Loan Approval

Final Project - Data Science

Almira Z.

A bachelor's degree in management, and was in an exchange program at the University of Applied Sciences Utrecht, **contributing to consulting projects for EDAG Group, Google Merchandise Store, and Ørsted. Hands-on experience in market research, strategy development, and competitor analysis as a Junior Consultant Intern at EDAG Group.** With strong analytical and communication abilities and proficiency in Microsoft Office, Google Suite, Canva, and Google Analytics.

About Dataset

This dataset is used **to analyze borrower demographics, financial, and credit behavior to predict loan approval outcomes.** By identifying key risk drivers and borrower patterns, financial institutions can **optimize credit decision processes, minimize default risk, improve portfolio quality, and enhance overall lending efficiency** while maintaining regulatory compliance.

Data Understanding

Column Name	Data Type	What It Tells Us
person_age	float64	Applicant's age and life stage.
person_gender	object (categorical)	Applicant's gender (demographic information).
person_education	object (categorical)	Highest education level attained.
person_income	float64	Annual income indicating financial capacity.
person_emp_exp	int64	Years of work experience and job stability.
person_home_ownership	object (categorical)	Housing status reflecting financial commitments.
loan_amnt	float64	Amount of loan requested.
loan_intent	object (categorical)	Purpose of the loan.
loan_int_rate	float64	Interest rate applied to the loan.
loan_percent_income	float64	Portion of income used for loan repayment.
cb_person_cred_hist_length	float64	Length of credit history.
credit_score	int64	Creditworthiness and repayment behavior.
previous_loan_defaults_on_file	object (categorical)	History of previous loan defaults.
loan_status	int64 (Target)	Loan decision outcome (approved or denied).

Pre-Processing

1. Duplicated & Missing Value

0 Duplicated Data and 0 Missing Values, *nothing to be dropped*

2. Outliers

7.451 Outliers, and we drop it.
Data Before : 45.000
Data After. : 37.549

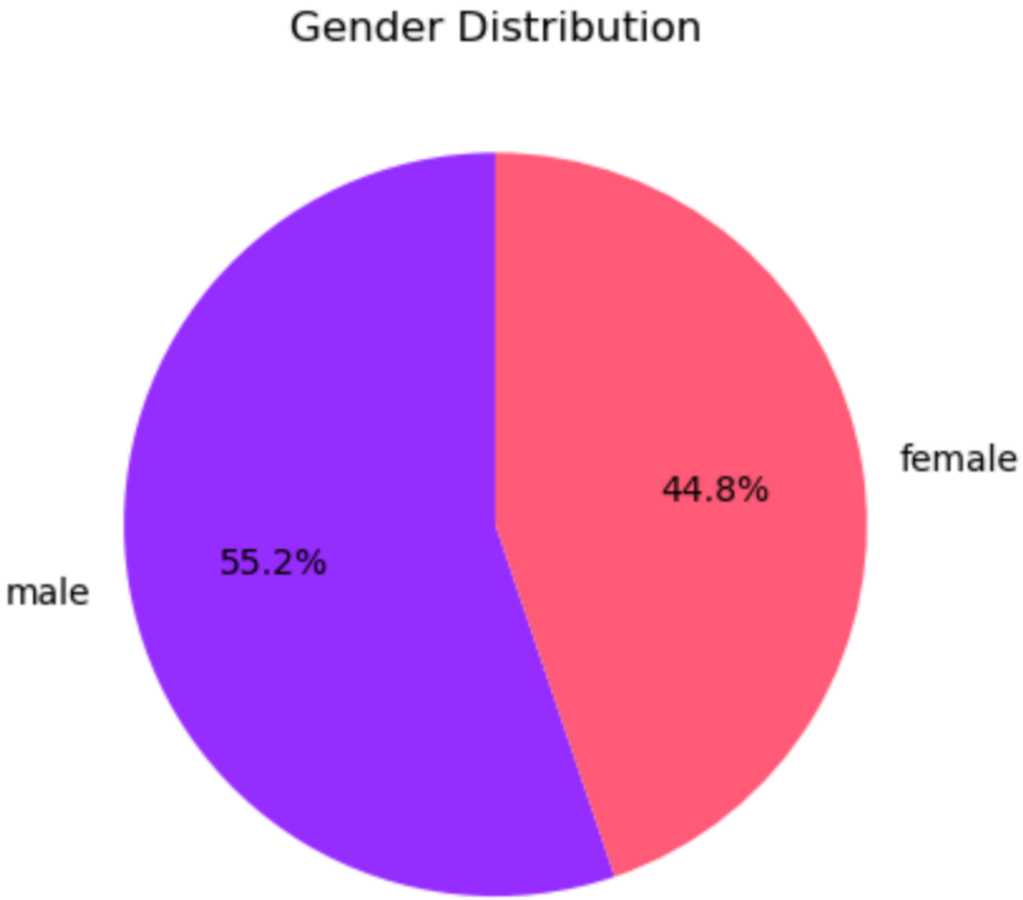
3. Normalization

*Using **Min-Max Scaler Method***

4. Featuring

*Using **Heatmap Correlation** to see the correlation between the numerical variables*

EDA



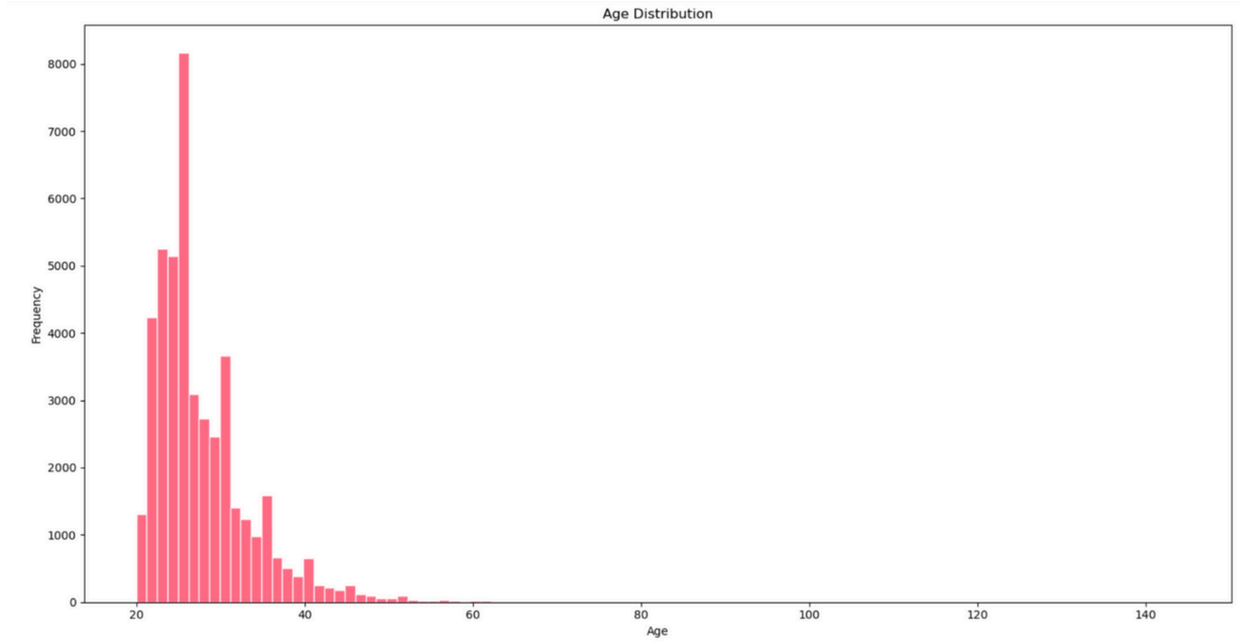
- **Male representation is higher at 55.2%,** compared to 44.8% female.
- The difference is 10.4 percentage points, which suggests a moderate imbalance, not an extreme one.
- The distribution is relatively close to parity, indicating fair diversity, **though still male-leaning.**

EDA

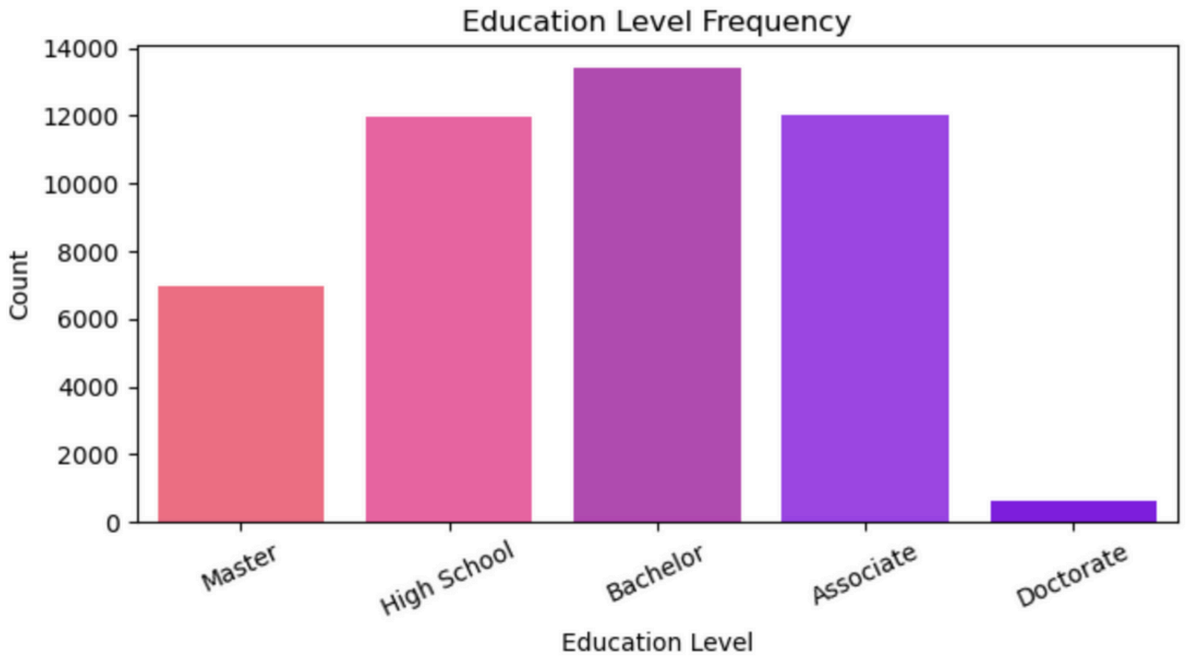
- **Peak Demographic**: The distribution is heavily right-skewed, with a significant peak in the mid-20s.

- **Primary Range**: Most applicants fall within the 20 to 40-year-old range, after which the frequency drops off sharply.

- **Presence of Outliers**: While the majority of data is concentrated under age 60, the x-axis extends toward 140, indicating the presence of extreme outliers or potentially erroneous data points that were likely addressed during the pre-processing stage.



EDA



- **Bachelor's degree holders are the most frequent applicants,** with a count exceeding 13,000.

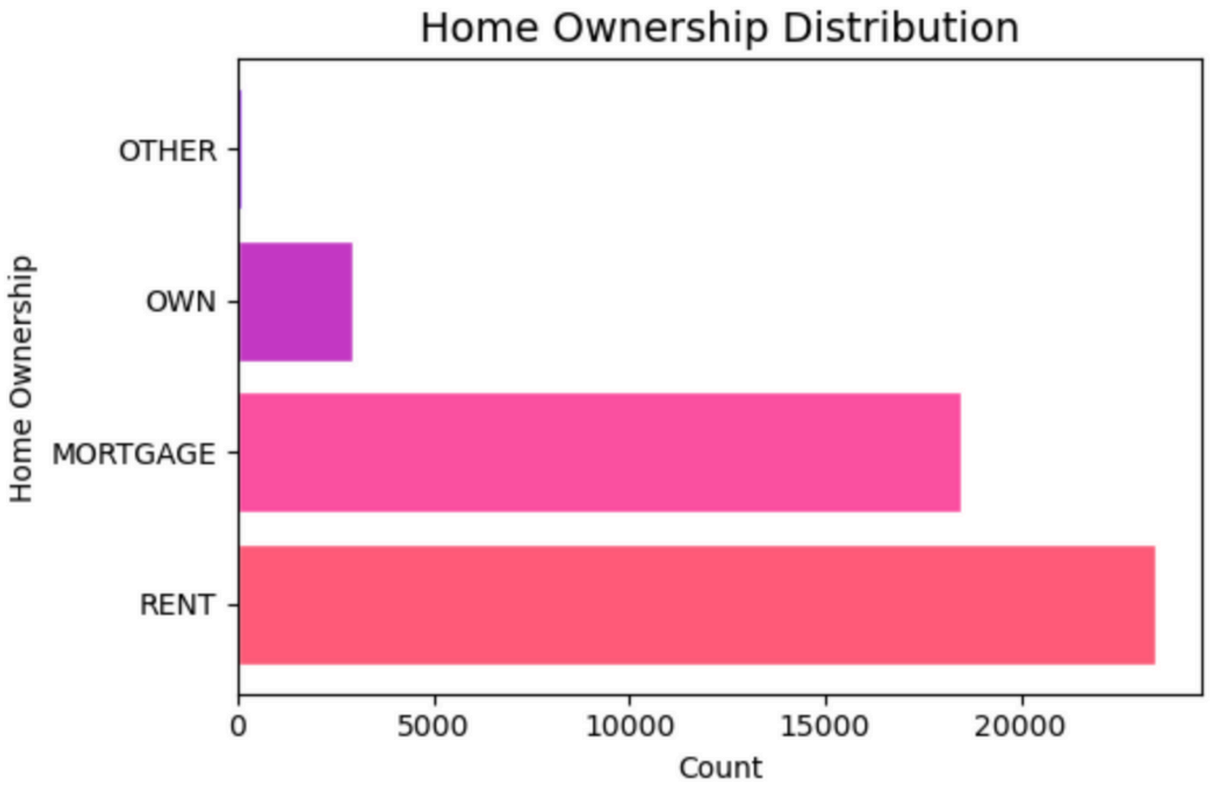
- **High School and Associate degree holders are nearly equal**, each representing approximately 12,000

- **Doctorate holders represent the smallest portion of the dataset,** with a count significantly lower than all other categories (less than 1,000).

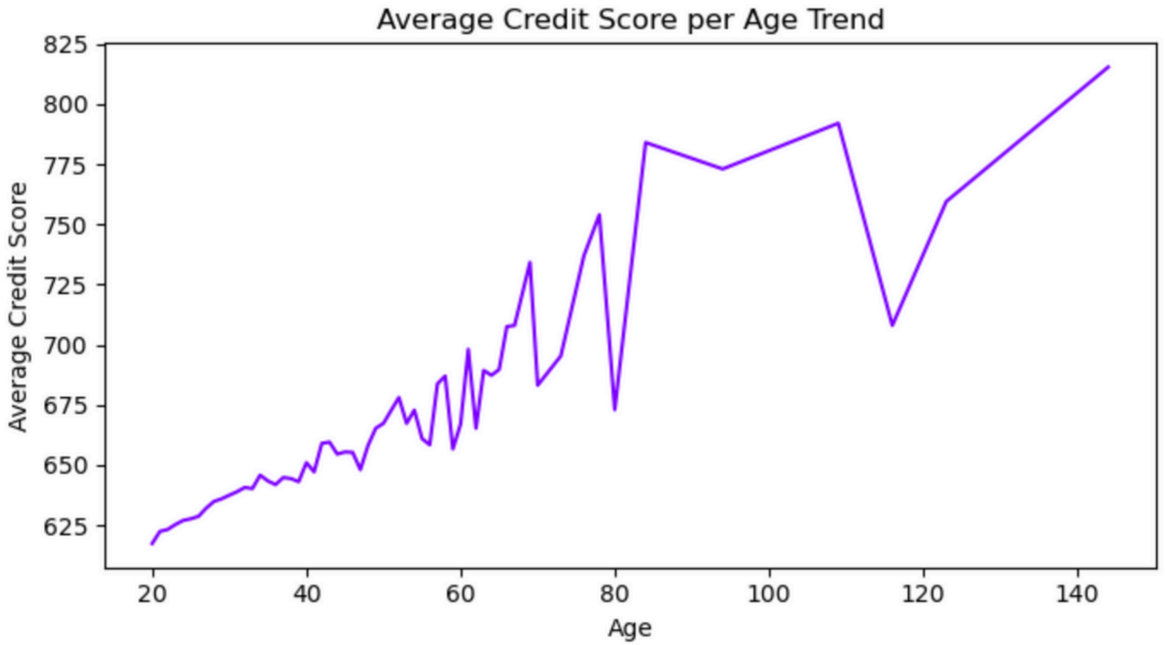
EDA

- **Most borrowers are renters.** This suggests the population is highly mobile or in a transitional life stage (e.g., young professionals, students).

- **Mortgage holders form the second-largest** group. This indicates financial stability and long-term planning, but not full ownership yet.



EDA



•**Positive Correlation**: Average credit scores generally increase as applicants get older.

•**Stable Growth**: There is consistent, steady improvement in scores between the ages of 20 and 50.

•**High Volatility**: Significant fluctuations occur after age 60, likely due to smaller sample sizes in older demographics.

•**Peak Creditworthiness**: The highest average scores (exceeding 800) are found in the oldest age brackets.

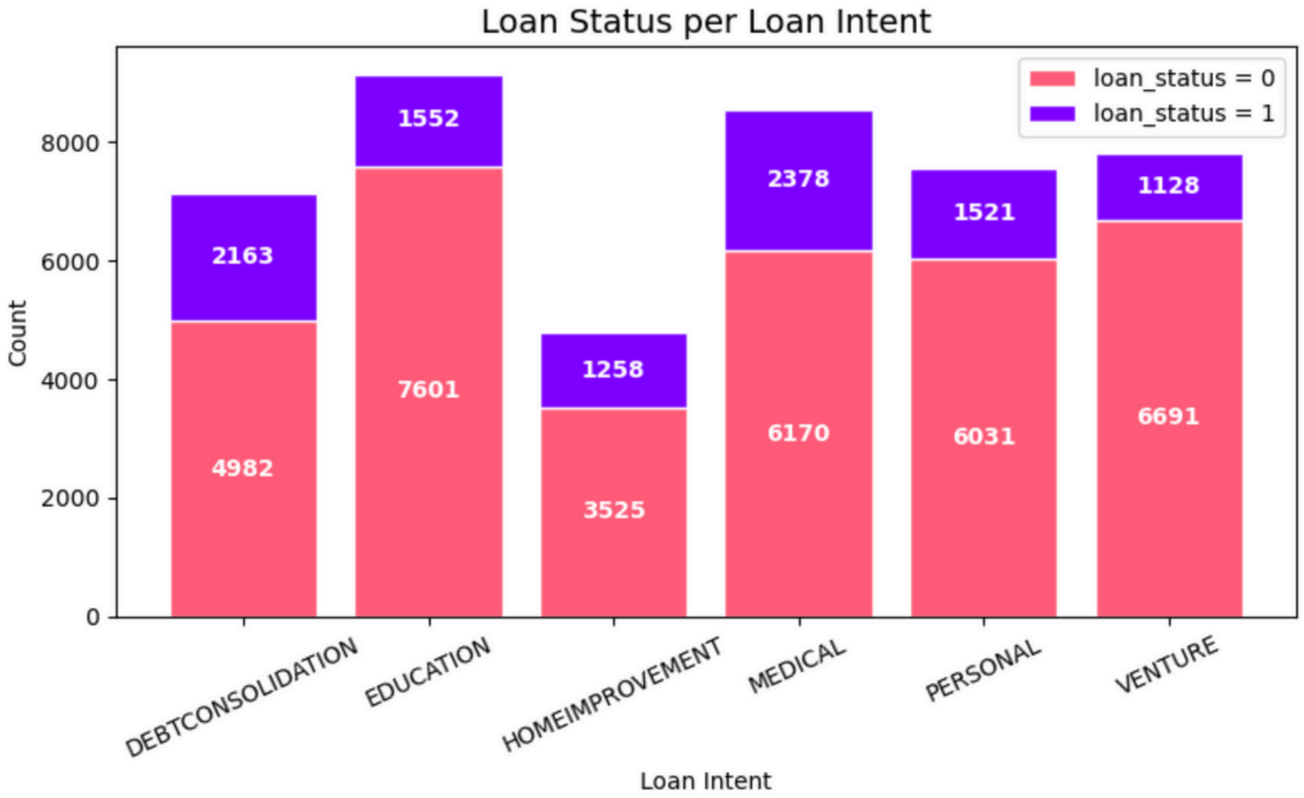
EDA

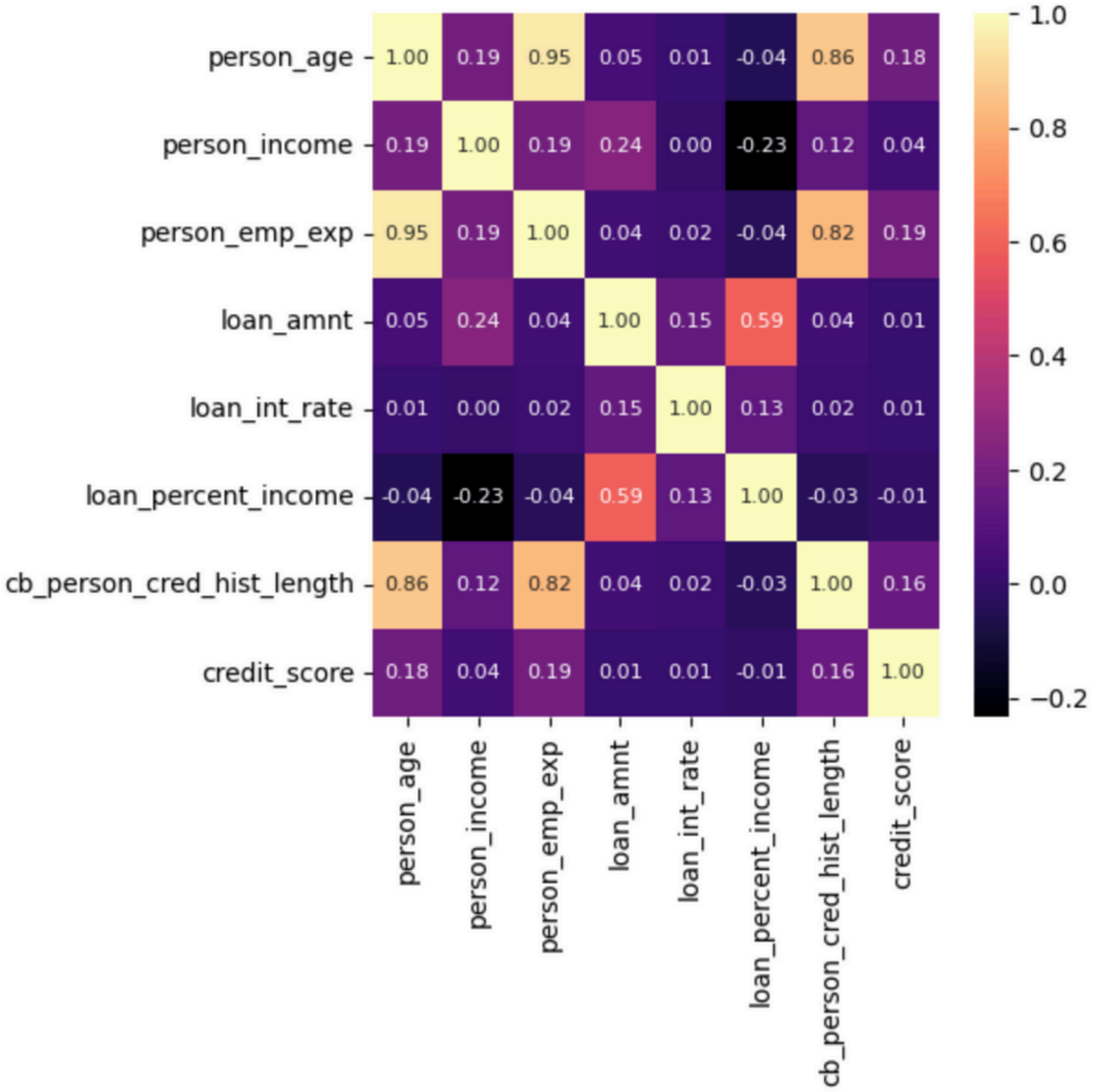
- **Highest Volume**: **Education** is the most frequent loan intent, with over 9,000 total applicants.

- **Highest Approval Rate: Medical** intent shows the highest number of approved loans (loan_status = 1) at 2,378.

- **Lowest Volume: Home Improvement** has the fewest applicants, totalling fewer than 5,000.

- **Rejection Dominance**: Across all categories, the number of denied loans (loan_status = 0) significantly outweighs approvals.





Correlation Heatmap

- **Strong Positive Age Correlation:** A very high correlation of 0.95 exists between person_age and person_emp_exp, as well as 0.86 between age and cb_person_cred_hist_length.

- **Loan-to-Income Relationship:** There is a moderate positive correlation of 0.59 between the loan_amnt and the loan_percent_income.

- **Weak Credit Score Ties**: The credit_score shows very low correlation with most other variables, with its highest relationship being only 0.19 with employment experience.

- **Inverse Relationship**: A weak negative correlation of -0.23 is observed between person_income and loan_percent_income, indicating that higher-income individuals typically use a smaller portion of their earnings for loan repayment.

Modeling

Train & Test Split

Data Train : 30.039
Data Test : 7.510

Tuning Hyperparameter

Only for the KNN Model
*with **K - Value : 3***

Class Imbalance

Label 0 : 29.562
Label 1 : 7.987

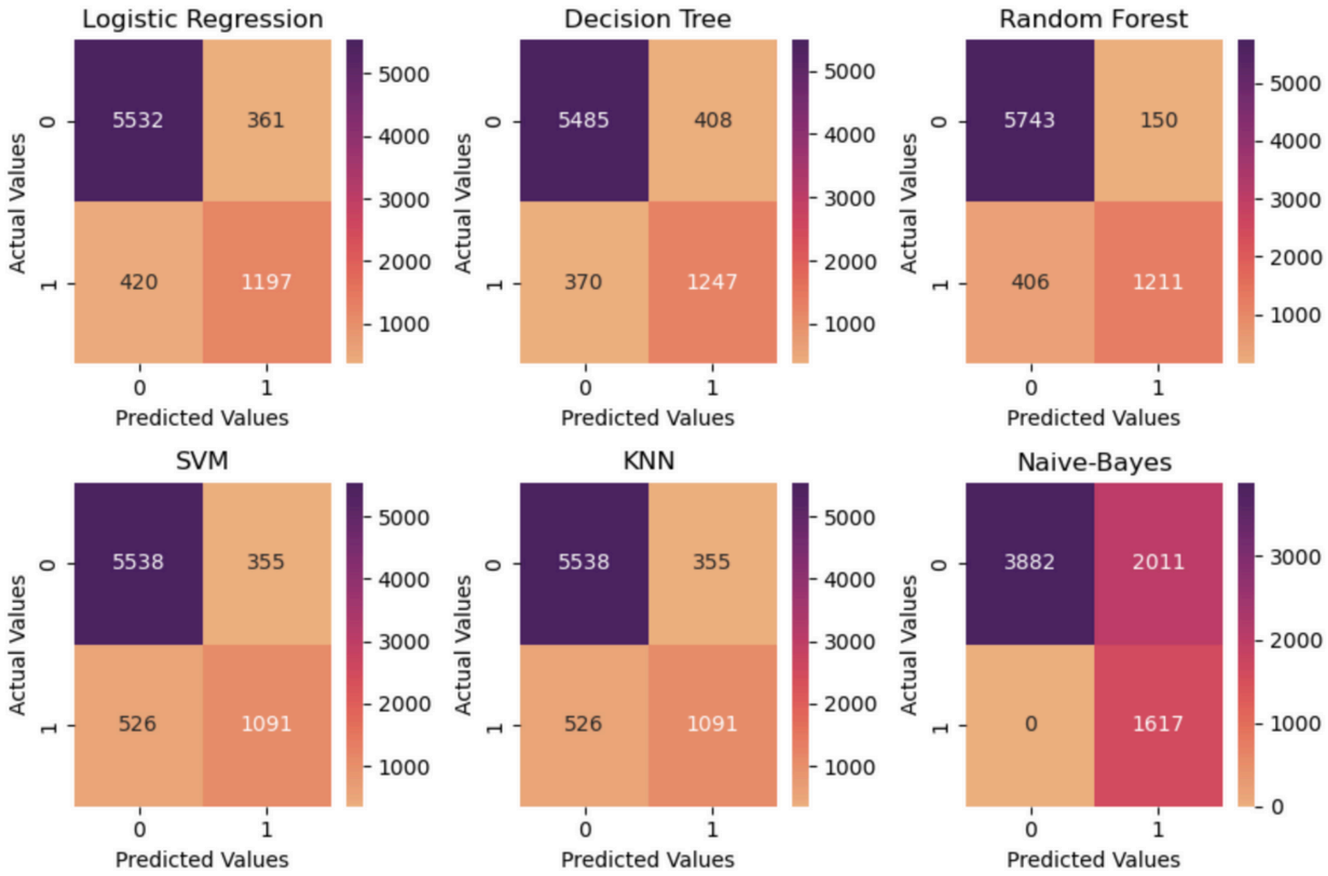
Needs Balancing by
Handle Class Imbalance:
23.669

Training a Modeling

KNN : 93,8%
Logistic : 89,5%
Decision Tree: 100%
Random Forest : 100%
Naive-Bayes: 72,8%
SVM: 91,4%

Modeling

	Model	Akurasi	Precision	Recall	F1 Score
0	Logistic Regression	0.896005	0.768293	0.740260	0.754016
1	Decision Tree	0.895473	0.750905	0.769944	0.760305
2	Random Forest	0.925699	0.886779	0.750773	0.813128
3	Support Vector Machine	0.907057	0.817993	0.730983	0.772044
4	K-Nearest Neighbor	0.882690	0.754495	0.674706	0.712373
5	Naive-Bayes	0.732224	0.445700	1.000000	0.616587



Evaluation

Random Forest

Best Model

Random Forest achieved the highest accuracy of 0.926, with precision 0.887, recall 0.751, and an F1-score of 0.813.
This shows the best overall balance between accuracy and reliability, making it the most suitable model for this task.

Logistic Regression

Stable and Interpretable Model

Logistic Regression achieved an accuracy of 0.896, with precision 0.768, recall 0.740, and an F1-score of 0.754.
The confusion matrix shows a relatively **high number of false negatives (420) and false positives (361),** indicating that some eligible applicants were not detected while some risky applicants were incorrectly approved.

Decision Tree

Moderate

The Decision Tree model achieved an accuracy of 0.895, with precision 0.751, recall 0.770, and an F1-score of 0.760.
The model performs well in identifying approved cases but produces more prediction errors compared to ensemble methods

Evaluation

K-Nearest Neighbors

Less Suitable Model

KNN achieved an accuracy of 0.883, with precision 0.754, recall 0.675, and an F1-score of 0.712.
The confusion matrix shows a high number of **false negatives (526) and false positives of 355, This indicates weaker performance,** particularly in identifying approved cases, making it less suitable for this problem.

Naive-Bayes

Not Suitable Model

Naive Bayes produced the lowest accuracy of 0.732, with precision 0.446, recall 1.000, and an F1-score of 0.617. The confusion matrix shows **zero false negatives but very high false positives (2011)**, indicating that the model predicts almost all cases as approved and is highly biased.

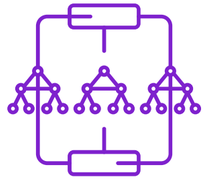
Support vector Machine

High-Precision Model

SVM achieved an accuracy of 0.907, with precision 0.818, recall 0.731, and an F1-score of 0.772.
The model is highly precise in predicting approved cases but slightly less effective in capturing all eligible applicants.

Evaluation

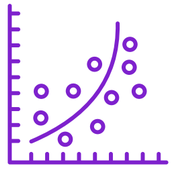
01



Random Forest

Primary model for production

02



Logistic Regression

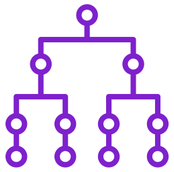
Stable baseline model

03

SVM

High-precision model

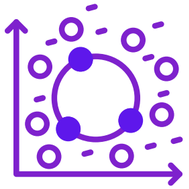
04



Decision Tree

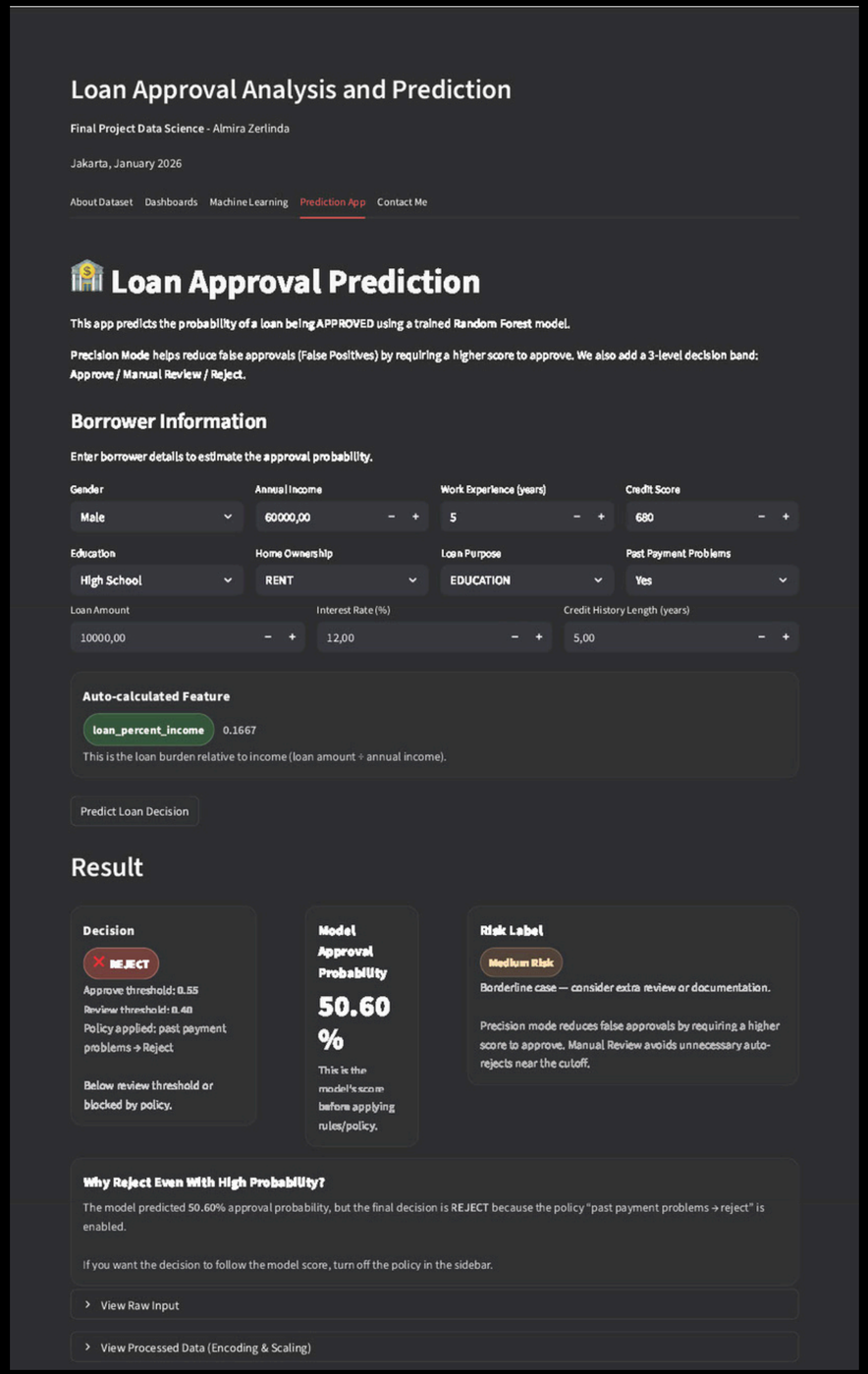
Moderately effective model

05



KNN dan Naive Bayes

Not suitable models



Prediction Test

This section presents the implementation of my loan approval prediction model using a **Random Forest classifier and Precision model.**

To demonstrate how borrower information is transformed into a **predicted approval probability, followed by a final decision based on predefined thresholds and business rules.**

What this demonstrates

- Application of machine learning to a real-world credit risk problem
- The impact of different borrower features on approval probability
- The trade-off between precision (risk control) and recall (approval coverage)
- How business rules can complement model predictions

Decision logic

- Approve → Approval probability exceeds the approval threshold
- Review → Approval probability falls within a review range
- Reject → Approval probability is below the review threshold or violates policy rules

Summary

This project **implements a Loan Approval Prediction System** using a **Random Forest model. The model prioritizes precision for risk control**, reducing false approvals and improving decision reliability. **By combining probability thresholds, risk labels, and business rules,** the system aligns machine learning predictions with real-world loan policies.

Recommendation

- **Use Random Forest as the primary model**

Random Forest is recommended for deployment due to its superior overall performance and balanced prediction capability.

- **Maintain Logistic Regression as a baseline model**

Logistic Regression should be retained for comparison purposes and for scenarios requiring model interpretability and transparency.

- **Apply SVM in precision-critical scenarios**

SVM is suitable when minimizing false approvals is a priority, such as in high-risk loan decisions.

- **Avoid using KNN and Naive Bayes for deployment**

Their lower performance and inconsistent predictions make them unsuitable for real-world loan approval systems.

Thank You

Link:

Canva **Streamlit**

Github