



Computação em Larga Escala

Projeto Final

Prof. André Filipe de Moraes Batista, PhD.
Prof. Michel Fornaciali, PhD.

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

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:** os dados serão disponibilizados na AWS, não se preocupem em baixá-los!

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

Data Explorer

7.1 GB

 [2009.csv](#)

 2010.csv

 2011.csv

 2012.csv

 2013.csv

 2014.csv

 2015.csv

 2016.csv

 2017.csv

 2018.csv

*Os dados estão
disponibilizados
por ano*

*Veja a lista completa na planilha anexada

O desafio – Diretrizes gerais

Detecção de atrasos em voos

- **Trabalho em grupo** com até 3 participantes
- **Entrega:** apresentação final no dia 26/junho, à tarde
 - Entrega via AWS, na data da apresentação
 - Uma entrega por grupo
 - Indicar os participantes do grupo
- **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

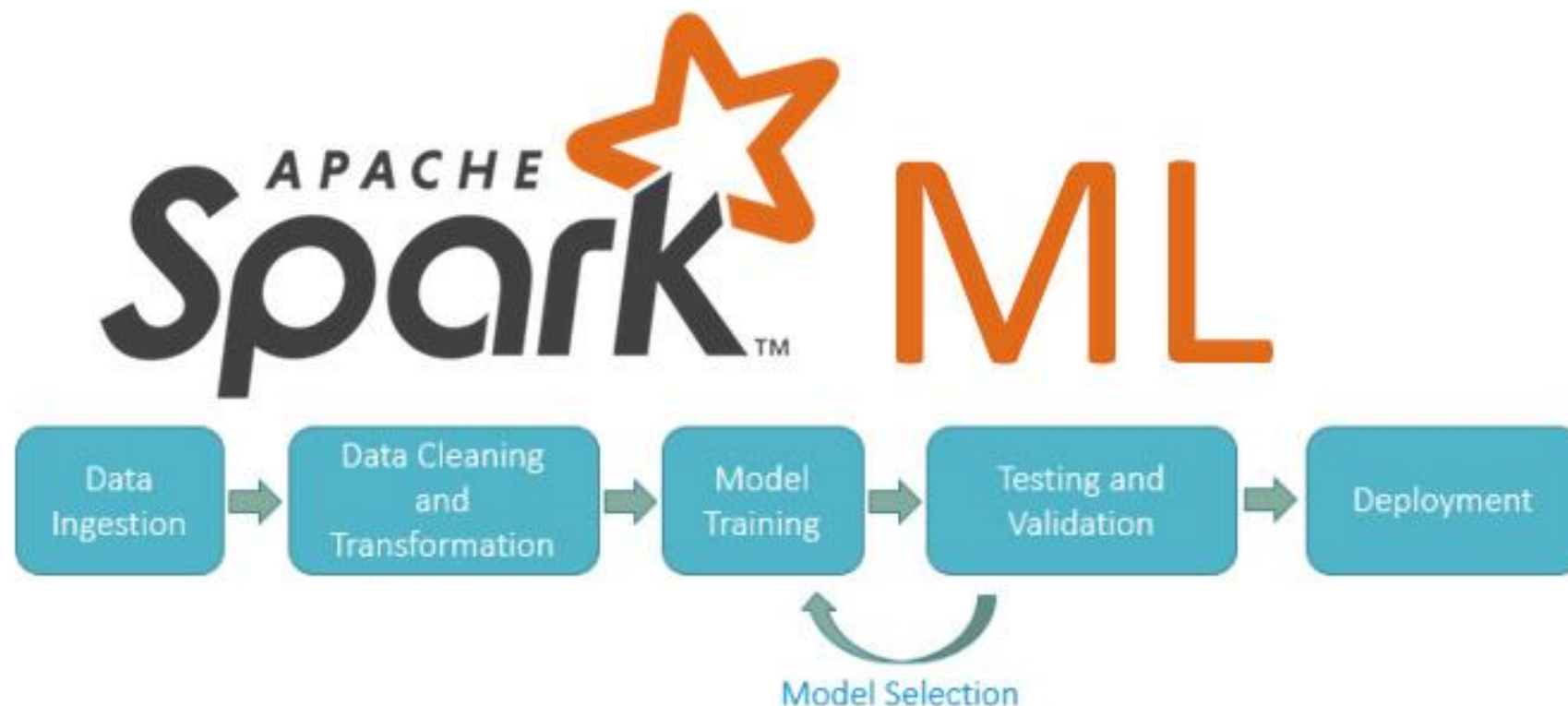


Imagem reproduzida de: <https://yurongfan.files.wordpress.com/2017/01/sparkml.jpg>

Inspire