Customer Transaction Prediction

Ambika Iragappa

Contents

Introduction

Problem Statement

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

At Santander, 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. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as:

- is a customer satisfied?
- Will a customer buy this product?
- Can a customer pay this loan?

According to past data and from the given problem the output is Classification and it comes under Supervised Machine Learning. We train the model with past data and when the new data is given, we predict the outcome

Data

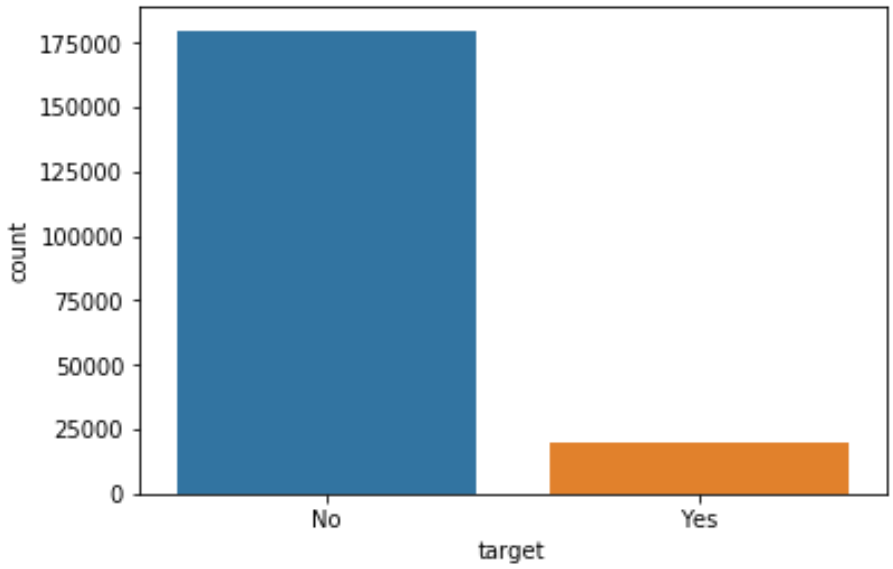
Given data contains numeric feature variables, the binary target column, and a string ID_code column. The task is to predict the value of target column in the test set.

- ID_code (string);
- Target;
- 200 numerical variables, named from var_0 to var_199;
- It has 201 predictors or independent variables and 1 target variable 'target'

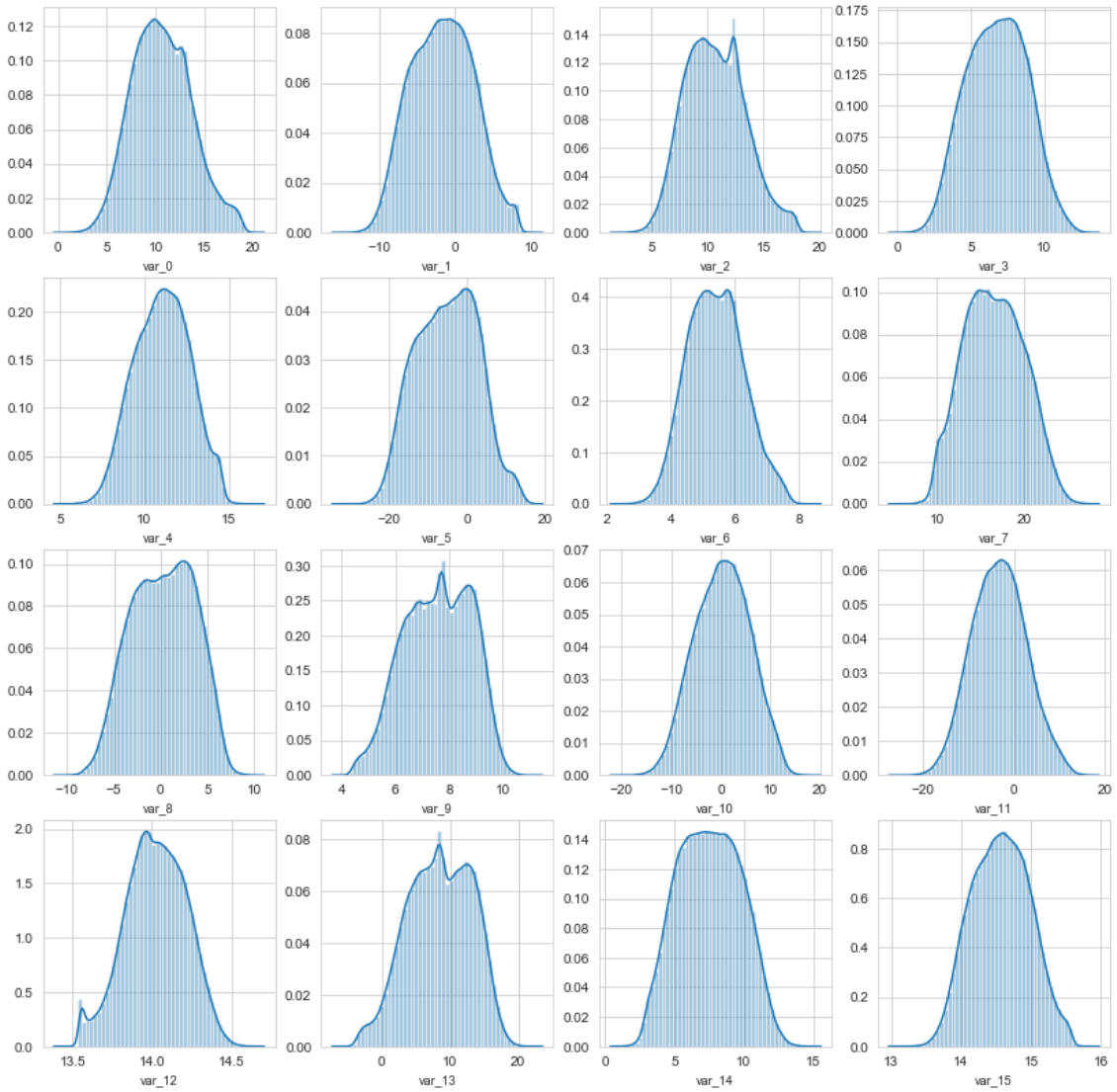
Methodology

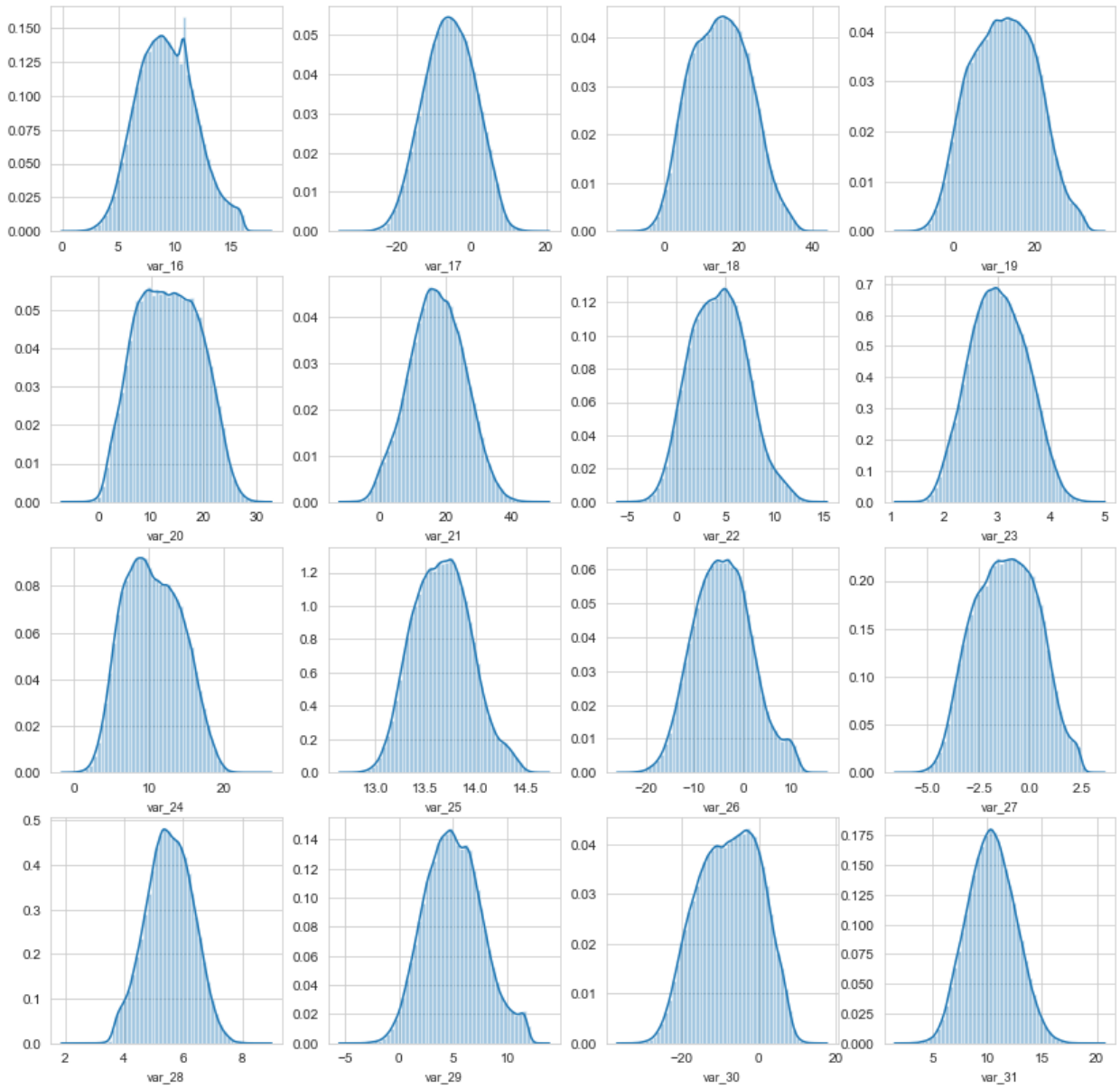
Pre-Processing

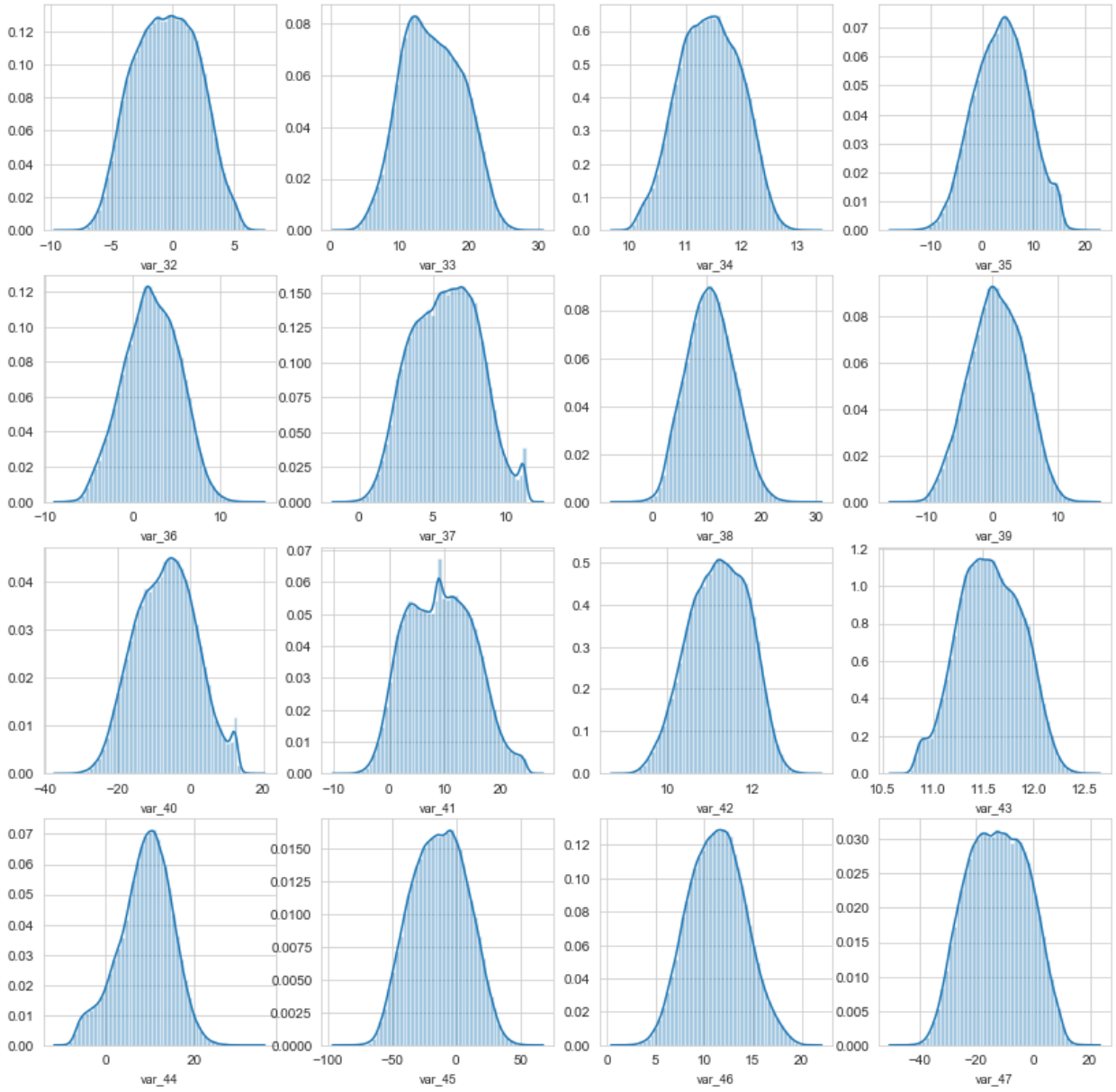
Checking the Distribution of target variable:



Checking the distribution of the independent variables




var_16
var_17
var_18
var_19
var_20
var_21
var_22
var_23
var_24
var_25
var_26
var_27
var_28
var_29
var_30
var_31


var_32
var_33
var_34
var_35
var_36
var_37
var_38
var_39
var_40
var_41
var_42
var_43
var_44
var_45
var_46
var_47

MISSING VALUE ANALYSIS:

Missing values are which, where the values are missing in an observation in the dataset. It can occur due to human errors, individuals refusing to answer while surveying, optional box in questionnaire.

IN OUR GIVEN DATASET WE ARE NOT HAVING ANY MISSING VALUES….SO WE ARE NOT PROCEEDING WITH ANY STEPS TO IMPUTE ANY MISSING VALUES.

OUTLIER ANALYSIS:

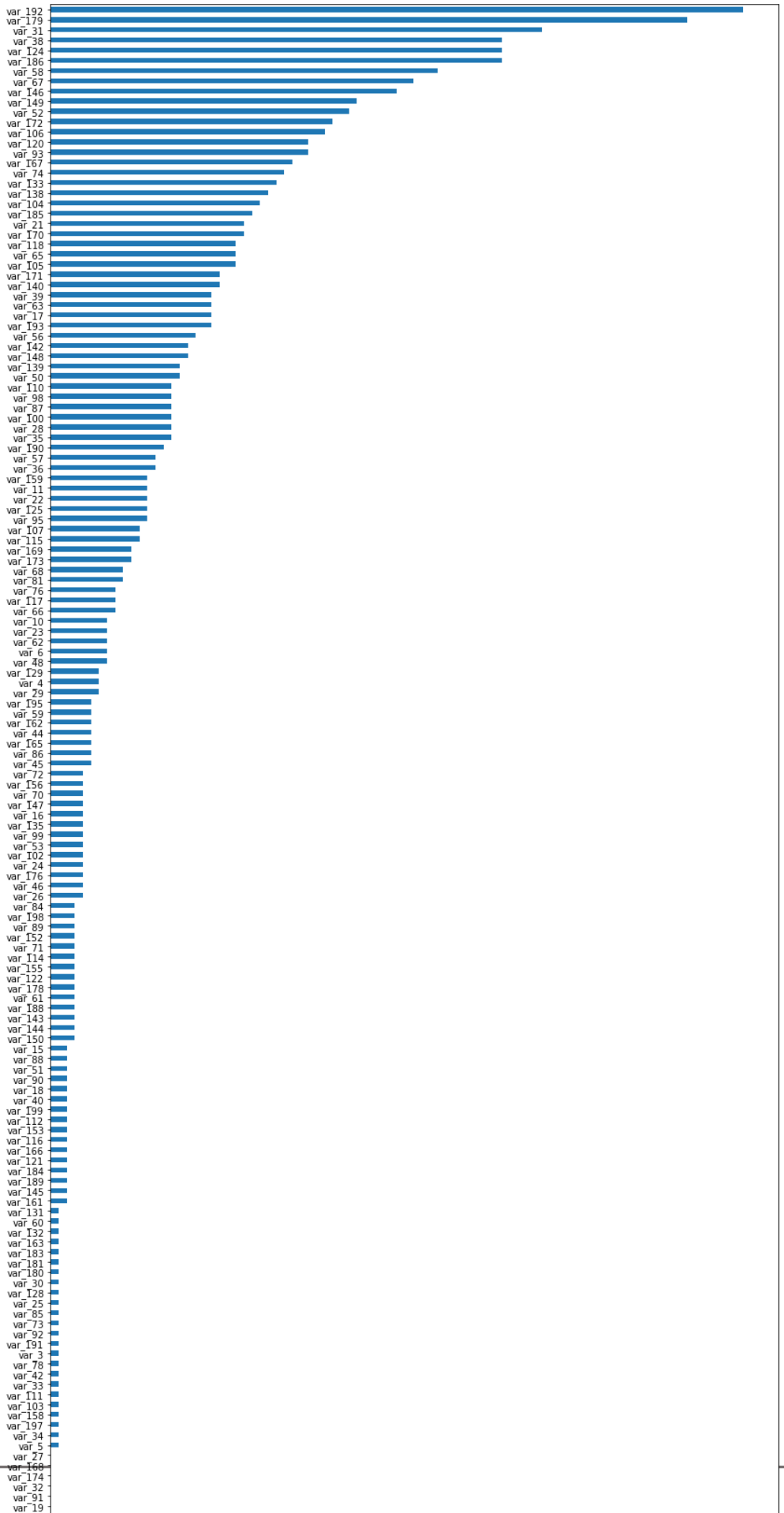
Inter Quartile Range (IQR) is calculated and the minimum and maximum value are calculated for the variables. Any value ranging outside the minimum and maximum value are discarded.

IQR = q75 - q25

min = q25 - (IQR*1.5)

max = q75 + (IQR*1.5)

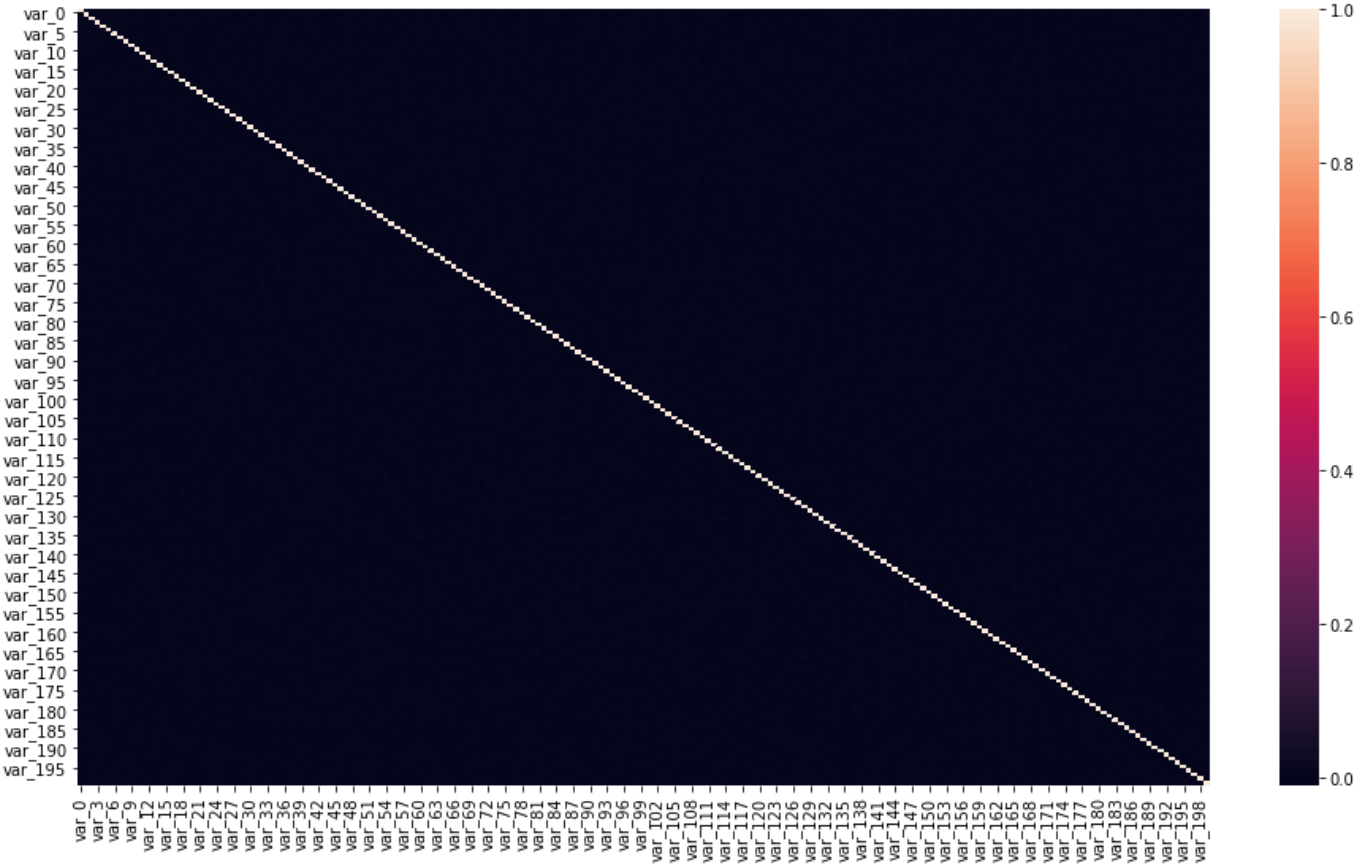
Customer transaction Prediction



FEATURE SELECTION

CORRELATION ANALYSIS:

Before performing any type of modeling, we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. This process of selecting a subset of relevant features/variables is known as feature selection. There are several methods of doing feature selection. I have used correlation analysis in our dataset, the correlation between the train attributes is very small. So, there is no need to remove variables



HANDLING IMBALANCED DATA:

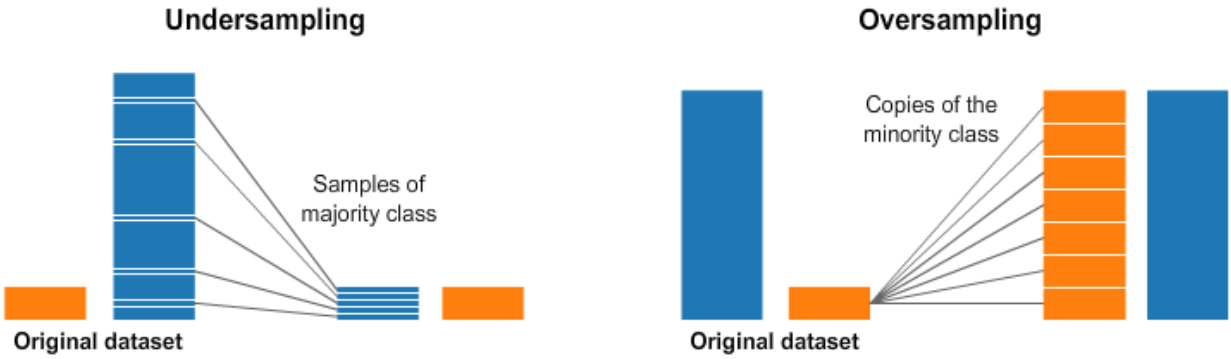
Imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. Class imbalance can be found in many different areas including medical diagnosis, spam filtering, and fraud detection. Some popular methods for dealing with class imbalance.

Change the performance metric Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be very misleading. Metrics that can provide better insight include:

- Confusion Matrix: a table showing correct predictions and types of incorrect predictions.
- Precision: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- Recall: the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- F1: Score: the weighted average of precision and recall.

Resampling Technique:

A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and / or adding more examples from the minority class (over-sampling).



Despite the advantage of balancing classes, these techniques also have their weaknesses (there is no free lunch). The simplest implementation of over-sampling is to duplicate random records from the minority class, which can cause overfitting. In under-sampling,

the simplest technique involves removing random records from the majority class, which can cause loss of information.

we can cluster the records of the majority class, and do the under-sampling by removing records from each cluster, thus seeking to preserve information. In over-sampling, instead of creating exact copies of the minority class records, we can introduce small variations into those copies, creating more diverse synthetic samples

Model Evaluation

Classification Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

Accuracy=Total correct prediction/Total no of prediction

It works well only if there are equal number of samples belonging to each class. For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get 98% training accuracy by simply predicting every training sample belonging to class A. When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the test accuracy would drop down to 60%.

Classification Accuracy is great, but gives us the false sense of achieving high accuracy. The real problem arises, when the cost of misclassification of the minor class samples are very high. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests.

Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model. 14 Let's assume we have a binary classification problem. We have some samples belonging to two classes: YES or NO. Also, we have our own classifier which predicts a class for a given input sample. On testing our model on 165 samples, we get the following result.

There are 4 important terms:

- True Positives: The cases in which we predicted YES and the actual output was also YES.
- True Negatives: The cases in which we predicted NO and the actual output was NO.
- False Positives: The cases in which we predicted YES and the actual output was NO.
- False Negatives: The cases in which we predicted NO and the actual output was YES.

Area Under Curve

Area Under Curve (AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining AUC, let us understand two basic terms:

- True Positive Rate (Sensitivity): True Positive Rate is defined as TP/ (FN+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

- False Positive Rate (Specificity): False Positive Rate is defined as FP / (FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. False Positive Rate and True Positive Rate both have values in the range [0, 1]. FPR and TPR both are computed at threshold values such as (0.00, 0.02, 0.04, …., 1.00) and a graph is drawn. AUC is the area under the curve of plot False Positive Rate vs True Positive Rate at different points in [0, 1].

F1 Score

F1 Score is used to measure a test's accuracy F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as:

F1 Score tries to find the balance between precision and recall.

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

• Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Precision = TP / TP + FP

• Recall: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

Recall= TP/TP+FN

MODEL DEVELOPMENT

LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Confusion Matrix:
[[35498 500]
[2954 1048]]

Accuracy: 0.913650

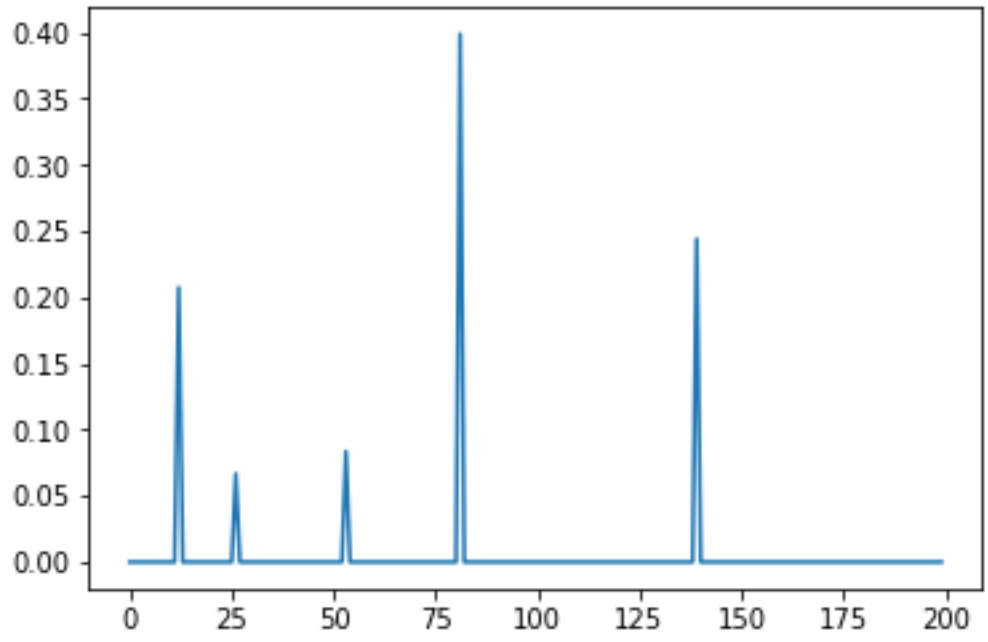
precision: [0.92317695 0.67700258]
recall: [0.98611034 0.26186907]
fscore: [0.95360645 0.37765766]

Decision Tree

Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

Accuracy score 0.899800
Confusion Matrix:
[[35969 29]
[3979 23]]

precision: [0.90039551 0.44230769]
recall: [0.9991944 0.00574713]
fscore: [0.94722566 0.01134682]



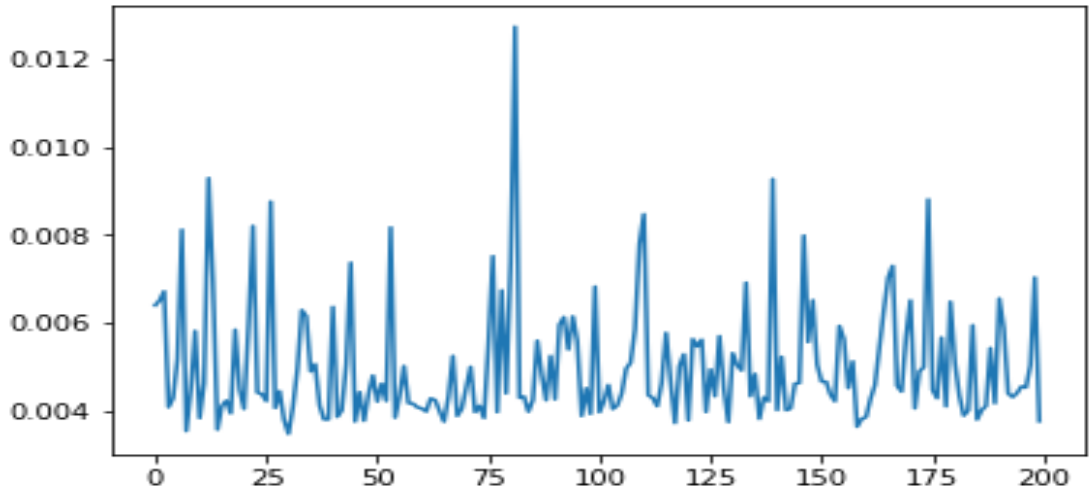
Decision tree feature importance

RANDOM FOREST

Random forests are based on a simple idea: 'the wisdom of the crowd'. Aggregate of the results of multiple predictors gives a better prediction than the best individual predictor. A group of predictors is called an ensemble. Thus, this technique is called Ensemble Learning. To improve our technique, we can train a group of Decision Tree classifiers, each on a different random subset of the train set. To make a prediction, we just obtain the predictions of all individuals trees, then predict the class that gets the most votes. This technique is called Random Forest. Random forest chooses a random subset of features and builds many Decision Trees. The model averages out all the predictions of the Decisions trees.

Accuracy score 0.900475
[[35992 6]
[3975 27]]

precision: [0.90054295 0.81818182]
recall: [0.99983332 0.00674663]
fscore: [0.94759429 0.0133829]



Random Forest feature importance

Reference

1. For Data Cleaning and Model Development - https://edwisor.com/career-data-scientist
2. For other code related queries - https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis- python/
3. For Visualization –
4. https://towardsdatascience.com/
5. https://stackoverflow.com/