הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

Final Project Report

Introduction to Machine Learning – Second Semester 2025
Tel Aviv University – Digital Sciences for High-Tech

Project name:

Classification of Mortgage Loan Applications – Florida HMDA 2013 Dataset

Group Number- 31:

Mellissa Ghandour: ID-326541521

Jude Younis: ID- 315868968

Project Summary – Loan Default Classification (Group 31)

This project focused on predicting whether a mortgage loan will be approved (1) or rejected (0) using structured applicant data. We approached it as a binary classification task, applying data preprocessing, model experimentation, and performance optimization techniques.

Project Goals

- Classify mortgage loan applications using known and anonymized features
- Experiment with different models, from basic classifiers to advanced ensemble methods
- Tune hyperparameters to maximize performance (measured by AUC). And win! But unfortunately, we lost by 0.0003 to first place.
- Deliver a clean, reproducible pipeline and a professional project report

Tools & Technologies Used

- **Language**: Python 3.11
- **Environment**: Jupyter Notebook (via Anaconda)
- **Libraries**: pandas, scikit-learn, LightGBM, XGBoost, matplotlib, seaborn
- **Automation**: We built a self-contained environment capable of testing multiple models across datasets and parameter settings. It automatically logs results and saves outputs for later analysis.

Files Included in Submission (31.zip)

File Name	Description
notebook_31.ipynb	Full Jupyter Notebook with code, experiments, plots, and model evaluations
report_31.pdf	PDF report summarizing preprocessing, modeling steps, and conclusions
results_31.csv	Prediction probabilities on the test set in the required format
example_results.csv	Provided by the course team as a formatting reference (not created by us)

All filenames and formats match the course submission requirements. The final submission is zipped as 31.zip.

Model Development Overview

- **Preprocessing**: Imputation, normalization, one-hot encoding, and ID feature analysis
- **Baseline Models**: Logistic Regression, KNN, Decision Trees
- **Advanced Models**: Random Forest, AdaBoost, XGBoost, LightGBM
- **Feature Engineering**: Feature selection and experiments with new variables
- **Automation**: A robust pipeline tested dozens of model configurations and stored all results
- **Best AUC Achieved**: 0.9807+ with LightGBM after tuning and feature selection

Reproducibility & Environment Setup

To reproduce our results or continue developing, follow these steps:

1. Clone the Repository

Open your terminal (or Command Prompt) and run the following:

Windows / macOS / Linux:

```
GIT CLONE HTTPS://GITHUB.COM/FEATHERLESS-BIPED/INTROTOML_FINALPROJECT.GIT
CD INTROTOML_FINALPROJECT
```

2. Create a Virtual Environment

We recommend using venv:

For Windows:

```
PYTHON -M VENV .VENV
.VENV\SCRIPTS\ACTIVATE
```

FOR MACOS / LINUX:

```
PYTHON3 -M VENV .VENV
SOURCE .VENV/BIN/ACTIVATE
```

3. Install Dependencies

```
PIP INSTALL -R REQUIREMENTS.TXT
```

4. Launch Jupyter Notebook

```
JUPYTER NOTEBOOK
```

How to Run the Final Model Only

If you're interested in running only the final trained model on the processed data:

- Scroll to the **bottom code cell** of notebook_31.ipynb
- This section:
    - Loads the cleaned dataset

- Applies all preprocessing steps
- Loads the trained model with tuned parameters
- Outputs predictions and evaluation metrics

There is **no need to re-run the entire notebook**. The final code block is self-contained and ready to use.

For additional documentation, development history, and version control, visit our GitHub repository:
https://github.com/Featherless-Biped/IntroToML_FinalProject