Kaggle Competition: Santander Customer Satisfaction

GA SEA DAT02

(THIS PERSON IS NAMED RACHEL CHEN)

Santander Bank

- Santander Group is a Spanish banking group.
- It’s pretty huge.
- Uuuuh.
- Santander Bank, N. A. is the North American subsidiary but this is irrelevant.

Which customers are happy customers?

- From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving.
- Santander Bank is asking Kagglers to help them identify dissatisfied customers early in their relationship. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late.
- In this competition, you'll work with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

TL;DR:

- Determine which customers are likely to leave Santander bank (binary response: 0 for happy, 1 for unhappy) based on given data containing demographic features and an account of each customer's finances and bank services usage over a period of three months.

The data at a glance

- It's at least been provided by Santander Bank.
- Training and testing data already split into train.csv and test.csv.
- There's no data dictionary provided.
- The features are also all in Spanish.
- Features include:
  - Nationality, age, use of various bank products, mortgages, balances, wages, fees, stocks(?), duration of implementation or use of product, etc.
- Different types of features, different variables for each
  - e.g.: imp_ent_varX, delta_imp_amort_varX_1y3, num_meses_varX, where X is a positive integer

The data at another glance

- train.csv original shape = (76020, 371)
- test.csv original shape = (75818, 370)
- According to train.csv, 4% of customers are discontent.
- “Delta” features have errors coded as 9999999999.
  - But mostly 0’s, occasional -1 and 9999999999.
- Nationality feature has errors coded as -999999.
- A few seemingly all-zero columns.
  - importe amortización, importe reembolso, número reembolso, saldo medio
- Some columns are the sums of other columns.

The data itself

'ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_comer_ult3	imp_op_var40_efect_ult1	imp_op_var40_efect_ult3	imp_op_var40_ult1	imp_op_var41_comer_ult1	imp_op_var41_comer_ult3	imp_op_var41_efect_ult1	imp_op_var41_efect_ult3	imp_op_var41_ult1	imp_op_var39_efect_ult1	imp_op_var39_efect_ult3	imp_op_var39_ult1	imp_sal_var16_ult1	ind_var1_0
1	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	2	37	0	195	195	0	0	0	0	0	195	195	0	0	195	0	0	195	0	0
10	2	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	2	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	2	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	2	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	2	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	2	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	2	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	2	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	2	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	2	33	600	1086.48	1952.91	0	0	0	0	0	1086.48	1952.91	360	750	1446.48	360	750	1446.48	0	0
34	2	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	2	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	2	36	0	55.2	70.95	0	0	0	0	0	55.2	70.95	0	0	55.2	0	0	55.2	0	0
42	229	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	2	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	2	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	2	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	2	54	0	117.75	122.25	0	0	0	0	0	117.75	122.25	0	0	117.75	0	0	117.75	0	0
56	2	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	2	40	0	1658.37	5722.77	247.56	835.86	0	960	1585.86	1410.81	4886.91	300	1320	1710.81	300	2280	3296.67	0	1
61	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	2	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	2	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	2	23	0	55.05	127.44	0	0	0	0	0	55.05	127.44	60	60	115.05	60	60	115.05	0	0
75	2	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
77	2	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
80	2	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
82	2	25	0	59.7	59.7	0	0	0	0	0	59.7	59.7	0	0	59.7	0	0	59.7	0	0
83	2	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	2	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	2	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	2	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

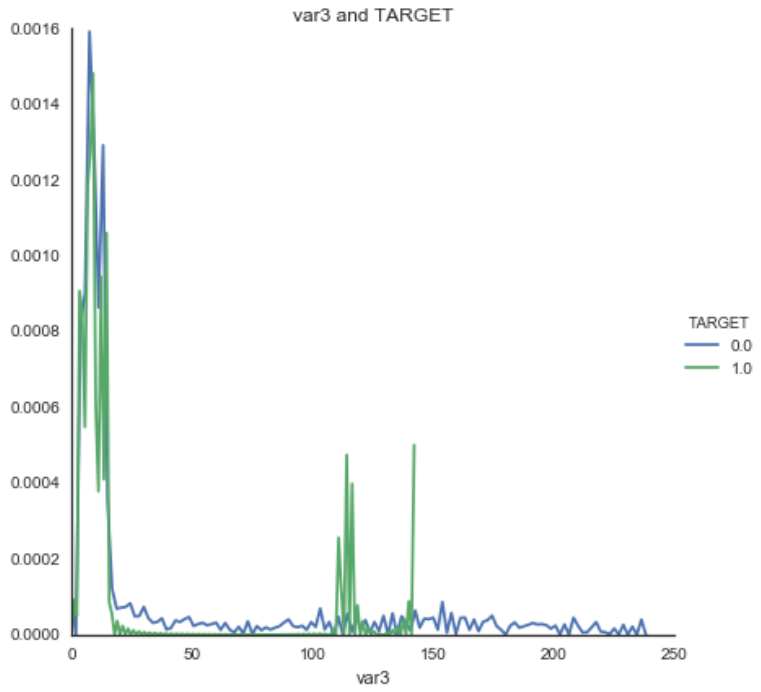
The data at a slightly deeper glance

	count	mean	std	min	25%	50%	75%	max
'ID	76020	75964.05	43781.95	1	38104.75	76043	113748.8	151838
var3	76020	-1523.2	39033.46	-999999	2	2	2	238
var15	76020	33.21287	12.95649	5	23	28	40	105
imp_ent_var16_ult1	76020	86.20827	1614.757	0	0	0	0	210000
imp_op_var39_comer_ult1	76020	72.36307	339.3158	0	0	0	0	12888.03
imp_op_var39_comer_ult3	76020	119.5296	546.2663	0	0	0	0	21024.81
imp_op_var40_comer_ult1	76020	3.55913	93.15575	0	0	0	0	8237.82
imp_op_var40_comer_ult3	76020	6.472698	153.7371	0	0	0	0	11073.57
imp_op_var40_efect_ult1	76020	0.412946	30.60486	0	0	0	0	6600
imp_op_var40_efect_ult3	76020	0.567352	36.51351	0	0	0	0	6600
imp_op_var40_ult1	76020	3.160715	95.2682	0	0	0	0	8237.82
imp_op_var41_comer_ult1	76020	68.80394	319.6055	0	0	0	0	12888.03
imp_op_var41_comer_ult3	76020	113.0569	512.1548	0	0	0	0	16566.81
imp_op_var41_efect_ult1	76020	68.20514	531.8979	0	0	0	0	45990
imp_op_var41_efect_ult3	76020	113.2251	950.0864	0	0	0	0	131100
imp_op_var41_ult1	76020	137.2428	697.7126	0	0	0	0	47598.09
imp_op_var39_efect_ult1	76020	68.61809	535.4737	0	0	0	0	45990
imp_op_var39_efect_ult3	76020	113.7924	953.5786	0	0	0	0	131100
imp_op_var39_ult1	76020	140.4035	712.7672	0	0	0	0	47598.09
imp_sal_var16_ult1	76020	5.477676	465.3911	0	0	0	0	105000
ind_var1_0	76020	0.011458	0.106425	0	0	0	0	1
ind_var1	76020	0.003762	0.061221	0	0	0	0	1
ind_var2_0	76020	0	0	0	0	0	0	0
ind_var2	76020	0	0	0	0	0	0	0
ind_var5_0	76020	0.958024	0.200535	0	1	1	1	1
ind_var5	76020	0.66376	0.472425	0	0	1	1	1
ind_var6_0	76020	0.000105	0.010258	0	0	0	0	1
ind_var6	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var8_0	76020	0.032833	0.178202	0	0	0	0	1
ind_var8	76020	0.028598	0.166674	0	0	0	0	1
ind_var12_0	76020	0.067522	0.250925	0	0	0	0	1
ind_var12	76020	0.045462	0.208316	0	0	0	0	1
ind_var13_0	76020	0.052249	0.222531	0	0	0	0	1
ind_var13_corto_0	76020	0.042936	0.202714	0	0	0	0	1
ind_var13_corto	76020	0.041476	0.19939	0	0	0	0	1
ind_var13_largo_0	76020	0.010168	0.100325	0	0	0	0	1
ind_var13_largo	76020	0.009997	0.099486	0	0	0	0	1
ind_var13_medio_0	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var13_medio	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var13	76020	0.050855	0.219703	0	0	0	0	1
ind_var14_0	76020	0.023652	0.151962	0	0	0	0	1
ind_var14	76020	0.005301	0.072617	0	0	0	0	1
ind_var17_0	76020	0.001802	0.042414	0	0	0	0	1
ind_var17	76020	0.001447	0.038012	0	0	0	0	1
ind_var18_0	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var18	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var19	76020	0.004196	0.064643	0	0	0	0	1
ind_var20_0	76020	0.003631	0.060146	0	0	0	0	1
ind_var20	76020	0.002697	0.05186	0	0	0	0	1
ind_var24_0	76020	0.04237	0.201434	0	0	0	0	1
ind_var24	76020	0.037885	0.190919	0	0	0	0	1
ind_var25_cte	76020	0.026427	0.160403	0	0	0	0	1
ind_var26_0	76020	0.024638	0.155021	0	0	0	0	1
ind_var26_cte	76020	0.027559	0.163705	0	0	0	0	1
ind_var26	76020	0.024638	0.155021	0	0	0	0	1
ind_var25_0	76020	0.023639	0.151921	0	0	0	0	1
ind_var25	76020	0.023639	0.151921	0	0	0	0	1
ind_var27_0	76020	0	0	0	0	0	0	0
ind_var28_0	76020	0	0	0	0	0	0	0
ind_var28	76020	0	0	0	0	0	0	0
ind_var27	76020	0	0	0	0	0	0	0
ind_var29_0	76020	0.000105	0.010258	0	0	0	0	1
ind_var29	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var30_0	76020	0.995488	0.06702	0	1	1	1	1
ind_var30	76020	0.732833	0.442483	0	0	1	1	1
ind_var31_0	76020	0.004275	0.065245	0	0	0	0	1
ind_var31	76020	0.00367	0.06047	0	0	0	0	1
ind_var32_cte	76020	0.00121	0.034767	0	0	0	0	1
ind_var32_0	76020	0.001079	0.032826	0	0	0	0	1
ind_var32	76020	0.001079	0.032826	0	0	0	0	1
ind_var33_0	76020	0.00075	0.027372	0	0	0	0	1
ind_var33	76020	0.000631	0.02512	0	0	0	0	1
ind_var34_0	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var34	76020	2.63E-05	0.005129	0	0	0	0	1
ind_var37_cte	76020	0.072297	0.25898	0	0	0	0	1
ind_var37_0	76020	0.065259	0.246984	0	0	0	0	1
ind_var37	76020	0.065259	0.246984	0	0	0	0	1
ind_var39_0	76020	0.880755	0.324079	0	1	1	1	1
ind_var40_0	76020	0.011418	0.106244	0	0	0	0	1
ind_var40	76020	0.003723	0.060901	0	0	0	0	1

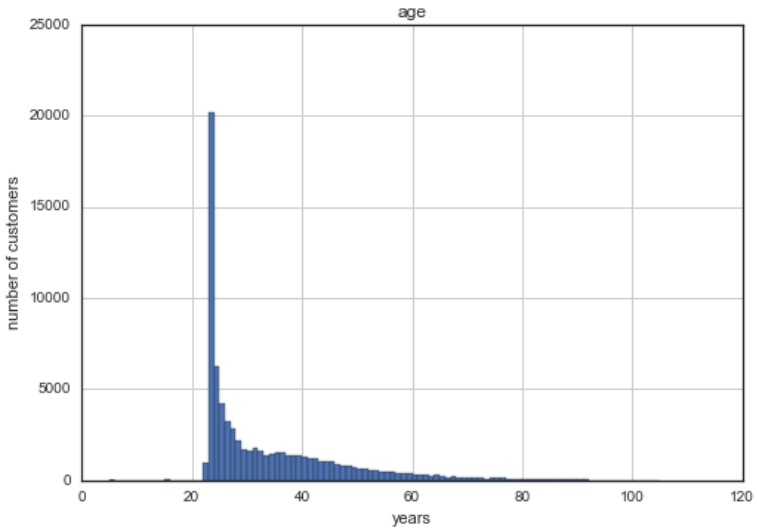
The data at another slightly deeper glance

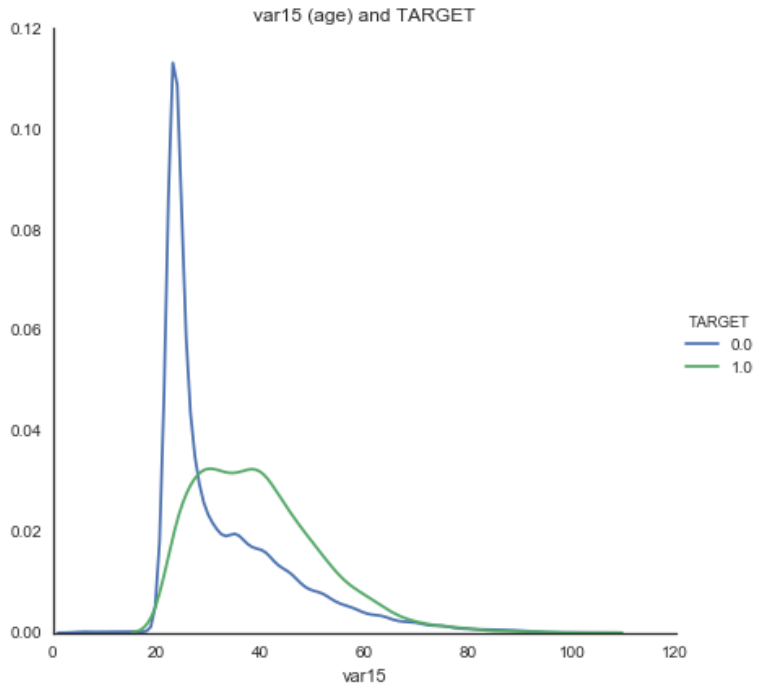
```
In [21]: # var3
         train.var3.value_counts()[0:20]

Out[21]:  2         74165
          8           138
         -999999      116
          9           110
          3           108
          1           105
          13           98
          7            97
          4            86
          12           85
          6            82
          0            75
          10           72
          11           66
          5            63
          14           61
          15           34
          18           10
          16            9
          17            7
         Name: var3, dtype: int64
```



The data at some more slightly deeper glances





What now

- Do everything else.
- Determine how other features relate to one another if there are more relationships I haven't figured out yet.
- Narrow down potential features as much as possible.
- Cross validate within train.csv before attempting to predict test.csv.