Données prises en compte dans cette étude Pourcentage de 15-24 dans la population Taux d'internautes

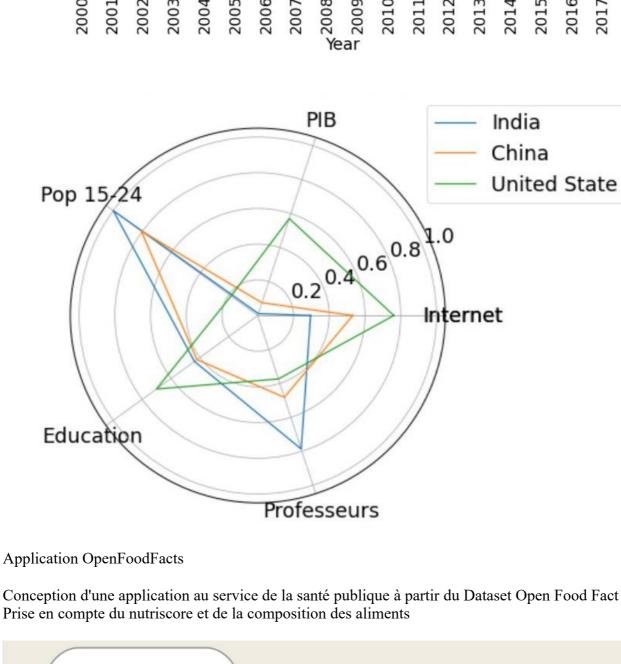
PIB normalisé Implication des jeunes dans l'éducation

Implication du pays (état) dans l'éducation

- Réalisation de divers visuels d'aide à la décision

80

```
60
Internet
                                                                     India
                                                                     China
                                                                     United States
                                                                     Turkey
                                                                     France
  20
                                                                     Brazil
                                                                     Japan
                                                                     Germany
    0
                                        Year
                                              2009
                                 2006
                                                  2010
                         2004
                             2005
                                     2007
                                                               2013
                                                      2011
                                                          2012
                                        PIB
                                                                  India
                                                                  China
                                                                  United State
    Pop 15/24
                                      0.2 0.4 0.6 0.8
                                                         1.0
                                                        Internet
```



latitude, longitude année de construction

comme le montre la matrice de Kendal

NumberofBuildings

PropertyGFATotal

LargestPropertyUseTypeGFA

condLargestPropertyUseTypeGFA

ThirdLargestPropertyUseTypeGFA

ENERGYSTARScore

SiteEnergyUse(kBtu)

TotalGHGEmissions

ou non) et plusieurs modèles testés :

Random Forest ElasticNet

Démarche générale

SVM: support vector machine

Mon repas

100 g steak haché 100 g haricots verts 100 g de salade

+ ajoutez un aliment

Score Montignac

Historique Apport glucidique, lipidique et protéine Apport en fibre, sodium Historique

Saisie des repas

Liste des repas

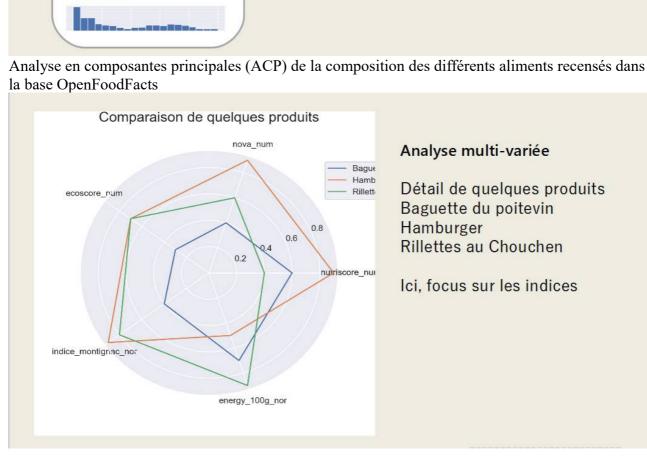
Score Montignac / Kéto

Une application qui aide à suivre au

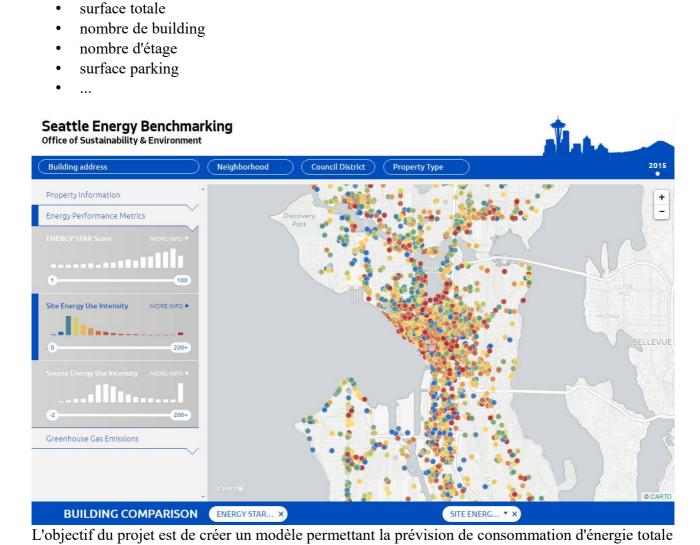
Liste des produits consommés

Divers indicateurs sur les repas

quotidien le régime Montignac ou Cétogène



Emission de CO2 et consommation d'énergie Réalisation d'une application de prévision de consommation d'énergie et d'émission de CO2. L'application se base sur le Dataset "Seatle Energy Benchmarking" qui recense la consomation d'énergie et les émissions de CO2 de nombreux batiments de la région de Seatle. Ce Dataset recense de nombreuses données concernant ces batiments :



et d'émission de CO2 des futures constructions. Ces deux grandeurs sont fortement corrélées

0.6

- 0.4

- 0.2

-0.2

- -0.4

-0.6

Matrice de corrélation de Kendall entre les différentes indicateurs

Enregistrement des données de performance dans un dataframe pour comparaison des différents modèles pour les différents Dataset Score (R2), MSE, RMSE, RMSE normalized, MAE, medianAE

ENERGYSTARScor

Quatre Datasets nettoyés ont été tiré du Dataset original selon quatre variantes (passage par le log

Préparation des Dataset pour ESU (EnergySiteUse) et GHG Avec et sans le log1p, avec et sans le EnergyStarScore Séparation dataset en X_train, y_train, X_test, y_test Création d'une baseline avec DummyRegressor(strategy='mean')

Définition des dictionnaires d'hyper-paramètres pour GridSearchCV Pour chaque modèle testé (ElasticNet, SVR, RandomForest)

Training des modèles sur X_train, y_train et évaluation des Performances sur X_test, y_test, mesure du temps de prédiction

Score

0,00

0,58

0,62

-0,13

0,30

0,28

0,23

0,76

0,84

-0,11

0,74

0,78

0,80

0,57

0,75

nRMSE

1,4390

0,9303

0,8812

1,5300

0,9022

0,9606

0,8785

0,6605

0,5514

1,4324

7,6724

0,5590

1,0169

0,9453

0,6216

Distribution des Erreurs

TimePredict

0,0030

0,0211

0,0450

0.6639

0,0092

0,1933

0,2718

0,0199

0,2232

0,3842

0,0077

0,1702

0,1416

0,0112

0,3262

RandomForest sur X3 9 SVM sur X3 10 ElasticNet sur X4 11 RandomForest sur X4

RandomForrest sur X3

Target sans le log1p

12 SVM sur X4

DummyRegressor

RandomForest sur X1

RandomForest sur X2

ElasticNet sur X1

ElasticNet sur X2

ElasticNet sur X3

SVM sur X1

SVM sur X2

Modèle

13 SGDR sur X1 14 Stacking avec KNNRegressor Modèle final retenu

Synthèse des résultats

3

5

6

7

160 140 Score 120 1.5 Predictions 10 80 0,84 60 40 20 0.0 1.0 1.5 0.8 1.0 le7 2.0 3.0 Segmentation clients sur une plateforme de e-commerce L'objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles. Le projet exploite les données du site de commerce Olist avec notamment les données suivantes : ID unique du client Nombre d'achat du client dans chaque catégorie, et le montant Nombre de produits achetés au total Nombre de jours depuis le dernier achat Montant du panier moyen Moyen de paiement Utilisation des facilités de paiement ou pas

L'objectif est de fournir à l'équipe d'e-commerce une segmentation des clients qu'elles pourront

{'max_features': 'sgrt', 'min_samples_leaf': 2, 'n_estimators': 40}

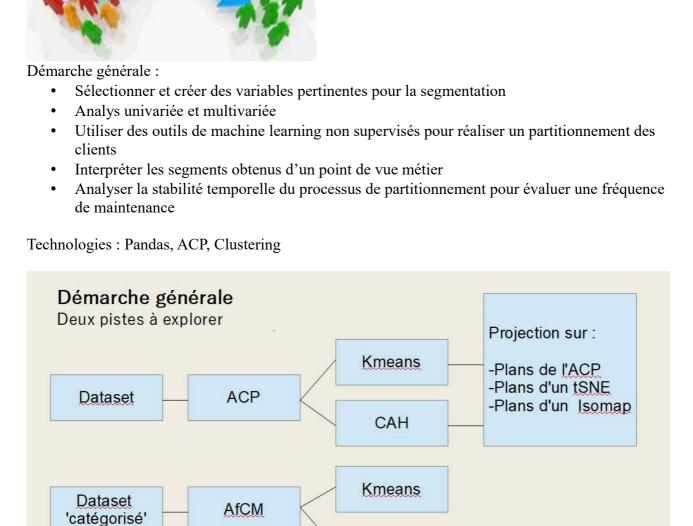
Features avec le log1p, avec le EnergyStarScore

Random Forrest sur X3

Numéro du mois du dernier achat

utiliser au quotidien pour leurs campagnes de communication.

Moyenne des notes laissées Moyenne des délais de livraison



CAH

Préparation de différents Dataset (X1, X2, X3, X4, X5) Feature enginering OneHotEncoder, StandardScaler

Test de Kmeans pour différentes valeurs de n_clusters, ...

Test de CAH pour différentes valeurs de n_clusters, linkage ['complete', 'average', 'single'] Affinity ['euclidian', 'manhattan', 'cosine']

Elbow method, Silhouette, Davies_bouldin

Projections : PCA, tSNE, Isomap Analyse visuelle du résultat

Le corpus de texte rassemble les transcripts des sketchs de 12 comiques Américain. L'objectif est de

Bill Burr

Bo Burnham

faire une analyse de type "Sentiment Analysis" sur ce corpus pour mettre en évidence les

Ce projet met en oeuvre les techniques courantes du Natural Language Processing (NLP)

Document-Term Matrix - word counts in matrix format

Anthony Jeselnik

Sentiment Analysis

différences entre les personnes.

Tockenisation, Stop words

Regex

Stemming / lemmatization

Technologies: Nltk, Textblob.sentiment

Corpus de mots pour chaque acteur :

Ali Wong

F-Word et S-Word (Bad words)

Ali Wong

Louis C.K.

Anthony Jeselnik

John Mulaney

20

Joe Rogan

Dave Chappelle

Bill Burr

0.02

0.04

Anthony Jeselnik

les données images selon des techniques comme ORB, SIFT

comme extracteur de features avec un XGBOOST

Evolution dans le temps de cette analyse (les sketchs dans le corpus commencent par les plus

Mike Birbiglia

Hasan Minhaj

40

Textblob.sentiment nous donne deux axes, la polarité et la subjectivité :

Bo Burnham Ricky Gervais

60

50

40

20

10

0

Ò

Sentiment-Analysis

opinion.

0.56

0.54

0.52

0.50

0.48

0.46

0.3

0.00

anciens jusqu'au plus récent)

Ali Wong

les données textuelles

Pré-traitement

Des données

textuelles

Pré-traitement

Des données

visuelles

Démarche pour les données textuelles

Démarche générale

-- Opinions -->

-- Facts --

Number of S Words

Exploratory Data Analysis (EDA)

Résultat : on arrive à séparer les clients en trois classes dans le tSNE

Louis C.K. EDA: exploration des données Diversité du vocabulaire et vitesse Number of Unique Words Number of Words Per Minute Joe Roga 1200

Number of Bad Words Used in Routine

Dave Chappelle

Number of F Bombs

Polarity: How positive or negative a word is. -1 is very negative. +1 is very positive. Subjectivity: How subjective, or opinionated a word is. 0 is fact. +1 is very much an

Sentiment Analysis

Jim Jefferies

Anthony Jeselnik

Louis C.K.

0.06

<-- Negative ----- Positive -->

Bo Burnham

Ricky Gervais

Ali Wong John Mulaney

0.08

Bill Burr

Hasan Minhaj

0.3

0.10

Essais de

Classification

Non supervisée

Essais de

Classification Transfert learning

Bo Burnham

0.12

Bill Burr

Joe Rogan

140

Jim Jefferies

120

Mike Birbiglia

0.2 0.2 0.2 0.2 0.1 0.1 0.1 -0.1 -0.1 -0.1 Jim Jefferies Dave Chappelle Hasan Minhaj Joe Rogan 0.3 0.3 0.2 0.2 0.1 0.1 0.1 -0.1 -0.1 -0.1 -0.1 John Mulaney Louis C.K Mike Birbiglia Ricky Gervais 0.3 0.3 0.2 0.2 0.2 0.1 0.1 0.1 0.1 0.0 0.0 -0.1 -0.1 -0.1 Classification automatique d'objet de consomation On dispose des données d'un catalogue de vente d'objets de consomation (textes et images). L'objectif est de faire une catégorisation automatique selon 7 classes (furnishing, watches, kitchen, ...) en se basant sur les données textuelles et visuelles Axes d'exploration :

les données images en utilisant un CNN pré-entraîné : RestNet50, VGG16, YOLO, VGG16

Etude des

fréquences

Essais de

Classification

Non supervisée

Extraction du corpus à partir du Nom et de la Description

Ponctuation, Tokenisation (unigram, bi-gram) Stop words Lematization (WordNetLemmatizer) (POS_tags) Bags of words (CountVectoriser et TF-IDF)

Réduction de dimension

Projections: PCA, tSNE Analyse visuelle du résultat

Entraînement des modèles (3 epochs)

Démarche pour les données images

Resnet50: 45% d'accuracy

VGG16 + XGBOOST : 81%

VGG16:84%

Résultat:

YOLO

Détecteur d'objets dans une image avec Yolo -Détection d'objets **■** Image -Contours

Conclusions : seul le CNN transfer learning donne des résultats corrects

Samples of majority class

Original dataset

Réalisation d'un training set de 840 images de 224*224

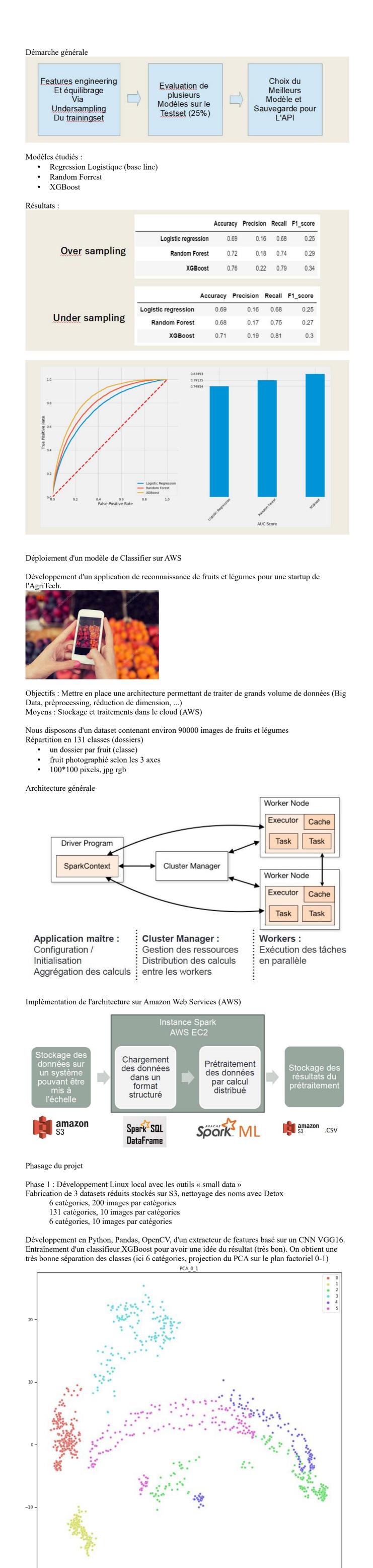
Evaluation des modèle avec le Testset de 210 images

Scoring pour l'octroi de crédit Nous disposons d'une base de données de 500000 clients ayant contracté un prêt auprès d'une

> Copies of the minority class

Original dataset

banque, certains (une minorité) ont connu des incidents de paiement, d'autres non. La base contient 240 données pour chaque client (âge, revenu, temps de travail, montant du prêt, situation de famille, nombre de personnes à charge, ...) L'objectif est de développer une appliction de scoring en ligne (et une API hébergée chez Heroku) permettant si on peut accorder un prêt à une personne en limitant le risque pour la banque. Fort déséquilibre : TARGET 91,9% de clients 0 8,1% de clients 1 100000 150000 count Il faut rééquilibrer le jeu de données : - undersampling - oversampling - smote Undersampling Oversampling



Boto 3 Amazon EC2

Phase 3 : Portage sur machine Amazon EC2

-10

On utilise les 3 même datasets réduits stockés sur S3

Test sur machine t2.xlarge, 16Go de RAM, 100Go DD: OK

RDD

Transformation du Notebook Jupyter en script Python et Pyspark en utilisant au maximum les

Node

Driver Node

Phase 2 : Développement Linux local avec les outils Spark et PySpark

30

-20

spark=sparkSession.builder ...

df_img = spark.read.format ...

Mise en œuvre de Jupyter via SSH Test sur machine t2.micro: échec

sc=spark.sparkContext

Phase 4: Cluster EMR

primitives PySpark

Node

Exécution sur un cluster EMR

Présentoir de tête de gondole