Capstone Project - 3

Bank Marketing Effectiveness Prediction

Presented By

Rupesh Sharma

Contents

❑ Introduction

❑ About the dataset

❑ Data Dictionary

❑ Data Analysis & Visualisation

❑ Evaluation metric

❑ Data Preprocessing

❑ Model Implementation

❑ Feature Importance & Final model evaluation on test set

❑ Challenges Faced

❑ Conclusion

Introduction

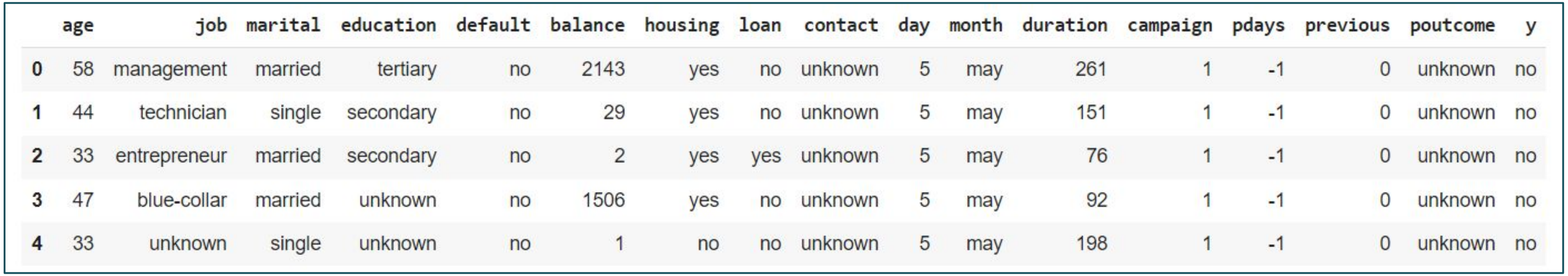
- ❑ A common approach for increasing business is to use marketing selling campaigns. Businesses use direct marketing to reach certain segments of their customers in order to achieve a specific goal. With the use of technology, these marketing methods can be improved even more.

- ❑ The objective of my project is to enhance these marketing strategy by leveraging a dataset which contains telemarketing campaign data from a Portuguese bank, which aids in the development of a predictive model that can target and identify certain clients who might be interested in a term deposit. It also examines the various factors that influence a telemarketing campaign's success.

About the dataset

The bank marketing dataset is about the marketing campaigns, which aim to promote term deposits among existing customers, of a Portuguese banking institution from May 2008 to November 2010.

It consists of 45,211 rows and 17 columns.



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Data Dictionary

Column	Data Type	Description
age	Integer	Age of the bank's customer
job	Object	Type of job of customer
marital	Object	Marital status of the client
education	Object	Education of the client
default	Object	Do customer has credit in default?
housing	Object	Does the customer has housing loan?
loan	Object	Does the customer has a personal loan?
contact	Object	Type of communication contact
month	Integer	Last contact month of the year

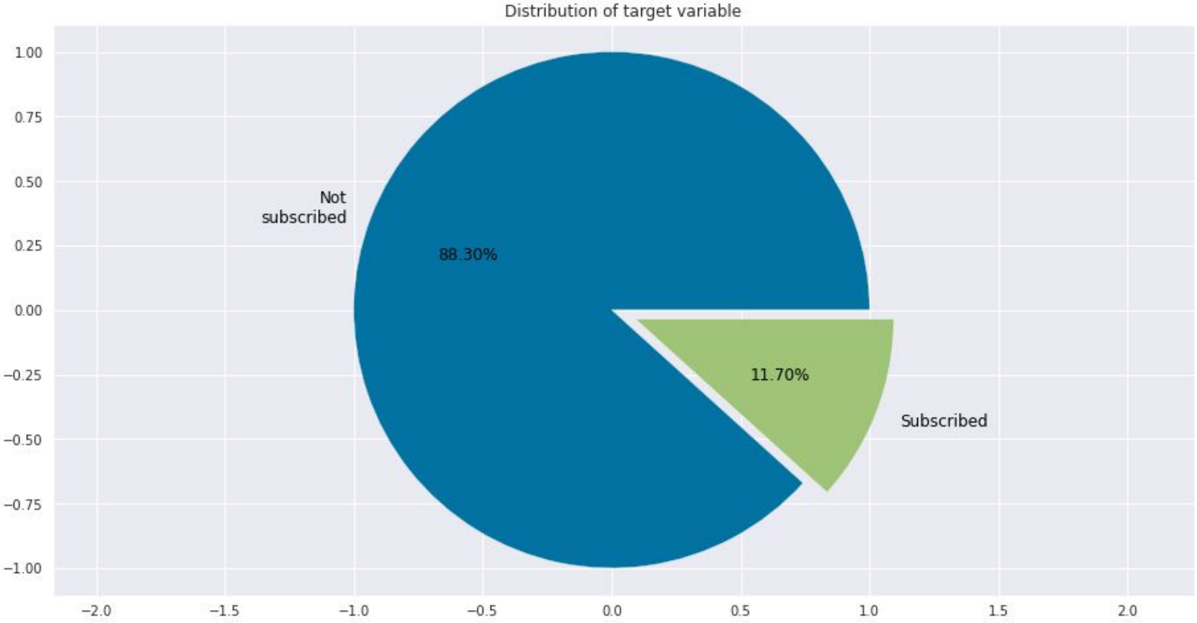
Data Dictionary

Column	Data Type	Description
day	Integer	Last contact day of the week
duration	Integer	Last contact duration in seconds
campaign	Integer	Number of contacts performed during this campaign and for this client
pdays	Integer	Number of days that passed by after the client was last contacted from a previous campaign
previous	Integer	Number of contacts performed before this campaign and for this client
poutcome	Object	Outcome of the previous marketing campaign
y	Object	Has the client subscribed to a term deposit?

Data Analysis & Visualisation

- The distribution of the class variable is shown on the right.

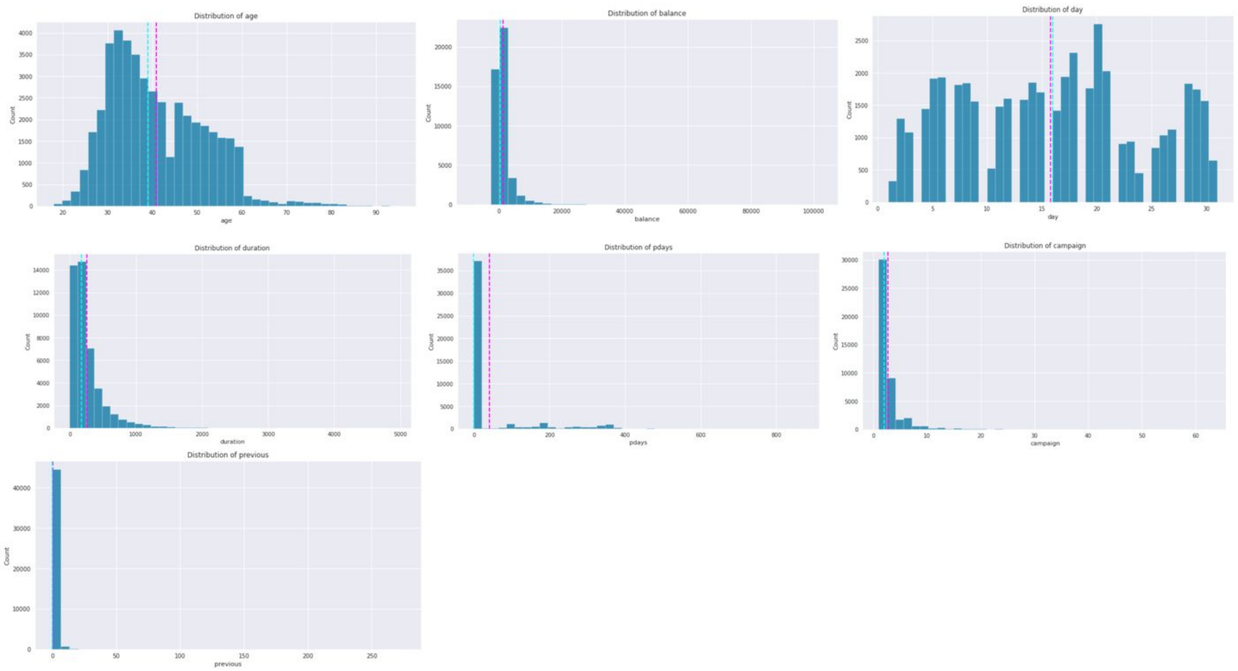
- It indicates that the class labels are highly imbalanced.



Data Analysis & Visualisation

- The histogram on the right depicts the distribution of continuous features.

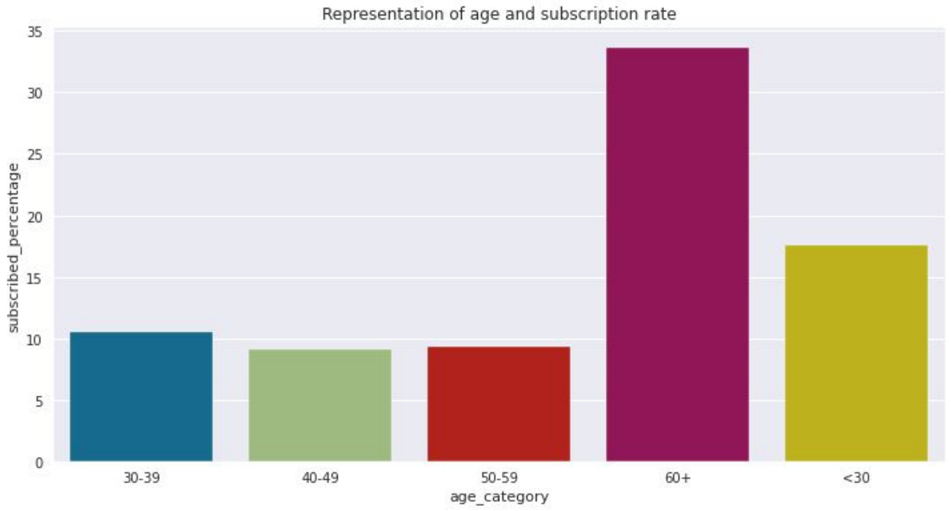
- It indicates that the continuous features are skewed.



Data Analysis & Visualisation

- ❑ The visualisation on the right depicts the representation of subscription rate with respect to the age of the customer.

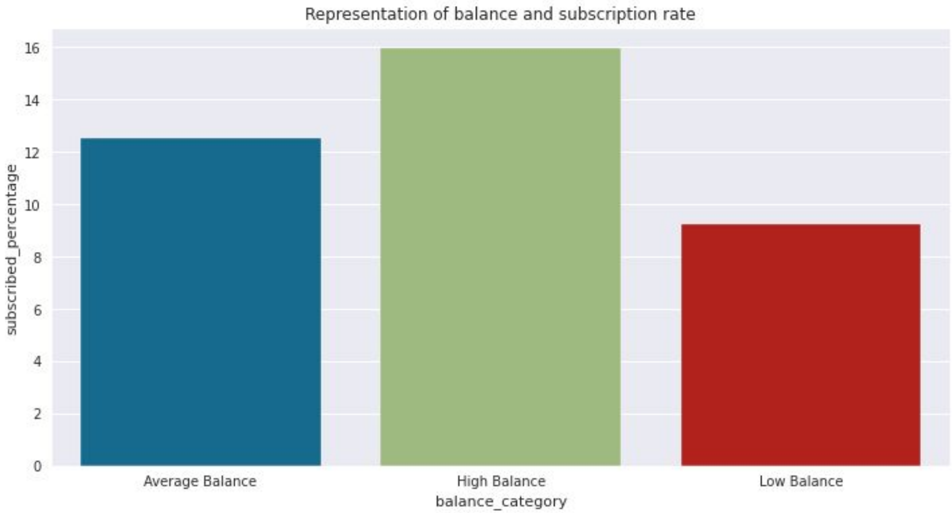
- ❑ From the visualisation, we can infer that customers above the age of 60 and under the age of 30 are more interested in term deposits.



Data Analysis & Visualisation

- The visualisation on the right depicts the representation of subscription rate with respect to bank balance of the client.

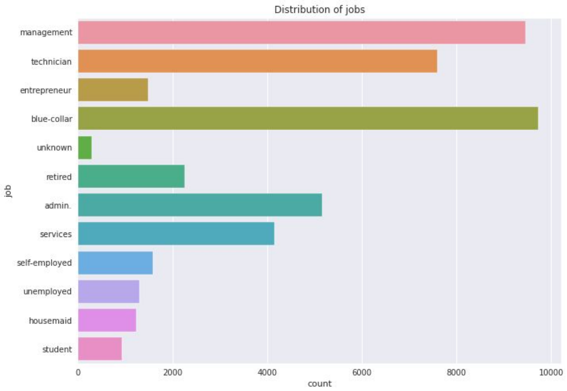
- From the visualisation, we can infer that customers with a medium to high amount are more likely to opt for a term deposit.

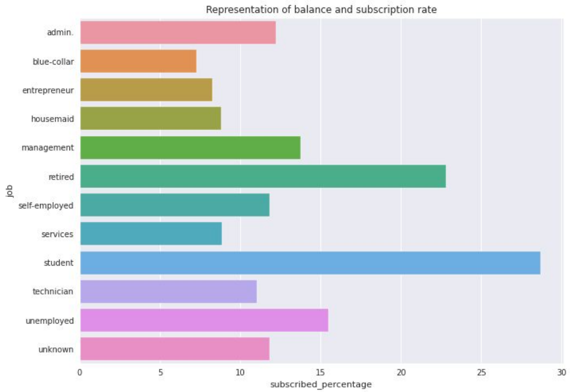


Data Analysis & Visualisation

- The visualisation on the right depicts the representation of subscription rate with respect to the job type of the client.

- From these visualisations, we may infer that although the majority of the bank's customers are employed in blue collar and management roles, term deposits are more popular among students and retirees.

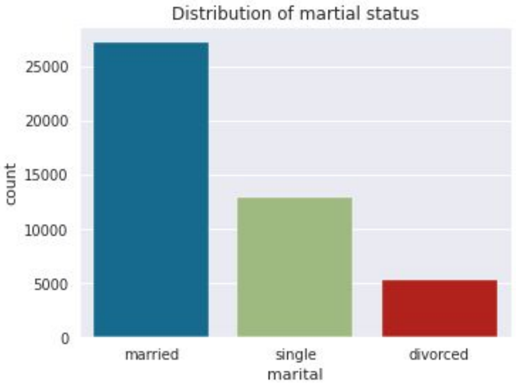


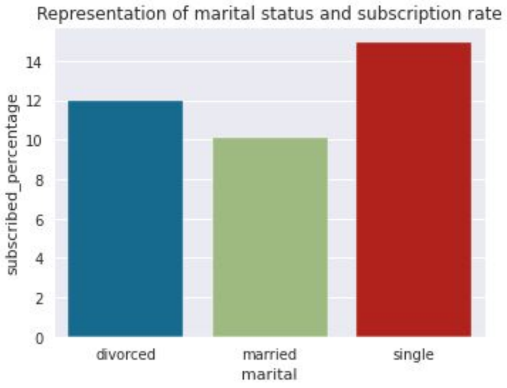


Data Analysis & Visualisation

- **The visualisation on the right depicts the representation of subscription rate with respect to the marital status of the client.**

- **From these visualisations, we may infer that despite the majority of customers are married, singles have a higher subscription rate.**


Distribution of martial status

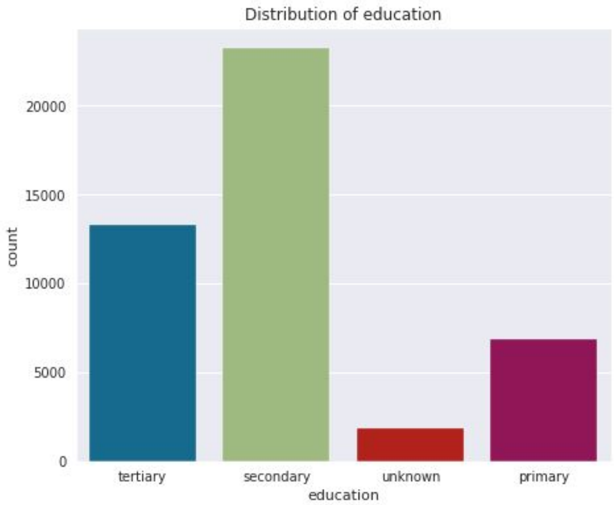

Representation of marital status and subscription rate

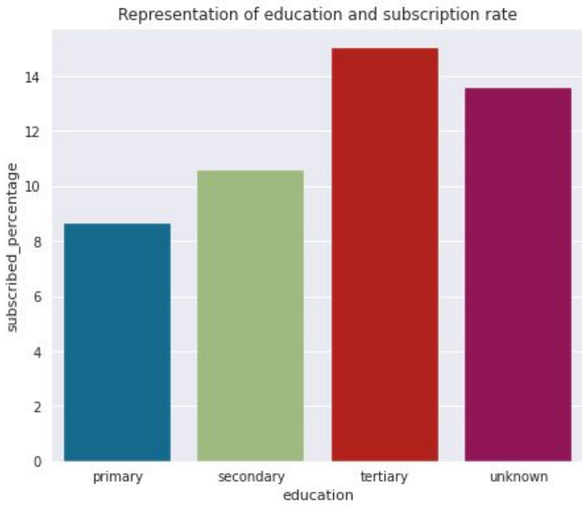
Data Analysis & Visualisation

❑ The visualisations on the right shows the representation of subscription rate with respect to education of the customer.

❑ It appears most of the bank clients are secondary passed.

❑ It seems that as the level of education goes up, more clients subscribe to the term deposit.

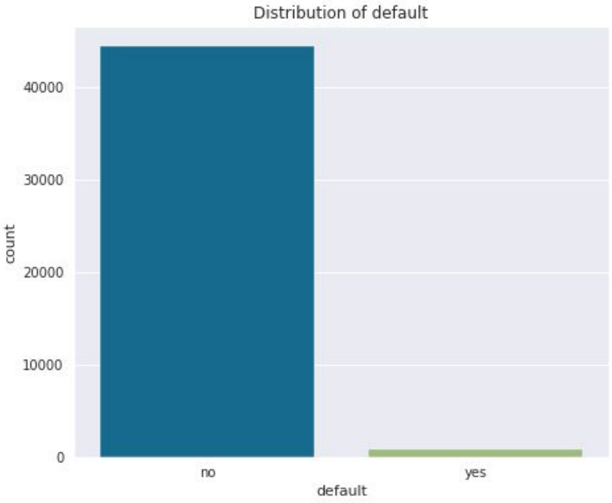


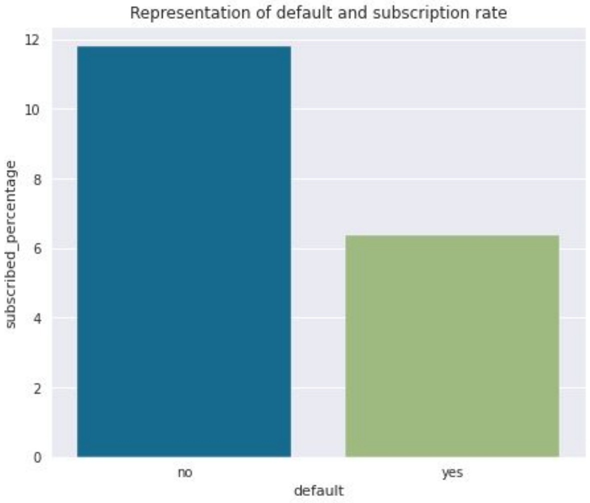


Data Analysis & Visualisation

- The visualisations on the right shows the subscription rate with respect to the client's credit in default.

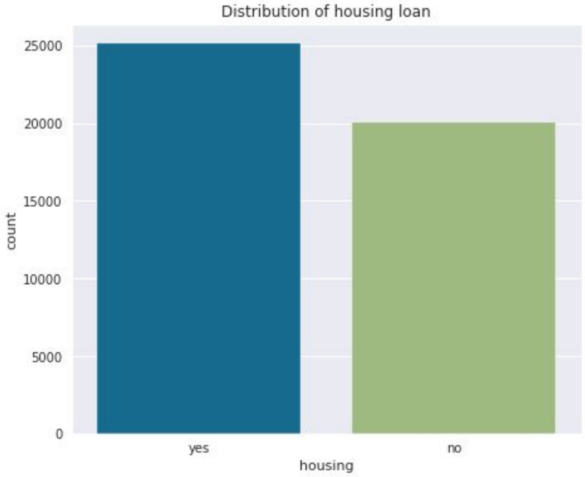
- It appears that customers with no default are more interested in term deposits.

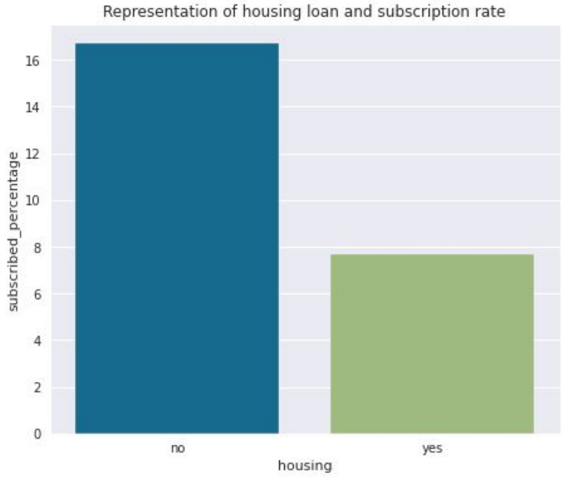




Data Analysis & Visualisation

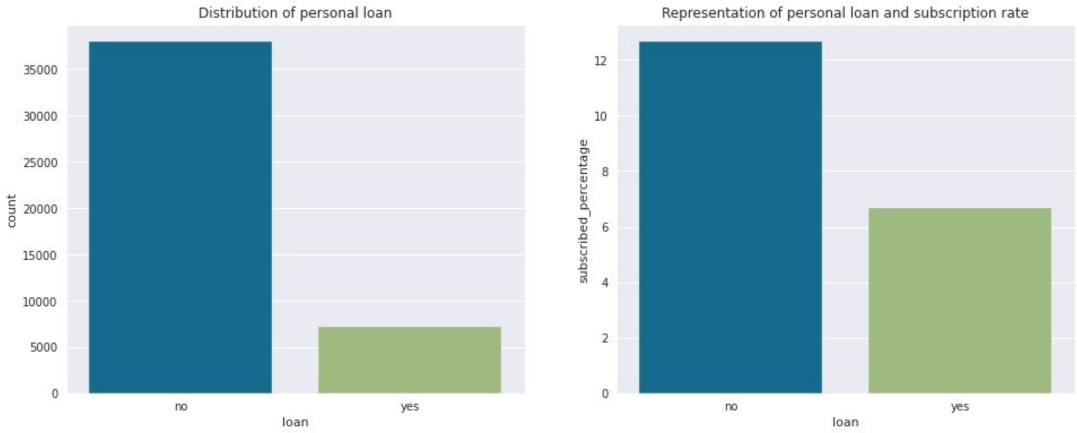
- The representation of subscription rate with respect to housing loan is shown on the right.
- We can deduce from the visualisation that despite the fact that majority of clients have a home mortgage, no-loan clients have a higher subscription rate.


Distribution of housing loan


Representation of housing loan and subscription rate

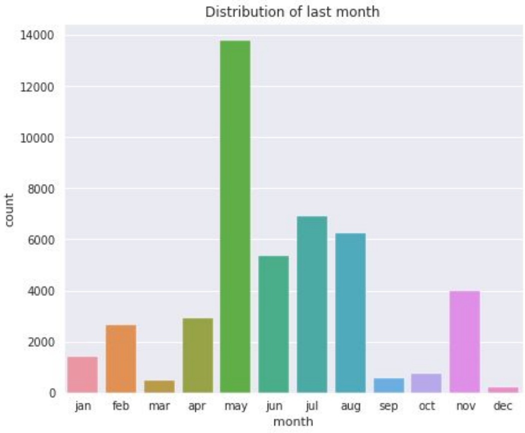
Data Analysis & Visualisation

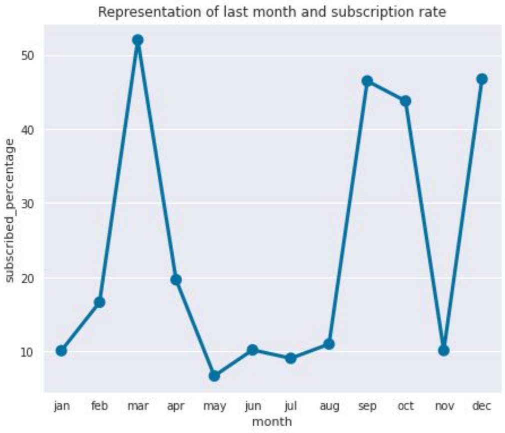
- **The representation of subscription rate with respect to personal loan is shown on the right.**
- **We can deduce from the visualisation that customers with no personal loan have higher subscription rate.**



Data Analysis & Visualisation

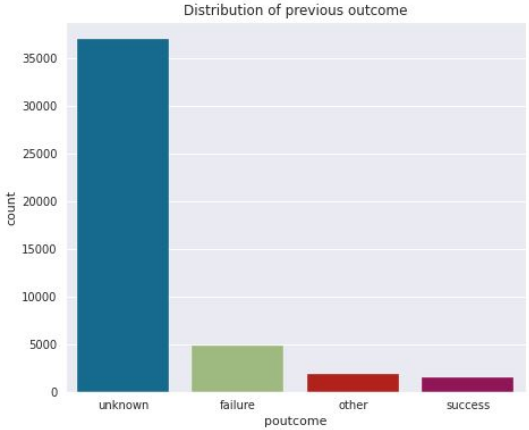
- The visualisations on the right depict the representation of subscription rate with respect to the last month of contact with the client.
- It seems that clients that were approached in March, September, and October have a higher subscription rate.

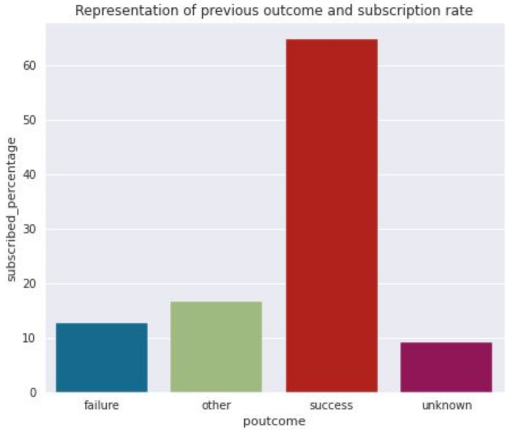

Distribution of last month


Representation of last month and subscription rate

Data Analysis & Visualisation

- ❑ The visualisations on the right depict the representation of subscription rate with respect ot the outcome of previous campaign.
- ❑ It seems that The majority of the preceding campaign's outcomes are unknown.
- ❑ The majority of clients who availed in previous services are more likely to apply for a term deposit.


Distribution of previous outcome


Representation of previous outcome and subscription rate

Evaluation Metric

- The balanced accuracy score is the evaluation metric used in the project. It is the most widely used metric for evaluating a classification task’s performance on an unbalanced dataset.
- It accounts for both the positive and negative outcome classes due to which it doesn’t mislead with unbalanced data.
- The balanced accuracy score is calculated as follows:

$$\text{Balanced Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

where,

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

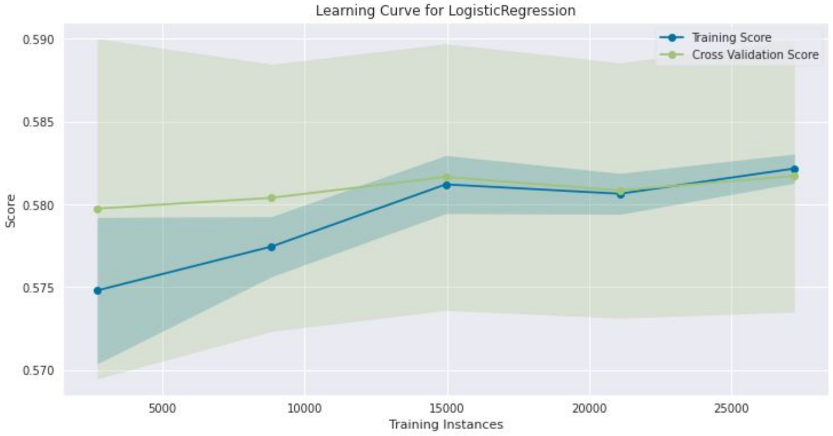
Data Preprocessing

- ❑ The threshold values are used to limit the skewed features.
- ❑ Performed sine and cosine transformation on month feature.
- ❑ Performed Target encoding on job feature and ordinal encoding on education feature, while one hot encoding on remaining categorical features.
- ❑ MinMax scaling is used to transform the numerical variables.
- ❑ Used TomekLinks combined with penalised hypertuned models to handle class imbalance.

Model Implementation

Logistic Regression:

- ❑ **Performance:**
  - ❑ **Balanced accuracy score on training set : 0.58**
  - ❑ **Balanced accuracy score using cross-validation set: 0.58**



- ❑ **Remarks:**
  - ❑ **Model has underfitted the data.**
  - ❑ **To improve the model's performance, hyperparameter tuning is required.**

Model Implementation

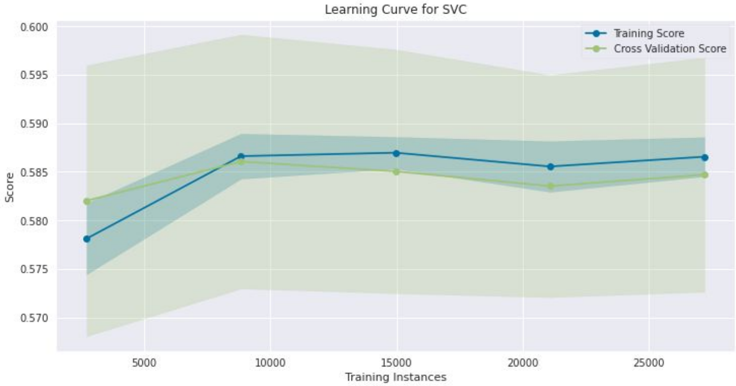
Logistic Regression:

- ❑ **Hyperparameter tuning:**
  - ❑ **Best parameters:**
    - ❑ C : 0.1
    - ❑ class_weight : {0: 1, 1: 7},
    - ❑ penalty: l1,
    - ❑ solver: liblinear
  - ❑ **Performance:**
    - ❑ Balanced accuracy score on training set : 0.69
    - ❑ Balanced accuracy score using cross-validation set: 0.69
- ❑ **Remarks after tuning:**
  - ❑ Although hyperparameter tuning has improved the model's performance, there is still room for improvement. Due to which I continued with more complex models.

Model Implementation

Support Vector Classifier:

- ❑ **Performance:**
  - ❑ **Balanced accuracy score on training set: 0.58**
  - ❑ **Balanced accuracy score using cross-validation set: 0.58**



- ❑ **Remarks:**
  - ❑ **Model has underfitted the training data.**
  - ❑ **To improve the model's performance, hyperparameter tuning is required.**

Model Implementation

Support Vector Classifier:

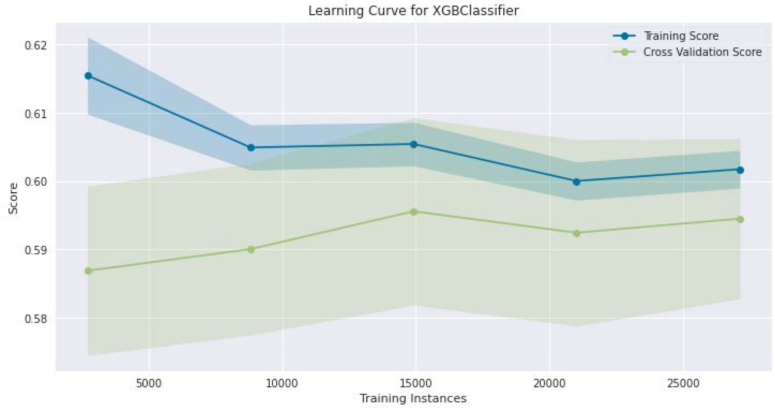
- Hyperparameter tuning:
  - Best parameters:
    - class_weight: {0: 1, 1: 7}
  - Performance:
    - Balanced accuracy score on training set : 0.73
    - Balanced accuracy score using cross-validation set: 0.71

- Remarks after tuning:
  - After hyperparameter tuning the performance of Support Vector Classifier has been improved and it outperformed Logistic Regression. But I continued to experiment with a boosting model with the goal of better performance.

Model Implementation

XGBoost Classifier:

- ❑ **Performance:**
  - ❑ Balanced accuracy score on training set : 0.60
  - ❑ Balanced accuracy score using cross-validation set: 0.59



- ❑ **Remark:**
  - ❑ Model has also underfitted the training data, but performs slightly better as compared to previous models.
  - ❑ To improve the model's performance, it must be hypertuned.

Fitting and Evaluating the Model

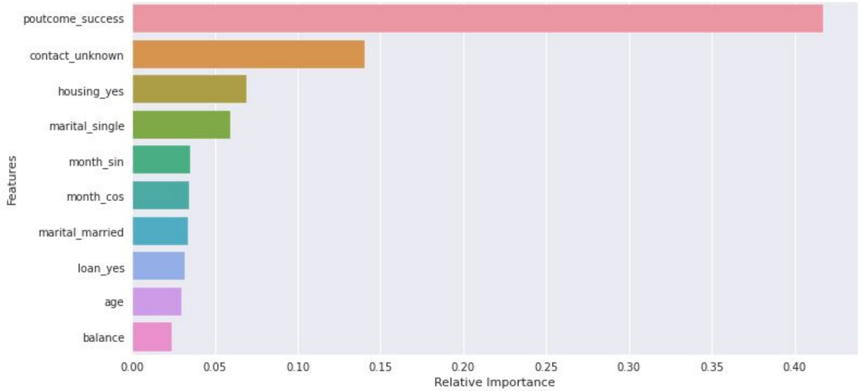
XGBoost Classifier:

- ❑ Hyperparameter tuning:
  - ❑ Best parameters:
    - ❑ reg_alpha : 0.01
    - ❑ scale_pos_weight: 7
  - ❑ Performance:
    - ❑ Balanced accuracy score on training set : 0.74
    - ❑ Balanced accuracy score using cross-validation set: 0.73

- ❑ Remarks after tuning:
  - ❑ The XGBoost Classifier's performance has improved significantly through hyperparameter tuning. The model does not appear to underfit or overfit the data, since it produced a decent score and surpassed the performance of both the previous models. As a result, I have selected XGBoost Classifier as my best model.

Feature Importance & Final Model evaluation on test set

❑ The visualisation depicts the most important features on which marketing campaign depends.



❑ The hypertuned XGBoost Classifier achieved 0.73 balanced accuracy score on the test set.

Challenges

- ❑ Deciding the correct categorical encoding technique for categorical features.
- ❑ Choice of evaluation metric due to imbalance class labels.
- ❑ Choosing the right technique for handling imbalance.
- ❑ Selecting the range of hyperparameter values.

Conclusion

❑ The bank should highly focus on customers who are already enrolled in their previous schemes as they are most likely to opt in for term deposits via marketing campaign.

❑ The bank should target clients whose marital status are single as they are more likely to avail term deposit.

❑ For the next marketing campaign, it will be wise for the banks to focus the marketing campaign during the months of March, September and December.

❑ The marketing campaign should be more focused on clients that neither have house nor personal loan.

Conclusion

- **The customer's age affects campaign outcome as well. The next marketing campaign of the bank should target potential clients in their 30s or younger and 60s or older. This will increase the likelihood of more term deposits subscriptions.**
- **Clients with average and high balances are more likely to subscribe a term deposit.**

Thank You