Springboard

Leveraging customer information for strategic telemarketing in the banking industry

Data Science Career Track

Capstone Project 3

Author
Isaac Ghebregziabher

September 13, 2021

Contents

Chapter 1

Problem statement and objective

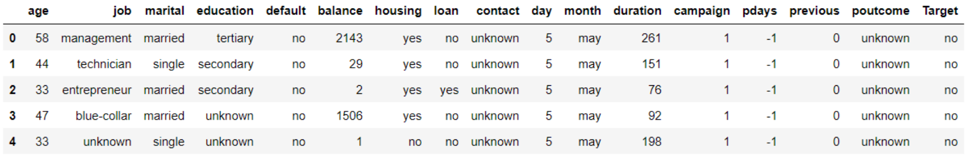
1.1 Introduction

Many business accumulate a wealth of data related to their technical operations and customer information. Descriptive analysis and visualization of these albeit complex data is essential not only to gauge the current state of business but also make strategic changes to improve business. In the case of banking industry, banks possess a wealth of data pertaining to customer information subscribed to the variety of portfolios available by the bank. Hence, systemic data analysis of customer information is essential for banks for a successful marketing campaign. In fact the banking industry spends large amounts of money and resources for telemarketing campaigns. Therefore, it is essential for banks to develop optimized marketing campaigns to reduce costs while maximizing effectiveness. One way to achieve this is to understand customer needs based on the available customer information.

1.2 business problem and stakeholders

Our client is a Portuguese banking institution. They have brought us a dataset directly related to their marketing campaigns conducted through phone calls. The dataset is a *CSV* file named bank-full.csv and is publicly available in the UCI Machine Learning Repository, which can be retrieved from here. The campaign was conducted over the period of time extending from May 2008 to November 2010 and collected data consist of:

- Demographics (age, job, education, marital status),
- Financial data (credit, housing loan, personal loan),
- Contact details (method of contact, month client was contacted, day client was last contacted, duration of last contact in seconds, campaign – number of contacts performed during this campaign and for this client)
- Previous campaign data (pdays: number of days that passed by after the client was last contacted from a previous campaign, previous: number of contacts performed before this campaign and for this client, poutcome: outcome of the previous marketing campaign)



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

Figure 1.1: Preview of the first five rows of the dataset.

- Campaign outcome (y - has the client subscribed a term deposit?)

The data is at the client account level; in other words, there is one row for each client account where rows are labeled by whether the client has subscribed to a term deposit or not for a sample of 42511 clients contacted during the campaign.

A brief preview of the dataset is shown below.

1.3 Objective

In this project, our goal is to develop a predictive model for whether a client will subscribe to a term deposit or not. A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits. The developed model will help the bank:

- Understand its customers and cluster them into meaningful groups based on their demographic and transaction information
- Predict customer response to its telemarketing campaigns
- Identify target customer groups for its future tele-marketing campaigns.

Chapter 2

Data wrangling and organizing

Now that we have a defined business problem, in this section we need to dig deeper into the dataset we received from our client. In addition to exploring the given dataset numerically and graphically, we need to ensure whether the dataset makes sense and matches to what we were told by our client, in this case to what is written on UCI website. The main focus of Data wrangling is organizing and cleaning our dataset making it ready for building the best performing Machine Learning model. In what follows below are essential guiding questions that need to be addressed during data wrangling and exploration:

- How many samples (rows)?
- How many features (columns)
- Types of features? - Which are Categorical? Which are numerical?
- What Does the features data look like?
- Range of values for numeric features
- Frequency of classes for categorical features
- Are there any missing values?

2.1 Size and type of dataset

After loading our dataset into a dataframe df, we obtained the number of rows and columns with the following pandas commands.

```
df␣=␣pd.read_csv('bank-full.csv')
df.shape
output:(45211,␣17)
```

As can be seen from the output above our loaded dataset consists of 45211 rows and 17 columns which was also confirmed by loading the dataset with an excel spreadsheet.

As can be seen from the output of the pandas code below, our data set consists of 17 features with 7 integer type features and 10 object type (categorical features). In what follows below is a list of the 7 integer type features and 10 categorical features.

```
  Integers: ['age','balance', 'day', 'duration', 'campaign', 'pdays', 'previous']
```

```
Categorical:['job', 'marital', 'education', 'default', 'housing', 'loan',
 'contact', 'month', 'poutcome', 'Target']
```

2.2 Visualizing the categorical features

In this section we want to understand how many unique values each of the categorical features possess and how many customers with each categorical feature values were contacted by the bank. For ease of our analysis the categorical features were grouped into demographic, financial, and campaign.

2.2.1 Visualizing demographic distribution of customers

Demographic segmentation is a commonly used technique in marketing where target categories based on one or more of socio-economic variables like job, age, gender, marital status, income, education, etc. In our case the categorical demographic content of the customers contacted by the bank includes job, marital, education.

The customer distribution contacted by the bank during this telemarketing campaign for each customer demographic category is shown in figure 2.1.

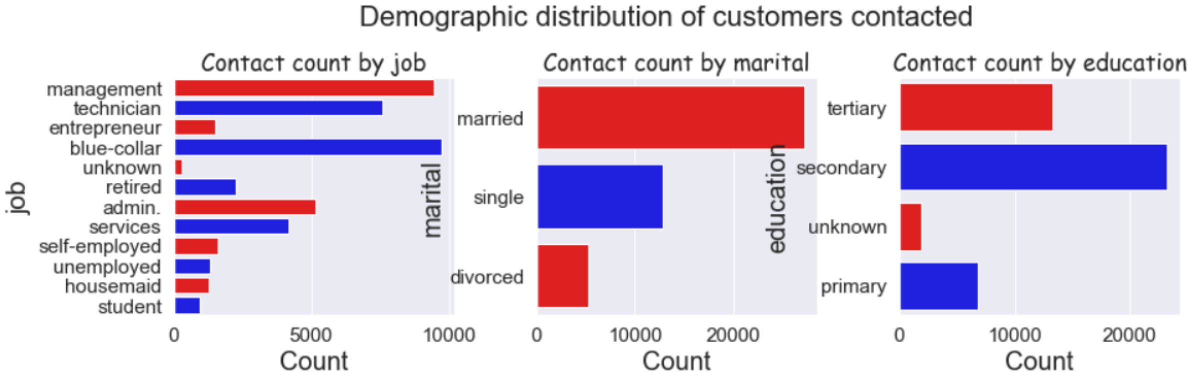


Figure 2.1: Frequency of customers contacted by the bank during the telemarketing for each job category, marital status, and educational level.

Customer jobs

Customers with **blue-collar** jobs were contacted the most during the marketing campaign. Customer's whose jobs is unknown (either not recorded or customer did not want to specify) are targeted the least during the marketing campaign. Following customers with blue-collar managers, technicians, and administrators, respectively in decreasing order were contacted during the marketing campaign. As one might expect students were targeted the least during the marketing campaign.

Marital status

Marital status is another important trait used in demographic segmentation. It provides useful information to target audience during marketing communication through promotional messages. Usually people in the same marital status group have common tastes and can can be easily targeted based on marital status. For example, generally people who are married need more financial strategy and are likely expected to respond to the term deposit marketing campaign. The largest number of people contacted during the marketing campaign are married. The list contacted people are divorced individuals.

Educational level

The level of education of an individual is also an important trait when segmenting population for marketing campaigns. For example educational level determines the choice of communication channel in passing across your message. During the marketing campaign customer who completed high school are contacted the most. There are also a few number of contacts made where customer's education level is UNKNOWN.

2.2.2 Visualizing financial data segmented distribution of customers

In addition to demographic segmentation, customers could be grouped together based on their account related to their finances. That is whether a customer has credit or not ('default'), has housing loan or not ('housing'), and has personal loan or not ('loan'). In figure 2.2 we plot the distribution of customers segmented by their financial status.

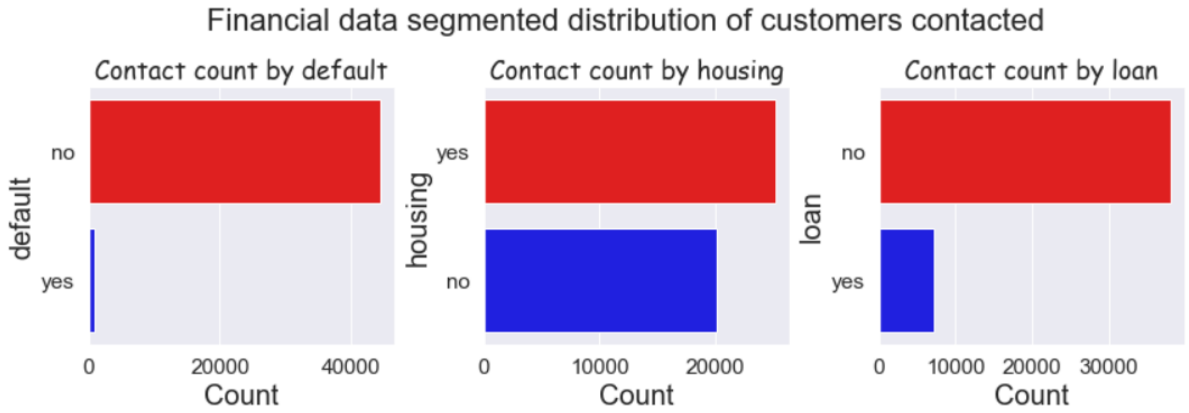


Figure 2.2: Frequency of customers contacted by the bank during the telemarketing for customers with/without credit, housing loan, and personal loan.

Has credit (default)

Majority of customers contacted by the bank do not have credit or did not default (98.2%). Only a few fraction (1.8%) of all the customers contacted had credit. Hence, credit might not be an important feature in determining whether a customer will subscribe to a term deposit or not.

Housing loan

During this telemarketing campaign, the bank called an almost balanced number of customers depending whether they have housing loan or not. About 55% of the customers contacted by the bank during this telemarketing possess housing loan. The remaining about 45% of the customers contacted did not have housing loan.

Personal loan

The majority of customers (84%) called during this telemarketing did not have any personal loans. However, a non-negligible number of customers (16%) had personal loans.

2.2.3 Distribution of customers versus past and current campaign details

In addition to demographic and financial segmentation, customers could be grouped together based on the details of the past and present campaign parameters. That is how was the customer contacted (for example iPhone or phone), the particular month the customer was contacted and the what was the outcome of the previous telemarketing campaign?

In figure 2.3 we plot customers segmented by details of the past and present campaign.

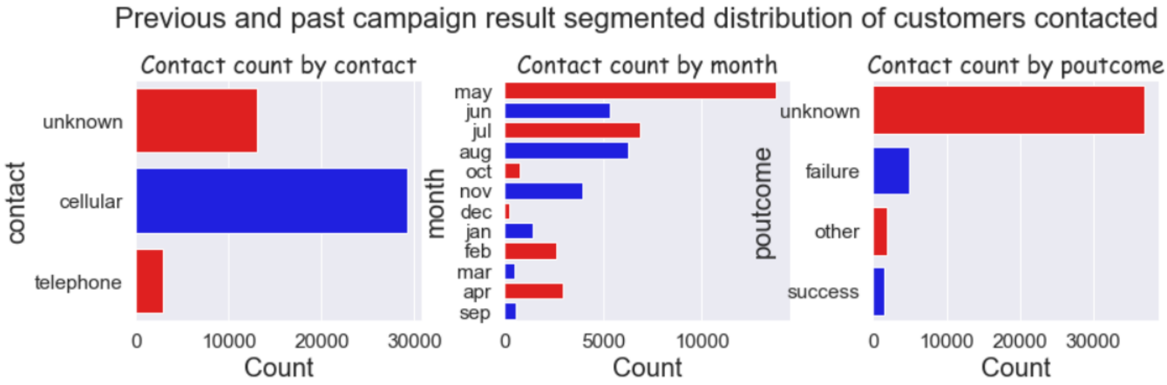


Figure 2.3: Frequency of customers contacted by the bank during the telemarketing for each method of contact, contact month, and outcome of previous telemarketing.

Method of contact

During this telemarketing, the majority of the customers (65%) were contacted using cellular phones. While contacts made through land lines amounted only to 6.5%, a significant fraction of customers (29%) were contacted through mode of communication that was not registered and marked as unknown. Different segments of the populations prefer different methods of contact. The communication channel used during the marketing campaign is essential. The largest number of customers were contacted through cellular phone. The least popular method of contact was through telephone. During the marketing campaign, 65% of the population contacted by the bank was through cellular phones. Only 6.4% contacted through a telephone line while about 29% contacted through an unknown (unspecified) mode of communication.

Month customer contacted

The season when the bank should contact target population segment is also an important trait for the success of the marketing campaign. The largest number of contacts were made during spring and summer seasons with combined contact rate of 79 %. This is understandable since spring and summer are the beginning and end dates of the tax season; where customers may be motivated for financial plans due to expected tax returns. Barely any telemarketing was conducted during the winter which is understandable as it is the holiday and end of year season. However, whether the choice of these high contact seasons maximizes efficiency requires farther analysis and will be discussed during the exploratory data analysis part of this project.

Previous marketing outcome

During this telemarketing, 3.3% of all the customers contacted have subscribed in a previous campaign while 10.8% did not subscribe in the past. The bank decided to contact more customers who did not subscribe in the past.

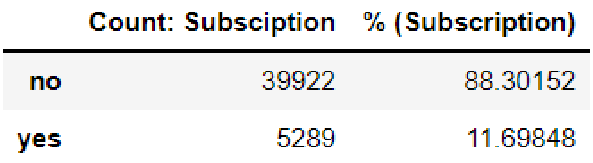
- **‘Issue:‘**

There are two ambiguous values in this feature, namely ‘unknown‘ and ‘other‘. We need to change the ‘unknown‘ to ‘other‘ for ease of analysis and doing so does not change anything since ‘unknown‘ is practically ‘other‘.

2.2.4 Class imbalance

To determine whether we are dealing with a balanced dataset or we have to deal with the difficulties in model evaluation for an imbalanced dataset, we need to look at the fraction of the two target classes, namely: 1) term-deposit subscribed 2.) term-deposit NOT subscribed.

In figure 2.4 we plot the distribution of the two classes for our dataset.



	Count: Subsciption	% (Subscription)
no	39922	88.30152
yes	5289	11.69848

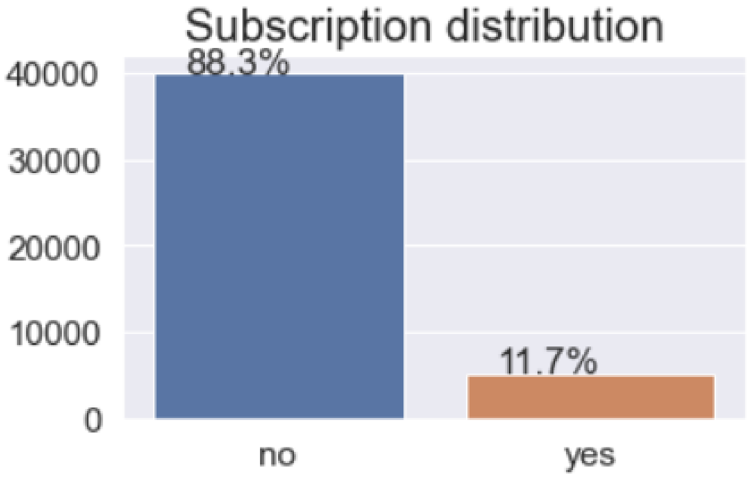


Figure 2.4: Frequency of the two target classes revealing an imbalanced dataset.

2.3 Range of numerical features

Let us know turn to investigating the range and type of the numerical features of our dataset.

2.3.1 Age range and account balance

The number of customers contacted during this telemarketing by the range of customer ages as well as their account balance is shown in figure 2.5.

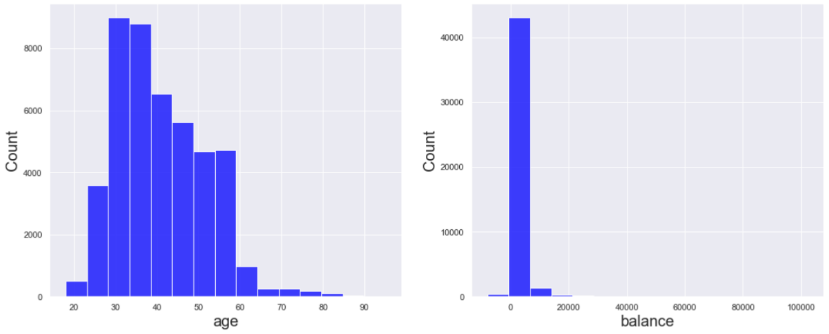


Figure 2.5: Frequency customers contacted during this telemarketing per age and account balance.

Distribution of age

The age of customers contacted by the bank during the telemarketing extends from a minimum age of 18 years to a maximum age of 95 years old. The majority of customer contacted by the bank were in their 30s and 40s (33 to 48 years land within 25^{th} and 75^{th} percentiles). With an average age of 41 years old and a standard deviation of 10 years, the distribution of age is close to normal distribution.

Distribution of balance

The customers contacted include those with positive and negative account balances. The distribution of account balances is highly skewed. The account balance of the customers called by the bank extends from -8019.0 to 102127.0 euros, resulting to a range of 110146 euros. With a mean of 1362 euros and standard deviation of 3044 euros, the balance distribution is highly skewed and far from normal. There are significant number of ateliers as can be seen by looking how far the minimum and maximum values from the mean.

To see the inter-dependence of age on balance, a scatter plot of age versus account balance is shown in figure 2.6. The figure shows no clear dependence of account balance on the customer's age. However, in general customers over the age of 60 and less than 20 years old seem to possess smaller account balances. This may be due to the fact the younger customers are in the process of establishing themsleves while the older customers have already retired and may not have any reliable source of income.

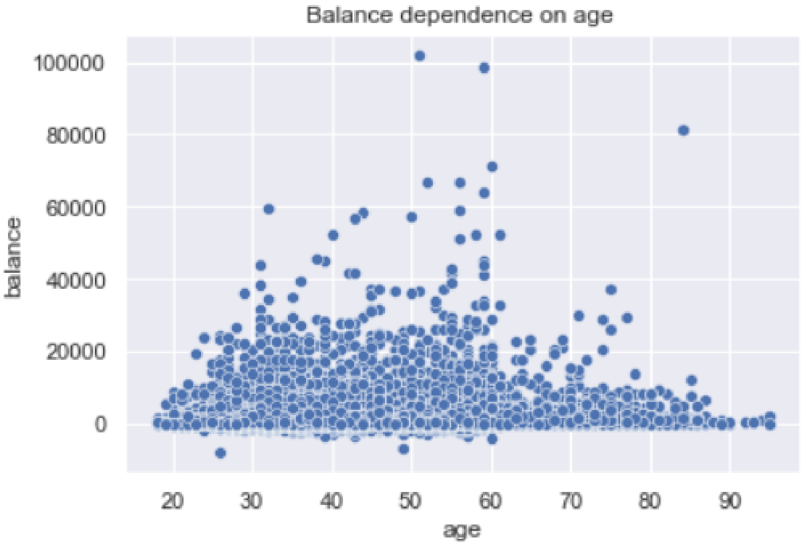


Figure 2.6: Customer age versus account balance.

2.3.2 Duration of contact and number of contacts

The distribution of the duration of contact in minutes and the number of contacts is shown in figure 2.7

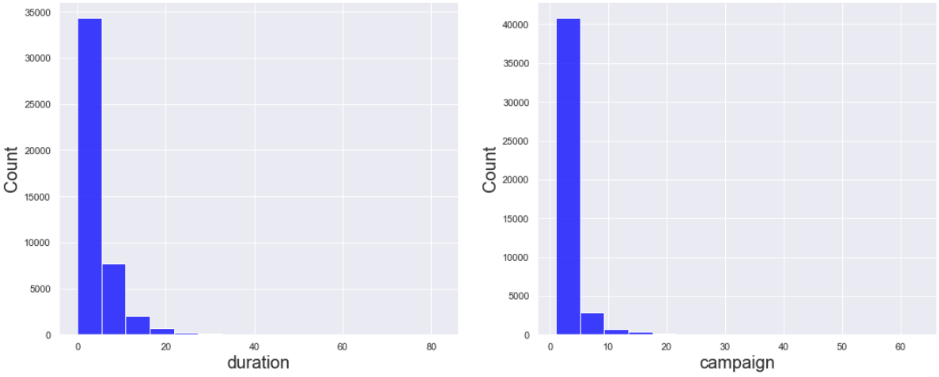


Figure 2.7: Customer age versus account balance.

Distribution of duration

The duration of a phone call made during the telemarketing extends from a minimum of 0 minutes to a maximum of 81.9 minutes. With mean duration of 4.3 minutes and median duration of 3 minutes, the distribution of the duration of all the calls during the telemarketing is left skewed; meaning most of the calls are of short duration, typically less than 5 minutes.

Distribution of campaign

The number of contacts performed during this campaign and for this client is plotted in the histogram plot shown above. The number of contacts made a client ranged from a minimum of 1 call to an extreme maximum of 63 calls. However, the majority of the calls required 1 to 3 contacts only (25 percentile to 75 percentile); indicating contacts more than 10 are outliers in the distribution.

To see whether the duration of contact is the result of many contacts, in figure 2.8 we show a scatter plot of duration versus number of contacts. Generally, the call duration decreases as the number of contacts during the campaign increases. The largest number of contacts made resulted to lowest call duration.

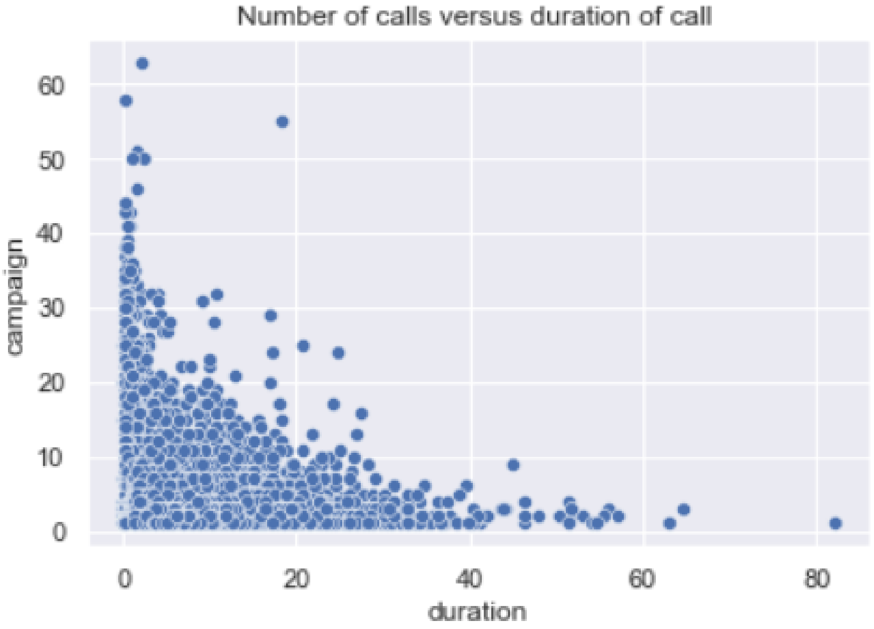


Figure 2.8: Customer age versus account balance.

2.3.3 Correlation matrix and feature interdependence

To check if any of the numerical features are linearly independent between each other, in figure 2.9 we plot a heatmap of the correlation matrix. clearly, the average number of days that passed after a client was contacted from the previous campaign (pdays) is positively correlated with the average number of contacts made before this campaign previous.

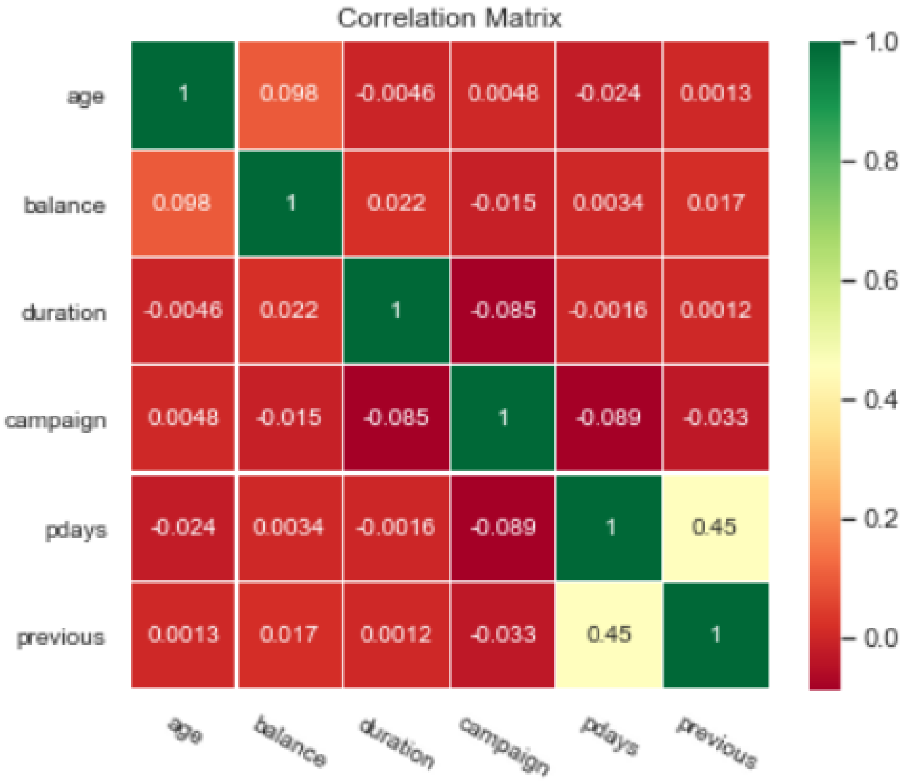


Figure 2.9: Heatmap plot of the correlation matrix of the numerical features.

Chapter 3

Exploratory data analysis

Now that we have organized our dataset and have gained the nature of the features, we are ready to start examining the values of the features and response (target) feature. Understanding the effect of the values of the features on the decision of the customer to subscribe to a term-deposit or not is essential for a predictive Machine Learning model development. In the previous chapter we gained an understanding of our dataset by visualizing the frequency of customers the bank contacted during this telemarketing. However, our analysis in data wrangling did not reveal the effect of each feature value on the success of the telemarketing campaign. In this chapter we perform detailed analysis of feature values versus response (target) variable to reveal the importance of each feature as well as its unique value on the success rate of the telemarketing campaign. As in the previous chapter, we group the categorical and numerical features into related groups such as demographics, financial data, and campaign.

3.1 Subscription rate per job category

The rate of customer subscription as well as the total number of subscribed customers per each job category is shown in figure 3.1. The figure shows that though the bank contacted the largest number of people in job segment category of blue-collar, and the number of students the bank contacted was the smallest, the percentage of students who subscribed to term deposit is the largest. We recommend the bank contact more students in the next telemarketing campaign.

3.2 Subscription rate per marital status and educational level

Customer subscription rate as well as the total number of subscribers per each marital status and educational level is shown in figure 3.2

Marital status

The largest number of customers who subscribed are married. This is due to the bank contacted more married people during this telemarketing. Nevertheless, unmarried people are more likely to subscribe (50% more probable than that of married people) though the

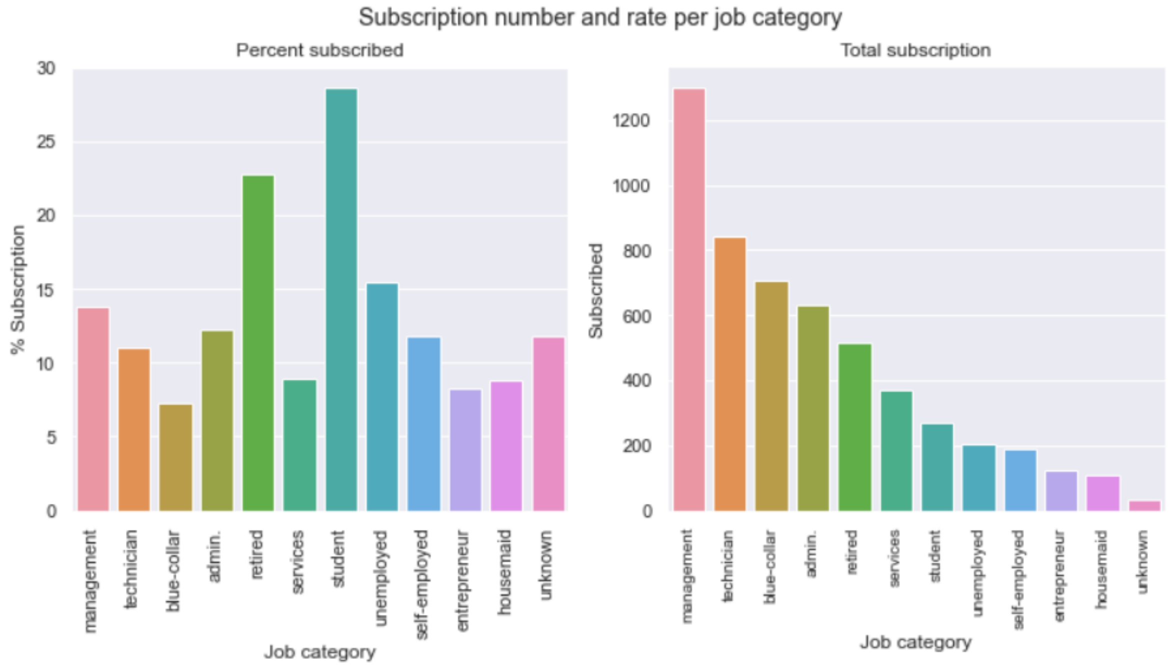


Figure 3.1: Rate of customer subscription and total subscribed number per job category

total number of single subscribers is less than the total number of married subscribers. the rate of subscription for divorced people is also greater than that of married people. We recommend the bank target more single and divorced people in its next telemarketing campaign for term deposit subscription. In contrast, the most probable customers to enroll for the term deposit have completed tertiary level of education. The educational level of the second largest probable group to enroll for the term deposit is unknown; with subscription rate of approximately equal to the average of the three remaining groups. This corroborates the claim the telemarketer forgot to record the educational level of some of the customers contacted.

Educational level

The largest number of subscribers have completed secondary education. Customers with unregistered unknown educational level are the least subscribers. The telemarketer might have forgotten to register or the customers might have not been willing to expose their educational level. In contrast, the most probable customers to enroll for the term deposit have completed tertiary level of education. The educational level of the second largest probable group to enroll for the term deposit is unknown; with subscription rate of approximately equal to the average of the three remaining groups. This corroborates the claim the telemarketer forgot to record the educational level of some of the customers contacted. Customers who completed primary education are the least subscribers and are least probable to subscriber. In its next telemarketing we recommend the bank target more customers who completed tertiary education.

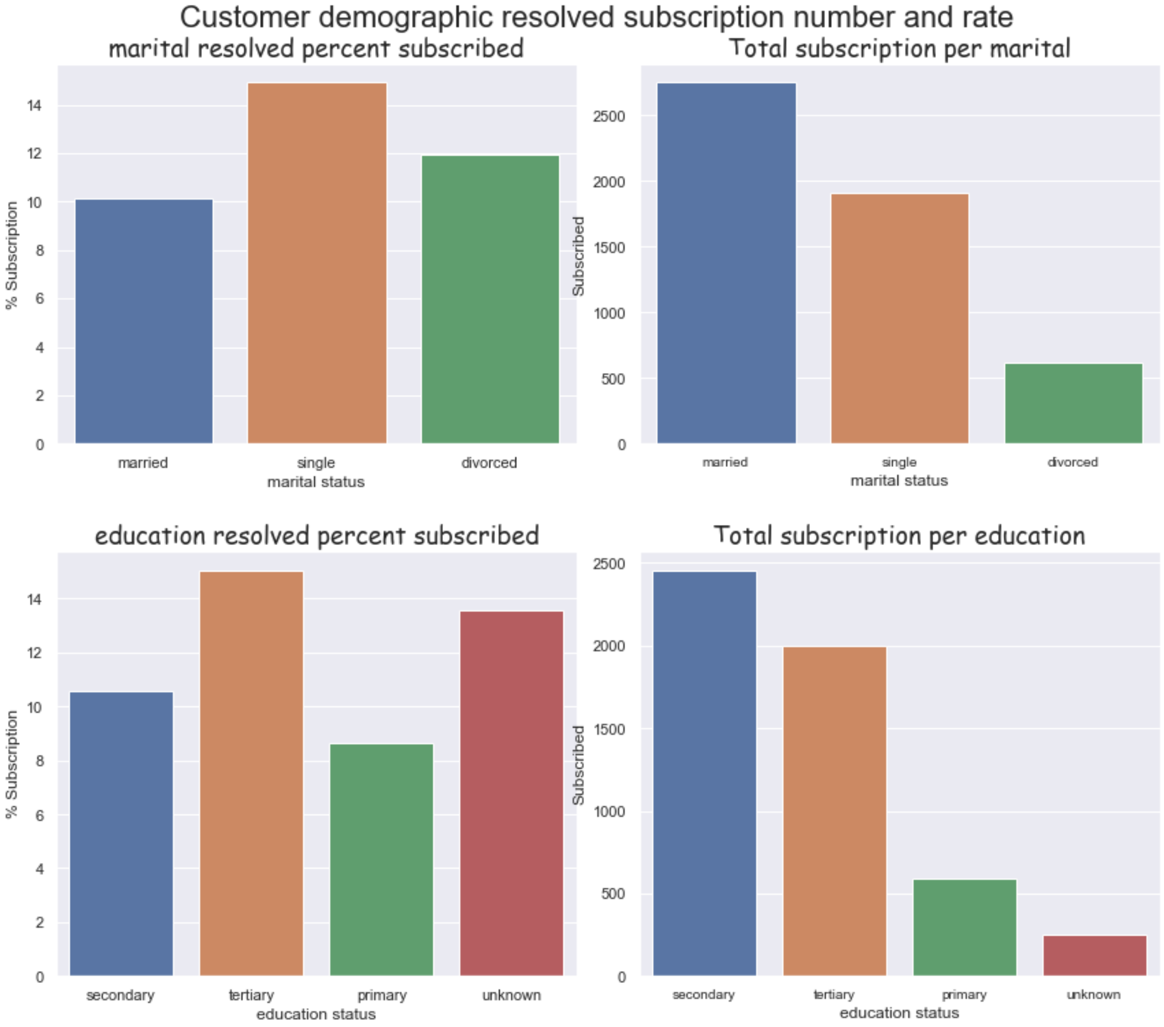


Figure 3.2: Rate of customer subscription and total subscribed number per customer marital status and educational level.

3.3 Subscription rate per customer financial segment

Customer subscription rate as well as the total number of subscribers per customer's financial account is shown in figure ?? each marital status and educational level is shown in figure 3.2

Credit

The number of customers who have credit and subscribed to term deposit is negligibly small. Most subscribers do not have credit. Moreover, the subscription rate of customers with no credit is about 2 times higher than those with credit. In the next telemarketing, we recommend the bank focus on customers with no credit.

Housing loan

The total number of subscribed customers as well as the rate of subscription is significantly higher for people with no housing loan as compared to customers that possess housing loan (about 2 times higher). This is consistent with our previous observation based on marital status of customers where the subscription rate is the highest for single (unmarried) people.

Personal loan

Similarly, the number as well as rate of subscription is the largest for customer groups with no personal loans.

3.4 Subscription rate per campaign details

Customer subscription rate as well as the total number of subscribers per categories based on the details of the campaign such as the method of contact and the month the contact was conducted is shown in figure **??**customer's financial account is shown in figure **??** each marital status and educational level is shown in figure 3.2

Method of contact

Though the majority of the subscribed customers were enrolled throuth talling over cellular phones, the subscription rates for contacts with land line phones and cellular phones are comparable. The larger number of contacts and hence subscription via cellular phones could simply be due to the larger number of customers using cellular phones. Since, whether using cellular or land line does not yield to an appreciable change in subscription rate; the bank should not worry on the type of communication during its telemarketing. We will drop the feature contact during our machine learning developlment as it carries no significant impact on the targeting customers.

Contact season

During this telemarketing the bank conducted the majority of its contacts during summer and spring seasons and obtained a large number of subscribers. Considering summer as a primary vacation season and spring the tax return season, the bank was able to contact a large number of people and get them subscribed. However, the subscription rate of the contacted customers is the highest during fall and winter season. Though this might be studies farther by including more historical data, during this campaign the bank would have better benefitted if it were to rigourously initiate its telemarketing during fall and winter seasons.

Previous outcome

From the histogram plot of the previous outcome status of a customer as, we find that a customer who responded favourably during past campaign is more likely to respond favourably during the current campaign.

3.5 Subscription rate per age group and balance group

For ease of analysis, we have grouped the numerical features age, balance into different practical groups. We grouped customers into groups of no balance, low balance, avg balance, and high balance based on how high a customer's balance is compared to the average balance of the populations. Similarly, customers are also grouped into 4 different age groups with 15 years bin size. The subscription rate of customers versus age group and balance group is shown in figure 3.5

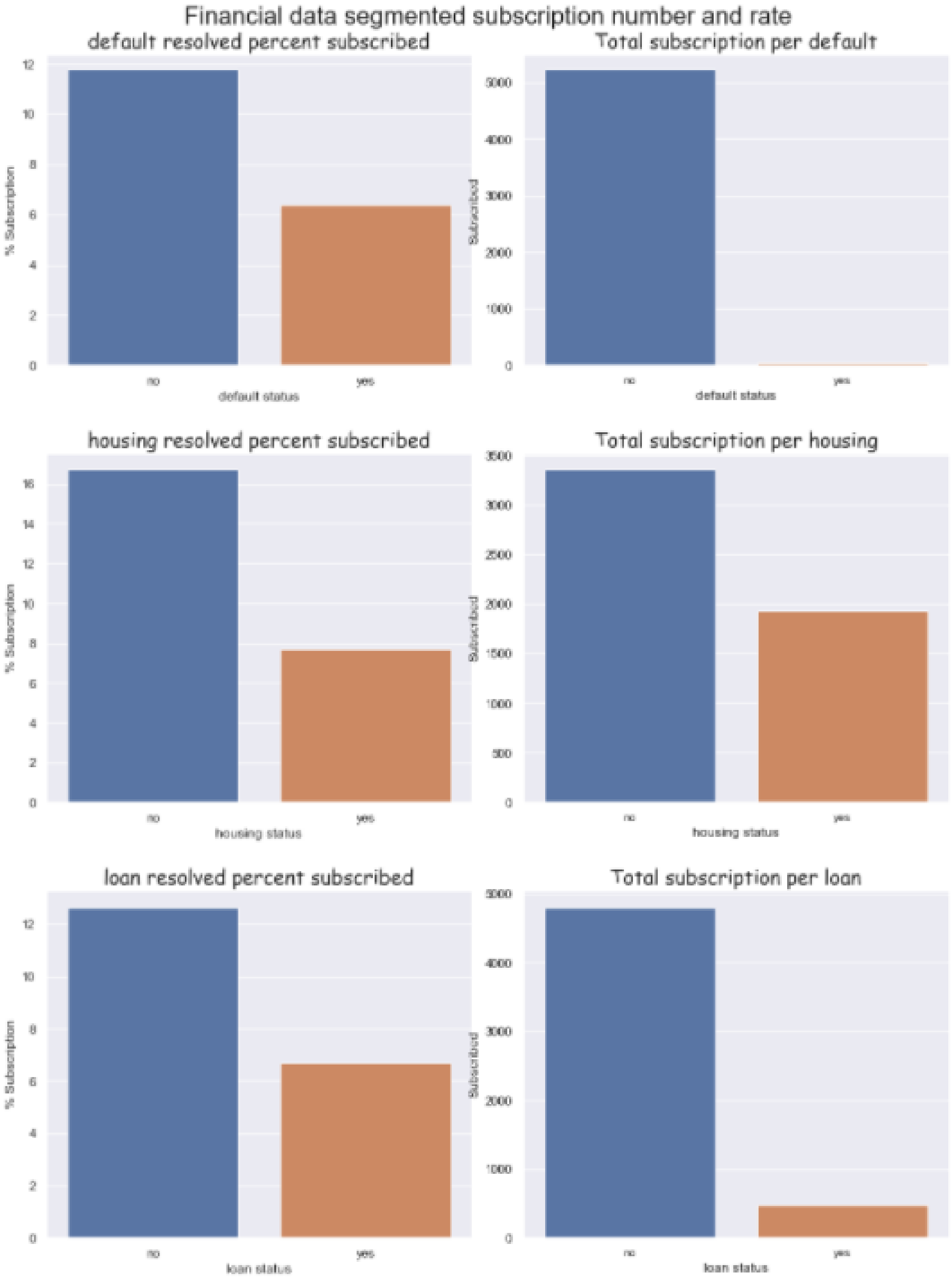


Figure 3.3: Rate of customer subscription and total subscribed number per customer's status of financial account.

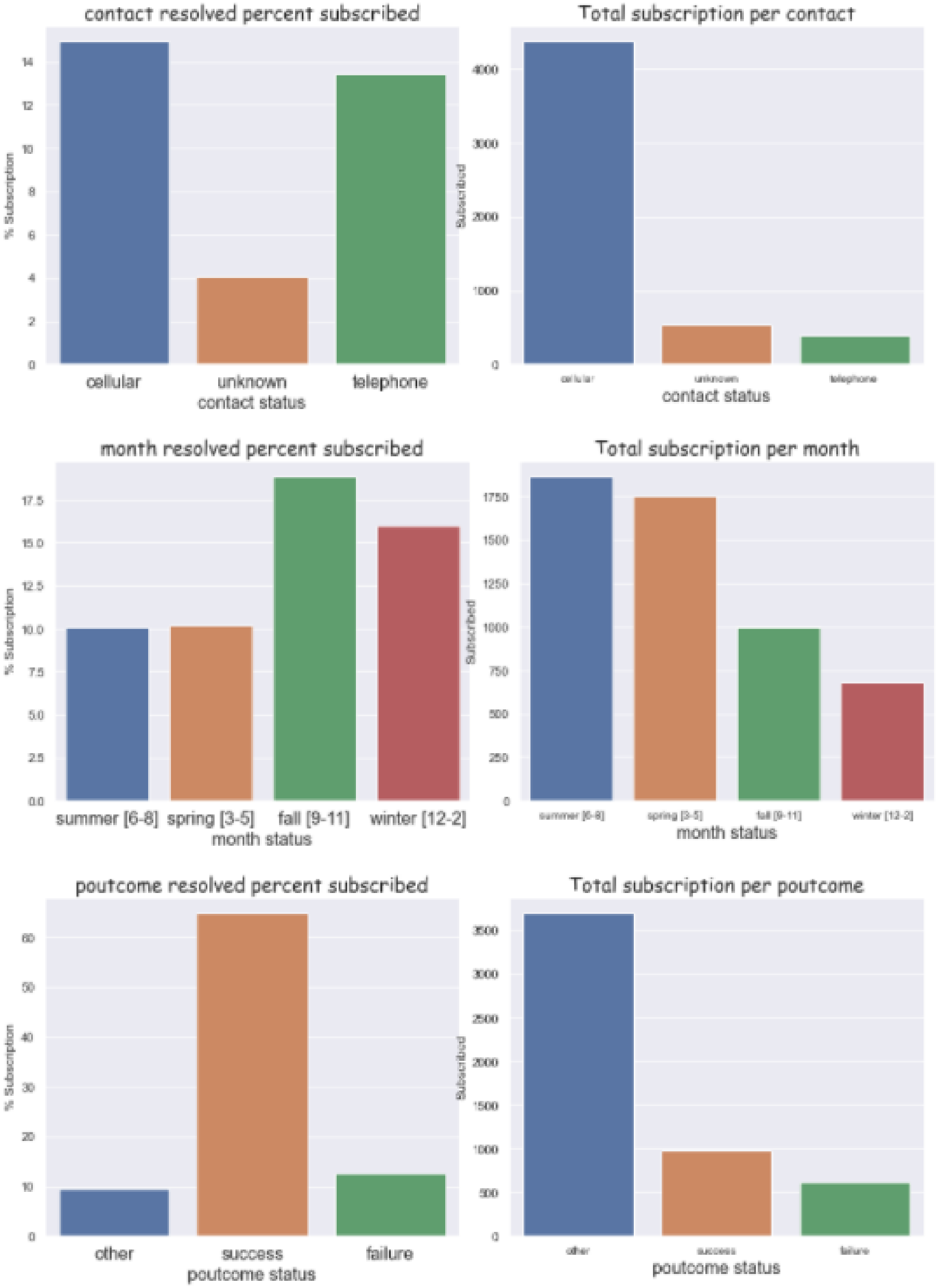


Figure 3.4: Rate of customer subscription and total subscribed number per details of campaign categories.

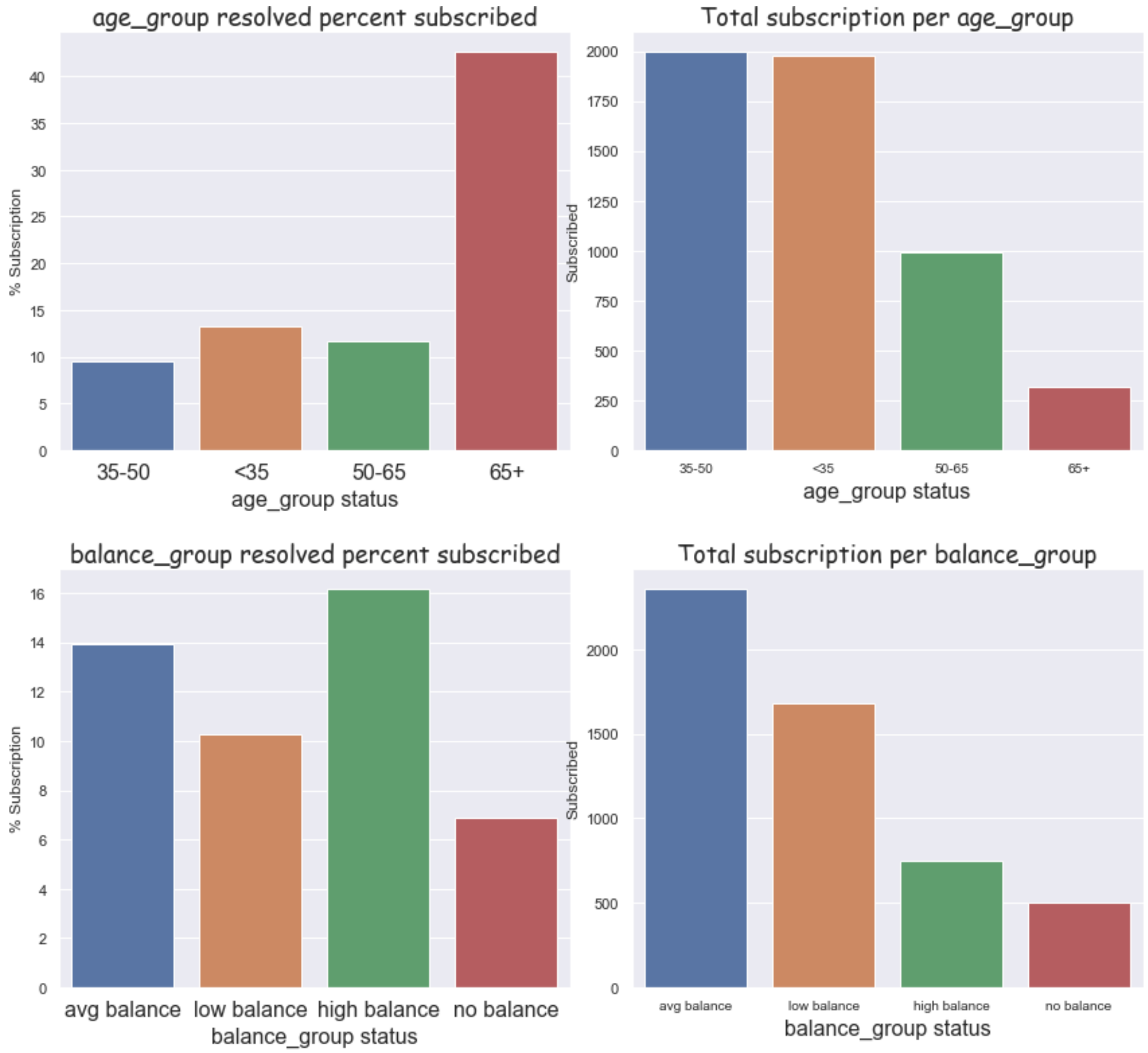


Figure 3.5: Rate of customer subscription and total subscribed number per customer age group and customer balance level.

Chapter 4

Exploratory data analysis

During the previous chapter, we saved a clean and organized dataset. In this chapter we process the cleaned dataset to make it ready for immediate use for machine learning model training. We perform the following list of tasks get a machine learning ready dataset.

- Remove features that do not have predictive value
- Transform categorical features to numerical features
- Select dependent and independent features
- Split dataset into training and testing parts

4.1 Select features relevant to customers

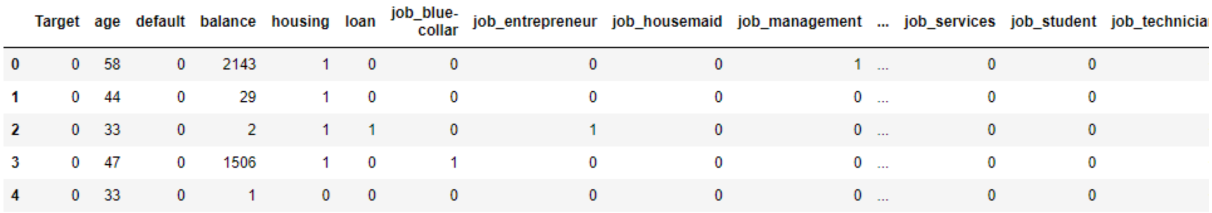
In this project our objective is to predict response of customers to future telemarketing campaign given a given set of customer information, features related to the campaign and other features not related directly to the customers themselves should be excluded when developing our predictive machine learning model. In our dataset, the feature relevant to the customers are the demographics (age,job,marital,education) and cucstomer financial data (default,balance,housing,loan). Only these 8 features and the response (target) variable should be considered when we develop our model.

4.2 Transform categorical features into numerical features

Machine learning algorithms do not work with categorical features. Only features with numberical values are employed by all machine learning algorithms. Hence, categorical features have to be transformed to numbers. As shown above our dataset consists of 7 categorical features. For example the feature education consisted of text labels tertiary, secondary, primary, unknown. One may map the numbers 1, 2, 3, ,4 to each of the text labels of the feature and thereby transform the categorical feature to numerical feature called **ordinal feature**. At first, this kind of mapping may seem to make some sense with 1 corresponding to the highest educational level and 4 to the lowest level. Nevertheless, when in employed in a machine learning model it will be treated just like any other numerical. Specifically, for models that seek to find a linear relationship between features and response the encoding might might lead to undesirable effect. Depending on how linear the relationship of the features and the response variable, such ordinal encoding might work well or not. Nevertheless, the limitation imposed by ordinal transformation could be avoided by the use of another versatile and popular method of categorical encoding called one-hot encoding (OHE).

One-hot encoding (OHE) is a way to encode categorical features without intoducing unintended feature/response relationships like an ordinal encoding, by expanding the categorical feature into as many new features as the number of distinct feature values. For example, in our case OHE will split the feature education into 4 columns corresponding to the number of text labels. Every row will have a value equal to '1' in exactly one column and a '0' elsewhere.

In figure 4.1 a preview of one hot encoded of our dataset is shown.



	Target	age	default	balance	housing	loan	job_blue-collar	job_entrepreneur	job_housemaid	job_management	...	job_services	job_student	job_technicia
0	0	58	0	2143	1	0	0	0	0	1	...	0	0	
1	0	44	0	29	1	0	0	0	0	0	...	0	0	
2	0	33	0	2	1	1	0	1	0	0	...	0	0	
3	0	47	0	1506	1	0	1	0	0	0	...	0	0	
4	0	33	0	1	0	0	0	0	0	0	...	0	0	

Figure 4.1: Preview of our dataset where categorical features are transformed into numerical features with mapping and OHE.

It must be noted that in the case of a categorical feature with at most 2 unique values, the use of of ordinal encoding is justified as two points always define a straight line. Hence, any model including a model that seeks a linear relationship should work well.

4.3 Feature selection and Train/test split

The first column named Target represent the outcome of the campaign and is selected as our independent feature while remaining 21 columns, which contain customer information are selected as dependent features.Also, in order to assess the predictive power of machine learning models it is important to split the data before training the model. The data split labeled as train is used to train the model, while the test split is reserved to assess the performance of the developed model. We split our dataset with train/test ratio of 80:20.

Chapter 5

Machine learning modeling

Now that we have prepared our dataset in the form that could be used by any machine learning algorith, in this chapter we will train a number of classification machine learning alogorithms with grid search parameter optimization and cross validation for evaluation during the training stage. Best perfoming model selected based an appropriate model metric will be tested with the test set of our data set which was held back during model training.

Classification models

The following list maching learning classification models will be tested.

- Logistic regression
- Decision tree
- Random Forest
- Extreme Gradient boosting
- K-Nearest Neighbourhood classifier

5.1 K-fold training data split and cross-validation

For each of the models listed above we use cross-validation to compare the performance of each model during the training. Cross validation is an essential technique when using the training data. It opens up a better of use the training data set through multiple model evaluation (K-fold) during the training phase. Briefly, it involves K-fold splitting of the training dataset where the training dataset is randomly split into K-folds or groups. During the training the first fold is kept for testing while the remaining K-1 group are used for training. The training is repeated k times and each time a different split group is used for validation. However, it is worth noting that each fold should be a good representative of the whole training dataset. Stratified K-fold attempts to solve this issue by preserving class ratio when fold groups are created.

5.2 The imbalanced dataset problem

Clearly our data set is imbalanced and we have to re-sample it before model training. The impacts of imbalanced data are implicit, i.e. it does not raise an immediate error when you build and run your model, but the results can be delusive.

5.2.1 How to deal with the imbalanced data

Several solutions have been suggested in the literature to address this problem, amongst which are:

- **'Data-level techniques'** — At the data level, solutions work by applying re-sampling techniques to balance the dataset. These can be done by oversampling the minority class, which is to synthetically create new instances from existing ones; or under-sampling the majority class, which eliminates some instances in the majority class. However, both techniques can have their drawbacks. Oversampling new data can cause the classifier to over-fit; whereas under-sampling can discard essential information. A combination of both techniques with a heuristic approach can be found in specialized literature with excellent results.

- **'Algorithmic-level techniques'** —Algorithmic level solutions can be done by adjusting weighted costs accordingly to the number of training instances in each class. In parametric classifier like Support Vector Machine, grid search and cross-validation can be applied to optimise the C and γ values. For non-parametric classifier like the decision tree, adjusting the probabilistic estimate at the tree leaf can improve the performance.

- **'A combination of both'** — A hybrid approach is also constantly being explored in various literature, including AdaOUBoost (adaptive over-sampling and undersampling boost) proposed by Peng and Yao and Learning By Recognition, using the concept of auto association-based classification approach proposed by Japkowicz.

Dealing with imbalanced data in Python

One of the most popular libraries for sampling methods in Python is none other than the imbalanced-learn package. It provides several methods for both over- and under-sampling, as well as some combinations methods.

- **Random under-sampling** with 'RandomUnderSampler'

- **'Oversampling'** with 'SMOTE' (Synthetic Minority Over-sampling Technique)

- **'Combination':**a combination of both random under-sampling and oversampling using pipeline

Reference for this notes are from a write up by Jack tan. The link is here https://towardsdatascience.com/how-to-deal-with-imbalanced-data-in-python-f9b71aba53eb

We balanced our dataset using SMOTE. Figure 5.1 shows a histogram plot for the two classes (subscribed and not subscribed). It is clear that using SMOTE resulted to a well balanced dataset with 50:50 class ratio.

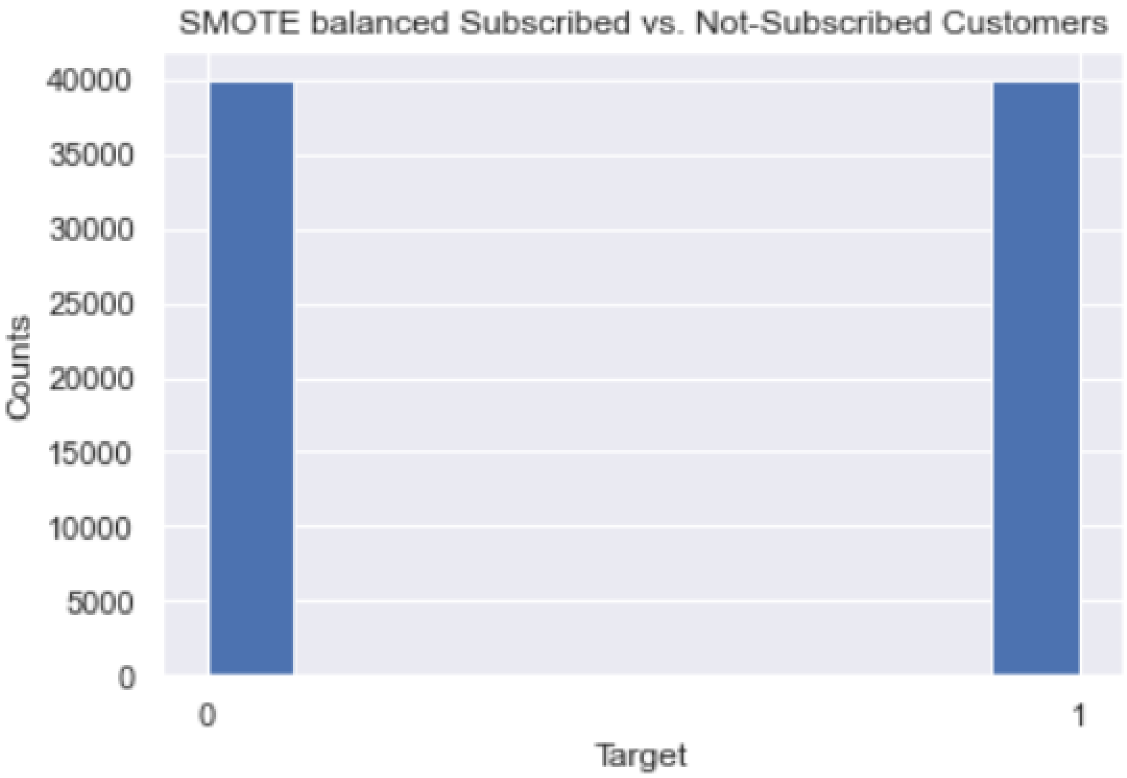


Figure 5.1: Table of evaluation metric values for 5 machine learning models.

5.3 Evaluation metric selection

There are a number metrics that are routinely used to evaluate machine learning models. The metrics include:

- **Accuracy** — the number of correct predictions made as a ratio of all predictions made.

- **ROC** — calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the curve is the ROC$_{S}core$.

- **Recall** — quantifies the number of true positives found

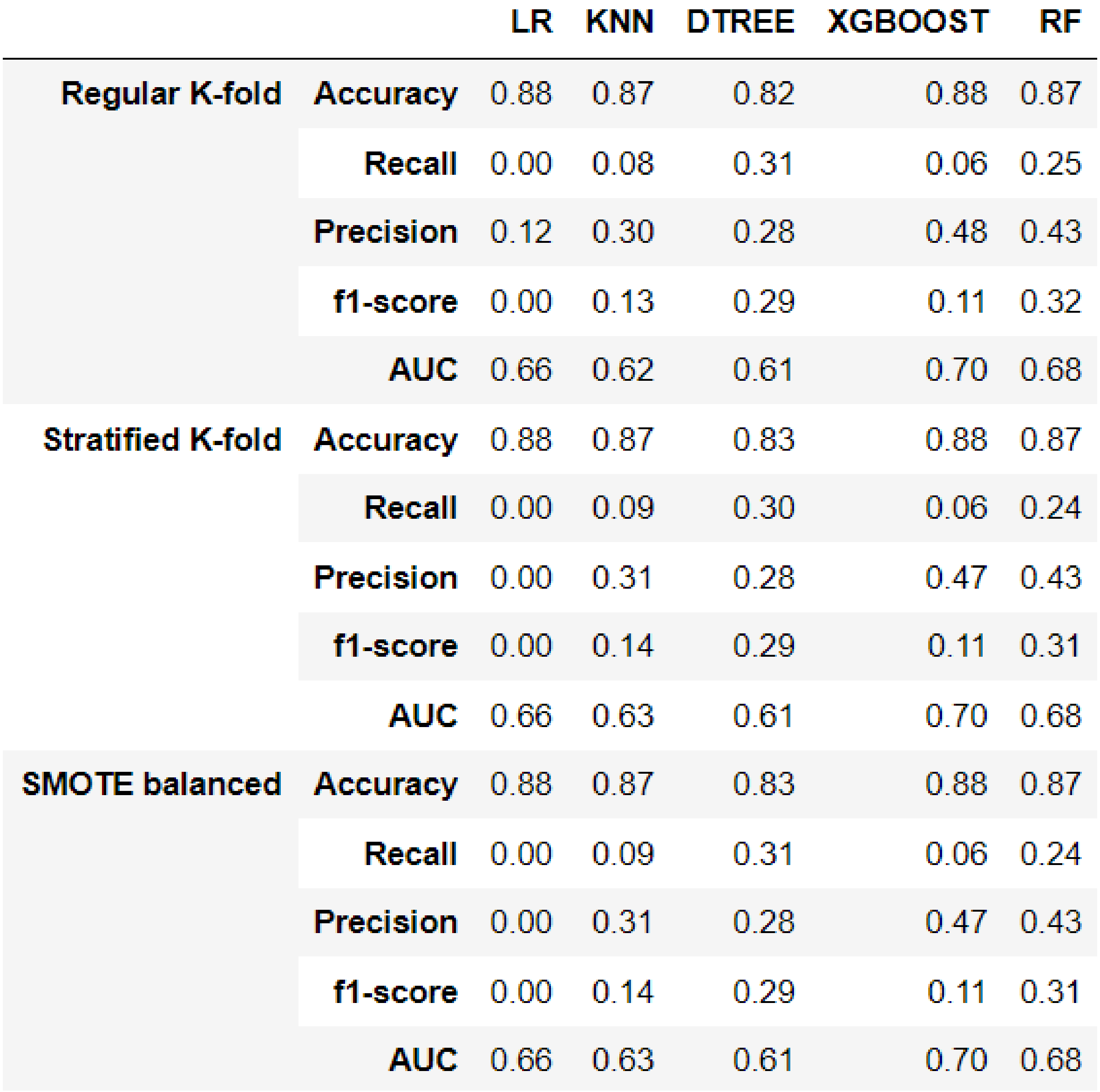
- **Precision** — quantifies the number of positive class predictions that actually belong to the positive class.

Since our dataset is imbalanced, using accuracy as metric would not be appropriate as any dummy could produce with a decent accuracy. I choose ROC$_{A}UCasmymetricofchoicebecauseoftheimb$

ROC analysis does not have any bias toward models that perform well on the majority class at the expense of the minority class — a property that is quite attractive when dealing with imbalanced data. ROC is able to achieve this by looking into both the True positive rate (TPR) and False positive rate (FPR). We will get a high ROC if both the TPR and FPR are above the random line.

5.4 Model evaluation

After splitting our training dataset into 80:20 ratio using regular K-fold splitting and stratified K-fold splitting, we trained and evaluated each of the five classification machine learning algorithms listed above. Figure 5.2 shows a table of the results of the metrics obtained using the two methods of k-fold splitting. We also have trained our model using the smote balanced dataset and the results are also incorporated in the figure. We see that using either stratified K-fold or balancing our dataset using smote resulted to marginal improvement in the value of AUC score. The improvement is only for the case of K-Neighbourhood (KNN) classifier (an improvement from 0.62 to 0.63). For all the other models, when judged based on the AUC score, there is no real improvement in using stratfied k-fold or smote balancing our dataset.



		LR	KNN	DTREE	XGBOOST	RF
Regular K-fold	Accuracy	0.88	0.87	0.82	0.88	0.87
	Recall	0.00	0.08	0.31	0.06	0.25
	Precision	0.12	0.30	0.28	0.48	0.43
	f1-score	0.00	0.13	0.29	0.11	0.32
	AUC	0.66	0.62	0.61	0.70	0.68
Stratified K-fold	Accuracy	0.88	0.87	0.83	0.88	0.87
	Recall	0.00	0.09	0.30	0.06	0.24
	Precision	0.00	0.31	0.28	0.47	0.43
	f1-score	0.00	0.14	0.29	0.11	0.31
	AUC	0.66	0.63	0.61	0.70	0.68
SMOTE balanced	Accuracy	0.88	0.87	0.83	0.88	0.87
	Recall	0.00	0.09	0.31	0.06	0.24
	Precision	0.00	0.31	0.28	0.47	0.43
	f1-score	0.00	0.14	0.29	0.11	0.31
	AUC	0.66	0.63	0.61	0.70	0.68

Figure 5.2: Table of evaluation metric values for 5 machine learning models.

Figure 5.3 shows a scatter plot of AUC versus Accuracy for the five models. As can be seen XGBOOST resulted with the highest AUC score (70%) while still maintaining a high accuracy of 88%. We therefore, choose XGBOOST as our base model.

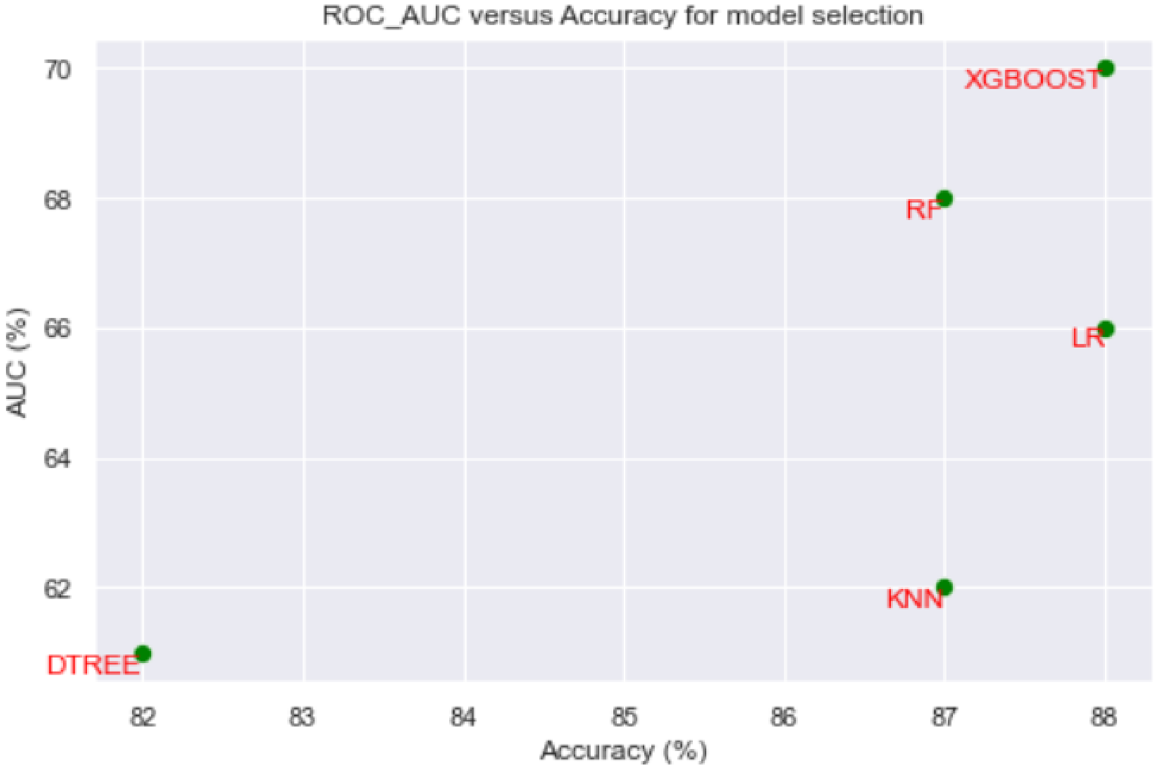


Figure 5.3: AUC score versus Accuracy showing XGBOOST performing the best.