Risky Business: ML’s Guide to German Credit Scores"

João, Toan, Lora & Eileen

21.06.2024

PROJECT OVERVIEW

Business Problem

A small bank in Germany wants to automize the process of credit risk evaluation.

Objective

Using supervised machine learning to predict German credit card approval, incorporating vintage analysis and addressing data imbalance

DATA SELECTION & PREPARATION

DATA SET

Dataset includes personal and credit card applicant information for machine learning model development.

Shape:

1000 rows × 11 columns

- . Unnamed: 0
- Age
- Sex
- Job
- Housing
- Saving accounts
- Checking account
- Credit amount
- Duration
- Purpose
- Risk

FEATURES

Feature Engineering and Selection

Dataset Exploring

Numerical Features

- age, saving account, checking account, credit amount, duration, unnamed

Categorical Features (Ordinal)

- job

Categorical Features (Nominal)

- sex, housing, purpose, risk (target)

Feature Encoding

Feature Selection

- dropping saving account and checking account due to their high number of missing values and lack of correlation with target variable
- dropping column unnamed as it was an index

Feature Encoding

Label Encoder

- target variable Risk was label-encoded to binary values (0 for bad, 1 for good)

One-Hot Encoding

- transforming categorical variables by using one-hot encoding to convert them into numerical format
- sex: (male, female) -> (0,1)
- housing (own, rent, free) -> (housing_own, housing_rent, free)
- purpose: (car, education, ...) -> (purpose_car, purpose_education, ...)

Feature Scaling

- MinMaxScaler: applied MinMaxScaler to normalize numerical features to bring them within range [0,1]

MACHINE LEARNING MODELS

Model Building and Evaluation

K-NN

Accuracy: 0.60

Random Forest

Accuracy: 0.69

Bagging & Pasting

Accuracy: 0.67 & 0.65

**with/without sample replacement*

Adaptive Boosting

Accuracy: 0.70

HYPERPARAMETER TUNING

Hyper parameter Tuning and Model Optimization

Grid search

Accuracy: 0.71

Random search

Accuracy: 0.715

Both Grid Search and Random Search improved the performance of the AdaBoost model from the baseline accuracy of 70.5% to around 71-71.5%.

Key Findings

Overall, our machine learning prediction is ***solid****, which shows* ***a good model performance*** *with* ***reasonable error rates*** *and a* ***high proportion of explained variance****.*

Best Performance Model

Adaptive Boosting

Stability and Error Consistency

MAE: 0.29 & RMSE: 0.54

Stable and Consistent

Accuracy

70.5% ~ 71.5%

Explained Variance

R^2 score: 0.71

Strong relationship between the input features and the target variable.

REAL-WORLD

*An accuracy of 71% indicates the model can **serve as a basic tool** for preliminary screening of credit card applicants' risk. However, in real-world business applications, **a higher accuracy** may be needed to **minimize potential erroneous decisions.***

Application

- **Automated Approval**

--> Automatically approving low-risk users

--> Flagging high-risk users for manual review

--> Human intervention for users predict as high risk

- **Risk Management**

--> Optimizing processes

--> Reducing bad debt rates

Improvement

- **Continuous Learning**

--> Continuously collecting new user data

--> Regularly updating the model

- **Model Monitoring**

--> Monitoring the performance in real-time

--> Retraining or adjusting it promptly if performance drops.

Challenges and Learnings

- **Insufficient Dataset**

--> Quality: noisy or incomplete data
--> Quantity: not have enough diverse and representative data

- **Feature Selection and Engineering**

--> Relevance of Features: important features might be missing

- **Model Complexity and Capacity**

--> Model Over fitting: Adaptive Boost model is too complex relative to the amount of data available?

OUR

TEAM

THANK YOU!