Model Efficacy in Credit Risk Predictions

Foodies: Bosia N’dri, Zanderz McCluer, Sailesh Pulukuri, Reagan Todd
AI/Data science track

Introduction

Credit Risk Estimation:

- Used to assess the likelihood that a borrower will repay a loan.[1,2]
- Based on the 3 C's: Character, Capacity, and Collateral.[3]
- These estimations cannot discriminate based upon race, gender, or age.[1]

Current Methods:

- Traditional estimation includes Logistic Regressions or Linear Discriminant Analysis (LDA), which focus on a linear scale.[4]
- Generally easier to interpret, but are not able to accurately assess each case and are not suitable for large datasets.[4,5]

Problem Statement

How can machine learning (ML) models improve the credit risk prediction process?

- Models such as K-NN, Decision Trees, Boosting, Neural Networks, and Random Forest can enhance accuracy of predictions, and are scalable to help with larger datasets.[4,5]
- The adaptation of these models to credit risk predictions do require transparency, accountability, and human oversight.[5]

Methodology

Data Preprocessing and Visualizations

- Data obtained from OpenML and converted to CSV for manipulation.
- 4,011 missing values removed.
- Outliers were detected using descriptive statistics as well as histograms and boxplots.
- Engineered ratios such as loan-to-income, loan-to-employment length, and interest-rate-to-loan amount for deeper interpretability.
- One-hot encoding used for categorical variables like home ownership, loan intent, and default history.

Methodology

Model testing and evaluation

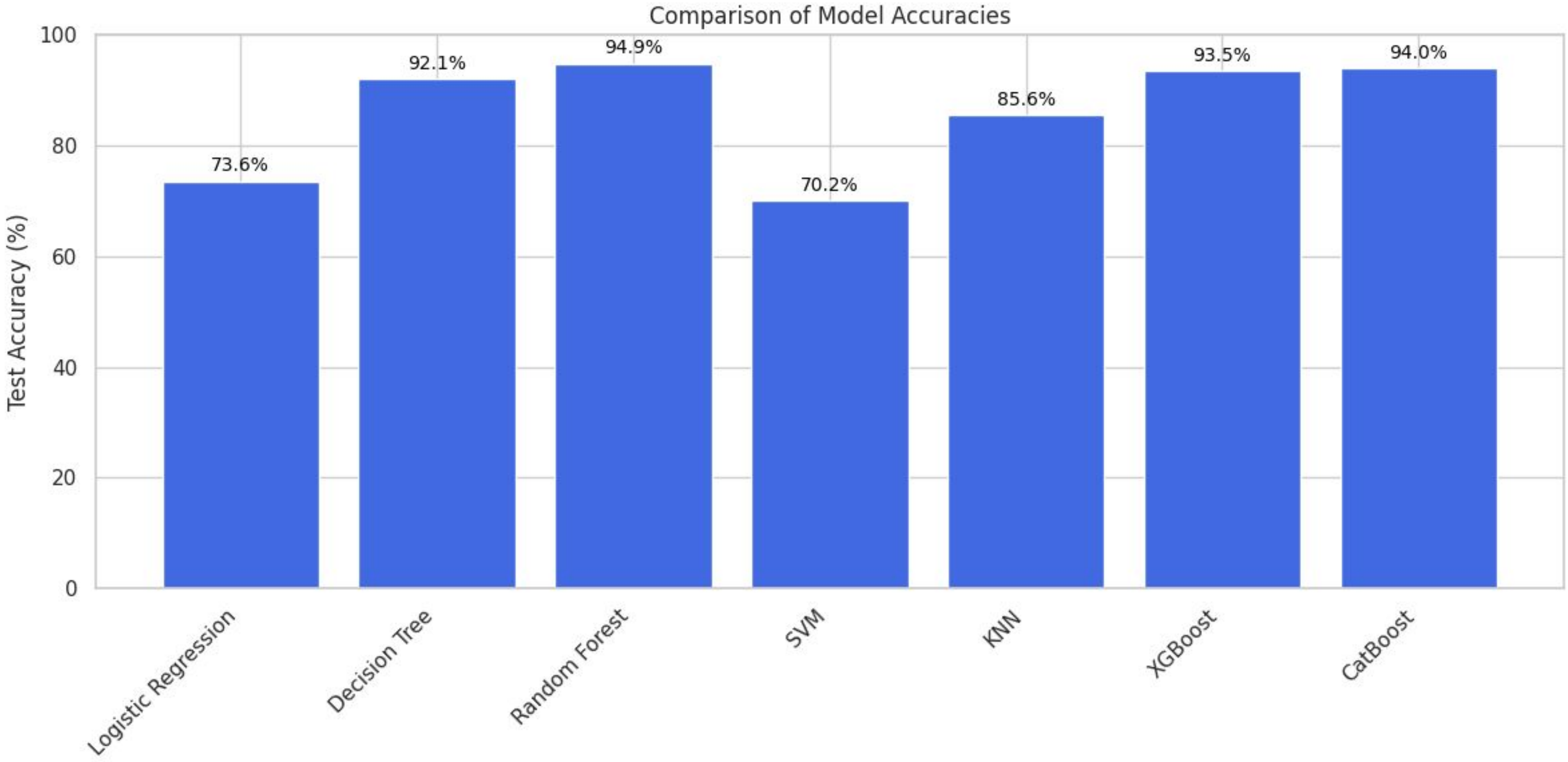
- Data split into 80 % training and 20 % testing.
- Evaluated 7 ML models: *Logistic Regression, Decision Tree, Random Forest, SVM, KNN, XGBoost,* and *CatBoost*.
- Metrics used: Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.
- Random Forest delivered the best results (Train = 0.869, Test = 0.868, F1 = 0.868).
  Final model exported as loan_rf_pipeline.pkl and deployed via a Streamlit web app.
- Integrated chatbot assistant for real-time feedback and explainability.

Results

Characteristics of the Study Sample			
	Mean (SD)	**Min**	**Max**
Age	28 (6)	20	84
Income	66,435 (51,522)	4,000	2,039,784
Employment Length	5 (4)	0	41
Loan Amount	9,651 (6,318)	500	35,000
Interest Rate	11% (3%)	5%	23%
Loan Percent Income	0.2 (0.1)	0.0	0.8
Credit History Length	6 (4)	2	30

Calculated Ratios for Study Sample			
	Mean (SD)	**Min**	**Max**
Loan to Income	0.2 (0.1)	0.001	0.83
Loan to Employment Length	0.001 (0.001)	0	0.02
Interest Rate to Loan Amount	0.002 (0.002)	0	0.02

Results


Comparison of Model Accuracies
Test Accuracy (%)
100
80
60
40
20
0
73.6%
92.1%
94.9%
70.2%
85.6%
93.5%
94.0%
Logistic Regression
Decision Tree
Random Forest
SVM
KNN
XGBoost
CatBoost

References

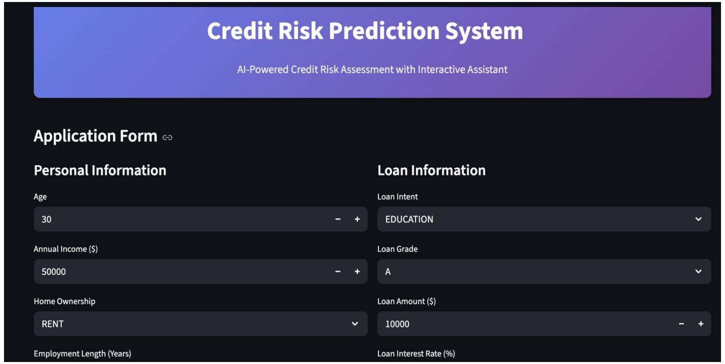
1. FRB: Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit. August 2007. Accessed October 16, 2025. https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/demographics.htm
2. Thomas LC. *Consumer Credit Models: Pricing, Profit and Portfolios*. OUP Oxford; 2009.
3. The Three C's of Credit (Lesson 9A). Accessed October 16, 2025. https://www.federalreserveeducation.org/en/teaching-resources/personal-finance/managing-credit/the-three-cs-of-credit-lesson-9a
4. Shi S, Tse R, Luo W, D'Addona S, Pau G. Machine learning-driven credit risk: a systemic review. *Neural Comput Appl*. 2022;34(17):14327-14339. doi:10.1007/s00521-022-07472-2
5. Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable Machine Learning in Credit Risk Management. *Comput Econ*. 2021;57(1):203-216. doi:10.1007/s10614-020-10042-0

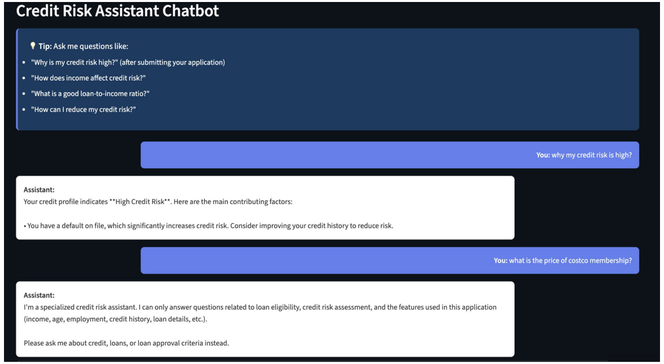
Analysis of Results

- Random Forest achieved the best balance between accuracy (~87% post-SMOTE) and reliability in predicting credit risk.
- SVM performed weakest (~70%), while boosting models (XGBoost, CatBoost) also achieved high accuracy (~93–94%) without resampling.
- SMOTE balancing slightly reduced accuracy but improved fairness and minimized bias toward majority classes.
- The final model was deployed through a Streamlit web application integrated with a chatbot assistant for enhanced accessibility and interpretation.
- Future work will focus on testing larger datasets, refining feature selection, and exploring LLM-based explainability for more transparent and user-friendly insights.

Conclusion and Future Direction

- Our trained Random Forest model was deployed as an interactive Streamlit web app that predicts credit risk and assists users in understanding their results.
- The integrated chatbot enhances accessibility by providing instant insights into loan eligibility and key risk factors.
- Translating accurate models into applications for borrower or lender use:





Live App: https://creditrisk-zsxirqtspn3csrdbntf5j6.streamlit.app/