Vincent1237 /
Loan-Default-Prediction-ML

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

0 stars 0 forks 0 watching Branches Activity
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ···

Vincent1237 presentation document updated 62c48de · now

.gitignore	Initial commit	3 days ago
Data.csv	define business problem for Loan de...	3 days ago
Final_Presentation.pptx	presentation document updated	now
Notebook - Jupyter Noteboo...	Final Notebook	yesterday
Notebook.ipynb	Final Notebook	yesterday
README.md	Final updated Read Me File	4 minutes ago

README

Loan Default Prediction Project

Author

Vincent Barchok Ngochoch
Full-time Data Science Student | Moringa School

Project Overview

This project presents an end-to-end machine learning pipeline to **predict loan default risk** using historical customer data. It is designed with business impact in mind and is targeted at financial institutions seeking data-driven solutions to minimize Non-Performing Loans (NPLs).

Business and Data Understanding

Stakeholder Audience

The primary audience for this project includes:

- Credit Risk Managers
- Lending Officers
- Banking Executives
- Data Strategy Teams in Financial Institutions

These stakeholders aim to enhance their credit assessment process and reduce losses arising from loan defaults.

Dataset Choice

The dataset used includes **10,000 customer records** from **KAGGLE DATASET** with financial and demographic variables relevant to credit risk evaluation. Key variables include:

- `Employed` : Employment status (binary)
- `Bank Balance` , `Loan Amount` , `Annual Salary` , `Savings Rate`
- `Defaulted?` : Target variable indicating if a customer defaulted

Problem Statements

1. Can we predict whether a customer will default on a loan using historical financial and demographic data?
2. What features most influence the likelihood of loan default?
3. Which model performs better between **Logistic Regression** and **Decision Tree**?

Modeling

The following steps were used to develop and train the models:

- Data cleaning and handling of missing values
- Exploratory Data Analysis (EDA)
- Feature Engineering (e.g., encoding employment, creating savings rate)
- Train-test split using stratification
- Model training with:
  - Logistic Regression (baseline)
  - Decision Tree Classifier
- Addressed class imbalance using **SMOTE**
- Hyperparameter tuning for Decision Tree

Evaluation

Model performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC AUC

Key results:

- Logistic Regression showed high accuracy but failed to detect defaulters (low recall).
- Decision Tree, after applying SMOTE, achieved significantly better recall and F1-score.

Conclusion

The Decision Tree model provided the best balance of performance and interpretability, especially after handling class imbalance with SMOTE. This model is well-suited for deployment in credit risk workflows to:

- Flag high-risk borrowers early
- Support risk-based pricing
- Reduce non-performing loans

This project demonstrates how machine learning can enhance lending strategies through data-driven insights.

File Structure

- `loan_default_prediction.ipynb` – Main Jupyter Notebook
- `resampled_train_data.csv` – (Optional) Exported dataset with SMOTE applied
- `README.md` – Project documentation

Releases

No releases published
Create a new release

Packages

No packages published

Publish your first package

Languages

- ● **Jupyter Notebook** 100.0%