Introduction:

I have got 3 datasets Churn,Adult income and Credit Card. MY task is to build 2 Models Adaboost and Logistic Regression from Scratch.In order to Train the Features I have to Preprocess the datasets.So i am going to do a lot of data preprocessing to get the max Accurate results from models , I will create a general model which i can save and could use in future for other puproses.

Churn Dataset:

The Churn dataset has 7043 Rows and 21 features and The Target Columne is Churn.This is Binary Classification problem.The data set is pretty clean .There is no null values but there still alot of work to do on this dataset like remove outliers , Normalize the data and Balance the classes of target column.

Adult Income Dataset:

The Adult Income dataset have 32561 Rows and 15 features , The target colume is income.This is binary classification problem.There are 9 object columns and 6 numeric columns.

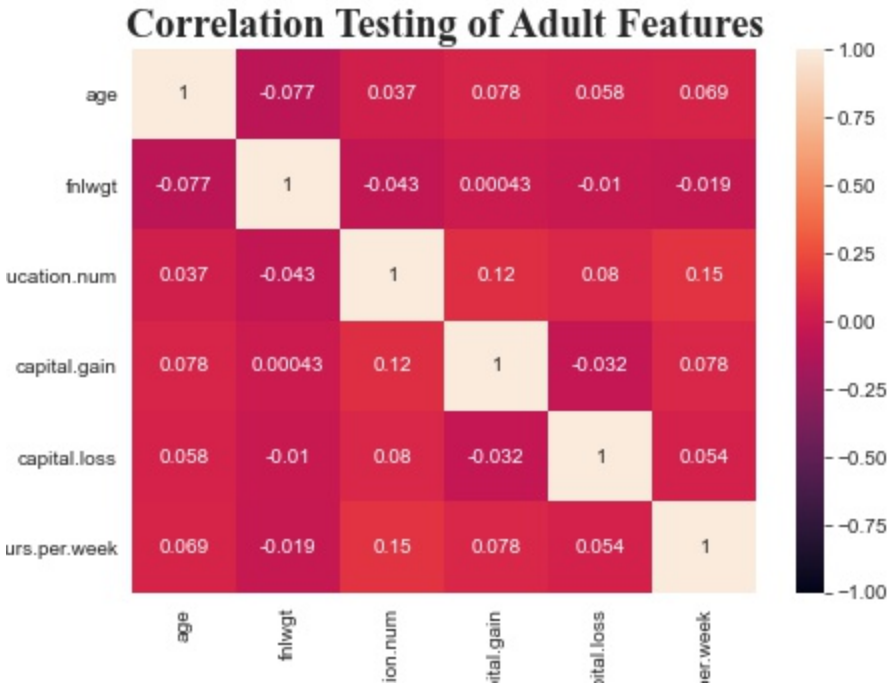
CreditCard Dataset:

The dataset contains 284807 Rows and 31 features and the target column is class.The dataset is very huge and the classes of target column are also imbalace so when i will over sample the minority class then the dataset will get bigger so i will use the the half part of the dataset.

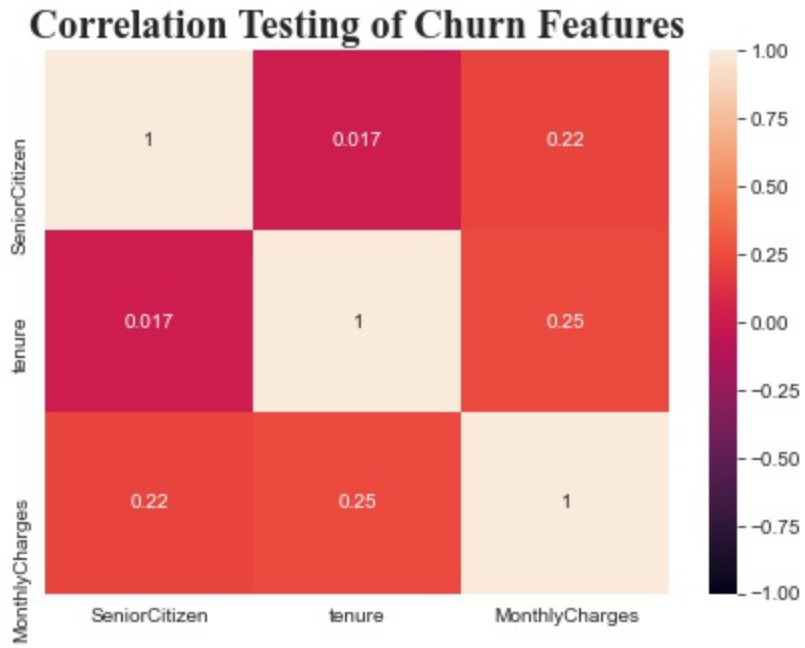
The Steps Which i will follow:

- Import the Datasets.
- Do some Exploratory data Analysis.
- Remove the Outliers.
- Labelize The Data.
- Split the dataset.
- Balance the Target Column Classes by Oversampling Them.
- Build The Adaboost and LogisticRegression model From Scratch.
- Train The Datasets with Model.
- Predict the Outcomes and check for Accuracy.

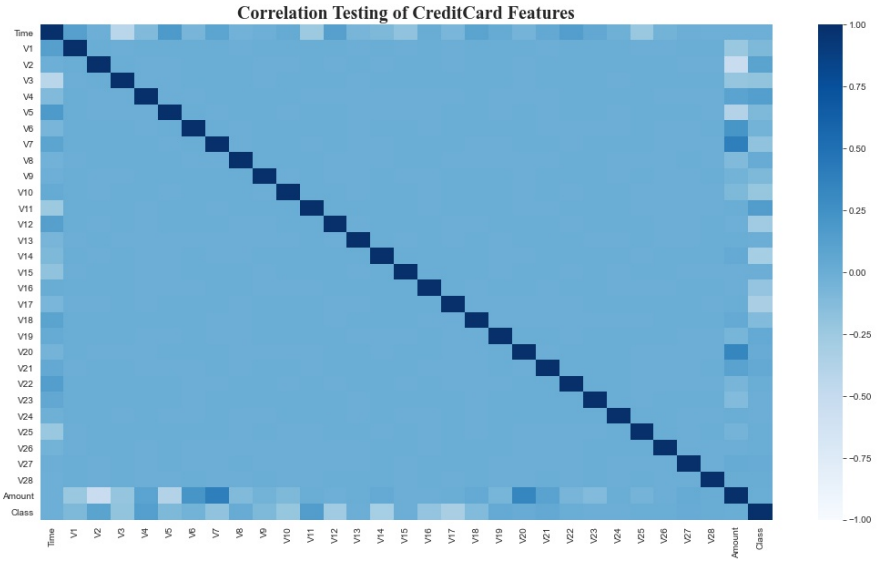
Heatmap of Correlation



- age is Postively Correlated with capital.gain.
- capital.gain , capital.loss and hours.per.week is Positively Correlated with education.num.

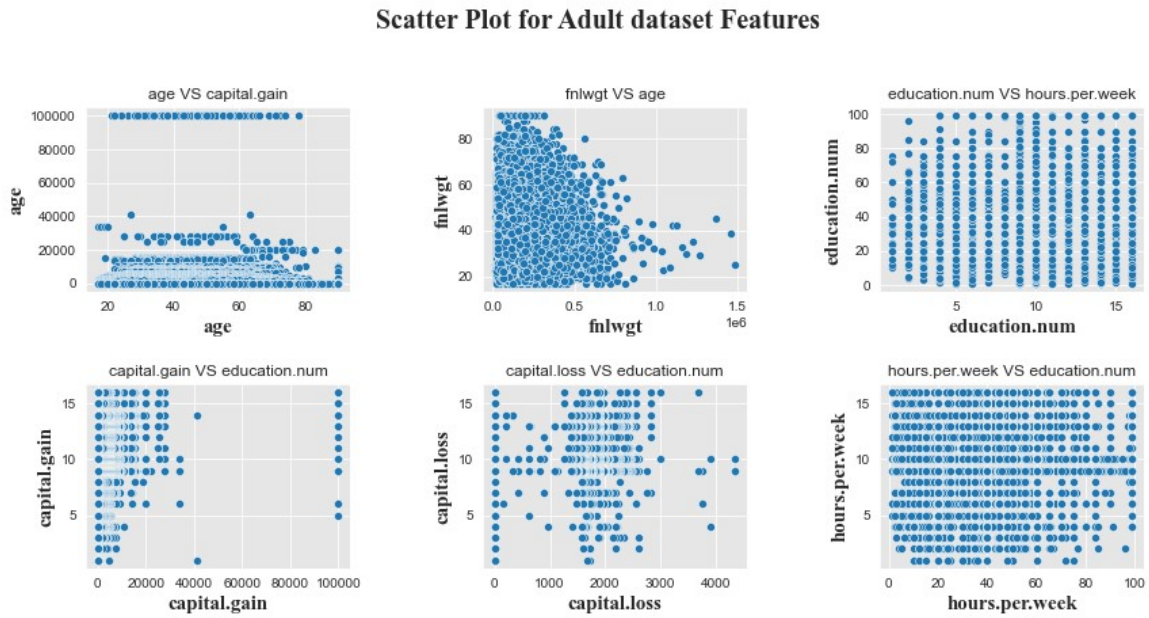


SeniorCitizen is Positively correlated with MonthlyChange and tenure is Positively Correlated with MonthlyChange column

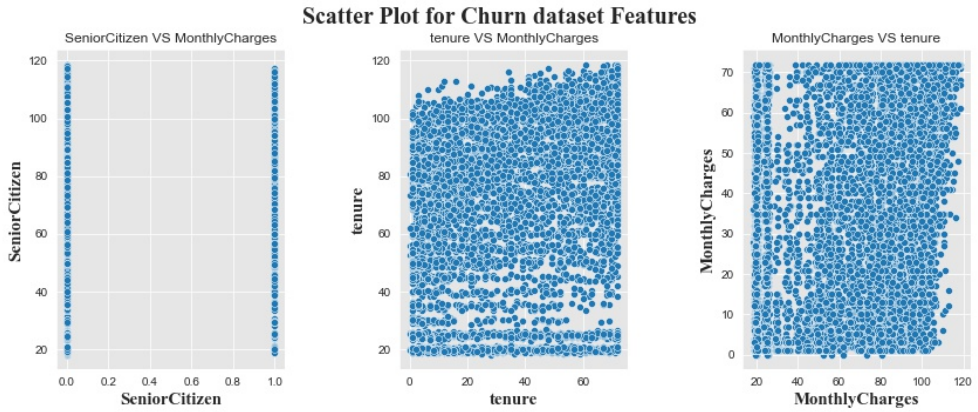


- Time is negatively Correlated with V3.
- V6,V7,V20 and V21 is Positiively Correlated with mounts.

- V4 annd V11 is Positively Correlated with Class.

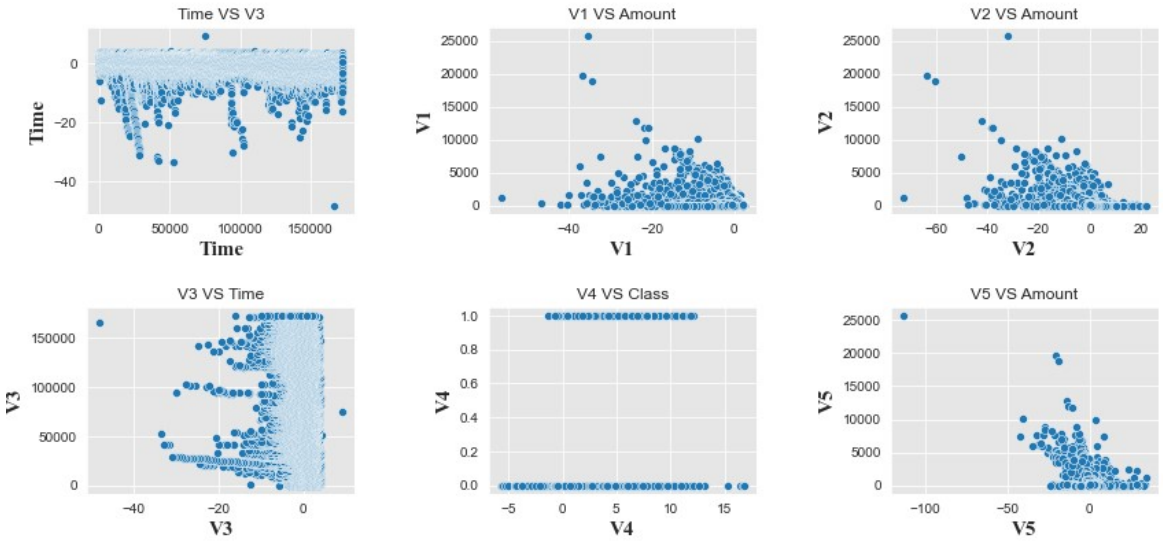


- there is still no strong correlation between the Adult features. Features are Only Uniformly distributed.



- As you can see there is no strong correlation between the features. Features are Uniformly distributed.

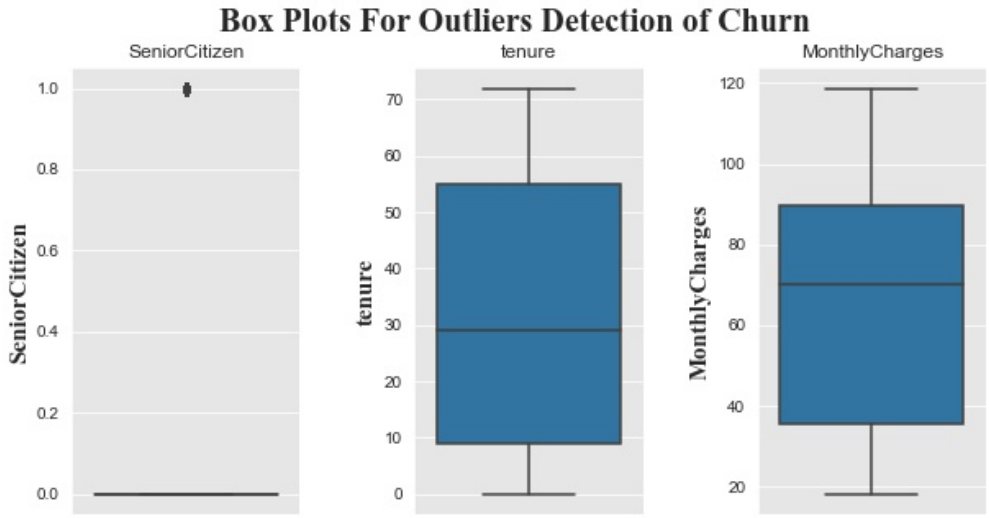
Scatter Plot for CreditCard dataset Features



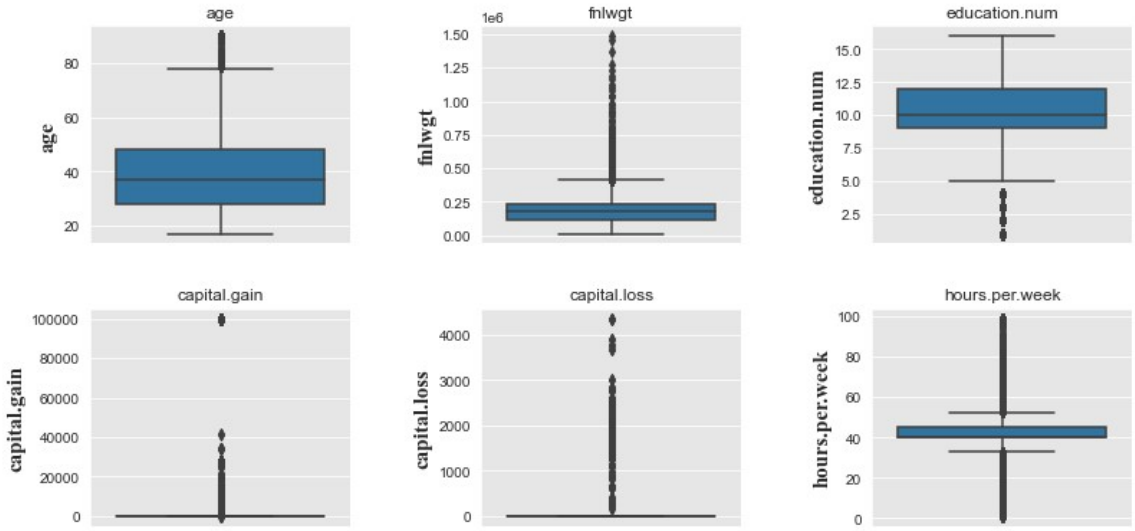
- The CreditCard Features Also does not have a strongly relationship with each other.So Every dataset Features are not strongly correlated with each other so this will affect the model accuracy.

Outliers Detection:

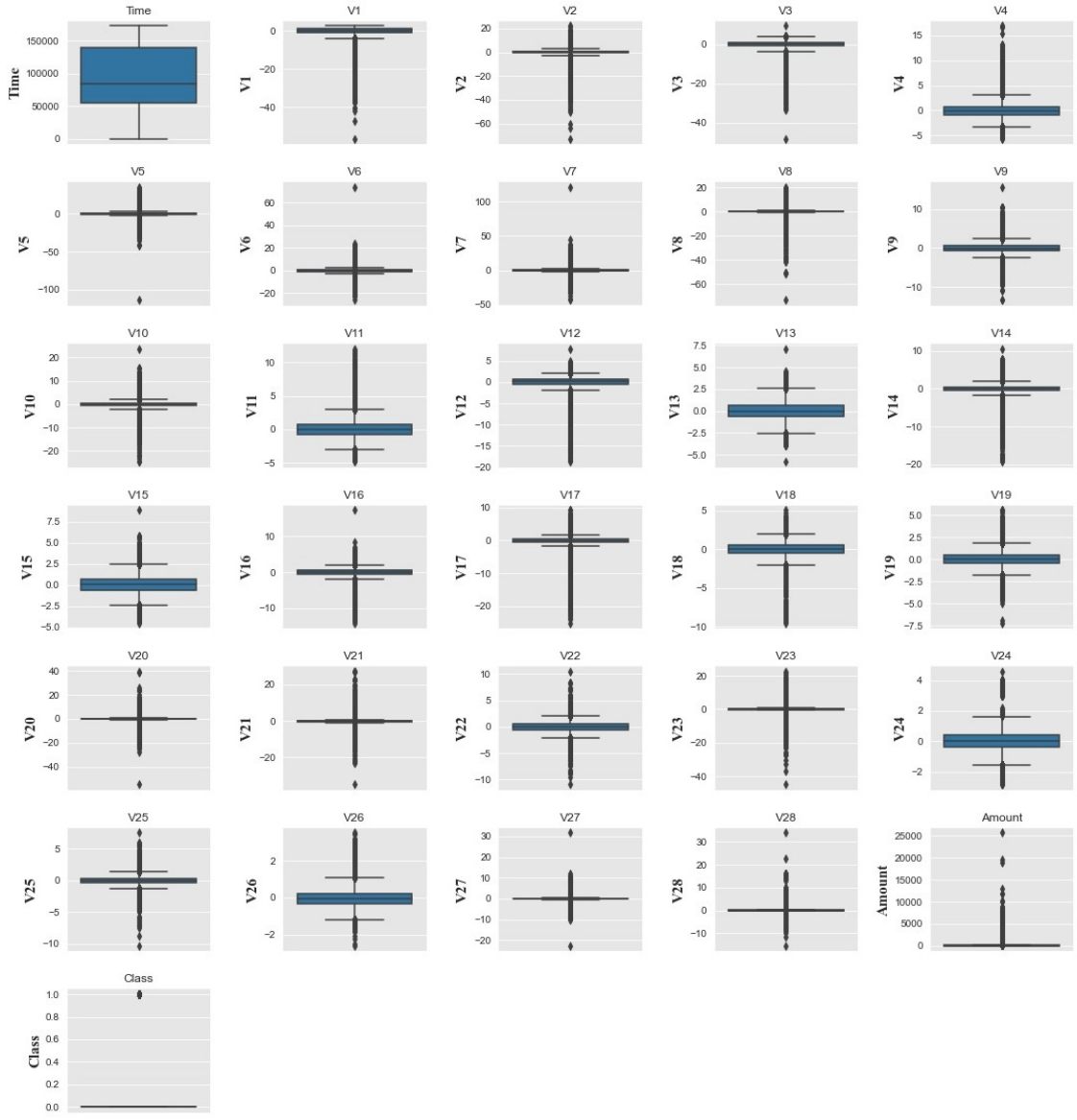
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Outliers can be bad for boosting because boosting builds each tree on previous trees' residuals/errors. Outliers will have much larger residuals than non-outliers, so Ada boosting will focus a disproportionate amount of its attention on those points.And Logistic Regression is alos very sensitive towards the outliers. So i have to get rid of them.In the Boxplot The dots which are far away from the box are the outliers.



Box Plots For Outliers Detection of Adult

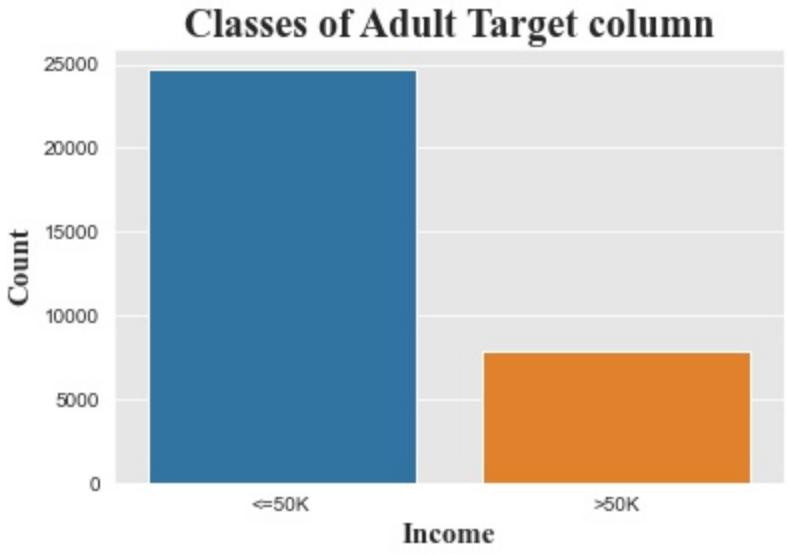


Box Plots For Outliers Detection of CreditCard

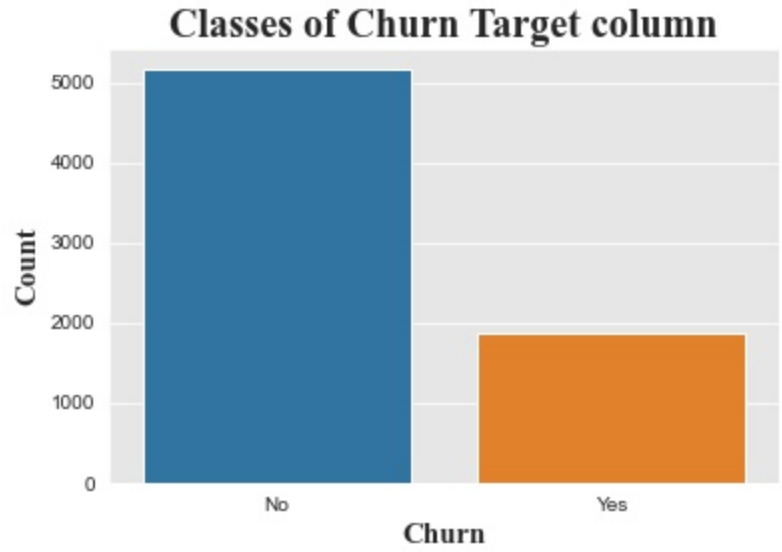


Imbalance Classes

Imbalanced classes are a common problem in machine learning classification where there are 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.Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class.I will Increase the data points of Minority class until it reach to the Majority class.I just Oversample them.



- As you can see Class 1 is higly in majority and the data points of class 0 is very less



- Class no is dominant than class yess.

- In this case class Yes is very less.

Adaboost Results:

	AdaBoost on CreditCard
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.907143
Recall	0.803279
Precision	0.980000

	AdaBoost on Adult
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.491736
Recall	0.803279
Precision	0.980000

	AdaBoost on Churn
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.493237
Recall	0.803279
Precision	0.980000

LogisticRegression Resuts:

	Logistic on Churn
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.493237
Recall	0.803279
Precision	0.980000

	Logistic on Adult
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.500108
Recall	0.803279
Precision	0.980000

	Logistic on CreditCard
True Positive	49.000000
True Negative	78.000000
False Positive	1.000000
False Negative	12.000000
Accuracy	0.350000
Recall	0.803279
Precision	0.980000