Credit Card Fraud Detection

This python solution works on kaggle dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud.

Dataset characteristics

- Highly imbalanced dataset with only 0.172% of transactions being classified as fraud.
- The features are PCA transformed
- The features are highly skewed.

Approach

- Dataset is divided in train and test with 70:30 ratio keeping in consideration that both the classes are proportionally divided.
- Power transformation is done on train and test dataset to treat the skewness.
- Following classification models are applied with roc-auc as scoring mechanism:
  XGboost Classifier: 0.9019
  Logistic Regression: 0.8916
  SVM with linear kernel: 0.8917
  Decision Tree Classifier: 0.8578
  SVM with polynomial kernel of degree 2: 0.8546
  RandomForestClassifier: 0.7634
  SVM with RBF kernel: 0.5337
- Hyper-parameter tuning is done using cross-validation to choose optimal parameters
- Data imbalance is treated using 3 techniques and different models are applied to gauge performance on test split:
  Technique 1 => Random oversampling
  Logistic Regression: 0.9276
  SVM with linear kernel: 0.9246
  XGboost Classifier: 0.9154

  Technique 2 => SMOTE
  SVM with linear kernel: 0.9344
  Logistic Regression: 0.9274
  XGboost Classifier: 0.9187

  Technique 3 => ADASYN
  Logistic Regression: 0.9222
  SVM with linear kernel: 0.9171
  XGboost Classifier: 0.9153