PCA6      PCA7  \
0      -2.599776 -1.246868 -1.837059 -0.742336 -1.708003  2.189196 -1.426307
1      -2.602924 -0.096356 -1.338966 -1.990454 -0.554874  0.291946 -2.040331
2      -2.110501  0.420912 -1.836986  0.021031  1.210460 -0.610510 -1.597083
3      -2.452442  1.677184 -0.561673 -0.596445 -0.566860 -1.172531 -0.018067
4      -1.525413 -0.316246 -1.421657  1.037045  1.580300  0.633594 -0.881192
...          ...       ...       ...       ...       ...       ...       ...
59029 -1.190894  0.015652 -1.030380  0.240217  0.495814 -0.413239  1.328986
59030 -0.139330  2.080669 -0.031139 -0.084284  1.916472  0.335619  0.657579
59031 -0.351539 -1.195355  0.993478  1.079955  1.242054  0.713600  0.881376
59032  1.701512  0.387775  1.561698  1.128387 -0.500600  0.338469 -1.311516
59033 -1.012210  1.690684  0.304928  1.072834 -0.295672  0.245890  0.091557

            PCA8      PCA9     PCA10
0       0.039358 -0.111194  0.919188
1      -0.019331 -1.864706 -0.724248
2       1.018291 -0.845233  0.691168
3      -1.785106 -0.829281  0.553452
4      -1.513282 -1.466277  1.109709
...          ...       ...       ...
59029  0.025523 -0.537919  0.015915
59030  1.378857  0.073325  0.853799
59031 -0.288624 -0.991088 -0.304123
59032  1.161417 -0.979492  0.158688
59033 -0.782141  0.712449 -0.026529

[59034 rows x 10 columns]

Explained Variance Ratio: [0.20841608 0.15411905 0.1173012  0.08887963 0.07811967 0.07449347
```

- For this dataset, before applying PCA, the dataset underwent one-hot encoding to convert categorical features into numerical format, enabling compatibility with the PCA algorithm.
- The explained variance ratio listed below the table quantifies how much information (variance) each component retains from the original data. The first component explains 22.07% of the total variance, followed by 15.18% from the second, and so on

PCA VARIANCE EXPLAINED GRAPHICALLY



This pattern is typical in PCA and suggests that while the first **few components are highly informative,** later components add diminishing returns. This insight can guide dimensionality reduction strategies by selecting only the **first 6 to 8 components** to maintain a balance between model simplicity and retained information.

DATA PREPROCESSING

- **Encoding Categorical Variables:** Label encoding applied to categorical columns.
- **Handling Missing Values:** Imputation techniques used where necessary.
- **Feature Scaling:** Standardization applied to numerical columns.

```
Training Set Size: (47227, 10)
Testing Set Size: (11807, 10)
```

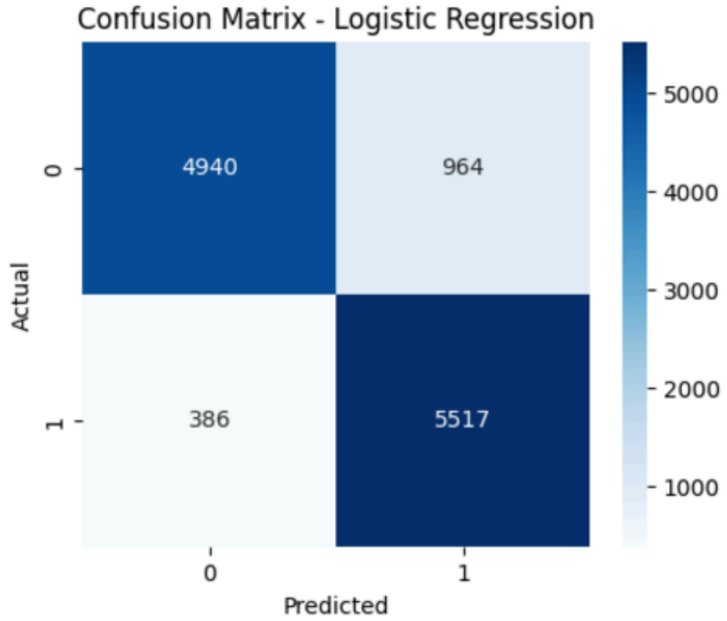
Data Splitting: **80% training** and **20% test** split for model evaluation.

MODELS IMPLEMENTED

Logistic Regression

🔷 Logistic Regression Performance:
Accuracy: 0.8856610485305327
ROC-AUC: 0.8856651938957133

	precision	recall	f1-score	support
0	0.93	0.84	0.88	5904
1	0.85	0.93	0.89	5903
accuracy			0.89	11807
macro avg	0.89	0.89	0.89	11807
weighted avg	0.89	0.89	0.89	11807



Key Insights:

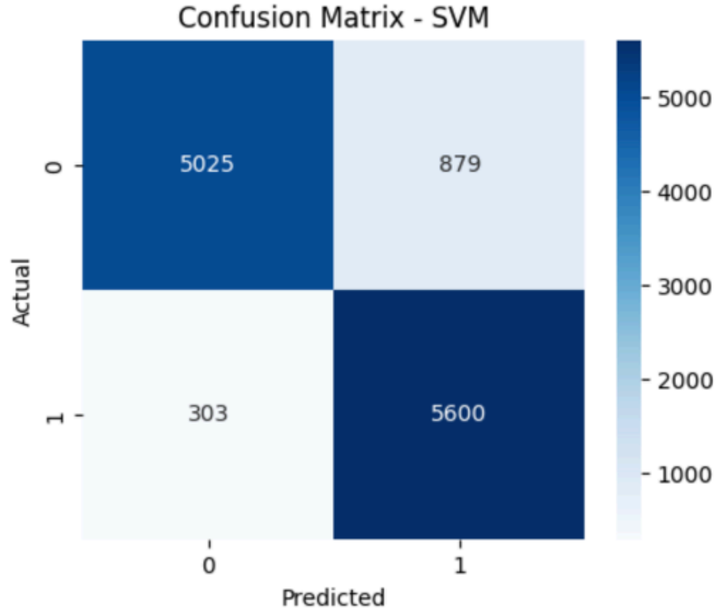
- **Decent Accuracy (89%) –** Performs well as a baseline model but is outperformed by tree-based models.
- **Higher False Positives (964 cases) –** More cases where loans were predicted as approved but should have been rejected.
- **Strong Recall for Class 1 (93%) –** Captures most of the actual approved loans but sacrifices precision slightly.

MODELS IMPLEMENTED

Support Vector Machine

🔷 SVM Performance:
Accuracy: 0.8998898958245108
ROC-AUC: 0.8998940269450973

	precision	recall	f1-score	support
0	0.94	0.85	0.89	5904
1	0.86	0.95	0.90	5903
accuracy			0.90	11807
macro avg	0.90	0.90	0.90	11807
weighted avg	0.90	0.90	0.90	11807



Key Insights:

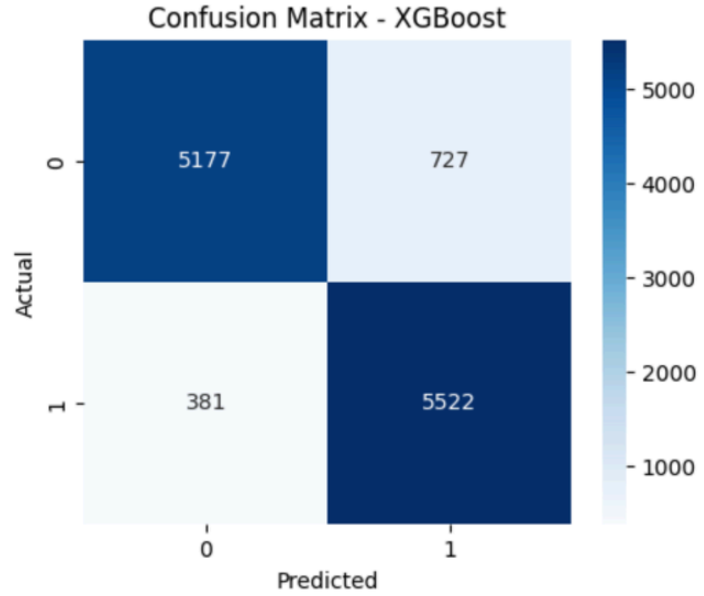
- **Competitive Accuracy (89.98%) –** Slightly lower than Random Forest but still strong.
- **High False Positives (879 cases) –** More false approvals than Random Forest and XGBoost.
- **Best False Negative Reduction (303 cases) –** Has the lowest number of misclassified actual approvals.

MODELS IMPLEMENTED

XGBoost Classifier

🔷 XGBoost Performance:
Accuracy: 0.9061573642754298
ROC-AUC: 0.9061598455748237

	precision	recall	f1-score	support
0	0.93	0.88	0.90	5904
1	0.88	0.94	0.91	5903
accuracy			0.91	11807
macro avg	0.91	0.91	0.91	11807
weighted avg	0.91	0.91	0.91	11807



Key Insights:

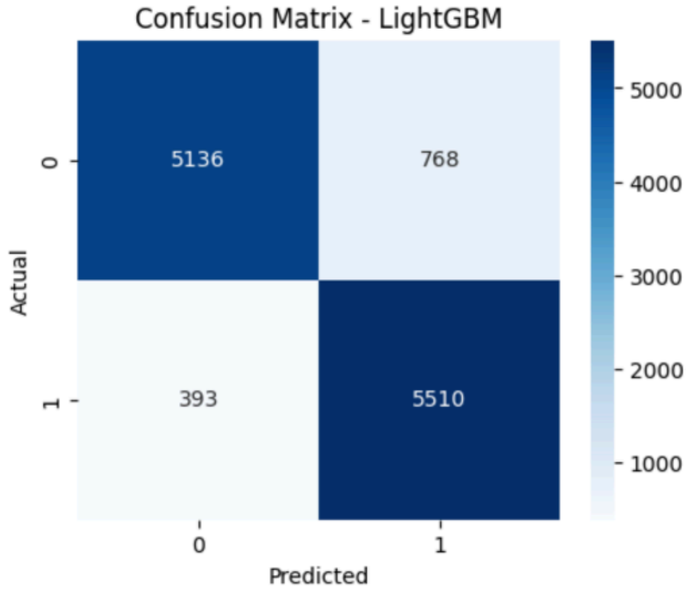
- **High Accuracy (90%) –** Slightly better than SVM and Logistic Regression.
- **Improved False Positive Rate (727 cases) –** Reduces incorrect approvals compared to Logistic Regression.
- **Balanced Precision & Recall (94%) –** Well-suited for complex data with feature interactions.

MODELS IMPLEMENTED

LightGBM Classifier

🔷 LightGBM Performance:
Accuracy: 0.9016685017362581
ROC-AUC: 0.9673091790633306

	precision	recall	f1-score	support
0	0.93	0.87	0.90	5904
1	0.88	0.93	0.90	5903
accuracy			0.90	11807
macro avg	0.90	0.90	0.90	11807
weighted avg	0.90	0.90	0.90	11807



Key Insights:

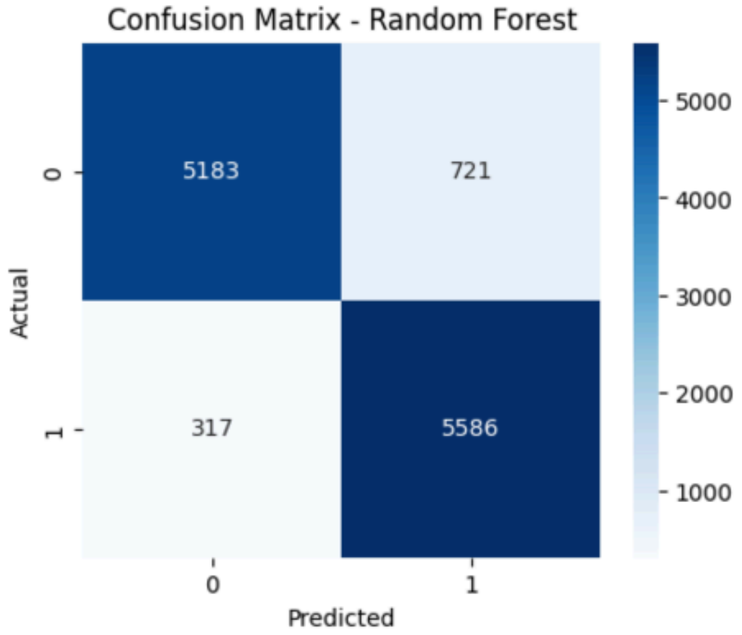
- **Highest AUC (96.73%) –** Best model for ranking loans by risk rather than just classification.
- **Good Accuracy (90.16%) –** Performs almost as well as XGBoost and Random Forest.
- **Strong Recall (93%) –** Captures nearly all loan approvals while maintaining a decent precision level.

MODELS IMPLEMENTED

Random Forest Classifier

🔷 Random Forest Performance:
Accuracy: 0.9120860506479207
ROC-AUC: 0.9120889480430464

	precision	recall	f1-score	support
0	0.94	0.88	0.91	5904
1	0.89	0.95	0.91	5903
accuracy			0.91	11807
macro avg	0.91	0.91	0.91	11807
weighted avg	0.91	0.91	0.91	11807



Key Insights:

- **Best Accuracy (91.20%) –** Outperforms other models in terms of overall correctness.
- **Lower False Positives (721 cases) –** Improves upon Logistic Regression by reducing incorrect approvals.
- **Balanced Precision & Recall (91%) –** Effectively classifies both approved and rejected loans with minimal bias.

MODEL PERFORMANCE

Model Name	Accuracy
Logistic Regression	89%
Random Forest Classifier	91.20%
Support Vector Machine (SVM)	89.98%
XGBoost Classifier	90%
LightGBM Classifier	90.16%

CONCLUSION

- Based on the ROC curve analysis and the provided model performance metrics, the **Random Forest Classifier** emerges as the best overall model for this classification task. It offers the **highest accuracy (91.06%),** a balanced precision-recall trade-off (~91%), and lower false positives (824 cases) compared to other models.
- **LightGBM and XGBoost** also perform exceptionally well with **AUC = 0.97** and strong recall, making them ideal for scenarios where ranking the risk of loans is more important than pure classification.
- **SVM has the lowest false negatives**, making it a good choice if minimizing missed approvals is crucial.
- **Logistic Regression,** while interpretable, falls **short in accuracy** and **false positive reduction.**

THANK YOU