Relatório de Análise de Churn em Streaming

1. Introdução

Este relatório apresenta o processo de análise de dados e modelagem preditiva realizado com a ferramenta Orange Data Mining para identificar clientes com maior probabilidade de churn em uma empresa de streaming. O objetivo é propor um fluxo de trabalho que permita prever o risco de cancelamento de clientes e apoiar a tomada de decisão estratégica.

2. Base de Dados

Foi utilizada uma base de dados simulada denominada 'streaming_churn_dataset.csv', composta por 158 registros de clientes, contendo variáveis como idade, gênero, tipo de assinatura, número de horas assistidas por semana, atrasos em pagamento, tickets de suporte e a variável alvo 'churn'.

3. Pré-processamento dos Dados

Foram aplicadas as seguintes etapas de preparação de dados no Orange:
- Tratamento de valores ausentes com o widget Impute;
- Normalização dos atributos numéricos para padronizar a escala;
- Seleção das variáveis mais relevantes para a modelagem.

4. Modelagem Preditiva

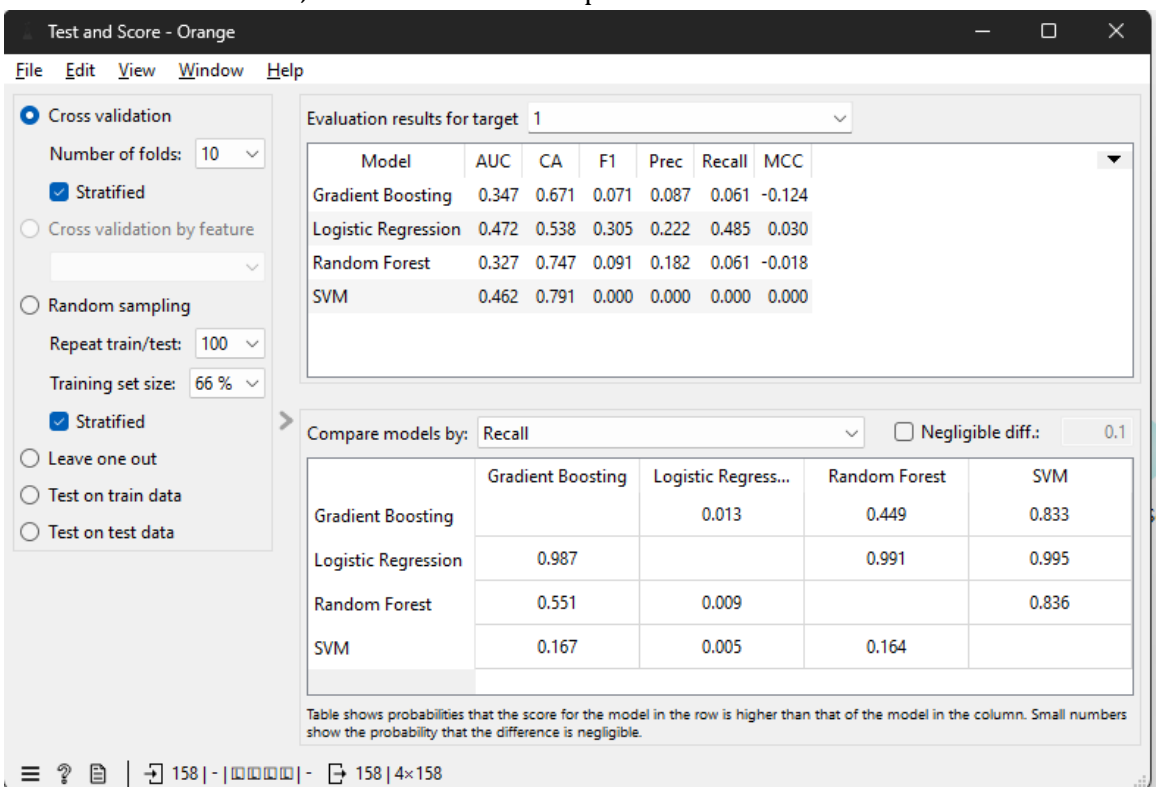
Foram testados diferentes algoritmos de classificação no Orange, incluindo:
- Regressão Logística;
- Random Forest;
- Gradient Boosting;
- Support Vector Machine (SVM).

5. Avaliação dos Modelos

A avaliação foi realizada utilizando o widget Test & Score com validação cruzada estratificada (10 folds). As métricas analisadas foram AUC, Acurácia (CA), F1 Score, Precisão, Recall e MCC.

Resumo dos principais resultados:
- Regressão Logística: Recall de 48,5% e F1 de 0,305, sendo o modelo que melhor identifica clientes churners.
- Random Forest: Alta acurácia (74,7%), porém baixo Recall (6,1%), mostrando que o modelo praticamente classifica todos como não churn.
- Gradient Boosting: Recall muito baixo (6,1%) e F1 de 0,071.

- SVM: Recall e F1 nulos, indicando falha em capturar clientes churn.



6. Matriz de Confusão

A matriz de confusão da Regressão Logística apresentou:
- 69 verdadeiros negativos;
- 16 verdadeiros positivos;
- 56 falsos positivos;
- 17 falsos negativos.

Apesar de erros de classificação, o modelo conseguiu capturar uma quantidade relevante de clientes churn em comparação com os demais algoritmos.

7. Curva ROC

A análise da curva ROC mostrou que a Regressão Logística e o SVM possuem desempenho semelhante em termos de separação (AUC em torno de 0,46), porém a Regressão Logística se destacou por apresentar maior Recall.

8. Ajuste de Threshold

Através do widget ROC Analysis, foi possível simular ajustes no threshold de decisão. Reduzindo o threshold abaixo de 0,5, aumentou-se a sensibilidade (Recall), o que é essencial em problemas de churn, pois é preferível identificar mais clientes com risco mesmo que ocorram falsos positivos.

9. Exportação de Resultados

As previsões geradas pelo modelo foram exportadas utilizando os widgets Predictions e Save Data. O arquivo resultante em formato CSV contém a probabilidade de churn para cada cliente, possibilitando ao time de marketing priorizar ações de retenção.

10. Conclusão

A análise demonstrou que, entre os modelos testados, a Regressão Logística apresentou o melhor desempenho para prever churn, principalmente por alcançar maior Recall. Esse modelo pode ser usado para apoiar estratégias de retenção, segmentando clientes de acordo com suas probabilidades de cancelamento. Como trabalhos futuros, sugere-se aplicar técnicas de balanceamento de classes (como SMOTE) e testar ensembles de modelos para aprimorar ainda mais os resultados.