Santander Customer Transaction Prediction

Bhumika Dathiya

18th Sep, 2019

Abstract

Santander Bank wants to help identify dissatisfied customers early in their relationship. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late.

Contents

1 Introduction

1.1 Problem Statement

At Santander bank, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. As a data scientist, our aim is to help customers using their previous data

1.2 Data

Our task is to build classification models which will classify the customers who will buy the services of the bank depending on multiple factors. Given below is a sample of the data set that we are using to predict the customer target column:

2 Pre Processing

Data pre processing is an important step in data mining. The phrase 'garbage in, garbage out' is applicable in this concept. Data gathering methods are often loosely controlled.

Data preprocessing includes instance selection, cleaning, normalization, transformation, feature extraction,etc. The product of the data preprocessing is the final training set.

Tasks of data preprocessing:

1 Data Cleansing

2 Data Editing

3 Data Reduction

4 Data Wrangling

2.1 Outlier Analysis

In statistics, outlier is a data point that differs significantly from other observations. An outlier occur may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the dataset. An outlier can cause serious problems in data analysis.

There is no rigid mathematical definition of what constitues an outlier; determining whether or not an observation is an outlier is a subjective exercise.

There are various methods of an outlier detection such as normal probability plots, box plots. In our project, we have used box plot to detect

the outliers.

2.2 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Generally speaking, there are three main approaches to handle missing data:

(1) Imputation—where values are filled in the place of missing data

(2) Omission—where samples with invalid data are discarded from further analysis

(3) Analysis—by directly applying methods unaffected by missing values.

2.3 Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Agenda of Feature selection:

To identify and remove irrelevant attributes from the dataset that do not contribute much information about the target variable.

Because fewer attributes reduces complexity.

Two methods:

1 Dimesionality reduction

2 Curse of Dimesionality

Correlation Analysis is used to determine association between two continuous variables.

Chi Square test is used to determine dependencies between two categorical testing using Hypothesis testing.

2.4 Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. It is performed only on continuous variables. This can be achieved using 2 methods:

1 Normalization: In statistics and applications of statistics, normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging. In our project, we have

used histogram plot to determine the scale of the variables.

Converts all the data values between 0-1. Sensitive to outliers.

2 Standardization/ Z-score: In statistics, standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable. If the data is uniformly distributed then it is adopted.

If Z-score is negative, then the raw score is below mean.

If Z-score is positive, then the raw score is above mean.

3 Modeling

Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Types of learning algorithms

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

1 Supervised ML(With target variable)

Regression(Continuous Variable)

Classification(Categorical Variable)

2 Unsupervised ML (No target variable)

Clustering

3 Recommender systems

Collaborative Filtering

3.1.1 Logistic Regression Introduction

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Logistic regression can be binomial, ordinal or multinomial.

It has some assumptions:

1 Independent of errors.

2 No outliers.

3 Absence of Multicollinearity.

4 Data normalization.

3.1.2 Model Evaluation

Accuracy: 89.54

FNR: 52.37

Recall: 48.53

Alkaline info criteria: 1245.7

3.1.1 KNN model Introduction

The model representation for KNN is the entire training dataset.KNN makes predictions using the training dataset directly.Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this is the mean output variable, in classification this is the mode (or most common) class value.

KNN works well with a small number of input variables (p), but struggles when the number of inputs is very large. As the number of dimensions

increases the volume of the input space increases at an exponential rate, but this general problem is called the “Curse of Dimensionality“.. KNN performs much better if all of the data has the same scale. Normalizing your data to the range [0, 1] is a good idea.

3.2.2 Model Evaluation

Accuracy: 90.64

FNR: 61.09

Recall: 39.91

3.3.1 Naive Bayes Introduction

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

3.3.2 Model Evaluation

Accuracy: 92.16

FNR: 42.71

Recall: 58.29

4 Conclusion

4.1 Model Acceptance and Rejection

As we have calculated the error metrics of Logistic Regression model, KNN model and Naive Bayes model, we have observed that the accuracy of Naive Bayes model is high as compared with the accuracy of other two models.

Also, the FNR is less of Naive Bayes model. In this case, we have high recall rate.

So, we have freezed Naive Bayes model for this problem.

4.2 Effectiveness of Model

Using a generative model such as Naive Bayes makes dealing with missing values a lot easier.

Very simple, easy to implement and fast.

If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.

Even if the NB assumption doesn't hold, it works great in practice.

Need less training data.

Highly scalable. It scales linearly with the number of predictors and data points.

Can be used for both binary and mult-iclass classification problems.

Can make probabilistic predictions.

Handles continuous and discrete data.

Not sensitive to irrelevant features.

Using this model can provide a quite accurate assumptions about the customers because this model provides the less false predictions of the customers and lot of correct predictions about the customers.