Data Science - Bank Marketing Campaign

1. Group Information

Group Name: Datalux

Group Members: 3

Name	Email	Country	Uni/Company
Huu Thien Nguyen	nguyenhuuthien27296@gmail.com	Sweden	Skövde University
Asmaa Alqurashi	asmaa.idk@gmail.com	Saudi Arabia	Taif University
Deepak Rawat	deepakrawat68@gmail.com	Ireland	Dublin Business School

Specialisation: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: LISUM09

2. Problem Description

2.1. Introduction

A common approach for increasing business is to run marketing and selling campaigns. Companies use direct marketing to reach certain categories of clients to achieve a specific goal. Customer distant interactions may be centralised in a contact centre, making campaign administration easier. Technology allows us to reimagine marketing by optimising customer lifetime value through analysing accessible data and customer KPIs, allowing us to develop longer and closer relationships in line with company needs.

The ABC Bank wants to market its term deposit product to clients in this project. Before doing so, they want to construct a machine learning model that will assist them in determining whether a particular consumer would buy their product based on the customer's previous interactions with the bank or other financial institution.

To solve the problem mentioned above, the bank wants to use machine learning modelling to identify the customers who are more reluctant to buy their services so that their marketing channels will only focus on these customers, which in turn will save time and resources and finally leads to optimised cost for this campaign.

3. GitHub Link

The link for GitHub: https://github.com/AndrewNguyen27296/DataGlacier

4. EDA performed on the data

4.1. Thien

Methods

In order to deliver the knowledge and insight to decision-makers, we will analyse the KPIs and their interrelations using the exploratory data analysis (EDA) methodology. The pandas profiling module was implemented to offer an accelerated summary of each attribute in the dataset for simplistic univariate analysis. Contingency charts with heatmaps were appointed to take advantage of the connection between categorical factors because the banking data involves a variety of variables. The box plot was also created to examine any irregularities in the amount within the features. The multivariable analysis was completed by using the Tableau tool. Below are three methods used for the EDA process.

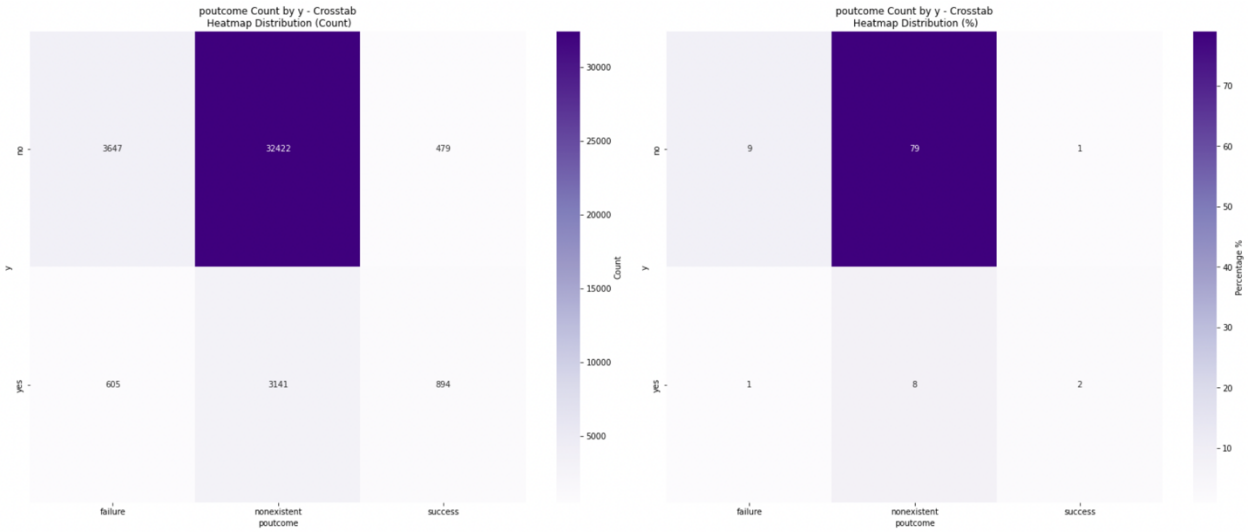


Figure 1: Crosstab heatmap of poutcome and y

The data distribution between two category variables was evaluated using a crosstab heat map in the image above. Due to the fact that 79% of the values are in the nonexistent sector, it showed a severely unbalanced output feature. Because this strategy aimed to examine pair-wise interactions between the two non-integer features, an automated function was applied to all categorical attributes. With the help of the procedure, more than fifty heatmaps were produced. The absolute count of samples is shown on the left side, and the percentage of the sub-sector segment is presented on an exemplary chart. The gradient colour scheme was chosen to draw attention to the low-to-high distribution.

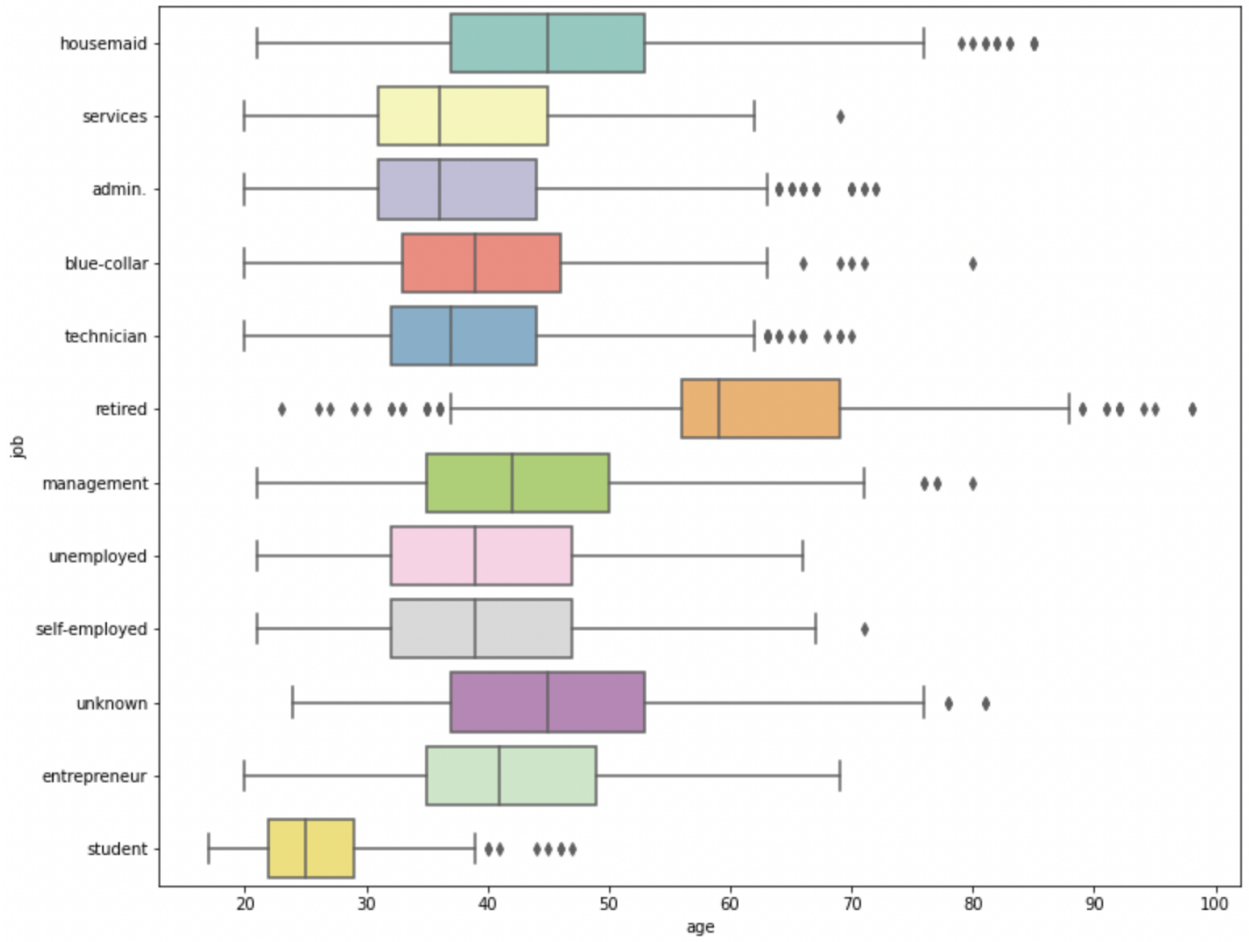
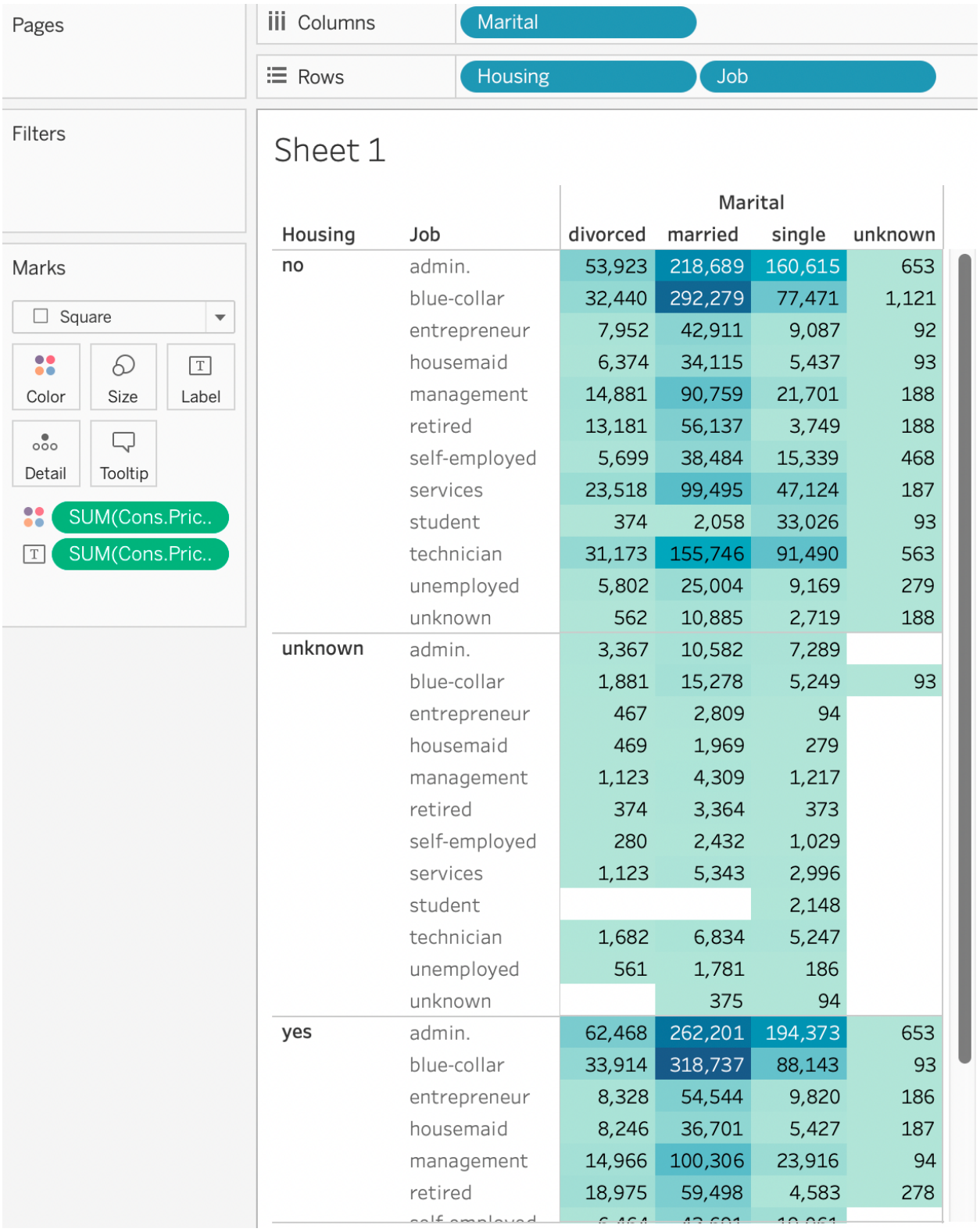


Figure 2: Boxplot showcases the outliers and the distribution for pair-wise analysis between categorical and numeric

According to the graph above, the applicant is typically between 30 and 50, except for students and retirees (naturally). Additionally, it is necessary to look into a few dubious values in more detail. For instance, some clients retire around age 20 to 40, which is too soon in real terms. It can be explained because several are wealthy or have financial inheritances for early retirement. Students in their 40s to 50s are another group. Given that they are PhD students, this occurrence can be understood. However, the plot provided a summary of the outlier drill-down in each dimension.



Sheet 1

Housing	Job	Marital			
		divorced	married	single	unknown
no	admin.	53,923	218,689	160,615	653
	blue-collar	32,440	292,279	77,471	1,121
	entrepreneur	7,952	42,911	9,087	92
	housemaid	6,374	34,115	5,437	93
	management	14,881	90,759	21,701	188
	retired	13,181	56,137	3,749	188
	self-employed	5,699	38,484	15,339	468
	services	23,518	99,495	47,124	187
	student	374	2,058	33,026	93
	technician	31,173	155,746	91,490	563
	unemployed	5,802	25,004	9,169	279
	unknown	562	10,885	2,719	188
unknown	admin.	3,367	10,582	7,289	
	blue-collar	1,881	15,278	5,249	93
	entrepreneur	467	2,809	94	
	housemaid	469	1,969	279	
	management	1,123	4,309	1,217	
	retired	374	3,364	373	
	self-employed	280	2,432	1,029	
	services	1,123	5,343	2,996	
	student			2,148	
	technician	1,682	6,834	5,247	
	unemployed	561	1,781	186	
	unknown		375	94	
yes	admin.	62,468	262,201	194,373	653
	blue-collar	33,914	318,737	88,143	93
	entrepreneur	8,328	54,544	9,820	186
	housemaid	8,246	36,701	5,427	187
	management	14,966	100,306	23,916	94
	retired	18,975	59,498	4,583	278

Figure 3: The advanced highlights table, which provides an analytic view of four variables at the same time.

For in-depth feature analysis, the Tableau program is a fantastic visualisation tool. Multiple charts and an advanced analytics view are included. The application's strength is that non-technical users can perform exploratory visual analytics by dragging and dropping objects instead of programming in a notebook or coding environment. The table in this example shows a matrix of four variables: marital status, housing stock, employment status, and consumer price index.

Results

Insight #1: Sampling techniques and ensemble methods are required when building classification due to an imbalanced dataset.

The critical imbalance between the dependent variables is 88.7% for refusal (no) and 11.3 % for customers who subscribe to a term deposit. This is a common issue in the banking industry and is depicted in the figure below. When a machine learning model is developed using the original information, it might result in a significant misclassification. Techniques for sampling can be used to solve this problem. It comprises strategies for under-, over-, and random selection. Additionally, the k-folding method can produce a reliable AI model. Finally, the ensemble learning family bagging, boosting, and stacking method can improve the model's predictive model.

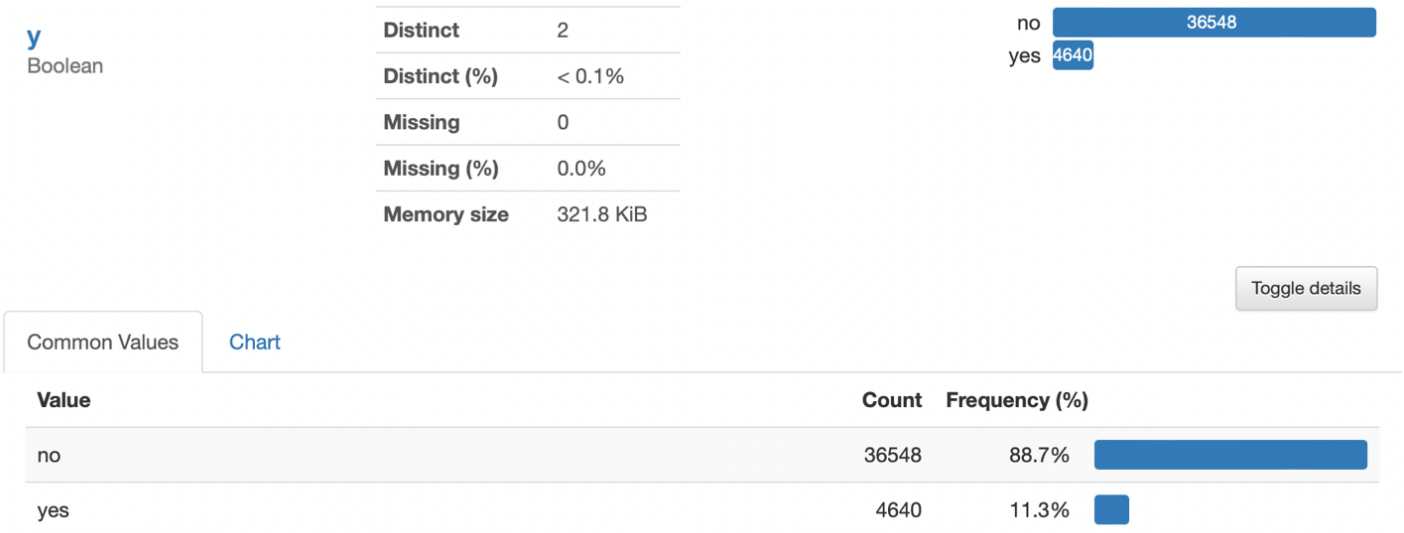


Figure 4: The imbalance of dependent variables in the banking dataset

Insight #2: Feature engineering with highly correlated numeric variables.

Pdays feature should be excluded when modelling or re-examining in the data collection step. There are two primary reasons that led to this decision. Firstly, the distribution is imbalanced; 96,3% is a "999" value, which does not provide any valuable information. Secondly, it correlated to the "previous" feature and will cause noises when building the machine learning model. Nevertheless, there is a proposal technique that can improve this situation. The feature engineering method can bin the variables into two groups, "contacted" and "not contacted". More importantly, domain knowledge in the banking sector is required for this process since several variables are difficult to engineer without business understanding. For instance, with the insight from the figure below, "cons.conf.idx", "euribor3m", and "nr.employed" can be combined since they are highly correlated.

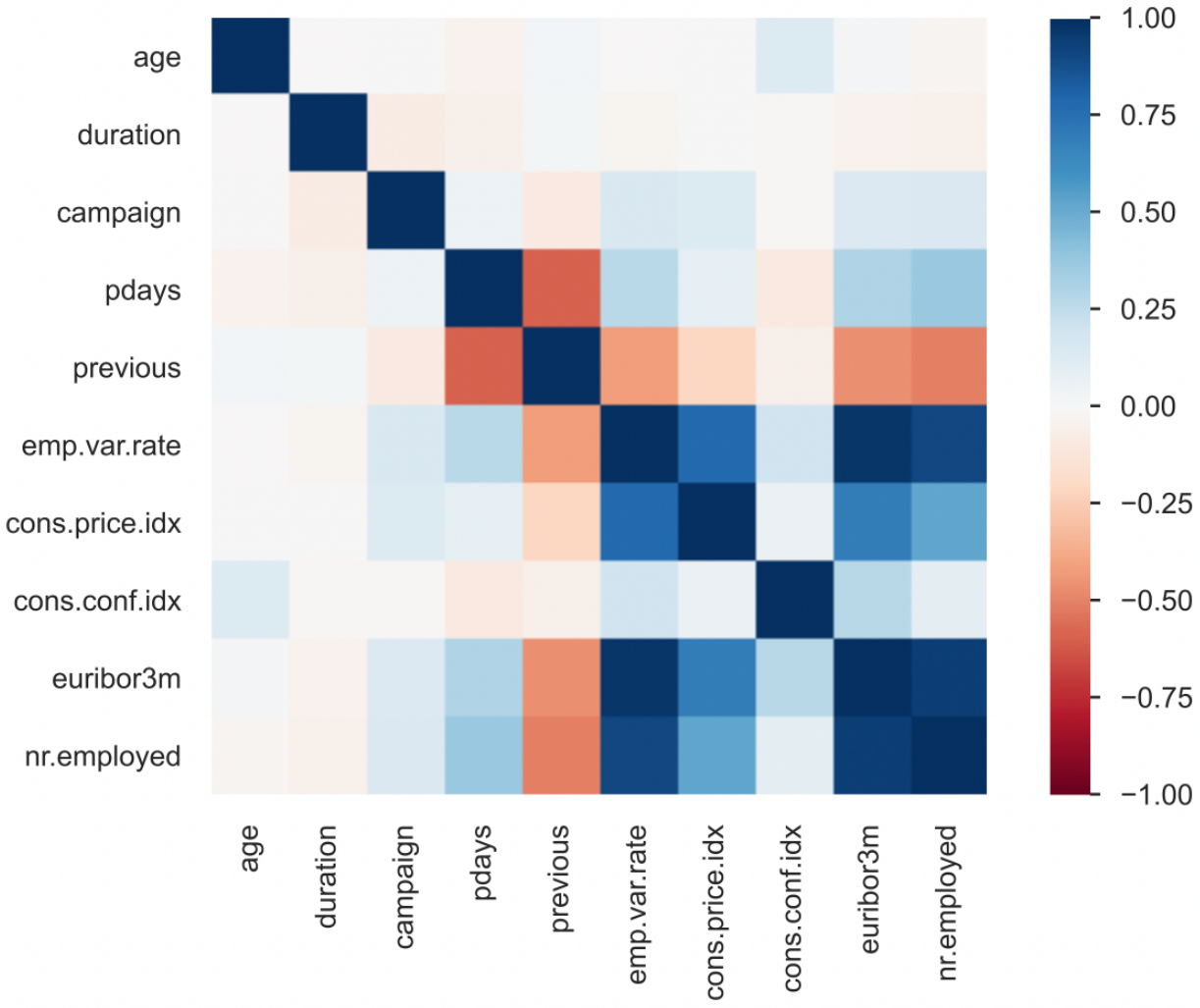


Figure 5: Correlation matrix of numeric variables in the banking dataset

Insight #3: Deploying marketing campaign on primary client segment (subscribed term deposit customers), which are married/single, non-existent poutcome, and do not have loans.

Since the subscribers are what created the profit in the banking sector, the figure below summarises the segment of customers based on work profession, marriage status, poutcome, and loan history. The filter of subscribers (Y=yes) and count of subscribed cases> 110 was applied to help the decision-maker to focus on the essential groups (which can be modified using the domain insight). If a client is single, the bank-firm needs to target admin and technicians to maximise the profit. Vice versa, admin, blue-collar, managers, retired, and technicians are the leading group of customers to focus on. Customer service can involve in the process of improving the user experience, thus increasing the bank's reputation.

Subscribe a term deposits (Primary customer segment)

			Job				
Loan	Poutcome	Marital	admin.	blue-coll..	manage..	retired	technician
no	nonexistent	married	348	289	123	160	204
		single	324				153

Figure 6: Primary customer segment, which needed to focus on the marketing/customer service

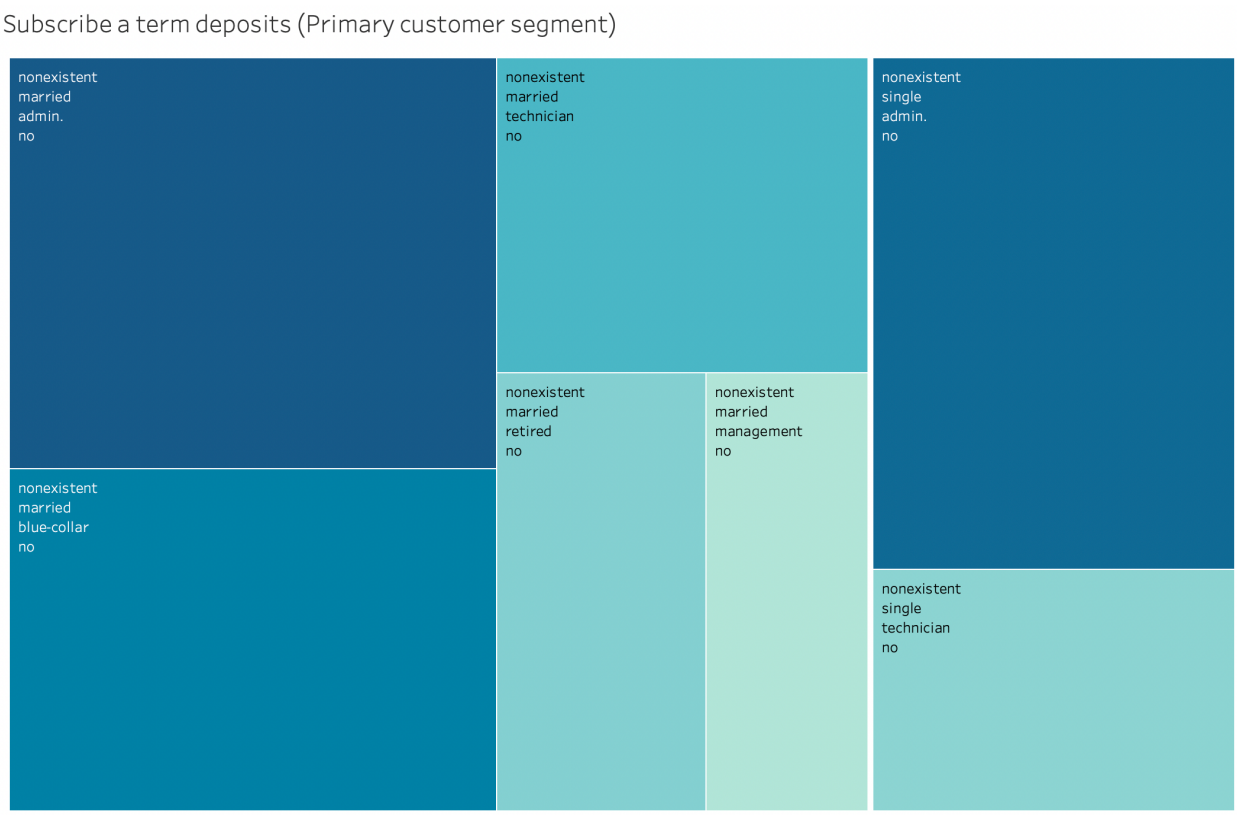


Figure 7: Segmentation of customers who subscribed to a term deposit

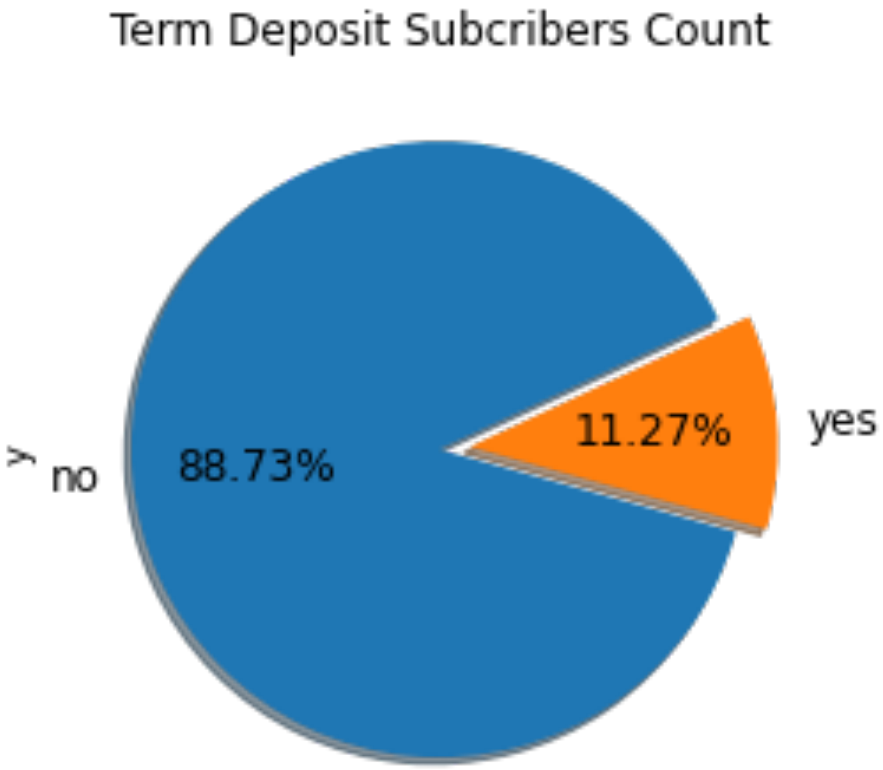
4.2. Asmaa

Methods

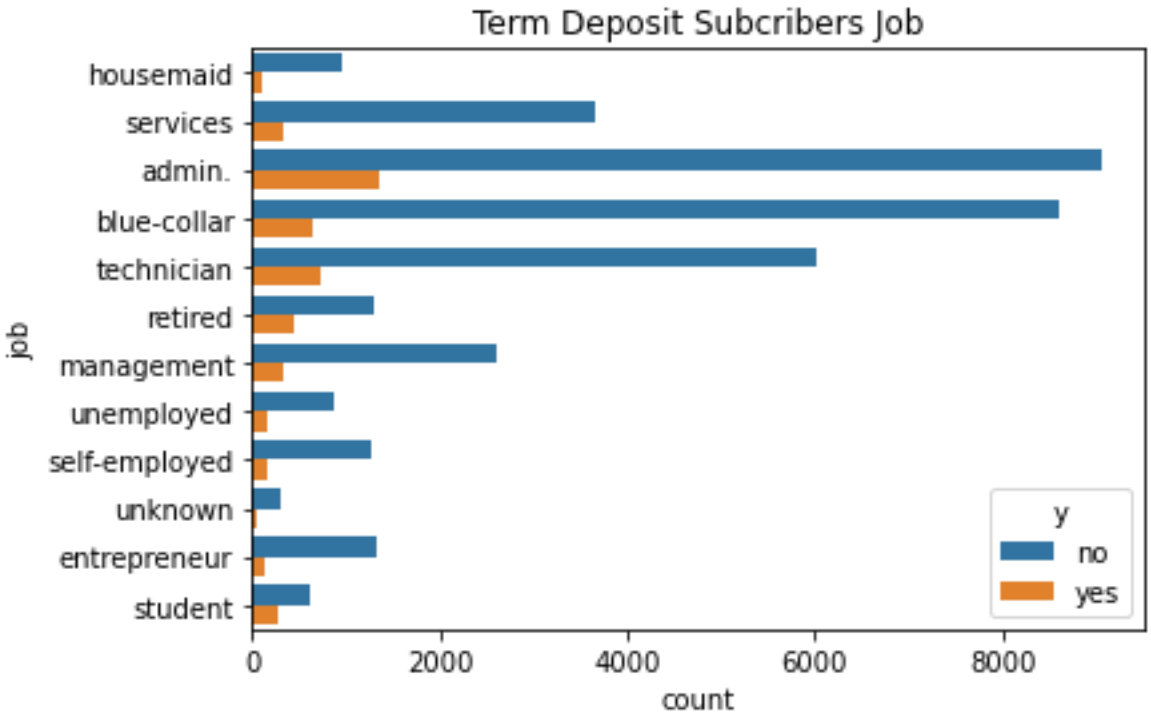
- Bar plots to explore Categorical features.
- Histograms to explore numeric features.
- Explore each feature with the target column to find the relations.
- A heatmap after changing the target column to numeric values to see the correlation between the target and other numeric features.

Results

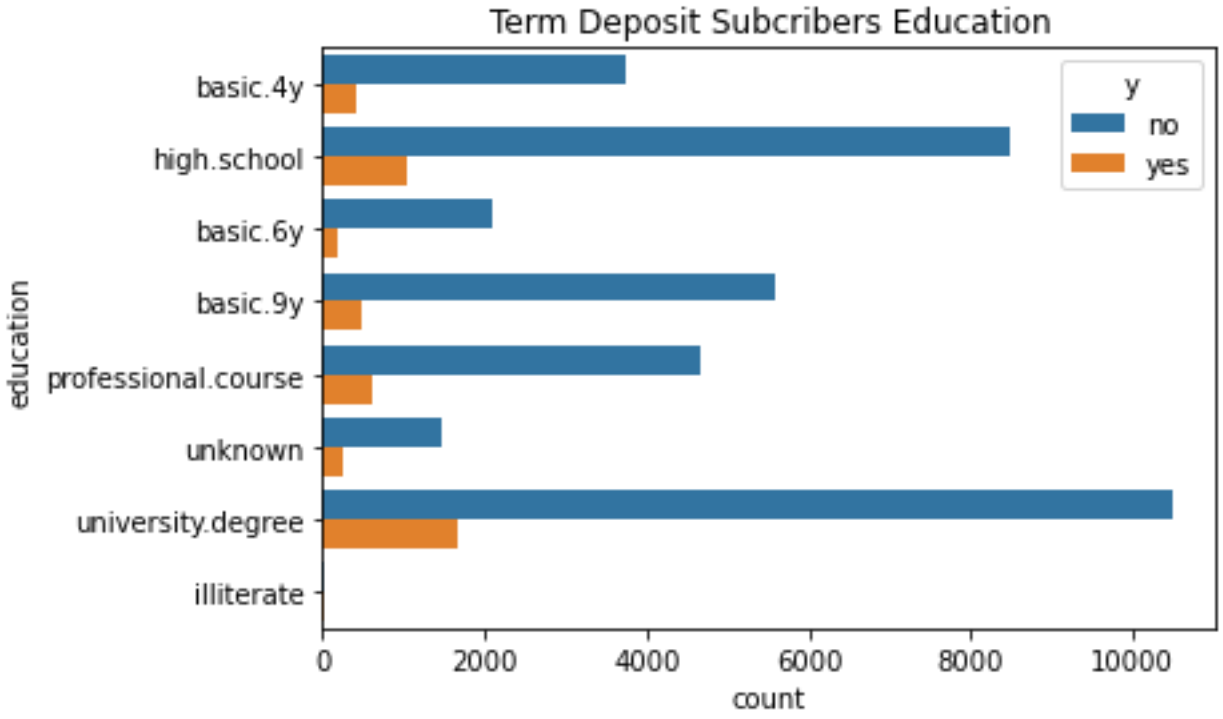
- Only 11.27% of the clients are subscribed to the term deposit.



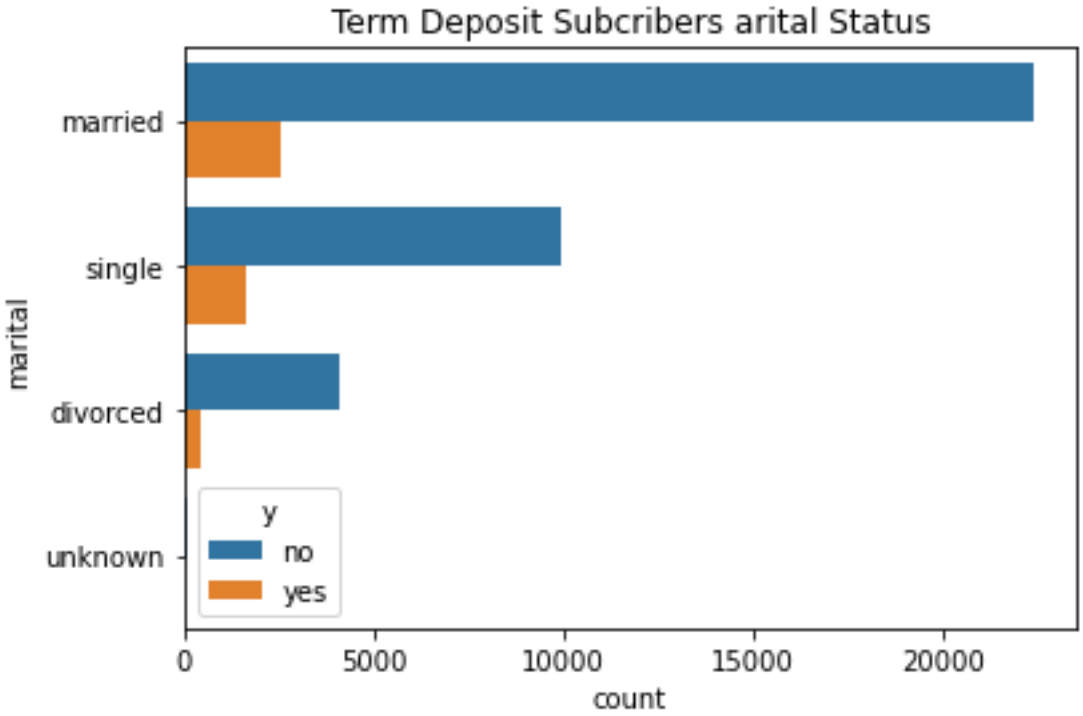
- Administer and technician are the jobs with most subscribers to the term deposit.



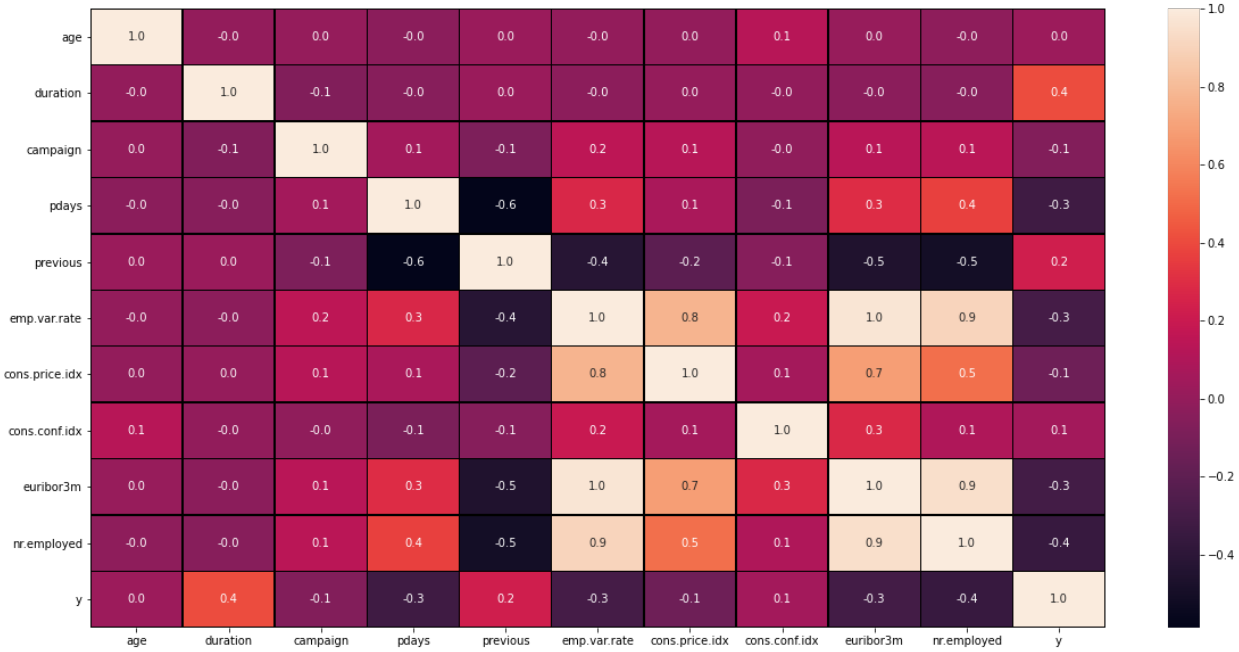
- More subscribers to the term deposit with a university degree.



- Most clients are married, so the number of married subscribers is higher, but relatively, singles are less but subscribed more to the term deposit.



- Duration has a high correlation with the target.



4.3. Deepak

Methods

- Barplots are used to explore categorical values
- Plots are created between two categorical variables and analysed in the data.
- The correlation heat map is used to find the relation between numerical values.
- Few graphs and insights are created using MS Excel

Results

Insight #7: From the above graph, we can conclude that the target variable is highly imbalanced, with a 'yes' count of 4640 and a 'no' count of 36548. So, to work with various machine learning models, we need to sample this column to predict both outcomes correctly. We can do up-sampling or downsampling to overcome this challenge.

Insight #8: The variables having a high correlation with one another provide no extra information for the machine learning models. So, for machine learning modelling purposes, we need to remove the features with high correlation and use only one. The features like 'emp.var.rate', 'euribor3m', 'cons.priceidx', and 'nr.employed' have a very high correlation.

Insight #9: The duration of most calls is under 600 seconds (10 minutes). The average call duration is 4 minutes; it is recommended to target the customers to have a call duration of approx. 5 minutes.

5. Final Recommendation

Insight #1: Sampling techniques and ensemble methods are required when building classification due to an imbalanced dataset.

Insight #2: Feature engineering with highly correlated numeric variables.

Insight #3: Deploying marketing campaign on primary client segment (subscribed term deposit customers), which are married/single, non-existent poutcome, and do not have loans.

Insight #4: Profile the clients to target the right group for the campaign: job (admin, technician), education (university degree) since we can see from the plots the high number of subscribers of these clients.

Insight #5: The month of the contact impacts the response of the clients (more clients subscribed in May), and more calls were made in May.

Insight #6: Increasing the call duration impacts the response to the campaign.

Insight #7: Lots of customers from the previous outcome are not contacted, which are categorised as nonexistent, so this time we can get these customers for the campaign.

Insight #8: The average call duration to the customers is around 4 minutes, which does not seem enough time to convince them to understand new services and products so that it can be increased.

Insight #9: From the correlation heat map, we can infer that the numerical variables having a high correlation with one another do not provide additional information, so these variables (‘emp.var.rate’, ‘euribor3m’, ‘cons.priceidx’, ‘nr.employed’) can be dropped (According to Tabachnick, B. G., & Fidell, L. S. (1996)) while selecting features.