**Prevendo Customer Churn em Operadoras de Telecom**

**Introdução:**

Customer Churn (ou Rotatividade de Clientes, em uma tradução livre) refere-se a uma decisão tomada pelo cliente sobre o término do relacionamento comercial. Refere-se também à perda de clientes. A fidelidade do cliente e a rotatividade de clientes sempre somam 100%. Se uma empresa tem uma taxa de fidelidade de 60%, então a taxa de perda de clientes é de 40%. De acordo com a regra de lucratividade do cliente 80/20, 20% dos clientes estão gerando 80% da receita. Portanto, é muito importante prever os usuários que provavelmente abandonarão o relacionamento comercial e os fatores que afetam as decisões do cliente.

**Objetivo do Projeto:**

Criar um modelo de aprendizagem de máquina que possa prever se um cliente pode ou não cancelar seu plano e qual a probabilidade de isso ocorrer.

**Etapas do projeto:**

**Análise Exploratória:**

O dataset tem um bom número de registros, porém tem que ter cuidado ao realizar remoção de registros para não afetar o objetivo final.

As variáveis categóricas do dataset estão com o tipo texto, devem ser transformadas para números.

Dataset não contem nenhum valor nulo.

O dataset tem muitos valores unicos, o que indica que temos bastantes variáveis numéricas, pode-se fazer quantization (Técnica para transformar valores em categorias) das variáveis numéricas para melhor performance do modelo.

**Análise Exploratória das Variáveis Numéricas**:

A média e mediana estão muito aproximadas, o desvio padrão está com valor baixo, indica que os dados estão próximos da média.

A variável total\_day\_charge tem uma correlação positiva com a variável total\_day\_minutes, ou seja ao aumentar os minutos falados no dia aumenta-se o custo.

A variável total\_eve\_charge tem uma correlação positiva com a variável total\_eve\_minutes, ou seja ao aumentar os minutos falados na vespera aumenta-se o custo.

De acordo com os histogramas, as variáveis numéricas estão aparentemente em uma distribuição normal com exceção da variável "number\_vmail\_messages".

De acordo com os boxplots, todas as variáveis tem valores outliers, a maioria dos outliers estão localizados no 1º e 3º quartil.

Foi realizado teste de hipótese para validar a distribuição normal, porém as variáveis account\_length, number\_vmail\_messages, total\_day\_calls, total\_eve\_calls, total\_intl\_minutes, total\_intl\_calls, total\_intl\_charge e number\_customer\_service\_calls não estão em uma distribuição normal, será realizado normalização nas variáveis para melhor performance do modelo preditivo.

Realizado plot de um heatmap com a correlação de todas as variáveis numéricas e pode-se concluir que o dataset tem poucas variáveis correlacionadas, porém as correlações existentes são fortes.

As variáveis estão simétricas, algumas com uma leve distorção para a cauda esquerda e outras para a cauda direita, mas o valor da distorção é muito baixo.

**Análise Exploratória das Variáveis Categóricas**:

De acordo com os gráfico de barras, os estados com mais registros são WV e MN, o código de área com mais registro é area\_code\_415, a maioria dos registros não tem plano internacional e não tem plano de correio de voz. A variável TARGET (churn) está desbalanceada, com mais de 2500 registros com churn “no” e aproximadamente 500 registros com churn “yes”, para melhor performance do modelo preditivo, precisa-se balancear a variável churn.

Realizado LabelEncoder nas variáveis para realizar análise de correlação. A correlação das variáveis preditoras com a variável target é fraca, muito próxima de zero, a variável preditora que tem mais correlação com a variável target é International\_Plan.

**Manipulação dos Dados:**

Removido a coluna "Unnamed: 0", transformado variáveis categóricas que estão em texto para número, normalizado e padronizado as variáveis numéricas e removido outliers do dataset de treino.

Ao finalizar a manipulação dos dados o dataset continua com uma boa quantidade de registros para o treinamento do modelo de machine learning, foram removidos 514 registros ao eliminar os outliers.

**Feature Selection:**

Utilizado a função SelectKBest do SKLearn e método estatístico ANOVA para selecionar as melhores variáveis para o modelo de machine learning. As variáveis selecionadas foram international\_plan, total\_day\_minutes, total\_day\_charge, number\_customer\_service\_calls, total\_eve\_minutes, total\_eve\_charge, voice\_mail\_plan, number\_vmail\_messages.

**Preparar dataset de treino e teste para o treinamento:**

Tratado o dataset de teste utilizando as mesmas técnicas da etapa de manipulação de dados no dataset de treino. Balanceado a classe target (churn), o dataset de treino ficou com 4940 registros, 8 variáveis preditoras e uma variável target.

**Treinamento do Modelo de Machine Learning:**

Treinado os algoritmos Naive Bayes, Regressão Logística, KNN e SVM, os modelos que apresentaram melhores resultados foram KNN e SVM, foi escolhido os 2 modelos para seguir para as próximas etapas.

A métrica escolhida para avaliação do modelo foi "Recall", o recall é intuitivamente a capacidade do classificador de encontrar todas as amostras positivas.

O modelo SVM foi superior na métrica recall, foi o modelo escolhido para receber otimização de hiperparametros.

**Otimização do Modelo de Machine Learning:**

Treinado o modelo com algoritmo Random Forest e XGBoost, porém não tiveram a métrica recall superior ao modelo SVM.

Otimizado hiperparametros do modeloSVM com GridSearchCV, os parametros escolhidos foram C=0.1, class\_weight={0:1.0, 1:1.12}, gamma='scale', kernel='rbf' e 'tol'=0.001.

O modelo otimizado teve uma queda de 3% de recall para a classe 0 (No), porém teve um aumento de 5% de recall para a classe 1 (Yes).

Para mais informações sobre o algoritmo SVM, consultar documentação do SKLearn. https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html

**Salvar o Modelo de Machine Learning para Entrega Final do Projeto:**

Previsto churn do dataset de teste, mostrado o resultado no notebook e entregue o modelo de machine learning na pasta “modelo” na raiz do projeto.

**FIM.**