Análise de Dados de Inadimplência

O grande mistério deste negócio está em descobrir quais clientes são bons pagadores e quais não são e, para isso, vamos usar a estatística como principal ferramenta.

Variáveis Disponíveis:

- ID
- Estado
- Setor da Empresa
- Faturamento Informado
- Dívida Total PJ
- Score de Crédito (0 – 1000), sendo 1000 o melhor cliente
- Taxa
- Atraso Corrente
- Prazo
- Valor do Contrato
- Valor do Contrato mais Juros
- Valor em Aberto

Conceitos de negócio:

> Conceito de ticket, taxa e prazo médios:
O ticket médio é o valor médio dos contratos, ou seja, o somatório do valor dos contratos dividido pelo total de contratos.
O mesmo vale para taxa e prazo, com 1 exceção: taxa e prazo médios devem ser ponderados pelo valor contrato!

> Conceito de BAD:
"Definimos determinado empréstimo como Bad quando este ultrapassa 180 dias de atraso.
Dessa forma, Bad = 1 define um mal pagador; Bad = 0 define um bom pagador."

> Conceito de Loss:
[valor em aberto dos mal pagadores] / [valor principal + juros totais]

Análise Descritiva dos Dados:

Software utilizado: R

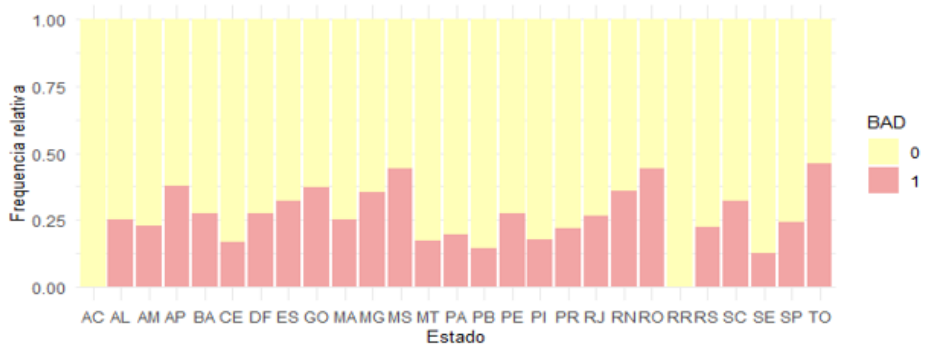
```
table(is.na(base))
table(is.na(base$setor))

#existem apenas 3 NAs na variável setor em toda a base.
base<- na.omit(base)
```

TICKET MÉDIO	TAXA MÉDIA *	PRAZO MÉDIO *
R$ 24.180,82	4,43%	11,82

*ponderados pelo valor do contrato.

```
# Ticket Medio
ticket_medio = mean(base$valor_contrato)

# Taxa Media Ponderada ponderado pelo valor do contrato
# criando nova variavel
soma_contratos = sum(base$valor_contrato)
base<- base %>% mutate(taxa_pond = (taxa*valor_contrato)/soma_contratos)
taxa_media_pond = sum(base$taxa_pond)
mean(base$taxa)

# Prazo Medio ponderado pelo valor do contrato
base<- base %>% mutate(prazo_pond = (prazo*valor_contrato)/soma_contratos)
prazo_medio_pond = sum(base$prazo_pond)
```

Analisando a relação entre Estado e BAD:

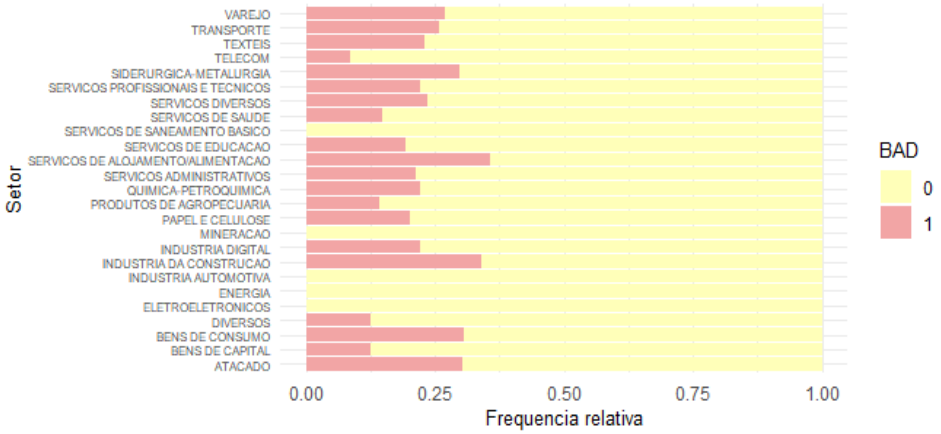


TESTE QUI-QUADRADO DE PEARSON: DEPENDÊNCIA ENTRE AS VARIÁVEIS, considerando um nível de significância de 10%.

```
# Criando a variável BAD
base<- base %>% mutate(BAD = ifelse(atraso_corrente>180,1,0))
base$BAD<- as.factor(base$BAD)

graf0 = base %>%
  ggplot(aes(x=estado)) + geom_bar(aes(fill=BAD), position = "fill") +
  labs(title="", y="Frequencia relativa",x="Estado") +
  theme_minimal()+
  scale_fill_manual(values=c("#FFFFB9","#F2A5A5")) +
  theme(title = element_text(size=10))
```

```
## Analisando a independência entre a variável BAD e outras
#Ho: Não há associação entre as variáveis
#H1: Há associação entre as variáveis
# Obs.: considerado um nivel de sign. de 10%

#Teste Qui-Quadrado de Independencia
tabela0<- table(base$estado,base$BAD)
chisq.test(tabela0)
#p-valor: 0.03767
#rejeitamos a hipotese. Logo podemos dizer que as variáveis estão associadas.
```

Analisando a relação entre o Setor e a BAD:



TESTE QUI-QUADRADO DE PEARSON: DEPENDÊNCIA ENTRE AS VARIÁVEIS, considerando um nível de significância de 10%.

Análise de correlação entre as variáveis:

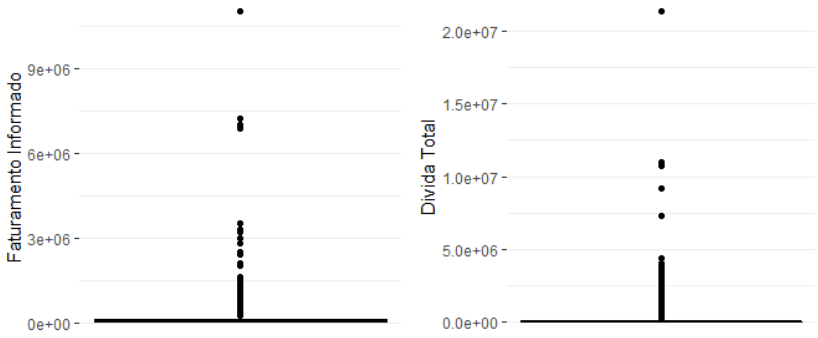
Não foi observada nenhuma correlação significativa, além do que já era esperado.

```
## Calcula a correlacao
base_q<- base %>% select(-c("id","estado","setor","taxa_pond","prazo_pond",
                           "atraso_corrente"))
base_q$BAD<- as.numeric(base_q$BAD)
correlacao <- cor(base_q, method="pearson")


## Gráfico com essas correlações
library(corrplot)
corrplot(correlacao, numbers=T, diag=F,type= "lower",tl.col = "darkblue",tl.cex=0.7,tl.srt=35)
```

Criação de um modelo de classificação (considerando a variável de interesse BAD) a fim de verificar quais são as características de um mau pagador:

AJUSTES A SEREM FEITOS:

- **OUTLIERS**



- **DESBALANCEAMENTO DA AMOSTRA**

Devido ao desbalanceamento da amostra, foi realizado um processo de oversampling a fim de igualar a proporção das classes da variável resposta, resultando então em um modelo mais confiável.

BAD	Frequência absoluta	Frequência relativa (%)
0	1,416	73.98
1	498	26.02

BAD	Frequência absoluta	Frequência relativa (%)
0	1,416	50
1	1,416	50

Por fim, foi realizada a padronização das variáveis e também foram avaliadas e retiradas, se existentes, as variáveis altamente correlacionadas e com variâncias próximas a zero.

- VARIÁVEIS ALTAMENTE CORRELACIONADAS:

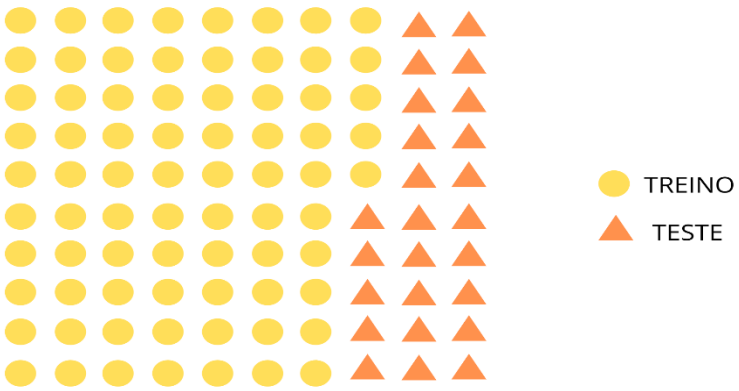
- Valor do Contrato

- VARIÁVEIS COM VARIÂNCIA PRÓXIMAS A ZERO

- Nenhuma

```
basemod<- base %>%
  select(-c("id","taxa_pond","prazo_pond","valor_contrato","atraso_corrente")) %>%
  filter(faturamento_informado < 6e+06 & divida_total_pj < 5.0e+06)

descrcor<- cor(base_q,method="pearson")

#PRE-PROCESSAMENTO
# todas variaveis com variancias proximas a zero
findCorrelation(descrcor,cutoff = 0.75,verbose=F,names=T)

#variaveis nao sao autocorrelacionadas
prop.table(table(basemod$BAD))
#variavel resposta desproporcional
# 74% de BAD = 0 (bom pagador)
# se nao balancear a amostra, podemos acabar criando um modelo
# ruim e classificando maus pagadores como bons.
# -> utilizar tecnica de rebalanceamento

# Separando aleatoriamente os dados treino (80%) e teste (20%)
library(caret)
library(ROSE)

#aumentando a amostra copiando dados existentes na base de forma aleatoria
over<- ovun.sample(BAD~.,data=basemod,method = "over",N=1416*2 )$data #N -> tamanho total da amostra
table(over$BAD)


set.seed(600)
inTrain<- createDataPartition(y=over$BAD,p=0.75,list=F)
training<- over[inTrain,]
testing<- over[-inTrain,]


modelo_over0<- train(BAD ~ ., data=training, method="glm",
                     preProcess=c("center","scale")) #prob de ser um mal pagador
summary(modelo_over0)


modelo_over1<- train(BAD ~ score + divida_total_pj + valor_contrato_mais_juros + valor_em_aberto,
                     data=training, method="glm",
                     preProcess=c("center","scale"))

summary(modelo_over1)

prediction_treino<- predict(modelo_over1,newdata = training)
```

Para o processo de criação do modelo, a amostra foi dividida em 2 partes de forma aleatória, uma delas contendo 75% dos dados, que representa a amostra de treino, utilizada para a construção do modelo e, a outra, contendo 25% dos dados, que representa a amostra de teste. Esta última tem como objetivo proporcionar a análise da capacidade preditiva do modelo.



A partir da amostra treino foi construído um modelo para inadimplência (se o cliente é mau pagador) com toda a base a fim de verificar quais variáveis são significativas para o modelo, ou seja, quais explicam as características de um mal e, consequentemente, de um pagador.

Variáveis Significativas:

- Dívida Total
- Score
- Valor do Contrato mais juros
- Valor em Aberto

Em seguida, foi construído um novo modelo apenas com as variáveis acima e, novamente, todas foram significativas.

Variável	Coeficiente Estimado
Intercepto	3.58108
Score	-0.27202
Dívida Total	-1.09393
Valor Contrato mais Juros	-1.48013
Valor em Aberto	11.68834

A partir desta tabela, observa-se que o valor em aberto e o valor do contrato mais juros são as variáveis que mais influenciam na inadimplência, sendo que o valor em aberto aumenta a chance do cliente ser um mau pagador, enquanto o valor do contrato mais juros diminui essa chance.

Então, foi desenhada a matriz de confusão a fim de analisar a capacidade do modelo.

```
confusionMatrix(prediction_teste,testing$BAD, positive='0')
confusionMatrix(prediction_treino,training$BAD, positive='0')
```

	0	1
0	Verdadeiro Positivo **346** (97.7%)	Falso Positivo **34** (9.6%)
1	Falso Negativo **8** (2.2%)	Verdadeiro Negativo **320** (90.4%)

- ACURÁCIA: **94.07%**
- SENSIBILIDADE: **97.74 %** (dar BAD=0 para quem realmente é bom pagador)
- ESPECIFICIDADE: **90.40%** (dar BAD=1 sendo que de fato é inadimplente)

Através das medidas observadas na matriz de confusão, podemos dizer que o modelo encontrado neste projeto acertará em 90.4% que o cliente é um mau pagador quando ele realmente é mau pagador. E em 9.6% dirá que ele é bom pagador, sendo que ele é mau pagador. Além disso, em 97.8% das vezes o modelo acerta quem é bom pagador.

OBS.: os códigos estão em uma versão resumida, ideal apenas para uma abordagem inicial. Qualquer dúvida ou sugestão é bem-vinda.