Bank Telemarketing Analysis Report

404 Brain Not Found

Yueru He
Shangxian Liu
Yisen Wang
Zichen Xiao
Yijing Xu
Wenhan Yang

I. Introduction

Telemarketing, a direct marketing method widely used in B2C industries, has long been loathed for its effectiveness in acquiring new customers. For example, a Portugal bank conducted a few telemarketing campaigns over the 2008 - 2010 period; however, the results have not gone as well as the firm would have hoped — over the 41188 prospective customers contacted, only 4640 of them converted to successful results.

In this research, we will serve as the bank's data analysts and find a way to provide solutions for future telemarketing campaigns based on data from these campaigns as well as some additional metrics such as the consumer price index and consumer confidence index during that period. We will start by building models to explain factors that might be significant to influence a customer's purchasing probability and predict the likelihood of subscribing to a term deposit, and then dig in further about this case from what we learn. By doing so, we would like to finally optimize this bank’s marketing efforts for better managerial decision-making in the future.

II. Data Cleaning and Preprocessing

As we receive the data, there are 21 independent variables as well as 1 binary dependent variable(Appendix 1). Among these independent variables, there are 7 variables related to customer characteristics, 7 variables recording facts related to each telemarketing, and 5 macroeconomic variables capturing general information about the market.

There are no null values across all variables, however, among categorical variables, the “unknown” category is viewed as missing values, and for 1 numerical variable, “pdays”, number 999 is used to record the fact that a specific client was not previously contacted. Because among all categorical variables other than the default, missing values only take less than 5% of the total, we decide to randomly generate values to missing values using the distribution of this particular attribute. And because over 96% of the value in “pday” is missing, we drop the “pday” column. We also temporarily remove the duration variable, as this attribute highly affects the output target. violin plots for continuous variables and bar plots for discrete variables can be seen in Appendix 2. We also added a “Year” column to the dataset because we are also interested in the effect of the subprime mortgage crisis in 2008.

We did one-hot encoding to all the categorical variables and produced 37 dummy variables. Most have limited explanatory power, so we reduce the number of dummy variables by running PCA only on dummy variables(Appendix 3). From the elbow plot(Appendix 4), we decide to use the first 6 principle components(Appendix 5). We took a closer look into the loadings of each principle component we chose, and concluded the characteristics each represents in the following table:

PCs	Variables With High Loadings	Interpretation
PC 1	job_technician, education_professional.course	Variables related to high education
PC 2	marital_married	Married people
PC 3	poutcome_success	People who were contacted in the previous campaign, succeeded
PC 4	education_high.school, job_services	Variables related to medium education
PC 5	job_blue-collar, education_basic.9y	Variables related to basic education
PC 6	contact_telephone, month_jun, month_may	Variables related to time and campaign

We also found that some variables, such as age, and the number of employers have a very large range, so we scaled our continuous variables.

III. Models for Explaining and Prediction

We built two models for prediction and explanation purposes: logistic regression and random forest. We started by visualizing the correlations among independent variables (Appendix 6). We observe a high correlation between the 5 macroeconomic variables, especially among the 3 months Euribor, number of employers, and employment variation rate. Therefore we dropped 2 of these variables and left 3 months Euribor in our model.

Logistic Regression

By running the logistic regression on 5 continuous variables and 6 principal components obtained in the data preprocessing section, we achieved an accuracy of 77.93% (Appendix 8). We found that all 11 factors can statistically significantly explain whether people will buy the product or not. We constructed a calibration curve to assess how well this model can be used for probability estimation (Appendix 8). The curve shows that the true probability lies below the model score, which means our model overestimated the probability of success.

From the model summary (Appendix 7), we conclude that while older people with higher education will be more likely to buy the banking product, an increasing number of calls (indicated by variable "campaign"), increasing 3 months Euribor rate, and the number of calls to the client before this campaign negatively affect the likelihood of campaign success. One counterintuitive observation is that the 3 months Euribor rate is negatively related to the success rate. We normally expect that when the interest rate increases, people are more likely to buy banking products such as long-term savings. We will explore this issue further in section V.

Additionally, we constructed several confusion matrices under different thresholds and observed that while the model precision rate is high when the threshold is low, the recall is very low. When the threshold is 0.2, the recall is almost 50%, which is close to randomly guessing. As we gradually increased the threshold, the recall didn't improve, and the precision decreases dramatically. As the threshold increased to 0.8, we have no more true positives or false positives. We concluded that our data has too many negative outcomes, and the big gap between recall and precision might be caused by this imbalance problem.

Random Forest

We performed learned random forests, using both random forest classifiers and bagging classifiers of bootstrapping, on data before and after factor analysis, respectively. For random forest classifiers, by trying different parameters through gridsearch, we find and use the best combination of parameters to discover the most important features and predict the outcome. For the two datasets (before and after factor reduction) from which random forest classifiers learned, the most important features are both 3 months Euribor (Appendix 9), followed by consumer confidence index, another macroeconomic indicator, and other demographic variables like age and previous campaign success (embodied in PC3). However, in both graphs of feature importances, the differences between the most important factor, 3 months Euribor, and the factors following it are huge, indicating that 3 months Euribor could be deterministic in predictions.

In fact, when we tested random forest classifiers, both models (before and after factor reduction) showed a relatively high accuracy score of around 0.89. Similarly to these results, when we conducted the decision tree algorithm using bootstrapping, we obtained very high accuracy scores of around 0.8868 on both before and after PCA data.

Such high accuracy led us to suspect that the 3-month-Euribor plays an extremely powerful role in the outcome classification, because the characteristic of random forest is that if data satisfies a certain criterion, it will be classified into certain categories. However, based on the confusion matrix of bootstrapping (Appendix 10) where we discovered that both true negatives and false positives are large, it is also possible that such high accuracy scores are due to the imbalance of data. When the actual successful subscription only takes less than 10% of the whole dataset, given the same amount of true positives and false positives, the true positive rate will be greatly larger than the false positive rate, making the accuracy score (tpr/fpr) higher than normal. Therefore, we need to address this data imbalance problem.

IV. Model Enhancement

From section III, both logistic regression and random forest reflected a strong data imbalance problem to be addressed. We therefore adopt an oversampling technique called Synthetic Minority Over-sampling Technique (SMOTE) on training data. As our dataset contains both continuous and categorical variables, we used its variation–SMOTE-NC. SMOTE-NC oversamples the minority class by KNN, and undersamples the majority. Before sampling, there were 36548 majority labels, and 4640 minority labels; after sampling, there are 25587 points in both classes.

We then ran the logistic regression on the balanced data and found a really surprising result of out-of-sample AUC equals 1. Double-checking the performance of predicting probabilities by a calibration curve, we found the perfect overlapping of the curves showing the model's wellness of fit to reality. We suspect this result is caused by some intrinsic reason with the data. One assumption we have is that the data is already linear separable by a hyperplane.

To test out our belief, we built a Support Vector Machine(SVC) model as it manually creates a margined hyperplane to separate classes. By changing the penalty level, we found that oftentimes SVC gave a before and after SMOTE out-of-sample AUC of 1 (Appendix 11). There indeed exists a hyperplane, or in other words, a certain dominating predictor could separate between "yes" and "no". From the previous evidence, we propose this predictor to be "Euribor3m". If this is true, it would make a prediction so much easier because we need to only consider the Macro Economical metric, which is already publicly available.

We've observed the counterintuitive effect of “Euribor3m” with logistic regression and proved its importance. We would like to explore the Euribor rate as well as other macroeconomic factors. We expected to gain some managerial insights from these explorations in Section V.

V. Economic Analysis

The economic analysis contains two parts. We needed to first understand the contradiction between our results and the common economic theory while secondly, we needed to find the best classifier (threshold) to maximize the company's profit.

The contradiction of our results and economic theory

A widely accepted economic rule states that with higher interest rates, more savings would be made due to higher profit generated. However, the results from our logistic regression showed -0.7234 coefficient between “Euribor3m” and subscription, stating that an increase in the 3-month Europe interest rate would result in reduced subscription and fewer deposits.

The possible three reasons are illustrated below. Firstly, most people make deposits for more than a year. The interest rate they consider is the long-term interest rate that would have effects on their return, such as annual interest or a 5-year interest rate, instead of a 3-month interest rate.

Secondly, our data covered a very special period, the economic crisis in 2008, that involved several changes in government policies and customer behaviors. One policy that should be addressed here is that the European government raised the deposit insurance cap for all financial institutions. In other words, the European government strengthened protection for depositors from bank failure by increasing the compensation for customers. Such strong compensation increased consumer confidence in banks and therefore encouraged deposition.

Lastly, weakened consumer confidence resulting from the economic crisis led to changes in customer behavior. Recall that the economic crisis started with a bubbling market with an incredibly high interest rate. Consumers after the crisis would therefore feel safer saving money with lowered interest rates.

Find the best threshold to maximize profit

In order to optimize our marketing strategy, we needed to find the best threshold to make telemarketing on the selected people who have a higher chance of subscribing. In this way, we could therefore help the company maximize its payoff. To begin with, we need to draft a cost matrix. Assuming A to be the marketing cost (sum of telemarketing cost and labor cost), m to be the percentage telemarketing effectively increased subscription, and B to be profit per customer subscription, we generated the cost matrix as shown below.

	Treatment - 0 (not call)	Treatment - 1 (call)
Outcome - 0 (non-subscribe)	0	-A
Outcome - 1 (subscribe)	+B	-A + B*(1+m)

Table.1 Cost matrix

Given the telemarketing and labor cost under 2022 US to be \$75 and \$25 per hour respectively, we used the wage difference vector to scale it back to 2022 Portugal.

$$Wage\ Difference\ Vector = \frac{US\ average\ monthly\ salary}{Portugal\ average\ monthly\ salary} = \frac{7892\ USD\ per\ month}{2676\ USD\ per\ month} = 2.94 \approx 3$$

$$2022\ Portugal\ Telemarketing\ Cost = \frac{2022\ US\ telemarketing\ cost}{wage\ difference\ vector} = \frac{\$\,75\ per\ hour}{3} = \$25\ per\ hour$$

$$2022\ Portugal\ Labor\ Cost = \frac{2022\ US\ labor\ cost}{wage\ difference\ vector} = \frac{\$\,25\ per\ hour}{3} \approx \$8\ per\ hour$$

Next, we used the GDP per capita of Portugal to scale it back to 2008 level. As shown in Appendix 12, we found no large difference between the 2008 and 2021 levels. Hence, we concluded the 2008 Portugal telemarketing cost and labor cost to be \$25 and \$8 per hour respectively. The overall telemarketing cost (A) = 8+15*average duration of the calls = \$ 9.79.

Due to the lack of reference data, we estimated B and m by making two sensitivity analyses, testing the best classifier on different selected B and m. By definition, m is tested in a reasonable range between 0 and 0.5. For profit (B), we estimated it as a certain multiplier of cost. However, setting profit as 1~3x unit cost results in the same threshold, implying that the marginal revenue doesn't cover the marginal cost. Therefore, we adjusted our multiplier to be 2.5~12x unit cost and found that expected changes.

Then, we try to find the best classifier. Given that the best prediction model for classification so far is logistic regression, we will use the confusion matrix result of logistic regression before computing the total payoff. The formula for computing the payoff is below:

$$\textbf{Total payoff} = \sum \textbf{cost matrix}_{(i,j)} \times \textbf{confusion matrix}_{(i,j)}$$

Thus, since the cost matrix is already assumed, the main job to do here is iterating through every possible value of threshold between 0 and 1, using the respective confusion matrix of threshold i multiplied by the cost matrix equals the total payoff, so that the best threshold will maximize the total payoff. As seen in Appendix 13, the best payoff is \$35,000 unit profit with a threshold of 0.79, under the assumption that A=9.79, B=25, m=0.2.

Sensitivity analysis: Trying various values of inputs

Finally, the real value of variables m and B in the cost matrix in the previous part cannot be estimated directly through online sources. Thus, the solution here is conducting a sensitivity analysis by trying values of m and B to see how best threshold and payoff change. From Appendix 14, as the unit cost is \$10, the unit profit should at least exceed \$50 so that the best threshold can be below 0.5. This means only when the unit profit is large enough (>=3 times the unit cost), can the bank profit from telemarketing and have a much larger preference to give subscriptions to customers (lower value of threshold that divides the result into 0 or 1).

VI. Conclusions

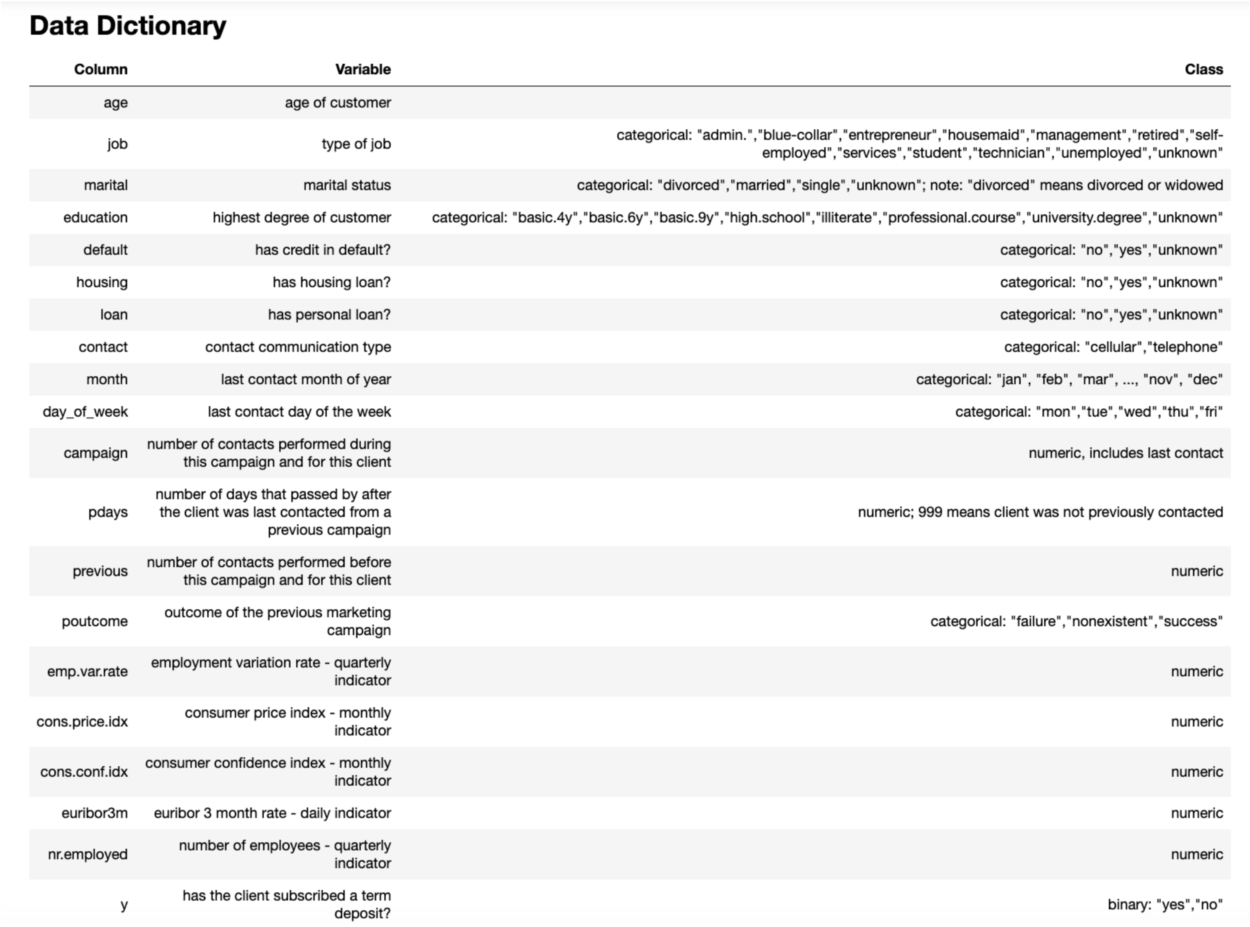
We highly agree with the current market trends that telemarketing is not a preferred perfect way to target customers and make profitable returns. Only when the profit is greater than 12 times the cost should the company be worth taking telemarketing strategies for business growth.

Bibliography

1. **S. Moro, P. Cortez and P. Rita.** A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
2. **S. Moro, R. Laureano and P. Cortez.** Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.
3. “Deposit Insurance.” FDIC, European Central Bank, 9 Oct. 2016, https://www.fdic.gov/resources/deposit-insurance/.
4. **Magellan Solutions.** “How Much Is the Cost of Telemarketing?” Magellan Solutions, 15 Sept. 2021, https://www.magellan-solutions.com/blog/cost-of-telemarketing/.
5. **Murphy, Eliza.** “Bank of America Accelerates US Minimum Hourly Wage to $22 as next Step to $25 by 2025.” Bank of America Accelerates US Minimum Hourly Wage to $22 as Next Step to $25 by 2025, Bank of America, 23 May 2022, https://newsroom.bankofamerica.com/content/newsroom/press-releases/2022/05/bank-of-america-accelerates-us-minimum-hourly-wage-to--22-as-nex.html.
6. **The World Bank.** *GDP per Capita (Current US$) - Portugal*, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=PT.
7. **Time Doctor.** “What Is the Average Salary in Portugal for 2022 ?” *Time Doctor Blog*, 10 Nov. 2022, https://www.timedoctor.com/blog/average-salary-in-portugal/.

Appendix

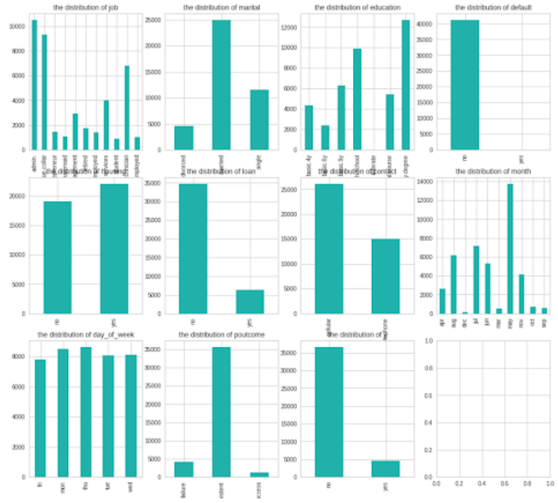
Appendix 1. Data Dictionary

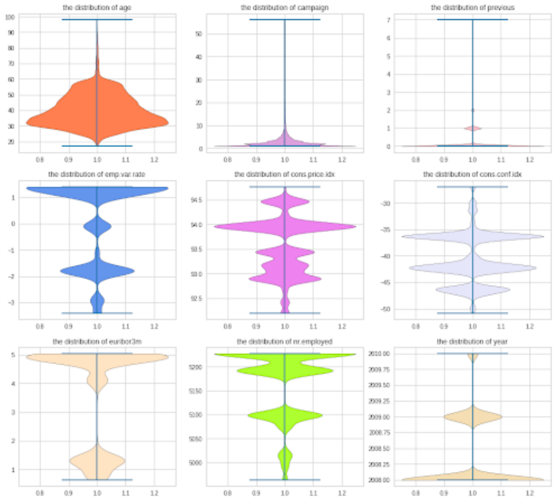


Data Dictionary

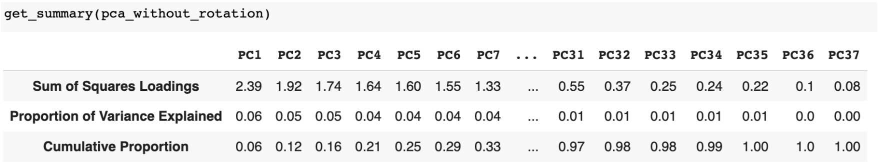
Column	Variable	Class
age	age of customer	
job	type of job	categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown"
marital	marital status	categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed
education	highest degree of customer	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"
contact	contact communication type	categorical: "cellular","telephone"
month	last contact month of year	categorical: "jan", "feb", "mar", ..., "nov", "dec"
day_of_week	last contact day of the week	categorical: "mon","tue","wed","thu","fri"
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"
emp.var.rate	employment variation rate - quarterly indicator	numeric
cons.price.idx	consumer price index - monthly indicator	numeric
cons.conf.idx	consumer confidence index - monthly indicator	numeric
euribor3m	euribor 3 month rate - daily indicator	numeric
nr.employed	number of employees - quarterly indicator	numeric
y	has the client subscribed a term deposit?	binary: "yes","no"

Appendix 2. Visualizing Categorical (Left) and Continuous Variables (Right)





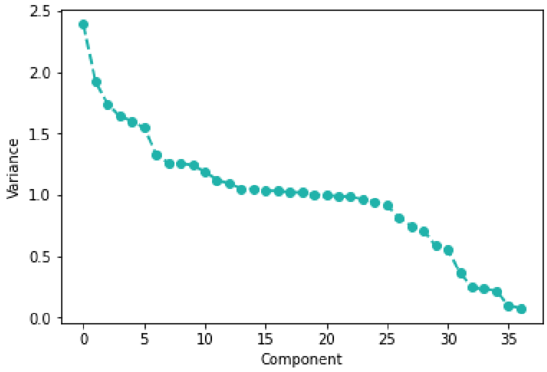
Appendix 3. PCA Results



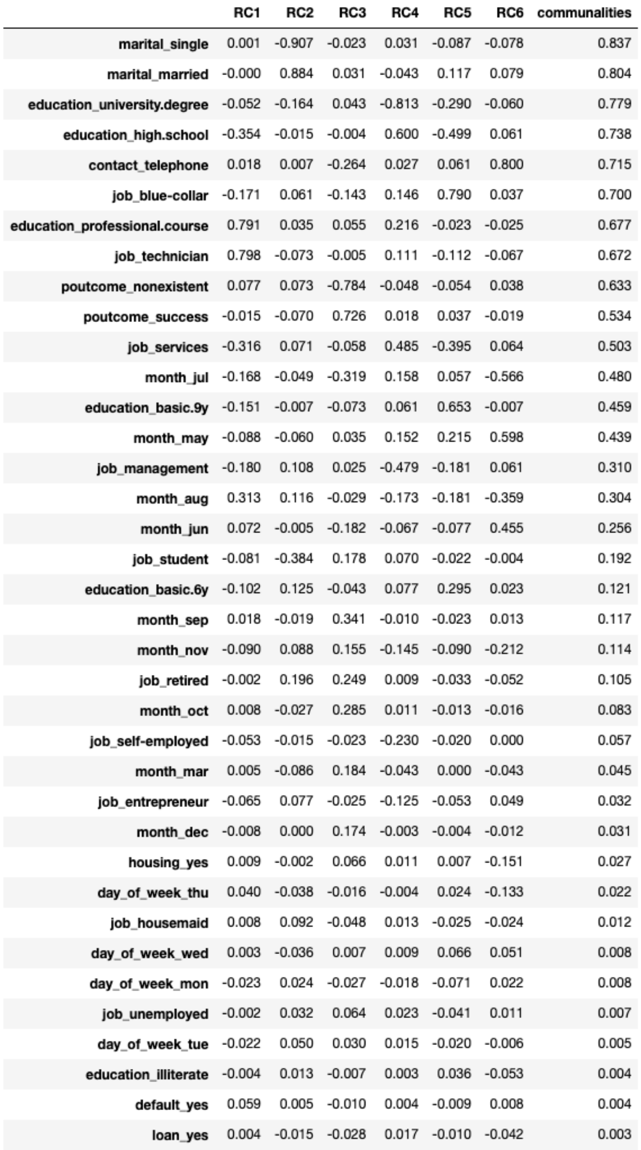
```
get_summary(pca_without_rotation)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	...	PC31	PC32	PC33	PC34	PC35	PC36	PC37
Sum of Squares Loadings	2.39	1.92	1.74	1.64	1.60	1.55	1.33	...	0.55	0.37	0.25	0.24	0.22	0.1	0.08
Proportion of Variance Explained	0.06	0.05	0.05	0.04	0.04	0.04	0.04	...	0.01	0.01	0.01	0.01	0.01	0.0	0.00
Cumulative Proportion	0.06	0.12	0.16	0.21	0.25	0.29	0.33	...	0.97	0.98	0.98	0.99	1.00	1.0	1.00

Appendix 4. Elbow plot - Variance Explained by k Components

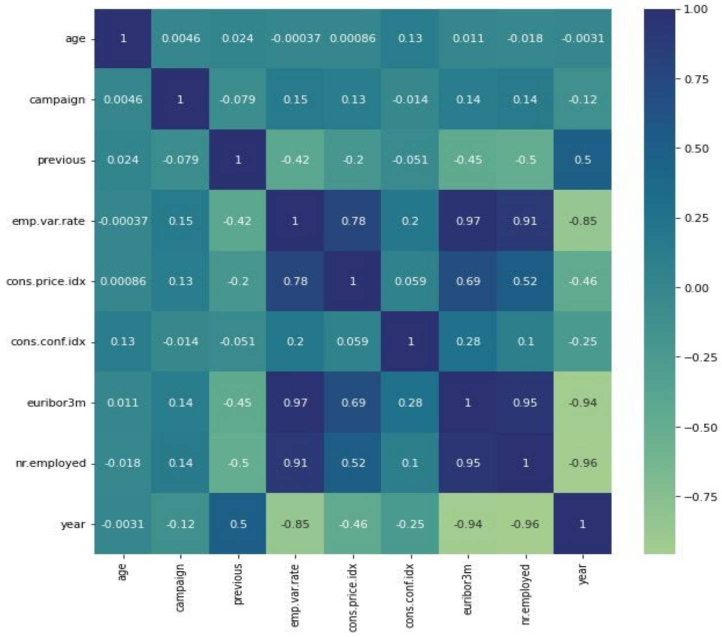


Appendix 5. Loading and Communality Table for our 6 Rotated Principal Components

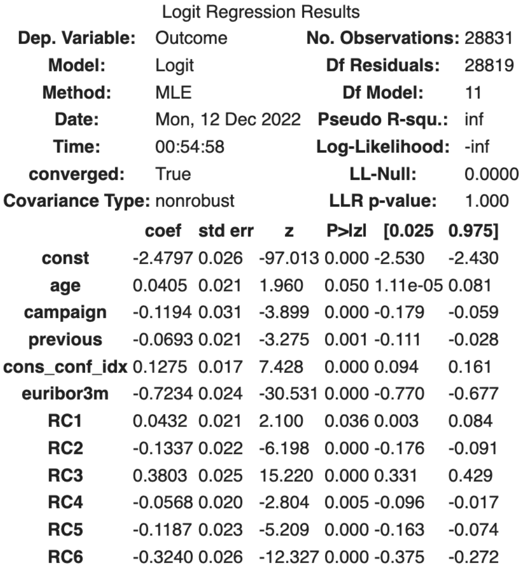


	RC1	RC2	RC3	RC4	RC5	RC6	communalities
marital_single	0.001	-0.907	-0.023	0.031	-0.087	-0.078	0.837
marital_married	-0.000	0.884	0.031	-0.043	0.117	0.079	0.804
education_university.degree	-0.052	-0.164	0.043	-0.813	-0.290	-0.060	0.779
education_high.school	-0.354	-0.015	-0.004	0.600	-0.499	0.061	0.738
contact_telephone	0.018	0.007	-0.264	0.027	0.061	0.800	0.715
job_blue-collar	-0.171	0.061	-0.143	0.146	0.790	0.037	0.700
education_professional.course	0.791	0.035	0.055	0.216	-0.023	-0.025	0.677
job_technician	0.798	-0.073	-0.005	0.111	-0.112	-0.067	0.672
poutcome_nonexistent	0.077	0.073	-0.784	-0.048	-0.054	0.038	0.633
poutcome_success	-0.015	-0.070	0.726	0.018	0.037	-0.019	0.534
job_services	-0.316	0.071	-0.058	0.485	-0.395	0.064	0.503
month_jul	-0.168	-0.049	-0.319	0.158	0.057	-0.566	0.480
education_basic.9y	-0.151	-0.007	-0.073	0.061	0.653	-0.007	0.459
month_may	-0.088	-0.060	0.035	0.152	0.215	0.598	0.439
job_management	-0.180	0.108	0.025	-0.479	-0.181	0.061	0.310
month_aug	0.313	0.116	-0.029	-0.173	-0.181	-0.359	0.304
month_jun	0.072	-0.005	-0.182	-0.067	-0.077	0.455	0.256
job_student	-0.081	-0.384	0.178	0.070	-0.022	-0.004	0.192
education_basic.6y	-0.102	0.125	-0.043	0.077	0.295	0.023	0.121
month_sep	0.018	-0.019	0.341	-0.010	-0.023	0.013	0.117
month_nov	-0.090	0.088	0.155	-0.145	-0.090	-0.212	0.114
job_retired	-0.002	0.196	0.249	0.009	-0.033	-0.052	0.105
month_oct	0.008	-0.027	0.285	0.011	-0.013	-0.016	0.083
job_self-employed	-0.053	-0.015	-0.023	-0.230	-0.020	0.000	0.057
month_mar	0.005	-0.086	0.184	-0.043	0.000	-0.043	0.045
job_entrepreneur	-0.065	0.077	-0.025	-0.125	-0.053	0.049	0.032
month_dec	-0.008	0.000	0.174	-0.003	-0.004	-0.012	0.031
housing_yes	0.009	-0.002	0.066	0.011	0.007	-0.151	0.027
day_of_week_thu	0.040	-0.038	-0.016	-0.004	0.024	-0.133	0.022
job_housemaid	0.008	0.092	-0.048	0.013	-0.025	-0.024	0.012
day_of_week_wed	0.003	-0.036	0.007	0.009	0.066	0.051	0.008
day_of_week_mon	-0.023	0.024	-0.027	-0.018	-0.071	0.022	0.008
job_unemployed	-0.002	0.032	0.064	0.023	-0.041	0.011	0.007
day_of_week_tue	-0.022	0.050	0.030	0.015	-0.020	-0.006	0.005
education_illiterate	-0.004	0.013	-0.007	0.003	0.036	-0.053	0.004
default_yes	0.059	0.005	-0.010	0.004	-0.009	0.008	0.004
loan_yes	0.004	-0.015	-0.028	0.017	-0.010	-0.042	0.003

Appendix 6. Variable Correlation Heatmap

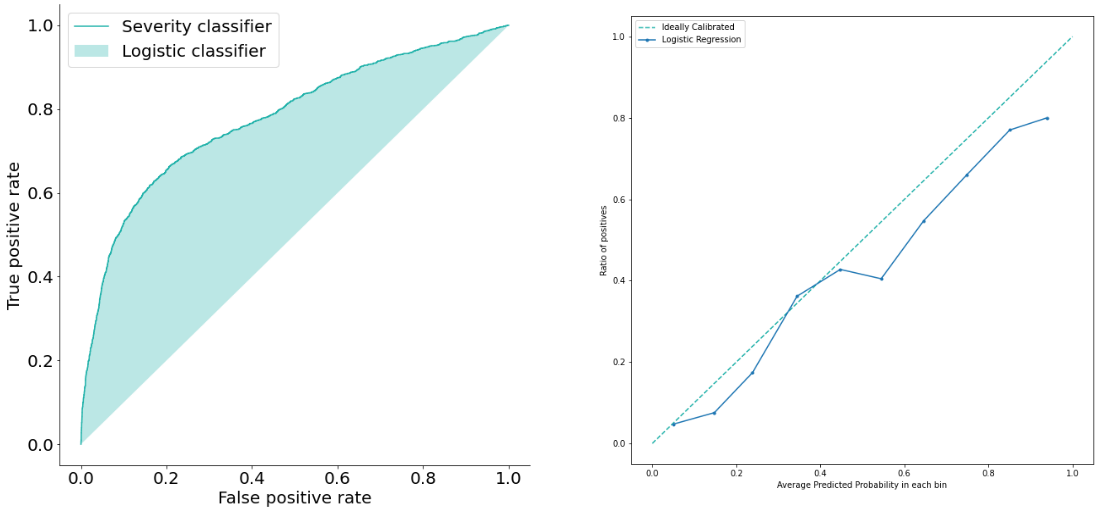


Appendix 7. Logistic Regression Model Summary

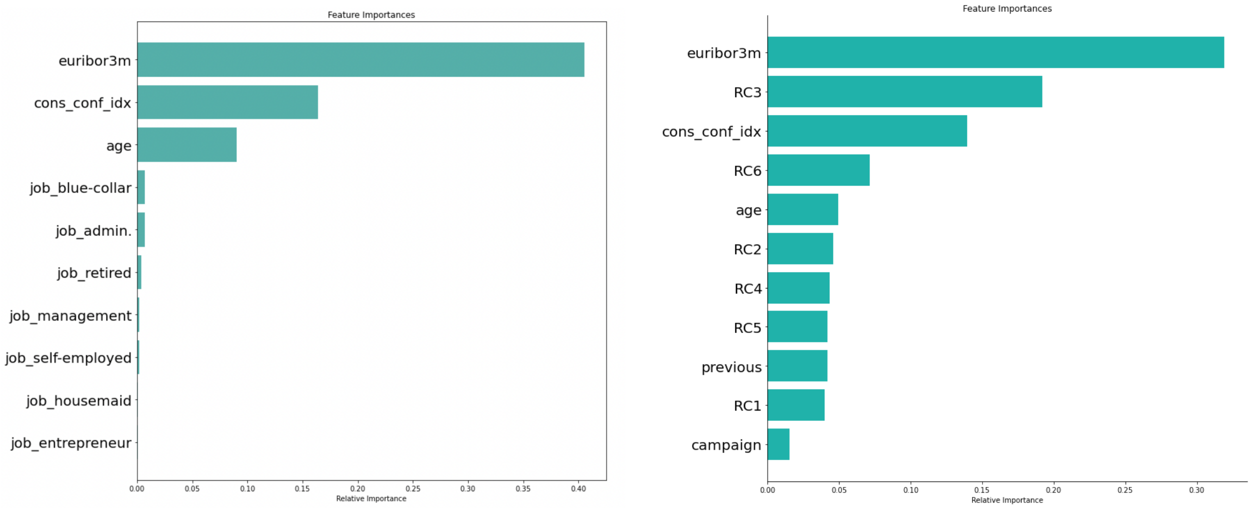


```
                    Logit Regression Results
  Dep. Variable:  Outcome            No. Observations: 28831
         Model:   Logit                  Df Residuals: 28819
        Method:   MLE                        Df Model: 11
          Date:   Mon, 12 Dec 2022    Pseudo R-squ.:   inf
          Time:   00:54:58           Log-Likelihood:   -inf
     converged:   True                      LL-Null:   0.0000
Covariance Type: nonrobust               LLR p-value:  1.000
                 coef    std err   z      P>|z|  [0.025   0.975]
       const    -2.4797 0.026  -97.013  0.000 -2.530    -2.430
         age    0.0405  0.021  1.960    0.050 1.11e-05 0.081
    campaign    -0.1194 0.031  -3.899   0.000 -0.179    -0.059
    previous    -0.0693 0.021  -3.275   0.001 -0.111    -0.028
cons_conf_idx 0.1275  0.017  7.428    0.000 0.094     0.161
   euribor3m    -0.7234 0.024  -30.531  0.000 -0.770    -0.677
         RC1    0.0432  0.021  2.100    0.036 0.003     0.084
         RC2    -0.1337 0.022  -6.198   0.000 -0.176    -0.091
         RC3    0.3803  0.025  15.220   0.000 0.331     0.429
         RC4    -0.0568 0.020  -2.804   0.005 -0.096    -0.017
         RC5    -0.1187 0.023  -5.209   0.000 -0.163    -0.074
         RC6    -0.3240 0.026  -12.327  0.000 -0.375    -0.272
```

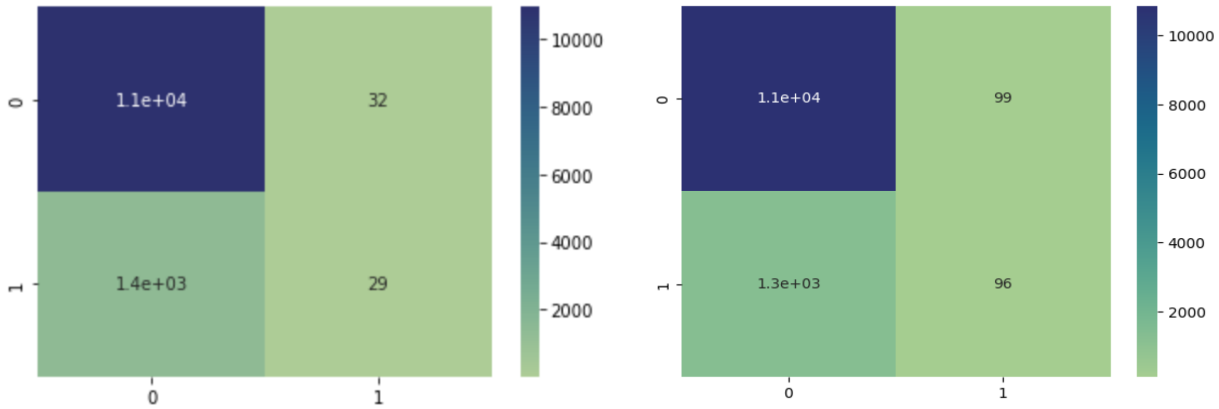
Appendix 8. Logistic Mode Out-of-Sample AUC (Left) vs. Calibration Curve (Right)



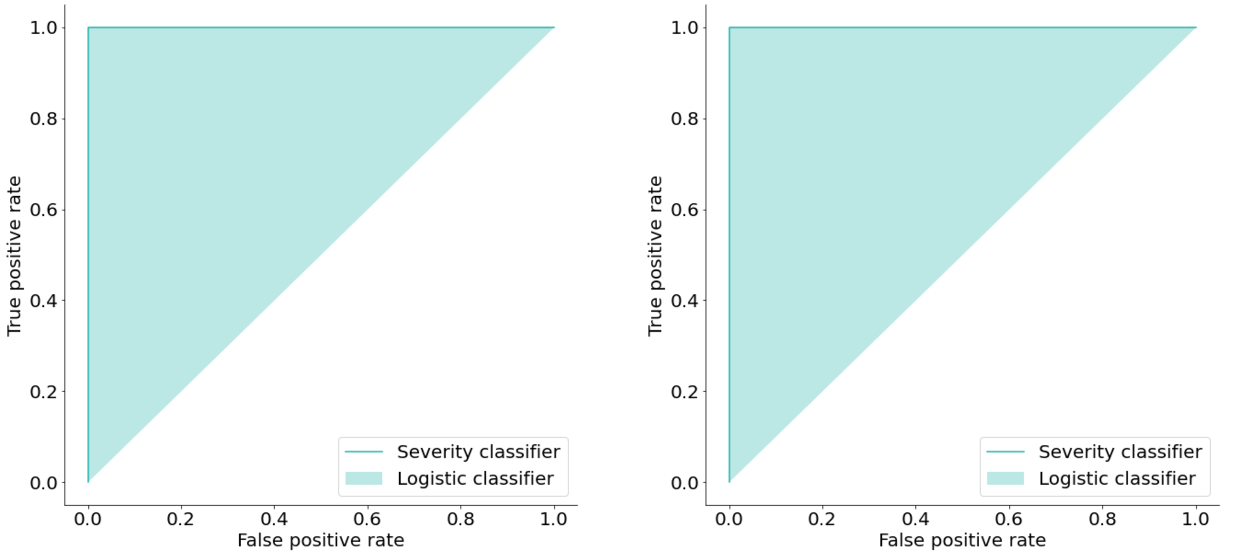
Appendix 9. Random Forest Feature Importance Before (Left) vs. After (Right) PCA



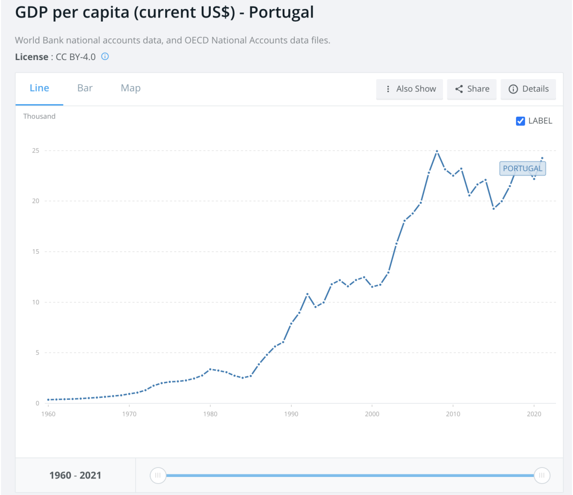
Appendix 10. Random Forest Confusion Matrix Before (Left) vs. After (Right) PCA with Bootstrapping



Appendix 11. SVC ROC Before (Left) vs. After (Right) SMOTE-NC



Appendix 12. Line Graph of Portugal GDP Per Capita Through 1960 to 2021

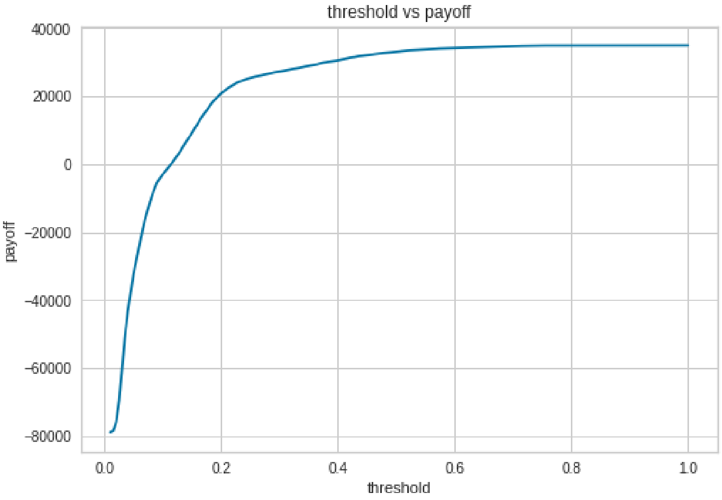


GDP per capita (current US$) - Portugal

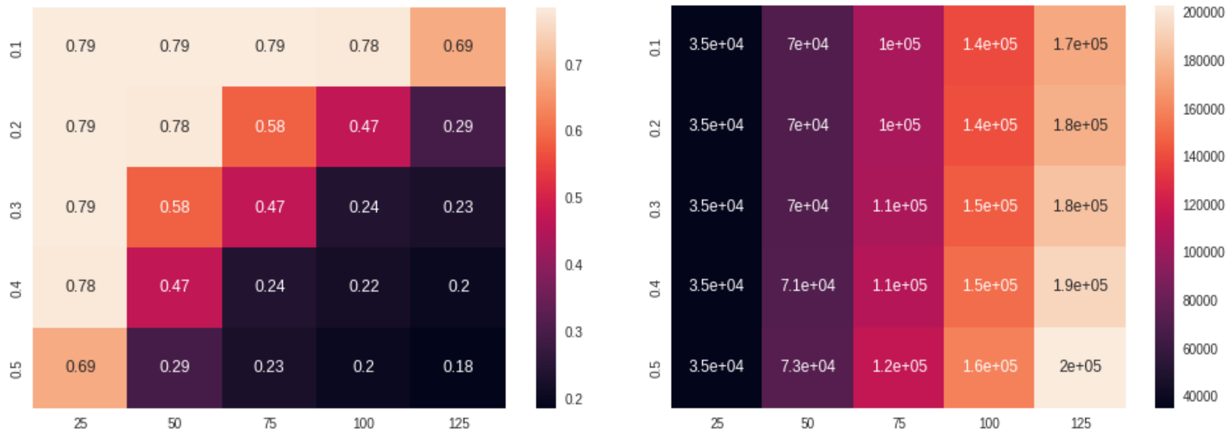
World Bank national accounts data, and OECD National Accounts data files.

License : CC BY-4.0

Appendix 13. Finding the Best Classifier Using Logistic Regression



Appendix 14. Sensitivity Analysis of Finding the Best Classifier: Threshold Result (Left) and Pay-off Result (Right)



	25	50	75	100	125
0.1	0.79	0.79	0.79	0.78	0.69
0.2	0.79	0.78	0.58	0.47	0.29
0.3	0.79	0.58	0.47	0.24	0.23
0.4	0.78	0.47	0.24	0.22	0.2
0.5	0.69	0.29	0.23	0.2	0.18

	25	50	75	100	125
0.1	3.5e+04	7e+04	1e+05	1.4e+05	1.7e+05
0.2	3.5e+04	7e+04	1e+05	1.4e+05	1.8e+05
0.3	3.5e+04	7e+04	1.1e+05	1.5e+05	1.8e+05
0.4	3.5e+04	7.1e+04	1.1e+05	1.5e+05	1.9e+05
0.5	3.5e+04	7.3e+04	1.2e+05	1.6e+05	2e+05