Zhejiang University
School of Mathematics and Statistics

Interpretable machine learning

Data visualisation

Professor   Student
Prof. Andre Python   Kevin Lin

Student ID:22235060   2023/03/05

1 Dataset Introduction

Problem description:

Dataset Description:

The dataset in question contains credit card transactions made by European cardholders in September 2013. The purpose of this dataset is to aid credit card companies in recognizing fraudulent transactions to ensure that customers are not charged erroneously for purchases they did not make.

The dataset includes transactions that occurred over a period of two days, where there were 492 instances of fraud out of 284,807 total transactions. It is worth noting that the dataset is highly unbalanced, as the positive class (i.e. fraud) accounts for only 0.172% of all transactions. This makes it particularly challenging to accurately predict instances of fraud.

The dataset consists solely of numerical variables resulting from a Principal Component Analysis (PCA) transformation. Unfortunately, due to confidentiality concerns, the original features and additional background information about the data cannot be provided. Aside from 'Time' and 'Amount', which were not transformed using PCA, features V1 through V28 represent the principal components obtained from the PCA analysis.

'Time' represents the elapsed time (in seconds) between each transaction and the first transaction in the dataset, while 'Amount' represents the amount of the transaction. The feature 'Class' is the response variable and takes a value of 1 if the transaction is identified as fraudulent, and 0 otherwise.

Given the class imbalance ratio in the data, it is recommended that accuracy be measured using the Area Under the Precision-Recall Curve (AUPRC), rather than through a confusion matrix accuracy score, which may not be meaningful for unbalanced classification.

To solve this problem, we need to build a model that can detect fraudulent transactions based on the features. One possible method is to use a logistic regression model, which is a type of supervised learning algorithm that can perform binary classification. The logistic regression model estimates the probability of a transaction being fraudulent using a sigmoid function:

$$p(y = 1|x) = \frac{1}{1 + e^{-\theta^T x}}$$

where x is the feature vector, θ is the parameter vector and y is the class label.

To train the logistic regression model, we need to find the optimal values of θ that minimize the cost function, which is defined as:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(p(y^{(i)} = 1|x^{(i)})) + (1 - y^{(i)})\log(1 - p(y^{(i)} = 1|x^{(i)}))]$$

where m is the number of training examples.

We can use an optimization algorithm such as gradient descent to update θ iteratively until convergence:

$$\theta := \theta - \alpha\frac{\partial J(\theta)}{\partial \theta}$$

where α is the learning rate.

However, since the dataset is highly unbalanced, we may need to apply some techniques to deal with the class imbalance problem, such as oversampling, undersampling or using a weighted cost function. These techniques can help improve the performance of the model and reduce the bias towards the majority class.

Example: The Distribution of Students' Grades

Compute the perceputal distortion:

2   Some Improvement Aspects

3   Summary