Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

By
Aakash Parwani
DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

OBJECTIVE

The objective of this project is to work with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience. Santander bank wants to 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.

DESIGN CONSIDERATIONS

- **Programming Language:** Python 3.5.
- **IDE:** Spyder, Jupyter.
- **Packages:** Numpy, Pandas, Sklearn, Json, Xgboost, Matplotlib, Scipy.

DATA STRUCTURE

List, Dictionary, Array, Dataframe.

ABOUT DATA

- Dataset contains a large number of numeric variables. The "TARGET" column is the variable to predict. It equals one for unsatisfied customers and 0 for satisfied customers. The task is to predict the probability that each customer in the test set is an unsatisfied customer.
- Project has two separate datasets one for training purpose and another one for testing purpose
- Training data set contains 76020 observations & 371 variables.
- Testing data set contains 75818 observations & 370 variables.
- Let us take glimpse of both the datasets.

Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

Training Data Set:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3
count	76020.000000	76020.000000	76020.000000	76020.000000	76020.000000	76020.000000
mean	75964.050723	-1523.199277	33.212865	86.208265	72.363067	119.529632
std	43781.947379	39033.462364	12.956486	1614.757313	339.315831	546.266294
min	1.000000	-999999.000000	5.000000	0.000000	0.000000	0.000000
25%	38104.750000	2.000000	23.000000	0.000000	0.000000	0.000000
50%	76043.000000	2.000000	28.000000	0.000000	0.000000	0.000000
75%	113748.750000	2.000000	40.000000	0.000000	0.000000	0.000000
max	151838.000000	238.000000	105.000000	210000.000000	12888.030000	21024.810000

8 rows × 371 columns

Testing Data Set:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3
count	75818.000000	75818.000000	75818.000000	75818.000000	75818.000000	75818.000000
mean	75874.830581	-1579.955011	33.138832	83.164329	74.312894	123.136448
std	43882.370827	39752.473358	12.932000	1694.873886	364.211245	606.431562
min	2.000000	-999999.000000	5.000000	0.000000	0.000000	0.000000
25%	37840.250000	2.000000	23.000000	0.000000	0.000000	0.000000
50%	75810.000000	2.000000	27.000000	0.000000	0.000000	0.000000
75%	113996.500000	2.000000	39.000000	0.000000	0.000000	0.000000
max	151837.000000	238.000000	105.000000	240000.000000	21093.960000	47943.960000

8 rows × 370 columns

ABOUT PARAMETERS

In this section we will checkout definitions of only important variables that we will use in the case study.

- **imp_ent_varX:** amount for the bank office.
- **imp_op_varX_comer:** amount for commercial option.

Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **imp_sal_varX:** Amount for wage.
- **ind_varX_corto:** short (time lapse?) indicator/dummy.
- **ind_varX_medio:** medium-sized (time lapse?) indicator/dummy.
- **ind_varX_largo:** long-sized (time lapse?) indicator/dummy.
- **saldo_varX:** Balance.
- **delta_imp_amort_varX_1y3:** Amount/price for redemption (?) 1 and 3.
- **delta_imp_aport_varX_1y3:** Amount/price for contribution (?) 1 and 3.
- **delta_imp_reemb_varX_1y3:** Amount/price for refund 1 and 3.
- **delta_imp_trasp_varX_out_1y3:** Amount/price for transfer 1 and 3.
- **imp_venta_varX:** Sale Fee.
- **ind_varX_emit_ult1:** Indicator of emission.
- **ind_varX_recib_ult1:** Indicator of reception.
- **num_varX_hace2:** Number [of variable X] done two units in the past.
- **num_med_varX:** Mean number [of variable X].
- **num_meses_varX:** Number of months [for variable X].
- **saldo_medio_varX:** Average balance.
- **delta_imp_venta_varX_1y3:** Fee on sales [for variable X] 1 and 3.
- **var4:** Number of bank products customer is enrolled.

And the products are:

Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

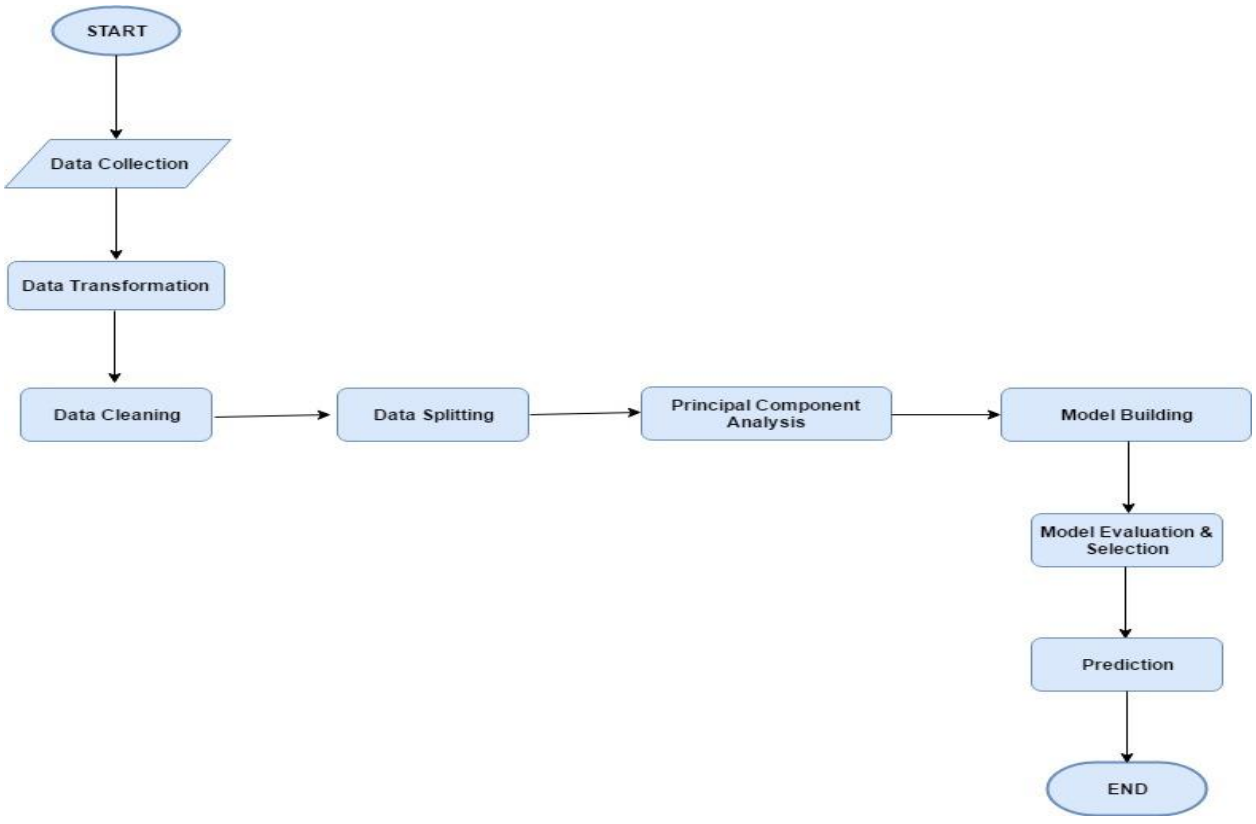
By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **Cash products:** var05, var08, var06/29, var20, var24, var14 and var13. These sum up to var30**.**
- **Credit products:** var17, var44, var33. These sums up to var31.
- **Card products:** var40, var41, var18, var34. These sums up to var01.
- **And name of some products are still unknown: var25, var32. These sums up to var26.**

ALGORITHM



Santander Customer Satisfaction

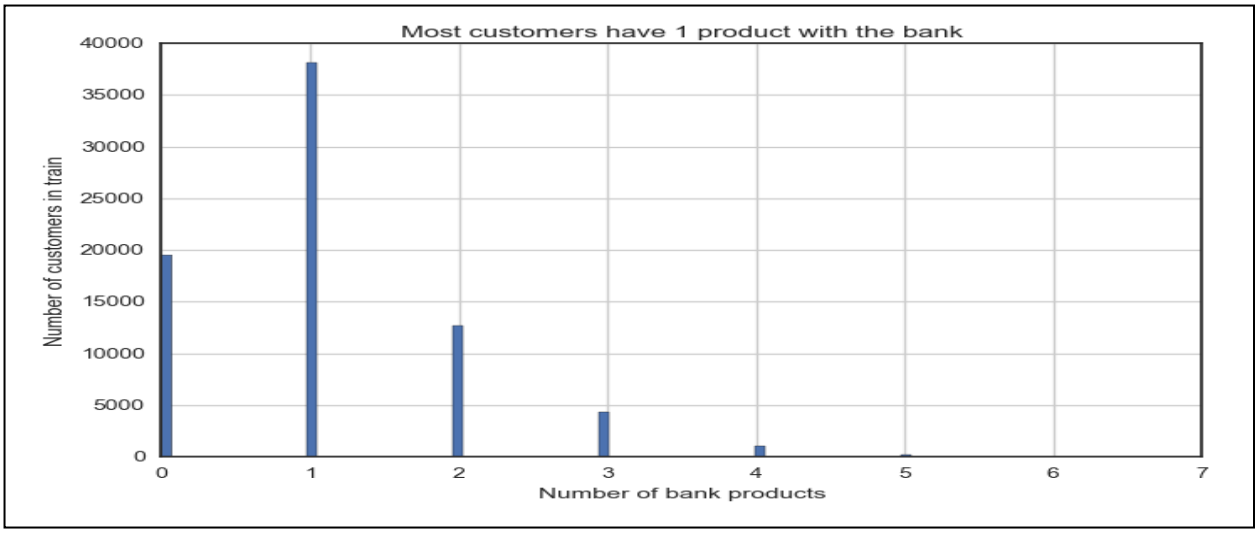
Predict customer satisfaction to improve relationship

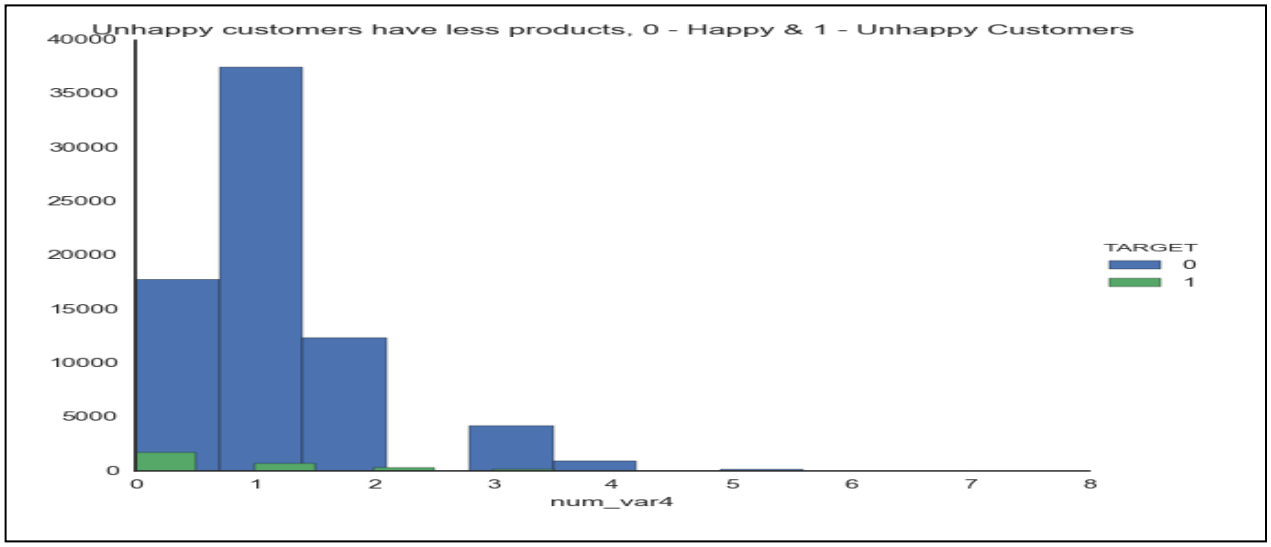
By

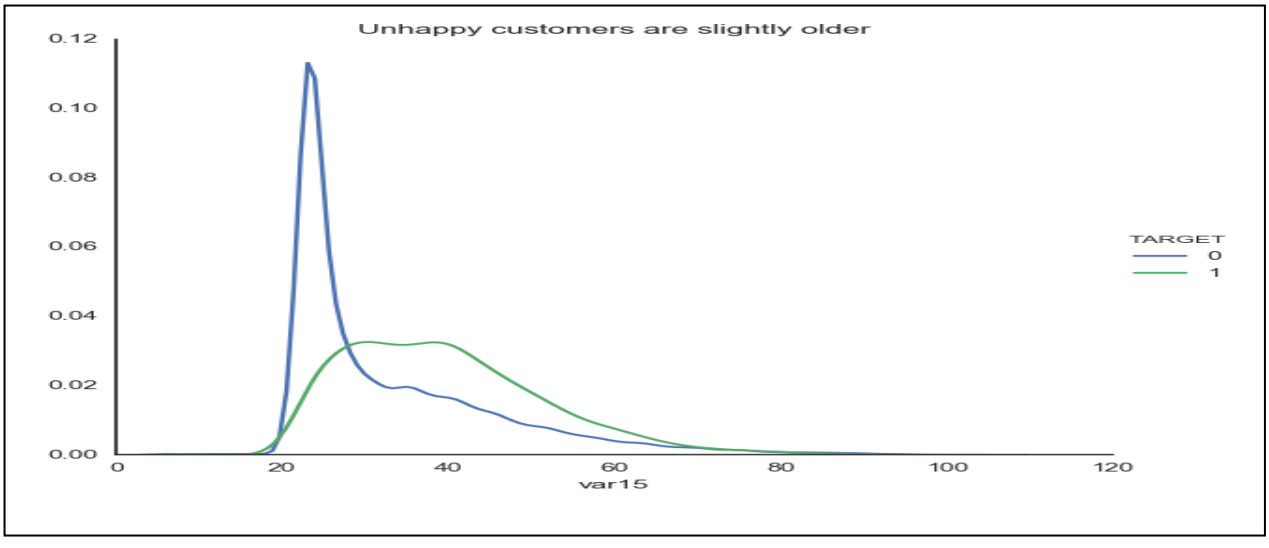
Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

DATA EXPLORATION







Santander Customer Satisfaction

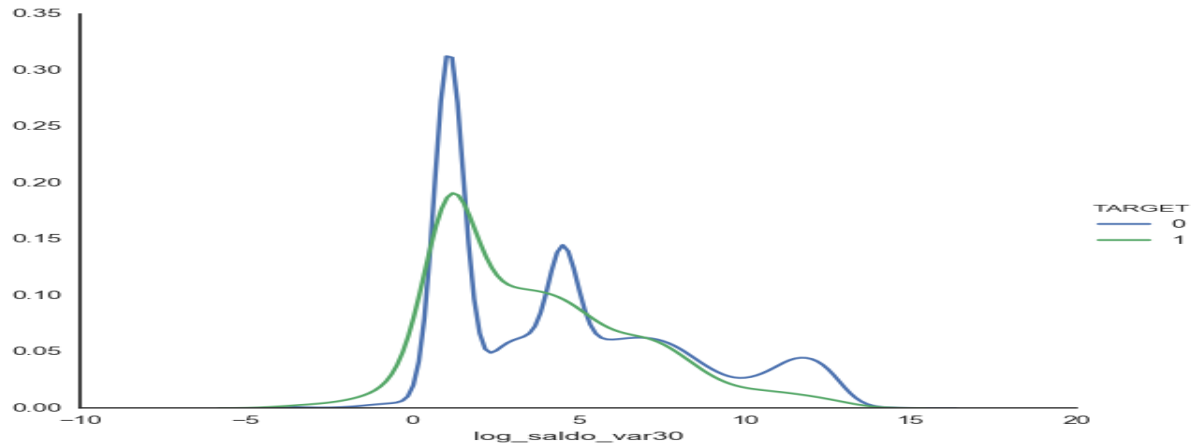
Predict customer satisfaction to improve relationship

By

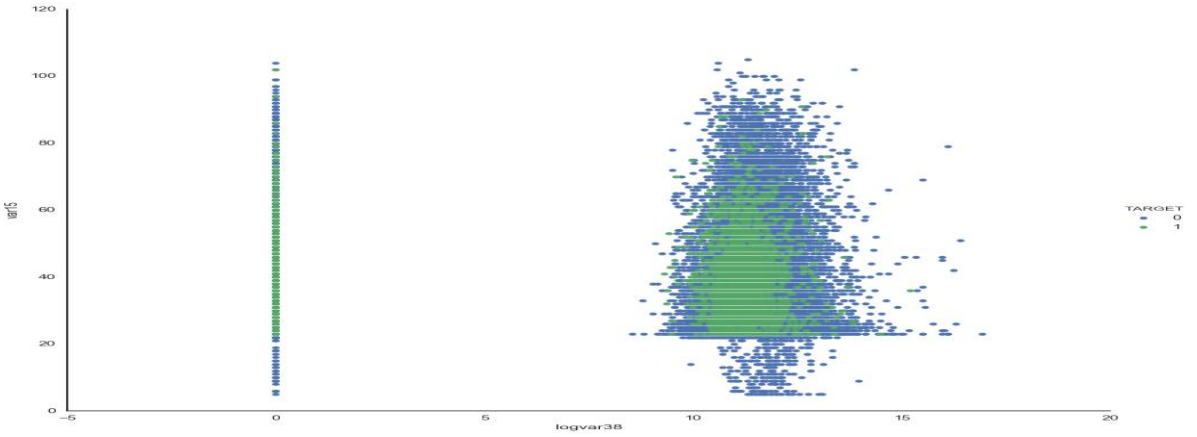
Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

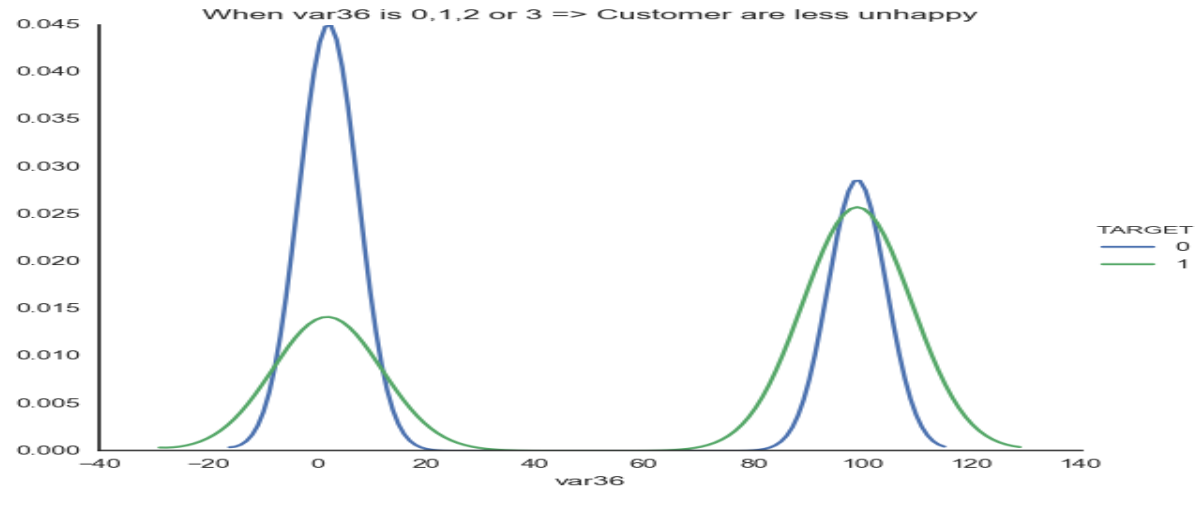
Density of the age of happy/unhappy customers for *saldo_var30* variable



Interaction of Var15 and Var38:



There is kind of cluster when we validate the customers reaction for different values of var36 feature. For 0,1,2,3 customers are mostly happy, on other hand for value 99 customers are mostly unhappy.



Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

By

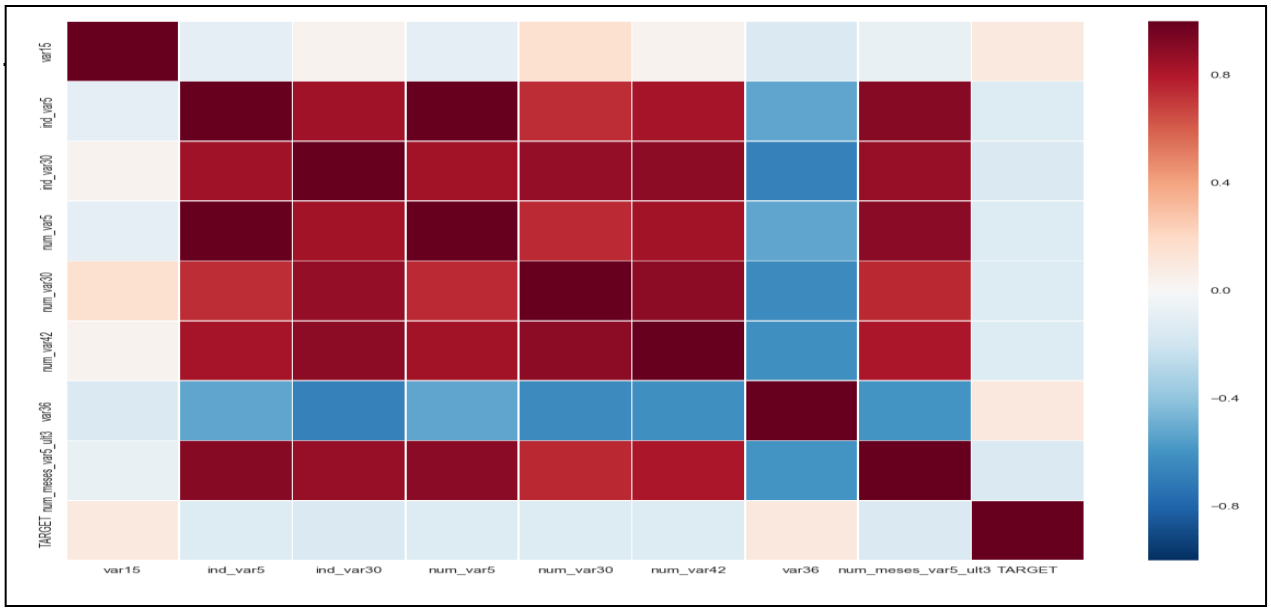
Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

PCA EVALUATION: Evaluation is performed for the selection of important features in the case study. **Chi Square and F-Classification** scores are considered for evaluation of important variables. Below are the test results.

```
Chi2 & F_classif selected 9 features
['var15', 'ind_var5', 'ind_var30', 'num_var5', 'num_var30', 'num_var42',
'var36', 'num_meses_var5_ult3', 'TARGET']
```

Now our future evaluations would be based on these 9 features only.



Correlation Stats

	attribute pair	correlation
6	(ind_var5, num_var5)	0.993709
14	(ind_var5, num_meses_var5_ult3)	0.908842
7	(num_meses_var5_ult3, num_var5)	0.903272
9	(num_var30, num_var42)	0.898119
11	(ind_var30, num_var42)	0.894182
13	(ind_var30, num_var30)	0.875812
0	(ind_var30, num_meses_var5_ult3)	0.869045
1	(ind_var30, ind_var5)	0.848338
4	(ind_var30, num_var5)	0.843001
2	(num_var42, num_var5)	0.839574
8	(ind_var5, num_var42)	0.832502
5	(num_meses_var5_ult3, num_var42)	0.813847
3	(num_meses_var5_ult3, num_var30)	0.756298
10	(num_var30, num_var5)	0.744330
12	(ind_var5, num_var30)	0.737867

Santander Customer Satisfaction

Predict customer satisfaction to improve relationship

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

MODEL BUILDING & EVALUATION:

For the prediction of **customer satisfaction** variable three statistical models are considered in this case study **RandomForest, RidgeRegression, & XGBoost.**

Accuracy Score of Models:

Model	Accuracy Score
RandomForest	0.741790
RidgeRegression	0.766448
XGBoost	0.800516

Result: - From scores it is clear that **XGBoost** model is performing well in terms of prediction of customer satisfaction.

Root Mean Square Error of Models:

['Root mean square error of rigde: 0.1912211',

'Root mean square error of randomforest: 0.1928076',

'Root mean square error of xboost: 0.1888888']

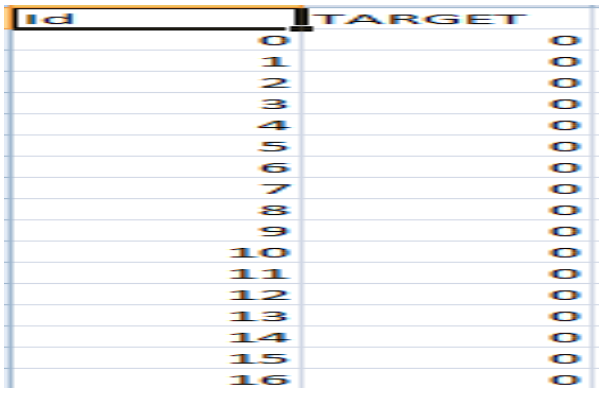
Result: - Now, **RMSE** score also confirms that **XGBoost** model is performing better than other models.

MODEL SELECTION:

From model evaluation section it is clear that **Xboost & Ridge** models are performing well**.**

Finally, **Xboost** model has been selected for the prediction of sales price on test data set.

Glimpse of values in submission final generated from the case study:



Id	TARGET
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0