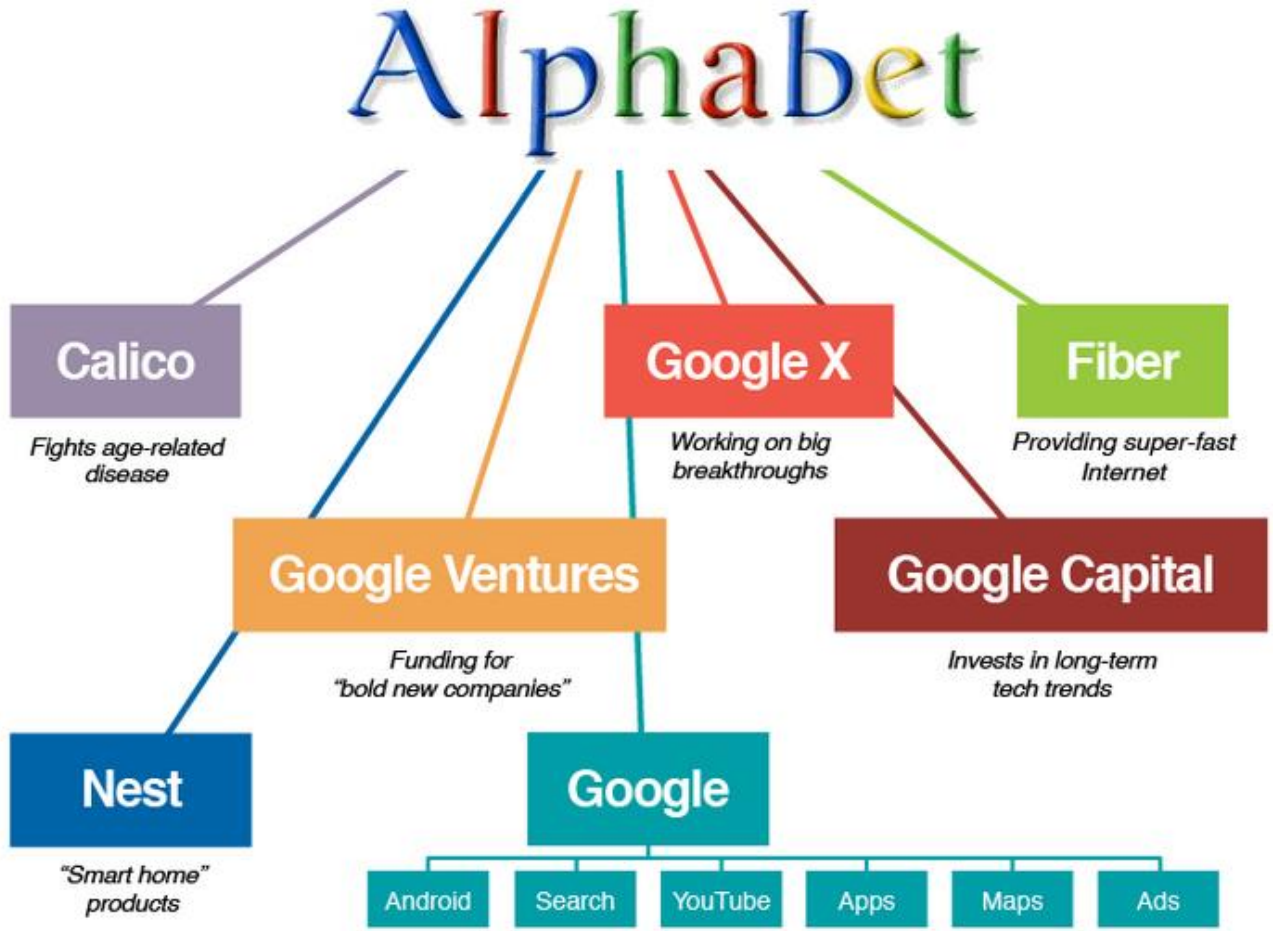
Dans un monde où mon modèle est parfait et où je gagnais le concours Kaggle…

Lucas Butscher,
Data analyst


Alphabet
Calico
Fights age-related disease
Google X
Working on big breakthroughs
Fiber
Providing super-fast Internet
Google Ventures
Funding for "bold new companies"
Google Capital
Invests in long-term tech trends
Nest
"Smart home" products
Google
Android
Search
YouTube
Apps
Maps
Ads


Mardi 9 oct. | 20 °C


Google
Google


Google

- Vente par internet
- Produits standarts

- "Petit" nombre de visite
  ~ 100 000 par mois
- Beaucoup de moyens alloués

kaggle.com

- Expliquer le comportement d'achat

Algorithme de prédiction d'achat

Modification des facteurs influants le comportement d'achat

- Modèle exportable
- Innovation dans les autres filiales

… Retour à la réalité

I. Le plan

Un plan chronologique...

- Suit la même trame que mon code
- Ordre logique

Table des matières

I. Le plan

… qui fait des allers-retours

Variable	État carré
SessionQualityDim	0.362
totalPageviews	0.142
hits	0.134
pageviews	0.133
timeOnSite	0.064
nbVisites	0.017
visitNumber	0.004

La variable **sessionQualityDim** est celle ayant l'état carré le plus élevé. Je suppose qu'elle sera indispensable pour la création du modèle prédictif. **TotalPageviews**, **hits** et **pageviews** ont aussi un état carré important. Cependant je suspecte une forte corrélation entre-elles et suppose n'en utiliserai qu'une.

Table des matières

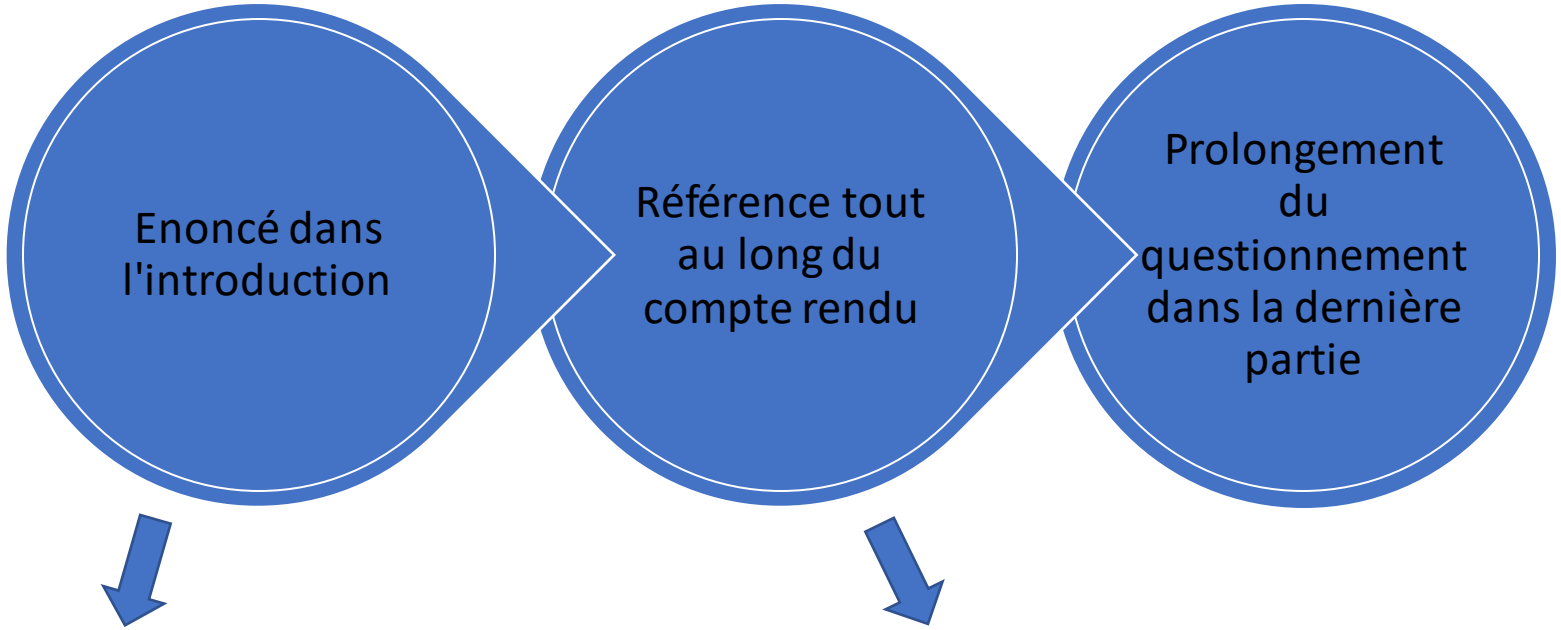
I. Le plan

Table des matières

Mauvais choix de variables

Aucune présente aucune présente dans le modèle prédictif

II. L'angle



"C'est pour cela que je cherche à déterminer les variables expliquant au mieux le comportement d'achat des utilisateurs de ce site"

" Cette étape va nous permettre à la fois de compléter le nettoyage de nos données et de percevoir ou non des possibles corrélations entre des variables et l'achat. Ainsi, cela se présente comme une pré-séléction des facteurs expliquant au mieux le comportement d'achat"

III. Les représentations visuelles

- Format spécifique

```
print(df_train.totals.loc[0])
```

```
{"visits": "1", "hits": "1", "pageviews": "1", "bounces": "1"}
```

III. Les représentations visuelles

- Fonction atypique

```python
def string_to_dict(dict_string):
    return json.loads(dict_string)
```

```python
def create_new_dataset(df,json_col):
    new = df.copy()
    colonnes = new.columns
    for col in json_col:
        print(col)
        new[col] = new[col].apply(string_to_dict)
        new = pd.concat([new,(pd.io.json.json_normalize(new[col]))], axis=1)

    for i in colonnes:
        new = new.drop(columns=i)

    return new
```

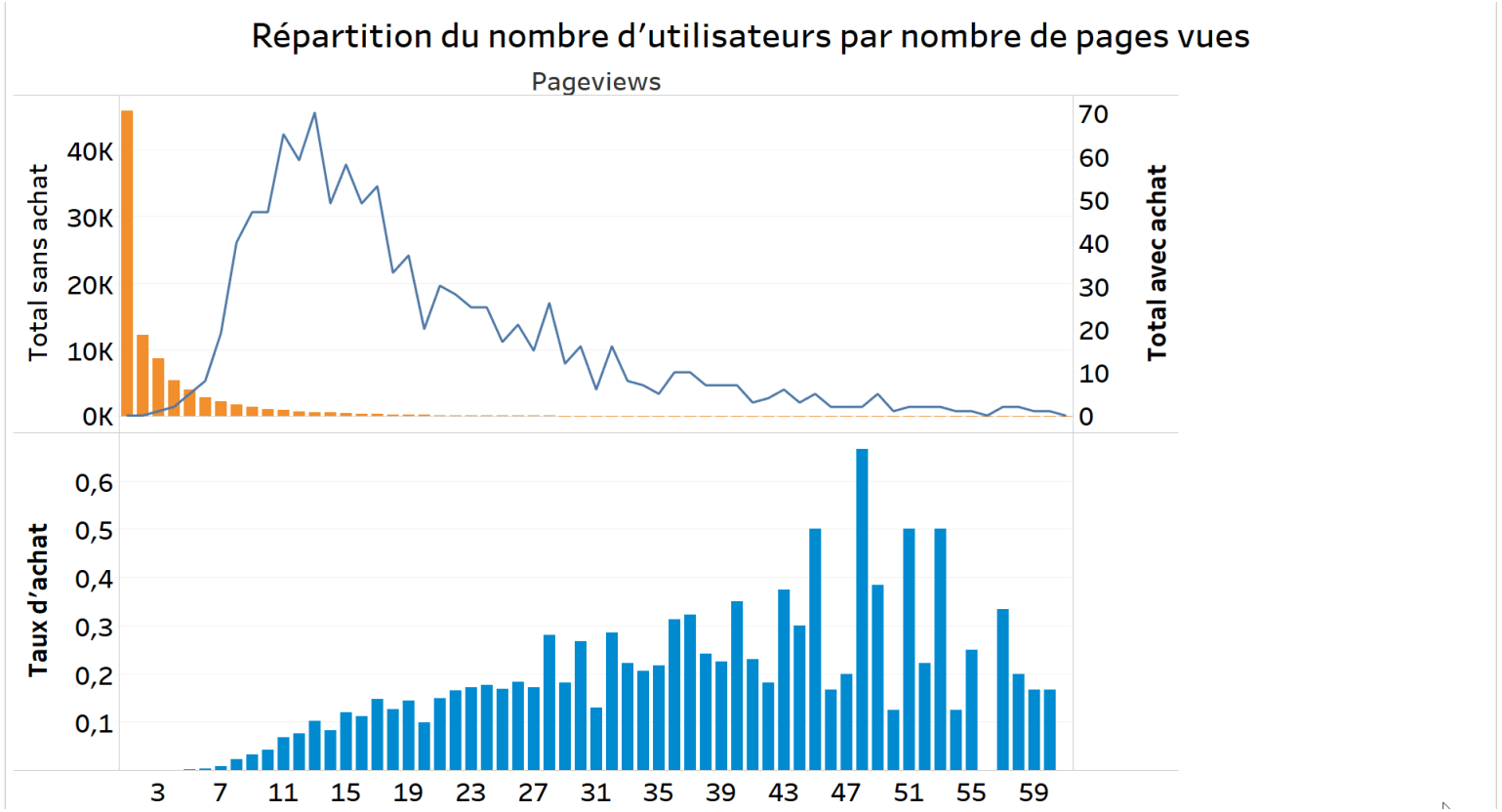
III. Les représentations visuelles

- Code important

```
X = df_acp[["sessionQualityDim","totalPageviews","nbVisites"]].values
y = df_acp["achat"].values
i=0
p=0

from random import shuffle
nb_erreur = 0

# découpe du jeu d'exemples en 80 % pour l'entraînement du modèle, le reste pour son test
p = 0.8 # 80%
listeDesIndices = [i for i in range (len (X))]
shuffle (listeDesIndices)

# la liste des indices des exemples utilisés pour l'entraînement
indices_entrainement = listeDesIndices [0:int(p*len(X))]

# la liste des indices des exemples utilisés pour le test
indices_test = listeDesIndices[int(p*len(X)):len(X)]

# on entraîne le modèle
lr.fit(X[indices_entrainement], y[indices_entrainement])

# on prédit la classe des exemples du jeu de test
classe_predite = lr.predict(X[indices_test])
classe_predite_entrainement = lr.predict(X[indices_entrainement])
lr_predict=pd.DataFrame(lr.predict_proba(X[indices_test]))


# on calcule le nombre d'erreurs de prédiction
for i in range (len(indices_test)):
    if y[indices_test [i]] != classe_predite [i]:
        nb_erreur += 1
```

III. Les représentations visuelles

- Graphiques



IV. Intéresser le lecteur

- Deux erreurs glissées au début

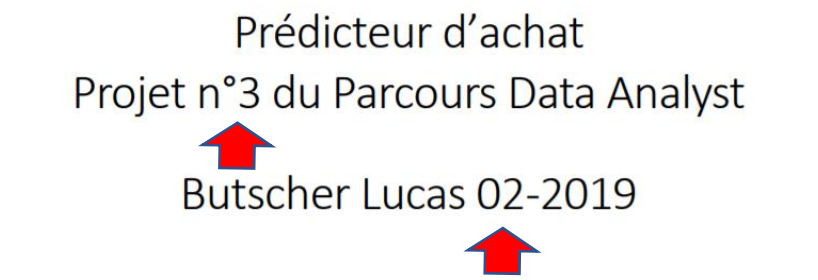


Table des matières

IV. Intéresser le lecteur

- Un peu d'humour

II. Nettoyage

II.A. Un problème de taille

La base de données sur laquelle nous allons travailler est contenue dans le fichier “train_v2.csv”. Cette base contient plusieurs millions de ligne et treize colonnes. Ainsi vint ma plus grande problématique : les performances de mon ordinateur. Celui-ci est trop âgé et trop peu performant pour effectuer des actions sur une base de données de cette ampleur. J’ai donc dû faire un choix. Plusieurs possibilités me sont donc apparues : soit changer de projet soit réduire la taille de mon DataFrame.

IV. Intéresser le lecteur

- Des citations pour prendre du recul

" Si vous torturez les données assez longtemps elles se confesseront ", Ronald Coase

" C'est une erreur capitale de théoriser avant que l'on ait des données. Insensiblement on commence à tordre les faits pour suivre des théories au lieu de tordre les théories pour suivre les faits ", Arthur Conan Doyle

" La visualisation nous donne des réponses à des questions que nous n'avions même pas ", Ben Shneiderman

Je vous remercie de votre attention