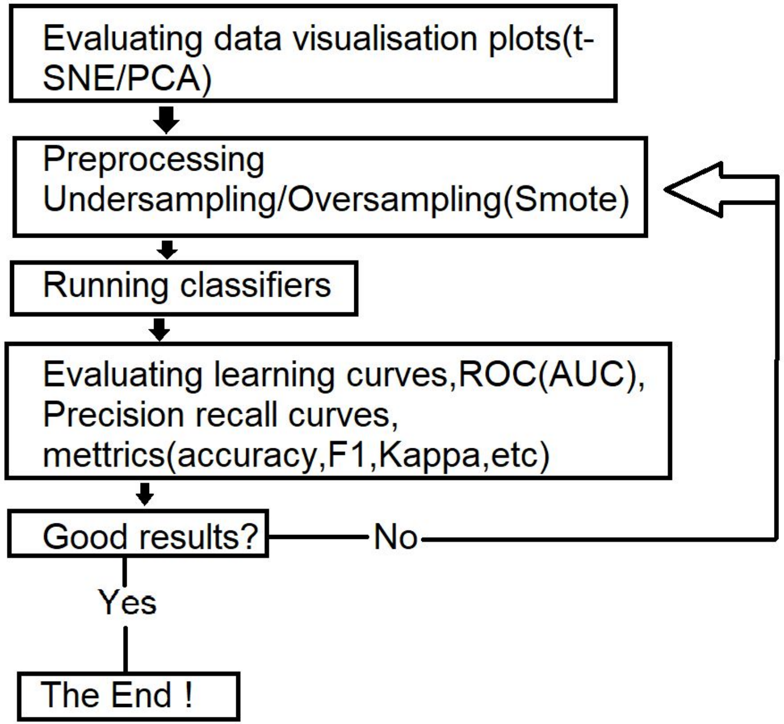
Detecting credit card fraud using machine learning techniques

Group No:15
Members- Aditya(2016127), Jai(2016154)

Problem statement

- Increasing digital transactions imply greater digital fraud.
- High losses incurred by banks.
- Customers become apprehensive to use digital transactions in countries where fraud is not insured.
- Automated system to detect fraud in real time is beneficial for both the bank and the consumer.

Pipeline

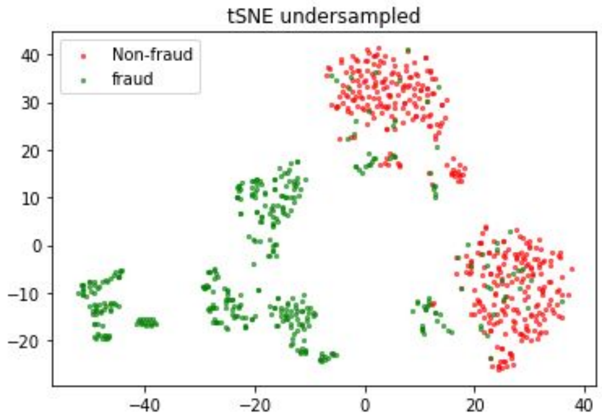

Evaluating data visualisation plots(t-SNE/PCA)
Preprocessing
Undersampling/Oversampling(Smote)
Running classifiers
Evaluating learning curves,ROC(AUC),
Precision recall curves,
mettrics(accuracy,F1,Kappa,etc)
Good results?
No
Yes
The End !

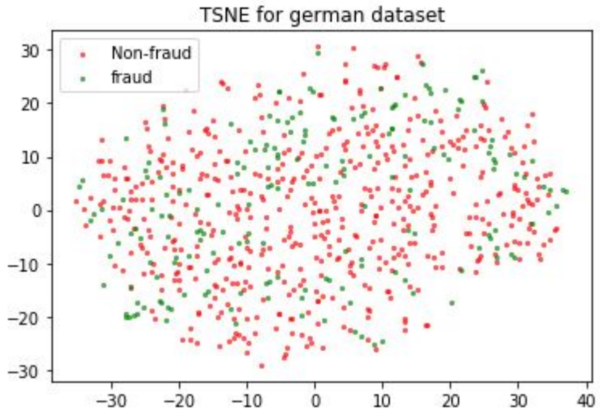
Approach/Formulation

- We visualized the dataset using t-SNE and PCA.
- We decided to use pre-processing techniques such as random Undersampling and SMOTE(synthetic minority oversampling).
- We decided to use Ensemble learning methods like Random forest and Neural networks as well as KNN for classification.
- We chose MLP with hidden layers : (100,70,50)
- Since we were dealing with unbalanced dataset, we decided to use precision,recall,F1 score,ROC_AUC, kappa to be the evaluation metrics rather than accuracy.

Datasets

- There are very few relevant datasets in this domain due to confidentiality issues.
- We will be using the european credit dataset and the German Credit fraud dataset to separately to build models.
- The former is a large dataset (284807, 31).It is very unbalanced since the frauds account only ~0.17% of the entire dataset and this dataset hides the feature details due to confidentiality .
- The german dataset is relatively smaller with 1000 records with well labelled features and is relatively balanced.

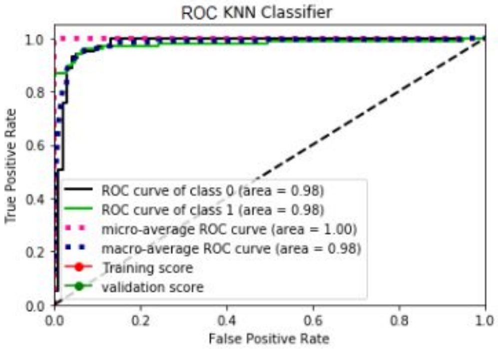




Results

TABLE I
PRE-MIDSEM RESULTS USING DATASET[1]

MODEL/METRICS	Precision	Recall	F1 score	Kappa
Naive Bayes	class 0: 0.8 class 1: 0.53	class 0: 0.8 class 1: 0.52	0.716	0.32
Logistic Regression	class 0: 0.75 class 1: 0.57	class 0: 0.90 class 1: 0.30	0.692	0.237



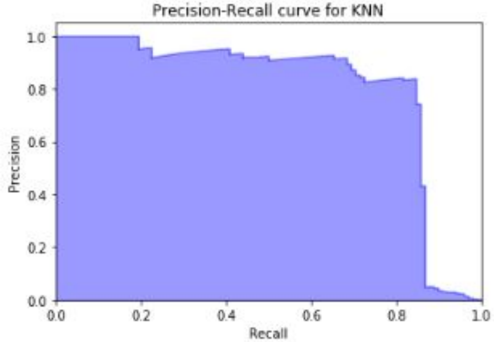
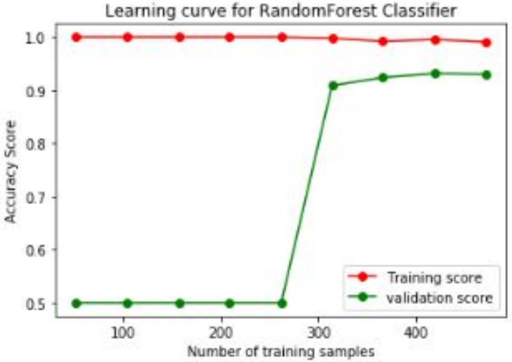
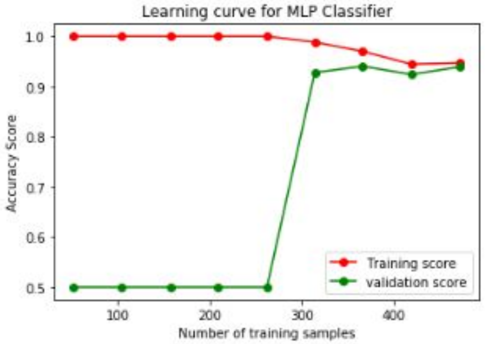


TABLE II
POST-MIDSEM RESULTS USING DATASET[1]

MODEL/METRICS	Precision	Recall	F1 score	Kappa
Random Forest	class 0: 1 class 1: 0.90	class 0: 1 class 1: 0.86	class 0: 1 class 1: 0.88	0.857
KNN	class 0: 1 class 1: 0.94	class 0: 1 class 1: 0.80	class 0: 1 class 1: 0.87	0.865
MLP	class 0: 1 class 1: 0.88	class 0: 1 class 1: 0.82	class 0: 1 class 1: 0.85	0.845





Analysis

- Preliminary analysis t-SNE plots
    - European dataset - visually separable
    - German dataset not visually separable
- Accuracy not a good metric for unbalanced datasets
- Learning curve plots(bias/variance) and other metrics. High bias for baseline models
- Oversampling using SMOTE gave the best results for European Dataset but increased time complexity.
- Random Forest was able to give the best results for both datasets