DATA SCIENCE PROJECT PROPOSAL

Project Name: Santander Product Recommendation

Project Group: Berke Aral Sönmez, Alper Köse

Kaggle Link of Project: https://www.kaggle.com/c/santander-product-recommendation

Project Description:

We selected our project from a competition in Kaggle which is called "Santander Product Recommendation". For this project, we need to predict the behavior of the new customers by looking at the behavior of the previous customers. We are given training data which has the size of 2.5 gb approximately and a test data which has the size 110 mb. Our aim is to predict the columns 25-48 for the test data as accurate as possible using the columns 1-24 which are the features of a given row.

Method:

As our features, we have 24 columns for each customer and some of those columns are categorical and the others are numerical. Furthermore, there are some unidentified feature points which we need to get rid of using data cleaning. Afterwards we need to find a way to transform categorical features to numerical and get our final feature points and use machine learning algorithms on them to find the correct class for each row. We have learned multiple methods to classify our data and we'll try all of them to see which method gives the best result in the competition.

Evaluation Metric of Results:

Submission in the Kaggle are evaluated using Mean Average Precision @7 which is the formula below:

$$MAP@7 = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{min(m,7)}\sum_{k=1}^{min(n,7)}P(k)$$

|U| is the number of rows, n is the number of predicted products, P(k) is the precision at cut-off k and m is the number of added products for the selected user at that time.