Assessment task 3: Exploration of data skills and issues

Course Title: Statistical Thinking for Data Science

Student Name: Yasaman Mohammadi
Student ID:24612626
Date of Submission: 11.59pm Sunday 5 of November

1 Introduction

1.1 Problem Statement

A telecommunications company has recently launched a marketing campaign to promote their new subscription plan among their customers. They are now seeking assistance in gaining a deeper understanding of their customer base and identifying the segments most receptive to their marketing efforts.

1.2 Rationale

In a rapidly changing industry with evolving customer preferences, this approach allows the company to optimize resources, improve ROI, and adapt to dynamic market conditions, ultimately fostering long-term growth and competitiveness.

1.3 Project aims and Objectives

This project aims to develop data science models to answer two vital business questions: first, identifying the customer segments with the highest response to marketing campaigns, and second, formulating effective business strategies based on these insights.

1.4 Research questions

Having previously conducted exploratory data analysis (EDA) for this project, we have uncovered valuable insights regarding potential influential factors for campaign success. These insights have paved the way for the research questions presented below:

- ***Analysing the Influence of Seasonal Customer Behaviour on Marketing Campaign Effectiveness***

  Does seasonality in customer behaviour (e.g., purchasing decisions) impact the marketing campaign's effectiveness?

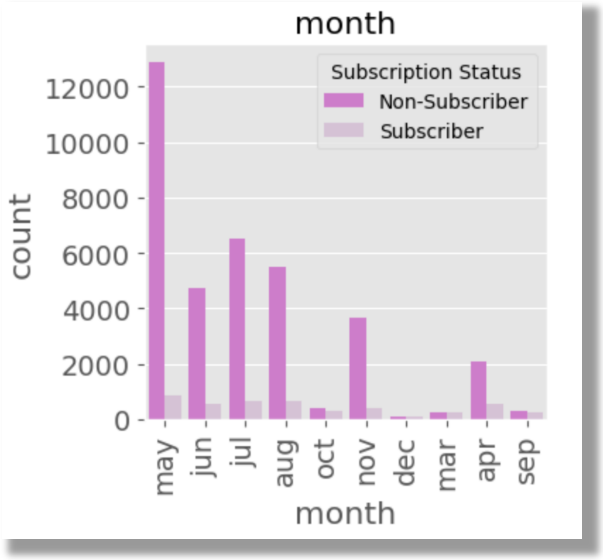


Figure 1| The bar chart illustrates the distribution of subscribers and non-subscribers across different months.

It appears that certain months yield less favourable results compared to others, suggesting the possibility of a variation in performance across different months. (Figure 1) This observation warrants further in-depth analysis to explore and understand any disparities associated with specific months.

- ***Exploring the Relationship Between Marital Status and the Marketing Campaign***

  Does the marketing campaign need to be tailored differently for customers with various marital statuses?

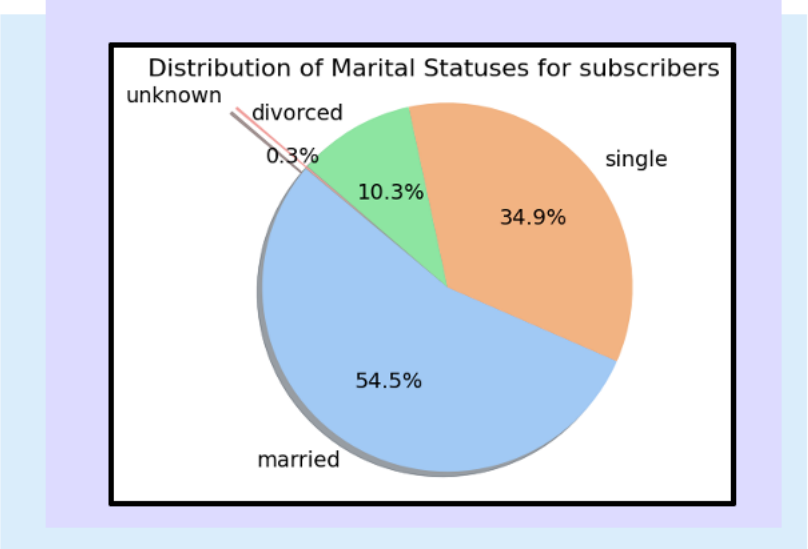


Figure 2 | The pie chart showcases the marital status distribution among subscribers.

The pie chart depicting subscribers to the marketing campaign suggests a potential preference for married individuals when compared to single and divorced individuals in terms of subscription. (Figure 2)

- ***Number of Days Since Last Contact from Previous Campaign***

How does the effectiveness of the marketing campaign vary for customers who were contacted recently versus those who were not contacted for an extended period?

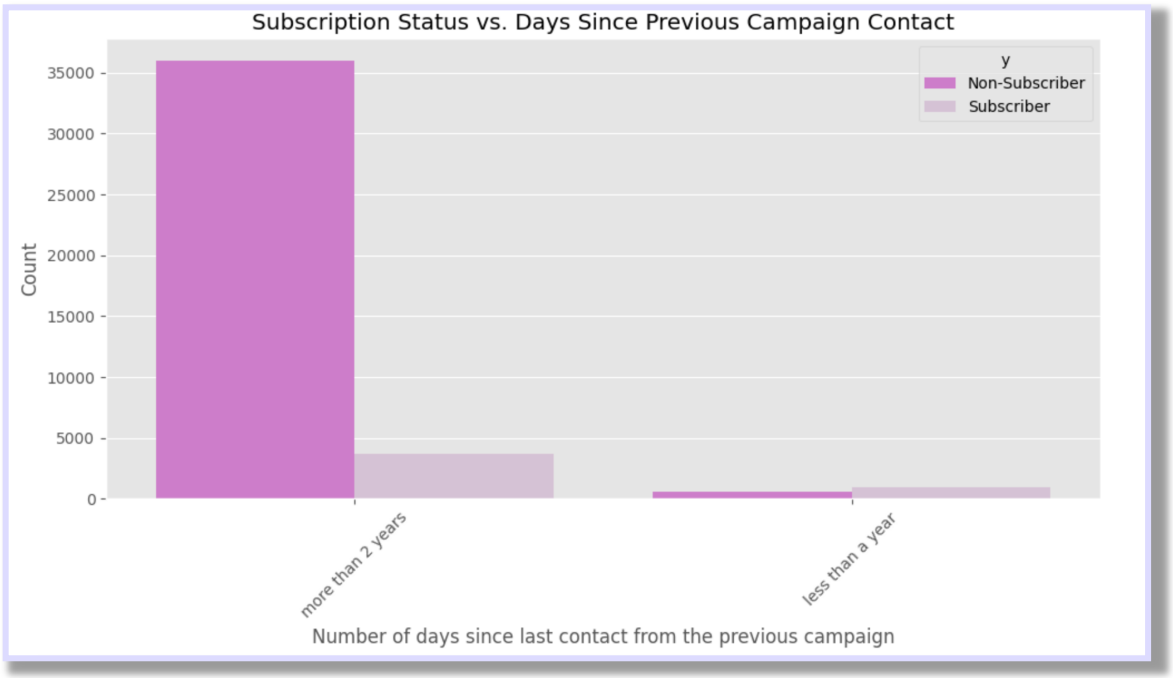


Figure 3| The bar chart illustrates the relationship between subscription status and the days since the previous campaign contact.

- ***Effect of Workforce Size on Campaign Success***

Does the quarterly variation in the number of employees ("nr.employed") impact the success of the marketing campaign in terms of customer subscriptions?

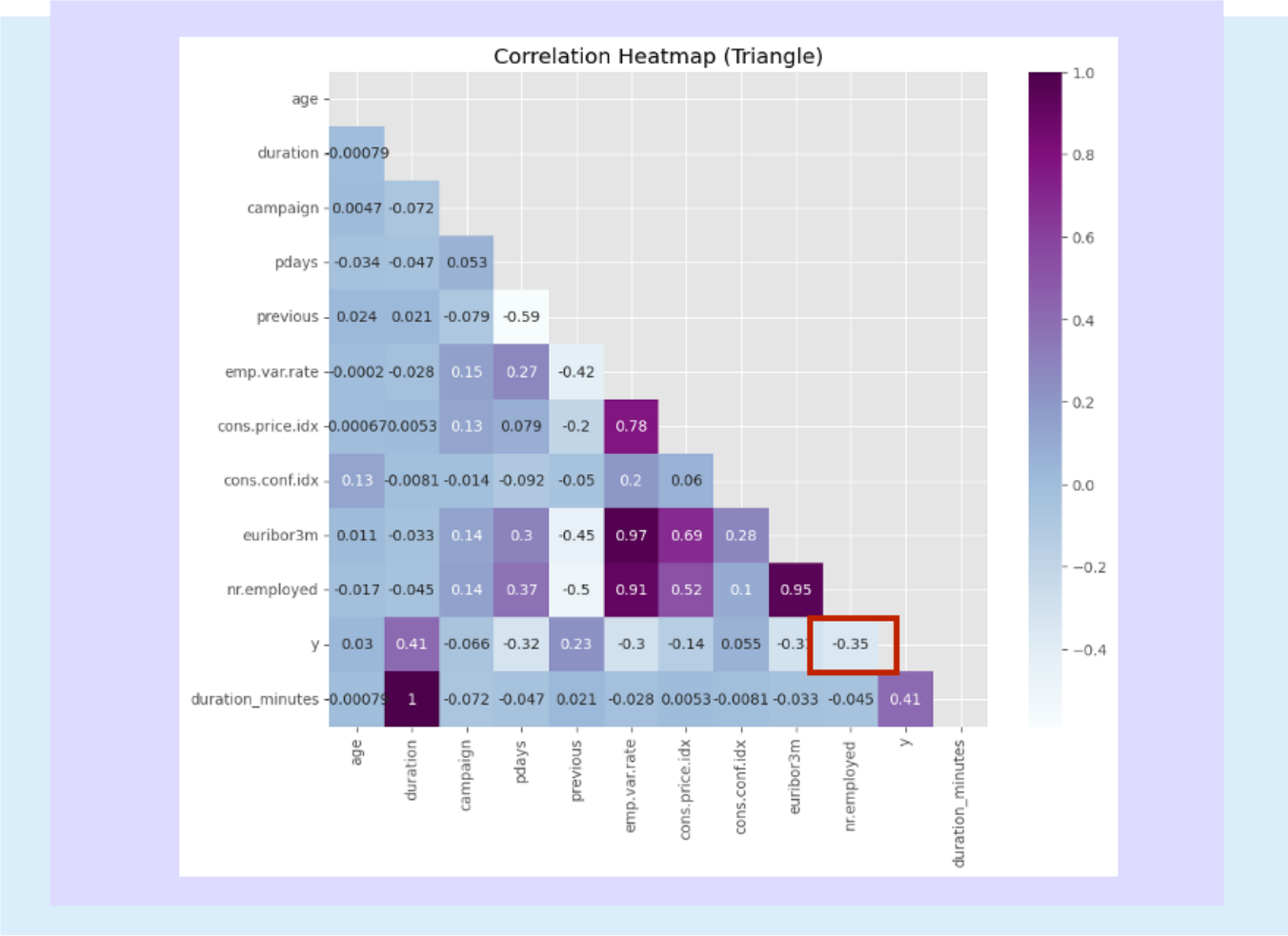


Figure 4|This heatmap represents the correlation between features and the target variable.

Based on the correlation heatmap above (Figure 4), it appears that there could be a connection between workforce size and subscription rates.

2 Methodology

2.1 Methods Overview

The Bayesian estimation was used for hyperparameter tuning to optimize model performance. The model chosen for this project was logistic regression, GBM (Gradient Boosting Machine), and KNN (K-Nearest Neighbors). The SMOTE technique was applied to address the imbalanced class.

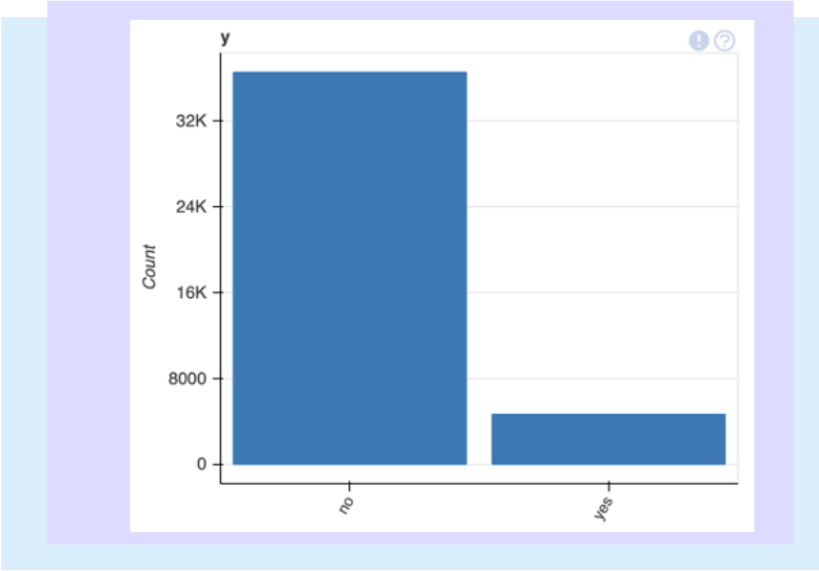


Figure 5|This bar chart illustrates the imbalanced dataset, with the majority belonging to the non-subscriber's category.

2.1 Methods details

Table 1| Methods details

	Method	Explanation
1	Feature engineering	To assess the impact of categorical values on the success of the marketing campaign, we employed feature engineering techniques. One-hot-encoding was applied to the 'marital,' 'job,' 'default,' 'housing,' 'loan,' 'contact,' and 'outcome' columns, while label encoding was utilized for the 'education,' 'month,' and 'day_of_week' columns due to their inherent ordinal sequences. This pre-processing approach allowed us to represent and incorporate categorical information (see appendix 1) into our analysis effectively.
2	Bayesian estimation	Utilizing Bayesian optimization for hyperparameter tuning streamlines finding an optimal parameter set, leading to significant time savings. This approach enhances the model's performance in generalizing to the test set by leveraging insights from past hyperparameter combinations to inform the selection of the next set to evaluate. The dataset at hand comprises 41,180 records with 21 features. Given our choice of models, which includes logistic regression, GBM, and KNN, and the modest size of the hyperparameter space explored for each model, we found that we had sufficient time and computational resources to train and evaluate these models effectively.
3	Scaling	Scaling was applied to the dataset to standardize the scale of all features. This step is crucial because specific machine learning algorithms are responsive to the input feature scaling.
4	Over sampling (SMOTE)	SMOTE, an oversampling technique, generates synthetic samples for the minority class. Given that our dataset predominantly includes non-subscribers (Figure 5), it is a worthwhile approach to balance the class distribution. However, it is essential to note that using SMOTE has its drawbacks. It may introduce instances in noisy and overlapping areas, which can be far from well-defined regions.
5	Cross-validation	K-fold cross-validation is a method used to assess predictive models. In this technique, the dataset is partitioned into 5 folds. The model is trained and assessed 5 times, with a distinct fold serving as the validation set in each iteration. The performance metrics obtained from each fold are then averaged to gauge the model's generalisation ability.

3 Results

	Classifier	f1 Test	f1 Train
0	KNN	0.548810	0.680041
1	Decision Tree	0.490185	0.967599
2	GBM	0.599639	0.607881
3	Random Forest	0.545351	0.967777
4	Logistic Regression	0.509165	0.487341

Figure 6| F1 score for different classifiers.

Five models were initially trained without hyperparameter tuning to assess their overall performance. The results reveal an interesting pattern: Decision Tree and Random Forest models exhibit strong F1 scores on the training set, but this high performance does not carry over to the test set. This discrepancy suggests that these models overfit the training data, capturing noise and randomness that hinder their generalization to unseen data. Additionally, given the dataset's imbalanced class, the models may excel in the majority class but struggle with the minority class.
The next step was to use the SMOTE technique and Bayesian estimation to train KNN, GBM and Logistic Regression.

3.1 Key findings

- ***F1 score for Logistic Regression (Parametric model) with oversampling and hyperparameter tuning with Bayesian method was 58%***

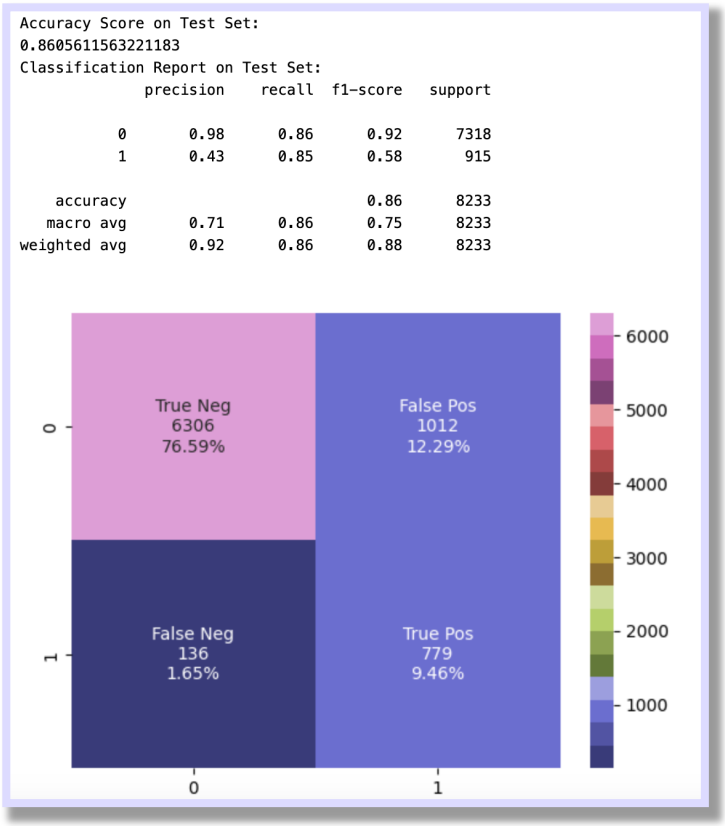


Figure 7| Confusion matrix for the Logistic Regression model

According to a classification report, the Logistic Regression model has an overall accuracy of 86%. The results indicate that it performs well when classifying instances into two groups. In contrast, the recall for the positive class (1) is slightly lower than that for the negative class (0), indicating that the model is less effective in identifying instances of the positive class. A precision of 0.43 is indicated for the positive class, which indicates that 43% of the instances classified as positive were, in fact, positive (0.57% were false positive).The recall for the positive class is 0.85, indicating that the model identified 85% of all positive instances. Therefore, the model was unable to predict approximately 15 percent of potential customers.

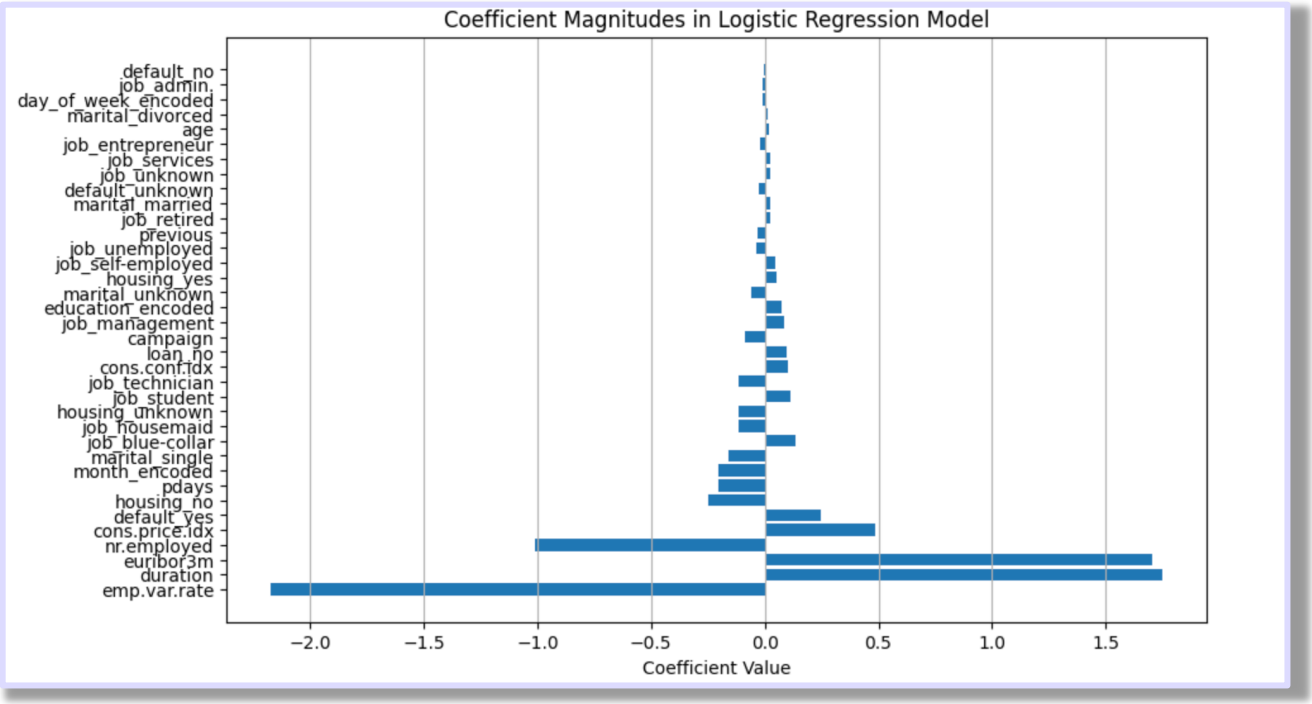


Figure 8| Coefficient magnitudes in Logistic Regression Model

- ***F1 score for GBM (Non-parametric model) using hyperparameter tuning with Bayesian method was 58%***

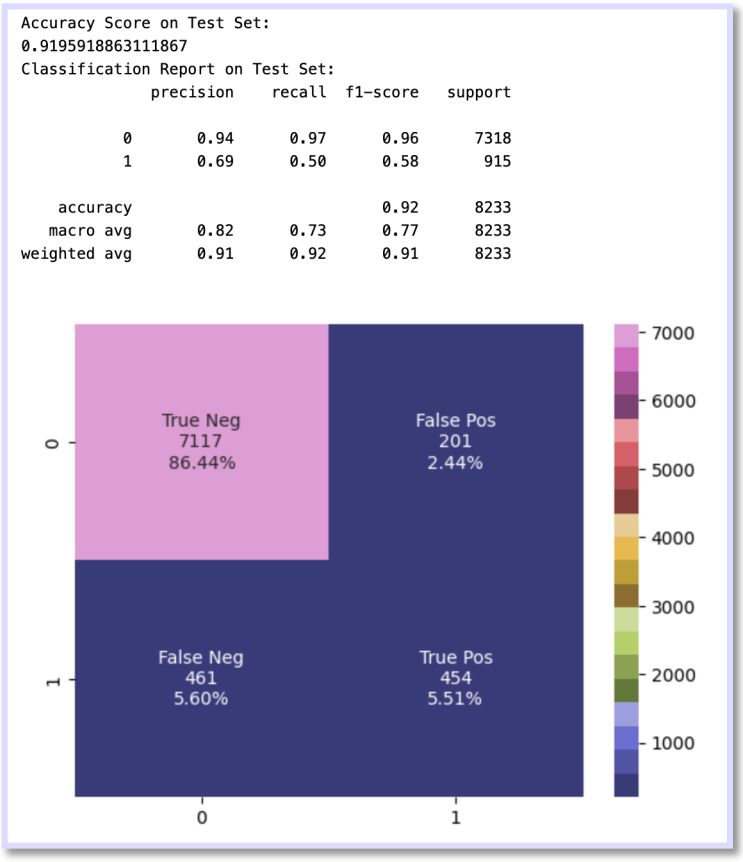


Figure 9| Confusion matrix for the GBM model

According to a classification report, the GBM model has an overall accuracy of 91%. The results indicate that it performs well when classifying instances into two groups. In contrast, the recall for the positive class (1) is lower than that for the negative class (0), indicating that the model is less effective in identifying instances of the positive class. A precision of 0.69 is indicated for the

positive class, which indicates that 69% of the instances classified as positive were, in fact, positive (31% were false positive). The recall for the positive class is 0.50, indicating that the model identified 50% of all positive instances. Therefore, the model was unable to predict approximately 50 percent of potential customers.

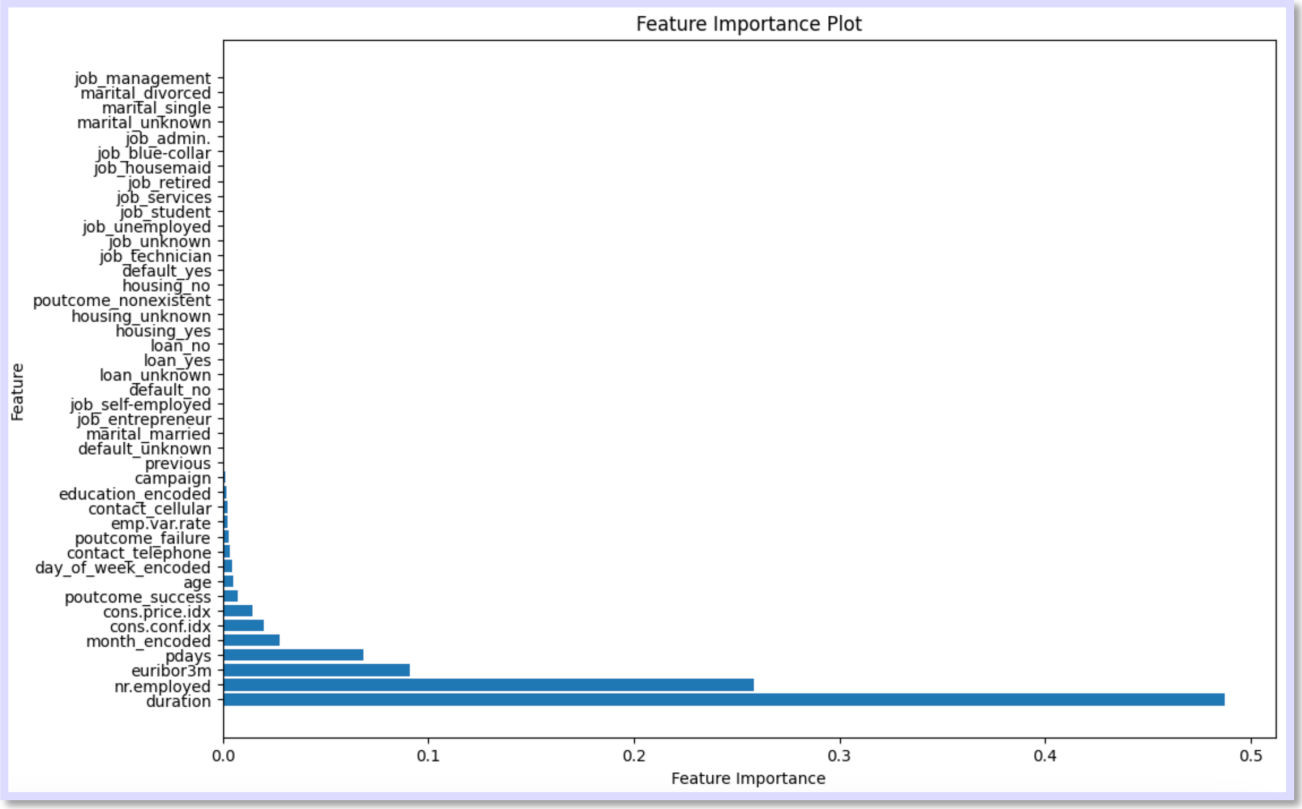


Figure 10| Feature importance for the GBM model

Due to the unsatisfactory performance of the KNN model (see appendix 3 and 4), we have decided not to pursue further exploration of this model.

4 Discussion

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

Equation 1| F1 score

For this project, the F1 score (Equation 1) was employed as the evaluation metric. An elevated F1 score signifies the robust overall performance of a binary classification model, demonstrating its effectiveness in accurately detecting positive cases while simultaneously minimizing the occurrence of false positives and false negatives.

- **Logistic Regression**

  The Coefficient Magnitudes Plot clearly illustrates that "emp.var.rate" and "nr. employed" exert a negative impact on the target variable. Conversely, "cons. price.idx," "duration," and "euribor3m" exhibit a positive influence on the target variable. (Figure)

- **GBM**

  From Feature importance plot it's obvious that "emp.var.rate", "duration", "euribor3m", "nr.employed", "cons.price.idx", "pdays", "cons.conf.idx","month_encoded" has influence on target value.

Main limitations

Despite employing an extensive search for hyperparameters to achieve the highest F1 score, we encountered computational constraints, preventing us from surpassing an F1 score of 58%. Furthermore, it is essential to acknowledge that more intricate models tend to demand additional computational resources, including memory, processing power, and time, for both training and prediction tasks. It is worth considering that overly complex models can lead to inefficiencies during the training and deployment phases.

Moreover, while using SMOTE improved the F1 score for the linear regression model, it is worth noting that this technique can potentially increase class overlap and introduce additional noise.

5 Business insights

Marital status has no significant impact on the success of the marketing campaign.

Seasonality plays a role in influencing the success of the marketing campaign.

The number of days since the last contact from the previous campaign has an impact on the success of the marketing campaign.

Economic parameters, including the three-month Euribor rate, consumer price index, and consumer confidence index, play a pivotal role in determining the success of the marketing campaign.

6 Conclusion

In summary, the analysis of the marketing campaign dataset revealed the significant influence of economic factors and timing on campaign success. This insight enables businesses to adjust their marketing strategies. It is important to note that there should be minimal time gaps between new and old campaigns. By doing so, they can time their campaigns to match periods of favourable economic conditions and tailor their messages accordingly. By understanding these influential factors, companies can allocate their resources more efficiently, focusing on marketing efforts during the best economic conditions and reducing them during less favourable times.

7 References

1. Forbes Business Development Council. (2023, February 1). 10 Top Tips for Businesses in a Recession. [Website]. https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2023/02/01/10-top-tips-for-businesses-in-a-recession/?sh=26a37e7274a5

2. HEC Paris. (n.d.). Marketing During Inflation: How to Adapt. [Website]. https://www.hec.edu/en/knowledge/instants/marketing-during-inflation-how-adapt

3. Jain, I. (n.d.). Model Complexity Explained Intuitively. [Website]. https://ishanjain-ai.medium.com/model-complexity-explained-intuitively-e179e38866b6

4. O8 Agency. (n.d.). B2B Marketing in a Recession. [Website]. https://www.o8.agency/blog/b2b-marketing-recession#:~:text=4.-,Customers%20Shift%20Focus%20to%20Value%2DBased%20Marketing,maximum%20value%20for%20money%20spent.

5. SeroKell. (n.d.). A Guide to F1 Score. [Website]. https://serokell.io/blog/a-guide-to-f1-score

6. Typeset.io. (n.d.). What Are the Pros and Cons of Using SMOTE? [Website]. https://typeset.io/questions/what-are-the-pros-and-cons-of-using-smote-2pzu32jb92

8 Appendix

Appendix Table 1 | Categorical features of the dataset

	job	marital	education	default	housing	loan	contact	month	day_of_week	poutcome
0	admin.	married	basic.6y	no	no	no	telephone	may	mon	nonexistent
1	services	married	high.school	no	no	yes	telephone	may	mon	nonexistent
2	services	married	basic.9y	unknown	no	no	telephone	may	mon	nonexistent
3	admin.	married	professional.course	no	no	no	telephone	may	mon	nonexistent
4	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	nonexistent

Appendix Table 2 | Numerical features of the dataset

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	40	151	1	999	0	1.1	93.994	-36.4	4.857	5191.0	0
1	56	307	1	999	0	1.1	93.994	-36.4	4.857	5191.0	0
2	45	198	1	999	0	1.1	93.994	-36.4	4.857	5191.0	0
3	59	139	1	999	0	1.1	93.994	-36.4	4.857	5191.0	0
4	41	217	1	999	0	1.1	93.994	-36.4	4.857	5191.0	0
...	...	...	...	...	...	...	...	...	...	...	...
41175	29	112	1	9	1	-1.1	94.767	-50.8	1.028	4963.6	0
41176	73	334	1	999	0	-1.1	94.767	-50.8	1.028	4963.6	1
41177	46	383	1	999	0	-1.1	94.767	-50.8	1.028	4963.6	0
41178	56	189	2	999	0	-1.1	94.767	-50.8	1.028	4963.6	0
41179	44	442	1	999	0	-1.1	94.767	-50.8	1.028	4963.6	1

41168 rows × 11 columns

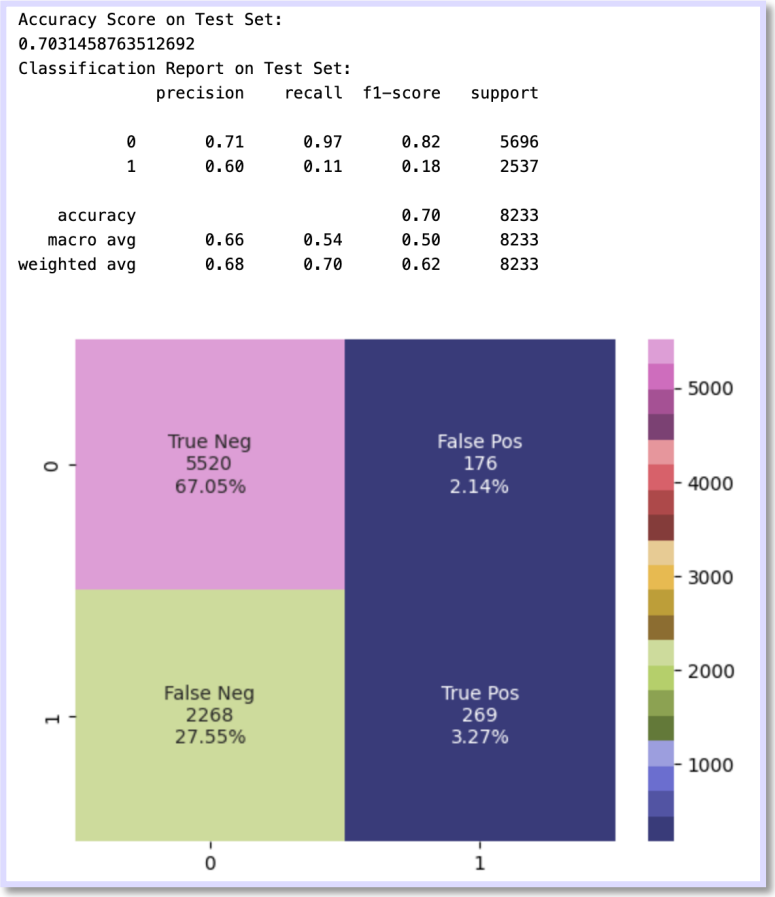
Appendix Table 3 |This table presents an overview of the performance of various machine learning models applied to the dataset. Each row corresponds to a different model, and the columns provide key performance metrics used for evaluation.

	Classifier	Sensitivity Test	Sensitivity Train	f1 Test	f1 Train	Precision Test	Precision Train	AUC Test	AUC Train	Cross-Entropy Test	Cross-Entropy Train
0	KNN	0.606289	0.475983	0.529769	0.664800	0.597260	0.735812	0.875017	0.964907	0.827565	1.366060e-01
1	Decision Tree	1.000000	0.532751	0.515584	1.000000	0.499488	1.000000	0.732965	1.000000	4.014092	2.220446e-16
2	GBM	0.543671	0.527293	0.586521	0.610625	0.660739	0.696386	0.947850	0.952130	0.176229	1.708863e-01
3	Random Forest	1.000000	0.500000	0.563346	1.000000	0.645070	1.000000	0.939786	1.000000	0.204128	4.739337e-02
4	Logistic Regression	0.384037	0.412664	0.513936	0.485312	0.681081	0.659133	0.929620	0.920671	0.212141	2.229970e-01

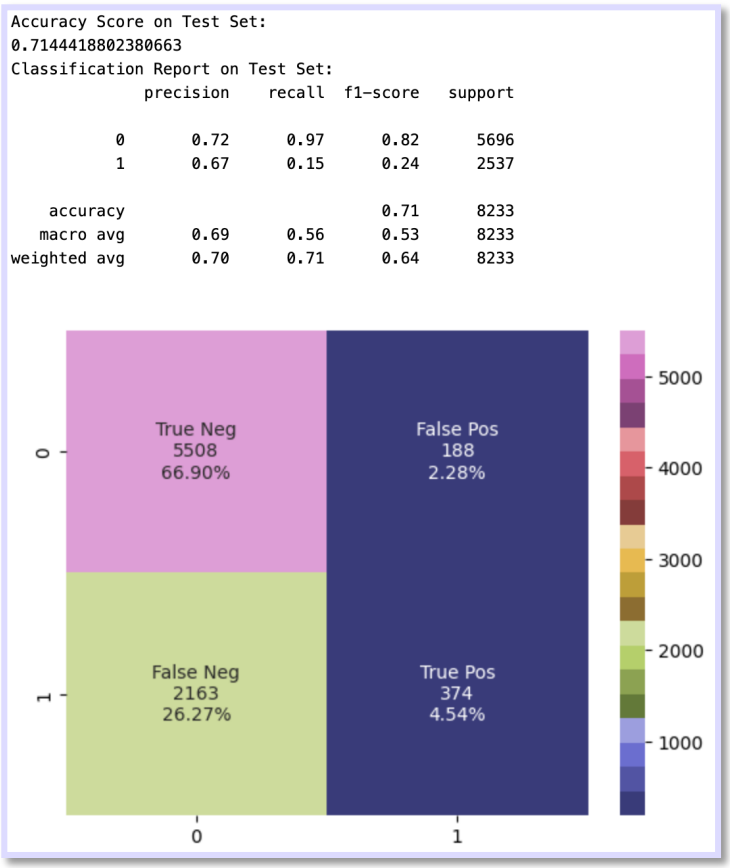
```
    # Use Bayesian optimization to find the optimal hyperparameters
    opt = BayesSearchCV(svm, param_space, n_iter=50, cv=5)
    opt.fit(X_train, y_train)

    # Get the best hyperparameters
    best_params = opt.best_params_

    # Train the SVM model with the best hyperparameters
    svm = SVC(**best_params, probability=True)
    svm.fit(X_train, y_train)

    # Generate probability scores for the training and test sets
    y_train_prob = svm.predict_proba(X_train)[:, 1]
    y_test_prob = svm.predict_proba(X_test)[:, 1]

    # Compute the AUROC scores for training and test sets
    train_auroc = roc_auc_score(y_train, y_train_prob)
    test_auroc = roc_auc_score(y_test, y_test_prob)

    train_auroc_scores.append(train_auroc)
    test_auroc_scores.append(test_auroc)

# Compute the mean of the AUROC scores across all folds
mean_train_auroc = np.mean(train_auroc_scores)
mean_test_auroc = np.mean(test_auroc_scores)

# Print the best hyperparameters and the mean AUROC scores
print("Best Hyperparameters:")
print(best_params)
print("\nTraining AUROC:")
print(mean_train_auroc)
print("\nTest AUROC:")
print(mean_test_auroc)

1221m 12.5s
```

Appendix Figure 1 | The SVM model exhibited significantly extended running times during training, which raised practical concerns about computational resources and time limitations. Given these constraints, it was deemed impractical to consider the SVM model for this analysis further.

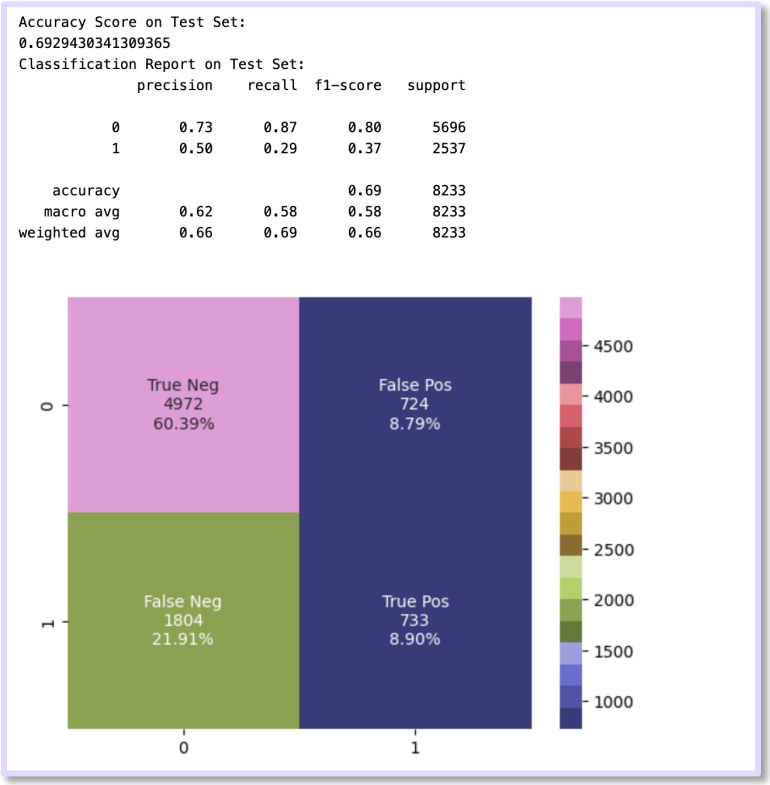
```
Accuracy Score on Test Set:
0.7031458763512692
Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.71      0.97      0.82      5696
           1       0.60      0.11      0.18      2537

    accuracy                           0.70      8233
   macro avg       0.66      0.54      0.50      8233
weighted avg       0.68      0.70      0.62      8233
```



Appendix Figure 2 | This confusion matrix shows the results of the test set for the Logistic Regression model without over sampling. This model demonstrates poor performance in the minority class.

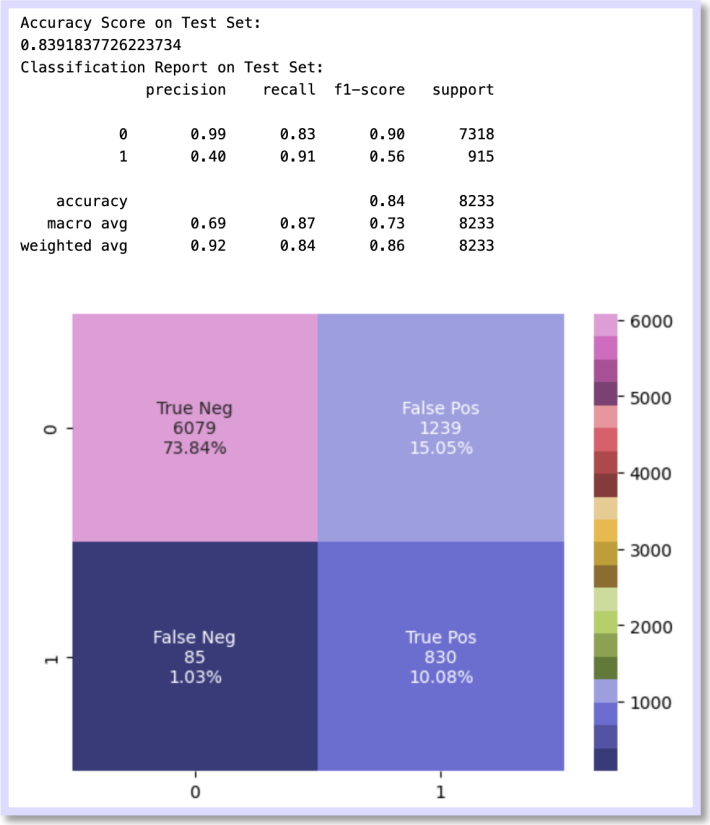


Appendix Figure 3 | This confusion matrix shows the results of the test set for the KNN model. Even using the Bayesian method and hyperparameter tuning, it still demonstrates poor performance in the minority class.



Appendix Figure 4 | This confusion matrix shows the results of the test set for the KNN model. Even using the Bayesian method, over sampling and hyperparameter tuning, it still demonstrates poor performance in the minority class.

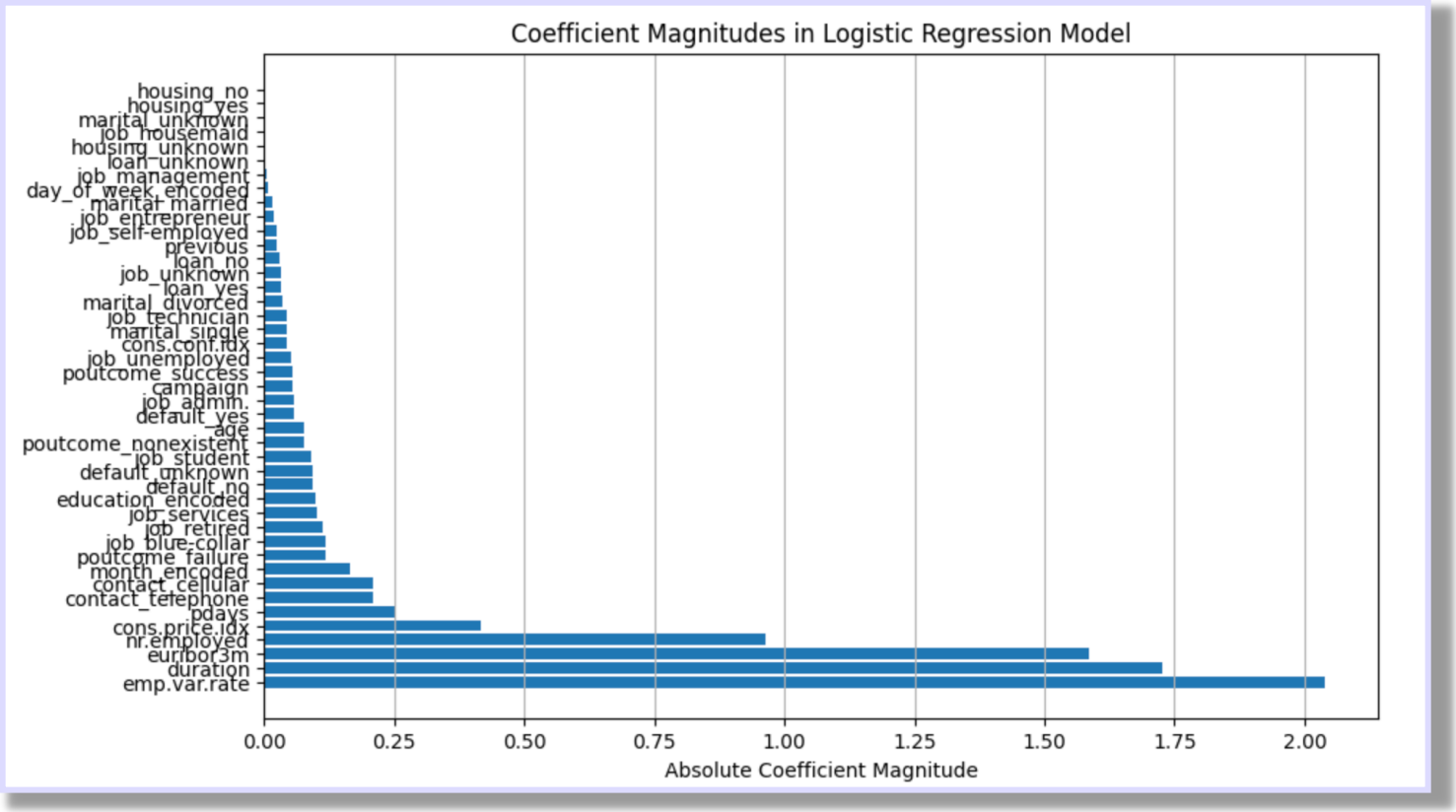
```
Accuracy Score on Test Set:
0.8391837726223734
Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.99      0.83      0.90      7318
           1       0.40      0.91      0.56       915

    accuracy                           0.84      8233
   macro avg       0.69      0.87      0.73      8233
weighted avg       0.92      0.84      0.86      8233
```



Appendix Figure 5| This confusion matrix illustrates the outcomes of the test set when utilizing the GBM model. However, it is worth noting that the inclusion of oversampling slightly degrades the model's performance compared to the same model without oversampling.

Variable Name	Description
age	Age
job	Type of job
marital	Marital status
education	Level of education
default	Has credit in default
balance	Average yearly balance
housing	Has a housing loan
loan	Has a personal loan
contact	Contact communication type
day	Day of contact
month	Month of contact
duration	Last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no").
campaign	Number of contacts performed during this campaign and for this client
pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous marketing campaign
emp.var.rate	employment variation rate - quarterly indicator (numeric)
cons.price.idx	consumer price index - monthly indicator (numeric)
cons.conf.idx	consumer confidence index - monthly indicator (numeric)
euribor3m	euribor 3 month rate - daily indicator (numeric)
nr.employed	number employed - quarterly indicator (numeric)
y	Did the client subscribe to a Telecom plan?

Appendix Figure 6| Data Dictionary



Appendix Figure 7 | Absolute Coefficient Magnitude plot in the Logistic Regression Model, clearly illustrates that "emp.var.rate", "nr. employed","cons. price.idx," ,"pdays","duration," and "euribor3m" has influence on the target variable.