Gabriel Augusto

Curso: Engenheiro de Machine Learning - Udacity

Histórico do assunto

A necessidade de aumentar a receita em voos e evitar prejuízos é algo comum para qualquer empresa aérea. Entre as principais causas de prejuízo existem os atrasos por problemas de manutenções não programadas e os atrasos e/ou cancelamentos devido ao "mau tempo".

Para ser permitido a partida ou pouso de um voo é necessário condições meteorológicas aceitáveis. Caso as condições meteorológicas não sejam aceitáveis isso poderá acarretar mudanças de rotas, indenizações para os passageiros. Por isso é necessário um planejamento por parte das companhias aéreas para como se portar no meio dessas condições adversas.

Em 2016 foi relatado que cerca de 20% dos voos sofreram atrasos e que cerca de 35% dos voos atrasados foram por causa do mau tempo [1]. Para ajudar a reduzir esses atrasos são utilizados modelos de predição a partir de Aprendizado de máquina [2]. O Google, por exemplo, já está adicionando ao Google Flights a funcionalidade de prever se o seu voo vai atrasar para decolar utilizando aprendizado de máquina [3] .

Descrição do problema

A partir de dados meteorológicos de aeroportos dados (METAR[1]) e de status de voos antigos Ex: realizado, atrasado por condições meteorológicas adversas, cancelado por condições meteorológicas adversas.

Seria um problema de classificação para prever se um voo terá atraso, será cancelado ou realizado com sucesso, inicialmente no Brasil isso para cada aeroporto individualmente, pois parâmetros da localização do aeroporto podem interferir na questão dos atrasos.

1. METAR - (METeorological Aerodrome Report - Informe meteorológico regular de aeródromo), é um informe codificado, associado às observações meteorológicas à superfície, e utilizado para fornecer informações sobre condições do tempo em um aeródromo específico.

Inputs

1º Conjunto de dados:

Entrada de dados METAR do Brasil retirada do site da Universidade de Iowa nos EUA [4]:

(Os arquivos dos aeroportos escolhidos estão no link do github enviado)

Todas as colunas: station,valid,lon,lat,tmpf, dwpf, relh, drct, sknt, p01i, alti, mslp, vsby, gust, skyc1, skyc2, skyc3, skyc4, skyl1, skyl2, skyl3, skyl4, wxcodes, ice_accretion_1hr, ice_accretion_3hr, ice_accretion_6hr, metar

2º Conjunto de dados

Dados retirados da Agência Nacional de Aviação Civil (ANAC) no Brasil de Janeiro de 2015 até Agosto de 2017 a partir de um projeto do Kaggle.

Os dados estão no site do Kaggle e como ia ficar muito grande para enviar segue o link na referência [5].

Todas as colunas:
Voos,Companhia.Aerea,Codigo.Tipo.Linha,Partida.Prevista,Partida.Real,Chegada.Prevista,Chegada.Real,Situacao.Voo,Codigo.Justificativa,Aeroporto.Origem,Cidade.Origem,Estado.Origem,Pais.Origem,Aeroporto.Destino,Cidade.Destino,Estado.Destino,Pais.Destino,LongDest,LatDest,LongOrig,LatOrig

O primeiro conjunto de dados que contém as informações dos aeroportos será entrelaçado com o segundo que contém informações do status de cada voo para predição de atraso ou cancelamento de voos

Esses conjuntos de dados seriam “concatenados” pela e pela Hora da observação (coluna valid do 1º conjunto de dados) Hora de Partida Prevista (coluna Partida.Prevista do 2º conjunto) mantendo colunas numéricas consideradas mais relevantes como:

tmpf: (Temperatura em Fahrenheit) -> Air Temperature in Fahrenheit, typically @ 2 meters

dwpf: (Ponto de orvalho em Fahrenheit) ->Dew Point Temperature in Fahrenheit, typically @ 2 meters

relh: (Umidade relativa) ->Relative Humidity in %

drct: (Direção do Vento em graus do norte) ->Wind Direction in degrees from north

sknt: (Velocidade do Vento em nós) -> Wind Speed in knots

alti: (Pressão do altimetro em polegadas) -> Pressure altimeter in inches

mslp: (Pressão do nível do mar em millibar) -> Sea Level Pressure in millibar

vsby: (Visibilidade em milhas) ->Visibility in miles

gust: (Velocidade do vento em nós) -> Wind Gust in knots

skyc1, skyc2, skyc3, skyc4 -> cobertura do céu, relacionado a visibilidade

skyl1, skyl2, skyl3, skyl4 -> nível do céu, relacionado a visibilidade

LongDest, LatDest, LongOrig, LatOrig

E também de datas para retirar o tempo de atraso caso seja necessário retirar o tempo de atraso:

Partida.Prevista, Partida.Real, Chegada.Prevista, Chegada.Real

Targets: Situação.Voo e Código.Justificativa (Do 2º conjunto de dados)

Como nas amostras de dados existem também voos atrasados/cancelados por outros problemas como falha técnica. Será feito um pré-processamento para tratar valores que podem acabar influenciando negativamente o modelo. Serão tratados também voos para o exterior já que os aeroportos escolhidos estão no Brasil.

E a partir de uma breve análise foi constatado que cerca de 0,03% dos voos sofrem atrasos devido ao mau tempo.

Obs: Vale ressaltar que o único código nulo que existe na tabela é quando o Voo é realizado com sucesso e não existe Código.Justificativa, pois é apenas para caso tenha tido algum problema.

Descrição da solução

Os dados de parâmetros do primeiro conjunto de dados (dados dos aeroportos a cada hora como temperatura) e os com status de cada voo desses mesmos aeroportos no mesmo horário serão utilizados sendo escolhidos inicialmente 8 aeroportos (Pela quantidade de voo e por ter estações do ano relativamente definidas: Guarulhos (SBGR), Congonhas (SBSP), Viracopos (SBKP) do Galeão (SBGL); Clima subtropical: o de Curitiba (SBCT), Porto Alegre (SBPA); por ter o clima tropical litorâneo: Recife(SBRF); Clima equatorial: de Manaus (SBMN):

Exemplo de identificador: SBKP, SBCT...

Para treinamento e classificação de 3 tipos de voos:

1. Os que poderiam decolar sem atrasos ocasionados pelo mau tempo

2. Os que podem atrasar

3. Os que tendem a ser cancelados

E fornecer uma estatística com os dados de teste de quantos acertos de atraso ou cancelamentos foram adquiridos a partir dessa base de dados treinada.

Os modelos de benchmark

Utilizar algoritmos como:

Árvores de decisão, random forest, Adaboost e KNN verificando o melhor modelo para minimizar os erros de treino e de teste. E por fim utilizar curvas ROC para comparar com os resultados obtidos em [6], para avaliar os dados no Brasil são mais fáceis de predizer do que o dos EUA que foi usado em [6].

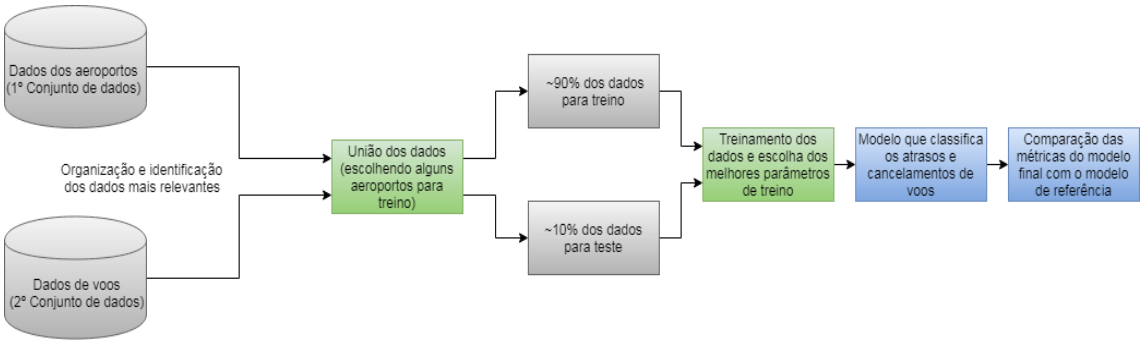
Métricas de avaliação

Taxa de erro/Acurácia (%)

Precisão, porcentagem de falso positivos em relação ao total (%)

Revocação, porcentagem de falsos negativos em relação à amostragem total (%), tempo(segundos) seguindo as métricas de [6] e de [7] , adicionando o FβScore para oferecer um score que demonstra a confiabilidade do resultado usando tanto a precisão e revocação.

Design do projeto:



Para auxiliar no entendimento do projeto foram lidos os seguintes artigos [2] e [8]

Referências

1. https://www.bts.gov/newsroom/february-2016-time-performance-previous-year-january-2016

2. http://www.mit.edu/~hamsa/pubs/GopalakrishnanBalakrishnanATM2017.pdf

3. https://futurism.com/google-thinks-accurately-predict-next-flight-delay

4. https://mesonet.agron.iastate.edu/request/download.phtml?network=BR__ASOS

5. https://www.kaggle.com/ramirobentes/flights-in-brazil/home

6. https://ieeexplore.ieee.org/document/7777956

7. http://cs229.stanford.edu/proj2012/CastilloLawson-PredictingFlightDelays.pdf

8. https://arxiv.org/pdf/1703.06118.pdf