Desafio - Estágio

Desafio consiste em propor um modelo de aprendizado de máquina que calcule a probabilidade de Churn de um cliente do conjunto de dados do link: https://extremedigital-my.sharepoint.com/personal/david_duarte_extreme_digital/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fdavid%5Fduarte%5Fextreme%5Fdigital%2FDocuments%2Fchurn%2Ezip&parent=%2Fpersonal%2Fdavid%5Fduarte%5Fextreme%5Fdigital%2FDocuments&ga=1

Para solução desse desafio, inicie com a visualização dos dados e o uso de alguns gráficos para compreensão como o todo. Depois gerei um conjunto de dados de treinamento e teste. Em seguida a comparação de Algoritmos de Classificação que chamei de Linha de Base 1ª Interação envolvendo os modelos Logistic Regression, SVC , Kernel SVM, KNN, Gaussian NB, Decision Tree Classifier e Random Forest. E utilizei também a métrica ROC (Receiver Operating Characteristic) para avaliar a qualidade de saída dos classificadores com os maiores escores médios de AUC (grau ou medida de separabilidade). Com essa iteração dos algoritmos de classificação de linha de base, pode-se ver que a Random Forest supera os outros seis modelos para o conjunto de dados escolhido com os maiores escores médios de AUC.

Para reconfirmar o resultado foi feita a segunda iteração, utilizando os modelos de aprendizagem de máquina como Regressão Logística,SVM, KNN, SVM do kernel, Byes ingênuos, Árvore de decisão e Floresta aleatória. Mas antes de executar a segunda iteração foram feitos a otimização dos parâmetros e finalizado as métricas de avaliação para a seleção do modelo, isso para KNN e Random Forest. Como resultado dessa segunda iteração obtermos novamente como melhor resultado o modelo Random Forest.

Após ter escolhido o modelo foi feito a avaliação dele e a conclusão foi que:

A partir desse conjunto de dados de rotatividade de clientes e utilizando o modelo de classificação Random Forest prevê que a propensão de qualquer cliente a desistir é de uma média de precisão de 94%.

Renata Santana