Simplified Data-Driven Approach to Predict the Success of Bank Direct Marketing Campaign

Project - SCS 3250 -029 – Prof. Wilson So
Group 3: Victoria Bateman - Cesar Juliano - Maristela Schiavo - Kaushik Sharma

Contents

Abstract

Data science offers banks a set of tools to identify customers most likely to sign-up to financial products, in this particular instance term deposits, which reduces the time, resources and costs required to reach and convert the highest possible number of customers with fewest number of calls.

Our primary goal was to understand the dataset, track down the most important attributes that have an outsized influence on the outcome and then prescribe alternatives that can further improve the success rate for the bank. By using basic statistical techniques, graphical representations and Random Forest algorithm, we were successful in building a model that can predict with an 89% accuracy, the success rate of this direct marketing campaign.

The dataset used was originally from a paper published by S. Moro, P. Cortez and P. Rita in March 2014, titled - "*A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems*". We accessed this dataset from the UCI Data Repository where the information has been truncated to include 41,188 unique rows and 21 attributes. The dataset is primarily a combination of demographics data, campaign operational data and economic indicators. The goal of the paper was to use machine learning algorithms to build a model, that can predict success or failure when a customer is targeted to open a term deposit account.

Age, Education and Job profile came up as the most important customer attributes that predicted success. In addition, we also reviewed economic indicators like Euribor rate, Consumer Confidence Index and Employment rate to further support the underlying assumptions. Overall, our suggestion is to not only focus on the most obvious success criteria, but also branch out and target underserved groups in order to drastically improve the success of the campaign.

1. Overview

1.1. Introduction

Due to the increasing number of direct marketing campaigns the effectiveness of such campaigns has decreased over time. In addition, drastically varying economic landscape and strong market competition across the world has led to marketing strategies that invest in campaigns that focus on obvious customer segments. By using data science, more specifically machine learning, it is possible to drastically improve the effectiveness of direct marketing campaigns and, therefore, the bank's return on investment.

Data science offers banks a set of tools to identify customers most likely to sign-up to financial products, in this particular instance term deposits, which reduces the time, resources and costs required to reach and convert the highest possible number of customers with fewest number of calls.

This project is about a series of direct marketing campaigns (phone calls to customers), executed by an unnamed Portuguese bank, to selected customers offering them a term deposit account. The bank conducted 41,188 potential between 2008 and 2010. The directing marketing campaign customer dataset used for this project contained 21 different data points information which includes a variety of demographics data, campaign operational data and economic indicators.

This report contains a more detailed explanation of the project problem statement as well as a high-level overview of the data science approach taken to explore this business problem. Overall, the recommendation is to not only focus on the most obvious success criteria, but also branch out and target underserved groups in order to drastically improve the success of the campaign. Underpinning this the report contains three recommendations: 1) a holistic demographic profile of the ideal target customer; 2) the most favourable economic circumstances for conducting this type of direct marketing campaign; and, finally, 3) the ideal time of the week to contact potential customers. In addition to the three key recommendations that should be immediately implemented, future considerations for optimising and operationalising this approach for long-term success.

An appendices has been added at the end of this report. The appendices includes a detailed explanation of the data and mathematical analysis used in this data science project, and additional data analysis carried out but not used to support the three key recommendations contained in this report.

1.2. Data Overview

Our main source of data is from a technical paper published by S. Moro, P. Cortez and P. Rita, titled - "*A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems*". This paper was officially published on March 13, 2014 and made available to the general public on the UCI Data Repository, starting June 2014.[1]

This data was collected by a Bank in Portugal, while conducting an direct marketing campaign. The goal of the campaign was to convince the targeted customer to sign up for a term deposit account. This dataset was collected during the months leading up to and including the worst recession that the world had seen in decades i.e. from 2008 - 2012. The authors eventually reduced the overall dataset to include 41,188 unique instances, which also included personal demographics, economic indicators and campaign logistics as the attributes for each individual customer. Additional information about the dataset is provided in the appendix for further review.

By accessing this data from the Bank telemarketing program, the authors were successful in extracting individual characteristics of the customers including age, education level, loan history, job that the individual is currently occupied in etc.,. We were also provided information on the campaign logistics such as day of the week the contact was made, month of the campaign, duration of each call and whether the customer had been contacted previously or not. Additionally, in order to take into account the external factors that might affect the outcome, the authors included major economic indicators such as Euribor rate on a quarterly timeframe, Consumer Confidence Index, Employment rate and Consumer Price Index data for the time period under consideration.

1.3. Problem Statement

Having reviewed the data in detail and further explored the circumstances during the telemarketing campaign, the team decided to ask a simple question:

"With the data collected, can we predict the willingness of a customer to open a term deposit account?"

By asking this question, we are now focusing our efforts on a classification problem. Essentially, the end goal for the team is to build a mathematical model that can successfully predict a customer's behavior, given the various factors under consideration.

This question is in line with the original authors objective as well. By attempting to predict the outcome, we hope to provide a data driven approach to organize and run future marketing campaigns across the bank.

[1] "UCI Machine Learning Repository: Bank Marketing Data Set." 14 Feb. 2012, https://archive.ics.uci.edu/ml/datasets/bank+marketing. Accessed 8 Aug. 2018.

2. Analysis

The team attempted to analyze the data with a strong focus on reducing bias. It was a multi-step process and included several attempts at building a viable model that can predict success or failure in a campaign. The steps taken are as follows:

- The team started out by tracking down any missing data points amongst the various data fields.
- We then attempted to replace missing information using standard statistical techniques.
- Once we were satisfied with the data quality, we then focused our efforts on converting all non-numeric data points into numerical data.
- This process gave us the ability to further review the data in detail, and come up with our recommendations that are stated below.
- We then passed the data set through several machine learning algorithms to find the best predictive model possible.
- We ended up picking the Random Forest was our preferred algorithm, and built a model based on the most significant attributes for each customer.
- The model is currently capable of predicting with an 89% accuracy rate, whether a customer is likely to sign up for a term deposit or not.
- Finally, we shifted our focus to exploring other areas that we might look at in the near future, so that we can further improve our accuracy rate.

The technical specifications, including a detailed review of the code is made available in appendix 1.

3. Recommendations

After going through the dataset and running mathematical algorithms to predict the final output, the team was pleasantly surprised by some of the correlations discovered, that led to the current campaigns success.

We have split our recommendations into two parts, with the express intent to focus on not only the most obvious customer demographics, but also the time in which to contact those customers. Our model was most successful when both of these factors were taken into account with over 89% accuracy achieved across the dataset.

There are a few other variables that were included in the dataset that added its own imprint on the overall success of the campaign. We have listed them all in appendix 3, in the form of graphs. This is to show that there are several factors, apart from the most obvious ones, which should be explored, in order to improve our ability to predict success from a marketing campaign.

Who to Call?

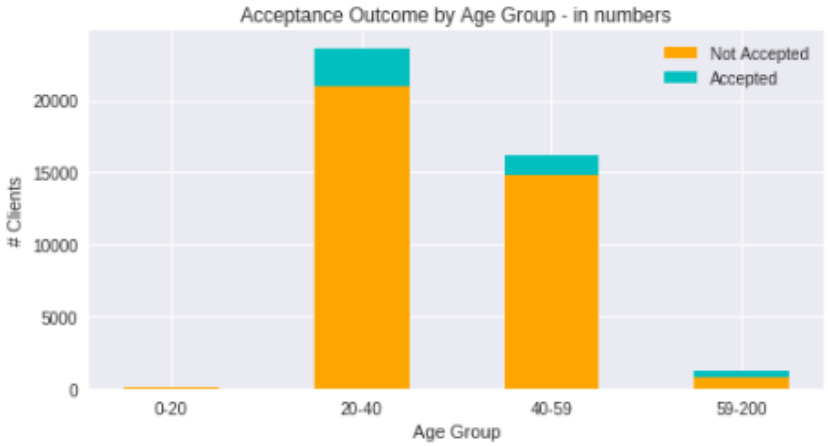
Age Range

Age has been identified as the top information associated with campaign success. Based on the data, it was very obvious that a serious attempt had been made to focus on customers in the age range of 20 - 40 years. This was inherently smart, as the output suggests, over 11% of this age group accepted the offer and signed up for a term deposit account.

By reviewing our data further, we realized that another age group had a much higher success rate than the 20 - 40 year olds. With almost 39.56% success rate within their group, customers in the age range of 59 years and above proved to be much more receptive to our message. This data point makes a lot of sense when you realize that a majority of the 59 years and older customers are probably retired, with a comfortable lifestyle and most importantly, use term deposit accounts as their primary source of income

Hence, our recommendation is to not only continue to focus on the 20 - 40 year olds, but also take special care to target the 59 years and over customers as well.

y	no	yes	acceptance_rate_total	acceptance_rate_group
age_break				
(0, 20]	83	57	0.138390	40.714286
(20, 40]	20964	2664	6.467903	11.274759
(40, 59]	14780	1447	3.513159	8.917237
(59, 200]	721	472	1.145965	39.564124



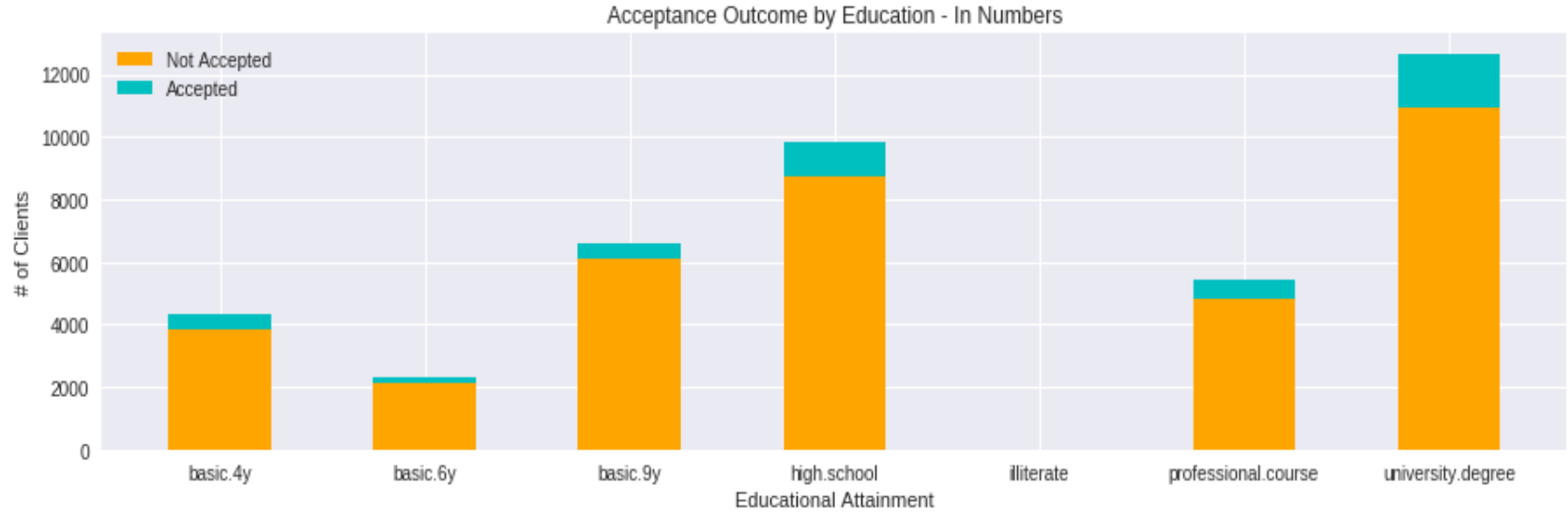
Education Level

The data had customers with varying levels of education, including basic 4 year school all the way to university degree holders. Incidentally, we also add customers who self-reported as being illiterate, though it was a very small subset. Again, there was a serious attempt, leading to a possible self- bias, in focusing on customers who had an university degree or higher. This group had over 13.75% success overall.

But upon further review, it became evident that we would have had a higher success rate if we had focused our efforts on customers with either a professional course, high school or basic 4-year grade school education. All three of these groups had over 11% success rate individually. By taking into consideration that educational proficiency does not always translate into financial success, we can broaden our approach thus leading to additional target audiences.

Hence, our recommendation is to not only focus on the customers that have an university degree but also spend additional time in targeting customers with either a professional course, high school or basic 4 year grade school education.

Education	Declined	Accepted	Acceptance Rate - Total	Acceptance Rate - Within Education Group
basic.4y	3859	474	1.15	10.94
basic.6y	2104	188	0.46	8.20
basic.9y	6093	504	1.22	7.64
high.school	8723	1109	2.69	11.28
illiterate	14	4	0.01	22.22
professional.course	4835	620	1.51	11.37
university.degree	10920	1741	4.23	13.75



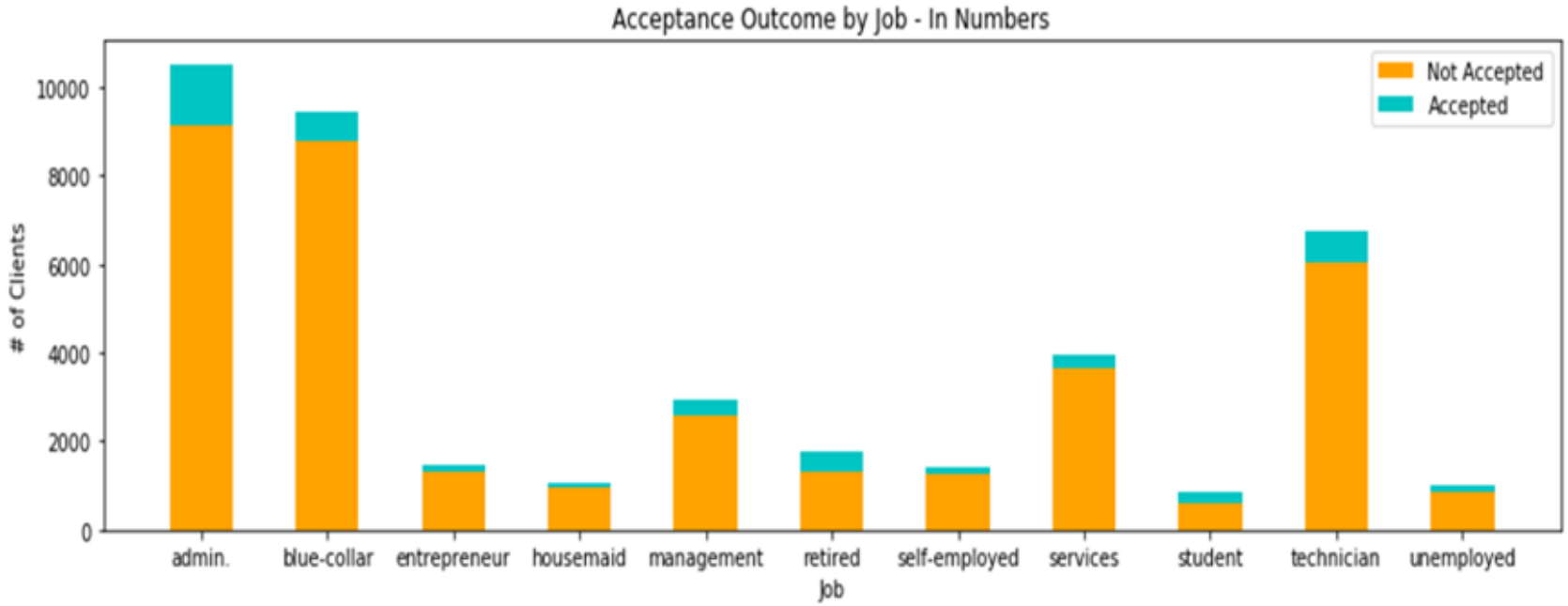
Job Type

By reviewing the data, we noticed our customers occupying a variety of job profiles including administrative jobs, blue collar jobs and some being unemployed as well. It's obvious that customers with an administrative job were targeted extensively in this campaign. And it worked really well, as we ended up with close to 13% in successful conversion for this job profile.

But by looking at the data a bit more closely, we were able to recognize that by focusing on customers who were retired (25%) and unemployed (14%), we might have had an even better success rate. Our understanding here is that just like with the age criteria, customers who are retired might have a higher than average savings, that they need to grow using term deposits. Hence, working with them is a viable idea. Regarding the unemployed, our conclusion is that these were people who were recently unemployed and have had some savings that they would like to further protect and grow during the recession. As a result, they were ideal candidates to target during the campaign as well.

Hence, our recommendation, while looking at customer job profiles, is to not only focus on the administrative job holders, but also focus on the retired, unemployed and students for additional conversions.

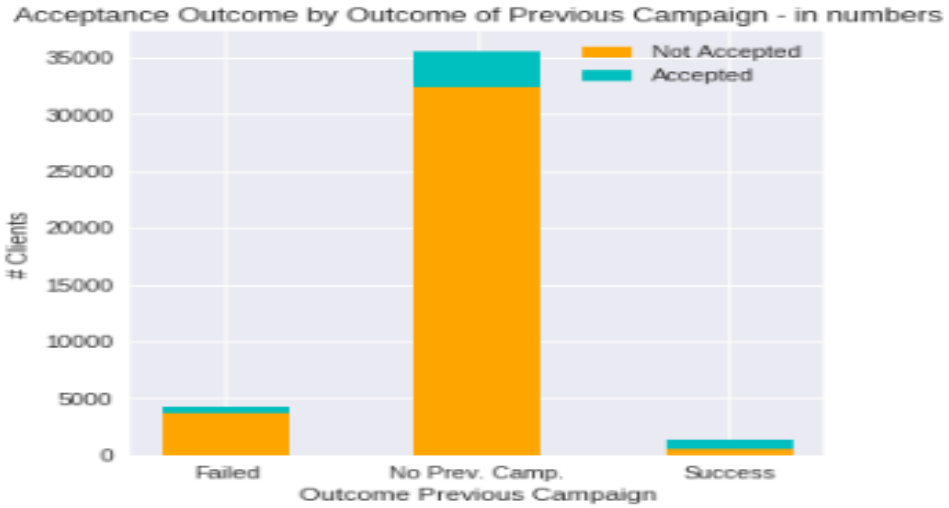
y	no	yes	acceptance_rate_total	acceptance_rate_group
job				
admin.	9155	1367	3.318928	12.991827
blue-collar	8803	652	1.582985	6.895822
entrepreneur	1332	124	0.301059	8.516484
housemaid	954	106	0.257357	10.000000
management	2596	328	0.796348	11.217510
retired	1307	442	1.073128	25.271584
self-employed	1272	149	0.361756	10.485574
services	3646	323	0.784209	8.138070
student	600	275	0.667670	31.428571
technician	6013	730	1.772361	10.826042
unemployed	870	144	0.349616	14.201183


Acceptance Outcome by Job - In Numbers
Not Accepted
Accepted
# of Clients
10000
8000
6000
4000
2000
0
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
Job

Previously Contacted Customers

This one was interesting mainly because, on first review, the assumption was that people who had signed up for such deposit accounts previously would not want to sign up for another account. By exploring the data further, we were able to determine that focusing on customers with existing term deposit accounts can lead to additional sales. Also, we are confident that the time and effort taken to close this sale would be considerably less when compared to customers who had never signed up for a term deposit account before.

y poutcome	no	yes	acceptance_rate_total	acceptance_rate_group
failure	3647	605	1.468874	14.228598
nonexistent	32422	3141	7.626008	8.832213
success	479	894	2.170535	65.112891



When to Call?

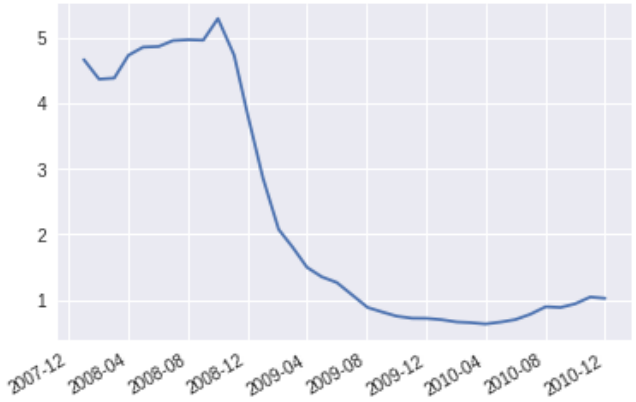
The dataset was also expanded by the authors to include national level economic indicators in an attempt to understand the dataset further. Four major indicators were taken into consideration including:

Euribor 3-Month Rate

Euribor rate is the interest rate that banks charge other banks when they borrow money. This rate is set amongst several major banks in the European Union. The essential idea here is, when the Euribor rate is low, banks will reduce the term deposit interest rates as well. And when the Euribor rates are high, term deposit interest rates offered to customers are increased accordingly. This is mainly to keep the liquidity in the bank at a condition that meets their fiduciary and legal obligations during any given period.

During the economic crisis, the Euribor rate was considerably lower than average. For almost the entire last quarter of 2008 cycle, the rate was close to zero percent. As a result, the banks had their term rates essentially close to zero. The obvious result of this should have been that customers that were contacted during this period should have refused to sign up for a term deposit. Though a majority of our customers followed this trend, almost 11% of our existing customers actually signed up , even when such a low rate was offered. Additionally, customers with existing term deposit accounts signed up again during this campaign cycle.

Hence, our recommendation is to not only follow conventional wisdom during campaigns, but also consider customer's willingness to save money in term deposits, even when the term deposit rates are not ideal.

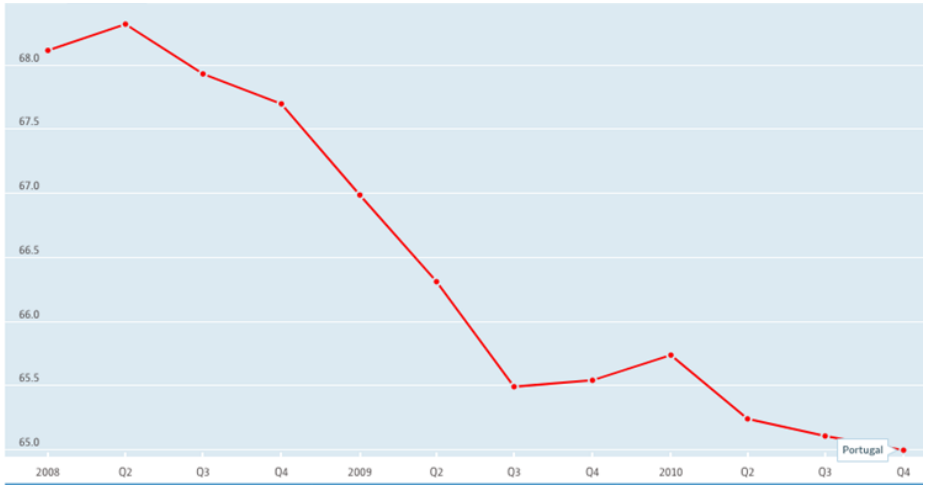


Employment Rate

Another economic indicator that was given high importance was the nation's employment rate. Again, very similar to the Euribor rate, the employment rate kept falling through 2008 and eventually settled down to an all-time low of 65% of working population who were employed. This should have been a major deterrent for people to invest in term deposits, according to conventional wisdom.

But again, our customers defied that wisdom by signing up for term deposits.

Hence, our recommendation is to not only focus on the employed customer base, but also market to the unemployed group, especially during economic downturns.

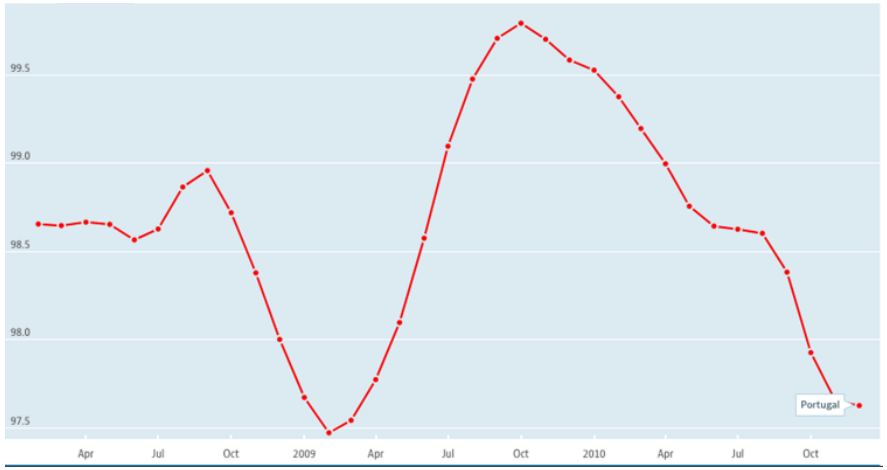


OECD (2018), Employment rate (indicator). doi: 10.1787/1de68a9b-en (Accessed on 07 August 2018)

Consumer Confidence Index

This was the one indicator that was the most surprising for our team. Knowing the economic conditions during the crisis in 2008, we were prepared to see the consumer confidence index take a nosedive. Surprisingly, the overall population held very close to the average throughout the time period that this campaign was run. Essentially, Portugal, including our customers, still had confidence in the economy. As a result, a majority of our customers did not sign up for the term deposit accounts.

Hence, our recommendation here is to monitor this index closely and launch campaigns when the index is higher than average in order to maximize our revenue.



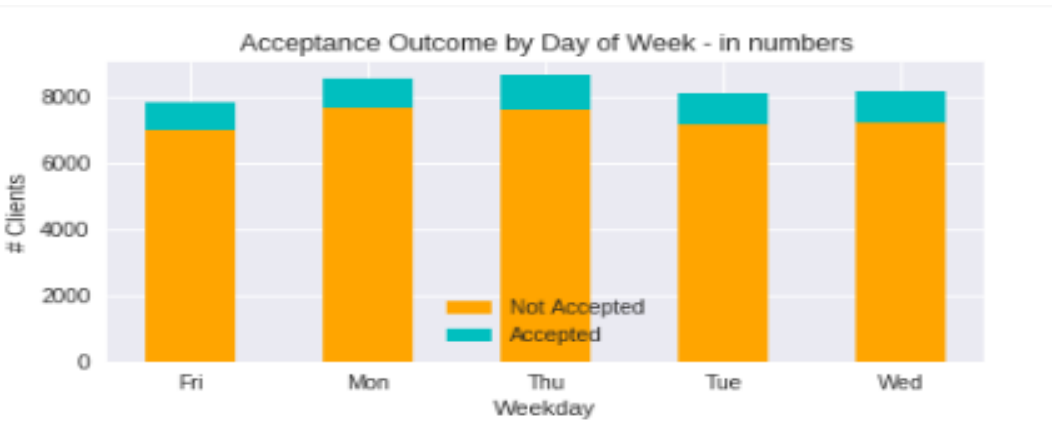
OECD (2018), Consumer confidence index (CCI) (indicator). doi: 10.1787/46434d78-en (Accessed on 07 August 2018)

Day of the Week

Based on preliminary review of the days of the week in relation to acceptance rate, there was no significant difference in response based on the days of the week.

Hence, targeting customers on all days of the week should be continued for all our outbound campaigns.

y	no	yes	acceptance_rate_total	acceptance
day_of_week				
fri	6981	846	2.053996	
mon	7667	847	2.056424	
thu	7578	1045	2.537147	
tue	7137	953	2.313781	
wed	7185	949	2.304069	



4. Future Considerations

Two key future considerations should be taken forward after the initial implementation and pilot: 1) operationalise data mining process; 2) expansion of data collection and feature engineering and the removal of redundant variables. The exercise of conducting a health check and review of existing and additional variables for data collection should be done at a regular interval, for example, once a year.

Substantial work will be required to effectively operationalise the end-to-end data mining process for future direct marketing campaigns. Initially the approach should be implemented as part of a pilot, as the existing code has been established for cleaning, feature engineering and running the model. This code can be used directly with a new sample of potential term deposit customers. This will require technical support from data analytics, data science and data engineering teams.

As a long-term practical solution the end-to-end data mining process can be automated to avoid the need to run a data project each time a new campaign is run or a new sample of potential customers is available.

A good place to start with this challenge is the CRISP-DM, see figure below, which is an industry data mining model.

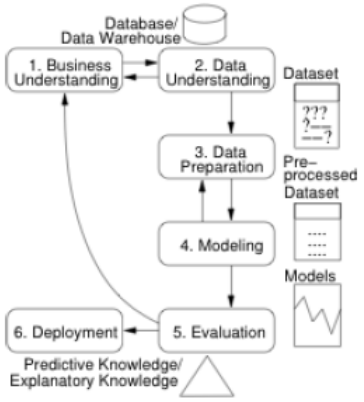


Figure 1 The CRISP-DM process model (adapted from Chapman et al., 2000)

The ambition behind this approach is to automate the process as much as possible thereby requiring minimal intervention by technical staff, as well as creating a continuous feedback loop. The feedback loop will require a data pipeline to feed the model evaluation results back into the model. This allows the expansion of the training data sample to retrain, update and retest the model. Whereas this is an ambitious undertaking, the reduction in costs and ROI investment should be high.

Even with the operationalisation of the data mining process, work will still be required to check the health of the model ensuring it continues to meet business requirements, including ROI targets. In addition, there may be opportunities to further improve the data mining process, by collecting additional customer data or economic indicator data. By collecting additional variables this should naturally open up opportunities for further feature engineering to enrich the data mining process.

One should also note that in addition to collecting other variables, it may also be prudent to consider dropping others that are not adding value to the data mining process

Appendix 1 – Technical Overview

The purpose of this appendix is to outline the technical details of this project from a data science perspective. It intended for those stakeholders who are interested in understanding more about the data and the predictive statistics underpinning the analysis and recommendations sections of this report. This appendix is broken down into subsections with each subsection representing a technical stage of the project.

Data Cleansing & Wrangling

Cleaning

During initial exploration of the data it was discovered that null values were coded as 'unknown'. To make data cleaning as efficient as possible 'unknown' values were replaced with 'NaN'. After cleaning up the null values there was much clearer picture of the completeness of the dataset. The following is a summary of the percentage of missing values by variable in the initial dataset:

Variable_Name	Null_Actual	Null_Percent
default	8597	20.87
education	1731	4.2
housing	990	2.4
loan	990	2.4
job	330	0.8
marital	80	0.19
age	0	0
campaign	0	0
cons.conf.idx	0	0
cons.price.idx	0	0

contact	0	0
day_of_week	0	0
duration	0	0
emp.var.rate	0	0
euribor3m	0	0
month	0	0
nr.employed	0	0
pdays	0	0
poutcome	0	0
previous	0	0
y	0	0

The following two variables were dropped from the dataset:

- 'Default' - more than 20% of the values were null. This in itself meant that imputation exercise may require the use of a model to provide a reasonable prediction of an appropriate value. In addition, since there are only three customers had actually been identified as defaulting on a loan this variable was probably used a pre-filter to get rid of any individuals who defaulted as a term deposit scheme may not be an appropriate financial scheme for these individuals. This variable should be used a pre-filter on the sample, but should then excluded from the input dataset to as it does not add any predictive value.

- 'Duration' - duration reflects the length of the call with the potential customer. As this cannot be predicted when trying to select potential customers that are most likely to sign-up to a term deposit this is not a useful variable to use in the model.

Imputing

As per section 4.1.1 the following four variables were deemed complete enough to retain in the dataset, but still required missing values to be imputed, i.e. replacing the null values with one of the valid unique values available for that variable:

- Job
- Marital
- Loan
- Housing

As the nature of each variable is different it was necessary to select an imputation method on an individual variable basis. A detailed explanation of the imputation method for each variable is given below:

Job:

An Age and Job lookup using the mode for a given age range was built to impute missing Job values using the following inputs. This approach was used due to the relationship between age and employment.

Age Range	Jobs
0 - 20	Student
20 – 40	Admin
40 – 59	Blue-collar
59 – 200	retired

Marital Status (marital):

Whilst there may have been an opportunity to use a stratified approach to filling the null Marital values in the end the mode was used due to its simplicity and efficiency.

Personal Loan (loan), Mortgage (housing):

A simple forward fill was used for the personal Loan and housing field.

One-Hot Encoding

There were a number of categorical variables in the dataset. To use categorical variables in a model the variables must be transformed into 'dummy variables'[2] as the model treat the number values assigned to the categorical labels as a measure of distance. However, that is not the case for categorical values.

Some implementations of Random Forest, for example, in R, handle the transformation of categorical variables within the Random Forest algorithm itself. However, during this project the Scikit-Learn implementation of Random Forest was used requiring the categorical variables to be explicitly transformed into dummy variables.

One of the most commonly approaches to transforming categorical variables to dummy variables is one-hot encoding. One-hot coding creates a new variable for each unique value within the original categorical variable, for example transforming a 'Days of the Week' variable into seven separate variables one for each day of the week (e.g. Monday, Tuesday). One natural consequence of the one-hot encoding is increases the number of variables in the transformed dataset.

The following 8 categorical variables were transformed using one-hot encoding resulting in a total of 41 new variables:

- job
- marital
- education
- housing
- loan
- contact
- month
- day_of_week
- poutcome

[2] Linear Regression by Geoffrey Hinton - https://www.cs.toronto.edu/~guerzhoy/321/lec/W04/onehot.pdf [Last accessed 07/08/2018]

Model Selection

The problem statement is essentially a binary classification problem. For this type of data science problem there are a plethora of models to choose from. The key decision making factors for selecting a model for this particular business problem were:

- Applicability - ease and speed of training and testing the model
- Adoption - ability to share and explain the model diagnostics to a lay audience, thereby, ensuring the results are accepted by the wider business
- Accuracy - an accuracy score good enough to provide the business with a high return on investment (ROI)

With the above criteria in mind, a decision was made to use Random Forest. The Random Forest algorithm provides the following key benefits[3]:

- Delivers high level of accuracy
- Provides the importance of variables in the model
- Easily combined with diagnostic evaluation tools such as confusion matrices and accuracy classification scoring
- Accessible implementation of the Random Forest algorithm using the open source Scikit-learn library for Python

Whilst there are some drawbacks, for example, the tendency for Random Forest models to be susceptible to overfitting, the benefits of implementing a Random Forest model far outweigh the benefits.

Model Training

To train and test the model the original sample size of 41,188 was split into two with two thirds as the training set (67%) and one third as the test set (33%). The train test split resulted in the following dataset shapes:

	Training Set	Test Set
X	(27595, 56)	(13593, 56)
Y	(27595,)	(13593,)

After running the model the top 20 most important variables in the model were:

[3] Random Forests by Leo Breiman and Adele Cutler - https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Last accessed 07/08/2018]

Rank	Variable	Importance Weight
1	age	0.1634
2	euribor3m	0.109052
3	campaign	0.078573
4	nr.employed	0.072198
5	pdays	0.039841
6	cons.price.idx	0.024378
7	housing__no	0.021564
8	cons.conf.idx	0.021334
9	housing__yes	0.019907
10	emp.var.rate	0.019646
11	poutcome__success	0.018907
12	job__admin.	0.018122
13	marital__married	0.017458
14	education__university.degree	0.016665
15	default__no	0.015953
16	day_of_week__tue	0.015874
17	day_of_week__mon	0.015684
18	education__high.school	0.015679
19	marital__single	0.015625
20	previous	0.014937

The two most important variables in predicting if a potential customer will sign-up to a term deposit is customer Age and Euribor3m (quarterly Euribor rate).

The model applied to the test dataset gave an accuracy score of 98%, meaning the model correctly 98% of the time if potential customers would or would not sign-up to a term deposit. This is a high accuracy score which suggests overfitting.

The following confusion matrix shows where the model using the test data correctly and incorrectly predicted a potential customer would sign-up to a term deposit.

	Predicted Success (actuals)	Predicted Success (%)	Predicted Failure (actuals)	Predicted Failure (%)
True Success	24434	98.22	58	2.13
True Failure	444	1.78	2659	97.87

For the test dataset the model predicted slightly more false negatives (2.13%), in other words predicted a customer wouldn't sign-up when in fact they did sign-up, than false negatives (1.78%).

Results

The model applied to the test data gave an accuracy score of 89%, meaning the model correctly 89% of the time if potential customers would or would not sign-up to a term deposit. The model achieved 9% less accuracy with the training dataset than it did with the training dataset. This confirms that the model was overfitting to the training dataset, however, 89% is an acceptable accuracy score. Further testing and validation should be carried on a new sample of customers to confirm that the model still achieves an acceptable accuracy rate.

The following confusion matrix shows where the model correctly and incorrectly predicted a potential customer would sign-up to a term deposit when applied to the training dataset:

	Predicted Success (actuals)	Predicted Success (%)	Predicted Failure (actuals)	Predicted Failure (%)
True Success	11703	91.14	353	46.88
True Failure	1137	8.85	400	53.12

The results of the confusion matrix for the test dataset appear in stark contrast to the training results. For the training dataset the model predicted substantially more false negatives (46.88%), in other words predicted a customer wouldn't sign-up when in fact they did sign-up, compared than false positives (8.85%).

Appendix 2 – Data Dictionary

Below is a list of variables with category, names, descriptions, data types:

Feature Category	Feature Business Name	Feature Business Description	Feature Information Provided	Feature name in notebook code
Bank customer data	Age	customer age	Numeric	age
Bank customer data	Job	customer job	Categorical: admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown	job
Bank customer data	Marital Status	customer marital status	Categorical: divorced, married, single, unknown; Note: 'divorced' means divorced or widowed	marital
Bank customer data	Level of Education	customer Level of Education	Categorical: basic.4y, basic.6y, basic.9y, high.school, illiterate, professional. course, university. degree, unknown	education
Bank customer data	customer in Default	Indicator of whether customer has been in default	Categorical: no, yes, unknown	default

Bank customer data	Mortgage	Indicator of whether customer has a mortgage (housing loan)	Categorical: no, yes, unknown	housing
Bank customer data	Personal Loan	Indicator of whether customer has a personal loan	Categorical: no, yes, unknown	loan
Last contact current campaign	Contact Type	Type of last customer communication / contact	Categorical: cellular, telephone	contact
Last contact current campaign	Month Last Contact	Month of customer last communication / contact	Categorical: jan, …dec	month
Last contact current campaign	Day of Week Last Contact	Day of the week of last customer communication / contact	Categorical: mon, tue, wed, thu, fri	day_of_week
Last contact current campaign	Duration Last Contact	Duration of customer last communication / contact	Numeric, in seconds	duration
Other	Campaign customer Calls	Number of contacts performed during this campaign for this customer	Numeric, includes the last contact	campaign
Other	Passed Days	Number of days that have passed by after the customer was last contacted from a previous campaign	Numeric Note - 999 means customer was not previously contacted	pdays
Other	Previous Calls	Number of contacts performed before this campaign and for this customer	Numeric	previous
Other	Previous Outcome	Outcome of the previous marketing campaign	Categorical: failure, nonexistent, success	poutcome
Social Economic	Employment Variation Rate	Employment variation rate - quarterly indicator	Numeric	emp.var.rate
Social Economic	Consumer Price Index	Consumer price index - monthly indicator	Numeric	cons.price.idx
Social Economic	Consumer Confidence Index	Consumer confidence index - monthly indicator	Numeric	cons.conf.idx

Social Economic	Euribor 3-Month Rate	EURo InterBank Offered Rate. Euribor 3-month rate is a daily indicator for a 3-month term	Numeric	euribor3m
Social Economic	Number of Employees	Employment Rate. The total number of people employed - quarterly indicator	Numeric	nr.employed
Customer's response	Outcome	Indicator whether the customer has subscribed for a term deposit with the bank via this call	Binary: yes, no	y

Below is a list of unique values by variable:

Feature name in notebook code	Unique values occurring in the dataset
age	56 57 37 40 45 59 41 24 25 29 35 54 46 50 39 30 55 49 34 52 58 32 38 44 42 60 53 47 51 48 33 31 43 36 28 27 26 22 23 20 21 61 19 18 70 66 76 67 73 88 95 77 68 75 63 80 62 65 72 82 64 71 69 78 85 79 83 81 74 17 87 91 86 98 94 84 92 89
job	Housemaid, services, admin., blue-collar, technician, retired, management, unemployed, self-employed, unknown, entrepreneur, student
marital	Married, single, divorced, unknown
education	basic.4y, high.school, basic.6y, basic.9y, professional.course, unknown, university.degree, illiterate
month	mar, apr, may, jun, jul, aug, sep, oct, nov, dec
day_of_week	mon,tue,wed,thu,fri
contact	telephone, cellular
pdays	999 6 4 3 5 1 0 10 7 8 9 11 2 12 13 14 15 16 21 17 18 22 25 26 19 27 20
previous	0 1 2 3 4 5 6 7
housing	no,yes,unknown
campaign	1 2 3 4 5 6 7 8 9 10 11 12 13 19 18 23 14 22 25 16 17 15 20 56 39 35 42 28 26 27 32 21 24 29 31 30 41 37 40 33 34 43
duration	261 149 226 ... 1246 1556 1868 (a wide range of call durations)
poutcome	Nonexistent, failure, success

Appendix 3 – Data Summaries and Visualizations

As mentioned there are a few other variables that were included in the dataset that added its own imprint on the overall success of the campaign. We have listed them all in the form of graphs. This is to show that there are several factors, apart from the most obvious ones, which should be explored, in order to improve our ability to predict success from a marketing campaign.

Data Summaries

Few things we can say about the descriptive statistics about this dataset:

AGE of customers targeted for campaign:
The average age was 40 years old, with a variation of +/- 10 years, being the youngest 17 years old and the oldest 98 years old.

DURATION of call the last time customer was contacted (in seconds):
The average was 4.3 minutes with a variation of +/- 4.4 minutes, being the longest 1hour and 36 minutes!

CAMPAIGN (total # contacts for a customer during this campaign):
The average was 2.56 times contacted, with variation of +/- 2.77 times, being the maximum # of contacts 56 times!!

PDAYS (# days passed after customer was last contacted from previous campaigns - 999: this is the first contact)
The average is 962 days (32 months / 2.67 years), with variation of +/- 187 days (6.23 months)

PREVIOUS (Number of contacts for a customer prior to this campaign)
There were a maximum number of 7 contacts.

Correlation Matrix

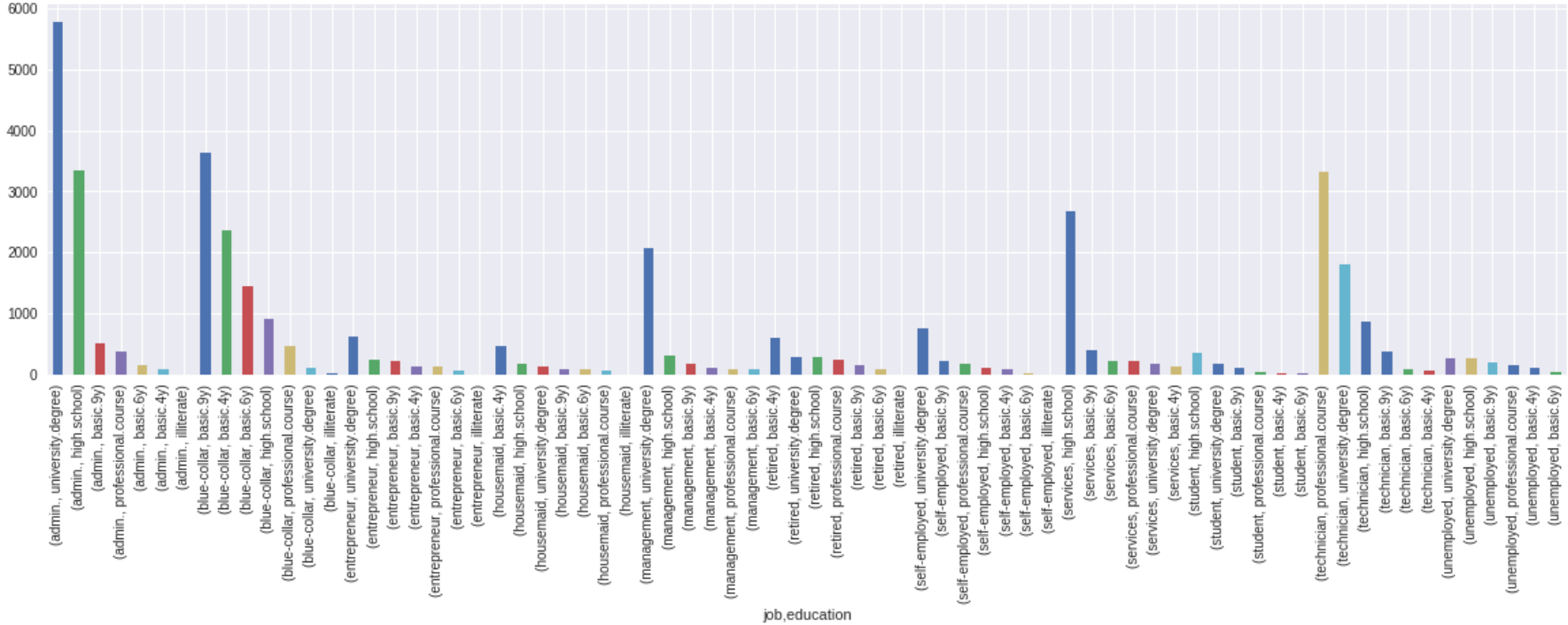
Based on the following correlation matrix we can see that, as expected, the economic index were highly correlated:

- Employment variation rate and euribor were highly correlated (0.9722)
- Employment variation rate and number of employees were highly correlated (0.9069)
- Euribor and number of employees were highly correlated (0.9451)

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1	-0.000865705	0.00459358	-0.0369009	0.0243647	-0.000370685	0.000856715	0.129372	0.0107674	-0.0177251
duration	-0.000865705	1	-0.0716992	-0.0448343	0.0206404	-0.0279679	0.00531227	-0.00817287	-0.0328967	-0.0447032
campaign	0.00459358	-0.0716992	1	0.0525466	-0.0791415	0.150754	0.127836	-0.0137331	0.135133	0.144095
pdays	-0.0369009	-0.0448343	0.0525466	1	-0.582679	0.270108	0.0846549	-0.0936773	0.292634	0.365305
previous	0.0243647	0.0206404	-0.0791415	-0.582679	1	-0.420489	-0.20313	-0.0509364	-0.454494	-0.501333
emp.var.rate	-0.000370685	-0.0279679	0.150754	0.270108	-0.420489	1	0.775334	0.196041	0.972245	0.90697
cons.price.idx	0.000856715	0.00531227	0.127836	0.0846549	-0.20313	0.775334	1	0.0589862	0.68823	0.522034
cons.conf.idx	0.129372	-0.00817287	-0.0137331	-0.0936773	-0.0509364	0.196041	0.0589862	1	0.277686	0.100513
euribor3m	0.0107674	-0.0328967	0.135133	0.292634	-0.454494	0.972245	0.68823	0.277686	1	0.945154
nr.employed	-0.0177251	-0.0447032	0.144095	0.365305	-0.501333	0.90697	0.522034	0.100513	0.945154	1

Descriptive Statistics

The following is the visualization of number of customers for each combination of education and job.



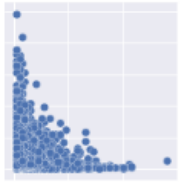
The list below outlines the most frequent combinations of education and job types (descending order); and list of most frequent combinations of job type and level of education, respectively.

Rank	Education	Job
1	University	Administrative
2	Basic 9 year	Blue-collar
3	Professional course	Technician
4	High school	Services
5	Basic 4-year	Blue-collar
6	University	Management

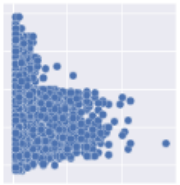
job	education
admin.	university.degree
blue-collar	basic.9y
entrepreneur	university.degree
housemaid	basic.4y
management	university.degree
retired	basic.4y
self-employed	university.degree
services	high.school
student	high.school
technician	professional.course
unemployed	university.degree

The following is a subset of the pair plot visualization, where the following observations are noted:

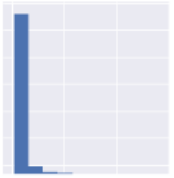
As the number of contacts to a customer during a campaign grow, the duration of the calls decreased (inverse correlated)
x-axis: number of contacts to a customer during a campaign (campaign)
y-axis: duration of the call (duration)



The number of contacts to a customer during a campaign were above 20 for customers in the age range 20 – 60
x-axis: number of contacts to a customer during a campaign (campaign)
y-axis: customer age (age)



Majority of the campaigns only contacted a customer up to about 5 times
x-axis: number of contacts to a customer during a campaign (campaign)
y-axis: number of customers



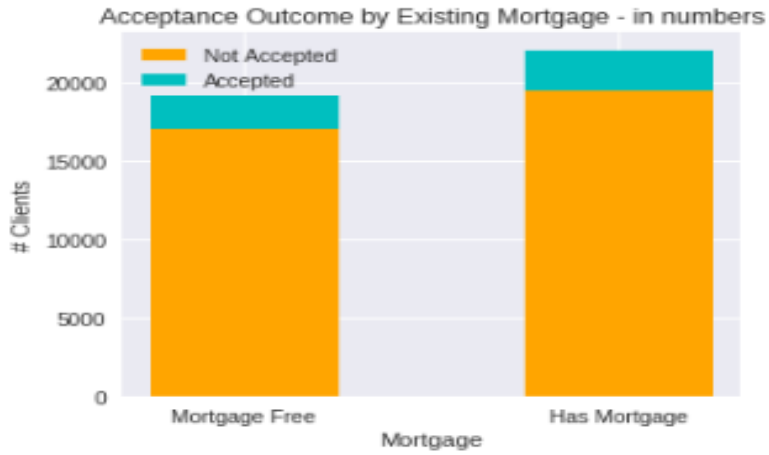
The following are visualizations of the main variables in relation to the outcome of the campaign.

Campaigns targeted mostly individuals with no personal loan. However the acceptance was low, regardless whether the target had personal loan. Hence, pre-existing personal loan does not impact the acceptance rate.

loan	y no	yes	acceptance_rate_total	acceptance_rate_group
no	30877	3942	9.570749	11.321405
yes	5671	698	1.694668	10.959334

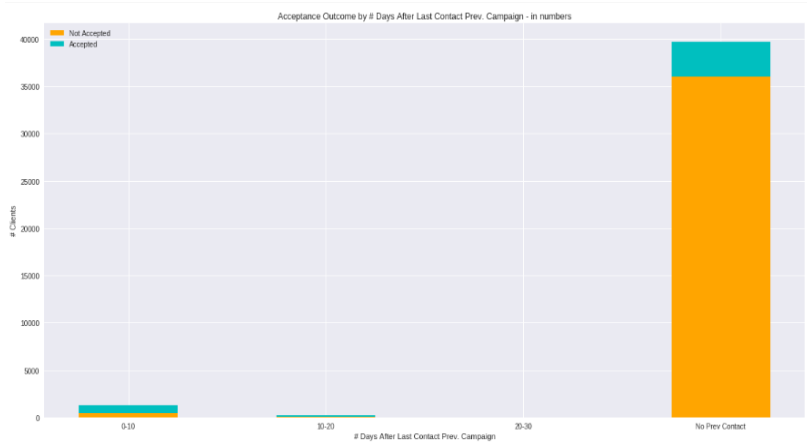
Campaigns targeted slightly more customers with mortgage, However the acceptance was the same in both groups, regardless. Hence, pre-existing mortgage does not impact the acceptance rate.

housing	y no	yes	acceptance_rate_total	acceptance_rate_group
no	17031	2085	5.062154	10.907094
yes	19517	2555	6.203263	11.575752



Campaign largely targeted customers who have not previously been contacted by other campaigns

pdays_break	y no	yes	acceptance_rate_total	acceptance_rate_group
(0, 10]	455	841	2.042601	64.891975
(10, 20]	87	109	0.264737	55.612245
(20, 30]	1	7	0.017001	87.500000
(30, 40]	36000	3673	8.920895	9.258186



Number of overall contacts to the customer by campaign is clearly under 10 calls, which intuitively makes sense.

campaign_break	y no	yes	acceptance_rate_total	acceptance_rate_group
(0, 10]	35706	4613	11.199864	11.441256
(10, 20]	686	26	0.063125	3.651685
(20, 30]	123	1	0.002428	0.806452
(30, 40]	27	0	0.000000	0.000000
(40, 50]	5	0	0.000000	0.000000
(50, 60]	1	0	0.000000	0.000000