Credit Card Fraud

1. INTRODUCTION TO DATASET

https://www.kaggle.com/mlg-ulb/creditcardfraud/home

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.

Credit Card Fraud

2. EXPLORATORY DATA ANALYSIS

- **PCA FEATURES-** Features V1, V2, ... V28 are the principal components obtained with PCA. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data.
- **TIME AND AMOUNT-** They're the only features which have not been transformed with PCA.
  - **'Time'** contains the seconds elapsed between each transaction and the first transaction in the dataset.
  - **'Amount'** is the transaction Amount.
- **CLASS-** It's the response variable and it takes value 1 in case of fraud and 0 otherwise.

Credit Card Fraud

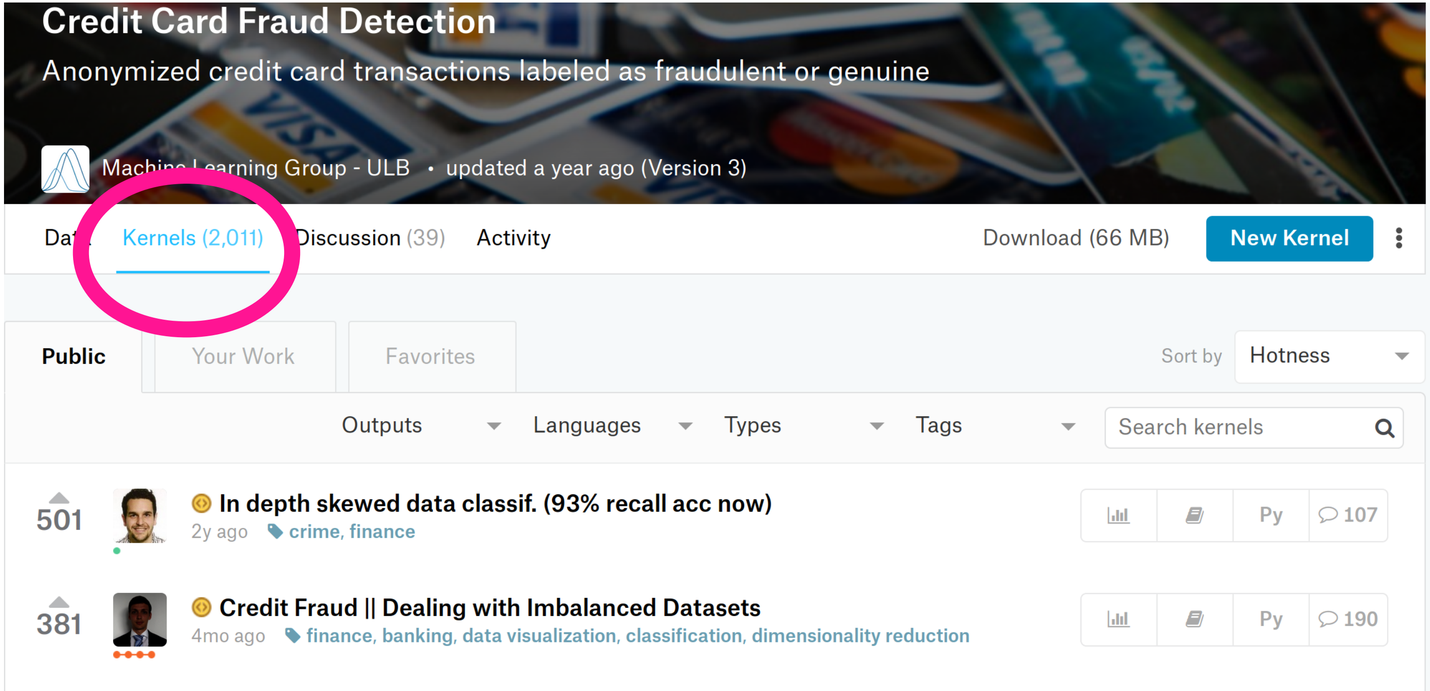
3. MODEL OUTCOME OF INTEREST (Finally)

- You should try several different approaches and really work to tune a variety of models before choosing what you consider to be the best performer.
- Make sure to think about explanatory versus predictive power and experiment with both.

Credit Card Fraud

4- RESEARCH QUESTION

- What is the best way to predict frauds? (show next)



Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

Machine Learning Group - ULB • updated a year ago (Version 3)

Data Kernels (2,011) Discussion (39) Activity Download (66 MB) New Kernel

Public Your Work Favorites Sort by Hotness

Outputs Languages Types Tags Search kernels

501 **In depth skewed data classif. (93% recall acc now)**
2y ago crime, finance
Py 107

381 **Credit Fraud || Dealing with Imbalanced Datasets**
4mo ago finance, banking, data visualization, classification, dimensionality reduction
Py 190

Credit Card Fraud

- **Focus on reducing false negatives.**

  VS
- **Focus on reducing false positives.**

  VS
- **Focus on a custom balance?**

Credit Card Fraud

5. HOW YOU CHOSE YOUR MODEL?

- Custom scoring function.
  - Recall vs Precision in a single score.
    - Tested with all most commonly used SKLearn Classifiers.
  - Optimum parameter combination.
    - Performance varies with settings on classifier.
  - Several iterations of model processing.
    - Outlier handling greatly influenced prediction scores.
    - Class-Balancing Techniques.

Credit Card Fraud

6. PRACTICAL USES FOR AUDIENCE OF INTEREST

- Bank’s fraud-prevention mechanisms.
  - *(Annoying: Transactions canceled when traveling)*
- Data Science students.
  - Addition to the pool of Kaggle’s forks on this Dataset.

Credit Card Fraud

7. WEAK POINTS OR SHORTCOMING?

- **Model Processing-** Involves many steps. Steps depend immensely on the data. Doesn't lend itself to quick iterations.
- **Need for Data Reduction-** 270,000 non-frauds were undersampled to 5,000… Definitely affected precision. A supercomputer might handle complete set without the need for reduction.
  - SVM and Kneighbors took the longest with larger model.