Credit Card Fraud Detection - Using machine learning

**Overview**

Credit card fraud has emerged as major problem in the electronic payment sector.

In this survey, I studied data-driven credit card fraud detection particularities and several machine learning methods to address each of its intricate challenges with the goal to identify fraudulent transactions that have been issued illegitimately on behalf of the rightful card owner.

**Data and Features**

A sample of 284807 was given , which included 30 columns as features.

Features are Time , Amount , V1 ,V2 , …. , V28.

Fraud detection gives each observation as value between 0 and 1 for these features.

i.e. Target variable which is Class .

**Exploratory Data Analysis**

During Exploratory data analysis (EDA) on the data I found that there was no null observations or misprint values in dataset.

And applied Standard scaler to make each column values as standard values i.e.

After Scaling unit of each column become same.

But by visualization count plot I found that data is imbalance.

So to make data balance I used Sampling technique.

Here data is in large quantity so to get better accuracy I used Random Over Sampling technique.

Before using Sampling Techniques (Random over sampling technique)



So I split data into two parts training and testing for modeling.

After using Sampling Techniques on training data set (Random over sampling technique)

****

After using Sampling Techniques on testing data set (Random over sampling technique)



**Models and Results :**

Given the unbalanced nature of the dataset, any model chosen would automatically yield high accuracy. So, in addition to aiming for high accuracy,

Here's a list of the models , Feature Selection Techniques and Ensembling techiniques I tested:

1. Logistic Regression
2. Decision tree classifier (Gini index)
3. Decision tree classifier (Gini index) max\_depth pruning technique
4. Decision tree classifier (Gini index) min\_samples\_leaf pruning technique
5. Decision tree classifier (Entropy)
6. Decision tree classifier (Entropy) max\_depth pruning technique
7. Decision tree classifier (Entropy) min\_samples\_leaf pruning technique
8. Support Vector Machine (Hard margin)
9. Support Vector Machine (Soft margin)
10. Naive aggregation method (Hard voting)
11. Random Forest Classifier
12. Bootstraping - Bagging Classifier
13. Bootstraping - Pasting
14. Stacking Classifier
15. Ada Boost Classifier
16. Gradient Boosting Classifier
17. Extreme Gradient Boosting Classifier (XGBClassifier)

**Model Summaries :**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.90 |
| Decision tree classifier (Gini index) | 0.72 |
| Decision tree classifier (Gini index) max\_depth pruning technique | 0.84 |
| Decision tree classifier (Gini index) min\_samples\_leaf pruning technique | 0.84 |
| Decision tree classifier (Entropy) | 0.70 |
| Decision tree classifier (Entropy) max\_depth pruning technique | 0.86 |
| Decision tree classifier (Entropy) min\_samples\_leaf pruning technique | 0.82 |
| Support Vector Machine (Hard margin) | 0.89 |
| Support Vector Machine (Soft margin) | 0.88 |
| Naive aggregation method (Hard voting) | 0.89 |
| Random Forest Classifier | 0.76 |
| Bootstraping - Bagging Classifier | 0.90 |
| Bootstraping - Pasting | 0.90 |
| Stacking Classifier | 0.84 |
| Ada Boost Classifier | 0.90 |
| Gradient Boosting Classifier | 0.87 |
| Extreme Gradient Boosting Classifier (XGBClassifier) | 0.79 |

**Conclusion :**

The best model or method after testing seems to be Ada Boost Classifier.