Fraud Detection and Philosophy and Religion

(but not all at once)

Bishal Sainju and Jack Kiefer

Fraud Detection

- Detecting whether a transaction is a normal payment or a fraud.
- The features provided were already scaled and the names of the features were not shown due to privacy reasons.
- Dataset is very unbalanced (492 frauds out of 284,807 transactions) i.e. 99.83% are non-fraud records.

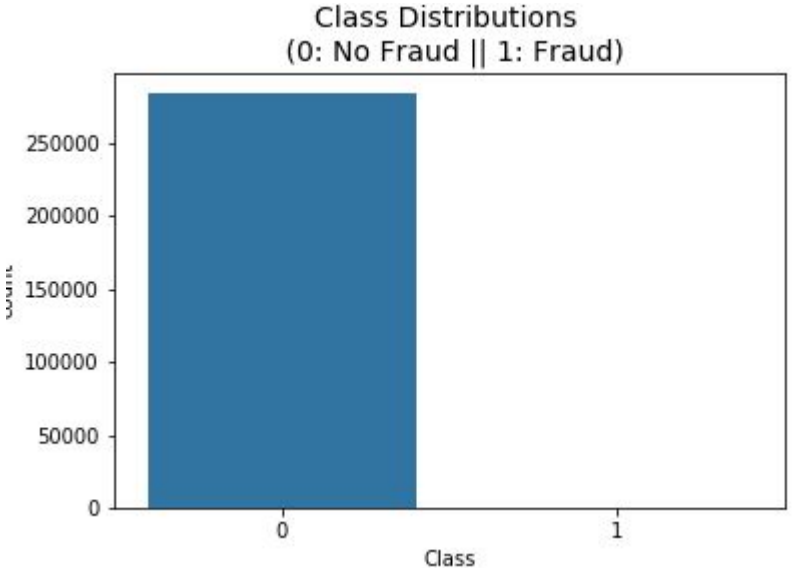
Motivation

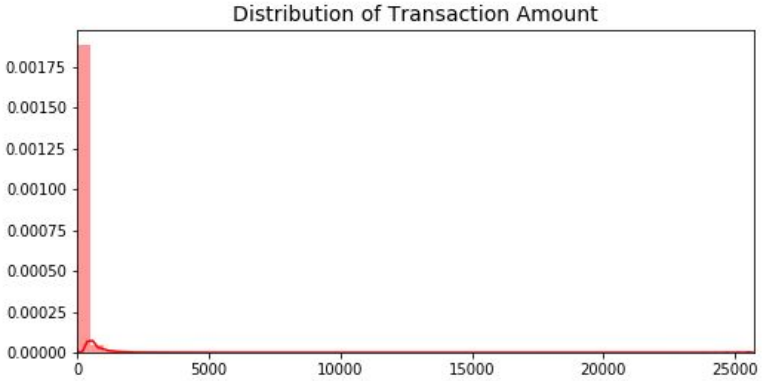
- Learning to handle unbalanced datasets.
- How well techniques used for handling unbalanced dataset generalize over the whole dataset?
- Which classifier, SVC or Logistic Regression works best for this kind of datasets?

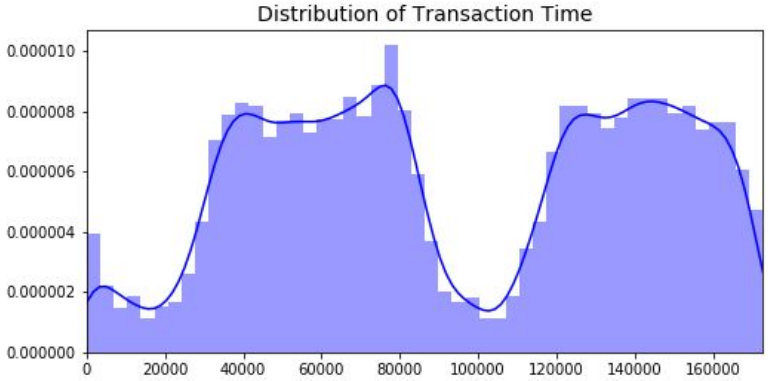
Dataset

- 28 features V1..V28 which are the principal components after the PCA transformation. Original features not provided due to confidentiality issue.
- Features ‘Time’ and ‘Amount’ has not been transformed.
- Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Analysis


Class Distributions
(0: No Fraud || 1: Fraud)
250000
200000
150000
100000
50000
0
0
1
Class

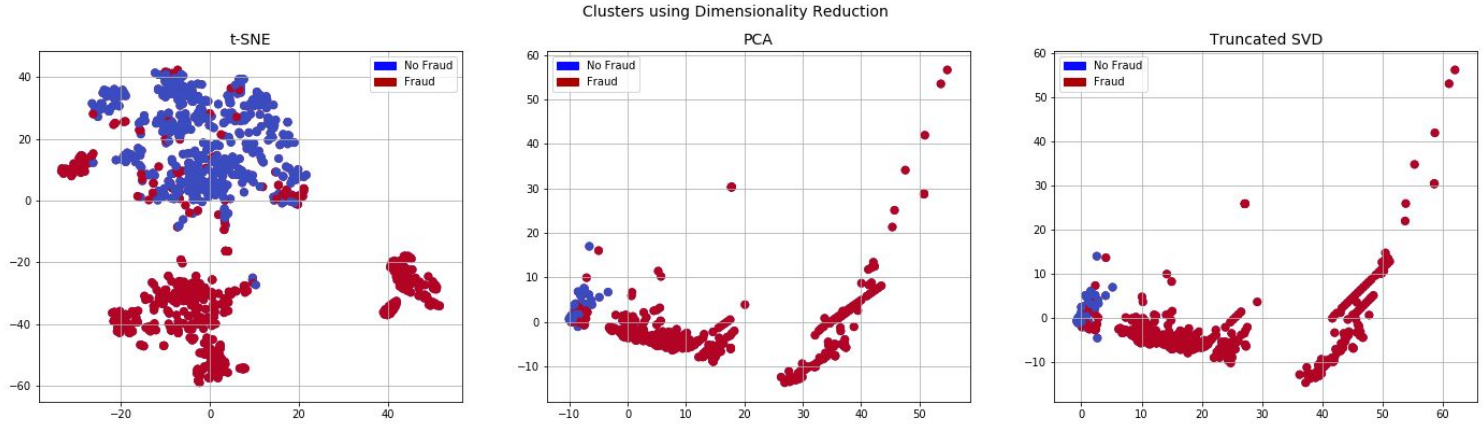

Distribution of Transaction Amount
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
5000
10000
15000
20000
25000


Distribution of Transaction Time
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
0
20000
40000
60000
80000
100000
120000
140000
160000

Analysis


Equally Distributed Classes
500
400
300
200
100
0
count
0
1
Class

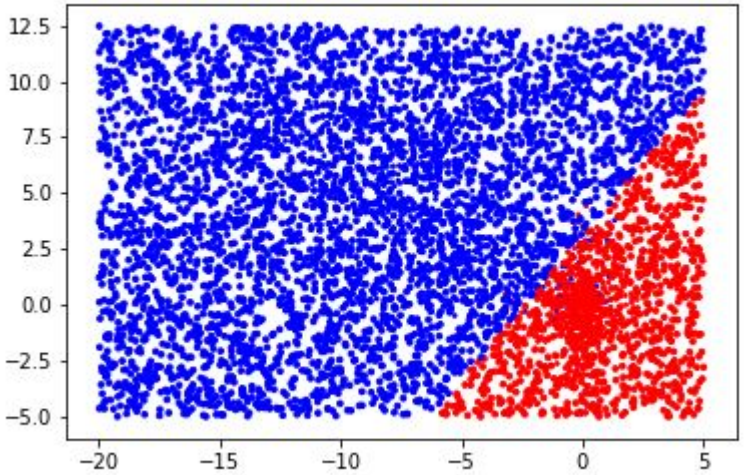
Analysis


Clusters using Dimensionality Reduction
t-SNE
PCA
Truncated SVD
No Fraud
Fraud

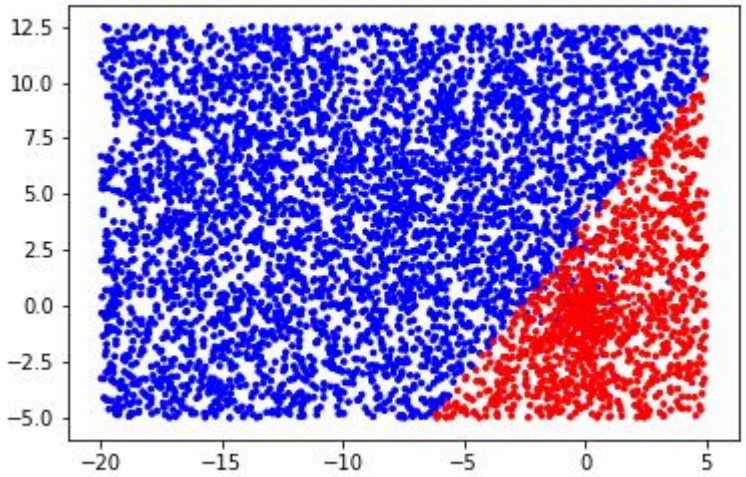
Analysis

```
['V14', 'V4', 'V12', 'V11', 'V10', 'V16', 'V3', 'V17', 'V9', 'V2', 'V7', 'V18', 'V1', 'V6', 'V5', 'V19', 'V20', 'V2
1', 'scaled_time', 'V28', 'V27', 'scaled_amount', 'V26', 'V8', 'V13', 'V24', 'V23', 'V25', 'V15', 'V22']
```

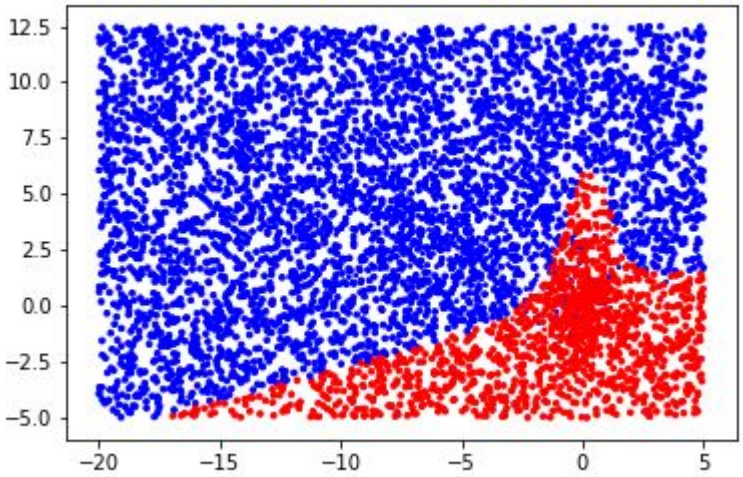
Logistic regression(V14 and V4)



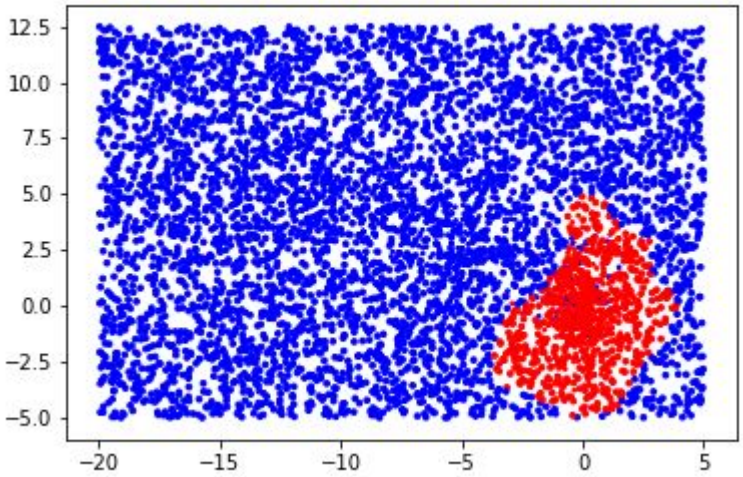
Linear SVM(V14 and V4)



SVM, kernel = poly (V14 and V4)



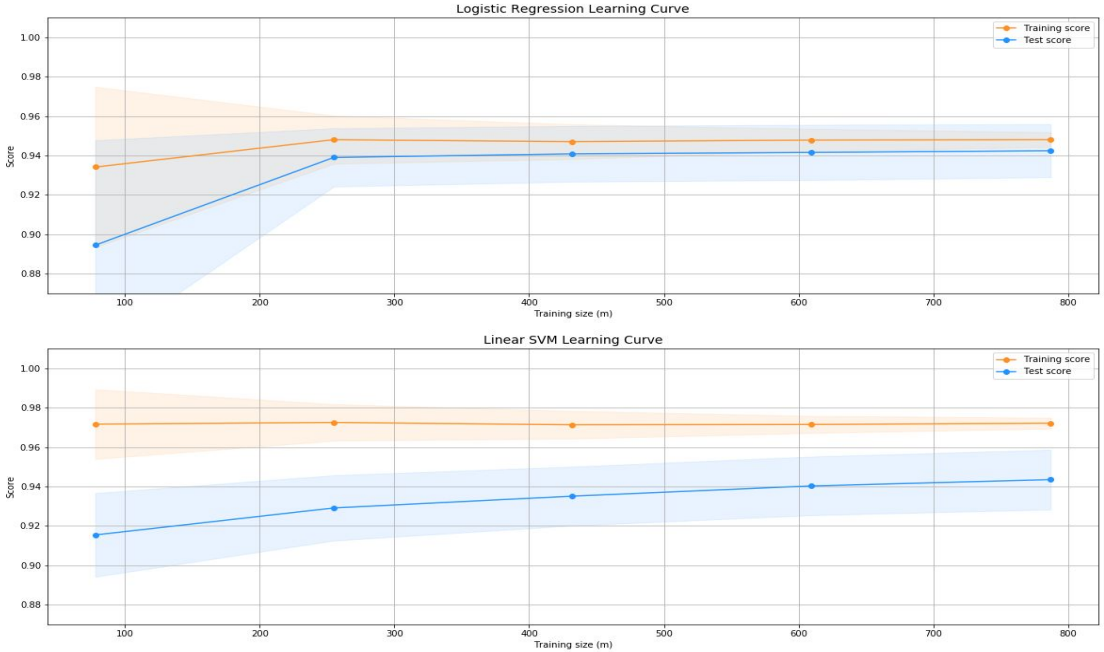
SVM, kernel = RBF (V14 and V4)



Best Parameters (Grid SearchCV)

```
Logistic Regression best parameter: LogisticRegression(C=0.1, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)

SVC best parameter: SVC(C=0.5, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='linear', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

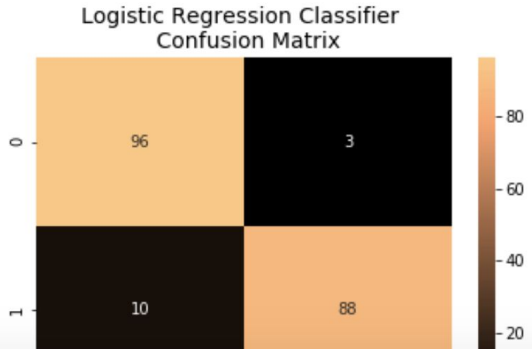
Learning Curve



Classification Report (Best Logistic Regresssion)

```
Recall Score: 0.90
Precision Score: 0.97
F1 Score: 0.93
Accuracy Score: 0.93
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94        99
           1       0.97      0.90      0.93        98

   micro avg       0.93      0.93      0.93       197
   macro avg       0.94      0.93      0.93       197
weighted avg       0.94      0.93      0.93       197

Confusion Matrix:
```

Out[26]: `Text(0.5, 1.0, 'Logistic Regression Classifier \n Confusion Matrix')`



Classification Report (Best SVC Classifier)

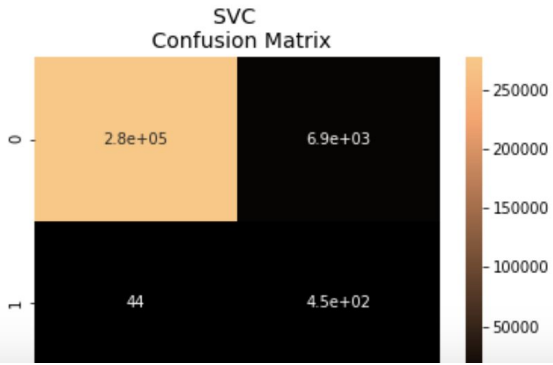
```
Recall Score: 0.88
Precision Score: 0.97
F1 Score: 0.92
Accuracy Score: 0.92
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.97      0.93        99
           1       0.97      0.88      0.92        98

   micro avg       0.92      0.92      0.92       197
   macro avg       0.93      0.92      0.92       197
weighted avg       0.93      0.92      0.92       197

Confusion Matrix:
```

Out[27]: `Text(0.5, 1.0, 'Suppor Vector Classifier \n Confusion Matrix')`

Classification Report (Generalization)

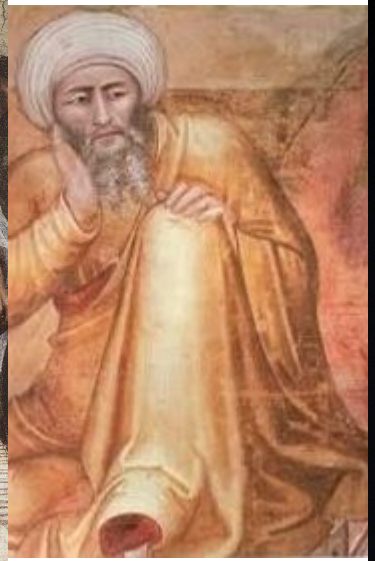
```
Recall Score: 0.91
Precision Score: 0.06
F1 Score: 0.11
Accuracy Score: 0.98
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99    284315
           1       0.06      0.91      0.11       492

   micro avg       0.98      0.98      0.98    284807
   macro avg       0.53      0.94      0.55    284807
weighted avg       1.00      0.98      0.99    284807

Confusion Matrix:
```

Out[65]: Text(0.5, 1.0, 'SVC \n Confusion Matrix')



Religion & Philosophy

- Can we determine if a society had developed a system of philosophy based on what we know of their religious structure and writing system?

Motivation

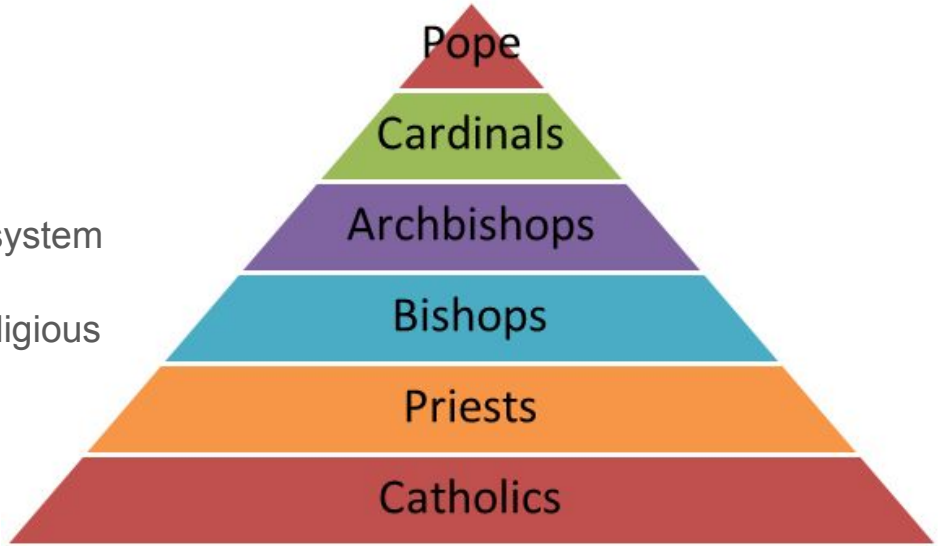
- Helps translators with probable contexts of untranslated texts
- Provides hints as to which social complexity variables can lead to the emergence of something as abstract as Philosophy.
    - (social complexity is a big deal in anthropology)

Dataset: *Seshat Global History Databank*

Quantitative data from 400 societies from 30 geographical areas around the globe across thousands of years of human history and pre-history.

Analysis

- Logistic regression
- Two input variables:
  - Whether or not a society had a writing system
  - The number of levels in the society's religious hierarchy



Note: Philosophy does *not* imply writing.

Results

Mean F1 score: **90%**