# Loan Approval Predictive Model


Presentation Group 5:
Artemis Lu
Yiling Ding
Sherry Xu
Ashley Sun
Yichen Yang


# Agenda

1. Introduction
2. Data EDA
3. Machine Learning
4. Interpretation
5. Conclusion

# Introduction

# Why Loan?

Loan approval is a cornerstone of the financial sector, directly tied to credit risk assessment. Building a predictive model addresses the critical need to:

- **Minimize default risk**: Financial institutions aim to approve loans for applicants with a low likelihood of default.
- **Streamline decision-making**: Automating loan approvals based on predictive models saves time and reduces human biases.
- **Enhance financial inclusion**: Such models can be tuned to ensure fairer evaluations, improving access to credit for underserved populations.

# Introduction Loan Approval Predictive Model

### Data Source

Dataset titled
*Loan Approval Classification Data*
sourced from **Kaggle**

### Project Objective

Develop a predictive model to classify loan applications as approved or rejected using historical data

### Significance

Enhance decision-making, promote fairness, and improve transparency in loan evaluation for financial institutions

### Methodology

Utilize logistic regression for binary outcome predictions:
approval (1) or rejection (0).

# Data EDA

# Dataset Overview

## The dataset contains 45,000 records and 14 variables

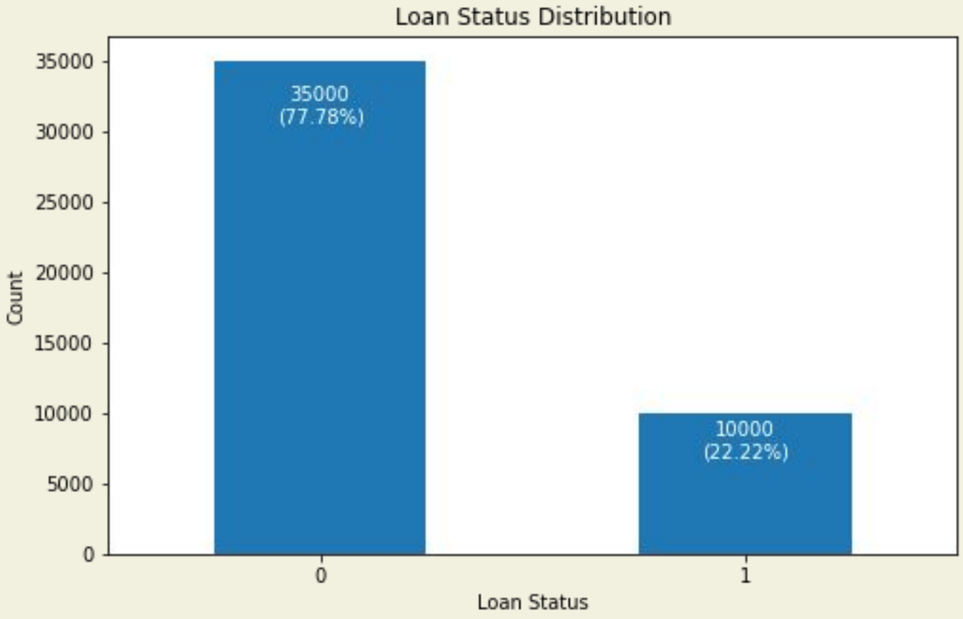
| Column | Description | Type |
|---|---|---|
| person_age | Age of the person | Float |
| person_gender | Gender of the person (female, male) | Categorical |
| person_education | Highest education level (Master, High School, Bachelor, Associate, Doctorate) | Categorical |
| person_income | Annual income | Float |
| person_emp_exp | Years of employment experience | Integer |
| person_home_ownership | Home ownership status (RENT, OWN, MORTGAGE, OTHER) | Categorical |
| loan_amnt | Loan amount requested | Float |
| loan_intent | Purpose of the loan (PERSONAL, EDUCATION, MEDICAL, VENTURE, HOMEIMPROVEMENT, DEBTCONSOLIDATION) | Categorical |
| loan_int_rate | Loan interest rate | Float |
| loan_percent_income | Loan amount as a percentage of annual income | Float |
| cb_person_cred_hist_length | Length of credit history in years | Float |
| credit_score | Credit score of the person | Integer |
| previous_loan_defaults_on_file | Indicator of previous loan defaults (No, Yes) | Categorical |
| loan_status (dependent variable) | Loan approval status: 1 = approved; 0 = rejected | Integer |

**Dependent variable !**

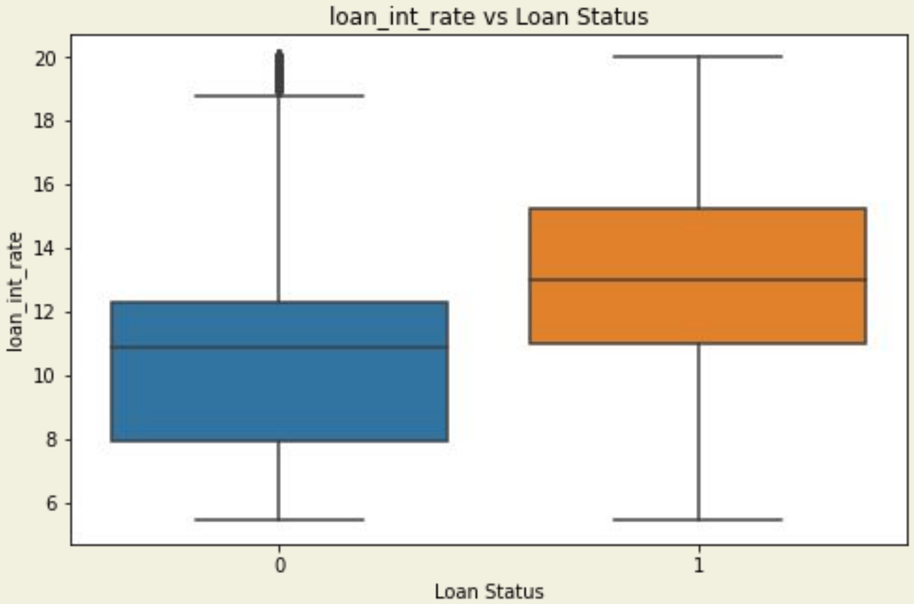
| | person_age | person_income | person_emp_exp | loan_amnt | loan_int_rate | loan_percent_income | cb_person_cred_hist_length | credit_score | loan_status |
|---|---|---|---|---|---|---|---|---|---|
| count | 45000.000000 | 4.500000e+04 | 45000.000000 | 45000.000000 | 45000.000000 | 45000.000000 | 45000.000000 | 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 |

# Loan Status Distribution

- Loans in category 0 account for 77.78% of the total dataset.
- Loans in category 1 account for 22.22% of the total dataset.
- The majority of the loans fall into category 0 (Rejected)

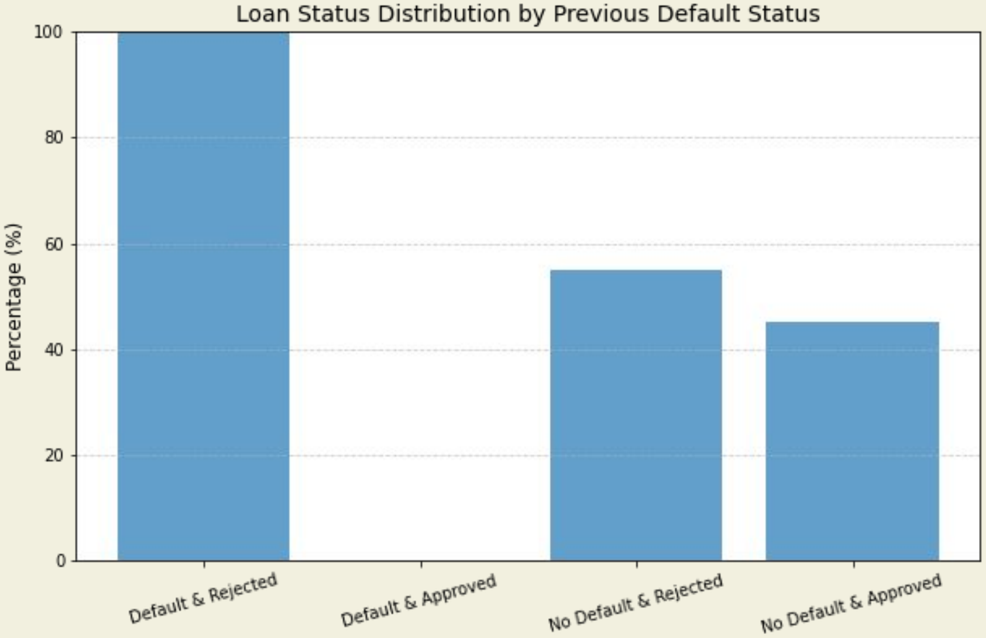


# Loan Status VS Loan Interest Rate




- Loan Status 0's median interest rate is lower compared to Loan Status 1.
- The maximum interest rate of Loan Status 1 appears to be higher than in Loan Status 0, with no visible extreme outliers.
- Loans with higher interest rate are more likely to be approved.

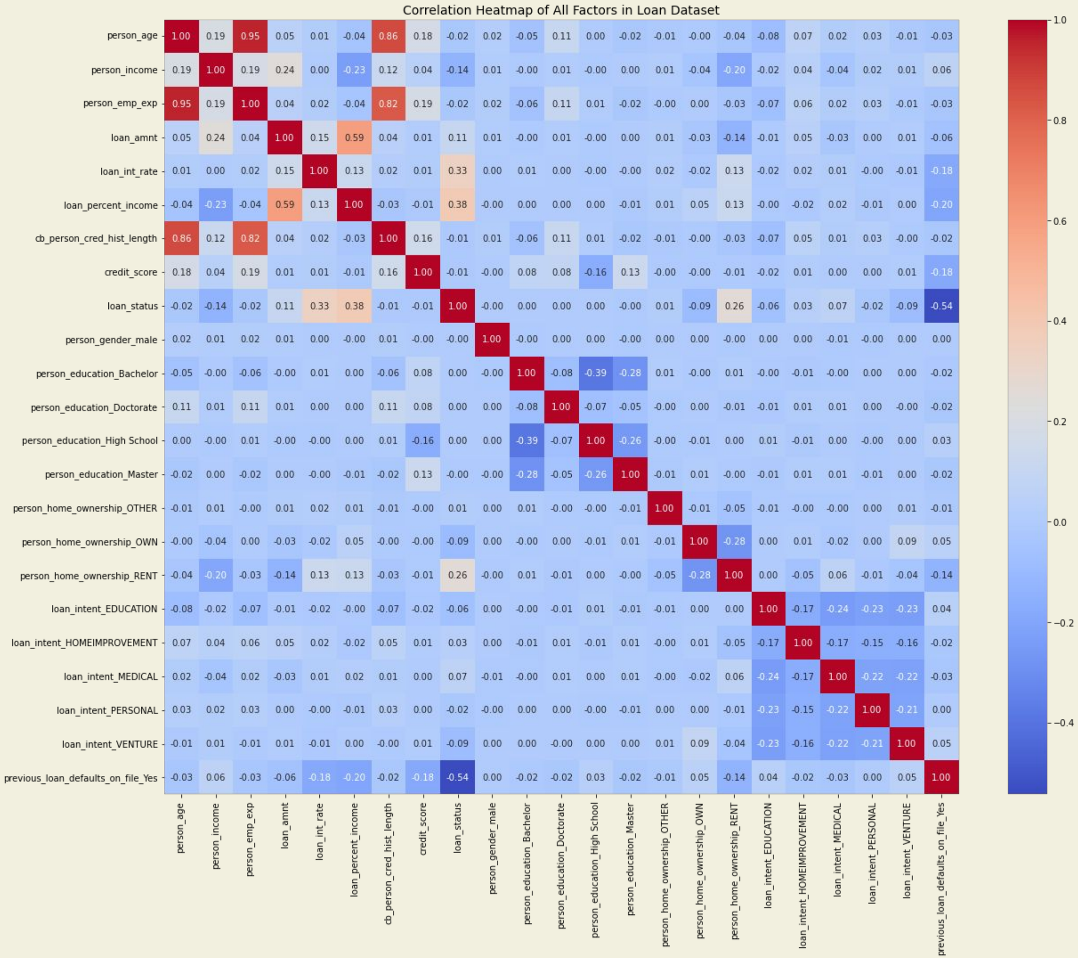
# Loan Status VS Previous Default Status




- Individuals with a history of default are significantly more likely to have their loan applications rejected.
- The approval rate for individuals with a default history is relatively low.

# Correlation Heat Map

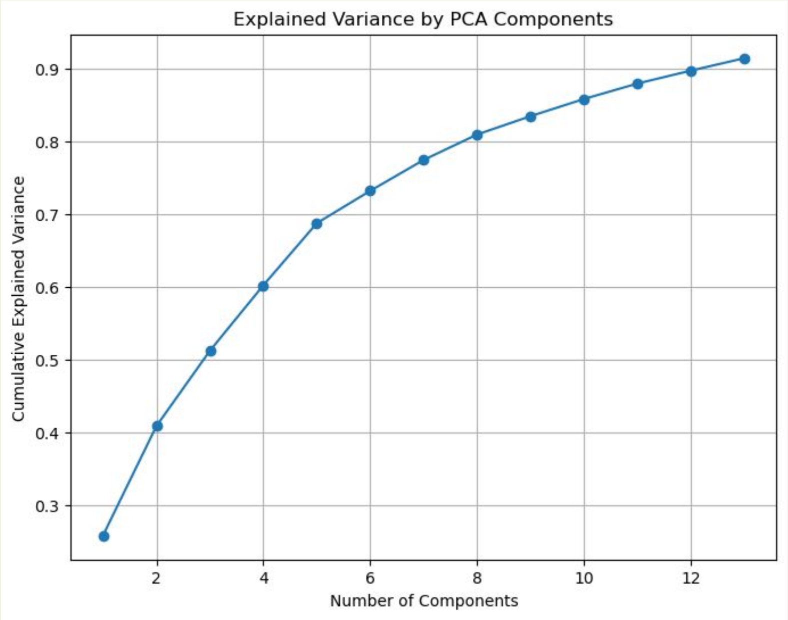
- A history of previous loan defaults strongly correlates with loan rejections.
- Higher credit scores are positively associated with loan approvals.
- Higher interest rates may be associated with approved loans, though the relationship is weak.



# Machine Learning

# Principal Component Analysis

- **Goal**: To reduce the dimensionality of the dataset while retaining the most important information. This helps to simplify the dataset for faster computations and improved model performance
- **Details**:
  - Applied Principal Component Analysis (PCA) to reduce features while retaining **91.49**% of the original dataset's variance
  - Selected **13** principal components out of 27 original features based on the explained variance threshold
  - Top 3 components explained **51.25%** of the total variance



# Binormal Model

- Target: Build a predictive model to classify whether a loan application will be approved or not
- Reasons to Choose Binomial Regression
  - **Binary Outcome**: Designed for binary (yes or no) classification tasks
  - **Interpretability**: Provides interpretable coefficients that explain the impact of each feature on the probability of approval
  - **Efficiency**: Computationally efficient and works well even with relatively large datasets
  - **Baseline Model**: Serves as a strong baseline to compare with other models

Statsmodels Logistic Regression Summary:

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | loan_status | No. Observations: | 36000 |
| Model: | GLM | Df Residuals: | 35986 |
| Model Family: | Binomial | Df Model: | 13 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -8007.0 |
| Date: | Fri, 06 Dec 2024 | Deviance: | 16014. |
| Time: | 18:41:57 | Pearson chi2: | 1.81e+04 |
| No. Iterations: | 11 | Pseudo R-squ. (CS): | 0.4588 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -12.4559 | 0.440 | -28.287 | 0.000 | -13.319 | -11.593 |
| x1 | 0.1847 | 0.016 | 11.698 | 0.000 | 0.154 | 0.216 |
| x2 | 3.2413 | 0.094 | 34.570 | 0.000 | 3.058 | 3.425 |
| x3 | -3.4267 | 0.108 | -31.849 | 0.000 | -3.638 | -3.216 |
| x4 | 3.4140 | 0.135 | 25.304 | 0.000 | 3.150 | 3.678 |
| x5 | 1.2284 | 0.026 | 47.917 | 0.000 | 1.178 | 1.279 |
| x6 | 0.4642 | 0.033 | 14.219 | 0.000 | 0.400 | 0.528 |
| x7 | 11.2772 | 0.403 | 27.953 | 0.000 | 10.486 | 12.068 |
| x8 | 10.3482 | 0.403 | 25.649 | 0.000 | 9.557 | 11.139 |
| x9 | 0.2174 | 0.038 | 5.786 | 0.000 | 0.144 | 0.291 |
| x10 | -0.2302 | 0.039 | -5.839 | 0.000 | -0.308 | -0.153 |
| x11 | -1.5711 | 0.096 | -16.322 | 0.000 | -1.760 | -1.382 |
| x12 | 0.2708 | 0.050 | 5.449 | 0.000 | 0.173 | 0.368 |
| x13 | -0.1710 | 0.048 | -3.546 | 0.000 | -0.265 | -0.076 |

# Formula for Binomial Logistic Regression

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}}$$

- $P(Y=1|X)$: Probability of the outcome $Y=1$ given predictors $X=(X_1, X_2, \ldots, X_p)$.
- $\beta_0$: Intercept term (bias).
- $\beta_1, \beta_2, \ldots, \beta_p$: Coefficients for the predictor variables.
- $X_1, X_2, \ldots, X_p$: Predictor variables (features).

# Model Methodology

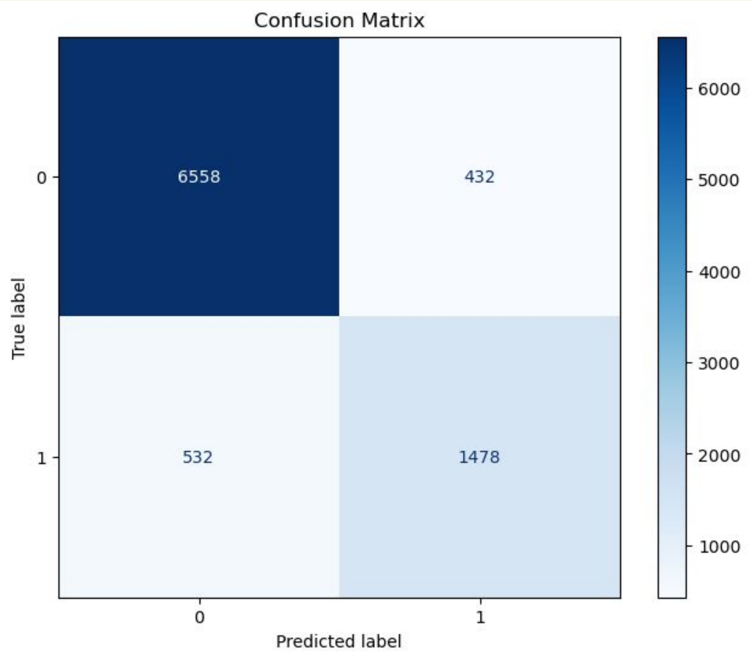
**Binomial Logistic Regression** is typically estimated using the **Maximum Likelihood Estimation (MLE)** method.

- Binomial regression models the probability of a **binary outcome** (Y=1 or Y=0) based on predictor variables. Unlike Ordinary Least Squares (OLS), the outcome is **not continuous** but **rather probabilistic**, making MLE the preferred method.

- MLE finds the set of parameters that **maximize the likelihood** of observing the given data.

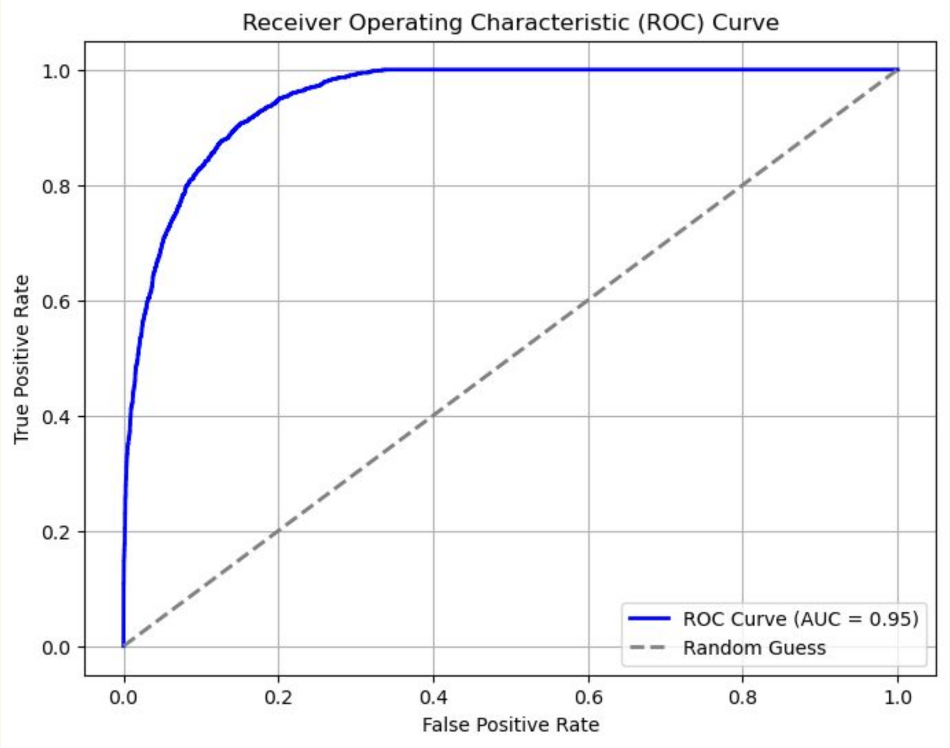
# Model Performance

```
Accuracy of Binomial Regression Model: 0.8928888888888888
ROC-AUC Score: 0.9516168086605599

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      6990
           1       0.77      0.74      0.75      2010

    accuracy                           0.89      9000
   macro avg       0.85      0.84      0.84      9000
weighted avg       0.89      0.89      0.89      9000
```




- Overall Accuracy: **0.89**

# ROC curve



Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.95)
Random Guess

# Interpretation

# Interpretation I: Our model is better and statistically sound

| Metric | Full Model | Our Model |
|---|---|---|
| Log-Likelihood | -10,880 | -8,007 |
| Deviance | 21,760 | 16,014 |
| AIC (Akaike Criterion) | 21,804 | 16,042 |
| BIC (Bayesian Criterion) | -4601,53 | -361,525 |

***Insights:***

***Our selected model performs better***

1. Variables included: 13 (91.5% explained variance)
2. Log-likelihood: Less negative; Better fit
3. Deviance:
   - Our model is lower & better by removing less useful features
4. AIC: Our model performs better
   - Simpler and more effective
   - The selected model has a better balance
5. BIC: Less negative; more fit

# Interpretation II: Train and Test Set

| Metric | Training Data | Test Data |
|---|---|---|
| Accuracy | 0.8949 | 0.8929 |
| ROC-AUC Score | 0.9533 | 0.9516 |
| Precision (Class 0) | 0.93 | 0.92 |
| Recall (Class 0) | 0.94 | 0.94 |
| F1-Score (Class 0) | 0.93 | 0.93 |
| Precision (Class 1) | 0.77 | 0.77 |
| Recall (Class 1) | 0.74 | 0.74 |
| F1-Score (Class 1) | 0.76 | 0.75 |
| Macro Avg (F1-Score) | 0.85 | 0.84 |
| Weighted Avg (F1-Score) | 0.89 | 0.89 |

# Interpretation II-Extra: Cross-Validation

To further validate our model performance, we conducted 5 folds

**Insights:**

The average C-V accuracy score of about 0.895:

- A good indication: the model generalize well.

The standard deviation C-V accuracy 0.003:

- Very low
- Model performance is stable

The accuracy scores for each of the 5 folds:

- Fold 1: `0.89125`
- Fold 2: `0.89986`
- Fold 3: `0.89375`
- Fold 4: `0.89583`
- Fold 5: `0.89264`

# Interpretation III: PCA-transformed Feature Importance

PCA-transformed Coefficient Comparison:

| | Feature | Coefficient | Absolute Coefficient |
|---|---|---|---|
| 6 | PC7 | 8.435353 | 8.435353 |
| 7 | PC8 | 7.504869 | 7.504869 |
| 2 | PC3 | -2.679868 | 2.679868 |
| 1 | PC2 | 2.592622 | 2.592622 |
| 3 | PC4 | 2.473071 | 2.473071 |
| 4 | PC5 | 1.133722 | 1.133722 |
| 10 | PC11 | -0.942629 | 0.942629 |
| 5 | PC6 | 0.343937 | 0.343937 |
| 9 | PC10 | -0.179319 | 0.179319 |
| 8 | PC9 | 0.165668 | 0.165668 |
| 11 | PC12 | 0.124084 | 0.124084 |
| 0 | PC1 | 0.105884 | 0.105884 |
| 12 | PC13 | -0.076440 | 0.076440 |

**Insights:**

1. PC7 and PC8 have **the largest positive** influence.
   a. A positive and large coefficient; When the value of PC7 is high, the model is much more likely to predict the positive class.
2. PC3 has a **strong negative** influence.
   a. PC3's coefficient is about −2.680
   b. *1-unit increase in PC3 decreases the probability of default by approximately 2.68%*

**Further Analysis Needed:**

Recognize which original features contribute most to the high-impact principal components (PCs) since *each PC is a linear combination of original features*

# Interpretation IV: Original Features Importance from PC7-Feature

| Original Feature Variable | Coefficient |
|---|---|
| loan_intent_HOMEIMPROVEMENT | 0.39 |
| person_home_ownership_OWN | -0.37 |
| person_age | 0.36 |
| loan_amnt | -0.23 |
| person_emp_exp | -0.20 |
| cb_person_cred_hist_length | -0.16 |
| loan_int_rate | 0.14 |
| loan_percent_income | -0.06 |
| previous_loan_defaults_on_file_Yes | 0.04 |
| loan_intent_MEDICAL | 0.03 |

## Insights:

1. Each coefficient indicates whether and how a feature influences the likelihood of a positive outcome **(loan_status=1).**
2. Features with positive coefficients raise the odds of the loan being in a positive state.
   a. A large positive coefficient i.e.: **`"loan_intent_HOMEIMPROVEMENT"`**, suggest this type of loan resulting in a higher chance of loan outcome.
   b. A negative coefficient i.e.: **`"person_home_ownership_OWN"`**, suggest owning a home is linked to a lower likelihood of that same positive result

# Interpretation V: Competing Model - GAM

```
GAM Results:
Accuracy: 0.8493333333333334
ROC_AUC_Score: 0.8579179211239938
Classification Report:              precision    recall  f1-score   support

           0       0.87      0.95      0.91      6990
           1       0.75      0.48      0.59      2010

    accuracy                           0.85      9000
   macro avg       0.81      0.72      0.75      9000
weighted avg       0.84      0.85      0.84      9000
```

- **Advantage:** Performed better in non-linear relationships between the features and the target variable
- **Result:** It achieved a high Recall score for group 0 but performed significantly worse across all indicators for group 1.

# Conclusion

The binomial logistic regression model achieves strong performance, with:

- **89% accuracy** and a **95% ROC-AUC score** on both training and test datasets.
- Balanced performance across groups, though stronger for rejected loans (class 0).

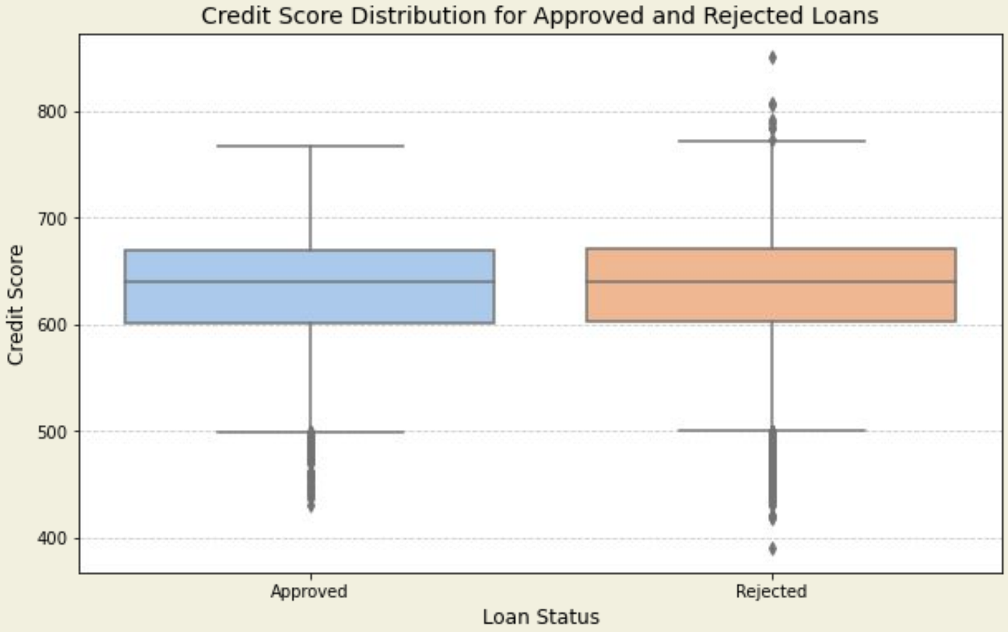
Our Model

- Identification of principal components and key features influencing loan approvals.
- Actionable recommendations for improving decision-making processes.

The model generalizes well, demonstrating robustness and reliability, making it suitable for real-world applications.
This approach enhances **transparency**, **fairness**, and **efficiency** in loan approval system

# Appendix

# Loan Status VS Credit Score



# Loan Status VS Loan Percent Income

- The median loan percent on income is relatively low in Loan Status 0
- People who with a lower loan percent on income are less likely to be approved
- **Weird**