PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

JÉSSICA FERREIRA SILVA

DETECÇÃO DE FRAUDE EM TRANSAÇÕES COM CARTÃO DE CRÉDITO

Belo Horizonte
2021

JÉSSICA FERREIRA SILVA

DETECÇÃO DE FRAUDE EM TRANSAÇÕES COM CARTÃO DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução

1.1. Contextualização

O cartão de crédito vem, dia após dia, se popularizando e aumentando sua frequência de utilização - e, consequentemente, é alvo de tentativas de fraudes e golpes. Por isso, é preciso conhecer os diferentes tipos de fraude de cartão de crédito para fortalecer os processos da empresa e melhor orientar os clientes.

Muitos dos casos de golpes e fraudes de cartão de crédito são resultado de pequenos descuidos ou incidentes por parte dos usuários, mas as organizações também podem construir ferramentas antifraude e de segurança para que compras indevidas não sejam aprovadas.

Mas como fazer isso de forma eficaz e sem atrapalhar o dinamismo e a agilidade de operações lícitas? É importante que as operações de cartões de crédito consigam reconhecer transações fraudulentas no momento exato em que elas estiverem ocorrendo, para que os clientes não sejam cobrados por itens que não compraram e não haja danos financeiros e de recuperação à operadora de cartão de crédito.

1.2. O problema proposto

O problema considerar em aplicar técnicas de Análise Exploratória e Modelos de Classificação para extração de informações importantes sobre transações fraudulentas.

Para isso, serão analisadas as variáveis contidas no dataset. Assim, tem-se como objetivos dessa análise:

- Verificar a qualidade dos dados;
- Análise exploratória dos registros;
- Amostragem – treino e teste;

- Balanceamento da base de treino;
- Geração dos modelos de classificação – Regressão logística, Árvore de decisão, Análise de discriminante, Naive Bayes e Support Vector Machine (SVM).
- Aplicação do modelo na base de treino;
- Comparação dos resultados (maior ROC);
- Análise dos decis de probabilidade no modelo campeão;
- Avaliação do KS no modelo campeão.

2. Coleta de Dados

Para o desenvolvimento deste trabalho foi utilizada a plataforma Kaggle para obtenção do datasets que contém transações realizadas utilizando cartão de crédito em setembro de 2013 por portadores de cartões europeus.

Esse conjunto de dados apresenta transações que ocorrem em dois dias e contém 284.807 observações.

Nome da coluna	Descrição	Tipo
Time	Número de segundos decorridos entre esta transação e a primeira transação do conjunto de dados.	float64
V1-V28	É resultado de uma redução da dimensionalidade (PCA) para proteger as identidades dos usuários	float64

	e recursos sensíveis.	
Amount	É o valor da transação.	float64
Class	É a variáveis resposta e assume o valor 1 em caso de fraude e 0 em caso contrário.	int64

Neste trabalho foi utilizada a variável **Class** como variável resposta.

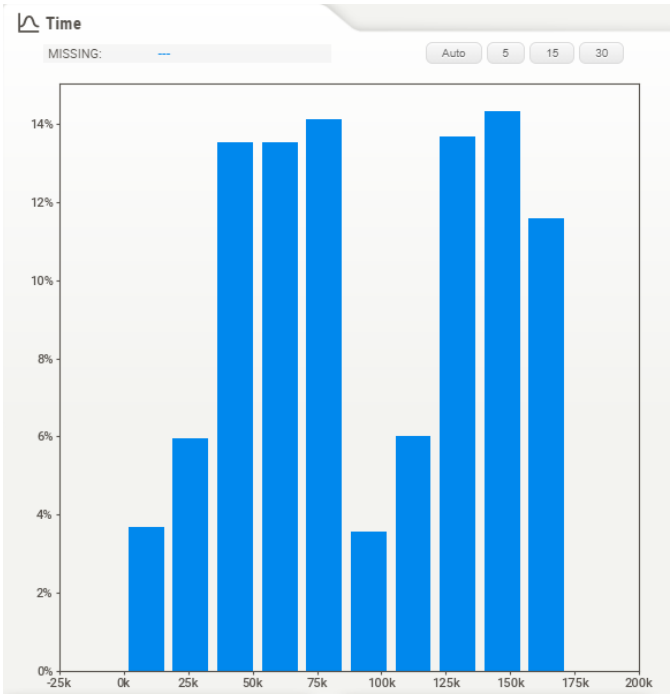
3. Análise e Exploração dos Dados

A análise exploratória foi obtida rodando a análise utilizando o pacote sweetviz.

A base de dados contém 284.807 observações, entre essas é possível observar 1.081 observações duplicadas (0,37%) e a base de dados contém 31 variáveis.

3.1. Variável Time

É possível observar a distribuição do tempo decorrido após a primeira transação da base de dados.



Métricas	Valor
Máximo	173K
95%	164K
3° quartil	139K
Média	95K
Mediana	85K
1° quartil	54K
5%	25K
Mínimo	0K

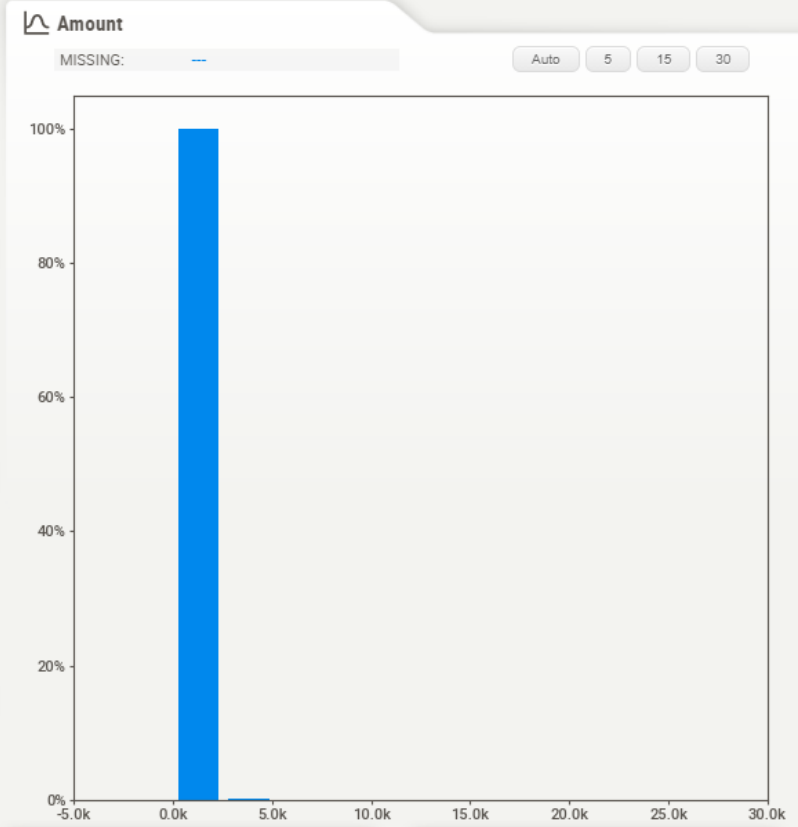
3.2. Variável V1-V28

A análise exploratória não se aplica as variáveis V1-V28, pois essas variáveis não apresentam uma interpretação, pois são resultado de uma transformação por componentes principais (PCA).

3.3. Variável Amount

É possível observar a distribuição dos valores das transações que contém na base de dados.

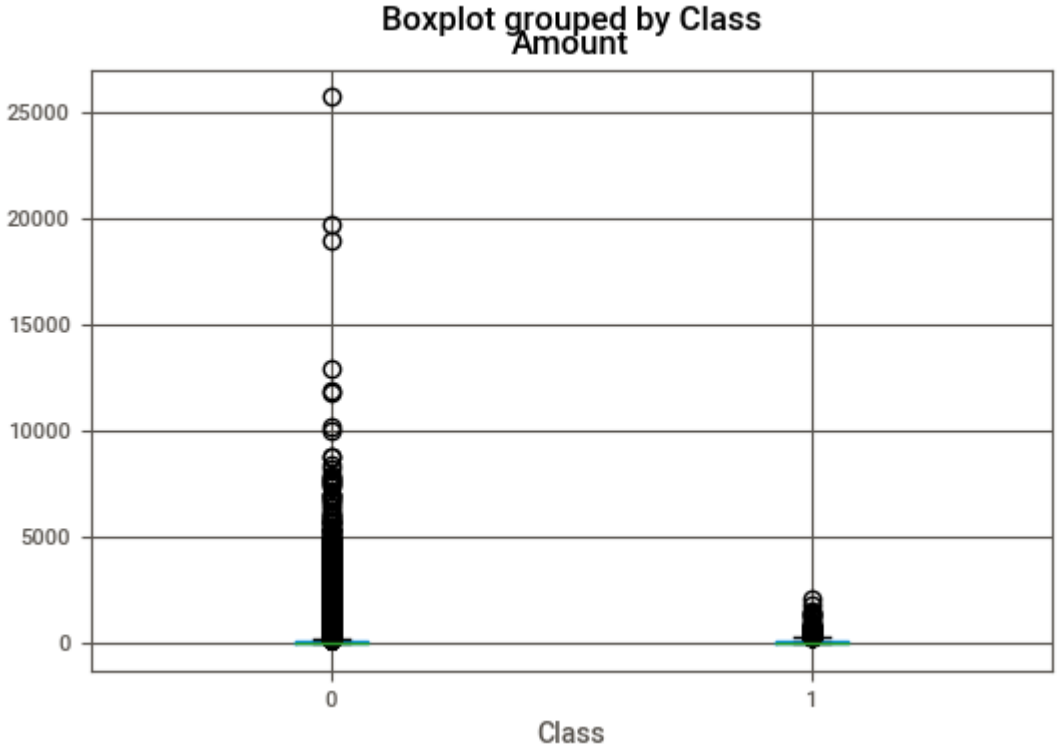
É possível notar que a maior partes dos valores estão entre 1 e 365.



Métricas	**Valor**
Máximo	25.691
95%	365
3° quartil	77
Média	88
Mediana	22
1° quartil	6
5%	1
Mínimo	0

Foi realizada também uma análise do valor das transações para cada uma das classes (0= não fraude, 1 = fraude).

Métricas	**Class = 0**	**Class = 1**
Máximo	25.691	2.125
3° quartil	77.07	105.89
Mediana	22.00	9.25
1° quartil	5	1
Mínimo	0	0
Média	88	122



Na análise não é possível fazer nenhuma inferência clara de relação entre a classificação da transação e o valor da transação. É possível notar que há uma diferença entre o valor média do valor das transações, comparando as classes.

3.4. Variável Class

Analisando a variável Class, que classifica a transação como fraudulenta e não fraudulenta, é possível observar que a base é desbalanceada.

Class	Quantidade de observações	% em relação ao total
0	284.315	>99%
1	492	<1%

4. Processamento/Tratamento de Dados

4.1. Remoção dos dados duplicados

Foram retiradas da análise 1.081 observações duplicadas que continham na base, que corresponde a 0,37% da base original.

4.2. Separação na base de treino e teste

A base original foi divida em dois datasets para realização do treinamento e teste da qualidade do modelo.

A divisão considerou 70% das observações para treinamento e 30% das observações para teste.

Nome do dataset	Descrição	Quantidade de observações
X_train, Y_train	É a base que vai ser utilizada para o treinamento do modelo.	198.608
X_test, Y_test	É a base que vai ser utilizada para o treinamento do modelo.	85.118

Nome do dataset	Class	Quantidade de observações
Y_train	1	335
	0	198.273
Y_test	1	138
	0	84.980

4.3. Balanceamento da base de treino

A análise exploratória da variável Class, indica que as observações estão bastante desbalanceadas e se o modelo fosse treinado sem nenhum tipo de balanceamento, o modelo aprenderia bem mais em relação às transações não fraudulentas.

Visando captar o final igualmente das duas classes é feito o balanceamento da base, utilizando técnicas de under sampling, ou seja, mantém a quantidade de observações da classe menos frequente (class=1) e amostra aleatoriamente a mesma quantidade de observação da classe mais frequente (class=0).

Nome do dataset	Class	Quantidade de observações
Y_train	1	335
	0	335

O balanceamento não é realizado na base de teste, porque é uma simulação da qualidade do modelo na realidade, onde a quantidade de transações não fraudulentas é muito superior à quantidade de transações fraudulentas.

5. Criação de Modelos de Machine Learning

Visando construir um modelo capaz de identificar transações fraudulentas, foram desenvolvidos os seguintes modelos: regressão logística, árvore de decisão, análise de discriminante e svm.

Para avaliar a qualidade do modelo foram ajustadas as seguintes métricas: roc, ks e acurácia.

A acurácia é usada apenas para fins de acompanhamento, como a base de teste é desbalanceada e a intenção do modelo é aprender a identificar as transações fraudulentas, não é a métrica ideal para classificar o melhor modelo.

5.1. Regressão logística

Métricas	Valores
ROC	0,75
KS	0,83
Acurácia	0,99

5.2. Árvore de decisão

Métricas	Valores
ROC	0,84
KS	0,78
Acurácia	0,99

5.3. Análise de discriminante

Métricas	Valores
ROC	0,90

KS	0,85
Acurácia	0,98

5.4. SVM

Métricas	**Valores**
ROC	0,79
KS	-
Acurácia	0,93

5.5. Seleção do melhor modelo

O melhor modelo foi escolhido considerando o que apresenta maior ROC e um KS superior à 0,25, logo, o modelo escolhido é o modelo de Análise de discriminante.

6. Apresentação dos Resultados

O modelo Análise de discriminante é o melhor modelo para detecção de fraude em cartão de crédito, considerando a base de treino e base de teste utilizada.

Avaliando a distribuição de probabilidade dos scores, é possível notar que o modelo começa a discriminar bem as transações fraudulentas a partir do 6° grupo.

	Faixa	%	Taxa de eventos positivos
0	(-0.001, 5.109999999999998e-149]	0.099762	0.000000
1	(5.109999999999998e-149, 1.3799999999999993e-148]	0.098674	0.000000
2	(1.3799999999999993e-148, 2.4199999999999988e-...	0.097972	0.000000
3	(2.4199999999999988e-148, 3.899999999999998e-148]	0.099528	0.000000
4	(3.899999999999998e-148, 6.189999999999997e-148]	0.100371	0.000000
5	(6.189999999999997e-148, 1.0899999999999994e-147]	0.102454	0.000000
6	(1.0899999999999994e-147, 2.5899999999999986e-...	0.100944	0.000348
7	(2.5899999999999986e-147, 1.0399999999999993e-...	0.100769	0.000232
8	(1.0399999999999993e-146, 1.8099999999999986e-...	0.098463	0.000832
9	(1.8099999999999986e-145, 1.0]	0.101062	0.017255

Analisando a matriz confusão é possível observar que:

- É fraude e o modelo classificou como fraude: 83.767 observações;
- É fraude e o modelo classificou como não fraude: 23 observações
- Não é fraude e o modelo classificou como fraude: 1.213 observações;
- É fraude e o modelo classificou como fraude: 115 observações.

Observando os resultados, a empresa pode adotar uma segunda verificação para os casos que o modelo classifica como fraude e algumas fraudes não são captadas pelo modelo (no teste representa 16% do total de fraudes).

Entre as transações fraudulentas, 89% são classificados como transações fraudulentas no modelo (sensibilidade) e entre as transações não fraudulentas, 99% são classificados como transações não fraudulentas no modelo (especificidade).

7. Links

Link para o vídeo: http://www.youtube.com/watch?v=t8qFwA2zx6o

Link para o repositório: https://github.com/jessiicafsilva/tcc_jessica.git

Link para fazer download dos dados: https://www.kaggle.com/mlg-ulb/creditcardfraud