Spring 2025: Strategic Capstone Project

Final Project Assignment

Bank Loan Approval Prediction

May 13th 2025

ADELPHI

Team Members

Roja Eslavath

Yamika Ratna Kadiyala

Group - 5

Amarnath Dasari

Suhitha Yalamanchili

Han Ru Wu

ADELPHI

Project Overview

This presentation will take you through our end-to-end capstone project based on the CRISP-DM framework.

We applied machine learning techniques to predict loan approval outcomes using real-world financial data.

Here's what we'll cover:

- Business Understanding: Problem definition, background, and project objectives
- Data Understanding: Exploratory Data Analysis and insights from the dataset
- Data Preparation: Cleaning, handling missing values, and feature engineering
- Modeling: Model selection, resampling, and feature selection
- Evaluation: Results, metrics comparison, and model performance
- Deployment: Actionable business recommendations based on data-driven insights

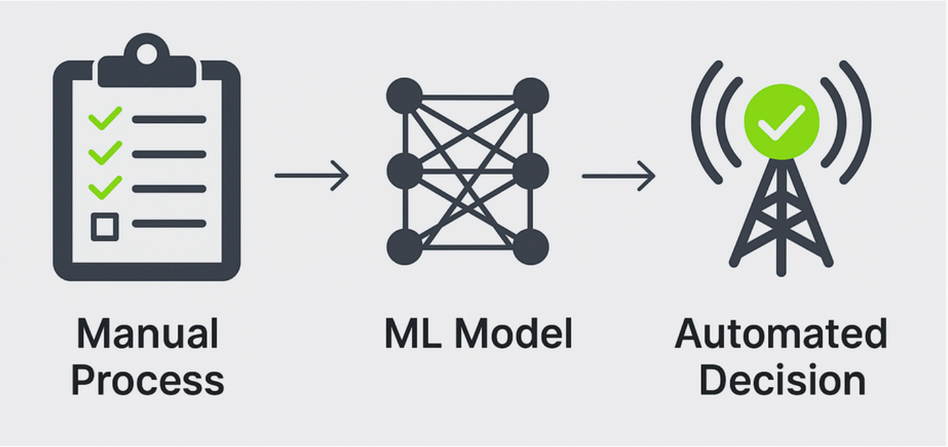
ADELPHI

Section 1 - Business Understanding

Problem & Objective

- Banks face challenges in processing loan applications quickly and fairly.
- Manual loan approvals are time-consuming and often inconsistent.
- Objective: Build a machine learning model to predict whether a loan should be approved or rejected.
- The model uses key applicant features:

1. Income
2. Employment length
3. Home ownership
4. Loan amount
5. Interest rate
6. Credit history



- This helps support faster, more accurate, and data-driven lending decisions.

Importance of the Project & Real-World Impact

- Reduces financial losses by detecting high-risk applicants early.
- Speeds up loan processing, improving customer experience.
- Brings fairness and transparency into the approval process.
- Enables banks to scale and automate lending operations effectively.
- Our work is structured using the CRISP-DM framework to ensure a clear, analytical approach.

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

ADELPHI

Section 2 – Data Understanding

Dataset Overview

- Dataset: 32,581 loan applications with 12 variables
- Target variable: loan_status (1 = good loan, 0 = bad loan)
- Features include personal and financial details like:
- Age, Income, Loan Amount, Employment Length
- Interest Rate, Loan Purpose, Home Ownership, Credit History
- Goal: Understand the behavior of each variable using EDA to guide modeling

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

Exploratory Data Analysis – Key Findings

Numerical Features:

- Most applicants are aged 20–30; some outliers exist (e.g., age 144)
- Income appears uniformly distributed (likely simulated)
- Most loans are between $5K–$12K with interest rates from 7%–13%
- Loan percent of income usually stays below 20% (suggesting responsible borrowing)

Categorical Features:

- Most applicants are renters or mortgage holders
- Most common loan purpose is Education, followed by Medical and Venture
- Grade A and B loans dominate the dataset
- Around 82% of applicants have no prior default

Variable	Mean	Std Dev	Min	25%	50%	75%	Max
person_age	27.73	6.35	20.0	23.0	26.0	30.0	144.0
person_income	66074.85	61983.12	4000.0	38500.0	55000.0	79200.0	6000000.0
person_emp_length	4.79	4.14	0.0	2.0	4.0	7.0	123.0
loan_amnt	9589.37	6322.09	500.0	5000.0	8000.0	12200.0	35000.0
loan_int_rate	11.01	3.24	5.42	7.9	10.99	13.47	23.22
loan_status	0.22	0.41	0.0	0.0	0.0	0.0	1.0
loan_percent_income	0.17	0.11	0.0	0.09	0.15	0.23	0.83
cb_person_cred_hist_length	5.8	4.06	2.0	3.0	4.0	8.0	30.0

Variable	Unique Values	Most Common Value
person_home_ownership	4	RENT
loan_intent	6	EDUCATION
loan_grade	7	A
cb_person_default_on_file	2	N

ADELPHI

Variable Relationships & Data Quality

- Renters showed higher default rates than mortgage holders
- Home improvement loans had higher default rates than others
- Prior default (cb_person_default_on_file = Y) was a strong predictor of loan rejection
- Employment length and loan amount had weaker impact on approval
- Missing values found in loan_int_rate and person_emp_length; handled using median imputation
- Outliers in income and employment length capped for cleaner model training.

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

ADELPHI

Section 3 – Data Preparation

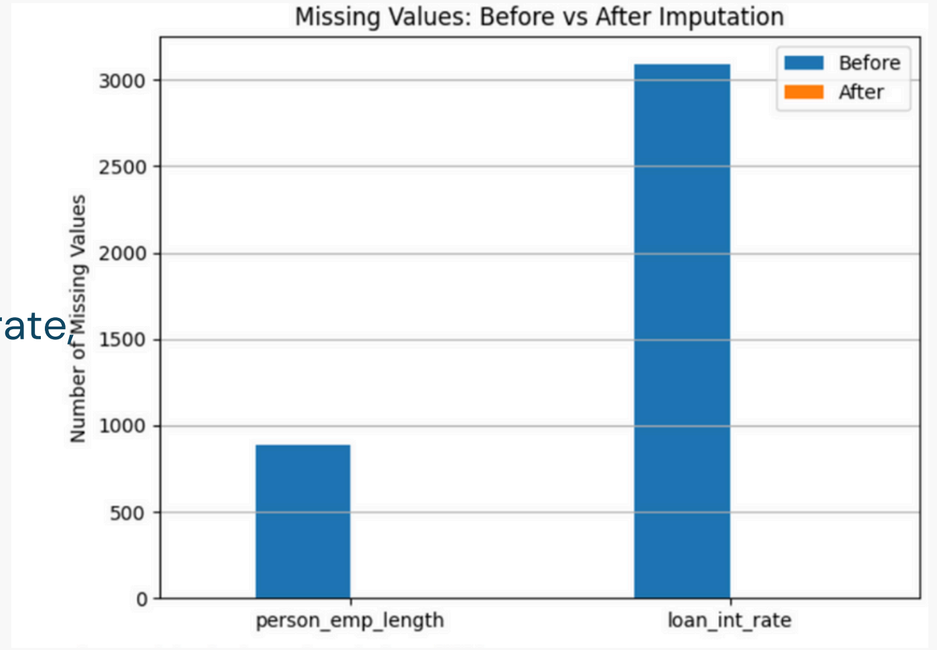
Cleaning & Structuring

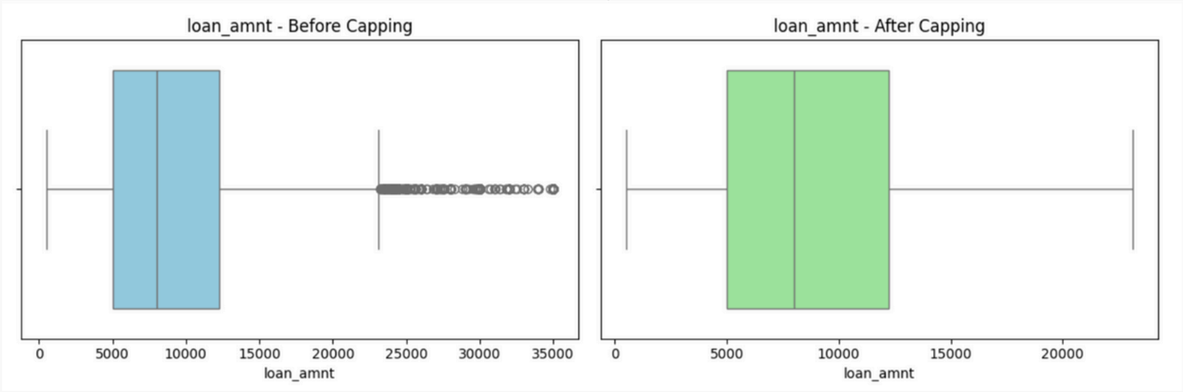
- Removed 165 duplicate entries to ensure data consistency.
- Treated missing values:

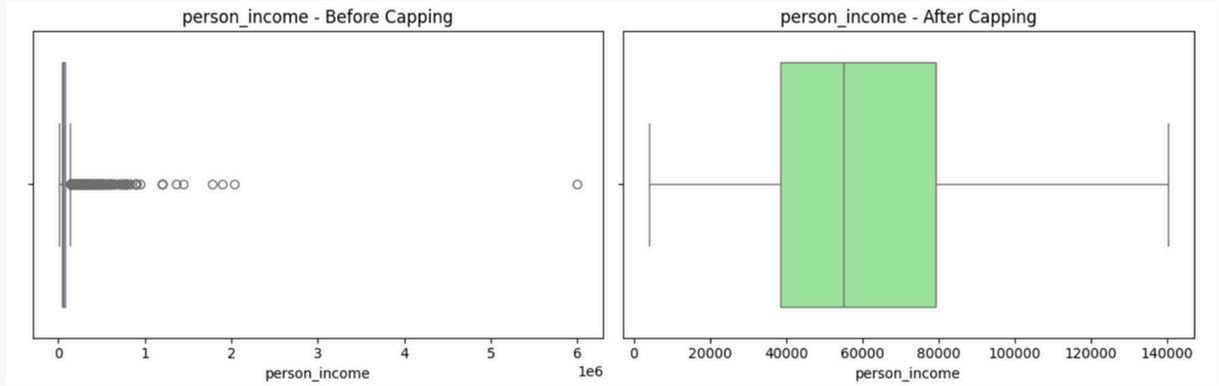
1. Categorical → Replaced with 'Unknown'
2. Numerical → Imputed using median (e.g., loan_int_rate, person_emp_length)

- Encoding applied:

3. Ordinal Encoding → loan_grade
4. One-Hot Encoding → home_ownership, loan_intent, default_flag







- Outliers handled in loan_amnt, person_income, and loan_int_rate using IQR method.

Feature Creation & Standardization

- Created new features to enhance model insight:

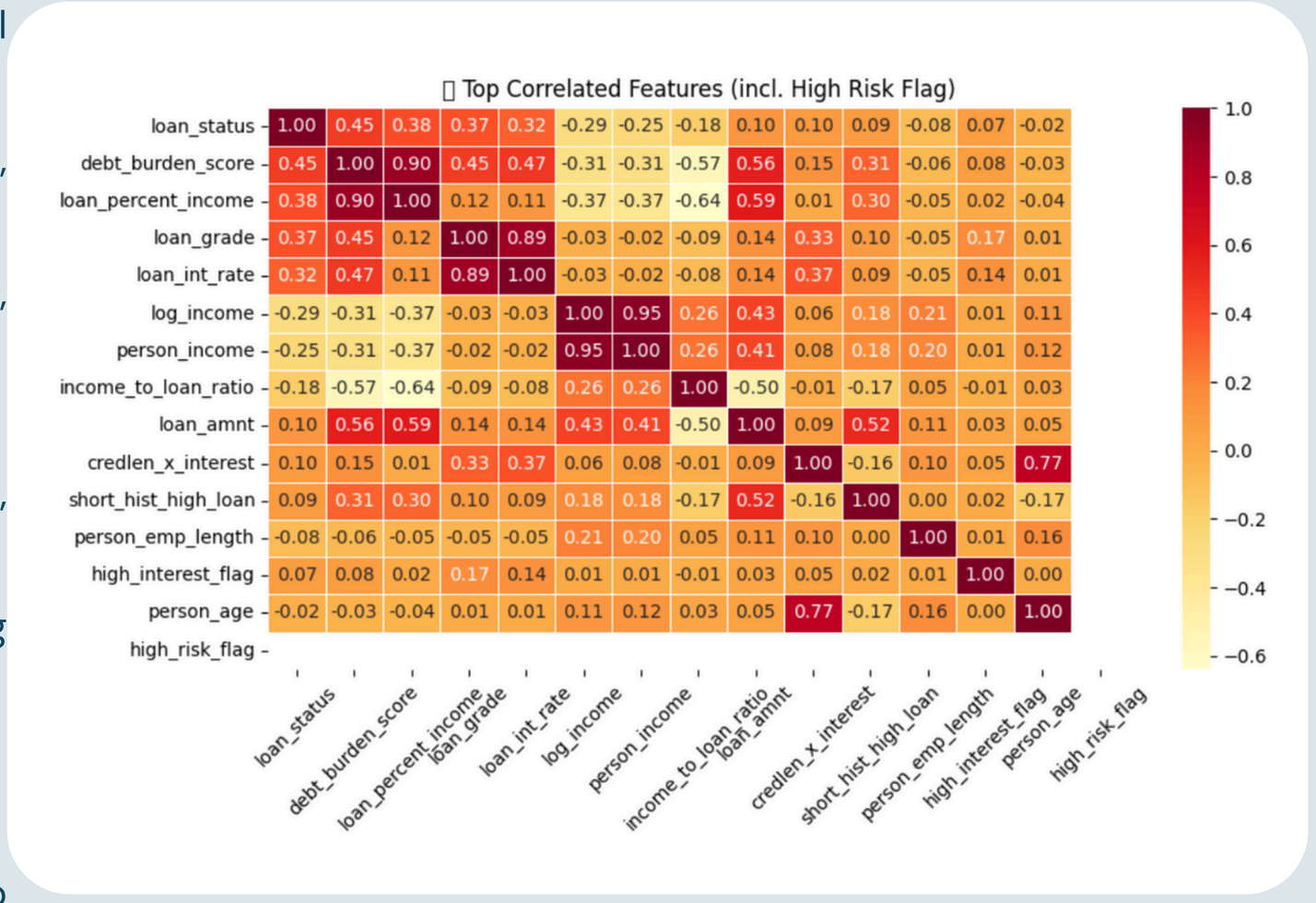
1. income_to_loan_ratio, debt_burden_score, log_income
2. Risk flags like high_risk_flag, short_hist_high_loan

- Binned features for interpretability:

3. cred_hist_cat, loan_size_cat, loan_exposure_cat

- Standardized numerical features using StandardScaler for equal model contribution.
- Smoothed noisy data by:

4. Rounding loan_int_rate
5. Grouping scores into bins (Very Low to Extreme)



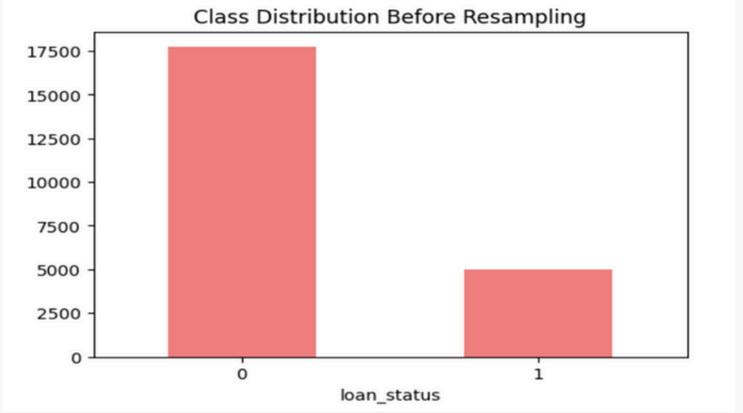
Balancing the Imbalanced Class:

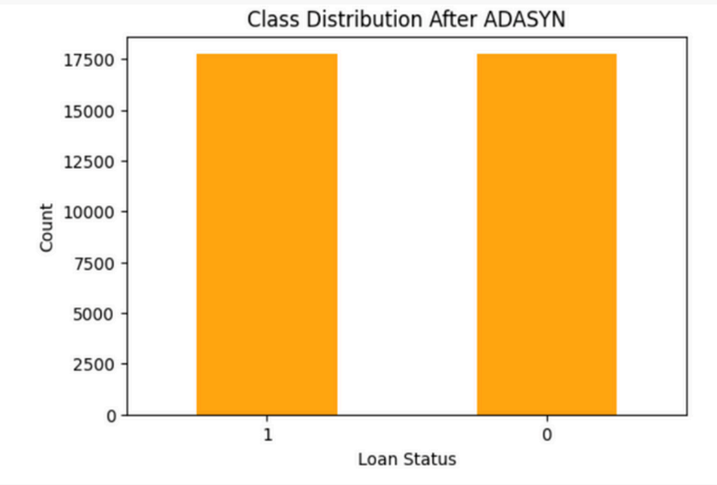
Class Imbalance Challenge: The dataset was highly imbalanced with only 22% of loans classified as good. Addressing this imbalance was crucial for improving model fairness and recall.

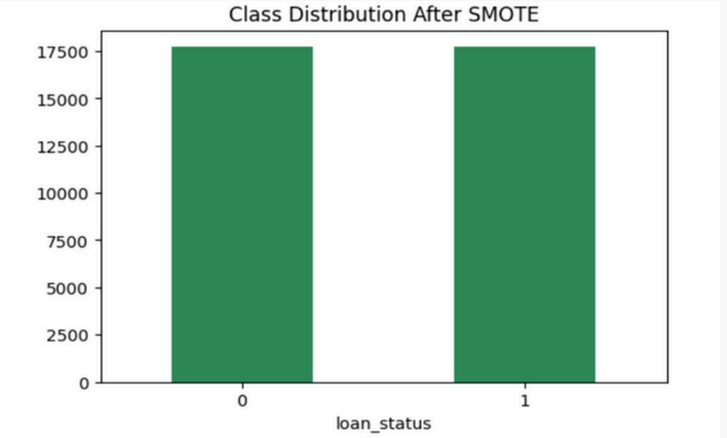
Balancing Techniques Applied:

- Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance class distribution
- SMOTE created synthetic samples for the minority class using nearest neighbors
- Resulted in 17,729 samples for both approved and not approved classes
- Addressed the imbalance to ensure equal representation in training data
- Visualized class distribution before and after SMOTE to confirm balance
- Resampling was done only on training data to prevent leakage
- Helped reduce model bias and improved fairness in predictions
- Applied **ADASYN (Adaptive Synthetic Sampling)** for a dynamic resampling approach
- Focused on generating more samples where the model struggled to learn
- Resulted in 18,890 samples for the minority class and 17,729 for the majority class
- Targeted difficult examples to enhance the model's learning capability
- Visualized updated class distribution after ADASYN application
- Resampling was applied only to training data for model integrity
- Improved classification performance on previously underrepresented cases

ADELPHI







Feature Selection

Feature Selection After SMOTE (LASSO & Random Forest):

•Applied LASSO Regression with cross-validation on SMOTE-balanced training data to identify top predictive features.

•LASSO automatically removed features with low impact by shrinking their coefficients to zero.

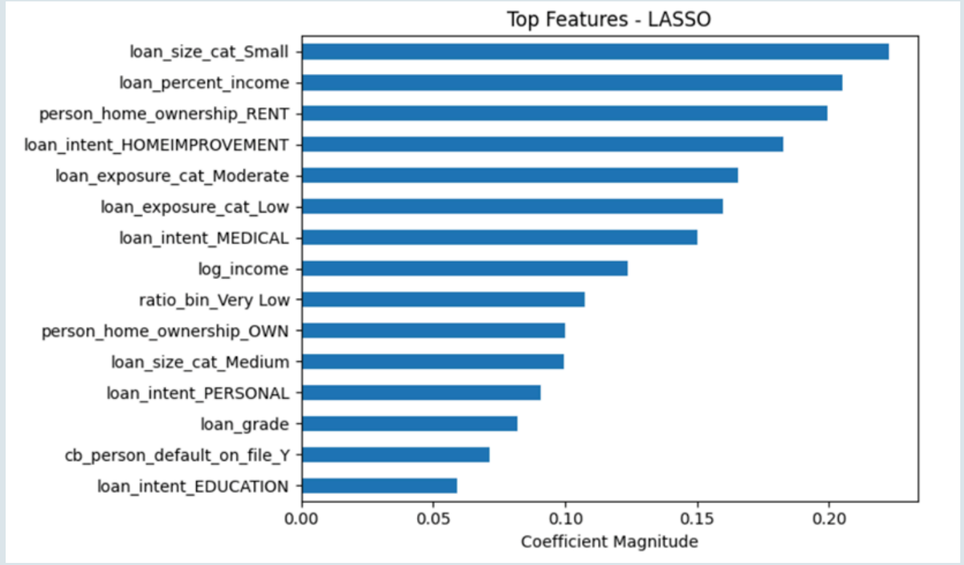
•Top LASSO features included:
loan_size_cat_Small, loan_percent_income, home_ownership_RENT, and loan_intent_HOMEIMPROVEMENT.

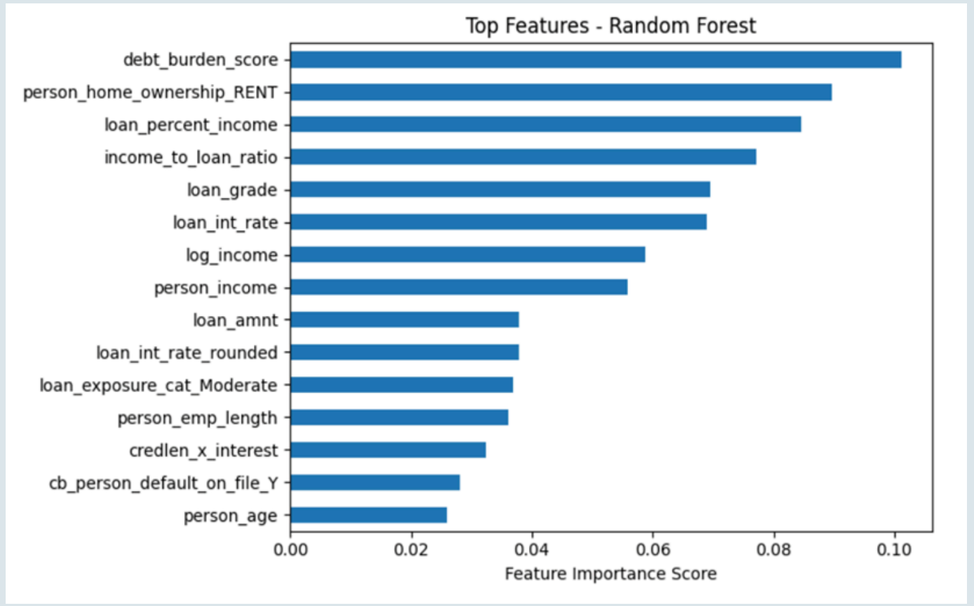
•Trained a Random Forest Classifier on the same SMOTE-balanced data to assess feature importance.

•Top Random Forest features included:
debt_burden_score, home_ownership_RENT, loan_percent_income, and income_to_loan_ratio.

•Generated feature importance visualizations for both LASSO and Random Forest to compare rankings.

•These selected features will be used for final model building and to enhance performance and interpretability.





Feature Selection After ADASYN (LASSO & Random Forest):

•Applied LASSO Regression with 5-fold cross-validation on the ADASYN-balanced training data to identify key predictive features.

•LASSO removed features with zero impact and retained the most relevant variables for prediction.

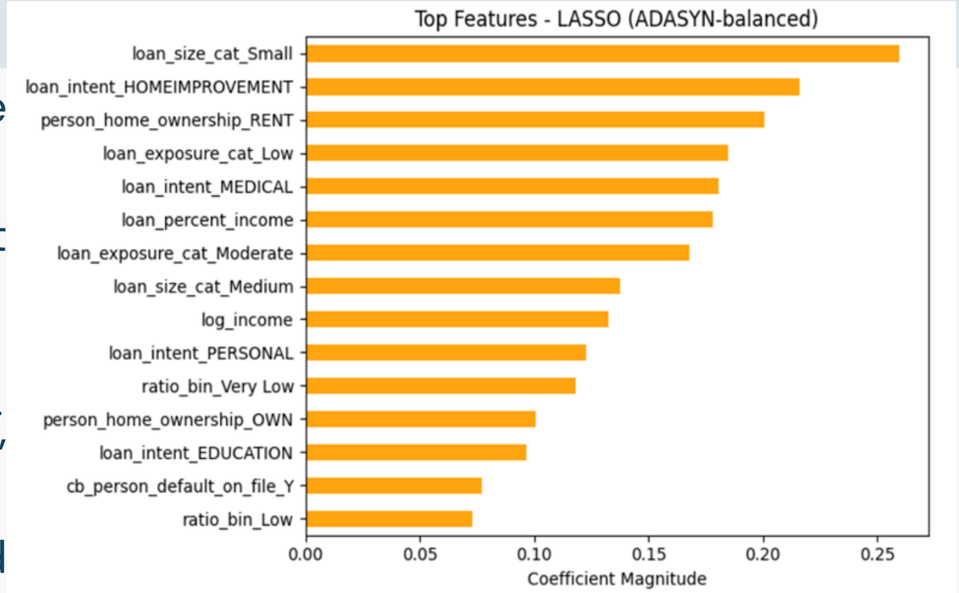
•Top LASSO features included:
loan_size_cat_Small, loan_percent_income, home_ownership_RENT, loan_intent_EDUCATION, and log_income.

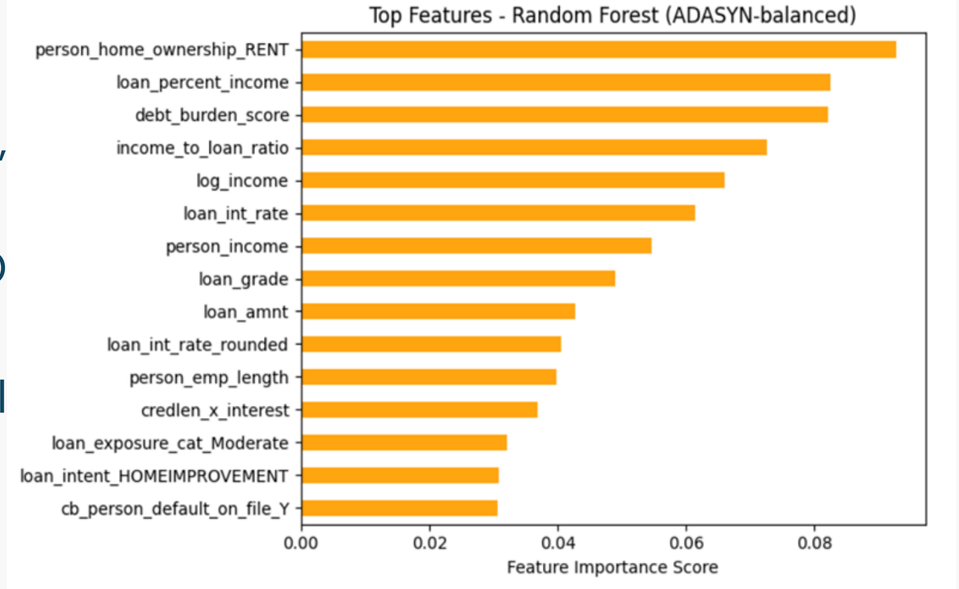
•Trained a Random Forest Classifier on the same ADASYN-balanced data to rank features by importance.

•Top Random Forest features included:
debt_burden_score, home_ownership_RENT, loan_percent_income, loan_int_rate, and income_to_loan_ratio.

•Generated bar plots comparing feature rankings from both LASSO and Random Forest.

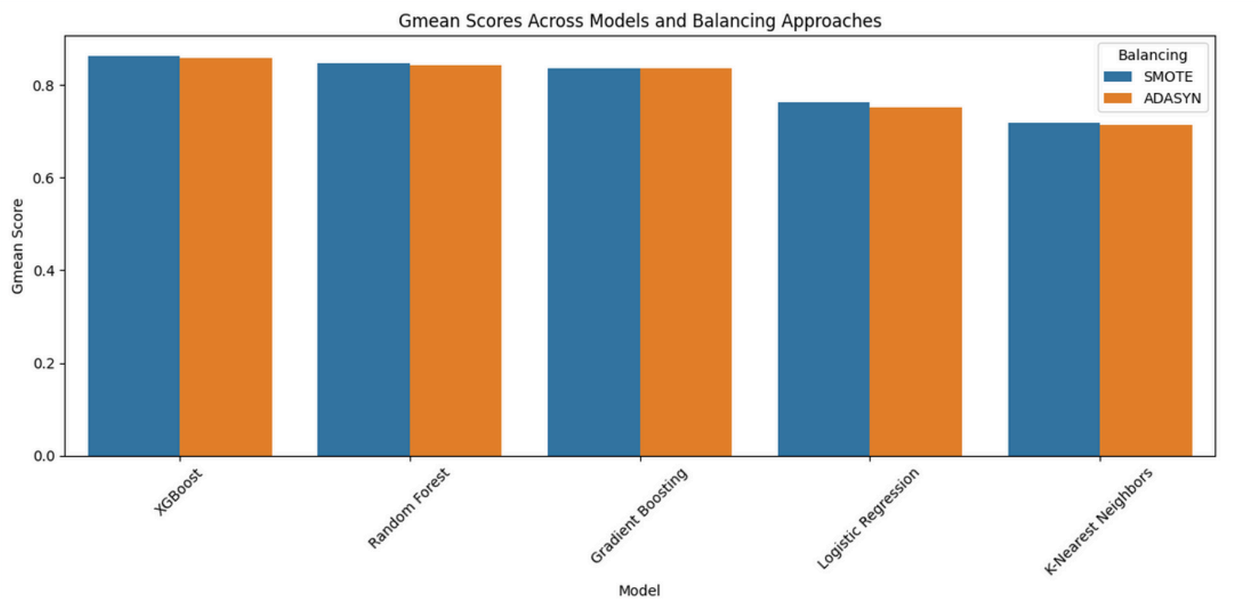
•These insights will support feature prioritization during final model development and help improve generalization.





Section 4 - Modeling

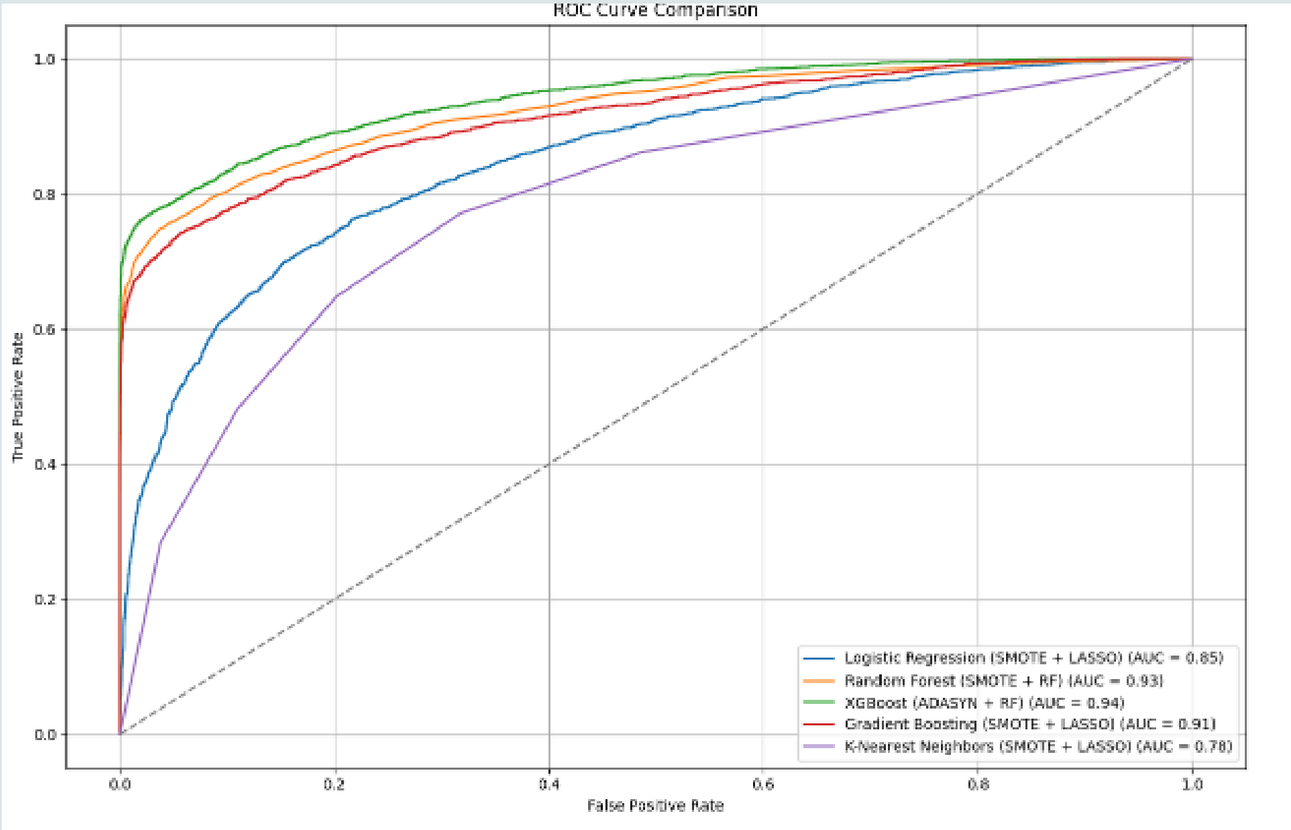
- Objective: Build robust and interpretable machine learning models to predict whether a bank loan should be approved, using balanced and feature-reduced data.
- Models Implemented:
  a. Logistic Regression – Baseline linear classifier.
  b. Random Forest – Bagging-based ensemble method.
  c. XGBoost – High-performing boosting algorithm.
  d. Gradient Boosting Classifier – Sequential optimization model.
  e. K-Nearest Neighbors – Simple distance-based learner.
- All models were trained using a stratified 70–30 train-test split and evaluated on multiple performance metrics.



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

ROC Curve Comparison

- *The ROC Curve illustrates model performance across different thresholds.*
- *XGBoost and Gradient Boosting achieved the highest AUC (0.91), showing excellent class separability.*
- *Random Forest also performed strongly (AUC = 0.93).*
- *Logistic Regression and KNN lagged behind in sensitivity at most thresholds.*
- *This visual confirms ensemble models are more confident and reliable in distinguishing approved vs. rejected loans.*



ADELPHI

Section 5 - Model Evaluation

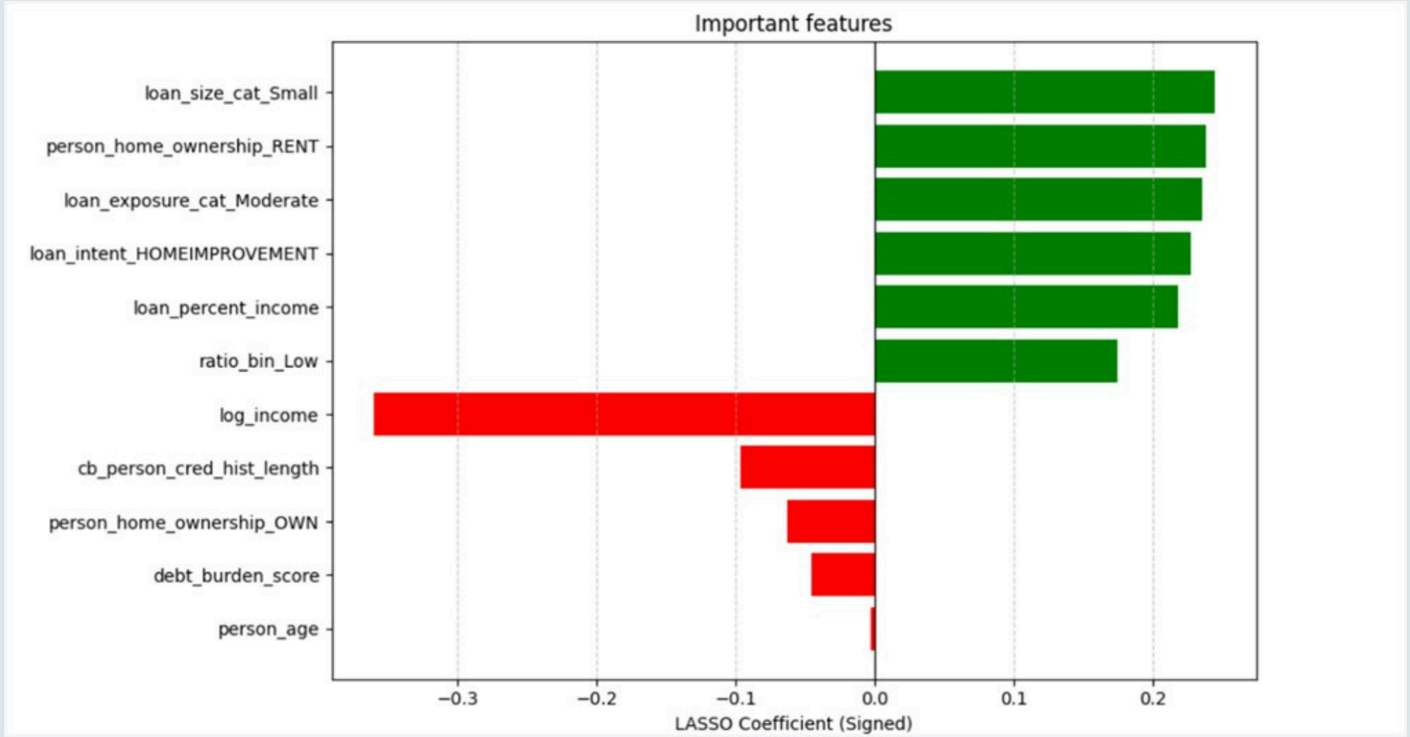
Evaluation Phase – Model Performance

- Evaluated 5 models: XGBoost, Random Forest, Gradient Boosting, Logistic Regression, and KNN
- Used SMOTE and ADASYN for class balancing
- Applied LASSO and Random Forest for feature selection
- Best Model: XGBoost + SMOTE + LASSO
- G-Mean: 0.863
- AUC: 0.91
- Precision: High across both classes
- Selected for balanced performance and interpretability

ADELPHI

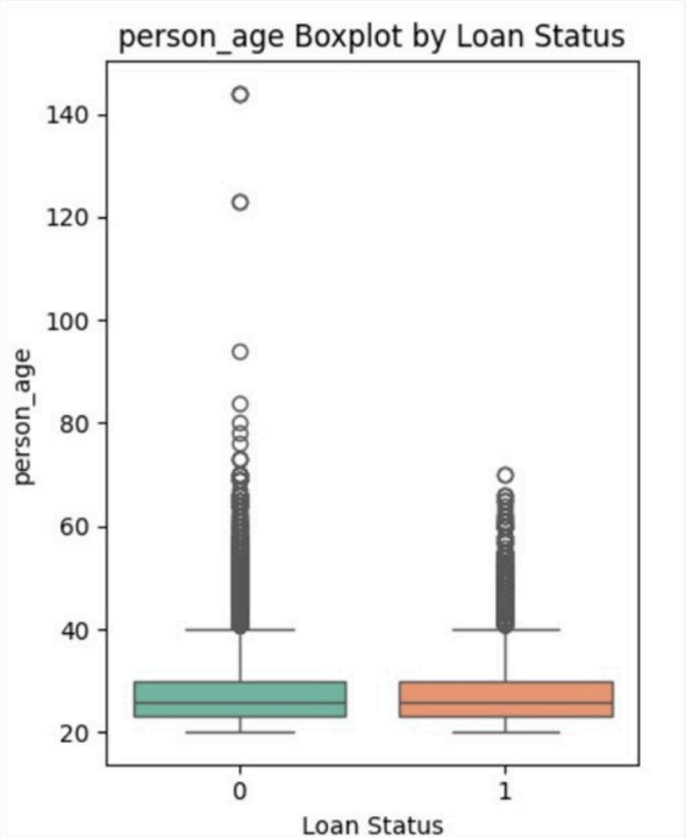
LASSO Feature Importance

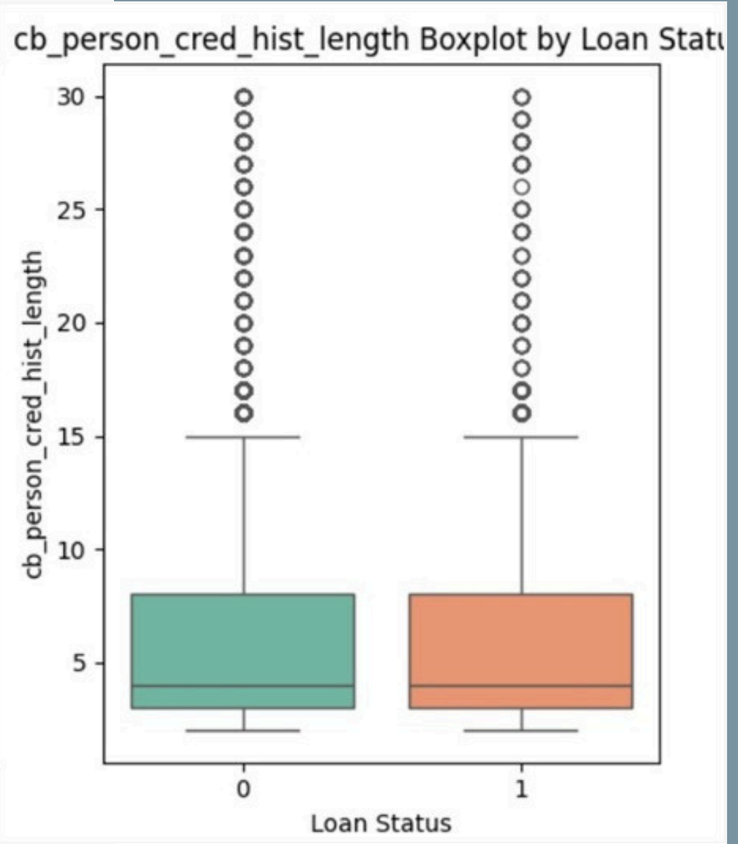
- LASSO selected features with the strongest influence on loan approval
- Bar plot created using absolute coefficient values
- Positive impact → Increased likelihood of approval
- Negative impact → Decreased likelihood of approval
- Top Positive Features:
- loan_size_cat_Small
- person_home_ownership_RENT
- loan_intent_HOMEIMPROVEMENT
- Top Negative Features:
- log_income
- cb_person_cred_hist_length
- person_home_ownership_OWN



Feature Analysis – Key Variables

- Person Income (Positive): Higher income = stronger repayment capacity
- Credit History Length (Mixed): Longer history shows reliability but may reflect legacy debt
- Loan Amount (Negative): Larger loans increase risk
- Loan Intent – Personal & Education: Lower approval rates; unsecured or delayed repayment
- Person Age (Negative): Slight decline in approvals for older applicants
- Business Insights:
- Promote financially stable profiles
- Use caution with high-risk loan intents and amounts
- Tailor policies for older or high-income applicants





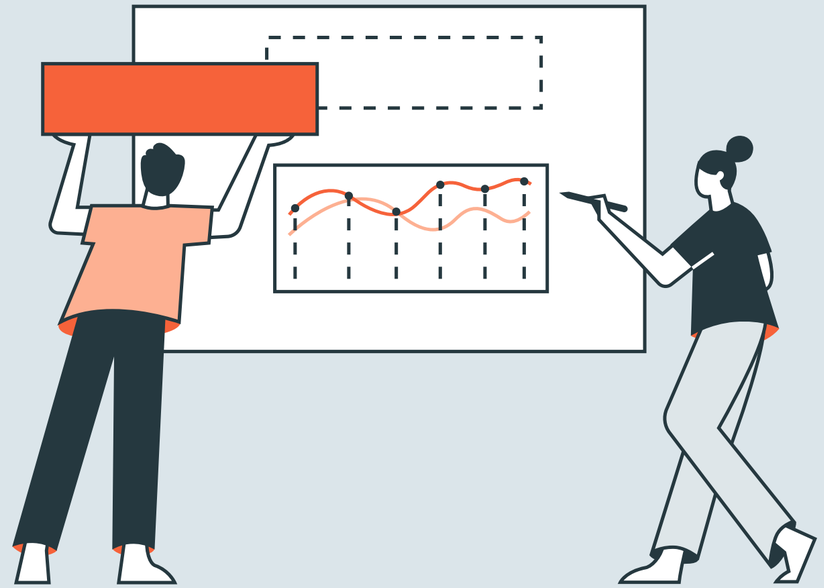
Key Takeaways from Evaluation

- XGBoost + SMOTE + LASSO was the best-performing and most interpretable model
- LASSO helped identify impactful features for transparent decision-making
- Smaller loans and moderate exposure → higher approval
- Some unexpected trends: high income & long credit history → lower approval
- Visualizations (bar plots, boxplots) enhanced stakeholder understanding

Section 6 - Deployment

Deployment Phase – Strategy & Business Impact

- This phase focuses on applying model insights to real-world lending decisions.
- Data-driven recommendations aim to reduce default risk and improve loan approval processes.
- Based on LASSO-selected features and detailed feature analysis, 13 key recommendations were developed.



ADELPHI

Top Actionable Recommendations

- Promote Small Loan Sizes (loan_amnt) – Lower risk, faster approval, ideal for safe lending growth.
- Target Renters (person_home_ownership_RENT) – Renters had higher approval odds; design custom offers.
- Encourage Moderate Exposure (loan_exposure_cat_Moderate) – Build products with optimal risk balance.
- Boost Home Improvement Loans (loan_intent_HOMEIMPROVEMENT) – Higher approval potential; promote actively.
- Use Income-to-Loan Ratio (loan_percent_income) – Integrate into prequalification tools to screen efficiently.
- Reward Low-Risk Borrowers (ratio_bin_Low, debt_burden_score) – Fast-track and incentivize stable applicants.

ADELPHI

Refined Strategies Based on Model Insights

- Screen High-Income Applicants Cautiously (log_income) – High income ≠ low risk; check over-leverage.
- Review Long Credit Histories (cb_person_cred_hist_length) – Identify outdated red flags.
- Tailor for Homeowners (person_home_ownership_OWN) – Design loans that offset liability risks.
- Adopt Age-Aware Policies (person_age) – Avoid bias; verify financial stability in older applicants.
- Tier Approvals by Employment (person_emp_length) – Use job history for loan customization.
- Limit High Loan Percent Requests (loan_percent_income) – Add checks for requests exceeding 30–40%.
- Use Explainable AI – Employ interpretable models like LASSO for trust and regulatory clarity.

ADELPHI

Conclusion

To conclude, our project focused on building a machine learning solution to predict loan approval outcomes using applicant data.

We followed the CRISP-DM framework throughout the entire process. In the business understanding phase, we defined the problem and objectives, highlighting the need for faster, more reliable loan decisions in banking.

In the data understanding stage, we performed exploratory analysis to understand key trends, missing values, and influential variables.

During data preparation, we cleaned the dataset, handled outliers, encoded categorical variables, created new features, and addressed missing data to prepare for modeling.

In the modeling phase, we trained five models using four scenarios combining different balancing and feature selection techniques. The best-performing model was XGBoost with SMOTE and LASSO, achieving an AUC of 0.9441 and a strong G-Mean.

We then evaluated model performance using metrics like sensitivity, specificity, precision, and AUC, and visualized the top features. These helped us understand which applicant traits were most predictive.

Finally, in the deployment and recommendation phase, we proposed actionable strategies for banks, such as prioritizing applicants with lower loan-to-income ratios and monitoring those with previous defaults or high interest rates.

This project not only achieved strong predictive performance but also demonstrated how data science can bring real business value — by enabling smarter, faster, and fairer loan approval processes.

ADELPHI

Thank you

ADELPHI