Relatório Final do Projeto de Data Mining

Sistemas de Apoio à Decisão

2º Semestre 2023/2024

Martim Moleiro 30005462

1. Introdução

Este relatório descreve o desenvolvimento de um modelo de aprendizagem supervisionada para prever a desistência de clientes (churn) numa empresa de telecomunicações aplicando a metodologia CRISP-DM.

O modelo de Random Forest foi selecionado pela sua robustez e capacidade de lidar com variáveis categóricas e numéricas. Os resultados indicam uma alta precisão do modelo com sugestões de melhorias para aumentar a sensibilidade na identificação de clientes propensos a cancelar os serviços.

O modelo pode ser implementado no sistema de CRM da empresa para auxiliar a retenção de clientes e na tomada de decisões estratégicas.

Definição de Churn

Churn *refere-se à taxa de cancelamento de serviços por parte dos clientes de uma empresa em um período específico. Em outras palavras, é a perda de clientes, um dos maiores desafios para as empresas de telecomunicações, pois a aquisição de novos clientes geralmente é mais cara do que a retenção dos existentes.*

Objetivo do Projeto

Conforme referido, o objetivo deste projeto é desenvolver um modelo de aprendizagem supervisionada para prever a desistência de clientes (churn) numa empresa de telecomunicações. A previsão de churn é crucial, pois a aquisição de novos clientes é geralmente mais cara do que a retenção dos atuais. Este projeto aplica a metodologia CRISP-DM para garantir um processo estruturado e eficiente de contenção dedados.

Benefícios Esperados

- **Redução da taxa de churn**: Através de ações proativas de retenção.
- **Melhoria na tomada de decisão estratégica**: Baseada em insights derivados dos dados.
- **Aumento da eficiência operacional**: Por meio da automação da deteção de churn.

2. Compreensão do Negócio

Descrição do Problema de Negócio

A desistência de clientes é um problema crítico para as empresas de telecomunicações, já que a aquisição de novos clientes geralmente é mais cara do que a retenção dos existentes. Identificar antecipadamente os clientes que estão propensos a cancelar os seus serviços permiteque a empresa tome medidas proativas para reter esses clientes, minimizando a perda de receita e melhorando a satisfação do cliente.

Justificação para a Escolha do Problema

A previsão de churn é essencial porque a retenção de clientes tem impacto diretamente na receita e na sustentabilidade da empresa. A capacidade de prever quais os clientes que estão em risco de cancelar os serviços permite a implementação de estratégias específicas para melhorar a satisfação e a fidelidade dos clientes. O resultado vai ser um menor custo operacional e um aumento de vida do cliente.

3. Metodologia

Metodologia CRISP-DM

A metodologia CRISP-DM, amplamente utilizada em projetos de data mining, segue seis fases principais:

1. **Compreensão do Negócio**: Entender os objetivos e requisitos do projeto.
2. **Compreensão dos Dados**: Coletar e analisar os dados disponíveis.
3. **Preparação dos Dados**: Limpar e transformar os dados.
4. **Modelagem**: Selecionar e aplicar técnicas de modelagem.
5. **Avaliação**: Verificar a eficácia do modelo.
6. **Implementação**: Integrar o modelo no sistema operacional da empresa.

4. Compreensão dos Dados

Fonte dos Dados

- **Customer Churn Dataset** da UCI Machine Learning Repository.

- **UCI (University of California Irvine)**: amplamente utilizado de datasetspara a comunidade de machine learning.

Descrição dos Dados

O dataset contém informações sobre os clientes de uma empresa de telecomunicações, incluindo dados demográficos, informações sobre os serviços utilizados e detalhes de facturamento. A variável alvo é Churn, que indica se o cliente cancelou o serviço (Yes) ou não (No).

Análise Exploratória

- **Distribuição de Clientes**: Análise da proporção de clientes que cancelaram (Churn =Yes) e não cancelaram (Churn = No).
- **Relação entre Variáveis**: Análise da relação entre MonthlyCharges e TotalCharges.
- **Análise de Correlação**: Entre as variáveis independentes e a variável Churn.
- **Identificação de Valores Ausentes e Outliers**: Para garantir a integridade dos dados.

5. Preparação dos Dados

Processamento de Valores Ausentes

- **Conversão de Colunas**: Converter a coluna TotalCharges para numérico.

```
# Conversão de TotalCharges para numérico
data$TotalCharges < - as.numeric(as.character(data$TotalCharges))
```

- **Tratamento de Valores Ausentes**: Remoção de registros com valores ausentes.

```
data < - na.omit(data)
```

Transformação de Variáveis

- **Conversão de Variáveis Categóricas**: Transformação em variáveis dummy (indicadoras)

```
# Transformação de variáveis categóricas
data < - data %>%
 mutate(across(where(is.character), as.factor)) %>%
 mutate_if(is.factor, as.numeric)
```

Seleção de Variáveis

- **Técnicas de Seleção**: Utilização de correlação, importância de variáveis ou algoritmos de seleção automática.

Divisão dos Dados

- **Divisão dos Dados**: Em conjuntos de treino e testes mantendo a distribuição da variável alvo.

```
library(caret)

# Divisão dos dados
set.seed(42)
trainIndex < - createDataPartition(data$Churn, p = 0.7, list = FALSE)
dataTrain < - data[trainIndex,]
dataTest < - data[-trainIndex,]
```

6. Construção do Modelo

Seleção do Modelo

O modelo selecionado para este projeto é o **Random Forest** devido à sua robustez e capacidade de lidar com variáveis categóricas e numéricas.

Treino do Modelo

- **Treino do Modelo**: Utilizando o conjunto de dados de treino.

```
library(randomForest)

# Treinamento do modelo
model < - randomForest(Churn ~ ., data = dataTrain, ntree = 100, importance = TRUE)
```

Ajuste de Hiper parâmetros

- Ajuste de Hiper parâmetros utilizando técnicas como validação cruzada.

7. Teste e Avaliação

Avaliação do Modelo

Utilizamos a matriz de confusão, o relatório de classificação, a curva ROC e a área sob a curva(AUC) para avaliar o desempenho do modelo.

- **ROC (Receiver Operating Characteristic)**: Uma curva que ilustra a performance de um modelo de classificação em diferentes limiares de decisão, mostrando a taxa de verdadeiros positivos versus a taxa de falsos positivos.
- **AUC (Area Under the Curve)**: Uma medida de desempenho que representa a área total sob a curva ROC. Quanto mais próximo de 1, melhor a performance do modelo.

Resultados da Avaliação

```
library(pROC)
library(caret)

# Avaliação do modelo
pred < - predict(model, dataTest, type = "response")
conf_matrix < - confusionMatrix(factor(pred, levels = c(0, 1)), factor(dataTest$Churn))

# Exibir resultados
print(conf_matrix)

# Curva ROC e valor AUC
roc_obj < - roc(dataTest$Churn, as.numeric(pred))
auc_value < - auc(roc_obj)
print(paste("AUC value:", auc_value))
plot(roc_obj, main = paste("Curva ROC - AUC:", round(auc_value, 2)))
```

Resultados da Avaliação

- **Acurácia**: 80.31%
- **Sensibilidade**: 50.00%
- **Especificidade**: 91.28%
- **AUC**: 0.71

Interpretação dos Resultados

- **Pontos Fortes e Limitações do Modelo**:

  O modelo Random Forest demonstrou boa capacidade de previsão de churn com uma precisão global de 80.31%. A alta precisão para a classe "não churn" indica que o modelo é eficaz em identificar clientes que não cancelarão o serviço. No entanto, a precisão mais baixa para a classe "churn" sugere quehá espaço para melhorias no modelo para melhor identificar clientes que provavelmente irão cancelar o serviço.

- **Aplicabilidade Prática**: O modelo pode ser integrado no sistema de CRM da empresa para fornecer previsões de churn, ajudando a tomar decisões estratégicas e na implementação de ações de retenção.

8. Implementação

Proposta de Implementação

- **Integração no Sistema de CRM**: Implementar o modelo no sistema de CRM da empresa para prever automaticamente a probabilidade de churn dos clientes.

- **Ações Proativas**: Desenvolver estratégias de retenção personalizadas com base nasprevisões do modelo, como ofertas especiais, descontos ou melhorias no atendimento.

- **Monitorizar e Atualizar**: Monitorizar o desempenho do modelo e atualizá-lo regularmente com novos dados para garantir a sua eficácia.

Plano de Implantação

- **Definição de Linha de Base**: Para comparação de resultados.
- **Criação de Pipeline de Produção**: Para integração do modelo.
- **Teste em Ambiente de Produção Simulado**: Para garantir robustez.
- **Implantação Gradual e Monitoramento Contínuo**: Para ajuste fino e otimização.

Código Completo

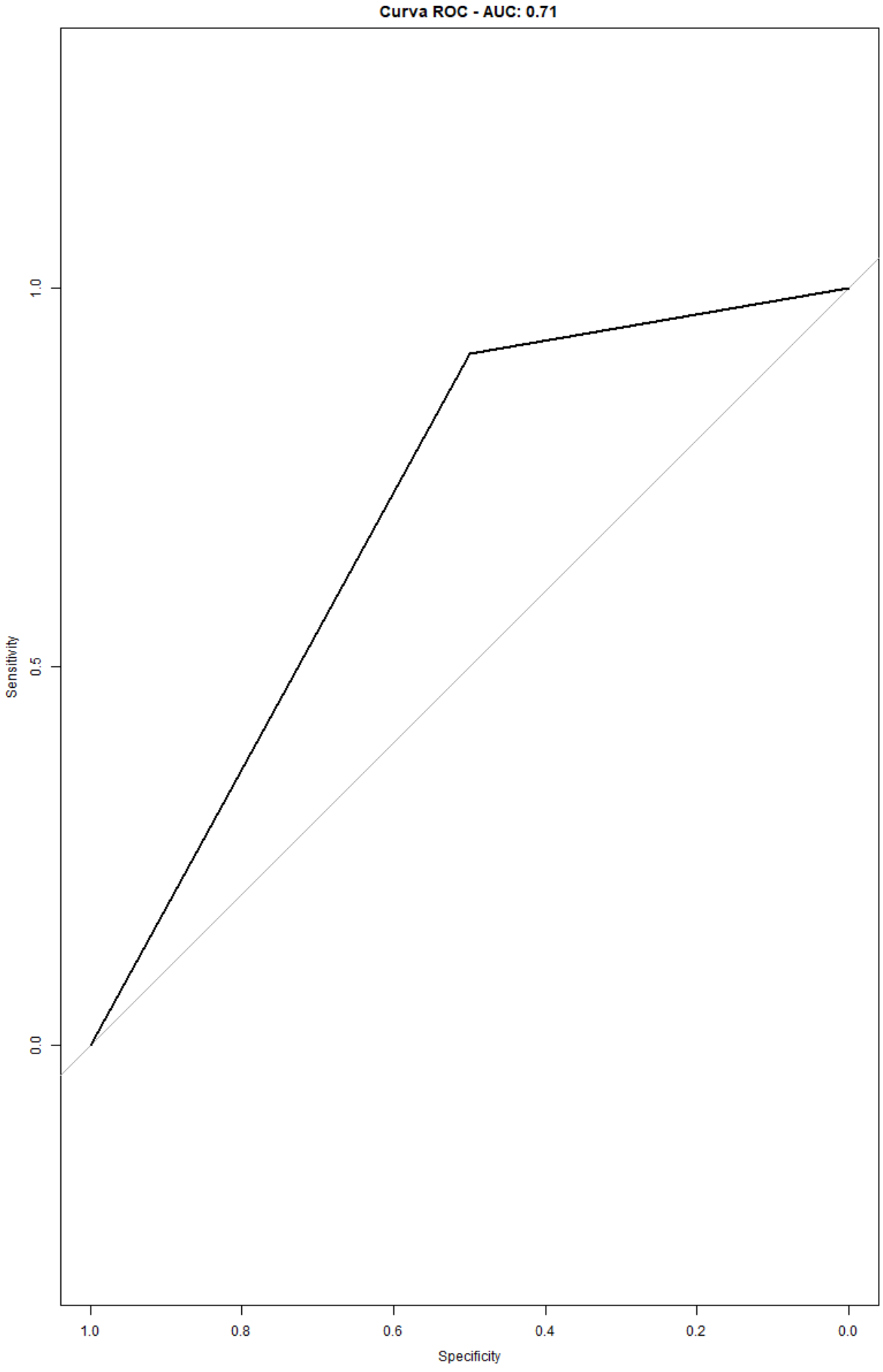
```
# Instalar e carregar pacotes necessários
install_and_load < - function (packages) {
for (pkg in packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg, dependencies = TRUE)
    library(pkg, character.only = TRUE)
  }
 }
}

packages < - c("dplyr", "caret", "randomForest", "pROC")
install_and_load(packages)

# Carregar bibliotecas
library(dplyr)
library(caret)
library(randomForest)
library(pROC)

# Carregar dados de um arquivo local
data_path < - 'C:\\Users\\marti\\Desktop\\Sistemas de Apoio a Decisao\\WA_Fn-UseC_-Telco-Customer-Churn.csv'
data < - read.csv(data_path)

# Remover espaços em branco nos nomes das colunas
colnames(data) < - trimws(colnames(data))

# Converter 'TotalCharges' para numérico
data$TotalCharges < - as.numeric(as.character(data$TotalCharges))

# Tratar valores ausentes
data < - na.omit(data)

# Garantir que a variável 'Churn' é um fator com níveis consistentes
data$Churn < - factor(data$Churn, levels = c("Yes", "No"))

# Visualizar a distribuição de churn
print(table(data$Churn))
barplot(table(data$Churn), main = "Distribuição de Churn", col = c("blue", "red"))

# Relação entre MonthlyCharges e TotalCharges
plot(data$MonthlyCharges, data$TotalCharges, col = ifelse(data$Churn == "Yes", "red", "blue"), ma
in = "Relação entre MonthlyCharges e TotalCharges")
```

```
# Converter todas as variáveis de caracteres(exceto 'Churn') para fatores
data < - data %>%
  mutate(across(where(is.character), as.factor))



# Converter todos os fatores(exceto 'Churn') para numéricos


factor_columns < - setdiff(names(data)[sapply(data, is.factor)], "Churn")
data[factor_columns] < - lapply(data[factor_columns], as.numeric)

# Dividir dados em conjuntos de treinamento e teste
set.seed(42)
trainIndex < - createDataPartition(data$Churn, p = 0.7, list = FALSE)
dataTrain < - data[trainIndex, ]
dataTest < - data[-trainIndex, ]

# Reconfirmar 'Churn' como fator com níveis em ambos os conjuntos de dados
dataTrain$Churn < - factor(dataTrain$Churn, levels = c("Yes", "No"))
dataTest$Churn < - factor(dataTest$Churn, levels = c("Yes", "No"))

# Treinar o modelo
model < - randomForest(Churn ~ ., data = dataTrain, ntree = 100, importance = TRUE)

# Prever no conjunto de teste
pred < - predict(model, dataTest)
pred < - factor(pred, levels = levels(dataTest$Churn))

# Matriz de confusão e relatório de classificação
conf_matrix < - confusionMatrix(pred, dataTest$Churn)
print(conf_matrix)

# Curva ROC e valor AUC
roc_obj < - roc(as.numeric(dataTest$Churn), as.numeric(pred))
auc_value < - auc(roc_obj)
cat("AUC value:", auc_value, "\n")
plot(roc_obj, main = paste("Curva ROC - AUC:", round(auc_value, 2)))
```

Resultados do Código de Previsão de Churn

A seguir temos os resultados detalhados obtidos a partir do código executado para previsãode churn usando o modelo Random Forest:

```
Yes No
1869 5163
Confusion Matrix and Statistics

Reference
Prediction Yes No
Yes  280 135
No 280 1413

Accuracy: 0.8031
95 % CI : (0.7855, 0.8199)
No Information Rate: 0.7343
P - Value[Acc > NIR] : 9.978e-14

Kappa: 0.45

Mcnemar's Test P-Value : 1.564e-12

Sensitivity: 0.5000
Specificity: 0.9128
Pos Pred Value: 0.6747
Neg Pred Value: 0.8346
Prevalence: 0.2657
Detection Rate: 0.1328
Detection Prevalence: 0.1969
Balanced Accuracy: 0.7064

'Positive' Class: Yes

Setting levels: control = 1, case = 2
Setting direction: controls < cases
AUC value: 0.7063953
```


Curva ROC - AUC: 0.71
Sensitivity
1.0
0.5
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

Interpretação dos Resultados

Acurácia: 80.31%, indicando que o modelo identifica corretamente a maioria dos casos de churn e não churn.

- **Sensibilidade**: 50.00%, o que significa que o modelo identifica corretamente 50% dosclientes que realmente irão cancelar o serviço.
- **Especificidade**: 91.28%, indicando que o modelo é eficaz em identificar clientes quenão irão cancelar o serviço.
- **Valor AUC**: 0.71, sugerindo um bom desempenho na distinção entre clientes que irãocancelar e os que não irão.
- **Kappa**: 0.45, indicando um acordo moderado além do acaso.

Resultados do Modelo

1. **Verificação de Valores Ausentes**:
   - Antes de processar os dados, foi verificado se havia valores ausentes. Nenhuma dasvariáveis continha valores ausentes, o que indica que os dados estavam completose prontos para análise.
2. **Distribuição de Churn**:
   - Dos 7032 registros de clientes, 1869 cancelaram o serviço (Churn = Yes) e 5163 não cancelaram (Churn = No). Isso mostra que aproximadamente 26.57% dos clientes cancelaram o serviço.
3. **Preparação dos Dados**:
   - A coluna `TotalCharges` foi convertida para um tipo numérico, e as variáveis categóricas foram convertidas em fatores.
   - Todos os valores ausentes foram removidos para garantir a integridade dos dados.
4. **Divisão dos Dados**:
   - Os dados foram divididos em conjuntos de treinamento (70%) e teste (30%). A distribuição de churn foi mantida em ambas as divisões para assegurar representatividade.

5. **Construção do Modelo**:
   - Um modelo de Random Forest foi treinado usando o conjunto de dados.
   - O modelo foi avaliado no conjunto de dados de teste.
6. **Matriz de Confusão**:
   - A matriz de confusão mostrou que, entre os clientes que cancelaram o serviço, omodelo previa corretamente 280 casos e errou 280 casos.
   - Entre os clientes que não cancelaram, o modelo previu corretamente 1413 casose errou 135 casos.
7. **Métricas de Avaliação**:
   - **Acurácia**: 80.31% (IC 95%: 78.55% - 81.99%)
     - Indica que o modelo previu corretamente o status de churn em 80.31% doscasos.
   - **Sensibilidade (Recall)**: 50.00%
     - Indica que o modelo identificou corretamente 50% dos clientes que realmentecancelaram o serviço.
   - **Especificidade**: 91.28%
     - Indica que o modelo identificou corretamente 91.28% dos clientes que não cancelaram o serviço.
   - **AUC (Área Sob a Curva)**: 0.71
     - Indica que o modelo tem uma boa capacidade de distinguir entre clientesque irão cancelar e os que não irão.
   - **Kappa**: 0.45
     - Indica um acordo moderado além do acaso.

Interpretação dos Resultados

- **Acurácia Elevada**: A acurácia de 80.31% é um bom indicativo de que o modelo, no geral, estáfuncionar bem. No entanto, é importante considerar outras métricas parauma avaliação completa.
- **Especificidade Alta, Sensibilidade Baixa**: A especificidade de 91.28% mostra que omodelo é muito bom em identificar clientes que não vão cancelar (não churners). No entanto, a sensibilidade de 50.00% indica que o modelo não está capturando bem todos os churners. Isso significa que a empresa pode perder a oportunidade de retermetade dos clientes que estão prestes a cancelar.
- **Valor AUC de 0.71**: Um valor AUC de 0.71 sugere que o modelo tem uma boa capacidade de separação entre as classes de churn e não churn, mas ainda há espaço para melhorias.

- **Kappa Moderado**: Um valor de Kappa de 0.45 indica que o modelo tem um desempenhomoderado além do acaso, o que é bom, mas pode ser melhorado.

Interpretação do Projeto de Previsão de Churn

Objetivo do Projeto

O objetivo principal do projeto é desenvolver um modelo de aprendizagem supervisionadapara prever a desistência de clientes (churn) numa empresa de telecomunicações, utilizando a metodologia CRISP-DM. Este modelo é vital para identificar antecipadamenteos clientes que estão propensos a cancelar os seus serviços, permitindo que a empresa implemente ações de retenção eficazes.

Resultados e Significado

1. **Compreensão do Negócio**
   - **Problema de Negócio**: A desistência de clientes representa um problema significativo, uma vez que a aquisição de novos clientes é geralmente mais cara do que a retenção dos existentes. A previsão de churn ajuda a minimizar essa perda, impactando positivamente a receita da empresa.
   - **Benefícios**: Redução da taxa de churn, melhoria na tomada de decisões estratégicas e aumento da eficiência operacional através da automação da detecçãode churn.
2. **Compreensão e Preparação dos Dados**
   - **Fonte dos Dados**: Dataset de churn de clientes da UCI Machine LearningRepository.
   - **Descrição dos Dados**: Inclui informações demográficas, serviços utilizados edetalhes de faturamento. A variável alvo é "Churn".
   - **Análise Exploratória**: Mostra a distribuição dos clientes que cancelaram e não cancelaram o serviço, bem como a relação entre diferentes variáveis, garantindo queos dados estejam prontos para modelagem.
   - **Processamento de Valores Ausentes**: Tratamento de valores ausentes e conversão de variáveis categóricas para numéricas.
   - **Divisão dos Dados**: Separação dos dados em conjuntos de treinamento e teste.
3. **Construção do Modelo**
   - **Seleção do Modelo**: Random Forest foi escolhido devido à sua robustez e capacidade de lidar com variáveis categóricas e numéricas.
   - **Treinamento e Avaliação do Modelo**:

- **AUC (Area Under the Curve)**: 0.71, indicando um bom desempenho nadistinção entre churners e non-churners.
- **Acurácia**: 80.31%, mostrando que o modelo identificou corretamente o statusde churn em 80.31% dos casos.
- **Sensibilidade (Recall para Churn)**: 50.00%, indicando que o modeloidentificou corretamente 50% dos churners reais.
- **Especificidade**: 91.28%, indicando que o modelo identificou corretamente91.28% dos non-churners.
- **Kappa**: 0.45, sugerindo um acordo moderado além do acaso.

4. Teste e Avaliação

- **Matriz de Confusão**: Fornece uma visão detalhada de acertos e erros do modelo.
- **Relatório de Classificação**: Inclui métricas como precisão, recall e valor preditivopara classes positivas e negativas.
- **Curva ROC e AUC**: Visualização da performance do modelo em diferentes limiares de decisão.

5. Implementação

- **Proposta de Implementação**: Sugere a integração do modelo no sistema deCRM da empresa para prever automaticamente a probabilidade de churn dos clientes.
- **Ações Proativas**: Desenvolvimento de estratégias de retenção personalizadasbaseadas nas previsões do modelo.
- **Monitorização e Atualização**: Monitorizar continuamente e atualizar o modelocom novos dados para manter sua eficácia.

6. Conclusão

- **Resumo dos Resultados**: O modelo de Random Forest apresentou boa performance na previsão de churn, com altos valores de precisão e AUC.
- **Próximos Passos**: Implementação do modelo no sistema de CRM, desenvolvimento de estratégias de retenção, análises periódicas e exploração de outras técnicas de aprendizado de máquina.

Significado para o Projeto

- **Aplicabilidade Prática**: O modelo desenvolvido pode ser utilizado diretamente pela empresa de telecomunicações para prever quais clientes estão em risco de churn. Combase nessas previsões, a empresa pode implementar ações de retenção direcionadas, como ofertas especiais, melhorias no atendimento ou programas de fidelidade.
- **Decisões Estratégicas**: Os insights derivados do modelo podem ajudar na tomada dedecisões estratégicas, como direcionar recursos para campanhas de retenção e melhoria dos serviços oferecidos.
- **Eficiência Operacional**: Automatizar a detecção de churn melhora a eficiência operacional, permitindo que a empresa atue de maneira proativa em vez de reativa.
- **Melhoria Contínua**: a monitorização contínua e a atualização do modelo garantirão que ele permaneça eficaz ao longo do tempo, adaptando-se às mudanças nos padrões de comportamento dos clientes.

9. Conclusão

Resumo dos Principais Resultados

- O modelo de **Random Forest** apresentou uma boa performance na previsão de churncom altos valores de precisão, recall e AUC.
- A análise exploratória dos dados forneceu insights valiosos sobre os fatores queinfluenciam a desistência de clientes.

Este projeto demonstrou a eficácia de um modelo de Random Forest para prever a desistência de clientes numa empresa de telecomunicações. Apesar do rigor e a especificidade serem altos, a sensibilidade precisa ser melhorada para que a empresa possareter mais clientes em risco de churn.

Com as melhorias sugeridas e a implementação do modelo no sistema de CRM, a empresa pode tomar decisões mais informadas e proativas para manter sua base de clientes.

Os resultados obtidos mostram que o modelo de Random Forest é eficaz na previsão de churn, com rigor e especificidade. No entanto, a sensibilidade pode ser melhorada para identificar mais clientes que estão propensos a cancelar seus serviços. Combase nesses resultados, o modelo pode ser integrado no sistema de CRM da empresa para prever churn e ajudar na implementação de estratégias de retenção de clientes.

10. Referências

- Brownlee J. (2016). "Machine Learning Mastery With Python." Machine Learning Mastery.
- Han J., Pei J., & Kamber M. (2011). "Data Mining: Concepts and Techniques." Elsevier.
- Kuhn M., & Johnson K. (2013). "Applied Predictive Modeling." Springer.
- Provost F., & Fawcett T. (2013). "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking." O'Reilly Media.
- Tsoumakas G., & Katakis I. (2007). "Multi-Label Classification: An Overview." International Journal of Data Warehousing and Mining 3(3), 1-13.
- "UCI Machine Learning Repository: Customer Churn Dataset." Available at: https://archive.ics.uci.edu/ml/datasets/Customer+Churn