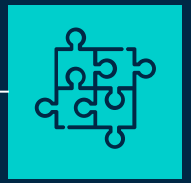
Etude de marché et Estimation des prix

Esrin ERDEM



01

PROBLEM

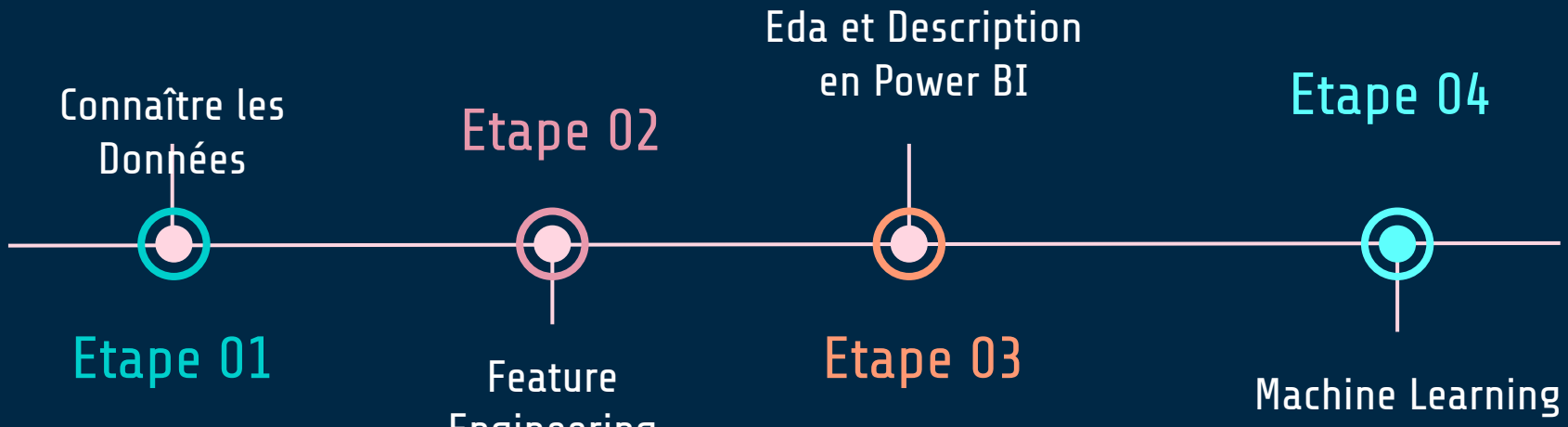
Le client, le Domaine des Croix, cherche à **définir le prix** de ses bouteilles de vin **pour le marché américain**.

02

OBJECTIF

Prédire les meilleurs prix pour les vins souhaités pour le marché américain à l'aide Machine Learning et visualiser les données du marché mondial du vin de manière significative.

LES ÉTAPES DE PROJET

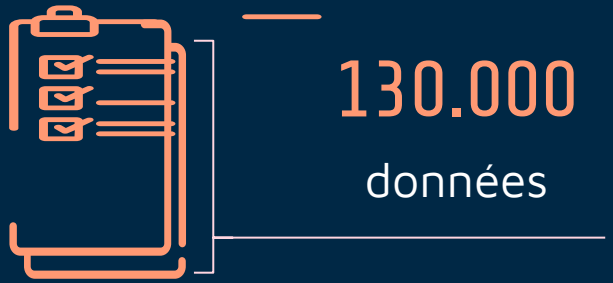


LES OUTILLES: PYTHON, POWER BI

CONNAITRE LES DONNEES

01

46 pays



En particulier, la note définie aux vins est une donnée comprenant le prix des vins, la description, et des colonnes telles que le nom, taster name, province etc.

FEATURE ENGINEERING

02

FEATURE ENGINEERING

- Étant donné que les données de titre contenant le nom des vins contiennent les informations sur le millésime, le millésime des vins a été trouvé à l'aide de Regex.

- De plus, le millésime a été soustrait de l'année en cours et l'âge du vin a été retrouvé.

- le millésime des vins souhaités à estimer: 2014-2019
- le millésime de données dont nous disposons: 1503-2021 (surtout ces dernières années)

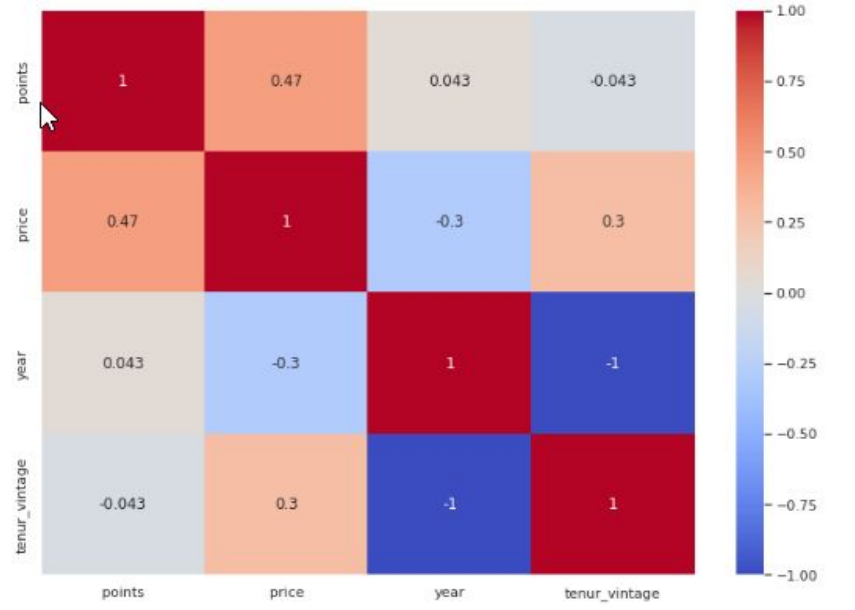
Exploratory Data Analysis(EDA) et Description sur power bi

03

EDA(Exploratory Data Analysis)

1. Les valeurs manquantes et les valeurs aberrantes sont trouvées.
2. Les valeurs vides dans Millésime ont été remplies de median.
3. Certaines variables qui avaient trop de valeurs vides et qui ont été décidées pour ne pas fonctionner pour nous en machine learning ont été supprimées. ('region_2','taster_twitter_handle','designation','taster_name','region_1')
4. Toutes les lignes restantes avec des valeurs vides ont été supprimées.

Analyse de corrélation entre les variables



Price- points:47 % de corrélation
Price- year:30 % de négatif corrélation
Price - tenure_vintage: 30 % de corrélation

Machine Learning

04

Part 1

SCORE

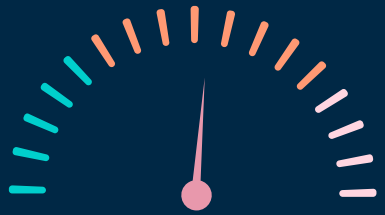
%44

L'estimation initiale des prix a été effectuée en utilisant uniquement des points et les durées des millésimes (tenure_vintage) .
L’estimation a été faite avec Random Forest

Part 2

SCORE

%50



Le modèle a été établi en fonction du point, la durée de millésime , du top 10 des pays et du top 10 des variétés.
L'estimation a été faite avec la régression KNN.

Next step

- Dans l'estimation, seul un score de 50% a été obtenu. Je n'ai pas pu utiliser certaines informations dans les données, il y avait trop de valeurs vides. On peut trouver ces informations par web scraping Ainsi, une estimation plus précise peut être faite.

- En faisant du NLP, les scores de prévision des prix peuvent être vérifiés.

- On peut essayer aussi les autres algorithmes de Machine learning (Lightgbm, xgboost et etc)

Les limites

Ces données n'ont pas pu être utilisées dans l'estimation car il y a trop de valeurs vides notamment dans les données telles que la région, la désignation. Étant donné que la corrélation entre les données était faible, une bonne estimation n'a pas pu être trouvée.

Merci