UNIVERSITÀ DI PISA

Master’s Degree in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning

Telco Customer Churn Dataset Analysis

LIU CHANG

Introduction

This analysis begins with the breakdown of the entire dataset that contains different parameters that describe the characteristics of churn decision taken by the customers from different telecommunication companies.

Our aim is to create different models built with the entire dataset, and to choose the most suitable one to handle the specific machine learning task which is to predict the decision to churn or not of the customer of a telecommunication company. This model will be exploited by the telecommunication company's manager to predict in advance the decision of churn of the customer and allow the manager to have time in advance to implement some marketing strategy to lower the churn desire.

Design and Implementation

There are performed a preprocessing task over the dataset.

By the first analysis of the dataset, we first drop the irrelevant features such as: customerID, churn rate, Churn, Country, State, Zip Code, Lat Long, Latitude, Longitude. Because them are always different since they are unique like the customerID is different for every customer, country is all USA, State is always California, and the coordinate of the house changes for every customer such as the id.

Then we dropped the Count feature since it contains only “1”.

We dropped the rows that contains “Joined” as Customer Status since this type of customer’s behavior is yet to be determined because are customers that are joined now in the company.

Then We transform categorical features like 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 into respectly :

- Yes → 1
- No → 0
- No Phone Service → 0
- No Internet Service → 0

Then the Customer status that are given by the 2 classes Stayed and Churned are transformed into 0 and 1 respectively.

By analyzing the Churn Reason we can discover that there are Deceased and Moved reasons that are caused by major forces so I will drop rows which are present these two values.

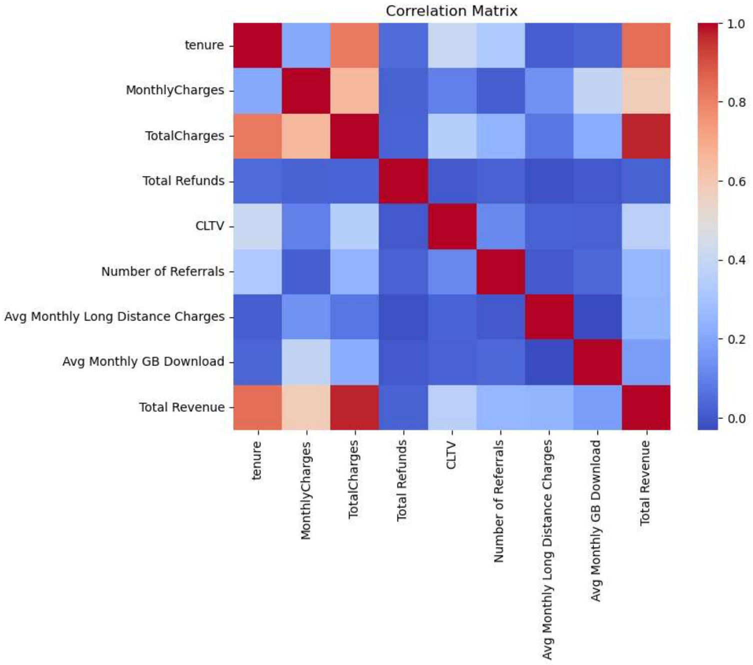
By the math we have an intuition about Long distance charge and Avg monthly long distance charge, I want to find their relationship.

I discover that these two are related thanks to the tenure feature by simply using the math Total=Avg*tenure, so I drop the total one that are presented by the long distance charge feature.

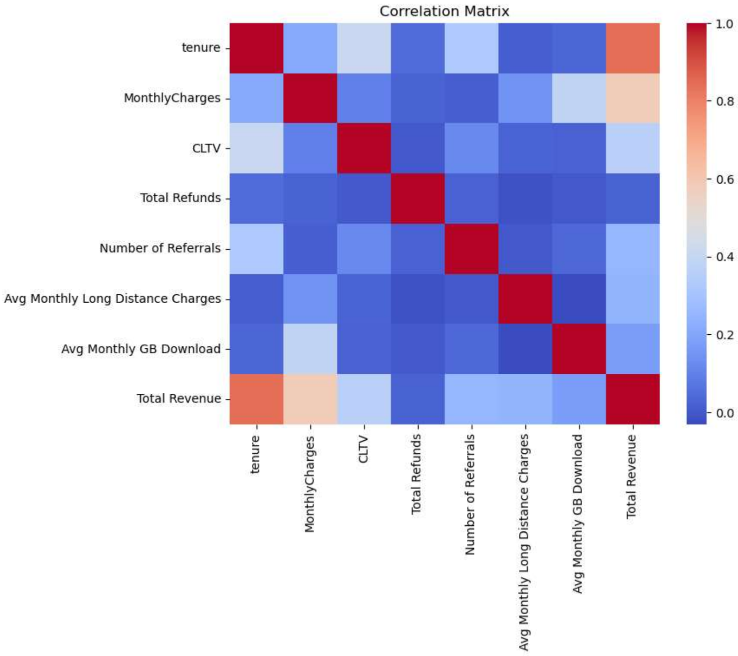
Now I want to divide in two types the people that has charged the data and the people that not charged so I create a new feature named "Has extra data charged" that for people that charged is true else is false

Then we provide to plot the confusion matrix related to the original numeric data, original because we operated some transformation before that in original they were categorical data so we don't consider them.

Numeric data considered are tenure, monthly charge, total charge, total refunds, CLTV(customer lifetime value),Number of rederrals, avg monthly long distance charges, avg monthly gb download, total revenue.



I want to eliminate redoundant features that gives me redundant informations, given by the correlation in range -0.75 to 0.75(chosen threshold), in this case the only one are Total charges and total revenue, since we know that the money spend to charge is equal to the revenue of the company referred to a customer, so we drop the Total Charges feature.



After that I want to verify if Married feature is equal to the Partner feature and the answer is yes so they are redundant so I provide to eliminate the Married one.

Then for InternetService, Offer, PaymentMethod and Contract we get dummy features from them and then when we have it, we drop these original features, then we map gender feature into “Male” to “0” and “Female” in “1”.

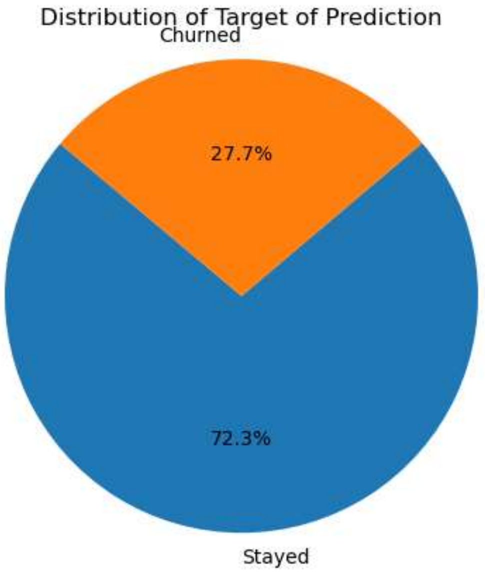
Then we pass to the model building.

Models that we choose are:

- RandomForestClassifier
- AdaBoostClassifier
- GaussianNB
- KneighborClassifier
- SVC.

We chose the X features and Y feature, the X features are all features except for Churn Reason, Churn Category and Customer Status, the other are 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 and for the Y features the Customer Status, that are Stayed or Churned.

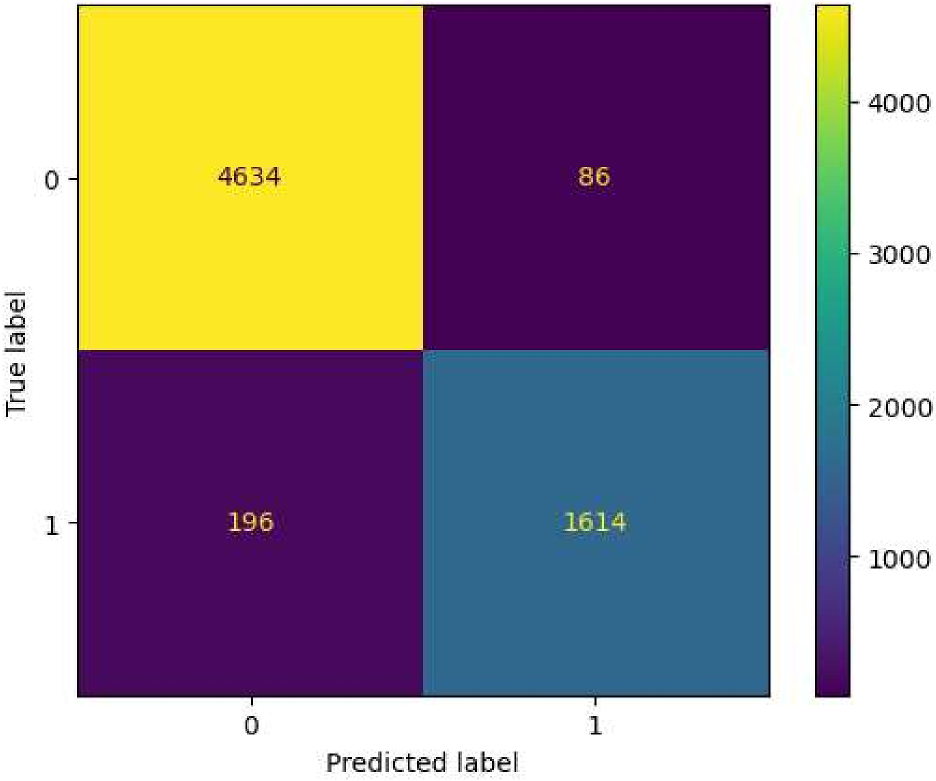
The distribution of the Y feature are:

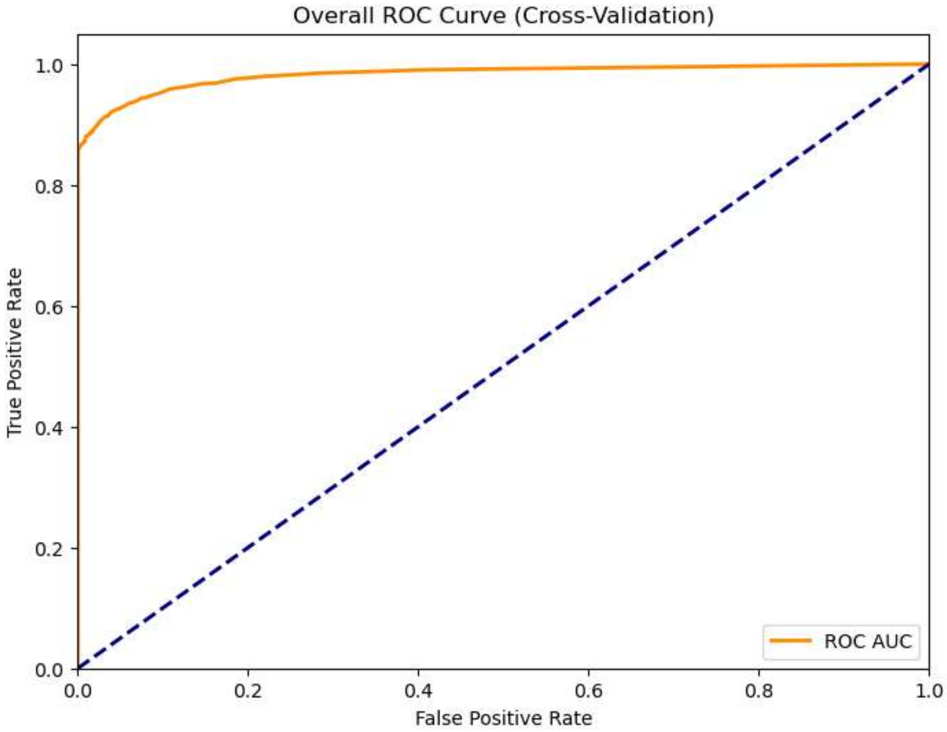


Then we perform a Cross-validation test on these classifiers by using 10 folds and for each classifier we build a pipeline that is composed by a MinMaxScaler for data normalization, SelectKBest for Feature selection that select the 10 best features based on f_classif that assess the ability of features on the discrimination of classes, and the classifier, the metrics used to valuate models are: accuracy, auc, recall, precision, f1 scores.

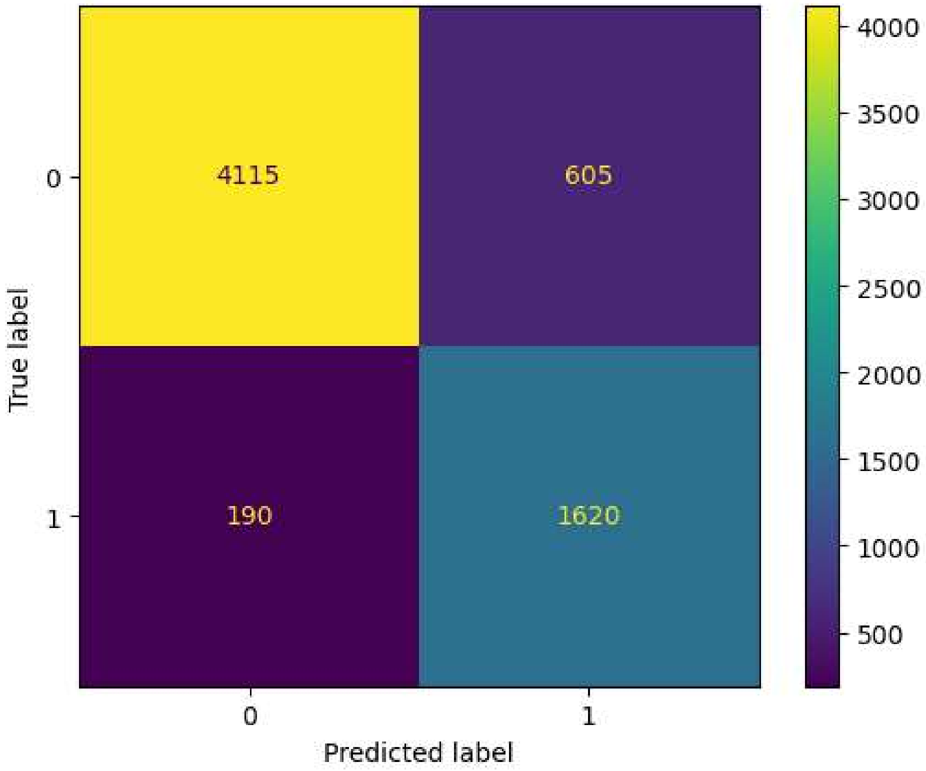
Results

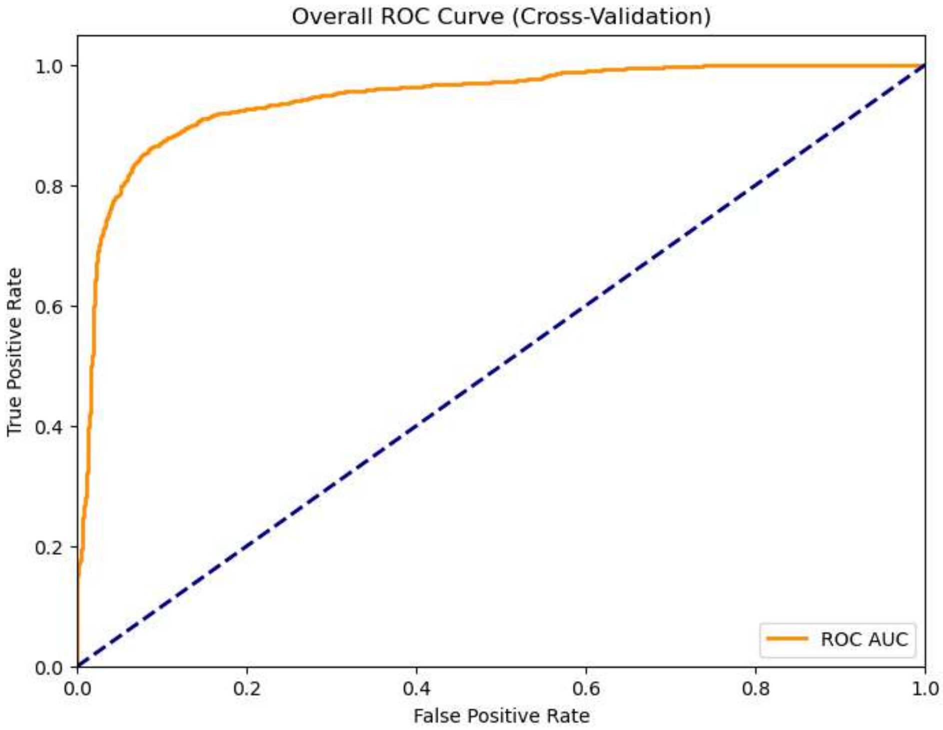
- Random Forest



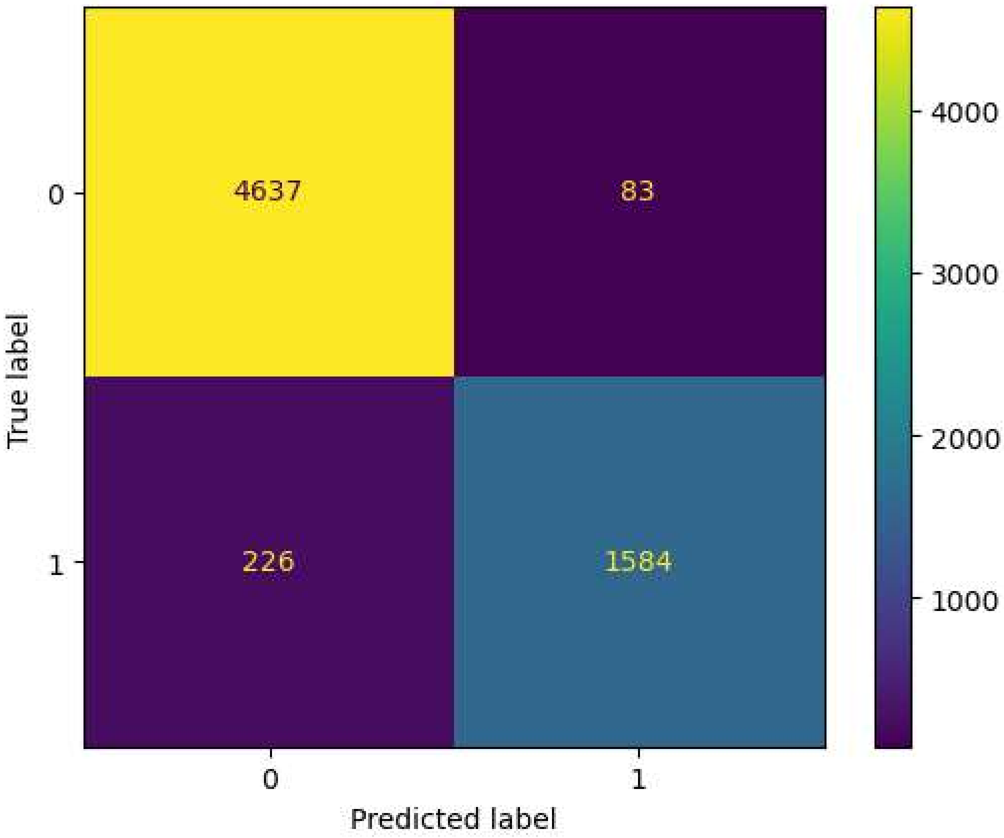


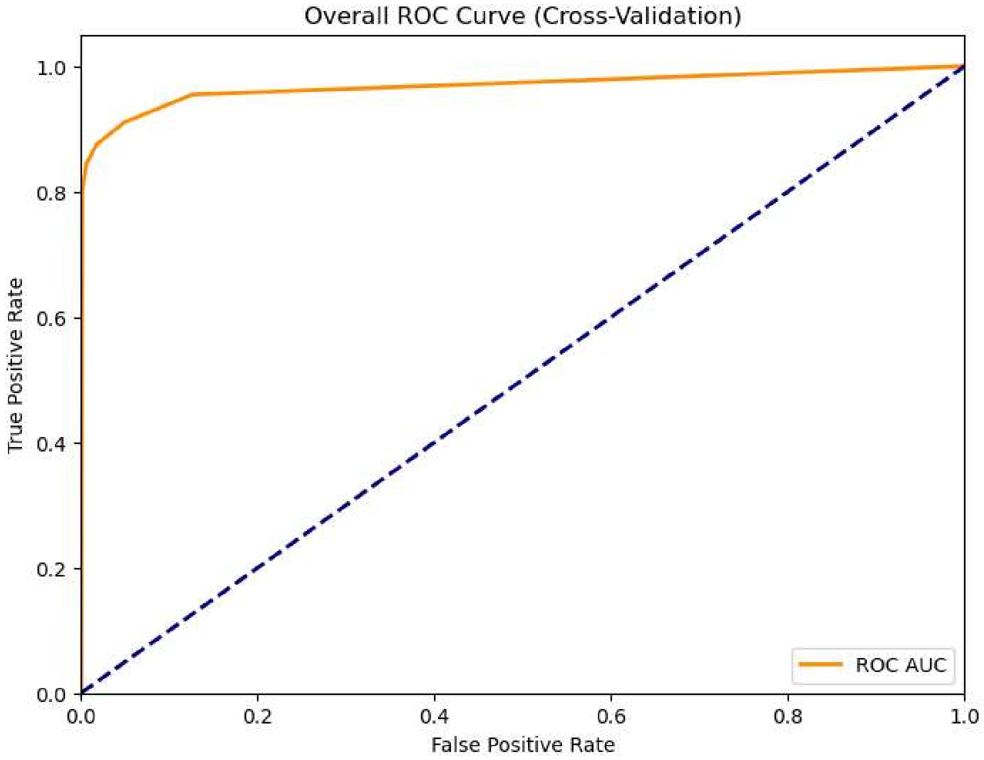
- Naive Bayes


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

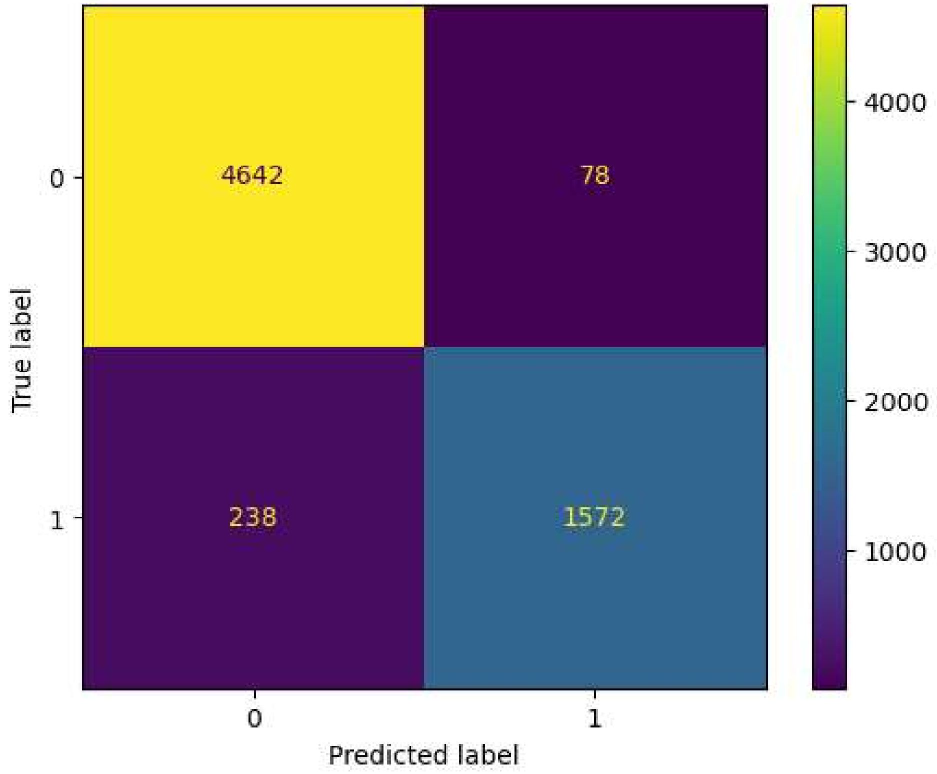

Overall ROC Curve (Cross-Validation)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC AUC

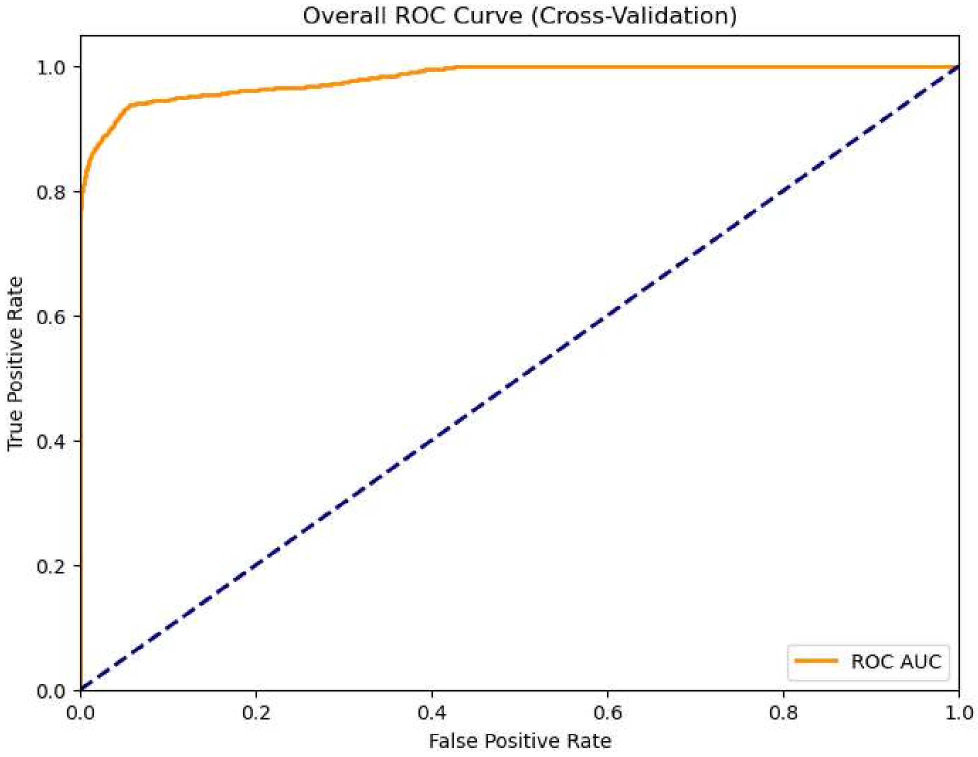
- KNeighbors


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

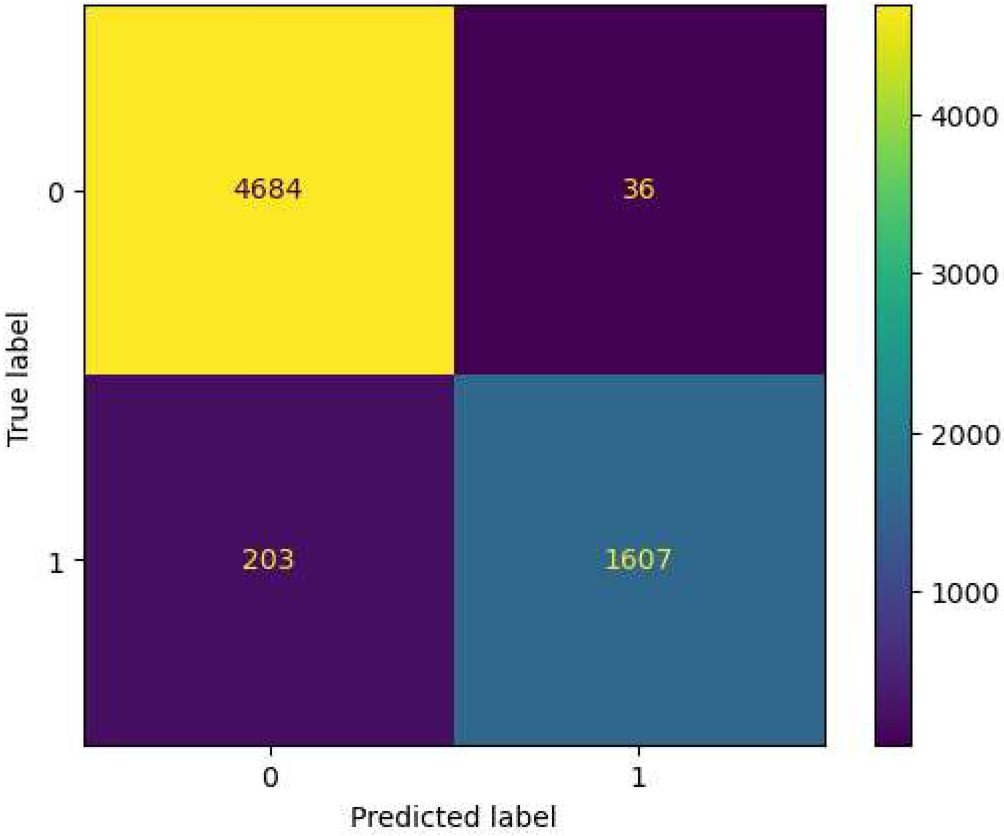

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC

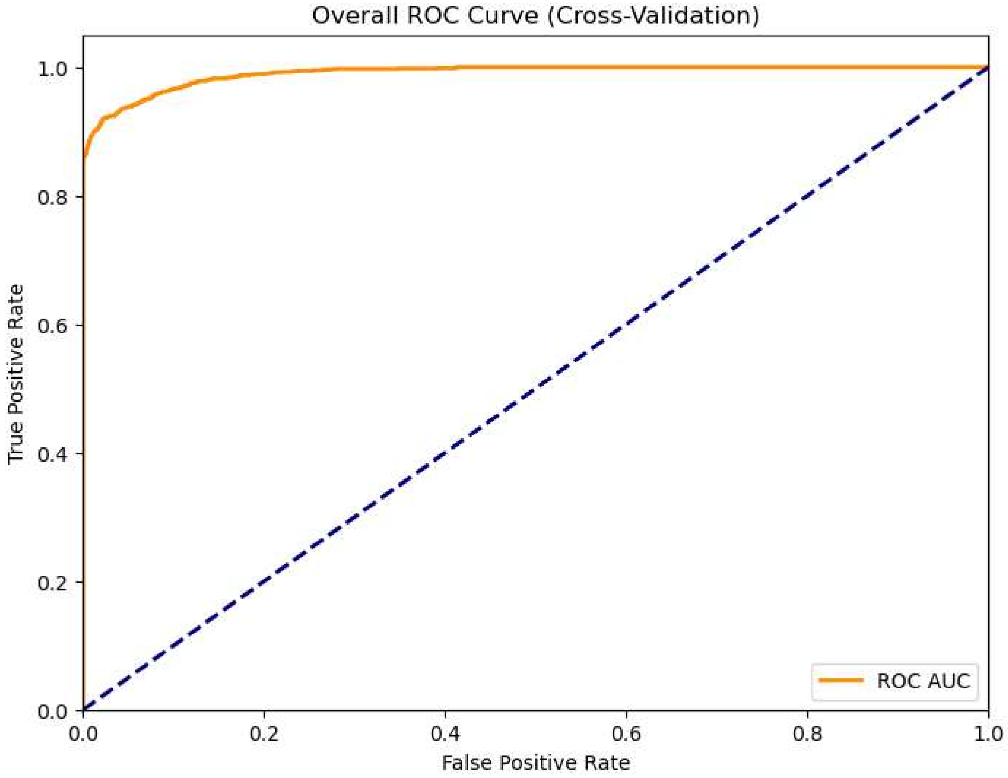
- SVM


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

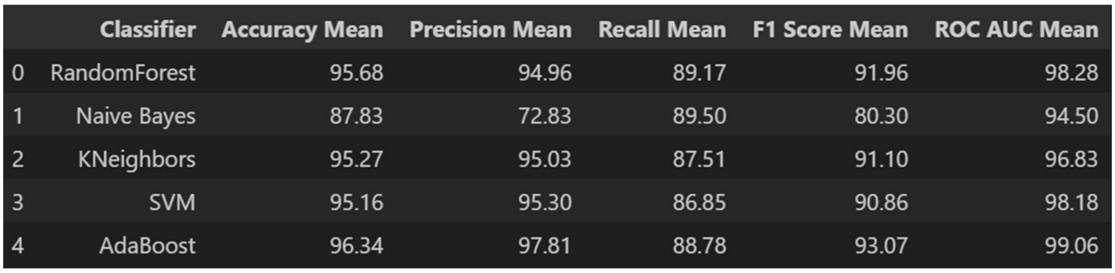

Overall ROC Curve (Cross-Validation)
True Positive Rate
False Positive Rate
ROC AUC
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

- AdaBoost


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


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

After Cross-Validation and we get the metrics for every fold, then we calculate the mean of the metrics.



	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

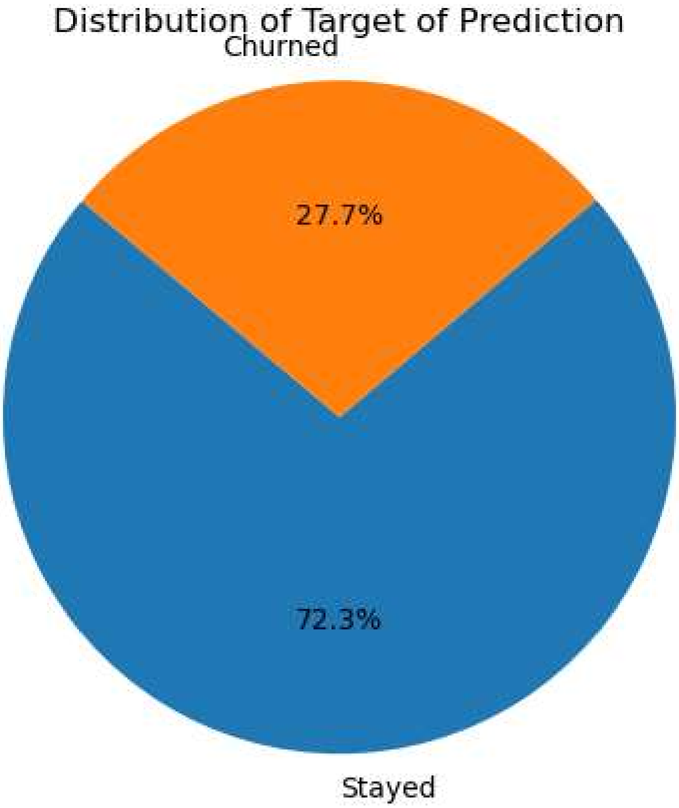
From the picture we can see that the top 4 based on F1-score are Random Forest, SVM, KNN and AdaBoost, then for these 3 classifiers I apply a statistical test to choose the best one, I chose only 4 over 5 because the performance difference are evident over 10% so even if we can reject the null hypothesis the f1 score of Naïve bayes is the worst, while the other 4 are similar, the metric chosen is the f1-score.We now perform a Wilcoxon statistical test with significance level of 5%, the result are

Classifier	P-value
ADA-SVM	0.001953125
RF-ADA	0.010862224704815628
RF-SVM	0.048828125
ADA-KNN	0.001953125
KNN-SVM	0.625
RF-KNN	0.009765625

With the significance level 0.05 we can say that in all the cases the null hypothesis are rejected except for the case ADA-SVM, Random forest is better than SVC and Adaboost, SVC, KNN and Adaboost is better than SVC and KNN, so in this case we have the best classifier which is Random Forest.

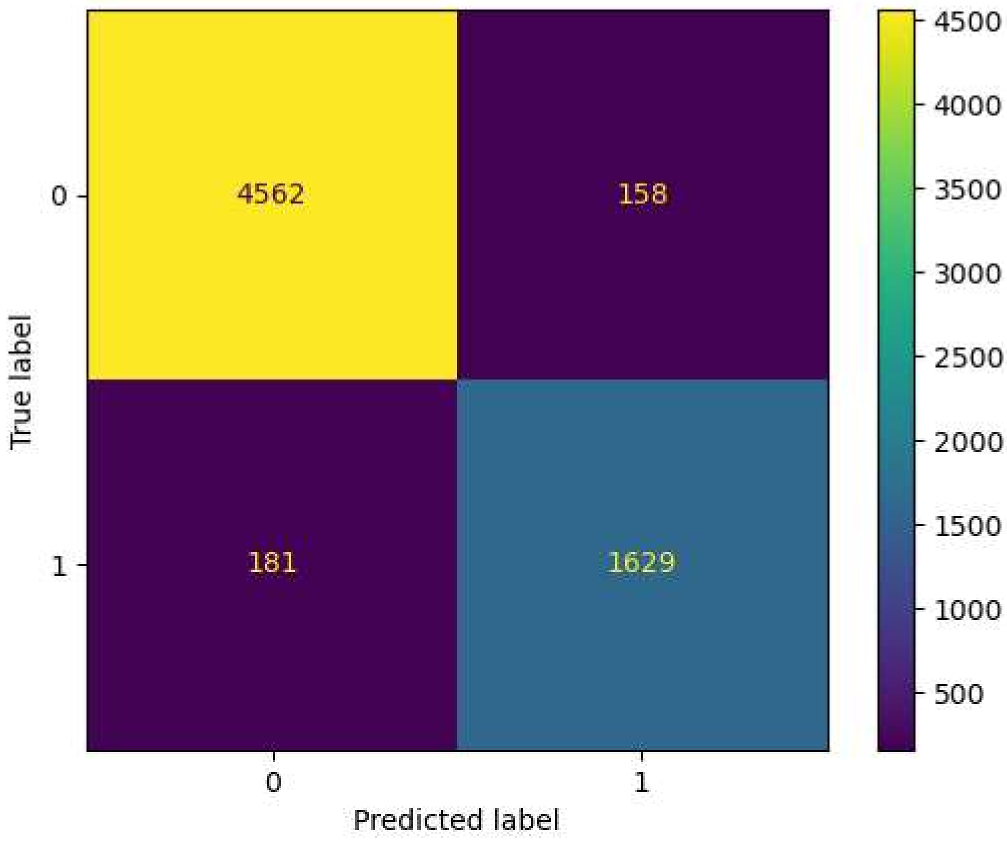
Results after sampling

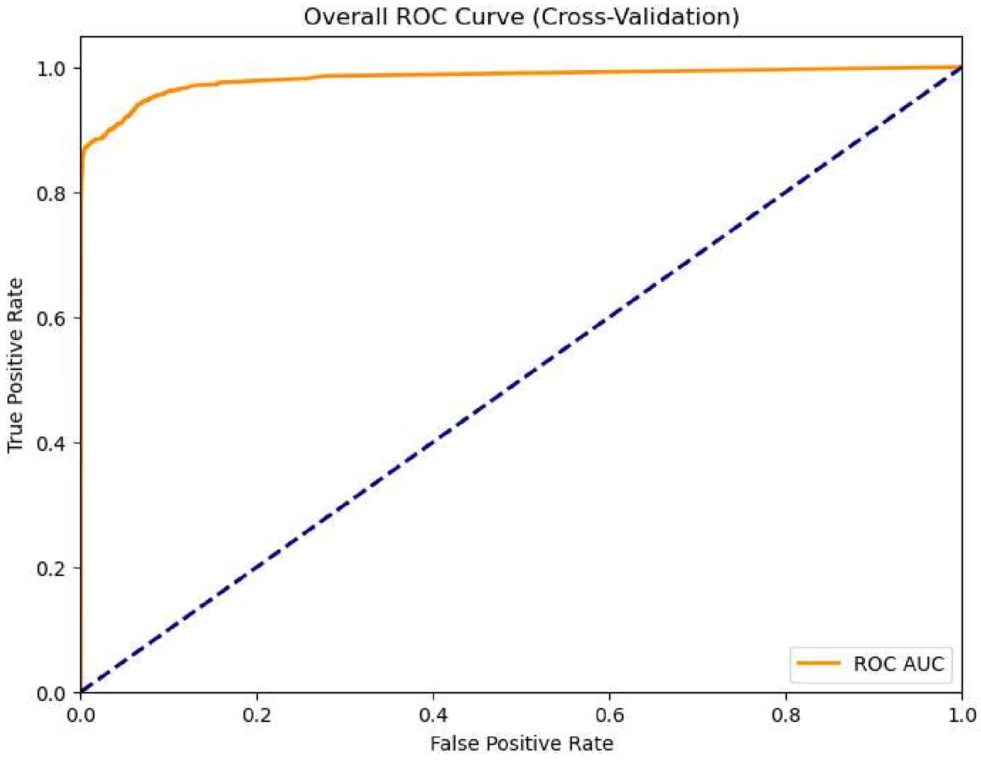
From a deep analysis of the class that we want to classify we see that the distribution of the 2 class are 72.3% and 27.7%, so they are imbalanced, so we want now try to perform a sampling with objective to increase the performance of classifier by rebalancing the dataset by using SMOTE oversampling technique.



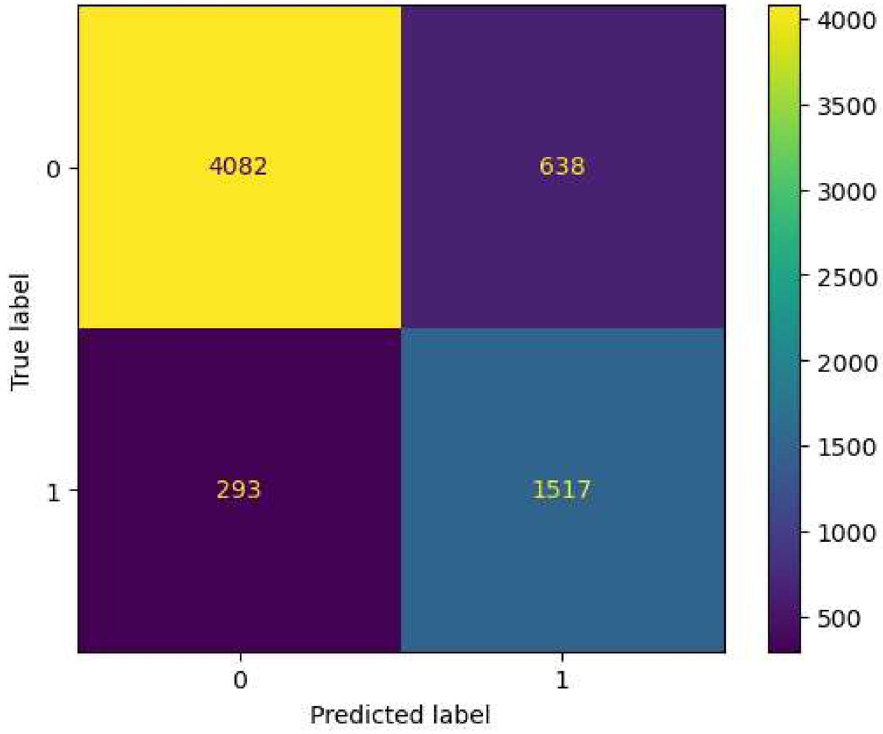
In this case the modify the pipeline used for the version without resampling, we add a stage before the scaler, that consists in the oversampling by using the SMOTE, so now the new pipeline is SMOTE-MaxMinScaler-SelectKBest-Classifier. Same as before we use a 10-fold cross validation, and calculate the same metrics as before, these now are done by the function eval_cross_validation.

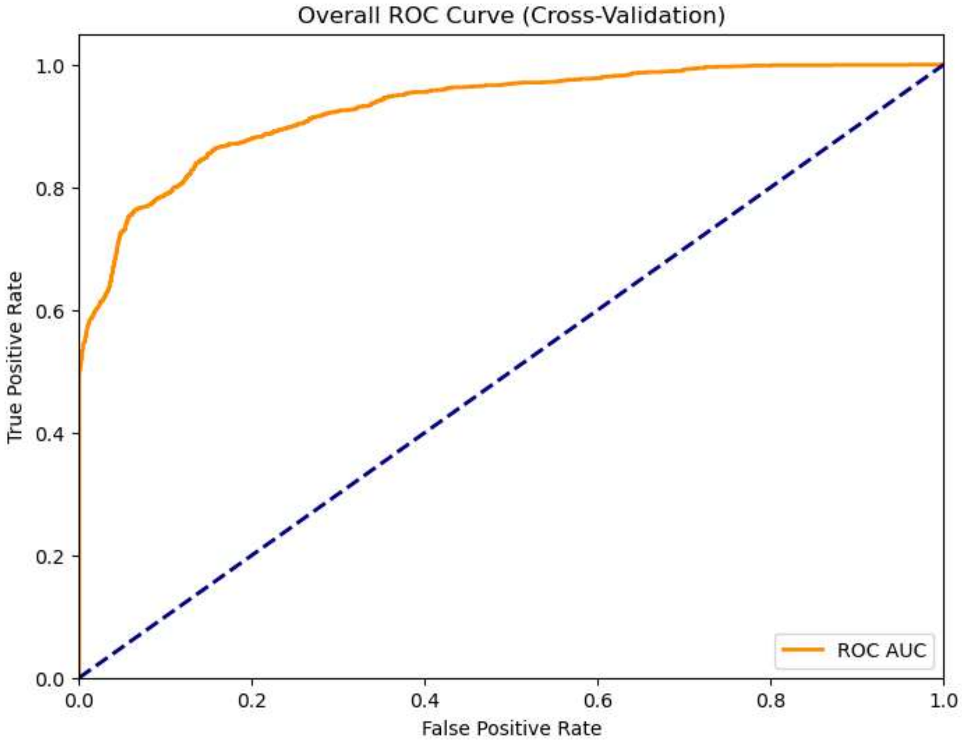
- Random Forest



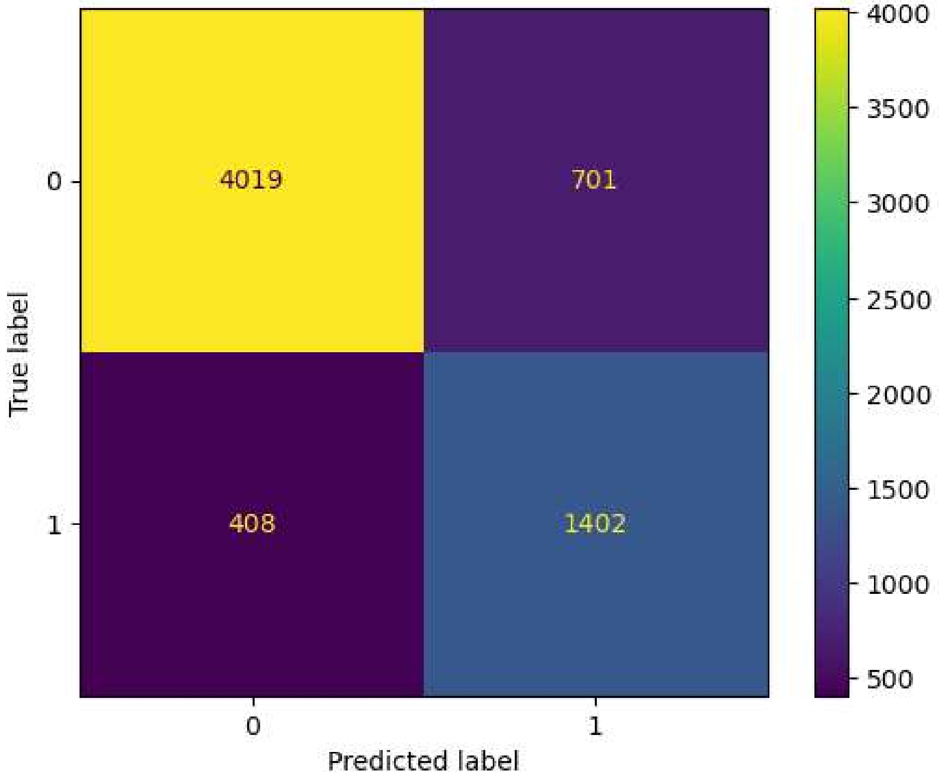


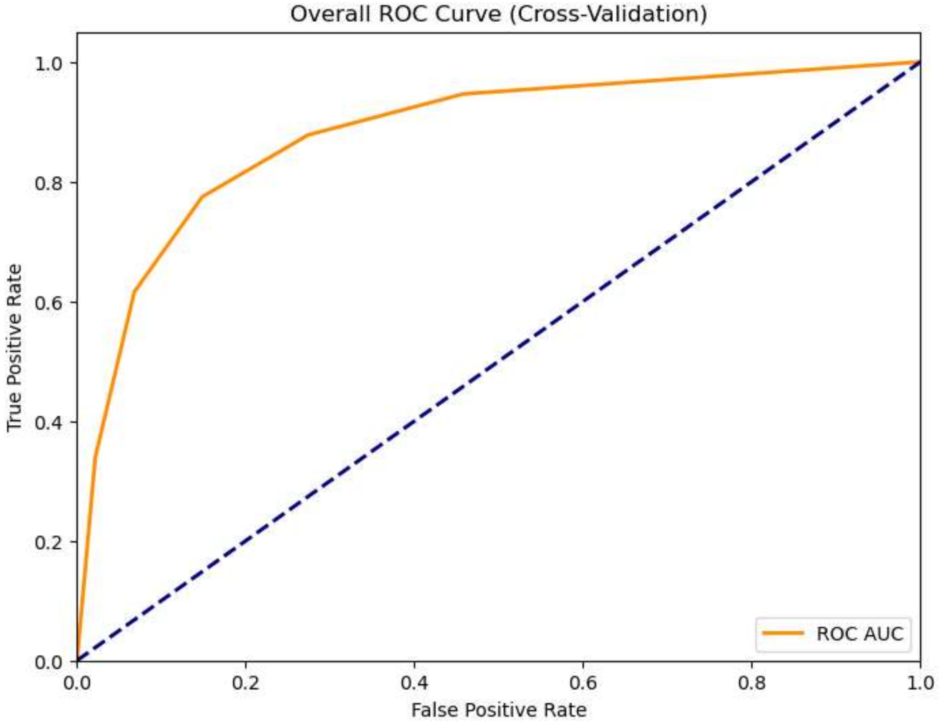
- Naive Bayes


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

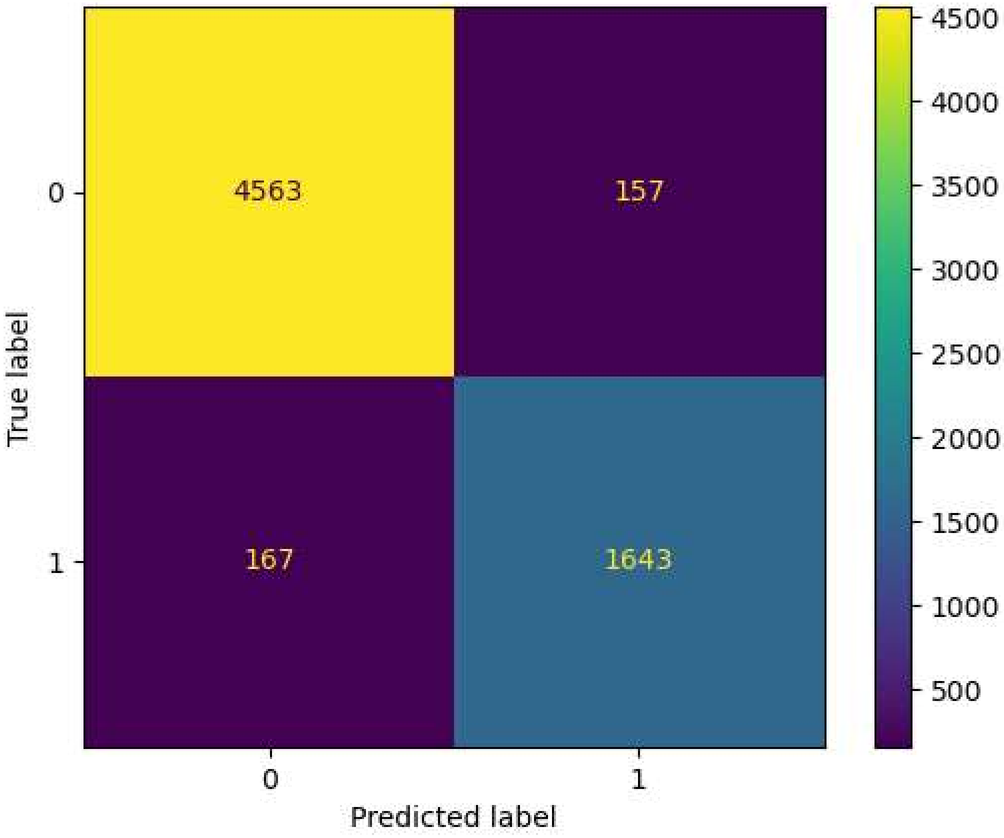

Overall ROC Curve (Cross-Validation)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC AUC

- KNeighbors

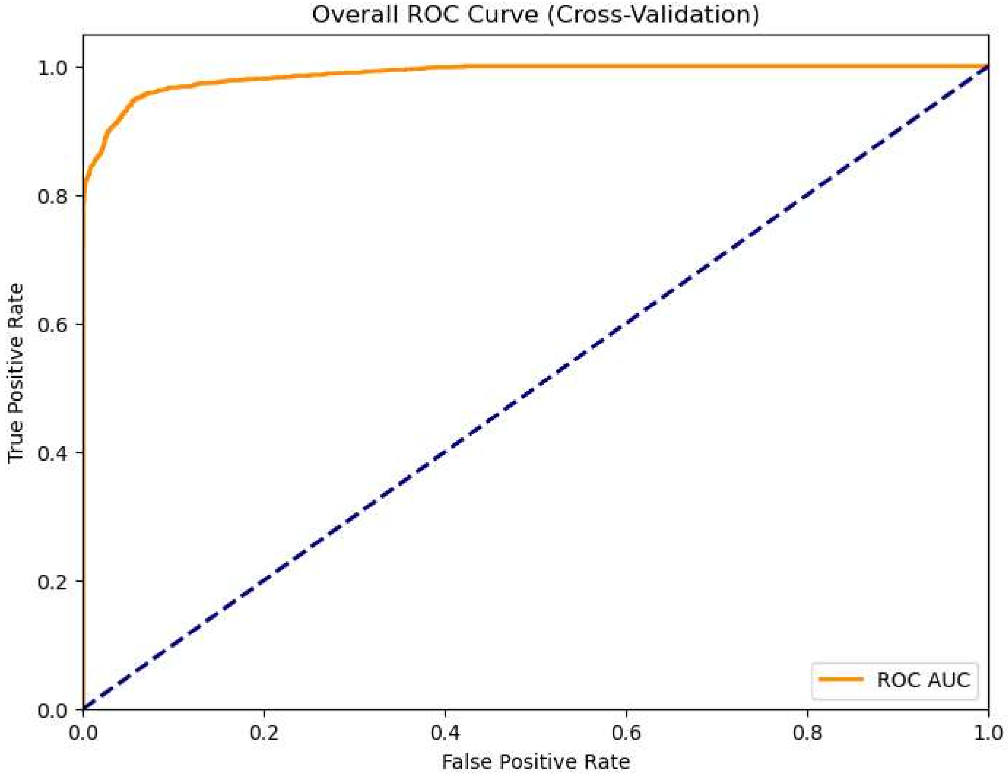

4019
701
408
1402
0
1
True label
0
1
Predicted label
4000
3500
3000
2500
2000
1500
1000
500


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

- SVC

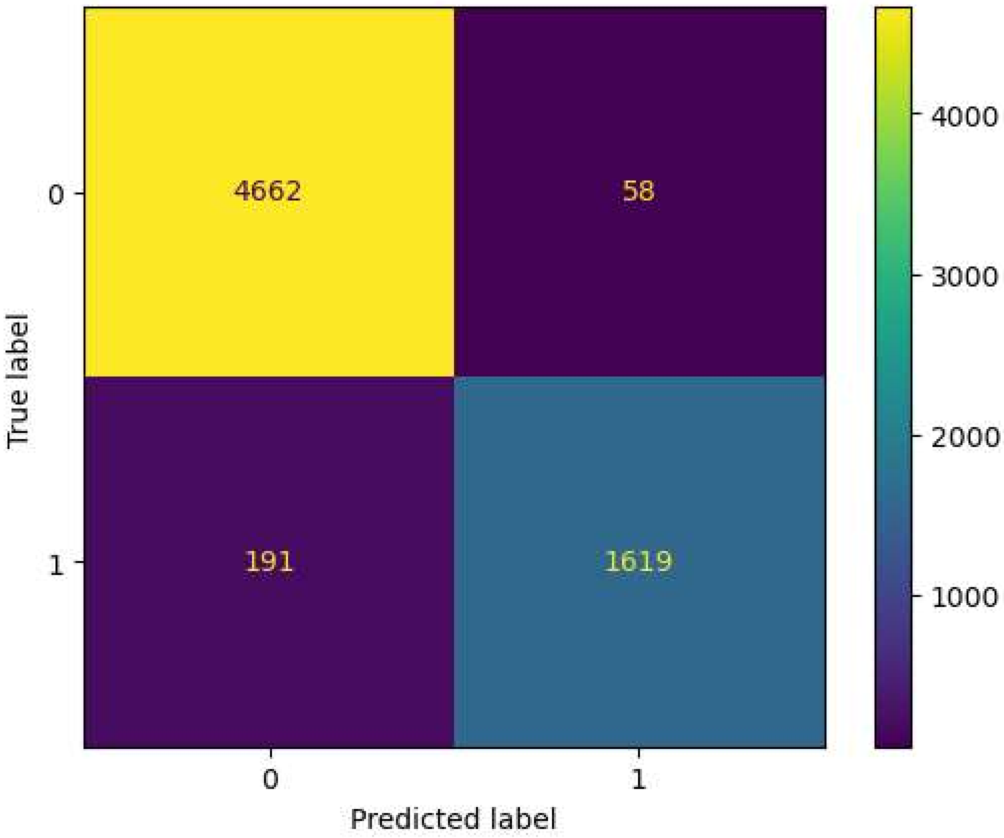


4563
157
167
1643
0
1
True label
0
1
Predicted label
4500
4000
3500
3000
2500
2000
1500
1000
500

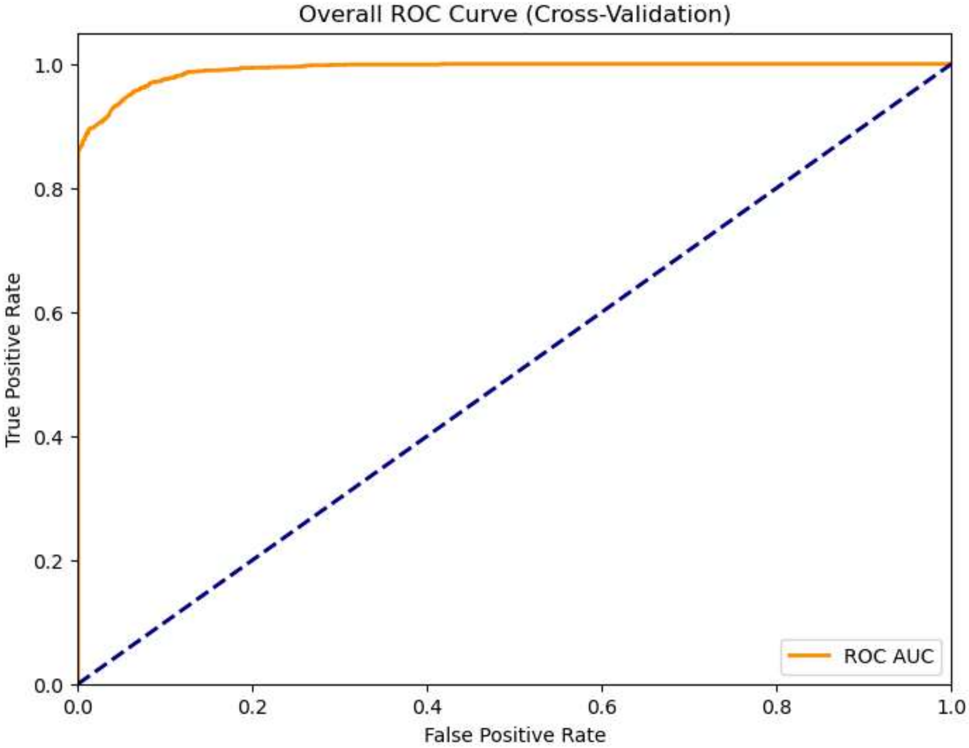


Overall ROC Curve (Cross-Validation)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC AUC

- AdaBoost

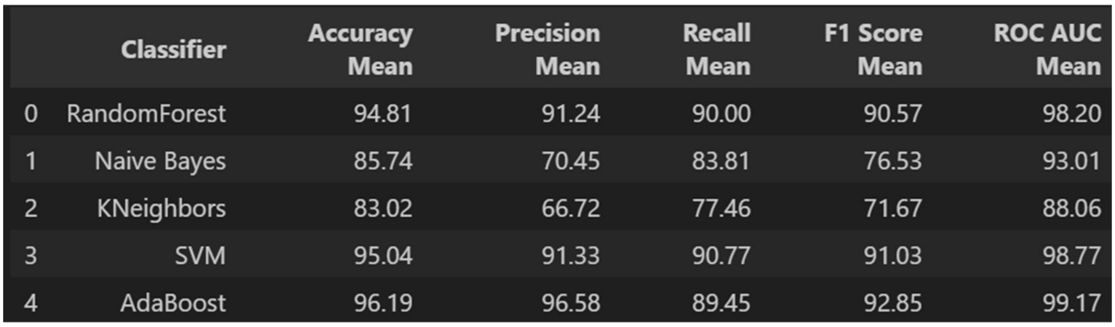


4662
58
191
1619
0
1
True label
0
1
Predicted label
4000
3000
2000
1000



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

After Cross-Validation and we get the metrics for every fold, then we calculate the mean of the metrics.



	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

From the picture we can see that the top 3 based on F1-score are Random Forest, SVM and AdaBoost, then for these 3 classifiers I apply a statistical test to choose the best one, the metric chosen is the f1-score, we chose these 3 over 5 for the same reason as before.

Based on the Wilcoxon test with 5% significance threshold, the result are

Classifiers	P-Value
ADA-SVC	0.001953125
RF-ADA	0.001953125
RF-SVC	0.375

With 0.05 as threshold we can conclude that in all the 3 cases we can reject the null hypothesis so Adaboost is better than SVC, Random forest is better than Adaboost, we can’t say nothing about SVC, so in this case the best classifier is the Random Forest classifier based on the f1-score metric compared to AdaBoost but with SVC we can’t say a thing.

Now we want a comparison between the result before and after the sampling, by doing that we perform the Wilcoxon statical test on f1-score before and after resampling, the test new is conducted as always on the 3 classifier over 5 with significance threshold of 5%

CLASSIFIER	P-VALUE
Random Forest	0.005859375
Naïve Bayes	0.845703125
SVM	0.275390625
KNN	0.001953125
AdaBoost	0.001953125

With 0.05 as threshold the we see that Random forest has a 1.5% in the deterioration of performance in term of f1-score, since we can reject the null hypothesis the SMOTE version of Random Forest is worst, while for naïve bayes and SVC we can’t say a thing because we can’t reject the null hypothesis, while for KNN and Adaboost we can say that for SVM has deterioration in terms of performance, while for AdaBoost the improvement of the performance is only 0.2%。

In general there are no improvement of performance, in the only case of improvement is only 0.2%, so compared to the cost of SMOTE operation is not a great improvement in term of performance, so the simples pipeline should be chosen (The one without SMOTE).

Bibliography

Dataset: https://www.kaggle.com/datasets/sibelius5/telco-customer-churn?datasetId=1372879&select=Telco_customer_churn.csv