Big Data e Computação em Nuvem

Projeto Final

Prof. Michel Fornaciali , PhD
Prof. Afonso Menegola , MSc

Contatos:
MichelSF@insper.edu.br
AfonsoM@insper.edu.br

Insper

Grupos

Turma 6

G1:

- Helena Funari
- Ives Fontes
- Felipe Tufaile

G2:

- Rodrigo Zamengo
- Wagner Neustaedter
- Douglas Batista
- Vinicius Camargo

G3:

- Diego Coelho
- Raul Aguiar
- Thais Ono

G4:

- Caio Cabral
- Pedro Leite
- Fernando Lima

~~G5:~~

- ~~Fernando Augusto~~
- ~~Carlos Neto~~

O problema

Detecção de atrasos em voos

- A aviação é em uma das maiores indústrias em receita em 2020. De acordo com um relatório da Forbes, se a aviação fosse um país, teria sido o 20º maior do mundo em PIB.

- Embora a indústria da aviação cresça rapidamente a cada ano, as perdas incorridas ainda são altas. Um dos maiores causas das perdas são os atrasos e cancelamentos ocorridos a cada hora.

- Qualquer pequeno ou grande atraso ou cancelamento de voo resulta na perda de milhares a milhões de dólares em receitas anuais para os aeroportos e também para as companhias aéreas.

- Sua missão é detectar se um determinado voo tem potencial para se atrasar ou não.

O dataset

Detecção de atrasos em voos

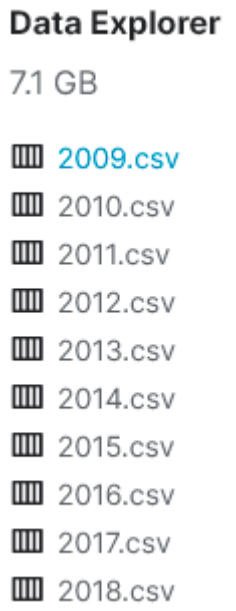
- Este dataset contém dados plurianuais de 2009 a 2018.
- O conjunto de dados possui quase 7 GB, com quase 68 milhões de linhas.
- Fonte: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018
- **Observação**:
  - não se preocupem em baixá-los, os dados já estão na AWS: **../10_dados/projeto**

Insper

O dataset – Dicionário de dados (exemplo)*

Detecção de atrasos em voos

Name	Description	Type(Format)	Example
FL_DATE	Date of the flight	DATE (yy/mm/dd)	02/05/2009
OP_CARRIER	Airline Identifier	STRING	9E
OP_CARRIER_FL_NUM	Flight Number	INTEGER	2216
ORIGIN	Starting Airport Code (IATA Code)	STRING	MLI
DEST	Destination Airport Code (IATA Code)	STRING	MEM
CRS_DEP_TIME	Planned Departure Time	INTEGER	600
DEP_TIME	Actual Departure Time	FLOAT	603.0
DEP_DELAY	Total Delay on Departure in minutes	FLOAT	3.0
TAXI_OUT	The time duration elapsed between departure from the origin airport gate and wheels off	FLOAT	14.0
WHEELS_OFF	The time point that the aircraft's wheels leave the ground	FLOAT	617.0
WHEELS_ON	The time point that the aircraft's wheels touch on the ground	FLOAT	757.0



Os dados estão disponibilizados por ano

*Veja a lista completa na planilha anexada

Insper

O desafio – Diretrizes gerais

Detecção de atrasos em voos

- **Trabalho em grupo** com até 3 participantes

- **Entrega**: apresentação final no dia 25/junho
  - Entrega via BB, na data da apresentação
    - Uma entrega por grupo
    - Indicar os participantes do grupo
    - Subam um ZIP do(s) código(s)

- **Entregável**: notebook com o processamento end-to-end, incluindo células markdown para explicações gerais e registro de análises mais profundas

- **Uso do Spark!**

O desafio – Critérios de avaliação

Detecção de atrasos em voos

Machine Learning end-to-end no Spark:

utilização do Spark desde a leitura dos dados até a modelagem, passando por todos os tratamentos pertinentes.

Utilização do Spark e boas práticas de programação:

utilização adequada do Spark, implementada corretamente com ferramentas pertinentes. Por exemplo, a utilização prematura do Pandas será considerado um redutor da nota, assim como a subutilização das funções vistas em sala de aula.

Robustez e criatividade:

considera a robustez do trabalho final (o modelo faz sentido?), bem como a criatividade na resolução do problema proposto (como utilizar os dados?). Importante: “simples > complexo”, mas “simples != simplório”

Nota geral e apresentação:

propôs uma solução cuja implantação faça sentido para o negócio? Tomou decisões baseadas em dados (tabelas? Gráficos? Métricas?) Fez uma apresentação clara da proposta?

Machine Learning end-to-end no Spark?

Detecção de atrasos em voos

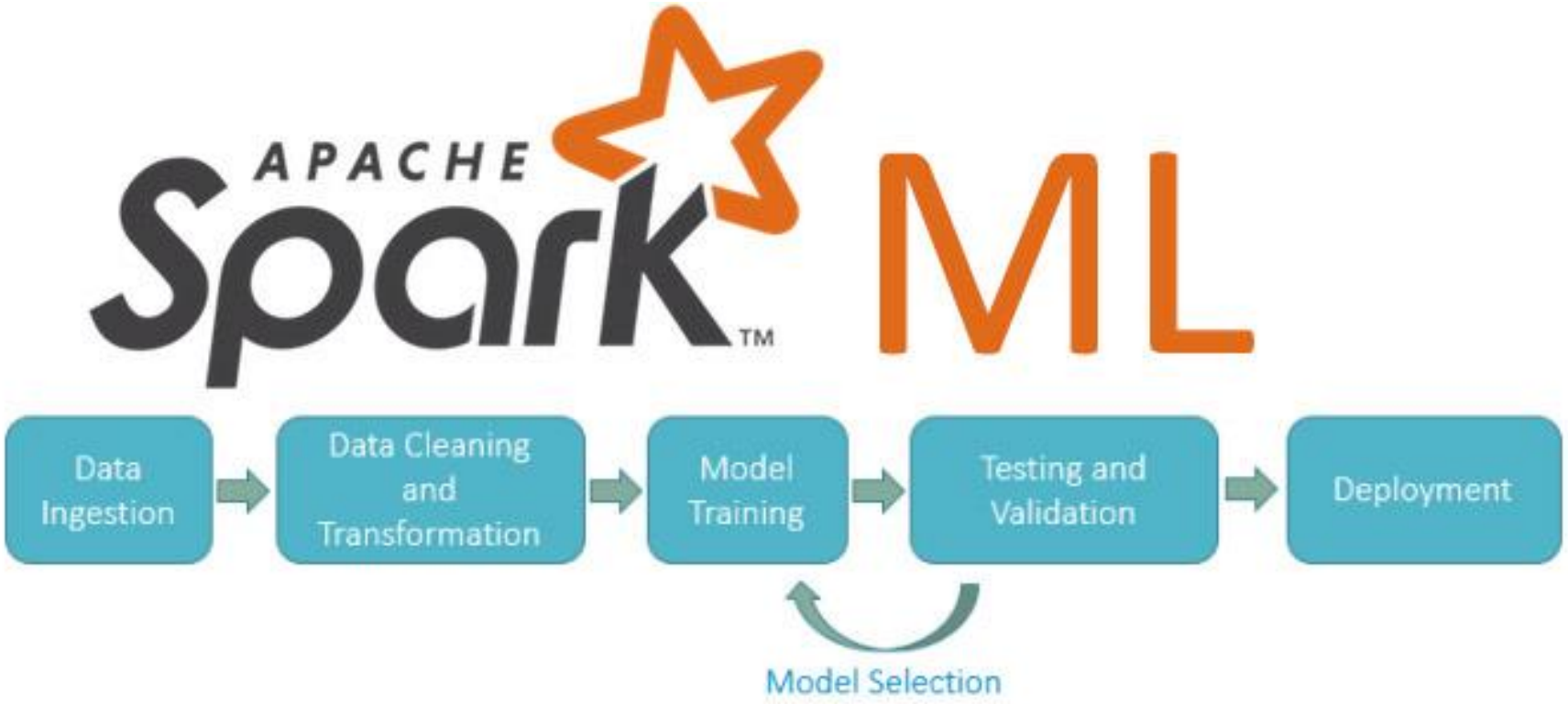


Imagem reproduzida de: https://yurongfan.files.wordpress.com/2017/01/sparkml.jpg

Insper

O desafio inicial

Detecção de atrasos em voos

- **Objetivos do dia**:
  - Entender o significado dos dados
  - Se familiarizar com o tratamento os dados (~NYC Taxi)
  - Pensar em features para o modelo
  - Criar o label

- **Importante**: para desenvolver, usem apenas 1 arquivo do dataset por grupo

- **Dica**: em tempo de desenvolvimento, *limit()* pode ser seu amigo

Insper