DATA ANALYSIS
WITH
MACHINE LEARNING

PROJECT TOPIC

CREDIT CARD FRAUD DETECTION

USING CONCEPT OF
NEURAL NETWORKS

SUBMITTED BY.:

ASHWANY KUMAR VERMA
18SDATVNS081

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my instructor **Mr. Chandan Verma Sir** who gave gave me golden opportunity to do this wonderful project on topic Credit Card Fraud Detection,using cocepts of Machine Learning, which also helped in doing a lot of Research and I came to know about so many new things.I am really thankful to him.

Secondly I would also like to thank my parents and my friends who helped me a lot in doing this project within the limited time.

This project helped a lot to me to understand the concepts more clearly.

THANKS AGAIN TO ALL WHO HELPED ME.

INFORMATION ABOUT DATASET

This data set is taken from the website. :
https://www.kaggle.com

The reference URL is.:
https://www.kaggle.com/mlg-ulb/creditcardfraud

Description of the DATASET:

- **Context**

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

- **Content**

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Since this dataset contains a large number of data so we would be taking only 10 % of the data for analyzing, so that it may not affect the runtime and complexity of the code(Data is chosen randomly).

It contains only numerical input variables which are the result of a PCA transformation. This dataset of total contains

31 columns. The first one is 'Time'. Then comes the 28 columns which refrence to the parameters. The other two columns are 'Amount' and 'Class'. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

- **Inspiration**

Identify fraudulent credit card transactions.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

- **Acknowledgment**

The data-set has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on http://mlg.ulb.ac.be/BruFence and http://mlg.ulb.ac.be/ARTML

Please cite: Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015.

CONCLUSION

This dataset contains 284207 number of data with 31 features out of which 1 is the target which represents whether the transaction is a valid or a fraud transaction.

Since this dataset has large number of data which increase the computational complexity and the run time so we take only 10% of the data which equals to 28421 which is enough to train a system and gain a high accuracy.

After the completion of project I reached to a conclusion that there are 0.172% of total fraud cases and the accuracy of the model is **0.9989467524868344**.

END