**Project Name – Santander Customer Transaction Prediction**

Background - At Santander, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan?

Problem Statement - In this challenge, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

The first process we started with was Missing Value analysis and find out the missing value

We did

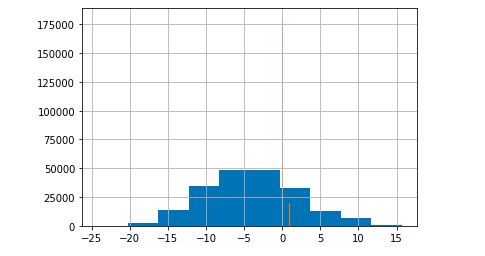
* Mean
* Median
* Knn imputation

Then we did the outlier analysis and find the values that are not the best and we take out the values unnecessary values which may make the aggregate incorrect

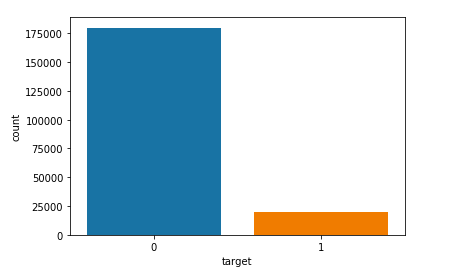
Correlation help us to find the correlation matrix present of the false positives and the false negatives present

When we find out the correlation In the descending order and we find the correct vales

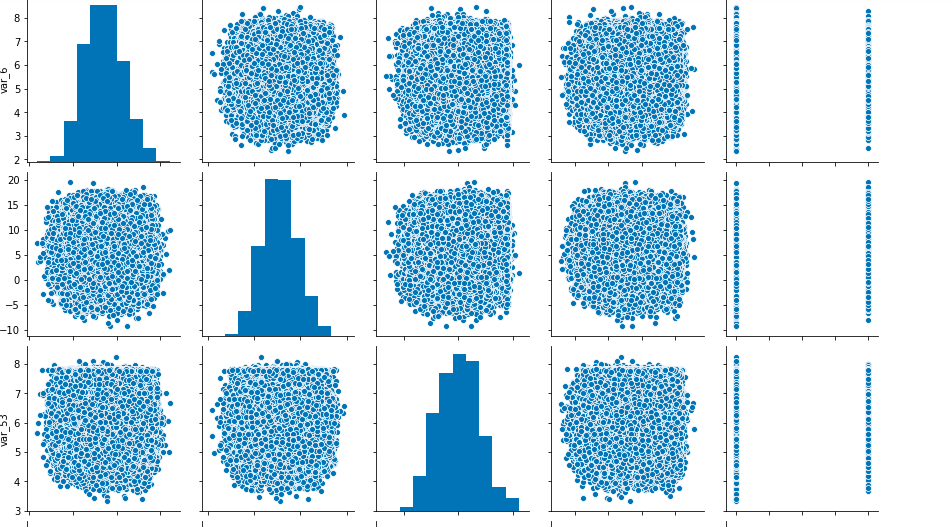
The highest correlated data with target.

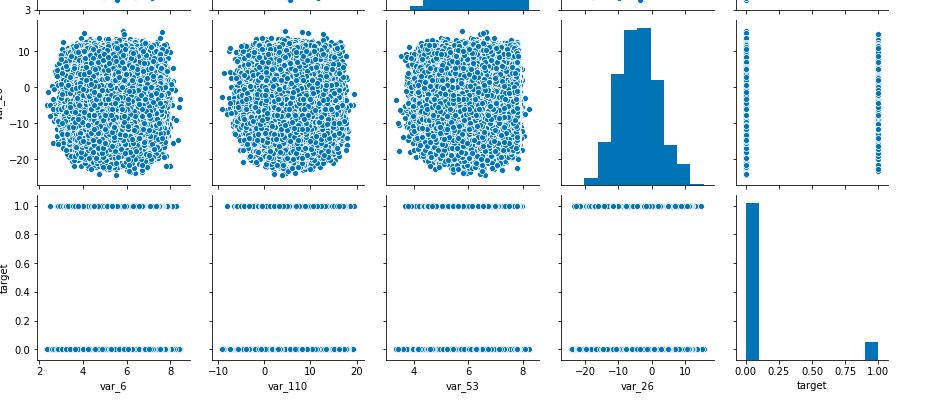


The number of target values the train dataset.



The pariplot defines it another way





The algorithms I used for classification is

* Decision Trees
* Random Forest
* SVM

Decision Trees

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems.

It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

Information Gain

• It represents the expected amount of information that would be needed

•Measure of purity

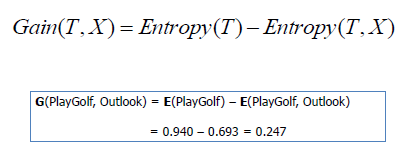
• Loss of entropy

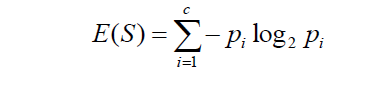
• IG = Entropy of the system before split – Entropy of the system after split

• Entropy: Uncertainty in the data/Measure of impurity, can be calculated as

• Selects the variable whose Information gain is hig

Eg.





ENTROPY FORMULAE

What is Random Forest?

• Random forest is an ensemble that consists of many decision trees

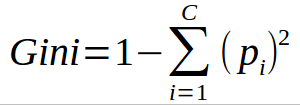
• The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995

. • The method combines Breiman's "bagging" idea and the random selection of features

• Outputs the class that is the mode of the class's output by individual trees.

• Mean for regression

• Can be used for classification and regression



SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

According to the SVM algorithm we find the points closest to the line from both the classes.These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

