Project Report on Default Prediction

Problem Identification

Loan default prediction is crucial for banks because it directly impacts their financial stability and profitability. Accurate prediction models allow banks to identify high-risk borrowers and mitigate potential losses by adjusting lending strategies, setting appropriate interest rates, and implementing robust risk management practices. This not only helps in maintaining the bank's credit quality and reducing the incidence of non-performing loans but also enhances overall operational efficiency by enabling better allocation of resources. Additionally, effective default prediction models contribute to maintaining regulatory compliance and preserving the bank's reputation, fostering trust among investors and customers. Overall, predicting loan defaults is fundamental to sustaining a healthy financial ecosystem within the banking sector.

In this data science project, I developed a predictive model to assess the probability of loan default based on comprehensive borrower and loan information. The model leverages machine learning techniques to analyze various factors such as credit scores, employment history, and loan amount, aiming to accurately forecast the likelihood of a borrower defaulting on a loan. The ultimate goal is to provide a robust tool that aids financial institutions in minimizing potential losses, optimizing loan portfolios, and improving overall financial stability.

Data Collection, Organization, and Definitions

Here is a preview of the raw dataset:

	customer_id	credit_lines_outstanding	loan_amt_outstanding	total_debt_outstanding	income	years_employed	fico_score	default
0	8153374	0	5221.545193	3915.471226	78039.38546	5	605	0
1	7442532	5	1958.928726	8228.752520	26648.43525	2	572	1
2	2256073	0	3363.009259	2027.830850	65866.71246	4	602	0
3	4885975	0	4766.648001	2501.730397	74356.88347	5	612	0
4	4700614	1	1345.827718	1768.826187	23448.32631	6	631	0

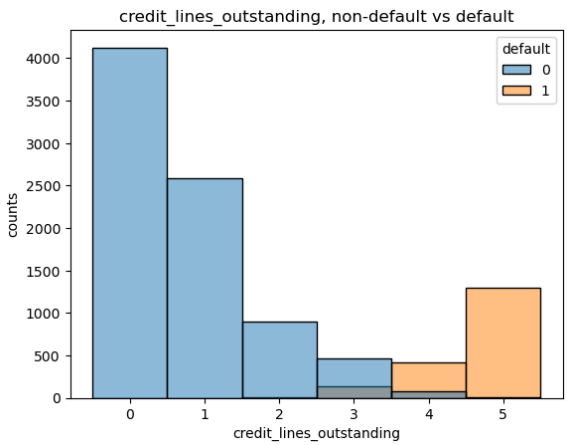
The dataset has 10,000 rows. The first a few columns will be used as predictor variables (features) and the last column, default, is the target variable. Here 0 indicates no default and 1 indicates a default event. Therefore we will also make probabilistic predictions of values between 0 and 1 and calculate performance metrics from these values.

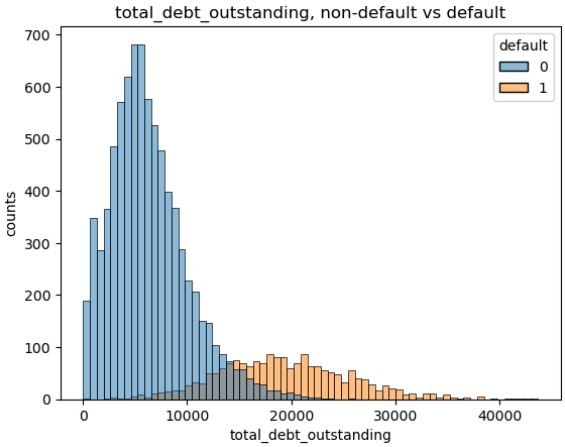
A brief overview of the predictor variables is as follows:

1. credit_lines_outstanding: Line of credit is one way the customer could borrow funds. They will receive a credit limit and make regular payments on principal and interest. They have continuous and repeated access to funds. ***Integer values, from 0 to 5. Although 41.3% has value zero, it is unclear if any of them indicate unknown values.***
2. loan_amt_outstanding: Loan is the other way the customer could borrow funds. They will only have access to the fund once and make payments on principal and interest until the loan is paid off. ***Floating point values, from roughly 47 to 10,751 with a mean of 4160. Bell-shaped. All unique values.***
3. total_debt_outstanding: potentially referring only to the debt through line of credit. ***Floating point values, from roughly 32 to 43,689 with a mean of 8719. Negatively skewed, although no negative values. All unique values.***
4. income: self-explanatory. ***Floating point values, from roughly 100 to 148,412 with a mean of 70,040. Bell-shaped. Values are mostly distinct except for 6 occurrencies of 1,000, which might be a place-holder value and might call for special considerations in later analysis.***
5. years_employed: self-explanatory. ***Interger values, from 0 to 10. Bell-shaped which makes the zeros likely meaningful values.***
6. fico_score: FICO scores take into account data in five areas to determine a borrower's credit worthiness: payment history, the current level of indebtedness, types of credit used, length of credit history, and new credit accounts. Scores range from 300 to 850. ***Integer values, from 408 to 850, agreeing with range previously stated.***

There are no missing values or obviously ‘wrong’ values (negative/out of expected range) so I proceeded to some exploratory data analysis.

Exploratory Data Analysis





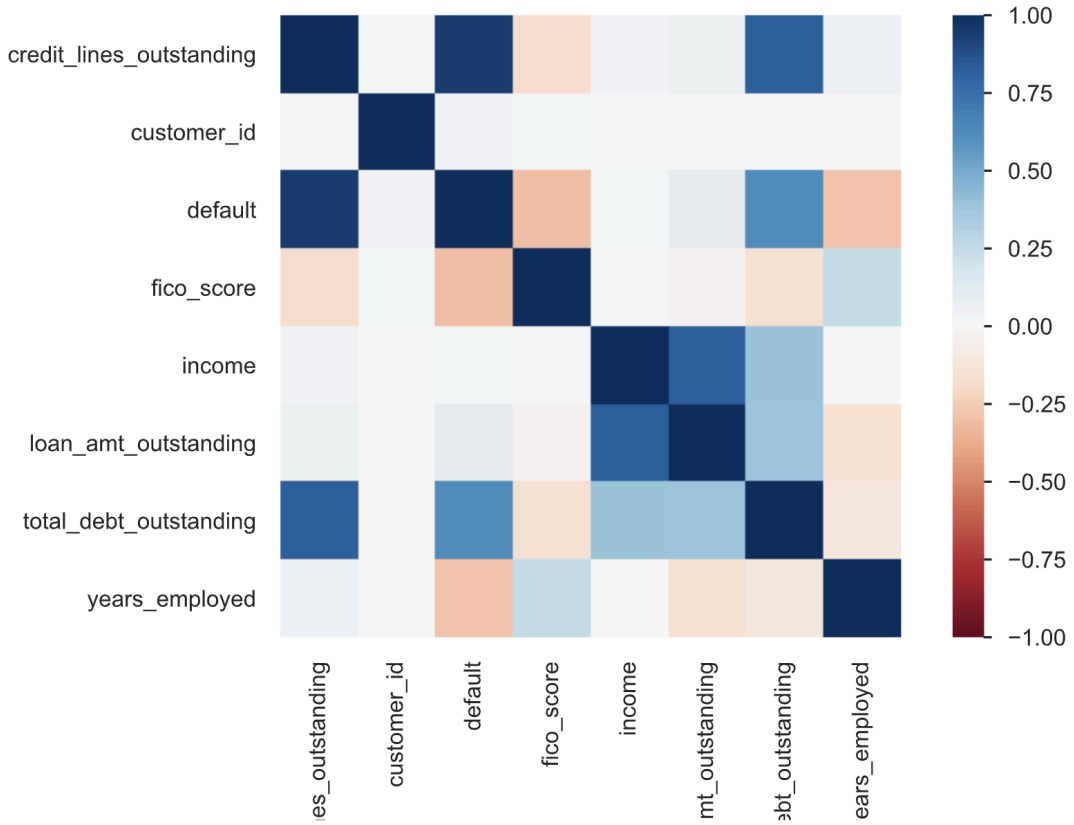
I first compared the default vs. no default classes for each predictor variable. I found that:

1. credit_lines_outstanding and total_debt_outstanding seem significant in explaining the probability of default (see graph above)

2. loan_amt_outstanding and income do not seem to be as prominent features in comparison
3. years_employed and fico_score share some similarities in distribution, which might be a result of how fico scores are calculated which is out of scope of this particular project but nonetheless interesting; Two-sample t-test reveals that the difference in population mean of default and non-default might be significant

I then investigated the correlations, and found some pairwise relationships:

1. credit_lines_outstanding, total_debt_outstanding (and both with default)
2. income, loan_amt_outstanding
3. years_employed and fico_score



Pre-processing and Training Data Development

During this step, we performed the training/validation split. Since the classes are imbalanced (80 to 20 no default to default), the split is stratified. We then fit a scaler on the training data as some algorithm are distance based, and finally used the scaler to transform the validation data.

Modeling

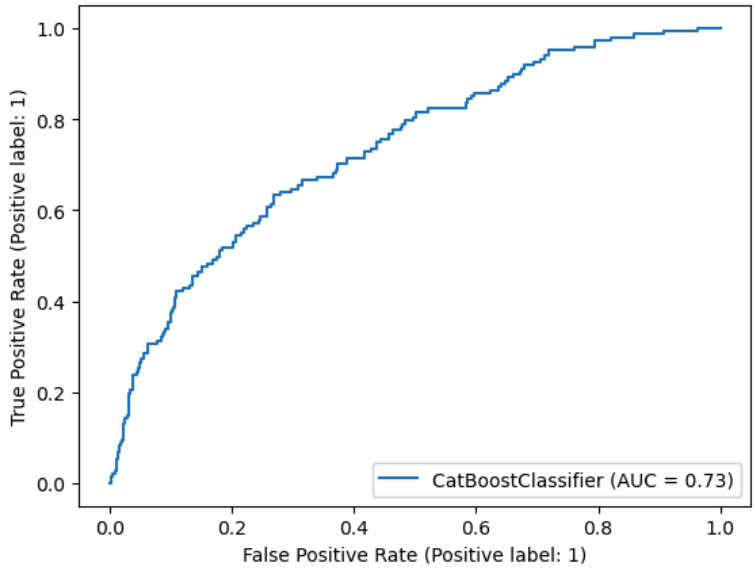
CatBoost

The first model tested was CatBoost (more information on the model could be found here: https://catboost.ai/en/docs/). A variety of metrics were considered in order to gain a more comprehensive understanding of model performance.

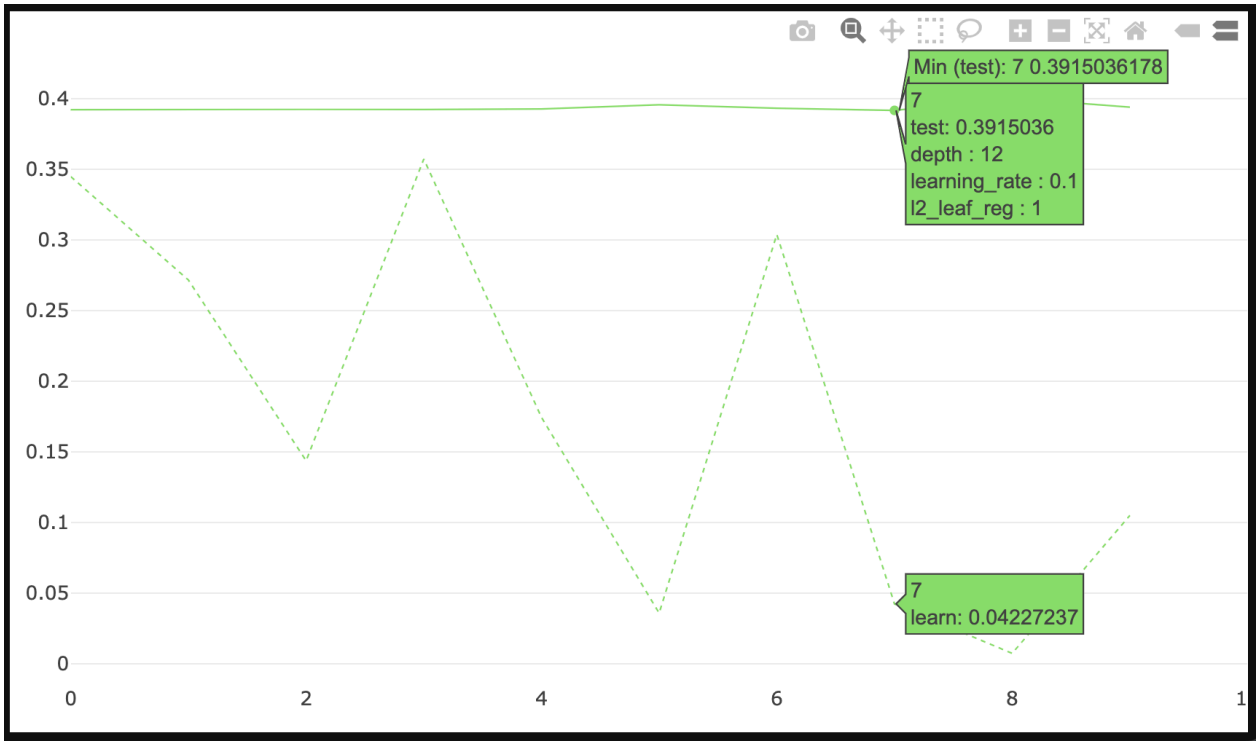
The first test model returned an almost perfect result (0.9998958235667927 AUC-ROC score), which shifted the goal of the project to training a model capable of predicting on fewer features. In order to do this we eliminated the features with the most predictive powers.

	Feature Id	Importances
0	credit_lines_outstanding	41.642504
1	years_employed	27.133336
2	total_debt_outstanding	10.848737
3	fico_score	10.598897
4	income	6.910185
5	loan_amt_outstanding	2.866341

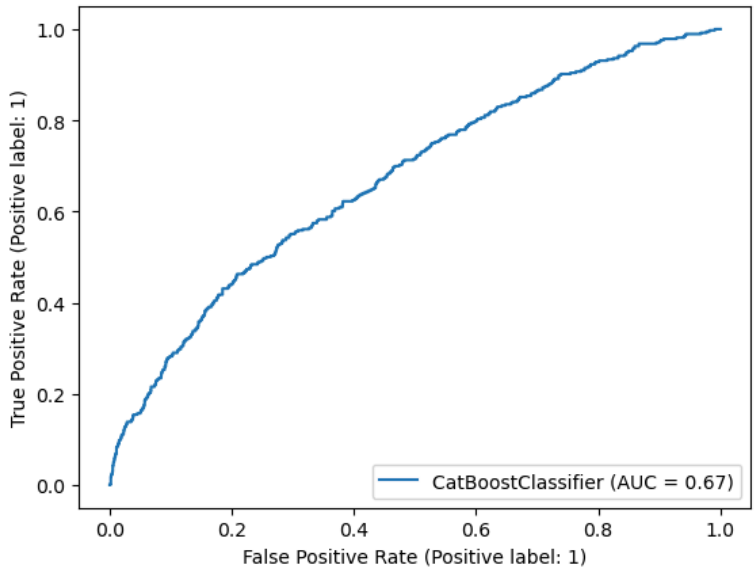
As was expected by exploratory data analysis, the credit_lines_oustanding is a very prominent feature. I chose to keep 'income,' 'years_employed,' and 'fico_score' to test the effect of parameter tuning. Here is the ROC curve after the aggressive feature reduction.



A random grid search was performed and the figure below shows the performance of different sets of hyperparameters. The best-performing set is used in the final model, trained using all of the training data.

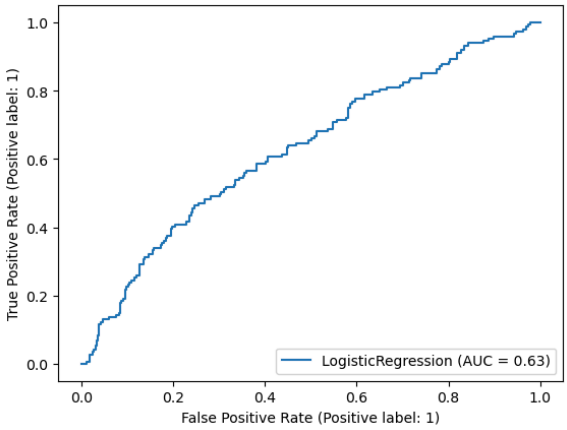


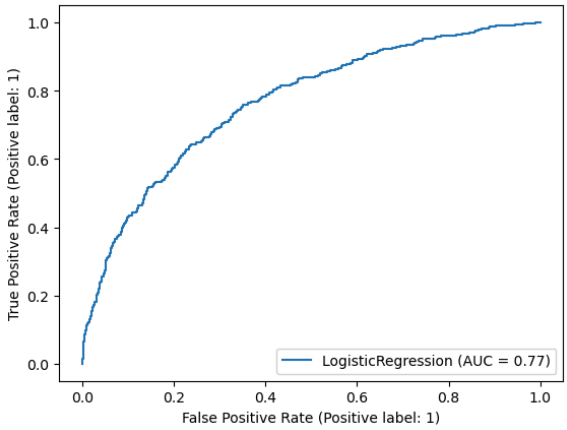
Here is the ROC curve of the final CatBoost model:



Logistic Regression

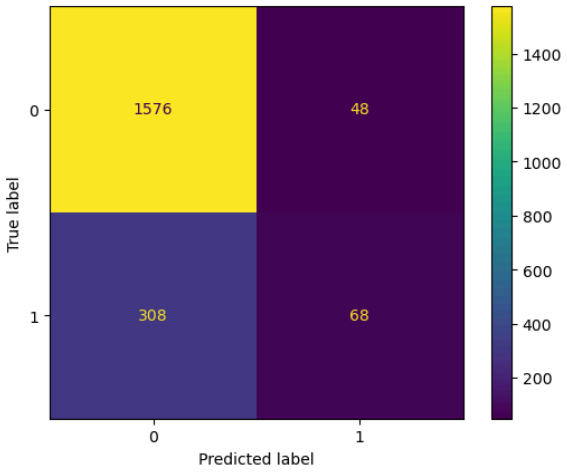
Similarly, we built a test model with the Scikit-learn Logistic Regression classifier then performed a random grid search. Here is a comparison of the default model and the final model:





Final Model Selection and Closing Thoughts

By AUC score, the final model is selected to be Scikit-learn's logistic regression classifier, with parameters {'solver': 'newton-cholesky', 'C': 0.75}. The model has a AUC score 0.77, and here are some more model metrics.



	precision	recall	f1-score	support
0(no default)	0.84	0.97	0.90	1624
1(default)	0.59	0.18	0.28	376
accuracy			0.82	2000
macro avg	0.71	0.58	0.59	2000
weighted avg	0.79	0.82	0.78	2000

Brier score loss: 0.12757509999134695
log loss: 0.4087386060695741
ROC_AUC score: 0.7668041216853578

Upon considering the use-case (help banks in decisions behind issuing a loan), it may be helpful to build a model that predicts potential loss, i.e. default probability times the loan amount. This would give insight into the overall risk of the portfolio.