Churn Analysis - KDD

Leonardo Borck da Silveira
Luiz Felipe Cipriani Morfelle
Victor Trindade de Carvalho

Sumário

Dataset

A database, provida por data.world que uma empresa que tem como propósito ser um catálogo de dados corporativos para a pilha de dados moderna, contém dados de dez mil clientes e ex-clientes de um banco, com isso temos como objetivo neste projeto foi desenvolver um modelo de previsão de churn usando algoritmos de aprendizado de máquina.

- .CSV
- 668 KB
- 14 atributos

Atributos - Dataset

- Surname: Sobrenome
- CreditScore: Score de Crédito
- Geography: País (Germany / France / Spain)
- Gender: Gênero (Female / Male)
- Age: Idade
- Tenure: Quantos anos é cliente do banco
- Balance: Saldo da conta
- NumOfProducts: Número de serviços que o cliente possui
- HasCrCard: Se possui cartão de crédito (0 = No, 1 = Yes)
- IsActiveMember: Se é cliente ativo no banco (0 = No, 1 = Yes)
- EstimatedSalary: Estimativa salarial anual
- Exited: Se abandonou o banco (0 = No, 1 = Yes)

Análise Exploratória (EDA)

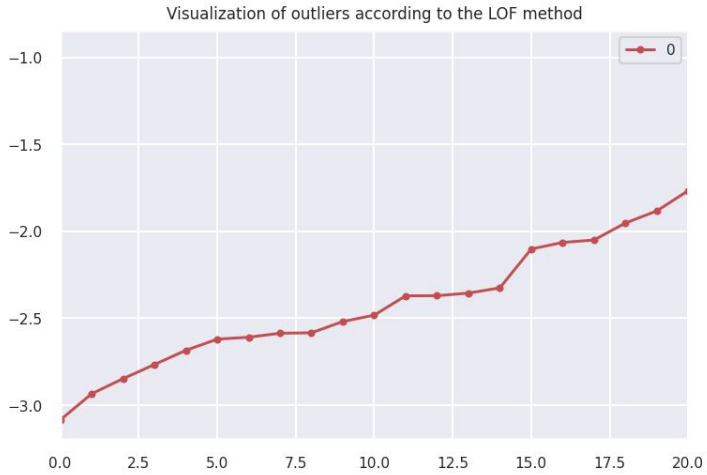
Durante a etapa de EDA (Exploratory Data Analysis) foram realizadas análises a fim de responder aos questionamentos do problema de negócio. Foi levantado algumas perguntas a serem respondidas ao longo do trabalho, sendo elas:

1. Qual a idade das pessoas que mais realizam churn no banco? (São os jovens?)
2. Existe correlação de algum dos serviços oferecidos pelo banco com o índice de churn?
3. Quais são os atributos com maior impacto (peso) na incidência de churn?
4. Podemos utilizar algum algoritmo de machine learning para desenvolver um modelo de predição do churn?
5. O score de crédito do cliente influencia no churn de alguma forma?

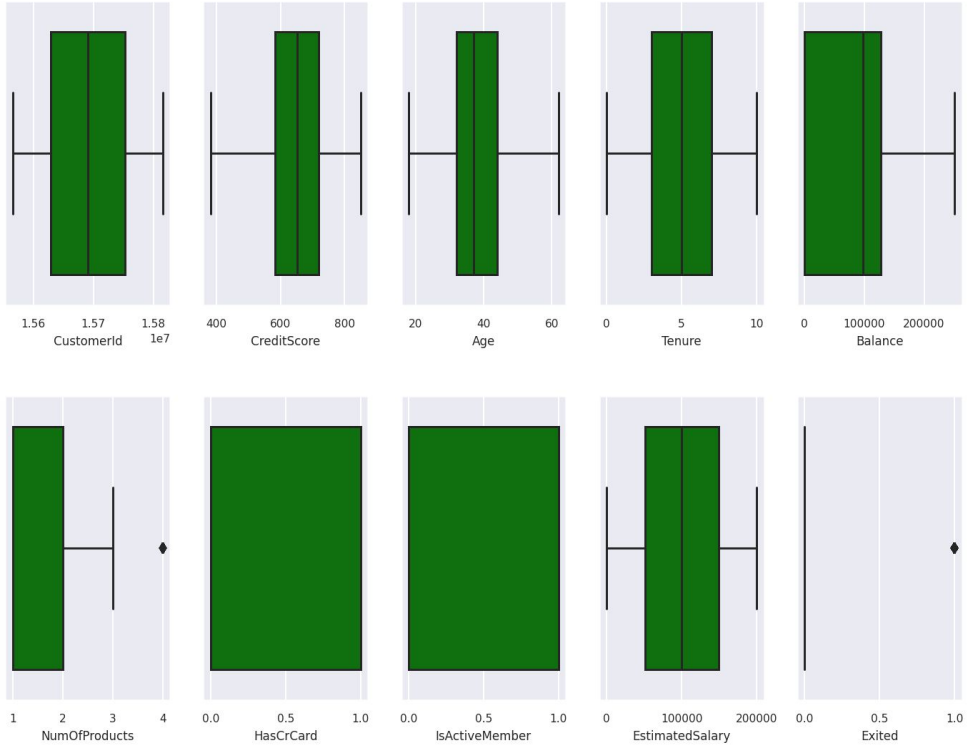
Pré-Processamento

- Análise e supressão de valores Missing
- Análise e supressão de valores Outliers
- Análise dos Dados
- Tabela de correlação
- Data Encoding

Análise Outliers



Observing Outliers



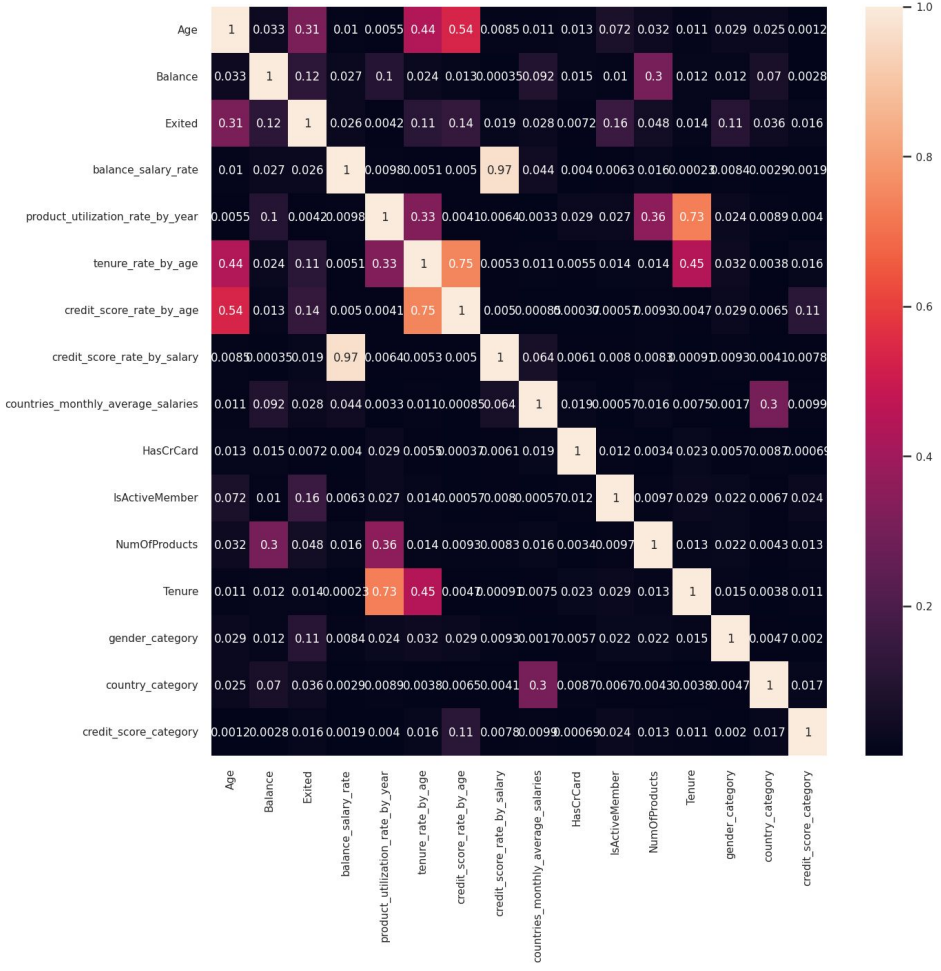
Visualização do balanceamento do dataset

Visualização dos atributos em função da saída

Tabela de Correlação

Gera uma lista dos atributos correlacionados com a saída:

1. Age 0.31
2. IsActiveMember 0.16
3. credit_score_rate_by_age 0.14
4. Balance 0.12
5. tenure_rate_by_age 0.11
6. gender_category 0.11
7. NumOfProducts 0.05
8. country_category 0.04
9. countries_monthly_average_salaries ...
10. balance_salary_rate ...
11. credit_score_rate_by_salary ...
12. credit_score_category ...
13. Tenure ...
14. HasCrCard ...
15. product_utilization_rate_by_year 0.00



Data Encoding

- Transforma valores qualitativos em quantitativos;
- Exemplo:

  Geography:

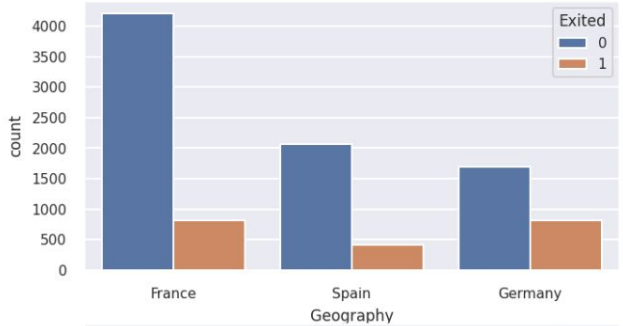
1. France
2. Spain
3. Germany

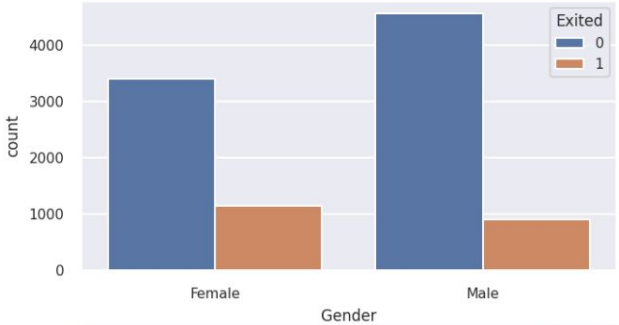
Clusterização

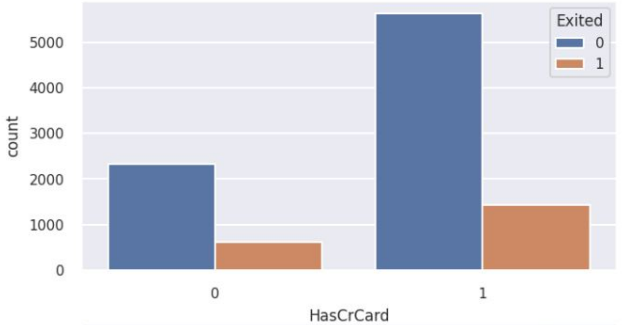
Inicializamos a categorização após analisar os dados das colunas (que achamos que podem ter um agrupamento) e depois olhamos 4 colunas em específicas que poderiam nos levar a novos atributos. Chegamos em alguns pontos:

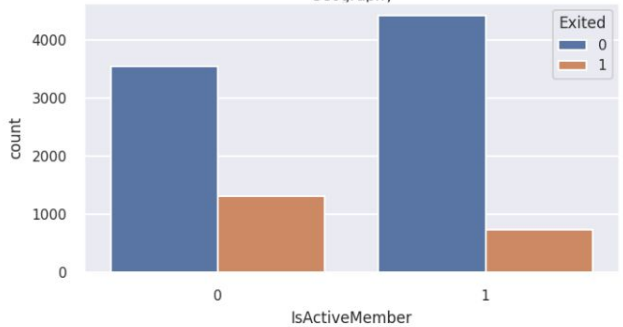
- Grupamento de Credit Score (min= 358 and max= 800)
- Utilização do Produto pela Taxa dependendo do tempo de serviço (Tenure)
- Utilização de produto (taxa) estimada pelo salário estimado
- Exibição do salário médio por país

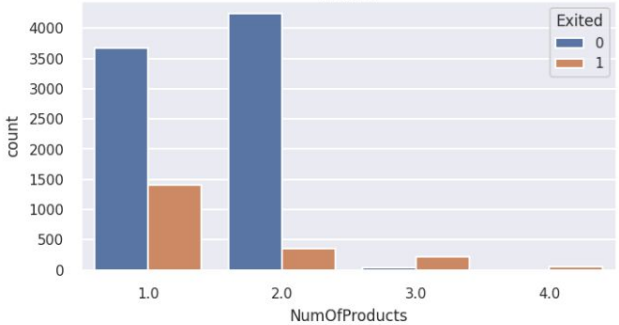
Exibição das ocorrências (Count) pelo output (churn) em cada atributo analisado

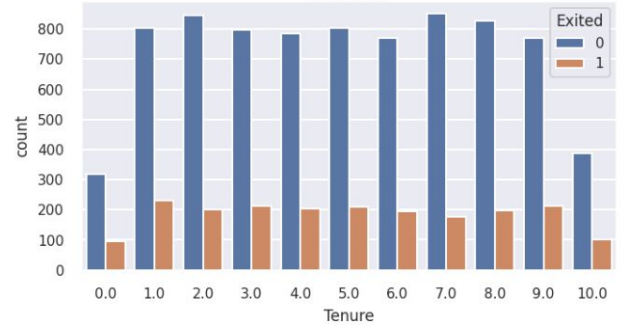












Treinamento (Classificação)

Teste: 20% da base (2 mil casos);

Treinamento: 80% da base (8 mil casos)

Aplicação de 8 algoritmos classificadores buscando gerar o melhor resultado possível.

- Light GBM Classifier : 0.87
- Random Forest Classifier : 0.86
- XGB Classifier : 0.86
- Cat Boost : 0.86
- Gradient Boosting Classifier : 0.86
- KNN : 0.84
- Logistic Regression : 0.81
- CART : 0.79

Resultados

- Análise de Acurácia
- Matriz de confusão
- Feature Importance

Análise de Acurácia

- Melhorando o modelo (tuning)
  - Feito tuning três modelos com melhor desempenho no treinamento;

- Accuracy score of tuned LightGBM model: 0.868
- Accuracy score of tuned Random Forest model: 0.8605
- Accuracy score of Tuned XGBoost Regression: 0.8655

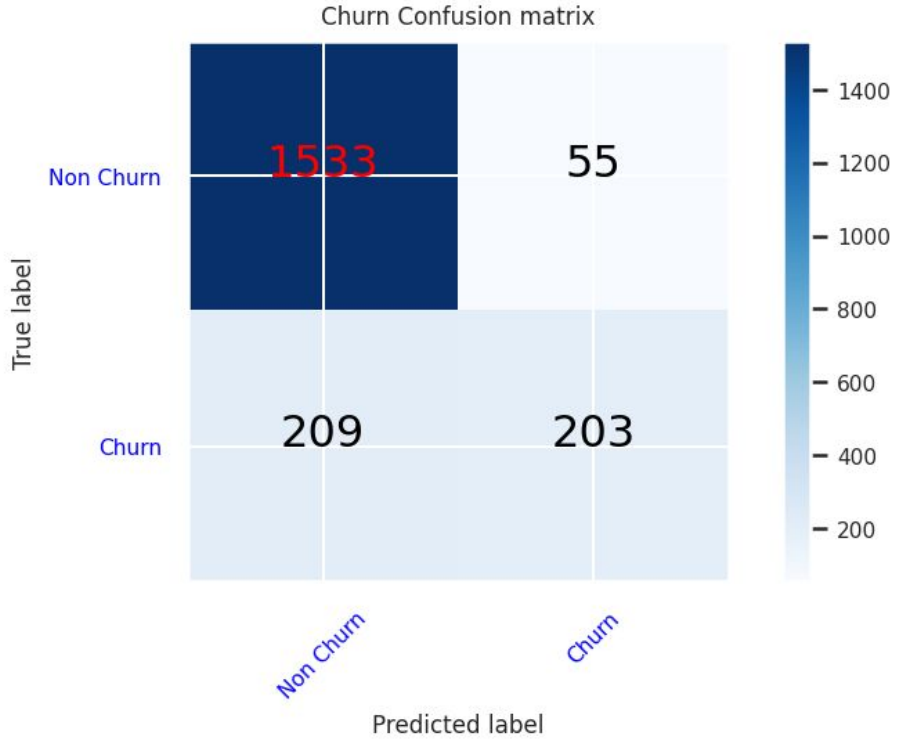
Matriz de Confusão

Negativo Verdadeiro: 1533

Falso Positivo: 55

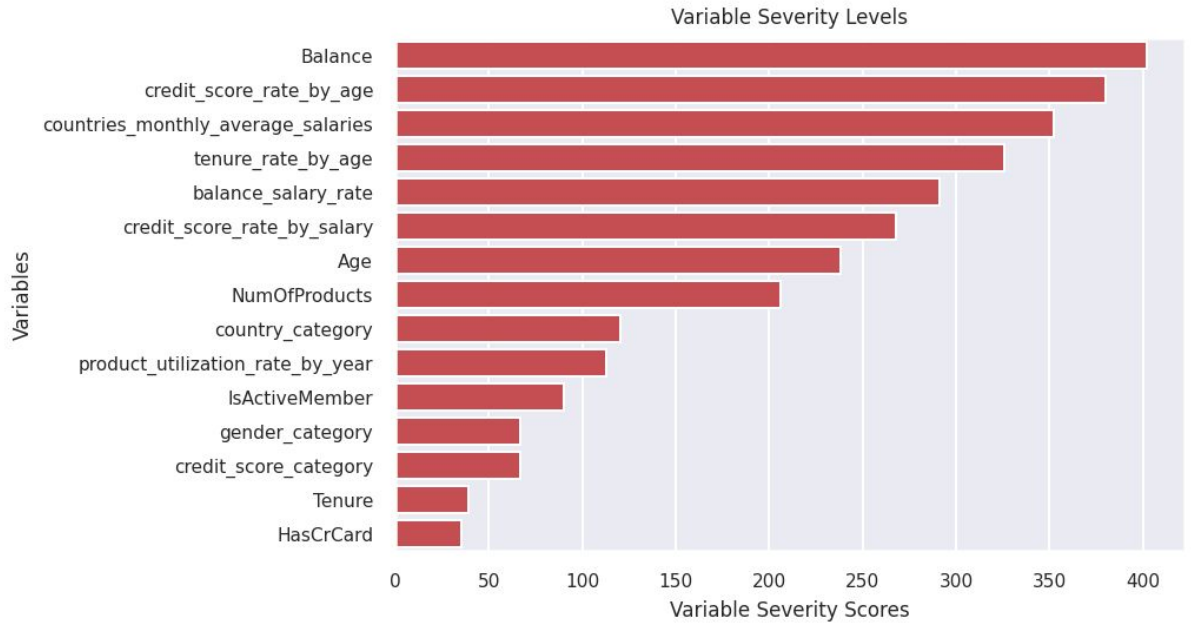
Falso negativo: 209

Positivo Verdadeiro: 203



Feature Importance

Ordena de acordo com a importância (peso) dos atributos



Conclusão

1. Qual a idade das pessoas que mais realizam churn no banco? (São os jovens?)

   R: Clientes com idade entre 40 e 65 anos são mais propícios a se desligar do banco, ou seja, são clientes de meia idade.

2. Existe correlação de algum dos serviços oferecidos pelo banco com o índice de churn?

   R: A maioria dos clientes que usam o produto 3 e 4 pararam de trabalhar com o banco. E além disso, todos os clientes que usaram o produto 4 já se desligaram.

3. Quais são os atributos com maior impacto na incidência de churn?

   R: Utilizando o algoritmo feature importance, definimos os atributos com maior impacto na saída, no slide anterior.

Conclusão

4. Podemos utilizar algum algoritmo de machine learning para desenvolver um modelo de predição do churn?

   R: Sim, Predições foram feitas com um total de 8 modelos de classificadores. O modelo com maior acurácia no nosso caso foi o método de LightGBM.

5. O score de crédito do cliente influencia no churn de alguma forma?

   R: Sim, aqueles com score de crédito abaixo de 450 tinham chances mais altas de sair do banco.

Repositório

https://github.com/leonardoborck/bankChurnPrediction