Credit Risk Prediction Using Random Forest

Objective:
To predict whether a bank customer will default on a loan using historical credit data.
The goal is to help banks make data-driven decisions when assessing loan applications.

Marek Wisniewski
April 2025

Dataset Overview

Source: UCI Machine Learning Repository – German Credit Data

Records: 1000

Features: 20 input attributes + 1 target variable

Target Variable:

- `1` = Good credit (no default)
- `0` = Bad credit (likely default)

Data Preparation

- Renamed ambiguous columns for better readability
- Separated categorical and numerical features
- Applied **one-hot encoding** to categorical variables
- Applied **MinMax scaling** to numerical variables
- Final feature matrix after transformation: **1000 rows × 61 columns**

Exploratory Data Analysis (EDA)

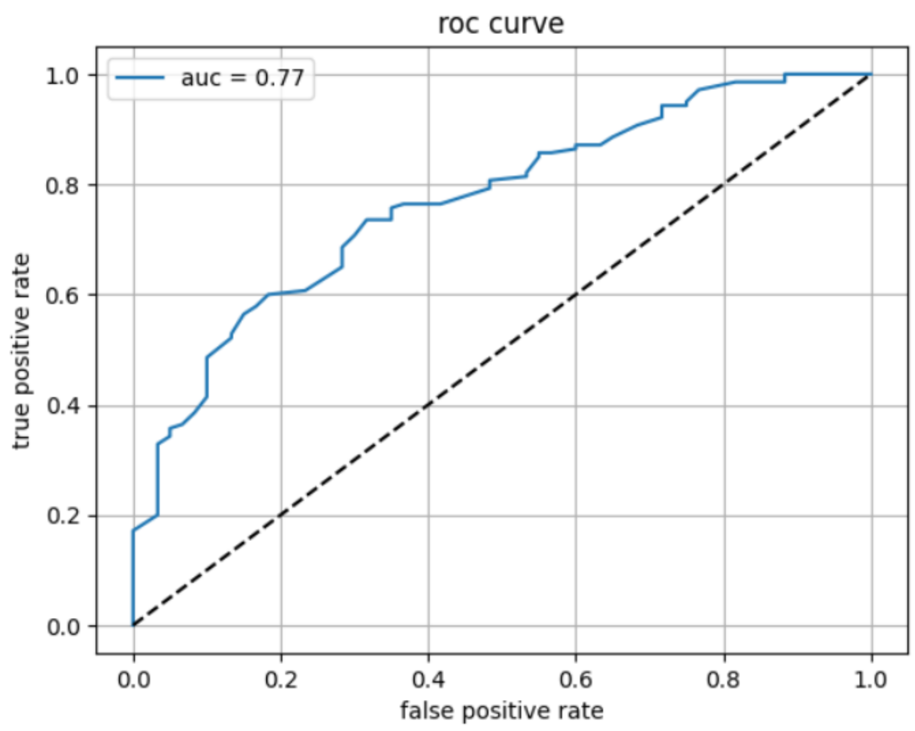
- Most customers are aged between **25 and 45 years**
- Class distribution: **70% Good credit**, **30% Bad credit**
- Higher loan amounts and longer loan durations tend to be associated with Bad credit
- Purpose of the loan and checking account status are also meaningful indicators

Modeling Approach

- **Model:** Random Forest Classifier
- Used `class_weight='balanced'` to compensate for class imbalance
- Data split: **80% for training, 20% for testing**
- No hyperparameter tuning was applied in this version

Evaluation Metrics

- **Accuracy:** ~73%
- **AUC Score:** ~0.79
- The model shows stronger performance for predicting **Good credit**
- ROC curve indicates decent separation between classes

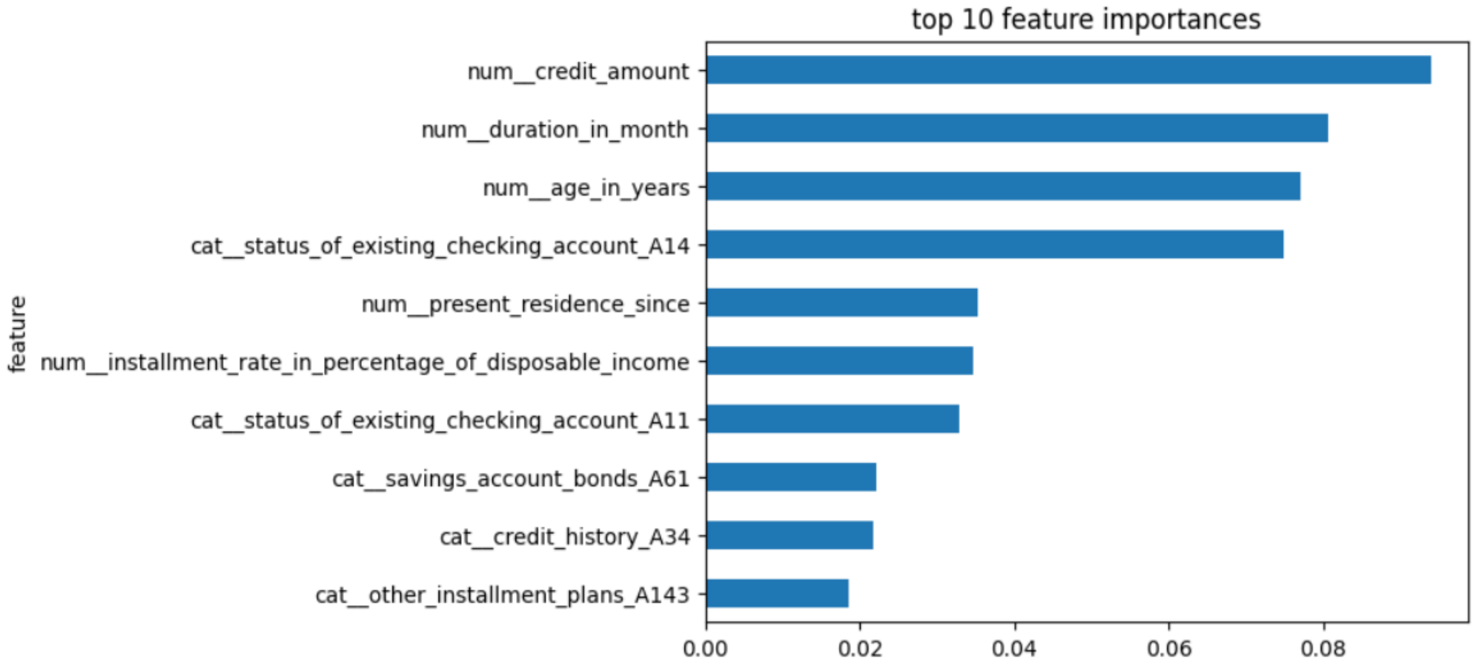


Feature Importance

Top predictive features:

- ***Credit amount***
- ***Duration in months***
- ***Checking account status***

The model gives higher importance to financial indicators than personal details



Why is "Checking Account Status" more important than "Age"?

- Feature importance in Random Forest is based on how well a feature splits data to reduce classification error.
- Checking Account Status provides direct insight into a customer's current financial situation (e.g. no account, low balance).
- It's a categorical variable that clearly separates “Good” vs “Bad” credit cases.
- Age, while informative, shows weaker separation — older doesn't always mean lower risk.
- The model identified stronger, more consistent patterns in account status than in age.

Conclusion

Random Forest performs well on a small but real-world dataset

AUC of **0.79** suggests that the model has strong classification capability

Key features are consistent with expectations in credit scoring

Future improvements could include:

- Hyperparameter tuning
- Model comparison
- Incorporating external data (e.g. income, region, etc.)

Thank you for your attention!

Feel free to connect or explore more:

- 📧 Email: marekwisniewskiuk@gmail.com
- 🔗 GitHub: https://github.com/M4R3K21
- 💼 LinkedIn: https://www.linkedin.com/in/marek-wisniewski-209930320/