Prepared by group 3

Predicting Loan Approval

Helping Financial Institutions Make Faster, Fairer, Data-Driven Decisions

27th June 2025

Team Members: Kalliopi Georgiou, Doga Hascelik, Sofia Fox, Wanling Cheng, Anahi Bautista

Business Understanding (CRISP-DM)

Key Message: Loan approval is time-consuming, inconsistent, and prone to human bias. We aim to automate and improve it using machine learning.

- Problem: How can financial institutions predict loan approvals faster and more fairly?
- Business Goal: Improve efficiency, reduce errors, and support fairness in lending decisions.

The Data Source

Where Did Our Data Come From?

- Source: Kaggle - Loan Approval Dataset
- Size: 614 observations, 13 features + 1 target variable (Loan_Status)
- Key Features: Income, Loan Amount, Credit History, Employment Experience, etc.
- Target: Loan_Status (1 = approved, 0 = rejected)

Loan Approval Classification Dataset

Synthetic Data for binary classification on Loan Approval

1. Data Source

This dataset is a synthetic version inspired by the original Credit Risk dataset on Kaggle and enriched with additional variables based on Financial Risk for Loan Approval data. SMOTENC was used to simulate new data points to enlarge the instances. The dataset is structured for both categorical and continuous features.

We used real-world-like data to reflect what financial institutions evaluate in applications.

Our Process: Clean, Train, Predict

Data Preprocessing

- Removed age outliers (>100)
- Verified no missing values
- Clean structure with a mix of categorical and numerical variables
- Dataset reflects moderate-income applicants realistically

CRISP-DM Phases: Data Preparation, Modeling, Evaluation

Objective: Predict loan approval to support faster, fairer, and more consistent decisions
Purpose: Help financial institutions reduce manual workload using a data-driven model
Benefits:

- Flag high-risk applicants
- Streamline approvals for reliable candidates
- Improve efficiency and reduce processing time

Impact:

- Minimize errors by reducing subjective judgment
- Promote fairness, objectivity, and credibility in loan decisions

Model Selection

In order to solve this problem, we explored **different versions of Logistic Regression** to test how various features influence loan approval.

We started with a simple model using only loan amount, and progressively added more features to improve prediction accuracy.

Model C used just the loan amount
Model A added a few financial indicators
Model B included all 8 available features.

This allowed us to test how feature richness improves prediction accuracy.

Model Results

A. Original Logistic Regression (3 features)

```
Optimization terminated successfully.
         Current function value: 0.468544
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:                31495
Model:                          Logit   Df Residuals:                    31491
Method:                           MLE   Df Model:                            3
Date:                Thu, 26 Jun 2025   Pseudo R-squ.:                  0.1155
Time:                        21:54:41   Log-Likelihood:                -14757.
converged:                       True   LL-Null:                       -16684.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const            -0.3414      0.183     -1.868      0.062      -0.700       0.017
person_income  -2.88e-05   6.04e-07    -47.705      0.000      -3e-05   -2.76e-05
loan_amnt         0.0001   2.69e-06     43.063      0.000       0.000       0.000
credit_score     -0.0001      0.000     -0.495      0.621      -0.001       0.000
=================================================================================
```

Takeaways:

- Negative person_income coefficient implies higher income increases the odds of loan approval
- Positive loan_amt coefficient implies larger loans reduce the odds of approval

Model Results

B. More complex Logistic Regression with 8 features

Logit Regression Results

Dep. Variable:	loan_status	No. Observations:	31495
Model:	Logit	Df Residuals:	31486
Method:	MLE	Df Model:	8
Date:	Thu, 26 Jun 2025	Pseudo R-squ.:	0.2615
Time:	22:23:39	Log-Likelihood:	-12321.
converged:	True	LL-Null:	-16684.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.8629	0.311	-22.062	0.000	-7.473	-6.253
person_age	0.0221	0.010	2.170	0.030	0.002	0.042
person_income	7.35e-07	4.37e-07	1.681	0.093	-1.22e-07	1.59e-06
person_emp_exp	-0.0199	0.009	-2.237	0.025	-0.037	-0.002
loan_amnt	-0.0001	4.32e-06	-25.675	0.000	-0.000	-0.000
loan_int_rate	0.3349	0.006	54.707	0.000	0.323	0.347
loan_percent_income	15.6056	0.306	50.951	0.000	15.005	16.206
cb_person_cred_hist_length	-0.0048	0.009	-0.543	0.587	-0.022	0.012
credit_score	-0.0003	0.000	-0.993	0.320	-0.001	0.000

Takeaways:

- Adding features has improved model fit
- Some variables are very strong predictors
- "person_income" lost significance in this model compared to the first
- "credit_score" is consistently non-significant in both this model and the first

Model Results

C. Linear Regression Model using only loan amount as numerical feature

```
Optimization terminated successfully.
         Current function value: 0.524130
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:                31495
Model:                          Logit   Df Residuals:                    31493
Method:                           MLE   Df Model:                            1
Date:                Thu, 26 Jun 2025   Pseudo R-squ.:                 0.01061
Time:                        21:22:40   Log-Likelihood:                -16507.
converged:                       True   LL-Null:                       -16684.
Covariance Type:            nonrobust   LLR p-value:                 5.537e-79
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.6407      0.025    -65.125      0.000      -1.690      -1.591
loan_amnt   3.874e-05   2.03e-06     19.055      0.000    3.48e-05    4.27e-05
==============================================================================
```

Takeaways:

- “loan_amt” is a statistically significant feature, but very weak as a sole predictor
- This model is useful only as a baseline.
- For real-world predictions, we would want to use model B

Model Results

Model Comparison Conclusion

Metric	Model C (loan_amt)	Model A (person_income, loan_amt, credit_score)	Model B (8 features incl. income, age, rate, etc.)
# Features	1	3	8
Pseudo R^2	0.0106	0.1155	**0.2615**
Log-Likelihood	-16507	-14757	**-12321**
LLR p-value	5.54e-79	0.000	0.000
Expected Accuracy	Low – likely near baseline	Moderate – likely decent improvement	**High – best performance**
Interpretability	Very high (1 var)	Medium	Lower, but can explain via SHAP or similar

Takeaways:

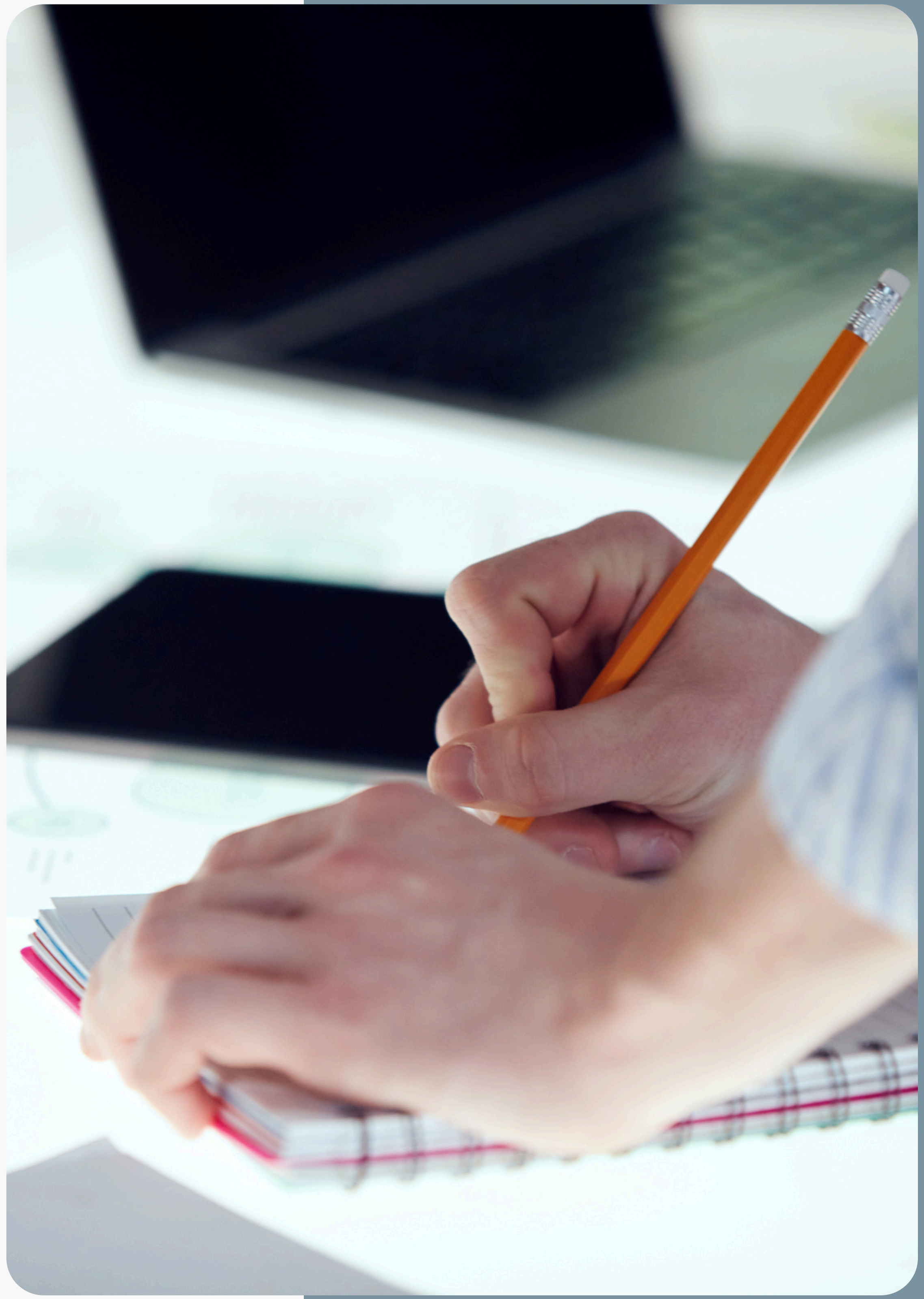
- Model C is useful only as a simple benchmark in comparison.
- Model A is a good trade-off between simplicity and predictive power
- Model B is the best in terms of fit and likely classification accuracy

Conclusion & Business Value

How Does This Help the Business?

- Speeds up the loan approval process
- Consistent, unbiased decision-making
- Identifies strong applicants faster
- Supports fairness in lending by relying on data, not opinion

You can use this tool to make confident lending decisions, reduce risk, and improve customer satisfaction.



How We Can Improve It

What's Next?

- Add more real-world data (ex: location, past banking history)
- Address class imbalance more robustly
- Improve model explainability (ex: use SHAP values)
- Integrate with a real-time dashboard for decision-makers

Our model is the foundation, with more data and feedback, it becomes even smarter

Answers to Relevant Course Questions

Applied course concepts: Followed full supervised learning pipeline: business framing → data prep → modeling → evaluation.

Ethics and real-world impact:

- Discussed false positives/negatives and fairness risks.
- Emphasized the need for human oversight in finance ML applications

Visualization as a tool:

- Histograms, correlation heatmaps helped shape decisions.
- Showed how storytelling and visuals improve data interpretation.

Chose better metrics than accuracy:

- Dataset was imbalanced (only 22.2% approvals), so accuracy alone was misleading.
- Used Pseudo R^2, Log-Likelihood, and LLR p-values — as taught in class.

Thank you