German-Hellenic Bank Predictive Analysis

**Created By: Hamza Rasheed**

Contents

[**2 Section B: Predictive analytics** 3](#_Toc124121363)

[**2.1 Introduction** 3](#_Toc124121364)

[**2.2 Process** 4](#_Toc124121365)

[**2.3 Test** 4](#_Toc124121366)

[**2.3.1 Logistics regression** 4](#_Toc124121367)

[**2.3.2 ROC plot:** 6](#_Toc124121368)

[**2.3.3 Linear regression** 8](#_Toc124121369)

[**3 Section C: Prescriptive Analytics** 9](#_Toc124121370)

[**3.1 Client: German-Hellenic Bank** 9](#_Toc124121377)

[**3.1.1 Executive summary** 1](#_Toc124121378)0

[**3.1.2 Aims & Objectives** 1](#_Toc124121379)0

[**3.1.3 Analysis:** 11](#_Toc124121380)

[**3.1.4 Recommendations** 11](#_Toc124121381)

[**3.1.5 Limitations:** 11](#_Toc124121382)

# **Section A: Predictive analytics**

## **Introduction**

This part involves predictive analytics. Anonymized data was provided by the bank which involved 17 variables, but one variable “duration” was excluded. Using the data and doing the right analysis, finally, the data-driven recommendations to the marketing department of the German-Hellenic bank will be provided. The dataset has variable 17 as the desired outcome variable. Based on that the research question will be

**Q:** Who are the ideal customers for the campaign?

## **Process**

Linear and logistic regression were used to derive the results for this dataset. Firstly, the data was loaded in R. the data was then cleaned by omitting the “duration” variable and removing the null values from the data. The columns were then converted into factors. For the logistic regression, the outcome variable “y” was taken and converted into “1” and ”0” – binary variables. The “set. seed” function was then used to get the same results even later. The data was then split into train and test sets using the 80:20 rule. A full model was created by taking the “y” variable. To not overinflate the model and to get the variables that are most important to the model, stepwise modeling was applied. To check the parametric assumptions for the logistic model, multicollinearity was checked on the reduced model. BIC was then performed to see how the new reduced model is better as compared to the previous one. Along with that, pseudo-R was used to see the model performance. The exponent of log ODDS was taken to see the prediction and for the analysis. To do the predictions and get the cut-off, the ROC curve was plotted on unseen(test) data. Finally, check the sensitivity, specificity, and accuracy of the thresholds and choose the desired one.

For the linear regression, different variables were checked (which variable explains more variance) but then the age variable was taken against the rest of the variables (multiple regression). A full model was created on a train set taking age as an outcome variable. The normality and parametric assumptions were then checked using qualitative and quantitative methods. The model was plotted (qualitative method) as well. The model was then reduced to variables that are most important to the model and did not overinflate the variance. Checking the Adjusted r square and multicollinearity using the “vif” function. The NCV test was also performed. Multiple models were made and checked how much better the new model was as compared to the full model. Lastly, the predictions based on the test set were extracted and plotted.

## **Test**

### **Logistics regression**

To get the answer to the research question – who is going to take the term deposit? The “y” variable was chosen. The full model was created on a train set data and then reduced to a stepwise model. to check the parametric assumptions, multicollinearity was checked. The results were less than 5, as shown in figure 1

Text

Description automatically generated with low confidence

***Figure 18: multicollinearity results***

Multiple models were made but the final reduced model came out **9.543719e+19** times more likely to fit the data than the full model. the results of multiple models are shown in table A.



***Table A: multiple models for logistic regression and BIC***

The **pseudo-R2** value (tells which model predicts the outcome better – higher the better) was **0.165** which was quite low. The results of the exponents which captured most of the variance in the stepwise model were as

* The married client has approx**. 21.2%** less chance to subscribe to term deposit than the divorced.
* Single clients are **8.8%** more likely to subscribe to a term deposit than divorced.
* With every unit increase in balance, the person is **0.002%** more likely to open an account.
* If the customer has a housing loan, he is approx. **45.15%** less likely to subscribe to a term account.
* If the person has taken a personal loan, he is **33%** less likely to subscribe.
* If the person is contacted on the telephone, **19.5%** are less likely to subscribe as compared to cellular.
* If the client is communicated through an unknown type, he is **75%** less likely to subscribe as compared to one with cellular.
* Client contacted in august is **57.7%** less likely to subscribe than the one contacted in April.
* Person contacted in December is **63%** more likely to subscribe as compared to the one in April.
* Client contacted in Feb is **38.6%** less likely to open a term acc than the one in April.
* A person contacted in Jan is **70%** less likely to say yes than in April.
* A person contacted in July is **52.8%** less likely to say yes than the one in April.
* If contacted in June, **13%** more likely to subscribe than in April.
* If contacted in March, approx. **216%** more likely to open than the one in April.
* If contacted in May, approx. **40%** less likely to subscribe to the April one.
* Month Nov- **59%** less likely to subscribe as compared to April.
* Month Oct - **97%** more likely to open than April.
* Month Sep- **111%** more likely to subscribe than April.
* With an increase in each unit of the number of contacts performed in the campaign, there is an **8.4%** less likely chance that the client subscribes.
* Client with outcome other, there is approx. **33.6%** chance that the client will open term acc as compared to the client with the previous outcome was a failure
* Clients with which previous marketing campaign was a success, **911%** more likely to open a term acc as compared to ones with which it was a failure.
* Clients with which previous marketing campaign was an unknown, **6%** more likely to open a term acc as compared to ones with which it was a failure.

### **ROC plot:**

Different thresholds were plotted using the ROC plot curve. Taking five assumptions from 0.1 to 0.3 and then choosing the best fit out of it.

Chart, scatter chart

Description automatically generated

***Figure 19: ROC plot curve***

The thresholds were chosen using the ROC curve. Out of the 5 thresholds, 0.3 gives the best highest accuracy and balance split of sensitivity and specificity. So, 0.3 was chosen



**Threshold 0.3**

There were 319 false positives, 323 True positives, 7660 true negatives 17 false positive



### **Linear regression**

For the linear regression, a full model was created on the train set where taking the “age” variable as the outcome variable. The adjusted R squared was 0.42(42% variance is captured) which is a bit acceptable. Plotting the diagnostic plots on the reduced model, some of the parametric assumptions were violated, so the results were presented with caution. Multicollinearity results were < 5 for the reduced model which was a good sign. The value of the NCV test was significant [p < 0.05, Chi-square = 429.4993].

Multiple models were built, the best fit reduced model was **3.58E+11 times** more likely to fit the data than the full previous model.



***Table B: Multiple linear models***

The adjusted R squared for the test model came out to be **41.2%** - 41.2% of the variance in the age variable was explained by 9 variables.

Chart, line chart

Description automatically generated

***Figure 20: model significance – age***

**Diagnostic Plot diagrams (Qualitative)**

Chart, scatter chart

Description automatically generated

***Figure 21: Residuals***

Chart, scatter chart

Description automatically generated

***Figure 22***

Chart, line chart

Description automatically generated

***Figure 23: Q-Q normality***

Chart, scatter chart

Description automatically generated

# **Section A: Prescriptive Analytics**

## **Client: German-Hellenic Bank**

### **Executive summary**

Predictive analysis was done using linear and logistic modeling and some key points were identified. The data provided include 45211 customer demographics well as psychographic data Logistic regression was used to find who is likely to subscribe to the term deposit while linear regression was done for age prediction. The key findings were then analyzed, and the key profile of the client was provided to the marketing department of the bank. Using these findings will help the bank in making the right decisions.

### **Aims & Objectives**

The aim of the analysis is to examine the market for the term deposit account for the German Hellenic bank and provide recommendations to the marketing department of the bank so that better decisions can be made using the analysis which will not only save time but cost as well and effective marketing can be done based on the analyzed data.

### **Analysis:**

Different findings came out after doing the analysis which will help the bank in doing effective marketing communication.

The analysis showed that the married person has approx. 21.2% less chance to subscribe to a term deposit than the divorced while Single clients are 8.8% more likely to subscribe to a term deposit than divorced. If the customer has a housing loan, he is approx. 45.15% less likely to subscribe to a term account. If the person has taken a personal loan, he is 33% less likely to subscribe. If the person is contacted on the telephone, 19.5% are less likely to subscribe as compared to cellular. If the client is communicated through an unknown type, he is 75% less likely to subscribe as compared to one with cellular. If contacted in March, approx. 216% more likely to open than the one in April. client with which a previous marketing campaign was a success, 911% more likely to open a term acc as compared to ones with which it was a failure. Clients with whom the previous marketing campaign was unknown, were 6% more likely to open a term acc as compared to ones with which it was a failure. There was a slight increase in balance as the increase in age as shown in the figure below.

Chart, line chart

Description automatically generated

### **Recommendations**

Based on the analysis, the following recommendations are provided to the marketing department

* Single people are best to target as they are more likely to subscribe as compared to the divorced and married.
* The person who does not have opted for both – the housing and personal loan is more likely to subscribe.
* The best way to communicate is cellular as clients respond more positively through that channel.
* For the April clients, the best time is march compared to the rest of the months where there are bright chances to get clients.
* Lastly, the clients with whom the previous marketing campaign was a success are the ones who will be more attracted to the account.

### **Limitations:**

This data does not include external factors like economic conditions, government policies, market fluctuations, or political factors of the country which are also main determiners, the bank should also consider those factors while starting the campaign because those will also affect the campaign’s success.