

CREDIT CARD FRAUD DETECTION

NAME: TANUSHREE ROY

REG. NO: 12104231

Abstract:

In the real world of credit card fraud detection, due to a minority of fraud related transactions, has created a class imbalance problem. Companies want to provide additional services to their customers. One of these Centers is the way to shop online. Customers can now buy the necessary goods online but this is also an opportunity for criminals to commit fraud. Criminals can steal the information of any cardholder and use it to make online purchases until the cardholder contacts the bank to block it. Fraud prevention technology is now critical to eliminating the losses of banks and other financial institutions.

Here machine learning algorithms are applied on a data set of credit cards frauds and the power of three machine learning algorithms is compared to detect the frauds accomplished using credit cards. The accuracy of linear regression machine learning algorithm is best as compared to Decision Tree and Polynomial algorithms.

Introduction:

In recent years, credit card users have lost countless millions of dollars to fraud. Many researchers are working on the early detection of credit card fraud. All information is available from a large number of sources such as followers on social media, customer ethics, liking, and sharing. Crime is a growing problem with consequences reaching the financial sector, business institutions and government. Fraud can really be described as illegal fraud in order to gain financial gain. Customers have no restrictions on their spending, unlike the limited amount of money in your wallet. In addition, many companies and institutions now tend to transfer their businesses to online services due to the rapid increase in the use of modern technology in all sectors. To solve this problem, credit card issuers have to use sophisticated methods to detect fraud. One of the major problems in this sector is the lack of good data sets as the available data sets of this problem are unbalanced data sets and have many unknown private insurance

sites. What makes it difficult for program planners is to understand the database and build the best model that solves this problem. There are various technics to do fraud detection in machine learning such as Regression, Classification and Clustering.

Classification - Classification find some conclusions from a huge amount of data. When given some input values from the data, the classification algorithms attempt to select one or more outputs on the basis of the input data. Machine learning algorithms are very useful in classification.

Regression - Regression is a supervised learning technique. It is used to predict output values from given input values. It is mostly used to predict continuous data. Regression techniques are machine learning techniques which are very useful in prediction.

Clustering- It refers to dividing the problem space into groups on the basis of the similarities between the data. The items in one cluster are very similar to each other. Items in different clusters are different to each other in their properties. Machine learning algorithms are very useful in clustering.

In this research we are using regression algorithms to detect fraud. The implementation for machine learning techniques as well as the estimation and evaluation of different performance measurement parameters are covered and then the findings of the entire research are covered and also suggested further enhancements.

Literature Review:

Many approaches have been proposed to bring solutions to detect fraud from supervised approaches, unsupervised approaches to hybrid ones; which makes it a must to learn the technologies associated in credit card frauds detection and to have a clear understanding of the types of credit card fraud.

- M.Puh and L.Brkcic presents a comparison of three supervised machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR), they used a dataset that contains credit card transactions made by European cardholders for two days. The environment of the fraud detection system is not static, where fraudulent behaviour is mutating over time to avoid detection, so the

predictive model should not be static. Their experiments were done using two approaches: static and increment, where the evaluated performance based on two measurements namely: AUC and average precision (AP). According to the mentioned results in this paper, SVM has achieved the lowest performance in static and increment setup through AUC and AP. [1]

- J.O.Awoyemi and A.O.Adetunmbi presents a comparison of three machine learning algorithms namely: naïve bayes, k-nearest neighbor and logistic regression on a dataset that is sourced from ULB Machine Learning Group made in September. They split it into 70% for training and 30% for testing and validating, the dataset consists of 284,807 transactions and its highly imbalanced and skewed data.[2]
- Related to the Fraud Detection. M. Zareapoor and P. Shamsolmoali in Presented an application based on a bagging ensemble for credit card fraud detection problems. The ensemble approach based on a decision tree algorithm that was used for the experimental step.[3]
- V. Dheepa and R.Dhanapal [11] proposed a model using Support Vector Machine (SVM). The performance of the SVM is affected by the number of features, which gives a good result when selecting a small number of features for training, so that they selected a small set of features that are relevant to customer behaviour for training, these features are transformed into numerical data before used. According to the mentioned results in this paper, accuracy achieved in this model more than 80 percent.[4]

Proposed Methodology:

- Reading the training data
- Data Pre-processing
- Applying Algorithms
- Check Accuracy and Comparison
- Creating Model

Data Set:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28
0	0.0	-1.358807	-0.072781	2.536347	1.378156	-0.338321	0.462388	0.239589	0.088698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.688281	-0.327642	-0.139097	-0.055353	-0.059752
3	1.0	-0.966272	-0.185226	1.752993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.582941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153

5 rows × 31 columns

The dataset is collected from Kaggle.com to evaluate the ml algorithms.

Algorithm Used:

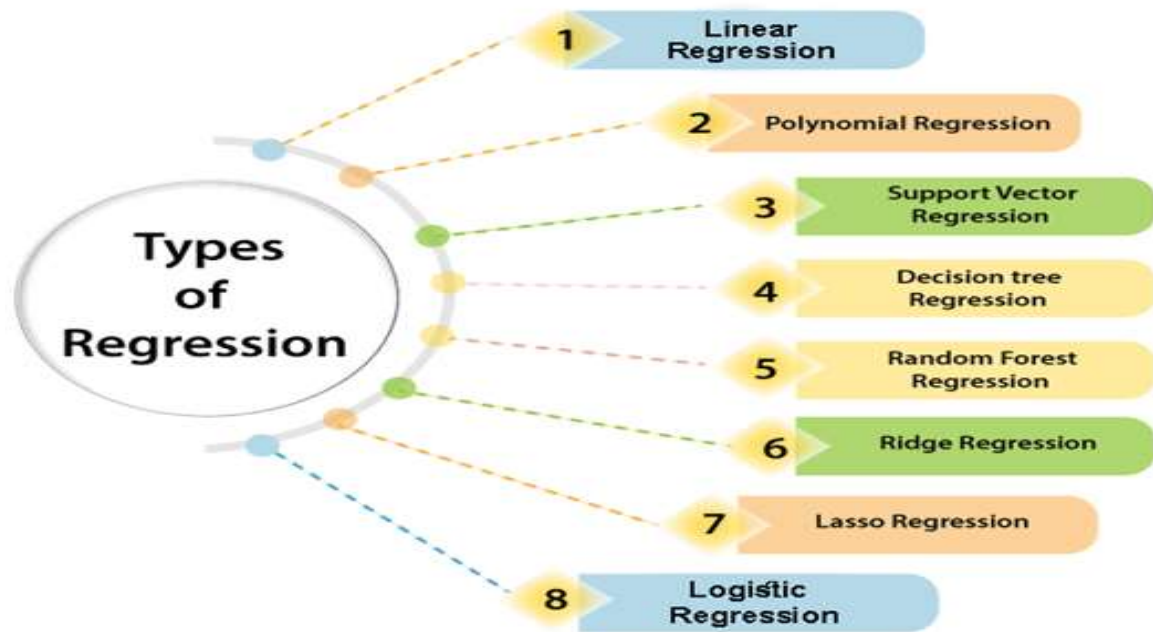
In this data we are using regression algorithms-----

1. Linear Regression-
2. Decision Tree
3. Polynomial

Regression-

Regression is a way of investigating the relationship between independent variables or factors and dependent variations or outcome. It is used as a model for predicting machine learning, in which the algorithm is used to predict continuous results.

The regression analysis is used to understand the relationship between different independent variables and the dependent variable or outcome. Models who are trained to predict or predict trends and results will be trained using retrospective strategies. These models will learn the relationship between inclusion and outgoing data in lab training data. It can then predict future trends or predict results from invisible input data, or be used to understand gaps in historical data.

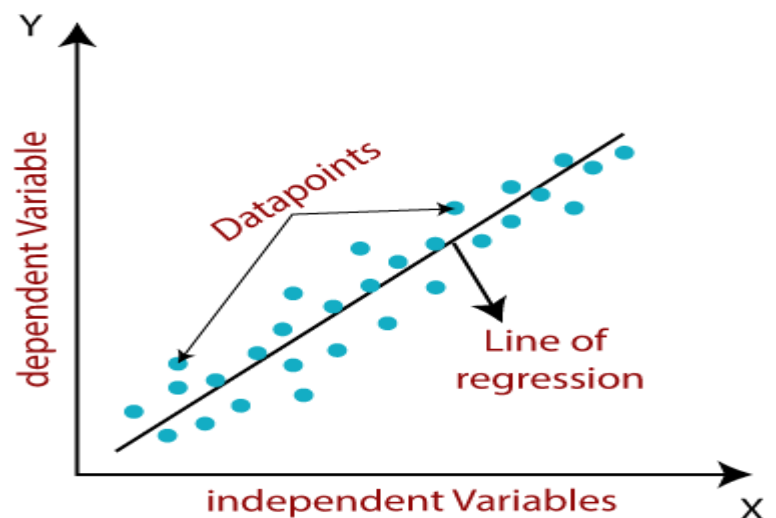


Linear Regression-

Linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. Here in this graph the black line refers the best fit straight line. Here the line will be modelled as below:

$$y = a_0 + a_1 * x$$

, where y is Dependent Variable and x is Independent Variable.



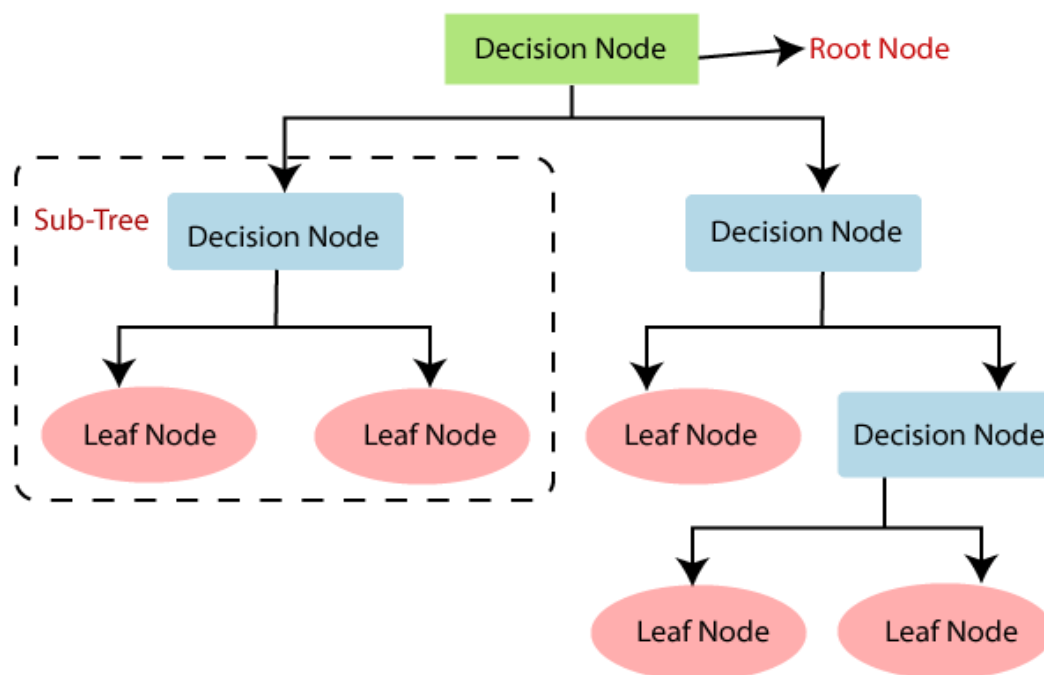
```
✓ [34] from sklearn.linear_model import LinearRegression  
0s lr=LinearRegression()  
lr.fit(train_x,train_y)  
train_pred_lr=lr.predict(train_x)  
test_pred_lr=lr.predict(test_x)  
  
✓ [36] from sklearn.metrics import mean_squared_error, r2_score  
0s  
  
✓ [36] print("Linear accuracy:",r2_score(train_pred_lr,train_y))  
0s print("Linear accuracy:",r2_score(test_pred_lr,test_y))  
  
Linear accuracy: 1.0  
Linear accuracy: 0.9995149108429403
```

Decision Tree-


The Decision Tree has many parallels in real life as well it arises, affecting the wider range of the Machine Learning, which includes both Classification and Descending. Sometimes decision trees too called CART, short for the Tree of Separation and Descent.

In the decision analysis, the decision tree can be used to see again they clearly represent decisions and decision making Tree-Based Algorithms is a popular family related to non-parametric and supervised methods for both separation and descent. If you are wondering what supervised reading is, this is the kind of a machine learning algorithm that includes training data models with both input and output labels.

There is another concept that is quite opposite to splitting. If there are ever decision rules which can be eliminated, we cut them from the tree. This process is known as Pruning and is useful to minimize the complexity of the algorithm.



```
✓ [40] from sklearn.tree import DecisionTreeRegressor  
0s dec=DecisionTreeRegressor()  
dec.fit(train_x,train_y)  
train_pred_dec=dec.predict(train_x)  
test_pred_dec=dec.predict(test_x)
```

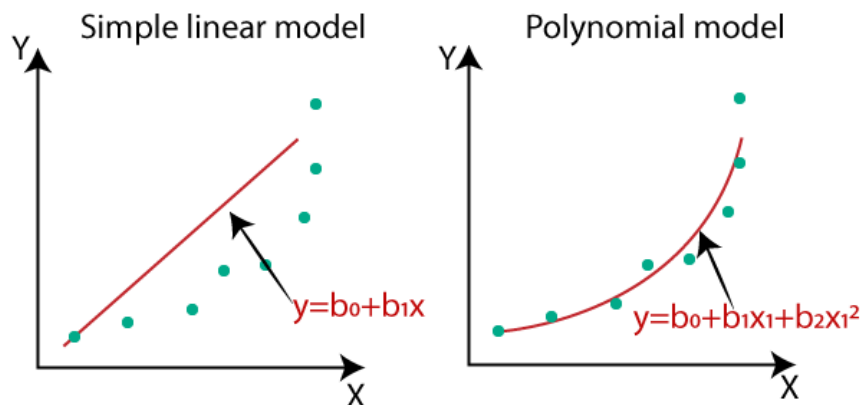
```
✓  print("Decision accuracy:",r2_score(train_pred_dec,train_y))  
0s print("Decision accuracy:",r2_score(test_pred_dec,test_y))
```

```
Decision accuracy: 1.0  
Decision accuracy: -3.0162099188269
```

Polynomial-

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \dots + b_nx_1^n$$



```
[45] from sklearn.preprocessing import PolynomialFeatures
deg_2 = PolynomialFeatures(degree=4)
lr_2 = LinearRegression()
x_deg2 = deg_2.fit_transform(x)
x_deg2.shape

(39, 46376)

[46] from sklearn.model_selection import train_test_split
train_x2, test_x2, train_y2, test_y2 = train_test_split(x_deg2, y, test_size=0.3)

[47] train_x2.shape

(27, 46376)

[48] lr_2.fit(train_x2, train_y2)
train_pred_lr_2 = lr_2.predict(train_x2)
test_pred_lr_2 = lr_2.predict(test_x2)

print("Polynomial Accuracy:", r2_score(train_pred_lr_2, train_y2))
print("Polynomial accuracy", r2_score(test_pred_lr_2, test_y2))

Polynomial Accuracy: 1.0
Polynomial accuracy -0.0634934495726165
```


Analysis-

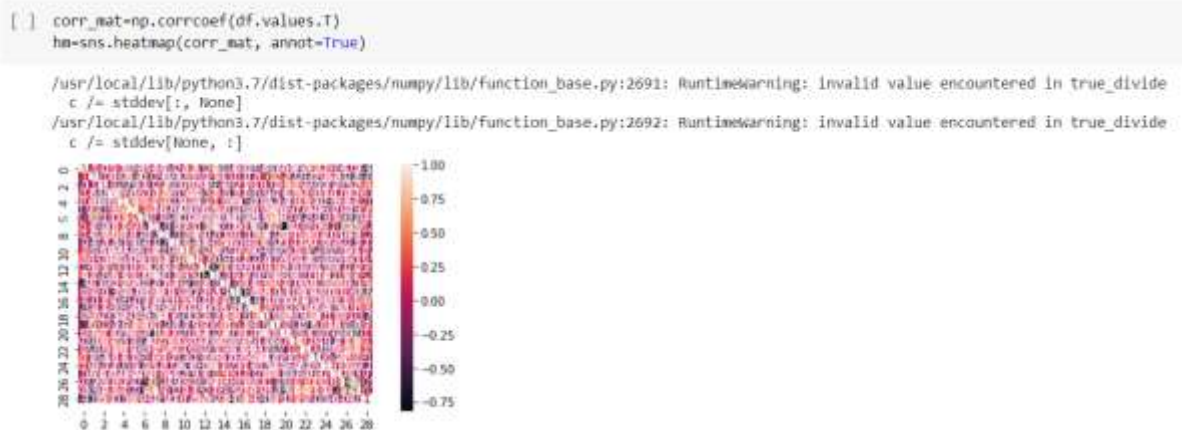
3 algorithms are train and tested individually.

The three algorithms are tested on the basis of their prediction accuracy and confusion matrix. These parameters are used for comparison of performance. The prediction accuracy of all the three techniques is very high and almost equal to each other but linear algorithm has highest prediction accuracy among the three.

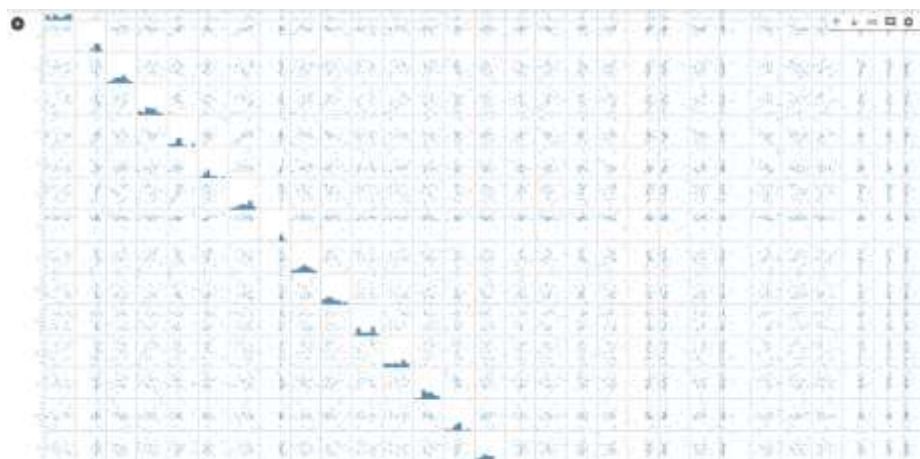
Linear Regression give 100% accuracy in train and test in both. Where in other algorithms less accuracy come through dataset.

Here other various parameters are also tested.

Co-relation Matrix



Plot graph



Conclusion-

In this paper machine learning algorithms are used for credit card fraud detection. The power of machine learning is used to detect credit cards frauds and the performance of different machine learning algorithms is compared. Three machine learning algorithms, Decision Tree, Linear regression and Polynomial are applied on a data set have the data of 33741 credit cards. The performance of Linear algorithm is found best with highest accuracy of 100 percent. The performance of Decision Tree is minimum with accuracy 100 and -3.01 and the performance of Polynomial algorithm is 0.06 percent.

The limitation of this paper are as follows on which work can be done in future:

1. The performance of other machine learning algorithm can be checked for credit card fraud detection.
2. The accuracy of Linear machine learning algorithm should also be tested on other data sets for credit card fraud detection.
3. The performance of Linear algorithm can also be tested on other data sets of different domains.

References:

1. M. Puh and L. Brkić, "Detecting credit card fraud using selected machine learning algorithms", 2019 42nd International Convention on Information and Communication Technology Electronics and Microelectronics (MIPRO), pp. 1250-1255, 2019.
2. J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9, 2017.
3. M. Zareapoor, P. Shamsolmoali et al., "Application of credit card fraud detection: Based on bagging ensemble classifier", Procedia computer science, vol. 48, no. 2015, pp. 679-685, 2015.
4. V. Dheepa and R. Dhanapal, "Behavior based credit card fraud detection using support vector machines", ICTACT Journal on Soft computing, vol. 6956, pp. 391-397, 2012.
5. <https://www.researchgate.net/profile/Dr-Kumar>
6. <https://ieeexplore.ieee.org/abstract/document/8776942>
7. Mittal, Sangeeta, and Shivani Tyagi. "Performance evaluation of machine learning algorithms for credit card fraud detection." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019.
8. Dornadula, V.N. and Geetha, S., 2019. Credit card fraud detection using machine learning algorithms. Procedia computer science, 165, pp.631-641.
9. Shukur, Hamzah Ali, and Sefer Kurnaz. "Credit card fraud detection using machine learning methodology." International Journal of Computer Science and Mobile Computing 8, no. 3 (2019): 257-260.
10. Trivedi, Naresh Kumar, et al. "An efficient credit card fraud detection model based on machine learning methods." International Journal of Advanced Science and Technology 29.5 (2020): 3414-3424.



MITTAL
SCHOOL OF BUSINESS