Credit card fraud detection

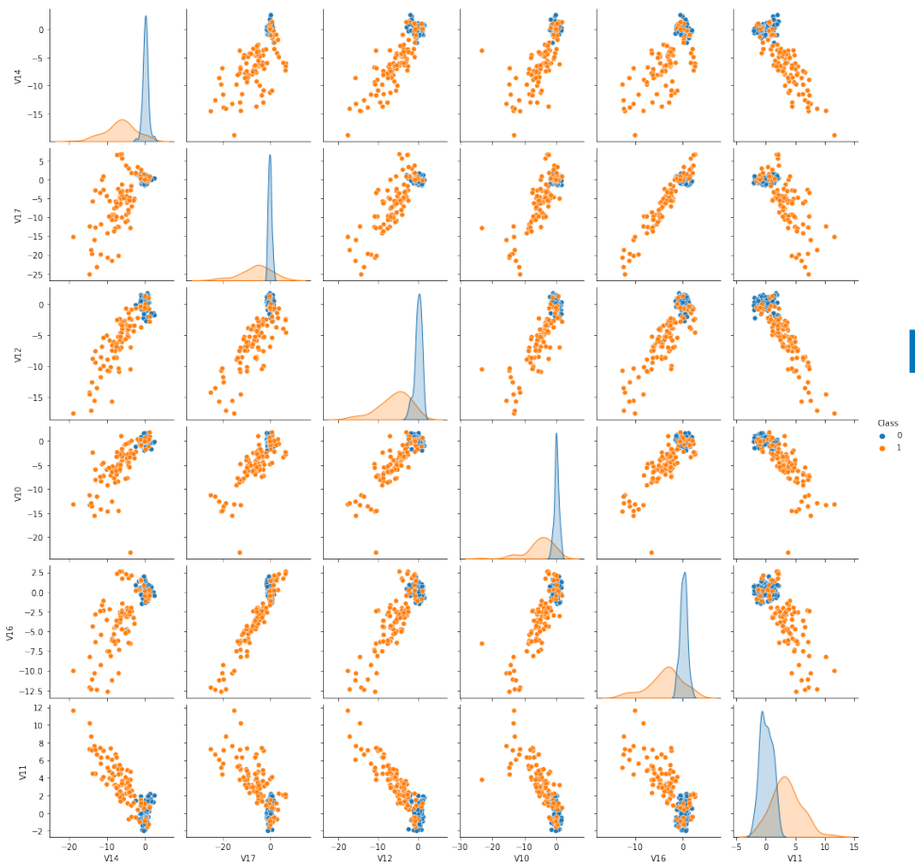
Minjung Kim

Overview

- Goal
    - Detect most of fraudulent transaction without losing too much precision
    - Build a simple intuitive model in order to tune easily whenever fraudulent trend changes
- Dataset
    - Provided by the Machine Learning Group of Université Libre de Bruxelles.
    - https://www.kaggle.com/mlg-ulb/creditcardfraud (144 MB, too large to upload to GitHub)
    - 284,807 transactions, 492 of them (0.172%) are frauds.
    - 30 numerical features: time, amount of money, 28 PCA transformed components of encrypted data
- Model
    - Supervised anomaly (outlier) detection based on Z-scores
    - Selected features having “high classification power”
    - Suggested best hyper parameters for two kind of “best choices”
- Result
    - Both models showed expected and excellent performance on the test set

Selecting features of High Classification Power (CP)

High CP



Fraud

Normal

Low CP

Metric: Z-score, using mean and standard deviation of normal sample for each feature

Drawn for 100 samples for both classes

Selected features of n-highest CP -> **“Leading features”**

Anomaly metric

- Anomaly (outlier) metric: squared sum of Z scores for leading features

$$Z_{fraud} = \sqrt{\frac{1}{n} \overset{n \text{ leading features}}{\sum_{i}} \frac{(Z_i - \mu_i)^2}{\sigma_i^2}} > \epsilon_{fraud}$$

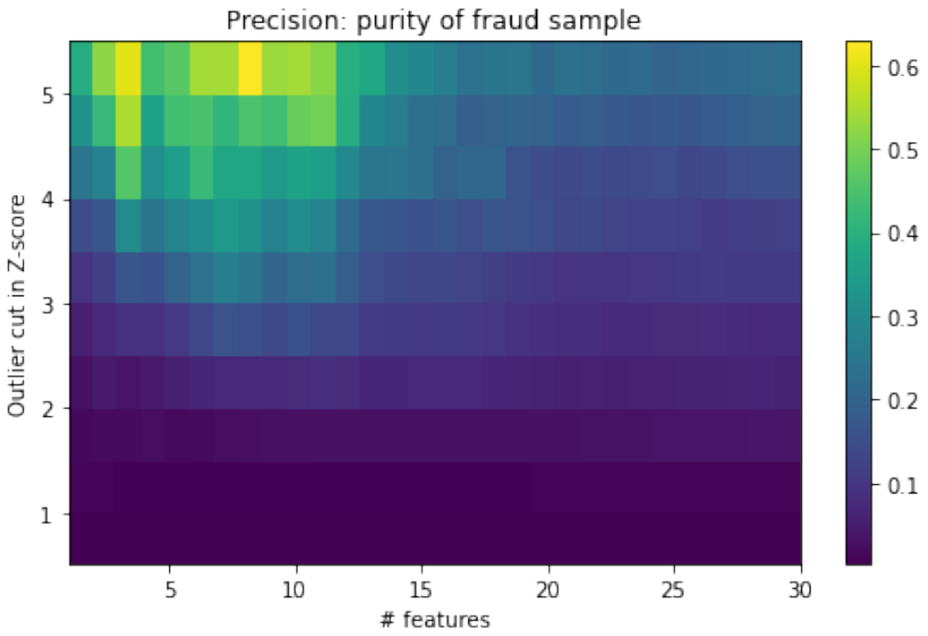
- μ_i, σ_i : mean and standard deviation of i-th feature of normal transaction training set

- Hyper parameters

  - n : number of leading features

  - ϵ_{fraud} : outlier cut in Z-score

Hyper parameter tuning - Precision



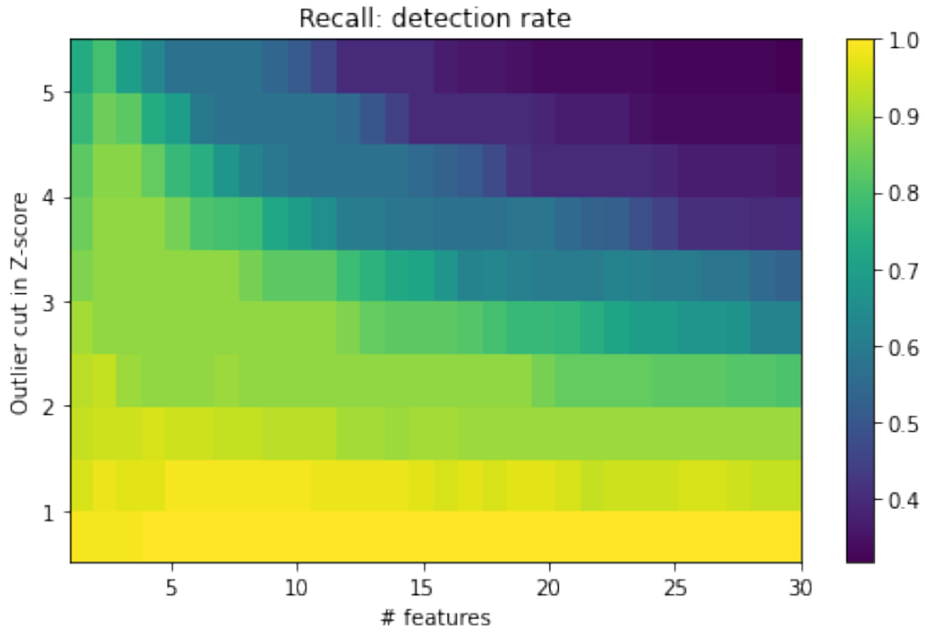
$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Purity of fraud sample

Customers will be annoyed if it's too low

- Precision goes higher as the number of leading features increases until it reaches about 10, then decreases
  - The more relevant features we use, the detection becomes more strict
  - When we start to use more irrelevant features, then the detection doesn't enhance fraud selection
- Precision goes higher as ϵ_{fraud} increases
  - Assuming normal samples have sharper distribution, higher ϵ_{fraud} always increases purity of fraud sample
- Even at the highest, the precision is much lower than 1, due to heavily skewed statistics of classification samples

Hyper parameter tuning - Recall



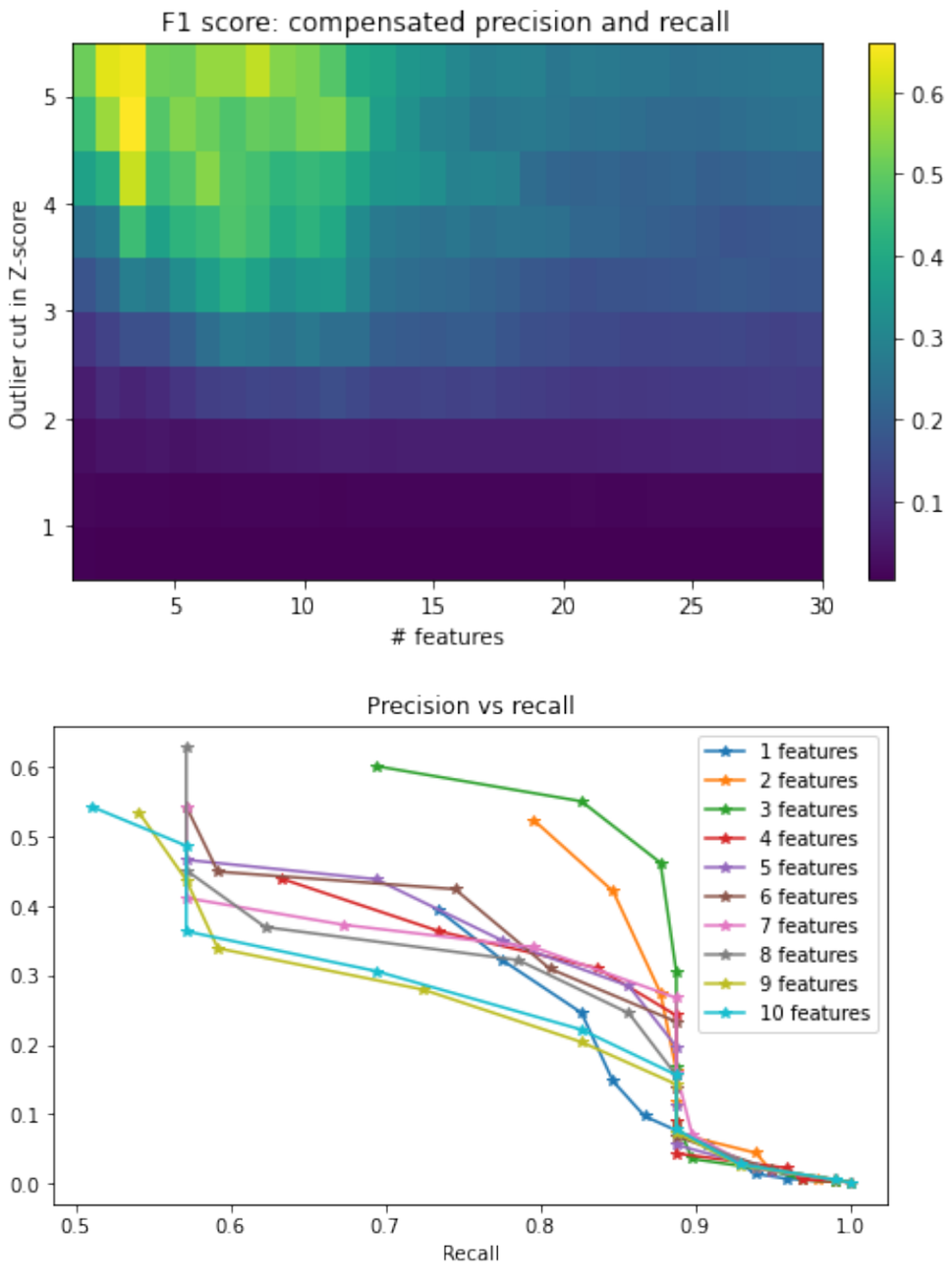
$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Detection rate of fraud sample

Customers will be scared if it's too low

- Recall goes higher as ϵ_{fraud} decreases
  - Not all fraud samples are separable from normal samples, so lower ϵ_{fraud} will only keep more fraud sample. However the precision is very small in the lowest ϵ_{fraud} area.
- Recall goes higher as the number of leading features decreases, except using only one feature
  - For a given ϵ_{fraud}, using a few multiple leading features detect more fraud samples than using too many features

Hyper parameter tuning - F1 score



$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

Harmonic mean of recall and precision

- The highest F_1 score is achieved when 2-3 leading features with highest ϵ_{fraud} are applied
  - Observation from precision and recall explain this result
  - 2-3 leading features do best on ROC curve, too

How to select final parameters

- Highest F_1 (or F_β) score is often the best selection.
- However, fraudulent transactions are so fatal, so you also want to keep a certain high recall value even though you lose some precision.
- I selected final parameters based on the overall trends, not relying on precise scoring too much. Here are reasons.
  - The number of fraud sample has 10% of high statistical error assuming that it follows common probability distributions. Besides, this kind of anomaly sample distributions are chaotic.
  - I can imagine that the precision won't be stable for each sampling or over time simply because the frequency of fraudulent transactions won't be regular (if so, it will be interesting…). Here, note that recall might be relatively stable than precision, assuming fraudulent transactions have much broader distribution than normal one, until fraudulent transaction techniques are evolved and become close to normal transactions
- As a conclusion, I suggest two parameter choices with two different kind of “best-overall" performance

Choice 1
Highest precision at recall > ~90%

- From cross validation set, top highest precision having recall > 90% were mostly achieved by parameters with a few multiple leading features (2-5 features) and moderately low ϵ_{fraud} (1.5-2.5) achieved.

- From final test, 88% recall and 3.6% precision is obtained with 2 leading features and ϵ_{fraud} = 2.

- Evaluation

    - System with ~90% recall means that this system catches 90% of fraudulent transaction, which sounds very safe.

    - From our model, the precision was a few percent, which is too low to halt (customers will be annoyed), however, quite high to be the first level of sampling.

    - The advantage of this sampling is that it reduces the size of "samples to be investigated" significantly without losing most of fraudulent samples, therefore, we can build a several levels of efficient detector.

- Application

    - One application example can be building multiple levels of fraud detector, combining faster first level and slower second (or higher) level

    - Level 1: Implement this model in as a fast hardware process, for example, add an operation which processes signals of two features and makes logic operation for comparison and decision in a card reader chip. If level1 detect fraud-like transaction, then it halt transaction and send feature signals to level 2

    - Level 2: Perform further and slower classification using software or information of higher-up security levels. If this decision says normal, then the transaction will be made (with delays introduced by level 2 processing time), and a customer will hardly notice anything. If this decision says fraud, then we can contact customer to ask security information.

Choice 2
Highest F1 score

- Highest F1 scores are achieved by parameters of 2-3 leading features and highest ϵ_{fraud}. From cross validation set, such parameters show good performance for both precision (40-60%) and recall (70-90%).

- From final test, 77% recall and 74% precision is obtained with 3 leading features and ϵ_{fraud} = 5.

- Evaluation

    - This system samples highly fraud-like transactions without losing too much of total fraud transactions. Considering only 0.17% are fraud, 74% of test precision is such a great focusing on crime.

- Application

    - One application can be a fast hardware level implementation in a card reader. If a transaction is marked as fraudulent, then the transaction is automatically halt, the transaction information is sent to security monitoring places and/or a staff in a shop can ask further questions to the credit card user.

Discussion and Conclusion

- Supervised anomaly detection model based on Z-score is built for fraud detection.

- Two best models are suggested, both perform excellent on test set

- As the trend of normal or fraud transaction changes, the parameters should be adjusted. This model provides intuitive way to tune the parameters.

- Some features might have skewed distribution. In such case, transformation to log might work better.

- Of course, knowing the meaning of each feature will significantly improve the model.