Prediction of Loan Default

Introduction

Financial institutions encounter a considerable challenge in predicting loan defaults. This poses an urgent dilemma for banks: identifying the customers most likly to defaulting on their loan obligations. Given the significant implications for a bank's financial stability and strategic decisions, accurately predicting defaults in the banking sector carries substantial importance.

The main objective of the project is to use dataset comprising historical data of customers who have availed bank loans, to develop a predictive machine learning model capable of accurately forecasting a customer's probability of default. This model leverages insights from diverse historical features associated with each individual customer. In this context, we are try to answer the following questions:

Given the German bank dataset, which machine learning model predicts loan defaults the best?

In order to reduce false negatives and more accurately identify possible defaulters, how can we optimize the performance of the model (recall)?

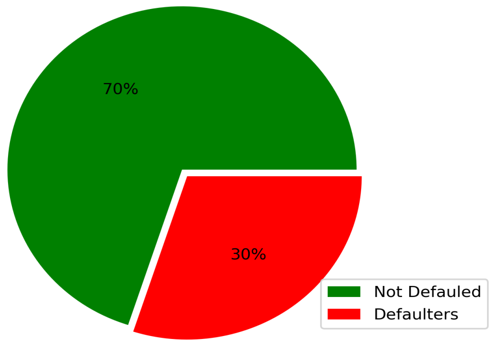
Does the chance of a loan default depend on financial related features? What about credit history, as it may be a significant indicator of default risk?

Methods and Materials

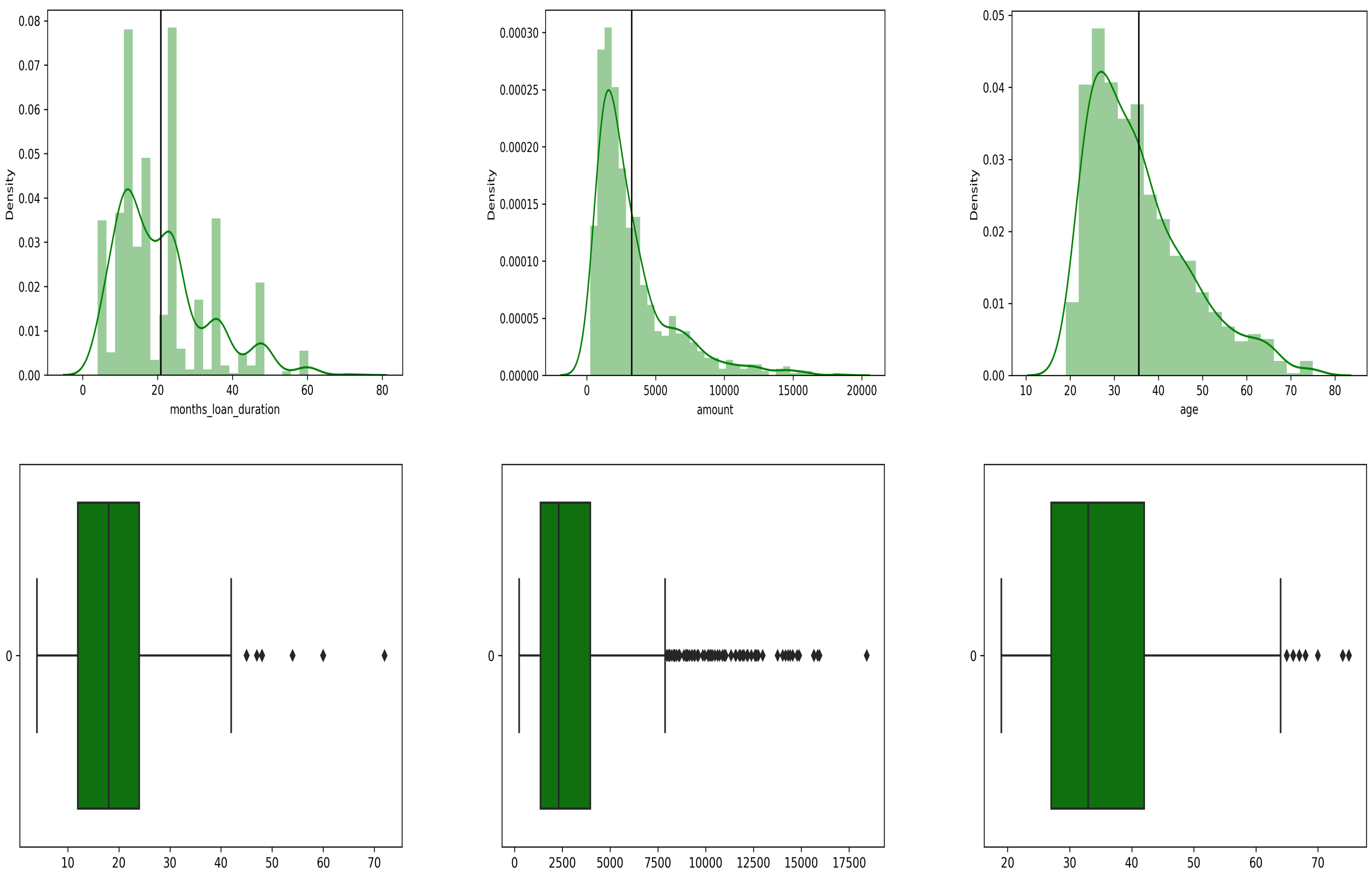
Start with conducting exploratory data analysis aiming to have insight from the dataset, first by looking at features data type and their values, no missing values are there, although some categorical attributes have 'unknown', 'none' or 'unemployment' that makes them nominal instead of ordinal, removing these categories could result in losing important information. There is also no redundancy in the dataset. Using graphs and statistics to study variables, then apply set of classification models and optimize their performances to select best possible model. Moreover, addressing some issue with the dataset in hand like imbalanced classes of the target.

Exploratory Data Analysis

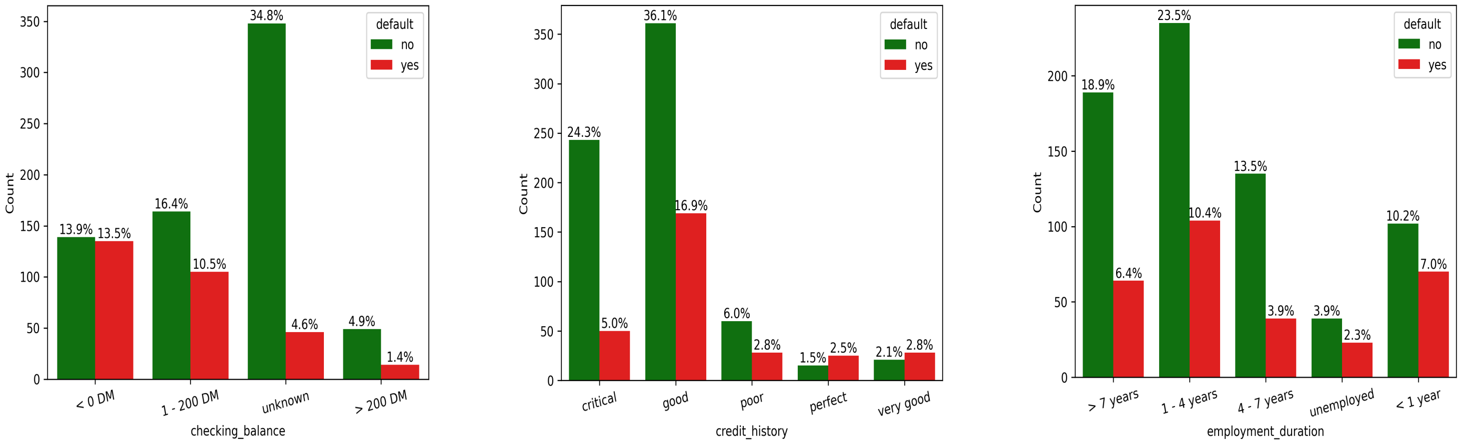
Through visualization and statistical summaries, we gain clear understanding of features and relationships between them. The target classes are imbalanced.



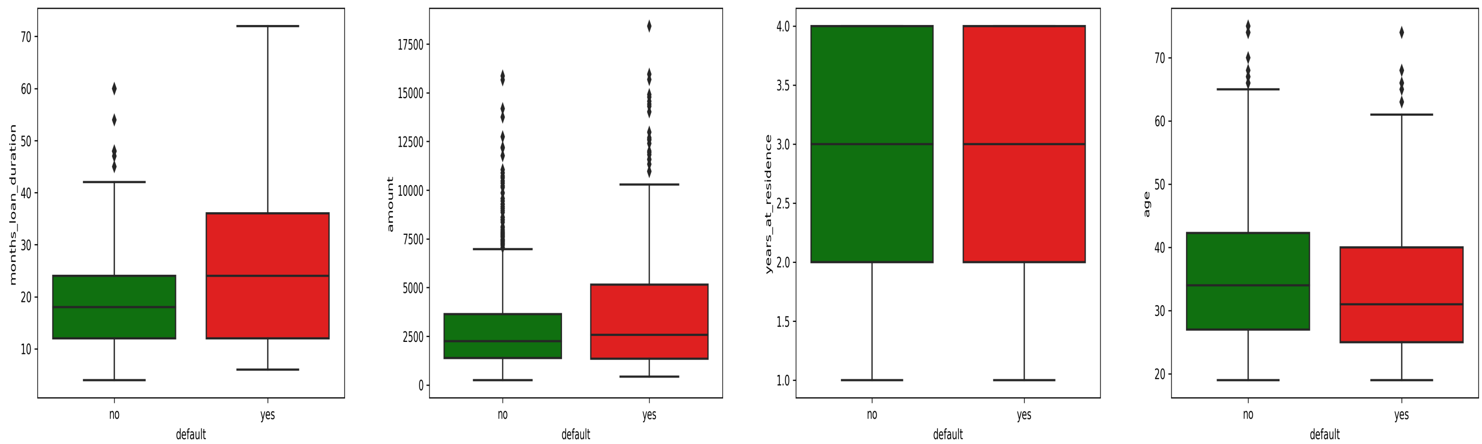
Continues variables (loan duration, amount of the loan, and age) are positively skewed and have different range of their values, box plots show that most of the amounts are between 1200 and 4000 dollars, most of the loan duration is from 11 to 25 months, and majority of the loan applicants have age between 28 - 42(below figures).



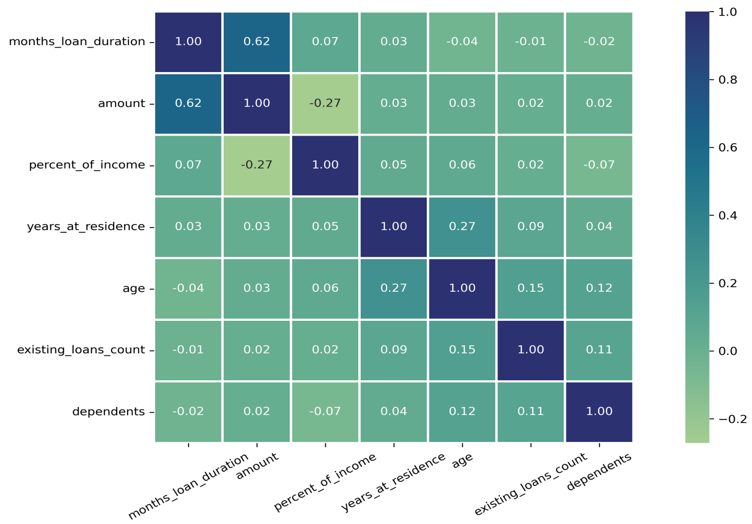
Also 'unknown' is the most frequent among checking balance categories which have the lower probability of been defaulted. Surprisedly, applicants with critical credit history have lower probability of been default. Also, Applicants with shorter employment duration tend to default more.



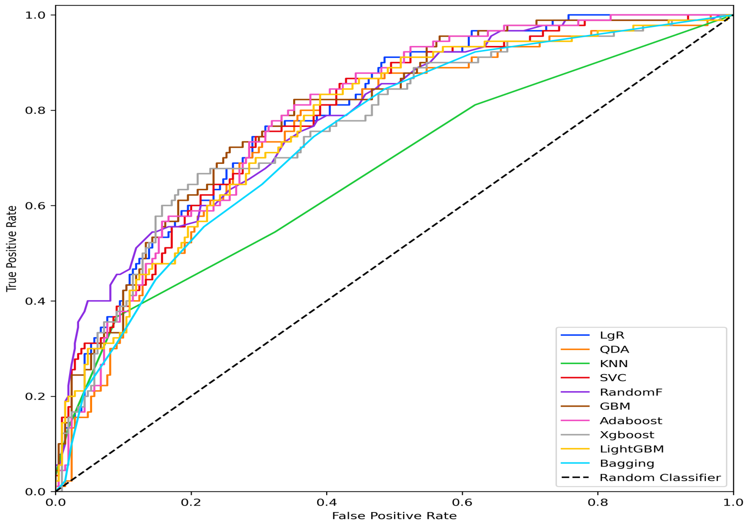
Duration of the loan has higher mean among those whom defaulted(yes), with bigger range. Mean of loan amount is in close proximity among the two default classes, same for years at residence, and younger customers tend to default.



From correlation matrix, loan amount and duration have strong positive correlation, which is expected, also amount have weak negative correlation with percentage of income, and age have weak positive correlation with yeas at residence. There is no other significant correlation.



Results



First, couple models were implemented with their default parameters. In term of overfitting, the Logistic regression is the only model that does not overfit, especially when we focus on recall score. All test recall score are poor, ranging from Quadratic Discriminant with 0.54 recall to the worse model K-nearest neighbours with 0.32. Three models had perfect fit on training data, Random Forest, XGBoost, and Light Gradient Boosting. Gradient Boosting has the highest test f1 score 0.56.

Metrics	LgR_Train	LgR_Test	QDA_Train	QDA_Test	KNN_Train	KNN_Test	SVC_Train	SVC_Test
Accuracy	0.77	0.74	0.82	0.72	0.83	0.72	0.83	0.75
Recall	0.47	0.41	0.71	0.54	0.55	0.32	0.52	0.36
Precision	0.67	0.61	0.7	0.53	0.82	0.58	0.85	0.64
f1	0.55	0.49	0.71	0.54	0.66	0.41	0.64	0.46

Metrics	RandomF_Trai	RandomF_Test	GBM_Train	GBM_Test	Adaboost_Trai	Adaboost_Test	Xgboost_Train	Xgboost_Test
Accuracy	1.0	0.77	0.92	0.77	0.81	0.74	1.0	0.76
Recall	1.0	0.4	0.76	0.5	0.58	0.47	1.0	0.5
Precision	1.0	0.73	0.96	0.67	0.72	0.6	1.0	0.62
f1	1.0	0.51	0.85	0.56	0.65	0.51	1.0	0.55

Metrics	LightGBM_Train	LightGBM_Test	Bagging_Train	Bagging_Test	Cat_Train	Cat_Test
Accuracy	1.0	0.76	0.98	0.74	0.88	0.77
Recall	1.0	0.48	0.95	0.39	0.64	0.41
Precision	1.0	0.63	1.0	0.62	0.93	0.7
f1	1.0	0.54	0.97	0.47	0.76	0.51

After tunning the important hyperparameters (below tables) still most models are overfitting, Logistic regression has improved in term of recall score now is 0.7 and it's not overfitting but the model and many other models test performance has overcome the performance on train set. Light Gradient Boosting has the highest recall score of 0.9, and still K-nearest neighbours has the worst performance.

Metrics	LgR train	LgR test	QDA train	QDA test	KNN train	KNN test	SVC train	SVC test	RF train	RF test	GB train	GB test
Accuracy	0.57	0.6	0.81	0.72	1.0	0.71	1.0	0.7	0.7	0.7	1.0	0.75
Recall	0.69	0.7	0.69	0.57	1.0	0.39	1.0	0.57	0.0	0.0	1.0	0.54
Precision	0.38	0.41	0.68	0.54	1.0	0.52	1.0	0.5	0.0	0.0	1.0	0.58
f1	0.49	0.51	0.69	0.55	1.0	0.45	1.0	0.53	0.0	0.0	1.0	0.56

Metrics	1_GB trair	1_GB test	AdaB trair	AdaB test	XGB train	XGB test	1_XGB tra	1_XGB tes	Light train	Light test	Bagg train	Bagg test
Accuracy	0.79	0.75	0.8	0.74	0.98	0.75	0.95	0.75	0.58	0.65	1.0	0.73
Recall	0.46	0.43	0.53	0.47	0.94	0.49	0.86	0.53	0.82	0.9	1.0	0.43
Precision	0.72	0.62	0.74	0.58	0.99	0.6	0.96	0.59	0.4	0.46	1.0	0.57
f1	0.56	0.51	0.61	0.52	0.96	0.54	0.9	0.56	0.54	0.61	1.0	0.49

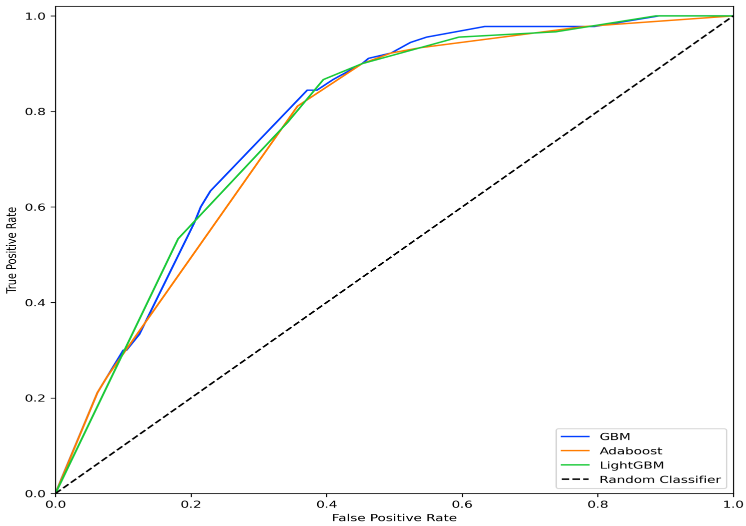
Since the target is imbalanced, oversampling is performed to address this issue. Random forest has the perfect recall score 1.0 but with poor f1 score. Best performance is achieved by Gradient Boosting, AdaBoosting, and light boosting with 0.9 recall score on test set and 0.61 f1. Moreover, support vector machine has 0.89 recall(below tables).

Metrics	LgR train	LgR test	QDA train	QDA test	KNN train	KNN test	SVC train	SVC test	RF train	RF test	GB train	GB test
Accuracy	0.79	0.7	0.8	0.7	1.0	0.69	0.68	0.54	0.5	0.3	0.68	0.65
Recall	0.8	0.64	0.86	0.68	1.0	0.52	0.92	0.89	1.0	1.0	0.88	0.9
Precision	0.78	0.5	0.76	0.5	1.0	0.48	0.62	0.38	0.5	0.3	0.63	0.46
f1	0.79	0.57	0.81	0.57	1.0	0.5	0.74	0.54	0.67	0.46	0.73	0.61

Metrics	1_GB train	1_GB test	AdaB train	AdaB test	XGB train	XGB test	Light train	Light test	Bagg train	Bagg test
Accuracy	0.68	0.65	0.68	0.65	0.87	0.72	0.68	0.65	1.0	0.73
Recall	0.88	0.9	0.88	0.9	0.91	0.67	0.88	0.9	1.0	0.57
Precision	0.63	0.46	0.63	0.46	0.84	0.53	0.63	0.46	1.0	0.55
f1	0.73	0.61	0.73	0.61	0.87	0.59	0.73	0.61	1.0	0.56

Discussion

From EDA we fund that there is no clear relationship between credit history and default status, applicants with critical credit history have the lower probability among all categories.



After oversampling the minority and tunning the models, these three model in the table below, relatively has the best performance with the highest recall score on test set 0.9, and moderate f1 score 0.61, but still there is some signs of overfitting since the f1 score difference between test and train sets is kind of big 0.12. Furthermore, recall score for test is higher than on train, we have witnessed this behaviour along the analysis of our models, it might be because of some difficult data points, this will lead us on discussing the limitations.

Metrics	GBM train	GBM test	Adaboost train	Adaboost test	LightGBM train	LightGBM test
Accuracy	0.68	0.65	0.68	0.65	0.68	0.65
Recall	0.88	0.9	0.88	0.9	0.88	0.9
Precision	0.63	0.46	0.63	0.46	0.63	0.46
f1	0.73	0.61	0.73	0.61	0.73	0.61
zero_1_loss_	0.32	0.35	0.32	0.35	0.32	0.35
AUC	0.68	0.72	0.68	0.72	0.68	0.72

Many issues have emerged during this study, starting with the relatively small dataset, which in somesituations cause overfitting, and in other cause test to overcome the training performance. Also in some predictors, there were some unknown categories. Additionally, the target is imbalanced, even though applying oversampling still this issue has it effect specially with the fact that the dataset is small.

Conclusions

In conclusion, while our analysis sheds light on effective prediction models for identifying risky loan applicants, it is essential to acknowledge the limitations inherent in our study. With the inclusion of more extensive datasets encompassing a broader range of samples and features, and ongoing model development, financial institutions can develop more robust prediction models to enhance their risk assessment processes and ultimately improve profitability.