Inteligência de Negócio - Trabalho 2 - Mineração de Dados

Caroline Rosa [17120082],
Débora Pires [16180233],
Franciele Constante [15180620],
Lucas Assumpção [20180356] e
Sylvio Correa [16180510]

2020/2

1 Justificativa do problema e tarefa escolhida

Para este segundo trabalho da disciplina de Inteligência de Negócio, a tarefa de mineração de dados escolhida foi a de classificação. O conjunto de dados escolhido para a tarefa foi o Telco Customer Churn, uma base de dados contendo informações sobre clientes do serviço de telefonia e internet da empresa Telco. A coluna alvo desta base de dados é a coluna Churn que contém as categorias Yes e No e indica se um cliente cancelou o serviço da companhia no último mês.

Cada uma das 7043 entradas da base de dados representa um cliente da companhia. A base possui os seguintes atributos:

- CustomerID: chave única que identifica cada cliente. Meta-atributo.
- Gender: gênero de cada cliente. Atributo categórico.
- SeniorCitizen: indica se o cliente é idoso. Atributo categórico.
- Partner: indica se o cliente está em um relacionamento com outra pessoa. Atributo categórico.
- Dependents: indica se o cliente possui dependentes. Atributo categórico.
- tenure: indica há quantos meses um cliente está no contrato com a companhia. Atributo numérico.
- PhoneService: indica se um cliente possui ou não serviço telefônico. Atributo categórico.
- MultipleLines: indica se um cliente possui múltiplas linhas telefônicas, só uma, ou não possui serviço telefônico. Atributo categórico.

- InternetService: indica se um cliente possui serviço de internet DSL, fibra ótica ou nenhum serviço de internet. Atributo categórico.
- OnlineSecurity: indica se o cliente contratou o serviço de segurança online, não contratou, ou não possui serviço de internet. Atributo categórico.
- OnlineBackup: indica se o cliente contratou o serviço de backup online, não contratou, ou não possui serviço de internet. Atributo categórico.
- DeviceProtection: indica se o cliente contratou o serviço de proteção de dispositivo, não contratou, ou não possui serviço de internet. Atributo categórico.
- TechSupport: indica se o cliente contratou o serviço de suporte técnico, não contratou, ou não possui serviço de internet. Atributo categórico.
- StreamingTV: indica se o cliente contratou o serviço de streaming para tv, não contratou, ou não possui serviço de internet. Atributo categórico.
- StreamingMovies: indica se o cliente contratou o serviço de streaming de filmes, não contratou, ou não possui serviço de internet. Atributo categórico.
- Contract: indica que tipo de contrato o cliente tem com a companhia: mensal, anual ou bienal. Atributo categórico.
- PaperlessBilling: indica se o cliente aderiu à opção de pagamento paperless. Atributo categórico.
- PaymentMethod: o tipo de pagamento contratado pelo cliente: cheque eletrônico, cheque por correio, transferência bancária (automática), cartão de crédito (automática). Atributo categórico.
- MonthlyCharges: mensalidade contratada pelo cliente. Atributo numérico.
- TotalCharges: total pago pelo cliente ao longo do período contratual. Atributo numérico.
- Churn: coluna alvo. Indica se o consumidor cancelou o serviço no último mês.

A previsão de churn é uma tarefa comum no campo de mineração de dados aplicada a negócios. Prever o cancelamento do serviço permite que o administrador do negócio possa tomar atitudes a fim de evitar a perda do cliente, direcionando recursos de infraestrutura e marketing de maneira precisa para a manutenção daquela conta ou mesmo, se preferir, evitar maiores gastos com um cliente que não dará retorno.

A ferramenta escolhida para realizar esta tarefa foi o Orange.

2 Análise e pré-processamento dos dados

Analisando os dados usando o node Feature Statistics do Orange, foram identificadas 11 entradas com valor faltante para o atributo TotalCharges. Usando o node Correlations, constatatou-se que nenhum dos atributos numéricos tinha correlação significativa o suficiente que justificasse a escolha de um deles para eliminação durante o pré-processamento.

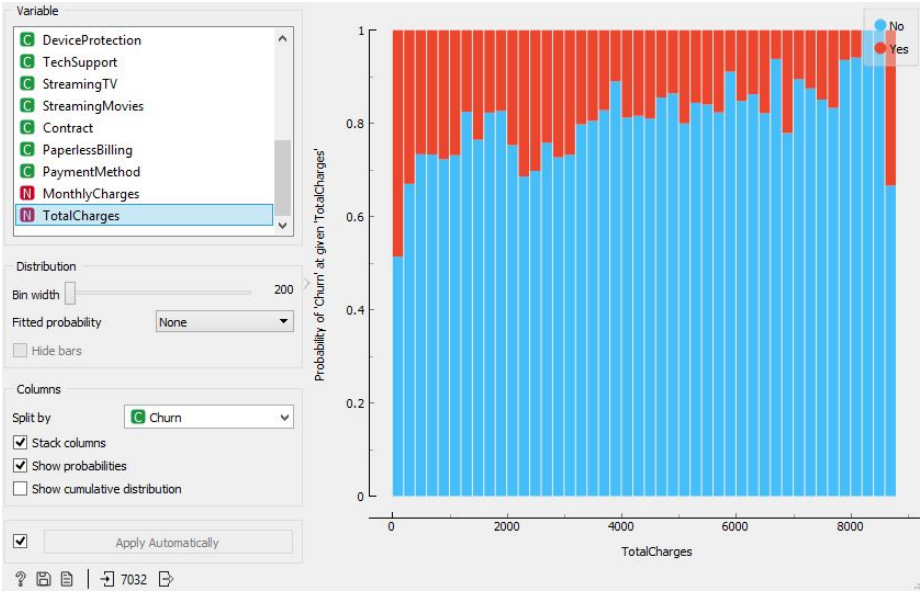

Figure 1: Distribuição de churn para o atributo TotalCharges. A cor vermelha indica Churn = yes.

Foi feita uma análise exploratória usando o node Distributions para verificar a distribuição do atributo alvo em relação a cada outro atributo presente no dataset. Verificou-se que este é um dataset desbalanceado para a coluna alvo: apenas 26% das entradas correspondem a clientes que cancelaram seus contratos com a companhia. Algumas correlações chamam atenção, como é o caso da distribuição de Churn para o atributo TotalCharges (imagem 1). Quase metade dos cliente que pagaram menos de 200 dólares ao todo cancelaram o contrato. Mais significativo ainda é a relação entre tempo de contrato e cancelamento (imagem 2). dos clientes com até 2 meses de contrato, 62% cancelaram seus planos com a companhia. É visível que esta proporção cai na medida em que aumenta a quantidade de meses de vigência de contrato.

Por outro lado, o gênero do cliente parece ter baixíssima influência na hora de prever o cancelamento (imagem 3). A proporção entre homens e mulheres no dataset é praticamente $1:1$ e, para ambos, a taxa de churn fica em torno de $26,5\%$.

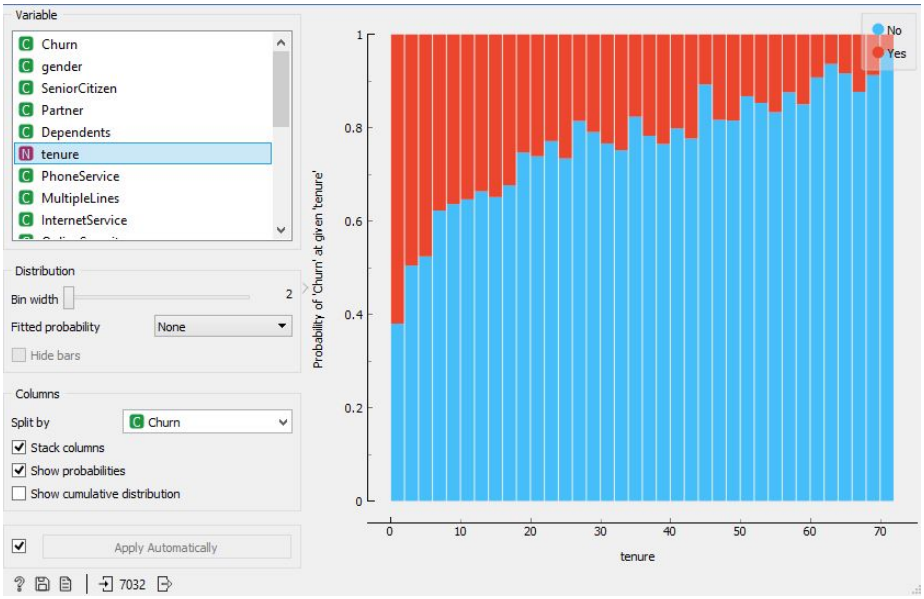

Figure 2: Distribuição de churn para o atributo Tenure. A cor vermelha indica Churn = yes.

Para fins de pré-processamento dos dados, foram eliminadas as 11 entradas com valores nulos. Então foi efetuada a normalização das colunas numéricas para que os valores ficassem com média 0 e variância 1. As colunas categóricas foram binarizadas usando one-hot encoding.

O resultado do one-hot-encoding para atributos com apenas duas categorias complementares gera dois novos atributos que são redundantes. Um exemplo é a coluna Gender. A binarização gera dois novos atributos: Male e Female. Qualquer um deles pode ser eliminado deixando apenas o outro sem que haja perda de informação. Uma curiosidade deste dataset é que ele possui muitos atributos com categorias que contém uma mesma informação. Um exemplo são os atributos InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV e StreamingMovies. Todos estes atributos possuem uma categoria que indica que o cliente não possui serviço de internet. Por este motivo, esta informação é mantida apenas pelo atributo InternetService=No. Usando o node Select Columns do Orange, todas as demais colunas equivalente foram eliminadas. Com isso resta ainda um tipo de atributo com informação redundante, o que identifica a ausência de um serviço em particular (ex: OnlineSecurity=no). Estes atributos também foram eliminados. O mesmo foi feito para atributos repetidos que indicavam a ausência de serviço de telefonia.

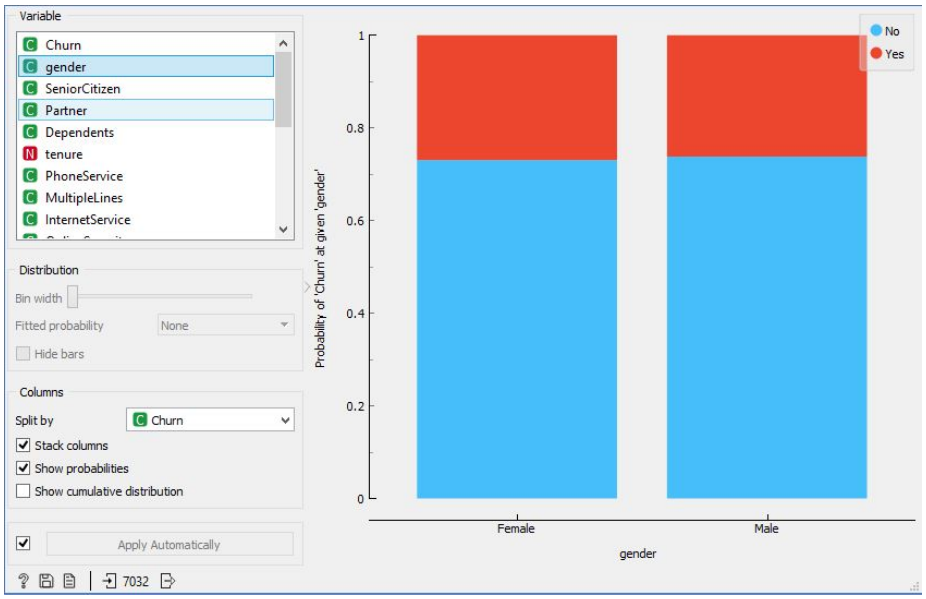

Figure 3: Distribuição de churn para o atributo Gender. A cor vermelha indica Churn = yes.

3 Algoritmo escolhido

O algoritmo selecionado foi a Regressão Logística.

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica. O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos. Em comparação com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo facto de a variável resposta ser categórica.

Seu funcionamento se dá através da função logística conhecida como Sigmóide. Independentemente do valor de entrada, a função sempre retorna valores entre 0 e 1, nunca sendo 0 ou 1, onde esse valor corresponde à probabilidade de o objeto pertencer à classe 0 ou à classe 1.

O processo tem como base a regressão linear, logo o objetivo do algoritmo é encontrar uma função que determine diferentes pesos aos atributos, minimizando uma função de custo que calcula a discrepância entre as predições e as classificações reais dos dados de treino.

Para calcular a entropia, selecionamos um objeto que desejamos classificar e o inserimos na função como sendo o valor original. Quanto mais próximo da fronteira de decisão, maior a possibilidade de erro da classificação. Por exemplo, se definirmos que a fronteira de decisão seja 0.6, todos os objetos cujo resultado seja maior que esse valor serão classificados na classe 1, bem como os objetos

cujo resultado seja menor do que esse valor serão classificado na classe 0. Se o resultado for superior a esse limite, mas próximo dele, como 0.65, por exemplo, ele será classificado como 1, mas com uma alta taxa de erro, que é a possibilidade de a classificação não ser precisa.

O custo médio se resume ao somatório das entropias dos objetos dividido pela quantidade de objetos da base de dados. O algoritmo de regressão logística itera ajustando os pesos usados na função de predição de modo a minimizar este custo médio.

4 Resultados encontrados

Os dados processados foram utilizados em três tipos diferentes de modelos: regressão logística, K-nn e árvore de decisão. O modelo resultante da regressão logística superou os demais em mais medidas, motivo pelo qual foi escolhido.

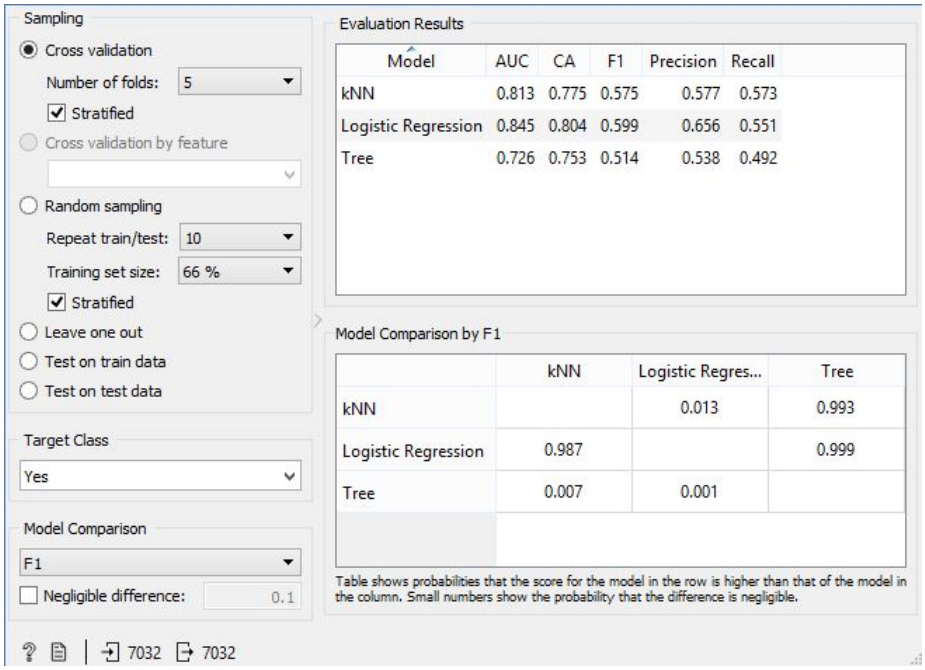

Figure 4: Resultados mostrados pelo node Test and Score para regressão logística, K-nn e árvore de decisão

Todos os modelos foram treinados e testados usando o mesmo protocolo de validação k-fold cross validation com $k = 5$. As métricas usadas na avaliação dos modelos foram área sob a curva ROC, acurácia, F-score, precisão e recall. São apresentadas figuras mostrando a comparação da curva ROC de cada modelo (imagem 5) e a matriz de confusão para a regressão logística (imagem 6.

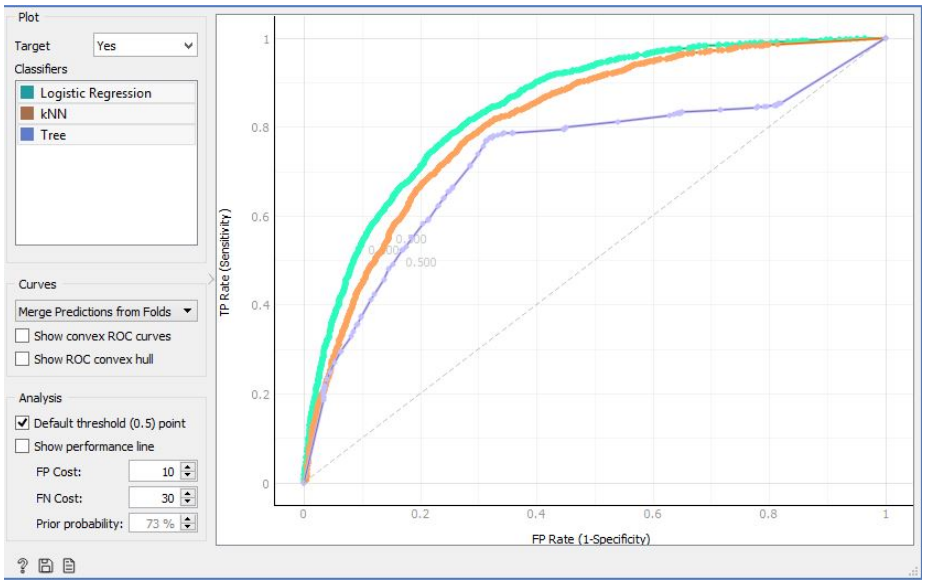

Figure 5: Curva ROC para os três modelos testados.

5 Conclusões sobre as descobertas

A curva ROC mostra que a regressão logística mantém um desempenho médio constantemente superior aos demais modelos quando levado em conta a proporção entre verdadeiros positivos e falsos positivos.

Cabe salientar que é de particular importância para a tarefa escolhida identificar aqueles clientes que irão cancelar o contrato (relativamente mais importante do que identificar aqueles que não irão cancelar o contrato). A medida que indica a proporção de elementos da classe positiva corretamente identificados é o recall. O recall do modelo de regressão logística é ligeiramente inferior ao do K-nn. Porém a precisão do modelo de regressão logística é significativamente superior ao do K-nn. Julgamos que a capacidade do modelo K-nn de encontrar verdadeiros positivos não é suficientemente superior para justificar sua falta de precisão.

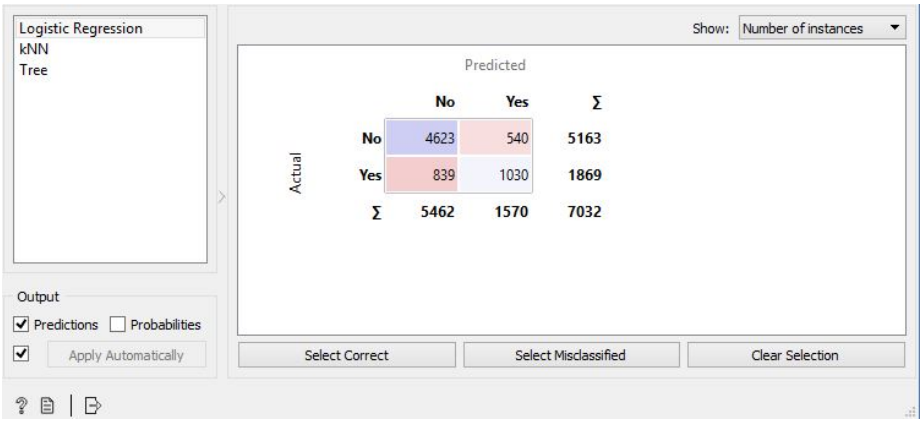

Figure 6: Matriz de confusão para o modelo de regressão logística.