**May**

2017

**Springboard Second Capstone project**

Predicting Customer Satisfaction at Santander

Yi Zhao

08

**Fall**

# Background and Description of Problem

Santander is a retail and commercial bank founded in 1856 with a headquarter in the U.K. They have 125 million customers worldwide. Customer satisfaction is a key measure of their success and is a strong indicator to future customer retention. Identifying unhappy customers, finding ways to improve their satisfaction will prevent them from leaving Santander. Yet, unhappy customers rarely voice their dissatisfaction before leaving.

Thus, it is crucial for Santander to predict and identify dissatisfied customers early in their relationship in order to take proactive steps to improve a customer's happiness before it's too late.

In a Kaggle competition, we are provided with 370 of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

One of the biggest challenge in solving this prediction problem is that no prior information was provided about those features. We do not know which variables were numerical and which were categorical, not to say the meaning behind the numbers.

# Data and description of features

There are in total 76020 observations and 370 predictive features. There is no information on the meaning and type of those features. It is assumed that are the observations are independent with each other and each represent a customer.

The "TARGET" column is the variable to predict. It equals 1 for unsatisfied customers and 0 for satisfied customers. We don't have information on what each feature variable actually means or refers to. There are 3008 out of 76020 dissatisfied customers, or only 4%.

Some of them have hidden NA values, such as "-999999" as a minimum value of var3, and “99” as a maximum value of a variable with all other vales being 0, 1, 2, 3, 4,5. Those missing values are replaced with numpy NA.

Variables with 20 or less unique values among all 76020 observations were assumed to be categorical variables. Even though it was unclear whether they were ordered variables or not, they were spread into columns of dummy variables, or binary variable with zero and one entries. For each original variable that was transformed, there was one last column omitted from all its binary columns.

Among features with greater than 20 unique values, only three have a relatively small number of "0", and the rest of the features have 20K+ or over 26% of observations being zero. And the remaining 100 of those numerical predictor variables (> 20 unique values) have 60K+ observations with zero values, which is 79% of the sample size.

After taking all feature transformation, there were 851 features. Since there was no information on the features, we couldn’t reasonably fill in missing values. Omitting all observations with missing values reduced the dataset from 76020 to 45908. Since there were only two features with missing values, those two features were omitted instead of omitting observations. In the rest of the analyses, there were 76020 observations and 849 predictive features used.

Usually, numeric variables are centered (subtract the mean) and scaled (divide by the standard deviation) in a regression analysis so that those with a larger range of values would not have an extremely small coefficient. The centering and scaling assume the variables to have a normal distribution. However, many of the features in this dataset did not have a normal distribution. Often, they resemble more of a bimodal distribution with some values a few thousand times as high as others. To resolve this issue, the min max scaling was used instead: where is a feature and is its index. In the rest of this report, a scaling transformation to the features would refer to a min max scaling.

# Feature selection and analysis

There were two major feature selection methods used: selecting K best out of chi-squared feature selection and Principal Component Analysis (PCA)

5-fold cross-validation and grid search was used to optimize for the K in chi-squared feature selection and the number of components in PCA. The model evaluation used was accuracy, or the ratio of total correct predictions over the total predictions. The optimal K was 100, and the optimal number of component was 5 on the original dataset and 60 on the scaled dataset. This would mean that instead of having 849 features, there would be only 100 from chi-squared method, and 5 features from PCA.

# Binary classification and model evaluation

Logistic regression is a most common method for binary classification that predicts the probability of an observation being 1. In this case, we are predicting the probability of a customer being dissatisfied. Usually, a threshold of 50% or 0.5 is used to classify 1 from 0. However, a higher or lower threshold may lead to better accuracy or model performance. The Receiver Operating Characteristics (ROC) curve is a good way of showing the change in model performances as the probability threshold or discrimination threshold changes. The curve plots the true positive rate (TPR) versus the false positive rate (FPR) at various probability thresholds. TPR is computed as , or the fraction of all actual positives that were predicted to be positive. FPR is computed as , or the fraction of all the actual negative values that were wrongly predicted as positive. The true-positive rate is also known as sensitivity, recall or probability of detection, while the false-positive rate is also known as the fall-out or probability of false alarm. If the classification is perfect, the TPR would be 1 when FPR is 0, and the area under the ROC curve would be 1. The higher the area under the ROC curve (AUC), the better the classification result is.

In order to benchmark the performance of our model, I first generated a randomized set of features. All the predicted values were the same, but all the features were resampled without replacement, so there shall be no relationship between the predictive features and the predicted values. When I split this randomized dataset into a training set and a testing set and fit a logistic regression to the training set, the resulting AUC was 0.49, or less than 0.5. The average of AUC from running a 5-fold cross validation was still 0.50. This would be expected, as in a random scenario, the probability of guessing the head or tail correctly from flipping a coin would be 50%.

In addition to logistic regression, there are a few methods for binary classification, namely Naïve Bayes and Support Vector Machine (SVM) that are similar because they all use one single linear boundary. Decision tree based ensemble models are different in that each tree partitions the feature space into half-spaces with each split and resulting in multiple linear decision boundaries. Those models are also used to compare with the results from logistic regression. The testing set mean AUC from 5-fold cross validation of different models and different feature space is summarized in the table below.

Table 1 Area under the ROC curve for the testing set of various model and feature selection combinations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Original features | Scaled features | Chi-squared chosen top 100 scaled features | PCA chosen top 5 original features | PCA chosen 60 scaled features |
| Logistic Regression | 0.599 | 0.786 | 0.774 | 0.595 | 0.776 |
| Random Forest | 0.686 | 0.670 | 0.705 | -- | 0.647 |
| XGBoost (max\_depth = 3) | 0.832 | 0.832 | 0.807 | -- | 0.804 |
| GradientBoost (max\_depth = 3) | 0.707 | 0.773 | 0.805 | -- | 0.803 |
| Bernoulli Naïve Bayes | 0.713 | 0.704 | 0.714 | -- | 0.757 |

Note that the maximum depth parameter in the XGBoost and GradientBoost models were optimized using grid search for values of 1,2, 3,4, 5, 10.

As shown in Table 1, the min max scale transformation significantly improved results of Logistic Regression, but not the other classifiers. This is expected because scaling would influence logistic regression more. Feature selection with Chi-squared or PCA didn’t improve result except of Bernoulli Naïve Bayes and XGBoost. Since there is built in bootstrapping resampling in ensemble trees, it is not surprising that feature selection was not helpful. Overall, the XGBoost had the best performances.

Even though the AUC metric is the best for XGBoost among all the classifiers used here, the recall score , or is about the same for XGBoost (0.49) and Logistic Regression (0.50), while the precision score () was lower for XGBoost ( 0.08) compared to Logistic Regression (0.11). These scores are computed for when both models had a probability threshold that maximizes its f1-score ( ). The XGBoost classifier used a threshold of 0.13, so observations with a probability of higher than 13% being dissatisfied customers were classified as dissatisfied customers. This probability threshold for logistic regression was 8%. These thresholds are small because there is only 4% of the raw dataset being dissatisfied customers. When I used bootstrap resampling to resample the same number of observations with a target of 0 and 1, the resulting optimal probability threshold for classifying dissatisfied customers became 67%. Nevertheless, the recall rate and f1-score were not different from without using the bootstrap resampling.

The sensitivity of our results to the way we defined categorical variables was tested by variable the minimum number of unique values for considering a variable as numerical. Using the same logistic regression model and the min max scaling transformation, the resulting AUC of ROC was similar for not distinguishing categorical versus numerical variables, using a threshold of 10 unique values, 20 and 30 unique values. The difference in AUC was within 0.003.

# Conclusion

Given an anonymized dataset with no information of observations and predictor variables, transformation of those predictor variables is crucial to building high performance predictive models. In this analysis, variables with less than 20 unique values out of 76K observations were considered to be categorical and transformed into binary formats. Overall, a simple logistic regression predicts dissatisfied customers of Santander’s at a reasonable accuracy level, outperforming Naïve Bayes and Random Forest models. The extreme gradient ensemble classifier has a better performance overall model performance in terms of AUC and f1-score, but the recall rate is no better than logistic regression. For Santander, recall is more important than precision or other measures, because the number of False Positives is less alarming than the number of False Negatives. Falsely identifying satisfied customers as dissatisfied and trying to improve service to those that were not going to leave would not hurt. However, falsely ignoring the dissatisfied customers, or identifying them as satisfied customers, would be much more hurtful.