XGBOOST

ALEXANDER VALVERDE GUILLÉN 2022

HACKHATON

DATA SETS AND THEIR CHARACTERISTICS

CustomerID	Surname	Geography	Gender	HasCrCard	IsActivemember	Estimated Salary	applicationdate	exitdate	birthdate	Years	Age
15745584	EIRLS	Germany	Female	0	1	0.00	2018-12-14		1997-09-18	2	24.82740
14990118	MOLOCK	Italy	Male	1	0	121219.28	2019-01-08		1980-08-03	2	41.95503
14648573	NALLS	Spain	Male	1	0	140827.98	2019-06-19		1979-02-27	2	43.38356
15638124	BRASHERS	Italy	Female	0	0	170661.45	2018-02-23		1983-01-13	2	39.50685
14611239	DOKKA	France	Male	0	1	72210.60	2019-02-24		1986-04-26	2	36.22466
15613805	KALATHAS	France	Male	0	0	47278.31	2019-02-19		1990-04-03	2	32.28767
15434700	STIMMELL	Germany	Male	1	0	138615.32	2018-06-06		1994-07-22	2	27.98630
14765889	FREEBURG	France	Male	1	1	133997.03	2019-04-22		1966-11-25	2	55.64110
15989348	LIE	Italy	Female	1	1	222666.00	2019-02-02		1989-03-24	2	33.31507
14928023	MARKEL	Spain	Female	0	0	136052.05	2018-05-16		1989-12-27	2	32.55342
14873630	ELZIE	Italy	Male	1	0	110883.36	2019-07-21		1984-09-03	2	37.87033
15933310	RIGNEY	Spain	Male	0	1	154401.04	2019-05-06		1980-01-30	2	42.46027
14844655	CALICA	Germany	Female	0	0	64408.92	2018-05-27		1971-03-11	2	51.35068
14845498	DRONKO	Germany	Male	1	1	148745.32	2019-06-19		1991-08-18	2	30.91233
15528567	GORBEA	Spain	Female	1	1	253177.47	2018-05-21		1983-06-26	2	39.05753
14921129	BALSIS	Spain	Female	0	1	158104.46	2019-09-27		1975-06-10	2	47.10137
15535518	YELDELL	Italy	Male	0	0	272214.21	2018-04-01		1976-05-12	2	46.18082
15713953	WIESMAN	France	Female	0	0	148824.96	2018-12-26		1979-06-10	2	43.10137
15610711	Eluemuno	Germany	Female	0	0	167673.37	2017-03-05	2019-11-27	1976-10-21	2	45.73918
15890394	RIEK	Germany	Female	0	0	122201.86	2017-12-20		1983-07-08	2	39.02466
15606983	JUARIQUI	Italy	Male	0	0	115931.19	2017-12-13		1976-03-28	2	46.30411
15724878	MATTERA	Germany	Female	1	1	112080.43	2019-01-30		1982-11-01	2	39.70685
15849554	ZUMWALT	Germany	Female	1	0	81357.34	2017-12-19		1996-02-08	2	26.43562
15014099	HANNIGAN	Germany	Male	1	0	144588.54	2018-01-07		1977-12-02	2	44.62192
15503180	SINNOTT	Germany	Female	1	0	104673.21	2017-12-19		1968-03-09	2	54.35616
15390001	WEAKLEY	Italy	Male	1	0	17518.46	2019-01-22		1977-10-01	2	44.79178

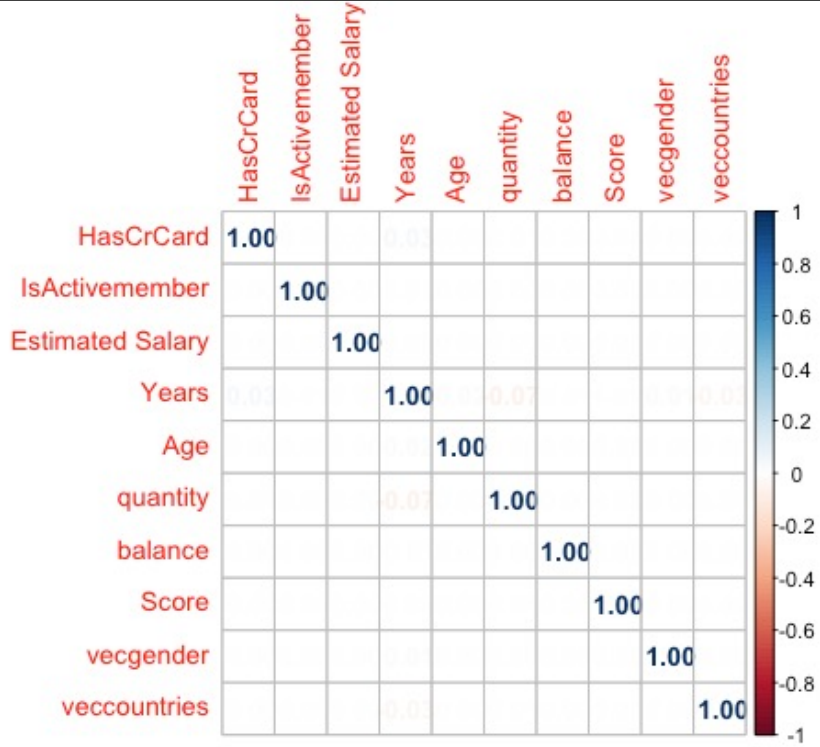
- There are 4 datasets divided in different variables.

The main one was the client data set with different variables regarding the company's clients and others such as entry date, exit date, etc.

The other 3 data sets have characteristics about their credit scores, the number of contracts, and financial balance. So its necessary to evaluate all these variables per Customer id, therefore, was necessary the use of SQL to develop new variables.

STATISTICS AND CORRELATION

Correlation Matrix



It can be determined that the dataset doesn't have important correlations between data, therefore, is not necessary to realize some cleaning related to the variables to avoid that impact.

Descriptive analysis

Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
HasCrCard	227496	0.503	0.5	0	0	1	1
IsActivemember	227496	0.502	0.5	0	0	1	1
Estimated Salary	227496	100984.891	55555.592	0	60936.663	139368.455	351311.15
Years	227496	2.008	0.104	2	2	2	4
Age	227496	42.604	10.276	20.688	35.444	49.529	98.219
quantity	227496	2.482	1.118	1	1	3	4
balance	227496	79557.399	56708.465	0	33730.54	118632.298	374633.66
Score	227496	649.714	96.619	177	585	715	1000
vecgender	227496	1.5	0.5	1	1	2	2
veccountries	227496	2.47	1.139	1	1	4	4

The table above indicates the descriptive statistics for almost variables, showing, that balance has approximately 80000 per customer, quantity indicates that at least people have 2 contracts at the company and a máximum of 4. Another important detail is that the age is around 40 years old, however, this service is being offered also for people that are 20 years old.

MODEL AND ANALYSIS

- To build a predictive statistical model based on the probability of a client dropping/canceling the product before 2 years was necessary to evaluate many models, however, the model that best fit the data for this data was a gradient boost model. Therefore was necessary to create an xgboost model.

- The model has as its predictive variable, the column of IsActiveMember, that indicates if the customer has an active account with the company. To explain this variables is neccesary to take in consideration more characteristics such as: age, balance account, score, quantity of products, years of service, application date, credit cards, etc.

- For launching the program was neccesary stablish an enough quatity of runs to test the best Mean Squared Error and see what is the correct number of rounds to have the least RMSE

TESTS AND MODEL

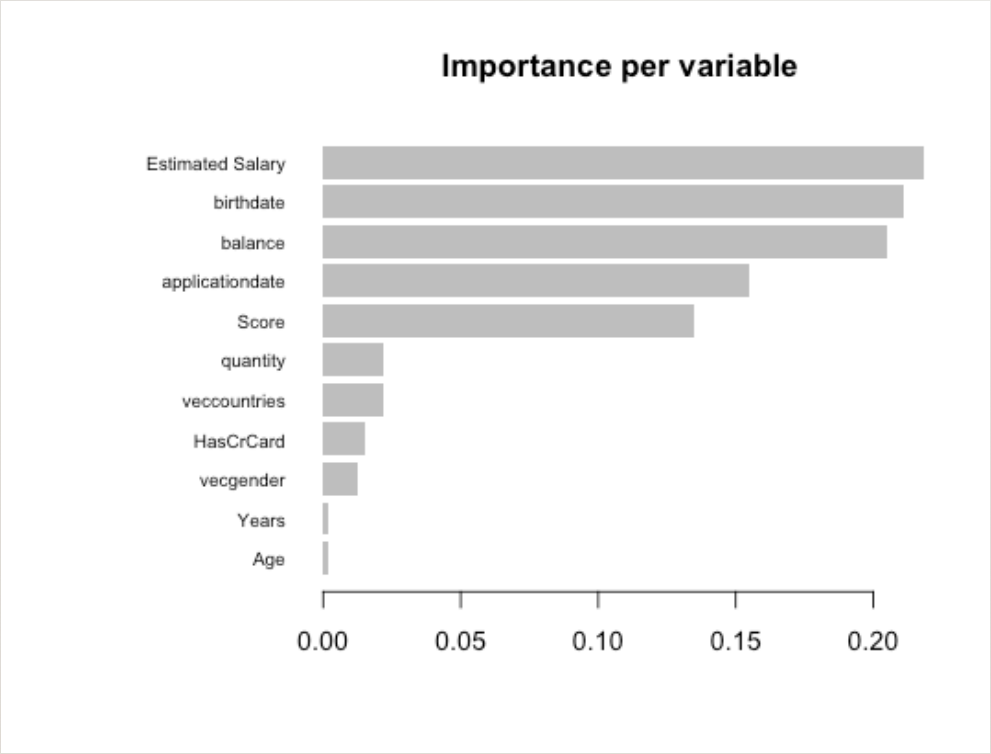
The Xgboost model was the following:

final = xgboost(data = xgb_train, max.depth = 3, nrounds = 27, verbose = 1)

According to the rmse was develop a model of 27 rounds and a max-Depth of 3, therefore, this model show the next tests when is compared to the y variable (IsActivemember)

MSE =0.250
MAE = 0.499
RMSE = 0.50

TOP 5 WEIGHER VARIABLES EXPLAINING THE Y VALUE



- The Xgboost model indicates that the top 5 variables that have more weight in the Y variable are birthdate, balance, application date, estimated salary, and score.
- The Salary and score are variables that could have a minimum relation because both are variables that are related to the economic situation of the person and birthdate or application date are dates variables, that one explains how old people are and the other one at what moment people decided to enter in the different contracts.