Telco Customer Churn Analysis

Academic Year 2022-2023

Liu Chang

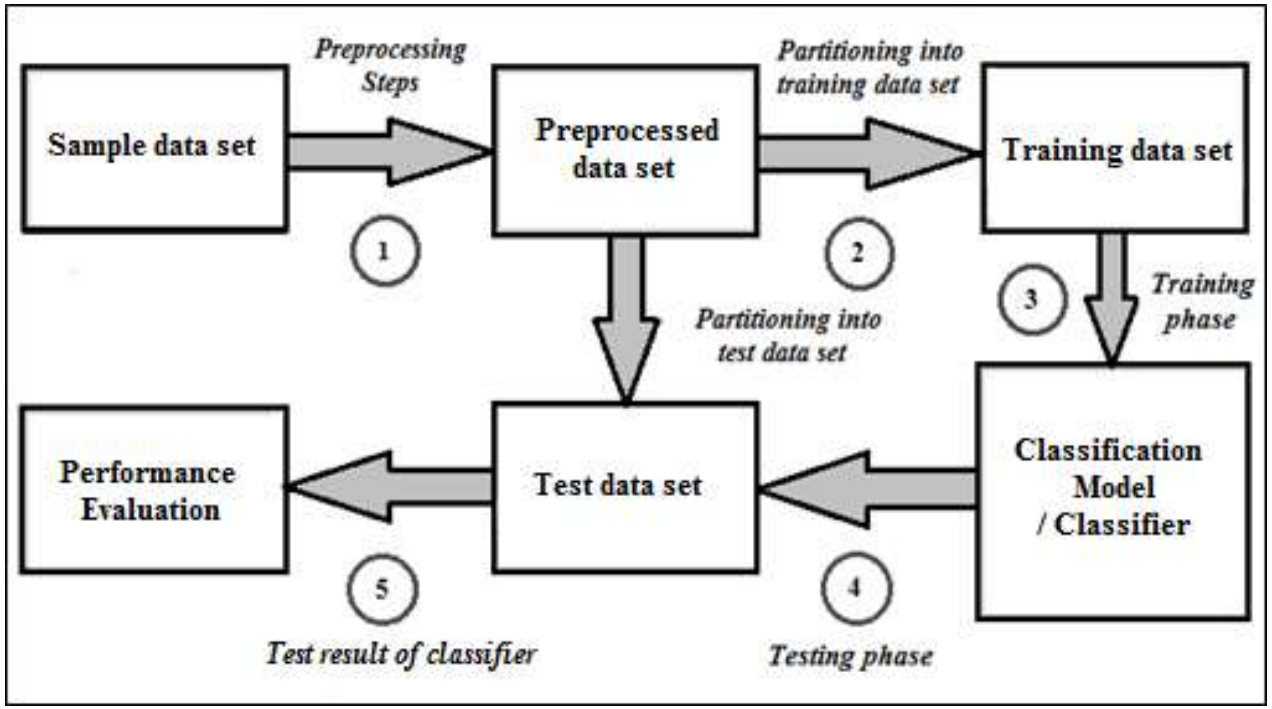
Introduction

Telco Customer Churn Predictor is a tool that allow the manager of a telecommunication company to prevent a possible churn from customers by letting the manager to know in advance and have a time to implement a market strategy to prevent the churn from customers

Dataset

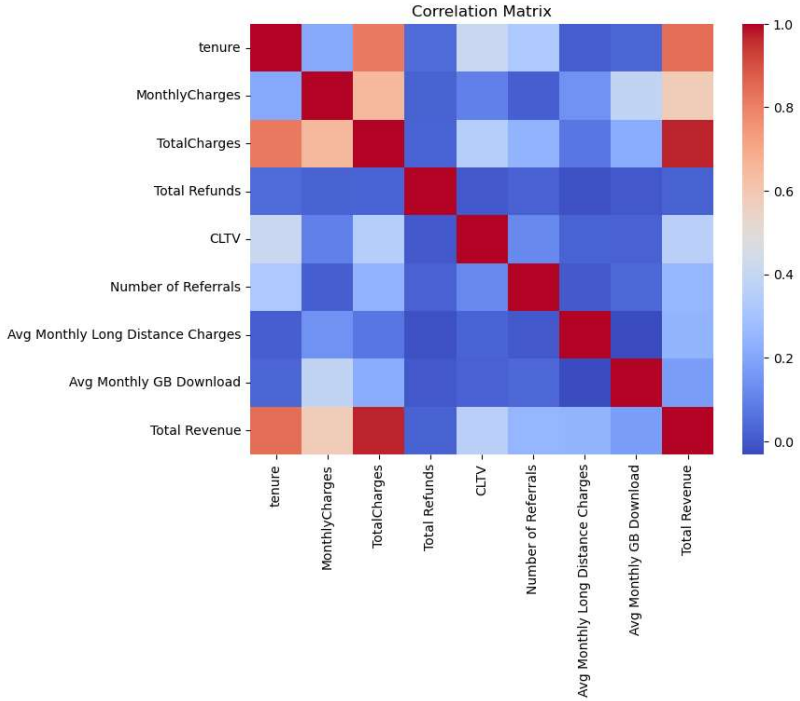
customerID	gender	SeniorCiti...	Partner	Dependents	tenure	PhoneSer...	MultipleLi...	InternetSe...	OnlineSec...	OnlineBac...
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No

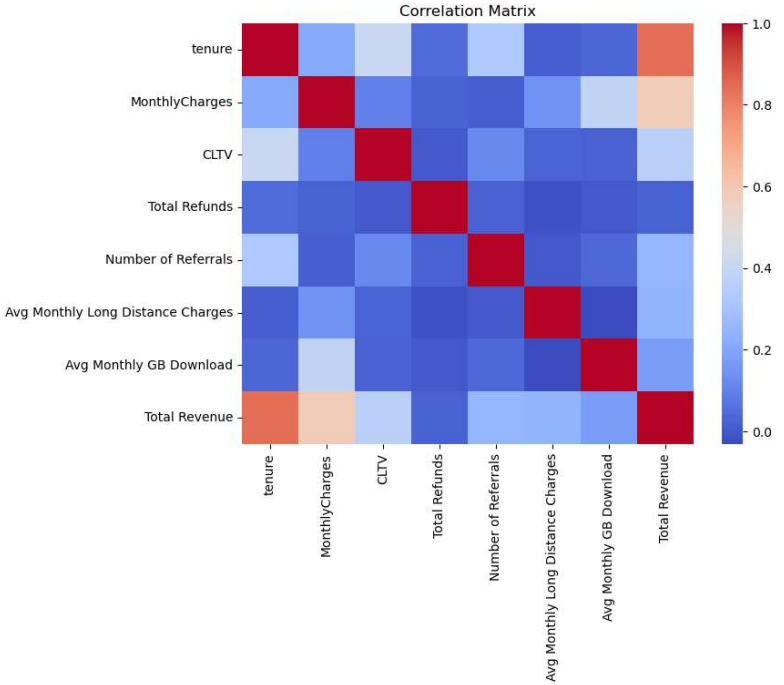
Analysis


Sample data set
Preprocessing Steps
Preprocessed data set
Partitioning into training data set
Training data set
1
2
3
Training phase
Partitioning into test data set
Performance Evaluation
Test data set
Classification Model / Classifier
5
4
Test result of classifier
Testing phase

Correlation matrix pre

Correlation matrix after





Pre-processing

- Categorical data : yes →1 , no →0, no phone service→0 , no internet service→0
- Partner Dependents PhoneService MultipleLines Online Security Online Backup StreamingTV StreamingMovie DeviceProtection Tech Support PaperlessBilling Streaming Music Under30 Maried Referred a Friend Unlimited Data and Premium Tech Support
- Drop by math long distance charge=Avg charge*tenure , Married and Partner by equality
- Drop feature that have equal value like count, or irrelevant like customerID and abitation (all things about it)
- Remove tuples where churn is caused by major forces(decease, moved) and joined now
- Map gender to 0 and 1
- Dummy cells for InternetService, Offer, PaymentMethod and Contract

X & Y Features

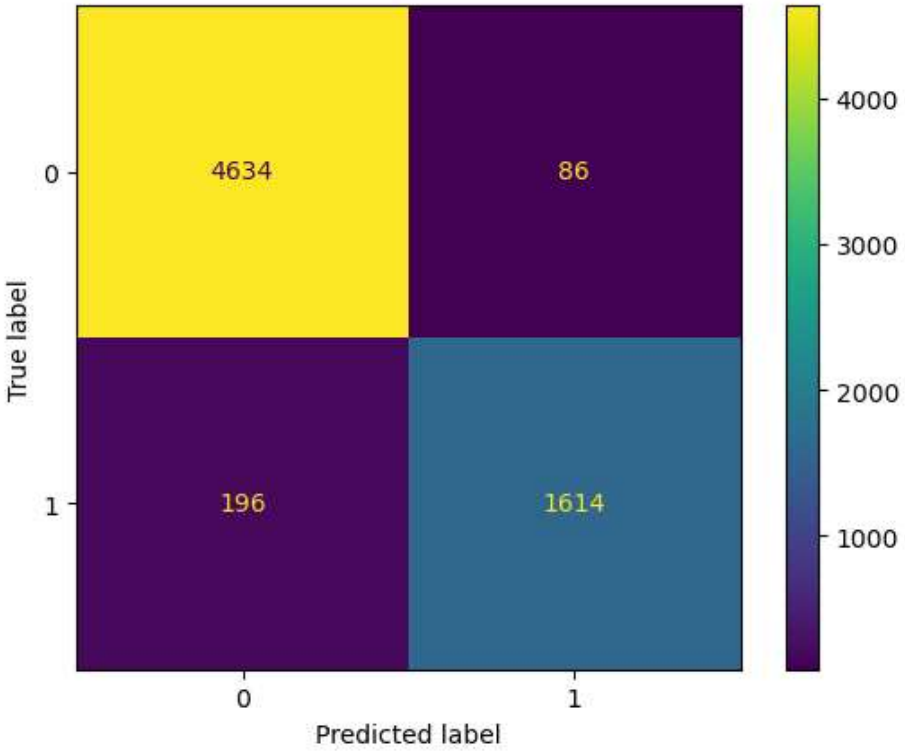
- X feature based on the gender of the person, the age, the type of services subscribed such as type of line and online services, or if they have tech support, the type of contract, payment method and if they requested an addition service like if they charged extra GB
- Y feature : churned or stayed

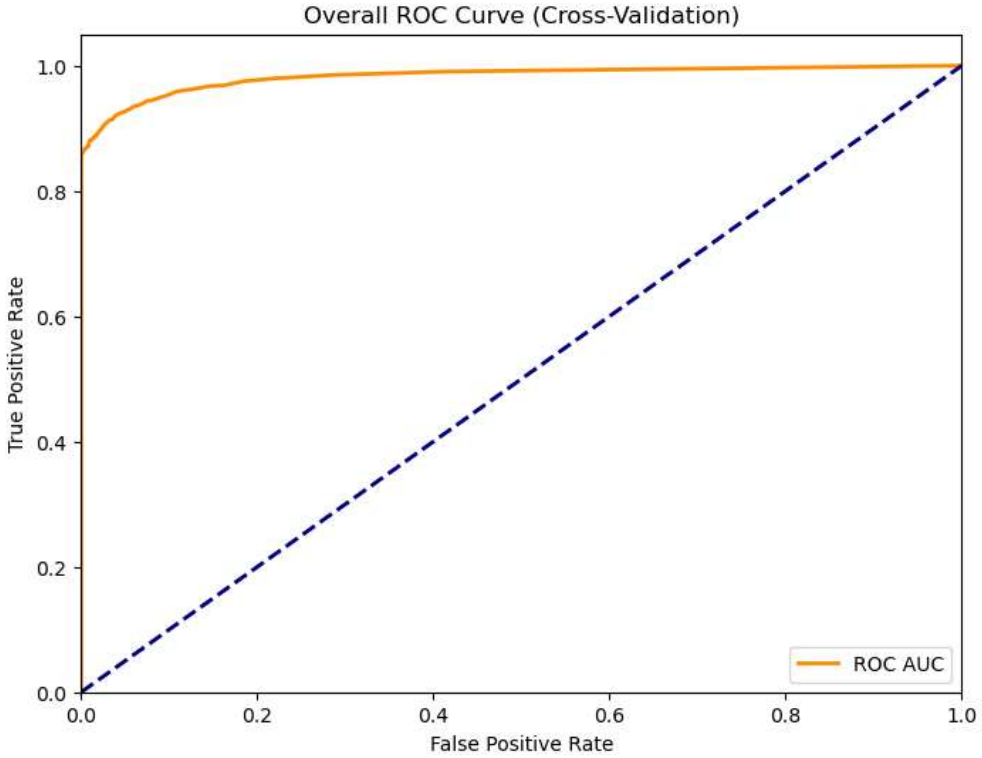
Classifiers

- Random Forest Classifier
- Naïve Bayes
- SVC
- K-Neighbors
- Ada Boost

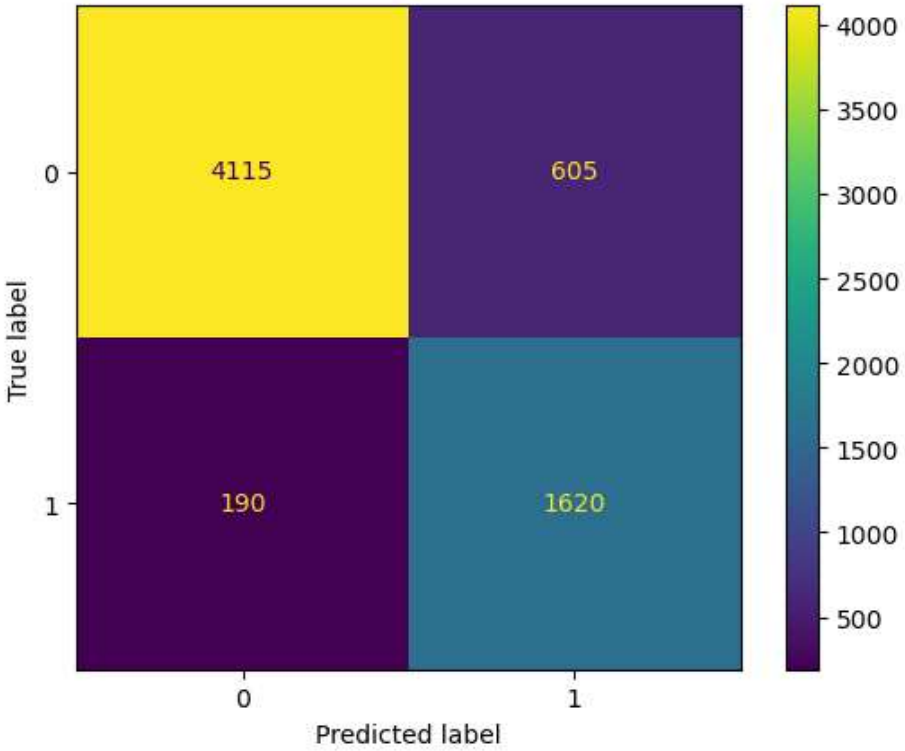
Pipeline: MinMaxScaler,SelectKBest → Classifier

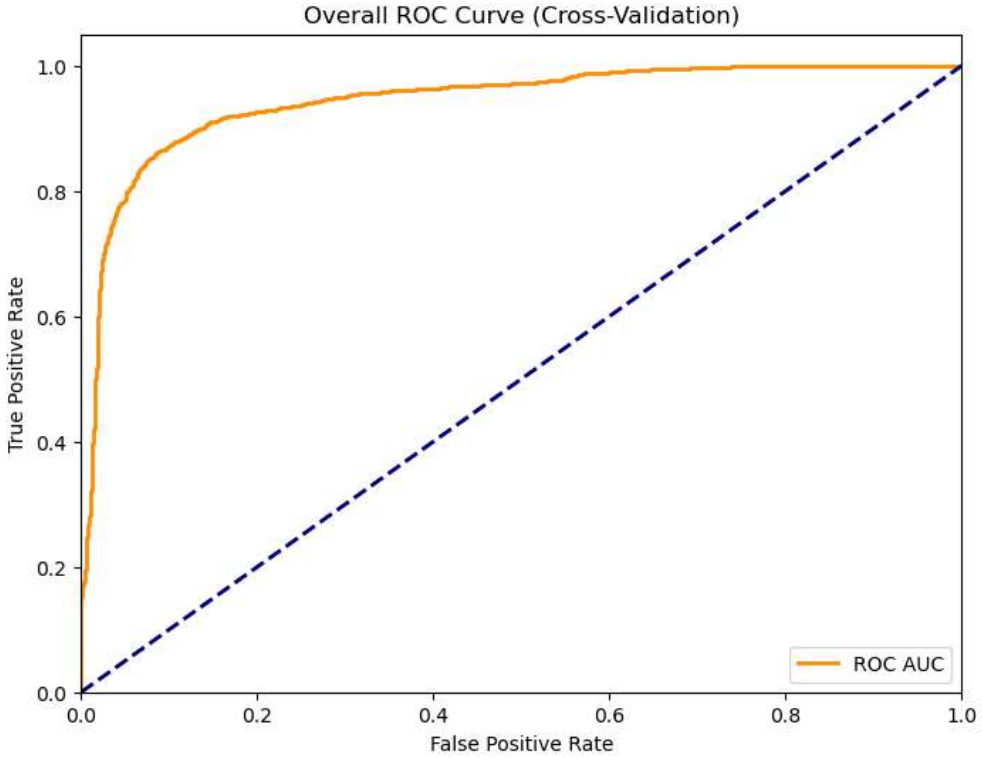
Random Forest



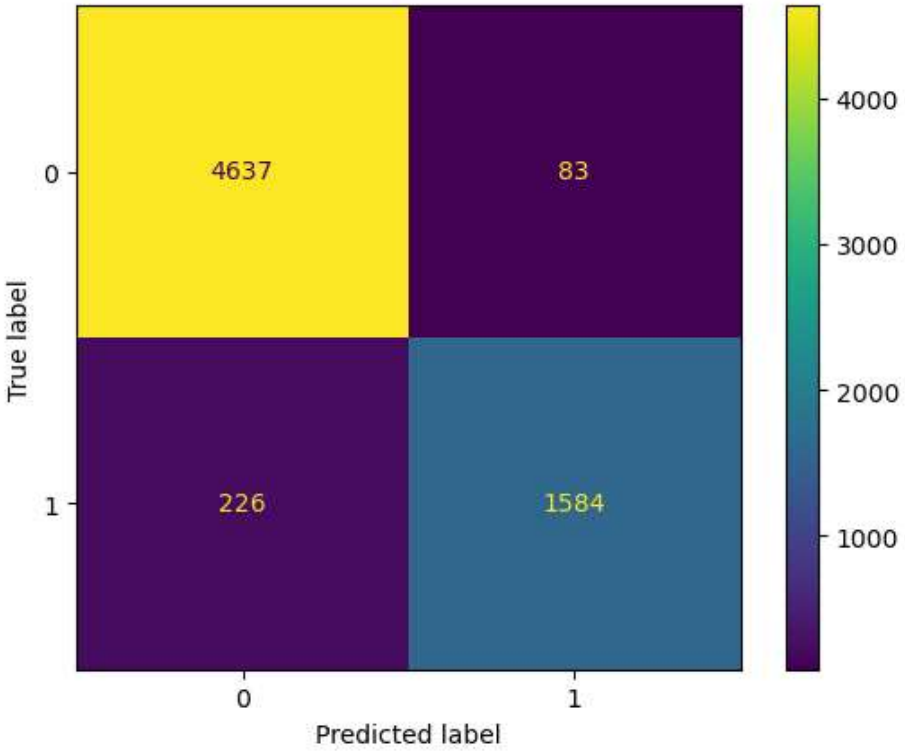


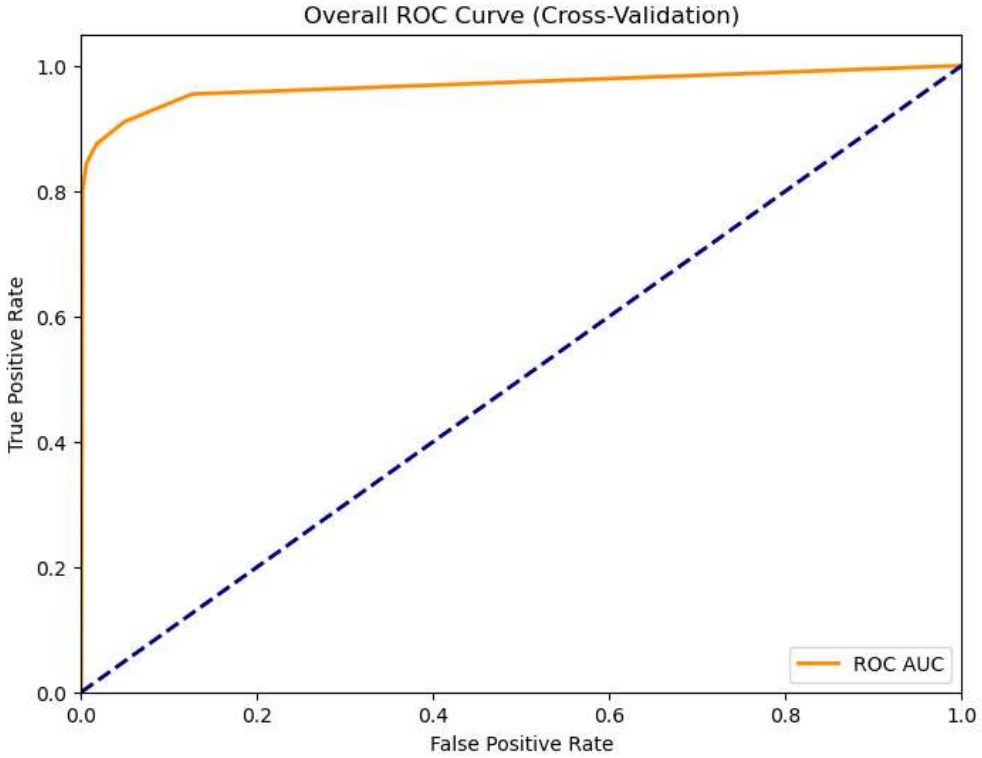
Naïve Bayes


4115
605
190
1620
0
1
True label
Predicted label
4000
3500
3000
2500
2000
1500
1000
500

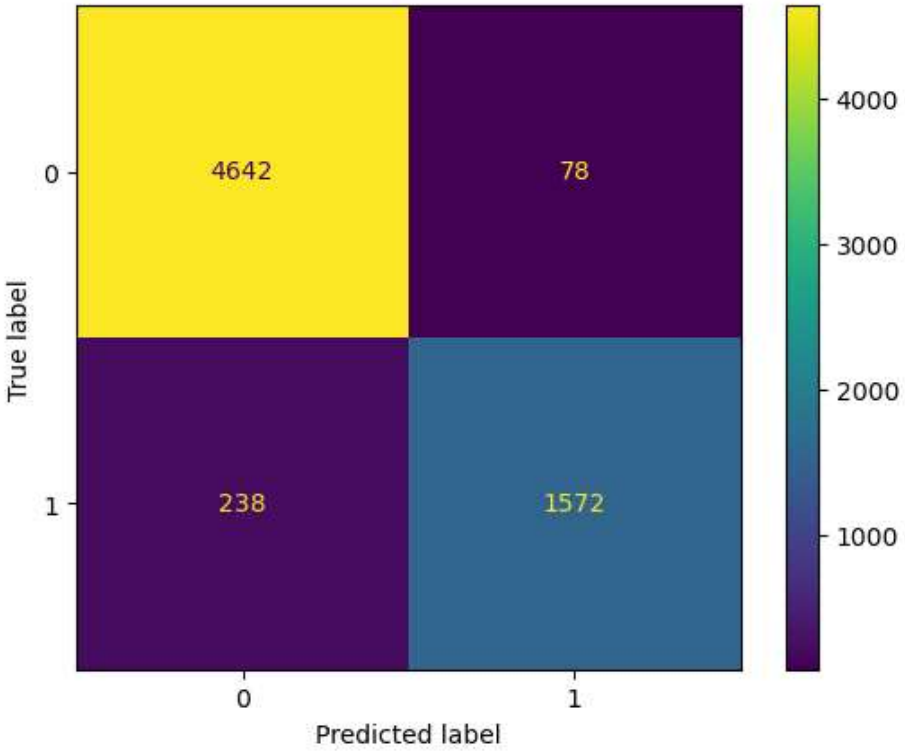

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

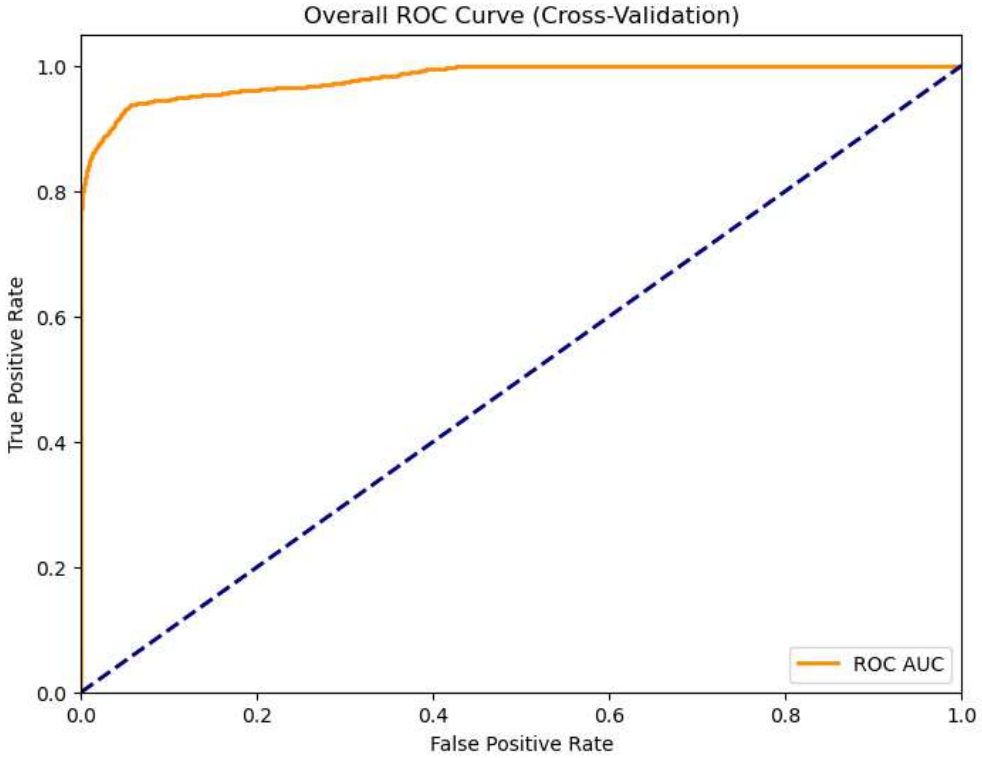
K-Neighbors


4637
83
226
1584
0
1
True label
Predicted label
4000
3000
2000
1000

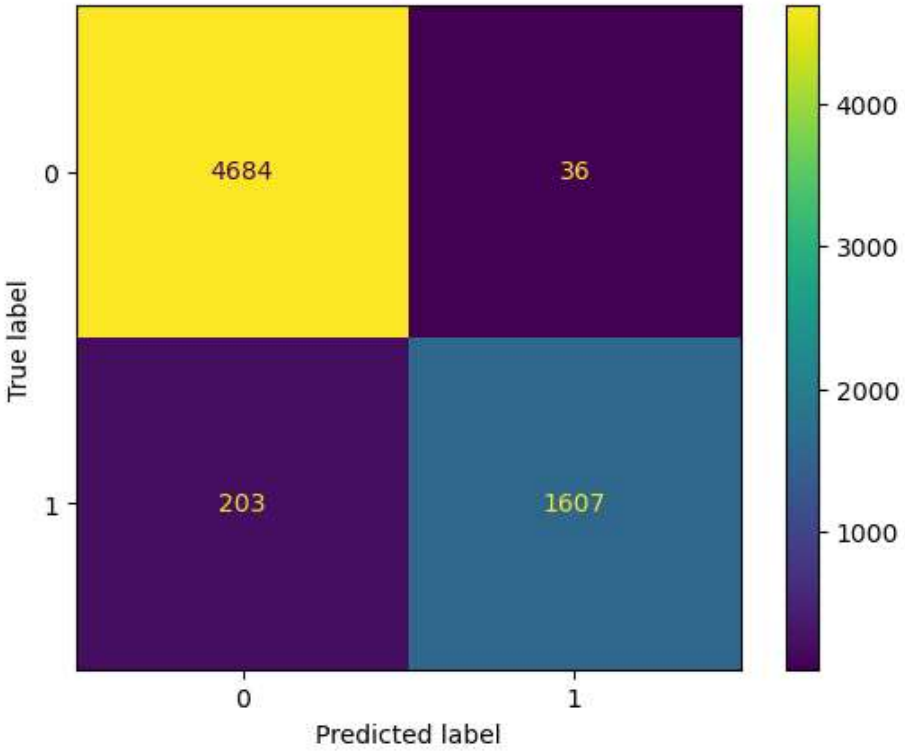

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
ROC AUC
0.0
0.2
0.4
0.6
0.8
1.0

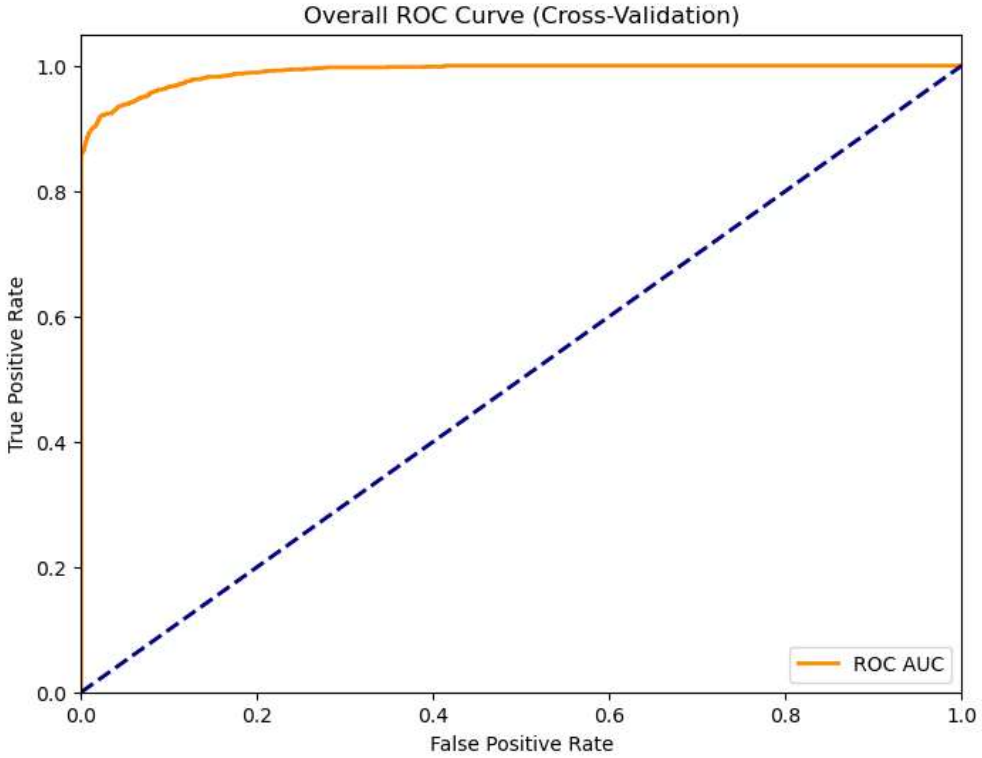
SVC


4642
78
238
1572
0
1
True label
Predicted label
4000
3000
2000
1000


Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
ROC AUC
0.0
0.2
0.4
0.6
0.8
1.0

Ada Boost


4684
36
203
1607
0
1
True label
0
1
Predicted label
4000
3000
2000
1000


Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

Result with 10-fold Cross validation

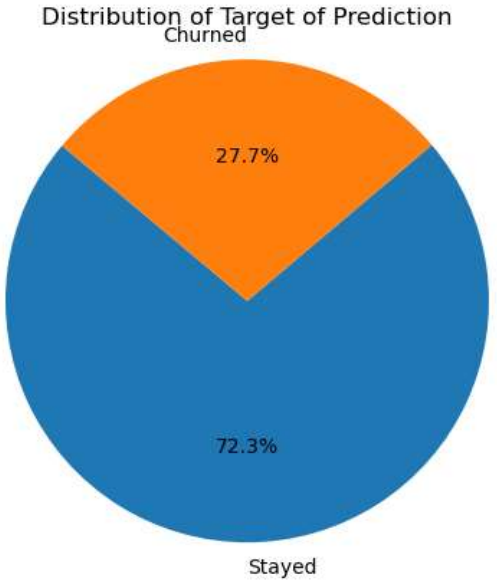
	Classifier	Accuracy Mean	Precision Mean	Recall Mean	F1 Score Mean	ROC AUC Mean
0	RandomForest	95.68	94.96	89.17	91.96	98.28
1	Naive Bayes	87.83	72.83	89.50	80.30	94.50
2	KNeighbors	95.27	95.03	87.51	91.10	96.83
3	SVM	95.16	95.30	86.85	90.86	98.18
4	AdaBoost	96.34	97.81	88.78	93.07	99.06

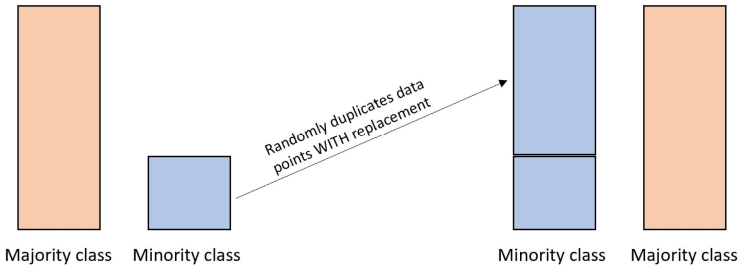
Wilcoxon Statical Test (Threshold=5% f1)

Classifier Comparison	P-value
AdaBoost-SVM	0.001953125
Random Forest-AdaBoost	0.010862224
Random Forest-SVC	0.048828125
AdaBoost-KNN	0.001953125
KNN-SVC	0.625
Random Forest-KNN	0.009765625

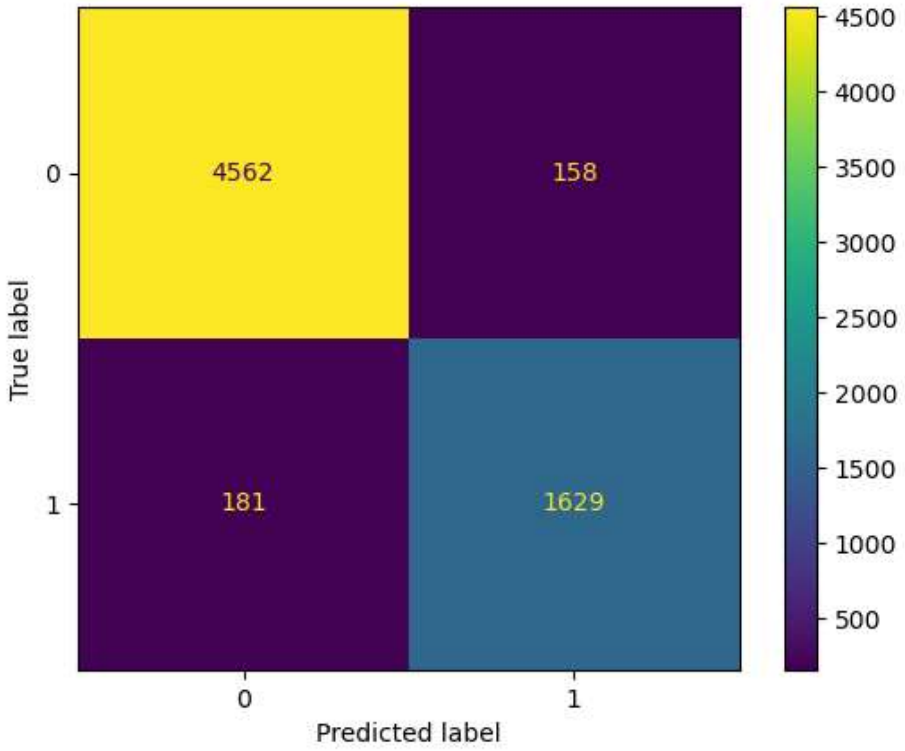
SMOTE Oversampling

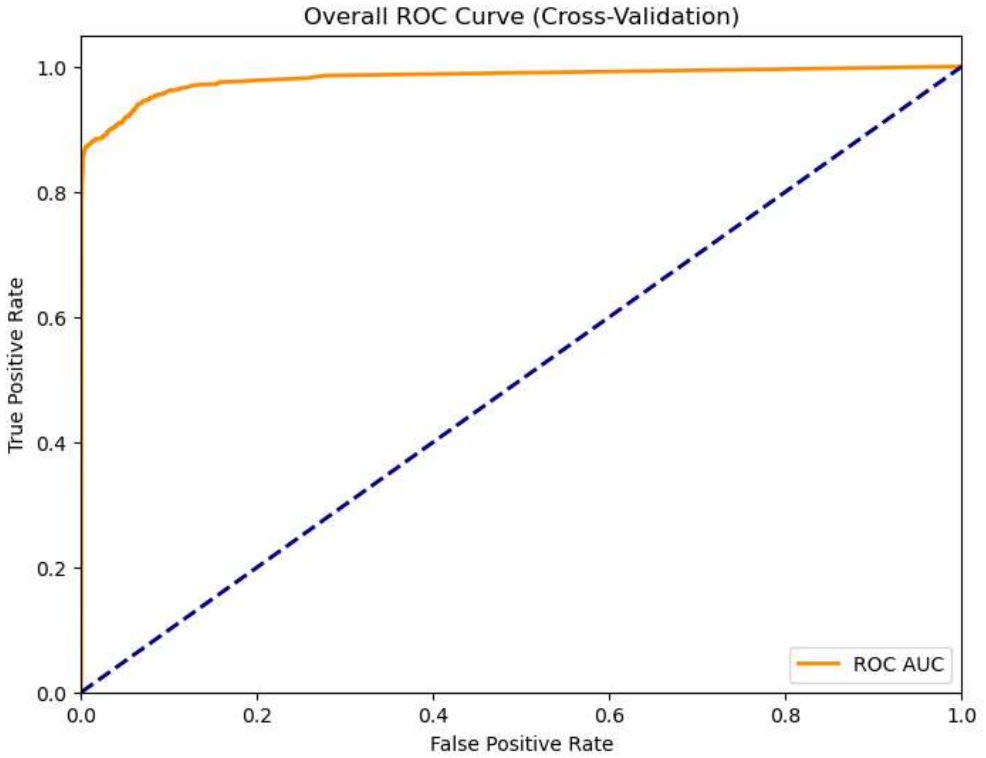
- Pipeline: SMOTE,MinMaxScaler,SelectKbest→Classifier



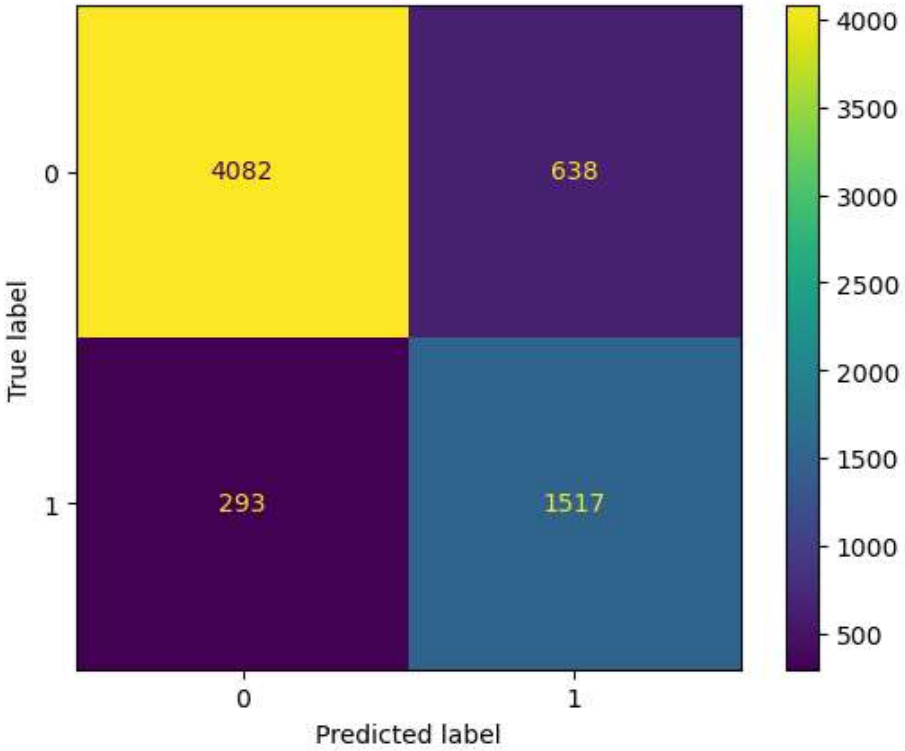


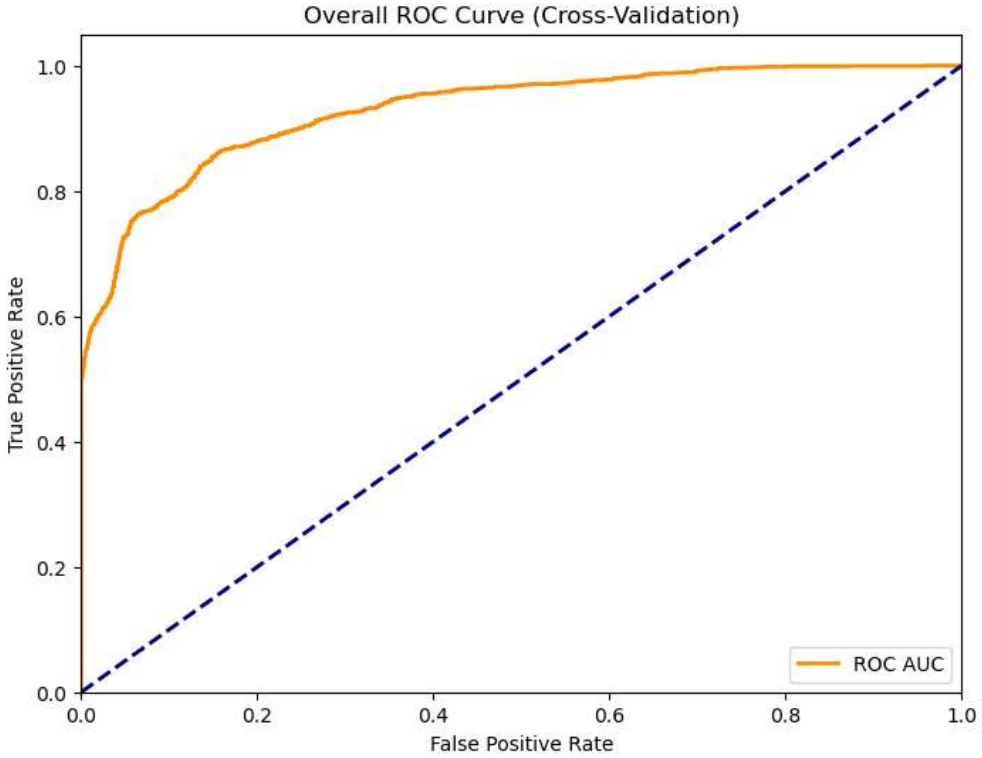
Random Forest



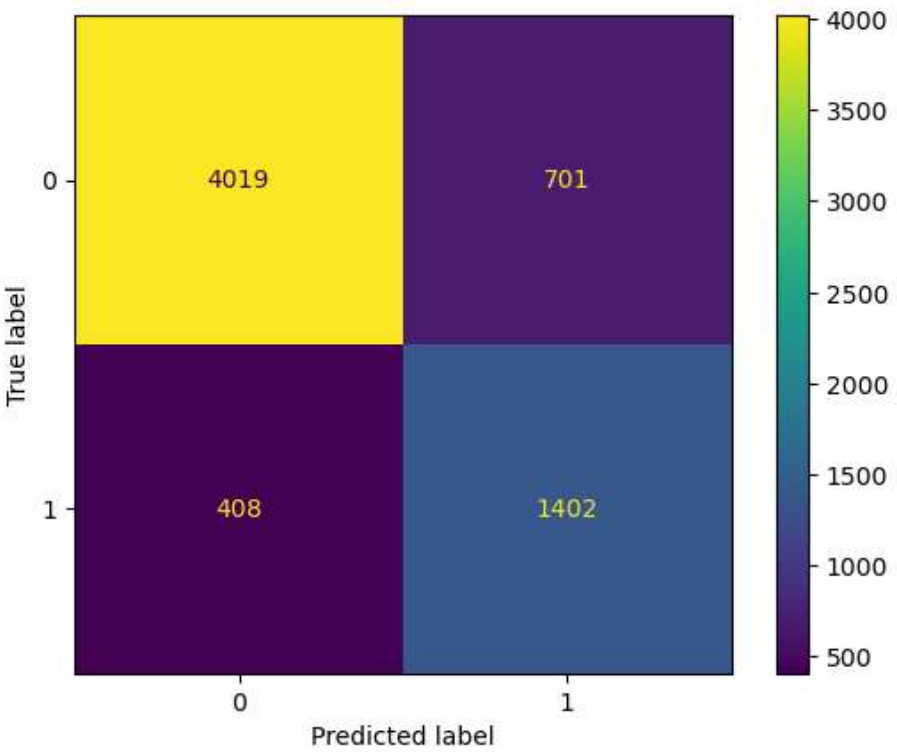


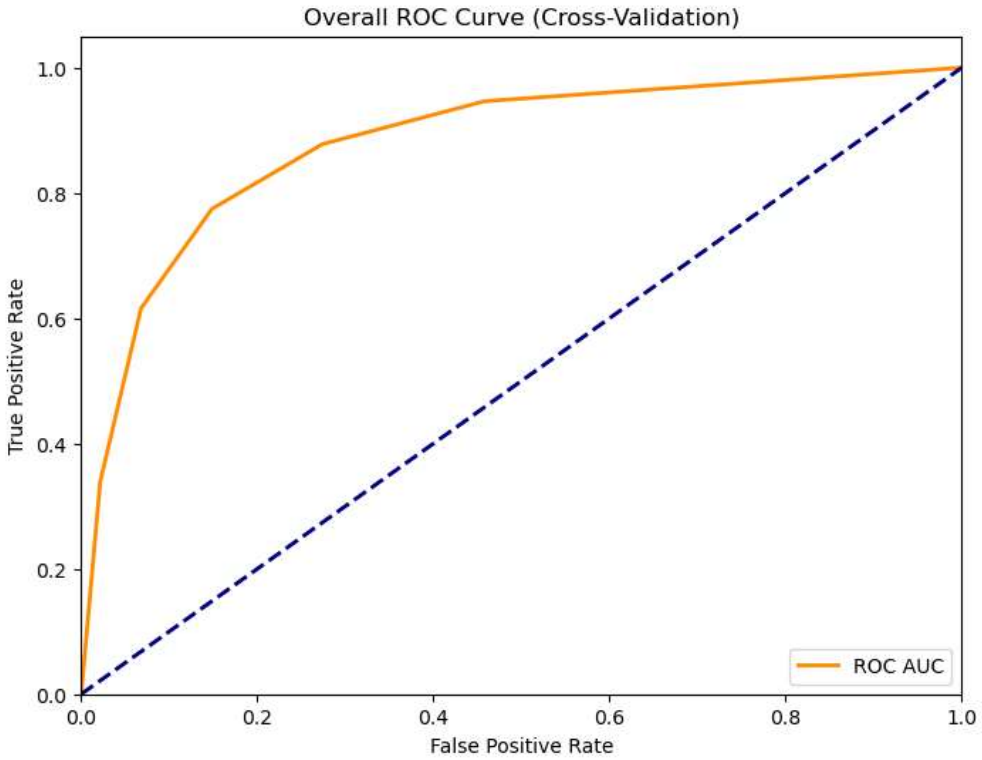
Naïve Bayes


4082
638
293
1517
0
1
True label
Predicted label
4000
3500
3000
2500
2000
1500
1000
500

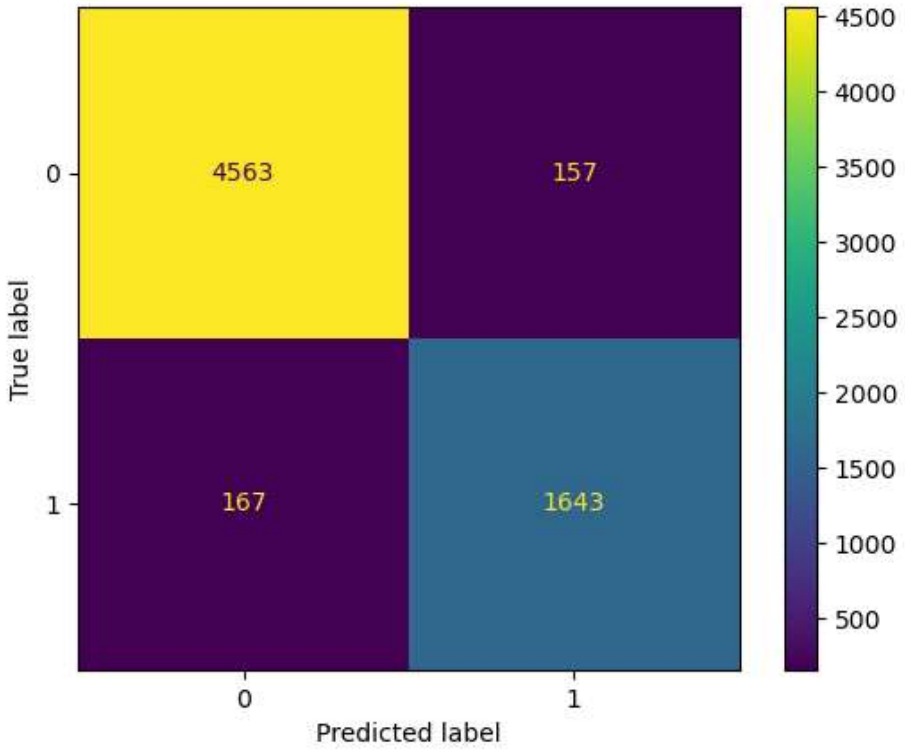

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

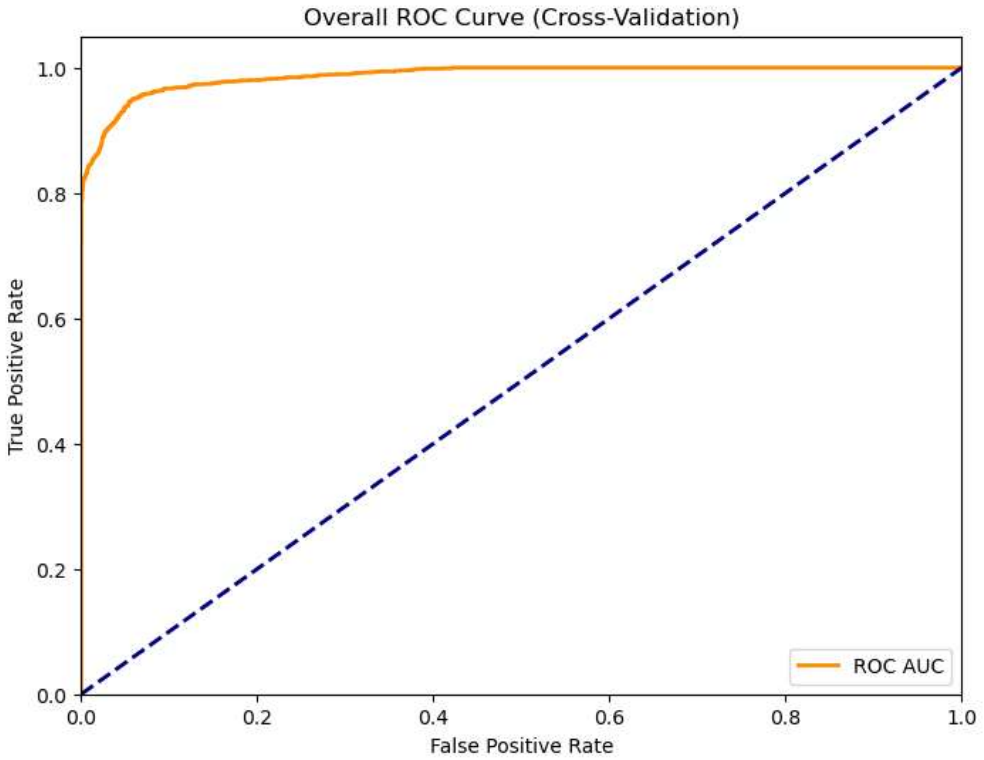
K-Neighbors


4019
701
408
1402
0
1
True label
Predicted label
4000
3500
3000
2500
2000
1500
1000
500

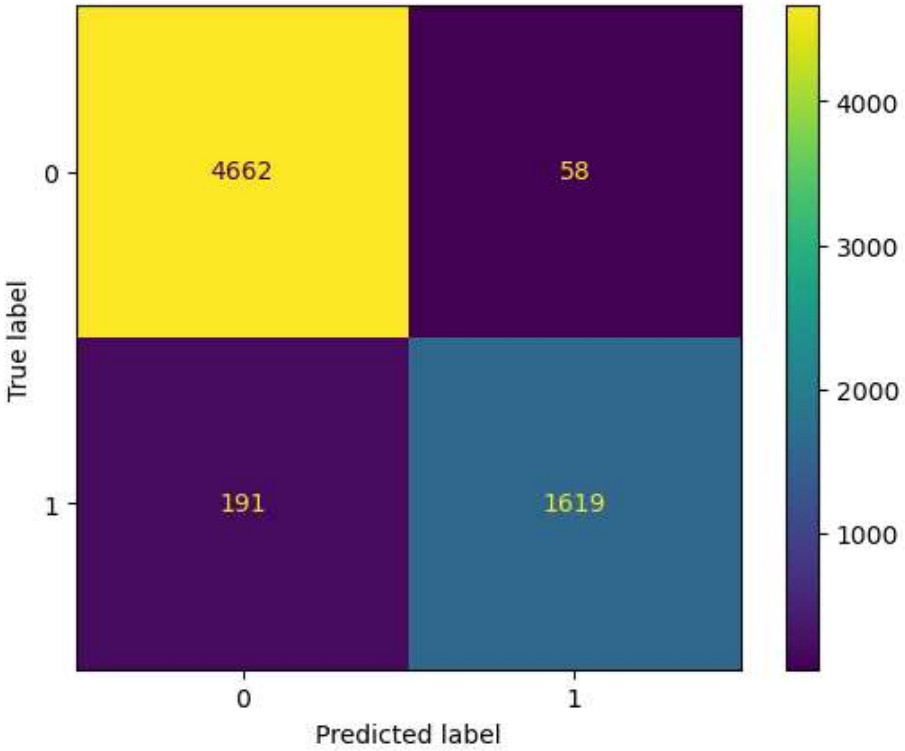

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

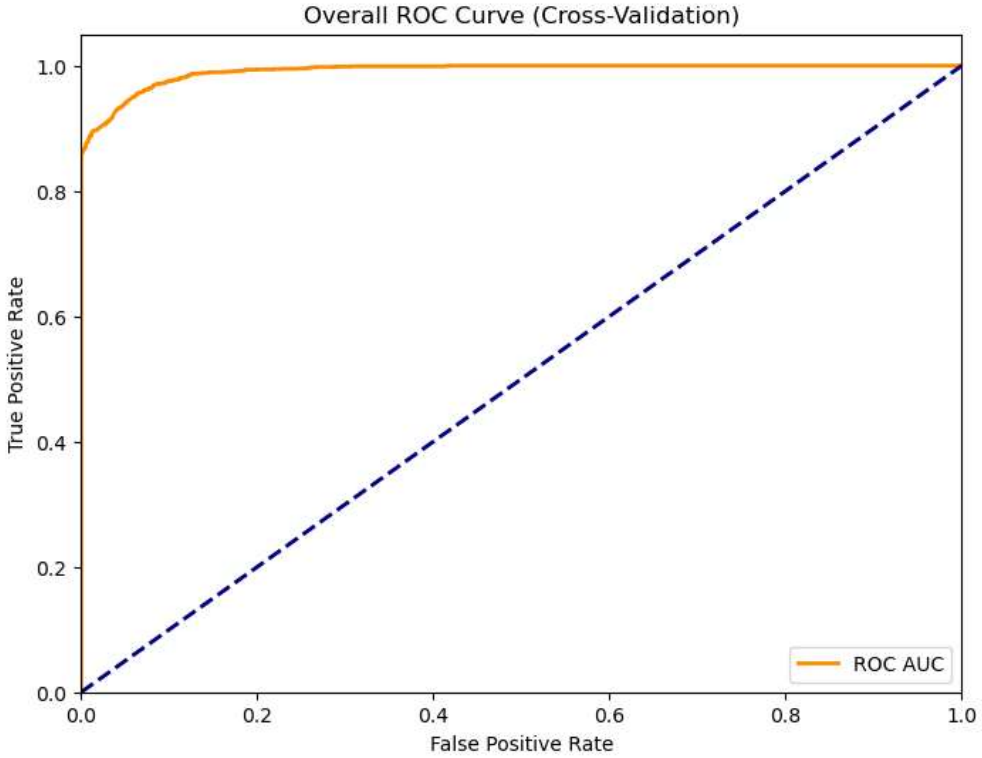
SVC


4563
157
167
1643
0
1
True label
Predicted label
4500
4000
3500
3000
2500
2000
1500
1000
500


Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

Ada Boost


4662
58
191
1619
0
1
True label
Predicted label
4000
3000
2000
1000


Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

Result with 10-fold Cross validation

	Classifier	Accuracy Mean	Precision Mean	Recall Mean	F1 Score Mean	ROC AUC Mean
0	RandomForest	94.81	91.24	90.00	90.57	98.20
1	Naive Bayes	85.74	70.45	83.81	76.53	93.01
2	KNeighbors	83.02	66.72	77.46	71.67	88.06
3	SVM	95.04	91.33	90.77	91.03	98.77
4	AdaBoost	96.19	96.58	89.45	92.85	99.17

Wilcoxon Statical Test (Threshold=5% f1)

Classifier Comparison	P-value
AdaBoost-SVM	0.001953125
Random Forest-AdaBoost	0.001953125
Random Forest-SVM	0.375

Wilcoxon Statical Test (Threshold=5% f1) Before and After Resampling

Classifier	P-value
Random Forest	0.005859375
Naïve Bayes	0.845703125
SVC	0.275390625
KNN	0.001953125
AdaBoost	0.001953125

Conclusion

- Best Classifier is Random Forest in the case without sampling, with SMOTE there are Random Forest and SVC because we can't reject the null hypothesis
- Only AdaBoost reject the null hypothesis for before after resampling, we can choose the version with SMOTE with improvement of 0.2% in f1
- The operation of SMOTE deteriorate the performance of almost all classifiers in term of f1 score, even if for Adaboost there are only 0.2% of improvement, so the cost of SMOTE is not justified so the simplest pipeline should be chosen