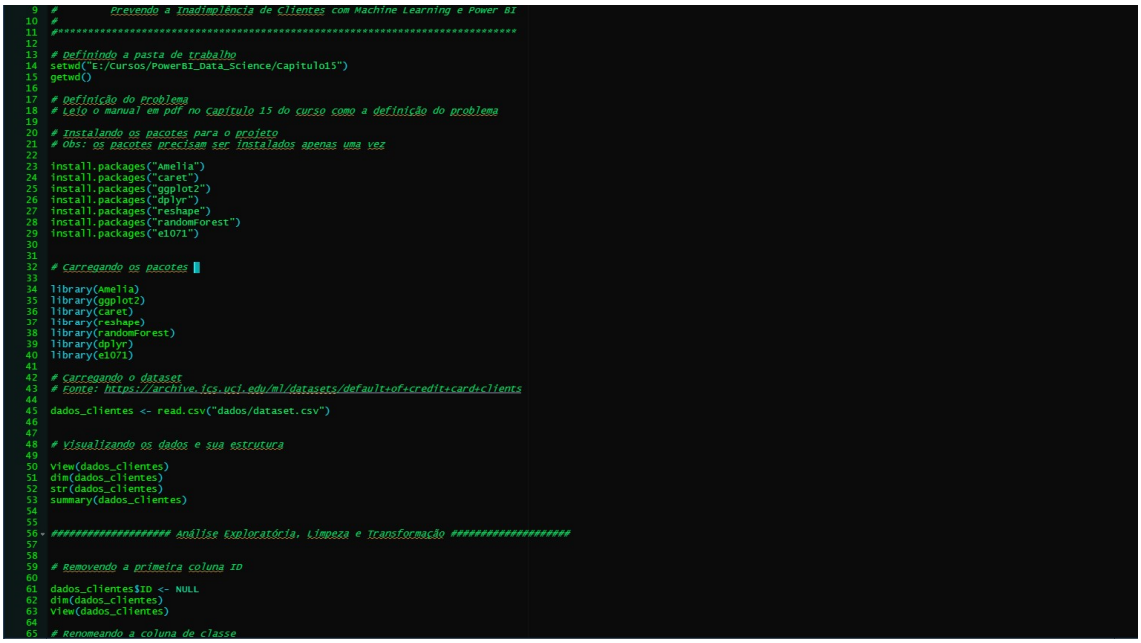
Previsão de Inadimplência com Machine Learning



```r
#         Prevendo a Inadimplência de Clientes com Machine Learning e Power BI
#
#*****************************************************************************

# Definindo a pasta de trabalho
setwd("E:/Cursos/PowerBI_Data_Science/Capitulo15")
getwd()

# Definição do Problema
# Leio o manual em pdf no Capitulo 15 do curso como a definição do problema

# Instalando os pacotes para o projeto
# Obs: os pacotes precisam ser instalados apenas uma vez

install.packages("Amelia")
install.packages("caret")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("reshape")
install.packages("randomForest")
install.packages("e1071")


# Carregando os pacotes

library(Amelia)
library(ggplot2)
library(caret)
library(reshape)
library(randomForest)
library(dplyr)
library(e1071)

# Carregando o dataset
# Fonte: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

dados_clientes <- read.csv("dados/dataset.csv")


# Visualizando os dados e sua estrutura

View(dados_clientes)
dim(dados_clientes)
str(dados_clientes)
summary(dados_clientes)


################### Análise Exploratória, Limpeza e Transformação ###################


# Removendo a primeira coluna ID

dados_clientes$ID <- NULL
dim(dados_clientes)
View(dados_clientes)

# Renomeando a coluna de classe
```

Visão Geral

Este script foi desenvolvido como parte do curso "Microsoft Power BI Para Data Science, Versão 2.0" da Data Science Academy.

O objetivo é prever a inadimplência de clientes utilizando técnicas de Machine Learning integradas ao Power BI.

Dependências

O script requer os seguintes pacotes R:

```R
install.packages(c("Amelia", "caret", "ggplot2", "dplyr", "reshape", "randomForest", "e1071", "devtools"))

devtools::install_github('cran/DMwR')
```

Fluxo de Trabalho

1. Pré-processamento dos Dados

-Carregamento: Dados de crédito de clientes (fonte: UCI Machine Learning Repository)

-Limpeza:

- Remoção de coluna ID
- Renomeação de colunas para melhor legibilidade
- Tratamento de valores ausentes

-Transformação:

- Conversão de variáveis categóricas (gênero, escolaridade, estado civil)
- Discretização de idade em faixas etárias
- Conversão de variáveis de pagamento para fatores

2. Análise Exploratória

- Visualização da distribuição de inadimplentes vs. não-inadimplentes
- Análise de proporções entre classes
- Geração de gráficos exploratórios com ggplot2

3. Modelagem Preditiva

-Divisão dos Dados: 75% treino / 25% teste (amostragem estratificada)

-Versões do Modelo:

1.Modelo Inicial: Random Forest com todas as variáveis

2.Modelo Balanceado: Aplicação de SMOTE para tratar desbalanceamento de classes

3.Modelo Otimizado: Utilizando apenas variáveis mais importantes

-Métricas de Avaliação:

- Matriz de confusão
- Precision, Recall e F1-Score

4. Implementação Final

- Seleção do melhor modelo (versão 3)

- Serialização do modelo para uso em produção

- Exemplo de aplicação com novos dados de clientes

Estrutura de Arquivos

```
/projeto
  /dados
    dataset.csv         Dados brutos dos clientes
  /modelo
    modelo_v3.rds       Modelo serializado
  script.R              Este script de análise
```

Como Executar

1. Definir diretório de trabalho (linha 8)

2. Instalar pacotes necessários (apenas na primeira execução)

3. Executar o script sequencialmente

Resultados Principais

- Modelo final alcançou F1-Score de [valor a ser preenchido após execução]

- Variáveis mais importantes identificadas:

1. PAY_0 (status de pagamento mais recente)
2. PAY_2
3. PAY_3
4. PAY_AMT1 (valor do pagamento mais recente)

Integração com Power BI

O modelo serializado (modelo_v3.rds) pode ser carregado no Power BI para:

- Criação de dashboards interativos
- Previsões em tempo real
- Análise de risco de crédito

Observações

- Para reproduzir exatamente os mesmos resultados, manter o set.seed(12345)
- O dataset original contém informações sensíveis - tratar com confidencialidade
- Versão dos dados utilizada: 17/04/2021

Referências

- Documentação original do curso: Capítulo 15
- Dataset: Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications.