LOAN DEFAULT PREDICTION

Tatiana Meshkova

Content

Overview of the problem

Loan default is a significant problem for the lending industry, resulting in financial losses and other negative consequences

Accurately predicting loan defaults can help banks minimize risk, allocate resources more effectively, comply with regulations, and promote fair lending practices

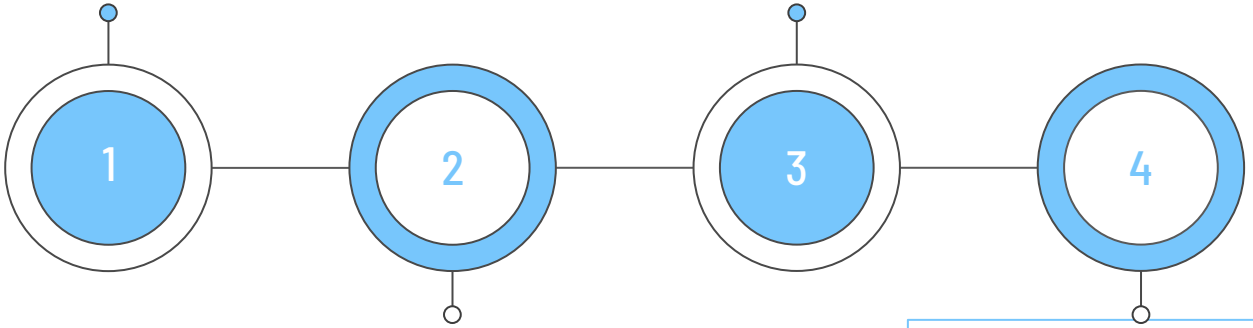
Developing an unbiased and data-driven loan approval process can enhance customer service quality and provide a competitive advantage

The goal is to build a classification model to predict loan defaults, considering important features and ensuring reliable performance and interpretability

Approach for the solution

Exploratory Data Analysis:
Analyzing the data to understand relationships and patterns between features and the target variable.

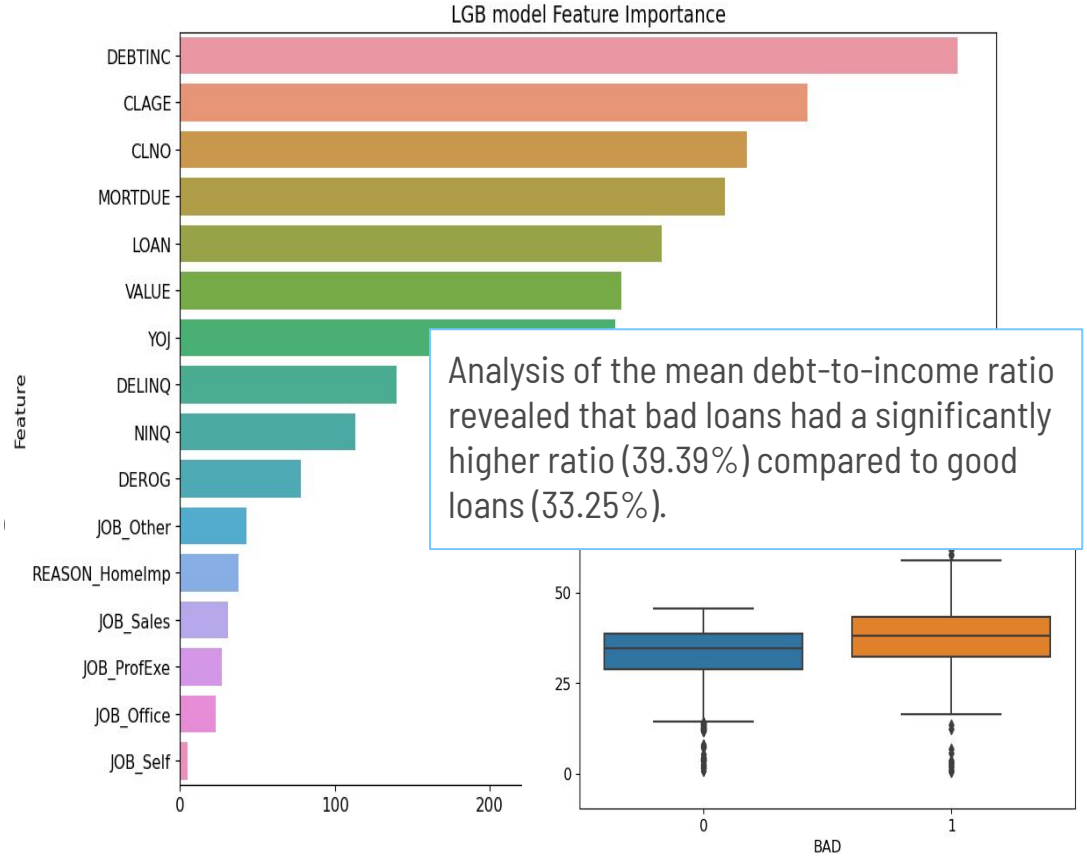
Model Evaluation:
Assessing the model's performance using evaluation metrics such as accuracy, precision, recall, and F1-score.



Model Selection and Training:
Comparing different machine learning models and selecting the best performing one, and fine-tuning the hyperparameters to improve its performance.

Choosing the optimal model for deployment and providing recommendations

Key findings & insights



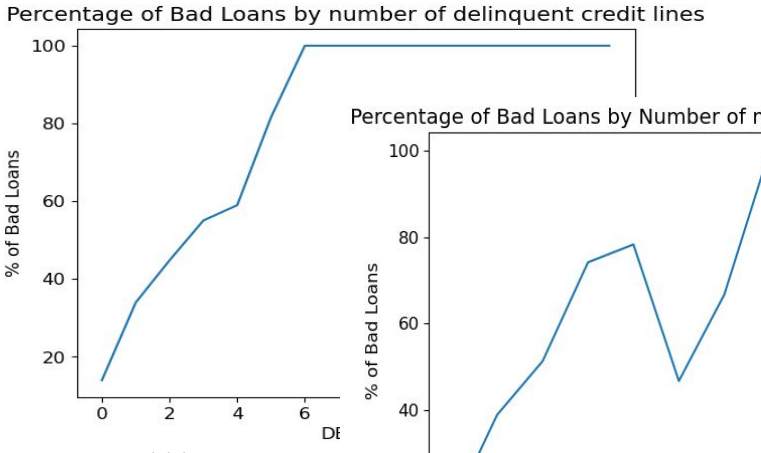
Analysis of the mean debt-to-income ratio revealed that bad loans had a significantly higher ratio (39.39%) compared to good loans (33.25%).

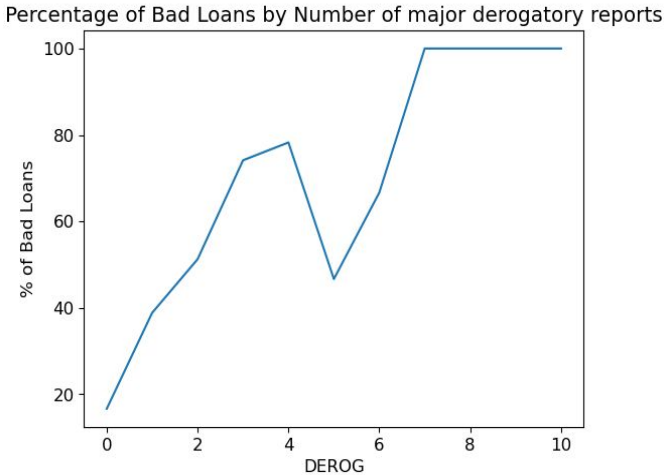
The most important features for predicting mortgage default are debt-to-income ratio (DEBTINC), age of oldest credit line in months (CLAGE), number of credit lines (CLNO), amount of mortgage owed (MORTDUE) and loan amount (LOAN),.

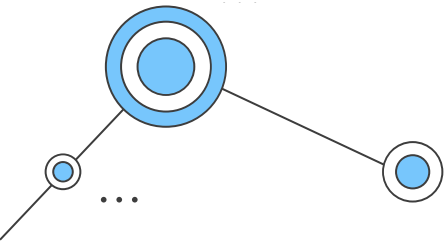
All TOP5 features were consistently identified across multiple models and the correlation matrix, indicating their strong impact on loan default prediction.

Key findings & insights

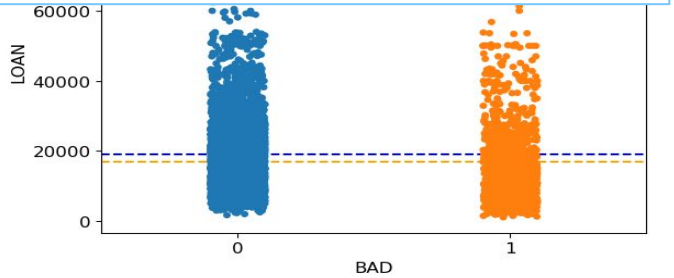
Analysis confirms the importance of delinquencies and derogatory reports as predictors of loan default.



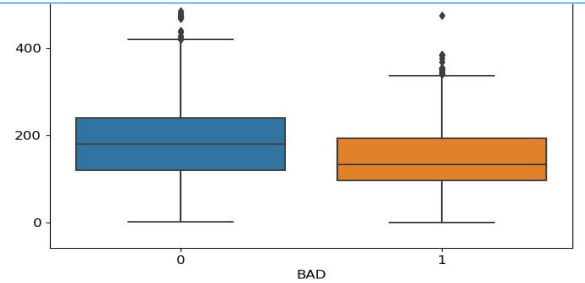




The average loan amount for customers who default on their loans is lower than for customers who pay back their loans.



Older credit lines are more reliable and may be a useful factor for lenders to consider when assessing loan applications.



Evaluation models

- Recall was chosen to evaluate the model's performance, as it tells us how many actual positive cases the model identified correctly, which is important in minimizing false negatives.
- Precision and F1-score are also considered to compare similar models, with Precision indicating the number of correctly identified positive cases and F1-score providing a balance between Precision and Recall.
- Considering all three metrics allows for a more informed decision on the best model for predicting loan default.
- In addition, the model's training time and ease of implementation were also taken into account when selecting the best model for deployment.

Choosing model

- Based on the analysis and evaluation of several models, I propose to adopt the **LGBM Classifier** as the final solution.
- This model outperformed the other models in terms of metric scores, as well as having a reasonable training time. It also provides feature importance ranking, which can help to identify the most important factors affecting the target variable.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	1431
1	1.00	0.82	0.90	357
accuracy			0.96	1788
macro avg	0.98	0.91	0.94	1788
weighted avg	0.97	0.96	0.96	1788

Next step – model implementation

- Collect data on borrowers, including both numerical and categorical variables, and split the data into training and testing sets.
- Train a predictive model using the LGBM Classifier on the training set, using the identified important predictors of loan default.
- Integrate the model into the bank's loan approval process, using the model's predictions as one of many factors to consider when making lending decisions.
- Set a threshold for the model's predicted probability of default and only approve loans for borrowers whose predicted probability of default falls below that threshold.

Recommendations

- Implement the LGBM Classifier into the loan approval process to more accurately assess the risk of loan default and take appropriate measures to mitigate that risk.
- Continue monitoring the model's performance in production and updating it as needed to ensure that it remains accurate and effective in identifying borrowers at risk of loan default.
- Establish a process for updating the model as needed, taking into account changes in economic conditions, regulatory requirements, and borrower profiles.
- Communicate the use of the model to customers to promote transparency and build trust, providing clear explanations of how the model works and how it informs lending decisions.

The potential benefits

01

Increase revenue: Our model can accurately predict loan defaults, allowing us to avoid risky borrowers and focus on profitable ones.

02

Mitigate risk: By avoiding loans that are likely to default, we can reduce the risk of financial losses and improve the overall health of our loan portfolio.

03

Better customer experience: With a faster and more accurate loan approval process, customers will have a better experience and are more likely to return for future loans.

04

Scalability: The solution can be easily scaled to handle an increasing volume of loan applications without requiring significant additional resources or manual labor.

Appendix

Correlation matrix

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	REASON_HomeImp	JOB_Office	JOB_Other	JOB_ProfExe	JOB_Sales	JOB_Self
BAD	1.00	-0.08	-0.03	-0.03	-0.06	0.27	0.35	-0.18	0.17	-0.00	0.26	0.04	-0.07	0.04	-0.04	0.05	0.05
LOAN	-0.08	1.00	0.23	0.33	0.10	0.01	-0.03	0.09	0.05	0.07	0.06	-0.15	-0.02	-0.06	0.02	-0.04	0.16
MORTDUE	-0.03	0.23	1.00	0.87	-0.07	-0.05	0.01	0.13	0.03	0.34	0.14	-0.04	-0.06	-0.27	0.26	0.02	0.10
VALUE	-0.03	0.33	0.87	1.00	0.01	-0.05	0.00	0.17	-0.00	0.27	0.10	-0.00	-0.06	-0.27	0.27	0.02	0.15
YOJ	-0.06	0.10	-0.07	0.01	1.00	-0.06	0.05	0.19	-0.06	0.03	-0.05	0.05	-0.03	0.03	-0.00	-0.04	-0.04
DEROG	0.27	0.01	-0.05	-0.05	-0.06	1.00	0.17	-0.08	0.15	0.05	0.05	-0.01	-0.05	0.03	-0.03	0.03	0.00
DELINQ	0.35	-0.03	0.01	0.00	0.05	0.17	1.00	0.03	0.06	0.16	0.09	0.00	0.01	-0.04	-0.00	-0.02	0.02
CLAGE	-0.18	0.09	0.13	0.17	0.19	-0.08	0.03	1.00	-0.11	0.24	-0.06	0.05	-0.00	-0.08	0.11	0.03	-0.01
NINQ	0.17	0.05	0.03	-0.00	-0.06	0.15	0.06	-0.11	1.00	0.09	0.14	-0.12	-0.06	0.05	-0.06	-0.03	0.01
CLNO	-0.00	0.07	0.34	0.27	0.03	0.05	0.16	0.24	0.09	1.00	0.16	-0.11	-0.00	-0.20	0.17	0.05	0.03
DEBTINC	0.26	0.06	0.14	0.10	-0.05	0.05	0.09	-0.06	0.14	0.16	1.00	-0.02	0.01	-0.02	-0.07	0.05	0.04
REASON_HomeImp	0.04	-0.15	-0.04	-0.00	0.05	-0.01	0.00	0.05	-0.12	-0.11	-0.02	1.00	0.02	-0.02	0.02	-0.06	0.12
JOB_Office	-0.07	-0.02	-0.06	-0.06	-0.03	-0.05	0.01	-0.00	-0.06	-0.00	0.01	0.02	1.00	-0.39	-0.23	-0.06	-0.08
JOB_Other	0.04	-0.06	-0.27	-0.27	0.03	0.03	-0.04	-0.08	0.05	-0.20	-0.02	-0.02	-0.39	1.00	-0.47	-0.12	-0.16
JOB_ProfExe	-0.04	0.02	0.26	0.27	-0.00	-0.03	-0.00	0.11	-0.06	0.17	-0.07	0.02	-0.23	-0.47	1.00	-0.07	-0.10
JOB_Sales	0.05	-0.04	0.02	0.02	-0.04	0.03	-0.02	0.03	-0.03	0.05	0.05	-0.06	-0.06	-0.12	-0.07	1.00	-0.02
JOB_Self	0.05	0.16	0.10	0.15	-0.04	0.00	0.02	-0.01	0.01	0.03	0.04	0.12	-0.08	-0.16	-0.10	-0.02	1.00

Test performance comparison

Model	Recall	Precision	Accuracy	Total Rank
Tuned Catboost classifier oversampling without one-hot encoding	0.916319	0.916319	0.916319	4
lgbm classifier	0.910015	0.910015	0.910015	6
Lgbm classifier without one-hot encoding	0.909665	0.909665	0.909665	12
Tuned Catboost classifier oversampling	0.912819	0.912819	0.912819	13
Tuned Catboost classifier	0.908265	0.908265	0.908265	16
Tuned lgbm classifier oversampling	0.909668	0.909668	0.909668	20
Tuned lgbm classifier	0.907566	0.907566	0.907566	22
Catboost classifier	0.904063	0.904063	0.904063	24
XGBoost classifier	0.905114	0.905114	0.905114	25
Decision Tree classifier	0.907589	0.907589	0.907589	32
Tuned Catboost classifier undersampling	0.908996	0.908996	0.908996	33
Random Forest classifier	0.896361	0.896361	0.896361	33
Tuned Random Forest classifier	0.872200	0.872200	0.872200	37
Tuned lgbm classifier undersampling	0.902369	0.902369	0.902369	42
Tuned Random Forest classifier 2	0.861016	0.861016	0.861016	43
Tuned Decision Tree classifier	0.883830	0.883830	0.883830	46

Test performance comparison

Model	Time	Rank
Lgbm classifier without one-hot encoding	0.35	**1**
Lgbm classifier	0.39	**2**
Tuned Catboost classifier oversampling without one-hot encoding	01.01	**3**
Tuned Catboost classifier oversampling	1.32	**4**

Set a threshold 20%

	yl_test	yl_pred_proba_lgbm	approved
4394	0	0.002379	1
5000	0	0.018923	1
2786	0	0.00344	1
2256	0	0.005746	1
114	0	0.307778	0
3787	0	0.000594	1
1289	0	0.00525	1
1189	1	0.997221	0
4715	0	0.002677	1
70	1	0.894823	0
1268	1	0.989448	0