Title : Credit card fraud detection analysis

Abstract : 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 the dataset is unbalanced I will use under sampling to make the data balanced, Then I will train my model with the balanced data and then I will use the original data set as testing dataset on my model. I will be using 3 or 4 algorithms as my models depending on time. I will show visually representation of the dataset before and after balancing as a part of data exploration and then after modeling I will show the accuracy of the algorithms on this dataset.

Alogrithms

- why these algorithms are important and how they are used.

I am using algorithms such as Tenserflow, Decision tree, Logistic regression, SVM.

Currently I am working on decision tree, logistic regression and SVM. I need to know more about tenser flow. In my project I am showing the accuracy of the output using different algorithms. So, no algorithm is corelated in my project to say few are important and few are not. But ultimately all the algorithms are used for classification problems and each algorithm has its own strength.

- how does the data have to manipulated or conform to use these algorithms (e.g. remove nulls, split data into training vs test sets,...)

I will remove the null data if there is any, then I need to either do undersampling or oversampling of data to balance my data, since fraud cases are very few, the model may not be build properly.

- How are results expected to be interpreted or used

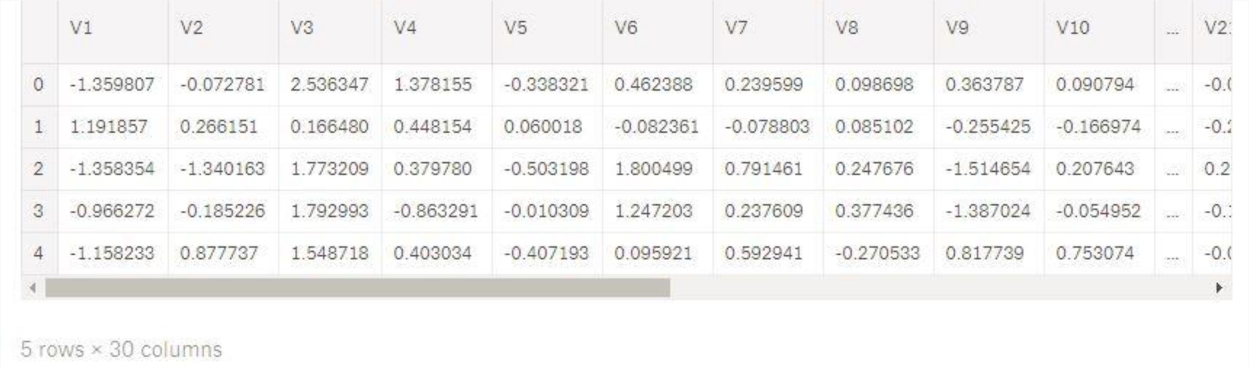
We can use these results to take a decision whether to give a loan or credit card to a particular person of certain criteria.

Data sources

I chose a problem statement from Kaggle and improvised in my own way. And data is provided by Kaggle.

Link - https://www.kaggle.com/dalpozz/creditcardfraud/data

- show samples of the data source (up to 10 lines of data (less if it is highly dimensional)



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V2
0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	...	-0.(
1	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	...	-0.:
2	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	...	0.2
3	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	...	-0.:
4	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	...	-0.(

5 rows × 30 columns

- show or describe the steps you will take to transform the sample and make it ready for the analysis

Step-1 – I will explore the data both statistically and analytically.

Step -2 – I will look for any missing values.

Step -3 – my data need to be oversampled or undersampled, since one class has 98% of the data.

Step – 4 – I will build model using different algorithms and check the accuracy.

Step-5 – I will remove outliers or any unwanted variables to improve my accuracy.

6. Graphics

what type of graphs, charts, tables are you planning to use to describe the end results

histograms, box plots, confusion matrix, classification report which has precision recall t- score.

What are your current challenges

I need to undersample my data correctly .

At least 5 reference URLs with phrases or paragraphs you are planning to cite

References : https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/learning-path-data-science-python/

https://www.kaggle.com/dalpozz/creditcardfraud

https://stackoverflow.com/questions/29204005/how-to-perform-under-sampling-in-scikit-learn

https://stackoverflow.com/questions/34831676/how-to-perform-undersampling-the-right-way-with-python-scikit-learn

https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now

https://gallery.cortanaintelligence.com/Experiment/Evaluating-and-Parameter-Tuning-a-Decision-Tree-Model-1

https://www.tensorflow.org/get_started/graph_viz