1. Problem motivation:

Identify dissatisfied customers from Santander bank.

1. Data:

There are a total of 76020 observations of customers’ survey responses about their experiences with Santander. Each customer has 370 attribute features, which is slightly over 2% of the number of observations. We don’t have information on what those features mean, and the type of those features are also unknown, such as categorical versus numerical variables.

The target variable is a binary dummy variable with 0 representing satisfied customers and 1 representing dissatisfied customers. The goal is to accurately predict dissatisfaction. In addition, this dataset has 73012 satisfied customers and

3008 dissatisfied customers, which is about 4% of the whole dataset.

1. Analysis:
2. Data Wrangling

* Remove features that have only one unique value
* Change features with two unique values into binary variables
* Replace obvious missing value with NA, such as -999999 and 99
* Distinguish categorical variables from numerical ones, and transform them into binary formats

1. EDA and feature selection

* Assess the amount of correlation among numerical variables
* Use F-regression model selection method to rank the significance level of individual features
* Use PCA to identify the top components that explain 95% of variation in the feature space and visualize the top two components

1. Logistic Regression

* Apply KFOLD cross-validation to assess the prediction accuracy using F-score
* Run Logistic regression on the full dataset, F-regression selected features, and PCA top components.
* Compare across those models after optimizing the number of top features from F-regression as well as the number of PCA components

1. Machine learning analysis

* Apply Random Forest, Adaboost and XGBoost methods on the full dataset and also on the subset feature space
* Compare results across models using KFold cross-validation

1. Model evaluation

* Finalize the model with best prediction accuracy
* Plot residuals in training and testing sets against the predicted values