Timothy Chuang

Professor Mobasher

PHYS 247

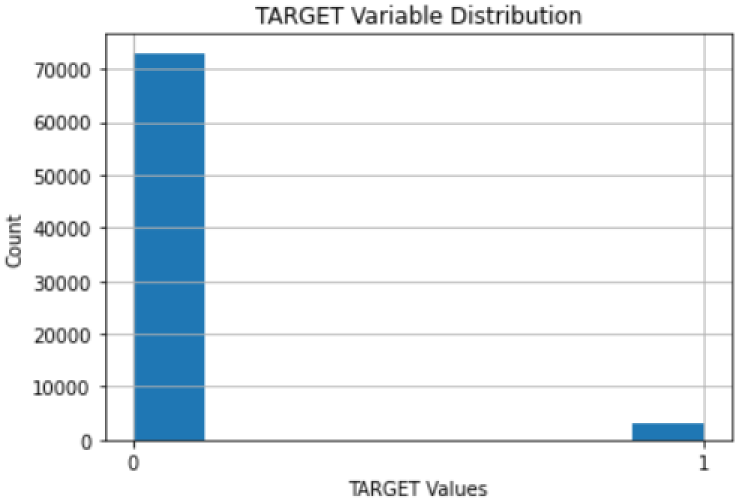
12/9/2022

Abstract

Our goal in the Santander Customer Satisfaction project is to build a model identifying satisfied and dissatisfied customers based on an anonymized dataset containing a large number of numeric variables. We are provided with two datasets; a TEST dataset that contains 370 features and a TRAIN dataset that consists of 370 features plus a TARGET column containing binary categorical data which represents satisfied/dissatisfied customers. Since there are only two possible categories of customer, we can identify the problem as a supervised binary-classification task. We used Logistic Regression to build our model, then trained and analyzed the efficiency of our model through a 70/30 train/test split within the TRAIN dataset before deploying the model on our TEST dataset.

Data Processing/ Cleaning

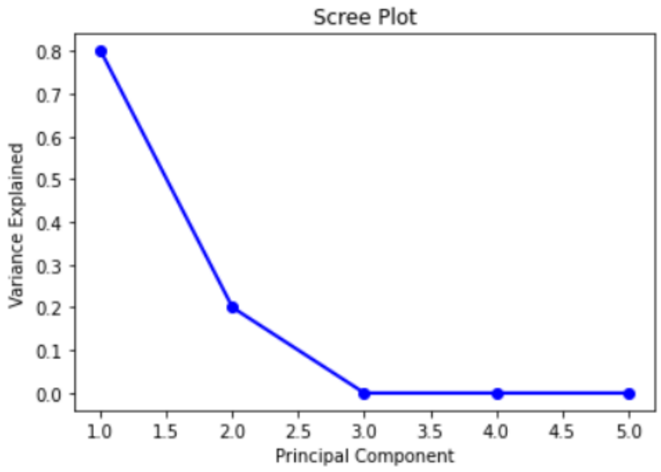
We are provided with two datasets, a TRAIN dataset with shape (76020, 371) and a TEST dataset with shape (75818, 370). The TRAIN dataset has an additional column because it contains our TARGET variable which will be used to train our model. To start our data cleaning process, we want to check for any null values in our dataset so that they can be replaced/removed. Next we want to make sure all of our columns are the correct types; we want all of our columns to be int64 or float64. Now that our data is clean we can run *test.head(50)* to gain an overview of our dataset by displaying the first 50 rows. We can find the dimensions of our dataset using *shape*, and find that our TRAIN set is 76020 x 371 and our TEST set is 75818 x 370. We then use *train.describe* to generate important statistics, and find that multiple columns have zero values for their 25th, 50th, and 75th percentiles but very large maximum values. This may suggest that these variables mostly consist of zero values, with very large outliers. Finally, we use a histogram plot to visualize the distribution of our TARGET variable.



We can see here that there are significantly more 0 values (satisfied customers) compared to 1 values (unsatisfied customers). We then calculate the proportions of TARGET and find that 96% of customers are satisfied while 4% of customers are dissatisfied.

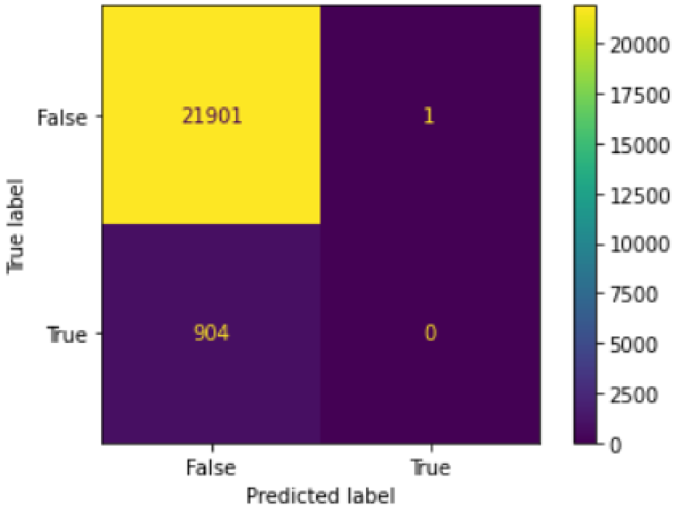
Model Building

Because the dataset we received was anonymized and there were a very large number of features, it would have been computationally expensive to perform feature selection based on correlation between features. Instead we opted to reduce the dimensionality of our dataset through PCA. We normalized our data using StandardScaler before performing PCA so that the PCA would not be affected by outliers. Initially we reduced the dimensionality of our dataset to 5, and produced the following Scree plot to measure variance dropoff.



We can see that PC3 and beyond accounts for zero percent of the variance, so we decided to change our dimensionality reduction to contain only three Principal Components.

Because of the binary nature of our TARGET variable, we chose to use Logistic Regression to build our Machine Learning model. We split our TRAIN dataset using a 70 / 30 train test split in order to train and evaluate our model. Using model.score, we evaluated a 96% model accuracy. To get a better understanding of our model's predictive power, we created a confusion matrix.



The results of the confusion matrix show that our model is very accurate at predicting Satisfied customers, but tends to mislabel dissatisfied customers as satisfied.