

CSE 454

Data Mining Term Project

Sedef Erdoğan
1801042102

Demonstration Video Link: <https://www.youtube.com/watch?v=vSwYQUsnRZw>

Contents

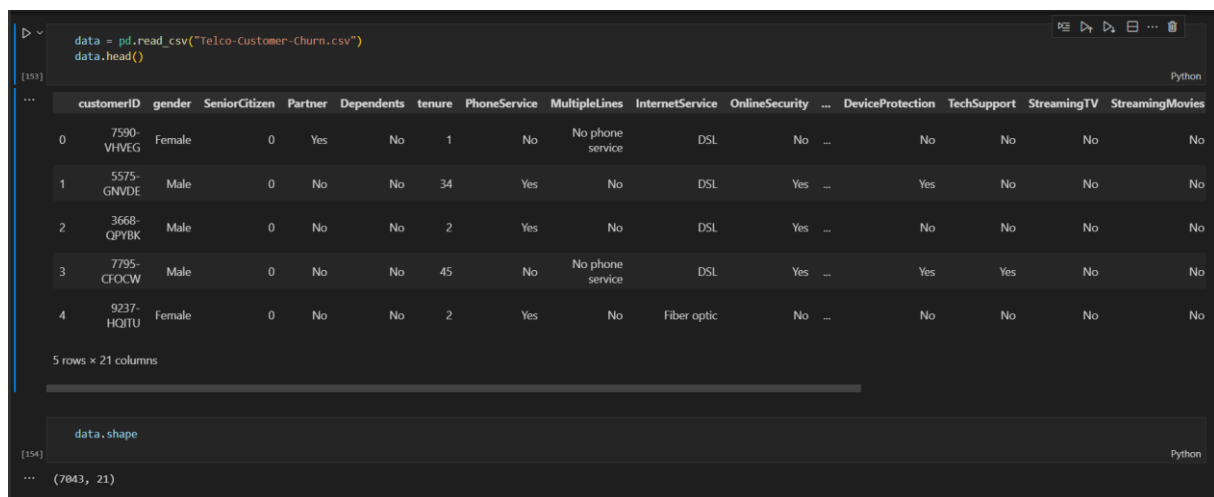
1. Problem Definition & Details.....	3
2. Data Preprocessing	4
3. Mean Values of All the Features & Target Variable Visualization (Churn)	5
3.1 Mean Values of All the Features	5
3.2 Target Variable Visualization (Churn)	6
4. Normalization.....	7
5. Correlation Matrix	7
6. Feature Selection	8
6.1 Feature Selection for Categorical Features.....	8
6.2 Feature Selection for Numerical Features	9
7. Evaluation Models.....	9
7.1 Decision Tree	10
7.1.1 Decision Tree with No Resampling	10
7.1.2 Decision Tree with RandomOverSampler.....	11
7.1.3 Decision Tree with RandomUnderSampler.....	12
7.1.4 Conclusion.....	13
7.2 Random Forest	13
7.2.1 Random Forest with No Resampling	14
7.2.2 Random Forest with RandomOverSampler.....	15
7.2.3 Random Forest with RandomUnderSampler.....	16
7.2.4 Conclusion.....	16
7.3 Logistic Regression Classification	17
7.3.1 Logistic Regression Classification with No Resampling	18
7.3.2 Logistic Regression Classification with RandomOverSampler	18
7.3.3 Logistic Regression Classification with RandomUnderSampler	18
7.3.4 Conclusion.....	19
8. Academic Paper Summary	19
9. References	20

1. Problem Definition & Details

In the telecom industry, customers are able to choose from a pool of companies to cater their needs regarding communication and internet. Customers are very critical about the kind of services they receive and judge the entire company based on a single experience. These communication services have become so recurrent and inseparable from the daily routine that a 30-minute maintenance break kicks in anxiety in the users highlighting our taken-for-granted attitude towards these services. Coupled with high customer acquisition costs, churn analysis becomes very pivotal. Churn rate is a metric that describes the number of customers that cancelled or did not renew their subscription with the company. Thus, higher the churn rate, more customers stop buying from your business, directly affecting the revenue! Hence, based on the insights gained from the churn analysis, companies can build strategies, target segments, improve the quality of the services being provided to improve the customer experience, thus cultivating trust with the customers. That is why building predictive models and creating reports of churn analysis becomes key that paves the way for growth.

Aim of this project is classifying the potential churn customers based on numerical and categorical features.

For this project, I found a dataset created from Kaggle. In this dataset, there are 7043 rows and 21 columns.



```
data = pd.read_csv("Telco-Customer-Churn.csv")
data.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	No	No	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	Yes	No	No	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	No	No	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	Yes	Yes	No	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No

5 rows x 21 columns

```
data.shape
```

[154] (7043, 21)

Dataset Attributes:

- customerID: Customer ID
- gender: Whether the customer is a male or a female
- SeniorCitizen : Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- tenure: Number of months the customer has stayed with the company
- PhoneService: Whether the customer has a phone service or not (Yes, No)
- MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)

- InternetService: Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: The amount charged to the customer monthly
- TotalCharges: The total amount charged to the customer
- Churn: Whether the customer churned or not (Yes or No)

2. Data Preprocessing

A heatmap is a graphical representation of data where values are represented as colors. It can be used to visualize the distribution of null (missing) values in a dataset. Figure 1 is the heatmap created for this dataset. According to this, no null values present in the data.

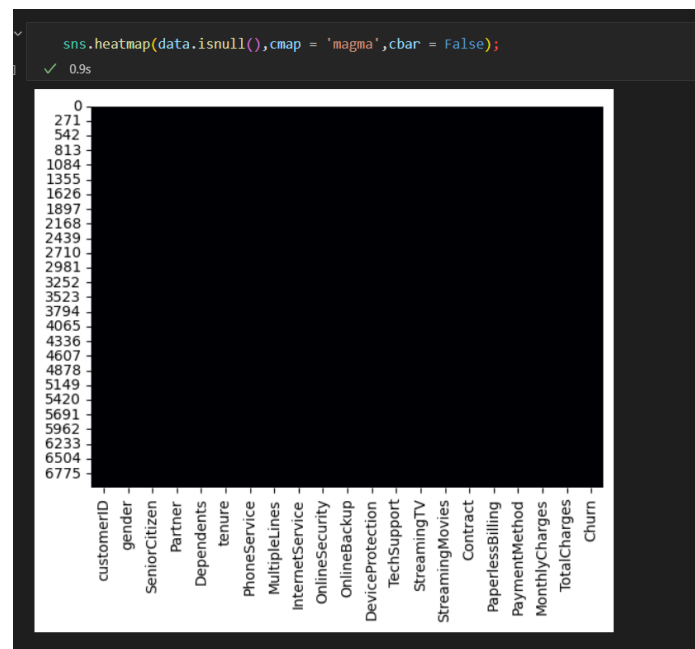


Figure 1. Heatmap

In this dataset, TotalCharges is a feature with numerical values but are stored in string datatype. While converting the 'TotalCharges' to float, an error occurred with the message describing that it could not

convert string to float. There is empty strings present in the some TotalCharges columns. As these elements were defined as string (E.g: a = ' '), they did not appear as null values and hence the heatmap for missing values did not display anything. After replacing the empty string with nan, I check null values. There are 11 null values. I preferred to fill these null values with the median of TotalCharges.

```
data.isnull().sum()
gender          0
SeniorCitizen  0
Partner         0
Dependents      0
tenure          0
PhoneService    0
MultipleLines   0
InternetService  0
OnlineSecurity  0
OnlineBackup    0
DeviceProtection 0
TechSupport     0
StreamingTV     0
StreamingMovies 0
Contract        0
PaperlessBilling 0
PaymentMethod   0
MonthlyCharges  0
TotalCharges    11
Churn           0
dtype: int64
```

Figure 2. Null Values

I drop the customerID column because it is not relevant to churn. I divide the features into numerical and categorical features. I also execute the label encoding transformation for categorical features.

3. Mean Values of All the Features & Target Variable Visualization (Churn)

3.1 Mean Values of All the Features

Mean Values of all the features for for churned and not-churned customers are shown in Figure 3.

Churned Customers		Not_Churned Customers	
gender	0.50	gender	0.51
SeniorCitizen	0.25	SeniorCitizen	0.13
Partner	0.36	Partner	0.53
Dependents	0.17	Dependents	0.34
tenure	17.98	tenure	37.57
PhoneService	0.91	PhoneService	0.90
MultipleLines	1.00	MultipleLines	0.92
InternetService	0.81	InternetService	0.89
OnlineSecurity	0.38	OnlineSecurity	0.94
OnlineBackup	0.62	OnlineBackup	1.01
DeviceProtection	0.64	DeviceProtection	1.00
TechSupport	0.39	TechSupport	0.94
StreamingTV	0.93	StreamingTV	1.00
StreamingMovies	0.94	StreamingMovies	1.01
Contract	0.14	Contract	0.89
PaperlessBilling	0.75	PaperlessBilling	0.54
PaymentMethod	1.76	PaymentMethod	1.51
MonthlyCharges	74.44	MonthlyCharges	61.27
TotalCharges	1531.80	TotalCharges	2552.88
Churn	1.00	Churn	0.00
mean		mean	

Figure 3. Mean Values

Clearly, the customers that churned had a low mean tenure of 17.98 months as compared to those who continued with an average tenure period of 37.57 months.

Mean values of OnlineSecurity, OnlineBackup, DeviceProtection and TechSupport are higher for not-churned customers than churn customers. This can serve as a good indicator or point to focus on.

Churned customer's Contract value is much smaller than those of not-churned customers.

Mean MonthlyCharges of the churn customers, 74.44, is more than that of not-churn customers, 61.27.

Not-churned customers TotalCharges, 2552.88, is higher than churn customers, 1531.80.

From these mean values, we can say that some of the features display a clear cut difference that can help to focus more churn customers to make sure they retain the services. The dataset has too many categorical features, hence mean values of the features are present in the vicinity of 0.

3.2 Target Variable Visualization (Churn)

As we can see in Figure 4, the dataset is imbalanced. Due to this, predictions will be biased towards Not-Churn customers.

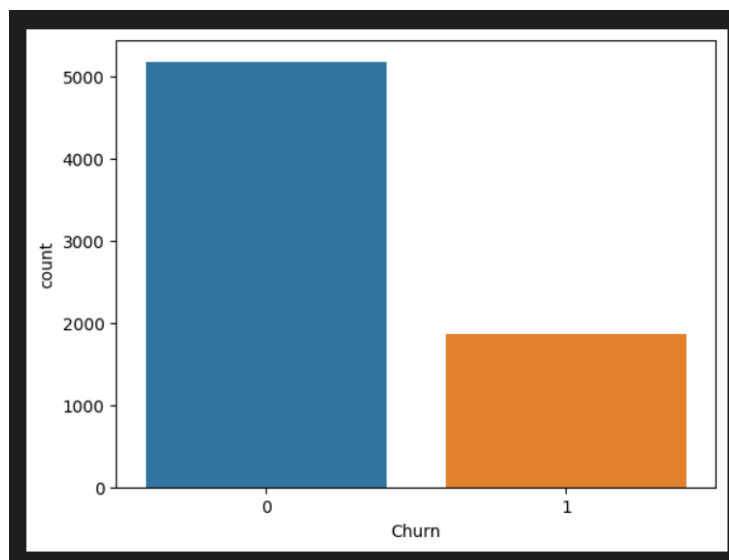


Figure 4. Target Variable Visualization

To fix class imbalance, RandomOverSampler and RandomUnderSampler will be used.

RandomOverSampler is a method for oversampling a dataset. The basic idea behind RandomOverSampler is that it selects examples from the minority class at random, and then duplicates them to increase the size of the minority class.

RandomUnderSampler is a method for under-sampling a dataset. The basic idea behind RandomUnderSampler is that it randomly selects examples from the majority class and removes them from the dataset.

4. Normalization

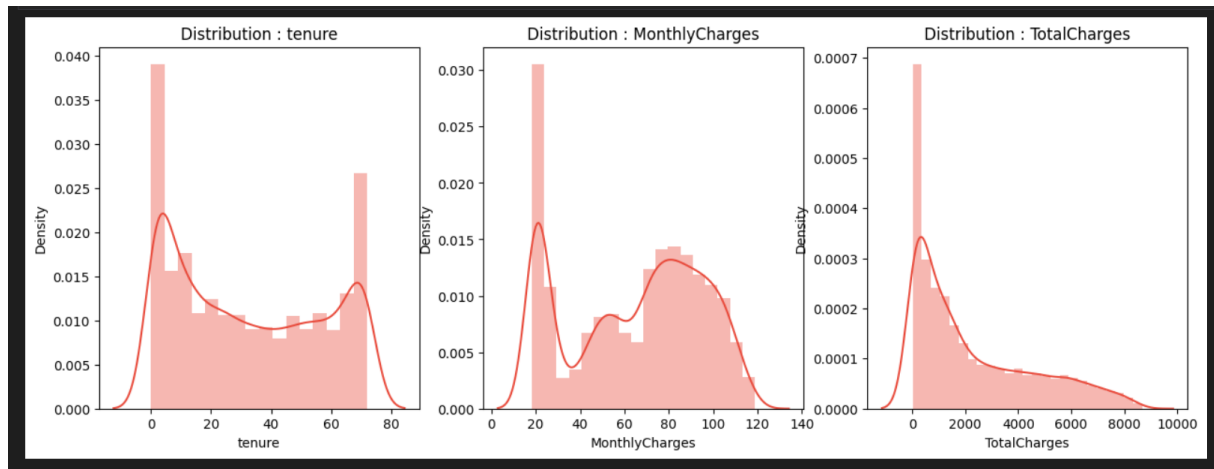


Figure 5. Distribution of Numerical Values

Models does not understand the units of the values of the features. It treats the input just as a simple number but does not understand the true meaning of that value. Thus, it becomes necessary to scale the data. We can scale data I 2 ways. First one is Normalization and second one is standardization. As most of the algorithms assume the data to be normally (Gaussian) distributed, Normalization is done for features whose data does not display normal distribution and standardization is carried out for features that are normally distributed where their values are huge or very small as compared to other features.

Tenure and MonthlyCharges kind of create a bimodal distribution with peaks present at 0 - 70 and 20 - 80 respectively. TotalCharges displays a positively or rightly skewed distribution. Tenure, MonthlyCharges and TotalCharges features are normalized as they displayed a right skewed and bimodal data distribution. None of the features are standardized for the above data.

5. Correlation Matrix

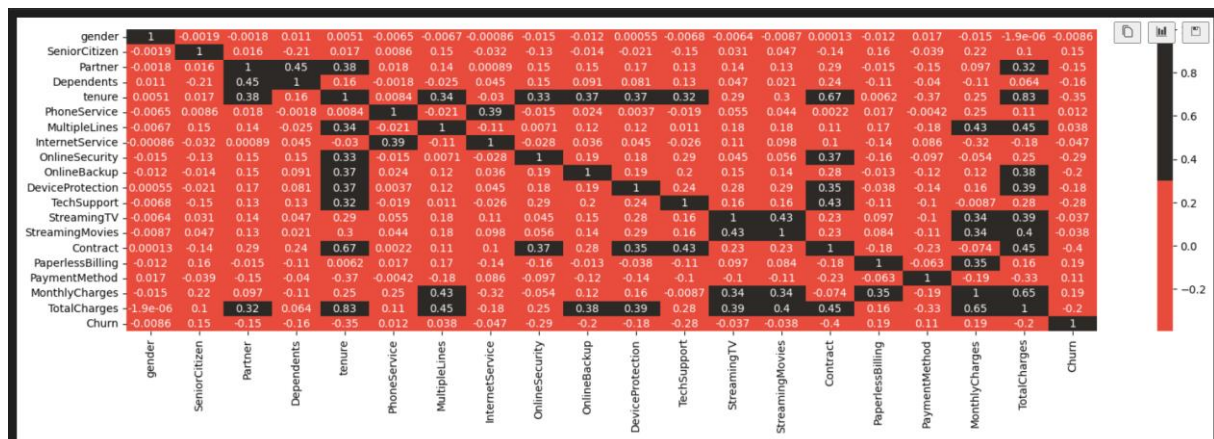


Figure 6. Correlation Matrix

As we can see in Figure 6, the correlation matrix is a huge matrix with too many features. I checked the correlation only with respect to Churn.

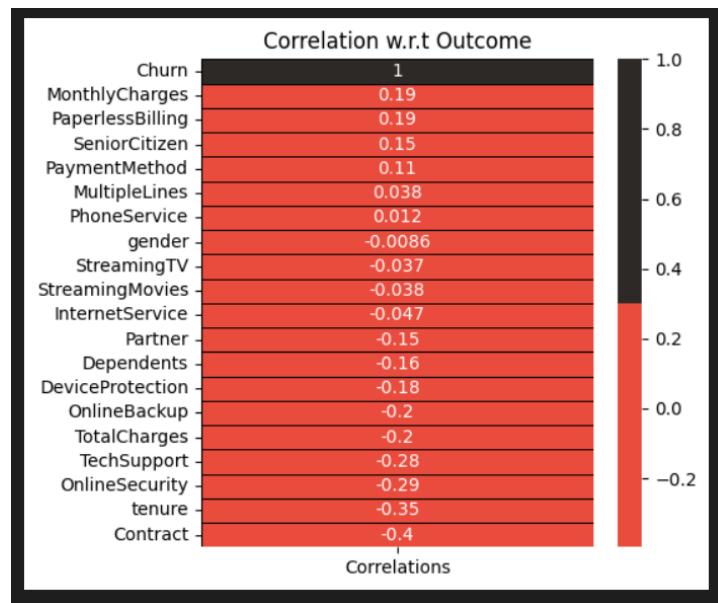


Figure 7. Correlation w.r.t. Churn

If we look at the Figure 7, we can say that MulipleLines, PhoneService, gender, StreamingTV, StreamingMovies and InternetService does not display any kind of correlation. I drop the features with correlation coefficient between $(-0.1, 0.1)$. Remaining features either display a significant positive or negative correlation.

6. Feature Selection

Feature selection is the process of selecting a subset of relevant features from the original set of features for a given task. The goal is to select the most informative and relevant features that can improve the performance of the model, while reducing the dimensionality and complexity of the data.

6.1 Feature Selection for Categorical Features

I apply chi-square test. As we can see in Figure 8, PhoneService, gender, StreamingTV, StreamingMovies, MultipleLines and InternetService display a very low relation with Churn.

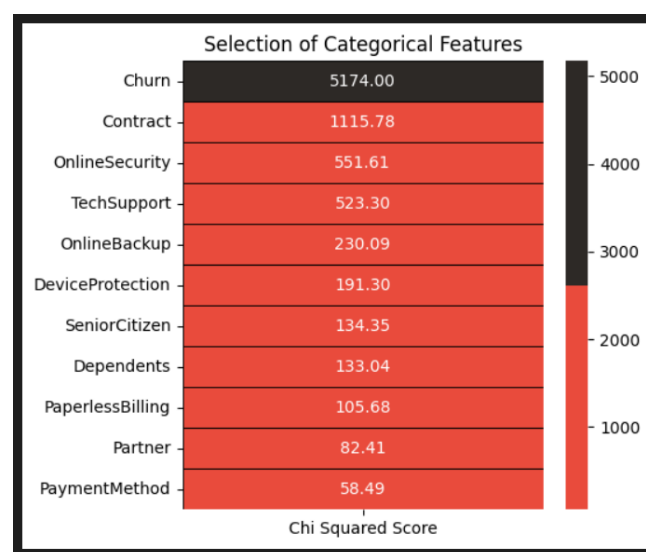


Figure 8. Selection of Categorical Features

6.2 Feature Selection for Numerical Features

I apply ANOVA test. According to the ANOVA test, higher the value of the ANOVA score, higher the importance of the feature. From the results in Figure 9, I include all the numerical features for modeling.

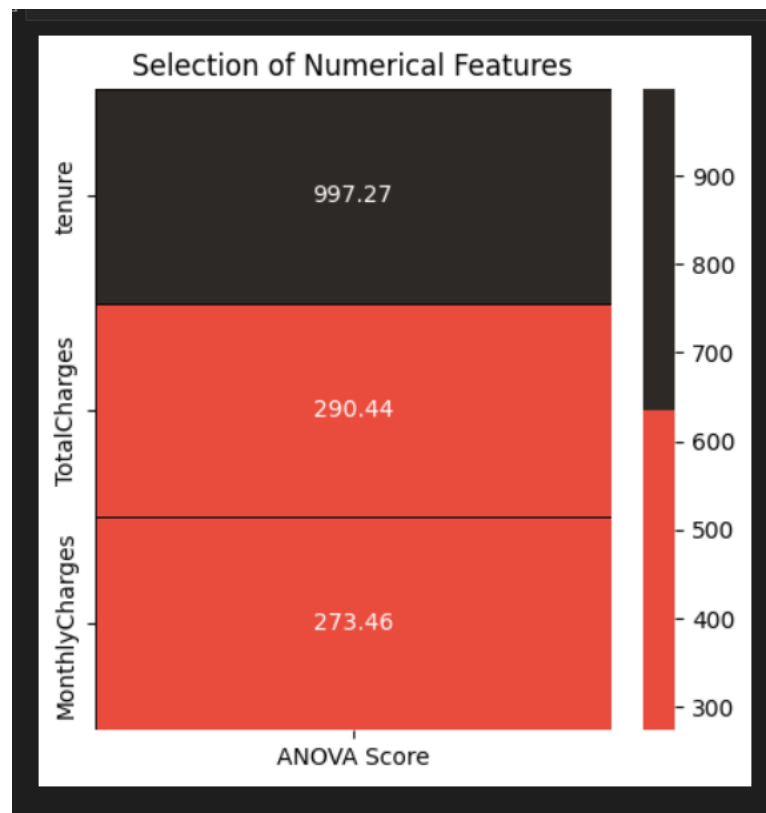


Figure 9. Selection of Numerical Features

7. Evaluation Models

I implement Logistic Regression Classification and I use Decision Tree and Random Forest models from sklearn library. I look also Cross Validation Score and ROC_AUC Score.

Cross-validation is a technique used to evaluate the performance of a machine learning model. It is used to estimate the performance of a model on unseen data. It helps to prevent overfitting by splitting the data into several subsets and training the model on one subset, while evaluating its performance on the other subset.

Cross-validation score is a metric that is used to evaluate the performance of a model using cross-validation. It is the average performance of the model across all the subsets. There are different types of cross-validation methods, such as K-Fold Cross-Validation, Leave-One-Out Cross-Validation, and Stratified Cross-Validation. Each method divides the data into different subsets with different sizes and patterns, and each one has its own advantages and disadvantages. I use RepeatedStratifiedKFold. Like the traditional K-Fold method, RepeatedStratifiedKFold divides the data into K subsets (or "folds") of roughly equal size. However, instead of training the model on K-1 of the folds and evaluating it on the remaining fold, it repeats the process multiple times, each time using a different fold as the evaluation set.

Stratified sampling is used to ensure that each fold contains roughly the same proportions of samples of each class. The final results are then averaged to give an estimate of the model's performance.

RepeatedStratifiedKFold is particularly useful when the dataset is imbalanced, as it can help to reduce the variability of the results by averaging multiple evaluations of the model. It is also useful when the sample size is small, as it helps to reduce the bias of the results by averaging multiple evaluations.

ROC_AUC score is a performance metric for binary classification problems. ROC stands for Receiver Operating Characteristic and AUC stands for Area Under the Curve. The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different threshold values. The AUC is the area under this curve. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

In general, a high ROC_AUC score indicates that the model is able to accurately distinguish between the positive and negative classes. It's a robust metric that takes into account both the true positive rate and the false positive rate, making it a useful metric for imbalanced datasets.

7.1 Decision Tree

A Decision Tree is a tree-like model of decisions and their possible consequences. It is used for both classification and regression tasks. At each internal node of the tree, a decision is made based on the values of one or multiple input features, and the sample is directed to one of the child nodes based on the outcome of that decision. The leaves of the tree represent the predicted class or value for the samples that reach that leaf. Decision Trees are easy to understand and interpret and can handle both categorical and numerical features.

I use Decision Tree from sklearn library.

7.1.1 Decision Tree with No Resampling

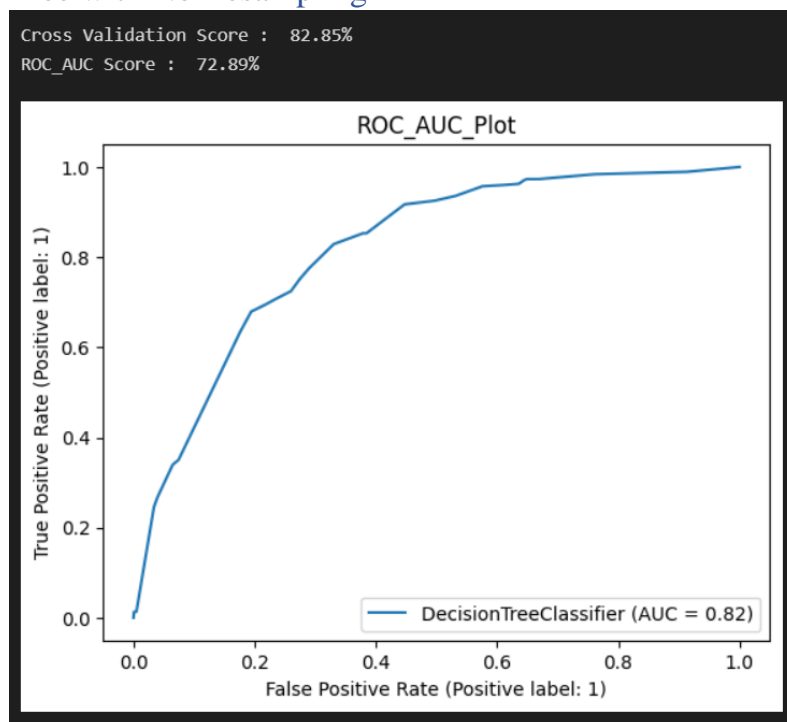


Figure 10. Decision Tree with No Resampling - Cross Validation Score and ROC_AUC Score

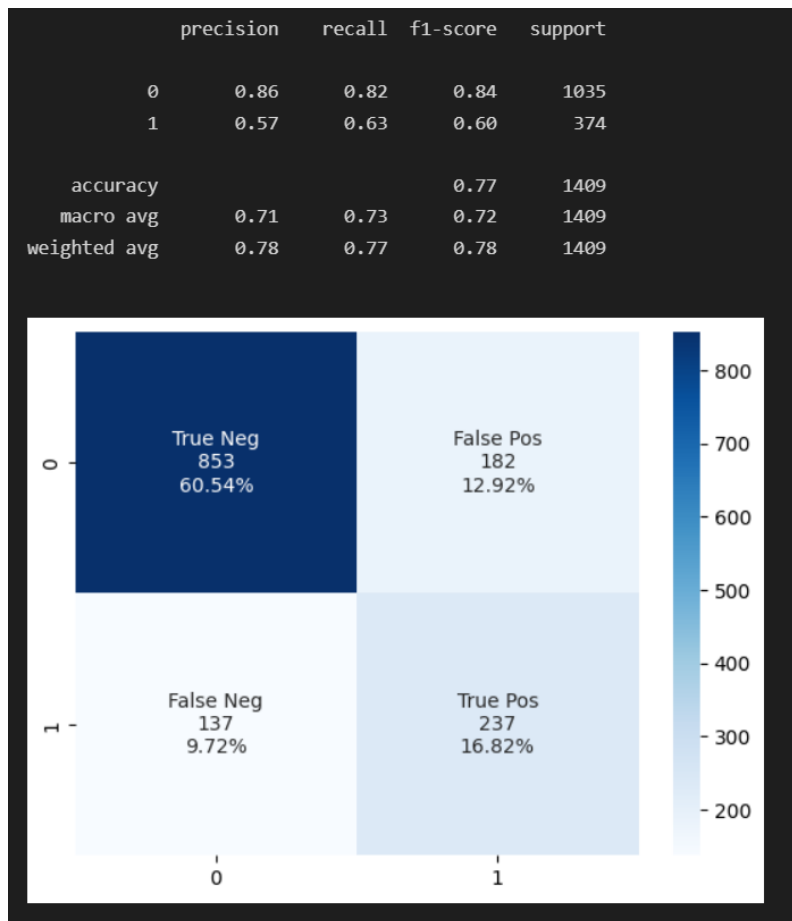


Figure 11. Decision Tree with No Resampling Result

7.1.2 Decision Tree with RandomOverSampler

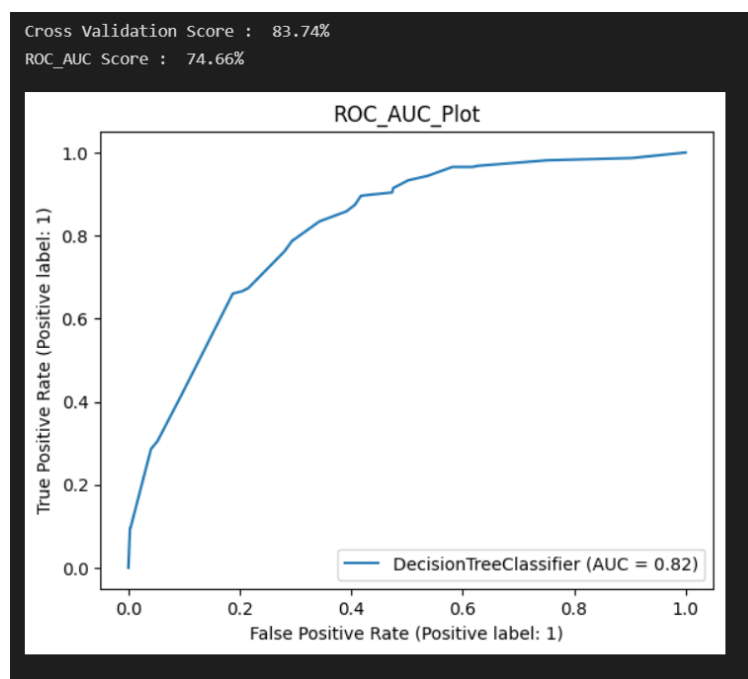


Figure 12. Decision Tree with RandomOverSampler- Cross Validation Score and ROC_AUC Score

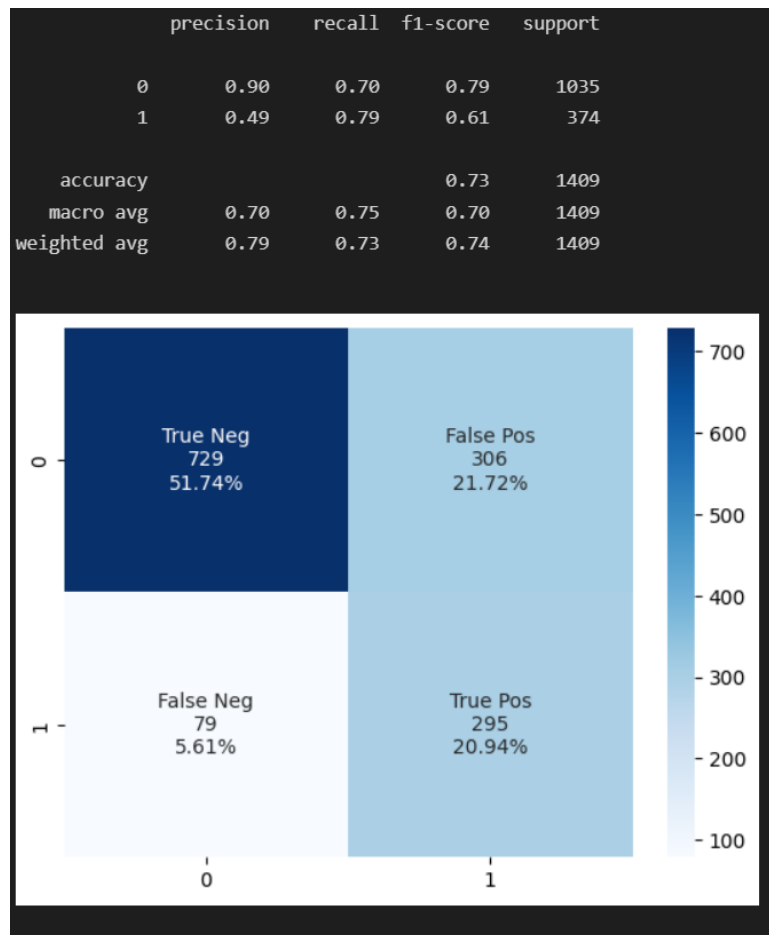


Figure 13. Decision Tree with RandomOverSampler Result

7.1.3 Decision Tree with RandomUnderSampler

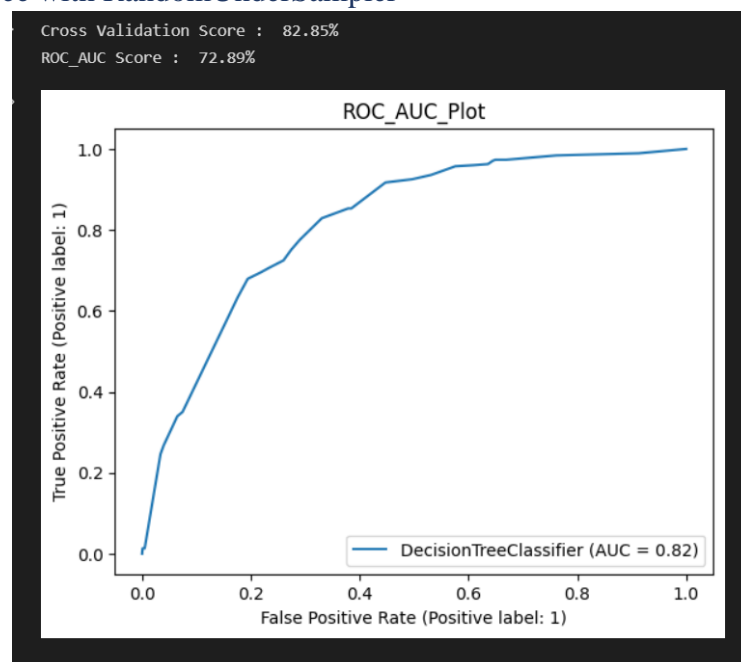


Figure 14. Decision Tree with RandomUnderSampler- Cross Validation Score and ROC_AUC Score

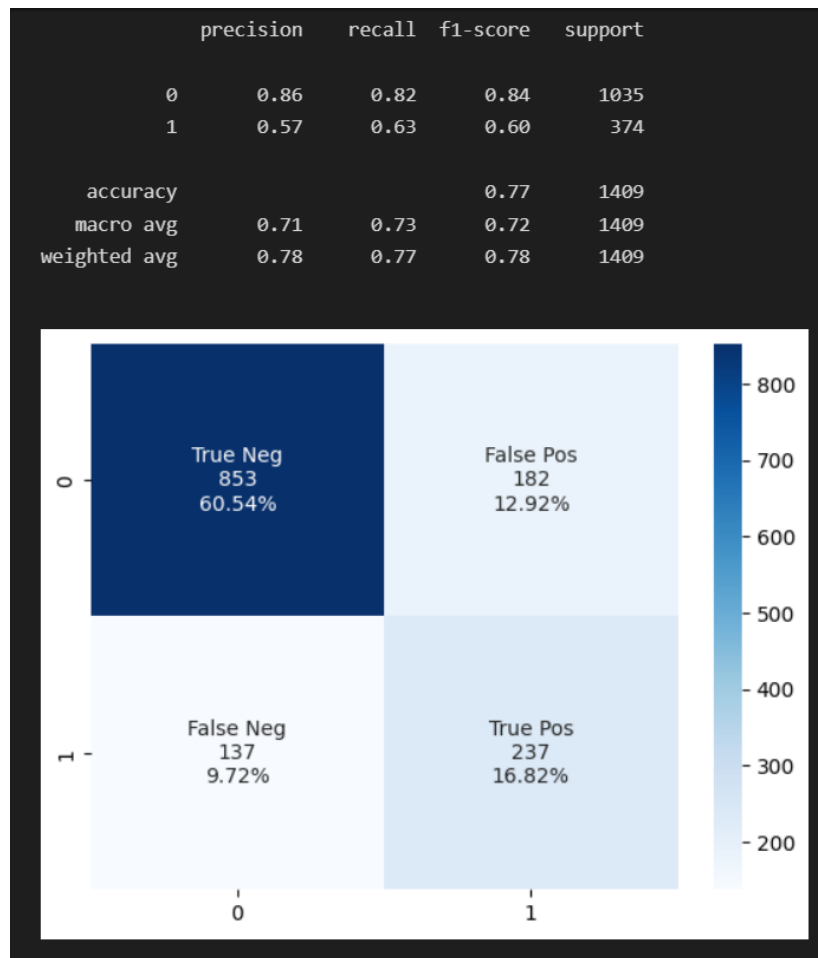


Figure 15. Decision Tree with RandomUnderSampler Result

7.1.4 Conclusion

Decision Tree with random oversampling gives the highest cross validation score and ROC_AUC score.

7.2 Random Forest

Random Forest is an ensemble learning method for classification and regression tasks. It is composed of multiple decision trees and it is a way to average multiple decision trees together to reduce overfitting and increase the overall performance. In a random forest, a new decision tree is created at each iteration, and a random subset of features is selected at each split, which helps to decorrelate the trees and decrease the variance. The final prediction is made by averaging the predictions of all decision trees. Random Forest is a robust and powerful method that can handle large datasets with many features and it's less prone to overfitting than a single decision tree.

I use Random Forest models from sklearn library.

7.2.1 Random Forest with No Resampling

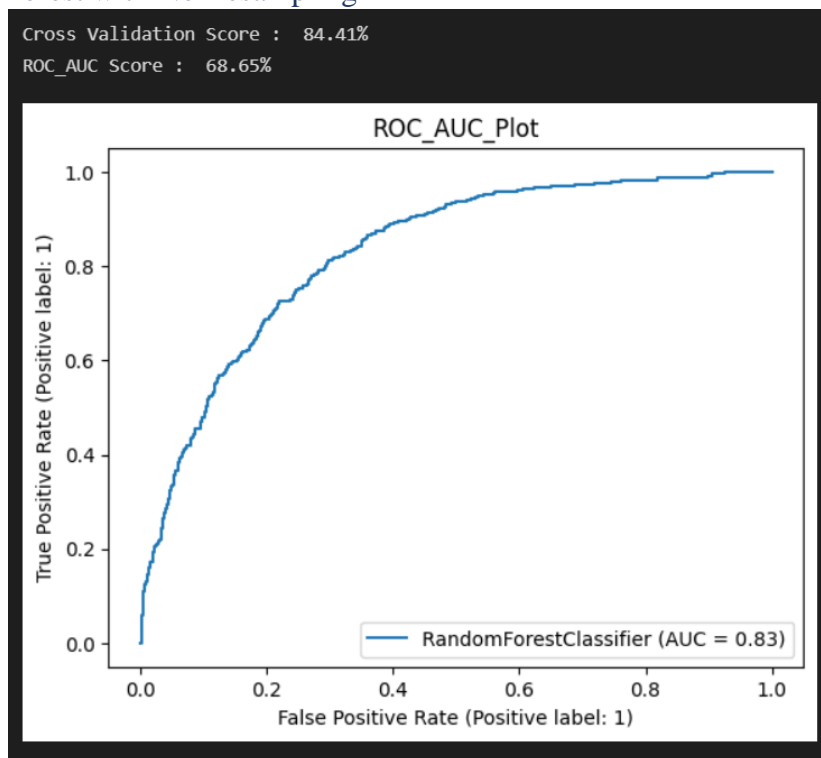


Figure 16. Random Forest with No Resampling- Cross Validation Score and ROC_AUC Score

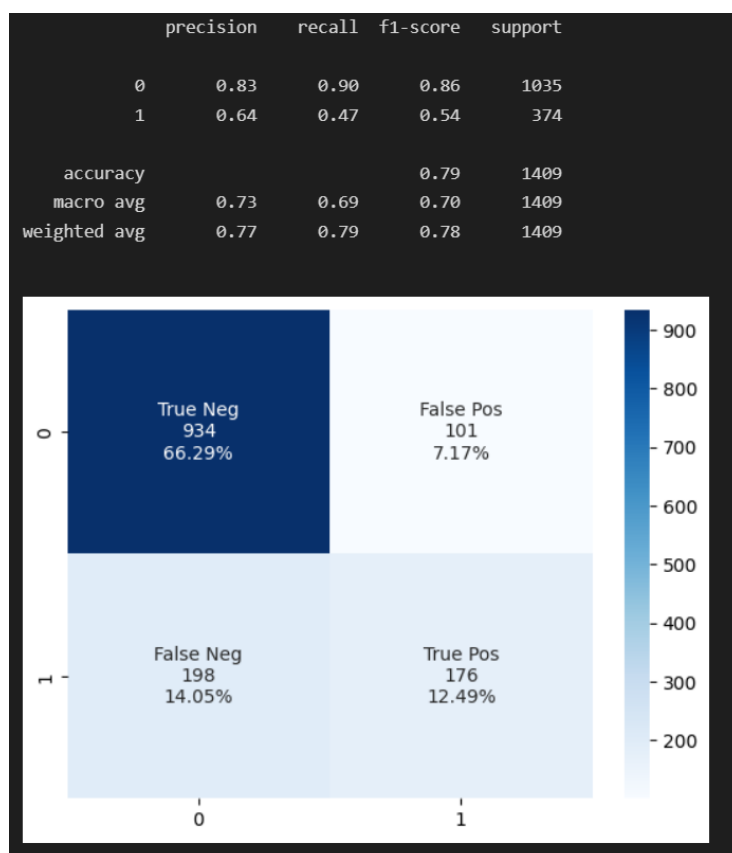


Figure 17. Random Forest with No Resampling Result

7.2.2 Random Forest with RandomOverSampler

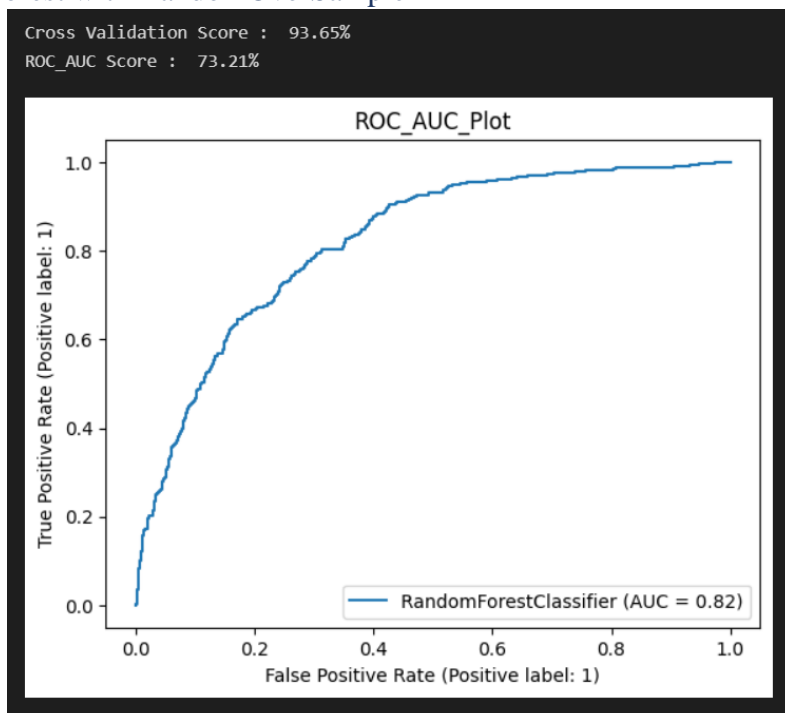


Figure 18. Random Forest with RandomOverSampler - Cross Validation Score and ROC_AUC Score

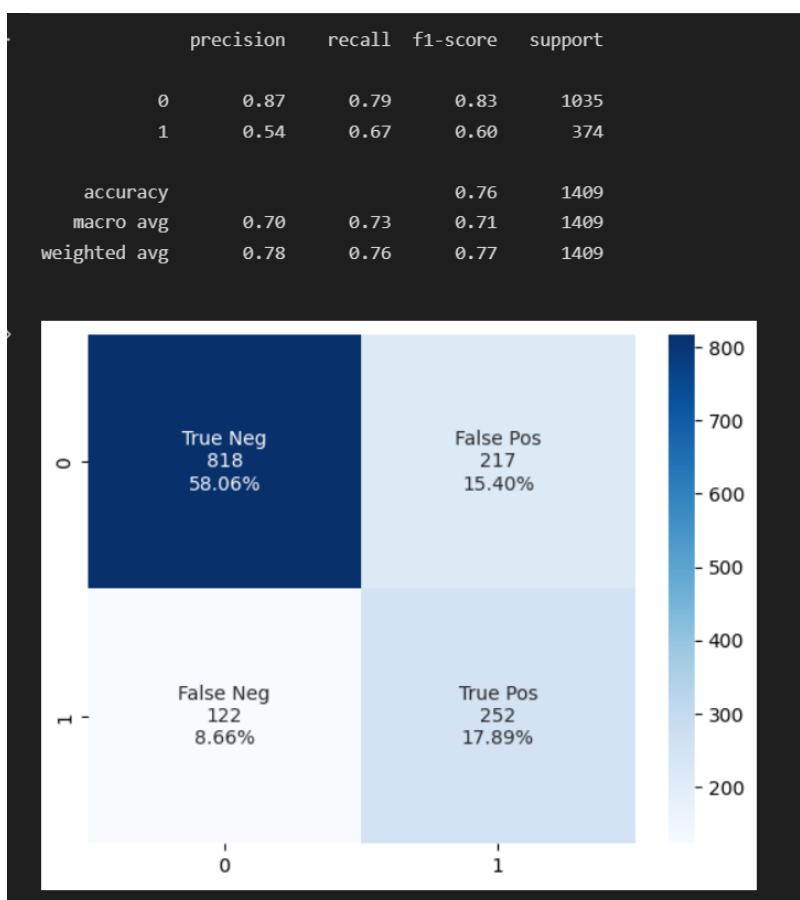


Figure 19. Random Forest with RandomOverSampler Result

7.2.3 Random Forest with RandomUnderSampler

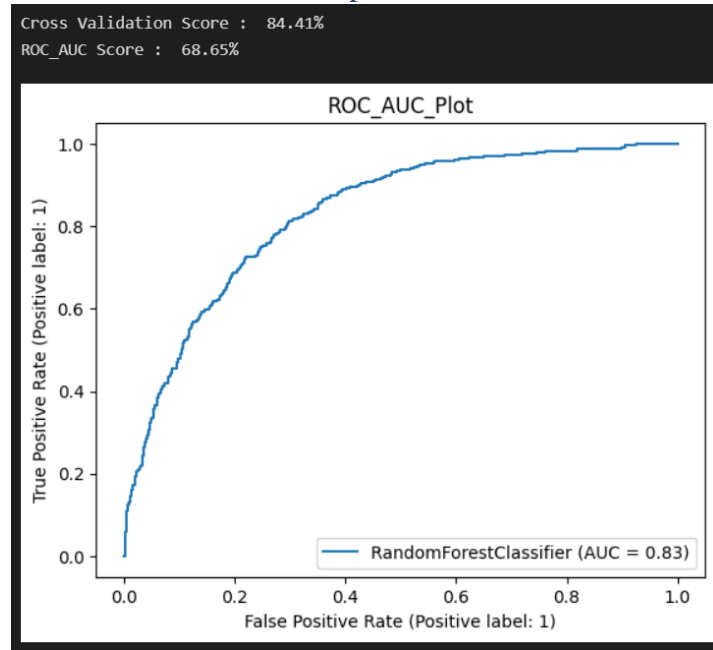


Figure 20. Random Forest with RandomUnderSampler - Cross Validation Score and ROC_AUC Score

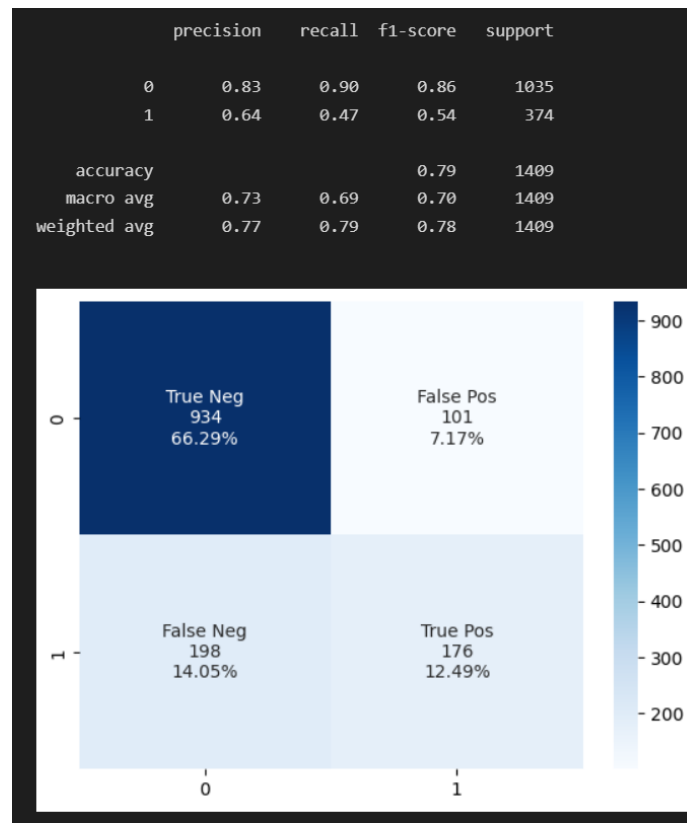


Figure 21. Random Forest with RandomUnderSampler Result

7.2.4 Conclusion

Random Forest with RandomOverSampler has a higher cross validation score of 93.65% and a higher ROC_AUC score of 73.21%.

7.3 Logistic Regression Classification

Logistic Regression Classification is a supervised machine learning technique that is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. Logistic Regression Classification is a simple and efficient method that uses the logistic function (also called the sigmoid function) to model the probability of an instance belonging to a certain class.

The logistic function is defined by a sigmoid function. Sigmoid function takes any real number input and output value between 0 and 1. Large positive values will output 1, and negative numbers will output 0.

Logistic regression includes these:

- Sigmoid Function

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

Figure 22. Sigmoid Function Formula

- Cost function
- Gradient descent function: This function returns the best parameters that minimizes cost function.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x), y)$$

$$\text{Cost}(h_{\theta}(\mathbf{x}), y) = -y \log(h_{\theta}(\mathbf{x})) - (1 - y) \log(1 - h_{\theta}(\mathbf{x}))$$

Gradient descent for logistic regression:

$$\begin{aligned} &\text{while not converged } \{ \\ &\quad \theta_j^{\text{new}} = \theta_j^{\text{old}} - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \text{ for } j = 0, 1, \dots, n \\ &\} \end{aligned}$$

Figure 23. Cost Function and Gradient Descent Function Formulas

- Predict Function: For predicting churn. When our sigmoid function outputs a value more than or equal to 0.5, it will output 1 for "Yes", and if the value is less than 0.5, it will output "0" for "No".
- Theta: I used scipy.optimize library for finding optimum parameters with fmin_bfgs()

```
#initializing theta
initial_theta = np.random.rand(x_train_lr.shape[1], 1)
|
optimized_theta = scipy.optimize.fmin_bfgs(costFunction, fprime = gradDescent, x0 = initial_theta, args=(x_train_lr, y_train_lr.ravel(), 1))
optimized_theta
```

✓ 225

Figure 24. Theta

7.3.1 Logistic Regression Classification with No Resampling

	precision	recall	f1-score	support
0	0.85	0.91	0.88	4139
1	0.68	0.56	0.62	1495
accuracy			0.81	5634
macro avg	0.77	0.73	0.75	5634
weighted avg	0.81	0.81	0.81	5634
ROC_AUC Score : 73.29%				

Figure 25. Logistic Regression Classification with No-Resampling Result

7.3.2 Logistic Regression Classification with RandomOverSampler

ROC_AUC Score : 77.92%				
	precision	recall	f1-score	support
0	0.81	0.73	0.77	4139
1	0.75	0.83	0.79	4139
accuracy			0.78	8278
macro avg	0.78	0.78	0.78	8278
weighted avg	0.78	0.78	0.78	8278

Figure 26. Logistic Regression Classification with RandomOverSampler Result

7.3.3 Logistic Regression Classification with RandomUnderSampler

ROC_AUC Score : 73.25%				
	precision	recall	f1-score	support
0	0.85	0.91	0.88	4139
1	0.68	0.56	0.61	1495
accuracy			0.81	5634
macro avg	0.77	0.73	0.75	5634
weighted avg	0.81	0.81	0.81	5634

Figure 27. Logistic Regression Classification with RandomUnderSampler Result

7.3.4 Conclusion

The ROC_AUC scores show that the random over sampler has the highest ROC_AUC score of 77.92%, followed by the random under sampler with 73.25% and without resampling the score is 73.29%. The results show that oversampling improves the ROC_AUC score and the precision, recall and f1-score of class 1.

8. Academic Paper Summary

In that time the use of logistic regression has increased in the social sciences. Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. Logistic regression is an alternative method to use other than the simpler linear regression.

Although logistic regression can do categorical outcomes that are polytomous, In this article they focus on dichotomous outcomes only. Logistic regression uses the concept of odds ratios to calculate the probability.

The simple logistic model has the form:

$$\text{logit}(Y) = \text{natural log(odds)} = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X.$$

Figure 28. Simple Logistic Model

For calculating effectiveness of the model we should do:

- overall model evaluation,
- statistical tests of individual predictors,
- goodness-of-fit statistics and
- validations of predicted probabilities.

The result gives an 'S' shaped curve to model the data.

9. References

- Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14.
10.1080/00220670209598786.