INTRODUCTION TO PRACTICAL MACHINE LEARNING

Final Report

Sofia Fox, Kalliopi Georgiou, Doga Hascelik, Anahi Bautista, Wanling Cheng

Business Problem: The goal of this project is to develop a machine learning model that predicts whether a loan application will be approved, helping financial institutions make faster, more consistent, and data-driven lending decisions.

Introduction and Objective

In this project, our purpose is to address a critical challenge in the financial services industry, which is the efficiency and reliability of the loan approval process. Financial institutions receive large volumes of loan applications, and manually reviewing each one can be both time-consuming and inconsistent. By using and leveraging supervised machine learning, we seek to build a model that predicts whether a loan application should be approved, enabling them faster and more data-driven decisions.

Loan Approval Classification dataset has been chosen due to its clear structure and relevance to real world banking practices. The dataset includes a range of applicant information such as education level, professional experience, home ownership, loan amount, etc. all of which are commonly evaluated in loan decisions. The target variable is the Loan_Status, a binary feature where:

- 1 indicates that the loan has been approved, and
- 0 indicates that the application has been rejected.

Our objective is to train a predictive model that can classify loan applications based on this target variable. By doing so we want to "Explore which features have the strongest influence on loan approval decisions", "Observe how different machine learning models perform on this classification task", "Evaluate fairness, accuracy, and the potential real world implications of such a model". To tackle this problem three supervised machine learning models will be used:

- Logistic Regression

We will assess the models using metrics such as pseudo R-squared, log-likelihood, and LLR p-values, which are more suitable for classification tasks like logistic regression than traditional R^2. These metrics will help us compare the model fit, significance, and overall predictive power across the different logistic regression models. By focusing on statistical strength and interpretability, we aim to select the model that not only performs well but also aligns with real-world decision-making criteria in financial institutions.

Beyond performance, we will also reflect on ethical concerns such as bias in financial data and the potential consequences of incorrect loan rejections.

Business Understanding (CRISP-DM)

The core business objective is to help financial institutions in making faster and more consistent, accurate loan decisions by predicting the likelihood of loan approval based on applicant characteristics. The model will serve as a decision support tool, reducing the manual workload and increasing fairness and efficiency in the process.

The predictive model will help institutions flag high- risk applicants and streamline approvals for reliable candidates, therefore balancing risk management with customer service. This project supports several business goals. By automating the loan process, it aims to improve operational efficiency and reduce the time required to evaluate each application. Moreover, the model can help minimize approval errors by relying on data-driven insights rather than subjective judgement. Ultimately, it promotes more objective and consistent decision making, which is critical in ensuring fair access to financial services and maintaining the institution's credibility.

Data Understanding

The selected dataset is publicly available on Kaggle and simulates the structure of a real loan approval system. It contains a total of 614 observations (loan applications) and 13 input features, along with the target variable Loan_Status. Here is a brief description of the key variables:

Variable	Description	Type
Gender	Applicant's gender (Male/Female)	Categorical
Married	Marital status (Yes/No)	Categorical
Dependents	Number of dependents (0, 1, 2, 3+)	Categorical
Education	Applicant's education level	Categorical
Self_Employed	Whether the applicant is self-employed	Categorical
ApplicantIncome	Monthly income of the applicant	Numerical
CoapplicantIncome	Income of co-applicant (if any)	Numerical
LoanAmount	Loan amount requested	Numerical
Loan_Amount_Term	Duration of loan repayment (in days)	Numerical
Credit_History	Whether applicant has a good credit record	Categorical (0 or 1)
Property_Area	Area where the property is located	Categorical
Loan_Status	Loan approved (1) or rejected (0)	Target

Dataset Description and Source (W/ URL)

The dataset used for this project is the "Loan Approval Classification Dataset", made publicly available by Ta-Wei Lo on Kaggle. This dataset was selected for its direct relevance to the project's objective: predicting whether a loan application will be approved or rejected based on

applicant data. It provides a structured and realistic representation of typical loan evaluation criteria used by financial institutions.
Categorical features in the dataset include variables such as Gender, Married, Dependents, Education, Self_Employed, Credit_History, and Property_Area. These represent key demographic and background characteristics of applicants. Numerical features include ApplicantIncome, CoapplicantIncome, LoanAmount, and Loan_Amount_Term, which provide quantitative insights into the applicant's financial status and the requested loan details.
While the dataset is synthetic, it has been designed to closely resemble real-world financial data, making it a suitable choice for experimentation and modeling within the context of supervised learning. Its moderate size and balanced feature diversity allow for effective model training, evaluation, and comparison. The dataset can be accessed through the following URL:

https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data?resource=download

Preprocessing Steps and Summary Statistics

Before building and training machine learning models, a review of the dataset's quality and structure was conducted. However, one of the most notable advantages of the dataset is it was clean and well structured. No missing values were found and all variables were already in a consistent format for analysis.

During preprocessing, we also identified a number of applicant records with an age above 100, which we considered highly improbable in the context of loan applications. While it's theoretically possible for someone of that age to apply for a loan, such cases are extremely rare and more likely to be the result of data entry errors or synthetic noise. To maintain the reliability and interpretability of our model, we decided to remove these records from the dataset. This decision helped us reduce distortion in the age distribution and improved the overall quality of our data, particularly in ensuring that the model was learning from plausible, real-world scenarios.

The dataset comprises 45,000 loan application records, each representing a potential custermer's profile, financial status, and loan-related attributes. A combination of numerical and categorical variables provides a view of applicant behavior and risk potential. Descriptive statistics were generated to understand the data's distribution, central tendencies, and variability.

The age of applicants (person_age) now ranges from 20 to 94 years, with a mean of approximately 27.7 years and the applicant income (person_income) spans a wide range, from $8,000 to over $2.4 million, indicating significant income diversity. Despite the high maximum, the median income is around $67,000, showing that the majority of applicants fall within a moderate-income bracket. The average income is slightly higher, at approximately $79,900, suggesting a right-skewed distribution due to a few high-income individuals.

Employment experience (person_emp_exp) ranges from 0 to 76 years, with an average of 5.4 years. This variable reflects a mix of early-career and more experienced applicants. The loan amount requested (loan_amnt) ranges from $500 to $35,000, with a mean of approximately $9,583, aligning with typical personal loan amounts offered by financial institutions.

The loan interest rate (loan_int_rate) averages around 11%, and the loan-to-income ratio (loan_percent_income) is about 14% on average, indicating that applicants generally request loans proportionate to their income. The credit history length (cb_person_cred_hist_length) varies from 2 to 30 years, with an average of 5.9 years, and the credit scores range from 390 to 784, centering around a mean of 633.

The target variable, loan_status, indicates whether a loan was approved (1) or rejected (0). The data shows that approximately 22.2% of applicants received loan approval, while the remaining 77.8% were declined. This imbalance will be important to address in the modeling phase through appropriate metrics and potentially through resampling techniques.

Description of All Three Models and Their Results

Logistic Regression models:

The chosen models for our loan approval classification employs logistic regression, classifying loan applications as either approved (1) or not approved (0) (target variable). In this dataset, the variables of this classification include age, gender, education, income, years of professional experience, and other values associated with the desired loan such as loan intent, income ratio, interest rate. The following models vary in features and complexity.

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
==============================================================================
```

No. of Observations: 31,495
Pseudo R-squared: 0.1155 – This indicates modest explanatory power (11.55% of variability in the outcome is explained by the predictors).
Log-Likelihood: -14757
LLR p-value: 0.0000 – Statistically significant model overall (rejects null hypothesis)

Significant Predictors: person_income and loan_amt are highly statistically significant.

Non-significant Predictors: credit_score and const (intercept) are not significant.

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.

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

No. of Observations: 31,495
Df Model: 8 (i.e., 8 predictors)
Pseudo R-squared: 0.2615 – higher than the simpler model (~0.11), indicating better explanatory power.
Log-Likelihood: -12321 – better than in previous models (less negative = better fit).
LLR p-value: 0.000 – the model is statistically significant overall.

Takeaways
- Adding features has improved model fit (higher pseudo R^2, better log-likelihood).
- Some variables (e.g., loan_percent_income, loan_amt, loan_int_rate) are very strong predictors.
- person_income **lost significance** in this model compared to the first, likely due to a higher correlation with other financial metrics like loan_percent_income or loan_amt.
- credit_score is **consistently non-significant** in both this model and the first, as well — suggesting it might be redundant or not predictive for this dataset.

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

Since this is a logistic regression, a positive coefficient increases the log-odds of the positive class (likely meaning loan denied, depending on how loan_status is encoded).

Even though loan_amt is statistically significant, its practical predictive **power is very weak** (as shown by low pseudo R^2).

- 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 a model with more predictors like in the 8-variable model (model B)

Metrics Used and Why They Were Selected

The target loan_status is binary — pseudo R^2 is standard here. With no continuous outcome R^2 in the traditional sense doesn't apply as in other models using linear regression (this is logistic).

Since logistic regression uses likelihood estimation (MLE), metrics like pseudo R^2 are the best evaluation tools we've learned. The closer the pseudo R^2 is to the value .5, the stronger the model, showing that actually Model B had a pseudo R^2 of 0.2615, suggesting moderate-to-strong classification ability for this dataset, making it potentially the strongest at loan prediction.

Model Comparison Table

Model C out of all of the 3 models is the weakest — useful only as a simple benchmark in comparison.

Model A is a good trade-off between simplicity and predictive power but misses some key features like loan interest rate or credit history length.

Model B is the best in terms of fit and likely classification accuracy, thanks to its higher pseudo R^2 and much better log-likelihood.

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

Insights, Conclusions, and Ethical Reflections

Given that this dataset deals with sensitive financial information, the privacy and protection of the information within can be seen as crucial for the safety of the users and for the reputation of the bank or financial institution that employs it.

In addition, the accuracy of the classification models we have generated are equally important. In the case that the models classify incorrectly or equally particular identifiers could cause issues for the user or institution. In the case of approving a false positive, for example, the financial institution granting the loan may be at risk. With this in mind, it is incredibly important to continuously check the model for any changes over time, or if a new numerical feature is added into the dataset that could potentially tweak the classification model.

Answers to Relevant Course Questions

This project closely mirrors the supervised learning process discussed throughout the course, from identifying a business problem and selecting relevant data, to building and evaluating models. In applying these techniques, we not only deepened our technical skills but also reflected on model implications, ethics, and interpretability, core themes emphasized in class.

One of the most important takeaways from the course was that accuracy is not always a reliable evaluation metric, especially when working with imbalanced datasets like ours. Since only 22.2% of our loan applications were approved, a model that predicted “rejected” for every case would appear accurate on paper but be completely useless in practice. Therefore, we used metrics such as pseudo R-squared, log-likelihood, and LLR p-values to evaluate our logistic regression models. These metrics were introduced in class as more appropriate for classification tasks, particularly when using likelihood-based models. Pseudo R^2, for example, helped us understand how much of the variability in the outcome could be explained by the model, a concept explored in our regression-focused labs.

In our logistic regression models, we also paid close attention to coefficient interpretation, another topic covered in depth during the course. For instance, we found that higher applicant income tended to increase the chances of loan approval, reflected in a negative coefficient for person_income. We also observed that loan_percent_income was a much more robust predictor than raw income alone. This insight echoes class discussions around feature engineering and the idea that transformed or ratio-based features can often capture patterns more effectively than raw values.

Beyond technical accuracy, the course consistently highlighted the importance of ethical modeling and real-world impact. Predictive systems, particularly in finance, carry serious implications. A false positive (approving a risky loan) could lead to financial losses for the institution, while a false negative (rejecting a qualified applicant) could harm individuals unfairly. These risks underscore the need for not only model validation but also human oversight and bias mitigation. We kept these principles in mind while evaluating model fairness and thinking about how our work might be applied in real-world settings.

Finally, we recognized the power of visualization in both exploratory data analysis and result communication. While our report includes detailed statistics, visual tools such as histograms and correlation matrices played a crucial role in shaping our understanding of the data. These practices were reinforced in our hands-on labs and served as a reminder that effective data science combines both numerical rigor and clear storytelling.

Team Contribution Breakdown

Each member of the team contributed in each stage of the project. The following is mainly how each member participated:
Kalliopi: Coding, model registration, dataset research
Doga: Report write up, term research, dataset research
Sofia: Coding, report write up, dataset research
Amanda: Report write up/PPT
Anahi: Dataset research, ethical reflections, presentation creation