Short description

Predict the Crude Oil production's trend base on the previous year Crude Oil data.

Challenge context

One of the key indicators of the Natural Resources market is the Crude Oil production. Understanding the variation of the production per region helps in predicting the Oil price trends in these regions and for different qualities of the Crude.
These indicators can be very useful for SOCIETE GENERALE teams since they will help on:

Predicting the revenue generated by the Trade Finance business line.
Building prospection watch-list for the Natural Resources front-office and anticipating clients needs.
Adjusting the forecasts of the Crude Oil trade back-office's workload.

Challenge goals

The objective of this challenge is to predict the probability of increase of Crude Oil production per quarter per country based on several indicators collected during the previous year.

For any question, feel free to contact us:
pejman(.)gohari at socgen(.)com
https://www.linkedin.com/groups/8579858

Data description

The datasets contain Crude Oil data including lease condensate – excluding NGL (Liquid or liquefied hydrocarbons). All provided data is open data coming from “The Joint Organisations Data Initiative (JODI)”.

The historical data covers the declarations coming from all Crude Oil producers all over the world for the period from January 2002 to August 2016.
Every line is defined by its unique ID and contains historical information concerning the Crude Oil sector of one country during the last year. Some of the features contain aggregated information on the worldwide Crude Oil sector.

The split between Training and Test has been done, such that the most recent data has been put in the Test dataset. Countries have been anonymized in the data.
The objective is to estimate for the Test dataset the probability of the increase of the Crude Oil production for the next quarter for every line.

The Train and the Test files contain the following features:
- ID: ID of the line that contains 12 months of data of a given country and other information detailed above.
- month: Month index. There is no indication regarding the year of the collect of the data .
- country: Country index. As said above the countries have been anonymized.
and features that are given for every month of the previous year:
- closing stocks(kmt): Represents the primary stock level at the end of the month within national territories; includes stocks held by importers, refiners, stock holding organisations and governments in Thousand Metric Tons.
- exports(kmt)/Imports(kmt): Amount of Crude Oil having physically crossed the international boundaries, excluding

transit trade, international marine and aviation bunkers in Thousand Metric Tons.
- refinery intake(kmt): Total amount of oil observed to have entered the refinery process in Thousand Metric Tons.
- WTI: West Texas Intermediate Price. This value is the close price at the last business day of the month in USD.
- SumClosing stocks(kmt), SumExports(kmt), SumImports(kmt), SumProduction(kmt) and SumRefinery intake(kmt): Sums of previous features over all countries on the same period in Thousand Metric Tons.

The prefix "diff" in columns names means that the columns are the difference between the month value and the value of the previous month. The prefix of the column refers the data record month. For example, the "12_diffExports(kmt)" is the closest value from the trend that we are predicting and "1_diffExports(kmt)" is the farthest value from the trend that we are predicting.
Semicolon is the columns separator used in all provided files.

The TrainOutput file contains the Target for each "ID", where the target is either:
- 1: if the production goes up for the next quarter.
- 0: if the production goes down for the next quarter.

Submission file has to be a CSV file with the following format (The first line of the file is the header):
"ID"; "Target"
"ID10160"; xxxx
...
"ID12159"; xxxx
Where xxx is a probability (number between 0 and 1 included), for instance 0.5.

The metric used for this challenge is the AUC (area under the ROC curve)