Data Science Challenge @ ITA 2023 - Previsão de ELDT

Overview

O *Estimated Landing Time* (ELDT) é o horário previsto para o pouso de aeronaves, sendo uma informação crucial para o planejamento do transporte aéreo. Com base nesta informação, várias ações são tomadas, desde a alocação do portão e esteira de bagagens no aeroporto, que tem impacto diretamente nos passageiros, ao reabastecimento das aeronaves e a alocação de espaço aéreo para os voos.

Podemos afirmar que o ELDT é muito importante para melhorar a previsibilidade na aviação e desta forma permitir o melhor uso da infraestrutura aeroportuária e melhorar a gestão do espaço aéreo. Uma melhor gestão do espaço aéreo, que amplie a capacidade e a previsibilidade, permitirá a introdução de novos vetores como os Drones e os eVTOLS, conhecidos como carros voadores.

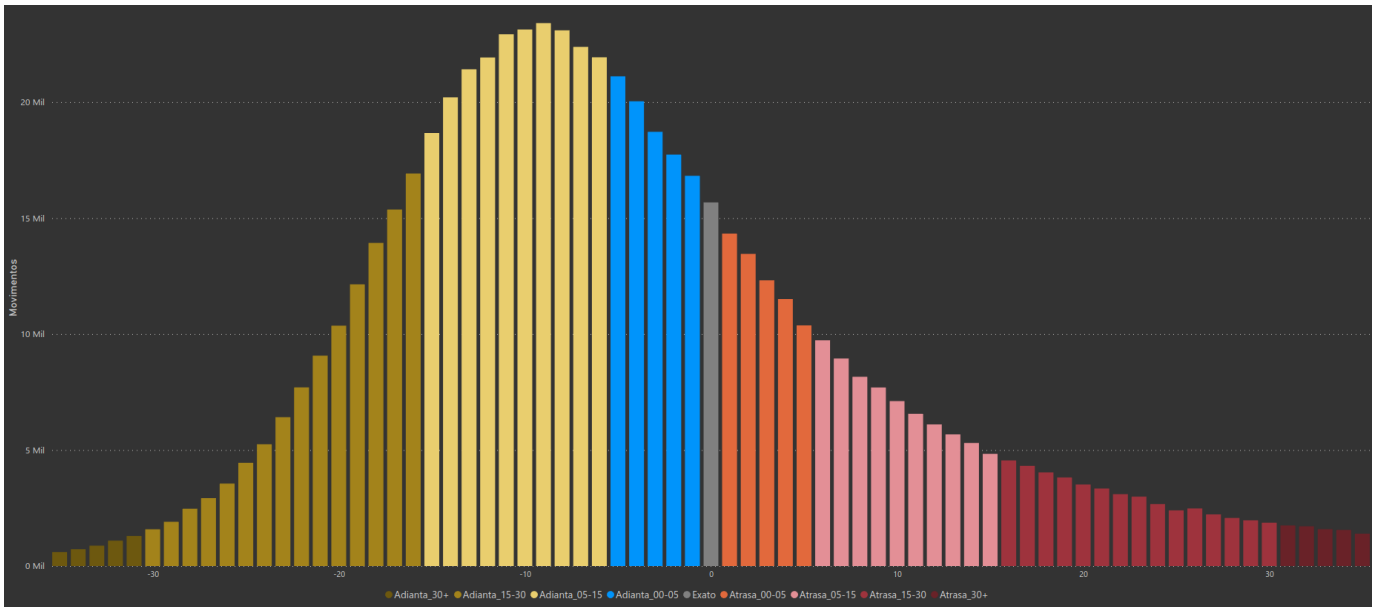

Figura 1: Histograma de atraso/adiantamento por minuto.

A Figura 1 apresenta o histograma de atrasos e adiantamento referente ao ELDT de voos em 2022. Na cor cinza (após as barras em azul), é possível perceber os voos em que o ELDT foi igual ao horário real de pouso. À esquerda da barra cinza, estão os voos que adiantaram o pouso em relação ao horário previsto, discriminados por minuto e, à direita, dos voos que atrasaram. Observa-se a grande frequência de voos não pontuais, isto é, de voos cujo ELDT foi diferente do horário real do pouso.

O ELDT pode ser estimado com base em vários fatores, incluindo a rota de voo, as condições meteorológicas, o desempenho da aeronave e outros fatores que possam afetar o voo. São levadas em consideração variáveis como a velocidade e direção do vento, a temperatura do ar, a altitude de cruzeiro, a velocidade do avião e outros parâmetros que podem afetar a velocidade do voo, como o congestionamento. Todas estas informações são usadas para estimar o tempo que a aeronave levará para percorrer a rota de voo e chegar ao destino.

Objetivo

O objetivo desta competição é desenvolver um modelo preditor do ELDT para voos comerciais com origem e destino nos 12 principais aeródromos do Brasil, a saber:

- Brasília (SBBR)
- Confins (SBCF)
- Curitiba (SBCT)
- Florianópolis (SBFL)
- Rio de Janeiro - Galeão (SBGL)
- Guarulhos (SBGR)
- Campinas (SBKP)
- Porto Alegre (SBPA)
- Recife (SBRF)
- Rio de Janeiro - Santos Dumont (SBRJ)
- São Paulo - Congonhas (SBSP)
- Salvador (SBSV)

Dados

Os participantes terão acesso a diversas bases de dados relacionadas à geolocalização, satélites, estações meteorológicas e preditores climáticos, para desenvolverem modelos precisos que possam aprimorar a previsibilidade do ELDT e contribuir para um transporte aéreo mais eficiente e seguro. Para isso, serão disponibilizadas as seguintes bases de dados:

- BIMTRA (Banco de Informações de Movimento de Tráfego Aéreo)
- Dados de Síntese Radar CAT-62
- Dados meteorológicos
    - METAR (Meteorological Aerodrome Report)
    - Imagens de satélite meteorológico
    - METAF (Terminal Aerodrome Forecast)
- Dados de ATFM (Air Traffic Flow Management)
    - Esperas em voo
    - Previsão e histórico de troca de cabeceira

O acesso pode ser realizado através da API via REST e é necessário o uso de token de acesso, que é individual para cada grupo. O token será enviado por email após o término das inscrições. O link anterior, possui um interface de teste para se familiarizarem com os formatos.

Qualquer problema de acesso, ou na requisição dos dados, por favor, entrar em contato conosco pelo Discord ou pelo e-mail jeanjpol@decea.mil.br.

BIMTRA

A base de dados **BIMTRA** possui informações dos movimentos nos aeródromos do Brasil. Assim, é possível encontrar dados como hora de decolagem estimada, origem, destino, tempo de voo previsto, etc. Para o problema em questão, uma versão resumida do BIMTRA será fornecida, contendo os seguintes campos:

- FlightID (Identificador único de um voo)
- Origem (Código ICAO do Aeroporto de Origem do voo)
- Destino (Código ICAO do Aeroporto de Destino do voo)
- Hora_Dep (Data/Hora de Decolagem do voo)
- **Hora_Arr (Data/Hora de Pouso do voo)**

O campo Hora_Arr corresponde ao atributo alvo do desafio proposto, isto é, o modelo desenvolvido por cada equipe deverá estimar os valores de Hora_Arr. Tomem cuidado com data leakage, ou seja, construindo modelos com informações do futuro.

CAT-62

A base de dados **CAT-62** consiste na síntese radar do espaço aéreo brasileiro. A cada 4 segundos é gerado um registro para cada aeronave que voa o espaço aéreo brasileiro, com os seguintes campos:

- FlightID
- Latitude
- Longitude
- Altitude
- Velocidade
- Data e hora

A Figura 2 ilustra um exemplo de voos no espaço aéreo brasileiro para um dado instante, ou seja, imagem que pode ser gerada a partir dos dados do CAT-62.

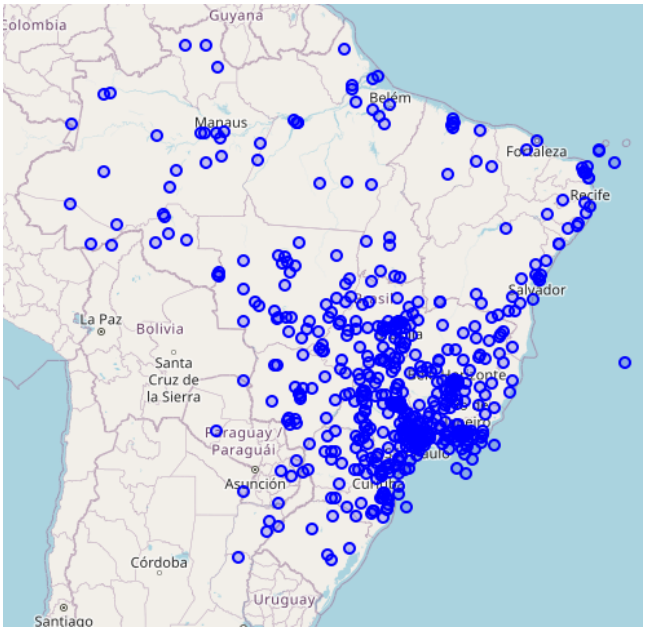

Figura 2: Voos no espaço aéreo brasileiro para um dado instante segundo o CAT-62.

Para a fase de treino e implementação dos modelos, serão fornecidos os dados do CAT-62 de minuto a minuto. Dessa forma, será possível analisar toda a trajetória feita por cada aeronave, bem como a densidade de uma região.

Dados Meteorológicos

Fenômenos meteorológicos exercem uma forte influência na pontualidade de um voo. Por esse motivo, serão fornecidos dados meteorológicos de 4 fontes:

- METAR (Meteorological Aerodrome Report)
- Imagens de satélite meteorológico
- METAF (Terminal Aerodrome Forecast)

Os dados de **Satélite Meteorológico** serão disponibilizados em formato de imagem, como ilustrado na Figura 3, em que é possível identificar tempos severos, potenciais precipitações, tendências de movimento de nuvens, etc.

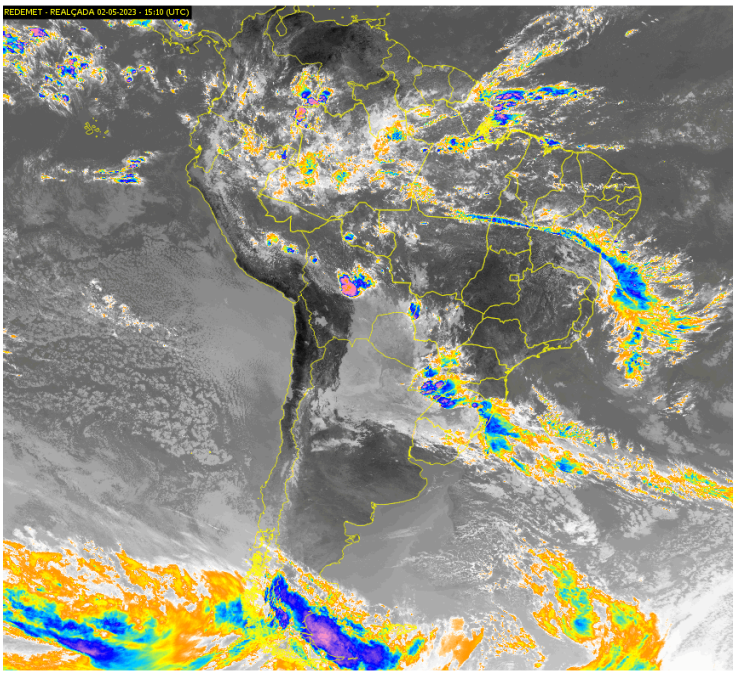

Figura 3: Exemplo de imagem de Satélite Meteorológico.

As bases **METAR** e **METAF** apresentam dados de telemetria de estações meteorológicas da região dos aeródromos, como temperatura, velocidade do vento, umidade, etc. Mais informações sobre quais dados existem e como interpretar os dados de METAR e METAF estão disponíveis aqui. Embora a codificação destas informações seja complexa, é possível utilizar bibliotecas python para o parsing. Um exemplo de uma biblioteca com essa funcionalidade está disponível aqui.

A diferença entre as bases METAF e METAR é que a METAR é o valor observado enquanto a METAF apresenta previsões para a próxima hora.

Dados de ATFM

Complementarmente, serão fornecidos dados de ATFM (Air Traffic Flow Management) que permitem auxiliar na análise comportamental do espaço aéreo, e compreendem:

- Esperas em voo
- Previsão de Troca de Cabeceira
- Histórico de Troca de Cabeceira

A base de dados **Esperas** contém o histórico de ocorrências de espera em voo por hora e aeródromo. Como temos 12 aeródromos, teremos 12 registros na base de dados por hora, onde cada registro indicará a quantidade de voos em espera para cada aeródromo.

A base de dados **Previsão de Troca de Cabeceira** traz as previsões por hora de troca de cabeceira nos aeroportos. Os campos dessa base compreendem a hora de referência, a informação de troca de cabeceira (variável booleana) e o aeroporto. Enquanto a base de **Histórico de Troca de Cabeceira** corresponde às observações de troca de cabeceira, que traz os campos hora de referência, quantidade de trocas de cabeceira nessa hora e o aeroporto.

Fases do desafio

Nesta edição teremos um único desafio que será avaliado em duas fases: Fase 1 é quantitativa e Fase 2 qualitativa.

Fase 1

Na Fase 1, cada time deverá submeter as previsões pela plataforma Kaggle através do link a ser disponibilizado após a abertura do evento. As submissões serão automaticamente ranqueadas pela plataforma gerando um ranqueamento preliminar, calculado pela métrica *mean squared error* (MSE).

Vale citar que este ranqueamento serve apenas de base para os times, como tradicionalmente é feito nos desafios pelo Kaggle. O ranqueamento final será calculado com base em uma base de teste a ser disponibilizada próximo do final da Fase 1.

Serão classificados para a Fase 2, apenas os 5 melhores times que:

1 - Tenham menor MSE para a base de teste

2 - Implemente a integração de coleta de dados diretamente via API disponibilizada pelo ICEA

3 - Submetam o código desenvolvido pela equipe, juntamente com um documento de instruções para que seja possível a verificação do funcionamento, via Google Form, a ser informado oportunamente.

Importante: observe que apenas as soluções integradas à plataforma do ICEA que poderão passar para a Fase 2.

É preferível que os times inicialmente concentrem seus esforços na geração do modelo, utilizando os dados disponibilizados pelo Kaggle, com os dados já integrados. Caso tenham potencial para serem classificados para a Fase 2, posteriormente, podem focar na integração do modelo com a interface do ICEA.

Dados do Kaggle

Para a avaliação, os dados serão fornecidos da seguinte forma:

- BIMTRA: Sem Hora_Arr (variável a ser estimada)
- CAT-62: Para cada voo contido no BIMTRA, será fornecido um snapshot do CAT-62 para o instante de decolagem do voo. Não incluirá o FlightID
- Satélite meteorológico: Será fornecida última imagem de satélite em relação a hora de decolagem
- METAR: Dados da última leitura do METAR em relação à hora da decolagem
- METAF: Dados da previsão futura mais próxima ao horário de decolagem.
- Dados de ATFM: Dados da última atualização em relação ao horário do voo.

Arquivo de Submissão

Os participantes devem submeter um arquivo .csv contendo uma coluna de FlightID e uma coluna com o atributo alvo previsto, da seguinte forma:

```
ID, solution
7945735584a3297121c4f5ae0de8ecd1, 3006
cd13f2d62720d6d43610ec777e7a213b, 5072
```

Repare que o atributo alvo não está explicitamente dado com o ELDT (timestamp), mas sim com o tempo em segundos entre a decolagem e o pouso da aeronave, isto é, o delta referente a duração do voo. Esta convenção foi feita para adequar-se à plataforma Kaggle, onde o desempenho das equipes será calculado.

Fase 2

Na Fase 2, as equipes qualificadas deverão apresentar a solução desenvolvida. Esta apresentação deve conter ao menos: técnicas de pré-processamento de dados utilizadas; pipeline da modelagem implementada; tecnologias e métodos utilizados; resultados obtidos.

Essa apresentação será ao vivo e de forma remota. A avaliação da Fase 2 será realizada por uma banca de especialistas formada por membros do ITA, ICEA e LATAM. A avaliação final levará em conta os resultados da Fase 1 e da Fase 2. Como critério de avaliação, os pontos considerados como principais são: criatividade da solução, estrutura metodológica e didatismo das apresentações.

A equipe vencedora será convidada a apresentar sua solução durante o Seminário de Performance ATM 2023, que acontecerá entre os dias 6 e 8 de novembro em São José dos Campos – SP.