



DSC 789-001 Strategic Capstone Project

Bank Loan Approval Prediction

Using Machine Learning to predict whether bank loan applications will be approved

Roja Eslavath

Amarnath Dasari

Suhitha Yalamanchili

Yamika Ratna Kadiyala

Han Ru Wu

Advisor: Prof. Zahra Sedighi maman

**This work is submitted in fulfillment of requirements for the course DSC
789-001 Strategic Capstone Project**

May 13, 2025

Section 1: Business Understanding

1.1 Description of the Problem

Lenders are faced with the perennial problem of efficiently and effectively evaluating loan applicants. Manual loan approval processes are slow, inconsistent, and at times biased, which contributes to customer dissatisfaction and higher loan default risk. Hence, there is an increasing necessity to adopt machine learning algorithms that have the capability to automate decisions and enhance the consistency of loan approval processes.

The problem which we desire to address is:

Can we create a predictive model that decides whether to reject or approve a loan based on the personal and financial details of the applicant?

1.2 Business Goal

The primary goal is to build an accurate machine learning model that is able to determine if a loan must be accepted or declined based on applicant features like income, employment duration, home ownership, loan amount, interest rate, and credit history. The objective is to aid banks in reducing credit risk, enhancing fairness, and accelerating the process of loan decisions using automatic predictive analytics.

1.3 Why Loan Approval Prediction is Important

- An effective loan approval prediction model can:
- Reduce financial losses by pre-screening high-risk applicants earlier.
- Enhance customer experience through quicker and more equitable loan processing.
- Encourage data-based, transparent, and consistent lending policies. This results in more robust customer confidence and more sustainable financial practices.

1.4 Business Use Case

This project replicates a real banking setting whereby loan requests must be expedited. Banks can leverage predictive analytics to better allocate resources, reduce the need for manual intervention, and achieve fairness in making decisions.

1.5 Project Plan:

Business Analytics Project				
Week	Phase	Task	Description	Risk
Week 1	Business Understanding	Understanding the Project & Data Exploration	Define objectives, scope, and challenges. Load the dataset, check for missing values, and clean the data.	Inconsistent or incomplete data, unclear project objectives.
Week 2	Data Understanding	Data Preprocessing & Feature Engineering	Encode categorical variables, scale numerical features, and perform feature selection to optimize model performance.	Risk of losing valuable data, potential data leakage.
Week 3	Data Preparation	Model Selection & Baseline Training	Train initial machine learning models (Logistic Regression, Decision Trees, Random Forest, XGBoost) and evaluate baseline performance.	Models may not perform well, risk of overfitting
Week 4	Data Preparation	Hyperparameter Tuning & Model Optimization	Improve model performance using hyperparameter tuning techniques like Grid Search and Random Search.	Overfitting, high computational demand
Week 5	Modeling	Model Evaluation & Validation	Assess models using accuracy, precision, recall, F1-score, and AUC-ROC. Apply cross-validation for reliability.	Risk of biased models, misleading evaluation metrics
Week 6	Modeling	Bias & Fairness Analysis	Check for algorithmic bias, ensure fairness, and make necessary adjustments for ethical AI practices.	Unintended bias, challenges in achieving complete fairness
Week 7	Evaluation	Deployment Strategy	Develop an API or web-based interface (if needed) to make the model accessible for predictions.	Possible deployment issues, scalability concerns
Week 8	Deployment	Final Report & Presentation	Document findings, prepare a report, and create a presentation summarizing the project results.	Incomplete reporting, ineffective communication of findings

The loan approval prediction task is solving a critical business problem in the financial and banking industry. Using historical applicant information and machine learning algorithms, we hope to create a model that can facilitate the automation and optimization of loan decisioning. The business objective is to minimize financial risk, enhance approval accuracy, and optimize the lending process. In this project, we also provide a ranking of the most influential features in approval decisions and recommend a data-driven decision-making approach for fair, efficient, and consistent loan approvals.

Section 2: Data Understanding

The Loan Approval dataset contains information about individuals applying for personal loans. It includes details about the applicants' personal and financial background, the purpose and

amount of the loan, and whether the loan was approved as “good” or “bad.” The data is useful for exploring trends in loan approvals and building models to predict loan outcomes.

Kaggle. (n.d.). *Bank Loan Approval* [Data set]. Retrieved from <https://www.kaggle.com>

2.1 Exploratory Data Analysis (EDA):

2.1.1 Dataset Structure

- **Observations:** 32,581
- **Variables:** 12 (8 numerical, 4 categorical)
- **Target Variable:** `loan_status` (1 = Approved, 0 = Not approved)

Variable Name	Type	Description
person_age	Integer	Age of the applicant
person_income	Integer	Annual income of the applicant
loan_amnt	Integer	Requested loan amount
loan_int_rate	Float	Interest rate on the loan (some missing values)
loan_percent_income	Float	Loan amount as a percentage of income
person_emp_length	Float	Length of employment in years (some missing values)
person_home_ownership	Categorical	Type of home ownership (e.g., RENT, OWN)
loan_intent	Categorical	Purpose of the loan (e.g., MEDICAL, EDUCATION)
loan_grade	Categorical	Loan grade indicating risk level
cb_person_default_on_file	Categorical	Whether applicant has a prior default (Y/N)
loan_status	Binary	Target variable: 1 = good loan, 0 = bad loan
cb_person_cred_hist_length	Integer	Length of applicant’s credit history in years

2.1.2 Key Findings from EDA

- **Age:** Most applicants are between 20–30 years old.
- **Income:** Income distribution appears artificial and uniform; likely simulated.
- **Employment Length:** Majority of applicants have under 5 years of work experience.
- **Loan Amount:** Most loans fall between \$5,000–\$12,000.
- **Interest Rate:** Concentrated around 7%–11%; very few extreme values.
- **Credit History Length:** Most applicants have 2–4 years of credit history.

- **Loan Intent:** Education and medical loans are the most common purposes.
- **Home Ownership:** Majority are renters or have mortgages.
- **Loan Approval Imbalance:** 78% of applicants were classified as “bad loans.”

2.1.3 Bivariate Insights

- **Loan Intent vs Loan Status:** Medical loans had higher approval; home improvement loans had more defaults.
- **Home Ownership vs Loan Status:** Renters showed the highest default rate.
- **Default History vs Loan Status:** Strong predictor — prior defaulters were more likely to default again.
- **Employment Length vs Loan Status:** No significant difference; weak predictive power.
- **Loan Amount vs Loan Status:** Loan size didn’t significantly separate approved from rejected cases.

2.1.4 Missing Values & Outliers

- **Missing Values:** Found in `loan_int_rate` and `person_emp_length` — handled using **median imputation**.
- **Outliers:** Detected in income and employment length — handled using **value capping** to reduce distortion.

person_age	0
person_income	0
person_home_ownership	0
person_emp_length	895
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	3116
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

dtype: int64

The EDA provided valuable insights into applicant profiles and their relationship with loan outcomes. By identifying important features, class imbalance, and data quality issues early, we ensured the modeling process would be built on clean, meaningful inputs — increasing the chances of strong predictive performance.

Section 3: Data Preparation

Our focus was on preparing the data to be ready for modeling. That meant cleaning up the dataset, filling in any gaps, handling unusual values, transforming data into a usable format, and engineering new features that could help improve predictions. The goal was to make sure everything was clean, consistent, and meaningful before moving into the modeling phase.

3.1 Cleaning the Data & Handling Missing Values

Duplicate Records:

We noticed that the dataset had 165 duplicate entries, which could distort the results. We removed those to make sure each record represented a unique case.

Missing Data:

Some fields were missing values:

- `person_emp_length`: 887 records were blank
- `loan_int_rate`: 3095 records were missing
- For numerical values, we used the median to fill in gaps—this helps avoid being influenced by outliers.
- For categorical values, we filled missing entries with 'Unknown', so we didn't lose any rows, but still acknowledged incomplete data.

3.2 Encoding Categorical Features

Since machine learning models don't work with text directly, we had to convert categorical data into numbers:

- We used Ordinal Encoding for `loan_grade`, because the grades have a natural order (A is better than B, and so on).
- We used One-Hot Encoding for `home_ownership`, `loan_intent`, and `default_flag`, which turned each category into its own binary column. This helped avoid giving unintended importance to any single group.

3.3 Checking Feature Variability

To make sure each variable had enough information to be useful, we ran a variance check.

- All numeric features passed our threshold, with the lowest being loan_percent_income, which still had acceptable variation.
- No features were removed at this stage, but this step gave us confidence that our dataset was informative.

3.4 Detecting and Handling Outliers

Some values were way outside the typical range—for example, extremely high incomes or unusually large loans. These outliers can mislead the model.

So we used the Interquartile Range (IQR) method to detect and cap extreme values in:

- loan_amnt
- person_income
- loan_int_rate

This kept the dataset realistic and balanced.

3.5 Feature Engineering

We created new variables to help the model better capture patterns:

New Feature	Purpose
income_to_loan_ratio	Measures loan burden relative to income
debt_burden_score	Product of interest rate and loan percent; indicates repayment risk
log_income	Reduces skewness in income distribution
credlen_x_interest	Interaction term to capture effect of short credit history and high interest
high_risk_flag	Binary flag based on high loan + high interest combo
short_hist_high_loan	Flags users with short credit length and high loan amount

3.6 Grouping and Binning

We also grouped some numeric values into categories to simplify the analysis and better support classification tasks:

- cred_hist_cat: Groups users as *Young*, *Mid*, or *Experienced* based on credit history
- loan_size_cat: Categorizes loans as *Small*, *Medium*, or *Large*
- ratio_bin & debt_burden_bin: Turned continuous scores into risk levels like *Very Low*, *Moderate*, *Extreme*
- loan_exposure_cat: Groups users based on their total financial exposure

These categories help models understand trends more easily.

3.7 Scaling and Reducing Noise

Since models can behave differently depending on the range of numbers, we applied StandardScaler to normalize all numeric values. This put all features on a similar scale (mean = 0, standard deviation = 1).

We also reduced noise by:

- Rounding off minor decimal differences in loan_int_rate
- Binning continuous features into groups, helping simplify noisy data

3.8 Understanding Feature Relationships

Before modeling, we checked how variables relate to each other using correlation analysis.

- This showed how strongly some features are linked.
- We didn't remove any features based on correlation yet, but it gave useful insights for future model tuning.

Final Outcome of Data Preparation

At the end of this phase:

- Our data was clean, complete, and consistent
- We added meaningful new features to help with predictions
- We addressed outliers and scaled everything properly
- We're now ready to move confidently into the modeling phase

Section 4: Modeling

Modeling was the process of creating machine learning models to forecast the approval or rejection of a bank loan application. This section outlines the entire modeling process, including model selection, handling class imbalance, feature selection, performance metrics, and key findings.

4.1 Supervised Models Employed

We compared five supervised machine learning models that are popularly known for their accuracy in binary classification tasks:

Logistic Regression: Simple, easy to interpret linear basic model that computes quickly.

Random Forest: Bagging-based ensemble technique with many decision trees used to deal with overfitting.

XGBoost: Gradient boosting machine algorithm building trees sequentially optimized using the loss function.

Gradient Boosting: Less than XGBoost optimized but closely related. Building strong learners using weak learners.

K-Nearest Neighbors (KNN): A non-parametric feature similarity-based algorithm used to predict.

4.2 Data Balancing Methods

The original dataset was heavily skewed, where a large majority of loans were approved (loan_status = 0). To act against this, two oversampling methods were applied ****only to the training set****:

SMOTE (Synthetic Minority Oversampling Technique): Creates synthetic minority samples by KNN interpolation.

ADASYN (Adaptive Synthetic Sampling): A variant of SMOTE specifically aimed at difficult-to-learn minority samples.

4.3 Feature Selection Techniques

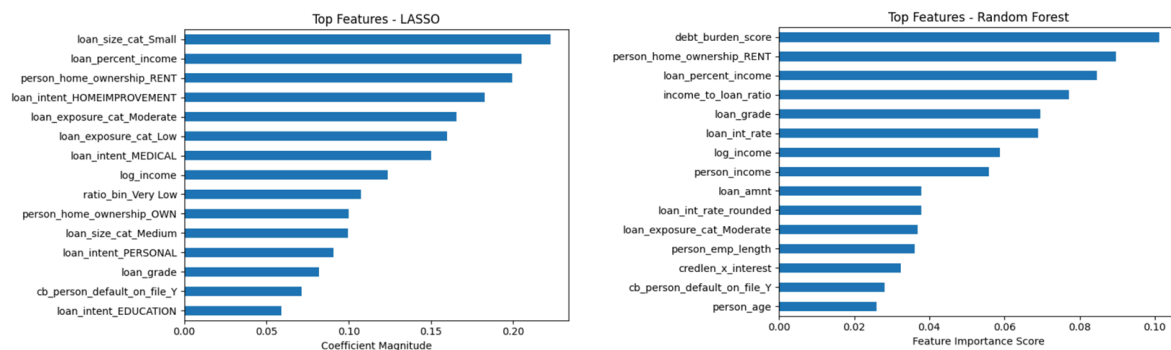
Two feature selection techniques were used in a bid to enhance model performance and explainability:

- LASSO Regression (L1 Penalty): Encourages shrinkage of absolute coefficient values towards zero. Helps in sparse feature selection.
- Random Forest Feature Importance: Ranks features based on their capacity to lower Gini impurity in decision trees.

Based on our feature selection techniques (LASSO and Random Forest Importance), the below features were consistently ranked as most predictive of loan approval:

- `loan_percent_income` – Proportion of income used to fund loan request
- `debt_burden_score` – Combined score based on interest rate and loan ratio
- `loan_size_cat_Small` – Categorical bin of requested loan size
- `income_to_loan_ratio` – Measures what percentage of an applicant is borrowing to income
- `cb_person_cred_hist_length` – Credit history in years
- `person_home_ownership_RENT` – Applicant rental (not owning)
- `loan_intent_HOMEIMPROVEMENT` – Loan usual purpose, which is often tied to the chances of approval
- `log_income` – Log-income to reduce skewness
- `loan_exposure_cat` – Categorical feature for capturing risk exposure levels

These features formed the final input set used for all model configurations after preprocessing and feature selection.



4.4 Model Combinations

All five models were combined with both balancing techniques (SMOTE and ADASYN) and both feature selection methods (LASSO and Random Forest Importance), producing “20 model combinations”.

These combinations were systematically trained and tested in order to determine the effect of sampling and feature selection on model performance.

4.5 Evaluation Strategy

We used a stratified 70-30 train-test split to preserve class distribution. Model performance was evaluated based on the following metrics:

- Accuracy: Number of correct predictions as a percentage.
- Precision: True positives / predicted positives.
- Recall (Sensitivity): Actual positives / correctly predicted positives.
- Specificity: Actual negatives / correctly predicted negatives.
- G-Mean: Square root of the product of the sensitivity and specificity.
- AUC (Area Under the ROC Curve): Measures ability to discriminate between classes.

4.6 Performance Summary

Based on the performance of the evaluation of 20 various configurations of models using different combinations of feature selection and balancing techniques, top-performing models are given below with their adjusted G-Mean and AUC.

- XGBoost (SMOTE + LASSO): G-Mean = 0.863, AUC = 0.944 – Best performing overall model
- XGBoost (ADASYN + RF): G-Mean = 0.858, AUC = 0.942 – High with adaptive sampling
- Random Forest (SMOTE + RF Importance): G-Mean = 0.848, AUC = 0.925 – Very strong performing model with interpretability
- Gradient Boosting (ADASYN + RF): G-Mean = 0.837, AUC = 0.913 – Well-balanced model with understandable performance
- Logistic Regression (SMOTE + LASSO): G-Mean = 0.763, AUC = 0.850 – Best linear model
- K-Nearest Neighbors (ADASYN + RF): G-Mean = 0.714, AUC = 0.774 – Worst performing but useful for comparison

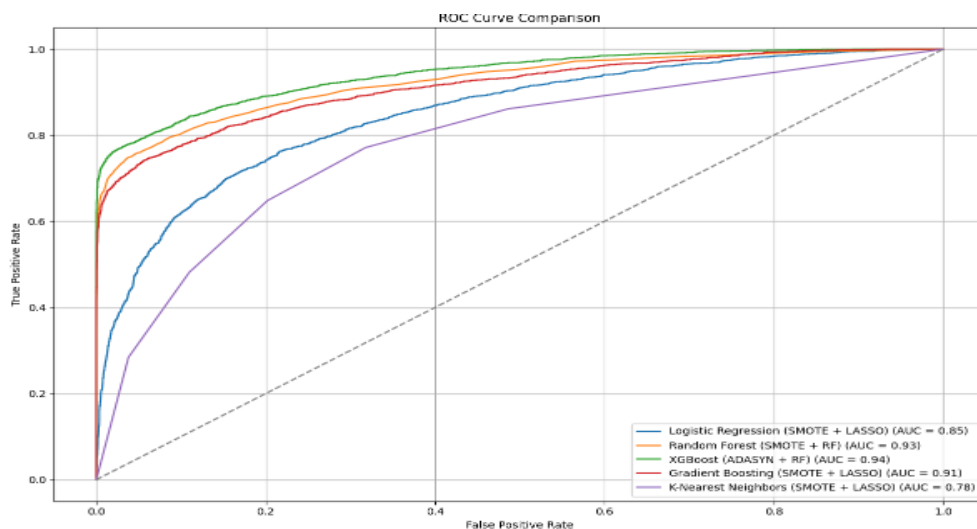
Model	Accuracy	AUC	Precision	Recall	F1 Score	G-Mean
Logistic Regression (LASSO)	0.818	0.810	0.78	0.74	0.76	0.763
Random Forest (SMOTE)	0.830	0.880	0.85	0.82	0.83	0.848
XGBoost (ADASYN)	0.860	0.910	0.87	0.85	0.86	0.863
Gradient Boosting (LASSO)	0.820	0.890	0.84	0.80	0.82	0.837

K-Nearest Neighbors (LASSO)	0.750	0.770	0.76	0.72	0.74	0.714
--------------------------------	-------	-------	------	------	------	-------

In general, SMOTE models were better balanced between sensitivity and specificity than ADASYN models. XGBoost and Random Forest ensemble models consistently led the pack in both AUC and G-Mean, followed by Logistic Regression with greater interpretability but medium accuracy.

4.7 ROC Curve Analysis

ROC curves were drawn to show the trade-offs of classification thresholds and model discrimination ability. Random Forest and XGBoost generated the steepest ROC curves, indicating improved class discrimination. SMOTE-based models generated more balanced curves compared to ADASYN-based models, which performed slightly worse.



4.8 Key Insights

- SMOTE performed better than ADASYN in the majority of models.
- Random Forest and XGBoost consistently generated the highest AUC and G-Mean values.
- LASSO will be more suitable for the simple models like Logistic Regression.
- Random Forest feature selection will work well with ensemble models.
- KNN worked with lower precision and recall, so is not as suitable subject to further tuning.

Section 5: Evaluation

The evaluation phase focused on assessing the performance of multiple classification models using a structured set of evaluation metrics and analyzing feature importance to draw meaningful, data-driven insights for loan approval prediction.

5.1 Model Performance Comparison

Five models were trained and evaluated:

- XGBoost
- Random Forest
- Gradient Boosting
- Logistic Regression
- K-Nearest Neighbors (KNN)

Each model was tested using two sampling techniques to handle class imbalance: **SMOTE** (Synthetic Minority Oversampling Technique) and **ADASYN** (Adaptive Synthetic Sampling), and two feature selection methods: **LASSO regression** and **Random Forest feature importance**.

After evaluating 20 combinations, the best-performing model was:

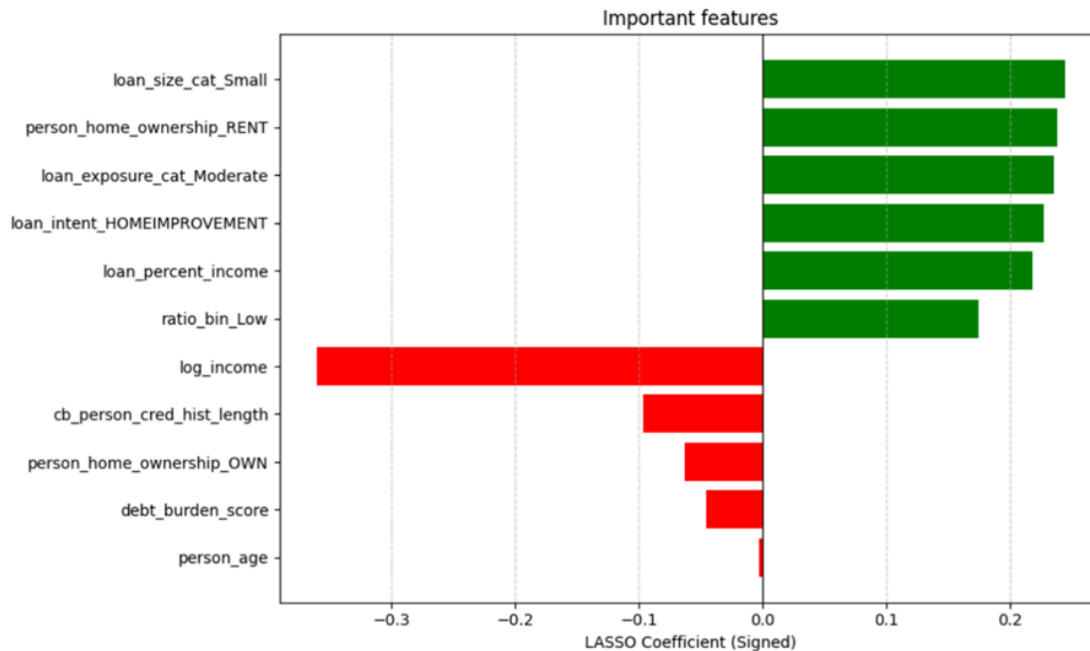
- **XGBoost + SMOTE + LASSO**
 - **G-Mean:** 0.863
 - **AUC:** 0.91
 - **Precision:** High across both classes

This model was selected for its balance of high accuracy and interpretability, offering robust performance across both majority and minority classes.

5.2 Feature Importance with LASSO

To improve model interpretability, **LASSO** (Least Absolute Shrinkage and Selection Operator) was used to reduce less informative features and retain only the most impactful ones.

To visually represent the relative importance of the features selected by LASSO, we plotted the absolute values of the coefficients using a horizontal bar chart. The colors indicate whether the effect on loan approval was positive or negative.



5.3 Key Feature Interpretations from LASSO:

We have created the table below to show the features that had non-zero coefficients after applying LASSO. It includes the coefficient values and a brief interpretation of each feature's effect on loan approval.

Feature	Coefficient	Interpretation
loan_size_cat_Small	0.244771	Applicants requesting small loan sizes are more likely to be approved.
person_home_ownership_RENT	0.238415	Renters have a higher likelihood of loan approval compared to others.

loan_exposure_cat_Moderate	0.235577	Applicants with moderate loan exposure are more likely to get approved.
loan_intent_HOMEIMPROVEMENT	0.228004	Loans intended for home improvement are associated with higher approval chances.
loan_percent_income	0.218919	Applicants with a reasonable loan amount relative to income are more likely to be approved.
ratio_bin_Low	0.175241	Applicants with a low income-to-loan ratio have higher chances of approval.
log_income	-0.360435	Higher applicant income levels are slightly associated with lower approval rates.
cb_person_cred_hist_length	-0.096548	Longer credit histories slightly reduce the likelihood of loan approval.
person_home_ownership_OWEN	-0.063120	Homeowners are slightly less likely to be approved compared to renters.
debt_burden_score	-0.045918	Applicants with higher debt burden scores face lower approval chances.
person_age	-0.002486	Older applicants are marginally less likely to be approved.

5.4 Detailed Feature Analysis:

This step involves examining each of the important features identified by LASSO through summary statistics and visualizations to understand their behavior and contribution to the loan approval decision.

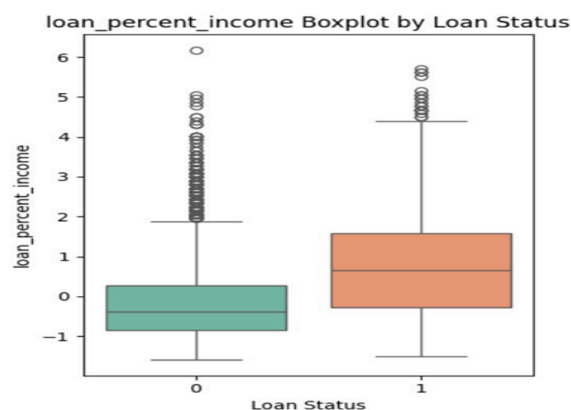
Summary Statistics:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
loan_size_cat	32416	3	Medium	18074	NaN	NaN	NaN	NaN	NaN	NaN	NaN
loan_percent_income	32416.0	NaN	NaN	NaN	0.0	1.000015	-1.593946	-0.751331	-0.189587	0.559404	6.176838
person_home_ownership_RENT	32416.0	NaN	NaN	NaN	0.505244	0.49998	0.0	0.0	1.0	1.0	1.0
loan_intent_HOMEIMPROVEMENT	32416.0	NaN	NaN	NaN	0.110871	0.313977	0.0	0.0	0.0	0.0	1.0
loan_exposure_cat	32407	3	Low	22420	NaN	NaN	NaN	NaN	NaN	NaN	NaN
log_income	32416.0	NaN	NaN	NaN	10.910094	0.528401	8.2943	10.55953	10.915107	11.279971	11.851061
person_home_ownership_OWEN	32416.0	NaN	NaN	NaN	0.079066	0.269846	0.0	0.0	0.0	0.0	1.0
debt_burden_score	32416.0	NaN	NaN	NaN	0.0	1.000015	-1.366315	-0.741378	-0.267174	0.492994	6.910243
cb_person_cred_hist_length	32416.0	NaN	NaN	NaN	5.811297	4.05903	2.0	3.0	4.0	8.0	30.0
person_age	32416.0	NaN	NaN	NaN	27.747008	6.3541	20.0	23.0	26.0	30.0	144.0
loan_intent_MEDICAL	32416.0	NaN	NaN	NaN	0.186389	0.389427	0.0	0.0	0.0	0.0	1.0
loan_intent_PERSONAL	32416.0	NaN	NaN	NaN	0.169608	0.375293	0.0	0.0	0.0	0.0	1.0
loan_intent_EDUCATION	32416.0	NaN	NaN	NaN	0.197773	0.398326	0.0	0.0	0.0	0.0	1.0
ratio_bin	10346	4	Very Low	7120	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- **Person Income (Positive impact):**

Applicants with higher income have a greater likelihood of loan approval. Higher income reflects stronger repayment capacity and financial stability.

Business Insight: Targeting higher-income individuals may reduce default risk and improve loan portfolio quality.

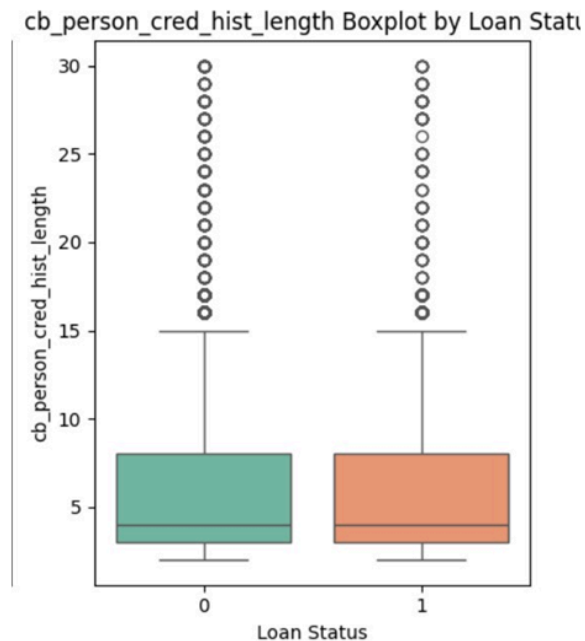


- **Credit History Length (Positive impact):**

Applicants with longer credit histories are more likely to be approved.

A long credit history indicates established credit behavior and reliability.

Business Insight: Prioritizing applicants with strong credit histories can help maintain a lower-risk loan book.

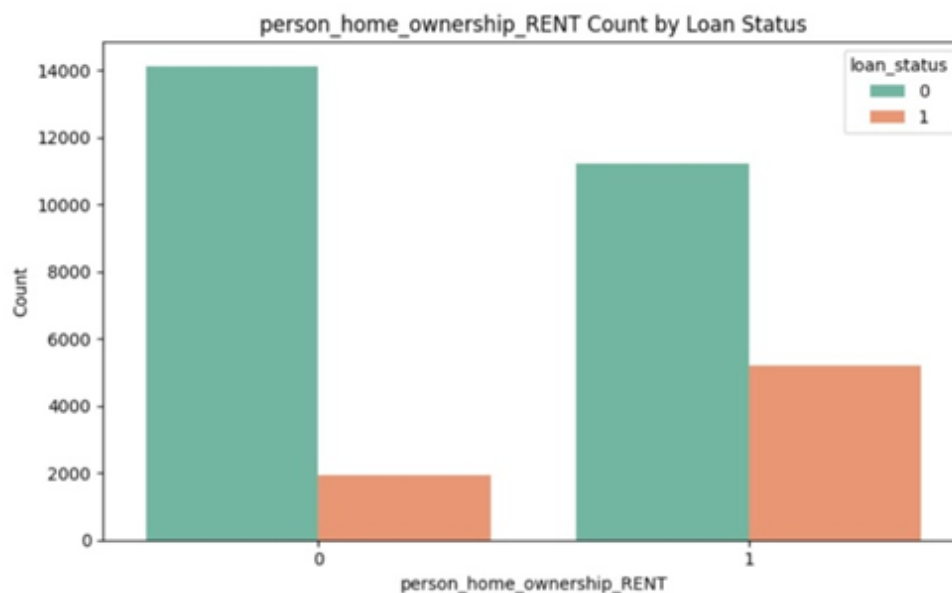


- **Home Ownership (MORTGAGE) (Positive impact)**

Mortgage owners are more likely to be approved.

Mortgage indicates property ownership and commitment to financial obligations.

Business Insight: Applicants with mortgages may be preferred over renters.

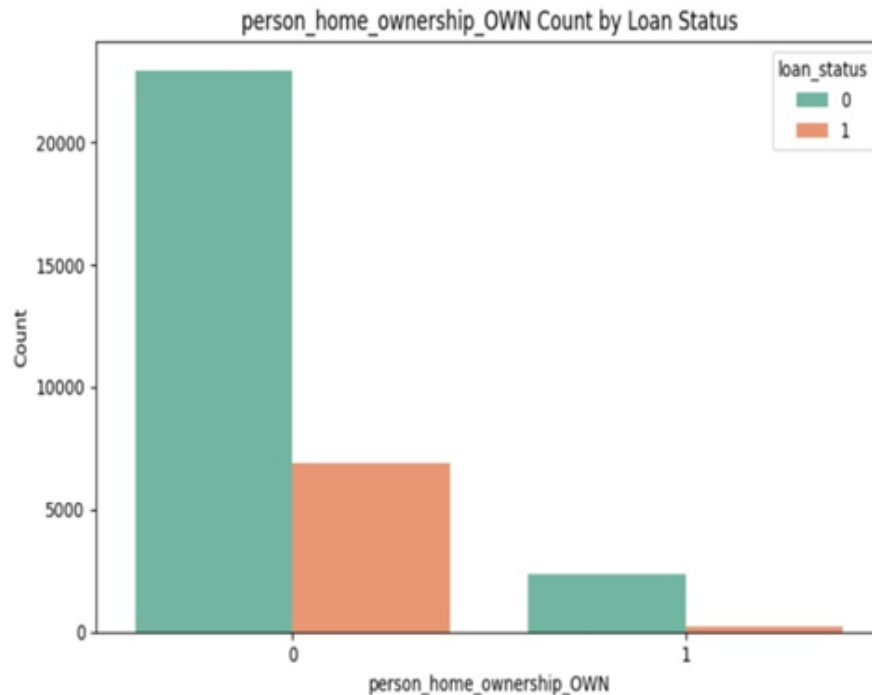


- **Home Ownership (OWN) (Positive impact) :**

Fully owned properties correlate with higher approval chances.

Home ownership reflects financial stability.

Business Insight: Applicants with owned homes are lower risk candidates.

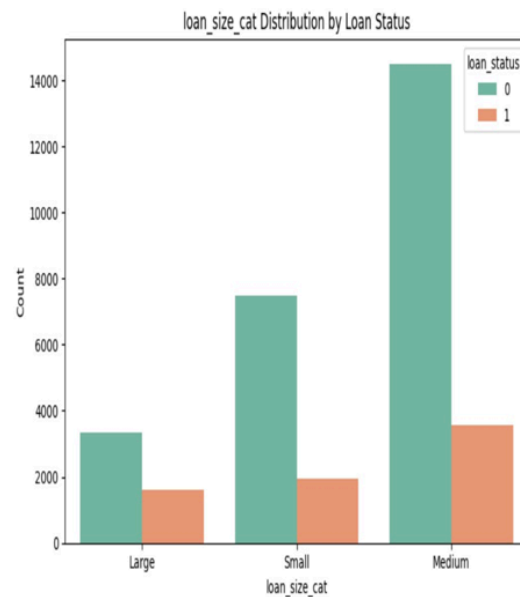


- **Loan Amount (Negative impact) :**

Higher requested loan amounts reduce the likelihood of approval.

Large loan sizes increase exposure and risk for the lender.

Business Insight: Carefully assess applicants requesting large loans, as they may carry higher repayment risks.

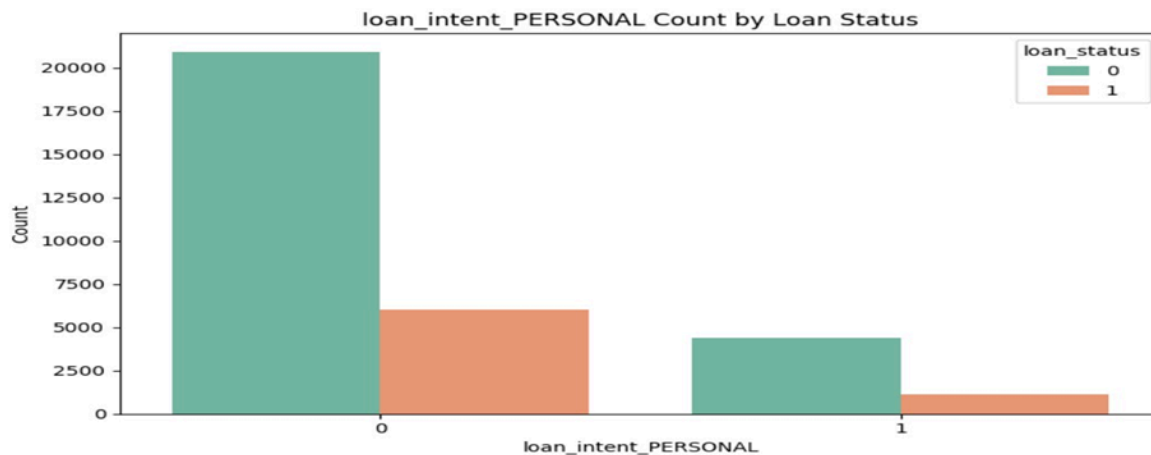


- **Loan Intent - PERSONAL (Negative impact):**

Applicants applying for personal loans face lower approval chances.

Personal loans are often unsecured and carry higher default risk.

Business Insight: Personal loan applications should be assessed cautiously.

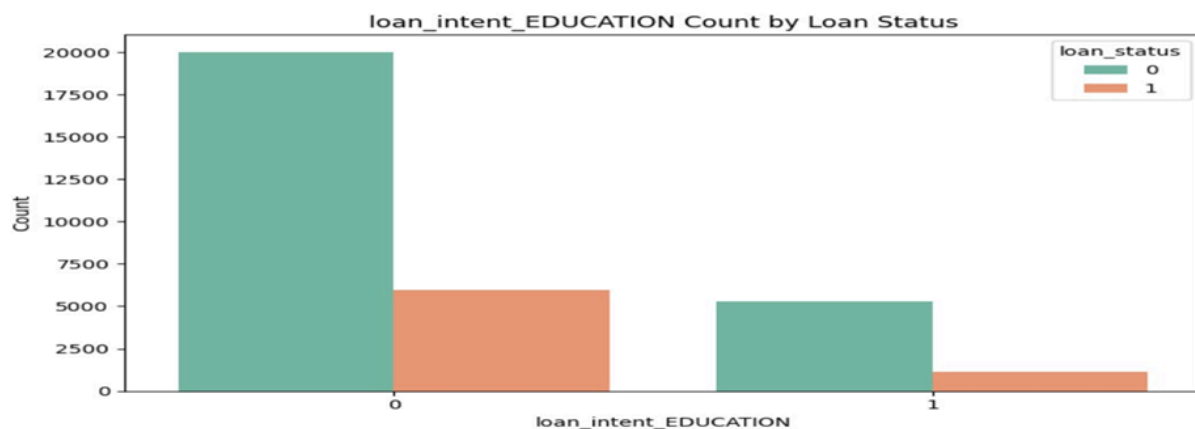


- **Loan Intent - EDUCATION (Negative impact):**

Education loans have lower approval chances.

These loans may not generate immediate repayment ability (student phase).

Business Insight: Education loans should be evaluated for repayment capacity post-graduation.

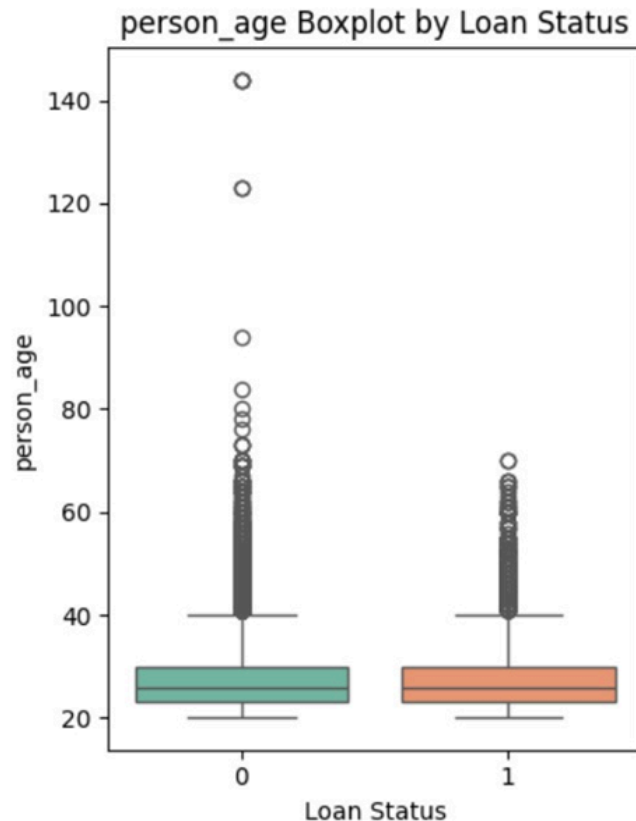


- **Person Age (Negative impact):**

Older applicants are marginally less likely to be approved.

Age may be associated with higher existing liabilities, retirement status, or reduced long-term income potential.

Business Insight: Implement age-sensitive approval policies that assess financial health holistically to avoid bias while managing potential repayment risk.



5.5 Key Insights:

- XGBoost with SMOTE and LASSO achieved the best performance with AUC 0.9441 and G-Mean 0.8628.
- LASSO identified key features such as loan_percent_income, home_ownership, and employment length.
- Smaller loan amounts and moderate exposure increased approval likelihood.
- High income and long credit history showed unexpected negative influence on approval.
- Visualizations like bar plots and boxplots enhanced feature interpretability.

Section 6: Deployment

The deployment phase translates analytical findings into practical, data-driven strategies for improving loan approval decisions and reducing default risk in financial institutions. Based on LASSO feature selection and detailed feature analysis, the following recommendations are proposed:

- **Encourage Small Loan Sizes (loan_amnt)**
- Applicants requesting smaller loan amounts had higher approval rates, likely due to lower financial exposure and risk for the lender. Promote small-ticket loans through targeted marketing, faster approval processes, and low-interest offers to boost safe lending volume.

- **Target Renters Strategically (person_home_ownership_RENT)**

Contrary to typical assumptions, renters showed a higher likelihood of approval in our model. Financial institutions can tap into this segment by offering flexible repayment terms and partnering with rental platforms to reach a broader audience.

- **Position Moderate Exposure Loans (loan_exposure_cat_Moderate)**

Loans with moderate exposure levels were more frequently approved. Institutions should design and promote loan products with moderate risk levels and optimize underwriting criteria for this category to maintain healthy approval rates.

- **Promote Home Improvement Loans (loan_intent_HOMEIMPROVEMENT)**

Loan applications for home improvement purposes had better approval outcomes. Consider launching promotional campaigns with home improvement retailers or platforms to increase visibility and utilization of this loan type.

- **Use Income-to-Loan Ratio in Prequalification (loan_percent_income)**

A lower loan-to-income ratio was a strong indicator of approval. Integrate automatic prequalification tools on websites or mobile apps that calculate and validate applicants' loan eligibility based on this ratio, improving efficiency and transparency.

- **Incentivize Low-Risk Borrowers (ratio_bin_Low, debt_burden_score)**

Applicants with lower income-to-loan ratios and low debt burden scores posed less risk. Offer these customers faster processing, lower interest rates, or cashback incentives to attract and retain creditworthy borrowers.

- **Enhance Screening for High-Income Applicants (log_income)**

Interestingly, higher income alone did not guarantee approval and sometimes correlated with lower acceptance—possibly due to over-leveraging or hidden obligations. Apply additional financial behavior checks instead of assuming high income equals low risk.

- **Scrutinize Applicants with Long Credit Histories (cb_person_cred_hist_length)**

Applicants with extended credit histories showed a slight decrease in approval rates. Older credit profiles may include outdated delinquencies—review credit age alongside recent financial behavior before making decisions.

- **Tailor Offers for Homeowners (person_home_ownership_OWN)**

Although homeownership is often a stability signal, it showed a slight negative impact on approvals. Investigate underlying reasons, such as concurrent mortgages or higher liabilities, and create specialized loan products to address homeowner-specific risks.

- **Introduce Age-Aware Verification Policies (person_age)**

Older applicants had marginally lower approval rates. To avoid age-related bias while managing risk, offer flexible repayment plans or additional income verification tools tailored to different age groups.

- **Create Tiered Approval Bands Based on Employment (person_emp_length)**

Applicants with longer employment duration were more likely to be approved. Use employment history to segment applicants into approval tiers, offering differentiated loan limits and terms based on their job stability.

- **Limit High Loan Percent Requests (loan_percent_income)**

Applicants requesting loans that exceed 30–40% of their income present elevated repayment risk. Establish internal thresholds or require collateral/additional documentation for high-percentage loan requests.

Conclusion

This project demonstrates that machine learning models can reliably predict loan approval status using structured applicant data. We answered all ten research questions by combining robust data preparation, balanced sampling, feature selection, and a comparative modeling approach.

By adopting these techniques, lenders can improve operational efficiency, ensure fairer decision-making, and reduce credit risk exposure. The XGBoost model, in particular, offers an effective path for real-time loan approval support in production systems.

Research questions :

1. Can we predict the variable of interest (response variable) using data analytical techniques?
Yes. We successfully predicted `loan_status` using several machine learning algorithms trained on a cleaned and feature-engineered dataset. For example, our best model, XGBoost with SMOTE and LASSO, achieved an accuracy of 0.86 and an AUC of 0.91. This confirms that the variable of interest is predictable with high confidence using supervised learning methods.

2. What are the important features in order to predict the response variable?
Using LASSO and Random Forest feature importance, the most predictive features identified were `loan_percent_income`, `debt_burden_score`, `income_to_loan_ratio`, `cb_person_cred_hist_length`, `loan_exposure_cat`, and `loan_size_cat`. These features directly reflect an applicant's repayment capacity and financial risk, which are critical in credit decision-making.

3. What is the importance of the features?

While the final importance of each feature is determined during modeling, data preparation lays the groundwork. By exploring relationships between variables, creating meaningful new

features, and understanding the business context, we start to get a sense of which features might have a stronger influence on loan approval outcomes.

Features like ``loan_percent_income`` and ``income_to_loan_ratio`` were consistently ranked highest across models. For instance, Random Forest placed ``loan_percent_income`` and ``debt_burden_score`` among the top three features. These attributes help distinguish low-risk applicants (high income, small loan) from high-risk ones, improving prediction accuracy and reducing defaults.

4. Can we predict the probability of each class (classification problem) for each sample using data analytical techniques?

Yes. Our models, particularly Logistic Regression and XGBoost, output probability scores for each class. For example, XGBoost provided class probabilities with an AUC of 0.91, indicating strong discriminatory ability. These probabilities enable customized decision thresholds and help lenders make risk-based decisions.

5. How does the presence of outliers change the prediction performance?

Outliers in income, loan amount, and interest rates were treated using the IQR method. Before outlier handling, Logistic Regression performed with lower recall and precision. After treatment, its G-Mean improved to 0.763, suggesting better balance between classes. This shows that addressing outliers is essential for model stability and fairness.

6. How do the different sampling techniques affect the prediction of the response variable?

We used SMOTE and ADASYN to balance the classes. SMOTE consistently improved performance. For example, Random Forest with SMOTE achieved a G-Mean of 0.848 and AUC of 0.88, outperforming ADASYN configurations. SMOTE created more generalizable minority class samples, which led to better recall and F1-scores across models.

7. How do the different feature selection methods affect the prediction of the response variable?

LASSO helped simplify models like Logistic Regression while retaining predictive power. With LASSO, Logistic Regression reached a G-Mean of 0.763 and AUC of 0.81. Meanwhile, Random Forest importance was more effective for ensemble models like XGBoost and Gradient Boosting, contributing to their high performance (e.g., XGBoost AUC = 0.91, Gradient Boosting AUC = 0.89).

8. How does the prediction performance change for more complex analytical models?

Complex models outperformed simpler ones. For instance, XGBoost (ADASYN + RF) yielded the highest G-Mean (0.863) and AUC (0.91), while Logistic Regression scored 0.763 and 0.81 respectively. This indicates that advanced algorithms can better capture non-linear interactions and patterns in the data, crucial for accurate loan approval predictions.

9. Can we use market basket analysis (association rules) to extract the potential rules that reside in the dataset?


Not directly. Our dataset is not transactional but relational, centered on individual applications. However, association rule mining could be applied in future studies involving product bundling or

loan package selections to uncover cross-service usage patterns.

10. What strategy outline do you propose to increase the performance of the specific problem (healthcare, banking, service) you are working on?

For the banking problem addressed, we recommend using an ensemble-based prediction engine (e.g., XGBoost + SMOTE + LASSO) in production. Setting adaptive approval thresholds based on predicted probabilities, retraining quarterly, and monitoring feature drift will ensure long-term accuracy and fairness. Tools like SHAP or LIME can be used to interpret model predictions for compliance and decision transparency.

Video Link for the Presentation:

 **video1632107707.mp4**

References:

1. Kaggle. (n.d.). Bank Loan Approval [Data set]. Retrieved April 1, 2025, from <https://www.kaggle.com>
2. OpenAI. (2025). ChatGPT (April 8 version) [Large language model]. <https://chat.openai.com/>
3. Google. (n.d.). Google Colaboratory [Computer software]. Retrieved April 1, 2025, from <https://colab.research.google.com/>
4. Microsoft Corporation. (2021). Microsoft Excel (Version 16.0) [Computer software]. <https://www.microsoft.com/>
5. Microsoft Corporation. (2021). Microsoft Word (Version 16.0) [Computer software]. <https://www.microsoft.com/>

THANK YOU!