Loan Interest Rate Prediction

CIMB Data Analyst Test

Rizki Teguh Kurniawan

30/03/2021

Development Process

1. Research & recreate model for Lending Club’s Default Problem
2. Data prep & engineering
    1. Handle missing data
        1. Drop features
        2. Fill with median and “MISSING” for numerical and categorical features
    2. Feature synthesis
        1. Manual: 7
        2. One hot encoding categorical features
        3. Final # of features: 50
3. Build Model
4. Evaluation

Model Selection

Model	Accuracy Score
SGD	0.3679
Ridge	0.5120
Random Forest	0.5055
Gradient Boosting	0.5252
Hist-Gradient Boosting	0.5285
Auto-Sklearn	0.5266
XGBoost	0.5284
LightGBM	0.5295
CatBoost	0.5325

Model Selection (Top 3)

Model	Label	Precision	Recall	Accuracy
XGBoost	1	0.5073	0.2086	0.5284
	2	0.4960	0.6484	
	3	0.5845	0.5700	
LightGBM	1	0.5181	0.1988	0.5295
	2	0.4960	0.6531	
	3	0.5849	0.5732	
CatBoost	1	0.5281	0.2169	0.5325
	2	0.4980	0.6500	
	3	0.5876	0.5746	

Best Model

Model	Label	Precision	Recall	Accuracy
CatBoost	1	0.5281	0.2169	0.5325
	2	0.4980	0.6500	
	3	0.5876	0.5746	

Key features (top 5) by abs corr:

1. income_verified_not_verified
2. inquiries_last_6mo
3. loan_income_ratio
4. income_verified_verified_income
5. debt_to_income

Problem, Solution, & Further Development

- Problem:
  - Insufficient supporting features
- Solution:
  - Use shap library to inspect feature importance
  - Merge with Lending Club’s data (proceed with domain expert’s advice)
- Next:
  - If solution above approved, retrain CatBoost model with new data
  - Publish work on research paper or public repository