### Exploratory data analysis

# Checking the class distribution of the target variable in percentage

A blue circle with orange line in the middle

Description automatically generated A blue and white rectangle

Description automatically generated

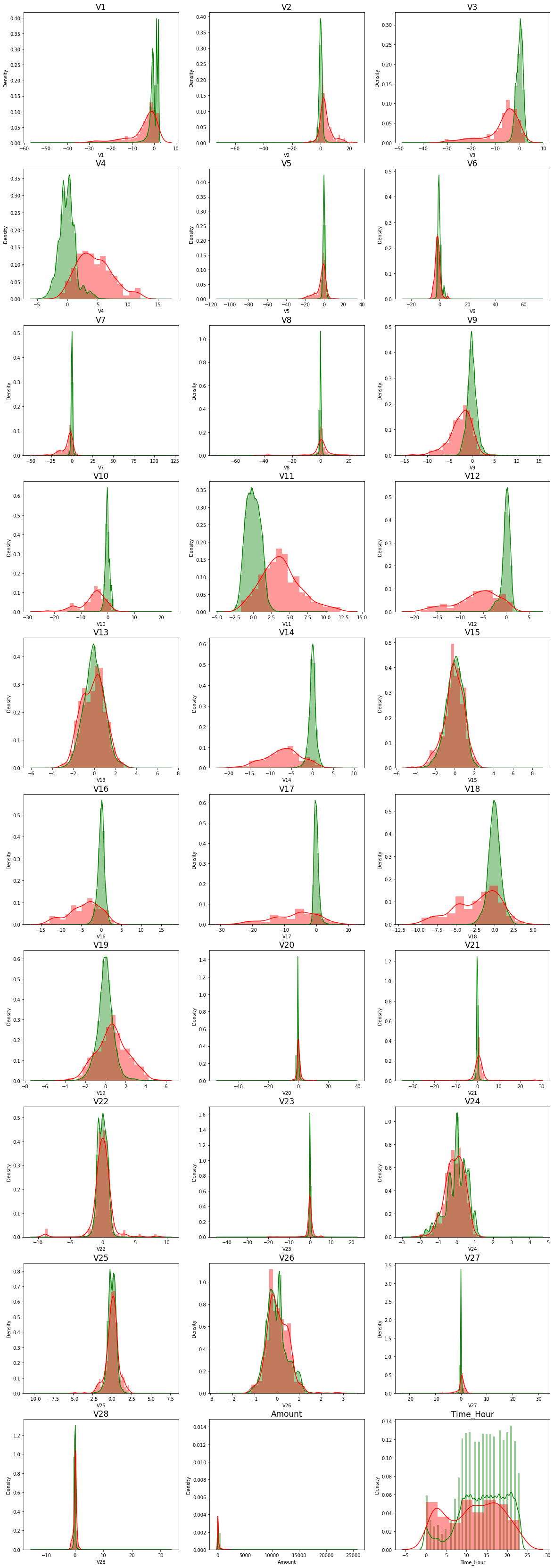
# Checking the correlation in heatmap

A screenshot of a computer screen

Description automatically generated

### Splitting the data into train & test data

# plot the histogram of a variable from the dataset to see the skewness



### Model Building

Model chosen:

1. Logistic Regression with L1 And L2 Regularisation
2. KNN model
3. Tree model
4. Random Forest
5. XGBoot Model
6. SVM model

Methodology:

1. Confusion matrix, auc/roc, threshold, accuracy value

### Model Validation

#Lets perfrom RepeatedKFold and check the results

**Logistic Regression** with L1 And L2 Regularisation

Max auc\_roc for l1: 0.9757741295537155

Max auc\_roc for l2: 0.9865741449266722

Accuarcy of Logistic model with l2 regularisation : 0.9989642035778866

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.82 0.53 0.64 100

accuracy 1.00 56961

macro avg 0.91 0.76 0.82 56961

weighted avg 1.00 1.00 1.00 56961

Accuarcy of Logistic model with l1 regularisation: 0.9989290918347641

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.78 0.54 0.64 100

accuracy 1.00 56961

macro avg 0.89 0.77 0.82 56961

weighted avg 1.00 1.00 1.00 56961

l2 roc\_value: 0.9690107455021895

l2 threshold: 0.0014225337256554363

ROC for the test dataset 96.9%

A graph on a white background

Description automatically generated

l1 roc\_value: 0.869288616098908

l1 threshold: 0.0472303427873661

ROC for the test dataset 86.9%

A graph with a line

Description automatically generated with medium confidence

Time Taken by Model: --- 344.7267997264862 seconds ---

------------------------------------------------------------

**KNN Model**

model score

0.9992626533944278

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.90 0.65 0.76 100

accuracy 1.00 56961

macro avg 0.95 0.82 0.88 56961

weighted avg 1.00 1.00 1.00 56961

KNN roc\_value: 0.8648265946782504

KNN threshold: 0.2

ROC for the test dataset 86.5%

A graph of a person with a blue line

Description automatically generated

Time Taken by Model: --- 1256.2385637760162 seconds ---

------------------------------------------------------------

**Decision Tree Models** with 'gini' & 'entropy' criteria

gini score: 0.9992450975228665

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.81 0.75 0.78 100

accuracy 1.00 56961

macro avg 0.90 0.87 0.89 56961

weighted avg 1.00 1.00 1.00 56961

gini tree\_roc\_value: 0.8748417192803503

Tree threshold: 1.0

ROC for the test dataset 87.5%

A graph with a blue line

Description automatically generated

entropy score: 0.9990870946788153

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.74 0.75 0.74 100

accuracy 1.00 56961

macro avg 0.87 0.87 0.87 56961

weighted avg 1.00 1.00 1.00 56961

entropy tree\_roc\_value: 0.8747625789205256

Tree threshold: 1.0

ROC for the test dataset 87.5%

A graph with a blue line

Description automatically generated

Time Taken by Model: --- 37.0897114276886 seconds ---

------------------------------------------------------------

**Random Forest Model**

Model Accuracy: 0.999490879724724

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.97 0.73 0.83 100

accuracy 1.00 56961

macro avg 0.99 0.86 0.92 56961

weighted avg 1.00 1.00 1.00 56961

Random Forest roc\_value: 0.9328157260688346

Random Forest threshold: 0.01

ROC for the test dataset 93.3%

A graph with a blue line

Description automatically generated

Time Taken by Model: --- 214.1870892047882 seconds ---

------------------------------------------------------------

**XGBoost Model**

Model Accuracy: 0.9994382121100402

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.93 0.74 0.82 100

accuracy 1.00 56961

macro avg 0.96 0.87 0.91 56961

weighted avg 1.00 1.00 1.00 56961

XGboost roc\_value: 0.9701069274194967

XGBoost threshold: 0.0014467865694314241

ROC for the test dataset 97.0%

A screen shot of a graph

Description automatically generated

Time Taken by Model: --- 53.274091482162476 seconds ---

------------------------------------------------------------

**SVM Model** with Sigmoid Kernel

accuracy\_score : 0.9980864099998245

Confusion Matrix

A green and grey squares with black text

Description automatically generated

classification Report

precision recall f1-score support

0 1.00 1.00 1.00 56861

1 0.09 0.01 0.02 100

accuracy 1.00 56961

macro avg 0.54 0.50 0.51 56961

weighted avg 1.00 1.00 1.00 56961

SVM roc\_value: 0.42160602170204536

SVM threshold: 0.002741365025150363

ROC for the test dataset 42.2%

A graph of a growth

Description automatically generated with medium confidence

Time Taken by Model: --- 633.070867061615 seconds ---

A screenshot of a computer

Description automatically generated

#Lets perfrom StratifiedKFold and check the results

A screenshot of a computer

Description automatically generated

### Model Optimization

*Proceed with the model which shows the best result*

* Apply the best hyperparameter on the model
* Predict on the test dataset

A green and blue rectangles

Description automatically generated

Hence it implies that V4, v5,V11 has + ve importance whereas V10, V12, V14 seems to have -ve impact on the predictaions

### Oversampling with RandomOverSampler with StratifiedKFold Cross Validation

A screenshot of a computer

Description automatically generated

### hyperparameter tuning

In the oversample cases, of all the models we build found that the XGBOOST model with Random Oversampling with StratifiedKFold CV gave us the best accuracy and ROC on oversampled data. Post that we performed hyperparameter tuning and got the below metrices :

XGboost roc\_value: 0.9815403079438694 XGBoost threshold: 0.01721232570707798

However, of all the models we created we found Logistic Regression with L2 Regularisation for StratifiedKFold cross validation (without any oversampling or undersampling) gave us the best result.