IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

# PROJET: FOUILLE DE DONNÉES

## Google Analytics Customer Revenue Prediction


**Présentée par**

**EL YOUNSI Abdellah**
**MORSLI Omar**

**Sous la tutelle de**

**Romain Billot**
**Sorin Moga**
**Philippe Lenca**

# SOMMAIRE



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Google


la règle 80/20

coûts marketing

client’s revenue

Objectif

**Train.csv**

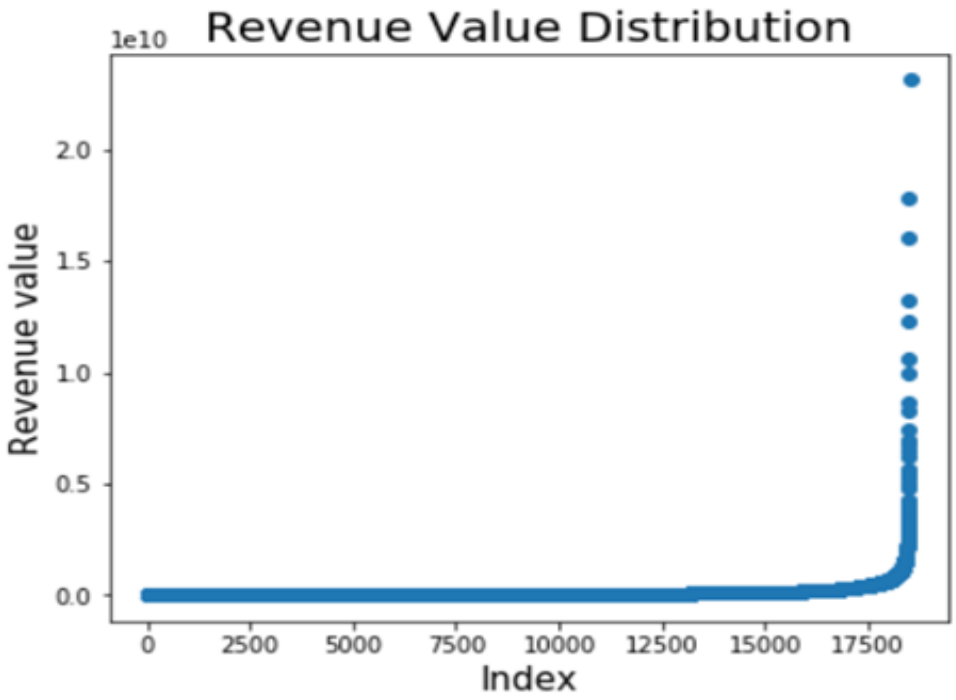
**1708337 transactions**

**Test.csv**

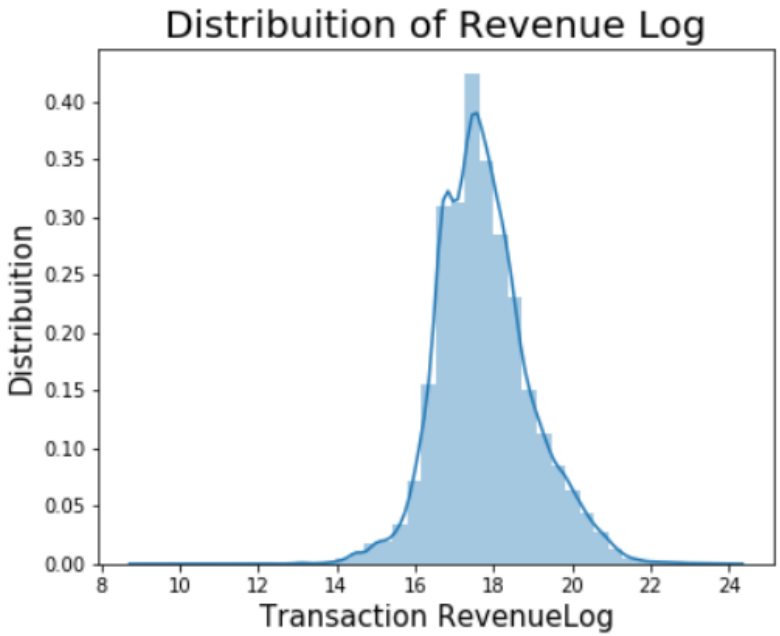
**401589 transactions**

## Qualité des Données:

- 23 colonnes constantes
- 16 variables avec plus que 40% de valeurs manquantes
- 0.7% de valeurs aberrantes

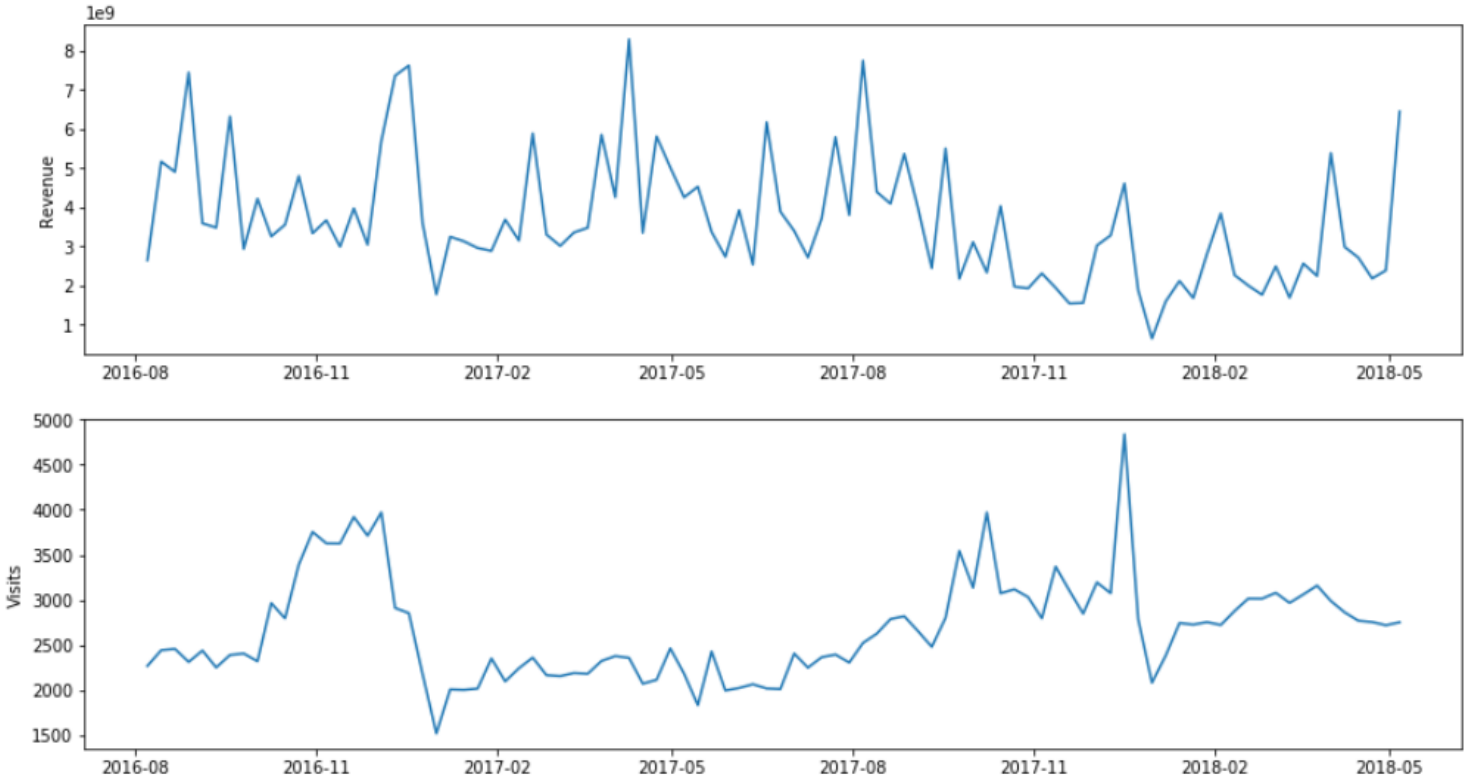



- 1.08% (18514) contribuent au revenu
- 98.91% (1689823) ne contribuent pas au revenu




**Distribution normale de log(revenu +1)**

# PARTIE 2 : DATA UNDERSTANDING (4/4)




**le nombre de visites a augmenté en Octobre et a diminué en Décembre 2016/2017.**

**Aucune relation de cause à effet entre le nombre de visites et le revenu.**

1) **sélection des variables significatives**

2) **Nettoyage des données**

| Variable | Valeur à remplacer | Valeur | Type |
|---|---|---|---|
| **totals.transactionRevenu** | missing values | 0 | Float |
| **trafficSource.adContent** | missing values | NoAdContent | String |
| **trafficSource.keyword** | missing values | NA | String |
| **totals.pageviews** | - | - | Integer |

## Transformation

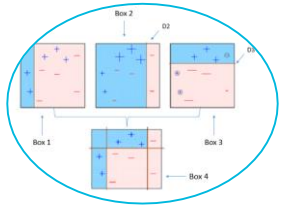
- One hot conversion : inférieur à 40 valeurs uniques
- Level conversion: supérieur à 40 valeurs uniques
- Logarithme de ‘totals.transactionRevenue’
- Normalisation : MinMaxScaler

## Les algorithmes testés

**Régression linéaire**

**Arbre de décisions**

**Random Forests**




**Gradient Boosting**

**y_client = ln(1+∑_(i≥1)⟦transaction(client,i)⟧)**

**Métrique:**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

**Résultats:**

| Modèle | RMSE |
|---|---|
| Régression linéaire | 1.8514 |
| Arbres de décisions | 1.6289 |
| Random Forests | 1.6199 |
| Gradient Boosting | 1.5922 |

## Amélioration de notre modélisation

- Introduire des modèles de forecasting.
- Changer la méthode d'encodage.
- Créer d'autres variables (Feature engineering)
- Utilisation des réseaux de neurons, meilleurs représentation des variables.

« Stacking » de modèles ?

Corrélation inter-modèles = 1

| | linearRegression | decisionTree | randomForest | gradientBoosting |
|---|---|---|---|---|
| linearRegression | 1 | 0.922581 | 0.92498 | 0.924177 |
| decisionTree | 0.922581 | 1 | 0.999227 | 0.999309 |
| randomForest | 0.92498 | 0.999227 | 1 | 0.999519 |
| gradientBoosting | 0.924177 | 0.999309 | 0.999519 | 1 |

Merci de votre attention

Avez-vous  des questions ?

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom