PROJET FOUILLE DE DONNEES

M2 IMPAIRS Université PARIS DIDEROT

Présenté par OUATTARI Abdelhaq

12/12/2018

Prédiction du Churn d'un Client

Dans le marché de la télécommunication

Qu'est ce que le **CHURN**

d'un client?

A qui on s'interesse Sur le Marché Français

orange™

kxen

Quel **Solution** ?

Pourquoi c'est important de

PREDIRE

Le CHURN **?**

Qui utilise ces méthodes ?

DATASET

- Source: KDD Cup 2009
- 100K ligne d'entées client splitté entre 50K train et 50K test avec 15000 variable. (100k x 15k)
- L'objectif du challenge tourné autour de la prédiction des trois valeurs suivantes : Churn, appetency, and upselling
- Moi, j'ai travaillé seulement sur le Churn.
- Les noms de variables et les valeurs catégoriques du dataset ont été anonymisé par Orange afin de protéger l'identité client.

Challenge Mondial – dataset important – la plus grande partie de mon expérience professionnelle est dans le domaine de la télécommunication – termes familiers – problème que j'ai déjà vécu par mes expériences précédentes – solution qui réponds a un grands problème qui ne cesse de grandir.

Nettoyage

- 1. Eliminer les N/A (? dans mon cas) de la Dataset

  → Toutes les lignes et les colonnes qui ne comportent que des N/A

  (213 Colonnes)

- Uniformiser les dtypes de ma Dataset

  → Toute valeur numérique = float;

  else = « category »

- Ne garder que les colonnes avec dtype = float

  (175 colonnes)

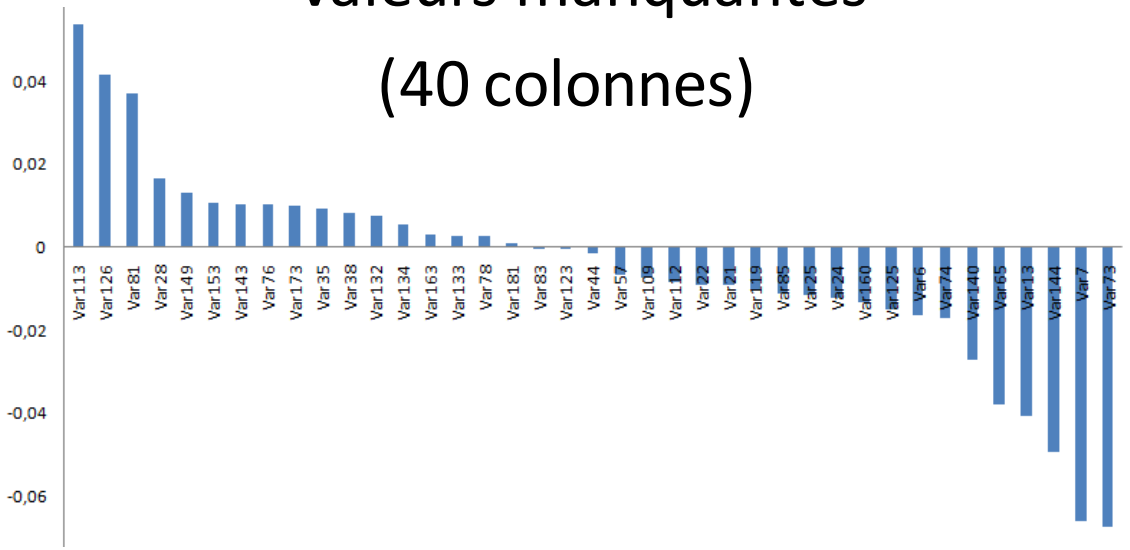
- Remplacer les éléments manquants pour chaque variable

Tuning

- Calculer la corrélation entre le Churn et les autres colonnes.
- Ne garder que les variables qui ont une corrélation supérieure a 40%

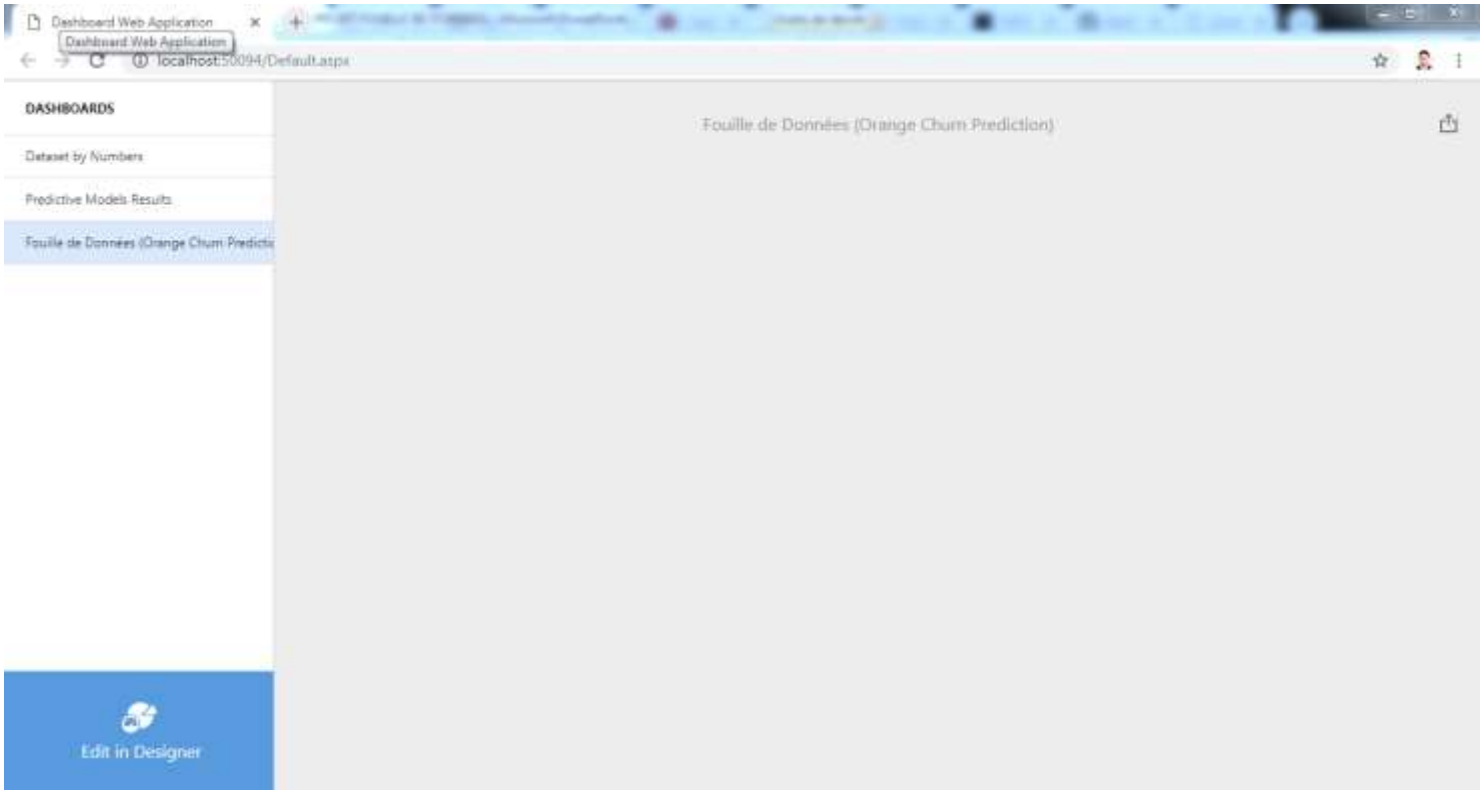
→ éliminer les variables qui ont beaucoup de valeurs manquantes

(40 colonnes)



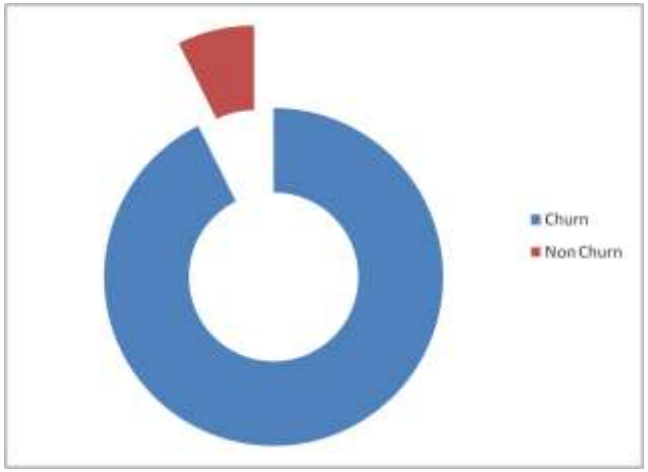
Data Exploration

- Exploration des données et des résultats via un Dashboard Web.



Train / Validation

- La Dataset que j'utilise est au préalable splitté en train et test sets.
- Toutefois j'ai préféré essayer l'utilisation de la bibliothèque Sklearn.train_test_split pour splitter ma Dataset en train & test avec un taux de test set de 20%.
- Le Taux de Churn global sur la Dataset est distribué comme suit



Modèles Prédictives (Algorithmes)

- En utilisation la bibliothèque Sklearn
- En s'appuyant sur une études faite par Neil Lawrence (Analysis of the KDD Cup 2009) qui donne une idée sur les modéles les plus utilisés par les participants au challenge et leurs efficacité

J'ai choisis l'application des trois modéles ci dessous:

- Naïve Bays
- Decision Tree
- Random Forest

Résultats

- Naive Bays Accuracy

0.8669

- Decision Tree Accuracy

0.9255

- Random Forest Accuracy

0.9280

D’aprés nos résultats ci dessus nous concluant que l’utilisation de Random forest donne une meilleur prédiction du Churn par rapport aux autres modèles qui restent eux aussi très proches de ca performance avec un petit écart.

Conclusion

- J'ai trouvé beaucoup de difficultés par rapport a la traduction et l'analyse des données comme les variables sont anonymisé
- Malheureusement je n'ai pas eu le temps pour exploiter toutes les pister et appliquer certain notamment sur l'évaluation des modèles utilisé et leurs performance
- J'ai trouvé beaucoup de fun dans la manipulation et l'exploitation de la donnée

MERCI Monsieur

pour cette formidable **opportunité**
&
excellente **expérience**