Data wrangling and Story telling from the Santander data set

The data set provided for Banco Santander Customer Prediction doesn't need cleaning with all records with unique data information.

Doing analysis on both the trained and test data frames, so much information was extracted from it which brings out so much insights from the data.

From the description of train and test data frames, the mean, min, and max are slightly different whereas the standard deviations of both are quite close but large. Standard Deviation being a measure that is used to quantify the amount of variation. This implies a low Standard Deviation value indicates that data points turn to be close to mean while a high Standard Deviation value shows data points are spread out over a wide range of values.

Visualizing the relationship between each feature and the target using density plots of all the 200 features shows sum of all probabilities with the peak of the density plot showing the highest probability of the data points. Most of these features have very low peak values which shows they have little or no influence on a customer making a transaction in the future. Also, their low standard deviation values also show the wide range of variability, so measures could be brought upon these anonymous features to make the values within a shorter range and eliminate lower value data points too. This will therefore decrease the standard deviation and an increased probabilistic value influence to the target variable 1.

Since the aim of this project is to predict customers that will make a transaction in the future, like 5% of the features have the peak value at a probabilistic level of 0.6 which is the highest i.e. they have lower values of Standard Deviation. These few features though unidentified for security purposes could therefore be used as the main target on how to best keep these customers.

To conclude, this analysis shows most of the features instead have high probabilistic values on a customer not doing a transaction in the future. Since this is not the aim of the project, those with lower Standard Deviation values and with higher peak values with probability values greater then 0.5 could therefore be used further though the features are unidentified, the target of the company should be keeping those values or how to increase them while increasing the other probability values contained in the other features as well.