Challenge data

Challenge Data 2016-2017

From 21/10/2016 to 01/01/2018

Challenge : Prédire la tendance de la production de pétrole brut.

Par GROUPE SOCIETE GENERALE

SOCIETE GENERALE

Description courte

Prédire la tendance de la production de pétrole brut.

Contexte du challenge

L'un des principaux indicateurs du marché des ressources naturelles est la production de pétrole brut. Comprendre la variation de la production par région aide à prédire l'évolution du prix du pétrole dans ces régions et cela pour les différentes qualités du brut.

Ces indicateurs peuvent être très utiles pour les équipes de SOCIETE GENERALE, car elles nous permettront de faire :

1/ Une prévision des revenus générés par le métier de Trade Finance.

2/ Une mise en place d'une liste de prospection pour le front-office NAT et anticipation des besoins des clients.

3/ Un ajustement des prévisions pour la charge de travail du back-office traitant les produits de financement de pétrole brut.

L’objectif du challenge

L'objectif de ce défi est de prédire la probabilité d'augmentation de la production de pétrole brut par trimestre par pays à partir de plusieurs indicateurs recueillis au cours de l'année précédente.

Description des données

Les ensembles de données fournis contiennent des données sur le pétrole brut, y compris lease condensate - à l'exclusion des NGL (hydrocarbures liquides ou liquides).
Toutes les données fournies sont des Open Data provenant de "The Joint Organizations Data Initiative (JODI)".

Les données historiques couvrent les déclarations de tous les producteurs de pétrole brut du monde entier pour la période allant de janvier 2002 à août 2016.

Chaque ligne est définie par son identifiant unique et contient des informations historiques concernant le secteur du pétrole brut d'un pays au cours de la dernière année. Certaines de ces caractéristiques contiennent des informations agrégées sur le secteur mondial du pétrole brut.

La séparation entre le training et le test a été faite, de sorte que les données les plus récentes ont été mises dans le jeu de données de test.

Les pays ont été anonymisés dans les données.
L'objectif est d'estimer pour l'ensemble des données de test la probabilité de l'augmentation de la production de pétrole brut pour le trimestre suivant et cela pour chaque ligne.

Les fichiers Train et Test contiennent les fonctions suivantes :
- ID : ID de la ligne qui contient 12 mois de données d'un pays donné et d'autres informations détaillées ci-dessus.
- month (mois) : Indice de mois. Il n'y a aucune indication concernant l'année de la collecte des données.
- country (pays) : Indice des pays. Comme indiqué ci-dessus, les pays ont été anonymisés.

Et les caractéristiques qui sont données pour chaque mois de l'année précédente :

- closing stocks (stocks de fermeture) (kmt) : représente le niveau de stock primaire à la fin du mois dans les territoires nationaux; Comprend les stocks détenus par les importateurs, les raffineurs, les organisations boursières et les gouvernements en milliers de tonnes.
- Exports (exportations) (kmt) / Imports (Importations) (kmt) : Quantité de pétrole brut ayant traversé physiquement les frontières internationales, à l'exclusion du commerce de transit, des bunkers internationaux de la marine et de l'aviation en milliers de tonnes métriques.
- refinery intake (consommation de raffinerie) (kmt) : quantité totale de pétrole observée pour entrer dans le processus de raffinage en milliers de tonnes.
- WTI : West Texas Intermediate Price. Cette valeur correspond au cours de clôture du dernier jour ouvrable du mois en USD.
- SumConsing stocks (kmt), SumExports (kmt), SumImports (kmt), SumProduction (kmt) et SumRefinery intake (kmt) : Sommes des features précédentes sur tous les pays sur la même période en milliers de tonnes.

Le prefix "diff" signifie que les colonnes représentent la différence entre la valeur du mois et celle du mois précédent. Le préfixe de la colonne fait référence au mois d'enregistrement. Par exemple, "12_diffExports (kmt)" est la valeur la plus proche de la tendance que nous essayons de prédire et "1_diffExports (kmt)" est la valeur la plus éloignée de la tendance que nous essayons de prédire.

Le point-virgule est le séparateur de colonnes utilisé dans tous les fichiers fournis.

Le fichier TrainOutput contient la cible pour chaque «ID», où la cible est soit:
- 1: si la production augmente pour le trimestre suivant.
- 0: si la production baisse pour le trimestre suivant.

Le fichier de soumission doit être un fichier CSV au format suivant (la première ligne du fichier est l'en-tête):
"ID"; "Target"
"ID10160"; xxxx
...
"ID12159"; xxxx

Lorsque xxxx est une probabilité (nombre compris entre 0 et 1 inclus), par exemple 0.5

La métrique utilisée pour ce challenge est l'AUC (aire sous la courbe ROC).

Fichiers / Jeux de données

Training output file : *A file containing a series of responses for the training part*

Training input file : *A set of files necessary for the training part.*

Testing input file : *A set of files necessary for the testing part.*

GROUPE SOCIETE GENERALE

Société Générale est l'un des tout premiers groupes européens de services financiers. S'appuyant sur un modèle de banque universelle, le Groupe allie solidité financière et stratégie de croissance durable, afin de mettre sa performance au service du financement de l'économie et des projets de ses clients.

http://www.societegenerale.fr/