Terceira Atividade - Parte complementar

Rennan Guimarães

24/09/2024

Contents

Introdução:

O objetivo desse trabalho utilizar as técnicas de regressão, agrupamento e mineração de regas.

Caso queira acessar o código fonte ou a versão em HTML, acesse o repositório da atividade e busque pelos arquivos nomeados “final_plus”.

Pré-processamento e Configurações Iniciais

Dividindo os Dados em Treino e Teste

```
set.seed(17)
data_split <- initial_split(german_data, prop = 0.8, strata = "Good_loan")
train_data <- training(data_split)
test_data <- testing(data_split)
```

Análise de Correlação e Seleção de Variáveis

```
numeric_vars <- train_data |>
  select(where(is.numeric))

correlation_matrix <- cor(numeric_vars, use = "complete.obs")

high_cor_vars <- findCorrelation(correlation_matrix, cutoff = 0.9, names = TRUE)
train_data <- train_data |> select(-all_of(high_cor_vars))
test_data <- test_data |> select(-all_of(high_cor_vars))
```

Definindo os Folds para Validação Cruzada

```
set.seed(17)
folds <- vfold_cv(train_data, v = 10, strata = "Good_loan")
```

Receitas de Pré-processamento

```
rec <- recipe(Good_loan ~ ., data = train_data) |>
  # Imputação
  step_impute_mode(all_nominal_predictors()) |>
  step_impute_median(all_numeric_predictors()) |>
  # Tratamento de outliers usando winsorization
  step_mutate_at(all_numeric_predictors(), fn = ~scales::squish(.x, quantile(.x, c(0.01, 0.99), na.rm =
  # Codificação
  step_dummy(all_nominal_predictors(), one_hot = TRUE) |>
  # Remover variáveis com variância zero
  step_zv(all_predictors()) |>
  # Tratamento de desbalanceamento
  step_smote(Good_loan)
```

```
rec_normalized <- recipe(Good_loan ~ ., data = train_data) |>
  # Imputação
  step_impute_mode(all_nominal_predictors()) |>
  step_impute_median(all_numeric_predictors()) |>
  # Tratamento de outliers usando winsorization
  step_mutate_at(all_numeric_predictors(), fn = ~scales::squish(.x, quantile(.x, c(0.01, 0.99), na.rm =
  # Codificação
  step_dummy(all_nominal_predictors(), one_hot = TRUE) |>
  # Remover variáveis com variância zero
  step_zv(all_predictors()) |>
  # Normalização
  step_normalize(all_numeric_predictors()) |>
  # Tratamento de desbalanceamento
  step_smote(Good_loan)
```

Função para Coletar Métricas de Resample

```
collect_resample_metrics <- function(tune_results, best_params, model_name) {
  tune_results |>
    collect_metrics(summarize = FALSE) |>
    inner_join(best_params, by = names(best_params)) |>
    mutate(Model = model_name)
}
```

Técnicas

Técnica 1: Regressão Logística com Dados Não Normalizados

```
log_reg_model <- logistic_reg() |>
  set_engine("glm") |>
  set_mode("classification")

workflow_log_reg <- workflow() |>
  add_model(log_reg_model) |>
  add_recipe(rec)

set.seed(17)
log_reg_fit <- workflow_log_reg |>
  fit_resamples(
    resamples = folds,
    metrics = metric_set(
      roc_auc, accuracy,
      precision = yardstick::precision,
      recall = yardstick::recall,
      f_meas = f_meas
    ),
    control = control_resamples(save_pred = TRUE)
  )

log_reg_resample_metrics <- log_reg_fit |>
  collect_metrics(summarize = FALSE) |>
  mutate(Model = "Regressão Logística")
```

Técnica 1b: Regressão Logística com Dados Normalizados

```
log_reg_model_norm <- logistic_reg() |>
  set_engine("glm") |>
  set_mode("classification")

workflow_log_reg_norm <- workflow() |>
```

```
  add_model(log_reg_model_norm) |>
  add_recipe(rec_normalized)

set.seed(17)
log_reg_fit_norm <- workflow_log_reg_norm |>
  fit_resamples(
    resamples = folds,
    metrics = metric_set(
      roc_auc, accuracy,
      precision = yardstick::precision,
      recall = yardstick::recall,
      f_meas = f_meas
    ),
    control = control_resamples(save_pred = TRUE)
  )

log_reg_resample_metrics_norm <- log_reg_fit_norm |>
  collect_metrics(summarize = FALSE) |>
  mutate(Model = "Regressão Logística (Normalizada)")
```

Avaliando Resultados do Treinamento

Vamos comparar as métricas de desempenho dos dois modelos no conjunto de treinamento (utilizando validação cruzada):

```
resample_metrics_combined <- bind_rows(
  log_reg_resample_metrics,
  log_reg_resample_metrics_norm
)

metrics_summary <- resample_metrics_combined |>
  group_by(Model, .metric) |>
  summarize(
    mean = mean(.estimate),
    variance = var(.estimate),
    sd = sd(.estimate),
    .groups = 'drop'
  ) |>
  mutate(
    formatted = glue("{round(mean, 4)} ± {round(sd, 4)}")
  )

metrics_table <- metrics_summary |>
  select(Model, .metric, formatted) |>
  pivot_wider(names_from = .metric, values_from = formatted)

print(metrics_table)
```

```
## # A tibble: 2 x 6
##   Model                           accuracy     f_meas precision recall roc_auc
##   <chr>                           <glue>       <glue> <glue>    <glue> <glue>
## 1 Regressão Logística             0.6912 ± 0.~ 0.572~ 0.4925 ±~ 0.687~ 0.7409~
## 2 Regressão Logística (Normalizada) 0.68 ± 0.05~ 0.561~ 0.4795 ±~ 0.679~ 0.7408~
```

Conclusão

Apesar dos resultados da regressão logística sem normalização terem sido maiores, a variancia também é maior o que pode prejudicar na consistencia da técnica no ambiente de produção, principalmente pelo resultado da acuracia e do roc terem uma variancia maior e um ganho pequeno de resultado, mas vamos analisar no conjunto de teste.

Ajuste Final e Avaliação no Conjunto de Teste

```
log_reg_last_fit <- last_fit(
  workflow_log_reg,
  split = data_split,
  metrics = metric_set(
    roc_auc, accuracy,
    precision = yardstick::precision,
    recall = yardstick::recall,
    f_meas = f_meas
  )
)

log_reg_last_fit_norm <- last_fit(
  workflow_log_reg_norm,
  split = data_split,
  metrics = metric_set(
    roc_auc, accuracy,
    precision = yardstick::precision,
    recall = yardstick::recall,
    f_meas = f_meas
  )
)

log_reg_metrics <- collect_metrics(log_reg_last_fit) |>
  mutate(Model = "Regressão Logística")

log_reg_metrics_norm <- collect_metrics(log_reg_last_fit_norm) |>
  mutate(Model = "Regressão Logística (Normalizada)")

test_metrics_combined <- bind_rows(log_reg_metrics, log_reg_metrics_norm) |>
  select(Model, .metric, .estimate)

test_metrics_table <- test_metrics_combined |>
  pivot_wider(names_from = .metric, values_from = .estimate)
```

Visualização da Curva ROC e Matriz de Confusão

```
log_reg_predictions <- collect_predictions(log_reg_last_fit)
log_reg_conf_mat <- log_reg_predictions |>
  conf_mat(truth = Good_loan, estimate = .pred_class)
log_reg_roc_curve <- log_reg_predictions |>
  roc_curve(truth = Good_loan, .pred_yes)
```

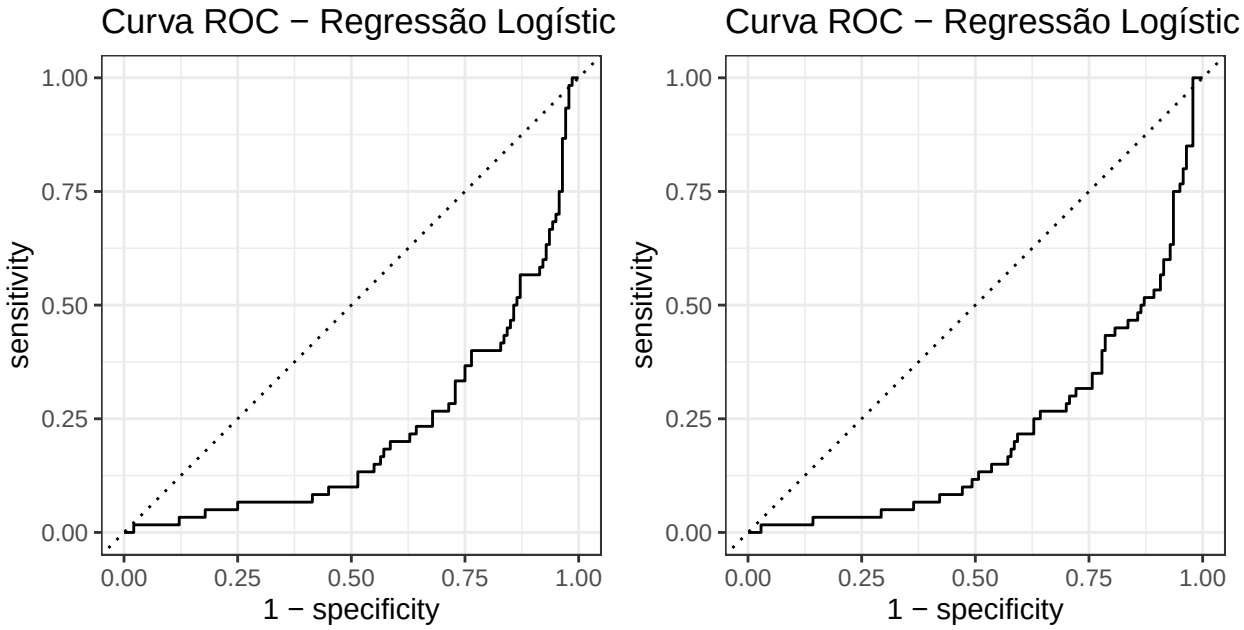
```r
log_reg_roc_plot <- autoplot(log_reg_roc_curve) +
  ggtitle("Curva ROC - Regressão Logística")

log_reg_predictions_norm <- collect_predictions(log_reg_last_fit_norm)
log_reg_conf_mat_norm <- log_reg_predictions_norm |>
  conf_mat(truth = Good_loan, estimate = .pred_class)
log_reg_roc_curve_norm <- log_reg_predictions_norm |>
  roc_curve(truth = Good_loan, .pred_yes)
log_reg_roc_plot_norm <- autoplot(log_reg_roc_curve_norm) +
  ggtitle("Curva ROC - Regressão Logística (Normalizada)")

print(test_metrics_table)
```

```
## # A tibble: 2 x 6
##   Model                              accuracy precision recall f_meas roc_auc
##   <chr>                                 <dbl>     <dbl>  <dbl>  <dbl>   <dbl>
## 1 Regressão Logística                    0.7      0.5    0.733  0.595   0.779
## 2 Regressão Logística (Normalizada)      0.67     0.468  0.733  0.571   0.780
```

```r
grid.arrange(log_reg_roc_plot, log_reg_roc_plot_norm, ncol = 2)
```



Conclusões

1. Desempenho Geral: Ambos os modelos têm um desempenho razoável, com ROC AUC em torno de 0.78, indicando uma capacidade discriminativa boa, mas não excelente.

2. Impacto da Normalização: A normalização dos dados não resultou em uma melhoria significativa do modelo. De fato, em algumas métricas (acurácia, precisão, F-measure), o modelo não normalizado teve um desempenho ligeiramente superior.
3. Trade-off Precisão-Recall: Ambos os modelos mostram um recall relativamente alto (0.7333) em comparação com a precisão (cerca de 0.5), sugerindo que são mais eficazes em identificar casos positivos, mas à custa de alguns falsos positivos.
4. Escolha do Modelo: Dado que o desempenho é muito similar, a escolha entre os dois modelos pode depender de outros fatores, como interpretabilidade ou eficiência computacional. O modelo não normalizado pode ser preferível por sua simplicidade e ligeira vantagem em algumas métricas, entretanto, comparado com os modelos apresentados na atividade principal eu não optaria por esse modelo de regressão.

Técnica 2: Agrupamento (Clustering)

Preparação dos Dados para Agrupamento

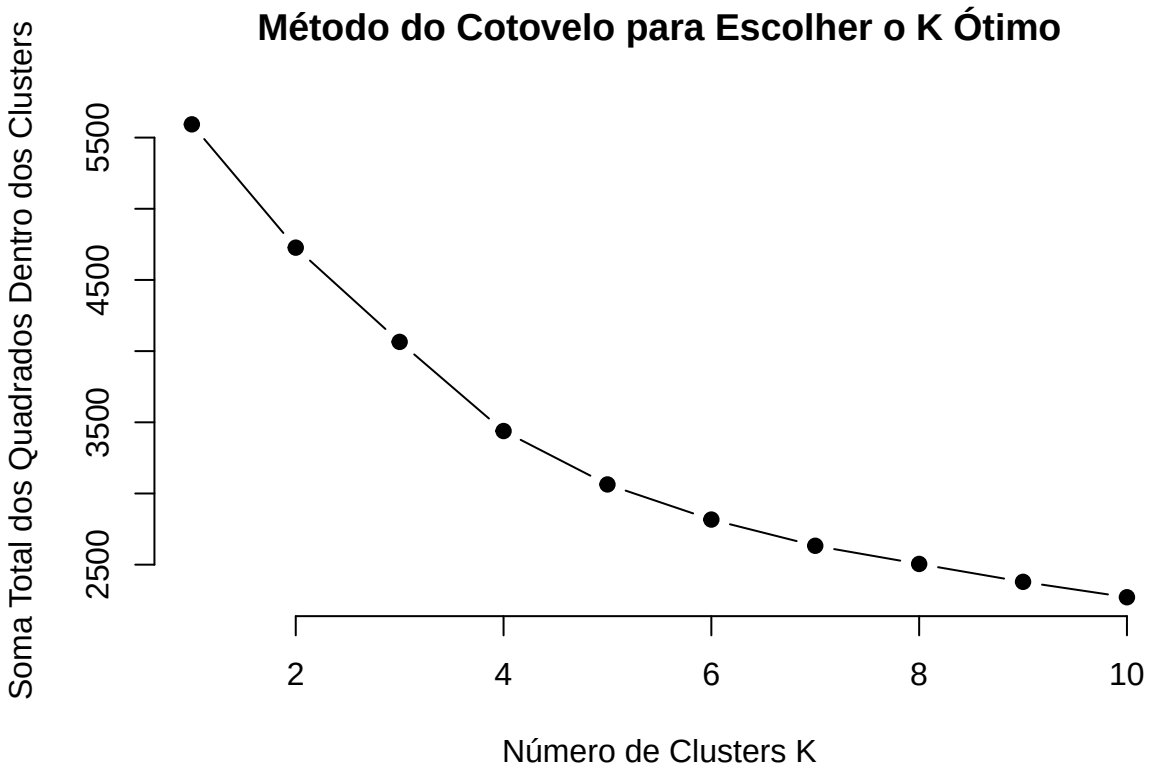
```
numeric_vars <- train_data |>
  select(where(is.numeric))

scaled_numeric_vars <- scale(numeric_vars)
```

Determinando o Número Ótimo de Clusters

```
set.seed(17)
wss <- sapply(1:10, function(k){
  kmeans(scaled_numeric_vars, centers = k, nstart = 10)$tot.withinss
})

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Número de Clusters K",
     ylab = "Soma Total dos Quadrados Dentro dos Clusters",
     main = "Método do Cotovelo para Escolher o K Ótimo")
```



Analisando o gráfico, optamos por K = 4.

Aplicando o K-Means com K Ótimo

```
set.seed(17)
kclust <- kmeans(scaled_numeric_vars, centers = 4, nstart = 25)

train_data_clust <- train_data |>
  mutate(Cluster = factor(kclust$cluster))
```

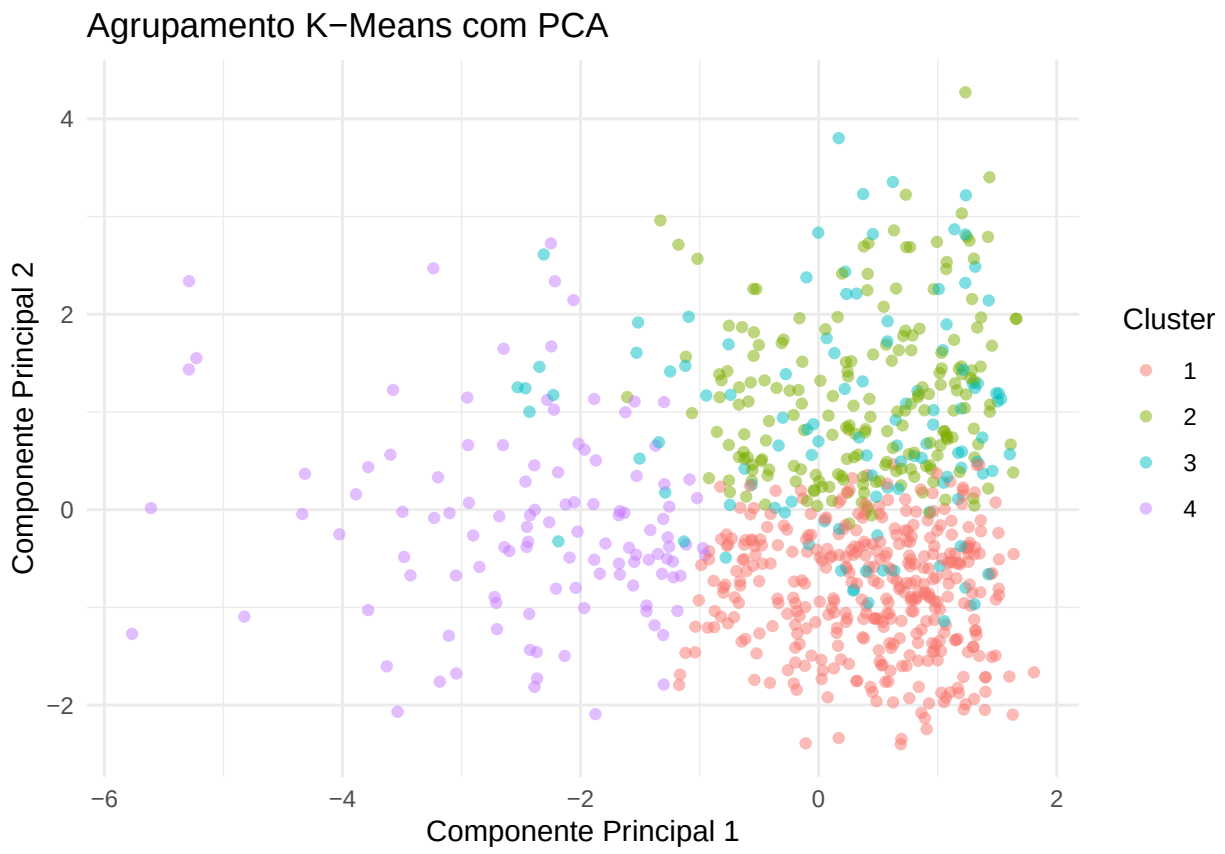
Visualização dos Clusters

Utilizamos PCA para reduzir a dimensionalidade e visualizar os clusters.

```
pca_res <- prcomp(scaled_numeric_vars)

pca_df <- data.frame(PC1 = pca_res$x[,1],
                     PC2 = pca_res$x[,2],
                     Cluster = factor(kclust$cluster))

ggplot(pca_df, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(alpha = 0.5) +
  labs(title = "Agrupamento K-Means com PCA", x = "Componente Principal 1", y = "Componente Principal 2
  theme_minimal()
```



Análise dos Clusters

Podemos analisar as características de cada cluster para entender melhor os grupos identificados.

```
# Resumo estatístico por cluster
cluster_summary <- train_data_clust |>
  group_by(Cluster) |>
  summarise(across(where(is.numeric), list(mean = mean, sd = sd), .names = "{col}_{fn}"))

print(cluster_summary)
```

```
## # A tibble: 4 x 15
##   Cluster Duration_in_month_mean Duration_in_month_sd Credit_amount_mean
##   <fct>                   <dbl>                <dbl>              <dbl>
## 1 1                        17.3                 7.61              2247.
## 2 2                        17.8                 8.56              2326.
## 3 3                        18.7                11.5               3047.
## 4 4                        39.1                12.1               8176.
## # i 11 more variables: Credit_amount_sd <dbl>,
## #   Installment_rate_of_disposable_income_mean <dbl>,
## #   Installment_rate_of_disposable_income_sd <dbl>,
## #   Present_residence_since_mean <dbl>, Present_residence_since_sd <dbl>,
## #   Age_in_years_mean <dbl>, Age_in_years_sd <dbl>,
## #   Number_of_existing_credits_at_this_bank_mean <dbl>,
## #   Number_of_existing_credits_at_this_bank_sd <dbl>, ...
```

Conclusão:

Características dos Clusters

Cluster 1: Clientes de Baixo Risco

- **Duração média do empréstimo:** 17,27 meses
- **Valor médio do crédito:** 2247,46
- **Idade média:** 28,42 anos
- **Taxa de parcela/renda:** 2,95% (mais baixa)
- **Perfil:** Clientes jovens com empréstimos menores e de curta duração

Cluster 2: Clientes de Risco Médio-Baixo

- **Duração média do empréstimo:** 17,82 meses
- **Valor médio do crédito:** 2325,81
- **Idade média:** 47,47 anos
- **Residência atual:** Mais longa (3,66 anos)
- **Perfil:** Clientes mais velhos, estáveis, com empréstimos moderados

Cluster 3: Clientes de Risco Médio

- **Duração média do empréstimo:** 18,73 meses
- **Valor médio do crédito:** 3047,11
- **Idade média:** 38,87 anos
- **Característica:** Maior variabilidade na duração do empréstimo
- **Perfil:** Clientes de meia-idade com empréstimos de valor intermediário

Cluster 4: Clientes de Alto Risco/VIP

- **Duração média do empréstimo:** 39,07 meses
- **Valor médio do crédito:** 8176,12
- **Idade média:** 35,37 anos
- **Pessoas responsáveis:** Maior número (média de 2)
- **Perfil:** Clientes mais jovens com empréstimos maiores e de longa duração

Insights Principais

1. **Segmentação Clara:** Há uma distinção nítida entre os clusters baseada em idade, duração do empréstimo e valor do crédito.
2. **Cluster Diferenciado:** O Cluster 4 se destaca significativamente, possivelmente representando clientes de alto valor ou alto risco.
3. **Base de Clientes:** Os Clusters 1 e 2 parecem representar a maioria dos clientes, com perfis mais conservadores.
4. **Histórico de Crédito:** O número de créditos existentes no banco é relativamente similar entre os clusters, com o Cluster 2 apresentando uma média ligeiramente superior.

Modelo 3: Mineração de Regras de Associação

```
numeric_vars <- names(select(german_data, where(is.numeric)))

german_data_discrete <- german_data |>
  mutate(across(all_of(numeric_vars), ~ arules::discretize(.x, method = "frequency", categories = 5)))
  mutate(across(everything(), as.factor))
```

```
## Warning: There were 11 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning in `arules::discretize()`:
## ! Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!
## i Run `dplyr::last_dplyr_warnings()` to see the 10 remaining warnings.
```

```
transactions <- as(german_data_discrete, "transactions")

min_support <- 0.01
min_confidence <- 0.5

rules <- apriori(
  transactions,
  parameter = list(supp = min_support, conf = min_confidence, minlen = 2),
  control = list(verbose = FALSE)
)

top_20_rules <- head(sort(rules, by = "lift"), 20)

print_rules <- function(rules) {
  for (i in 1:length(rules)) {
    rule <- rules[i]
    cat(paste0("\nRegra ", i, ":\n"))
    cat("SE ", paste(labels(lhs(rule)), collapse = " E "), "\n")
    cat("ENTÃO ", paste(labels(rhs(rule)), collapse = " E "), "\n")
    cat("Suporte: ", quality(rule)$support, "\n")
    cat("Confiança: ", quality(rule)$confidence, "\n")
    cat("Lift: ", quality(rule)$lift, "\n\n")
  }
}

print_rules(top_20_rules)
```

```
##
## Regra 1:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.011
## Confiança:  0.7857143
## Lift:  15.10989
##
##
## Regra 2:
```

```
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.011
## Confiança:  0.7857143
## Lift:  15.10989
##
##
## Regra 3:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.011
## Confiança:  0.7857143
## Lift:  15.10989
##
##
## Regra 4:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.011
## Confiança:  0.7857143
## Lift:  15.10989
##
##
## Regra 5:
## SE  {Duration_in_month=[4,12),Purpose=car (new),Sex=male,Present_residence_since=[2,4],Property=real
## ENTÃO  {Foreign_worker=no}
## Suporte:  0.01
## Confiança:  0.5555556
## Lift:  15.01502
##
##
## Regra 6:
## SE  {Duration_in_month=[4,12),Purpose=car (new),Sex=male,Present_residence_since=[2,4],Property=real
## ENTÃO  {Foreign_worker=no}
## Suporte:  0.01
## Confiança:  0.5555556
## Lift:  15.01502
##
##
## Regra 7:
## SE  {Duration_in_month=[4,12),Purpose=car (new),Sex=male,Present_residence_since=[2,4],Property=real
## ENTÃO  {Foreign_worker=no}
## Suporte:  0.01
## Confiança:  0.5555556
## Lift:  15.01502
##
##
## Regra 8:
## SE  {Duration_in_month=[4,12),Purpose=car (new),Sex=male,Present_residence_since=[2,4],Property=real
## ENTÃO  {Foreign_worker=no}
## Suporte:  0.01
## Confiança:  0.5555556
## Lift:  15.01502
##
```

```
##
## Regra 9:
## SE  {Present_residence_since=[2,4],Other_installment_plans=none,Job=unemployed/unskilled - non-reside
## ENTÃO  {Present_employment_since=unemployed}
## Suporte:  0.012
## Confiança:  0.9230769
## Lift:  14.88834
##
##
## Regra 10:
## SE  {Present_residence_since=[2,4],Other_installment_plans=none,Job=unemployed/unskilled - non-reside
## ENTÃO  {Present_employment_since=unemployed}
## Suporte:  0.012
## Confiança:  0.9230769
## Lift:  14.88834
##
##
## Regra 11:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 12:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 13:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 14:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 15:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
```

```
## Lift:  14.7929
##
##
## Regra 16:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 17:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 18:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 19:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
##
##
## Regra 20:
## SE  {Status_of_existing_checking_account=0 <= ... < 200 DM,Credit_history=existing credits paid back
## ENTÃO  {Other_debtors_guarantors=guarantor}
## Suporte:  0.01
## Confiança:  0.7692308
## Lift:  14.7929
```

Conclusão:

1. Garantias são cruciais: Ter um fiador é um fator muito importante na concessão de crédito.
2. Perfil de bom pagador: Inclui conta corrente com saldo moderado, histórico de crédito positivo, empréstimos para eletrônicos, pouca poupança, taxa de prestação moderada e propriedade de imóveis.
3. Propriedade imobiliária: Forte indicador positivo na avaliação de crédito.
4. Trabalhadores estrangeiros: Há associações específicas para não-estrangeiros, como empréstimos para carros novos.
5. Desemprego: Forte relação entre ser desempregado/não qualificado e duração do desemprego.

6. Estabilidade residencial: Tempo de residência entre 2-4 anos é relevante.
7. Gênero e finalidade: Associação entre homens e empréstimos para carros novos.
8. Tamanho da família: 1-2 dependentes é frequentemente mencionado nas regras.
9. Força das regras: Alta confiança e lift, mas baixo suporte, indicando nichos específicos de clientes.