Weekly Feedback Slidedoc

Zhengqi Wang

June 26, 2024

Table of Contents

Overview of Home Credit Risk Model Stability

- The goal of this research is to predict which clients are more likely to default on their loans. The evaluation will favor solutions that are stable over time.
- My participation may offer consumer finance providers a more reliable and longer-lasting way to assess a potential client's default risk.

List of Researched Methods

1. Supervised Learning Techniques
    1. Logistic Regression
    2. Random Forest
    3. Extreme Gradient Boosting (XGBoost)
2. Linear Analysis Techniques
    1. Linear Regression
    2. Logistic Regression
3. Unsupervised Learning Techniques
    1. Isolation Forest Algorithm
    2. Local Outlier Factor (LOF)
    3. Principal Component Analysis (PCA)
    4. K-Nearest Neighbor (KNN)
    5. Histogram-based Outlier Score (HBOS)

Description

- Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

- Currently, consumer finance providers use various statistical and machine learning methods to predict loan risk. These models are generally called scorecards. In the real world, clients’ behaviors change constantly, so every scorecard must be updated regularly, which takes time. The scorecard’s stability in the future is critical, as a sudden drop in performance means that loans will be issued to worse clients on average. The core of the issue is that loan providers aren’t able to spot potential problems any sooner than the first due dates of those loans are observable. Given the time it takes to redevelop, validate, and implement the scorecard, stability is highly desirable. There is a trade-off between the stability of the model and its performance, and a balance must be reached before deployment.

- Our work in helping to assess potential clients' default risks will enable consumer finance providers to accept more loan applications. This may improve the lives of people who have historically been denied due to lack of credit history.

Evaluation

Submissions are evaluated using a *gini stability metric*. A gini score is calculated for predictions corresponding to each `WEEK_NUM`:

$$\text{gini} = 2 \times \text{AUC} - 1$$

A linear regression, $\alpha \cdot x + \beta$, is fit through the weekly gini scores, and a *falling_rate* is calculated as $\min(0, \alpha)$. This is used to penalize models that drop off in predictive ability.

Finally, the variability of the predictions are calculated by taking the standard deviation of the residuals from the above linear regression, applying a penalty to model variability.

The final metric is calculated as:

$$\text{stability metric} = \text{mean(gini)} + 88.0 \cdot \min(0, \alpha) - 0.5 \cdot \text{std(residuals)}$$

Data overview

Data Overview

Dataset Description

Objective: Predicting default of clients based on internal and external information.

- Scoring uses a custom metric evaluating both AUC and the stability of prediction models across the data range.
- Refer to the Evaluation tab for more detailed understanding of this metric.

Table Description

Dataset Characteristics:

- Contains multiple tables from diverse data sources.
- Available in both `.csv` and `.parquet` formats.

Base Tables:

- Store basic observation info and `case_id`.
- `case_id` is essential for joining with other tables.

File Overview

File Types:

- **Base Tables:** `train_base.csv`, `test_base.csv`
- **Static Files:** `train_static_0_0.csv`, etc.
- **Credit Bureau Data:** `train_credit_bureau_a_1_0.csv`, etc.

Note:

- The `test_base.csv` contains approx 90% of the numbers of `case_id` values of `train_base.csv`.

Credit Default Prediction using Supervised and Unsupervised Learning Techniques

Supervised Learning

Methods of Supervised

We use the following Machine learning algorithms and built Classification models for our supervised classification task.

- Logistic Regression
- Random Forest
- Extreme Gradient boosting

Supervised Learning: Random Forest

Random forests

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree ”votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m¡¡M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

In the original paper on random forests, it was shown that the forest error rate depends on two things:

- The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate.
- The strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.

In the original paper on random forests, it was shown that the forest error rate depends on two things:

- The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate.
- The strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.

Reducing m reduces both the correlation and the strength. Increasing it increases both. Somewhere in between is an ”optimal” range of m - usually quite wide. Using the oob error rate (see below) a value of m in the range can quickly be found. This is the only adjustable parameter to which random forests is somewhat sensitive.

Features of Random Forests

- It is unexcelled in accuracy among current algorithms.
- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.

Supervised Learning: Extreme Gradient Boosting

XGBoost

XGBoost is one of the most popular and efficient implementations of the Gradient Boosted Trees algorithm, a supervised learning method that is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques.

Linear Analysis: Linear Regression

Linear Analysis: Logistic Regression

Logistic Regression

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1).
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

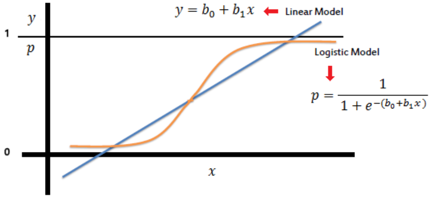

Figure: Comparison of Linear Model and Logistic Model. Image source:

In the logistic regression the constant b_0 moves the curve left and right and the slope b_1 defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio, like 1.

$$\frac{p}{1-p} = exp(b_o + b_1 x) \tag{1}$$

Finally, taking the natural log of both sides, as 2, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient b_1 is the amount the logit (log-odds) changes with a one unit change in x.

$$ln(\frac{p}{1-p}) = b_0 + b_1 x \tag{2}$$

Logistic regression 3 can handle any number of numerical and/or categorical variables.

$$p = \frac{1}{1 + e^{-(b_o + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)}} \tag{3}$$

Unsupervised Learning Techniques: Isolation Forest Algorithm

Isolation Forest Algorithm

One of the newest techniques to detect anomalies is called Isolation Forests. The algorithm is based on the fact that anomalies are data points that are few and different. As a result of these properties, anomalies are susceptible to a mechanism called isolation.

This method is highly useful and is fundamentally different from all existing methods. It introduces the use of isolation as a more effective and efficient means to detect anomalies than the commonly used basic distance and density measures. Moreover, this method is an algorithm with a low linear time complexity and a small memory requirement. It builds a good performing model with a small number of trees using small sub-samples of fixed size, regardless of the size of a data set.

Typical machine learning methods tend to work better when the patterns they try to learn are balanced, meaning the same amount of good and bad behaviors are present in the dataset.
How Isolation Forests Work The Isolation Forest algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The logic argument goes: isolating anomaly observations is easier because only a few conditions are needed to separate those cases from the normal observations. On the other hand, isolating normal observations require more conditions. Therefore, an anomaly score can be calculated as the number of conditions required to separate a given observation.

Unsupervised Learning Techniques: Local Outlier Factor(LOF) Algorithm

LOF Algorithm

The LOF algorithm is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outlier samples that have a substantially lower density than their neighbors.

The number of neighbors considered, (parameter n_neighbors) is typically chosen 1) greater than the minimum number of objects a cluster has to contain, so that other objects can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by objects that can potentially be local outliers. In practice, such informations are generally not available, and taking n$_{n}eighbors = 20 appearstoworkwellingeneral$.

Unsupervised Learning Techniques: Principal Component Analysis

PCA

We have PCA to learn the underlying structure of the Loan application dataset. The more anomalous the data is, the more likely it is to be fraudulent, assuming that fraud is rare and looks somewhat different than the majority of application data, which are normal. Once we learn this structure, we will use the learned model to reconstruct the Loan application data and then calculate how different the reconstructed data are from the original data. Those transactions that PCA does the poorest job of reconstructing are the most anomalous (and most likely to be fraudulent).

The algorithms will have the largest reconstruction error on those data points that are hardest to model—in other words, those that occur the least often and are the most anomalous. Since fraud is rare and presumably different than normal observations, the fraudulent observations should exhibit the largest reconstruction error. So let's define the anomaly score as the reconstruction error. The reconstruction error for each transaction is the sum of the squared differences between the original feature matrix and the reconstructed matrix using the dimensionality reduction algorithm. We will scale the sum of the squared differences by the max-min range of the sum of the squared differences for the entire dataset, so that all the reconstruction errors are within a zero to one range.

Unsupervised Learning Techniques: K-Nearest Neighbor

KNN

K-nearest neighbor: k-NN is a simple, non-parametric lazy learning technique used to classify data based on similarities in distance metrics such as Eucledian, Manhattan, Minkowski, or Hamming distance.
It uses the math behind the classification algorithm KNN. Indeed, for any data point, the distance to its kth nearest neighbor could be viewed as the outlying score. PyOD supports three KNN detectors: largest, mean and median, which use as outlying score, respectively, the distance of the kth neighbor, the average of all the k neighbors and the median distance to k neighbors.

Unsupervised Learning Techniques: Histogram-based outlier score

Histogram-based outlier score

It is only a combination of univariate methods not being able to model dependencies between features, its fast computation is charming for large data sets. Histogram-Based Outlier Score (HBOS) is an efficient unsupervised method. It assumes the feature independence and calculates the degreeof outlyingness by building histograms.

For each single feature (dimension), an univariate histogram is constructed first. If the feature comprises of categorical data, simple counting of the values of each category is performed and the relative frequency (height of the histogram) is computed. For numerical features, two different methods can be used:

1. Static bin-width histograms
2. Dynamic bin-width histograms

The first is the standard histogram building technique using k equal width bins over the value range. The frequency (relative amount) of samples falling into each bin is used as an estimate of the density (height of the bins). The dynamic binwidth is determined as follows: values are sorted first and then a fixed amount of N.

Credit Default Prediction using Supervised and Unsupervised Learning Techniques