



DSA – PROJECT – 03

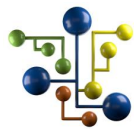
Data Science e Big Data



**Data Science
Academy**

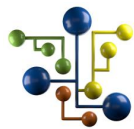
21 DE MAIO DE 2020

WALTER TREVISAN
São Paulo - SP



Sumário

1	Objetivo	2
2	Avaliação	2
2.1	Métrica de Avaliação	2
2.2	Ferramenta	2
3	Problema de Negócio	2
4	Dataset	3
4.1	Descrição do Dataset	3
4.2	Descrição do Arquivo	3
4.3	Descrição das Variáveis Predictoras (Independentes)	3
4.4	Descrição da Variável Target (Dependente)	3
5	Solução	3
6	Referência	6



1 Objetivo

O objetivo deste projeto é criar um modelo de “*Machine Learning*” que seja capaz de prever o “Nível de Satisfação dos Clientes do Santander”.

2 Avaliação

2.1 Métrica de Avaliação

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

2.2 Ferramenta

Neste projeto utilizaremos a “**Linguagem Python**” para o desenvolvimento da solução.

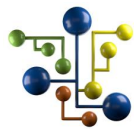
3 Problema de Negócio

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.

Neste projeto de aprendizado de máquina, você trabalhará com centenas de recursos anônimos para prever se um cliente está satisfeito ou insatisfeito com sua experiência bancária.





4 Dataset


Fonte dos dados:

<https://www.kaggle.com/c/santander-customer-satisfaction>

4.1 Descrição do Dataset

Foi fornecido um conjunto de dados anonimizados contendo um grande número de variáveis numéricas. A coluna "**TARGET**" é a variável a ser prevista. É igual a "1" para **clientes insatisfeitos** e a "0" para **clientes satisfeitos**.

4.2 Descrição do Arquivo

 **train.csv**: é o conjunto de dados de treinamento, incluindo a variável "*target*".

4.3 Descrição das Variáveis Predictoras (Independentes)

Temos um total de "369" variáveis predictoras "**anônimas**" em nosso dataset.

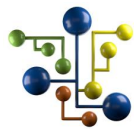
4.4 Descrição da Variável Target (Dependente)

✓ **TARGET**: "0" – para o **cliente satisfeito** / 1 – para o **cliente insatisfeito**).

5 Solução

A solução do projeto foi desenvolvida em "**06 fases**" (explicadas abaixo) sendo executada na seguinte "**estrutura de diretórios (pastas)**":

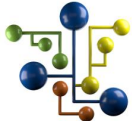
- ❖ **BusinessProblem**: contém a descrição do projeto (este documento);
- ❖ **Datasets**: contém o conjunto de dados do projeto;
- ❖ **Solution**: contém todos os programas (códigos fontes – arquivos "**ipynb**") de cada "**fase**" do projeto, a saber:
 - ✓ **01-Business-Problem.ipynb**: contém uma explicação do "**problema de negócio**" a ser resolvido, assim como os objetivos a serem alcançados para o projeto;



- ✓ **02-Get-The-Data.ipynb:** contém as atividades relacionadas a **leitura e preparação dos dados** que serão utilizados no projeto;
- ✓ **03-Explore-The-Data.ipynb:** contém todas as atividades relacionadas a **análise exploratória dos dados**;
- ✓ **04-Preprocessing.ipynb:** contém todas as atividades relacionadas ao **pré-processamento dos dados**, ou seja, limpeza e transformação ("*Feature Engineering*") dos dados;
- ✓ **05-Machine-Learning.ipynb:** contém todas as atividades relacionadas ao treinamento dos **algoritmos de classificação** que foram utilizados no projeto;
- ✓ **06-Tests.ipynb:** contém todas as atividades relacionadas aos testes de validação dos **algoritmos de classificação selecionados** no projeto.

Nota: os programas devem ser executados nesta ordem:

- 01-Business-Problem.ipynb;
- 02-Get-The-Data.ipynb;
- 03-Explore-The-Data-Step-01.ipynb
- 03-Explore-The-Data-Step-02.ipynb
- 03-Explore-The-Data-Step-03.ipynb
- 03-Explore-The-Data-Step-04.ipynb
- 03-Explore-The-Data-Step-05.ipynb
- 04-Preprocessing.ipynb;
- 05-Machine-Learning-Step-01.ipynb;
- 05-Machine-Learning-Step-02.ipynb;
- 05-Machine-Learning-Step-03.ipynb;
- 06-Tests-Step-01.ipynb;
- 06-Tests-Step-02.ipynb;
- 06-Tests-Step-03.ipynb.

 <p>Data Science Academy</p>	<p align="center">Project – 03</p> <p align="center">Prevendo o Nível de Satisfação dos Clientes do Santander</p>
---	---

Temos ainda outros arquivos armazenados nos seguintes diretórios (pastas):

- ❖ **Data:** contém todos os dados (arquivos “**csv**”) e objetos (arquivos “**pickle**”) criados na execução do projeto;
- ❖ **Images:** contém todos os gráficos/visualizações do projeto, separados por “**fase**”, além de algumas imagens utilizadas no projeto;
- ❖ **Library:** contém todas as bibliotecas (códigos fontes em python) utilizadas nas “**fases**” do projeto;
- ❖ **Models:** contém todos os objetos (arquivos “**pickle**”) criados na fase “**05-Machine-Learning**” do projeto;
- ❖ **Objects:** contém outros objetos (arquivos “**pickle**”) criados na execução do projeto;
- ❖ **Tests:** contém todos os objetos (arquivos “**pickle**”) criados na fase “**06-Tests**” do projeto.
- ❖ **StoryTelling:** contém uma apresentação do projeto, explicando os resultados e insights encontrados, assim como as estratégias utilizadas no desenvolvimento do projeto.

Fase 1: Business Problem

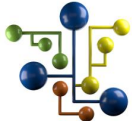
Nesta fase temos uma explicação do “**problema de negócio**” a ser resolvido, o(s) **objetivo(s)** a serem alcançados para o projeto, e de que forma o projeto será avaliado, ou seja, quais serão as **métricas de avaliação** do projeto.

Fase 2: Get The Data

Nesta fase iremos **obter** e **preparar** os **dados** que serão utilizados durante o desenvolvimento e execução do projeto.

Fase 3: Explore The Data

Nesta fase realizaremos algumas atividades com o objetivo de fazer a “**análise exploratória**” dos dados, ou seja, nesta fase estudaremos cada atributo e suas características, criaremos visualizações dos dados, estudaremos as correlações entre os atributos e identificaremos transformações promissoras que

 <p>Data Science Academy</p>	<p align="center">Project – 03</p> <p align="center">Prevendo o Nível de Satisfação dos Clientes do Santander</p>
---	---

poderão ser aplicadas. Também identificaremos atributos (variáveis) **relevantes** e **irrelevantes** para alcançarmos o objetivo do projeto.

Fase 4: Preprocessing

Nesta fase temos todas as atividades relacionadas ao **pré-processamento dos dados**, ou seja, limpeza e transformação (*“Feature Engineering”*) dos dados.

Fase 5: Machine Learning

Nesta fase temos todas as atividades relacionadas a preparação e treinamento dos **algoritmos de classificação** que foram utilizados no projeto. Nesta fase os melhores algoritmos serão selecionados para a próxima fase (testes e avaliação final).

Fase 6: Tests

Nesta fase temos todas as atividades relacionadas aos testes de validação dos **algoritmos de classificação selecionados** na fase anterior. Nesta fase definiremos o **melhor modelo preditivo** que resolve o problema de negócios apresentado, de acordo com as métricas de avaliação definidas.

Fase Final: Story Telling

Nesta fase faremos uma apresentação da solução do projeto, explicando os resultados e insights encontrados, assim como as estratégias utilizadas no desenvolvimento do projeto.

6 Referência

Projeto da *“Formação Cientista de Dados”* da *“Data Science Academy”*.