PCA v/s LDA v/s SVD

Riya Goyal 2148054

Importing libraries

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import math
from sklearn.model_selection import train_test_split
```

Data Description:

- The dataset 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.
- Feature **'Time'** contains the seconds elapsed between each transaction and the first transaction in the dataset.
- Feature **'Amount'** is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning.
- Feature **'Class'** is the response variable and it takes value 1 in case of fraud and 0 otherwise.

In [3]:
```
# importing the dataset
df = pd.read_csv('D:/Downloads/creditcard.csv/creditcard.csv')

# displaying the first 5 rows of the dataset
df.head()
```

Out[3]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	(
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	-(
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	-(
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	(

5 rows × 31 columns

In [4]:
```
df.shape
```

Out[4]: (284807, 31)

Dropping unnecessary columns, which contains null values

In [5]:
```
# columns containing null values
df.columns[df.isnull().any()]
```

Out[5]: Index([], dtype='object')

There is no missing data in the dataset

In [6]:
```
df['Class'].value_counts()
```

Out[6]:
```
0    284315
1       492
Name: Class, dtype: int64
```

There is strong imbalance in the target value counts. To solve this issue, we will oversample the dataset.

Oversampling the dataset

In [7]:

```
def _oversample_positives(df, target):
    """ Oversample the minority classes to match
    the majority class.

    :param df: pandas dataframe - input df.
    :param target: string - classification target column.

    :return: pandas datframe - oversampled version
    """

    class_count = df[target].value_counts()

    print("Before oversampling: %s" % class_count)

    for i in range(1,len(class_count)):
        df_i = df[df[target] == i]
        oversampling_factor_i = class_count[0] / float(class_count[i])
        print(len(df_i))
        print("Oversampling factor for class %i: %s" %(i, str(oversampling_factor_i)))

        # Integer part of oversampling
        df = df.append(
            [df_i] * int(math.floor(oversampling_factor_i) - 1),
            ignore_index=False)

        # Float part of oversampling
        df = df.append(
            [df_i.sample(frac=oversampling_factor_i % 1)],
            ignore_index=False)

    print("After oversampling: %s" % df[target].value_counts())
    print("Shape after oversampling: %s" % str(df.shape))

    return df
```

In [8]:

```
df_oversampled = _oversample_positives(df, 'Class')
```

```
Before oversampling: 0    284315
1       492
Name: Class, dtype: int64
492
Oversampling factor for class 1: 577.8760162601626
After oversampling: 0    284315
1    284315
Name: Class, dtype: int64
Shape after oversampling: (568630, 31)
```

Splitting the dataset into Independent and Dependent variables

In [9]:

```
X = df_oversampled.drop(['Class'],axis = 1)
y = df_oversampled['Class']
```

In [10]:

```
y.value_counts()
```

Out[10]:

```
0    284315
1    284315
Name: Class, dtype: int64
```

Training data and testing data

In [11]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,random_state=1)
```

Standardizing the data

In [12]:

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Linear Discriminant Analysis

Linear Discriminant Analysis as its name suggests is a linear model for classification and dimensionality reduction. Most commonly used for feature extraction in pattern classification problems.

Assumptions:

LDA makes some assumptions about the data:

- Assumes the data to be distributed normally or Gaussian distribution of data points i.e. each feature must make a bell-shaped curve when plotted.
- Each of the classes has identical covariance matrices.

However, it is worth mentioning that LDA performs quite well even if the assumptions are violated.

In [13]:

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
lda = LDA(n_components=1) #n_components to be less than the n_classes count
X_train = lda.fit_transform(X_train, y_train)
X_test = lda.transform(X_test)
LDA_df = pd.DataFrame(X_test)
LDA_df.head(10)
```

Out[13]:

	0
0	1.756693
1	1.403826
2	-0.689900
3	-2.580457
4	1.359025
5	-0.806420
6	0.152485
7	0.525899
8	-0.154159
9	0.276191

Modelling - Logistic Regression

In [14]:

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
y_pred2 = classifier.predict(X_train)
```

Model Evaluation

In [15]:

```
from sklearn.metrics import confusion_matrix,accuracy_score

cm = confusion_matrix(y_test, y_pred)

print(cm)
print('Accuracy: ' + str(accuracy_score(y_test, y_pred)))
```

```
[[54629  2297]
 [ 5740 51060]]
Accuracy: 0.9293301443821114
```

Principal Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

In [16]:

```python
# Importing standardscalar module
from sklearn.preprocessing import StandardScaler
df = pd.DataFrame(X)
y_target = pd.DataFrame(y)
scalar = StandardScaler()

# fitting
scalar.fit(df)
scaled_data = scalar.transform(df)

# Importing PCA
from sklearn.decomposition import PCA

# Let's say, components = 2
pca = PCA(n_components = 0.95)
pca.fit(scaled_data)
x_pca = pca.transform(scaled_data)
print(x_pca.shape)
PCA_df = pd.DataFrame(x_pca)
PCA_df.head()
```

```
(568630, 17)
```

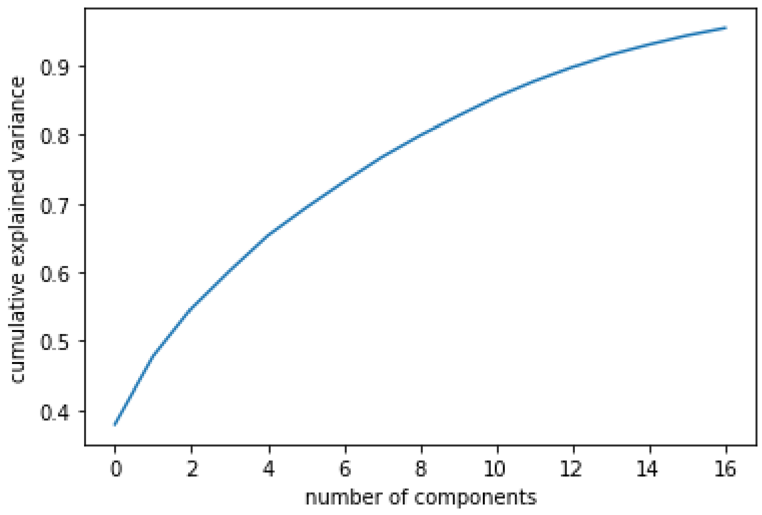
Out[16]:

	0	1	2	3	4	5	6	7	8	9	10	11	1
0	-2.017422	-0.079591	0.270550	0.674383	-0.385741	1.409463	-1.026928	0.507152	-1.805759	-0.222482	0.500370	0.295596	-0.03115
1	-1.938631	-0.518955	0.009032	-0.092563	-0.703678	0.711102	-1.065389	-0.913978	-1.023532	-0.655152	0.747481	-0.203345	0.08095
2	-2.099587	0.313580	0.095831	0.335490	0.099113	2.515647	-0.768611	-0.864740	-1.587526	-1.435280	0.386839	-0.321086	1.40594
3	-2.340797	-0.038814	-0.452258	-1.047547	-1.109917	1.074414	0.101514	-1.381527	-0.728062	-1.592945	0.695028	0.530310	-0.65105
4	-1.898936	-0.053823	0.790057	0.433809	-0.245760	0.886058	-1.582719	-0.982657	0.132393	-0.345936	1.300697	-0.919169	-0.41544

17 variables justify 95% variability in my data

In [17]:

```python
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance');
```



Training data and testing data

In [18]:

```python
X1_train, X1_test, y1_train, y1_test = train_test_split(PCA_df, y, test_size=0.2,random_state=1)
```

Logistic Regression Model

In [19]:

```python
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=0)
classifier.fit(X1_train, y1_train)
y1_pred = classifier.predict(X1_test)
y1_pred2 = classifier.predict(X1_train)
```

Model Evaluation

In [20]:

```
from sklearn.metrics import confusion_matrix,accuracy_score

cm1 = confusion_matrix(y1_test, y1_pred)

print(cm1)
print('Accuracy: ' + str(accuracy_score(y1_test, y1_pred)))
```

```
[[55550  1376]
 [ 5481 51319]]
Accuracy: 0.9397059599388002
```

Inference:

LDA & PCA both performed similarly on the datatset contributing to the model accuracy which was observed to be **92.93%** and **93.96%** respectively. The **information contained within 60 features was retained in 17 features approximately** and we did not lose much of information as we saw good accuracy.

In [21]:

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()

#fitting
sc.fit(X)
X = sc.transform(X)
```

In [22]:

```
from sklearn.decomposition import TruncatedSVD

print("Original Matrix:")
print(X,'\n')

svd =  TruncatedSVD(n_components = 4)
X_transf = svd.fit_transform(X)

print("Singular values: \n")
print(svd.singular_values_, '\n')

print("Transformed Matrix after reducing to 2 features: \n")
print(X_transf)
```

```
Original Matrix:
[[-1.82317125  0.18495473 -0.50611358 ...  0.04759203 -0.12984457
   0.17479509]
 [-1.82317125  0.64671985 -0.41494358 ... -0.09265743 -0.05086622
  -0.40402063]
 [-1.82315048  0.18521769 -0.84702927 ... -0.13828151 -0.21527228
   1.07707476]
 ...
 [ 1.09027437  0.60313741  0.38847542 ...  0.08311508  0.1556491
  -0.41158427]
 [-0.35229583  0.46663122 -0.16650073 ...  0.04922196  0.48465339
  -0.41162367]
 [ 1.30211714  0.76634433 -0.19882401 ... -0.04222082  0.00619462
  -0.41067821]]

Singular values:

[2542.83877106 1295.96156021 1090.67602695  966.24091486]

Transformed Matrix after reducing to 2 features:

[[-2.01742172 -0.07958723  0.27057454  0.67440123]
 [-1.93863053 -0.518954    0.00904835 -0.09269215]
 [-2.09958693  0.31359854  0.0959183   0.33562026]
 ...
 [-1.82230282 -0.11685556 -1.42931602 -2.3445905 ]
 [-1.96656279  0.29998635 -0.47543002 -0.77941364]
 [-2.24188402  0.52321109 -0.995532   -2.0965485 ]]
```

In [23]:

```
X_train, X_test, y_train, y_test = train_test_split(X_transf, y, test_size=0.2,random_state=1)
```

In [24]:

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=1)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
y_pred2 = classifier.predict(X_train)
```

In [25]:

```python
from sklearn.metrics import confusion_matrix,accuracy_score

cm = confusion_matrix(y_test, y_pred)

print(cm)
print('Accuracy: ' + str(accuracy_score(y_test, y_pred)))
```

```
[[56398   528]
 [ 9427 47373]]
Accuracy: 0.9124650475704764
```

In [32]:

```python
var_explained = svd.explained_variance_ratio_
var_explained
```

Out[32]:

```
array([0.37904138, 0.09845397, 0.06973335, 0.05472931])
```

In [33]:

```python
def select_n_components(var_ratio, goal_var: float) -> int:
    # Set initial variance explained so far
    total_variance = 0.0

    # Set initial number of features
    n_components = 0

    # For the explained variance of each feature:
    for explained_variance in var_ratio:

        # Add the explained variance to the total
        total_variance += explained_variance

        # Add one to the number of components
        n_components += 1

        # If we reach our goal level of explained variance
        if total_variance >= goal_var:
            # End the loop
            break

    # Return the number of components
    return n_components
# Run function
```

In [34]:

```python
select_n_components(var_explained, 0.95)
```

Out[34]:

```
4
```

The model performance with **SVD** had an accuracy of **91.24%** since we scaled down the whole dataset into 4 features, while PCA gave us an accuracy of **93.96%** with 17 components.**LDA** performed with accuracy of **92.93%**. With all the three techniques, we did not witness any major loss of information of our original data but we can say that LDA and SVD work better than PCA, given the number of features in each technique.

In [ ]:

```python

```