**DATA SCIENCE PROJECT**

|  |  |
| --- | --- |
| **Kaggle Competition: Santander Customer Satisfaction**  **Website:** <https://www.kaggle.com/c/santander-customer-satisfaction>  **Date: March to Mayl 2016** | **Autor: MSc. Oswaldo F. Domejean**  **Email:** [**ofigue@gmail.com**](mailto:ofigue@gmail.com)  **Lugar: La Paz – BOLIVIA** |

**Business understanding**

From the competition site description:

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.You are provided with an anonymized dataset containing 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.

**Data**

The data was around 75000 records with 370 features, one more for training, the dependent features called TARGET. Most of data was anonymised and besides from that most of the features have zeros and ones, kind of sparce matrix.

**Data exploration and Feature Engineering**

The data was mainly anonymized, and he name of the features made us think that it is an enterprise where Spanish is the base language, like in the case of “saldo” that means balance appear in same of them, and also “amort” that probably means amortization. But besides from that specific set of features most of them were very difficult to identify their meaning. It had been found that some of the features were constants, which were removed from the dataset. Something that showed up was the high correlation we could find in various features, something very unusual when you have a lots of features like in this case. Something that drawn my attention was the fact that this dataset had a very high level of imbalance in the predicted features TARGET.

Because of the variety of features and most of them impossible to identify their meaning, I created some new features that were combinations of existing ones using almost randomly functions like sqrt, log, multiplications and some others, this is a problem when you have mainly anonymized data, because it is not possible to identify some new features if you do not have the meaning of them.

**Models and Evaluation**

Mainly it had been used XGBoost, something that was amazing was that it could manage this imbalance data very good with come parameter values that I got from the kaggle posts.

As a conclusion the main aspect that I learned from this competition is that the problem with imbalance data can be managed with models like XGBoost in a smooth way.