Identifying Potential Customers of a Banking Institution for a Marketing Campaign

DS 5110: Introduction to Data Management and Processing

Nandavardhan Chirumamilla chirumamilla.n@northeastern.edu

Pranav Bansal bansal.pran@northeastern.edu

Sri Harika Cherukuri cherukuri.s@northeastern.edu

Abstract

Our objective is to identify the potential customers that are likely to set up a term deposit at a Portuguese banking institution along with the key characteristics that add value to the marketing campaign. After preprocessing the data, that includes treating the imbalanced class through sampling techniques, we have evaluated the performance of several classification models such as Logistic regression, Decision Trees, Random Forests, Gradient Boosting Classifier, K-NN and SVM. Random Forest ended up as the best performing model among all of them, where F1-Score has been chosen as the evaluation criteria.

1. Summary

1.1 Background

Banks use depositor's money to make loans. Term deposits which are deposits in a financial institution with a specific maturity rate are one of the major sources for a bank which they can use to make loans[1]. Banks follow various marketing strategies like email marketing, SMS/Phone call marketing, advertisements etc. to reach out to the customers. Phone call campaigning is one of the traditional forms of marketing and, when done suitably, can have the best results. Most of the businesses follow a priority queue where they shortlist the customers, they believe are likely to convert. Organizations allot a huge number of resources towards organizing such campaigns which makes the task of identifying potential customers a crucial one. Our aim is to identify such potential customers that the bank could target and help banks make optimal usage of their resources .

1.2 Related work

'Who will Subscribe a Term Deposit?'[2] by Zhao Hu and others is related to our work where they suggested Neural Networks as the optimal model with ~90% accuracy. However, rather than considering accuracy score as an evaluation metric they could have considered F1-Score or a similar metric to account for the class imbalance in the test data. In another such impressive work 'Bank Marketing Campaign – Opening a Term Deposit'[3], Janio Bachmann suggested an ideal set of months during which the bank could focus it's marketing activity to achieve maximum results. He also suggested an optimal approach to target customers based on their duration of the call, number of calls, etc. This work has also evaluated the classification models based on accuracy and suggested Gradient Boosting Classifier as the best performing model with ~84% accuracy.

1.3 Dataset

The data contains information about a marketing campaign conducted by the *Central Bank of the Portuguese Republic*[4] with phone calls as the medium of communication. There are about 41,188 observations and 20 variables/features including *consumer price index*, *marital status*, *employment variation rate*, average yearly balance etc[5]. The dataset is split into train and test sets in the ratio 80:20.

1.4 Goal of the Project

The main goal of the project is to identify the potential customers who are likely to set up a term deposit using a robust classifier based on relevant features/predictors. We intend to identify the key characteristics that makes a customer, a potential customer. This kind of analysis may also reveal reasons that lead to customers not setting up term deposits, which in some cases may be resolved by the Bank.

1.5 How to classify the customers?

Before we train our classifiers, the data needs to be preprocessed. There are instances in a few categorical variables, where the value is unknown. Such instances are either removed or imputed with the highest occurring element relevant to a particular factor. After treating all such instances, behavior of each of the variables is analyzed and relationships between the variables is explored.

For instance, if the duration of call made to a customer is long, then the customer is likely to set up a term deposit. In addition to this, if the outcome of the previous marketing campaign of a customer was a success, there are higher chances that the customer would convert this time as well i.e., set up a term deposit. Along with these, if the customer has a high consumer confidence index, the probability of conversion might be higher. All such relevant features are identified, and the customers are classified accordingly using suitable classification algorithms.

2. Methods

2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) helps us in analyzing the data to visually summarize their characteristics. It helps us to see what the data can tell us beyond the formal modelling or statistical analysis task. Here, we have performed univariate and bivariate analysis to understand and interpret the different kinds of patterns that exist within the data. We have also inspected the data for any missing/null/unknown values during the EDA.

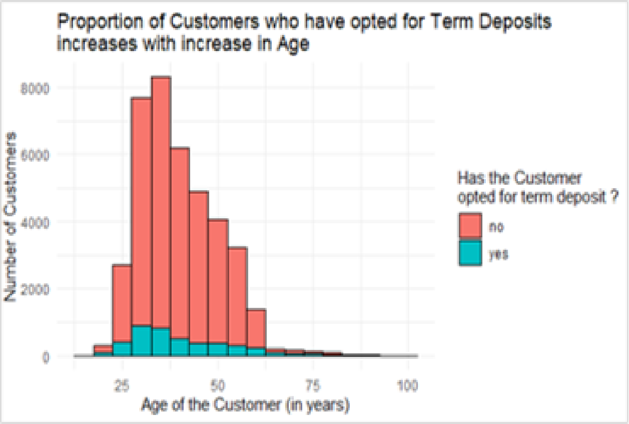


Figure 1 : Call Duration of Customers

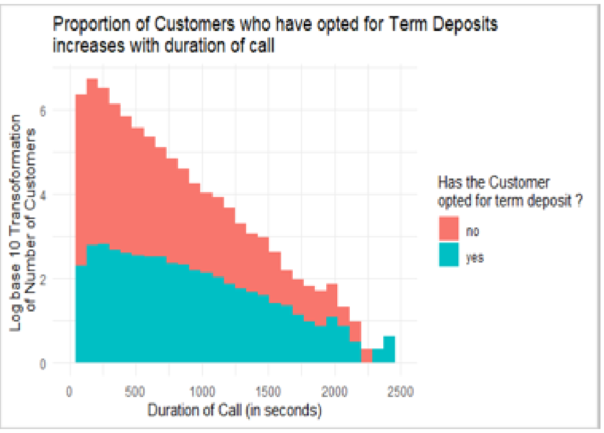


Figure 2 : Call Duration of Customers

The EDA revealed some interesting and useful results. As seen from **Figure 1**, majority of the customers reached out by the Bank were in the age range of 25-50 years. Moreover, there was an increase in numbers of subscribers to term deposits as the age increased. Similar trend can be seen with the duration of the call in **Figure 2**. Higher subscriptions to term deposits were observed with longer durations of phone call. The longer a customer talks on the call, more interested he is in setting up a term deposit.

To better understand the distribution of customers across different categories in a particular feature and how it affects the response/target variable – *y*, we plotted the Normalized Frequency Plots. Normalized frequency for a specific category is equal to:

$$NF = \frac{no.of\ positive\ results\ for\ the\ category}{total\ no.of\ positive\ results} - \frac{no.of\ negative\ results\ for\ the\ category}{total\ no.of\ negative\ results} \qquad \ldots \text{ Equation 1}$$

A positive value for the normalized frequency suggests that the category favors the client who will subscribe to the term deposit while a negative value favors the negative outcome.

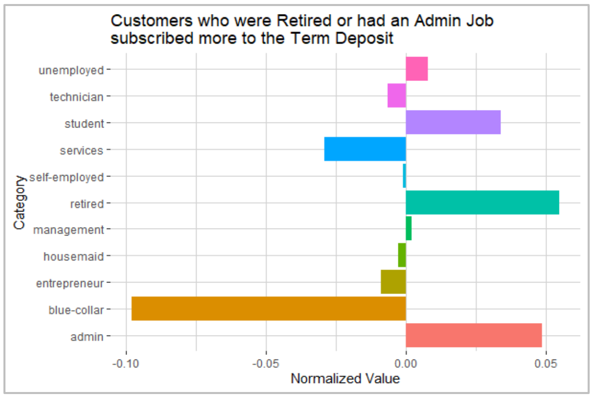


Figure 3: Normalized Frequency for Job

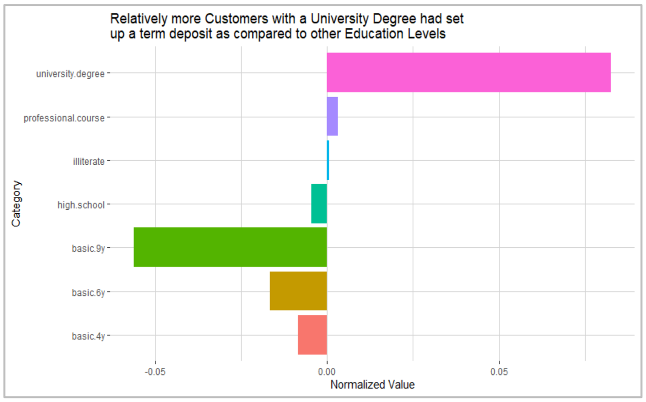


Figure 4: Normalized Frequency for Education

As observed from **Figure 3,** most of the customers with a *blue-collar* job gave a negative response while customers who were *retired*, had an *admin* job or were *students* said yes. In context of Education, **Figure 4** suggests that customers with just a middle school education level (*basic.9y*) were highest in the number to say no while customers with a *University Degree* were most likely to subscribe to a term deposit.

Visualizations for Socioeconomic factors indicated that Lower *Consumer Price Index* and higher *Consumer Confidence Index* is more favorable for a positive outcome. Moreover, the Bank received comparatively more positive responses from the customers who were a part of some previous campaign run by the Bank.

Bivariate Analysis

We wanted to know if there existed some pattern in the age of customers who were subscribing to a term deposit across different job categories. It can be observed from the plot in **Figure 5** that as general trend for a given job category, lower the age of the customer, more are the chances of them getting converted. It can also be seen that the median age of customers in the *retired* category is higher for the positive outcome which is interesting as it suggests that relatively older '*retired'* people wanted to set up a fixed term deposit with the Bank.

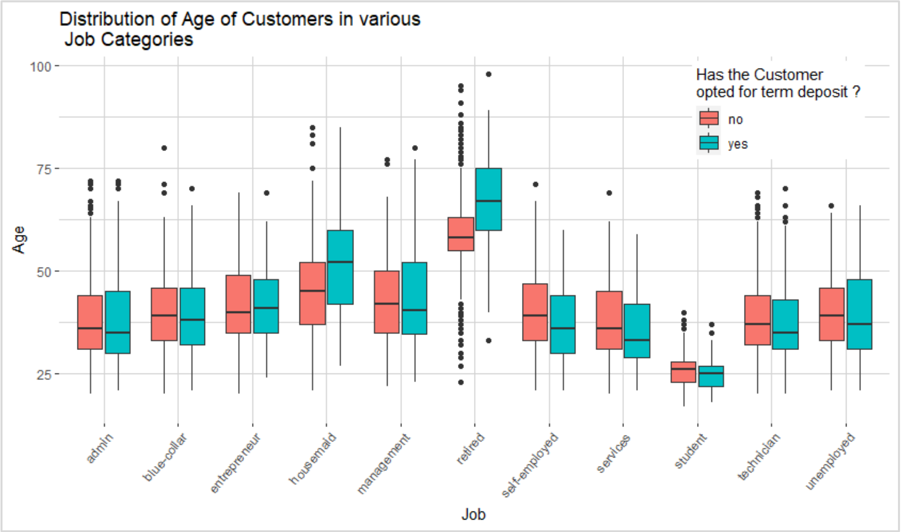


Figure 5 : Boxplot of Job vs Age

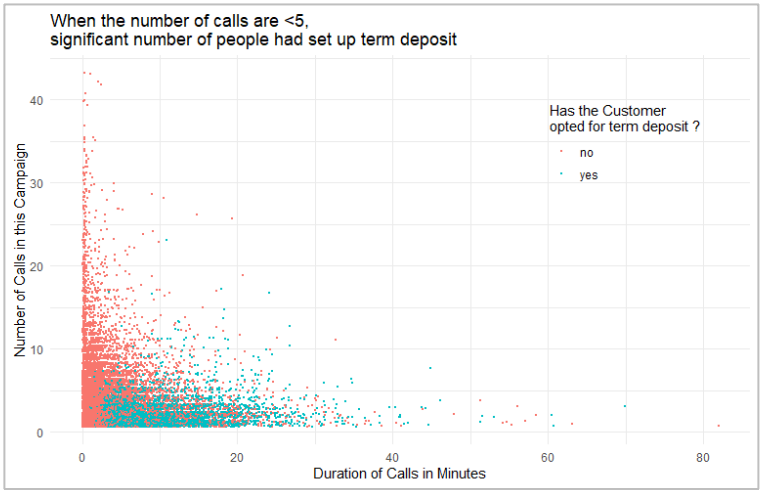
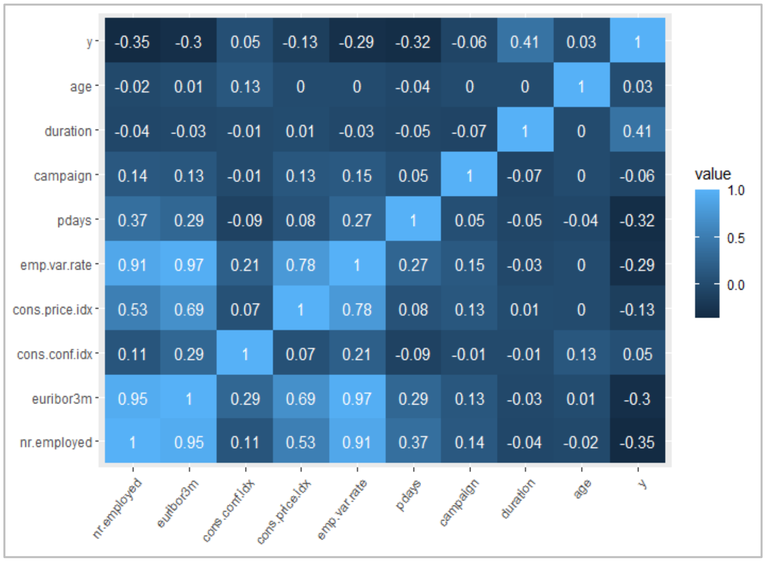


Figure 6 : Duration of Calls Vs No. of Calls

We were also interested in understanding how the duration of calls and number of calls to a particular customer together affected the final outcome. It can be observed from **Figure 6** that as the duration of call increases and the number of calls decreases, there is an increase in the number of conversions. So, we can optimal conversion rates for higher durations of calls with lower frequencies.

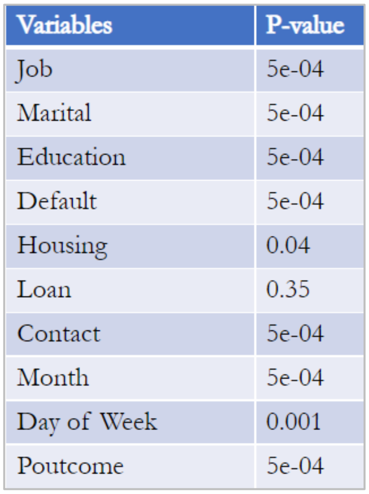
Correlation Matrix

Variables such as *no. of employees, euribor3m, consumer price index* and *employment variation rate* are strongly correlated. We have removed them from the training data as highly correlated features may negatively impact the model efficiency. The target variable *'y'* has a strong correlation with the *no. of employees* and *duration*. **Figure 7** shows the correlation values for all the pairs of continuous values in our data.



	nr.employed	euribor3m	cons.conf.idx	cons.price.idx	emp.var.rate	pdays	campaign	duration	age	y
y	-0.35	-0.3	0.05	-0.13	-0.29	-0.32	-0.06	0.41	0.03	1
age	-0.02	0.01	0.13	0	0	-0.04	0	0	1	0.03
duration	-0.04	-0.03	-0.01	0.01	-0.03	-0.05	-0.07	1	0	0.41
campaign	0.14	0.13	-0.01	0.13	0.15	0.05	1	-0.07	0	-0.06
pdays	0.37	0.29	-0.09	0.08	0.27	1	0.05	-0.05	-0.04	-0.32
emp.var.rate	0.91	0.97	0.21	0.78	1	0.27	0.15	-0.03	0	-0.29
cons.price.idx	0.53	0.69	0.07	1	0.78	0.08	0.13	0.01	0	-0.13
cons.conf.idx	0.11	0.29	1	0.07	0.21	-0.09	-0.01	-0.01	0.13	0.05
euribor3m	0.95	1	0.29	0.69	0.97	0.29	0.13	-0.03	0.01	-0.3
nr.employed	1	0.95	0.11	0.53	0.91	0.37	0.14	-0.04	-0.02	-0.35

Figure 7 : Correlation Matrix (only Numerical Variables)



Variables	P-value
Job	5e-04
Marital	5e-04
Education	5e-04
Default	5e-04
Housing	0.04
Loan	0.35
Contact	5e-04
Month	5e-04
Day of Week	0.001
Poutcome	5e-04

Table 1: Chi Square Test

We also performed chi square tests with the categorical variables, the results of which are shown in **Table 1**. We took the alpha value as 0.05 and performed hypothesis tests, during which it was revealed that p value for *Loan* variable is greater than alpha and it was not included for training.

2.2 Preprocessing

2.2.1 Missing Value Treatment

Before training the models on our data, we properly cleaned and preprocessed the data so that we can remove any bias that is already present in the data. Most of the features had very few unknown values which were dropped, but the variables education and job had almost 4% and 1% unknowns respectively. The unknown values in *education* variable were imputed by inferring the values from the *job* variable. The unknown values in *job* variable where age was >60 were considered retired and the rest were dropped.

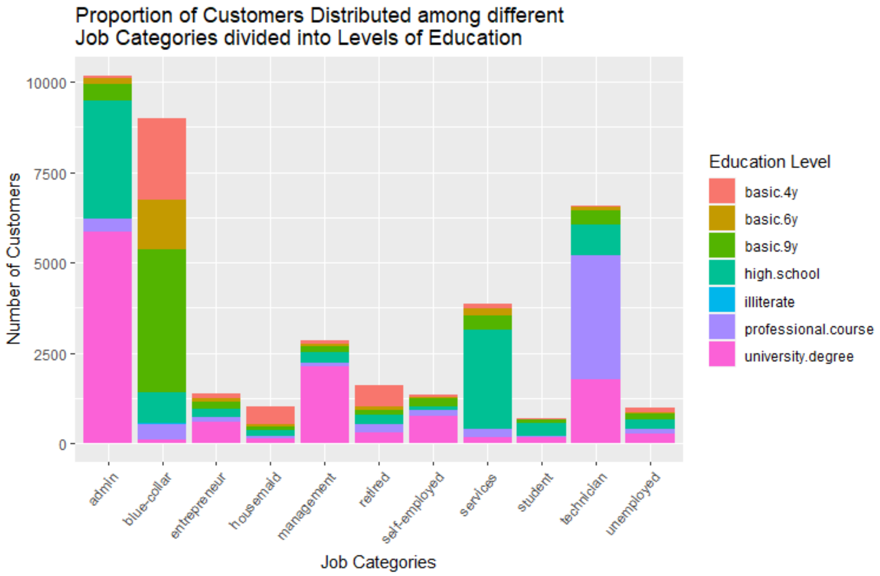


Figure 8 : Plot showing Distribution of Customers across different Job Categories

An example of the above imputation is if a customer had *unknown* education level and job profession as *admin* then we imputed the education value with "*university.degree*" which is basically the mode of the education categories occurring within the *admin* job.

2.2.2 Dealing with Class Imbalance

The dataset at hand is a highly imbalanced dataset with 89% negative and 11% positive responses. Having an imbalanced dataset, leads to bias towards the majority class in the trained model, which may lead to decreased performance. In order to deal with this problem, we used the **Synthetic Minority Oversampling Technique (SMOTE)** to make the dataset balanced. This is a statistical technique that makes the dataset balanced by sampling new instances from the feature space of the minority class.

Following the balancing of the dataset, we have encoded all the categorical variables to make them suitable for training the classification models.

2.3 Classification Models

Given the information about a customer, we are using a classification algorithm to predict whether a customer would set up a term deposit account or not. We have preprocessed all the data and split it into train and test datasets in the ratio 80:20. We have used 10-fold cross validation on the training data to estimate the performance of the training data. We have used different suitable classification algorithms to identify the best model that can capture all the variation in the data.

For most of the algorithms, there are multiple hyperparameters to choose from and identifying the optimal hyperparameters to attain an optimal model is a challenge. We have performed grid search for each of the algorithms limited to certain specific hyperparameters rather than extensive grid search on all hyperparameters and all combinations due to hardware and time constraints. We have used average f1 score as the evaluation metric during 10-fold cross validation. Logistic regression was implemented using step wise AIC in R while all the other algorithms were implemented using the *sklearn* library in python.

The models which we have implemented are:

2.3.1 Logistic Regression

Logistic Regression is one of the most used Supervised Machine Learning algorithms used for classification. In logistic regression, we intend to find a decision boundary that separates the two classes, and we assign the probability for a data point to belong to a particular class based on the distance of the point from the decision boundary. In Logistic regression, we assume that the model has no multicollinearity or that if it exists, it is minimal. This simply means that the independent variables should not be correlated with each other. Hence, all highly correlated features/predictor variables have been removed from the dataset during preprocessing. Logistic regression can be viewed as a linear regression model, whose output is transformed using a sigmoid function. We can extend logistic regression to multiple classes using the SoftMax function instead of the sigmoid function. Logistic regression achieved a f1 score of 0.48 on the testing data.

2.3.2 Decision Trees

A Decision Tree is a tree-like model of decisions where at each node, we have a decision to make and by the time we have reached the leaf node, the data point would have been classified. Different algorithms like Gini impurity and information gain are used to determine which input features/predictor variables at a node help us achieve the best split. We must ensure to limit the depth of a decision tree and ensure that the model does overfit. To control the depth of the decision tree we can either set the maximum depth using model building or using pruning. We achieved the best performing model with a f1 score of 0.54 at a depth of 9 levels.

2.3.3 Random Forests

Random Forests is an ensemble algorithm that is based on the concept of bagging. It is based on the intuition that a large number of uncorrelated models, will outperform any of the individual models. It builds multiple decision trees simultaneously and uses the prediction from all the trees to make a final decision. Random forests rely on the concept of bagging. In bagging, we make use of bootstrapped datasets to construct decision trees, so that each tree is constructed differently focusing on different aspects. Adding on to this, random forest adds in more variability by adding another constraint that each tree in a random forest can only pick from a random subset of features. In order to arrive at the optimal model, we have fine tuned variables like the number of trees, the depth of each tree and the minimum number of samples to split a node for which we have achieved 10,10 and 4 respectively as the optimal values. The f1 score on the test data for the optimal model is 0.58. This turned out to be the best performing model.

2.3.4 Gradient Boosting Classifier

Gradient Boosting Classifier is another ensemble-based algorithm that uses multiple weak learners to build a strong predictive model. Gradient Boosting Classifier is based on the concept of boosting, where the individual models are built subsequently. Decision trees are commonly used as the individual models in gradient boosting classifier. Each subsequent model learns from the mistakes of the previous model by reducing the errors. So instead of filling a predictor on the data at each iteration, it fits a new model to the residual errors made by the previous model. We achieved an optimal model with an f1-score of 0.56 using 200 estimators each with a maximum depth of 1 level and minimum number of samples as 2 for splitting at a node.

2.3.5 K Nearest Neighbors (KNN)

K Nearest Neighbors algorithm is a non-parametric Supervised classification algorithm. The KNN classification algorithm is used to classify a given data sample into one of the categories by identifying the K nearest samples in training data and then using the mode of all the labels of the K closest matches to predict. The KNN algorithm involves the fine tuning of several hyperparameters. For example, there are several distance metrics to choose from to calculate the K closest samples like the Euclidean, Minkowski, Chebyshev or Manhattan distance. We performed a grid search on the number of neighbors, the weight function and the algorithm used to compute the nearest neighbors in order to identify the optimal model. We found the optimal model with number of neighbors as 6 and uniform weights using ball tree as the algorithm for computing the distance and Minkowski as the distance metric.

2.3.6 Support Vector Machine

Support Vector Machine is another supervised classification algorithm whose objective is to find a hyperplane in an N-dimensional space, where N is the number of features/ input variables that classifies the data points. Many hyperplanes may exist that separate the data points into respective classes, but our goal is to find the hyperplane that has the maximum margin. Support vectors are the data points that are closer to the hyperplane, that influence the orientation and position of the hyperplane and help us build our SVM. Any change in the support vectors, may yield an entirely different hyperplane with different positions and origin. After fine tuning if different kernels, we have achieved an optimal model using a polynomial kernel and a kernel coefficient set to *scale* with a regularization penalty of 1.0 yielding an f1 score of 0.53 on test data.

2.4 Evaluation Metrics

We have used Accuracy, f1 score, Precision and Recall as the evaluation metrics for the classification problem at hand. Due to the imbalance of classes present in the dataset, accuracy works as a very poor evaluation metric on test data (SMOTE was performed only on the training data). We have used f1 Score as the main evaluation metric, upon which the best model has been selected. F1 score has been chosen as the main evaluation metric, as it controls the tradeoff between precision. So, with the current evaluation metric, the tradeoff between wasting resources and losing out on potential customers is controlled.

3. Results

3.1 Comparison of Models

Model Name	Cross Validation Avg f1 Score (Training)	F1 score (Testing)	Accuracy (Testing)	Precision (Testing)	Recall (Testing)
Decision Tree	0.84	0.54	0.87	0.43	0.73
Logistic Regression	0.47	0.48	**0.91**	**0.64**	0.38
Random Forest	0.85	**0.58**	0.82	0.44	**0.82**
Gradient Boosting	0.86	0.56	0.87	0.43	0.79
Support Vector Machines	0.86	0.53	0.85	0.40	0.80
KNN	**0.90**	0.51	0.84	0.39	0.75

Table 2 : Performance of Classifier on Test Data

We have achieved the best results on the test data, using Random Forest as the classification model. Logistic regression is a fast but simple classifier. Although, it has achieved good accuracy and precision , it has attained the least recall and f1-score.

3.2 Feature Importance – Show results are in line with EDA

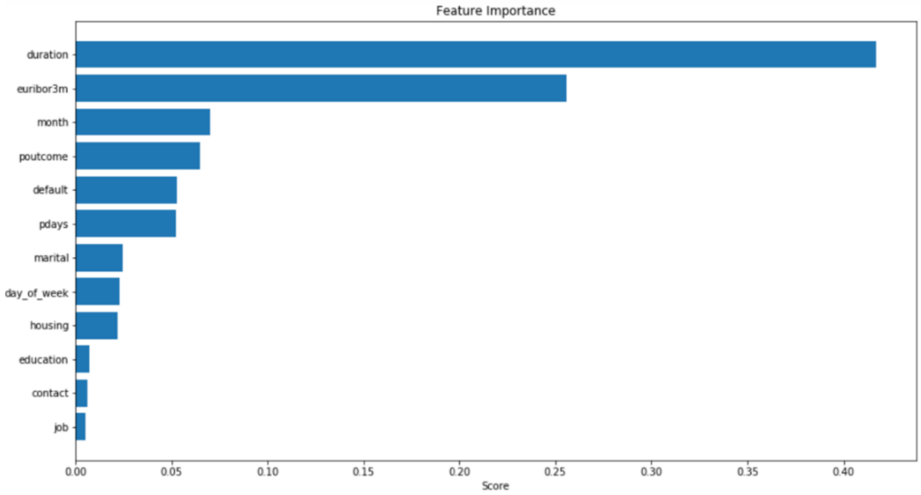


Figure 9 : Plot showing the feature importance of the top features as per the Random Forest Classifier

In **Figure 9**, we can clearly observe that *duration* of the call has achieved the highest feature importance according to the random forest classifier, which is in line with the insights generated during EDA. Similarly, the insights from EDA are consistent with the 3-month Euribor interest rate and month of the year, w.r.t having high importance in predicting whether a customer would set up a term deposit or not.

4. Discussion

We have already analyzed which features are important that help the bank to identify the customers who are likely to set up a term deposit in the above sections. Based on these results the bank can focus on its next marketing campaign. They can target the individuals whose duration of call is higher in this campaign. If customers already have a housing loan, it means they might have financial compromises whom the bank can give least priority to contact. They can target retired individuals as they tend to set up more term deposits. With such strategies the bank can identify the potential customers they can target in their next campaign. Using our current model, the Bank would be able to get a better idea of how much of its customers that have already been contacted would set up a term deposit and also be able to strategize on how to target customers for the next campaign. Since, they invest on call centers to carry out these phone call campaigns, with these strategies they can limit their resources thus maximizing their return on investment.

As part of our future work, we are looking forward to getting more data of the customer like – Daily Bank Account Balance, Average Bank Transaction Amounts per Week, Income, Credit Score, etc., that may explain the residual variability, which couldn't be explained by the current features that have been used in our model. We can also explore more advanced sampling techniques to deal with the class imbalance problem. For example, we can train a machine learning model to generate more realistic samples of the minority class. With all these strategies, we can expect improved efficiency of our current approach.

5. Statement of contributions

Nandavardhan Chirumamilla : Performed EDA, preprocessing of dataset and built Decision Trees and Gradient Boosting Model and optimized the models built. Prepared presentation and report

Pranav : Performed EDA , preprocessing of dataset and built Logistic Regression and Random Forest Model and optimized the models built. Prepared presentation and report

Harika : Performed EDA , Feature Engineering and built K-Nearest Neighbours and Support Vector Machines and optimized the models built. Prepared presentation and report

References

1. Chen, J. (2021, December 7). *Term deposit definition*. Investopedia. Retrieved December 13, 2021, from https://www.investopedia.com/terms/t/termdeposit.asp.

2. Chen, J., Han, Y., Hu, Z., Lu, Y., & Sun, M. (2014, December 7). *Ada Project - Columbia University*. Advanced Data Analysis, Department of Statistics. Retrieved December 13, 2021, from http://www.columbia.edu/~jc4133/ADA-Project.pdf.

3. Bachmann, J. (2019, March 16). *Bank marketing campaign || Opening a Term Deposit*. Kaggle. Retrieved December 13, 2021, from https://www.kaggle.com/janiobachmann/bank-marketing-campaign-opening-a-term-deposit/notebook.

4. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

5. Moro, S., Rita, P., & Cortez, P. (2014, February 14). *Bank Marketing Data Set*. UCI Machine Learning Repository: Data Set. Retrieved December 13, 2021, from https://archive.ics.uci.edu/ml/datasets/bank+marketing.

6. Dua, S. (2020, May 18). *Text classification using K nearest neighbors*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/text-classification-using-k-nearest-neighbors-46fa8a77acc5.

7. Gandhi, R. (2018, July 5). *Support Vector Machine - Introduction to Machine Learning Algorithms*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

8. Nelson, D. (2019, August 8). *Gradient Boosting Classifiers in Python with Scikit-Learn*. Stack Abuse. Retrieved December 13, 2021, from https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/.

9. Yiu, T. (2021, September 29). *Understanding Random Forest*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

10. Aliyev, V. (2020, October 7). *Gradient Boosting Classification Explained through Python*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d.

11. Gupta, P. (2017, November 12). *Decision Trees in Machine Learning*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

12. Vyas, A. (2020, September 1). *Understanding Logistic Regression*. Medium. Retrieved December 13, 2021, from https://medium.com/analytics-vidhya/understanding-logistic-regression-in-depth-intuition-99ad14724464#:~:text=Phew%E2%80%A6-,Assumptions%20of%20Logistic%20Regression,to%20the%20log(odds).

13. Swaminathan, S. (2019, January 18). *Logistic Regression - Detailed Overview*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

14. Gandhi, R. (2018, July 5). *Support Vector Machine - Introduction to Machine Learning Algorithms*. Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

15. Patil, P. (2018, May 23). *What is Exploratory Data Analysis?* Medium. Retrieved December 13, 2021, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15.

16. Wikipedia. (2021, November 15). *Exploratory Data Analysis*. Wikipedia. Retrieved December 13, 2021, from https://en.wikipedia.org/wiki/Exploratory_data_analysis.

17. Media, T. (n.d.). *3 month Euribor Interest Rate*. Euribor 3 months - 3 month Euribor interest rate. Retrieved December 14, 2021, from https://www.global-rates.com/en/interest-rates/euribor/euribor-interest-3-months.aspx.

APPENDIX

Github repo : https://github.com/Nandu960/Identifying-Potential-Customers-of-a-Banking-Institution-for-a-Marketing-Campaign

We share some additional plots that further support the insights shared in EDA.

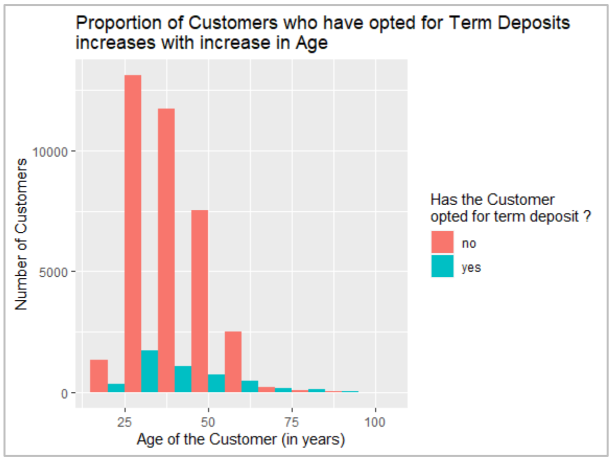


Figure 1 : Plot showing the Impact of Age on subscription to Term Deposits

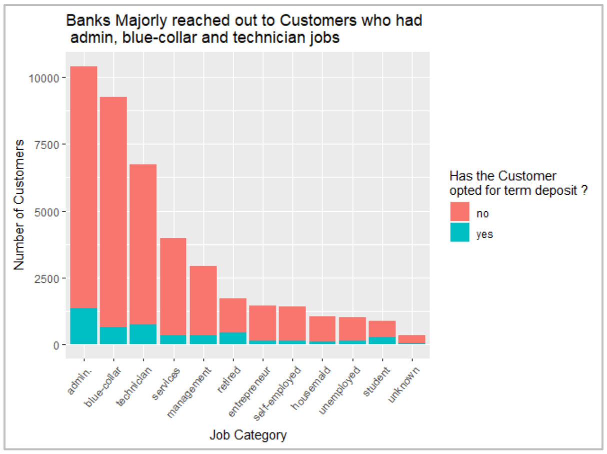


Figure 2 : Plot showing Impact of Job Category on subscription to Term Deposits

Figure 2 shows the impact of Job category on subscription to term deposits. The Job demographic of the customers reached out by the Bank was very diverse. Although, customers with an admin job were highest in the numbers but the ratio of customers who said yes was highest for the customers in the "student" category (approx. 31%) followed by retired customers (approx.. 25%).

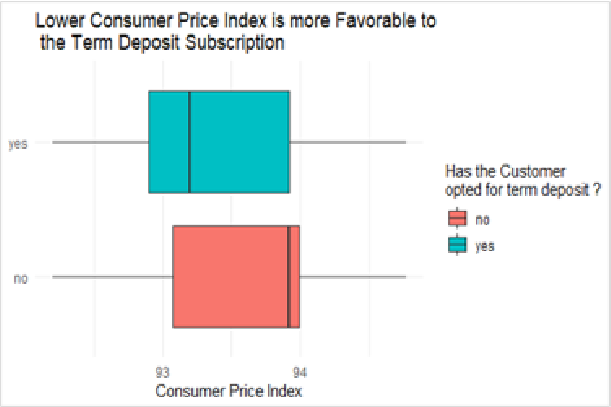


Figure 3 : Plot showing impact of Consumer Price Index on subscription to Term Deposits

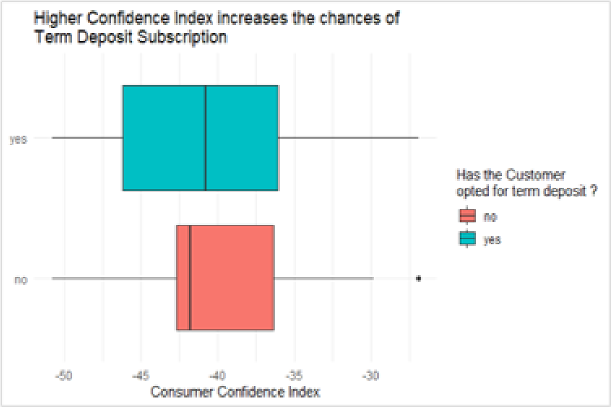


Figure 4 : Plot showing impact of Consumer Confidence Index on subscription to Term Deposits

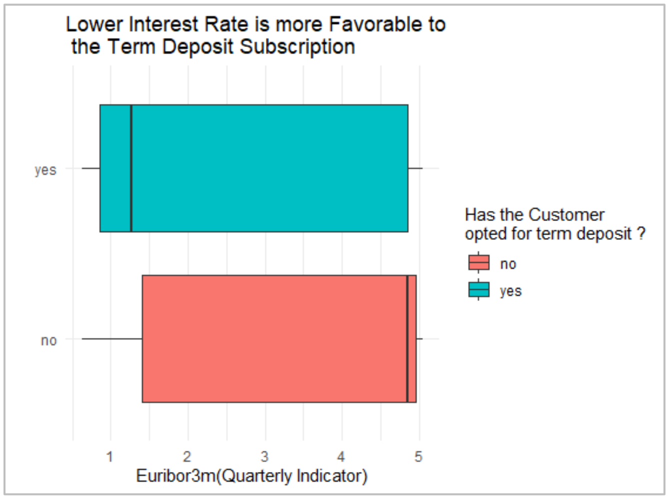


Figure 5 : Plot showing impact of 3 Months Euribor Interest Rate on subscription to Term Deposits

The 3-month Euribor interest rate is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months. From **Figure 5** it is clearly observed that for higher values of euribor3m we see that a majority of customers don’t subscribe to term deposits and vice versa for lower values of euribor3m.