Big Data Analytics com R e Microsoft Azure Machine Learning
Versão 3.0

Estudo Dirigido

Balanceamento de Classes em Dados de Fraudes Financeiras com ROSE (Random OverSampling Examples)

O objetivo do Estudo Dirigido é servir como material de auto-estudo, demonstrando na prática e de forma pontual, conceitos importantes em Data Science e Machine Learning, complementando assim o conhecimento que você está recebendo neste treinamento.

Leia **ATENTAMENTE** cada comentário no script em anexo, execute todas as linhas, analise os resultados, faça mudanças no script para compreender o impacto dessas mudanças e o mais importante: procure compreender o que está sendo feito e porque está sendo feito. Você não está aqui para ficar reproduzindo scripts, mas sim para aprender de verdade e por isso o conteúdo da Data Science Academy é denso e diferenciado.

No caso de dúvidas, nossos canais de suporte estão à disposição (o manual de suporte está no Capítulo 1 do curso).

Definição do Problema

É importante que as operadoras de cartão de crédito possam reconhecer transações fraudulentas no momento exato em que elas estiverem ocorrendo, para que os clientes não sejam cobrados pelos itens que não compraram.

Nosso objetivo neste trabalho de análise é identificar um problema comum quando trabalhamos com dados que apresentam anomalias (fraudes, nesse caso).

Em cenários assim, temos uma situação comum que precisa ser tratada:

Conforme você já sabe, usamos dados históricos para treinar modelos de Machine Learning. Esperamos que a operadora de cartão de crédito tenha muito mais exemplos históricos de transações corretas do que transações fraudulentas (se essa premissa não fosse verdadeira, a empresa já teria ido à falência, concorda?).

Mas se entregarmos os dados dessa forma ao modelo de Machine Learning, ele vai aprender mais sobre uma categoria de transações do que outra. Imagine por exemplo que a empresa tenha essa massa de dados de um dia de transações de cartão de crédito:

- 25.000 exemplos de transações corretas (classe majoritária)
- 314 exemplos de transações fraudulentas (classe minoritária)

Esse é tipicamente um problema de classificação, em que o modelo de Machine Learning deve analisar cada transação e classificar como fraudulenta ou não fraudulenta.

Cada modelo de Machine Learning procura pelo relacionamento matemático nos dados. Mas nosso dataset está desbalanceado e, nesse caso, o modelo vai aprender muito mais sobre uma transação normal do que sobre uma transação fraudulenta. Como resultado, o modelo pode classificar novas transações fraudulentas como se fossem transações normais, simplesmente porque aprendeu mais sobre uma classe do que sobre a outra.

Para minimizar esse problema, podemos aplicar uma de muitas técnicas de balanceamento de classes, criando dados sintéticos para aumentar o volume de transações fraudulentas (isso é chamado de oversampling) ou então podemos remover alguns registros da classe de transações normais (isso é chamado de undersampling).

O undersampling é mais fácil, mas reduz o tamanho do dataset, o que não é o ideal. O oversampling pode ser mais trabalhoso e mais complicado de explicar, porém aumenta o tamanho do dataset criando dados sintéticos com base em regras estatísticas e de forma aleatória, usando observações da classe minoritária como ponto de partida.

Neste estudo dirigido usaremos uma técnica de balanceamento de classes chamada **Randomly OverSampling Examples (ROSE)** e com um pacote R perfeito para essa tarefa, chamado.....ROSE.

Dada a taxa de desequilíbrio de classe, o ideal é medir a precisão usando a métrica Área Sob a curva Precision-Recall (AUPRC). Usar apenas a acurácia definida pela matriz de confusão não é significativo para a classificação com classes desbalanceadas e também veremos isso.

Usaremos o dataset público disponibilizado pelo Machine Learning Group. O dataset deve ser baixado do link abaixo (o dataset não será fornecido com o script, pois o arquivo é grande):

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

O conjunto de dados contém transações realizadas com cartões de crédito em setembro de 2013 por portadores de cartões europeus.

Esse conjunto de dados apresenta transações que ocorreram em dois dias, nas quais temos 492 fraudes em 284.807 transações. O conjunto de dados é altamente desequilibrado, a classe positiva (fraudes) representa 0,172% de todas as transações.

Ele contém apenas variáveis de entrada numéricas que são o resultado de uma transformação PCA. Devido a problemas de confidencialidade, não se pode fornecer os recursos originais e mais informações básicas sobre os dados. Recursos V1, V2,… V28 são os principais componentes obtidos com o PCA, os únicos recursos que não foram transformados com o PCA são 'Tempo' e 'Valor'. O recurso 'Hora' contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados. O recurso 'Valor' é o valor da transação. O recurso 'Classe' é a variável de resposta e assume o valor 1 em caso de fraude e 0 em caso contrário.

O script em anexo oferece conhecimento importante para seu aprendizado. Aproveite mais esse material fornecido a você pela Equipe DSA.

Bons estudos!