Fradulent Transaction Prediction with comparision Model

B Naga Prasanth Reddy
11907524 KM018

https://github.com/prasanth-battula/Int247-ca1-

Abstract

The detection of bank frauds is a topic which many financial sector companies have invested time and resources into. However, finding patterns in the methodologies used to commit fraud in banks is a job that primarily involves intimate knowledge of customer behavior, with the idea of isolating those transactions which do not correspond to what the client usually does. Thus, the solutions proposed in literature tend to focus on identifying outliers or groups, but fail to analyse each client or forecast fraud. This paper evaluates the implementation of a Linear regression,logistic regression and DecisioTree regression

Introduction

About The data set:

1. **step** - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
2. **type** - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
3. **amount** - amount of the transaction in local currency.
4. **nameOrig** - customer who started the transaction
5. **oldbalanceOrg** - initial balance before the transaction
6. **newbalanceOrig** - new balance after the transaction

7. **nameDest** - customer who is the recipient of the transaction
8. **oldbalanceDest** - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
9. **newbalanceDest** - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
10. **isFraud** - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.
11. **isFlaggedFraud** - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Approach:

The dataset is available in kaggale webiste a and . All the models here were created through Jupyter Notebook and its available software libraries, including SciKit-Learn [4], Numpy [8], Pandas [9], and more.

Work Flow:

1. Importing the required libraries:

Libraries that I have imported for this project is numpy,pandas,matplotlib,and from scikit learn packages we have to import required models Logisticregression, Linearregression, DecisionTreeregression And after that we have to ensemble the three algos we have to take the final accuracy of training and testing.

2. **Preprocessing the data**:

We will import the data and preprocess it by replacing the null values with empty string and reading the final data again, later we merge the author name and the news title followed by sepearting the data and label into x and y.

After Preprocessing the data

	step	type	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
0	1	3	9839.64	170136.0	160296.36	0.0	0.0	0
1	1	3	1864.28	21249.0	19384.72	0.0	0.0	0
2	1	4	181.00	181.0	0.00	0.0	0.0	1
3	1	1	181.00	181.0	0.00	21182.0	0.0	1
4	1	3	11668.14	41554.0	29885.86	0.0	0.0	0

3. Training and Spliting the data:

Splitting Now we will split the data into train data and test data representing them with x and y assigning a particular test size and random state. For training I have taken 80% data and for testing I have taken 20%

4. Using three algorithms to find best

Accuracy:

Logistic regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for

solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.

Logistic Regression Equation:

- We know the equation of the straight line can be written as:

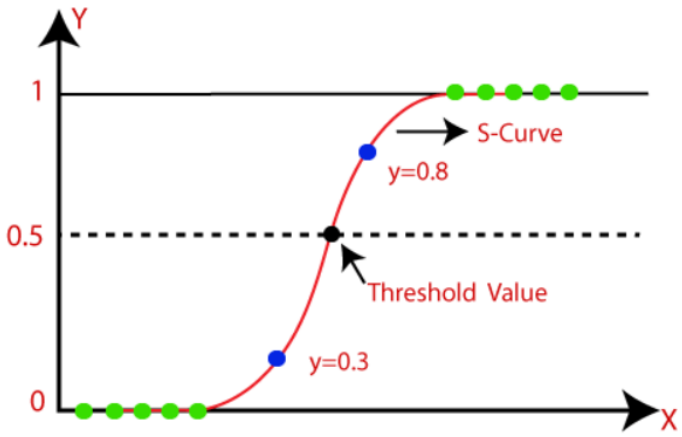
$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

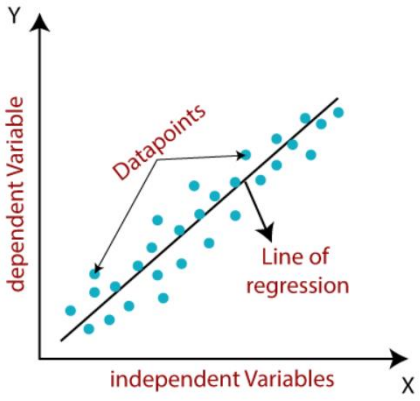
- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$



Liner Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.



Mathematically, we can represent a linear regression

$y= a_0+a_1x+ \varepsilon$

Here,

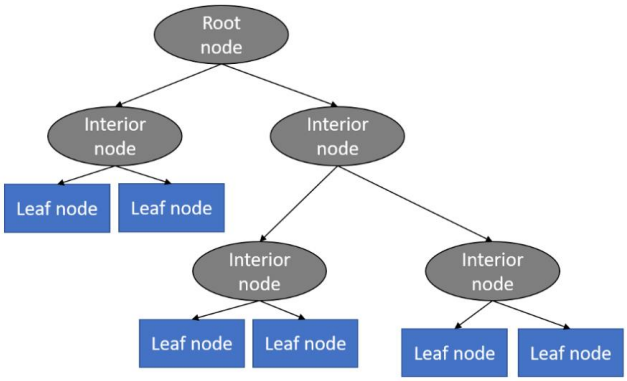
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

DecisionTreeRegression

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The ***Root Node*** is the initial node which represents the entire sample and may get split further into further nodes. The ***Interior Nodes*** represent the features of a data set and the branches represent the decision rules. Finally, the ***Leaf Nodes*** represent the outcome. This algorithm is very useful for solving decision-related problems.



With a particular data point, it is run completely through the entirely tree by answering *True/False* questions till it reaches the leaf node. The final prediction is the average of the value of the dependent variable in that particular leaf node. Through multiple iterations, the Tree is able to predict a proper value for the data point.

5. By using Ensemble bagging technique:

Bagging is **a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data**. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

Partitioning of data	**Random**
Goal to achieve	**Minimum variance**
Methods used	**Random subspace**
Functions to combine single model	**Weighted average**

Example	Random Forest

Conclusion:

This research focuses on the application of the following supervised ML algorithms for DecisionTree regression, Linear Regression, Logistic Regression. ML systems are trained and tested using large datasets. Finding the best accuarcy by using regresson models