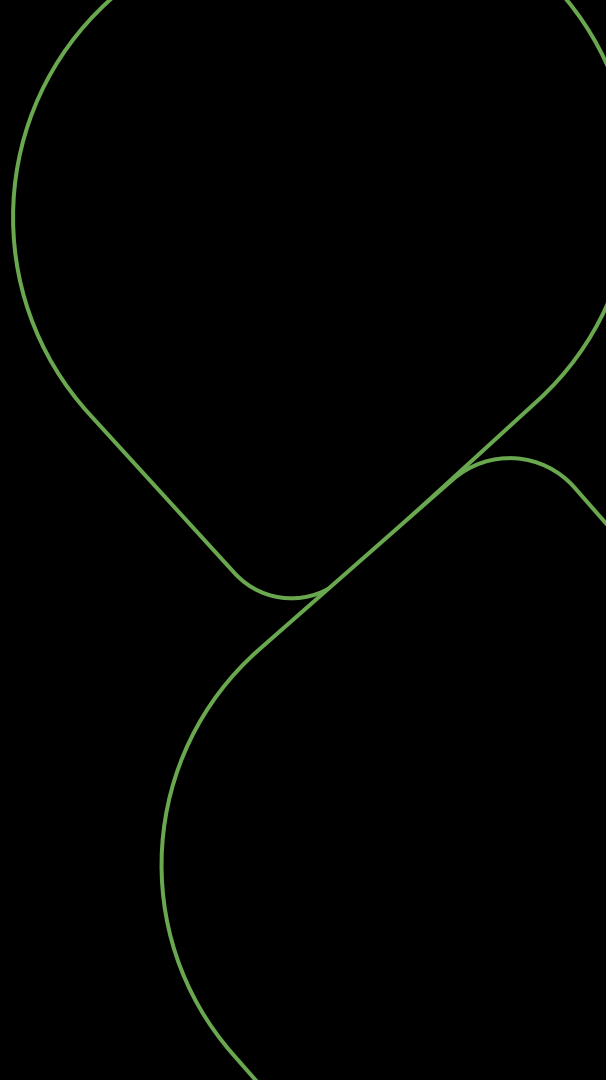


Desafio

Cientista de Dados Jr.

Lucas Cunha



Problema proposto

A Agência Nacional de Aviação Civil (ANAC) disponibiliza dados estatísticos sobre o transporte aéreo, como informações relacionadas a voos, operações aéreas, e estatísticas sobre companhias aéreas, aeroportos, rotas, passageiros, carga e etc.

O desafio consiste em fazer análises destes dados, responder perguntas de negócio e criar modelos de previsões do número de passageiros.

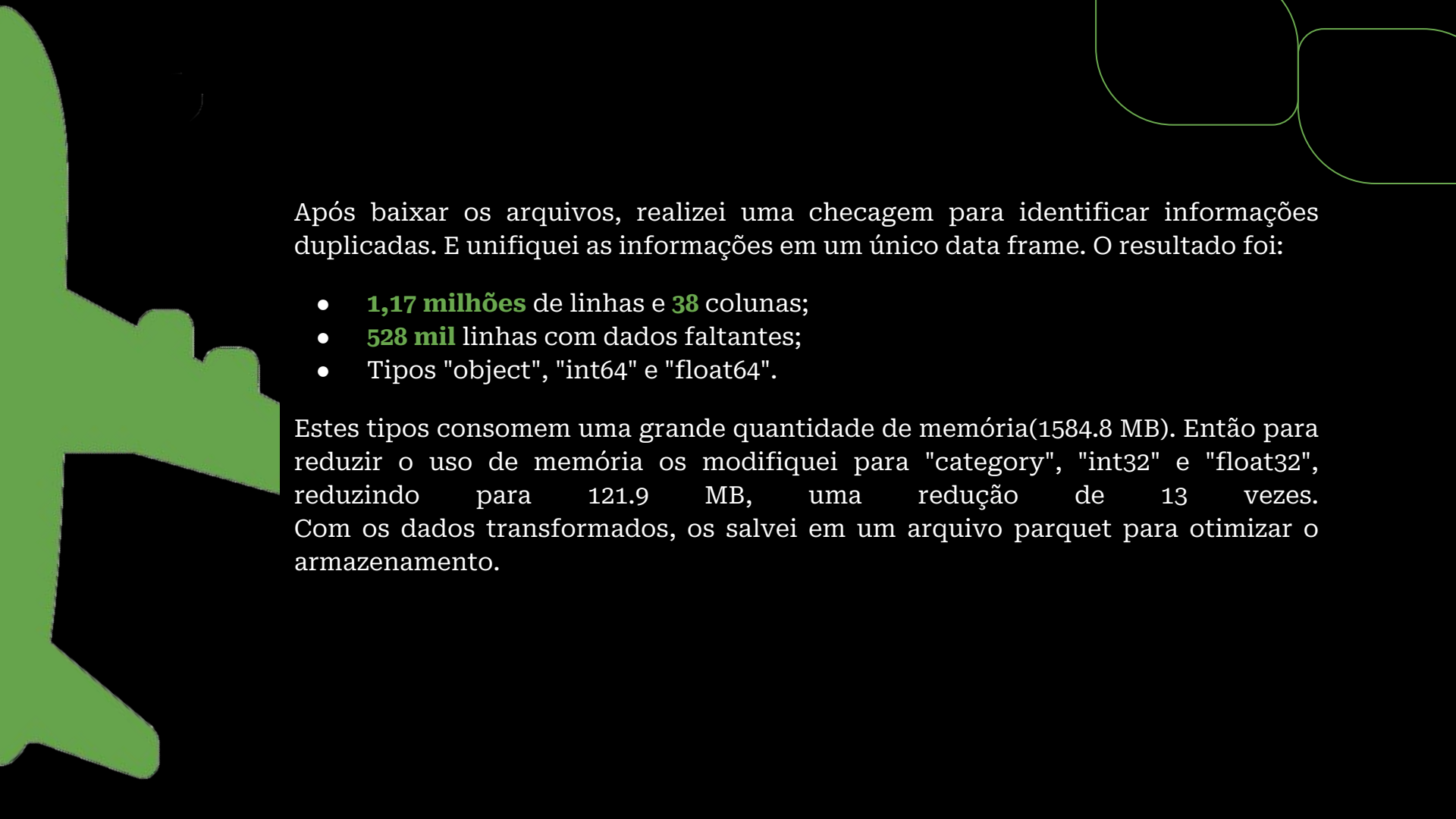




Entendendo o Contexto

Os dados estão separados em vários arquivos, de extensões diferentes. E apesar da informação de que os arquivos “.json” estão separados por décadas, existiam algumas inconsistências. Um outro ponto importante é que ao navegar pelo site da ANAC, entender como foram gerados esses dados é uma tarefa complicada, já que os textos ficam abertos a interpretação. Nesta etapa algumas **melhorias de gerenciamento de dados** já poderiam ser aplicadas. Como armazenar os dados em um Banco SQL, ou unificar o formato dos arquivos para um único, como por exemplo Parquet que é um formato otimizado para processamento e armazenamento eficiente de dados em estruturas de colunas.

../		
Dados Estatisticos.csv	24-Apr-2023 01:09	314890040
Dados Estatisticos 2000 a 2010.json	24-Apr-2023 01:09	482445937
Dados Estatisticos 2011 a 2020.json	24-Apr-2023 01:09	431654315
Dados Estatisticos 2021 a 2030.json	24-Apr-2023 01:10	77105521
Dados Estatisticos parte.csv	09-Mar-2023 10:13	62987814



Após baixar os arquivos, realizei uma checagem para identificar informações duplicadas. E unifiquei as informações em um único data frame. O resultado foi:

- **1,17 milhões** de linhas e **38** colunas;
- **528 mil** linhas com dados faltantes;
- Tipos "object", "int64" e "float64".

Estes tipos consomem uma grande quantidade de memória(1584.8 MB). Então para reduzir o uso de memória os modifiquei para "category", "int32" e "float32", reduzindo para 121.9 MB, uma redução de 13 vezes. Com os dados transformados, os salvei em um arquivo parquet para otimizar o armazenamento.

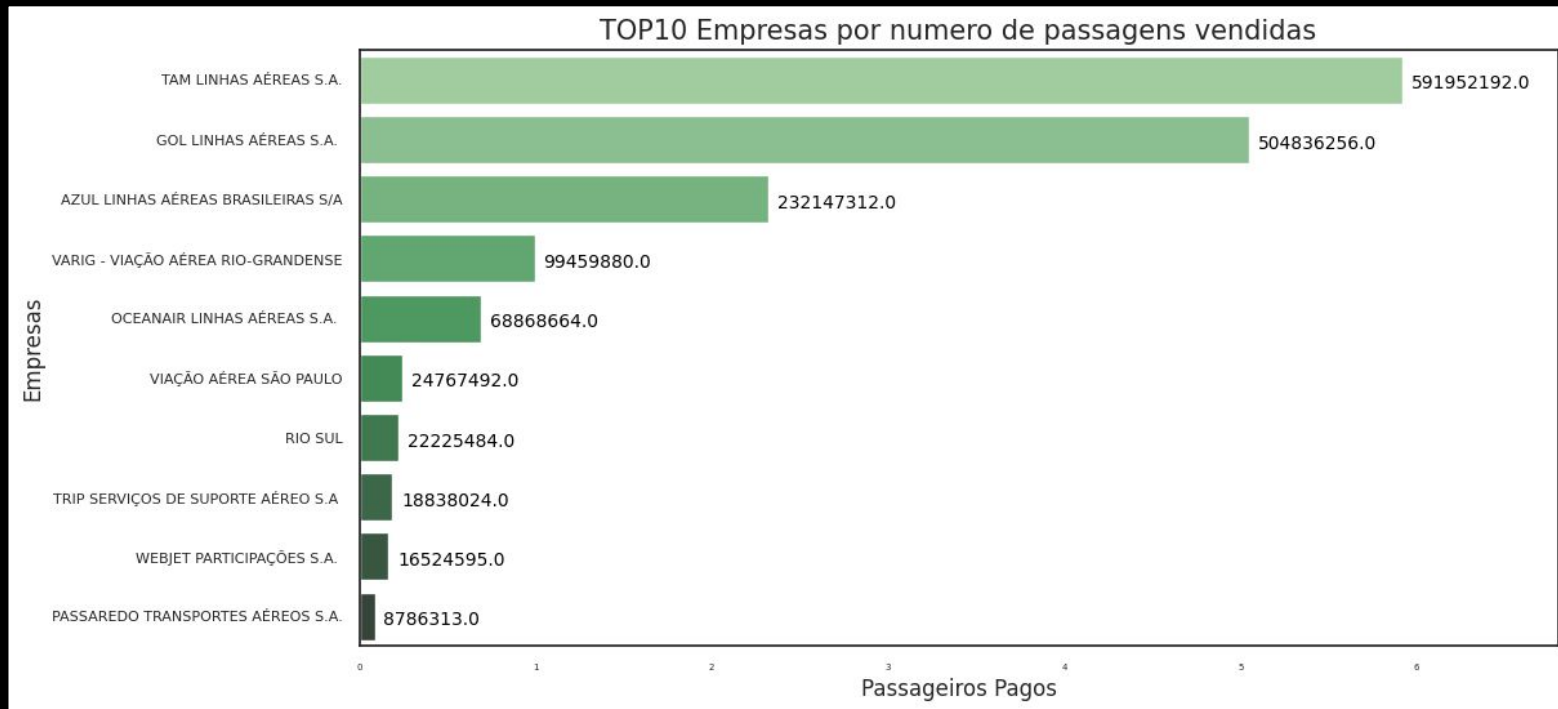


Exploratory Data Analysis

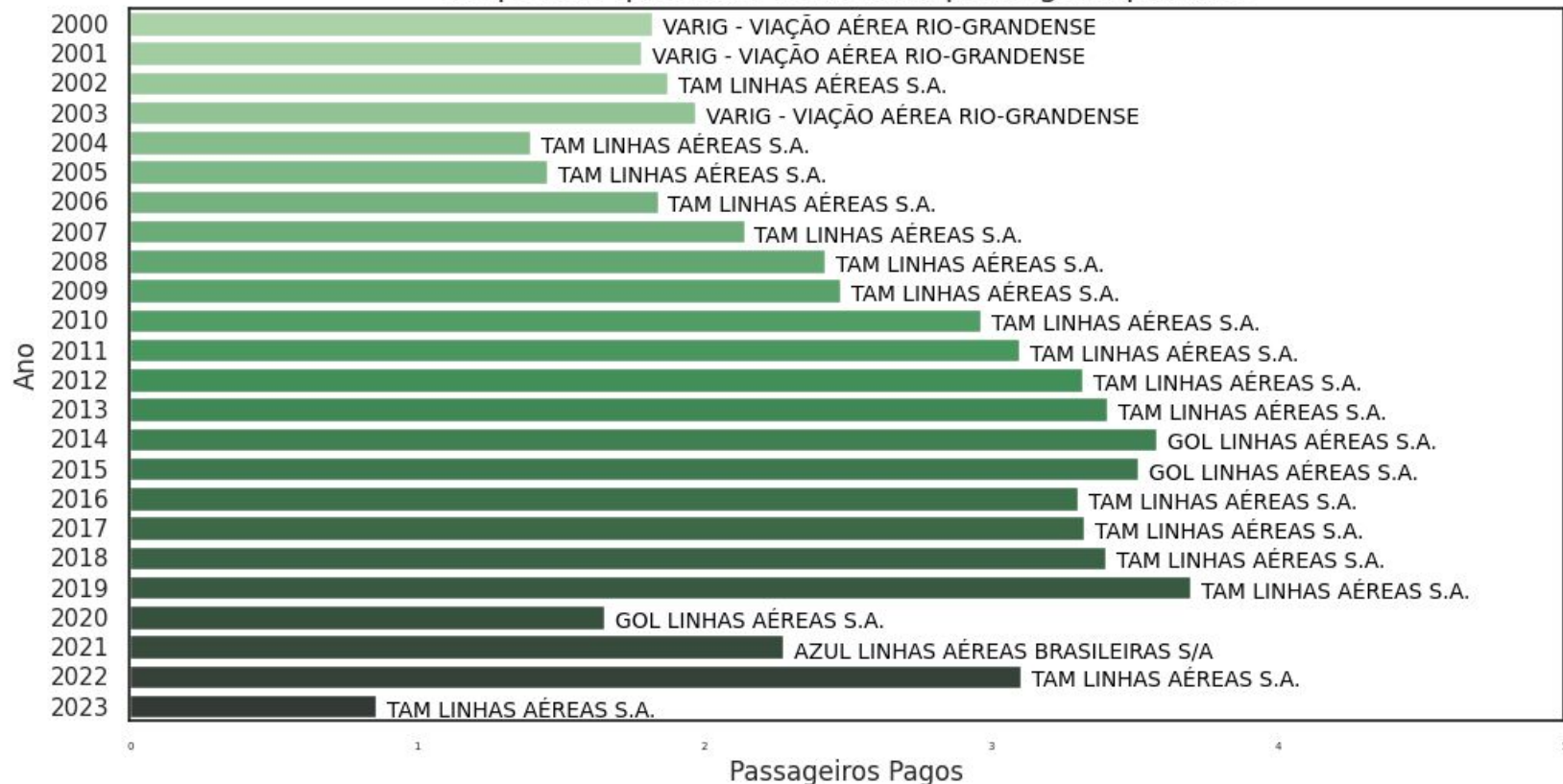
Nesta etapa é analisado os dados com objetivo de:

- Estrutura dos dados: Compreender a composição dos dados, os tipos de variáveis, unidades de medida e verificar se os dados estão no formato adequado;
- Padrões e tendências: Explorar a distribuição dos dados, identificar padrões de variabilidade, tendências temporais, sazonalidades e outros padrões relevantes;
- Insights e hipóteses: Analisar relacionamentos entre variáveis, identificar associações, tendências conjuntas ou contrastantes e formular hipóteses para explicar esses padrões.
- Anomalias e outliers: Identificar observações incomuns ou discrepantes que possam indicar erros de medição, valores extremos ou comportamento anômalo;
- Modificação de dados: Realizar tratamentos e transformações nos dados, como lidar com valores ausentes, normalizar variáveis, criar novas variáveis derivadas ou agrupar categorias;

Análise de Mercado



Empresas que mais venderam passagens por ano



Análise Comparativa

TAM x GOL

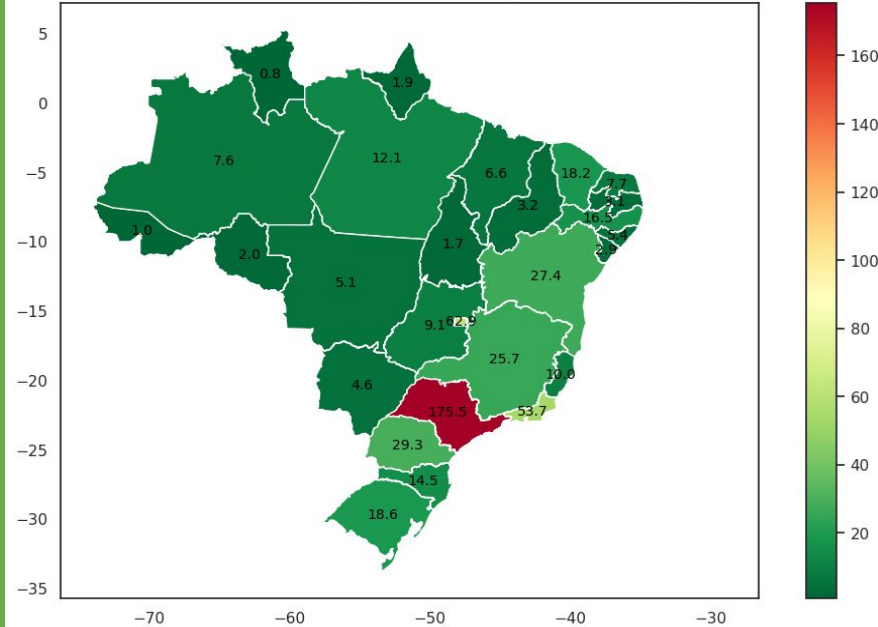
Com base nas análises realizadas em todas as empresas, meu objetivo foi identificar e compreender as diferenças entre a TAM e a GOL, que são as duas principais empresas do mercado.

Por meio desse estudo, onde busquei obter insights sobre o funcionamento das empresas e entender como essas diferenças impactam nos resultados alcançados por cada uma delas.

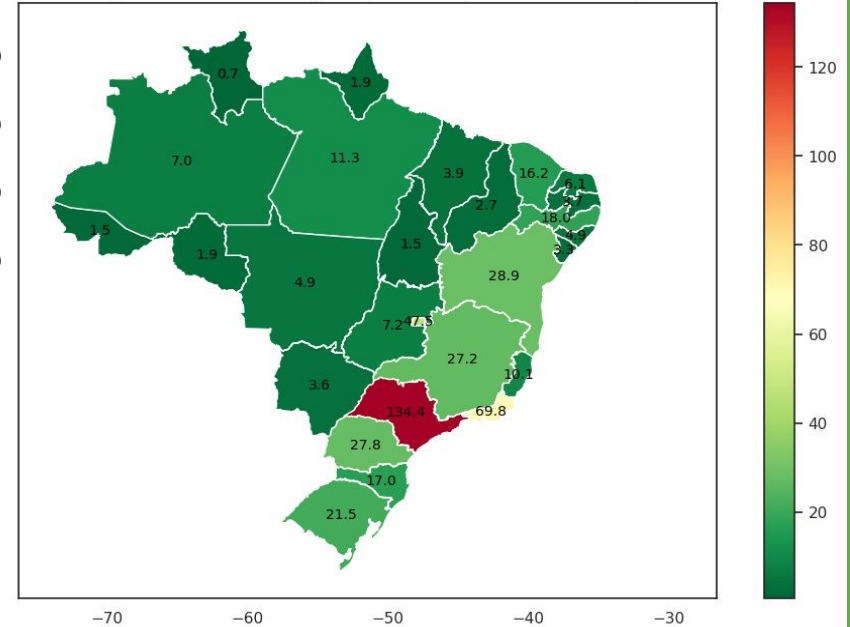
	TAM	GOL
Número de registros	122314	93485
Número de rotas	1815	1414
Quantos anos de registro	24	23
Média de decolagens por ano	247.081	211.235
Número total de decolagens	5.806.402	4.749.250
Média de passageiros por ano	25.189.453	22.453.798
Total de passageiros pagos	591.952.128	504.836.224
Média de passageiro pagos por voo	96	108
Total de passageiros grátis	11.410.875	11.773.455
Média de passageiros grátis por voo	1	2
Média de distância em km por voo	928	873
Médias de assentos por voo	166	172
Média de horas por voo	2	2

Densidade Geográfica - Passageiro por estado

Mapa de calor TAM - Passageiros por Estado (Numero em Milhões)



Mapa de calor GOL - Passageiros por Estado (Numero em Milhões)





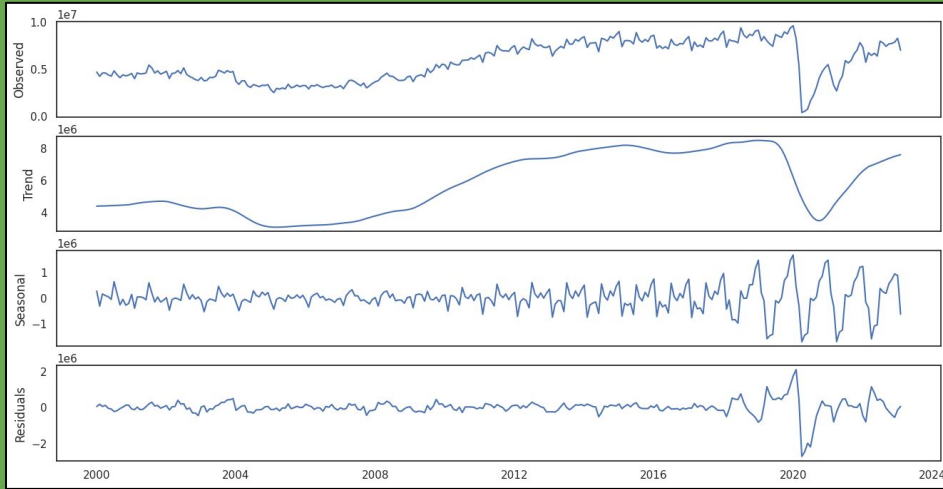
- Maior foco em número de rotas, sendo elas mais distribuídas geograficamente, consequentemente faz mais decolagens por ano.
- Devido as rotas mais diversas o número médio de passageiros por voo é reduzido.
- O resultado da estratégia de atender um número maior de locais é um impacto de maior número de passageiros pagos em relação a GOL.



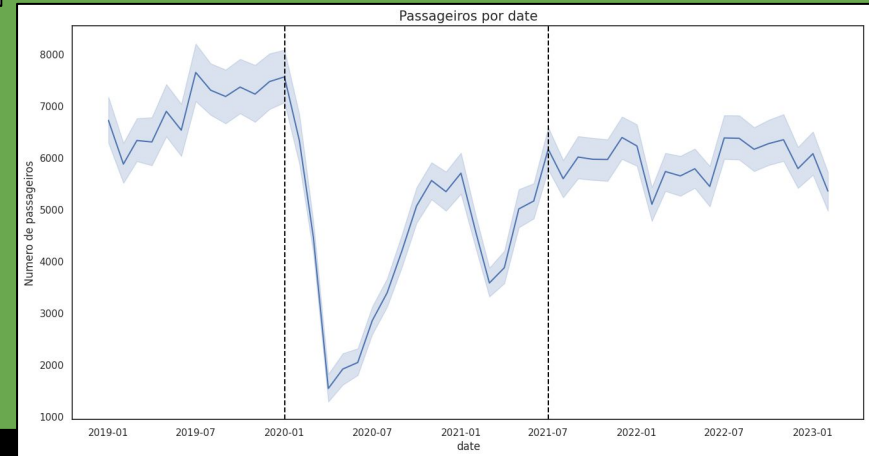
- Adota rotas mais estratégicas.
- O menor número de rotas é compensado com um foco maior em estados de maior densidade populacional, como o Rio de Janeiro, por exemplo.
- Como resultado, a GOL alcança um número médio maior de passageiros por voo.

Time Series Analysis

- É possível ver que a tendência atual é crescente.
- E que a sazonalidade se dá nos períodos finais de cada ano, tendo uma queda em fevereiro.



Como podemos ver a queda no número de passageiros causada pela covid é uma mudança muito grande de comportamento, e tende a não acontecer novamente. Por isso este espaço de dados onde ocorreu o Covid não entrará no modelo.



Modelo e Previsões

Data Cleaning

- Unicode para limpeza de textos;
- Split de informações;
- Preenchimento dos NA's nas localizações;

Feature Engineering

- Lag Features (Rolagem por ano);
- Transformações de seno;
- Transformações de cosseno;
- Rota (origem - destino);
- Features derivadas de operações entre outras colunas;

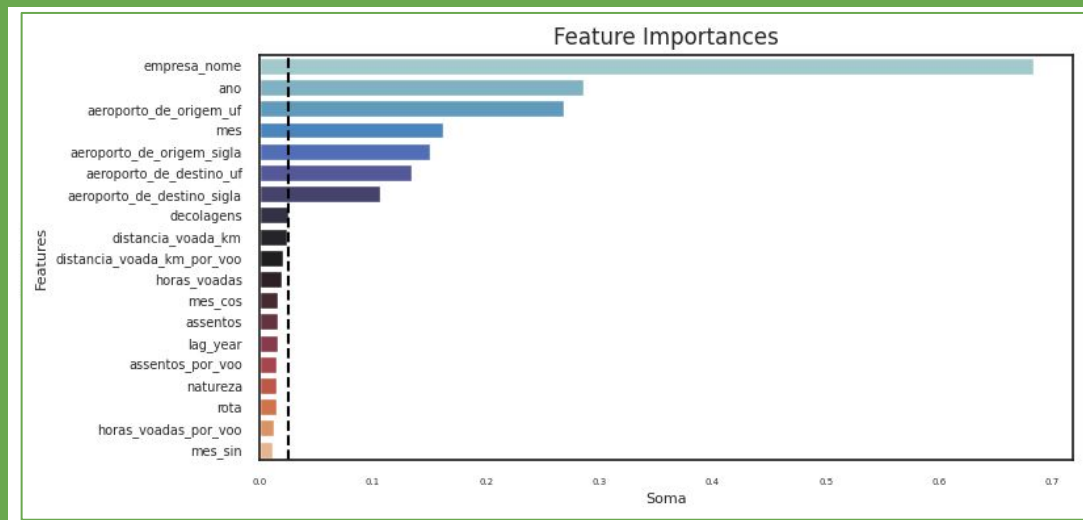
Pre Processing

- Substituição de NA's numéricos por 0;
- Substituição de NA's categóricos pelo valor mais frequente;
- Padronização dos dados para modelos lineares;
- Substituição dos dados categóricos pela frequência;

Modelo e Previsões

Feature Selection

Para a seleção de features, os resultados da importância de recursos (feature_importance) de um XGBoostRegressor e os valores dos coeficientes normalizados de uma Regressão Linear foram somados. A seleção final incluiu apenas os recursos cuja soma total resultou em um valor acima de 0.03.



Algoritmos e Métricas

O CatBoostRegressor e o LGBMRegressor são algoritmos de regressão baseados em árvores que são eficientes para lidar com conjuntos de dados grandes e complexos. Embora sejam semelhantes em muitos aspectos, existem algumas diferenças entre eles:

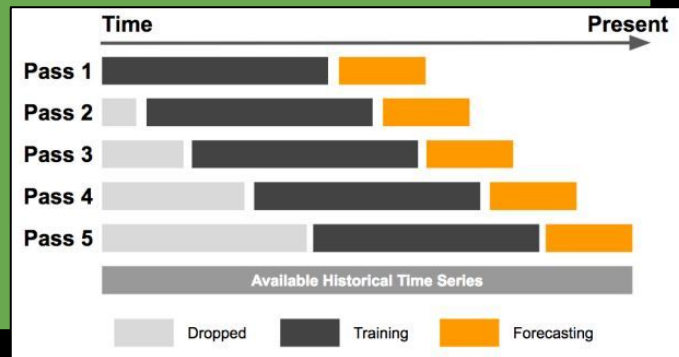
- CatBoost é especialmente projetado para lidar com variáveis categóricas;
- O CatBoost possui um mecanismo interno para lidar com valores ausentes nos dados de entrada, eliminando a necessidade de tratamento prévio desses valores;
- LGBMRegressor é conhecido por sua velocidade de treinamento mais rápida;

Para avaliação dos algoritmos a métrica escolhida foi o RMSE.

Modelo e Previsões

Cross Validation e Hyperparameter Fine Tuning

A validação cruzada é uma técnica fundamental na avaliação de modelos de aprendizado de máquina, pois fornece uma estimativa mais robusta do desempenho do modelo em dados não vistos. Em problemas de previsão, como séries temporais, é possível verificar se o modelo é capaz de generalizar bem e lidar com variações e padrões temporais presentes nos dados. Para realizar a validação cruzada neste projeto foi escolhido a Sliding TimeSeriesSplit. A Sliding TimeSeriesSplit divide o conjunto de dados em conjuntos de treinamento e teste de forma sequencial, permitindo-os percorrer os dados de maneira conjunta, garantindo que as informações futuras não sejam utilizadas no treinamento. Essa abordagem considera a natureza temporal dos dados e permite verificar se o modelo é capaz de lidar com a dinâmica e variação temporal dos dados de forma adequada.

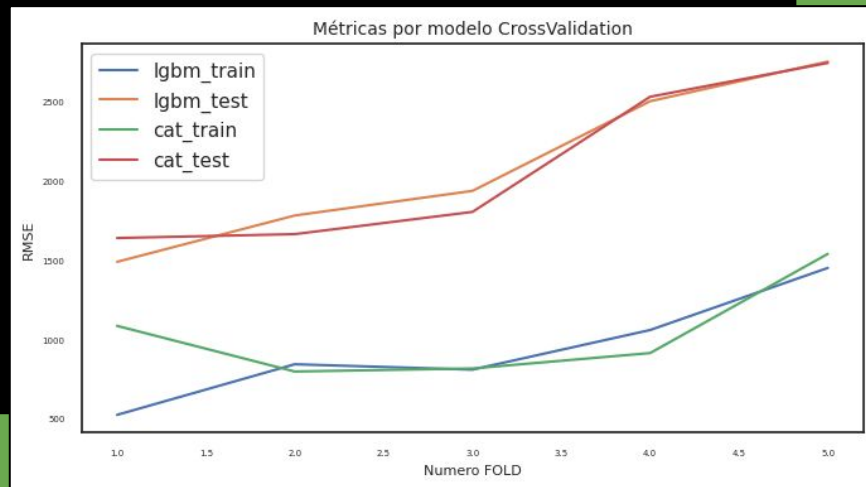


Modelo e Previsões

Cross Validation e Hyperparameter Fine Tunning

Hiperparâmetros controlam aspectos como a complexidade do modelo, regularização, taxa de aprendizado e outras configurações que influenciam seu desempenho. Para ajustar esses parâmetros, utilizei uma abordagem chamada Bayes Search Cross Validation. Essa técnica combina a busca de hiperparâmetros com a validação cruzada mencionada anteriormente. Dessa forma, a busca pelos melhores hiperparâmetros ocorre dentro das janelas do CV, garantindo um conjunto de parâmetros mais robustos que melhor generalizem os dados.

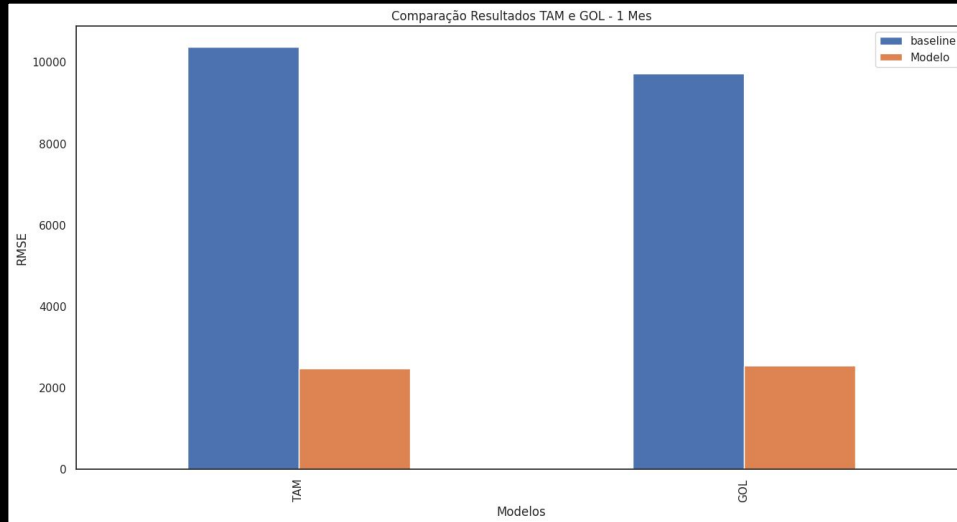
Como os dois modelos tiveram uma performance bem parecida, decido por juntar eles criando um ensemble.





Modelo e Previsões

Previsão de 1 mês

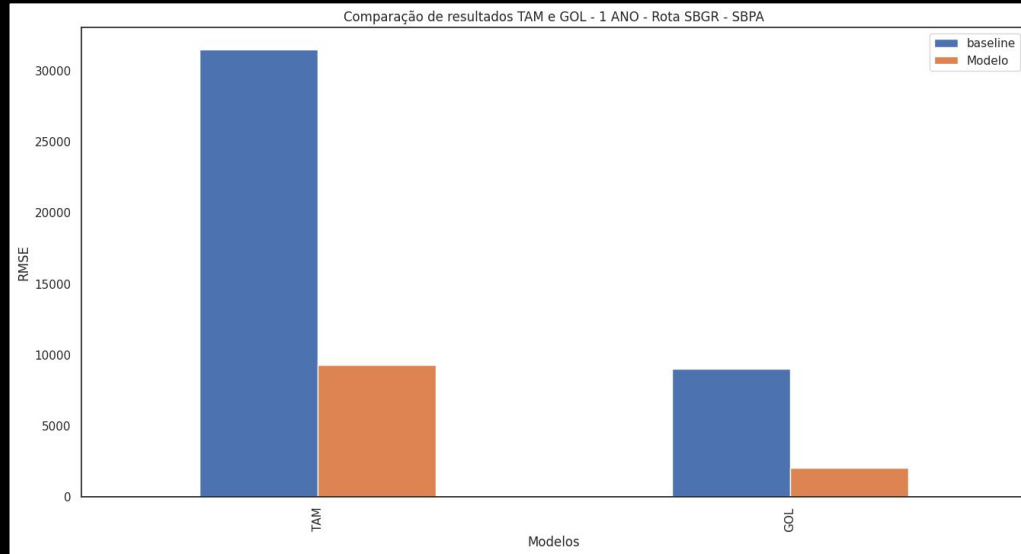


Realizando a previsão do próximo mês e comparando com um modelo baseline (média), o modelo se sobressai em ambos os casos, em mais de 4 vezes de performance.



Modelo e Previsões

Previsão de 1 ano para uma rota SBGR - SBPA

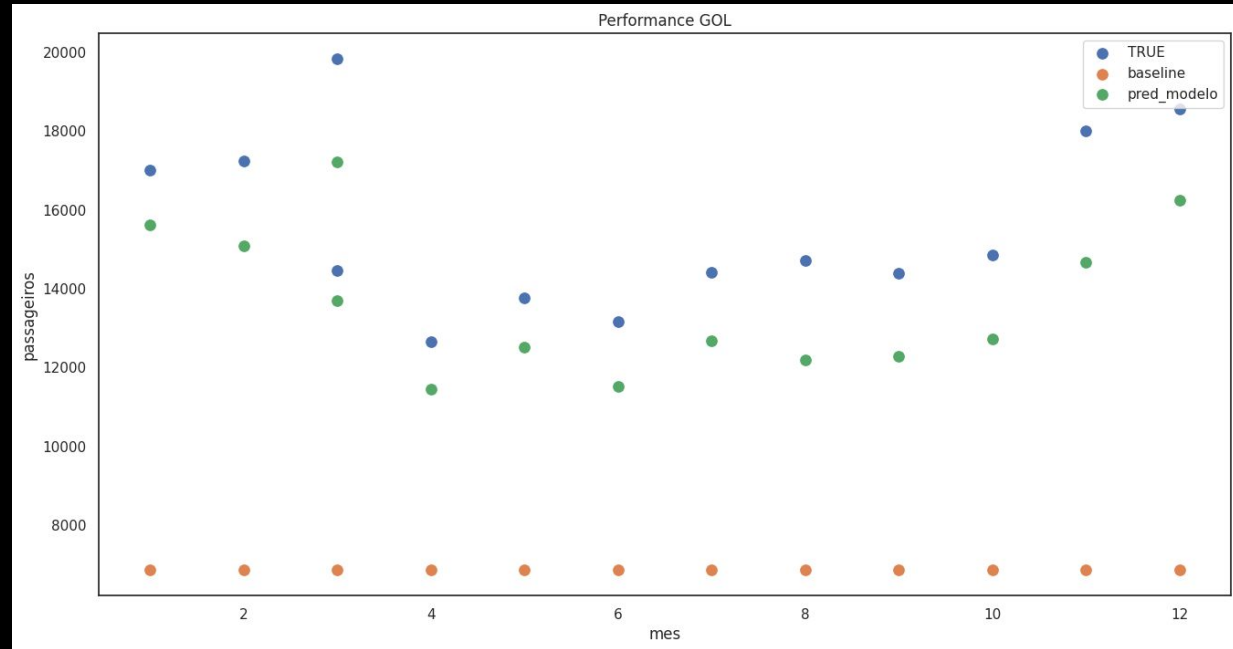


Utilizando o modelo para prever o próximo ano em uma rota específica, agora em um horizonte de tempo maior, ele também consegue se sair muito bem.

Modelo e Previsões

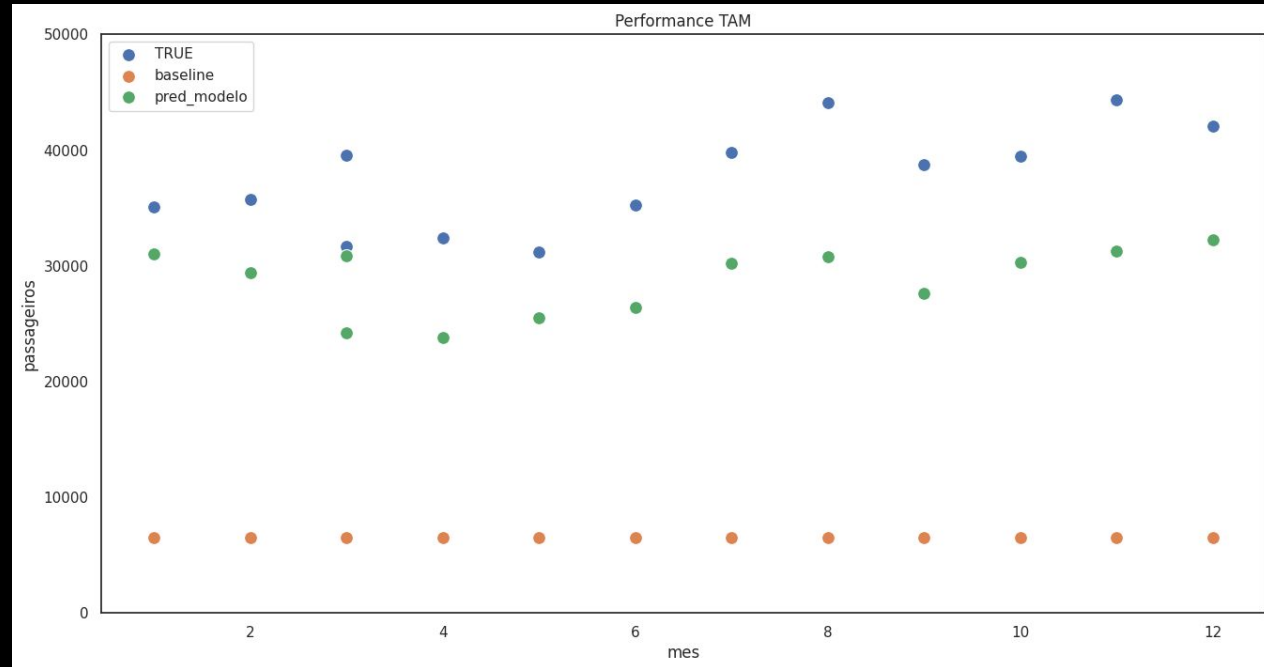
Previsão de 1 ano para uma rota SBGR - SBPA

GOL



Modelo e Previsões

Previsão de 1 ano para uma rota SBGR - SBPA



Obrigado!



<https://www.linkedin.com/in/lucas-dacunha/>



<https://jlcunha.github.io/ProjectPortfolio/>



<https://github.com/jlcunha>