Report for:

BANKING, FINANCIAL SERVICES AND INSURANCE (BFSI)

BANK MARKETING

Name: Surampudi Vnss Siva Krishna

Email: Surampudi3311@gmail.com

Phone: 8790464715

Contents

Part 1 Project Background

Nowadays, marketing expenditures in the banking industry are massive, meaning that it is essential for banks to optimize marketing strategies and improve effectiveness. Understanding customers' need leads to more effective marketing plans, smarter product designs and greater customer satisfaction.

Main Objectives: predict customers' responses to future marketing campaigns & increase the effectiveness of the bank's telemarketing campaign

This project will enable the bank to develop a more granular understanding of its customer base, predict customers' response to its telemarketing campaign and establish a target customer profile for future marketing plans.

By analysing customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviours and identify which type of customers is more likely to make term deposits. The bank can then focus its marketing efforts on those customers. This will not only allow the bank to secure deposits more effectively but also increase customer satisfaction by reducing undesirable advertisements for certain customers.

Part 2 About the Data

There are 45,211 observations in the dataset, with no missing values. Each represents an existing customer that the bank reached via phone calls. For each observation, the dataset records 17 input variables that stand for both qualitative and quantitative attributes of the customers.

There is a single binary output variable that denotes "yes"(1) or "no(0)" revealing the outcomes of the marketing phone calls. "Yes" means that a customer subscribed to term deposits.

Part 3 Data Cleaning

Several changes were made to the dataset to prepare it for analysis.

There are unknown values for many variables in the data set. Variables with unknown/missing values are: 'education', 'job', 'poutcome', 'contact'. By observing these features, we found way of doing an imputation where we use other independent variables to infer the value of the missing variable.

Therefore, we start with creating new variables for the unknown values in 'education', 'job'. We do this to see if the values are missing at random or is there a pattern in the missing values.

3.1 Imputation:

Now, to infer the missing values in 'job' and 'education', we make use of the cross-tabulation between 'job' and 'education'. Our hypothesis here is that 'job' is influenced by the 'education' of a person.

Hence, we can infer 'job' based on the education of the person. Moreover, since we are just filling the missing values, we are not much concerned about the causal inference. We, therefore, can use the job to predict the education.

Inferring education from jobs : From the cross-tabulation, it can be seen that people with management jobs usually have a university degree. Hence wherever 'job' = management and 'education' = unknown, we can replace 'education' with 'tertiary'. Similarly, 'job' = 'services' --> 'education' = 'secondary and 'job' = 'housemaid' --> 'education' = 'primary.

Inferring jobs from age : As we see, if 'age' > 60, then the 'job' is 'retired,' which makes sense.

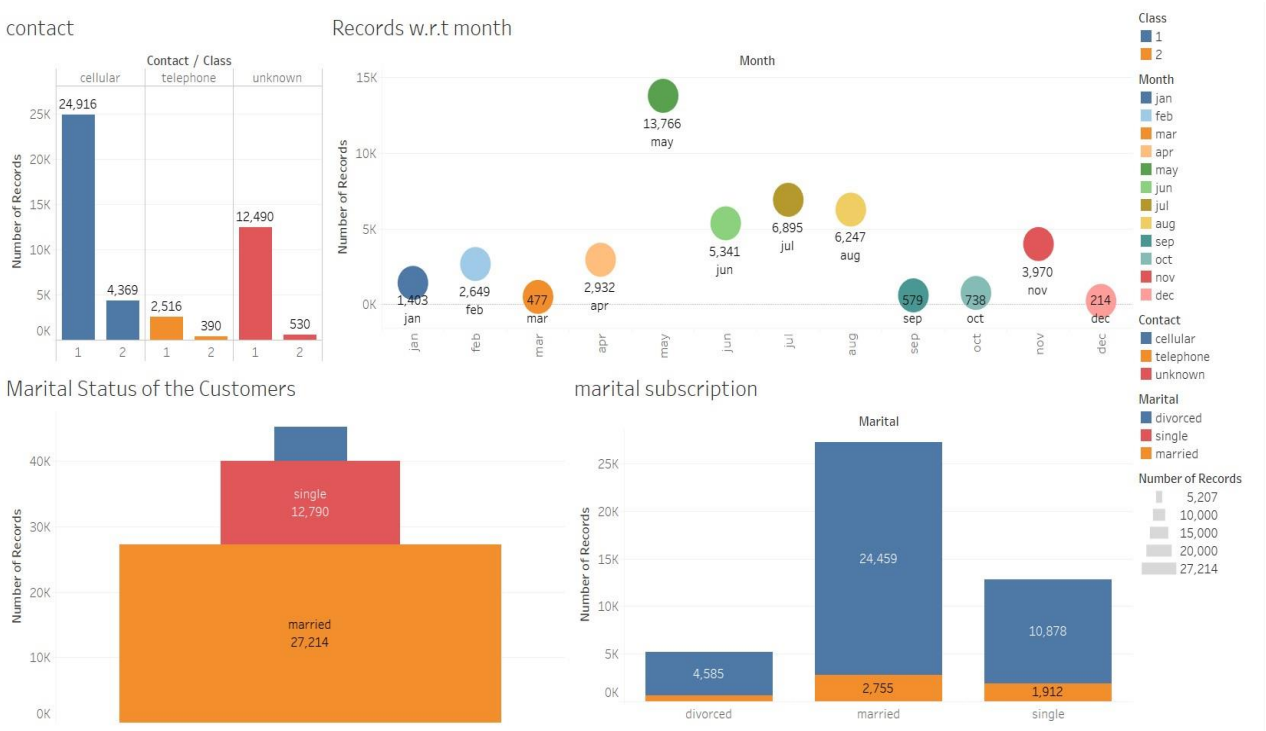
While imputing the values for job and education, we were cognizant of the fact that the correlations should make real world sense. If it didn't make real world sense, we didn't replace the missing values.

We haven’t found such any hidden pattern between poutcome and contact and we haven’t changed unknowns in these features.

Change the "response" variable (yes/no) to binary values (1/0) for easier analysis.

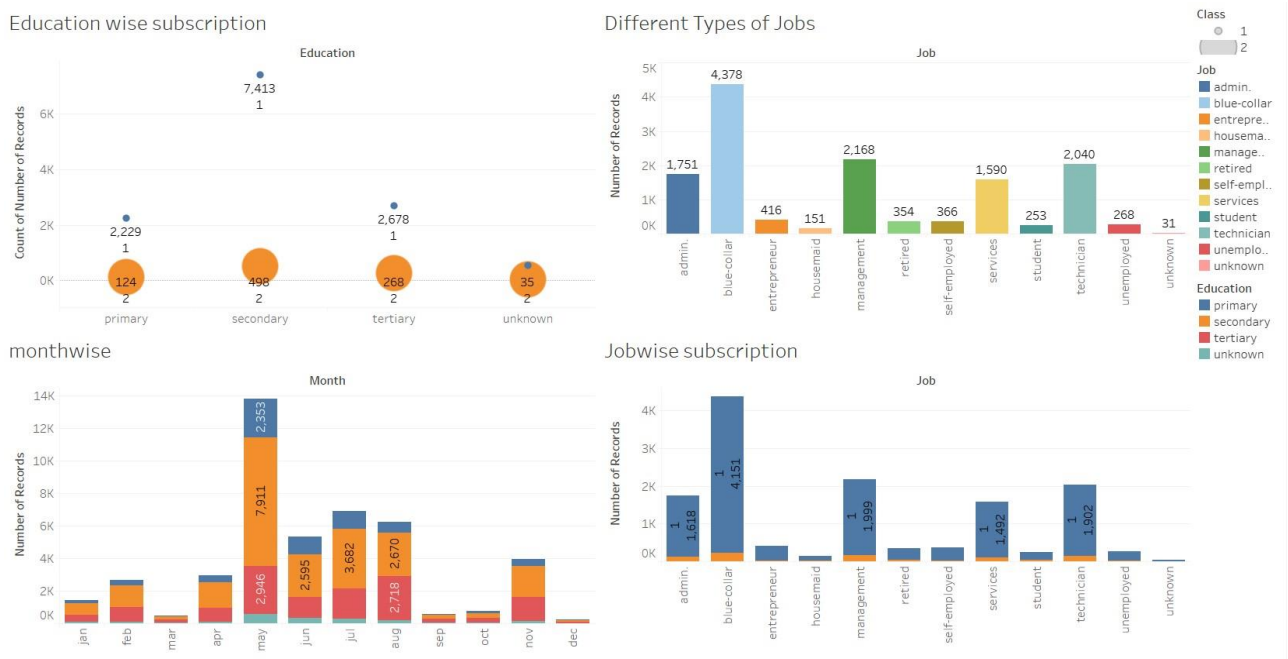
Part 4 Exploratory Data Analysis

4.1 Bivariate Analysis



The contact plot signifies that most of the people were contacted by the bank through cellular network more than 25,000 and among them more than 4,000 people have subscribed term deposit. From this we can infer that the bank should focus on the people who are available to contact them on cellular network.
Even unknown contact type records are also of high number approximately 12,400. So, bank should take some initiatives to identify which contact type those unknowns are using and separate them into cellular, telephone or make a new cluster.

We noticed a hidden pattern from “Records with respect time plot”. Majority of the records were identified in the month may followed by April, November and others. Similarly, lowest records were identified in the months March, December, June and others. So, bank should not focus directly on the months which are having highest records instead of that they have to focus on the months at which most the subscriptions happening.

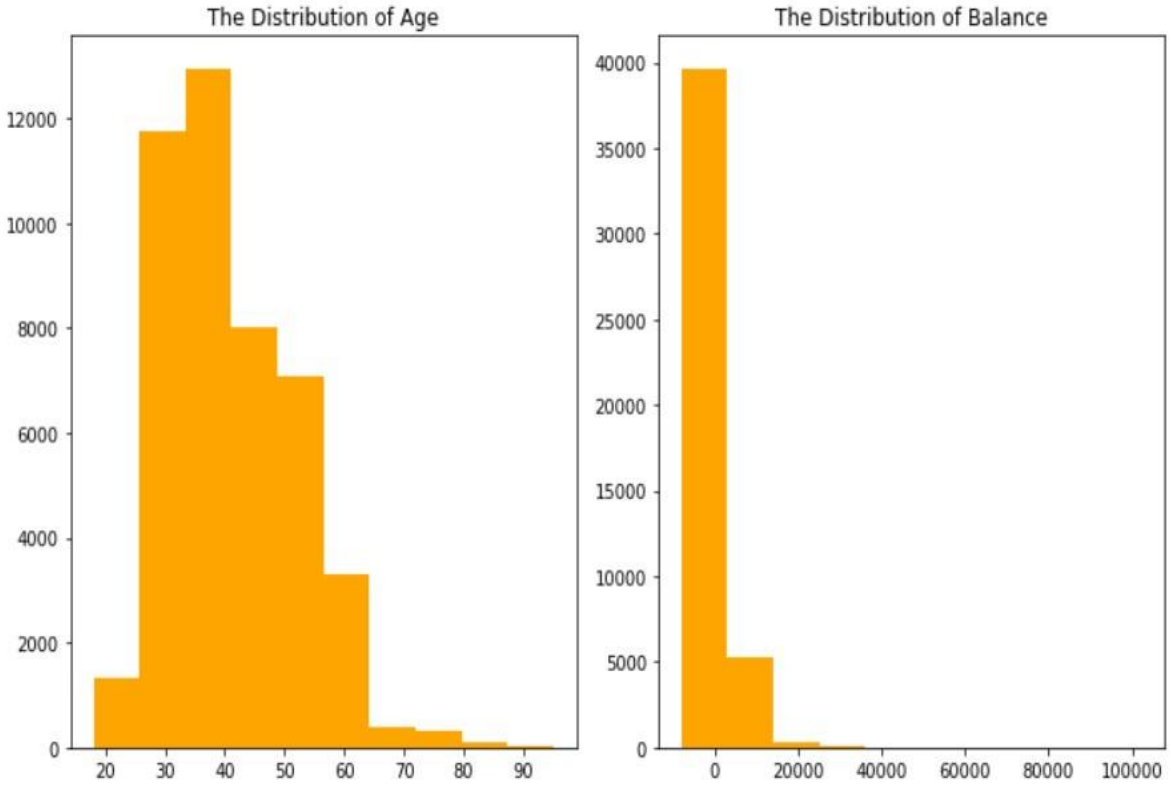


In the given data it is observed that there are three types of educated people (primary, secondary, tertiary) and unknown. Among these secondary educated people are doing more subscriptions followed by tertiary, primary. So, the bank should focus their marketing on the tertiary educated people first and in the month of the may followed by other months. In the given data it is observed that there are different types of jobs such as admin, management, student, blue collar and others. Among these blue collar job records are high approximately 4,000 and subscription wise also blue collar job holders are doing more term deposits. This signifies that bank should focus on doing marketing on blue collar job holder's followed by management, admin, and others which leads to good number of the people turn towards subscriptions.

In the given data it is observed that there are different types of jobs such as admin, management, student, blue collar and others. Among these blue collar job records are high approximately 4,000 and subscription wise also blue collar job holders are doing more term deposits. While considering these job categories bank should also consider housing and personal loan into consideration.

From the above graph it is observed that the people with no loan are doing high term deposits compared to people with loan burdens. This signifies that bank should focus on doing marketing on blue collar job holder's followed by management, admin, and others who are not having loans. This leads to good number of the people turn towards subscriptions.

4.2 Visualize the distribution of customer age and balance levels:



The distribution of age:

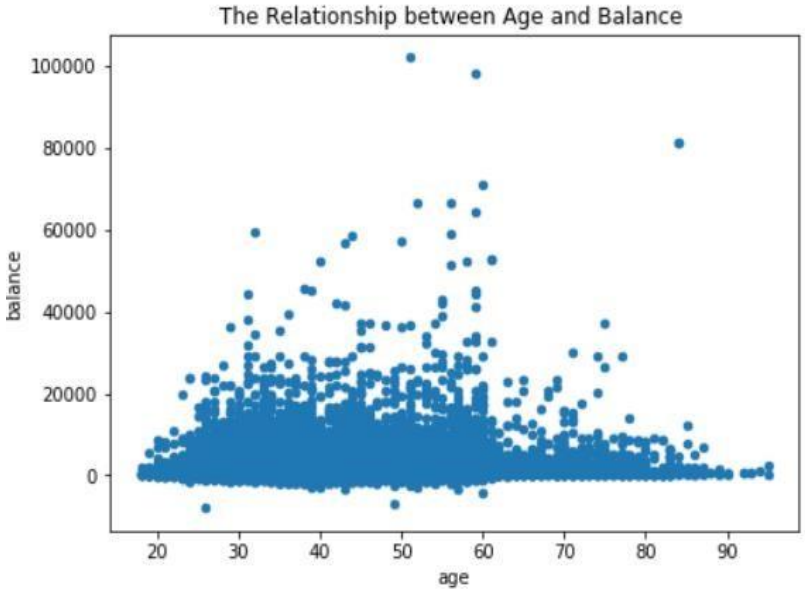
In its telemarketing campaigns, clients called by the bank have an extensive age range, from 18 to 95 years old. However, a majority of customers called is in the age of 30s and 40s (33 to 48 years old fall within the 25th to 75th percentiles).

The distribution of customer age is fairly normal with a small standard deviation.

The distribution of balance:

The distribution of balance having a minimum of -8019 to a maximum of 102127 euros. The distribution of balance has a huge standard deviation relative to the mean, suggesting large variabilities in customers' balance levels.

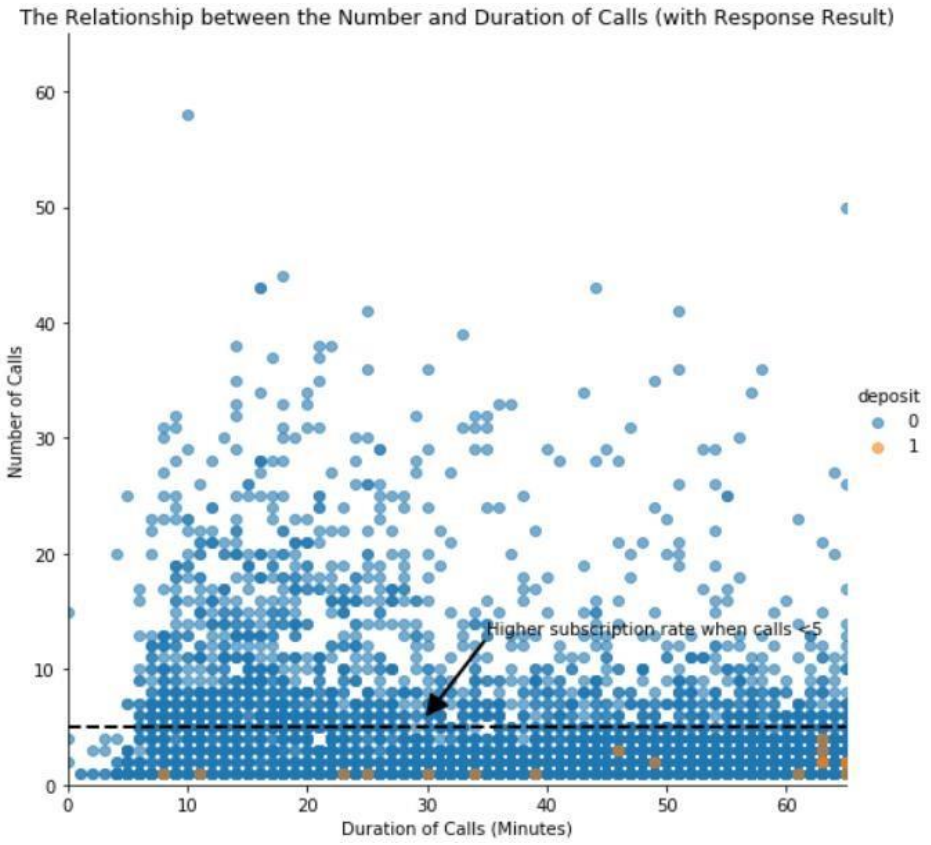
4.3 Visualize the relationship between customer age and balance



Based on this scatter plot, there is no clear relationship between client's age and balance level.

Nevertheless, over the age of 60, clients tend to have a significantly lower balance, mostly under 5,000 euros. This is due to the fact that most people retire after 60 and no longer have a reliable income source.

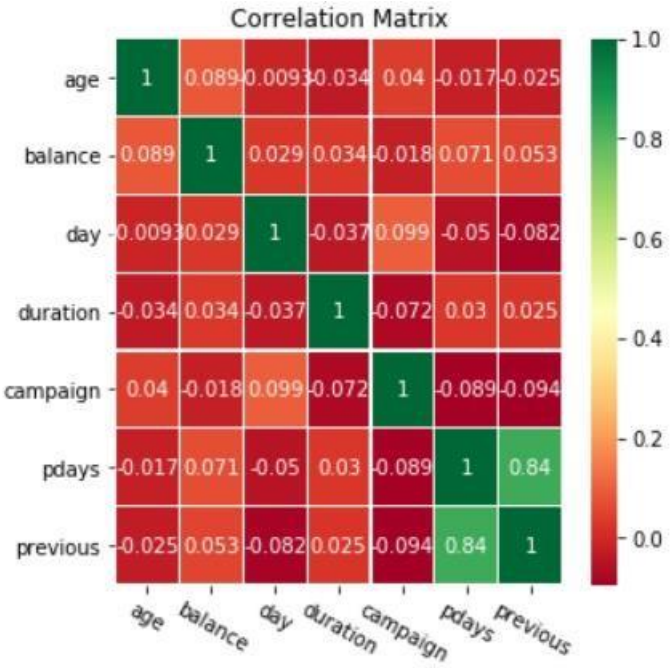
4.4 Visualize the relationship between phone call duration & the number of campaigns



In this scatter plot, clients not subscribed to term deposits are denoted as "0" while clients subscribed are denoted as "1". Compared to “no” clients”, “yes” clients were contacted by fewer times and had longer call duration. More importantly, after five campaign calls, clients are more likely to reject the term deposit unless the duration is high. Most “yes” clients were approached by less than 10 times.

This suggests that the bank should resist calling a client for more than five times, which can be disturbing and increase dissatisfaction.

4.5 Correlation matrix

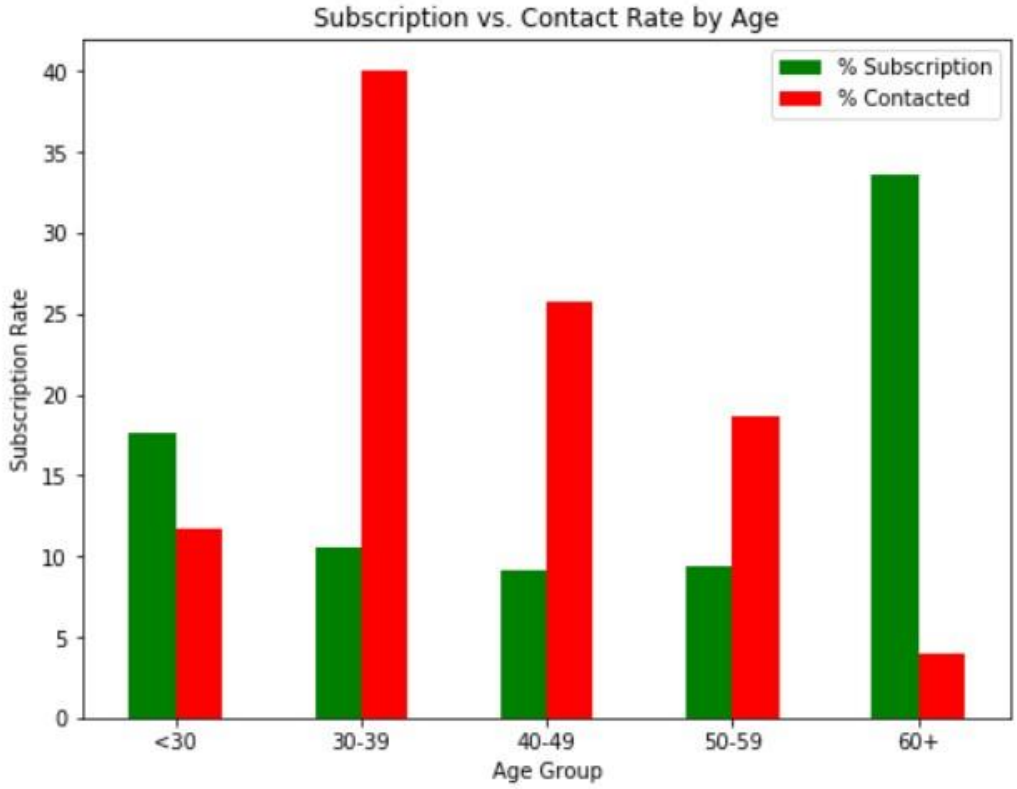


	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.089447	-0.009292	-0.033634	0.040172	-0.017421	-0.025499
balance	0.089447	1.000000	0.028636	0.033703	-0.018256	0.070765	0.053070
day	-0.009292	0.028636	1.000000	-0.037362	0.098733	-0.050017	-0.082465
duration	-0.033634	0.033703	-0.037362	1.000000	-0.071561	0.029902	0.025478
campaign	0.040172	-0.018256	0.098733	-0.071561	1.000000	-0.089031	-0.093580
pdays	-0.017421	0.070765	-0.050017	0.029902	-0.089031	1.000000	0.835455
previous	-0.025499	0.053070	-0.082465	0.025478	-0.093580	0.835455	1.000000

To investigate more about correlation, a correlation matrix was plotted with all qualitative variables. Clearly, this correlation matrix infers that there is no such strong correlation (multi collinearity) among the features and it resembles that there is no much noise in the data. We also checked the multi collinearity effect clearly through VIF in part 5.3.

4.6 Visualize the subscription and contact rate by customer age

Now we have a good understanding of the distribution of key variables. Five plots will be generated to further investigate the influence of different customer characteristics on the subscription rate.



Insights: target the youngest and the oldest instead of the middle-aged

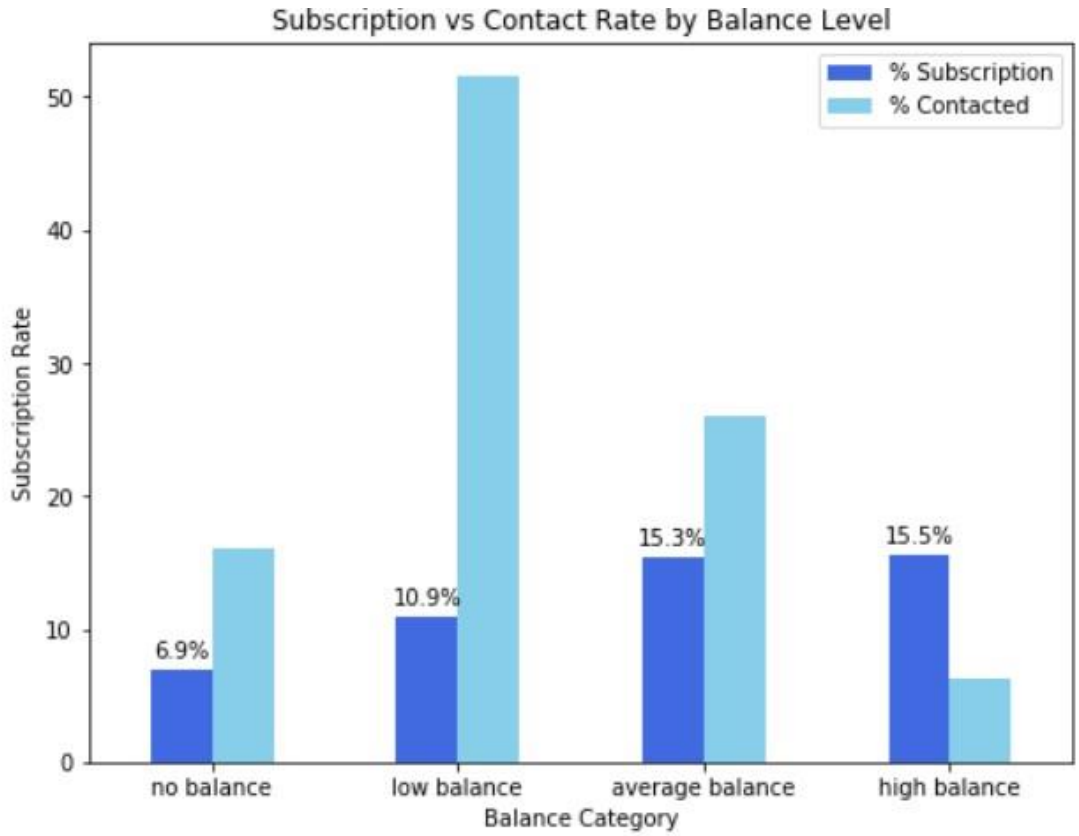
Green vertical bars indicate that clients with a age of 60+ have the highest subscription rate. About 17% of the subscriptions came from the clients aged between 18 to 29. More than 50% of the subscriptions are contributed by the youngest and the eldest clients.

It is not surprising to see such a pattern because the main investment objective of older people is saving for retirement while the middle-aged group tend to be more aggressive with a main objective of generating high investment income. Term deposits, as the least risky investment tool, are more preferable to the eldest.

The youngest may not have enough money or professional knowledge to engage in sophisticated investments, such as stocks and mutual funds. Term deposits provide liquidity and generate interest incomes that are higher than the regular saving account, so term deposits are ideal investments for students.

However, red vertical bars show that the bank focused its marketing efforts on the middle-aged group, which returned lower subscription rates than the younger and older groups. Thus, to make the marketing campaign more effective, the bank should target younger and older clients in the future.

4.7 Visualize the subscription rate by balance level



Insights: target clients with average or high balance

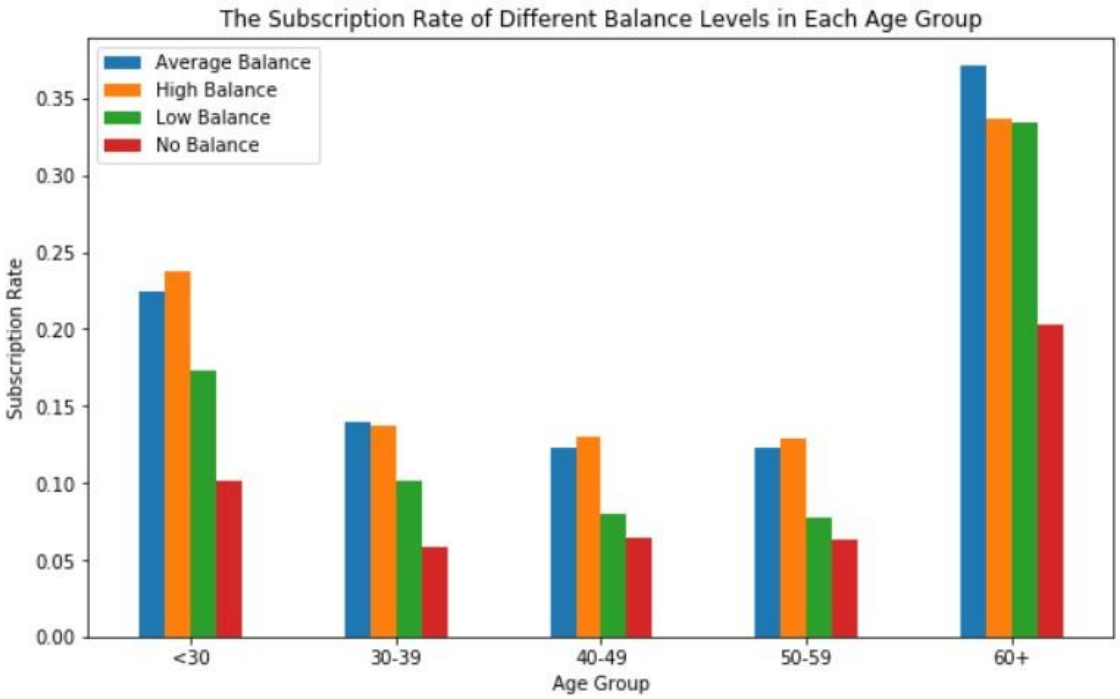
To identify the trend more easily, clients are categorized into four groups based on their levels of balance:

- ✓ No Balance: clients with a negative balance.
- ✓ Low Balance: clients with a balance between 0 and 1000 euros
- ✓ Average Balance: clients with a balance between 1000 and 5000 euros. ✓ High Balance: clients with a balance greater than 5000 euros.

Unsurprisingly, this bar chart indicates a positive correlation between clients' balance levels and subscription rate. Clients with negative balances only returned a subscription rate of 6.9% while clients with average or high balances had significantly higher subscription rates, more than15%.

However, in this campaign, more than 50% of clients contacted only have a low balance level. In the future, the bank should shift its marketing focus to high-balance customers to secure more term deposits.

4.8 Visualize the subscription rate by age and balance



Insights: target older clients with high balance levels

While age represents a person's life stage and balance represents a person's financial condition, jointly evaluating the impact of these two factors enables us to investigate if there is a common trend across all ages, and to identify which combination of client features indicates the highest likelihood of subscription.

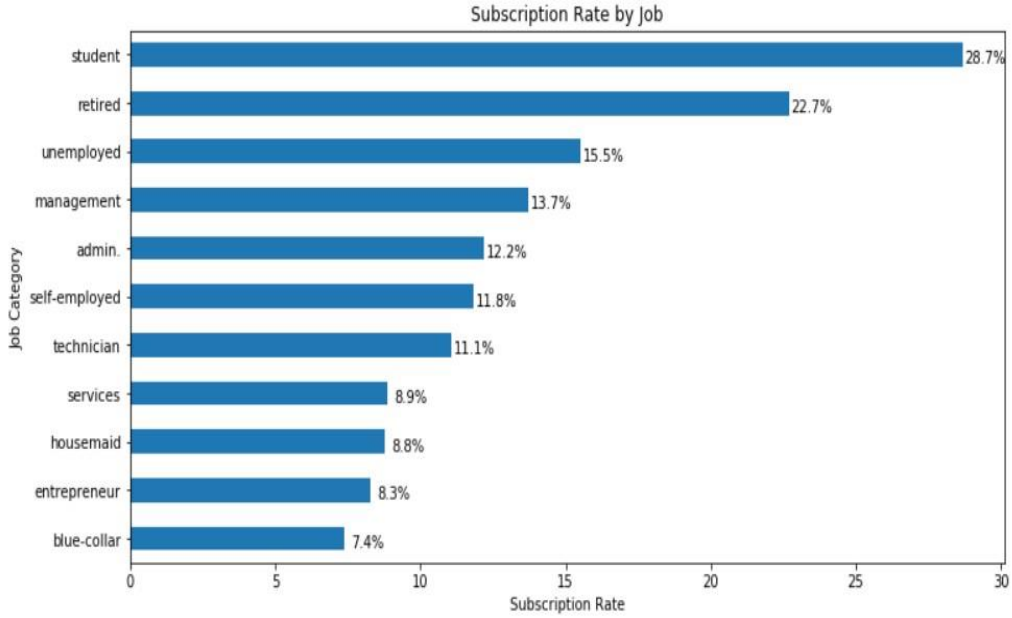
In order to investigate the combined effect of age and balance on a client's decision, we performed a two-layer grouping, segmenting customers according to their balance levels within each age group.

The graph tells the same story regarding the subscription rate for different age groups: the willingness to subscribe is exceptionally high for people aged above 60 and younger people aged below 30 also have a distinguishable higher subscription rate than those of other age groups.

Furthermore, the effect of balance levels on subscription decision is applicable to each individual age group: every age group shares a common trend that the percentage of subscription increases with balance.

In sum, the bank should prioritize its telemarketing to clients who are above 60 years old and have positive balances, because they have the highest acceptance rate of about 35%. The next group the bank should focus on is young clients with positive balances, who showed high subscription rates between 15% and 20%.

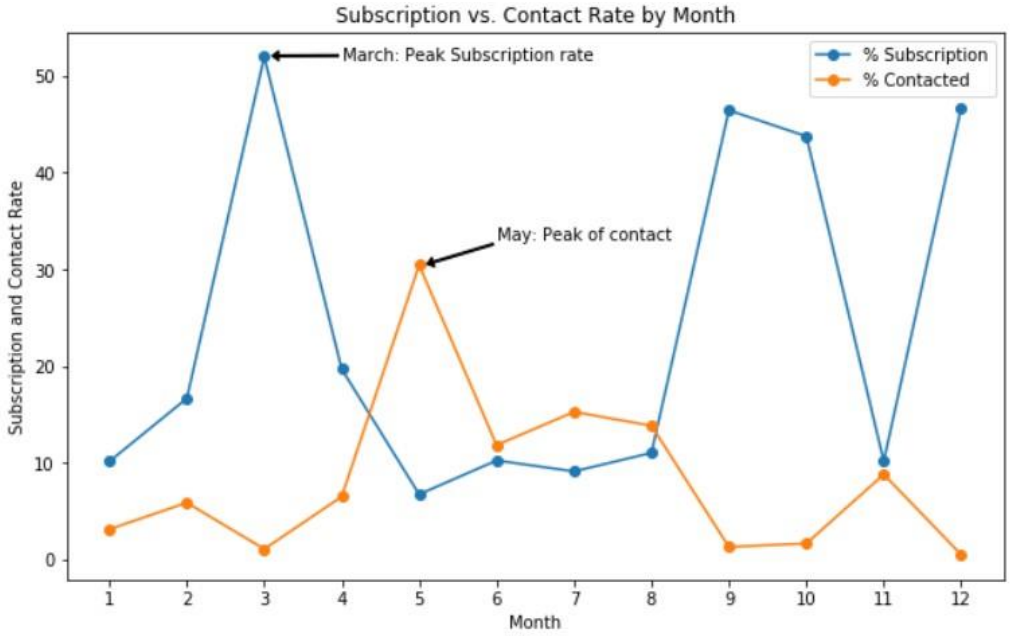
4.9 Visualize the subscription rate by job



Insights: target students and retired clients

As noted from the horizontal bar chart, students and retired clients account for more than 50% of subscription, which is consistent with the previous finding of higher subscription rates among the younger and older.

4.10 Visualize the subscription and contact rate by month



Insights: Initiate the telemarketing campaign in fall or spring

Besides customer characteristics, external factors may also have an impact on the subscription rate, such as seasons and the time of calling. So the month of contact is also analysed here.

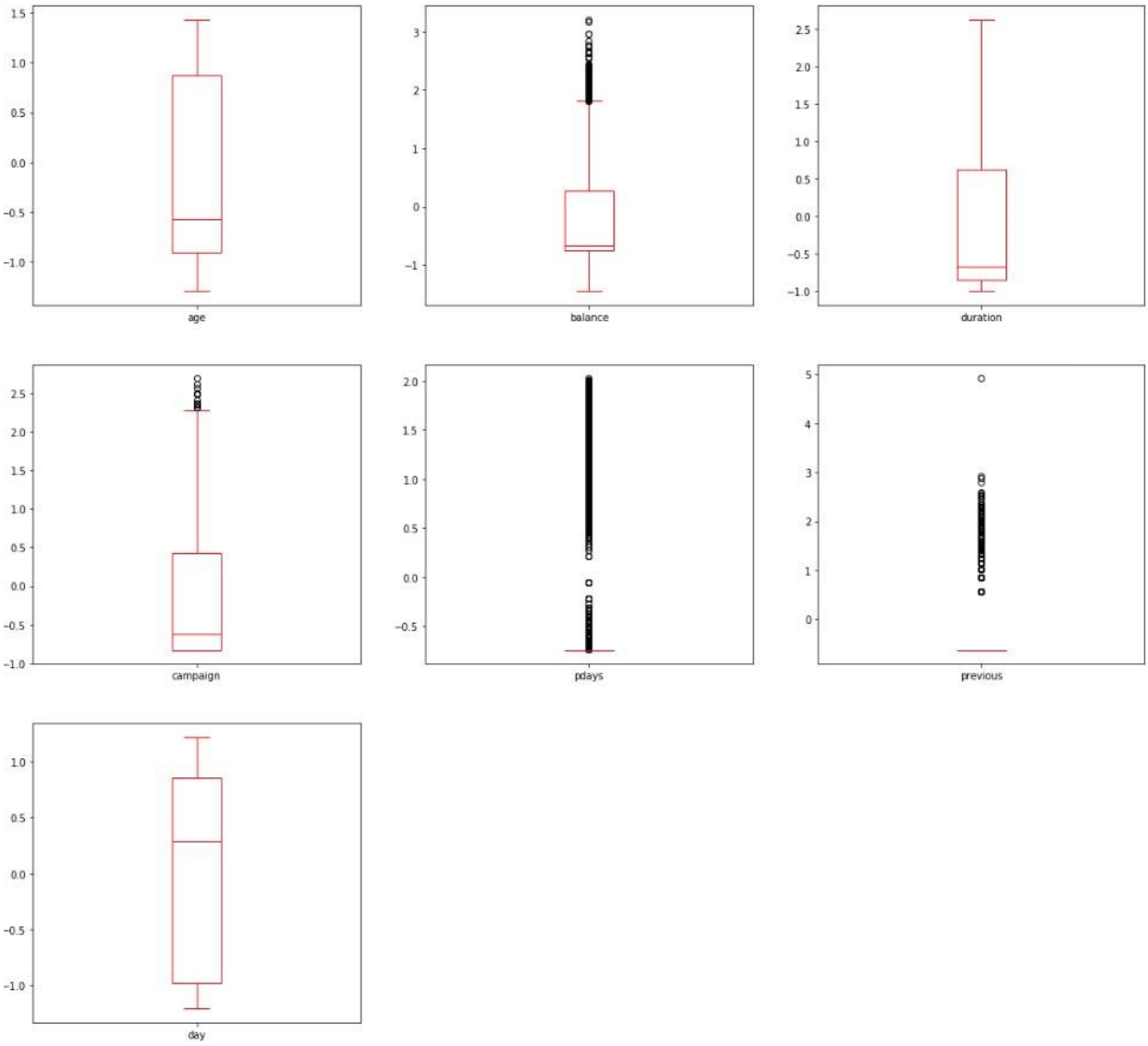
This line chart displays the bank's contact rate in each month as well as clients' response rate in each month. One way to evaluate the effectiveness of the bank's marketing plan is to see whether these two lines have a similar trend over the same time horizon.

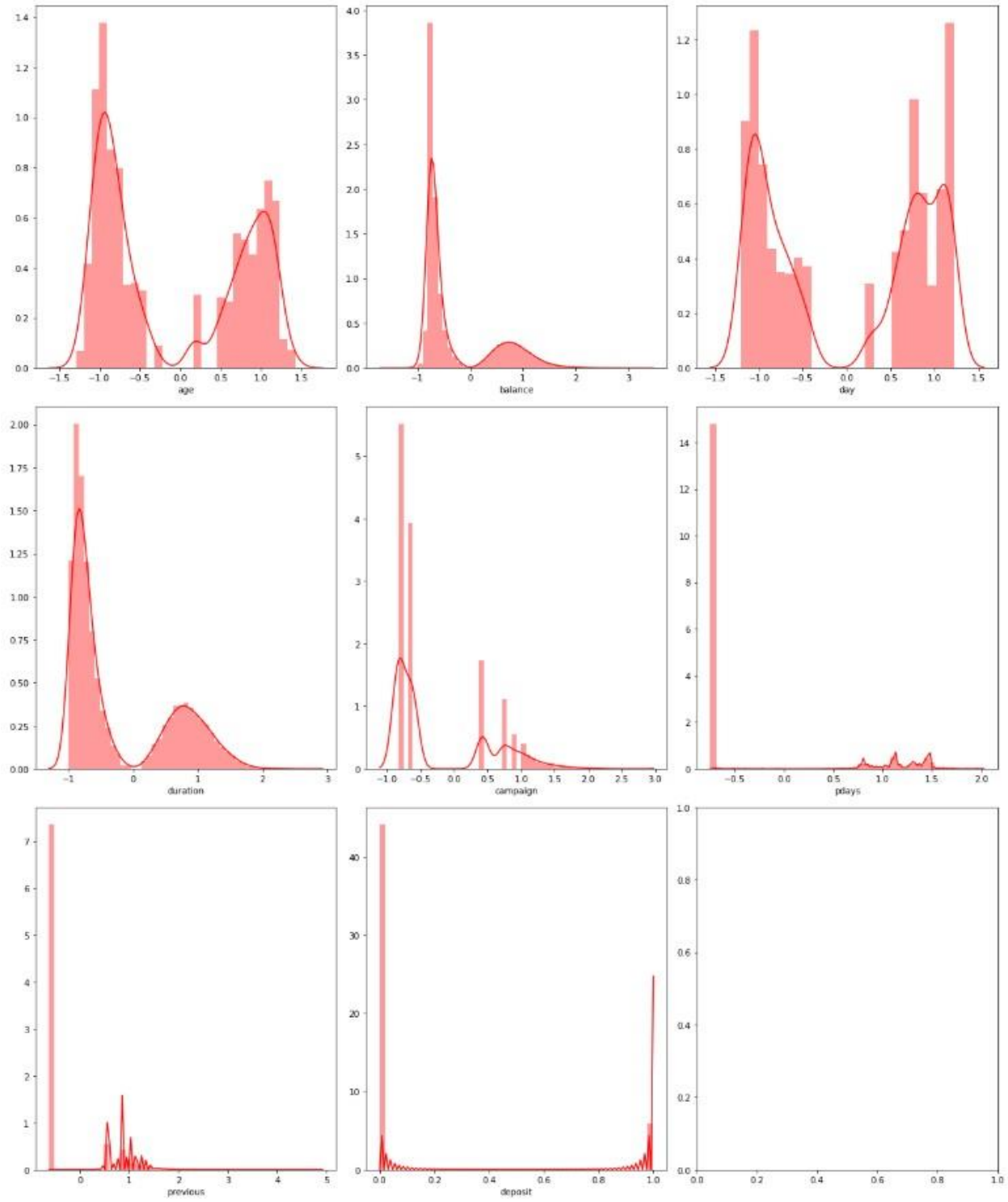
The bank contacted most clients between May and August. The highest contact rate is around 30%, which happened in May, while the contact rate is closer to 0 in March, September, October, and December. However, the subscription rate showed a different trend. The highest subscription rate occurred in March, which is over 50%, and all subscription rates in September, October, and December are over 40%.

Clearly, these two lines move in different directions which strongly indicates the inappropriate timing of the bank's marketing campaign. To improve the marketing campaign, the bank should consider initiating the telemarketing campaign in fall and spring when the subscription rate tends to be higher.

Nevertheless, the bank should be cautious when analysing external factors. More data from previous marketing campaign should be collected and analysed to make sure that this seasonal effect is constant over time and applicable to the future.

4.11 Normality & Outliers of the features after outlier's treatment





Before the outlier treatment it is observed form the data is not normally distributed and also many outliers are present in the data. In order to reduce the outliers and to make the data to be normally distributed we identified the outliers and performed knn- imputation on the outlier's and made outlier treatment.

In order to reduce the scale of the data we also performed cube root transformation on the data. We choose cube root transformation due to few reasons:

- ✓ As data in some of the features are left skewed and square root transformation of left skewed data make data NaN.

- ✓ Log transformation in order to reduce the skewness it is increasing the skewness of the some of the features such as “balance” due to the data in the balance feature is left skewed (i.e negative values).

This makes us to choose higher end transformation techniques such as cube root and finally makes our data normally distributed compared to the data before outlier treatment and transformation.

Part 5 Statistical Analysis

Statistical tests were performed to see the whether the independent variables have a significant relationship with the dependent variable, DEPOSIT

5.1 Chi-square Test

For the Categorical Columns, a Chi-square Test of independence was performed with the target variable, DEPOSIT which is also a categorical column.

Null Hypothesis H0: There is NO association between the two variables

Alternate Hypothesis Ha: There is an association between the two variables

Variable	p-value	Decision
Job	3.5547437743574746e-171	Reject H0
job (retired)	0.367	Fails to reject H0
Marital	2.1450999986791486e-43	Reject H0
marital single	0.259	Fails to reject H0
Education	6.897199498466207e-52	Reject H0
Education secondary	0.716	Fails to reject H0
Contact	1.251738325340495e-225	Reject H0
contact_telephone	0.124	Fails to reject H0
Month	0.0	Reject H0
Pout come	0.0	Reject H0
poutcome_other	0.583	Fails to reject H0
Default	2.4538606753508344e-06	Reject H0
default_yes	0.197	Fails to reject H0
Housing	2.918797605076633e-192	Reject H0
Loan	1.665061163492756e-47	Reject H0

Based on the above results it is observed that most of the features p-values is less than 0.05 which indicates that most of the features are statistically significant with the target variable and helps in the prediction of the term depositor’s.

5.2 Two-sample t test

For all the numeric variables, A two-sample unpaired t tests was performed between values of the variable for two classes of target variables to compare their means. Null Hypothesis H0: The means of the two samples are EQUAL

Alternate Hypothesis Ha: The means of the two samples are NOT EQUAL

If the means of the two samples are significantly different form each other, then we can conclude that the variable does have a significant relationship with the target variable.

Variable	p-value	Decision
Age	9.25408389720676e-05	Reject H0
Balance	1.6290605680244357e-72	Reject H0
Day	0.429	Fails to reject H0
Duration	0.0	Reject H0
Campaign	1.933381009701562e-62	Reject H0
Pdays	9.868576155361588e-221	Reject H0
previous	0.066	Fails to reject H0

Based on the above results it is observed that most of the features p-values is less than 0.05 which indicates that most of the features are statistically significant with the target variable and helps in the prediction of the term depositor's.

5.3 check for multicollinearity (VIF)

Variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

Features	vif
pdays	9.134602
month_may	4.226903
poutcome_unknown	27.344407

From the above results we can infer that the features which are having multicollinearity are only three features. This resembles that there is very less multicollinearity in the data. So there is no need to go for PCA (Principal Component Analysis).

Part 6 Machine Learning: Classification

The main objective of this project is to identify the most responsive customers before the marketing campaign so that the bank will be able to efficiently reach out to them, saving time and marketing resources. To achieve this objective, classification algorithms will be employed. By analyzing customer statistics, a classification model will be built to classify all clients into two groups: "yes" to term deposits and "no" to term deposits.

Prepare Data for Classification

Select the most relevant customer information: job, education, age, balance, default, housing and loan and other features whose p-value <0.05 and vif<4. Since machine learning algorithms only take numerical values, all five categorical variables (job, education, default, housing and others) are transformed into dummy variables. Dummy variables were used instead of continuous integers because these categorical variables are not ordinal. They simply represent different types rather than levels, so dummy variables are ideal to distinguish the effect of different categories. Feature selection: all customer statistics were selected as features while the deposit feature was set as target. 70% of the data was used to build the classification model and 30% was reserved for testing the model.

Part 7 Evaluation Metrics

The Evaluation Metrics that can be used for a Binary Classification problem are:

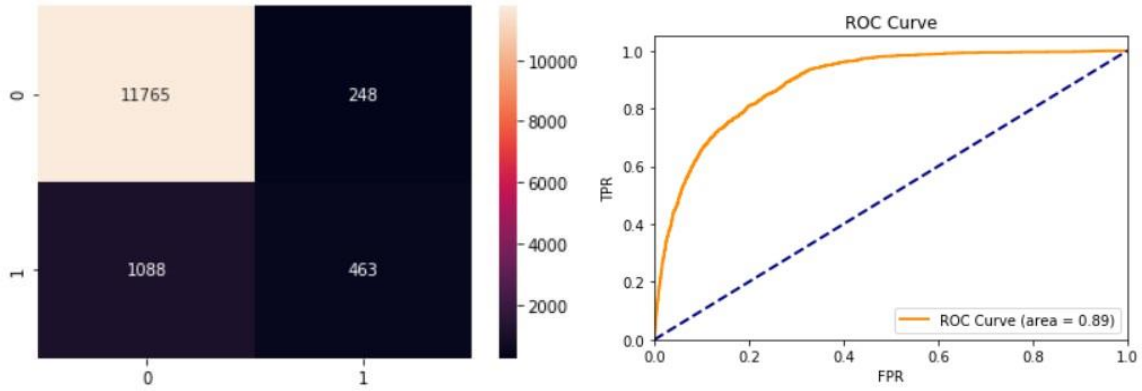
- ✓ Accuracy - Proportion of correctly identified instances
- ✓ Precision - proportion of positive predictions that are correct

- ✓ Recall - Proportion of Actual positives predicted correctly
- ✓ F1 Score - Harmonic mean of Precision and Recall
- ✓ ROC AUC - Area Under Receiver's Operating Characteristics Curve (tradeoff between sensitivity and specificity for different thresholds)

Part 8 Base Model with imbalanced target variable

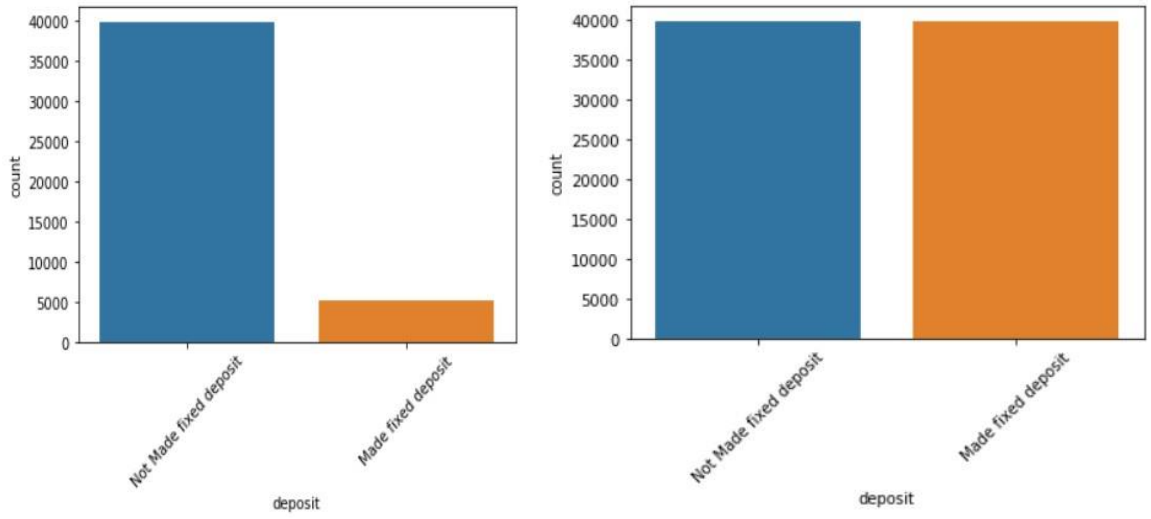
Base classification algorithms (Logistic Regression) was run on the dataset which was under sampled and identified the classification metrics such as Accuracy, Precision, Recall, F1-score.

Accuracy	0.9015039811265113
Precision	0.6511954992967651
Recall	0.2985170857511283
F1-Score	0.4093722369584439



From the above results it is observed that Accuracy is good but precision, recall and f1-score are very less compared to Accuracy. This is due to imbalance in the target variable. So by doing oversampling there is a chance to significantly increase the classification metrics (precision, recall, f1-score) along with accuracy.

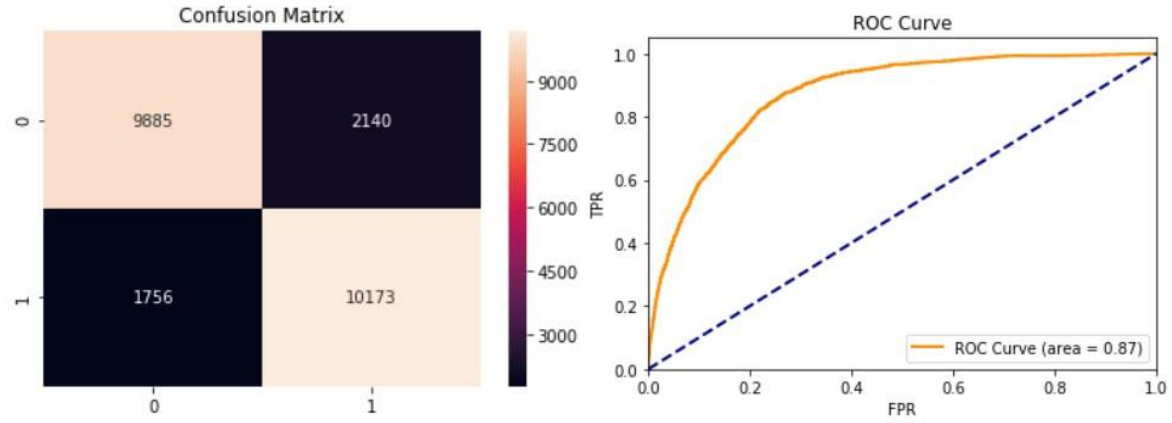
8.1 Oversampling the target variable by using SMOTE



Part 9 Base Model with oversampled data (SMOTE)

Base classification algorithms (Logistic Regression) was run on the dataset which is balanced by doing over sampling and identified the classification metrics such as Accuracy, Precision, Recall, F1-score.

Accuracy	0.8373549302830425
Precision	0.8261999512710144
Recall	0.8527957079386369
F1-Score	0.8392871875257817



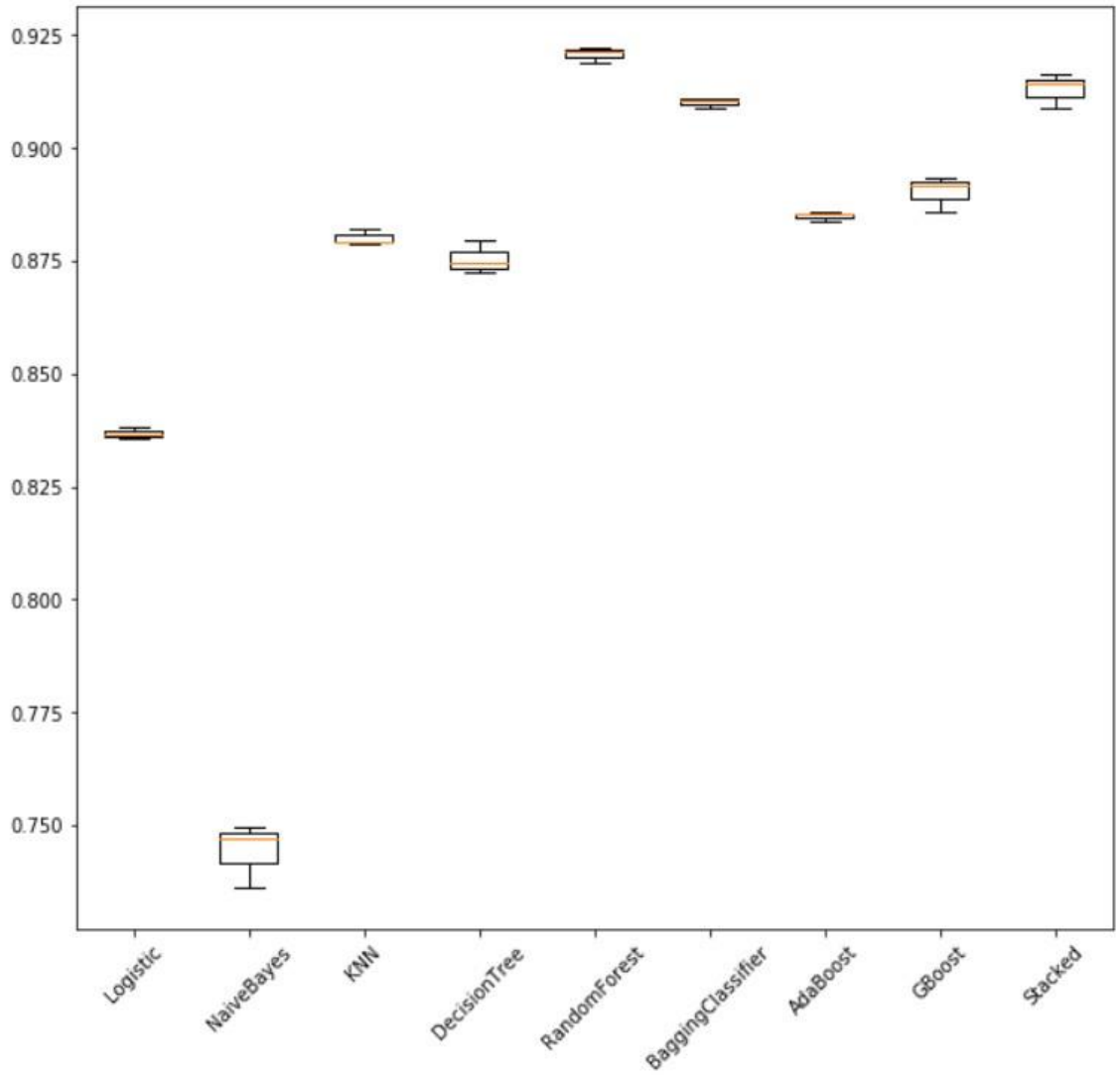
From the above results it is observed that Accuracy, precision, recall and f1-score are approximately same and all are around 80% which indicates that model is having less bias and variance errors. This is due to generating balance in the target variable by using smote technique. So by doing oversampling the data there is an increase in the classification metrics (precision, recall, f1-score) along with accuracy.

9.1 Algorithms Comparison (K-fold Cross Validation):

Four different classification algorithms (Logistic Regression, K-Neighbours Classifier, Decision Tree Classifier, and Gaussian NB, Random Forest, Ada boost, Gradient Boosting) were run on the dataset through K-fold cross validation and the best-performing one was (identified by observing bias and variance errors) and used to build the classification model.

Classification Algorithm	Bias Error	Variance Error
Logistic Regression	0.836783	0.000002
NaiveBayes	0.744287	0.000049
KNN	0.880156	0.000004
DecisionTree	0.875544	0.000013
RandomForest	0.920787	0.000003
BaggingClassifier	0.910145	0.000002
AdaBoost	0.884917	0.000001
GBoost	0.890304	0.000016
Stacking classifier	0.913070	0.000015

Algorithms Comparison through boxplots:

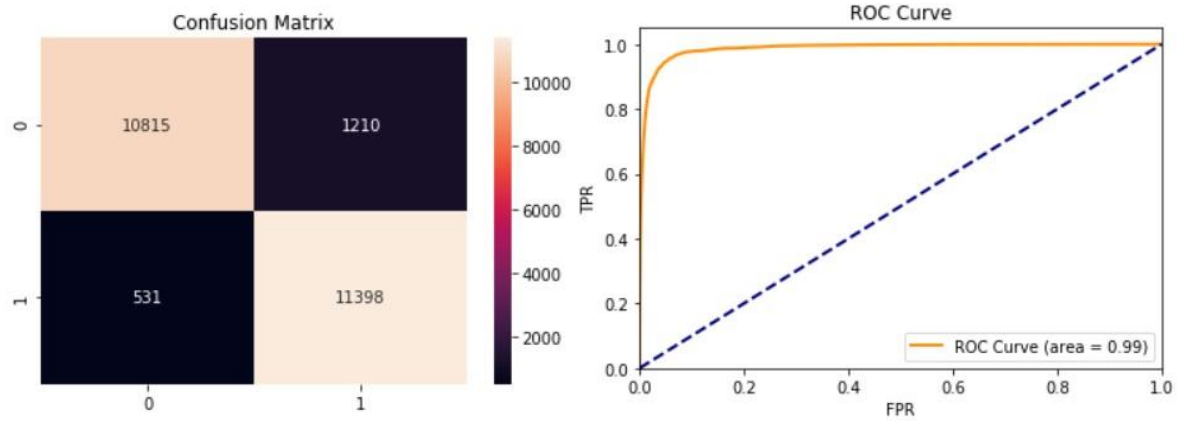


From the above results it is observed that RF Classifier is the best performing model.

By comparing all algorithms bias error and variance error, RF classifier is observed to be the best so it would be used to predict term depositor's. The test of RF classifier with base estimator (Decision Tree, n_estimators=13) is the best model with best bias & variance error trade off.

Best performing classification algorithms (RF classifier) was run on the dataset which was over sampled and identified the classification metrics such as Accuracy, Precision, Recall, F1-score.

Accuracy	0.9273190281372631
Precision	0.9040291878172588
Recall	0.9554866292229022
F1-Score	0.929045930635367



Classification Report:

	precision	recall	f1-score	support
0	0.95	0.90	0.93	12025
1	0.90	0.96	0.93	11929
accuracy			0.93	23954
macro avg	0.93	0.93	0.93	23954
weighted avg	0.93	0.93	0.93	23954

- Precision of 0 (the client said no) represents that for all instances predicted as no subscription, the percentage of clients that actually said no is 95%.
- Recall is the ability of a classifier to find all positive instances. Recall of 0 indicates that for all clients that actually said no, the model predicts 90% correctly that they would decline the offer.

Part 10 Conclusion

The main objective of this project is to increase the effectiveness of the bank's telemarketing campaign, which was successfully met through data analysis, visualization and analytical model building. A target customer profile was established while classification models were built to predict customers' response to the term deposit campaign.

According to previous analysis, a target customer profile can be established. The most responsive customers possess these features: Feature 1: age < 30 or age > 60

Feature 2: students or retired people

Feature 3: a balance of more than 5000 euros

By applying RF classifier algorithm, classification and estimation model were successfully built. With this model, the bank will be able to predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits, and call less to those who are unlikely to make term deposits.

In addition, predicting duration before calling and adjusting marketing plan benefit both the bank and its clients. On the one hand, it will increase the efficiency of the bank's telemarketing campaign, saving time and efforts. On the other hand, it prevents some clients from receiving undesirable advertisements, raising customer satisfaction. With the aid of RF classifier model, the bank can enter a virtuous cycle of effective marketing, more investments and happier customers.

Part 11 Recommendations

11.1. More appropriate timing

When implementing a marketing strategy, external factors, such as the time of calling, should also be carefully considered. The previous analysis points out that March, September, October and December had the highest success rates. Nevertheless, more data should be collected and analysed to make sure that this seasonal effect is constant over time. If the trend has the potential to continue in the future, the bank should consider initiating its telemarketing campaign in fall and spring.

11.2. Smarter marketing design

By targeting the right customers, the bank will have more and more positive responses, and the classification algorithms would ultimately eliminate the imbalance in the original dataset. Hence, more accurate information will be presented to the bank for improving the subscriptions. Meanwhile, to increase the likelihood of subscription, the bank should re-evaluate the content and design of its current campaign, making it more appealing to its target customers.

11.3. Better services provision

With a more granular understanding of its customer base, the bank has the ability to provide better banking services. For example, marital status and occupation reveal a customer's life stage while loan status indicates his/her overall risk profile. With this information, the bank can estimate when a customer might need to make an investment. In this way, the bank can better satisfy its customer demand by providing banking services for the right customer at the right time.