Problem Statement:

In this problem, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

Training data is an anonymous data set containing 200 numeric feature variables, the binary target column, and a string ID_code column and 2,00,000 observations. Test data includes 200 anonymous numeric variables and a string ID_code column and 2,00,000 observations. This is a binary classification problem under supervised machine learning algorithm. The task is to predict the value of target column in the test set.

General Business Significance

This project can help company in following ways-

1. Segmenting customers into small groups and addressing individual customers based on actual behaviors — instead of

hard-coding any preconceived notions or assumptions of what makes customers similar to one another, and instead of only looking at aggregated data which hides important facts about individual customers.

2. Accurately predicting the future behavior of customers (e.g., transaction prediction) using predictive customer behavior modeling techniques — instead of just looking in the rear-view mirror of historical data.

3. Using advanced calculations to determine the customer lifetime value (LTV) of every customer and basing decisions on it — instead of looking only at the short-term revenue that a customer may bring the organization.

4. Knowing, based on objective metrics, exactly what marketing actions to do now, for each customer, in order to maximize the long-term value of every customer.

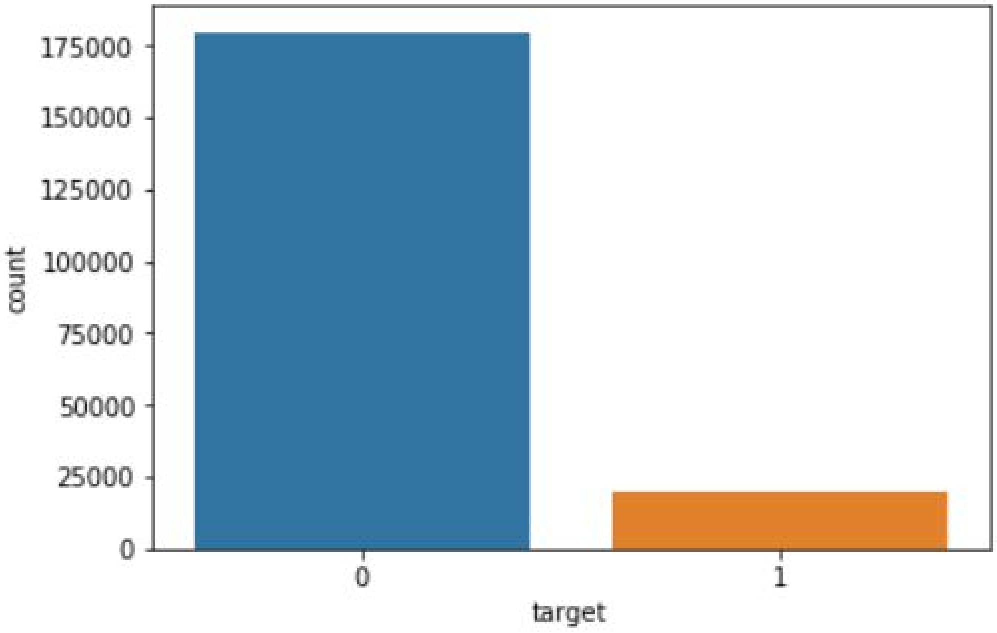
5. Using marketing machine learning technology that will reveal insights and make recommendations for improving customer

marketing that human marketers are unlikely to spot on their own.

Exploratory Data Analysis

Exploratory data analysis mainly includes missing value analysis, outlier analysis, correlation analysis, descriptive analysis and visualizations to gain insights from data.

First Let’s check the amount of minority and majority class in the training data.



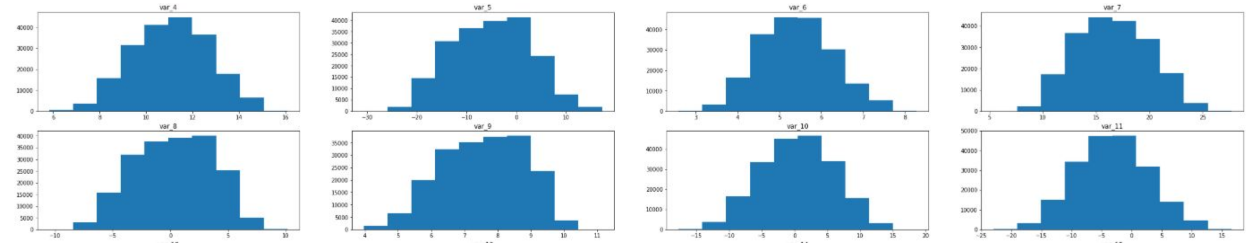
Dealing with Imbalanced data set

Before modelling for this data set let us understand how to deal with imbalanced data set for classification problem. Traditional Machine Learning algorithms tend to produce unsatisfactory classifiers when faced with imbalanced data sets. For any imbalanced data set, if the event to be predicted belongs to the minority class and the event rate is less than 10%, it is usually referred to as a rare event. The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced data sets. Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have large number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of mis-classification of the minority class as compared to the majority class. Performance of classification algorithm is measured by the Confusion Matrix which contains information about the actual and the predicted class. Thus we need to deal with this imbalanced data set.

Here are some methods to deal with imbalanced data for classification.

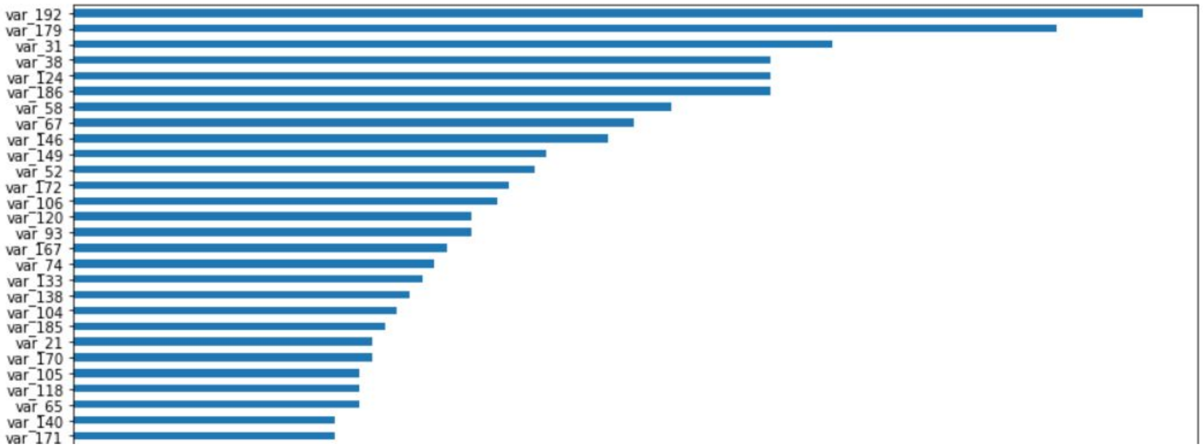
There is no missing values in both train and test data.

Let us check the distribution of first few numerical features in train data by plotting histograms of each variable.

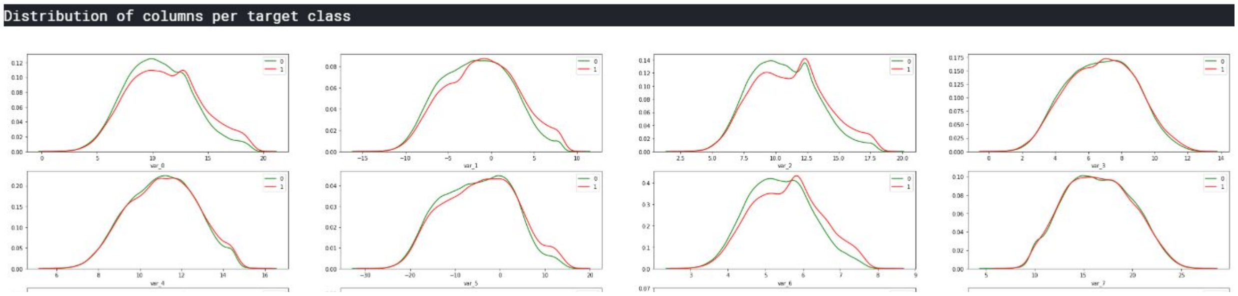


Looking at the shapes of histograms we can easily conclude that almost all numeric variables follow a normal distribution.

I have decided to see if there are any outliers in the data set according to Chauvenet’s criterion. After removing outliers (0.87% of total observations) we have 1,98,264 observations in train data and 1,98,250 observations in test data.



Now let’s check the distribution of first few variables in train data with both of the target classes.



```
Distribution of columns per target class
```