Project Report

on

Santander Customer Transaction Prediction

Submitted by:
Prashant Nookala
Date: 08/12/2019

Contents

Executive Summary

Prashant Nookala is pleased to submit a Project report on Santander Customer Transaction Prediction. This document states the approach to the Data Set of customer transactions at Santander Bank provided by Edwisor and the algorithms used to predict if customer is satisfied or not with the given transactions.

Assumptions:

- All the variables provided is anonymous so looking at the data, I assume it is in the same scale.
- In target variable; 0 is No & 1 is Yes

1. Introduction

1.1 Problem Statement

The objective of this Case is to predict if the customer is satisfied or not based on the transactional data given by the Bank.

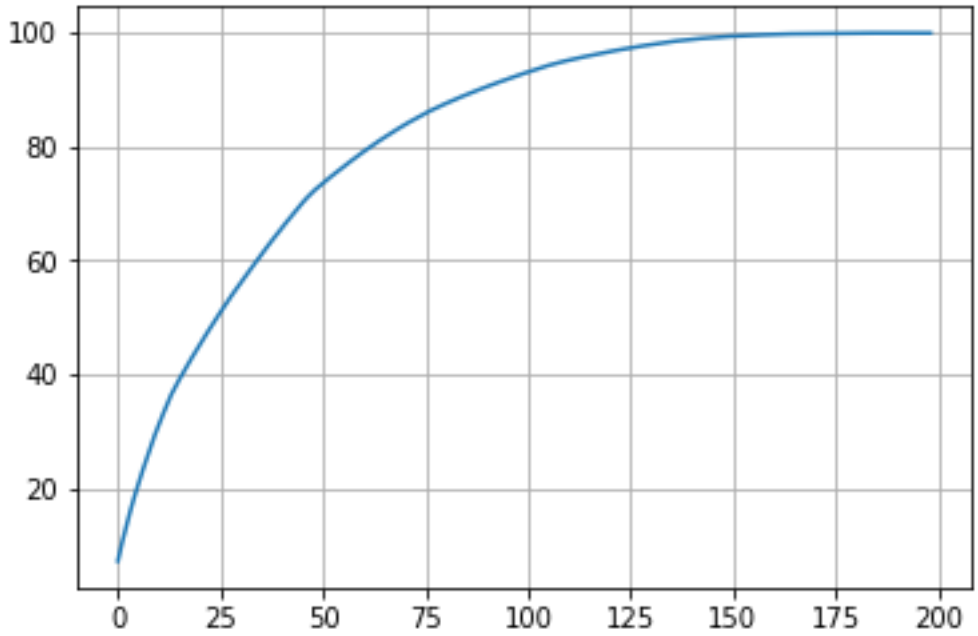
1.2 Data

Data was anonymous with 200 variables & a target and ID variable and has 200000 observations:

2. Methodology

2.1 Pre-Processing

As there were huge no. of variables, I used PCA to reduce the no. of variables. In the below graph we can see that 150 variables shows > 99% variance.

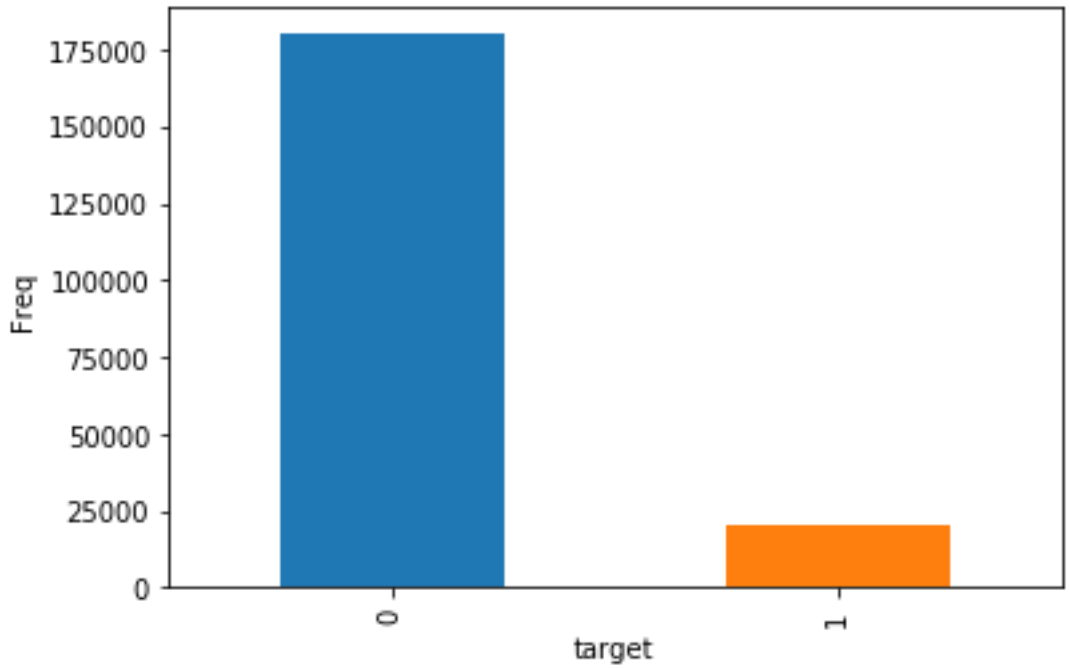


2.1.1 Missing Value analysis

After analysing the data, we can see that there are no missing values in this dataset

2.1.2 Distribution & Outlier Analysis

Distribution of target variable:



From the above figure we can say that 89.95% of customers are not satisfied or not interested in the product.

2.1.2 Feature Selection

As discussed in section 2.1.1, using Principal Component Analysis 150 components explains > 99% variance in the data and hence we have selected only 150 variables from 200. Remaining we don't need ID variable and target is our target variable.

2.2 Modeling

2.2.1 Model Selection

As we have our target variables as 1 (Yes) and 0 (No), we will be using classification models to predict the outcome. I have used stratified sampling technique to proportionally divide 1s & 0s to training and testing purposes.

Below are the models used and their error metrics:

I. **Logistic Regression:**
Accuracy = 58%
Recall = 88%
Precision = 18%
Specificity = 55%

II. **Decision Tree Classifier**
Accuracy = 83%
Recall = 18%
Precision = 17%
Specificity = 90%

III. **Random Forest Classifier**
Accuracy = 88%
Recall = 6%
Precision = 30%
Specificity = 98%

IV. **Naive Bayes**
Accuracy = 91%
Recall = 21%
Precision = 73%
Specificity = 99%

Note: I have tested these algorithms in Python and applied in both Python & R.

3. Conclusion

As we can see in the above results, although accuracy is high & specificity (TNR) is good for Naïve Bayes Classifier. We need to consider Specificity as the major metric because it calculates the ratio of how many cases are correctly predicted as negative compared to all negative cases in the training set. Here customer would be interested in knowing the TNR because, he would be more concerned on the non-satisfied customers & he can draw a strategy to retain them by providing better customer service.

3.1 Sample Data & Output

Naïve Bayes is run on the test.csv dataset with is transformed using PCA and the resulted value is stored in the target variable in the test dataset

4. Help for businesses

Customer satisfaction & retention is one of the major focus areas of the businesses. In this case, Santander Bank, who has large customers across USA and Europe offers several products and services to its customers. It is important for a business to understand the needs of customers and if they are satisfied with the services they are getting from the bank. By understanding the not satisfied customers, they can strategize their customer satisfaction and also give offers to the customers who are not likely to buy the product. This will help them in increasing their customer base and gaining more revenue.

With this project, we are helping Santander Bank to understand who will be the customers who are interested in buying a specific product/service or satisfied the services they are getting from the bank. It will allow Bank to understand why specific set of customers not satisfied with the service and they can put more efforts in retaining those customers who are not satisfied and inturn increasing their customer retention and satisfaction.