Machine Learning for Finance

Supervised Learning: Linear and Logistic Regression

Alireza Nouri - Mahdi Mastani

Sharif University of Technology

May 28 , 2024

Outline of talk

1. Linear Regression

2. Logistic Regression

Linear Regression

Linear Regression

- Linear regression is an important tool in machine learning, often used first by analysts in supervised learning.
- Used for predicting the value of a target from one or more features.
- Minimizes the mean squared error (MSE).
- Linear regression can handle both categorical and numerical features effectively.
- Assume that relationship between the target and the features are linear.

- Model: $Y = a + bX + \epsilon$
- Objective: Minimize MSE for the observations in the training set.
- Best fit values of a and b are those that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - a - bX_i)^2$$

Linear Regression: Multiple Features

- Extends single feature regression to multiple features.
- Model: $Y = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p + \epsilon$
- Requires solving for $p + 1$ coefficients (including the intercept).
- In machine learning, the parameter a is referred to as the **bias** and the coefficients b_j are referred to as the **weights**

Categorical Features

- Two types of categorical features:
  1. **Ordinal**
  2. **Nominal**

Handling Ordinal Features

- Ordinal features have a natural order (e.g., low, medium, high).
- Can be encoded using integer values reflecting their order.
- Example: Education level (High School $= 1$, Bachelor's $= 2$, Master's $= 3$).
- Ensure that the encoding captures the order relationship.

Handling Nominal Features

- Nominal features have no intrinsic order (e.g., red, green, blue).
- Handled by creating dummy variables (one-hot encoding).
- Each category is represented by a binary variable.
- One category can be dropped (e.g., drop the blue dummy variable).

Example: Categorical Features

- Consider a feature "City" with categories "New York", "San Francisco", and "Chicago".
- Create two dummy variables: $D1$ for "New York" and $D2$ for "San Francisco".
- Model:
  $\text{Price} = a + b_1 \times \text{Size} + b_2 \times \text{Bedrooms} + c_1 \times D1 + c_2 \times D2 + \epsilon$

Credit Card Balance Data

- Income: Income in \$1,000
- Limit: Credit limit
- Rating: Credit rating
- Cards: Number of credit cards
- Age: Age in years
- Education: Education in years
- Own: A factor with levels No and Yes indicating whether the individual owns a home
- Student: A factor with levels No and Yes indicating whether the individual is a student
- Married: A factor with levels No and Yes indicating whether the individual is married
- Region: A factor with levels East, South, and West indicating the individual's geographical location
- Balance: Average credit card balance in \$

Data Table

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
0	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
1	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
2	104.593	7075	514	4	71	11	Male	No	No	Asian	580
3	148.924	9504	681	3	36	11	Female	No	No	Asian	964
4	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
...	...	...	...	...	...	...	...	...	...	...	...
395	12.096	4100	307	3	32	13	Male	No	Yes	Caucasian	560
396	13.364	3838	296	5	65	17	Male	No	No	African American	480
397	57.872	4171	321	5	67	12	Female	No	Yes	Caucasian	138
398	37.728	2525	192	1	44	13	Male	No	Yes	Caucasian	0
399	18.701	5524	415	5	64	7	Female	No	No	Asian	966

400 rows × 11 columns

Multivariate Linear Regression Summary

OLS Regression Results

Dep. Variable:	Balance	R-squared:	0.955
Model:	OLS	Adj. R-squared:	0.954
Method:	Least Squares	F-statistic:	658.5
Date:	Sun, 26 May 2024	Prob (F-statistic):	9.50e-202
Time:	20:46:25	Log-Likelihood:	-1926.8
No. Observations:	320	AIC:	3876.
Df Residuals:	309	BIC:	3917.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-472.8236	41.035	-11.522	0.000	-553.568	-392.079
Income	-7.5486	0.277	-27.270	0.000	-8.093	-7.004
Limit	0.1998	0.037	5.380	0.000	0.127	0.273
Rating	0.9443	0.558	1.693	0.092	-0.153	2.042
Cards	19.2659	5.014	3.843	0.000	9.401	29.131
Age	-0.6210	0.343	-1.809	0.071	-1.296	0.054
Education	-1.1159	1.819	-0.613	0.540	-4.695	2.463
Gender	12.6153	11.410	1.106	0.270	-9.837	35.067
Student	419.1412	19.697	21.280	0.000	380.385	457.898
Married	-5.5100	11.836	-0.466	0.642	-28.800	17.780
Ethnicity	5.0384	6.898	0.730	0.466	-8.535	18.612

Omnibus:	18.537	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.759
Skew:	0.620	Prob(JB):	3.11e-05
Kurtosis:	2.861	Cond. No.	3.85e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.85e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Train and Test Evaluation

	MAE	MSE	RMSE	R2_squared
Multi_Linear_Regression_Train	0.040445	0.002488	0.049876	0.955179
Multi_Linear_Regression_Test	0.037307	0.002014	0.044877	0.951831

Extending Features

- Feature engineering is the process of creating new features from existing ones to improve model performance.
- Techniques include:
  1. Interaction features
  2. Polynomial features
  3. Domain-specific transformations

Interaction Features

- Interaction features capture the effect of one feature on another.
- Example: Interaction between age and income.
- Formula: $\text{Interaction} = \text{Age} \times \text{Income}$
- Created by multiplying or combining two or more features.
- Can help model complex relationships.

Polynomial Features

- Polynomial features are created by raising existing features to a power.
- Helps capture non-linear relationships.
- Example: If a feature is x, polynomial features include x^2, x^3, etc.
- Formula: $\text{Polynomial} = [\text{Income}, \text{Income}^2, \text{Income}^3]$
- Captures more complex patterns than linear relationships.
- Useful in polynomial regression.

Domain-Specific Transformations

- Features created based on domain knowledge.
- Example: Financial ratios in accounting (e.g., profit margin, return on equity).
- Original Features:Net Income , Net Sales Revenue
- Domain-Specific Feature: Profit Margin.
- Formula: Profit Margin $= \frac{\text{Net Income}}{\text{Net Sales Revenue}}$
- Helps incorporate expert insights into the model.

Polynomial Regression Summary

	MAE	MSE	RMSE	R2_squared
Polynomial Regression_2	1.956470e-02	5.846371e-04	2.417927e-02	0.989466
Polynomial Regression_3	3.669813e-03	2.674971e-05	5.172012e-03	0.999518
Polynomial Regression_4	4.728095e-15	4.302659e-29	6.559466e-15	1.000000
Polynomial Regression_5	8.130118e-15	1.279324e-28	1.131072e-14	1.000000

Train and Test Evaluation

	MAE	MSE	RMSE	R2_squared
Multi_poly_Regression_Train_4	4.728095e-15	4.302659e-29	6.559466e-15	1.000000
Multi_poly_Regression_Test_4	5.687804e-02	1.274558e-02	1.128963e-01	0.695156

Regularization

- Helps to prevent overfitting.
- Before using regularization, it is important to carry out **feature scaling**
- To ensure that the numerical values of features are comparable.

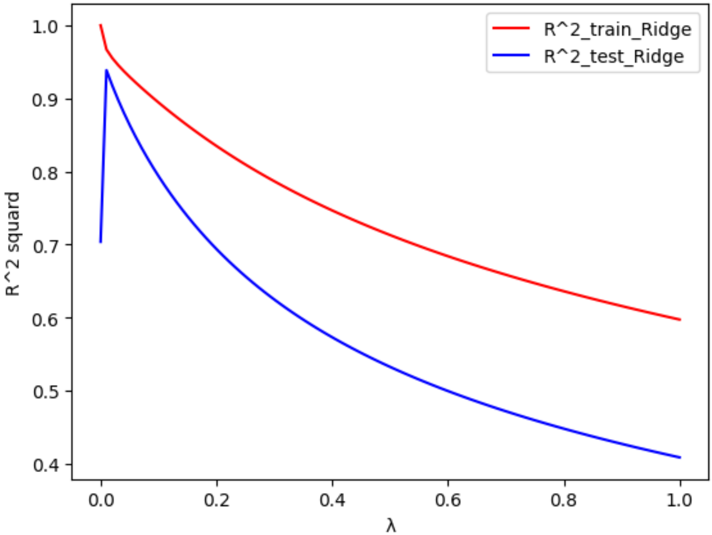
Ridge Regression

- Ridge regression is referred to as L2 regularization.
- Adds a penalty equal to the sum of the squared coefficients.
- Objective: Minimize

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{p} b_j^2$$

- Reduces the magnitude of the coefficients.
- Useful when features are highly correlated.
- The parameter λ is referred to as a hyperparameter.

Ridge regularization (L2)



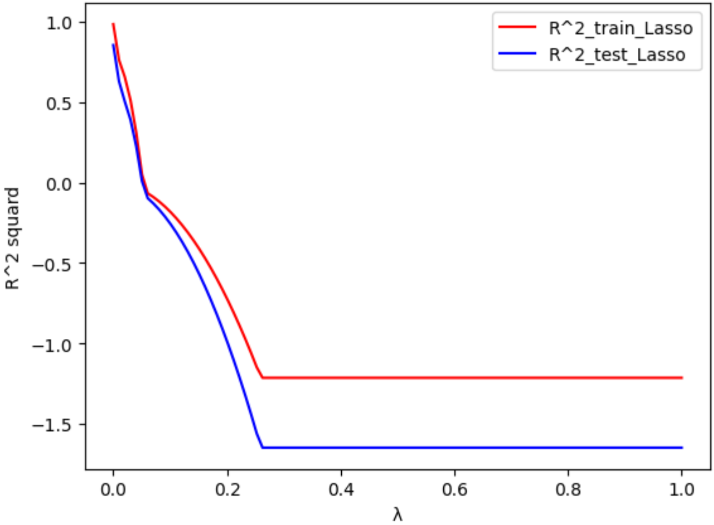
Lasso Regression

- Lasso regression is referred to as L1 regularization.
- Adds a penalty equal to the sum of the absolute values of the coefficients.
- Objective: Minimize

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda\sum_{j=1}^{p}|b_j|$$

- Can shrink some coefficients to zero, effectively selecting a simpler model.
- Useful for feature selection.

Lasso regularization (L1)



Elastic Net Regression

- Combines Ridge and Lasso penalties.
- Objective: Minimize

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda_1 \sum_{j=1}^{p} b_j^2 + \lambda_2 \sum_{j=1}^{p} |b_j|$$

- Balances between Ridge and Lasso properties.

Elastic Net regularization

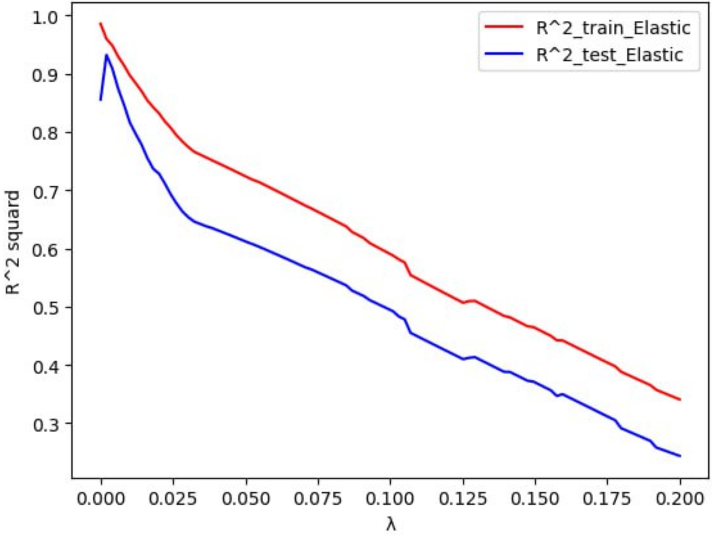


Figure: $\lambda(\alpha\|\beta\|_1^2 + ((1-\alpha)/2)\|\beta\|_2^2)), \alpha = 0.1$

The L2 norm of the coefficient

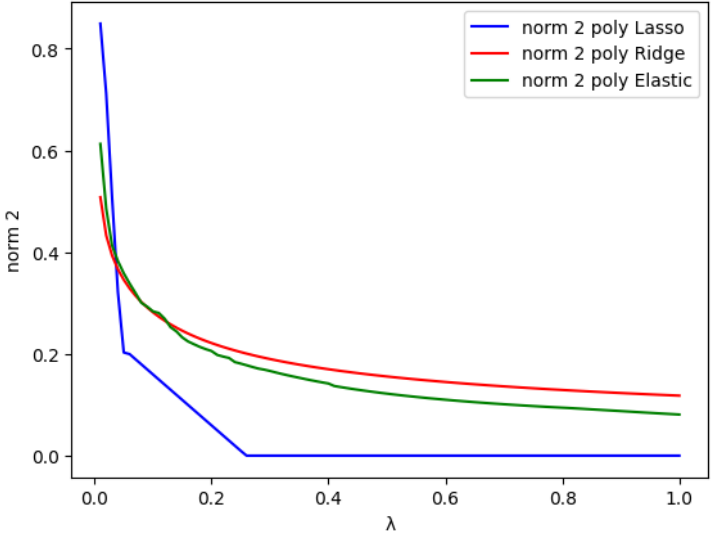


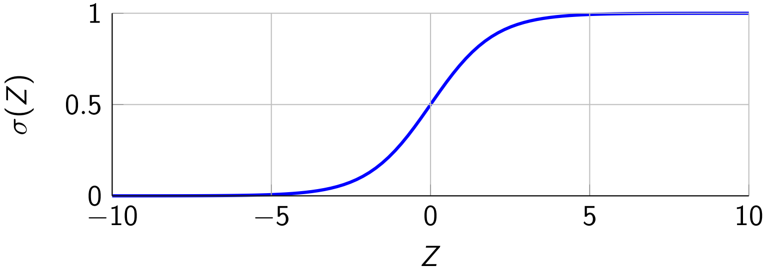
Figure: The L2 norm for the polynomial regression is 13.37

Logistic Regression

Logistic Regression

- Used for classification problems.
- Predicts the probability of an observation belonging to a category.
- $Z = a + b_1X_1 + \cdots + b_pX_p$
- Uses the sigmoid function:

$$P(Y = 1) = \frac{1}{1 + e^{-Z}}$$



Cross-Entropy Loss for Classification

The cross-entropy loss function for binary classification is given by:

$$L(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right)$$

Where:

- y is the true label (0 or 1) of the sample.
- $\hat{y}$ is the predicted probability of the sample belonging to class 1.

Decision Criteria

- Define a threshold probability t.
- If $P(Y = 1) > t$, classify as positive.
- If $P(Y = 0) \leq t$, classify as negative.
- Trade-off

Confusion Matrix Components

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **TP (True Positive):** Correctly predicted positive cases.
- **TN (True Negative):** Correctly predicted negative cases.
- **FP (False Positive):** Incorrectly predicted as positive.
- **FN (False Negative):** Incorrectly predicted as negative.

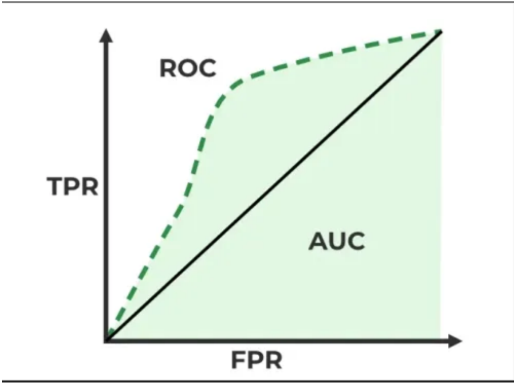
Model Evaluation

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- True Positive Rate(TPR): $\frac{TP}{TP+FN}$
- False Positive Rate (FPR): $\frac{\text{FP}}{\text{FP+TN}}$
- Precision: $\frac{\text{TP}}{\text{FP+TP}}$
- F1-score: $2x\frac{\text{Precision.TPR}}{\text{TPR+Precision}}$

ROC Curve

- Plots the true positive rate against the false positive rate.
- Area Under the Curve (AUC) measures the model's performance.
- AUC ranges from 0.5 (random) to 1 (perfect classification).



Example

The S&P 500 stock index between 2017 and 2024

- Year: The year that the observation was recorded
- Lag1: Percentage return for previous day
- Lag2: Percentage return for 2 days previous
- Lag3: Percentage return for 3 days previous
- Lag4: Percentage return for 4 days previous
- Lag5: Percentage return for 5 days previous
- Volume: Volume of shares traded for 1 days previous (number of daily shares traded in billions)
- Direction: A factor with levels 'Down' and 'Up' indicating whether the market had a positive or negative return on a given day

Data Table

	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
0	-0.206290	0.257143	0.036155	0.404576	0.278029	3.64056	-0.036131	Down
1	-0.036131	-0.206290	0.257143	0.036155	0.404576	3.62508	0.296217	Up
2	0.296217	-0.036131	-0.206290	0.257143	0.036155	3.46622	-0.030819	Down
3	-0.030819	0.296217	-0.036131	-0.206290	0.257143	3.09068	0.083595	Up
4	0.083595	-0.030819	0.296217	-0.036131	-0.206290	3.58695	-0.055087	Down
...	...	...	...	...	...	...	...	...
1855	-1.356112	-0.230671	0.428789	0.052396	0.003204	3.00551	0.440599	Up
1856	0.440599	-1.356112	-0.230671	0.428789	0.052396	3.75154	-0.185671	Down
1857	-0.185671	0.440599	-1.356112	-0.230671	0.428789	3.55275	-0.223156	Down
1858	-0.223156	-0.185671	0.440599	-1.356112	-0.230671	3.81875	-0.461808	Down
1859	-0.461808	-0.223156	-0.185671	0.440599	-1.356112	5.43716	0.654176	Up

1860 rows × 8 columns

Logistic Regression Summary

	coef	std err	z	P>|z|
const	1.6127	0.877	1.839	0.066
Lag1	-0.4008	0.648	-0.618	0.536
Lag2	0.0692	0.646	0.107	0.915
Lag3	-0.5883	0.659	-0.893	0.372
Lag4	-0.5198	0.649	-0.801	0.423
Lag5	-0.5108	0.659	-0.775	0.439
Volume	-1.3464	0.458	-2.940	0.003

Train and Test Evaluation

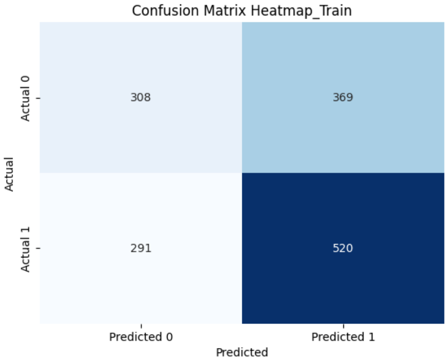


Figure: accuracy=0.56

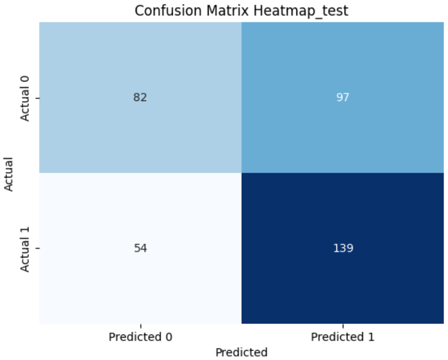
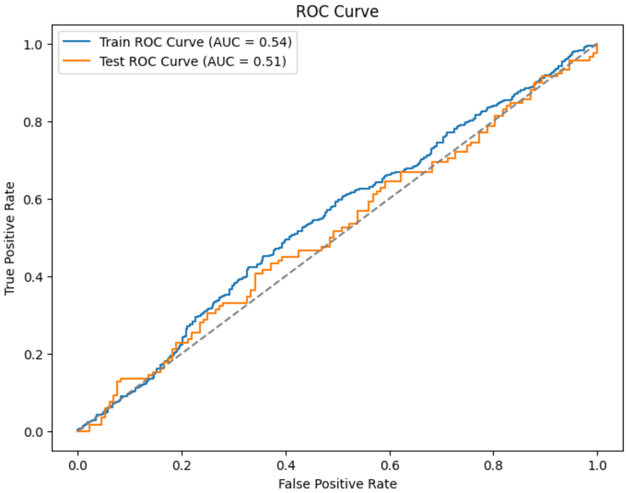


Figure: accuracy=0.59

ROC curve



ROC Curve
Train ROC Curve (AUC = 0.54)
Test ROC Curve (AUC = 0.51)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Features

Default of Credit Card Clients

- LIMIT_BAL
- SEX
- EDUCATION
- MARRIAGE
- AGE
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5
- PAY_6
- BILL_AMT1
- BILL_AMT2
- BILL_AMT3
- BILL_AMT4
- BILL_AMT5
- BILL_AMT6
- PAY_AMT1
- PAY_AMT2
- PAY_AMT3
- PAY_AMT4
- PAY_AMT5
- PAY_AMT6
- default

Imbalance Data

	YES(1)	No(0)
Number	6636	23364

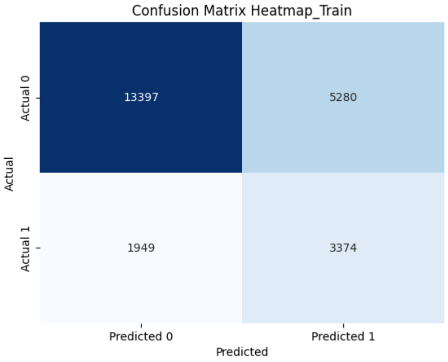
$$w_i = \frac{n}{k \cdot n_i} \tag{1}$$

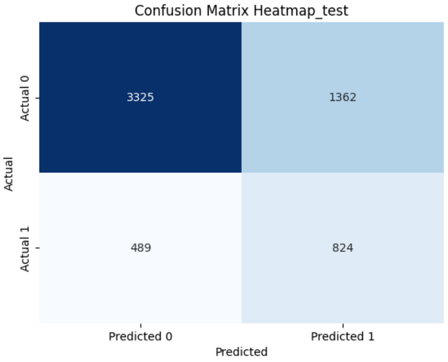
- w_i: Weight for class i
- n: Total number of samples
- k: Number of classes
- n_i: Number of samples in class i

$$L = -\frac{1}{N} \sum_{i=1}^{N} [w_0 \cdot y_i \log(\hat{y}_i) + w_1 \cdot (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

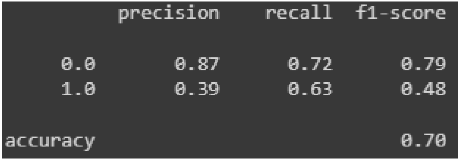
- N: Total number of samples
- y_i: True label of the i-th sample (0 or 1)
- $\hat{y}_i$: Predicted probability of the i-th sample being in class 1
- w_0: Weight for the class where $y_i = 1$
- w_1: Weight for the class where $y_i = 0$

Train and Test Evaluation



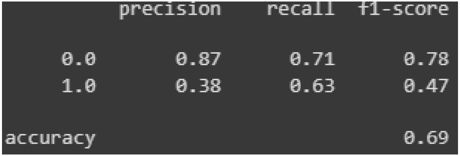


Train and Test Report



	precision	recall	f1-score
0.0	0.87	0.72	0.79
1.0	0.39	0.63	0.48
accuracy			0.70

Figure: Train



	precision	recall	f1-score
0.0	0.87	0.71	0.78
1.0	0.38	0.63	0.47
accuracy			0.69

Figure: Test

High false positive rates (FPR) can be dangerous

We can imagine that having a 10% FPR (90% TNR) can be considered perfect.

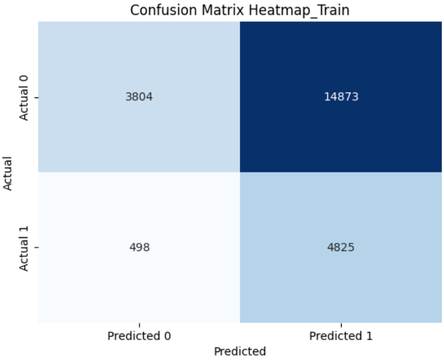


Figure: Train

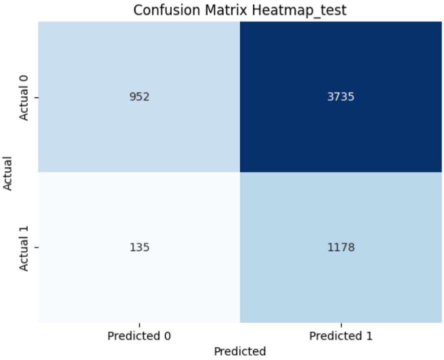


Figure: Test

Train and Test Report

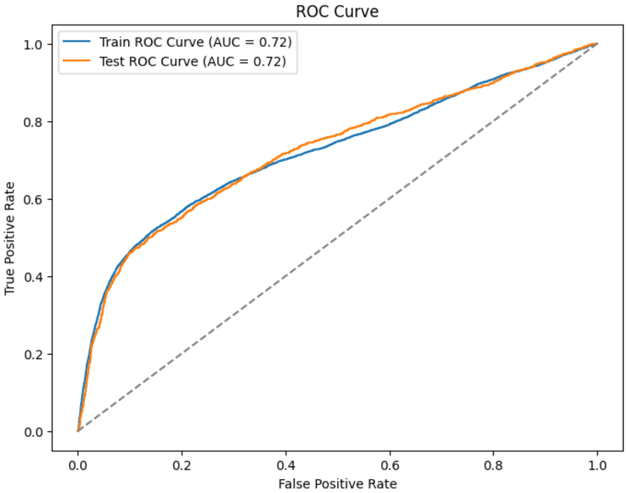
	precision	recall	f1-score
0.0	0.88	0.20	0.33
1.0	0.24	0.91	0.39
accuracy			0.36

Figure: Train

	precision	recall	f1-score
0.0	0.88	0.20	0.33
1.0	0.24	0.90	0.38
accuracy			0.36

Figure: Test

ROC curve



ROC Curve
Train ROC Curve (AUC = 0.72)
Test ROC Curve (AUC = 0.72)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Thank you for your attention

Any question?