My Submission

Daniel González Juclà

Code structure

1. **data_gathering_and_integration.ipynb:** Contains the gathering and the integration of the data from the different sources. PDF data was not considered because only a proportion could be parsed correctly in an automated way, so as it only represents a 1% of the data it has been ignored.
2. **data_preprocessing.ipynb**: This notebook contains the exploration and preprocessing of the data, to form the final training and testing datasets. The preprocessing consists on:
    2.1. **Drop** those c**olumns** that are not consider relevant
    2.2. Encode time as day of the year and year
    2.3. Scale the numerical variables in the [0,1] range
    2.4. One-hot encoding of the categorical variables
    2.5. Dimensionality reduction of the one-hot encoded columns with PCA.
3. **xgboost_modeling_pca.ipynb**: The final modeling has been carried out with an XGBoost model, with has performed better than another modeled MLP. This XGBoost modeling is explained next.

XGBoost modeling

The final modeling has been done by fine-tuning an XGBoost model, following this guide.

More precisely, the model has been fine-tuned with Cross-validation and the following steps:

1. Choose a relatively **high learning rate** and determine the **optimum number of trees** for this learning rate.
2. **Tune tree-specific parameters** (max_depth, min_child_weight, gamma, subsample, colsample_bytree) for decided learning rate and number of trees.
3. **Tune regularization parameters** (lambda, alpha) for xgboost which can help reduce model complexity and enhance performance.
4. **Lower the learning rate** and decide the optimal parameters .