Åbo Akademi
University

Machine Learning 2024

Mini Project 1

Banking Campaign Output Prediction

Project Report

Author: Halidu Abdulai

Student ID: 2301998

Contents

1 Introduction

In the context of the banking sector, where competition is fierce and customer loyalty is paramount, direct marketing campaigns have emerged as a fundamental strategy for banks to not only expand their customer base but also to improve the uptake of various banking services. Specifically, this project report focuses on an in-depth analysis of a particular bank's marketing efforts that utilize phone calls to encourage customers to subscribe to term deposits. The core of our study is a dataset that includes a wide range of information on clients, such as demographic characteristics, economic indicators, and the results of these marketing campaigns.

The primary aim is to apply sophisticated Machine Learning (ML) algorithms to predict the outcomes of these campaigns accurately. Additionally, the project seeks to extract valuable insights that could inform and optimize future marketing strategies. By conducting thorough data preprocessing, employing various ML models, and aiming for a high benchmark of predictive accuracy, this report endeavors to provide a detailed comparison of the effectiveness of different ML techniques in a real-world banking scenario.

2 Data Preprocessing and Exploration

The initial phase of the project involved a meticulous process of preparing the dataset for analysis, which is crucial for the success of any ML endeavor. This section outlines the steps taken to preprocess the data and the findings from the exploratory analysis.

2.1 Exploratory Data Analysis (EDA)

The first step involved a detailed examination of the dataset to gain insights into its structure and content. This phase revealed a significant imbalance in the campaign outcomes, with a large majority of the contacted individuals (88.7%) not subscribing to the term deposit. This initial finding pointed to a potential misalignment in the targeting strategy of the bank's marketing campaign. Further analysis of the categorical variables in the dataset highlighted that the majority of the contacted individuals were employed in specific sectors (such as administration, blue-collar jobs, technical roles, or services) and were predominantly married with a basic level of education. Interestingly, the analysis also showed that individuals without credit defaults or personal loans were more likely to subscribe to the term deposit, with the highest subscription rates occurring in May and among individuals aged between 20 and 60.

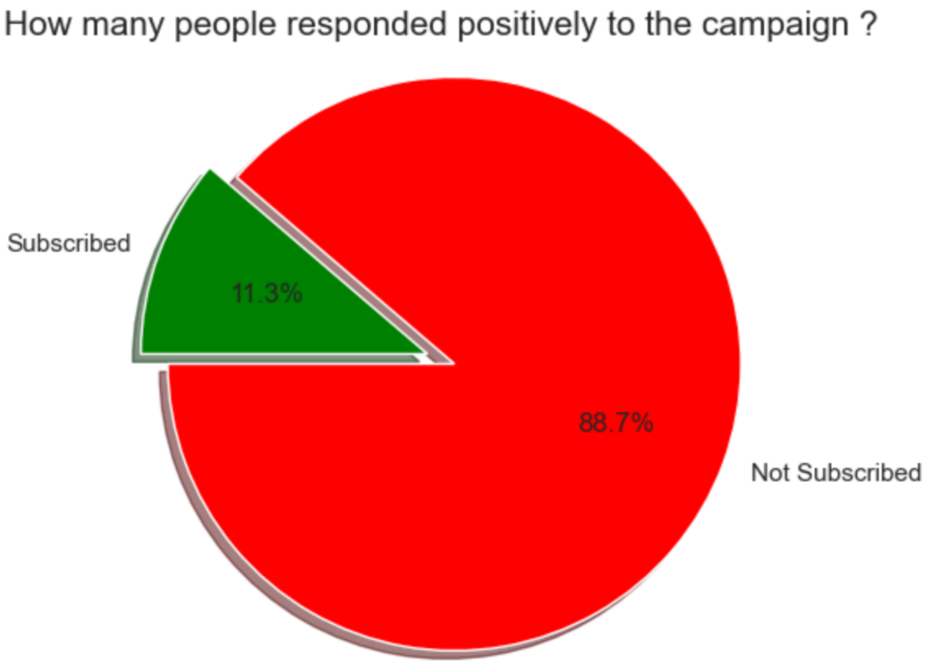


Figure 1: A pie chart of the percentage of contacted individuals who responded positvely to the campaign.

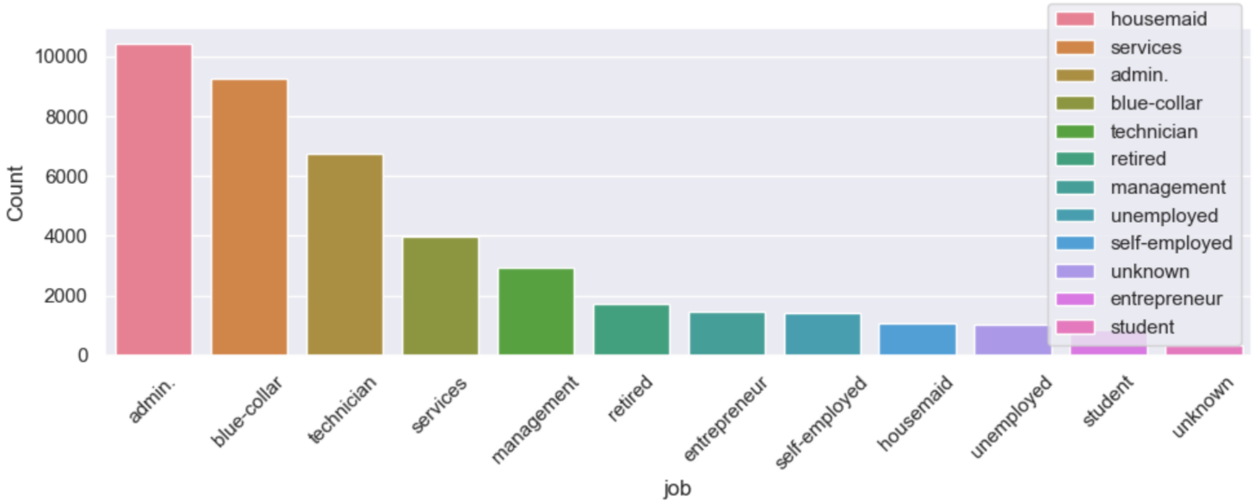


Figure 2: A bar chart of the profession of the contacted individuals.

2.2 Preprocessing and Feature Selection

Following the EDA, the next step was to prepare the categorical variables through one-hot encoding, integrating them with the numerical features to form a comprehensive dataset ready for ML analysis. However, this process resulted in a dataset with high dimensionality (53 features), necessitating the implementation of feature selection techniques to identify the most informative features for the ML models. Utilizing an embedded feature selection method via a random forest model, the dataset was effectively reduced to 35 key features, streamlining the data for more efficient analysis.

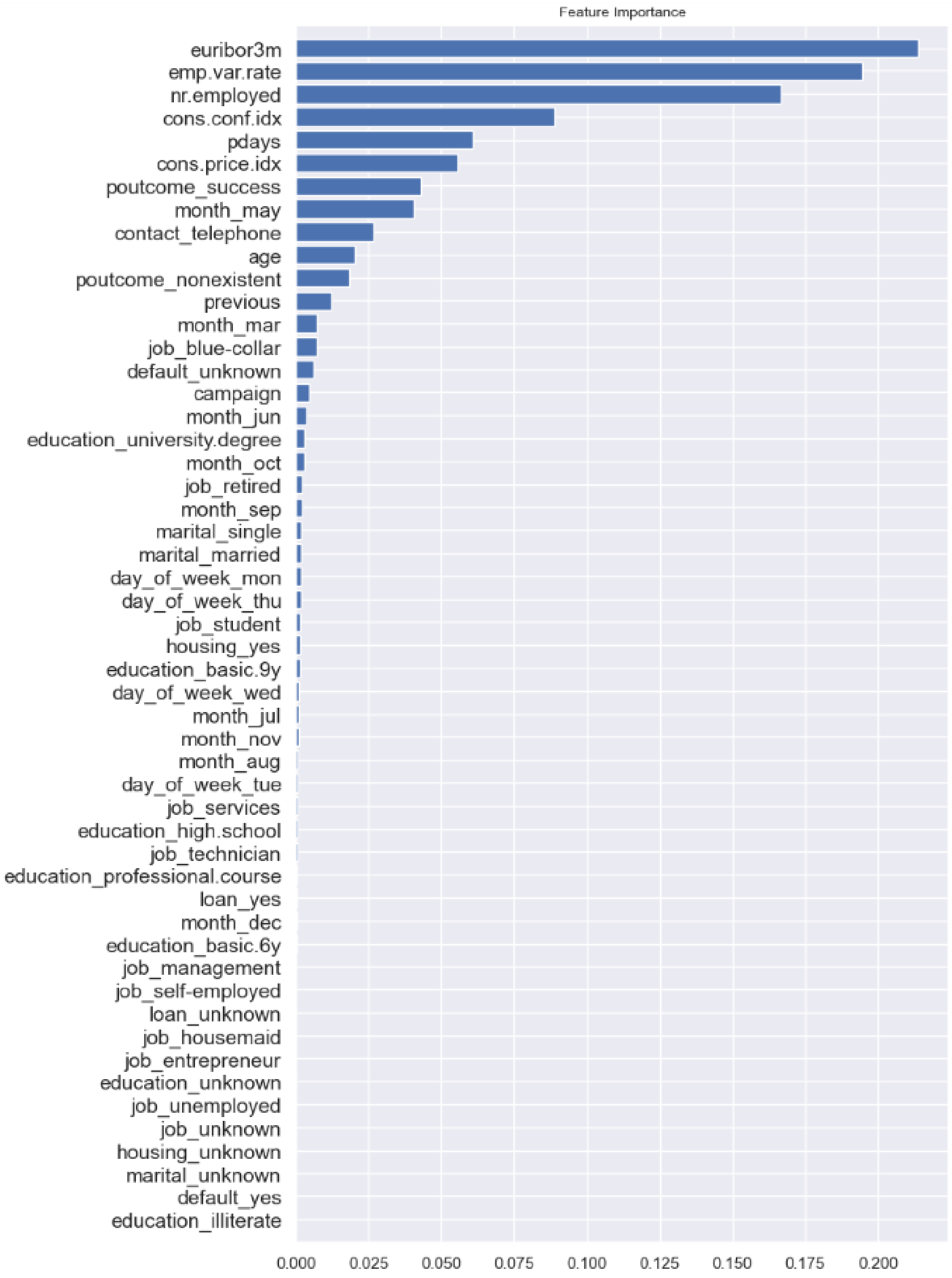
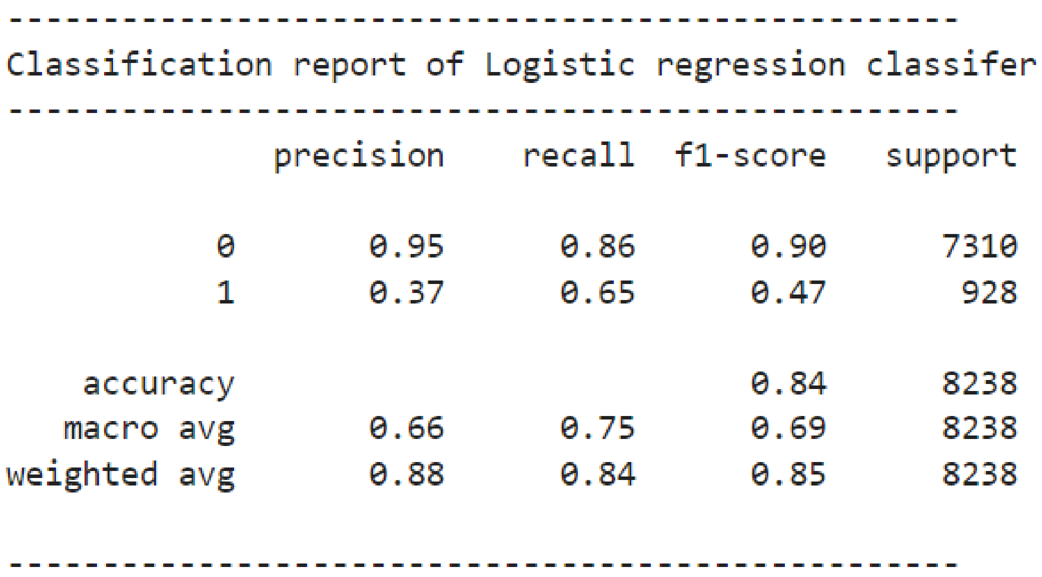


Figure 3: Utilizing a Random Forest model to select the most informative features.

3 Modeling

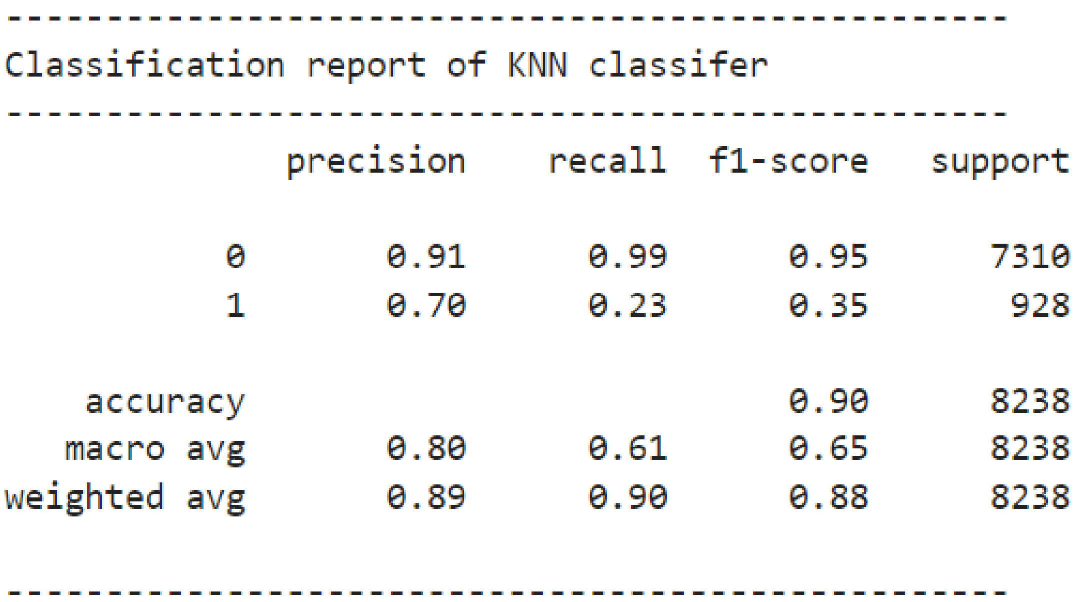
The choice of ML algorithms was guided by a combination of factors, including simplicity of implementation, historical performance in similar tasks, and the ability to handle imbalanced datasets. The selected algorithms included Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and XGBOOST classifiers. Each model was evaluated based on its ability to achieve a minimum accuracy of 80% on the test dataset. The decision to employ classical machine learning models in tackling this problem was due to the fact that the more powerful neural networks often require hundreds of thousands of training samples to thrive. Looking at the dataset at hand, it might not be sufficient to train neural networks hence the decision to go with classical machine learning algorithms.

3.1 Model Performance Evaluation



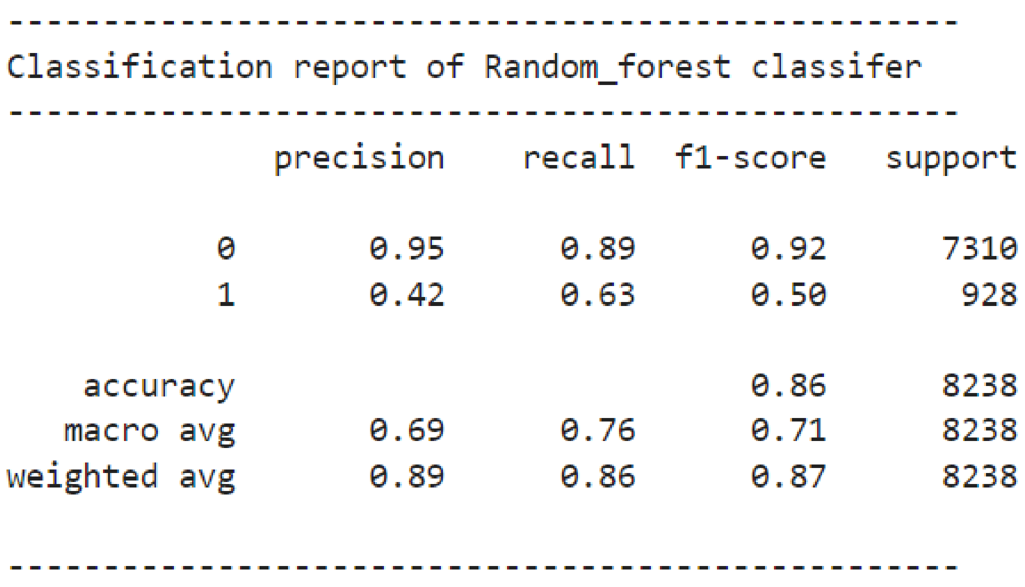
```
------------------------------------------------
Classification report of Logistic regression classifer
------------------------------------------------
              precision    recall  f1-score   support

           0       0.95      0.86      0.90      7310
           1       0.37      0.65      0.47       928

    accuracy                           0.84      8238
   macro avg       0.66      0.75      0.69      8238
weighted avg       0.88      0.84      0.85      8238

------------------------------------------------
```

Figure 4: Classification report of Logistic regression classifer.



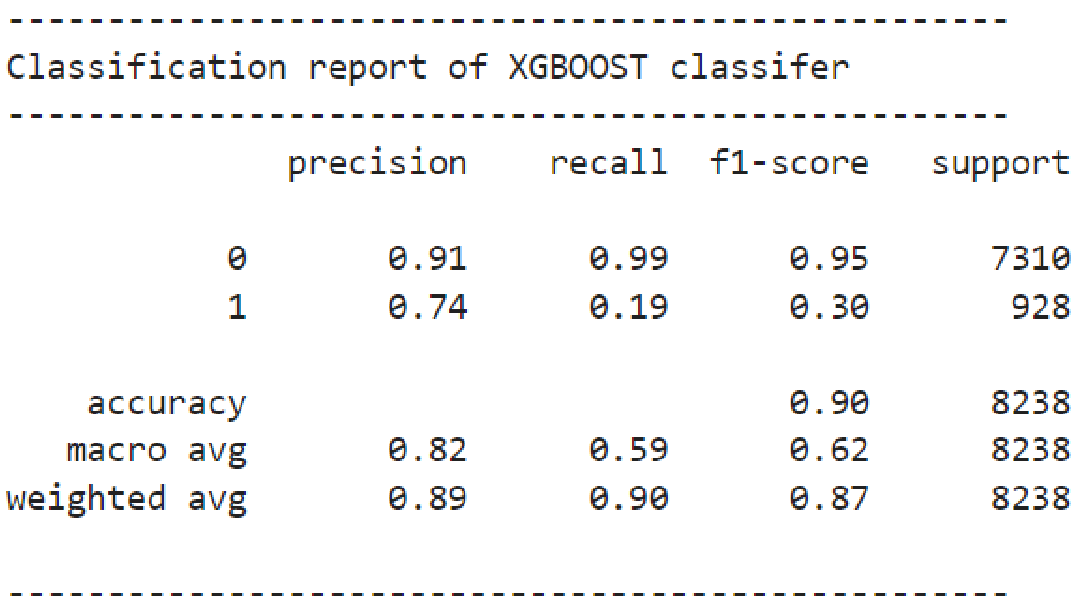
```
--------------------------------------------------
Classification report of KNN classifer
--------------------------------------------------
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      7310
           1       0.70      0.23      0.35       928

    accuracy                           0.90      8238
   macro avg       0.80      0.61      0.65      8238
weighted avg       0.89      0.90      0.88      8238

--------------------------------------------------
```

Figure 5: Classification report of KNN classifer.



```
--------------------------------------------------
Classification report of Random_forest classifer
--------------------------------------------------
              precision    recall  f1-score   support

           0       0.95      0.89      0.92      7310
           1       0.42      0.63      0.50       928

    accuracy                           0.86      8238
   macro avg       0.69      0.76      0.71      8238
weighted avg       0.89      0.86      0.87      8238

--------------------------------------------------
```

Figure 6: Classification report of Random Forest classifer.



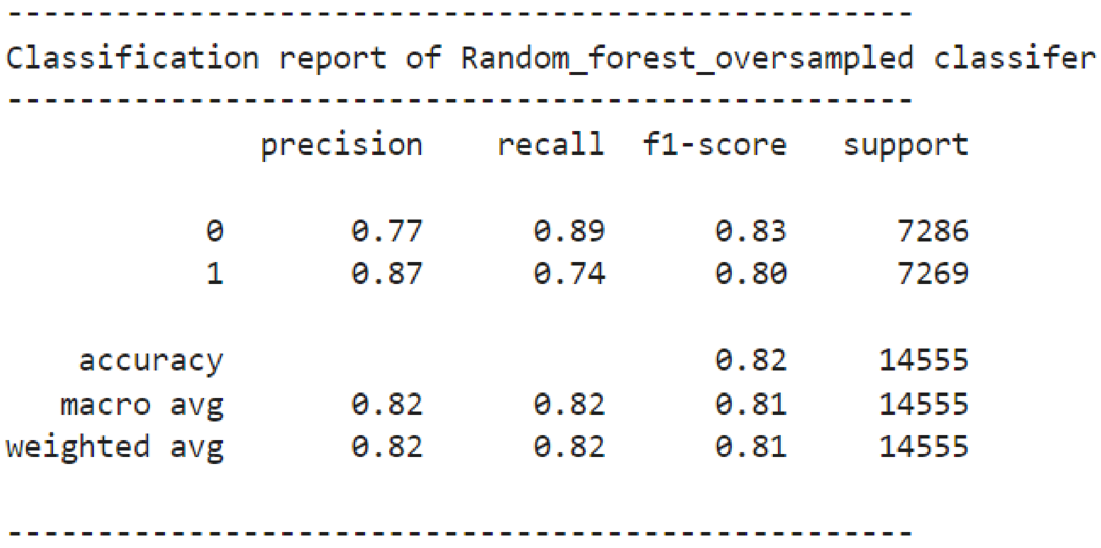
```
--------------------------------------------------
Classification report of XGBOOST classifer
--------------------------------------------------
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      7310
           1       0.74      0.19      0.30       928

    accuracy                           0.90      8238
   macro avg       0.82      0.59      0.62      8238
weighted avg       0.89      0.90      0.87      8238

--------------------------------------------------
```

Figure 7: Classification report of XGBOOST classifer.

Despite all models meeting the accuracy benchmark, the imbalanced nature of the dataset necessitated a re-evaluation of success metrics, with precision emerging as a more critical measure of model performance. This shift in focus was due to the recognition that accuracy alone might not adequately reflect the model's ability to identify potential subscribers accurately.

3.2 Improvement Strategies

To address the imbalance issue and improve model precision, an oversampling technique known as SMOTETomek was employed. This approach significantly enhanced the precision scores of the models, particularly for the random forest classifier, highlighting the effectiveness of addressing class imbalance in improving predictive performance. The improved precision score of the Random Forest classifier is shown in the figure below.

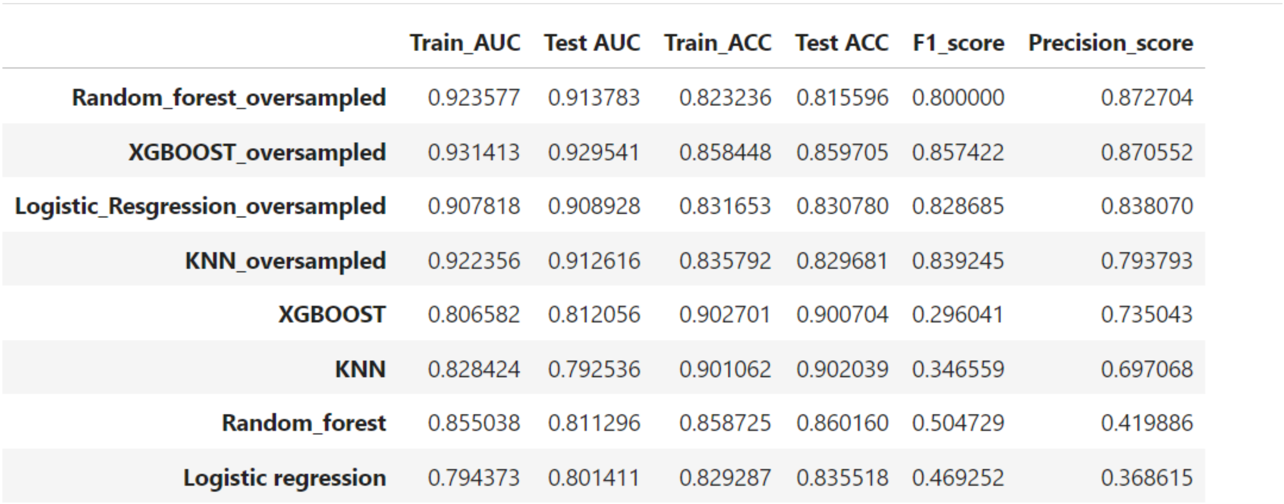


```
-------------------------------------------------
Classification report of Random_forest_oversampled classifer
-------------------------------------------------
              precision    recall  f1-score   support

           0       0.77      0.89      0.83      7286
           1       0.87      0.74      0.80      7269

    accuracy                           0.82     14555
   macro avg       0.82      0.82      0.81     14555
weighted avg       0.82      0.82      0.81     14555

-------------------------------------------------
```

Figure 8: Classification report of the Random Forest classifer trained on the oversampled dataset.

From the rersults above, even though the accuracy score dropped to 82%, the precision score which is of interest has increased significantly. A performance which is highly desirable.

4 Analysis of the results



	Train_AUC	Test AUC	Train_ACC	Test ACC	F1_score	Precision_score
Random_forest_oversampled	0.923577	0.913783	0.823236	0.815596	0.800000	0.872704
XGBOOST_oversampled	0.931413	0.929541	0.858448	0.859705	0.857422	0.870552
Logistic_Resgression_oversampled	0.907818	0.908928	0.831653	0.830780	0.828685	0.838070
KNN_oversampled	0.922356	0.912616	0.835792	0.829681	0.839245	0.793793
XGBOOST	0.806582	0.812056	0.902701	0.900704	0.296041	0.735043
KNN	0.828424	0.792536	0.901062	0.902039	0.346559	0.697068
Random_forest	0.855038	0.811296	0.858725	0.860160	0.504729	0.419886
Logistic regression	0.794373	0.801411	0.829287	0.835518	0.469252	0.368615

Figure 9: Comparison of the performance of the ML models sorted based on precision.

The results demonstrated that, while all models achieved satisfactory accuracy on the original dataset, their precision scores were initially low. The implementation of oversampling markedly improved these scores, with the random forest classifier achieving the highest precision, followed closely by the XGBOOST classifier. The analysis conclusively showed that precision is a more relevant metric than accuracy for evaluating the success of marketing campaign predictions, especially in the context of an imbalanced dataset.

5 Conclusion & Remarks

In conclusion, the chosen algorithms for predicting the outcome of the marketing campaign showed great performance. For the reasons explained earlier, the accuracy wasn't a good metric for this task and for that reason, the evaluation of the performance of the chosen algorithms were based on the precision score. The random forest algorithm had the best precision score when oversampling was performed (87.2%). The XGBOOST classifier followed closely with a precision score of about 87%. The KNN classifier had the best accuracy score (about 90% on the test data) followed by the XGBOOST (about 90%).

The imbalanced dataset posed a challenge on this project. To overcome this issue, oversampling of the dataset was performed to have the same distribution of the two classes. This technique helped improve the precision score of all the four chosen algorithms. Overall, the tree-based algorithms had a better performance compared to the Logistic Regression and KNN classifiers. In all, the Random Forest classifier emerged as the winner using precision as the score metric. In a future work, a different feature selection technique could be employed to determine the most informative features.