Sabancı Üniversitesi

CS 210 Project Report

Bank Marketing

Fatih Arda Zengin
28031

August 2022

Contents

1 Introduction

Abstract: The data obtained is from Kaggle and includes the marketing information of Portuguese Banking institutions.The main objective is to determine whether bank customers have opened a time deposit account (Variable y). [1]

2 Research Questions

1. Is the average age of people opening a time deposit account 40 ?
2. Is there a relationship between educational status and defaulted loans?
3. Is there a relationship between marital status and home loan?
4. What is relationship between age and housing/ personal loan?
5. What are the factors affecting time deposit accounts?

3 Data

Attribute Information:
Bank Client Data:

- **Age:** (Numeric)
- **Job** : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- **Marital**: marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- **Education**: (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- **Default**: has credit in default? (categorical: 'no', 'yes', 'unknown')
- **Housing**: has housing loan? (categorical: 'no', 'yes', 'unknown')
- **Loan:** has personal loan? (categorical: 'no', 'yes', 'unknown') Related with the last contact of the current campaign:
- **Contact**: contact communication type (categorical: 'cellular','telephone')
- **Month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- **Dayofweek**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- **Duration**: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- **Campaign:** number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **Pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **Previous:** number of contacts performed before this campaign and for this client (numeric)
- **Poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes:

- **Emp.var.rate**: employment variation rate - quarterly indicator (numeric)

- **Cons.price.idx**: consumer price index - monthly indicator (numeric)
- **Cons.conf.idx**: cons5mer confidence index - monthly indicator (numeric)
- **Euribor3m**: euribor 3 month rate - daily indicator (numeric)
- **Nr.employed**: number of employees - quarterly indicator (numeric)
- **Output variable (desired target): y** - has the client subscribed a term deposit? (binary: 'yes', 'no')

4 Data visualization and analysis

4.1 Description of Data:

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Table 1: Description of different variables [2]

As seen in the Table 1, we have many mathematical and statistical outputs of numerical variables.

4.2 Job Distribution:

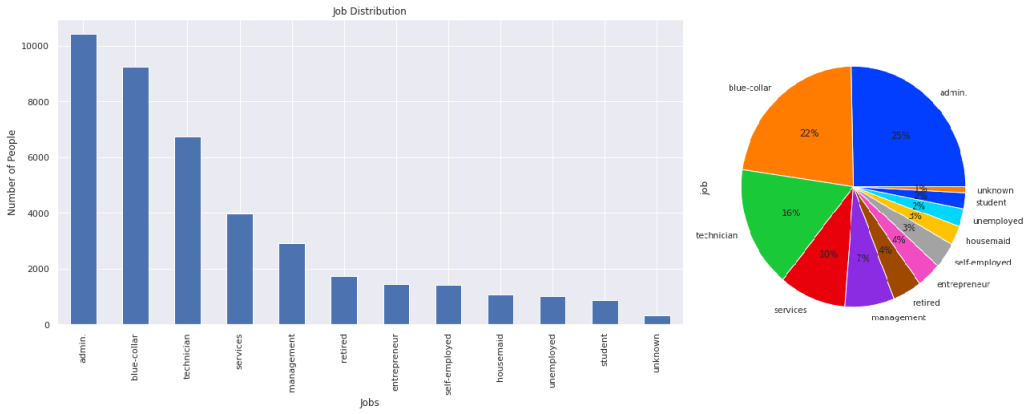

Table 2: Job distribution bar and pie charts [2]

According to our Table 2 , 25% of the employees are admin, 22% are blue collar and 16% are technicians.

4.3 Marital Distribution:

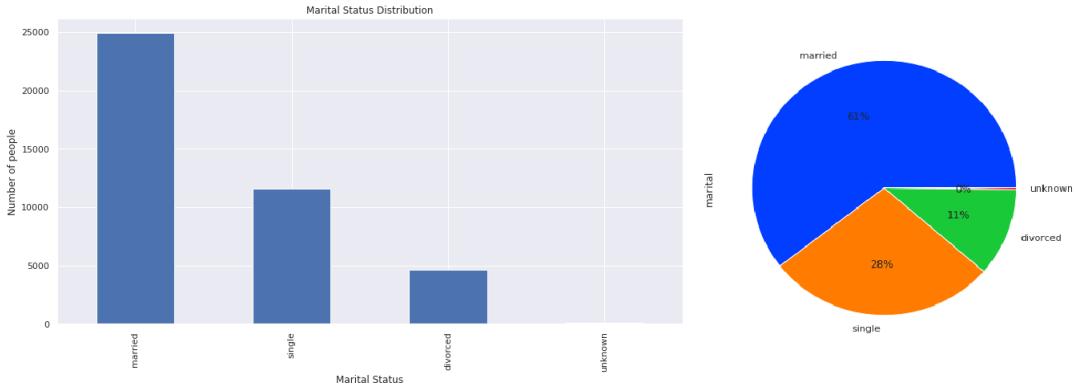

Table 3: Marital distribution bar and pie charts [2]

As can be seen from the Table 3 , while the rate of married people is 61 percent, 28 percent are single and 11 percent are divorced.

4.4 Education Distribution:

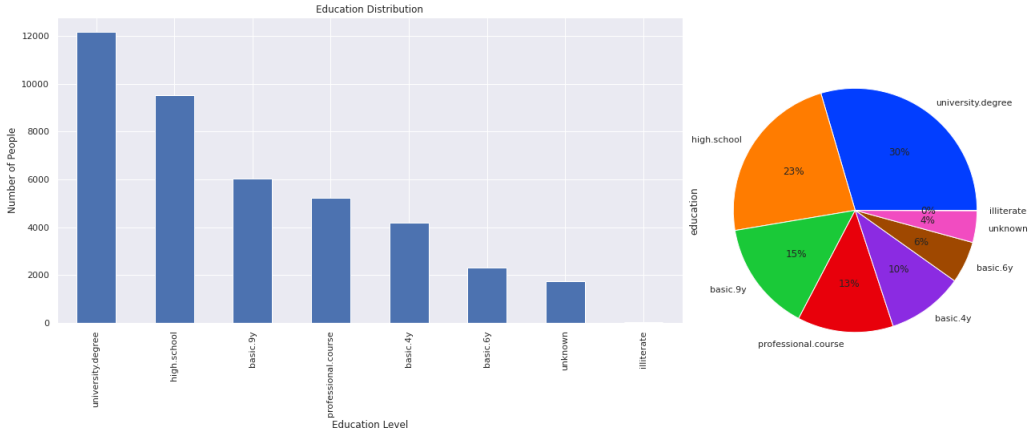

Table 4 :Education distribution bar and pie chart [2]

According to Table 4 in the education level of the participants, university graduates constitute the largest group with 30%. University graduates are followed by high school graduates with 23%.

4.5 Default Distribution:

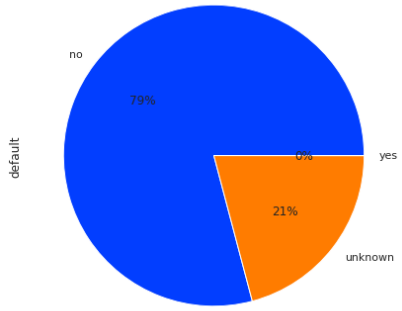

Table 5 : Default distribution pie chart [2]

According to Table 5, 79% of the participants did not default and the default status of 21% of the participants is unknown.

5 Kernel Density Estimate Plots:

5.1 Age / Default kernel density estimate plot:

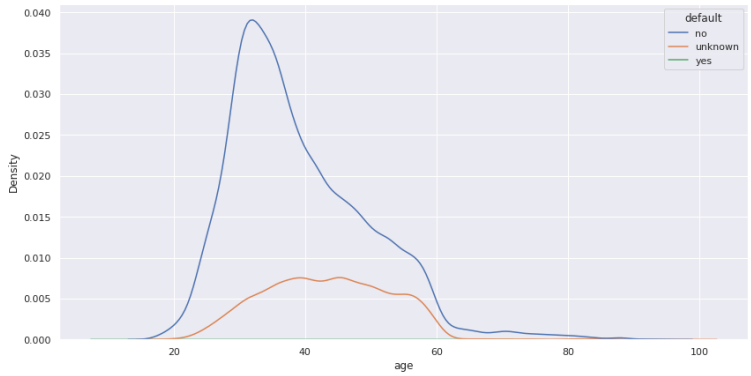

Table 6 : Age / Default kernel density estimate plot [2]

The rate of not falling into default increased the most between the ages of 20 and 40. This ratio can also be interpreted as the age range with the most loans.

5.2 Age / Housing kernel density estimate plot:

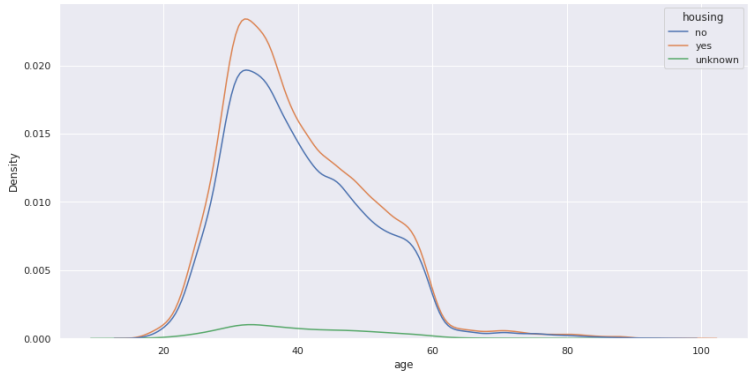

Table 7: Age / Housing kernel density estimate plot [2]

While the participants start to take out housing loans at the age of 20-25, it can be easily seen that the peak of the loan withdrawals is in the 30s. After the peak point of 30, the loan rates gradually declined, and they hit the bottom rapidly around the age of 60 according to Table 7. **This analysis also forms part of the answer to question 4.**

5.3 Age / Loan kernel density estimate plot:

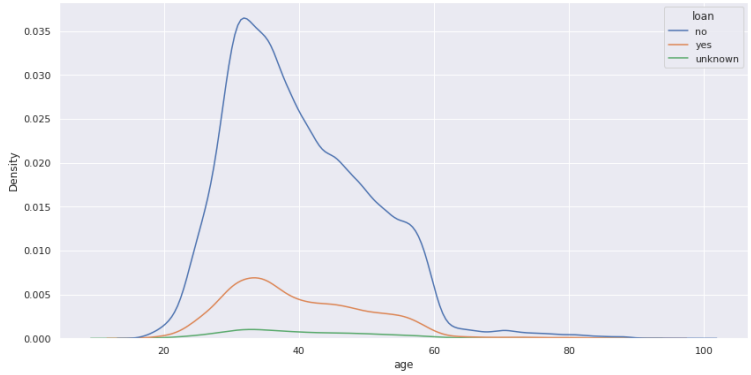

Table 8: Age / Loan kernel density estimate plot [2]

According to Table 8, although personal loans do not increase as fast as Home loans, they reach their peak at the age of 30-35. After the peak, it showed a gradual downward trend until the age of 60. At the same time,

personal loans have never been used as intensely as home loans in any age range. **This analysis also provides an answer for the remainder of question 4**.

5.4 Age / Marital kernel density estimate plot:

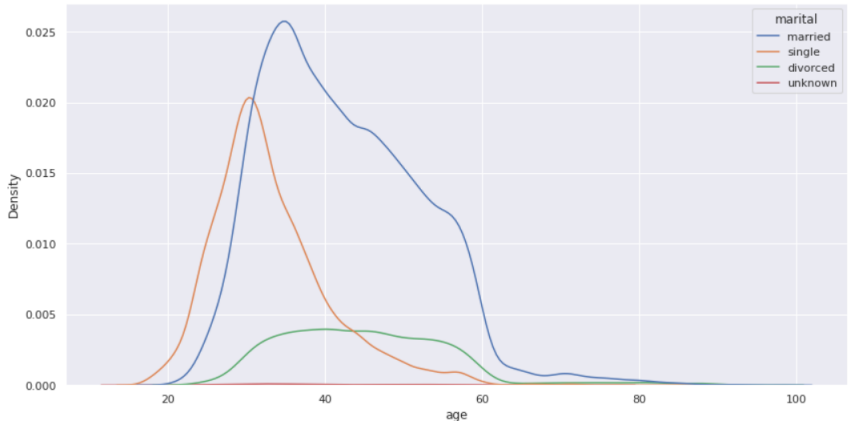

Table 9: Age / Marital kernel density estimate plot [2]

Most single respondents are in their early 30s, while most married respondents are concentrated in their late 30s. The density of divorced participants, on the other hand, increased significantly at the age of 35 and remained stable until the age of 60.

6 Age Box Plot:

6.1 Box plot grouped by marital:

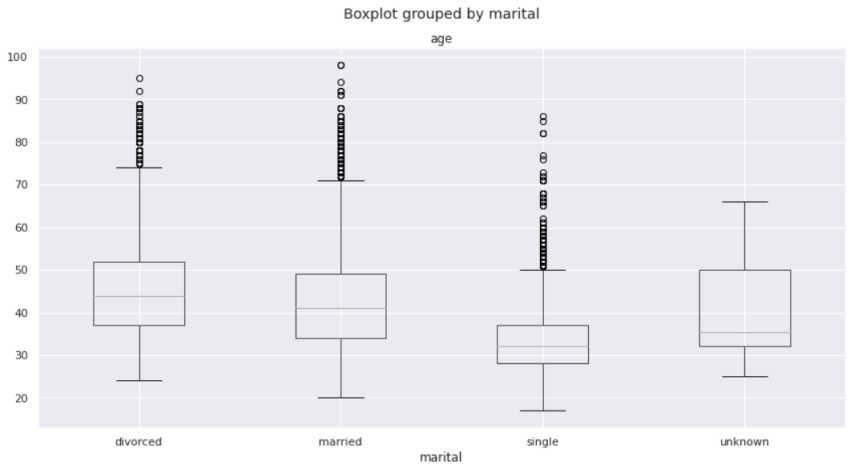

Table 10 : Box plot grouped by marital [2]

Divorced age median is about 45 and married median is about 42. Single age median is about 32 which is younger.

7 Bar Plots:

7.1 Distribution of default by education:

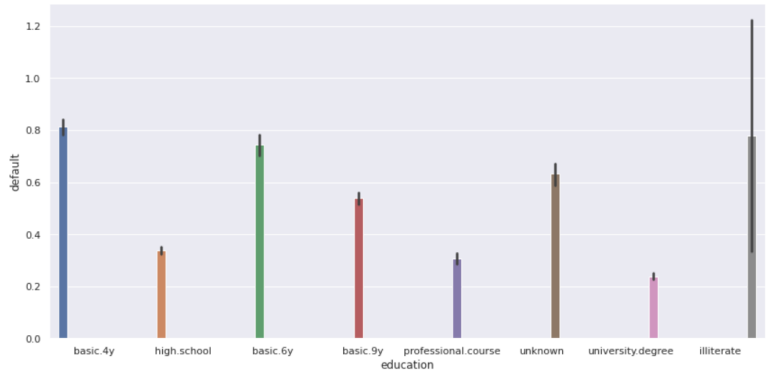

Table 11: Distribution of default by education [2]

According to Table 11, as the education level decreases, default rates increase. The people who default most are those who are illiterate, followed by basic 4-year school graduates. **This also answers question number 2**.

7.2 Distribution of housing by marital status:

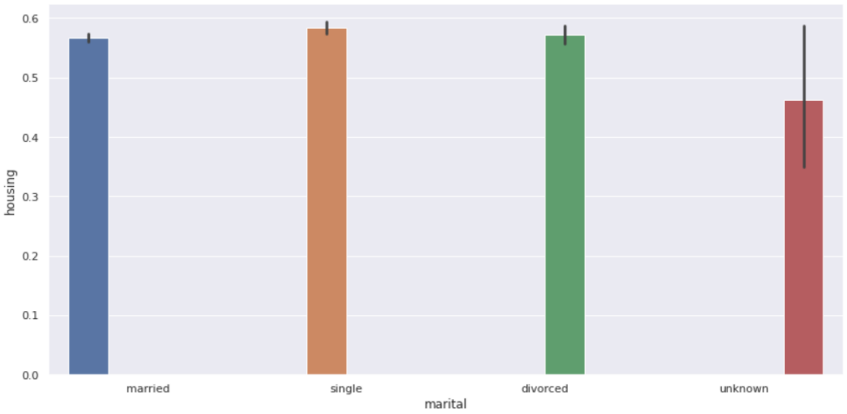

Table 12: Distribution of housing by marital status [2]

According to Table 12, people with home loans are mostly single, followed by divorced people. Married people follow divorced people with a very small margin. The first thing that comes to mind is that married people will generally have the most home loans, but this does not seem to be the case for Portugal. Perhaps the main reason for this is the increase in the habit of living alone in developed countries. **This analysis is also the answer to question 3.**

8 Statistical analysis of data:

8.1 T-Test :

In this section, we will try to **answer the 1st question** with the T-Test application.

1st Question : **Is the average age of people opening a time deposit account 40 ?**

So,
H_0 : pop. mean of age of people whose application approved = 40

H_1:pop mean of age of people whose application weren't approved !=40

According to the calculations in the notebook [2] ;

Limits are ; -1.9604754912823092 — 1.9604754912823088

Statistic: 4.49513340999622

p value: 7.1220451741420605e-06

Since p value is lower than 0.05 and statistical value is not inside our limits so, we reject H_0.
For this reason, it **cannot be said that** the average age of people who open a time deposit account is 40.

H0 : pop mean Age of people whose application approved =40,

H1: pop mean Age of people whose application werent approved !=40

+ Code + Markdown

```
data=bdata[bdata["y"]==1]
p_value=stats.ttest_1samp(data.age,40) #(data,population mean)

bottom_limit=stats.t.ppf(q=0.025,df=len(data)-1)
upper_limit=stats.t.ppf(q=0.975,df=len(data)-1)
print(f"Limits = {bottom_limit} --- {upper_limit}\nstatistic: {p_value[0]}\np value: {p_value[1]}")

if(p_value[1]>0.05):
    print("Since p value is bigger than 0.05 and statistical value is inside our limits we dont reject H0")
else:
    print("Since p value is lower than 0.05 and statistical value is not inside our limits we reject H0")
```

```
Limits = -1.9604754912823092 --- 1.9604754912823088
statistic: 4.49513340999622
p value: 7.1220451741420605e-06
Since p value is lower than 0.05 and statistical value is not inside our limits we reject H0
```

Table 13: T-test codes from the notebook [2]

8.2 Machine Learning:

In this section, the **5th question will be answered** using the *Random Forest Regressor* machine learning model.

5th Question : **What are the factors affecting time deposit accounts?**

```
from sklearn.model_selection import train_test_split

train_X, val_X, train_y, val_y = train_test_split(pd.get_dummies(bdata.drop(["y"],axis=1)), bdata["y"] ,random_state = 0)
```

+ Code + Markdown

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error

forest_model = RandomForestRegressor(random_state=31)
forest_model.fit(train_X, train_y)
preds = forest_model.predict(val_X)
print(mean_absolute_error(val_y, preds))
```

```
0.10876177527435175
```

Table 14: Machine Learning codes from notebook [2]

According to Table 14, the margin of error of the model is 0.10876177527435175, that is, **approximately 11 percent**.

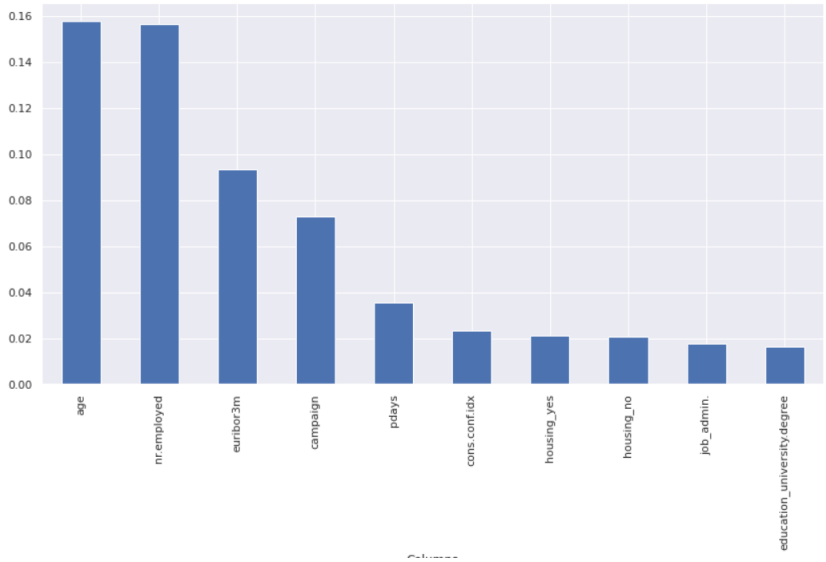

Table 15: Bar chart of the most important variables affecting time deposits according to machine learning. [2]

If we look at Table 15, the most important variable while opening a time deposit account is **age**, while the second most important variable is the **number of workers employed** in the interviewed quarter (Nr.employed), followed by the **3-month Euribor index** (Banks in the European Union take this index into account when lending to each other) appears as the the most important variables for time deposit bank accounts.

9 Discussion

In this data set, which was created with the marketing data of Portuguese bank institutions, it was seen that those who had the most profession were in contact. At the same time, most of the people who contacted stated that they were married.
When we look at the dividend rates, we observed that the most illiterate people receive dividends, followed by those with basic education of 4-6-9 years, respectively. This shows us that falling into dividends and educational status are closely related.
When we look at the home loan rates, contrary to expectations, singles mostly used home loans. Singles were followed by divorced and married, respectively. This circumstance has disproved the estimations that mostly married people take home loans.
When we look at the term account section, we see that the most important criterion is age, followed by the employment participation index in the current quarter.
The fact that the bank marketing data we have is obtained from phone calls limits the scope of the data in some respects. The fact that the information obtained from the phone interviews is not very verifiable and the participant wants to turn off the phone early are just a few of the variables that affect getting accurate data. It is obvious that if these data were collected face to face, it would be more reliable.

References

[1] Henrique Yamahata. Bank marketing dataset.
`https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing`, Jun 2018.

[2] Fatih Arda Zengin. Bank marketing notebook (in progress).
`https://www.kaggle.com/code/fatihardazengin/bank-marketing-in-progress`, Aug 2022.