EDA: Exploratory Data Analysis

1. Working Directory

Configurando o diretório de trabalho

setwd("C:/Users/Utilizador/repos/Formacao_cientista_de_dados/big_data_analytics_R_microsoft_azure_machine_lear
getwd()

2. Imports library(data.table) library("ggplot2") library(readr) library(plyr) library(caret) library(rpart) library(mboost) library(MASS) library(pamr) library(dplyr) library(naivebayes)

3. Data Loading dados <- fread("creditcard.csv", stringsAsFactors = F, sep = ",", header =T)

head(dados)

4. Data Cleaning

4.1 Variable Format Changes

Make the variable of interest (Fraud) a factor and re-level the variable for later interpretation in the models, this makes that the fraud level can be interpreted as Y=1

names(dados)[names(dados)=="Class"] <- "Fraud"

dados$Fraud <- -factor(dados$Fraud, labels = c("Normal", "Fraud")) dados$Fraud <- -relevel(dados$Fraud, "Fraud")

4.2 Missing values any(is.na(dados))

5.Descriptive statistics

Descreve, compreende, organiza e resumi os dados

5.1 Structure & simple statistics

DataViewer View(dados)

Object structure(str) str(dados)

summary summary(dados)

5.2 Central Trend Measures

Medidas de Tendência Central

resumo das variaveis do dataset

summary(dados$Fraud)summary(dados$Amount)

5.3 Dispersion Measures

Medidas de Dispersão

5.3.1 Variation

Na variância, números maiores indicam que os dados estão espalhados mais amplamente

em torno da média.

var(dados$Amount)

5.3.2 standard deviation

O desvio padrão indica, em média, a quantidade de cada valor diferente da média.

sd(dados$Amount)

6. Exploratory Data Analysis

6.1 Exploratory Data Analysis for Numerical Variables

Análise Exploratória de Dados Para Variáveis Numéricas

mean

mean(dados$Amount)

median

median(dados$Amount)

quartiles

quantile(dados$Amount) quantile(dados$Amount, probs = c(0.01, 0.99)) quantile(dados$Amount, seq(from = 0, to = 1, by = 0.20))

Difference between quartiles (Q3 and Q1)

IQR(dados$Amount) #Diferença entre Q3 e Q1

Minimum and maximum

range(dados$Amount)

Difference between ranges: Amplitude

diff(range(dados$Amount))

6.1.2 Plot

6.1.2.1 Univariate Analysis

Explora cada variável em um conjunto de dados, separadamente;

a) Boxplot

Diagrama de caixa (Bigode);

Leitura de Baixo para Cima - Q1, Q2 (Mediana) e Q3

valor mínimo, 1º quartil (25%), mediana(50%), 3º quartil(75%), valor máximo, outliers

boxplot(dados$Amount, main = "Boxplot para o Aumont", ylab = "Preço (R)")

b) Histogram

Distribuição de frequência dos valores dentro de cada bin (classe de valores)

Indicam a frequência de valores dentro de cada bin (classe de valores)

variavél amount

hist(dados$Amount, main = "Histograma Aumont", xlab = "Preço (R)")

variavel amount com bins

hist(dados$Amount, main = "Histograma Aumont", breaks = 5, ylab = "Preço (R)")

6.1.2.2 Bivariate Analysis

Explora como duas variáveis se comportam na presença uma da outra

a) Scatterplot

plot(x = dados$Amount, y = dados$Fraud, main = "Scatterplot - Aumont x Fraud", xlab = "Aumont", ylab = "Fraud")

6.2 Exploratory Data Analysis for Categorical Variables

Análise Exploratória de Dados Para Variáveis Categóricas

Criando tabelas de contingência - representam uma única variável categórica

Lista as categorias das variáveis nominais

table(dados$Fraud)

Calculando a proporção de cada categoria

model_table <- table(dados$Fraude) prop.table(model_table)

Arrendondando os valores

model_table <- table(dados$Fraud) model_table <- prop.table(model_table) * 100 round(model_table, digits = 1)

7. Visual Analysis

7.1 Fraud Distribution

This graph shows the high prevalence in the data, only 0.17% of the transactions in the data were fraudulent and the other 99.83% do not.

dados%>% group_by(Fraud)%>% summarise(n = n())%>% mutate(percentage = n/sum(n)*100)%>% ggplot(aes(Fraud, n, fill=Fraud))+ geom_bar(stat = "identity")+ geom_text(aes(label=n), vjust=-0.3, hjust = 1, size=4.5)+ geom_text(aes(label=paste0("(",round(percentage,2),"%)")), vjust=-0.3, hjust = -0.1, size=4.5)+ ylab("transactions")+ xlab(" ")

7.2 Fraud Distribution by Amount (scaled) dados%>% group_by(Fraud)%>% ggplot()+ geom_density(aes(Amount, fill = Fraud), alpha = .5)+ scale_x_log10()

7.3 Fraud Distribution by other variables

Here we can see that for some variables we can easily discriminate the fraudulent transactions

7.3.1 V2 dados%>% group_by(Fraud)%>% ggplot()+ geom_density(aes(V2, fill = Fraud), alpha = .5)

7.3.2 V3 dados%>% group_by(Fraud)%>% ggplot()+ geom_density(aes(V3, fill = Fraud), alpha = .5)

7.3.3 V4 dados%>% group_by(Fraud)%>% ggplot()+ geom_density(aes(V4, fill = Fraud), alpha = .5)

But for others this would be a little more difficult

7.3.4 V25 dados%>% group_by(Fraud)%>% ggplot()+ geom_density(aes(V25, fill = Fraud), alpha = .5)