Guião para apresentação DAA

Slide 1 – Capa [30 s]: Para reduzir o tempo, usamos este slide da capa para fazer logo uma introdução, assim já poupamos um slide. Então, este slide serve **de introdução dos 2 datasets**. Devem dizer algo como:

- O presente trabalho tem o objetivo de conceber e otimizar modelos de Machine Learning;
- O trabalho teve essencialmente duas fases: uma primeira de adequação dados e construção de modelos preditivos, e, uma segunda fase, de tuning destes modelos;

Slide 2 – Metodologia [15 s]: Aqui é **só** para dizer a metodologia e porque escolhemos esta:

- A Metodologia seguida foi o CRISP-DM, já que era bem conhecida por todo o grupo e ambos os datasets apresentam uma certa aplicabilidade à Indústria;

DATASET COMPETIÇÃO

Slide 3 – Data Understanding [1 min]: Aqui descrevem de forma geral o que fizemos para compreender os dados do dataset:

- Describe Data:
  - Começamos por perceber o tipo de dados e algumas informações estatísticas;
  - O objetivo do problema é prever o nível de incidentes rodoviários, através de features sobre as condições meteorológicas e na via, na tabela ao lado temos alguns exemplos de atributos;
- Explore Data:
  - Observamos que:
    - A coluna avg precipitation e city_name tinham o mesmo valor em todos os registos;
    - Apenas a coluna (affected roads) tinha missing values;
    - Algumas colunas apresentavam dados razoavelmente distribuídos normalmente, outras tinham skewness (delay in seconds, avg atm pressure, avg humidity);
    - Detetou-se a presença de outliers.

Slide 4 – Data preparation [1.5 min a 2 min]: Aqui devem descrever todo o processamento que fizemos, independentemente do cenário. Considera-se a globalidade:

- Ao longo dos vários cenários, efetuou-se:
  - Remoção das colunas com todos os registos iguais;
  - Transformação de colunas categóricas para numéricas com label encoding;
  - Tratamos a coluna affected_roads por dois métodos diferentes: o método 1 em que apenas contamos as ocorrências diferentes e o método 2 onde criamos novas colunas com o número de estradas afetadas para cada tipo;
  - Divisão da data e hora nas suas componentes;
  - Tratamos outliers por 2 métodos, um com deteção por IQR ou MAD e substituição/remoção (média, moda, mediana) e, outro em que fazíamos a mesma coisa apenas para as features distribuídas normalmente. As que não tinham skewness aplicamos a transformação logarítmica, reduz o impacto de outliers, por ser menos sensível a estes.
  - Criação de features como a estação, se é feriado ou não e se é fim de semana ou não;
  - Normalização usada foi a de 0-1

Slides 5 e 6 – Modeling & Evaluation [1.5 min]: Aqui é descrever os modelos usados e os resultados obtidos:

- Realizamos vários cenários e para cada um verificamos se houve melhoria ou não;
- Recorremos primeiro a técnicas de autoML para ter uma ideia melhor de quais modelos seguir;
- Realizamos uma experimentação depois para os top-6 modelos, com o processamento que referimos antes;
- Verificamos os top-2 modelos e fizemos um processamento individual para cada um;
- As métricas utilizadas foram a Acc, F1 Score, a matriz de confusão e Curva ROC.
- Foram comparando cada cenário com estas métricas e mantínhamos as alterações se o modelo experimentasse uma melhoria.

- No último cenário, como se observa, o RF, sem tuning obtém uma accuracy de 0.94, assim como o LGBM. Referir que para o LGBM usamos Cross Validation (CV – tem na imagem para não se esquecerem).
- Após encontrar os top-2 modelos, realizamos tuning com Random Search e Grid Search;
- Encontrando estes dois modelos finais.

DATASET GRUPO

Slides 7 – Data Understanding [1 min]:

- Maior ocorrência de casos para quando a ocupação é 0;
- Analisou-se a ocupação ao longo do dia, de acordo com as medições de sensores diferentes;
- A maioria dos dados apresentava skewness;
- Existia uma grande quantidade de outliers;
- Averiguamos a correlação entre as várias features;
- Ao lado, temos exemplos de algumas features e abaixo a distribuição anormal;

Slide 8 – Data preparation [1.5 min a 2 min]:

- Detetamos desbalanceamento, experimentamos algoritmos de data augmentation;
- Realizamos seleção de feature através do teste de hipóteses e com wrapper methods;
- Feature engeneering com as datas;
- Análise bivariada de outliers;
- Utilizamos MinMaxScaler para reduzir a influência da escala;

Slide 9 – Modeling & Evaluation [1 min]:

- Também realizamos vários cenários e verificamos se a alteração provocava uma melhoria;
- Fez-se um tratamento indicidual para os top-1 modelo (o RF);
- Também utilizamos autoML para ganhar uma primeira ideia;
- Desta vez, usámos métricas adaptadas ao nosso tipo de dados: O RMSLE (a raiz quadrada do logaritmo do erro médio ao quadrado entre as previsões do modelo e os valores reais. É uma medida do desvio médio das previsões do modelo em relação aos valores reais, mas é mais robusta em relação a outliers do que o RMSE. Utilzamos mais esta métrica porque ser uma medida mais estável do desempenho do modelo, que tem em conta algum desbalanceamento que possa estar presente. Após o tratamento de outliers, já se poderia recorrer ao R2 score apenas), o R2 não é sensível a outliers.
- No gŕafico, observa-se a presença de overfitting, e um melhor desempenho do RF em relação aos restantes modelos.

Slide 10 – Modeling & Evaluation [15 s]:

- Como forma de reduzir o impacto do overfitting, recorremos a técnicas de CV com STK;
- Observa-se uma melhoria do desempenho para 93.7, do RF.

Slide 11 - Conclusão [30 s]:

- Observando este gráfico, é possível observar que as fases mais demoradas foram as de processamento, modelação e de exploração de dados;
- Consideramos que o trabalho é completo, tendo em conta as várias técnicas utilizadas nas várias fases.
- Obrigado.

Considerando o máximo tempo: 30, 15, 60, 120, 60, 30, 60, 120, 60, 15, 30 = **10 min**