# Problem Statement: Fraud Detection

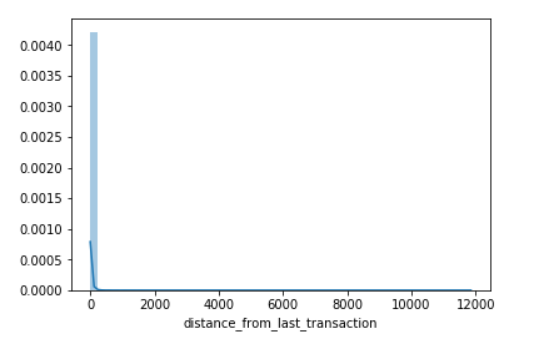
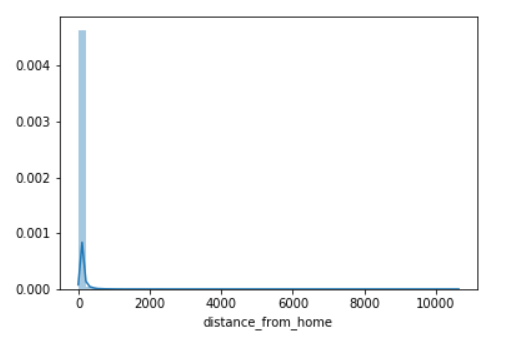
According to the Data Breach Index, more than 5 million records are being stolen on a daily basis, a concerning statistic that shows - fraud is still very common both for Card-Present and Card-not Present type of payments.

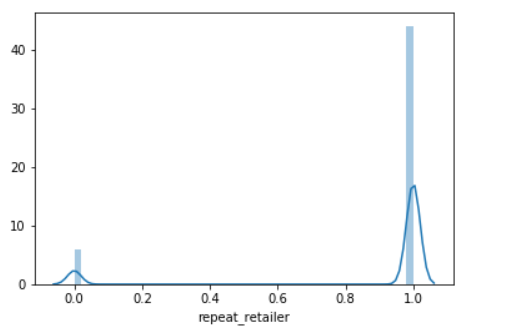
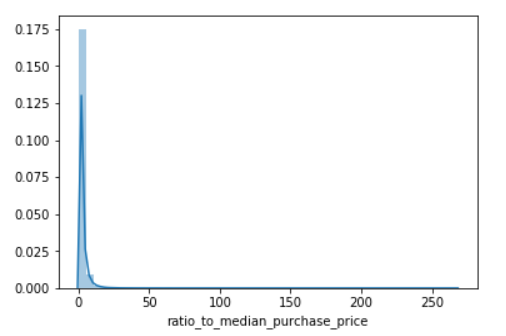
To construct a ML model for industry (bank- payment gateways or VISA / Mastercard) in categorization & authorization of transaction based on efficiency in **fraud detection**.

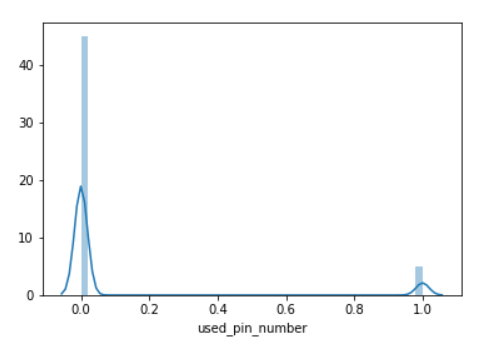
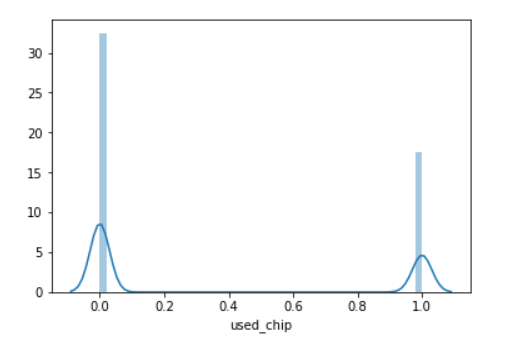
# Exploratory Data Analysis

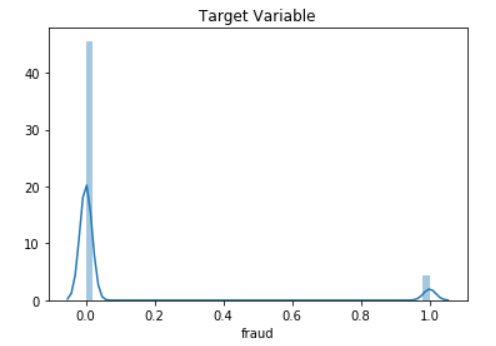
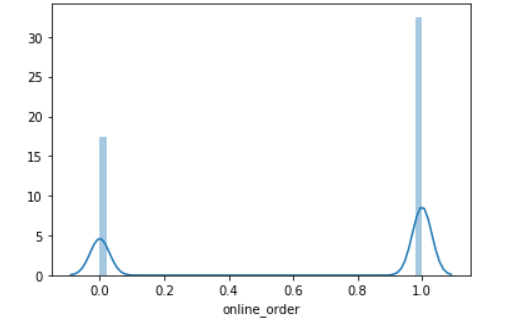
## Data Distribution

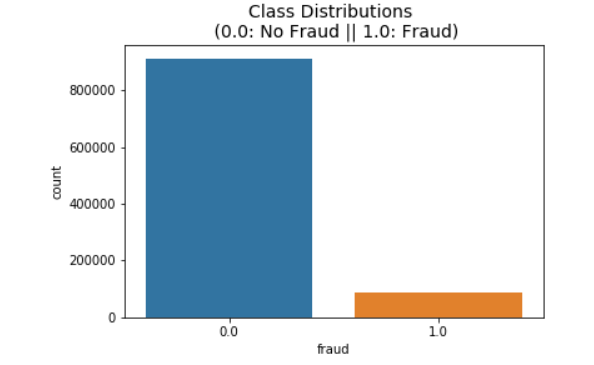
Below graph will show how the data is distributed among each variable











## Understanding the data

From Above graphs we have concluded that our data is highly imbalanced

1. We have 7 Independent Variables and 1 Dependent variable (Fraud)
2. We have float values for dependent variables
3. We have no null values
4. We have no duplicate values
5. We don’t have normal distribution of variables that is data is skewed but Random Forest can handle data without normalization
6. We have Highly imbalanced target class

i.e. **91 .6% Non-Fraudulent cases** and **8% Fraudulent cases**

## Correlation between Variables



# Model

We have tried two model

(1)Random Forest and

(2)Logistic Regression

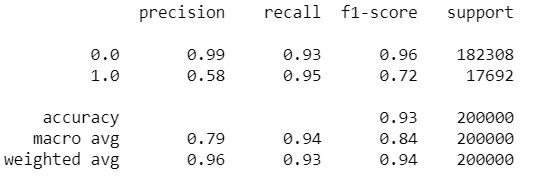
Creating a Train and Test Split..!

To make sure our model is not overfitted on training data we will split our data in

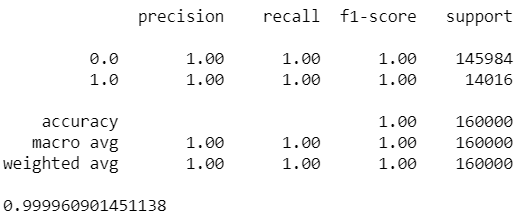
(1)Train -80%

(2)Test -20%

## Logistic Regression



## Random Forest



**Confusion Matrix for Random Forest**



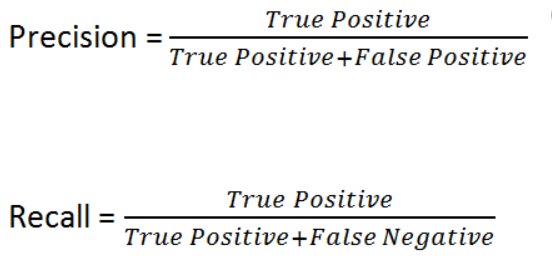
As our Target Variable is highly imbalanced we will need something more than accuracy .

# Precision and Recall

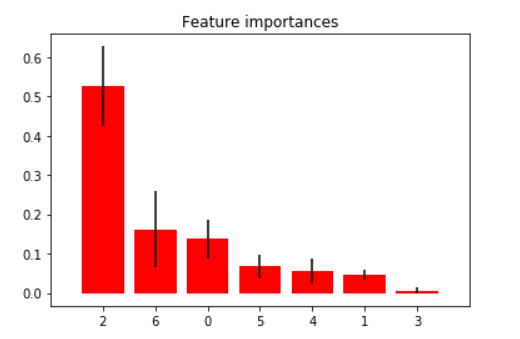
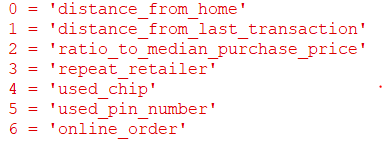
**(1)For the credit card fraud detection case recall will be a better option because the number of normal transactions are very high as compared to the number of fraud cases**

**(2)Recall will give us a sense of only fraud cases. Now, it is possible to build our model with 100% recall but the downside will be our precision will be worst and that will result into nothing but useless model.**

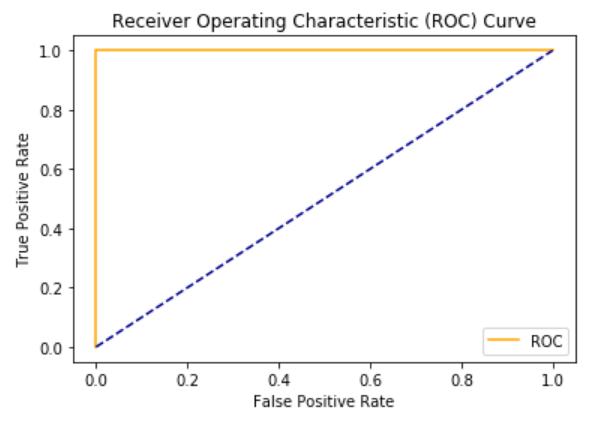
**(3)Thus our aim is to attain high recall with maintaining precision.**



# Feature Importance Created from Random Forest

# Region Under Curve



# Suggest ways of improving the model

## Issue1: Not Normally Distributed Data

As our data was not Normalized Logistic Regression model does not perform well but Logistic Performs can be improved by normalization of variables.

As we got good accuracy in Random Forest we have not use normalization of data

## Issue 2:Highly Imbalanced Data

(1) For that we have different techniques like oversampling, under sampling and SMOTE(Synthetic Minority Oversampling Technique) to create balance in our data.

(2) **Oversampling** can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model.

This can balance the class distribution but does not provide any additional information to the model.

(3) With **Undersampling** or **Downsampling**, the main disadvantage is that it will lose a lot of useful information from the dataset as it is removing more than 99% of non-fraudulent cases from the dataset. Thus, this should not be a practical approach here.

**SMOTE** uses a nearest neighbours algorithm to generate new and synthetic data we can use for training our model. The synthetic points are added between the chosen point and its neighbours. So it avoids adding duplicates in the data.

In the case of our model, we used SMOTE

**We have saved the model using pickle in .sav format for further analysis and attached the model along with the code**