Is Telemarketing actually an effective sales tactic?

By: Akshay Shembekar, Courtney Golding, Jonathan Littleton, Komal Handa, and Sambhavi Parajuli

Hello!

We are Team **JACKS**

Akshay Shembekar

Software Engineer

Courtney Golding

Digital Control Manager

Jonathan Littleton

RPA Developer

Komal Handa

UX Researcher

Sambhavi Parajuli

Strategic Business Analyst II

1

Decision Making with the Artificial Intelligence (AI) Canvas

AI Canvas Project Overview

Prediction

Predict if a customer will open an account as a result of being included in a telemarketing campaign.

Judgement

Incorrect predictions would lead to wasted time, resources and money of the bank. Falsely predicting that a customer will open an account would cause the wrong types of customers to be targeted in the campaign. Falsely predicting that a customer wouldn't open an account would cause the bank to miss out on a potential customer.

Action

To include a particular customer in a telemarketing campaign or not.
Also, to carry out the telemarketing campaign or not.

Outcome

Accuracy will look at the number of correctly predict values, and F1 score takes into account the precision and recall. F1 measure is also good for imbalanced datasets like this one. We will use F1 score to pick the best model.

Training

Data needed to train the model includes customer attributes (ex: age, job, etc.), when and how many times the customer was contacted, and whether or not the customer opened an account.

Input

Data needed to generate predictions includes customer attributes such as age, job, education, marital status, etc.

Feedback

As additional campaigns are carried out, additional customer data and outcomes would improve the model. The current data is imbalanced, so additional data for customers who chose to open an account would cause improvements.

How will this AI impact the overall workflow?

Telemarketing campaigns can be carried out in a more targeted way to reduce the amount of time and resources allocated for each campaign. Additionally, campaigns may be carried out more or less frequently depending on results.

2

Exploratory Data Analysis (EDA) and Preprocessing

Data Set & Features

- The data is related with direct marketing campaigns of a Portuguese banking institution.
- The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required.
- The goal of the campaign was to get consumers to subscribe to a bank term deposit.
- Link to data set: Telemarketing Data

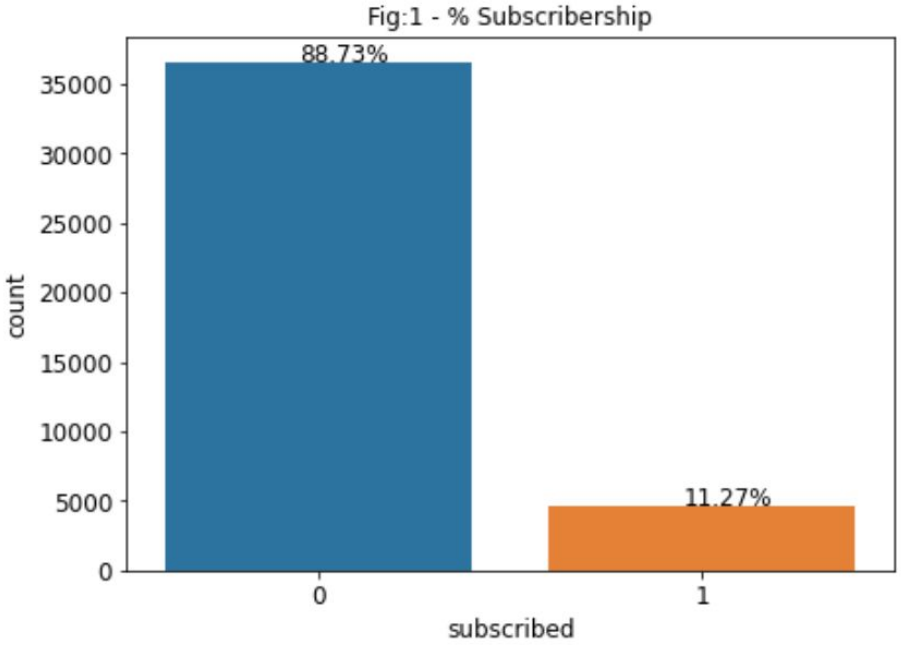
11 Categorical Features in the Data Set:

- **job** : Describes Job of the potential client
- **marital status**: Describes Marital status of the potential client
- **education** : Describes Education of the potential client
- **default** : Represents whether the potential client has credit in default
- **housing** : Represents whether the potential client has housing loan
- **loan** : Represents whether the potential client has personal loan
- **contact** : Represents type of communication channel used
- **month** : Indicates the month when the potential client was last contacted
- **day_of_week** : Indicates the last contact day of the week of the month when contacted
- **poutcome** : Indicates the outcome of the previous marketing campaign
- **y** : binary column indication whether has the 'potential' client subscribed a term deposit

10 Numerical Features in the Data set:

- **age** : Integer value for Age of the potential client
- **duration** : Indicates last contact duration (in seconds)
- **campaign** : Indicates the # of times potential client was contacted
- **pdays** : Indicates number of days since last contacted from a previous campaign
- **previous** : Indicates the # of times potential client was contacted, previously
- **emp.var.rate** : Shows the 'quarterly' employment variation rate
- **cons.price.idx** : Shows the 'monthly' consumer price index
- **cons.conf.idx** : Shows the 'monthly' consumer confidence index
- **euribor3m** : Shows the 'daily' euribor 3 month rate
- **nr.employed** : Shows the 'quarterly' number of employees employed

Distribution of Data



- 0: the customer did not subscribe to a term deposit
- 1: the customer did subscribe to a term deposit

- This data set is an imbalanced data set which has 88.73% of class variable 0 and 11.27% of 1.
  - Class distribution is not uniform among the classes in the imbalanced dataset, so the model will be biased towards the majority class

- Stratified sampling will be used to split the data into train and test data sets
  - The train and test sets will have an equal percentage of “subscribed” and “not subscribed” samples

Null Values

- The following features have null values:
  - Default (8597 null values)
  - Job (330 null values)
  - Marital (80 null values)
  - Education (1731 null values)
  - housing (990 null values)
  - Loan (90 null values)

- 20.8% of the total values for “Default” are null
  - Results of Chi-Square test indicate a dependency between the amount of unknown default values and the outcome, so the rows with null values should not be deleted

- Remaining variables have less than 5% null values

- Values for the features with null variables will be imputed using the most common values

Feature Removal and Manipulation

Variables Removed:

- Housing
  - Crosstab table shows no relationship between “housing” and “subscribed”
- Duration
  - Can not be used in the model for predicting outcomes, because the duration of the call will not be known until after the call is complete
- Pdays
  - The value "999" was used in the pdays columns when a customer was not previously contacted, which will impact the model.
  - Other features such as "Campaign" and "Previous" indicate if a customer was previously contacted, so pdays is not necessary

Variables Modified

- Loan
  - The new variable "Has Loan" will have a value of 1 if the customer has either a Housing loan or Personal loan, and a value of 0 if the customer has no loans
- Education
  - The education levels of basic.4y, basic.6y and basic.9y have been converged into one-single category called basic

3

Model Training, Tuning, and Selection

Baseline Prediction using DummyClassifier/Model Free Baseline/Naive Classifier

Baseline prediction is the simplest possible prediction and a dummy classifier is used to make predictions using simple rules.

For imbalanced dataset, **Stratified strategy** is used which generates predictions by respecting the training set's class distribution

Baseline Accuracy Score : 79.7%

Methods Tested

DECISION TREE

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed.

XGBOOST

Gradient Boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

RANDOM FOREST

The PCA does an unsupervised dimensionality reduction, while the logistic regression does the prediction.

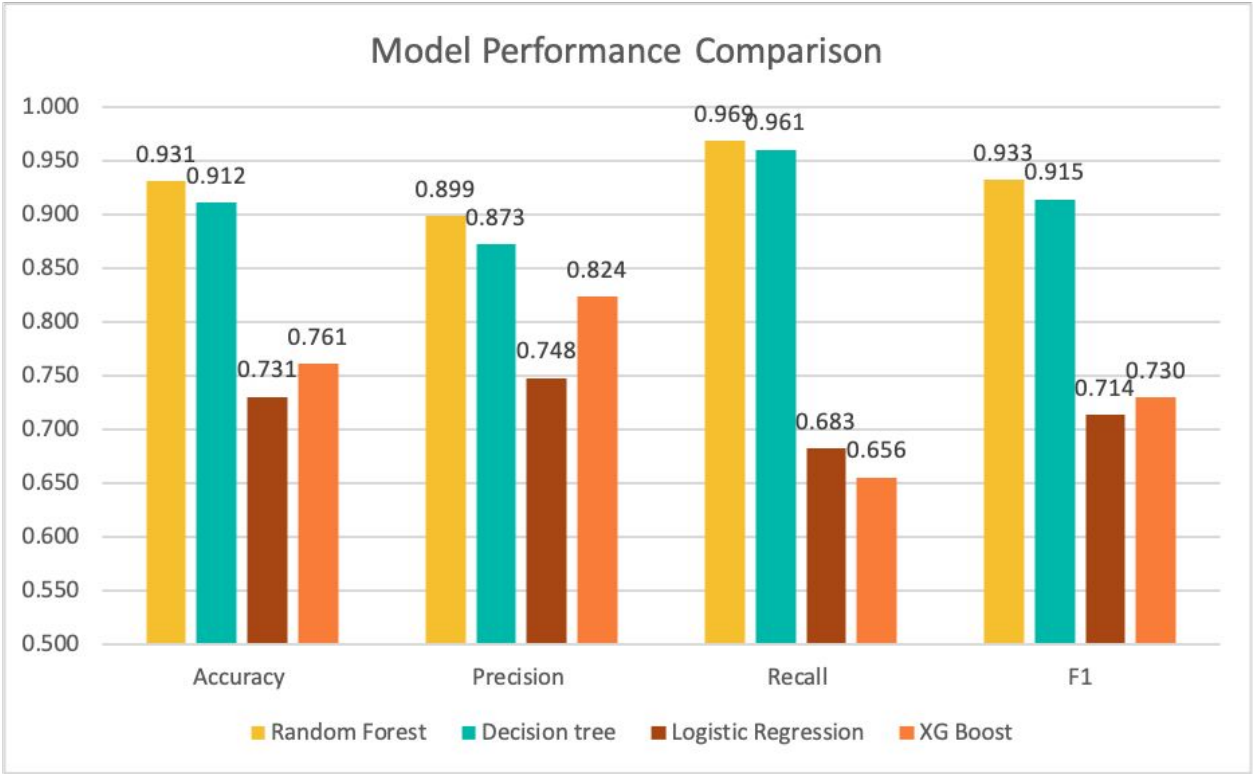
The PCA converts data from high dimensional space to low dimensional space by selecting the most important attributes that capture maximum information about the dataset and logistic regression method is used to train the model.

LOGISTIC REGRESSION WITH PCA

Model Performance Metrics

Random forest was the best performing model with all metrics.

F1 score, precision, and recall were more significant for this dataset than accuracy because it is imbalanced.



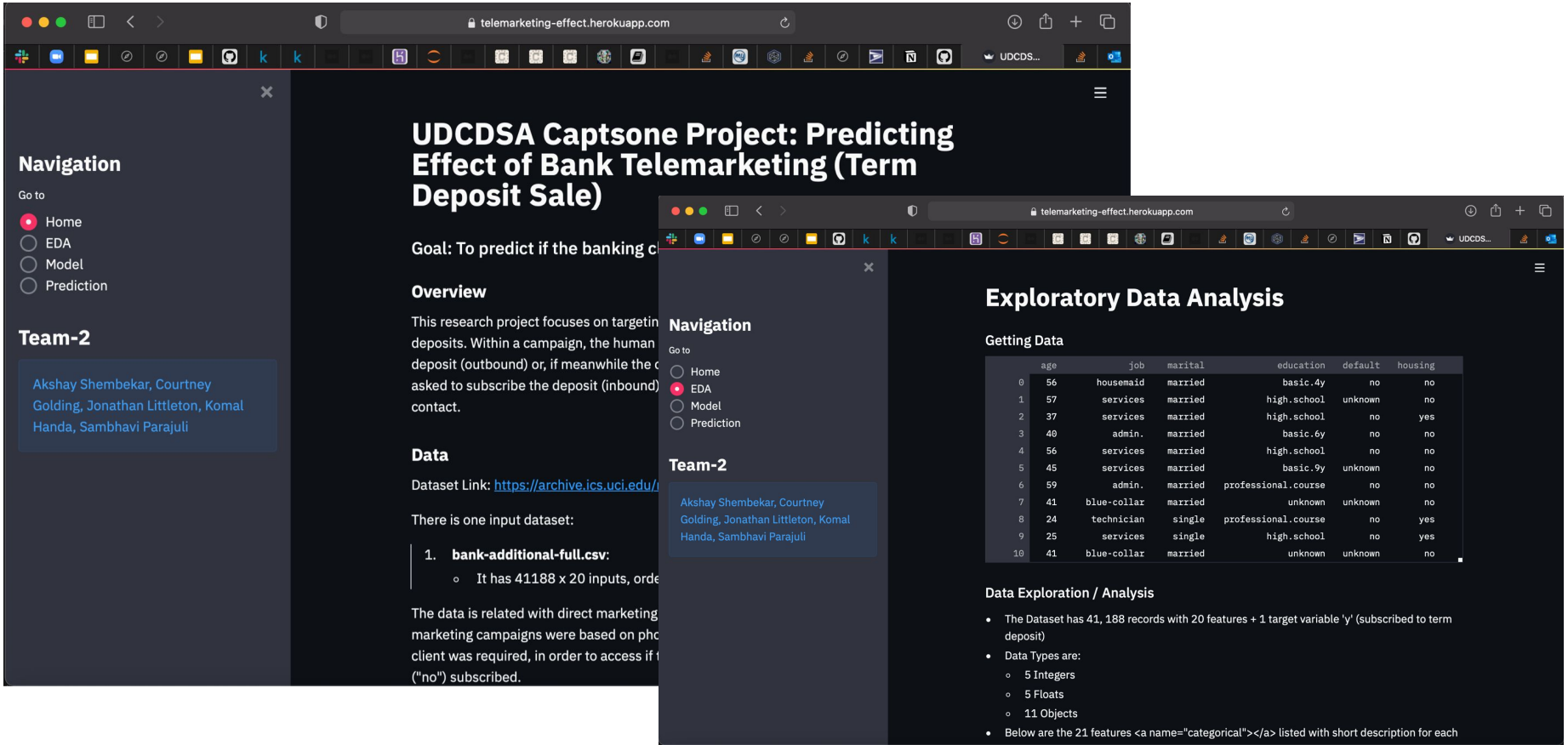
Hyperparameter Tuning

Parameter	Values	Purpose	Optimal Values
Criterion	Gini (impurity) or Entropy (information gain)	Measure the quality of a the split	Entropy
Num Estimators	integer	The number of trees in a forest	150
Max Depth	Integer	The maximum depth of the tree; if unspecified, will keep splitting until all leaves are pure	35
Min Samples Split	Integer	Minimum number of samples required to split an internal node	10

4 Demo

https://telemarketing-effect.herokuapp.com

4.1 Demo - Home & EDA



5

Insights and Conclusion

Key Insights

The average success rate of telemarketing campaigns is 2.5%. Based on this dataset, this bank had an above average success rate, at 11% of calls leading to a successful subscription of a term deposit.

While the bank had an above average success rate, email marketing campaigns have an average response rate of 10.6%. The bank may want to consider allocating resources towards email campaigns in addition or in place of telemarketing, as it can reach a higher volume of consumers quickly and using less resources.

Average Response Rates by Campaign Type:

- Email: 10.6%
- Mail: 0.5 - 2%
- Paid Search: 0.5 - 2%
- Display Ads: 0.05 - 0.1%

Assuming that the average telemarketer salary is €7 in Portugal and the average call duration is 4.3 minutes, these 41,000+ calls cost the bank approximately €20,678. Approx **€18,348** can be lost on unsuccessful telemarketing calls.

Conclusions

If the bank would like to optimize their telemarketing efforts, they should focus on these types of customers, which had a higher impact on the success rate.

Age: 20-40

Targeted in previous campaign: Yes

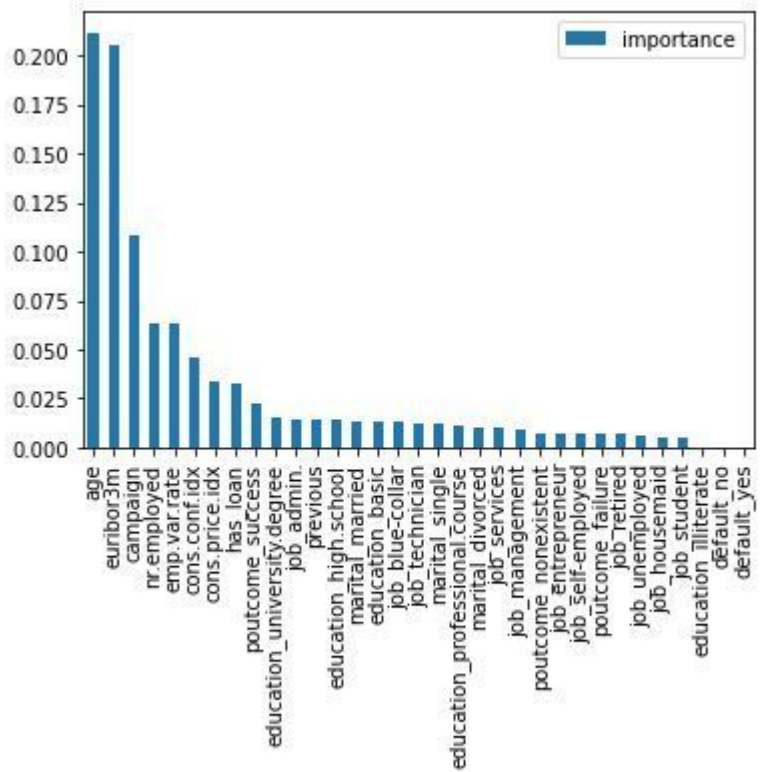
Number Employed: 5,000

Has loan: No

Previous success: Yes

Education: University Degree

Other socio-economic factors like the Euribor rate, Consumer Confidence Index, and Consumer price index also had high importances, but cannot be easily modified by the bank like the customer based features.



Thank you!

6 Appendix

Metric Definitions

Precision:

Out of all the positives predicted - the truly positive percentage

$$Precision = \frac{TP}{TP + FP}$$

Recall:

Out of the total positive, what percentage are predicted positive

$$Recall = \frac{TP}{TP + FN}$$

Accuracy

How often a point is classified correctly

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

F1 Score:

Harmonic mean of precision and recall. It takes both false positive and false negatives into account. Therefore, it performs well on an imbalanced dataset.

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Conclusion Cost Calculation

Average Salary for a telemarketer	€ 7.00
Salary/min	€ 0.12
Average minutes for a call	4.304750167
cost per call	€ 0.50
Unsuccessful calls	36535
Successful calls	4639
Unsuccessful call Cost	€ 18,348.64
Successful call cost	€ 2,329.80

Chi - Squared

- A Chi-Square test is a test of statistical significance for categorical variables. It helps to understand the relationship between the categorical variables of the dataset.
- It is used to analyze the dependence of one category of the variable on the other independent category of the variable. Chi-squared test is performed between categorical variable and class variable.
    - The null hypothesis -The grouping variables have no association or correlation amongst them.
    - The alternate Hypothesis-The variables are associated with each other and happen to have a correlation between the variables.