Teste Técnico Cientista de Dados - A3Data

Sillas Gonzaga • 26.02.2023

Link do notebook

Google Colab:
https://colab.research.google.com/drive/1Vk3bawzrzSC4lwFiExE8NqSIv88MtYUd?usp=sharing

Sumário Executivo

Após uma breve EDA, foi detectado que uma das possíveis entregas de valor a partir dos dados fornecidos seria um modelo de séries temporais de previsão de acidentes.

Foram testados dois modelos: Suavização Exponencial e ARIMA. Ambos obtiveram desempenhos semelhantes.

Ambos modelos não tiveram um bom desempenho preditivo.

EDA

EDA

Existem muitas colunas categóricas nos datasets que poderiam ser exploradas em uma EDA:

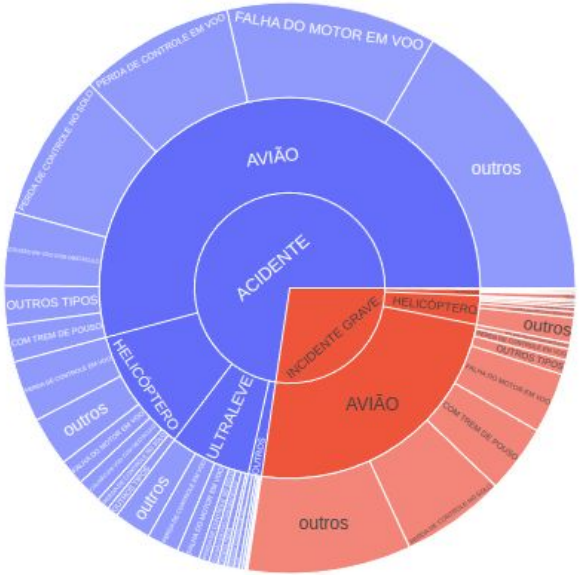
1. Existe uma associação entre tipo de motor e tipo de acidente?
2. Existem aeroportos (coluna origem_voo) mais problemáticos e propensos a acidentes?
3. Aviões mais antigos são mais propensos a acidentes?
4. Quais fases de operação tem mais risco de acidente?

EDA

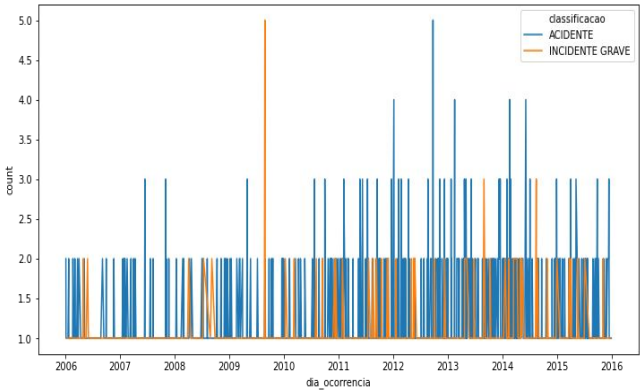
Contudo, para responder adequadamente essas perguntas, os dados fornecidos não são suficientes. Precisaríamos de um dataset também de operações que não tiverem acidentes para podermos modelar uma propensão a acidentes. Portanto, os dados fornecidos limitam as possibilidades de análise.

Como existe uma variável de dia da ocorrência, entendo que umas das entregas de maior valor seria construir um modelo de previsão de acidentes no tempo futuro. Este será o foco deste relatório

EDA

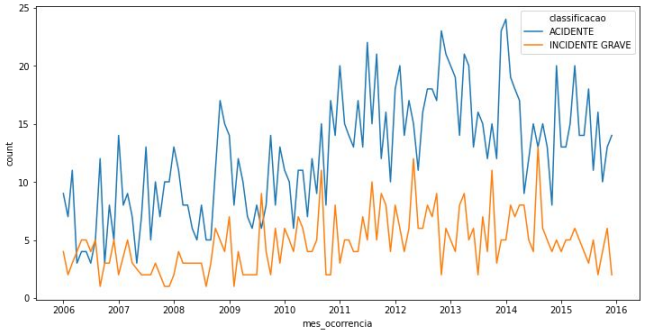

ACIDENTE
INCIDENTE GRAVE
AVIÃO
HELICÓPTERO
ULTRALEVE
FALHA DO MOTOR EM VOO
PERDA DE CONTROLE EM VOO
outros
OUTROS TIPOS
COM TREM DE POUSO
outros

EDA



O gráfico mostra que (felizmente) há pouquíssimos acidentes por dia, sendo o máximo valor apenas 5. Fazer um modelo nessas condições seria difícil, portanto os dados serão agregados por mês.

EDA

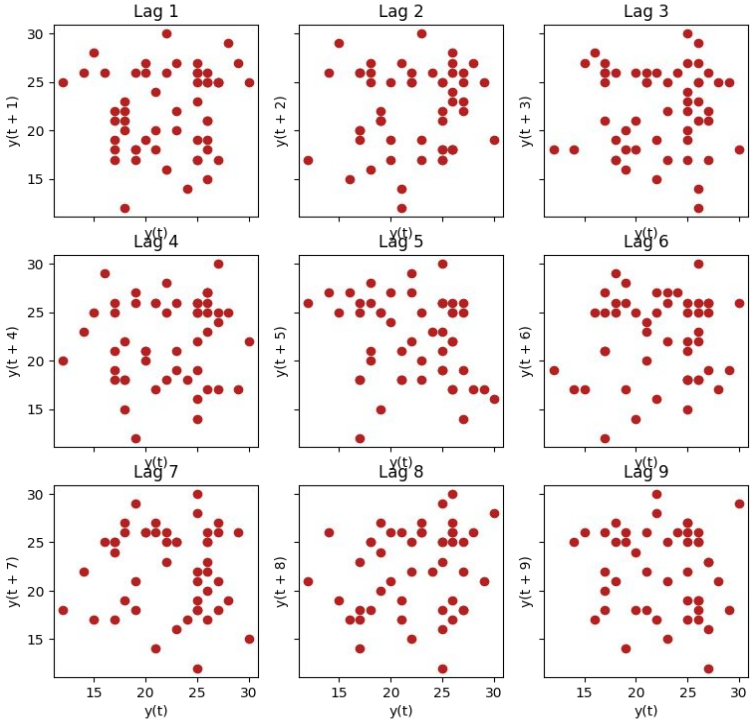


O gráfico mostra que mesmo agregando por mês a quantidade de incidentes graves não é alta, raramente ultrapassando a marca de 10 por mês.

Por isso, faz sentido agregar as duas classificações e fazer um único modelo de séries temporais mensais.

Modelagem

Análise de autocorrelação



Não há autocorrelação na variavel de acidentes por mês.

Modelos escolhidos

Foram escolhidos dois modelos para teste: Suavização Exponecial e Arima.

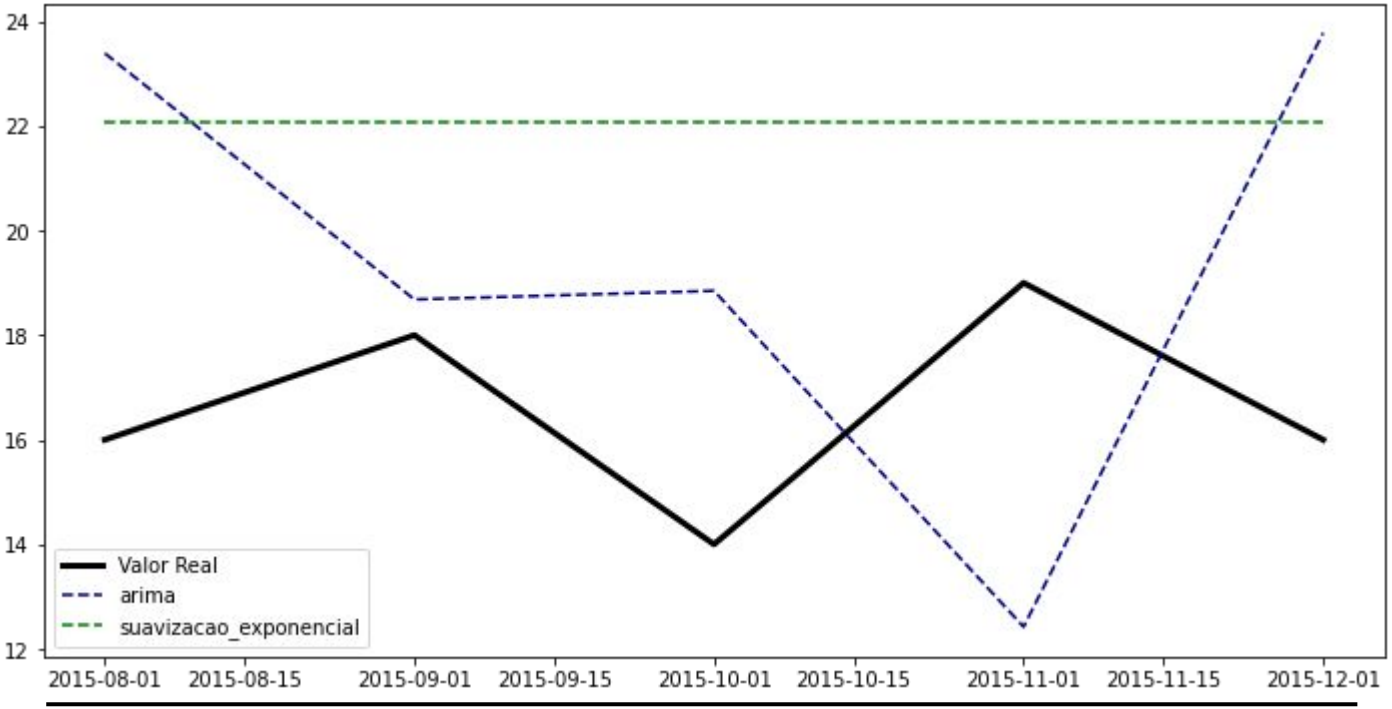
Foi construída a classe **ModelMetadata** para armazenar todas as informações importantes sobre o modelo e também para rodar as funções necessárias para ajustar o modelo e obter as previsões.

```
class ModelMetadata:
    """
    Classe para armazenar metadados do modelo e realizar previsões.

    Args:
        nome_modelo (str): Nome do modelo a ser utilizado. Deve ser "arima" ou "suavizacao_exponencial".
        base_treino (pd.Series): Série temporal contendo a base de treino do modelo.
        base_teste (pd.Series): Série temporal contendo a base de teste do modelo.

    Raises:
        ValueError: Caso o nome do modelo fornecido não seja suportado.

    Attributes:
        base_treino (pd.Series): Série temporal contendo a base de treino do modelo.
        base_teste (pd.Series): Série temporal contendo a base de teste do modelo.
        nome_modelo (str): Nome do modelo a ser utilizado. Deve ser "arima" ou "suavizacao_exponencial".
        modelo_ajustado (modelo ajustado): Modelo ajustado de acordo com a base de treino.
        forecast (pd.Series): Previsões geradas pelo modelo.

    Methods:
        construir_modelo(): Ajusta o modelo aos dados de treino.
        rodar_previsao(): Gera previsões utilizando o modelo ajustado.
        calcular_mape(): Calcula o Mean Absolute Percentage Error (MAPE) das previsões geradas.
        plot_previsao(cor_linha): Plota o gráfico das previsões geradas pelo modelo.
    """
```

Resultados:

MAPE ARIMA: 33,56%

MAPE Suavização Exponencial: 34,47%


24
22
20
18
16
14
12
2015-08-01
2015-08-15
2015-09-01
2015-09-15
2015-10-01
2015-10-15
2015-11-01
2015-11-15
2015-12-01
Valor Real
arima
suavizacao_exponencial

Conclusões

Ambos modelos tiveram desempenho insastisfatório, o que sugere o uso de outros modelos antes de assumir a imprevisibilidade dessa série temporal.