Credit Risk Prediction

Nutritional Label for the Automated Decision System

Mimi Chen & Kieren Gill

Background

Our ADS

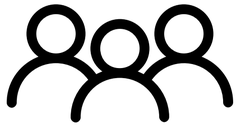
The ADS is a model that borrowers can use to help them make financial decisions. It works to improve credit scoring algorithms by predicting the probability that somebody will experience financial distress in the next two years.

With this model, borrowers can better understand what they need to do to improve their credit scores.

Background

Stakeholders

Banks



People

↓ False-negative rate,

fewer people default or experience financial distress

↑ False-positive rate,

borrowers who would have had a chance to receive a loan may not be entitled to do so anymore.

Inputs and Outputs

Inputs

#	Column	Data Type
0	SeriousDlqin2yrs	int64
1	RevolvingUtilizationOfUnsecuredLines	float64
2	age	int64
3	NumberOfTime30-59DaysPastDueNotWorse	int64
4	DebtRatio	float64
5	MonthlyIncome	float64
6	NumberOfOpenCreditLinesAndLoans	int64
7	NumberOfTimes90DaysLate	int64
8	NumberRealEstateLoansOrLines	int64
9	NumberOfTime60-89DaysPastDueNotWorse	int64
10	NumberOfDependents	float64

Figure 1.1 Features and Corresponding Data Types

Outputs

#	Column	Data Type
0	id	int64
1	probability	float64

Figure 1.2 Output Features and Corresponding Data Types

Exploratory Data Analysis

Pre-Processing

Dropping NaNs:	29,000 observations with >= 1 feature having NaN values, 20% of the dataset
Checking for Outliers:	Extreme outliers are 3x the interquartile range, these outliers are right-tail skewed
Checking for Multicollinearity:	30, 60, and 90 days past due are highly correlated with each other
Handling Duplicates:	There is not a large number of duplicates, so it will not bias the model significantly

Implementation and Validation

Baseline Models

Classifier	AUC	Standard Deviation	Time
GaussianNB	0.85	0.01	4 seconds
KNeighborsClassifier	0.74	0.00	25 seconds
LogisticRegression	0.85	0.01	5 seconds
RandomForestClassifier	0.84	0.01	46 seconds
XGBClassifier	0.85	0.00	17 seconds

Figure 1.3 Classifier Performance

Five different classification models are evaluated:

- GaussianNB
- KNeighborsClassfication
- LogisticRegression
- RandomForestClassifier
- XGBClassifier

XGBoost algorithm had the best performance in with the highest AUC and the lowest standard deviation.

Implementation and Validation

Hyperparameter Tuning — The model hyperparameters were tuned with randomized search

Model Evaluation — The XGBoost AUC ROC for the test results was 0.86083

Model Explainability — The output of the model is explained with SHAP

Fairness Evaluation — Analyzed the model's performance with the binary feature of SeriousDlqin2yrs

#	features	feature_importance
0	RevolvingUtilizationOfUnsecuredLines	0.260455
2	NumberOfTime30-59DaysPastDueNotWorse	0.208461
6	NumberOfTimes90DaysLate	0.194711

Figure 1.4 Top 3 Feature Performance

True Negative: 30423	False Negative: 11597
False Positive: 468	True Positive: 2512

Figure 1.5 Confusion Matrix

Outcomes

Fairness Metrics

Metric	Score
Accuracy	0.73188889
AUC	0.86083
Precision	0.213574327
Recall	0.777181208
Average Precision	0.1807415089

Figure 1.6 Fairness Metrics

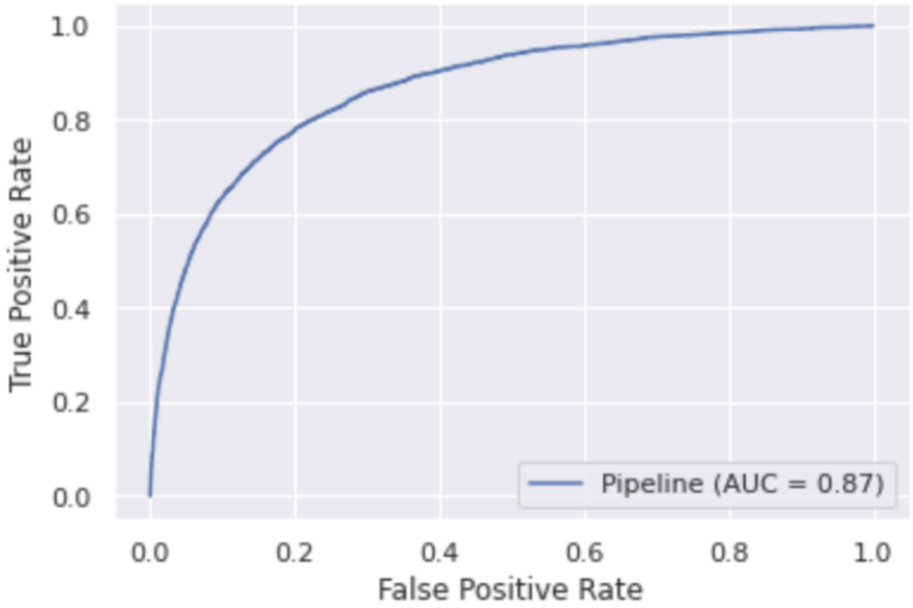


Figure 1.7 AUC-ROC Curve

Summary

Ingredients

Feature	Data Type
SeriousDlqin2yrs	int64
RevolvingUtilizationOfUnsecuredLines	float64
age	int64
NumberOfTime30-59DaysPastDueNotWorse	int64
DebtRatio	float64
MonthlyIncome	int64
NumberOfOpenCreditLinesAndLoans	int64
NumberOfTimes90DaysLate	int64
NumberRealEstateLoansOrLines	int64
NumberOfTime60-89DaysPastDueNotWorse	int64
NumberOfDependents	float64

Utensils

Model	AUC	STD
XGBClassifier	0.85	0.00
GaussianNB	0.85	0.01
LogisticsRegression	0.85	0.01
RFClassifier	0.74	0.01
KNeighborsClassifier	0.74	0.00

Product

Top 5 Features	Importance
RevolvingUtilizationOfUnsecuredLines	0.260
NumberOfTime30-59DaysPastDueNotWorse	0.208
NumberOfTimes90DaysLate	0.195
NumberOfTime60-89DaysPastDueNotWorse	0.123
NumberRealEstateLoansOrLines	0.064

Fairness Measures and Disparate Impact

Metrics	Score
Accuracy	0.795733333
AUC	0.787115116
Precision	0.213574327
Recall	0.777181208
Average Precision	0.213574327

Outcomes	
Median Age	52.405
Median Probability	0.2883
Undeprivileged Ratio (Age <25)	0.7274
Privileged Ratio	0.4953
Disparate Impact	1.4685