Springboard Foundations of Data Science - Capstone Final Report

Susan Joseph

November 23, 2017

## Introduction

An anonymous bank based in Spain has published 1.5 years of customer behavior data that consists of customer personal details (Customer Personal Details.csv) and the products of the bank they have subscribed to (Customer Products.csv). The different products that the customers subscribed to are Current Account, Particular Account, Direct Debit, E-account, Payroll Account, Taxes , Credit Cards, Pensions, Securities, Funds and Long Term Deposits. The major Customer attributes that are of interest are the Age, household income, the channel used by the customer to subscribe to a product and Segmentation 01 - VIP, 02 - Individuals 03 - college graduated. The goal of the project is to determine if the Customer attributes mentioned have any correlation to the different bank products so that the ###marketing department of the Bank could target new customers based on the analysis.

## Approach

The capstone project is approached by first identifying the problem statement. Next, the most important csv files are identified. Data wrangling is done to tidy the data. After this imputations , exploratory data analysis and some basic statistical correlations are performed. Finally, regressions and machine learning algorithms are used on the dataset.

## Problem Statement

The goal of the project is to determine if there are any correlations between the Customer attributes and the correlation to the different bank products so that the marketing department of the Bank could target new customers. The problem is important to the Bank in Spain to find out what customer attributes such as Age, Income and Segment would be responsible in influencing the subscription ###of the bank product by the customer.

## Data

The Bank has provided two .CSV files Customer Personal Details.csv and Customer Products.csv. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc. These products are the columns named: ind\_(xyz)\_ult1 etc. There are 1048576 rows and 24 observations in the Customer Peronal Details.csv file and 1048576 rows and 24 ###observations in the Customer Products.csv file.

## Structure of the Customer and Product files

## 'data.frame': 1183036 obs. of 24 variables:  
## $ ncodpers : int 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 ...  
## $ fecha\_dato : chr "2015-01-28" "2015-02-28" "2015-03-28" "2015-04-28" ...  
## $ ind\_empleado : chr "N" "N" "N" "N" ...  
## $ pais\_residencia : chr "ES" "ES" "ES" "ES" ...  
## $ sexo : chr "V" "V" "V" "V" ...  
## $ age : chr "23" "23" "23" "23" ...  
## $ fecha\_alta : chr "2011-12-29" "2011-12-29" "2011-12-29" "2011-12-29" ...  
## $ ind\_nuevo : chr "0" "0" "0" "0" ...  
## $ antiguedad : chr "43" "43" "43" "43" ...  
## $ indrel : chr "1" "1" "1" "1" ...  
## $ ult\_fec\_cli\_1t : chr NA NA NA NA ...  
## $ indrel\_1mes : chr "1.0" "1.0" "1.0" "1.0" ...  
## $ tiprel\_1mes : chr "A" "A" "A" "A" ...  
## $ indresi : chr "S" "S" "S" "S" ...  
## $ indext : chr "N" "N" "N" "N" ...  
## $ conyuemp : chr NA NA NA NA ...  
## $ canal\_entrada : chr "KHE" "KHE" "KHE" "KHE" ...  
## $ indfall : chr "N" "N" "N" "N" ...  
## $ tipodom : chr "1" "1" "1" "1" ...  
## $ cod\_prov : chr "52" "52" "52" "52" ...  
## $ nomprov : chr "MELILLA" "MELILLA" "MELILLA" "MELILLA" ...  
## $ ind\_actividad\_cliente: chr "1" "1" "1" "1" ...  
## $ renta : num 108623 108623 108623 108623 108623 ...  
## $ segmento : chr "03 - UNIVERSITARIO" "03 - UNIVERSITARIO" "03 - UNIVERSITARIO" "03 - UNIVERSITARIO" ...

## 'data.frame': 1183036 obs. of 26 variables:  
## $ ncodpers : int 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 1000013 ...  
## $ fecha\_dato : chr "2015-01-28" "2015-02-28" "2015-03-28" "2015-04-28" ...  
## $ ind\_ahor\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_aval\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_cco\_fin\_ult1 : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ ind\_cder\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_cno\_fin\_ult1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_ctju\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_ctma\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_ctop\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_ctpp\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_deco\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_deme\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_dela\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_ecue\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_fond\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_hip\_fin\_ult1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_plan\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_pres\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_reca\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_tjcr\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_valo\_fin\_ult1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_viv\_fin\_ult1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ind\_nomina\_ult1 : chr "0" "0" "0" "0" ...  
## $ ind\_nom\_pens\_ult1: chr "0" "0" "0" "0" ...  
## $ ind\_recibo\_ult1 : int 0 0 0 0 0 0 0 0 0 0 ...

The first step was to convert the columns to the correct data types.

## Data Wrangling

Data wrangling was performed on both the Customer and Product data frames.

There are 17 rows for one Customer where every row is a month of data of a customer. The below ouput displays the number of rows for each customer. The rows are summarized so there is only one row per customer and columns are eliminated that do not add value to the analysis. There are 78,359 customers after the summarization. The column canal\_entrada (Channel) was replaced with 'Other' values where count of observations is less than 200. Hence all the categories in the Channel column have more than 200 elements.

The product data frame had the same situation as in Customer where there are many product rows for each Customer. So a simplification was made that either a customer has or does not have a product. The rows where the sum of products was zero was eliminated for each customer.

Finally both the Customer and product data frames 78359 observations and 25 variables and the data frames were joined thereby having a single record for each Customer.

Customer

## # A tibble: 78,359 x 2  
## # Groups: n [17]  
## ncodpers n  
## <int> <int>  
## 1 15895 17  
## 2 15918 17  
## 3 15937 17  
## 4 15941 17  
## 5 15952 17  
## 6 15957 17  
## 7 15964 17  
## 8 15969 17  
## 9 15976 17  
## 10 15997 17  
## # ... with 78,349 more rows

Product

##   
## KAA KAB KAE KAG KAS KAT KAY KBZ KCC KEY KFA KFC   
## 233 300 208 305 444 16112 355 241 206 241 1836 14556   
## KHD KHE KHK KHL KHM KHN OTHER RED   
## 540 23189 1728 346 399 406 2104 477

## Finding out the number of NA's by column

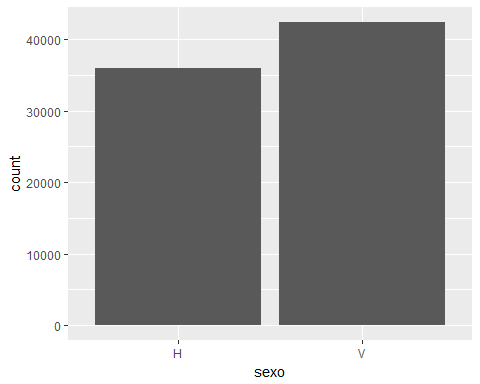
## ncodpers ind\_empleado pais\_residencia   
## 0 514 514   
## sexo age fecha\_alta   
## 515 514 514   
## antiguedad indrel indrel\_1mes   
## 514 0 735   
## tiprel\_1mes indresi indext   
## 735 514 514   
## ind\_actividad\_cliente renta segmento   
## 0 20830 807   
## canal\_entrada\_clean   
## 14133

## Impute NA's for each column

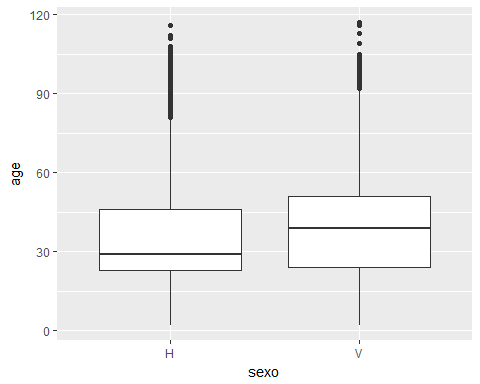
The NA's in the columns Sex, Income, Segment, Channel, Employee indicator, Country, Customer type, Customer Relation type, Age, Customer residence indicator, Contract date, Customer foreign index were all imputed. In all these cases the most commonly occuring value was selected for the categorical columns and the mean of the values were taken for the numeric columns.

## Descriptive and Exploratory Analysis

Absolute and relative frequencies were computed for categorical columns and descriptive statistics performed for the numeric variables. The number of males are more than females at 42,403. The percentage of males are more than females at 54.12%. Plotting the sex in a Histogram shows that the count of males are more than females.



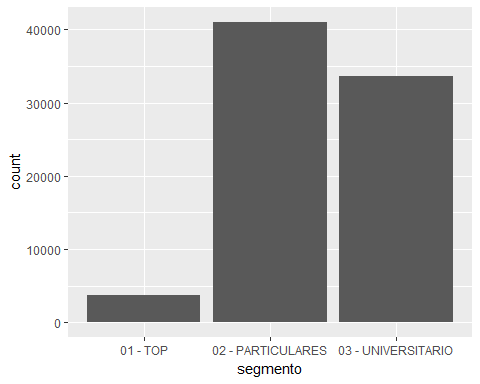
Age: It was found that the minumum age of the customer is 2 and maximum is 117. The mean or average age of the customer is around 38 years and the standard deviation is only 16.98 which means they are closer to the mean age. Plotting the Customer Sex against Age in a Box plot shows that 50% of the females are in the age range 25 to 45 and males are between 27 and 55. The median age for females is around 30 and that of males is around 40.



Channel: Customers use KHE Channel the most with a total of 37,322 and account for 47.63% of the total. Gross income of household: The mean income is 121,388 . The minimum salary of the Customers is 5341 and the maximum salary is 15,711,716. Customer Segmentation: The highest segment of customers are 02 - PARTICULARES(Individuals) with 41047 and accounting for 52.38% as shown by the bar graph.

##   
## 01 - TOP 02 - PARTICULARES 03 - UNIVERSITARIO   
## 3675 41047 33637

##   
## 01 - TOP 02 - PARTICULARES 03 - UNIVERSITARIO   
## 4.690 52.383 42.927



## Machine Learning Techniques

Since the goal of the project was to determine the effect of the Customer attributes on the banking products, the first step was to determine the products that was most subscribed to by the Customers. It was found that the Current Accounts had 67,246 customers whereas the Direct Debit has 16,985 customers. Hence these two banking products were selected for the analysis. Since the outcome of a customer subscribing to the products is either a 1 (subscribing) or 0 (not subscribing), Logistic Regression was chosen to predict the probability of a customer subscribing to a product.

## [,1]  
## ind\_ahor\_fin\_ult1 11  
## ind\_aval\_fin\_ult1 2  
## ind\_cco\_fin\_ult1 67246  
## ind\_cder\_fin\_ult1 51  
## ind\_cno\_fin\_ult1 9134  
## ind\_ctju\_fin\_ult1 818  
## ind\_ctma\_fin\_ult1 1549  
## ind\_ctop\_fin\_ult1 13394  
## ind\_ctpp\_fin\_ult1 2607  
## ind\_deco\_fin\_ult1 858  
## ind\_deme\_fin\_ult1 165  
## ind\_dela\_fin\_ult1 4511  
## ind\_ecue\_fin\_ult1 8885  
## ind\_fond\_fin\_ult1 1793  
## ind\_hip\_fin\_ult1 566  
## ind\_plan\_fin\_ult1 885  
## ind\_pres\_fin\_ult1 134  
## ind\_reca\_fin\_ult1 4759  
## ind\_tjcr\_fin\_ult1 5863  
## ind\_valo\_fin\_ult1 2468  
## ind\_viv\_fin\_ult1 329  
## ind\_nomina\_ult1 7585  
## ind\_nom\_pens\_ult1 8129  
## ind\_recibo\_ult1 16985

### Analyzing product ind\_recibo\_ult1(Direct Debit)

The first step in Logistic regression is to to formulate a Baseline model. It is found that 61374 customers do not subscribe to the product whereas 16985 customers only subscribe to the product. Since predictions are based on more occurence the baseline model predicts that only 61374/78359 = 78.32 % of the customers do not subscribe to the product. A Logistic regression model is created with the variables age,income and segment. Based on the significance of the coefficients (stars) it is seen that age , rent and "segmento02 - PARTICULARES" and "segmento03 UNIVERSITARIO" are significant in the model. The AIC value of this model is 75734. Lower the AIC value better is the model.

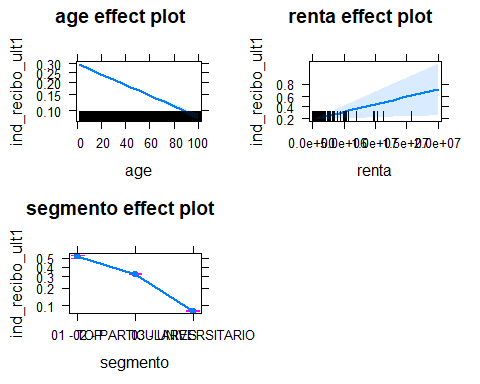
## Model 1

##   
## Call:  
## glm(formula = ind\_recibo\_ult1 ~ age + renta + segmento, family = binomial,   
## data = na.omit(main\_df))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5841 -0.8137 -0.4595 -0.4384 2.4531   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.509e-01 5.174e-02 12.580 <2e-16 \*\*\*  
## age -1.534e-02 7.109e-04 -21.573 <2e-16 \*\*\*  
## renta 1.217e-07 5.399e-08 2.255 0.0242 \*   
## segmento02 - PARTICULARES -8.162e-01 3.553e-02 -22.976 <2e-16 \*\*\*  
## segmento03 - UNIVERSITARIO -2.523e+00 4.399e-02 -57.343 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 81928 on 78358 degrees of freedom  
## Residual deviance: 75724 on 78354 degrees of freedom  
## AIC: 75734  
##   
## Number of Fisher Scoring iterations: 4

The model is then used with the predict() function to make direct statements about the predictors in our model. For example questions like "How much more likely are the different segments of customers likely to subscribe to the product at an average age of 38 and an average rent of 122,000". It is seen that the 01 - TOP segment has a 53% probability of subscribing to the Direct debit product than 02 - PARTICULARES segment with 34% and 03 - UNIVERSITARIO segment with 8%.

## age renta segmento fit se.fit residual.scale  
## 1 34 99710.07 01 - TOP 0.53533566 0.009094735 1  
## 2 34 99710.07 02 - PARTICULARES 0.33745711 0.003190378 1  
## 3 34 99710.07 03 - UNIVERSITARIO 0.08462419 0.001529871 1

Plotting the graph it is seen that the probability of subscribing to the product decreases as age increases or has a negative intercept. The probability of subscribing to the product increases as income increases or has a positive intercept. The probability of subscribing is the highest in the 01 - TOP segment with 53%.



### Model 2

Adding another variable aniguedad(priority of customer in number of months) lowers the AIC value to 73371 and hence is a better model. Antiguedad has a positive intercept and is significant in the model.

##   
## Call:  
## glm(formula = ind\_recibo\_ult1 ~ age + antiguedad + renta + segmento,   
## family = binomial, data = na.omit(main\_df))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4174 -0.8169 -0.4599 -0.4374 2.4500   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.336e-01 5.193e-02 12.201 < 2e-16 \*\*\*  
## age -1.698e-02 7.698e-04 -22.064 < 2e-16 \*\*\*  
## antiguedad 8.401e-04 1.460e-04 5.752 8.81e-09 \*\*\*  
## renta 9.652e-08 5.375e-08 1.796 0.0725 .   
## segmento02 - PARTICULARES -8.072e-01 3.559e-02 -22.679 < 2e-16 \*\*\*  
## segmento03 - UNIVERSITARIO -2.486e+00 4.450e-02 -55.860 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 81928 on 78358 degrees of freedom  
## Residual deviance: 75691 on 78353 degrees of freedom  
## AIC: 75703  
##   
## Number of Fisher Scoring iterations: 4

### Model 3

Since renta i.e income is insignificant a third model is created excluding renta which brought down the AIC value to 73369 which is not very different from the second model. Hence the second model log\_reg2\_ind\_recibo\_ult1 was retained for predicting probabilities.

Now lets predict the probability of the selected model that predicts if the customers would subscribe to the product i.e 1 or not subscribe to the product i.e 0. We see that the probabilities are between 0.008 and 0.913.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.04972 0.09933 0.23714 0.21676 0.31088 0.68176

Translating the probabilities to actual numbers i.e compute the actual numbers of how many customers will subscribe to the product or not rather than the probabilities. Let us select a Threshold value of 0.5. The accuracy of the model is True Positive + True Negative/N= 2045+59560/78359=78.61% which is nearly the same as the baseline model. The sensitivity or true positive rate of the model is TP/TP+FN = 2045/2045+14940= 12.04% The speicificity or the true negative rate of the model is TN/TN+FP=59560/59560+1814 = 97.04 %

Hence the prediction that the probability of customers who do not subscribe to the product ind\_recibo\_ult1 is greater than those who subscribe to the product.

##   
## FALSE TRUE  
## 0 61020 354  
## 1 16417 568

Lowering the threshold to 0.3 we see that the accuracy of the model is True Positive + True Negative/N= 8109+49783/78359=73.88% which is lower than the baseline model. The sensitivity or true positive rate of the model is TP/TP+FN = 8109/8109+8876= 47.74% The specificity or the true negative rate of the model is TN/TN+FP=49783/49783+11591= 81.11%.

Hence by decreasing the threshold the number of people not subscribing to the Direct Debit product is again higher at 81.11% than at 47.74%.

##   
## FALSE TRUE  
## 0 45729 15645  
## 1 8131 8854

### Analyzing product ind\_cco\_fin\_ult1(Current Accounts)

We then analyze the next product i.e Current Accounts. The first step is to formulate a Baseline model. About 67246 customers subscribe to the product and hence the model accuracy is 67246/78359=85.81%. A Logistic regression model is created with the variables age, income and segment. Based on the significance of the coefficients (stars) we see that age , rent and "segmento02 - PARTICULARES" and "segmento03 - UNIVERSITARIO" are significant in the model. The AIC value of this model is 54675. Lower the AIC value better is the model.

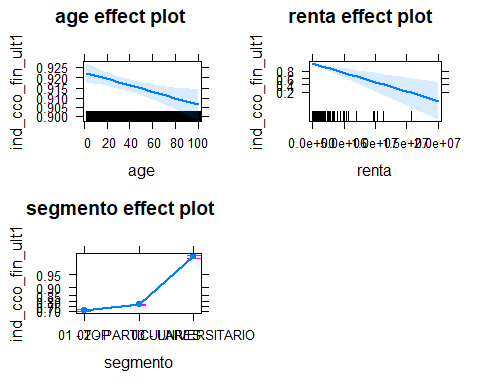
### Model 1

##   
## Call:  
## glm(formula = ind\_cco\_fin\_ult1 ~ age + renta + segmento, family = binomial,   
## data = na.omit(main\_df))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8483 0.1882 0.1902 0.7178 2.3357   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.000e+00 5.500e-02 18.181 < 2e-16 \*\*\*  
## age -2.071e-03 7.370e-04 -2.811 0.004943 \*\*   
## renta -2.478e-07 6.382e-08 -3.883 0.000103 \*\*\*  
## segmento02 - PARTICULARES 3.449e-01 3.823e-02 9.022 < 2e-16 \*\*\*  
## segmento03 - UNIVERSITARIO 3.091e+00 5.947e-02 51.973 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 63981 on 78358 degrees of freedom  
## Residual deviance: 54665 on 78354 degrees of freedom  
## AIC: 54675  
##   
## Number of Fisher Scoring iterations: 6

Next the `predict()' function to make direct statements about the predictors in our model. For example, we can ask "How much more likely are the different segments of customers likely to subscribe to the product at an average age of 34 and an average rent of 99,710". We see that the 03 - UNIVERSITARIO segment has a 98% probability of subscribing to the ind\_cco\_fin\_ult1 product than 02 - PARTICULARES segment with 78% and 01 - TOP segment with 71%.

## age renta segmento fit se.fit residual.scale  
## 1 34 99710.07 01 - TOP 0.7119541 0.0080862757 1  
## 2 34 99710.07 02 - PARTICULARES 0.7772681 0.0027688428 1  
## 3 34 99710.07 03 - UNIVERSITARIO 0.9819393 0.0007439738 1

Plotting the graph we can see that the probability of subscribing to the product decreases as age increases or has a negative intercept. The probability of subscribing to the product decreases as income increases or has a negative intercept. The probability of subscribing is the highest in the 03 - UNIVERSITARIO segment with 98%.



### Model 2

Adding another variable aniguedad lowers the AIC value to 52264 and hence is a better model. Rent is not significant in this model

##   
## Call:  
## glm(formula = ind\_cco\_fin\_ult1 ~ age + antiguedad + renta + segmento,   
## family = binomial, data = na.omit(main\_df))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9255 0.1814 0.1973 0.6356 1.1678   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.220e+00 5.698e-02 21.412 < 2e-16 \*\*\*  
## age 9.171e-03 8.272e-04 11.087 < 2e-16 \*\*\*  
## antiguedad -6.185e-03 1.680e-04 -36.827 < 2e-16 \*\*\*  
## renta -6.468e-08 6.020e-08 -1.074 0.283   
## segmento02 - PARTICULARES 2.737e-01 3.903e-02 7.012 2.34e-12 \*\*\*  
## segmento03 - UNIVERSITARIO 2.755e+00 6.104e-02 45.140 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 63981 on 78358 degrees of freedom  
## Residual deviance: 53230 on 78353 degrees of freedom  
## AIC: 53242  
##   
## Number of Fisher Scoring iterations: 6

### Model 3

Since rent is insignificant let us create a third model excluding rent which brings down the AIC value to 52262 which is not very different from the second model. Hence we will retain the second model log\_reg2\_ind\_cco\_fin\_ult1 for predicting probabilities.

Now lets predict the probability of the selected model that predicts if the customers would subscribe to the product i.e 1 or not subscribe to the product i.e 0. We see that the probabilities are between 0.25 and 1.00.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.5057 0.7403 0.8565 0.8582 0.9820 0.9861

Translating the probabilities to actual numbers i.e compute the actual numbers of how many customers will subscribe to the product or not rather than the probabilities. Let us select a Threshold value of 0.5. The accuracy of the model is True Positive + True Negative/N= 67179+72/78359=85.82% which is nearly the same as the baseline model The sensitivity or true positive rate of the model is TP/TP+FN = 67179/67179+67= 99.94% The specificity or the true negative rate of the model is TN/TN+FP=72/72+11041 = 0.64%

Hence the prediction that the probability of customers who subscribe to the product ind\_cco\_fin\_ult1 is greater than those who do not subscribe to the product.

##   
## TRUE  
## 0 11113  
## 1 67246

Lowering the threshold to 0.3 we see that the accuracy of the model is True Positive + True Negative/N= 67245+1/78359=85.81 which is nearly the same as the baseline model, The sensitivity or true positive rate of the model is TP/TP+FN = 67245/67245+1= 99.99%. The speicificity or the true negative rate of the model is TN/TN+FP=1/11112+1 = 0.00%.

Hence by decreasing the threshold the number of people subscribing to the Current Accounts product is again higher at 99.99%

##   
## FALSE TRUE  
## 0 45729 15645  
## 1 8131 8854

## Conclusions

Results from the Capstone Project show that

1. There is a higher probability of customers subscribing to the product Current Accounts(ind\_cco\_fin\_ult1).
2. There is a higher probability of customers not subscribing to the product Direct Debit(ind\_recibo\_ult1).
3. It is predicted that the highest number of customers who subscribe to the product Current Accounts(ind\_cco\_fin\_ult1) are from the 03 - UNIVERSITARIO segment.
4. It is also predicted that the highest number of customers who subscribe to the product Direct Debit(ind\_recibo\_ult1) are from the 01 - TOP segment.
5. The customer attributes like Age and an Income have a negative correlation to subscribing to the Current Accounts as most of them are University students.
6. The customer attributes like Age has a negative correlation to the Direct Debit Account and Income has a positive correlation to the Direct Debit account that comprises mostly VIP customers.