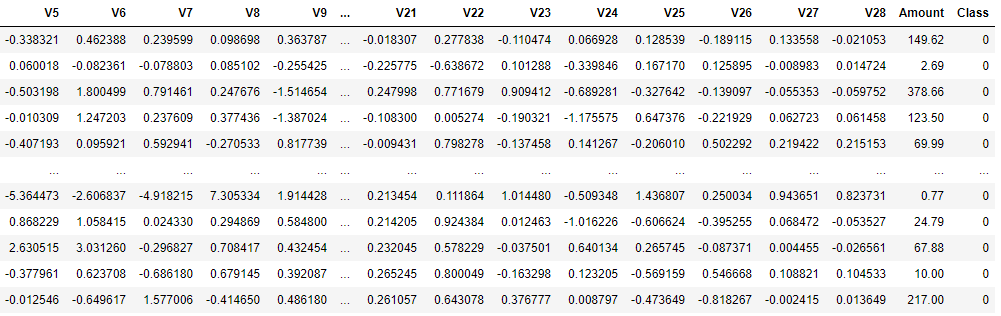
Análise do Credit Card Fraud Detection Dataset

A análise a seguir se refere ao Credit Card Fraud Detection Dataset, um dataset público que contém informações sobre transações, normais e fraudulentas, feitas por cartões de créditos da Europa em setembro de 2013. Download feito no site <https://www.kaggle.com/mlg-ulb/creditcardfraud>*.* Utilizou-se o método de *machine learning – logistic regression* feito na linguagem de programação *python* para a análise do *dataset*.

O *dataset* contém mais de 280.000 transações, das quais apenas 492 são fraudulentas. Isso representa 0.172% de todas as transações, ou seja, temos um *dataset* desbalanceado, um possível problema para o modelo. O *dataset* possui 30 *features*, V1 a V28 (por razões de confidencialidade, não foi possível obter mais informações sobre as *features*), *Time* e *Amount*, o valor da transação. A *feature* Time contém o tempo decorrido entre a primeira transação e as demais. Não é uma informação relevante para análise, então foi desconsiderada. *Amount* é o quanto foi gasto na transação, portanto é uma *feature* possivelmente importante para a análise. V1 a V28 são os restantes das *features* que foram analisadas. A última coluna do *dataset* (*Class*) se refere se a transação foi ou não fraudulenta. Ela é o *target* do estudo.

Tabela 1 – Dados carregados no jupyter notebook

Com os dados carregados no *jupyter notebook*, como visto na tabela 1 (V1 a V5 não apareceram na tabela 1, pois são muitos dados), foi separado em 2 variáveis os *inputs* e o *target*. I*nputs* foram as *features* V1 a V28 e *Amount*. Já o target foi Class.

Para uma melhor análise, os dados foram embaralhados randomicamente, já que eles coletados e armazenados numa mesma janela de tempo. No fim do estudo também foi feito um modelo sem o embaralhamento para comparação dos resultados.

Em seguida, foi feito o balanceamento do *dataset*, pois o número de transações fraudulentas em relação ao total é muito baixo e isso pode prejudicar o algoritmo do modelo, levando a interpretações e resultados errados.

O próximo passo foi separar o *dataset* em 80% dados de treino e 20% dados de teste que foram usados para checar a precisão final do modelo com dados novos.

|  |  |
| --- | --- |
| ***Feature Name*** | ***Coefficient*** |
| V22 | 1,1664 |
| V4 | 0,7353 |
| V28 | 0,7039 |
| V23 | 0,5223 |
| V18 | 0,5008 |
| V26 | 0,2715 |
| V17 | 0,2032 |
| V7 | 0,2017 |
| V1 | 0,1029 |
| V11 | 0,0952 |
| V16 | 0,0666 |
| V13 | 0,0567 |
| V5 | 0,0131 |
| Amount | 0,0003 |
| V9 | -0,0996 |
| V19 | -0,1653 |
| V21 | -0,2211 |
| V2 | -0,2831 |
| V14 | -0,3245 |
| V10 | -0,4452 |
| V6 | -0,5153 |
| V8 | -0,5528 |
| V15 | -0,6523 |
| V20 | -0,6758 |
| V24 | -0,7697 |
| V25 | -0,8224 |
| V27 | -0,8461 |
| V3 | -0,8540 |
| V12 | -1,1498 |
| Intercept | -2,3413 |

Tabela 2 – Valores dos coeficientes e *intercept* - dados randomizados

Figura 1 - Valores dos coeficientes - dados randomizados

Na tabela 2, temos os valores dos coeficientes para o modelo *logistic* *regression* e na figura 1 a representação visual dos valores. Pode-se perceber que o valor de *Amount* é próximo de zero e não é impactante para o modelo, ou seja, na prática, o valor das transações não é um fator relevante para classificar se a transação é fraudulenta ou não. Pode ser uma transação de R$5,00 ou R$10.000,00, não é relevante. Outras *features* que não são relevantes são V9, V5. Não é possível verificar a razão desse comportamento, pois não foram fornecidas as informações sobre o que representa cada *feature*. V22 e V12 se motram as features mais importantes para o modelo, mas, pelo menos motivo apresentado anteriormente, não é possível comentar o motivo do seu impacto.

|  |  |
| --- | --- |
| *Accuracy* | |
| Dados Treino | 96,57 % |
| Dados Teste | 95,43 % |

Tabela 3 – Valores da precisão do modelo – dados randomizados

Os resultados obtidos da *accuracy* do modelo podem ser vistos na tabela 3. A diferença obtida com os dados de treino e teste são esperados e, como a diferença entre eles é pequena, não é necessária uma preocupação inicial com *overfitting,* já que o modelo se comportou bem na presença de novos dados. Então, foi obtido um bom modelo e sem a presença de *overfitting*, podendo classificar transações fraudulentas com uma *accuracy* de 95 %, o que é um resultado ótimo para uma operação vital e sensível nos modelos econômicos atuais.