Projeto: Prevendo o Nível de Satisfação dos Clientes do Santander

Walter Trevisan

Problema de Negócio

A **satisfação do cliente** é uma medida fundamental de sucesso. Clientes insatisfeitos cancelam seus serviços e raramente expressam sua insatisfação antes de sair. Clientes satisfeitos, por outro lado, se tornam defensores da marca!

O Banco Santander está pedindo para ajudá-los a identificar clientes insatisfeitos no início do relacionamento. Isso permitiria que o Santander adotasse medidas proativas para melhorar a felicidade de um cliente antes que seja tarde demais.

Objetivo do Projeto

O objetivo deste projeto é criar um **modelo preditivo** de "***Machine Learning***" que seja capaz de prever o "**Nível de Satisfação dos Clientes do Santander**", ou seja, dado um conjunto de atributos sobre um determinado cliente, o "**modelo preditivo**" deverá prever se o cliente está "**Satisfeito**" ou "**Insatisfeito**".

Avaliação do Projeto

A métrica principal de avaliação que utilizaremos para este projeto é a "***accuracy***" (**acurácia**), com um índice de no mínimo "**90%**". Também desejamos que neste projeto a taxa de "***recall***" (**revocação**) seja no mínimo de "**85%**".

A "**acurácia**" é a porcentagem de clientes "**satisfeitos**" e "**insatisfeitos**" que foram classificados corretamente pelo "modelo preditivo".

A "**revocação**" é a porcentagem de clientes "**insatisfeitos**" que o "modelo preditivo" conseguiu detectar.

Análise Exploratória dos Dados

- A “base de dados” (***dataset***) utilizada possui um total de **60816** observações, com **369** atributos **anônimos** e a variável ***target***, que o “modelo preditivo” deverá prever.
- Identificamos **45** tipos de variáveis preditoras, numeradas de **1** a **46** (**var1**, **var2**, **var3**, ...). O tipo de variável **var23** não existe no *dataset* fornecido.
- Agrupamos cada tipo de variável de acordo com a sua quantidade de atributos relacionados, e com isso identificamos 12 grupos. Por exemplo, grupo de variáveis que possuem **1** atributo, **2** atributos, **18** atributos, e assim por diante;
- Portanto, temos **45 variáveis** distribuídas em **12 grupos** de tamanhos (contagens) diferentes de atributos.

Análise Exploratória dos Dados

Analisando as **correlações** entre os atributos e as suas medidas **estatísticas descritivas**, classificamos todos os atributos da seguinte forma:

- Temos **88** atributos do tipo **categórico**;
- Temos **134** atributos do tipo **numérico** (quantitativo discreto);
- Temos **147** atributos do tipo **numérico** (quantitativo contínuo).

Portanto, em nosso *dataset*, a maior parte dos atributos são **numéricos** (**76%**).

Também identificamos e concluímos que apenas **23** atributos são **relevantes** em nosso *dataset* para serem analisados na fase de "***machine learning***", onde treinamos e testamos os "**modelos preditivos**".

Análise Exploratória dos Dados

Os **23** atributos **relevantes** em nosso dataset foram classificados da seguinte forma:

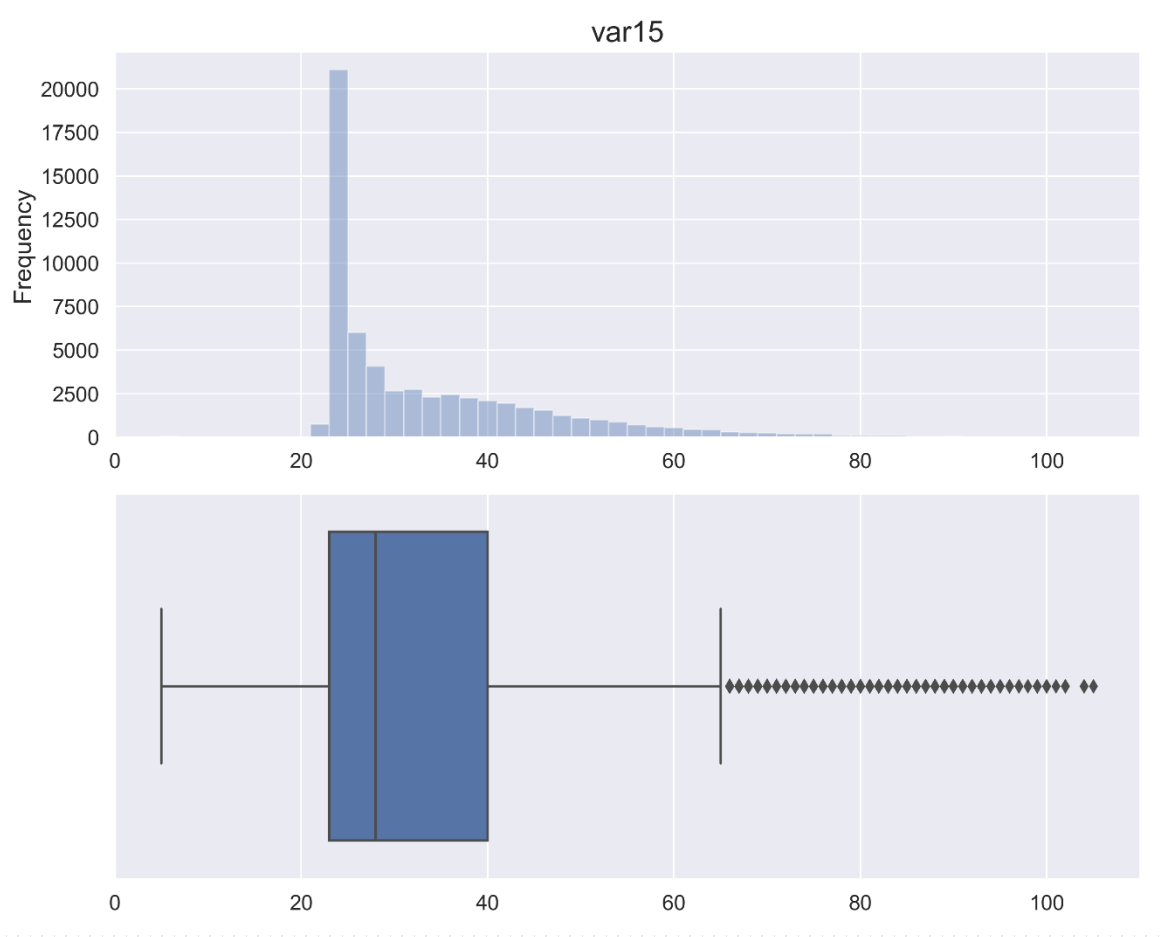
- Temos **05** atributos do tipo **categórico**;
- Temos **10** atributos do tipo **numérico** (quantitativo discreto);
- Temos **8** atributos do tipo **numérico** (quantitativo contínuo).

As atributos "**var15**" e "**var38**" são os **previsores mais importantes** para a classificação do nível de satisfação dos clientes do Santander.

Também identificamos outros dois previsores importantes: "**num_var4**" e "**var36**".

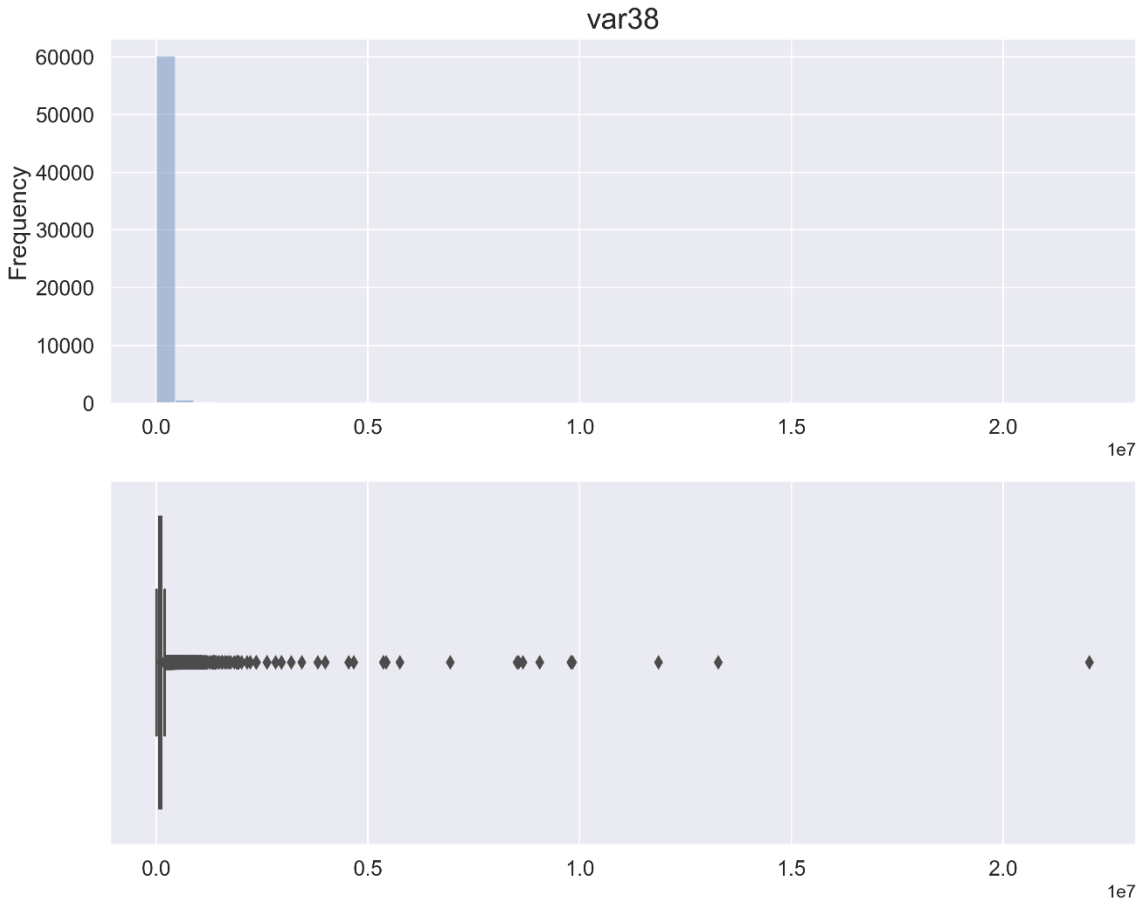
Análise Exploratória dos Dados: "**var15**"

- Podemos observar no **histograma** (1º gráfico), que a sua distribuição de frequência está concentrada entre os valores **23 a 40**. Temos um pico de frequências no valor **23;**
- Também identificamos uma pequena variabilidade nos dados ocasionada pelos **outliers**, conforme podemos observar no **boxplot** (2º gráfico). Os outliers estão concentrados nos valores acima de **65**.



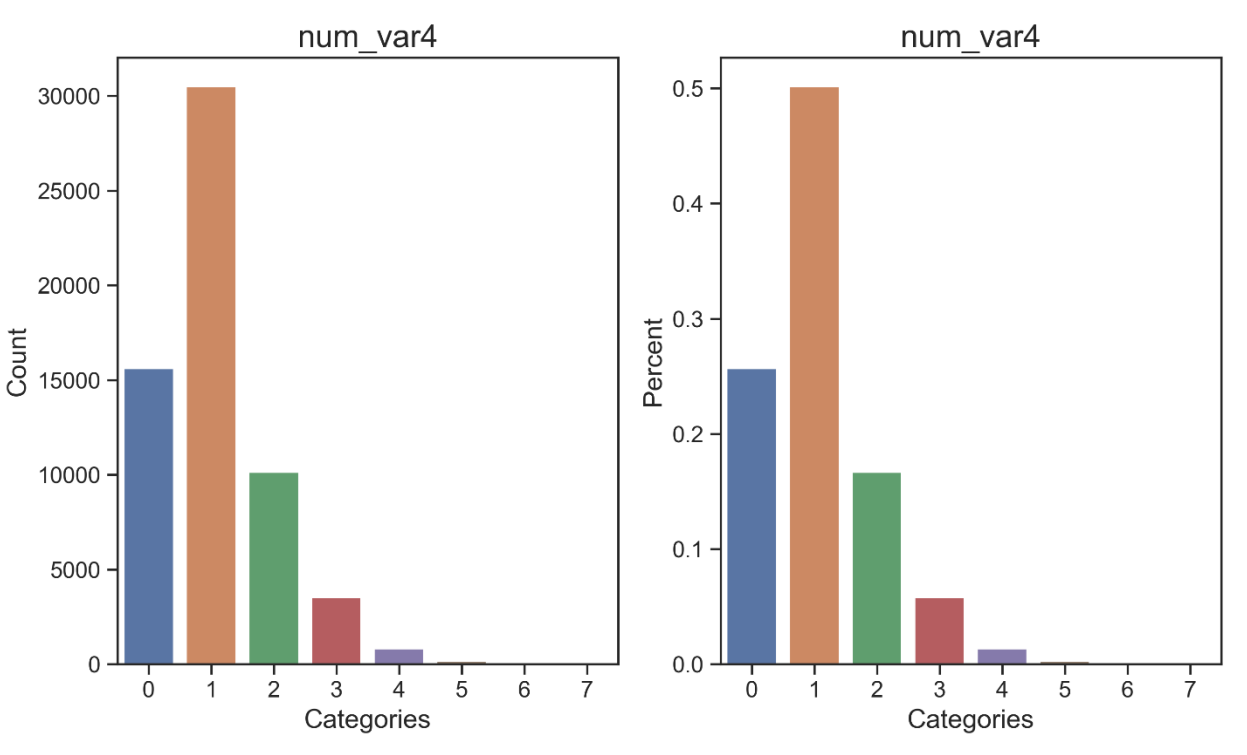
Análise Exploratória dos Dados: "**var38**"

- Observamos que existe uma alta variabilidade nos dados ocasionada pelos **outliers** (**Box Plot**). Os outliers estão concentrados nos valores acima de **500 mil**;
- A alta variabilidade nos dados também pode ser observada em seu **histograma** (1º gráfico).



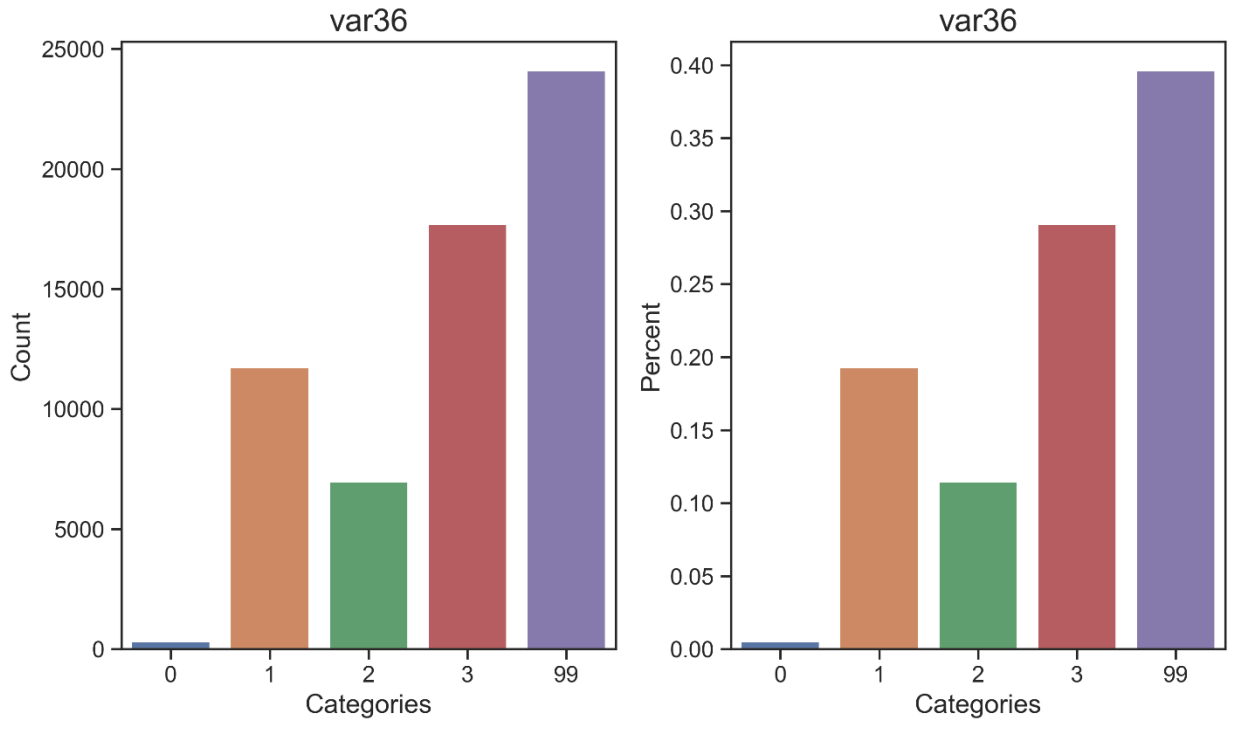
Análise Exploratória dos Dados: "**num_var4**"

- Podemos notar que **50%** das observações do nosso *dataset* foram classificados na **categoria** "**1**";
- As **categorias** "**5**", "**6**" e "**7**" praticamente não possuem proporções significativas no nosso *dataset*.



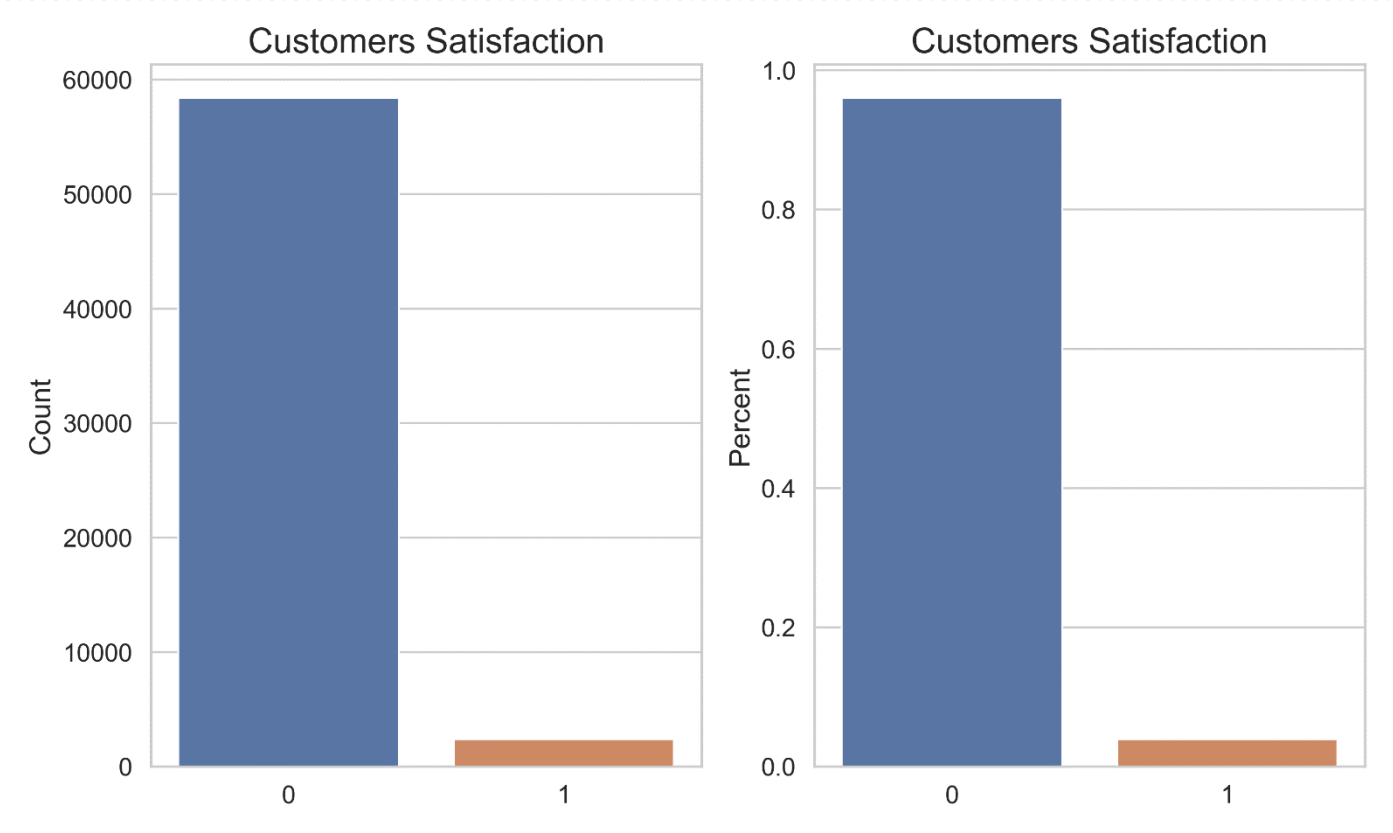
Análise Exploratória dos Dados: "**var36**"

- Podemos notar que **40%** das observações do nosso *dataset* foram classificados na **categoria** "**99**";
- Apenas a **categoria** "**0**" não têm proporções significativas no nosso *dataset*.



Análise Exploratória dos Dados: "*target*"

- A nossa variável ***target*** é do tipo **categórica nominal binária**, ou seja, possui apenas duas classes: "**0**" para clientes **satisfeitos** e "**1**" para clientes **insatisfeitos**;
- Podemos observar, que **96%** dos clientes estão "**satisfeitos**" com a sua experiência bancária e apenas **4%** dos clientes estão "**insatisfeitos**" com a sua experiência bancária, ou seja, o nosso *dataset* **não está balanceado** (proporções iguais ou próximas de cada "classe").



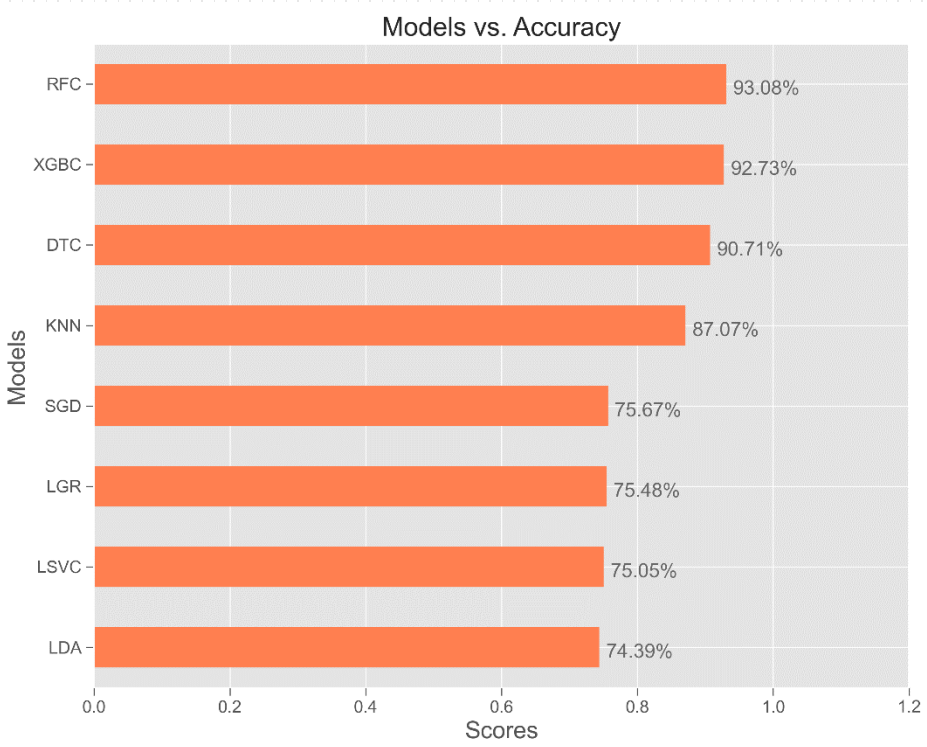
Machine Learning: treinamento

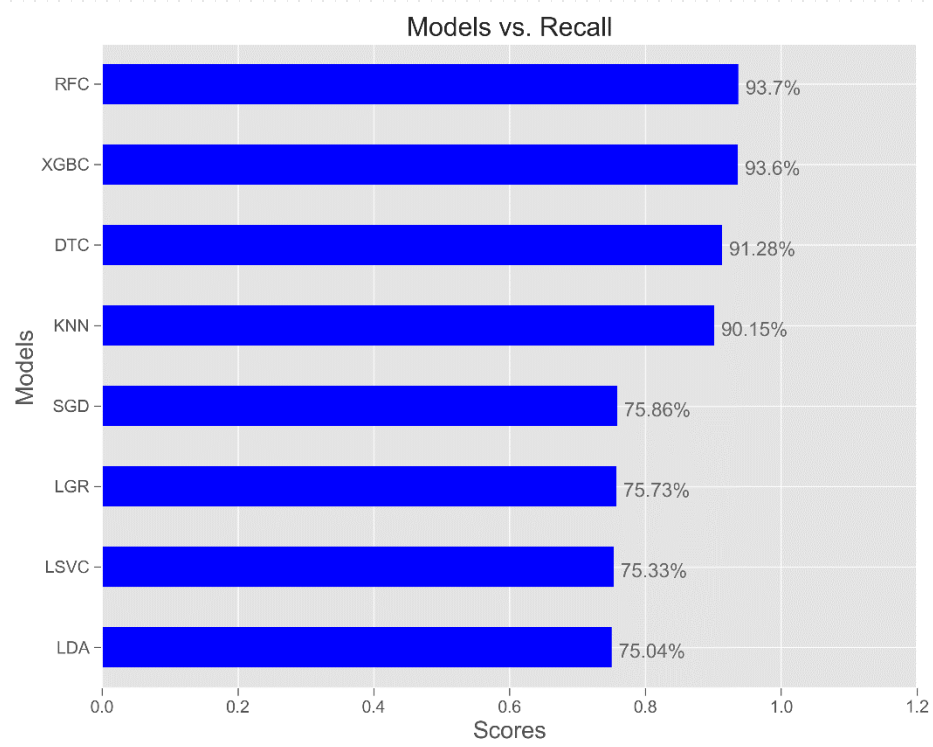
Escolhemos "**8**" algoritmos para serem treinados:

- **SGD**: Stochastic Gradient Descent;
- **KNN**: K Nearest Neighbors;
- **LGR**: Logistic Regression;
- **DTC**: Decision Tree Classifier;
- **RFC**: Random Forest Classifier;
- **LDA**: Linear Discriminant Analysis;
- **LSVC**: Linear Support Vector Classification;
- **XGBC**: XGBoost Classifier.

Machine Learning: treinamento

Podemos observar que **04** algoritmos se destacaram e obtiveram as melhores pontuações (**Scores**) nas métricas de acurácia (***Accuracy***) e revocação (***Recall***):

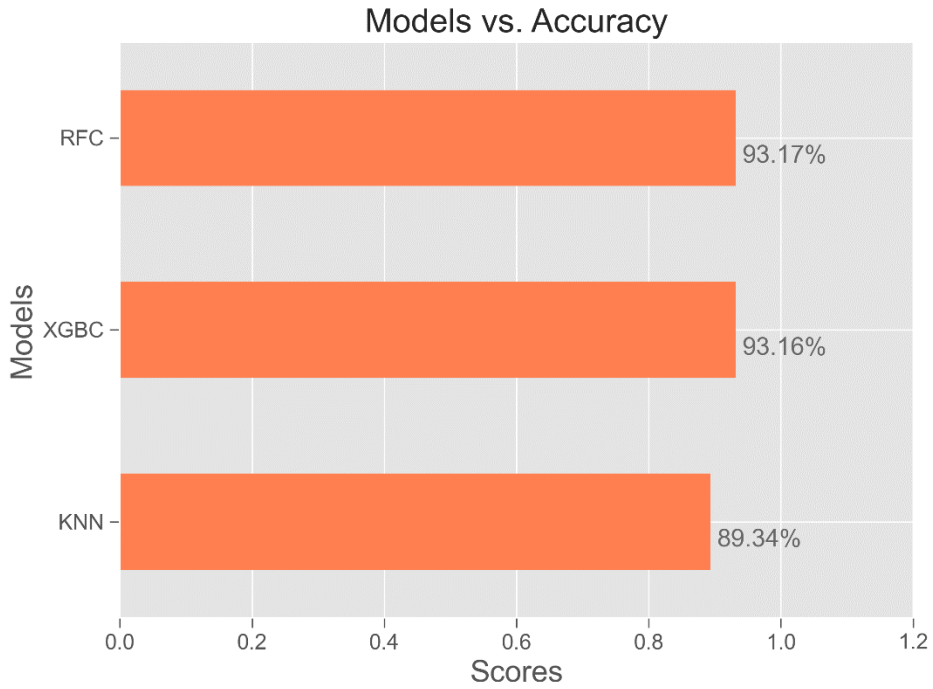


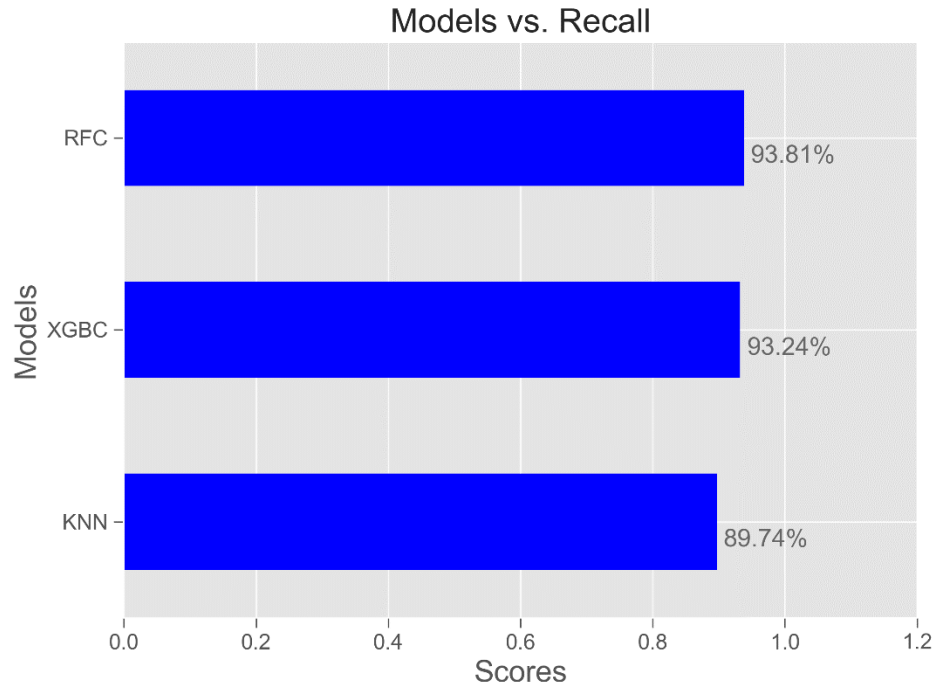


Machine Learning: ajuste fino

Nesta etapa selecionamos os melhores "**hiperparâmetros**" de cada algoritmo que obteve os melhores **scores** na etapa de treinamento. Conseguimos melhorar as métricas de acurácia (***Accuracy***) e revocação (***Recall***).

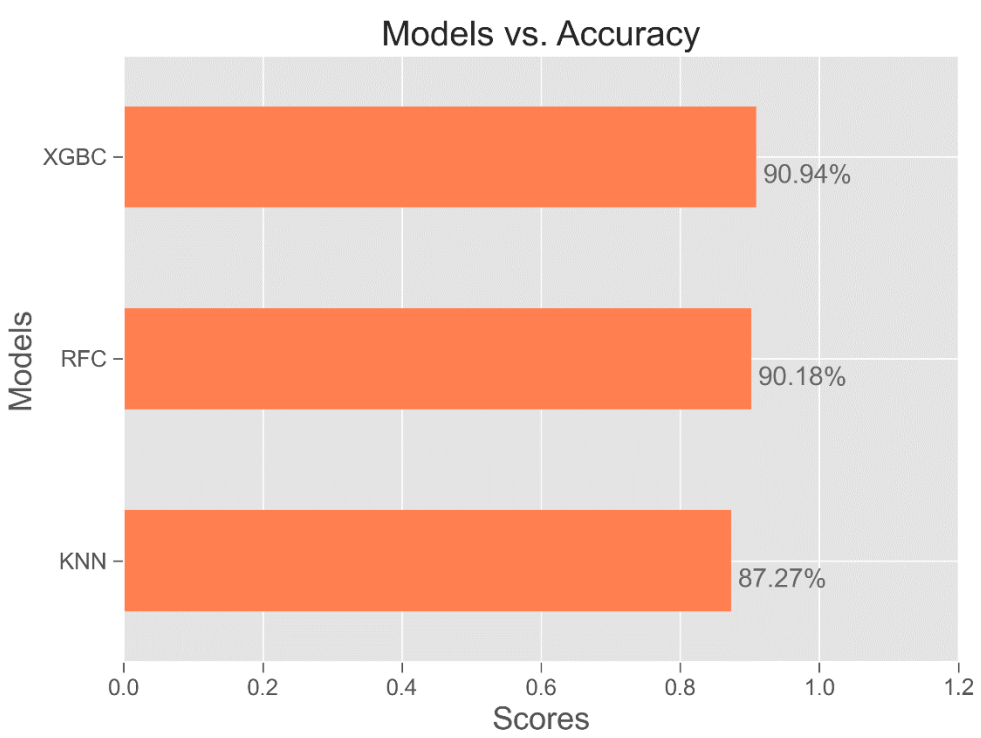
Nota: embora o algoritmo **DTC** tenha apresentado ótimas pontuações (**Score**) na etapa de treinamento, ele não foi selecionado para esta etapa, porque o algoritmo **RFC** é um método **ensemble** mais poderoso que o DTC e apresentou uma performance melhor.

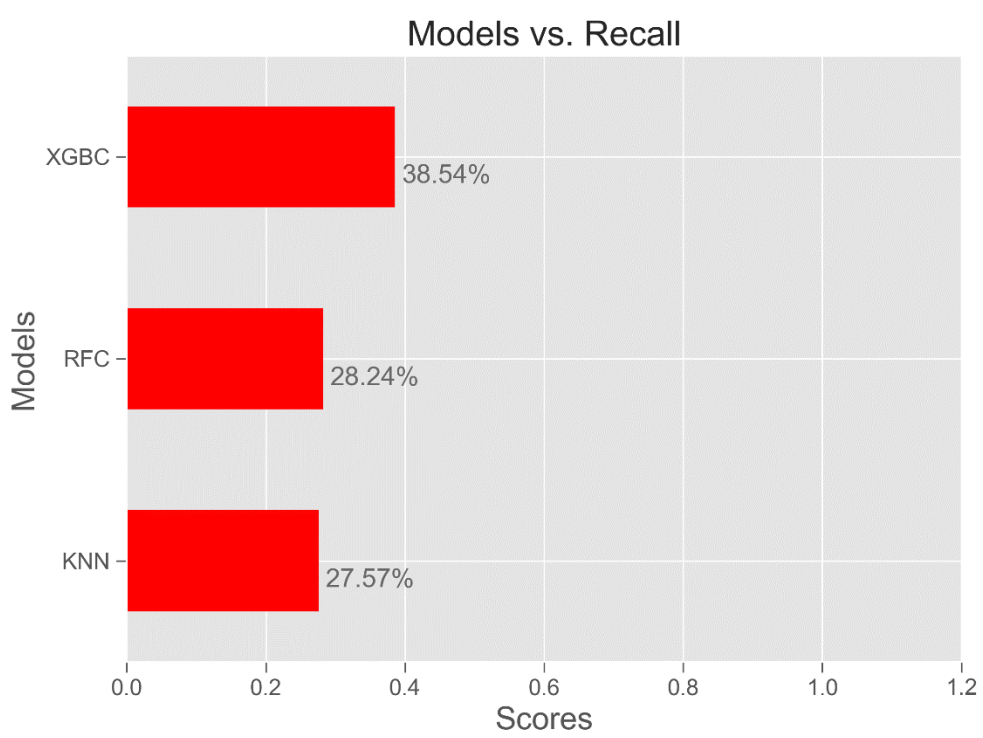




Machine Learning: testes

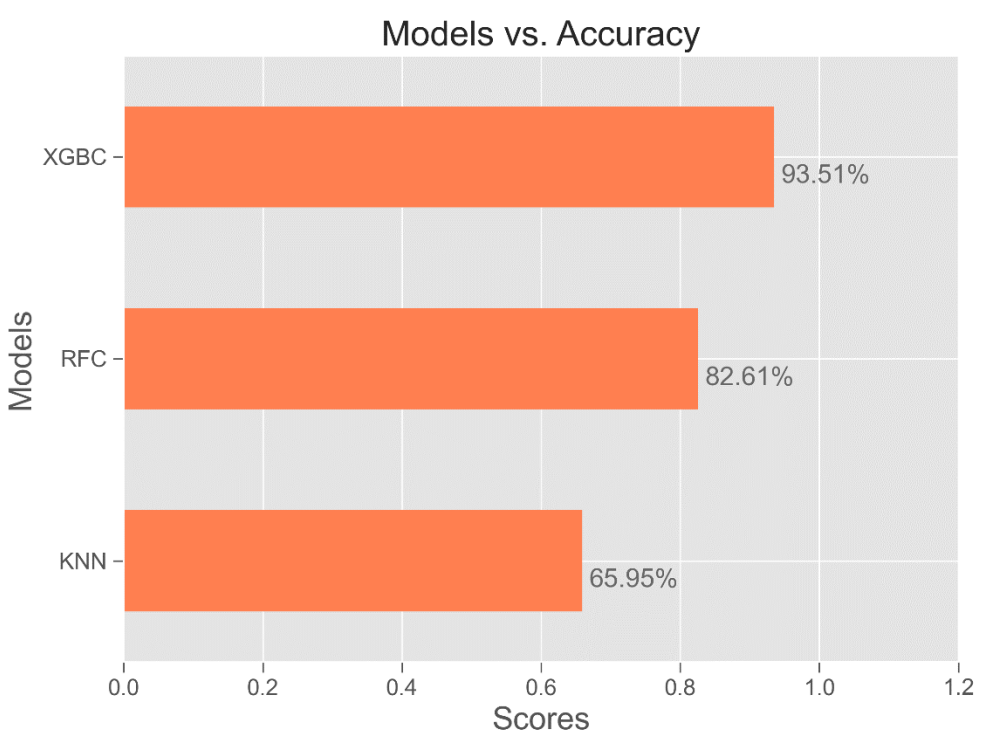
Nesta etapa, realizamos os testes em um *dataset* com apenas **4%** dos clientes classificados como "**Insatisfeitos**". Apenas dois modelos preditivos apresentaram resultados satisfatórios na acurácia (***Accuracy***):

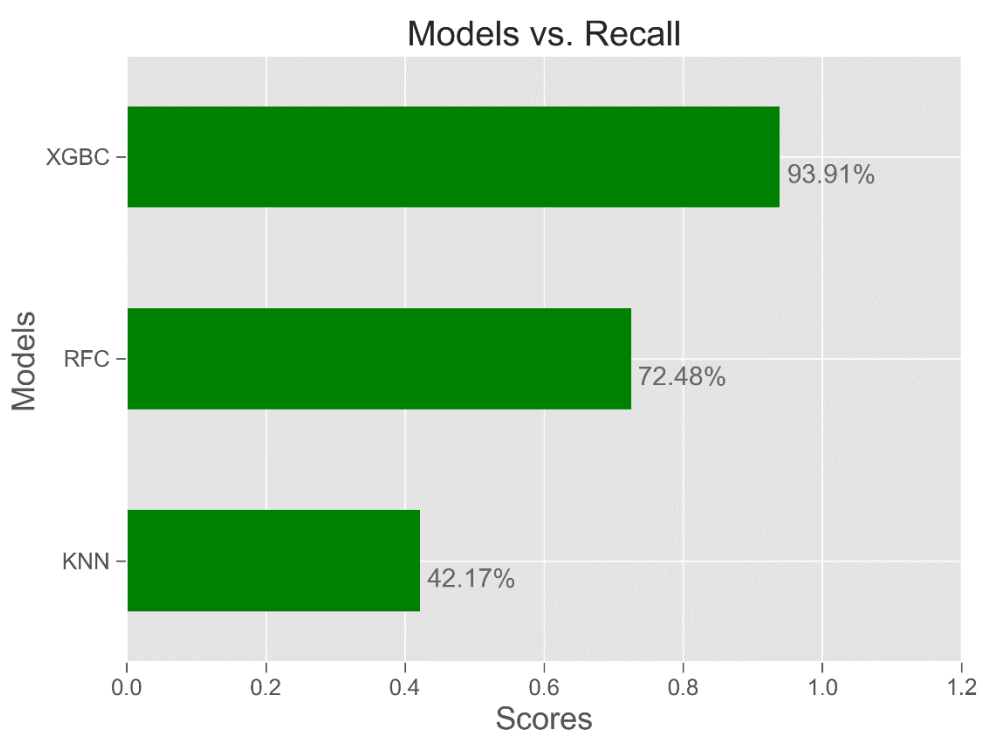




Machine Learning: testes

Nesta etapa, realizamos os testes em um *dataset* com **50%** dos clientes classificados como "**Insatisfeitos**". Apenas o modelo preditivo "**XGBC**" apresentou resultados satisfatórios para as duas métricas:





Conclusão

- ✓ ***Accuracy***: de acordo com os objetivos do projeto (***Accuracy* >= 90%**), apenas o modelo **XGBC** (***XGBoost Classifier***) apresentou um resultado satisfatório, obtendo uma **taxa de acurácia excelente** (**93.51%**), **acima do mínimo desejado**;
- ✓ ***Recall***: de acordo com os objetivos do projeto (***Recall* >= 85%**), podemos apenas o modelo **XGBC** (***XGBoost Classifier***) apresentou um resultado satisfatório, obtendo uma **taxa de revocação excelente** (**93.91%**), **muito acima do mínimo desejado**.
- ✓ Importante ressaltar que o modelo preditivo **RFC** (**Random Forest Classifier**) não conseguiu obter resultados satisfatórios quando aumentamos muito o número de observações de clientes classificados como "**Insatisfeitos**", ou seja, o modelo **RFC** não conseguiu **generalizar** para os clientes classificados como "**Insatisfeitos**".

Portanto, recomendamos para este projeto o modelo preditivo:

XGBoost Classifier