OM386 Advanced Data Analytics in Marketing

Final Project

Campaign Outcome Prediction

Vishal Gupta & Vishu Agarwal

Background & Problem Statement

Marketing campaigns are characterized by focusing on customer needs and their overall satisfaction. Nevertheless, different variables determine whether a marketing campaign will be successful or not. Some important aspects of a marketing campaign are as follows:

1. *Segment of the Population:* To which segment of the population is the marketing campaign going to address and why? This aspect of the marketing campaign is extremely important since it will tell which part of the population should receive the message of the marketing campaign.
2. *Distribution channel to reach the customer's place:* Implementing the most effective strategy to get the most out of this marketing campaign. What segment of the population should we address? Which instrument should we use to get our message out? (Ex: Telephones, Radio, TV, social media, etc.)
3. *Promotional Strategy:* This is the way the strategy is going to be implemented and how potential clients are going to be addressed. This should be the last part of the marketing campaign analysis since there has to be an in-depth analysis of previous campaigns and their outcome (if possible) to learn from previous mistakes and to determine how to make the marketing campaign much more effective.

In this project, we will analyze the campaign effectiveness for a banking institution and bring insights on what features of a customer can help us understand their conversion after the campaign. There has been a revenue decline for the bank and the root cause is that their clients are not depositing as frequently as before. Term deposits allow banks to hold onto a deposit for

a specific amount of time, so banks can lend more and thus make more profits. In addition, banks also hold a better chance of persuading term deposit clients into buying other products such as funds or insurance to further increase their revenues. Details about the dataset and research objective are below -

- *Dataset:* Dataset is obtained from the open-source platform Kaggle (link), and it contains the details of marketing campaigns done via phone with various details for customers such as demographics, last campaign details, etc. and the goal is to predict accurately whether the customer will subscribe to the focus product for the campaign - Term Deposit after the campaign?
- *Research Objective:* Identifying the customers beforehand that have a high propensity to subscribe to a term deposit before the campaign can help run more effective campaigns as instead of reaching out to a bigger audience the marketing team can focus more on a smaller group thereby increasing the effectiveness of the campaign. The analysis can also help understand what the key attributes of a customer and the engagement level during the previous campaign are to predict whether the customer will lead to a positive outcome or not.

Data Summary & Exploratory Analysis

We want to predict the outcome of the campaign – term deposit subscribed or not – hence we will treat the problem as a binary classification problem. Before doing the modeling, we need to do feature engineering to ensure that the data is in the correct format so that it can be fed to the model. Feature engineering for the problem primarily includes outlier detection & removal, missing value treatment & one-hot encoding.

Dataset summary

The data contains 31.6K records and 18 features. The features provide information about the customers such as their age, job type, marital status, education attained, default status, loans taken, and previous campaign information. Based on these features, we need to predict whether the customer subscribes to a term deposit or not.

Below table provides a quick description of all the features.

Feature Name	Description
ID	Unique ID for a customer
Customer Age	Age of the customer
Job Type	Job Type of the customer – Blue collar, technician, Services, housemaid, unemployed, etc.
Marital	Marital Status of the customer – Single, Married, Divorced
Education	Education level attained by the customer – Primary, Secondary, Tertiary, Unknown
Default	Indicates if the customer has defaulted on loans historically
Balance	Financial account balance of the customer
Housing Loan	Indicates if the customer has a housing loan
Personal Loan	Indicates if the customer has a personal loan
Communication Type	Preferred mode of communication of the customer – Cellular, Telephone, Unknown
Day of Month	Day of the month of the previous contact with customers
Month	The month of the previous contact with customers
Last Contact Duration	Duration of the last contact made with the customer
Num Contacts in campaign	Number of contacts the customer has in the current campaign
Days since prev campaign contact	Number of days since a customer was contacted during the previous campaign
Num contacts prev campaign	Number of contacts the customer had in the previous campaign
Prev campaign outcome	The outcome of the given customer during the previous campaign
Term deposit subscribed	Indicates whether the customer subscribed for a term deposit

Post spending some time understanding the features and the data, we started looking at all the features for missing values. We found that the following features had missing values –

Feature Name	% Records with Missing Values
Customer Age	2%
Marital	0.5%
Balance	1.3%
Personal Loan	0.5%
Last Contact Duration	1%
Num Contacts in campaign	0.4%
Days since prev campaign contact	81.6%

Missing values treatment

1. Customer Age: It was difficult to obtain age using other features, and there were only 2% of records with missing age. Hence, we excluded the records with missing age
2. Marital Status: Imputed the missing values with 'Unknown'
3. Balance: Since there are just 1.3% of records with missing values of account balance, we excluded those records
4. Personal Loan: There are just 0.5% of records with missing information about Personal Loan. Hence, we excluded those records
5. Last Contact Duration: Imputed the missing values with the median value of the feature
6. Num contacts in the current campaign: Imputed the missing values with the median value of the feature
7. Days since prev campaign contact: This feature had missing values for more than 80% of the records. Hence, we created a new binary column indicating whether this column has a missing value or not

Now that we have treated the missing values, our next step was to look at the distribution of all the columns and assess if outlier treatment was required on the features.

Outlier detection and treatment

None of the features except 'Num contacts in prev campaign' seem to have an outlier. 'Num contacts in prev campaign' has one outlier with a value of > 250 while other values are well below 100. Hence, we have removed the record with the outlier value.

While looking at the distribution of all the features to detect outliers, we also observed that some of the features have a sparse distribution and hence we performed some transformations to them to enable better modeling.

Feature Transformation

All the numerical variables are highly skewed since this can affect the performance of any model, we have tried to use log transformation to make the independent variable less skewed

and more normal-like. Also, other transformations and new variables are created based on the understanding of the dataset.

Below is a summary of the transformations applied to a few of the columns –

Feature Name	Transformation Applied
Balance	1. Log Transformation 2. Created another feature to indicate positive/negative balance
Last contact duration	Log Transformation
Number of contacts in the current campaign	
Number of contacts in prev campaign	

1. Balance

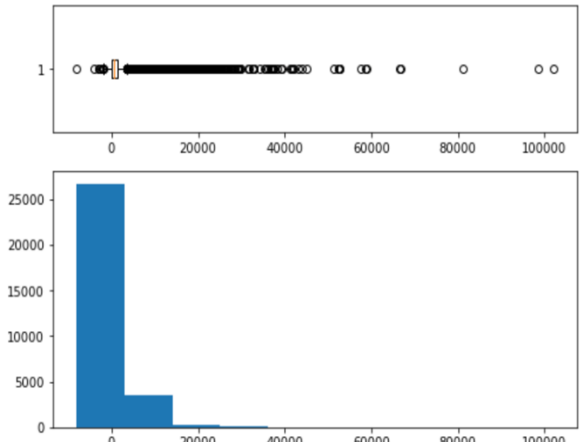


Figure 1: Before Log Transformation

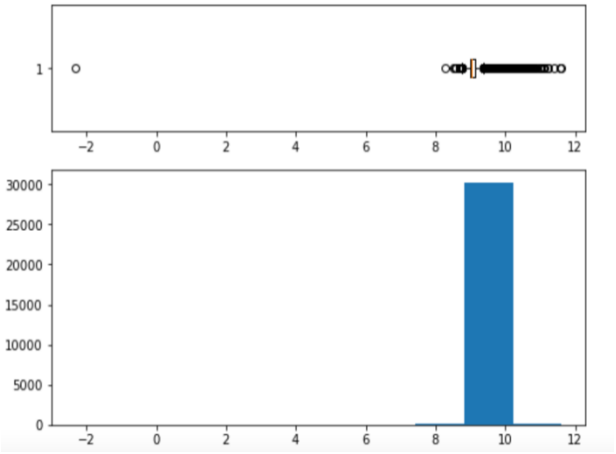


Figure 2: After Log Transformation

2. Last Contact Duration

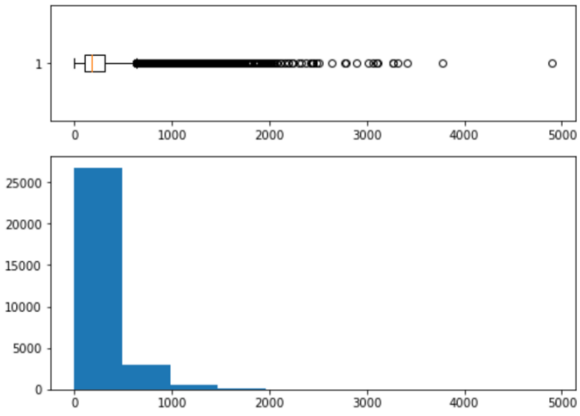


Figure 3: Before Log Transformation

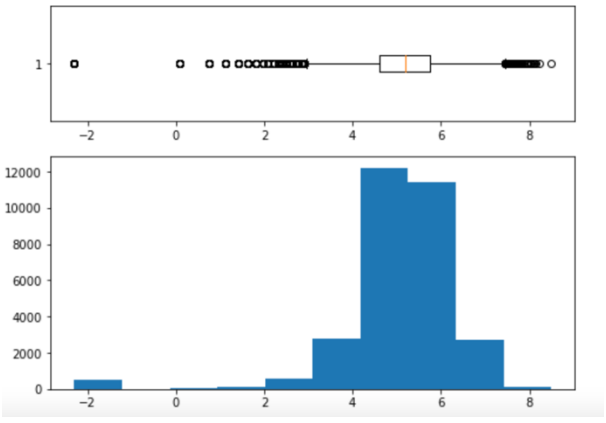


Figure 4: After Log Transformation

3. Number of contacts in the campaign

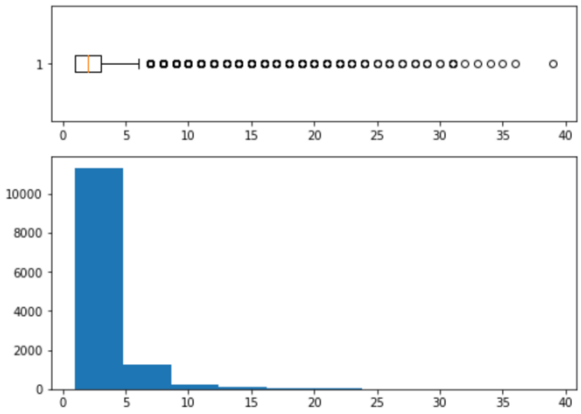


Figure 5: Before Log Transformation

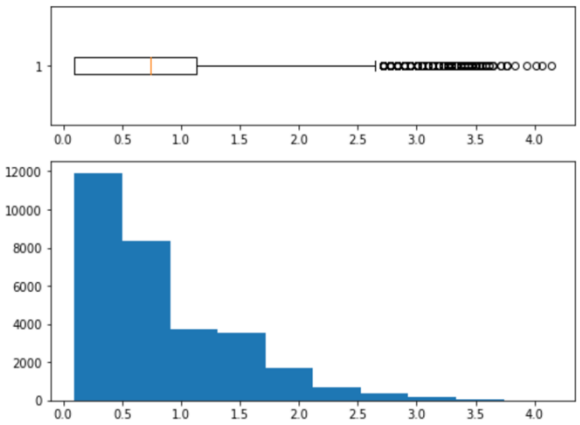


Figure 6: After Log Transformation

5. Number of contacts in the previous campaign

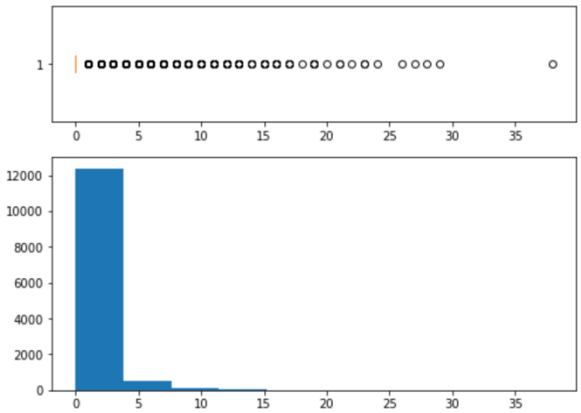


Figure 7: Before Log Transformation

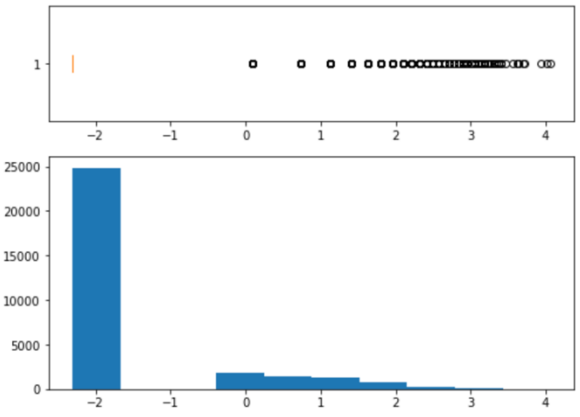


Figure 8: After Log Transformation

Feature relationships

After missing values treatment, outlier detection & removal, and feature transformation, our next step was to –

1. Assess the correlation among the features
2. Analyze bi-variate plots between the features and y – variable to assess the qualitative importance of features

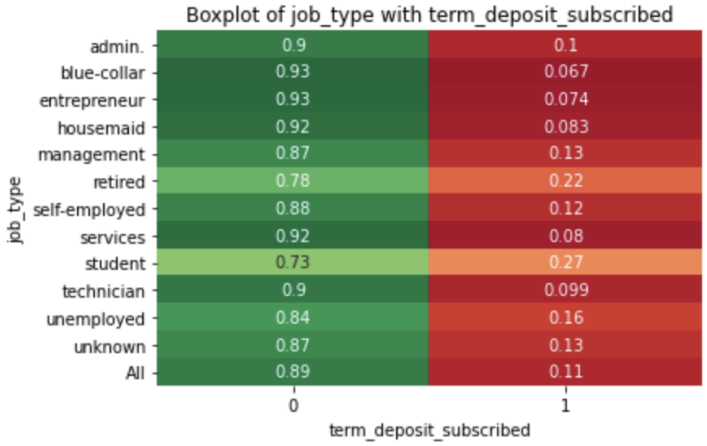
1. Correlation among the features

We can observe that the correlation is only high between the original variable and their transformed versions, otherwise, it is well below 0.5. Since we will not be using the non-transformed versions of the features in the model, we should be good from a multi-collinearity point of view.

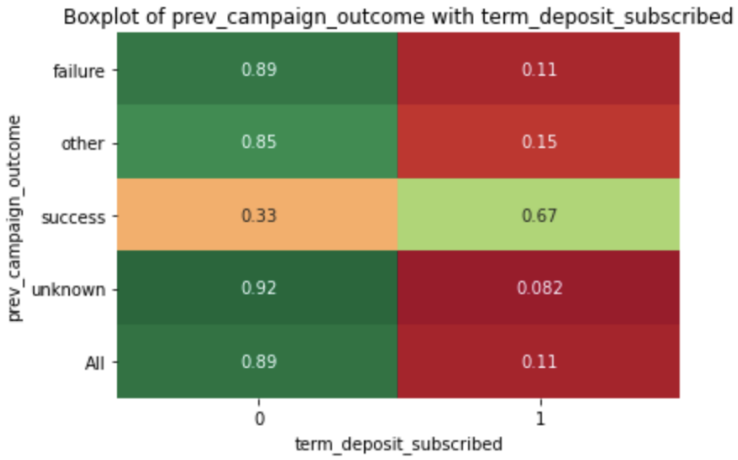
2. Bi-variate plots between the features and Y-variable

Across all the bi-variate plots, few significant ones have been shown here. The rest of the plot has been placed in the appendix A.1.

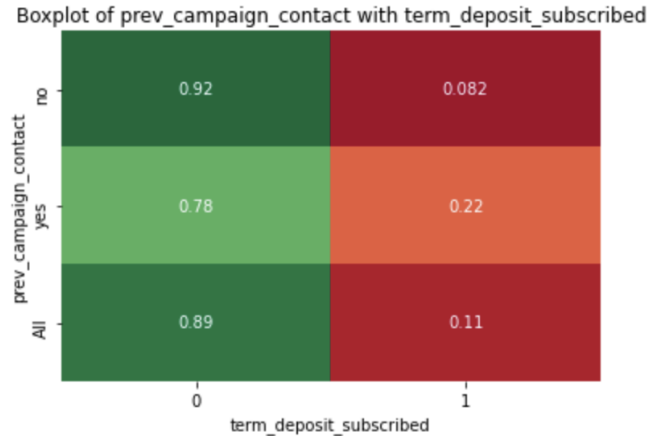
Job Type vs. Term Deposit Subscribed- Student & retired customers tend to subscribe to a term deposit more than customers with other job types.



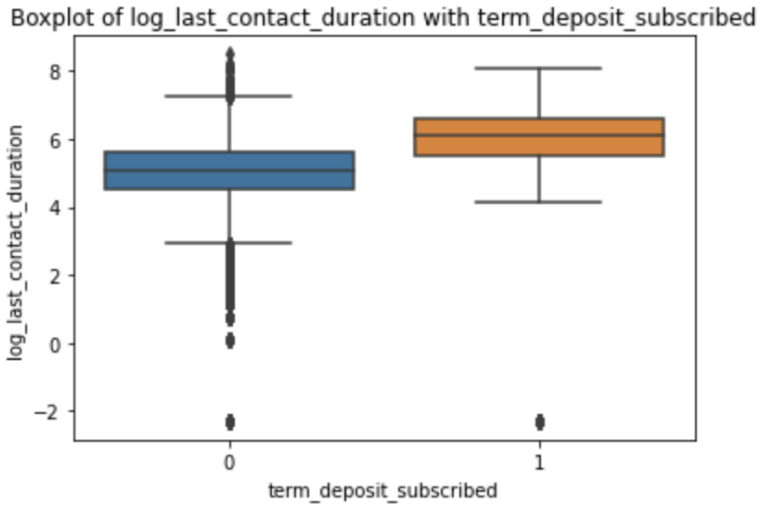
Previous Campaign Outcome vs. Term Deposit Subscribed- The probability of success of the current campaign increases significantly if the customer has subscribed to a term deposit after the previous campaign.



Previous Campaign Contact vs. Term Deposit Subscribed- The probability of success of the current campaign increases significantly if the customer has been contacted in the previous campaign.



Log Last Contact Duration vs. Term Deposit Subscribed - The probability of success of the current campaign increases significantly if the customer has been in contact for a longer duration.



Methodology & Analysis

Once data is prepared for the model, we initially tried some of the tree-based ensemble machine learning models - random forest and xGBoost, followed by statistical models - logit and mixed-effects linear models.

Evaluation metric for modeling

We want to focus more on the customers who are more likely to create a term deposit after the campaign, as it can help identify the attributes in a customer that can make future campaigns more effective. Accuracy might not be the best metric as we have an unbalanced class case where most of the customers (>80%) do not subscribe to the term deposit. *Recall* (sensitivity or true positive rate) will be the primary metric as it evaluates the model on reducing false negative (where customers made a term deposit, but the model predicted otherwise). The second metric of focus will be *Precision*, which evaluates false positives for model performance, since there are very few customers that subscribe to a term deposit, predicting them as positive will not increase the campaign load significantly. Also, these can be the customers that have the potential to subscribe to the term deposit in future campaigns.

To enable this, we down sampled the records with Class 0 which resulted in a balanced class. One potential downside to down sampling is that we lose some data. However, reduced data also helped us overcome computational obstacles faced during tuning the models.

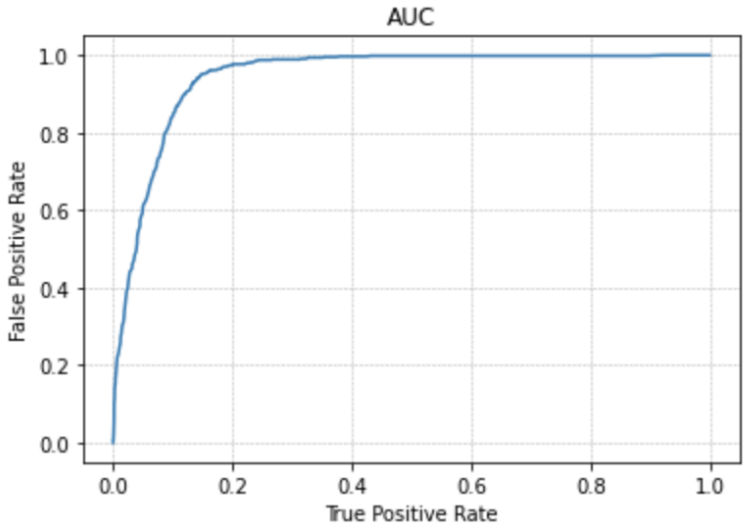
Class	% Records (before down sampling)
0	90%
1	10%

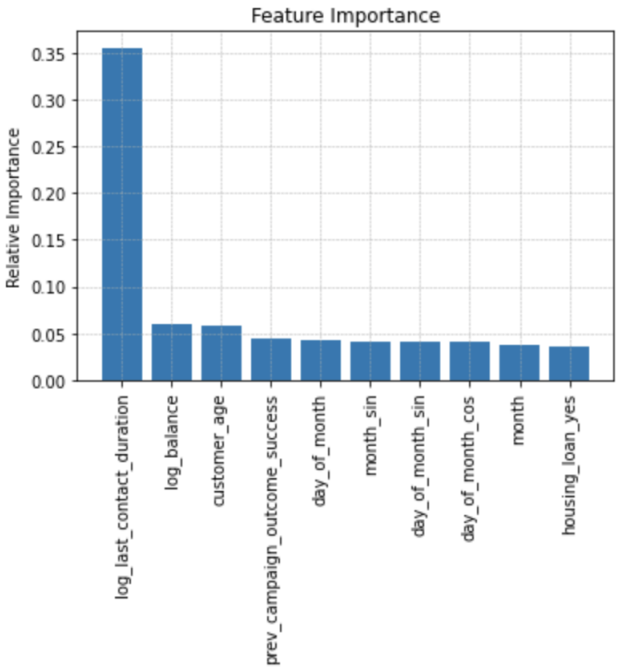
Tree-based ensemble models

1. *Random Forest*

We trained the model on down-sampled data and tuned the hyperparameters as well by maximizing the recall – obtained 95% recall and 43% precision. This means that of all the potential subscribers to term deposits, our model was able to predict 95% of them. Also, of all the customers that our model predicted as subscribers to the term deposit, 43% of them are actual subscribers.

Metric	Measure
Recall	95%
Precision	43%
Accuracy	86%
AUC	95%

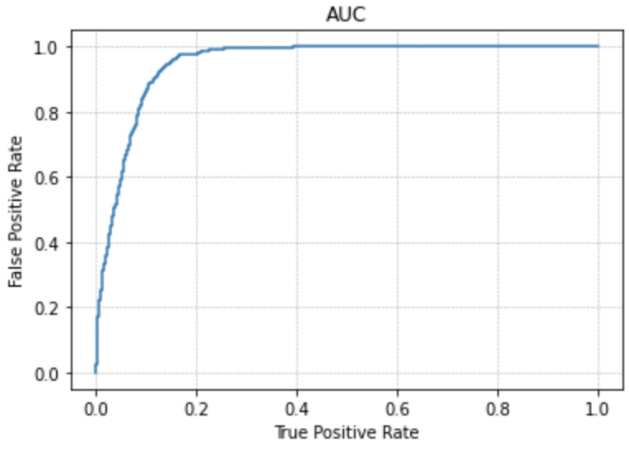


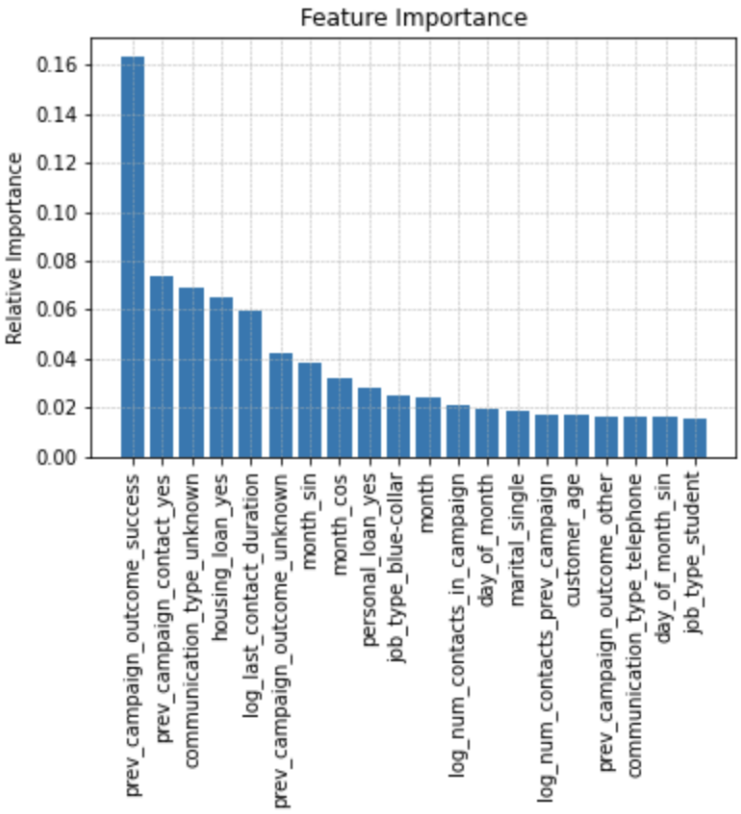


2. *<u>XgBoost</u>*

We tried XgBoost as well to see if we can further improve precision and recall. By tuning the hyperparameters, we obtained marginally improved accuracy, with slightly reduced recall.

Metric	Measure
Recall	94%
Precision	45%
Accuracy	87%
AUC	95%





Logit Models – Fixed effects and Random effects

Logit models are used to model the probability of an event happening by assuming log-odds to be a linear function of independent variables. Logit models are easier to understand and can help understand the effect of dependent variables on the probability of the event happening, in our case, the probability of a customer subscribing term deposit after being contacted through the campaign.

Due to a high number of variables (numeric and multi-level categorical) in the dataset, we have first tried a few formulations to set up the baseline and understand the importance of the variable in prediction. After that, we will introduce random effects and interaction terms as features as well as reduce some variables based on the baseline model's feature importance. For the mixed-effects model, we have used Bayesian estimation using the MCMC package as likelihood estimation was not converging even after hours of execution due to high cardinality.

Different variations of the logit model tried are in A.2, the performance of those models is in the table below -

Model variant	Recall	Precision
Baseline	71.98%	46.5%
Transformed variables - balance indicator and log-transformed	72.45%	42.66%
Cyclicity– month and day	71.98%	42.39%
Random effect on intercept – Job type	85.14%	34.33%
Random effect on intercept – Marital	84.67%	34.14%
Random effect on intercept – Education	84.06%	33.87%
Random effect on intercept, marital status & education – job type	82.97%	34.63%
Random effect on intercept, marital status & education – job type, variable selection, and the interaction term of balance and previous campaign contact	98.3%	15.89%

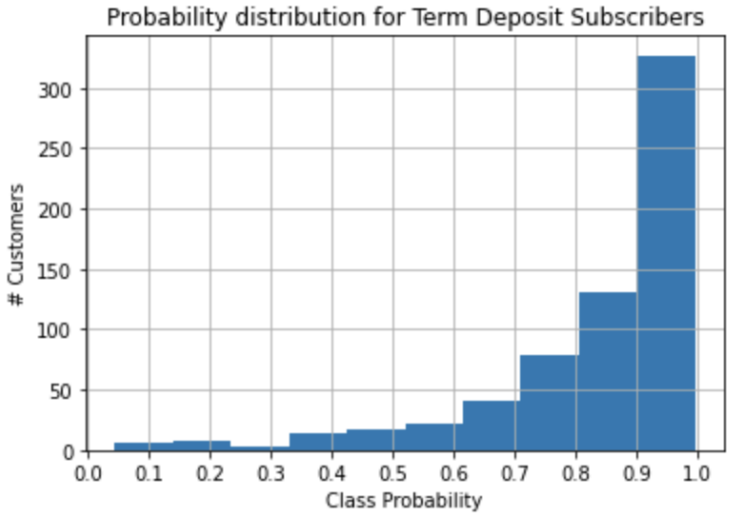
Results & Insights

Tree-based ensemble models

Random Forest and XgBoost are decision tree-based ensemble methods based on bagging and boosting methods, respectively. Both are widely used in classification problems due to their high performance and explainability.

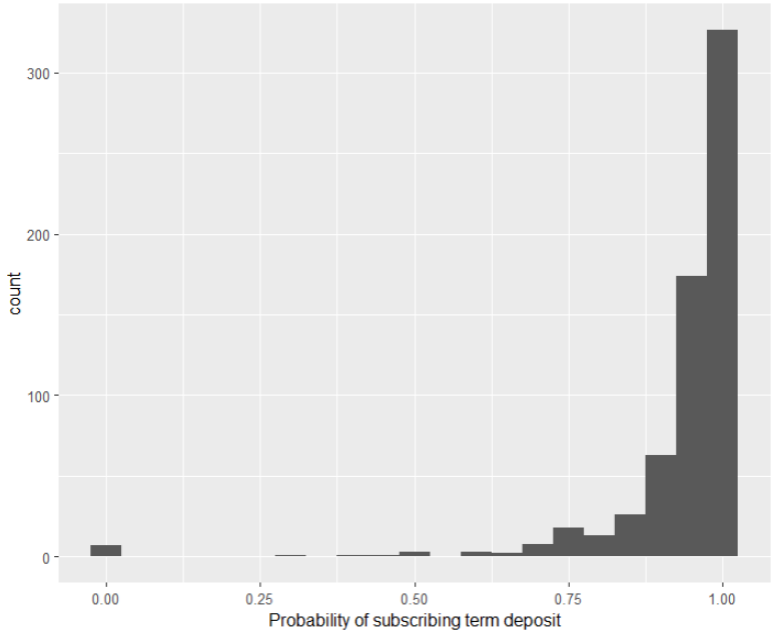
Both the models yield comparable results for recall and precision (~94% and ~45% respectively) with Random Forest being much faster in terms of computation. Also, features corresponding to the last campaign – duration, success, mode of communication, number of contacts, etc. seem to be the most notable features in determining whether the customer will subscribe for a term deposit.

For the test dataset, out of a total of 646 consumers that subscribed to a term deposit, we can accurately classify 605 of them with the model assigning them a probability of more than 50% missing 41 of such customers. The probability distribution curve of the positive class in test data is as follows -



Logit Models

As we can observe, the final improvements with variables selection, adding interaction terms of customer balance & previous campaign contact with random effect is performing the best among both tree-based models and logit models with 98.3% recall. The probability distribution curve of positive classes is as below –



For the test dataset, out of a total of 646 consumers that subscribed to a term deposit, we can accurately classify 635 of them with the model assigning them a probability of more than 50%, missing only 11 of such customers. Based on the model results we can expect to reduce the

customer base by 34.5% though losing only 1.7% of customers who would have subscribed to term deposits after the campaign. Considering we use the model to support a similar campaign, instead of reaching out to 6098 customers via this campaign where only 646 customers ended up subscribing to the term deposit, we can use the model probability estimate to focus only on 3996 customers but still get 635 customers to subscribe the term deposit. This can help in reducing costs and creating a more directed and relevant campaign for both banking institutions and their customers.

Interpretation & Insights

To identify the most significant variable in the model, the independent variables that affect the log odd ratio the most, we will look at the absolute coefficient values. A total of 12 variables are selected with 6 having high positive coefficient values and 6 with high negative values. We will also look at the distribution of Markov chains to evaluate the significance of these variables, as just having a high positive or negative value may not be sufficient. The Markov chain distribution of all the 12 coefficient values is available in A.3.

Top 12 variables with a high negative and positive value of coefficient are as below –

Independent Variables	Coefficient values
prev_campaign_outcomesuccess	2.844
log_last_contact_duration	1.578
intercept.retired	0.893
intercept.self_employed	0.792
intercept.housemaid	0.584
intercept.student	0.568
intercept	-8.660
intercept.job_type_unknown	-2.328
communication_typeunknown	-0.933
housing_loanyes	-0.931
intercept.management	-0.518
personal_loanyes	-0.487

Based on the chain distribution, we can infer that only the *Job-type: Student* variable cannot be significant as 0 lies well within the distribution range, but the rest of the other variables in the table above can be considered statistically significant. The high intercept value indicates that by

default the probability of subscribing to the term deposit will be lower, and it could increase significantly for the following cases –

1. if the outcome of the previous campaign with the customer is positive
2. The current campaign contact duration was high
3. if the customer is retired, self-employed, a housemaid, or a student

Similarly, the probability of the customer subscribing to the term deposit will go down for the following cases –

1. if the job type is unknown or management
2. if the communication channel is unknown for the campaign
3. if the customer already has housing or a personal loan

Recommendation & Future Scope

Based on the mixed-effects logit model we have formulated recommendations for the banking institutions to focus on certain demographics and run a more directed campaign rather than reaching out to more customers with ineffective conversion.

The focus group for the campaign can be –

- Retired customers or students
- Low-income group customers – self-employed, housemaids, etc.
- The customers that have been part of the previous campaign and have subscribed to a term deposit after the campaign ended
- Customers that do not have a personal or housing loans
- Avoid customers who are in management or entrepreneurs - those who know how to manage their money a little better

Campaign execution improvements –

- Reduce the total number of customers to focus on by using the model output to only target those customers with a higher probability to convert

- Focus more time on each customer, as the overall base is reduced by leveraging the model output since higher contact duration can lead to a higher probability of a customer subscribing to the term deposit
- Communication type for each customer to be made via telephone
- Contact the customers converted in the previous campaign as they are low hanging fruits

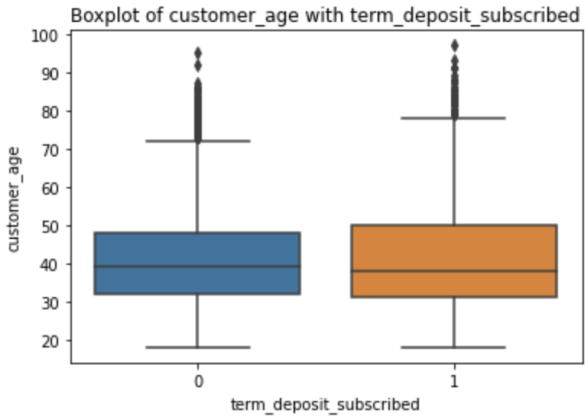
Mixed-effects models have performed a lot better, although the estimation of the parameters is done using Bayesian methods, the number of iterations has been restricted to only 1000 due to run time constraints. In the future, with more computational power, the Markov chain should be allowed to run for more iterations to have higher confidence in the estimated parameters' value.

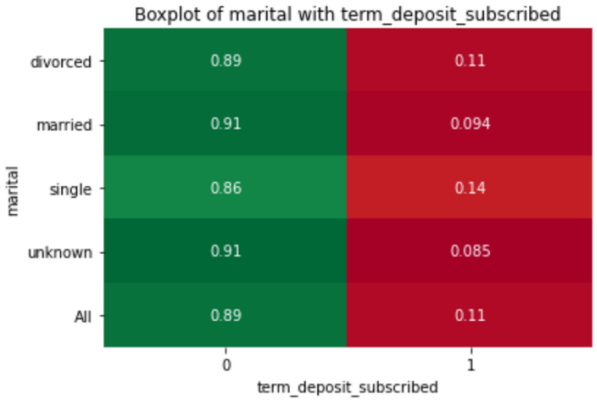
Under sampling is done to reduce the training data size which has allowed us to try out multiple iterations of the logit model faster, with higher computation power, we can also explore oversampling techniques to resample or synthetically generated minority samples that can help learn the outcome of positive class a lot better.

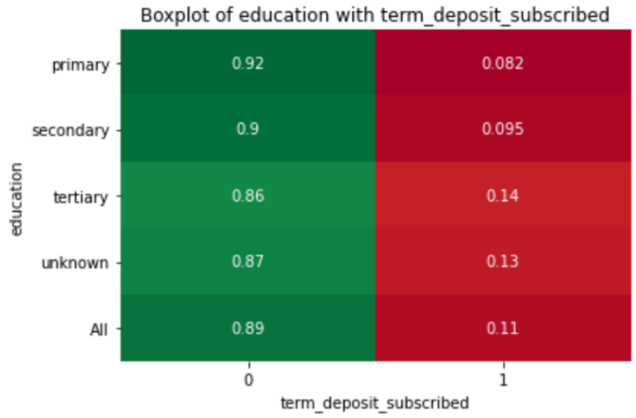
Finally, the dataset has a high number of numeric variables as well as categorical variables with multiple levels, this has led to high cardinality and can affect the estimation of parameters. Correlation analysis of independent variables can also be done, along with the possible interaction of different independent variables, to identify the variables that are highly correlated with outcome variable – term deposit subscribed, this can help reduce the cardinality in the dataset and allow us to explore more interaction terms that can also help improve upon the recall and precision further.

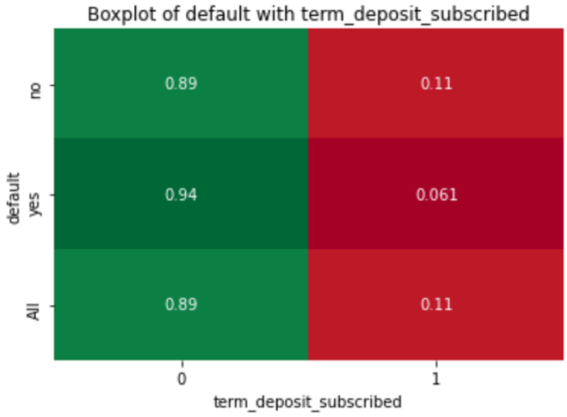
APPENDIX

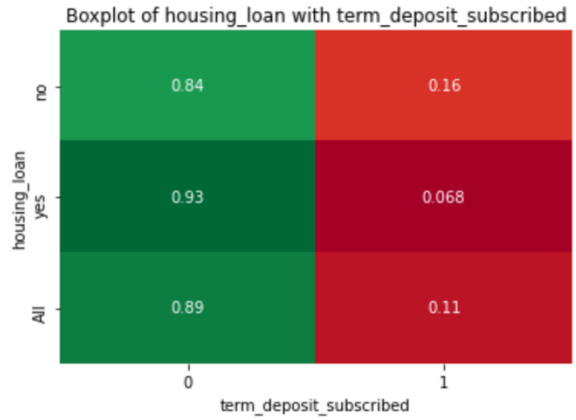
A.1 Bivariate plots between features and Y-variable

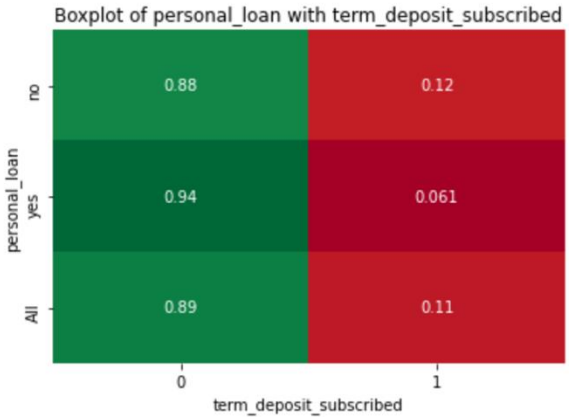


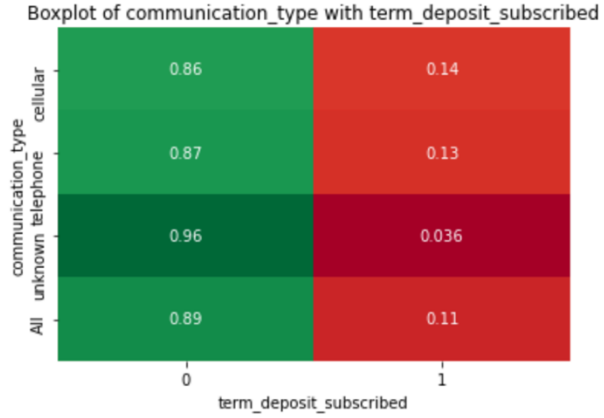


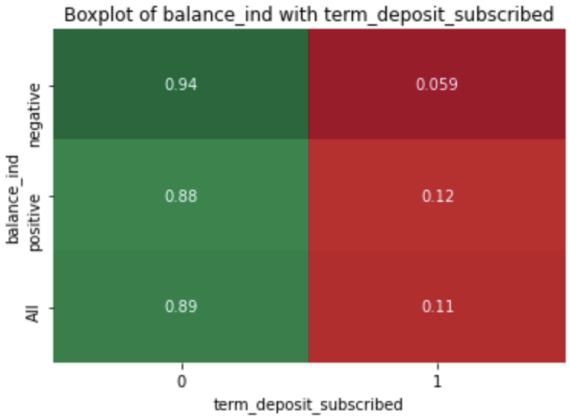


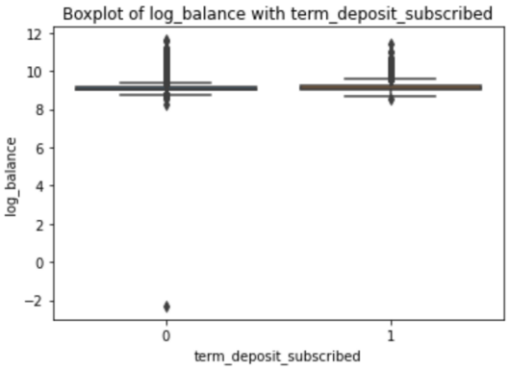


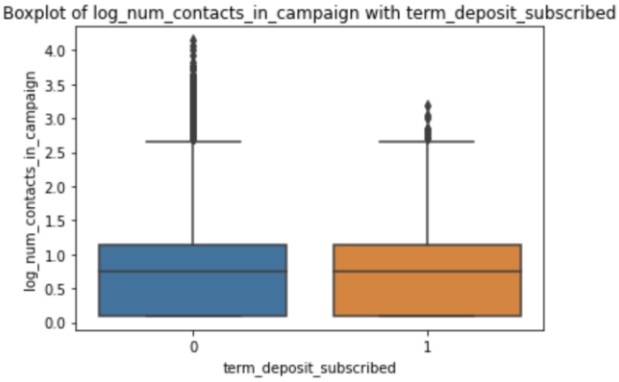


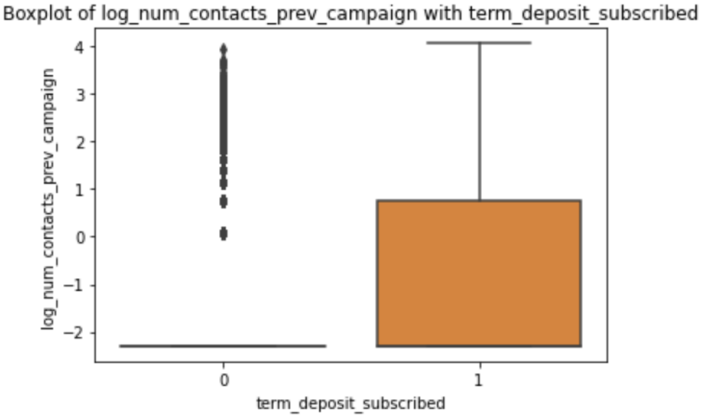



Boxplot of communication_type with term_deposit_subscribed
communication_type
cellular
telephone
unknown
All
0.86
0.14
0.87
0.13
0.96
0.036
0.89
0.11
0
1
term_deposit_subscribed


Boxplot of balance_ind with term_deposit_subscribed
balance_ind
negative
positive
All
0.94
0.059
0.88
0.12
0.89
0.11
0
1
term_deposit_subscribed


Boxplot of log_balance with term_deposit_subscribed
log_balance
12
10
8
6
4
2
0
−2
0
1
term_deposit_subscribed


Boxplot of log_num_contacts_in_campaign with term_deposit_subscribed
log_num_contacts_in_campaign
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
term_deposit_subscribed


Boxplot of log_num_contacts_prev_campaign with term_deposit_subscribed
log_num_contacts_prev_campaign
4
3
2
1
0
−1
−2
0
1
term_deposit_subscribed

A.2 Different model variants for logit models

Model variant	Formulation
Baseline	term_deposit_subscribed ~ job_type + marital + education + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance + last_contact_duration + num_contacts_in_campaign + num_contacts_prev_campaign + month + day_of_month + customer_age
Transformed variables - balance indicator and log transformed	term_deposit_subscribed ~ job_type + marital + education + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind + log_balance + log_last_contact_duration + log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month + day_of_month + customer_age
Cyclicity– month and day	term_deposit_subscribed ~ job_type + marital + education + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind + log_balance + log_last_contact_duration + log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month_sin + month_cos + day_of_month_sin + day_of_month_cos + customer_age
Random effect on intercept – Job type	term_deposit_subscribed ~ marital + education + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind + log_balance + log_last_contact_duration + log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month_sin + month_cos + day_of_month_sin + day_of_month_cos + customer_age + (1| job_type)
Random effect on intercept – Marital	term_deposit_subscribed ~ job_type + education + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind + log_balance + log_last_contact_duration + log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month_sin + month_cos + day_of_month_sin + day_of_month_cos + customer_age + (1| marital)
Random effect on intercept – Education	term_deposit_subscribed ~ job_type + marital + default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind + log_balance + log_last_contact_duration +

	log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month_sin + month_cos + day_of_month_sin + day_of_month_cos + customer_age + (1| education)
Random effect on intercept, marital status & education – job type	term_deposit_subscribed ~ default + housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact + balance_ind_positive + log_balance + log_last_contact_duration + log_num_contacts_in_campaign + log_num_contacts_prev_campaign + month_sin + month_cos + day_of_month_sin + day_of_month_cos + customer_age + (1 + marital + education|job_type)
Random effect on intercept, marital status & education – job type, variable selection and interaction term of balance and previous campaign contact	term_deposit_subscribed ~ housing_loan + personal_loan + communication_type + prev_campaign_outcome + prev_campaign_contact_yes:log_num_contacts_prev_campaign + balance_ind_positive:log_balance + log_num_contacts_in_campaign + log_last_contact_duration + month_sin + day_of_month_cos + (1 + marital + education|job_type)

A.3 Markov chain coefficient value for important coefficients – high positive or negative value in the final model