Predictive Analysis of Fixed Term Deposit Subscriptions in Portuguese Banking

Exploring Predictive Models for Client Subscription Behavior

Data Analytics Project

Author:	Michèle Fille, Rahel Häusler, Jervin Ureath
Lecturer:	Prof. Andreas Reber
City, Date:	Basel, 24th of May 2024

Table of Contents

1 Introduction

Direct marketing methods, notably telemarketing, play a crucial role in customer engagement for banks and insurance companies. Thanks to advancements in technology and extensive data collection, marketing campaigns have evolved into highly targeted initiatives. This shift enables a systematic approach to contacting high-potential prospects for specific marketing initiatives. Banks leverage these targeted marketing strategies to promote their financial products more effectively, including fixed term deposits, as part of the broader revenue generation strategy.

A fixed term deposit is a type of savings account where you deposit a sum of money for a specific period, usually ranging from a few months to several years, at a fixed interest rate. During this period, you typically cannot withdraw the funds without incurring a penalty. Fixed term deposits are considered low-risk investments as they provide a guaranteed return on your investment, making them popular for people looking to save money securely over a set period.

In this case study, a Portuguese bank conducted a direct marketing campaign over the phone to sell customers fixed term deposits. The success of the marketing campaign is measured whether a client would choose to accept ('yes') or reject ('no') the proposed fixed term deposit. The data presented was collected over the period of May 2008 to November 2010, a period significantly impacted by the Global Financial Crisis respectively Euro Crisis.

The aim of this data analysis project is to create three predictive models that determine whether or not a customer will subscribe to a fixed term deposit, leveraging the data and insights obtained from the "marketing.csv" dataset. This project applies the Cross Industry Standard Process for Data Mining (CRISP-DM) model and its phases. Figure 1 shows the different phases of the CRISP-DM Model. These phases are iterative, meaning it is often necessary to revisit previous phase(s) or perform additional iterations across all phases to further refine and enhance the process. This paper summarizes the CRISP-DM phases conducted to analyze and define the most promising model for successfully assessing whether a client will subscribe to a fixed term deposit. Since this project focuses on analysis and model development, no application is being developed. Therefore, the final phase of 'Deployment' is not executed and thus not included in this paper (*What Is CRISP DM? - Data Science Process Alliance*, n.d.).

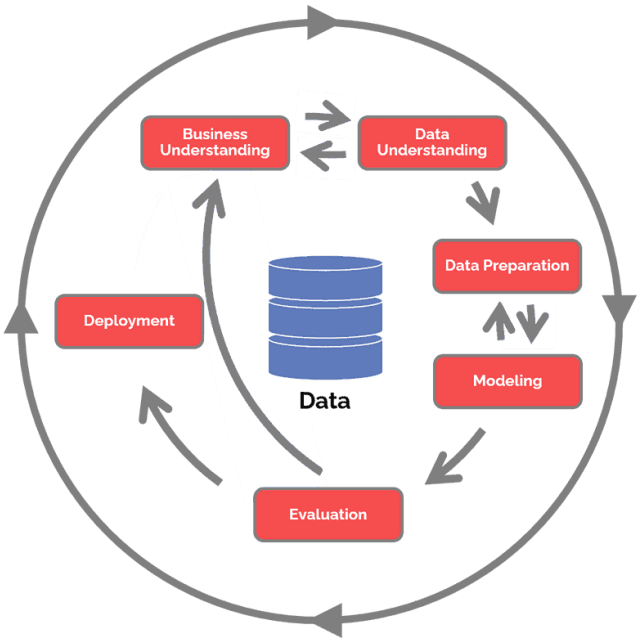

Figure 1: CRISP-DM Model with its 6 Phases (*What Is CRISP DM? - Data Science Process Alliance*, n.d.)

2 Business Understanding

The Business Understanding phase is pivotal for the success of the data analytics project as it lays the foundation of the project's objectives and requirements. Furthermore, it includes a thorough understanding of the business landscape. These aspects provide the necessary guidance for subsequent steps and therefore strongly influence the result of the project.

According to the CRISP-DM model, this phase comprises four tasks: 1) Determine business objectives, 2) Assess situation, 3) Determine data mining goals, and 4) Produce project plan. We have tailored these tasks to align with the specific requirements of our project.

2.1 Background Information

The dataset was collected between May 2008 and November 2010, a period deeply impacted by the financial and Euro crises (Neubäumer, 2011).

Even prior to the collapse of Lehman Brothers in August 2008, which ultimately lead to the financial crisis, banks had begun reducing exposures to crisis-affected countries like Portugal (Lindner, n.d.). This reduction could trigger deep recessions in countries with current account deficits and high net foreign debt, as they are abruptly cut off from external financing and forced to reduce essential expenses. Moreover, such countries typically have short-term liabilities but hold long-term assets. In crises, if investors refrain from extending loans, liquidity issues arise, forcing the borrower to liquidate assets (Lindner, n.d.).

The impact of the global financial crisis was highly evident in Portugal. As the country's credit-worthiness decreased, banks faced rising financing costs for lending. By November 2010, the Portuguese national bank alerted to liquidity problems, urging banks to reduce loans, promote savings, and seek alternative financing (*Portugals Zentralbank Warnt Vor Risiken Für Bank-ensektor | Tagesschau.De*, n.d.). This press release underscored the severity of the financial crisis in Portugal.

Amidst the financing challenges, fixed term deposits emerged as a viable funding solution for banks. Fixed term deposits offer banks a stable and predictable source of funding. By locking in funds for specific periods, banks can rely on this stability for their business operations. Furthermore, offering fixed term deposits enables banks to efficiently manage their liquidity. With foresight into the availability of funds, banks can plan their activities, mitigating liquidity risks during turbulent times (*Attraktive Festgeld Zinsen Sichern | UBS Schweiz*, n.d.).

2.2 Business Objective, Risks and Benefits

Businesses in general aspire to operate on high level of efficiency and effectiveness to maximize returns. Especially during an economic crisis, it is essential to fully leverage existing tools and workforce capabilities. Consequently, the Portugues bank decided to enhance the efficacy of its telemarketing campaigns through the utilization of predictive models for future marketing campaigns.

By incorporating predictive models into its telemarketing strategies, the Portuguese bank seeks to optimize the allocation of resources and increase the success of its marketing efforts. However, the use of customer data for such purposes raises concerns regarding data privacy and protection. It is imperative that customers explicitly consent to the use of their data, which should ideally be outlined in the General Terms and Conditions. Failure to address these concerns could not only compromise customer trust but also damage the bank's reputation, especially if the model is found to have biases and employees rely solely on its results. Therefore, ensuring the privacy and integrity of the model is crucial to maintaining the bank's credibility and fostering positive customer relationships.

By harnessing a robust prediction model, the bank aims to streamline resource allocation and to improve the success rate of their marketing campaign. Therefore, the business objective is a heightened conversion rate for fixed term deposits thanks to a more targeted approach. Another benefit provided by the marketing efforts is the strengthening of the client relationship.

In context of the financial crisis, the main benefit provided through the improved marketing campaign will be the securing of fundings for the bank. In the event of a false positive prediction, the negative impact lies in the personal cost for the operator, as the operator could have been allocated to a successful marketing prospect instead.

2.3 Success Criteria and Data Mining Goals

Success in this data analytics project is defined by the predictive model’s ability to accurately classify whether a client will subscribe to a fixed term deposit. In our project, the key performance metrics used for the evaluation of the prediction model is the classification accuracy (CA). We have defined a specific target for this metric at 0.800, signifying our aim for a high level of accuracy in classification. Moreover, we will examine the ROC curve of each designed model to assess its discriminative ability across varying thresholds.

Moreover, scalability and efficiency are crucial factors to consider during the model development process, particularly given the anticipated size and complexity of future datasets.

The project aims to develop a model capable of seamlessly handling massive volumes of data, facilitating rapid predictions and actionable insights.

2.4 Project Phases and Tools

The project will be executed using the CRISP-DM approach. The project will be split into the following phases: Business Understanding, Data Understanding, Data Preparation, Modelling, and Evaluation (*What Is CRISP DM? - Data Science Process Alliance*, n.d.). The final Deployment is not target to this project and therefore not considered.

Business Understanding

The initial phase includes understanding the project objectives, requirements, and the business success criteria. By gaining a comprehensive understanding of the project and its business context, through thorough research, the efforts can be directed towards achieving the project's objective.

Data Understanding

In the following phase, the statistical programming language R, R Studio and Orange, an effective data mining and visualization tool, will be used to perform a comprehensive analysis of the dataset. In this step, insight of the nature, quality, and potential relevance of the data will be gained.

Data Preparation

The subsequent phase focuses on preparing the data for analysis by addressing any issues identified during the previous phase. The objective of this stage is the creation of a clean, structured and properly formatted dataset, suitable for modeling. For this step, Orange has been used as main tool, following attempts to perform this phase using R Studio and R.

Modelling

In this phase, suitable models for building a predictive model will be developed. This phase is inherently iterative, as adjustments to the data preparation may be necessary, along with fine-tuning the parameters within the model. The aim of this phase is to develop three models that effectively address the project's objectives, assess the models and compare the model's performance. Similarly to the Data Preparation phase, the capabilities of Orange as well as the add-on "Explain" has been used. Furthermore, classification accuracy (CA) will be used as main evaluation metrics, while also considering ROC in the model's evaluation.

Evaluation

After the development of the models and the assessment according to their performance, the effectiveness in meeting the project objectives is evaluated. Additionally, recommendations of possible enhancements for future improvements are made. No specific tools were used, this step was performed based on the outcome of the previous phase.

3 Data Understanding

According to the CRISP-DM model, the data understanding phase serves to support business understanding. In this phase, the focus is on identifying, collecting, and analyzing data sets that can contribute to and help the project.

This phase has four tasks 1) Collect initial data, 2) Describe data, 3) Explore data, and 4) Verify data quality. The initial data comes from the 'marketing.csv' file, which was provided to us. The next three tasks will be looked at in more detail in the following paragraphs (*What Is CRISP DM? - Data Science Process Alliance*, n.d.).

3.1 Description of the Data

This paragraph concerns the examination of data and the documentation of its superficial characteristics, such as its data format, the quantity of records, or the identities of fields.

After downloading the dataset, it was imported into the R analysis tool R Studio. The dataset is a csv file, therefore the data is separated by semicolon and thus divided into various variables. This dataset has 41,188 records and contains 17 variables, including one target variable. The information included in the dataset ranges from financial factors like housing and loan status to demographic details like age, job, marital status, and education level. What is notable is that the campaign's day and month are provided, but the year is not. This is a crucial variable, even if only to help us comprehend the data better. Table 1 gives and overview of the 16 features and the target variable, which are contained in the dataset.

It was found that 1,980 rows of the dataset were duplicates, of which 55 rows had the target variable y set to 'yes'. It was also found that 10,700 rows contained missing data, particularly in fields associated with the characteristics of job, marital status, education level, housing, loan and default. Table 1 presents both the absolute number of NA values and their corresponding percentages for each attribute.

The analysis is documented in the Quarto document 'DataUnderstandingwithR.qmd', which can be executed in R Studio. Additionally, a Word document 'DataUnderstandingwithR.docx' is generated during the rendering process of the Quarto file.

Variable name	Data Format	Variable Type	Number of NA	Description and unique values (UV)/ numeric range of values
age	Integer	Numerical	-	Age of the client | Range: 17 – 98
job	Character	Categorical	330 (0.8% of job)	Type of job | UV: *'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 're-tired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'NA'*
marital	Character	Categorical	80 (~0.2% of marital)	Marital status | UV: *'divorced', 'married', 'single', 'NA'* (Note: 'divorced' means divorced or widowed)
education	Character	Categorical	1731 (4.4% of education)	Highest level of education | UV: *'basic.4y', 'basic.6y', 'basic.9y', 'high school', 'illiterate', 'professional course', 'university degree', 'NA'*
default	Character	Categorical	8597 (26.4% of default)	Has the client ever defaulted on previous debts | UV: *'yes', 'no', 'NA'*
housing	Character	Categorical	990 (2.5% of housing)	Does the client possess a housing loan | UV: *'yes', 'no', 'NA'*
loan	Character	Categorical	990 (2.5% of loan)	Does the client possess a personal loan | UV: *'yes', 'no', 'NA'*
day_of_week	Character	Categorical	-	Last contact day of week | UV: *'mon', 'tue', 'wed', 'thu', 'fri'*
month	Character	Categorical	-	Last contact month of year | UV: *'may', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'mar', 'apr', 'sep'*
campaign	Integer	Numerical	-	Number of contacts performed during this campaign and for this client | Range: 1 – 56
previous	Integer	Numerical	-	Number of contacts performed before this campaign and for this client | Range: 0 – 7
poutcome	Character	Categorical	-	Outcome of the previous marketing campaign | UV: *'failure', 'nonexistent', 'success'*
emp.var.rate	Numerical/Float	Numerical	-	Employment variation rate – quarterly indicator | Range: -3.40 –1.40
cons.price.idx	Numerical/Float	Numerical	-	Consumer price index – monthly indicator | Range: 92.20 – 94.77
cons.conf.idx	Numerical/Float	Numerical	-	Consumer confidence index – monthly indicator | Range: -50.80 – -26.90
euribor3m	Numerical/Float	Numerical	-	Euribor 3-month rate – daily indicator | Range: 0.634 – 5.045
y	Character	Categorical	-	Target variable: Has the client subscribed a fixed term deposit | UV: *'yes', 'no'*

Table 1: Feature Overview

3.2 Further Exploration of the Data

In this section, a deeper dive into the dataset is conducted through various queries and visualizations, allowing for the identification of potential relationships among the data.

Exploration of the target variable 'y'

First, we examine the distribution of the target variable 'y'. Figure 2 shows that there is a strong bias towards 'no'. As stated in Table 2, overall, 88.73% of the observations of the target variable in the dataset are 'no', revealing that the dataset is imbalanced. We will refrain from mentioning this imbalance again unless it provides additional insights.

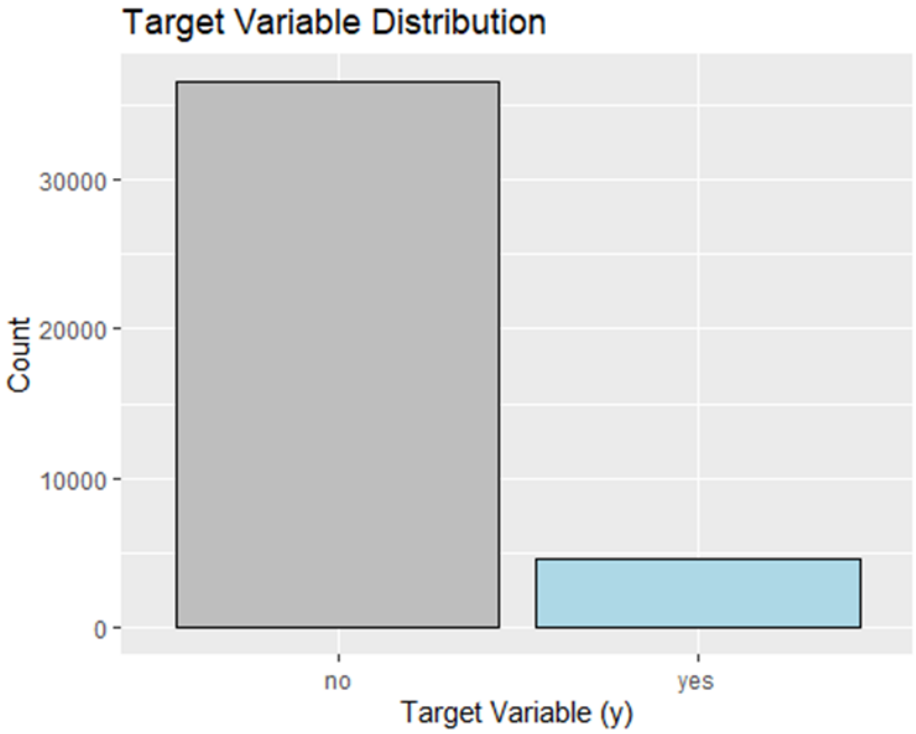

Figure 2: Distribution of Target Variable 'y'

unique values of target variable 'y'	total count	percentage
no	36,548	88.73%
yes	4,640	11.27%

Table 2: Counts and Percentages per Unique Values of Target Variable 'y'

Exploration of the feature ‘age’

The values of ‘age’ are distributed in the range from 17 to 98 years. Analyzing the distribution of ‘age’ in Figure 3 reveals a peak at about the age of 35 years, with the mean age across all clients hovering around 40 years. The median value is 38 years, which means that half of the contacted people are younger than 38 years and the other half is older than 38 years. The mean is 40.02 years, which is close to the median and indicates that the distribution of the data is relatively symmetrical. A peak in subscriptions can be observed at about the age of 30 years.

There is a drastic drop in the number of customers contacted over the age of 60. Only five individuals who were 17 years old were contacted in the entire dataset. In principle, individuals of this age are not legally permitted to make such decisions. However, it is assumed either that they were close to their 18th birthday at the time of the phone call or that the customers’ age had not been updated. Figure 3 also illustrates the imbalance of the target variable 'y', which will also be observed in the subsequent analyses of other features.

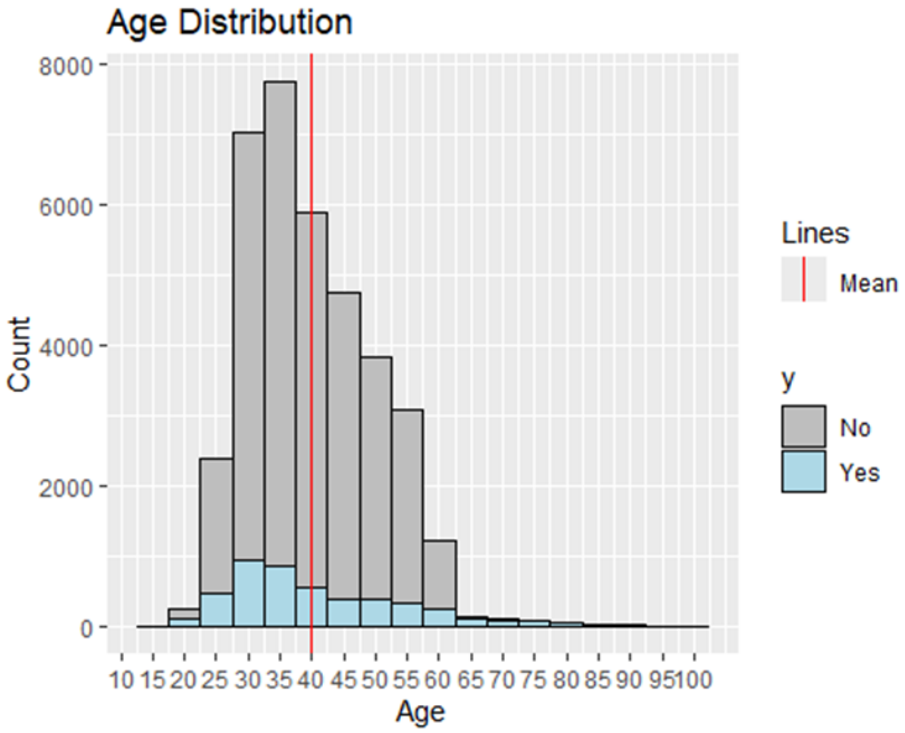

Figure 3: Distribution of Variable ‘age’

Exploration of the feature 'job'

We assume that job categories such as 'housemaid', 'retired', 'student', and 'unemployed' have a lower or no income. Figure 4 indicates that customers who earn less or have no salary tend to be contacted less frequently. Looking at Table 3, we observe that the percentage of individuals who agreed to a fixed term deposit varies across job categories. Particularly noteworthy is that 'student' shows the highest rate at around 31.4%, followed by 'retired' individuals at around 25.2%. This indicates that, despite the assumption that job categories such as 'retired' and 'student' are associated with low or no income, people in these categories are still very interested in setting up a fixed term deposit.

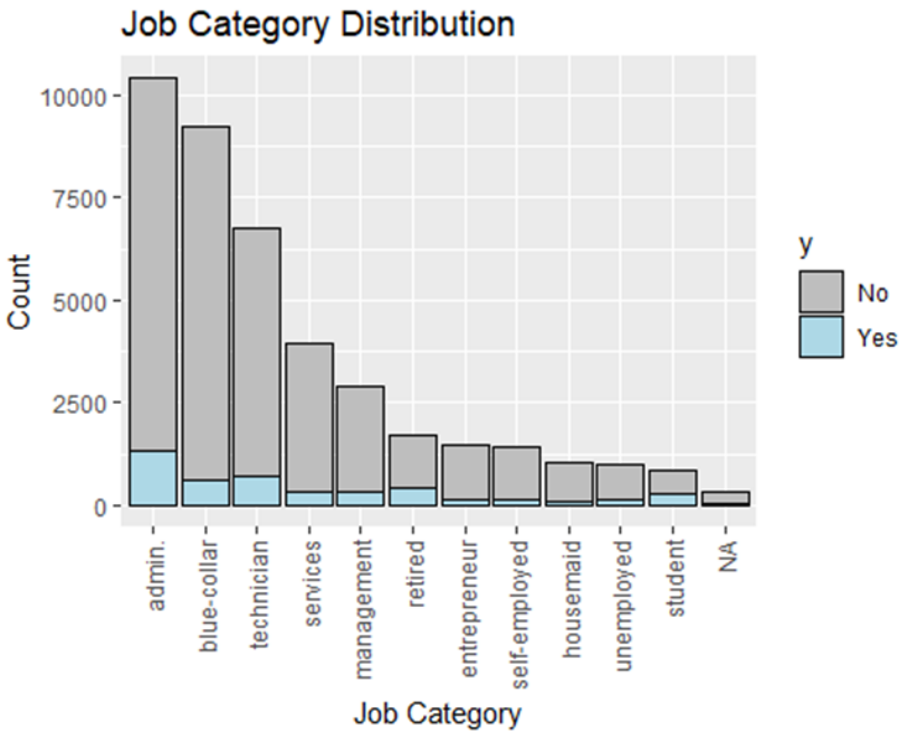

Figure 4: Distribution of Variable 'job'

unique values of 'job'	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
student	875	275	600	31.43%	68.57%
retired	1,720	434	1,286	25.23%	74.77%
unemployed	1,014	144	870	14.20%	85.80%
admin.	10,422	1,352	9,070	12.97%	87.03%
management	2,924	328	2,596	11.22%	88.78%
NA	330	37	293	11.21%	88.79%
technician	6,743	730	6,013	10.83%	89.17%
self-employed	1,421	149	1,272	10.49%	89.51%
housemaid	1,060	106	954	10.00%	90.00%
entrepreneur	1,456	124	1,332	8.52%	91.48%
services	3,969	323	3,646	8.14%	91.86%
blue-collar	9,254	638	8,616	6.89%	93.11%

Table 3: Counts and Percentages per Unique Values of 'job'

Exploration of the feature 'marital'

Figure 5 demonstrates that most of the customers contacted are married. Out of 41,188 observations, 24,928 have a marital status of 'married', which accounts for approximately 60.5% of the contacted clients. As stated in Table 4, observations with the marital status 'single' or 'NA' have the highest percentage of individuals who agreed to a fixed term deposit per status. Furthermore, 14% of the singles and 15% of customers with unknown marital status (NA) said yes to a fixed term deposit.

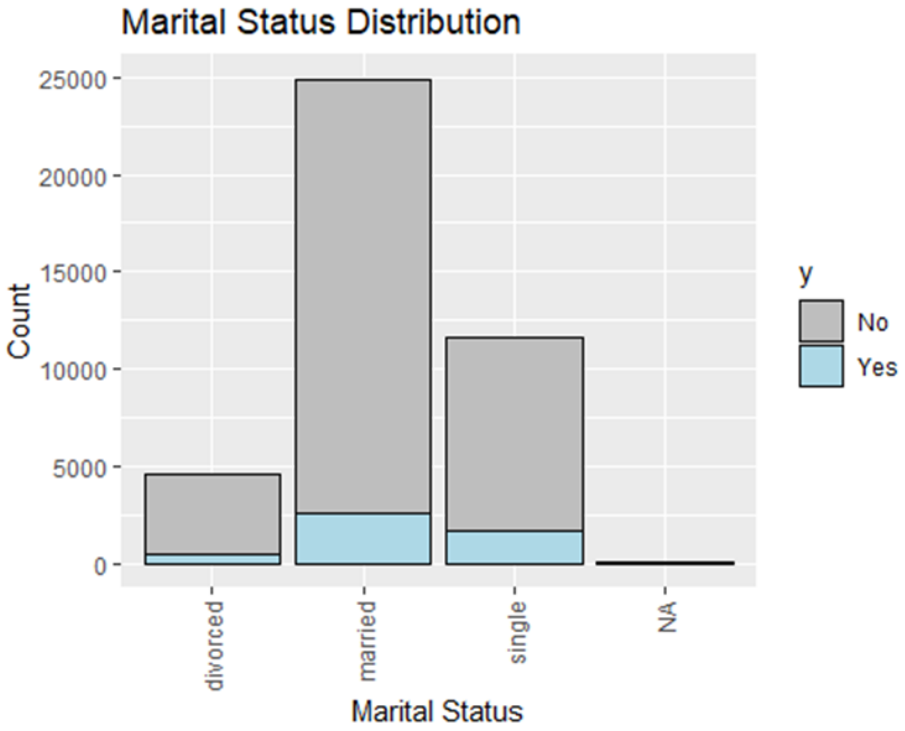

Figure 5: Distribution of Variable 'marital'

unique values of 'marital'	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
NA	80	12	68	15.00%	85.00%
single	11,568	1,620	9,948	14.00%	86.00%
married	24,928	2,532	22,396	10.16%	89.84%
NA	80	12	68	15.00%	85.00%

Table 4: Counts and Percentages per Unique Values of 'marital'

Exploration of the feature 'education'

Figure 6 shows the distribution of the education categories. It can be seen that people with a higher education, namely 'high.school', 'professional.degree' and 'univeryits.degree' are more often contacted. However, when considering the percentage of individuals who agreed to a fixed term deposit per education category, see Table 4, it is notable that 'illiterate' shows the highest conversion rate at around 22.2%, followed by individuals with unknown education level (NA) at around 14.5%, 'university.degree' with around 13.7% and 'professional.degree' with around 11.4%. This indicates that customers with a higher level of education are more likely to subscribe to fixed term deposits. The significance of the percentage of 22.2% illiterates agreed to a fixed term deposit is uncertain, considering only 18 out of 41,188 contacted individuals were categorized as 'illiterate'.

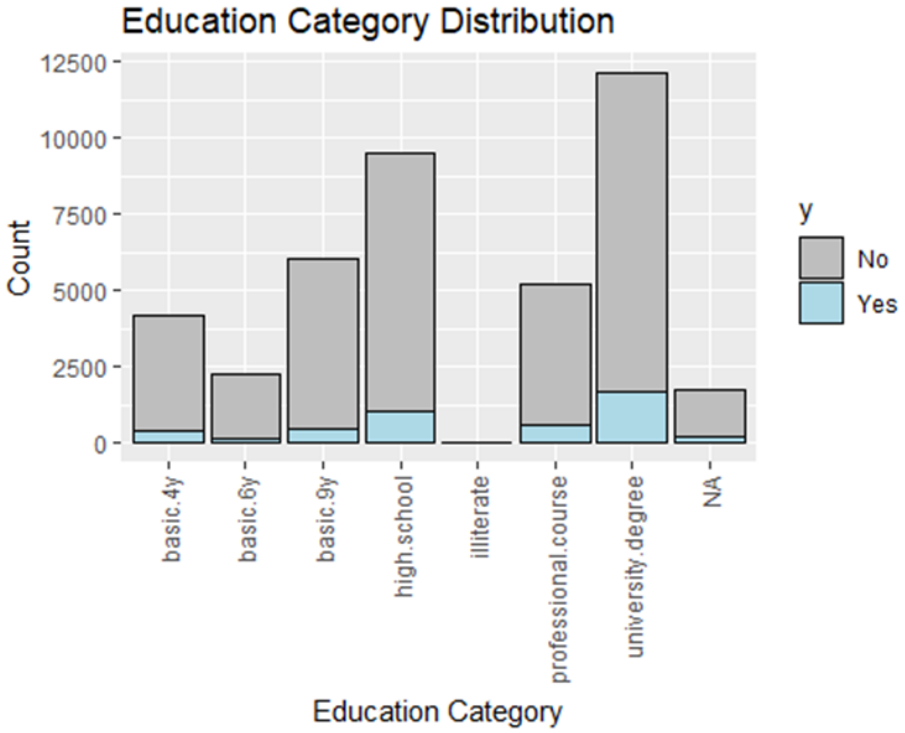

Figure 6: Distribution of Variable 'education'

unique values of 'education'	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
illiterate	18	4	14	22.22%	77.78%
NA	1,731	251	1,480	14.50%	85.50%
university.degree	12,168	1,670	10,498	13.72%	86.28%
professional.course	5,243	595	4,648	11.35%	88.65%
high.school	9,515	1,031	8,484	10.84%	89.16%
basic.4y	4,176	428	3,748	10.25%	89.75%
basic.6y	2,292	188	2,104	8.20%	91.80%
basic.9y	6,045	473	5,572	7.82%	92.18%

Table 5: Counts and Percentages per Unique Values of 'education'

Exploration of the feature 'deafult'

A closer look at the default status variable indicates an imbalance, visualized in Figure 7.Only three instances are classified as 'yes', meaning this person has a default. None of these three customers have subscribed to a fixed term deposit. Furthermore, as indicated in Table 1, about 26.4 % of the data is missing (NA).

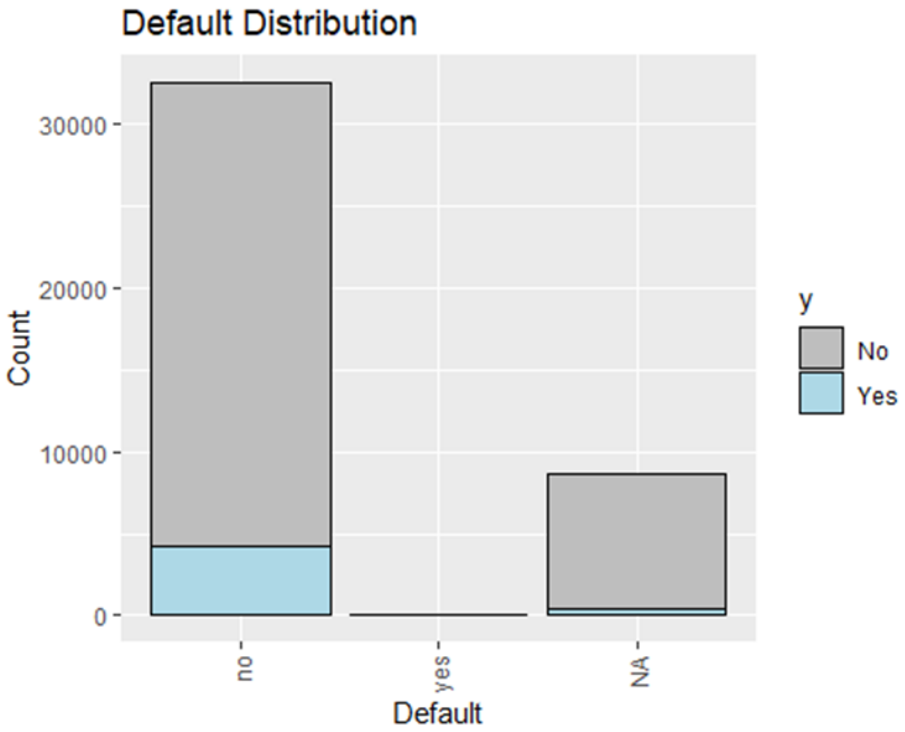

Figure 7: Distribution of Variable 'default'

Exploration of the feature 'housing'

Figure 8 displays the distribution of housing loans. The data appears fairly balanced; however, slightly more people with a housing loan were contacted. Regarding housing status, clients with housing loans show a slightly higher inclination towards fixed term deposit subscriptions compared to those without. Looking at Table 6, 11.6% of contacted individuals with a housing loan agreed to a fixed term deposit, whereas 10.8% of those without a housing loan agreed.

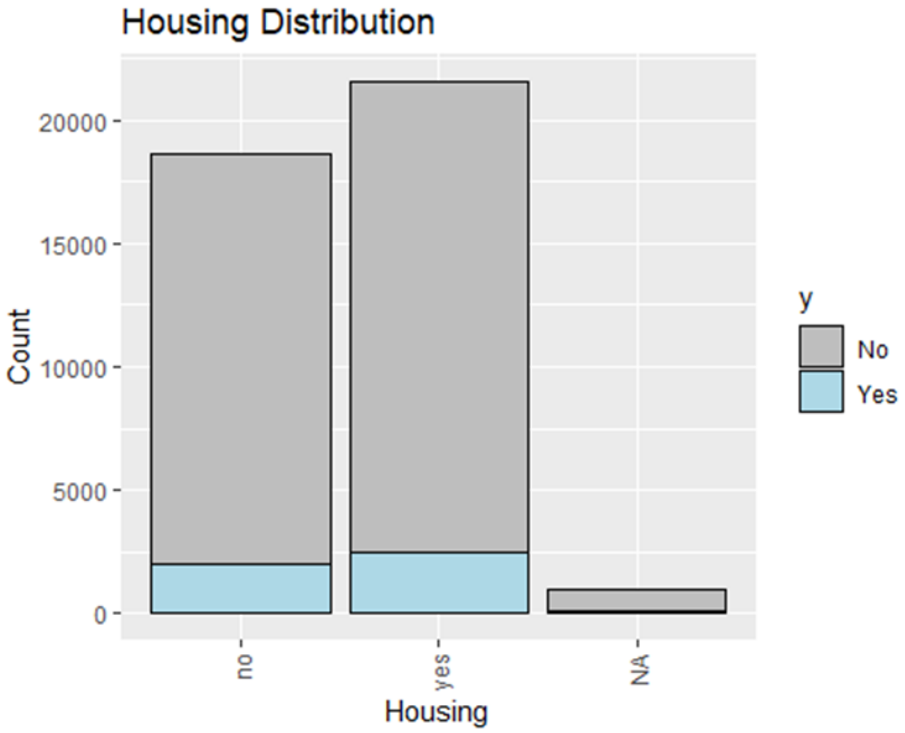

Figure 8: Distribution of Variable 'housing'

unique values of 'housing	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
yes	21,576	2,507	19,069	11.62%	88.38%
no	18,622	2,026	16,596	10.88%	89.12%
NA	990	107	883	10.81%	89.19%

Table 6: Counts and Percentages per Unique Values of 'housing'

Exploration of the feature 'loan'

Figure 9 reveals an imbalance in the data; more people without loans were contacted. Regarding the loan status, clients without loans show a slightly higher inclination towards fixed term deposit subscriptions compared to those with loans. As stated in Table 7, 11.3% of contacted individuals without loans agreed to a fixed term deposit, whereas 10.9% of those with loans agreed. Approximately the same percentage, 10.8%, of people contacted with no information on their loan status (NA) subscribed to a fixed term deposit.

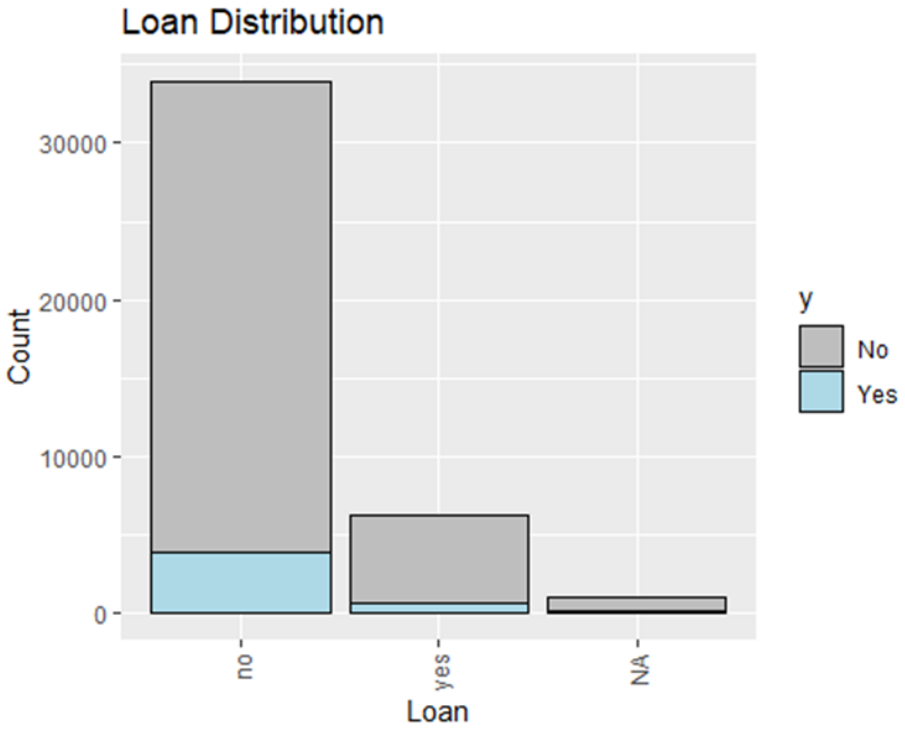

Figure 9: Distribution of Variable 'loan'

unique values of 'loan	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
no	33,950	3,850	30,100	11.34%	88.66%
yes	6,248	683	5,565	10.93%	89.07%
NA	990	107	883	10.81%	89.19%

Table 7: Counts and Percentages per Unique Values of 'loan'

Exploration of the feature 'month'

Figure 10 illustrates that the bank did not initiate any contacts regarding this marketing campaign in January and February. The peak, where the highest number of contacts occurred, was in May, while September, October, and December witnessed notably fewer contacts. If we consider that September and October are typically months when many people in Portugal take their summer vacations, and December is the Christmas holiday season, during which many individuals may also take extended breaks, these assumptions may explain the reduced number of contacts during these months. However, it is important to note that the observations only include the month of collection without mentioning the specific year. Therefore, for May, we have data for three years, namely May 2008, May 2009, and May 2010, whereas for December, we only have data for 2 years, which are December 2008 and December 2009, as shown in Table 8. This variability in available data may also contribute to fluctuations in contact numbers.

Furthermore, when looking at Table 9 and analyzing the percentage of individuals who agreed to a fixed term deposit per month, March exhibits the highest rate at around 50.6%, followed by December at around 48.9%, September with around 44.9%, and October with around 43.9%. This observation suggests that in March, individuals may have been more inclined to subscribe for a fixed term deposit or even might have been waiting to get contacted regarding fixed term deposits, especially considering the absence of calls in January and February. Similarly, in September and October, individuals may have had more time to inform themselves about fixed term deposits due to vacation or time off. The same assumption could be made for December. An additional aspect to highlight is the fact that months with a high percentage of fixed term deposit subscriptions demonstrate a significantly low total of contact, as shown in Table 9.

year	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
2008					x	x	x	x	x	x	x	x
2009			x	x	x	x	x	x	x	x	x	x
2010			x	x	x	x	x	x	x	x	x	
Total			**2**	**2**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**

Table 8: Monthly Distribution Over Three Years (2008-2010)

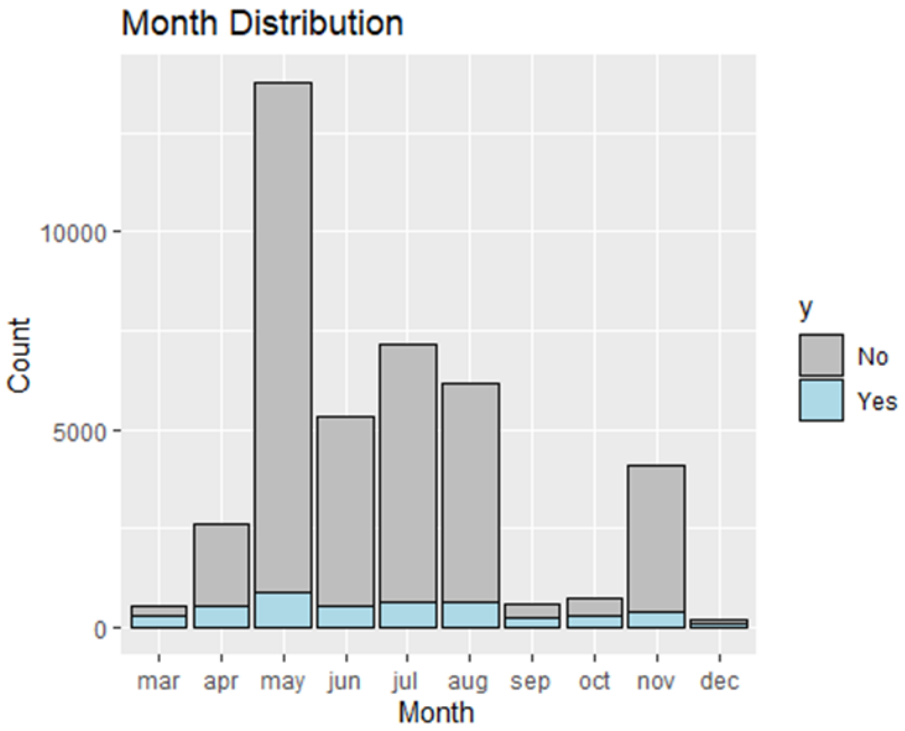

Figure 10: Distribution of Variable ‘month’

unique values of ‘month	total count	count ‘yes’	count ‘no’	% ‘yes’	% ’no’
mar	546	276	270	50.55%	49.45%
dec	182	89	93	48.90%	51.10%
sep	570	256	314	44.91%	55.09%
oct	718	315	403	43.87%	56.13%
apr	2,632	539	2,093	20.48%	79.52%
aug	6,178	655	5,523	10.60%	89.40%
jun	5,318	559	4,759	10.51%	89.49%
nov	4,101	416	3,685	10.14%	89.86%
jul	7,174	649	6,525	9.05%	90.95%
may	13,769	886	12,883	6.43%	93.57%

Table 9: Counts and Percentages per Unique Values of ‘month’

Exploration of the feature 'day_of_week'

The number of customers contacted and fixed term deposits concluded is distributed roughly equally across the days of the week, as can be seen in Figure 11. Furthermore, when analyzing the percentage of individuals who agreed to a fixed term deposit per day of week, Thursday exhibits the highest rate at around 12.1% and Monday the lowest rate at around 9.95%, as stated in Table 10.

It is worth noting that there is no data available for weekend days (Saturday and Sunday), as the bank is likely closed, and its employees do not work during these days. Additionally, it is important to mention that the dataset only includes data from May 2008 to November 2010, with information categorized by weekdays rather than specific dates.

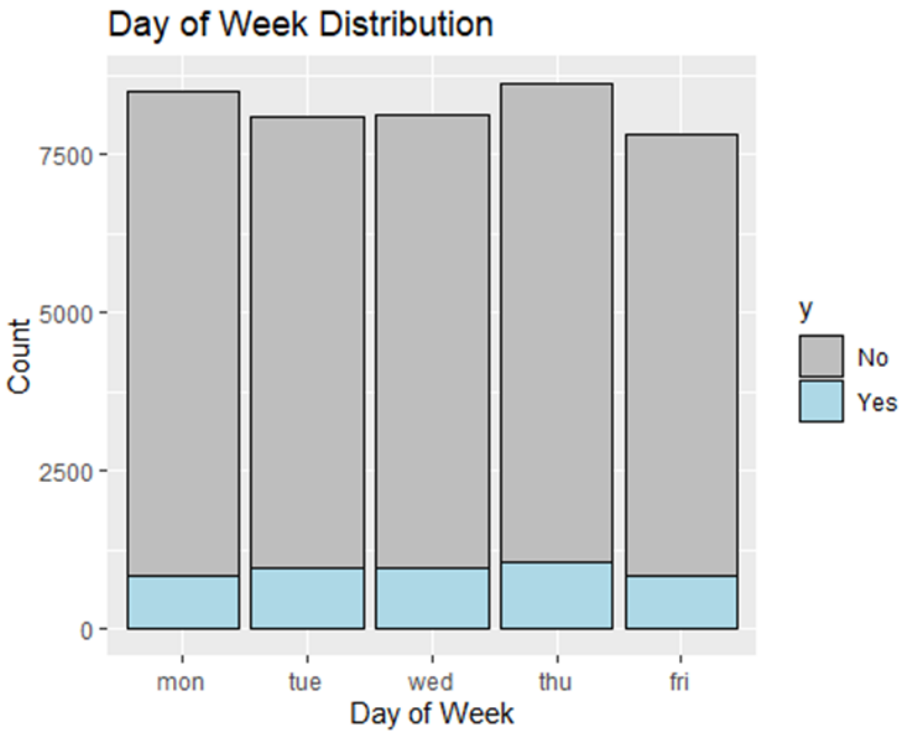

Figure 11: Distribution of Variable 'day_of_week'

unique values of 'day_of_week'	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
thu	8,623	1,045	7,578	12.12%	87.88%
tue	8,090	953	7,137	11.78%	88.22%
wed	8,134	949	7,185	11.67%	88.33%
fri	7,827	846	6,981	10.81%	89.19%
mon	8,514	847	7,667	9.95%	90.05%

Table 10: Counts and Percentages per Unique Values of 'day_of_week'

Exploration of the feature 'campaign'

Examining the number of contacts made during a campaign in relation to subscriptions suggests a point of diminishing returns beyond 6 contacts, as visualized in Figure 12. As shown in Figure 13, several outliers in the feature 'campaign' can be observed. Furthermore, when analyzing the percentage of individuals who agreed to a fixed term deposit per number of contacts per campaign, one contact exhibits the highest rate at around 13%, followed by two contacts at around 11.5%, and then three contacts at around 10.8%. As stated in Table 11 the percentage further declines with an increasing number of contacts. Optimizing resource allocation by limiting campaign contacts, particularly focusing efforts on the first three or four interactions, could potentially enhance subscription rates.

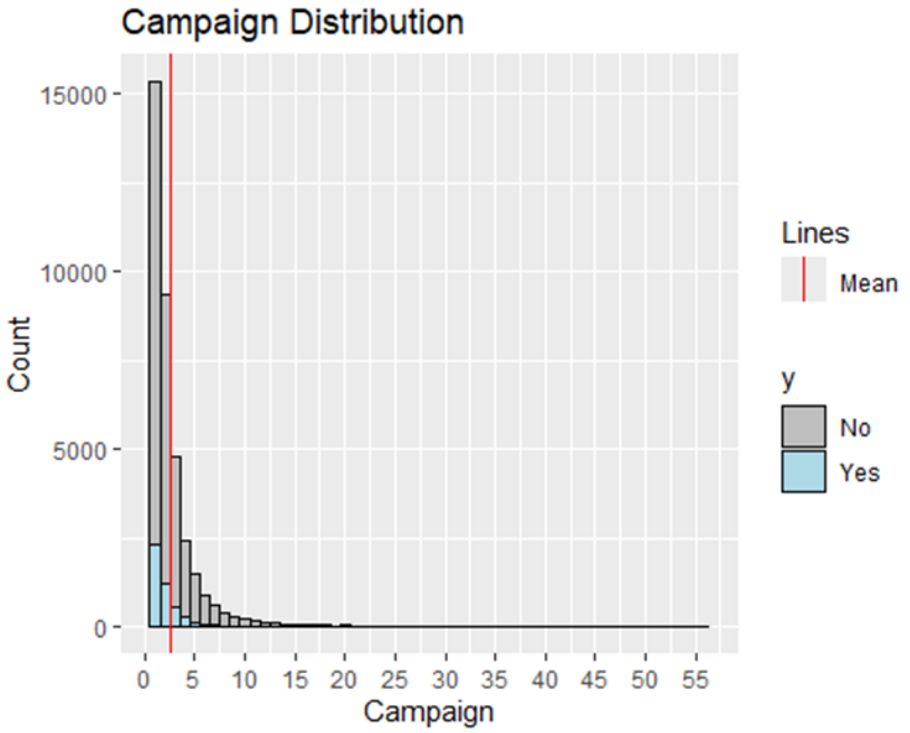

Figure 12: Distribution of Variable 'campaign'

unique values of 'campaign'	total count	count 'yes'	count 'no'	% 'yes'	% 'no'
1	17,642	2,300	15,342	13.04%	86.96%
2	10,570	1,211	9,359	11.46%	88.54%
3	5,341	574	4,767	10.75%	89.25%
4	2,651	249	2,402	9.39%	90.61%
…	…	…	…	…	…
43	2	0	2	0%	100%
56	1	0	1	0%	100%

Table 11: Counts and Percentages per Unique Values of 'campaign'

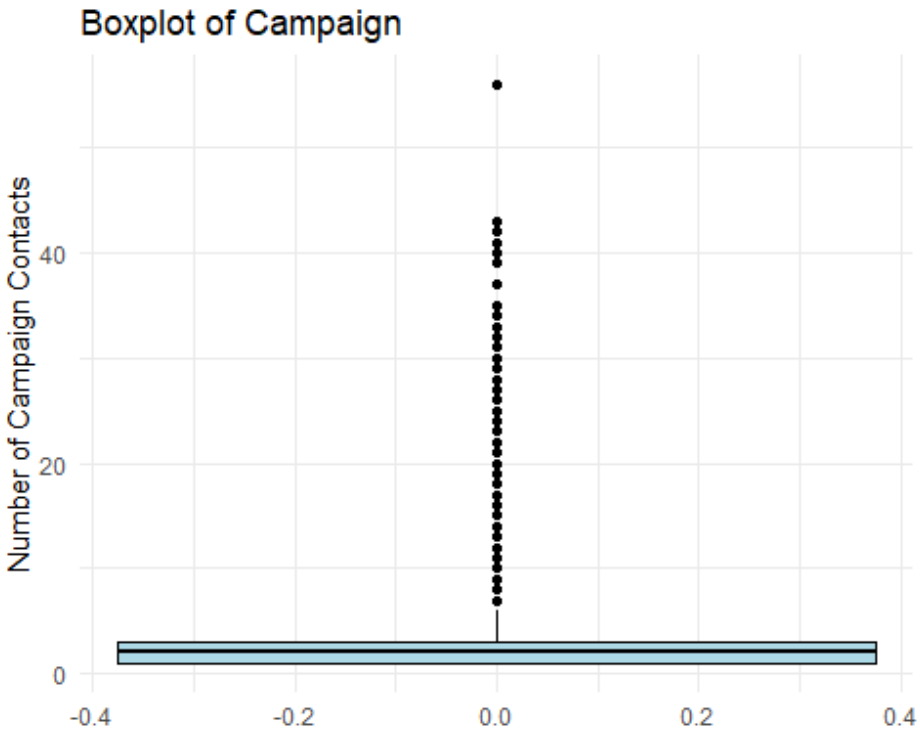

Figure 13: Boxplot of Variable ‘campaign’ revealing several outliers

Exploration of the feature 'previous'

As visualized in Figure 14, 35,563 customers, or about 86.3% of the observations, have not been contacted before for this campaign. Only a minority have been already contacted one or more times before this campaign.

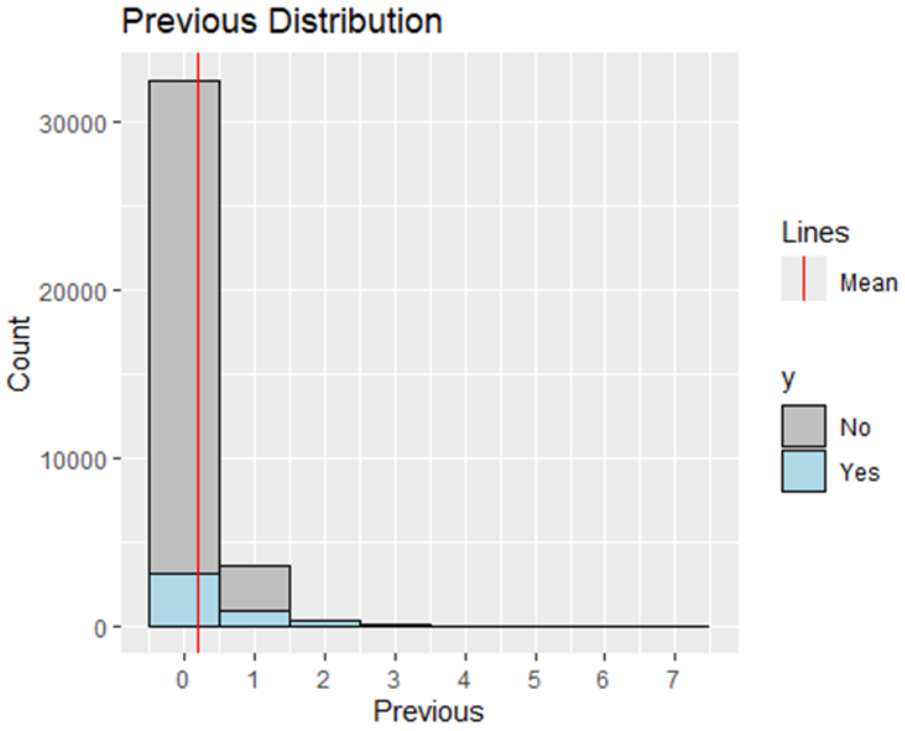

Figure 14: Distribution of Variable 'previous'

Exploration of the feature ‘poutcome’

The feature 'poutcome' is closely associated with the 'previous' attribute, where the absence of 'previous' contact results in a nonexistent outcome. As depicted in Figure 15 most of the data falls into the 'nonexistent' category, in numbers 35,563 entities. This is about 86.3% of the data, the exact same numbers as for no ‘previous’ contacts. This relationship can also be seen in Figure 16. The depicted correlation underscores the importance of considering previous interactions when predicting the outcomes of current encounters. Particularly noteworthy is the finding that individuals who have previously subscribed to a fixed term deposit ('poutcome = success') exhibit a conversion rate of approximately 65.1% for subscribing again.

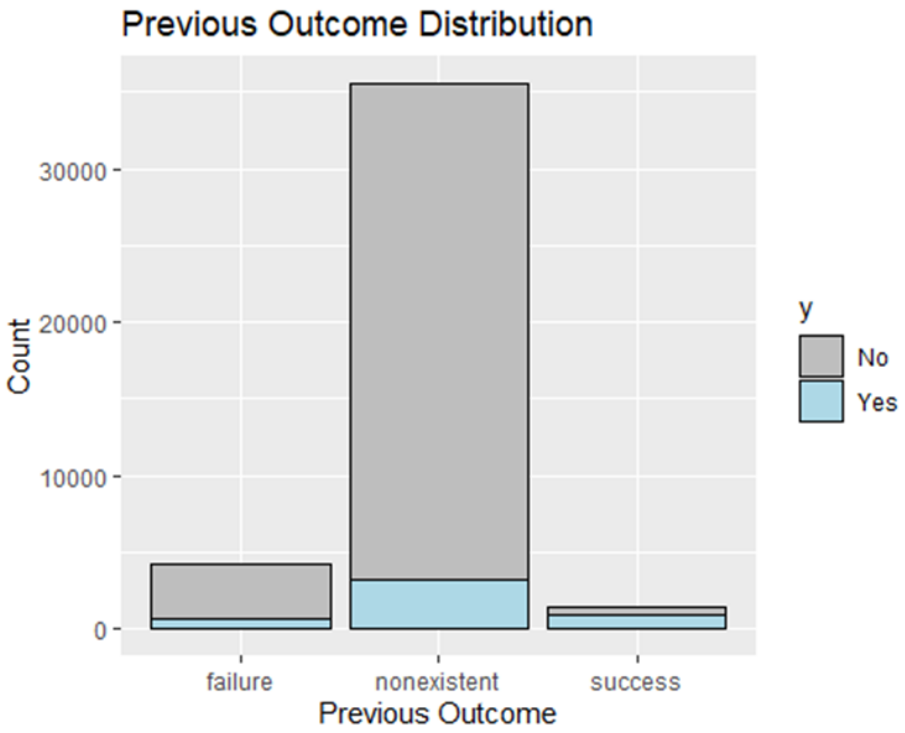

Figure 15: Distribution of Variable ‘poutcome’

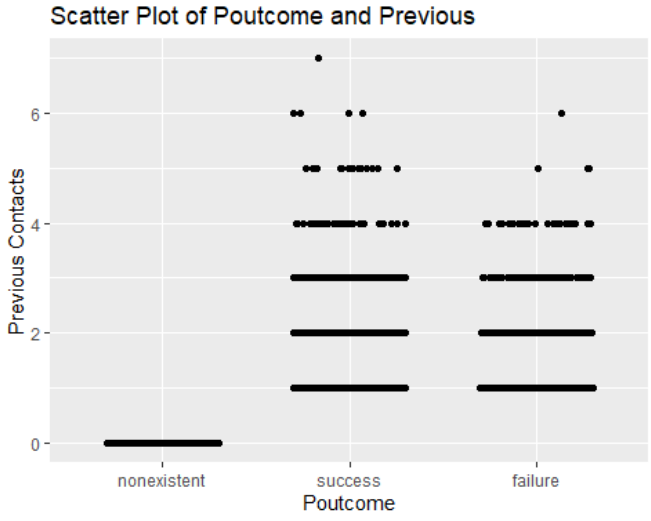

Figure 16: Scatterplot showing the relationship of 'previous' and 'poutcome'

Exploration of the feature 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', and 'euribor3m'

Figure 17 to Figure 20 depict various economic rates and indices. Figure 17 illustrates the distribution of the employment variation rate, Figure 18 visualizes the distribution of the consumer price index, Figure 19 illustrates the distribution of the consumer confidence index, and Figure 20 shows the distribution of the Euribor 3-month rate. Identifying any patterns or trends is challenging since we lack exact dates or a time series. However, these attributes can help in interpreting the economic situation, particularly considering that the data was collected during the Euro Crisis in Portugal. Additionally, the Euribor 3-month rates, as a daily indicator, can be utilized to derive a new attribute 'year,' but this will be addressed in the subsequent section Data Preparation.

General Interpretation of Employment Variation Rate

An employment variation rate of over 1% typically indicates a positive trend in employment, suggesting an increase in the number of people finding jobs. This can be a sign of economic growth, increased business activity, and overall optimism in the job market. It implies that more individuals are entering the workforce or that existing workers are finding new job opportunities (*Beschäftigungsquote – Wikipedia*, n.d.).

Conversely, a negative employment variation rate indicates a decrease in employment levels, signifying a decline in the number of people employed. This could be indicative of economic contraction, layoffs, or other factors contributing to job losses. A negative employment variation rate often reflects economic challenges, such as recession, industry downturns, or structural changes in the labor market (*Beschäftigungsquote – Wikipedia*, n.d.).

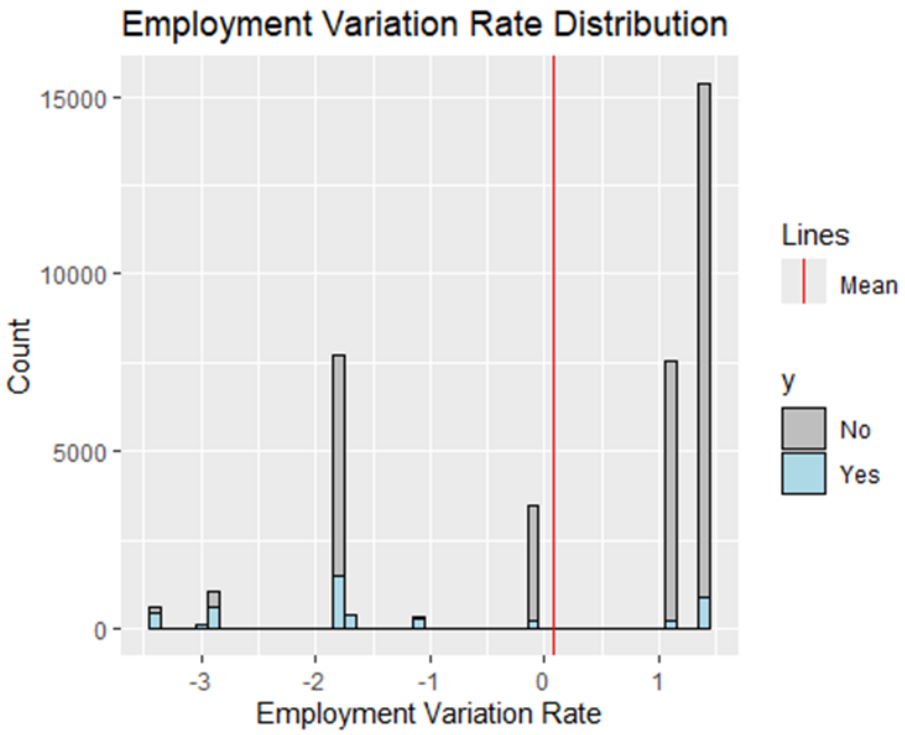

Figure 17: Distribution of Variable 'emp.var.rate'

General Interpretation of Consumer Price Index

A consumer price index (CPI) of over 94 typically indicates a relatively high level of inflation. This means that, on average, the prices of goods and services included in the CPI basket have increased significantly compared to a base period. A CPI above 94 suggests that consumers are experiencing rising costs for essential goods and services, which can reduce purchasing power over time. High inflation can also lead to uncertainty, as consumers may expect further price increases in the future. Economically, high inflation can indicate robust demand, increased consumer spending, and potentially overheating in the economy (*Consumer Price Index - Wikipedia*, n.d.).

Conversely, a CPI of under 93 suggests relatively low inflation or even deflation. In this case, the prices of goods and services may be decreasing on average compared to the base period. A CPI below 93 indicates that consumers may be experiencing stable or falling prices for goods and services, which can potentially boost purchasing power and stimulate consumer spending. However, persistent deflation can also signal economic weakness and may pose challenges for businesses and policymakers (*Consumer Price Index - Wikipedia*, n.d.).

Figure 18: Distribution of Variable ‘cons.price.idx’

General Interpretation of Consumer Confidence Index

A negative consumer confidence index indicates that consumers are pessimistic about the state of the economy and their financial well-being. It suggests that consumers are less confident about current economic conditions, future economic prospects, and their own financial situations. From an economic perspective, a negative consumer confidence index can have several implications:

When consumers are pessimistic about the economy, they are more likely to cut back on their spending. This decrease in consumer spending can have a significant impact on economic growth, as consumer spending is a major driver of economic activity (*Consumer Confidence - Wikipedia*, n.d.).

Negative consumer sentiment can also affect investment decisions by businesses, leading to businesses scaling back their investment plans, delaying expansion projects, or reducing production levels. This can lead to lower levels of capital investment, decreased business confidence, and slower economic growth (*Consumer Confidence - Wikipedia*, n.d.).

To summarize, a negative consumer confidence can contribute to a broader sense of economic uncertainty and pessimism. A prolonged period of negative consumer confidence can lead to a self-reinforcing cycle of economic downturn, as decreased spending and investment further weaken economic conditions.

As visualized in Figure 19, all observations exhibit a negative value for the feature consumer confidence index.

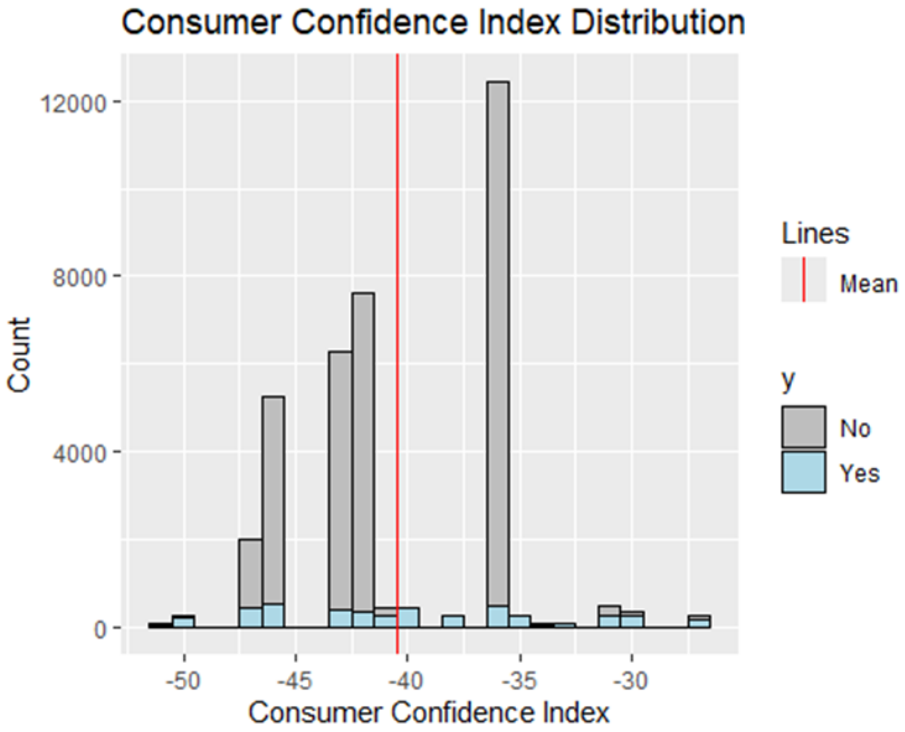

Figure 19: Distribution of Variable ‘cons.conf.idx’

General Interpretation of Euribor 3-month Rate

The Euribor 3-month Rate is an interest rate at which European banks lend to one another on the interbank market. It serves as a benchmark for short-term interest rates in the eurozone and can provide insights into the economy of a country in several ways (*Euribor - Wikipedia*, n.d.).

The Euribor 3 Month Rate can also reflect broader economic conditions. A declining Euribor rate may indicate weak economic growth or recession, as central banks lower rates to stimulate borrowing and spending. Conversely, a rising Euribor rate may suggest stronger economic growth or inflationary pressures, leading central banks to raise rates to cool down the economy (*Euribor - Wikipedia*, n.d.).

Furthermore, fluctuations in the Euribor 3 Month Rate can impact financial stability. A sudden increase in the Euribor rate may indicate liquidity problems in the banking system or concerns about credit risk, which can undermine confidence and stability in financial markets (*Euribor - Wikipedia*, n.d.).

Therefore, it is essential to consider as well other economics indices to understand the economic context.

As illustrated in Figure 20, the values for the feature 'euribor3m' are distinctly separated into two groups. This observation suggests the possibility of either a sudden increase or a drop in the Euribor 3-month rate.

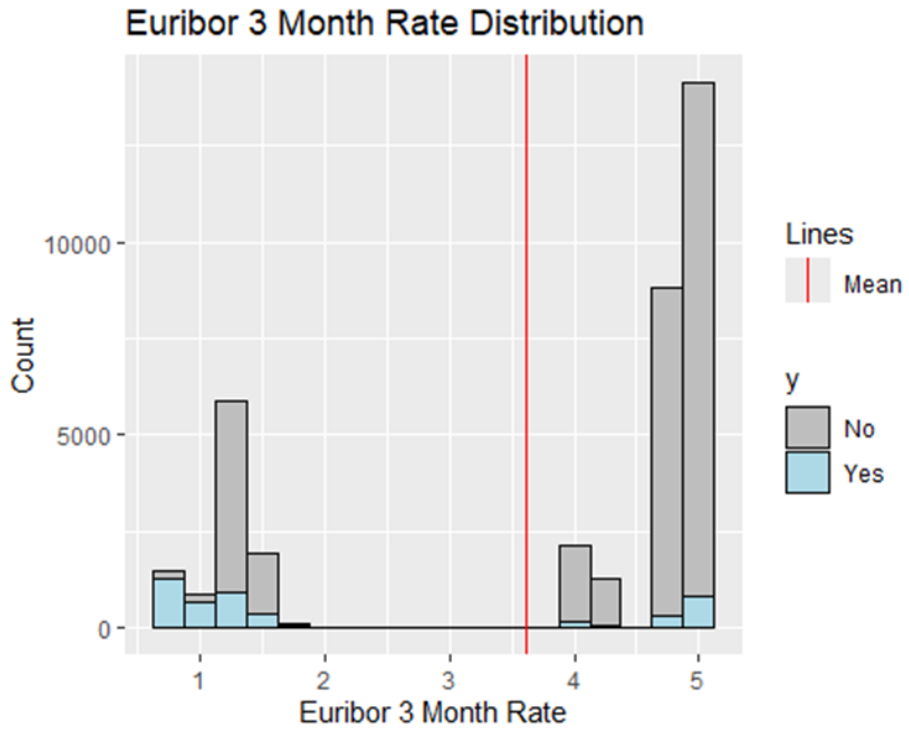

Figure 20: Distribution of Variable ‘euribor3m’

The analysis is documented in the Quarto document ‘DataUnderstandingwithR.qmd’, which can be executed in R Studio. Additionally, a Word document ‘DataUnderstandingwithR.docx’ is generated during the rendering process of the Quarto file. In the Quarto document ‘DataUnderstandingwithR.qmd’ more analyses were made to find further relationships. However, no further relationships have been found. By applying the widget correlations in the Orange workflow ‘Data Preparation FINAL.ows’, highly correlated attributes have been identified.

After loading in the initial dataset from ‘marketing.csv’ file into Orange, we applied the Correlations widget, the output is shown in Figure 21. It can be seen that the employment variation rate and the Euribor 3-month rate have a high positive correlation of 0.972. Additionally, both the consumer price index and the employment variation rate with 0.775, as well as the consumer price index and the Euribor 3-month rate with 0.688, exhibit a high correlation rate.

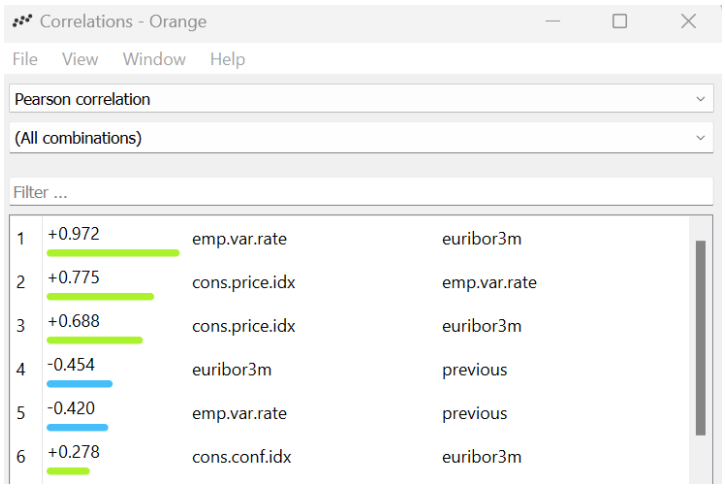

Figure 21: Output of correlations on initial data (Orange)

3.3 Verification of the Data Quality

This paragraph describes how clean or dirty the data is and documents any quality issues.

According to our analysis, the dataset comprises 41,188 entries spread across 17 features, including the target variable. We identified 1,980 duplicated rows, accounting for approximately 4.8% of the dataset. Furthermore, we discovered that 10,700 rows have at least one missing value, representing about 26% of the rows. Considering the total number of entries (41,188) and features (17), we have a total of 700,196 values. Among these, there are 12,718 missing values, which equates to approximately 1.8% of the total values.

Additionally, we observed that the dataset exhibits an imbalance regarding the target variable 'y', as stated in Table 2. Specifically, 36,548 records have the value 'no', constituting about 88.7% of the dataset, while only 4,640 records have 'y' = 'yes'. This imbalance is noteworthy and could potentially introduce a bias towards the 'no' class.

Further we also found that the values of the features ‘default’ and ‘loan’ are imbalanced. For ‘default’ there are 32,588 records without a default, what is about 79.1%, 3 records with a default, and 8,597 records without a value (NA), what is about 20.9%. And for ‘loan’, there are 33,950 records without a loan, what is about 82.4%, 6,248 records with a loan, and 990 records have no value (NA). To fully interpret the distribution of these attributes, as well as others, we would have needed a representative sample of the Portuguese population, or the distribution of these characteristics within the entire population.

We also identified that the 'year' attribute could provide valuable insights, not for modeling purposes but for gaining a deeper understanding of the data and its relation to the prevailing economic conditions during the data collection period. In the Data Preparation phase, we incorporated the 'year' attribute using information from the Euribor 3-month rate, a process that will be elaborated on later. However, while developing this 'year' attribute, we observed that the data follows a sequential pattern. Specifically, the dataset begins with entries from May 2008 and concludes with entries from November 2010. Moreover, the data between these time points is organized chronologically by both year and month.

The data has several quality issues and can be considered relatively dirty. The presence of duplicated rows, a high percentage of missing values, and imbalanced distributions in certain features, especially the target variable, indicates that the data requires cleaning. The steps taken to clean the data are documented in the following paragraph.

The analysis is documented in the Quarto document ‘DataUnderstandingwithR.qmd’, which can be executed in R Studio. Additionally, a Word document ‘DataUnderstandingwithR.docx’ is generated during the rendering process of the Quarto file.

4 Data Preparation

According to the CRISP-DM model the Data Preparation phase develops the final data set for modeling. This phase has five tasks: 1) Select data, 2) Clean data, 3) Construct data, 4) Integrate data, and 5) Format data.

All tasks were taken into consideration; however, the following paragraphs align with the order in which we conducted the data preparation process. To prepare the data, we utilized Orange and constructed a workflow designed to clean the data, which is saved in the file 'Data Preparation FINAL.ows'. Before using Orange for data preparation, we attempted to do so using R Studio and a Quarto script named ‘DataPreparationwithR.qmd’, which also generates a Word document, ‘DataPreparationwithR.docx’. Due to the complexity involved in data preparation using R, we opted for the Orange version.

4.1 Handling Duplicates and Missing Values (NA)

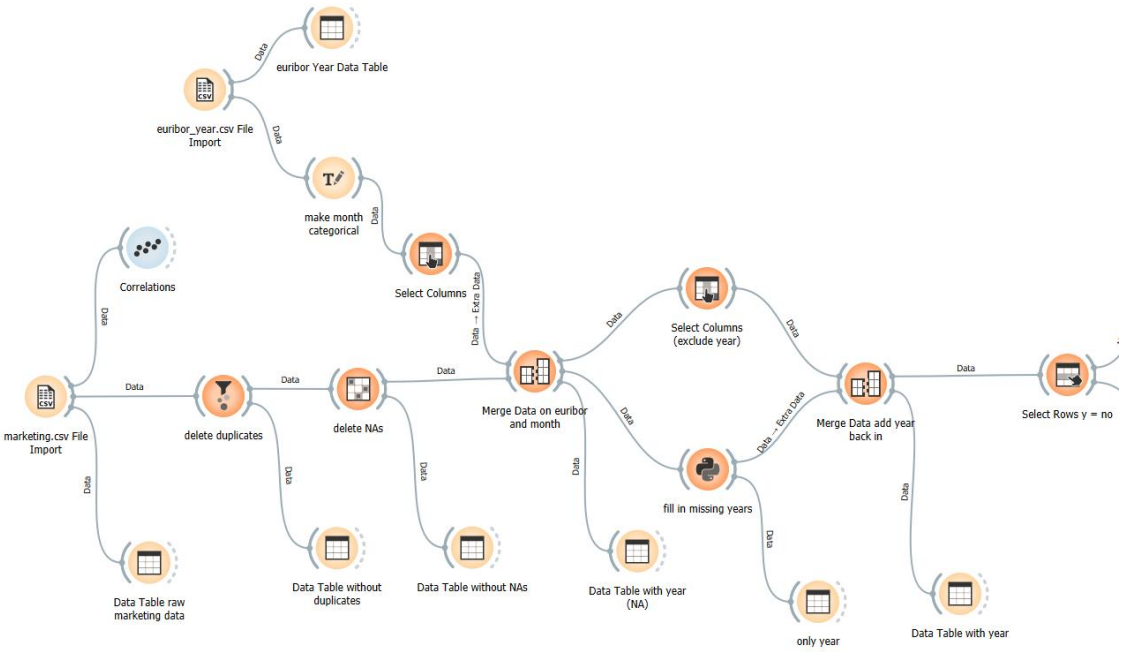

Figure 22: Extract 1/3 of Orange workflow for Data Preparation

Figure 22 shows an extract of our Orange workflow, and how we applied the widgets to exclude the duplicated rows and rows containing NA values.

We used the Unique widget to show occurrences in which every entry was the same across all features to identify the duplicates. By excluding the duplicate rows, we reduced the number of instances in the dataset to 39,612.

The next important step was to address missing values (NA). Applying the Impute widget made it easier to exclude cases where the data was missing, which again decreased the number of records to 28,912.

In the Quarto document ‘DataPreparationwithR.qmd, which can be executed in R Studio, we analyzed if we should exclude all rows that have at least one NA value or if we should consider some and fill in the NA value with the mean or most popular category. We found that excluding all rows containing NA values is the best option, as we want to reduce the number of records and get a more balanced dataset. Furthermore, we observed that by deleting observations with a missing value, an improvement of the percentage of the target variable ‘y’= ‘yes’ was visible. Additionally, a Word document ‘DataPreparationwithR.docx’ is generated during the rendering process of the Quarto file.

4.2 Integration year

One issue we encountered was the absence of a specific 'year' attribute in the dataset, as it only includes the 'month' attribute spanning from May 2008 to November 2010. While the 'year' attribute may not significantly influence predictive models, having a complete dataset is advantageous. Despite encountering missing values when directly integrating the year based on the Euribor 3-month rate dataset (*Euribor-Werte pro Jahr*, n.d.) and the month attribute (due to identical values across different months and years), we could still identify breakpoints, as the dataset appears to follow a sequential pattern.

To address this, we initiated a comprehensive data analysis to pinpoint these sequence breakpoints. Utilizing the Python Script widget, we dynamically assigned years to rows with missing values by generating the requisite Python code. Specifically, rows indexed from 1 to 18,096 were assigned to the year 2008, rows 18,097 to 27,131 to the year 2009, and rows 27,132 to 28,912 (comprising the end of the dataset) to the year 2010. This methodology ensures the temporal consistency of our data, enabling a deeper understanding of both the data itself and the economic crisis context.

Figure 22 shows the first extract of our Orange workflow, we imported the years with the Euribor 3-month rate dataset and merged it with the existing data, then as explained above, we fill in the missing values with the Python Script widget, containing the above described python code.

4.3 Balancing Data Set

For a model to be unbiased, it is crucial to have a balanced dataset regarding the target variable. Figure 23 shows the second extract of our Orange workflow. After selecting the rows, where the target variable ‘y’ is ‘no’, we apply the Data Sampler widget. We used the ‘Stratify sample’ option to keep the structure of the original dataset, as well as the option ‘Replicable (deterministic) sampling’ to ensure that each user receives identical datasets, enabling consistent outcomes. We preserved proportionality and correlations by imitating the original dataset’s variable distribution, maintaining the fundamental structure of the data while minimizing its size. With the widget Data Sampler, we reduce the number of rows, where the target variable is ‘no’, to 3,808, the same number of rows with the target variable ‘y’ = ‘yes’. By concatenating the reduced data sample with target variable equals no to the rows with ‘y’ = ‘yes’, a balanced dataset was created. Where 50% of the target variable equals ‘no’ and 50% equals ‘yes’.

4.4 Standardize Economic Rates/Indices and Construct Data

Figure 23 shows that we have standardized the attributes ‘euribor3m’, ‘emp.var.rate’, ‘cons.price.idx’ and ‘cons.conf.idx’ with the Continuize widget. Here the mean was subtracted from each value and the result then gets divided by the standard deviation of the variable. Data standardization ensures consistency and compatibility across datasets, facilitating efficient processing and analysis. By transforming data into a standard format, it becomes easier to compare and interpret, benefiting both humans and machines.

Next, we combined the attributes 'cons.price.idx' and 'cons.conf.idx' into a single attribute called 'prod_indices' as it represents the product of the two indices. We combined these attributes as the Consumer Price Index and the Consumer Confidence Index are interconnected, with changes in one often directly influencing the other. For instance, rising inflation, reflected by an increase in the CPI, can erode consumer confidence, leading to reduced spending. Conversely, stable prices can enhance consumer sentiment and encourage spending. Therefore, these two economic indicators have been combined by multiplying them to represent the overall economic sentiment in Portugal, using the Aggregate Columns widget. Afterwards the newly created feature was inserted back into the dataset as a single attribute ‘prod_indices’ using the Merge Data widget.

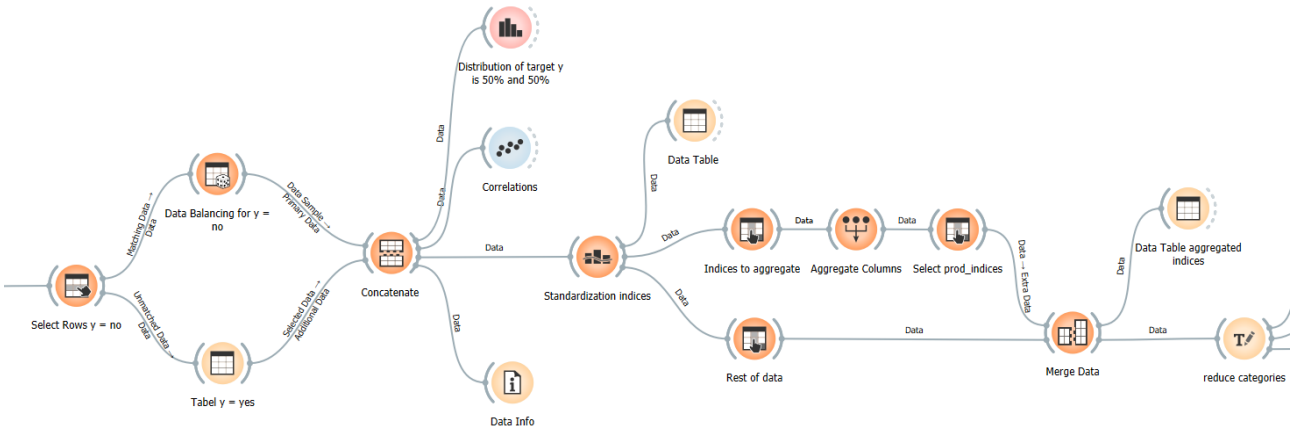

Figure 23: Extract 2/3 of Orange workflow for Data Preparation

4.5 Reducing Categories / Grouping

Figure 24 shows how we have applied the Edit Domain widget in Orange to reduce the number of categories in the attributes that are categorical. We have simplified the categories for attributes such as job and education.

We divided the groupings in the education characteristic into three different categories: basic, intermediate, and high. The ‘illiterate’ category is included in the basic category since it is only present in a limited number of records and corresponds to basic educational levels such as ‘basic.4y’, ‘basic.6y’ and ‘basic.9y’. Levels of education above basic education but below university education are included in the intermediate category; which are ‘professional.course’ and ‘high.school.’ Graduates of ‘university.degree’ are represented by the high category.

In the same way, we divided the job attribute into three categories: executives, employees, and non-employed. Executives are higher-level positions and include professions such as ‘technician’, ‘manager’, and ‘entrepreneur’. Employees include those in ‘admin.’, ‘blue-collar’, ‘self-employed’, ‘housemaid’ and ‘service’ roles; here ‘self-employed’ people are assumed to have comparable salary levels. Individuals who are not actively employed in the workforce comprise ‘retired’, ‘unemployed’, and ‘student’.

Categorical variables were binarized, allowing for easier interpretation and analysis by machine learning algorithms that require numerical input. In this case, we binarized ‘default’, ‘housing’, ‘loan’ and the target variable ‘y’, assigning ‘1’ for ‘yes’ and ‘0’ for ‘no’. In the same step we renamed the ‘y’ attribute to ‘TargetVariable’.

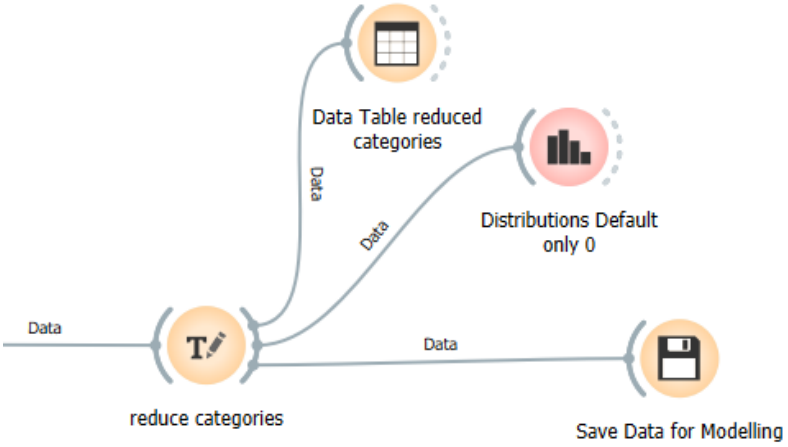

Figure 24: Extract 3/3 of Orange workflow for Data Preparation

4.6 Feature Selection

Initially we did the feature selection in the Orange workflow ‘Data Preparation FINAL.ows’, where we selected the 10 best ranked features using the widget Rank, while also considering high correlations between features. However, we found that the models work best and produce the best results when individually selecting for each model the best features. Therefore, we did the feature selection in another Orange workflow ‘Modelling FINAL.ows’.

After the step Reducing Categories / Grouping we save the dataset locally as a csv file. This file was then imported into the ‘Modelling FINAL.ows’ workflow, what can be seen in Figure 26.

After importing our dataset, we utilize the Data Sampler widget to create two samples. One sample contains 80% of the data, which will be used to train the models, while the other sample contains 20% of the data and will be used to test the models. When applying the Data Sampler widget we opt for ‘Replicable (deterministic) sampling’, to ensure consistent results across various users. Additionally, we employ the ‘Stratify sample’ option to maintain the composition of the input dataset. Afterwards, we trained the models on the training data and tested it on the test data, considering all features. With the Add-On widget Feature Importance, we were able to identify the 10 to 12 most important features for each model.

When analyzing the best features according to Feature Importance we also considered the output from the Correlations Widget, which can be seen in Figure 25.

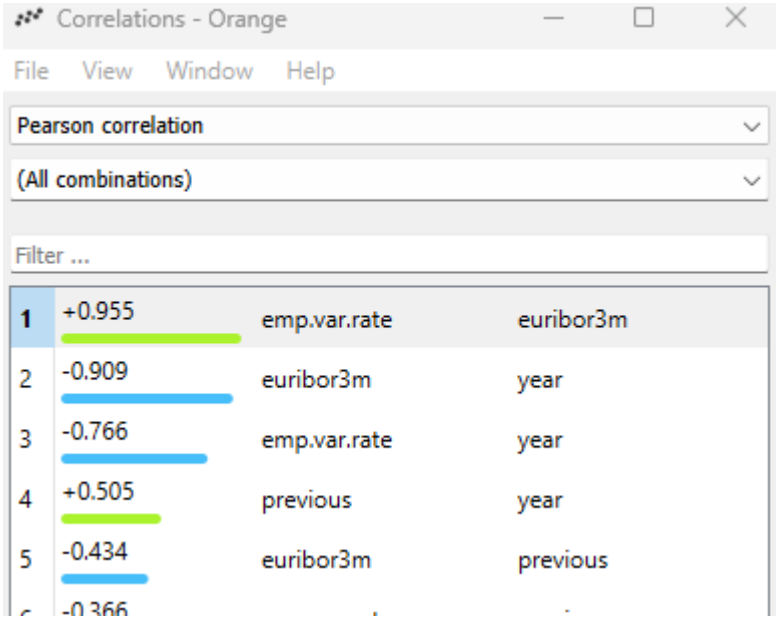

Figure 25: Correlations output (on cleaned data)

The following paragraphs explain the feature selection for each model.

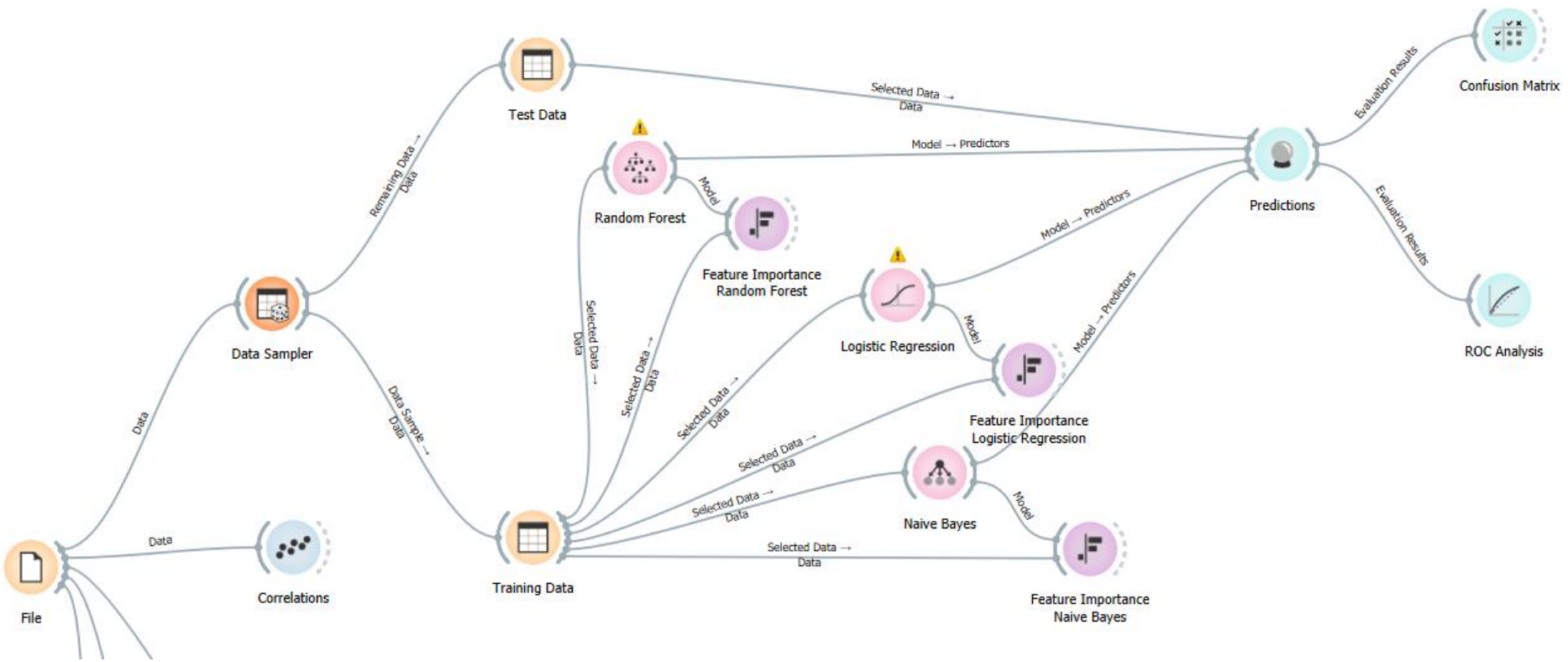

Figure 26: Extract 1/2 of Orange workflow Modelling, showing the Feature Selection

4.6.1 Random Forest

When looking at the 10 best features for Random Forest we notice that both ‘emp.var’rate’ and ‘euribor3m’ are included. As these two features have a high positive correlation of 0.955, as shown in Figure 25, we checked the 11 best features according to Feature Importance, what are the following:

1. month
2. prod_indices
3. emp.var.rate
4. euribor3m
5. education
6. age
7. poutcome
8. day_of_week
9. job
10. campaign
11. marital

After analyzing different combinations of features, always considering 9 to 10 features, we found that the following features allow the Random Forest model to perform best:

1. month
2. prod_indices
3. euribor3m
4. education
5. age
6. poutcome
7. day_of_week
8. job
9. campaign
10. marital

We did only consider ‘euribor3m’ and excluded the ‘emp.var.rate’, these 10 features are then selected in the lower part of the Orange workflow ‘Modelling FINAL.ows’, before training and testing the model, this can be seen in Figure 27.

4.6.2 Logistic Regression

When looking at the 10 best features for Logistic Regression we notice that both ‘emp.var’rate’ and ‘euribor3m’ are included. As these two features have a high positive correlation of 0.955, as shown in Figure 25, we checked the 11 best features according to Feature Importance, what are the following:

1. euribor3m
2. month
3. poutcome
4. previous
5. emp.var.rate
6. job
7. campaign
8. loan
9. marital
10. age
11. education

After analyzing different combinations of features, always considering 9 to 10 features, we found that the following features allow the Logistic Regression model to perform best:

1. euribor3m
2. month
3. poutcome
4. previous
5. job
6. campaign
7. loan
8. marital
9. age

We did only consider ‘euribor3m’ and excluded the ‘emp.var.rate’, these 9 features are then selected in the lower part of the Orange workflow ‘Modelling FINAL.ows’, before training and testing the model, this can be seen in Figure 27.

4.6.3 Naïve Bayes

When looking at the 10 best features for Naïve Bayes we notice that ‘emp.var’rate’, ‘euribor3m’ and ‘year’ are included. As ‘emp.var’rate’ and ‘euribor3m’ have a high positive correlation of 0.955 and ‘euribr3m’ and ‘year’ have a high negative correlation of -0.909, as shown in Figure 25. Additionally, we must consider that ‘year’ is not a good feature to predict on, as we cannot travel back in time- Therefore, the feature year should only be considered to understand and interpret the data better. Therefore, we checked the 12 best features according to Feature Importance, what are the following:

1. month
2. year
3. euribor3m
4. emp.var.rate
5. poutcome
6. age
7. job
8. previous
9. campaign
10. education
11. loan
12. marital

After analyzing different combinations of features, always considering 9 to 11 features, we found that the following features allow the Logistic Regression model to perform best:

1. month
2. euribor3m
3. poutcome
4. age
5. job
6. previous
7. campaign
8. education
9. loan
10. marital

We did only consider ‘euribor3m’ and excluded the ‘emp.var.rate’ as well as the ‘year’, these 10 features are then selected in the lower part of the Orange workflow ‘Modelling FINAL.ows’, before training and testing the model, this can be seen in Figure 27.

4.7 Formatting the Data

Initially we considered binning some numerical features and one-hot encoding the categorical features, but as we have learned, the models in Orange do their own data preprocessing. So, when binning or one-hot encoding is needed, the model would do that automatically. How each model does the data preparation will be described in the following paragraph.

5 Modelling

According to the CRISP-DM model the Modelling phase is about building and assessing different models. This phase has the following four tasks: 1) Select modeling techniques, 2) Generate test design, 3) Build model, and 4) Assess model.

5.1 Selection of Modeling Techniques

In selecting the modeling techniques for our classification task of predicting whether individuals will agree to a fixed term deposit, we assessed various models, such as Neural Network and SVM. Among them, Random Forest, Logistic Regression and Naïve Bayes showed the most promising performance based on our evaluation metrics. While considering all performance scores, our primary evaluation metric was classification accuracy (CA), as our target was that about 80% of the input data gets correctly predicted. Therefore, we chose to focus on these three models for further analysis and optimization.

5.2 Generation of the Test Design

Here the prepared data set from the previous stages is used. For each model we create a separate branch, as we select different features for each model, as stated above. After selecting the determined feature per model using the Select Columns widget, we use the Data Sampler widget to partition the data into training data and test data. 80% of the data is allocated to the training set, while the remaining 20% is assigned to the test set. To ensure consistent results across various users, we opt for ‘Replicable (deterministic) sampling’. Additionally, we employ the ‘stratify sample’ option to maintain the composition of the input dataset. This was applied to all branches, so for each model, as shown in Figure 27.

To ensure effective training, the larger portion is allocated to the training of the models. To test the model on the test data the widget Predictions is used. This widget is connected to the model and the test data. The Predictions widgets output are the predictions, a comparison to the true value as well as the performance scores, such as AUC (Area Under the ROC Curve), Classification Accuracy (CA), F1 Score, Precision, Recall and MCC (Matthews Correlation Coefficient).

For the evaluation of the performance of the models, we decided to use the Classification Accuracy (CA) as primary indicator for the performance, but also considering the other performance scores, as listed above. To further assess the models, we connected the ROC Analysis widget and the Confusion Matrix widget to the Predictions widget. The ROC Analysis widget produces a graph of the ROC curve. For all three models we set the Target to 1, meaning agreement to a term deposit, for the ROC curve. The Confusion Matrix widget produces for each model and its predictions on the test data a confusion matrix, indicating the number of True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). The outputs from the evaluation widgets Predictions, ROC Analysis, and Confusion Matrix are further discussed in the Assessing the Models part below.

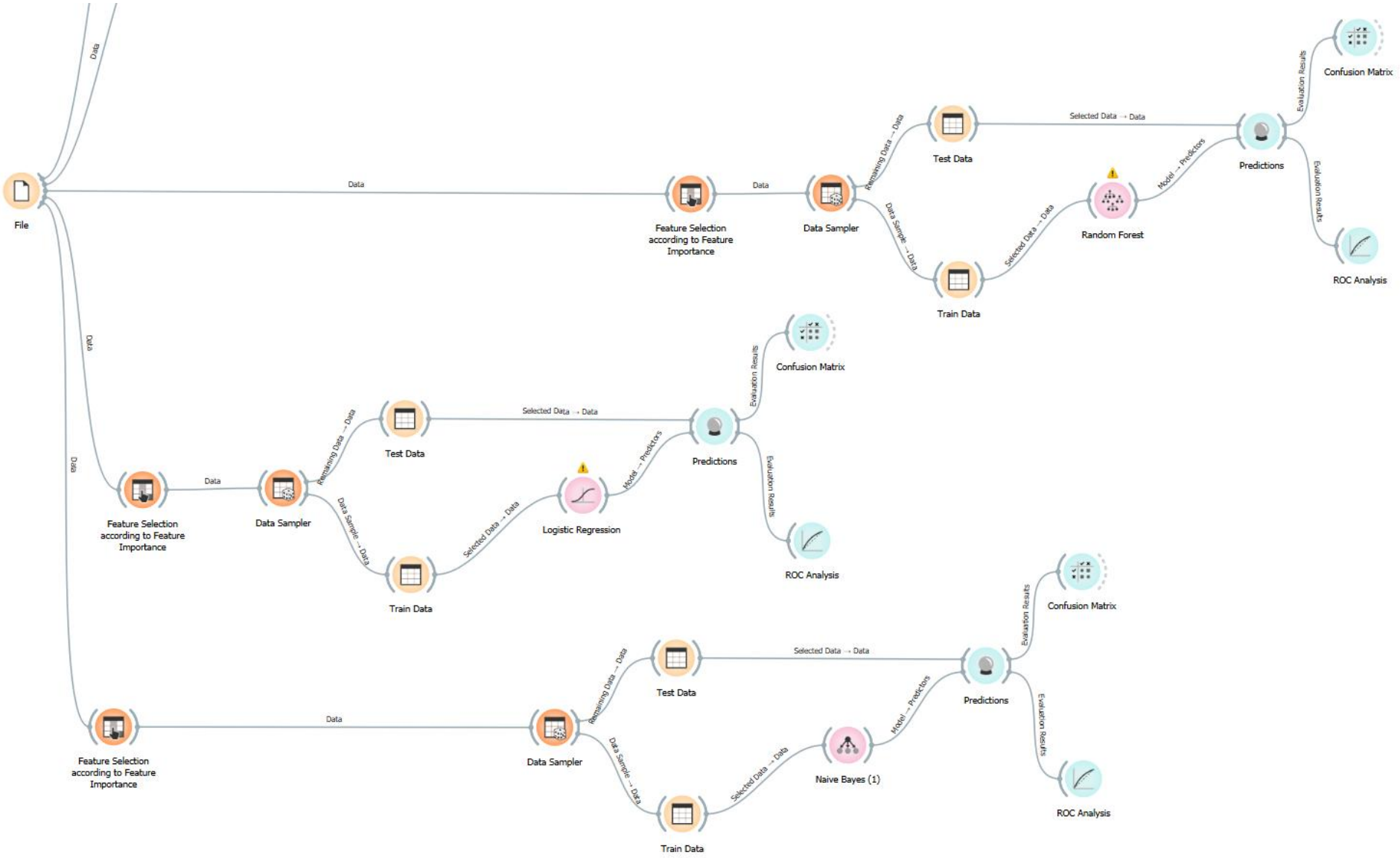

Figure 27: Extract 2/2 of Orange workflow Modelling

5.3 Explanation, Preprocessing, and Building of Models

Each model, its preprocessing, and how we built it will be further explained in the following paragraphs.

5.3.1 Random Forest

Random Forest is a powerful machine learning algorithm that constructs a collection of decision trees to make predictions. During the creation of individual trees, the Random Forest model randomly selects a subset of attributes from which the best feature is chosen for the split. This randomness helps to diversify the trees, making them less susceptible to overfitting and more robust to variations in the data. The majority decision of all trees is the base for the resulting model (*Orange Data Mining - Random Forest*, n.d.).

When no other preprocessors are specified, Random Forest uses default preprocessing: first records with unknown target class are removed, secondly categorical features get one-hot-encoded, in the third step empty columns are removed, and lastly NA values get imputed with the mean value. As we have prepared our data, there are no records with unknown target values, no empty columns, and no NA values, therefore the default preprocessing of the Random Forest model only continuizes the categorical values using one-hot-encoding (*Orange Data Mining - Random Forest*, n.d.).

The following parameters can be applied and customized for the Random Forest model.

In Orange, the number of trees can be adjusted within the widget. An increase of the number of trees to create a larger forest can improve the predictive performance and robustness. We found that 200 trees create a large enough forest, that increases the performance. Further increasing the number of trees would result in much more computational power needed than the performance would increase.

Furthermore, the number of attributes considered at each split can be changed. In general, the default is set to the square root of the total number of attributes. By increasing the number of attributes to be considered at each split to 6, an increase of the performance indexes has been visible.

For the collaboration between team members and to ensure replicability of results, the box for ‘replicable training’ was ticket to fix the seed for tree generation.

Moreover, the option Balance class distribution was ticked, to prevent the model from being biased towards the majority class. This is especially helpful when working with an imbalanced data set. Even though we balanced our data during the Data Preparation step, we ticked this option to be sure to not have any bias in the model, however this option ticked could negatively influence the computational performance.

To prevent overfitting, the limit depth of individual trees should be considered. After careful consideration of different values such as for example 10 or 20, the depth of individual trees has been limited to 10 due it is overall better performance.

Furthermore, subsets smaller than 5 will not be further split. Increasing the value can help to control the complexity of the trees and could also prevent overfitting. We chose the number 5, as this gave us the best performance.

5.3.2 Logistic Regression

Logistic Regression is a statistical algorithm only used for classification tasks. It predicts the probability that of a given input belong to a particular class (*Orange Data Mining - Logistic Regression*, n.d.).

When no other preprocessors are specified, Logistic Regression uses default preprocessing: first records with unknown target class are removed, secondly categorical features get one-hot-encoded, in the third step empty columns are removed, and lastly NA values get imputed with the mean value. As we have prepared our data, there are no records with unknown target values, no empty columns, and no NA values, therefore the default preprocessing of the Logistic Regression model only continuizes the categorical values using one-hot-encoding (*Orange Data Mining - Logistic Regression*, n.d.).

The following parameters can be applied and customized for the Logistic Regression model.

The Logistic Regression widget in Orange allows the user to choose between Lasso Regression (L1) and Ridge Regression (L2). As the model performs best with the Lasso (L1), we opt for this Regression.

Furthermore, the user can fine-tune the cost strength parameter C, which controls the strength of regularization. A smaller value of C results in stronger regularization, while a larger value of C results in weaker regularization. The default value is set to C = 1. We kept the default value, therefore applying a moderate strength of regularization to the model, as this results in the best performance of the model.

Moreover, the option Balance class distribution was ticked, to prevent the model from being biased towards the majority class. This is especially helpful when working with an imbalanced dataset. Even though we balanced our data during the Data Preparation step, we ticked this option to be sure to not have any bias in the model, however this option ticked could negatively influence the computational performance.

5.3.3 Naïve Bayes

Naïve Bayes algorithm learns a Naïve Bayesian model from a given dataset. It is based on Bayes' theorem, assuming feature independence. This classifier is especially designed for classification tasks (*Orange Data Mining - Naïve Bayes*, n.d.).

When no other preprocessors are specified, Naïve Bayes uses default preprocessing: first empty columns are removed, and then numerical features get binned into 4 bins. As we have prepared our data, there are no empty columns, therefore the default preprocessing of the Naïve Bayes model only discretizes the numerical features into 4 bins (*Orange Data Mining - Naïve Bayes*, n.d.).

For the Naïve Bayes model, no parameters can be applied and customized.

5.4 Assessing the Models

Before assessing each model, it is worth checking the different performance parameters and their interpretation.

The *ROC (Receiver Operating Characteristic)* curve is a metric used to assess the performance of a model. The y-axis represents the true positive rate (sensitivity), while the x-axis represents the false positive rate (1-specificity). If an ROC curve is very similar to the diagonal, this indicates a random process: Proximity to the diagonal means a balanced ratio between the true positive rate and the false positive rate, which is similar to the outcome of a random process. An optimal ROC curve shows an initial steep slope where the true-positive rate approaches 100% while the false-positive rate remains minimal. Subsequently, the false positive rate begins to increase. A curve that is consistently below the diagonal indicates a misinterpretation of the data (*ROC-Kurve – Wikipedia*, n.d.).

The A*rea Under the ROC curve (AUC*) is computed to assess the performance of the ROC curve. Its range is from 0 to 1, with 0.5 representing the lowest possible value. As mentioned earlier, an ROC curve near the diagonal indicates a random process, yielding an AUC of 0.5. Conversely, an AUC between 0.5 and 1 is indicative of a more optimal curve (*ROC-Kurve – Wikipedia*, n.d.).

The *Classification Accuracy (CA)* measures the proportion of correctly classified instances out of all instances in the dataset. It provides an overall measure of the model's correctness in predicting class labels. However, it may not be suitable for imbalanced datasets since it does not consider class distribution. A CA of 0 means the model's predictions are completely incorrect, respectively a CA of 1 means the model's predictions are perfect, with no errors (*Precision and Recall - Wikipedia*, n.d.).

The *F1 Score* is the harmonic mean of precision and recall. It balances between precision and recall and is especially useful when the class distribution is uneven. The F1 Score ranges from 0 to 1, where 1 is the best score indicating perfect precision and recall, and 0 is the worst score indicating poor performance (*Precision and Recall - Wikipedia*, n.d.).

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. A precision of 0 means that none of the positive predictions made by the model are correct (i.e., all positive predictions are false). And respectively a precision of 1 means that all positive predictions made by the model are correct (i.e., there are no false positives) (*Precision and Recall - Wikipedia*, n.d.).

Recall measures the proportion of true positive predictions out of all actual positive instances in the dataset. A recall of 0 means that the model fails to identify any of the actual positive instances in the dataset (i.e., all positive instances are missed). And respectively a recall of 1 means that the model correctly identifies all actual positive instances in the dataset (i.e., there are no false negatives) (*Precision and Recall - Wikipedia*, n.d.).

MCC (Matthews Correlation Coefficient) measures the correlation between the predicted and true binary classifications. It considers true positives, true negatives, false positives, and false negatives. MCC ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation (*Precision and Recall - Wikipedia*, n.d.).

5.4.1 Random Forest

Table 12 shows an overview of the performance metrics, which will be interpreted in the following paragraphs.

The AUC value of 0.810 indicates that the model's ability to distinguish between positive and negative classes is relatively good. This gets confirmed when looking at the ROC curve in Figure 28, where the curve shows an initial steep slope. In the ROC Analysis widget in Orange we considered the target 1.

The CA value of 0.762 indicates that approximately 76.2% of all instances in the dataset were correctly classified by the model, the effective numbers can be seen in the Confusion Matrix in Table 13.

The F1 score is the harmonic mean of precision and recall and with a value of 0.761, it suggests a balance between precision and recall, indicating that the model performs reasonably well in both areas. This is confirmed when looking at the values of precision and recall.

The precision value of 0.770 indicates that out of all instances predicted as positive by the model, approximately 77.0% were truly positive.

The recall value of 0.762 suggests that the model correctly identifies approximately 76.2% of all actual positive instances in the dataset.

The MCC value of 0.532 indicates a moderate level of correlation between the predicted and true binary classifications.

AUC	CA	F1	Precision	Recall	MCC
0.810	0.762	0.761	0.770	0.762	0.532

Table 12 Performance Scores of Random Forest

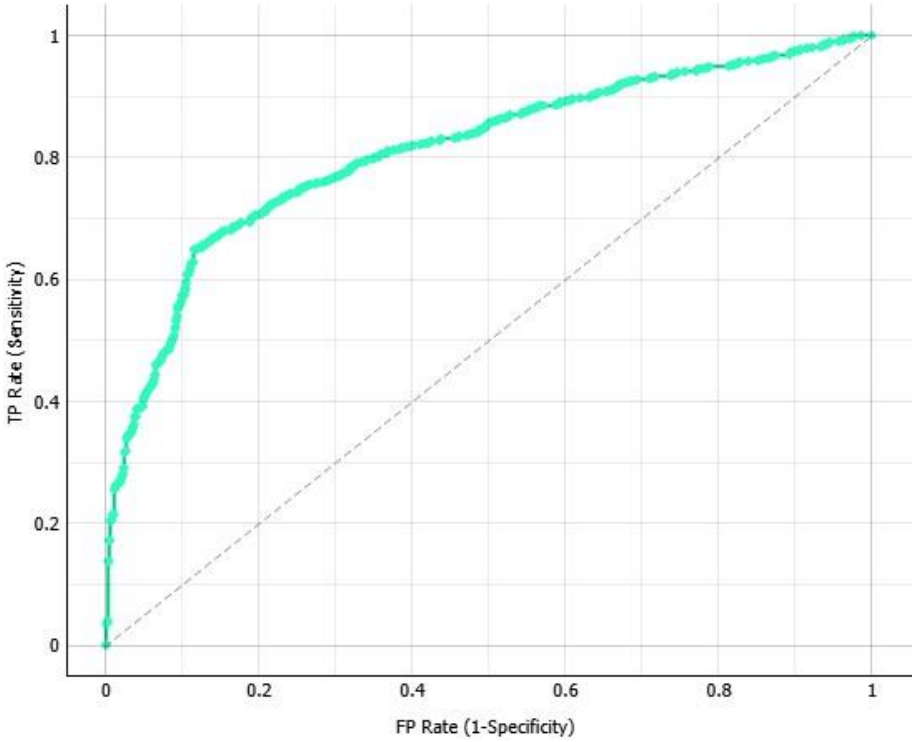

Figure 28: ROC Curve of Random Forest

		Predicted		
		0	**1**	∑
Actual	**0**	643 *(84.5%)*	118 *(15.5%)*	761
	1	244 *(32.0%)*	518 *(68.0%)*	762
	∑	887	636	1,523

Table 13 Confusion Matrix of Random Forest *(Proportion of actual)*

5.4.2 Logistic Regression

Table 14 shows an overview of the performance metrics, which will be interpreted in the following paragraphs.

The AUC value of 0.801 indicates that the model's ability to distinguish between positive and negative classes is relatively good. This gets confirmed when looking at the ROC curve in Figure 29, where the curve shows an initial steep slope. In the ROC Analysis widget in Orange we considered the target 1.

The CA value of 0.750 suggests that 75% of the instances in the dataset are correctly classified by the model, the effective numbers can be seen in the Confusion Matrix in Table 15.

The F1 score is the harmonic mean of precision and recall and with a value of 0.750, it suggests a balance between precision and recall, indicating that the model performs reasonably well in both areas. This is confirmed when looking at the values of precision and recall.

The precision value of 0.753 indicates that out of all instances predicted as positive by the model, approximately 75.3% were truly positive.

The recall value of 0.750 suggests that the model correctly identifies approximately 75% of all actual positive instances in the dataset.

The MCC value of 0.504 indicates a moderate level of correlation between the predicted and true binary classifications.

AUC	CA	F1	Precision	Recall	MCC
0.801	0.750	0.750	0.753	0.750	0.504

Table 14: Performance Scores of Logistic Regression

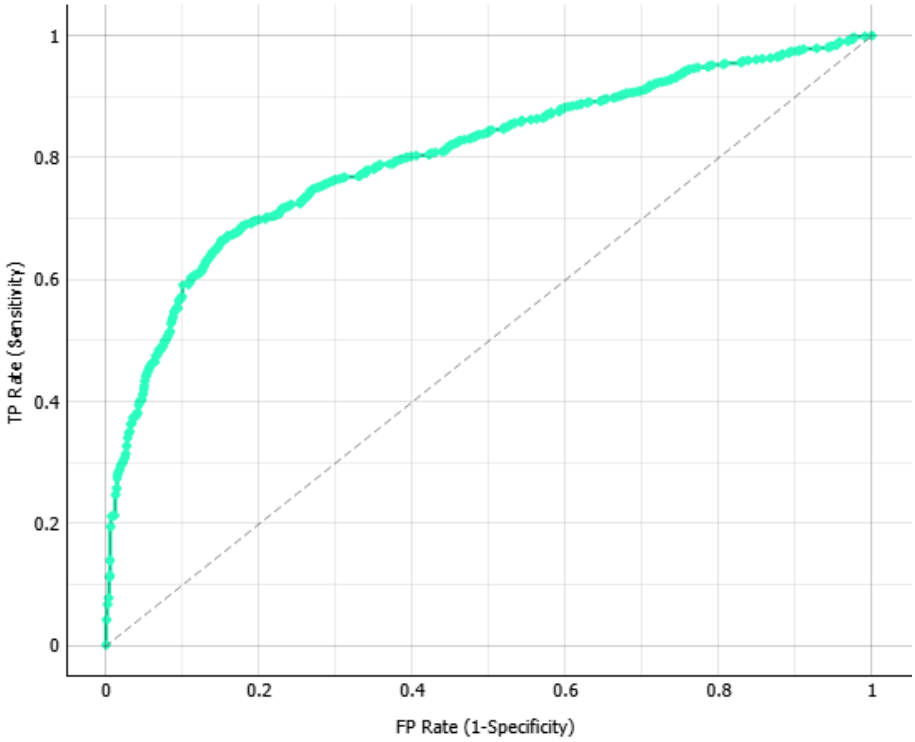

Figure 29: ROC Curve of Logistic Regression

		Predicted		
		0	**1**	∑
Actual	**0**	612 (80.4%)	149 (19.6%)	761
	1	231 (30.3%)	531 (69.7%)	762
	∑	843	680	1,523

Table 15 Confusion Matrix of Logistic Regression *(Proportion of actual)*

5.4.3 Naïve Bayes

Table 16 shows an overview of the performance metrics, which will be interpreted in the following paragraphs.

The AUC value of 0.782 indicates that the model's ability to distinguish between positive and negative classes is relatively good. This gets confirmed when looking at the ROC curve in Figure 30, where the curve shows an initial steep slope. In the ROC Analysis widget in Orange we considered the target 1.

The CA value of 0.731 suggests that 73.1% of the instances in the dataset are correctly classified by the model, the effective numbers can be seen in the Confusion Matrix in Table 17.

The F1 score is the harmonic mean of precision and recall and with a value of 0.729, it suggests a balance between precision and recall, indicating that the model performs reasonably well in both areas. This is confirmed when looking at the values of precision and recall.

The precision value of 0.742 indicates that out of all instances predicted as positive by the model, approximately 74.2% were truly positive.

The recall value of 0.731 suggests that the model correctly identifies approximately 73.1% of all actual positive instances in the dataset.

The MCC value of 0.473 indicates a moderate level of correlation between the predicted and true binary classifications.

AUC	CA	F1	Precision	Recall	MCC
0.782	0.731	0.729	0.742	0.731	0.473

Table 16: Performance Scores of Naïve Bayes

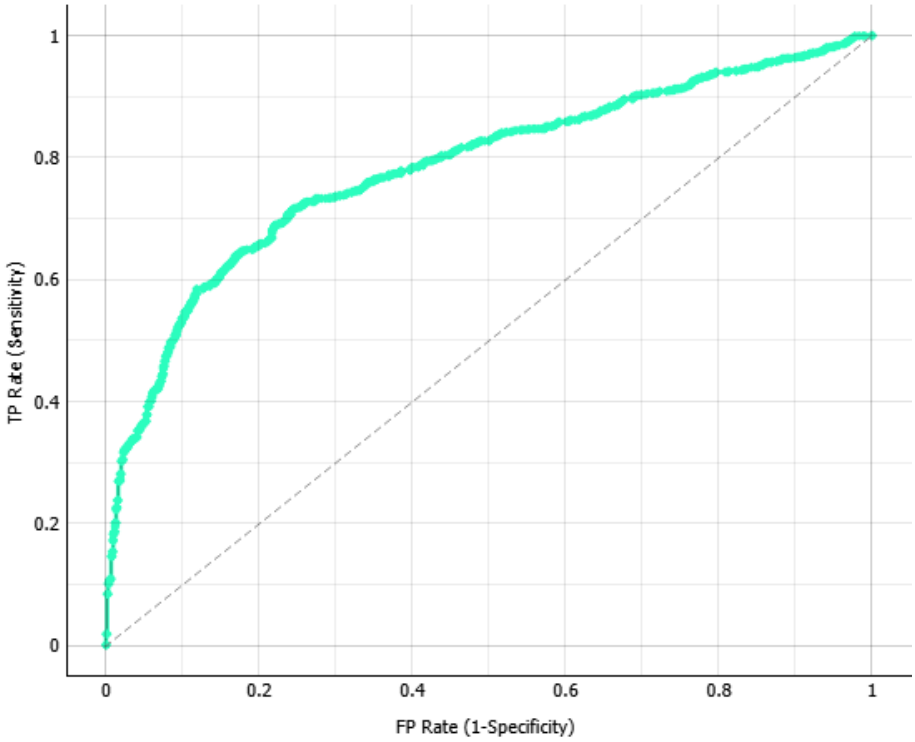

Figure 30: ROC Curve of Naïve Bayes

		Predicted		
		0	**1**	∑
Actual	**0**	636 (83.6%)	125 (16.4%)	761
	1	284 (37.3%)	478 (62.7%)	762
	∑	920	603	1,523

Table 17 Confusion Matrix of Naïve Bayes *(Proportion of actual)*

5.4.4 Model Performance Comparison

Based on Classification Accuracy (CA), the Random Forest model outperforms the others with a CA of 0.762, followed by Logistic Regression with 0.750, and Naïve Bayes with 0.731. Additionally, the Random Forest model exhibits the highest values for AUC, F1, Precision, Recall, and MCC, indicating its superior performance across all metrics. While Logistic Regression performs moderately well, Naïve Bayes lags slightly behind. Thus, considering CA as the primary metric along with other evaluation metrics, the Random Forest model emerges as the preferred choice for this classification task.

6 Evaluation

According to CRISP-DM model the Evaluation phase involves a comprehensive assessment to determine which model aligns most effectively with the business objectives and to determine the subsequent steps. This phase has 3 tasks: 1) Evaluate results, 2) Review process, and 3) Determine next steps.

6.1 Evaluating Results

As the primary business criterion is achieving an 80% prediction accuracy, we rely on the Classification Accuracy (CA) performance metric. As indicated earlier, the Random Forest model outperforms the other models with a CA of 0.762, indicating that 76.2% of the input data is predicted correctly.

Since there is no explicit mention of the costs associated with false positives and false negatives, we suggest approving the model for business use. As illustrated in Figure 2, approximately 88.73% of the recorded observations declined to subscribe to a fixed term deposit. Implementing the model is expected to enhance the conversion rate. Essentially, the model could act as a kind of pre-filter, allowing the bank to focus its efforts on individuals who are more likely to agree to a fixed term deposit, provided the required information, the required features, is available.

6.2 Process Review

Throughout applying the CRISP-DM model in an iterative manner, we frequently found it necessary to revisit previous steps and make adjustments based on new insights.

We made two adjustments in the Data Preparation process: Firstly, instead of selecting the same features for all models using the Rank widget, we opted to select the best features for each model. Therefore, we used the Feature Importance widget, while still considering high correlations between features, such as 'emp.var.rate' and 'euribor3m', or 'euribor3m' and 'year'. Secondly, as we discovered that the models apply default preprocessing, we decided against one-hot encoding categorical features and refrained from applying binarization to numerical features.

Throughout the modelling phase, we continuously revisited the outcomes of the data understanding and data preparation phases and made adjustments to the models accordingly. For instance, we observed a pronounced bias towards ‘no’ in the ‘default’ feature. If this feature would have been identified as significant by the Feature Importance analysis, we would have excluded it due to its known bias towards 'no'. By consistently monitoring the results, we were also able to consider and address high correlations, as mentioned earlier, thereby optimizing the models.

In our final review of the work conducted, we did not identify any oversights or corrections. However, regarding the feature 'campaign', additional insight into the counting and recording method would have been beneficial. This information would have been particularly useful for handling outliers and could have led to adjustments in the data preparation measures taken, as outliers are currently not considered or handled.

6.3 Determining next Steps

The Random Forest model performs as the top-performing model, although it falls slightly short of fully achieving the data mining goal of predicting approximately 80% of the data accurately. While achieving a prediction accuracy of 76.2% is commendable and very close to the target, it does not quite reach the 80% threshold. To enhance the model further, an in-depth analysis of falsely predicted values (false positives and false negatives) could reveal potential patterns. Subsequently, the model could be refined and optimized accordingly.

Despite not fully meeting the data mining objective, this model is expected to significantly enhance the success of the marketing effort of the bank by effectively selecting high-potential prospects for fixed term deposits for the campaign. As a result, the primary business objective is deemed to be achieved.

7 Findings

In reflecting on our project and its process, it is evident that we encountered a fair bit of uncertainty on how to approach the project and if we were on the right path. Especially the fact that there is not one true solution of developing a predictive model has led to the feeling of doubt. With limited prior experience in data analytics, we found ourselves revisiting certain steps multiple times to ensure we were on the right track. Reviewing on the project, it was definitely a lot of learning by doing. Throughout the project, we became much more comfortable with the tools we used, such as Orange, R, and R Studio. This familiarity and the gained understanding will undoubtedly benefit us for a future data analytics project, allowing for a quicker start.

In terms of improvement of this specific use case, we identified a couple of areas which could enhance the predictive capabilities of the model while also providing additional assets for concluding fixed term deposits. Firstly, by including client income data and analyzing transfer patterns, we could refine our predictions further. For instance, if a high amount of assets has been transferred to a different financial institution, the client may be persuaded to transfer the assets back if an attractive interest rate of the fixed term deposit is offered. Furthermore, by analyzing the income and spending habits of a client, the bank could analyze in advance if the client has free assets which could be invested in a fixed term deposit. Additionally, by collecting information of customers in advance of a marketing campaign, a predictive model could be applied directly, leading to a more successful marketing campaign from the beginning. Considering the development of the regulations in the financial sector in the previous years as well as the technological advancements, this data collection should be easily possible as of today.

In regard of the team collaboration, it was essential that each team member understands each step and stays up to date with the status of the project. Especially due to these two points, we faced challenges in ensuring each team member was equally invested and contributing the same amount of time to the project. Furthermore, a regular exchange between the members and a high degree of communication was key for the delivery of the project.

References

Attraktive Festgeld Zinsen sichern | UBS Schweiz. (n.d.). Retrieved May 4, 2024, from https://www.ubs.com/ch/de/private/investments/fixed-term-deposit.html

Beschäftigungsquote – Wikipedia. (n.d.). Retrieved May 5, 2024, from https://de.wikipedia.org/wiki/Besch%C3%A4ftigungsquote

Consumer confidence - Wikipedia. (n.d.). Retrieved May 5, 2024, from https://en.wikipedia.org/wiki/Consumer_confidence

Consumer price index - Wikipedia. (n.d.). Retrieved May 5, 2024, from https://en.wikipedia.org/wiki/Consumer_price_index

Euribor - Wikipedia. (n.d.). Retrieved May 5, 2024, from https://en.wikipedia.org/wiki/Euribor

Euribor-Werte pro Jahr. (n.d.). Retrieved May 4, 2024, from https://www.euribor-rates.eu/de/euribor-werte-pro-jahr/

Lindner, F. (n.d.). *Banken treiben Eurokrise Standard-Nutzungsbedingungen*. https://nbn-resolving.de/urn:nbn:de:101:1-2014030715981

Neubäumer, R. (2011). Eurokrise: Keine Staatsschuldenkrise, sondern Folge der Finanzkrise. *Wirtschaftsdienst*, *91*(12), 827–833. https://doi.org/10.1007/s10273-011-1308-5

Orange Data Mining - Logistic Regression. (n.d.). Retrieved May 4, 2024, from https://orangedatamining.com/widget-catalog/model/logisticregression/

Orange Data Mining - Naïve Bayes. (n.d.). Retrieved May 4, 2024, from https://orangedatamining.com/widget-catalog/model/naivebayes/

Orange Data Mining - Random Forest. (n.d.). Retrieved May 4, 2024, from https://orangedatamining.com/widget-catalog/model/randomforest/

Portugals Zentralbank warnt vor Risiken für Bankensektor | tagesschau.de. (n.d.). Retrieved May 4, 2024, from https://www.tagesschau.de/wirtschaft/eurokrise-ts-116.html

Precision and recall - Wikipedia. (n.d.). Retrieved May 4, 2024, from https://en.wikipedia.org/wiki/Precision_and_recall

ROC-Kurve – Wikipedia. (n.d.). Retrieved May 4, 2024, from https://de.wikipedia.org/wiki/ROC-Kurve

What is CRISP DM? - Data Science Process Alliance. (n.d.). Retrieved May 4, 2024, from https://www.datascience-pm.com/crisp-dm-2/

List of Figures

List of Tables

Appendix

The following documents are submitted with this documentation:

Data input/output files:

- marketing.csv
- euribor3m_ratesMAY2008toNOV2010.csv
- OUTPUT Data Preparation.csv

Quarto scripts to be run in R Studio with its associated wordfile outputs:

- DataUnderstaningwithR.qmd
- DataUnderstaningwithR.docx
- DataPreparationwithR.qmd
- DataPreparationwithR.docx

Orange workflows:

- Data Preparation FINAL.ows
- Modelling FINAL.ows