Rapport Projet Apprentissage

Abdallah ESSA, Ismaïl Hour MAHAMAT, Kevin SYLLA

Présentation du Dataset

Le dataset est un dataset de données de visites de sites provenant de Google Analytics. La taille du dataset est de 903 653 individus pour 427 variables. La variable-objectif est la variable TransactionRevenue. L’objectif est donc d’étudier les différentes variables explicatives des sessions d’utilisation d’un site web pour prédire la somme qu’il va dépenser sur ce site. Après un peu de recherches, nous avons une description du dataset précise à ce lien : <https://wpengine.com/blog/the-ins-and-outs-of-google-analytics/> qui permet de comprendre les intitulés des variables et qui se révèlera très utile au moment où l’on voudra créer et/ou modifier des variables que nous aurons identifié comme importantes.

Contexte

Pour nous guider vers un objectif cohérent et appliquer des méthodes et transformations cohérentes et pour un objectif concret, nous avons ajouté du contexte. Dans la tâche de prédiction de somme d’achat en fonction des caractéristiques d’une visite sur un site, on peut émettre l’hypothèse que l’objectif est de trouver les variables qui permettent d’expliquer une décision d’achat. Ainsi, il est aussi très intéressant non pas d’estimer la somme dépensée sur le site mais si oui ou non un utilisateur va dépenser sur le site et pourquoi. On peut donc ramener le problème à une tâche de classification visant à trouver les variables qui déterminent une décision d’achat. C’est pourquoi nous avons fait de la régression et de la classification.

Méthodologie

Pour résoudre ce problème, nous avons appliqué les étapes suivantes.

* Acquisition et exploration des données

Une des contraintes dans l’acquisition et le traitement des données que nous avons constaté lors de nos premiers tests d’analyse est la lenteur de tous les traitements dus à la taille du dataset (presque 430 millions de données). Nous avons donc décidé de conserver toutes les données et d’utiliser un cluster H20 pour accélérer nos traitements grâce à du calcul parallèle. Ainsi, nous ne tronquons pas nos données pour conserver toute l’information disponible et obtenir des résultats plus pertinents.

* Prétraitement et gestion des NA

Ensuite, nous explorons visuellement les données via des plots et analysons le ratio de NA dans nos données. Nous observons que beaucoup de colonnes sont des colonnes indisponibles dans la version qui nous a été donnée. Ces colonnes ne possédant aucune information, nous décidons de les supprimer du dataset. Nous binarisons les données booléennes pour les traiter plus facilement pour les modèles de Machine Learning. Les colonnes étant partiellement remplies sont conservées et nous remplissons les NA par l’information la plus pertinente (« non fourni », 0 ou la médiane si la colonne est beaucoup remplie). Par exemple, la variable à expliquer TransactionRevenue a beaucoup de NA, mais nous ne pouvons clairement pas supprimer toutes les lignes ou cette colonne. Pour pallier ce problème, nous remplaçons les NA par 0, car les lignes remplies sont toujours différentes de 0 ; nous avons donc conclu que les NA signifiait simplement qu’il n’y avait pas eu de transaction effectuée. L’information devient ainsi exploitable et les NA ont en réalité un sens. Nous appliquons la fonction logarithme sur la variable transaction revenue pour avoir des scores compréhensibles et non spécifiques à la disparité contenue dans cette variable.

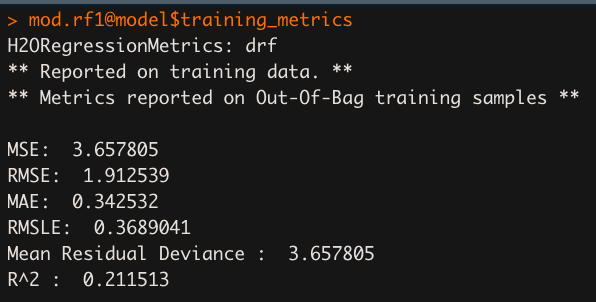
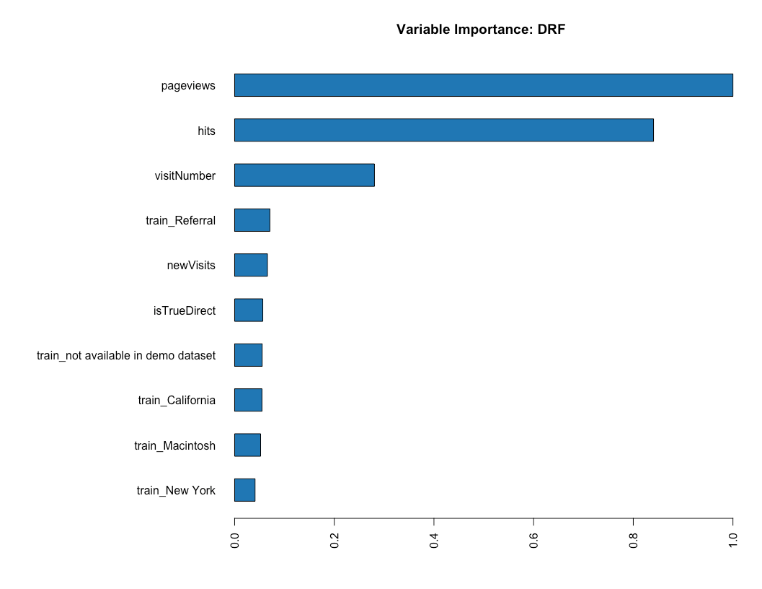
Nous avons également observé des dates dans le dataset. Nous avons donc directement créé des variables « month » et « dayofweek » qui peuvent se révéler intéressantes. De plus, nous avons remarqué que les sessions avaient parfois le même utilisateur, nous avons donc fait une agrégation en fonction de l’id de l’utilisateur.

* Méthodologie

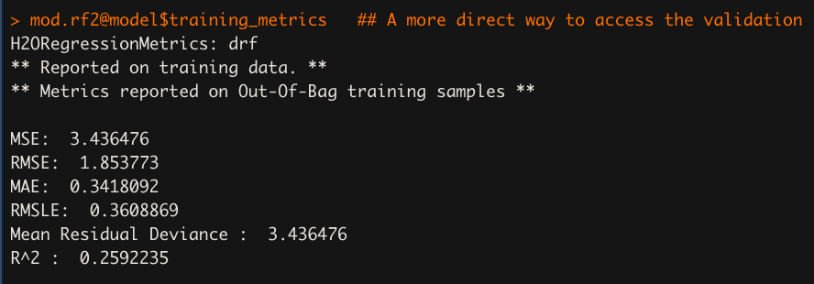
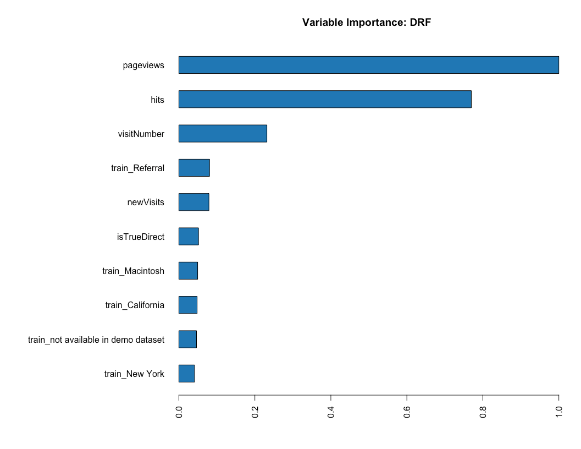
Pour analyser les données et exploiter nos résultats, voici le processus que nous avons utilisé et que nous avons itéré :

* Utilisation de modèle
* Evaluation du modèle
* Identification des variables importantes
* Analyse des variables importantes pour d’éventuels features engineering et modifications
* Réitération de la boucle
* Modèles et Résultats
  + En régression
    - Random Forest

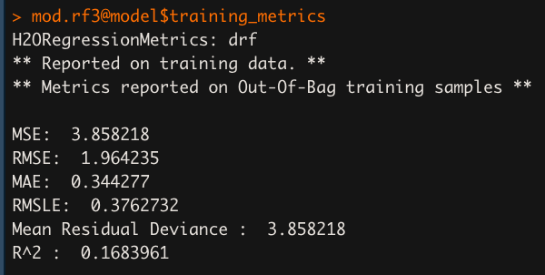
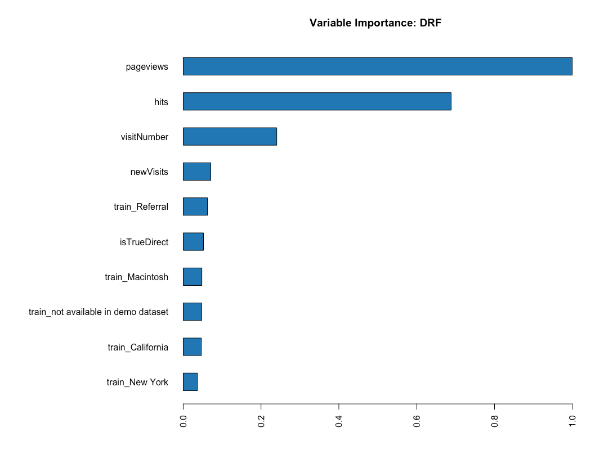
Random Forest 1 :



Random Forest 2 :



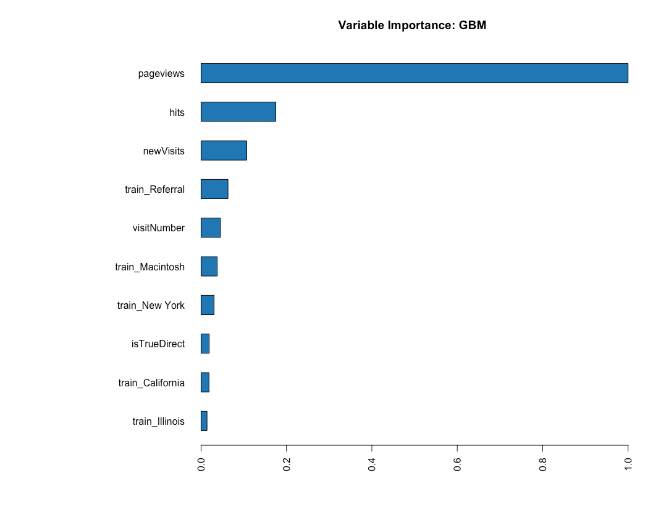
Random Forest 3 :



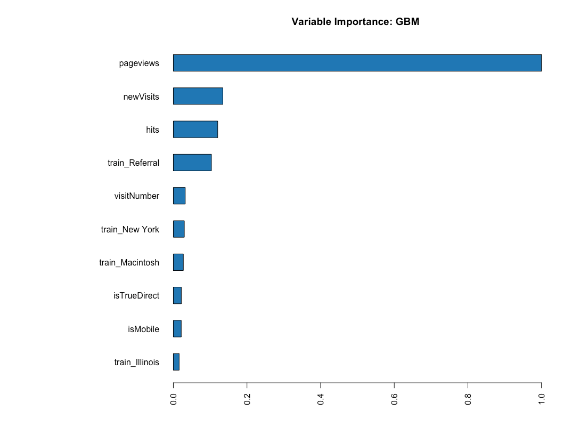
Observations : On voit que les variables les plus importantes selon ce modèle sont pageviews, hits et visitnumber. En consultant le lien qui détaille le dataset, on observe que ces colonnes sont le nombre de pages consultés, le nombre de clics sur le site et le nombre de visites sur le site.

* + - GBM

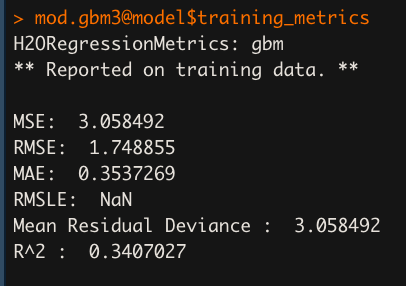
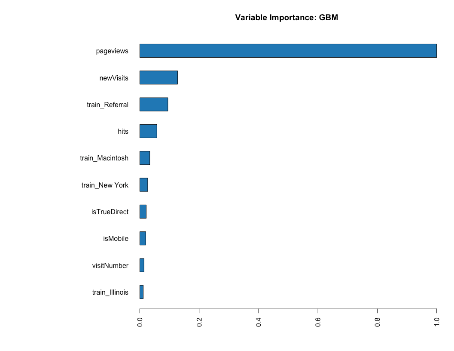
GBM 1 :



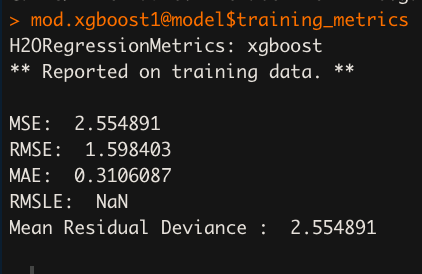
GBM 2 :



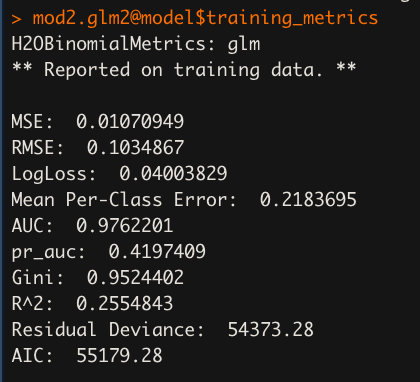
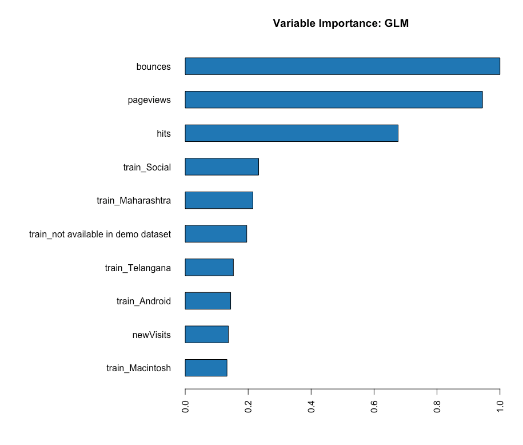
GBM 3 :



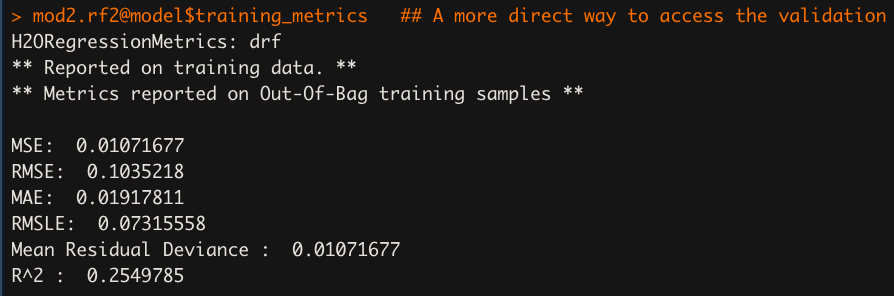
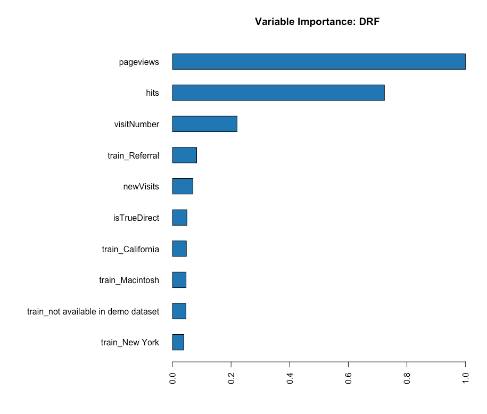
XGBoost :



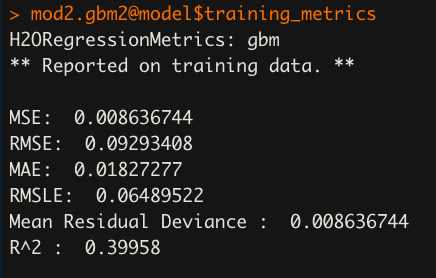
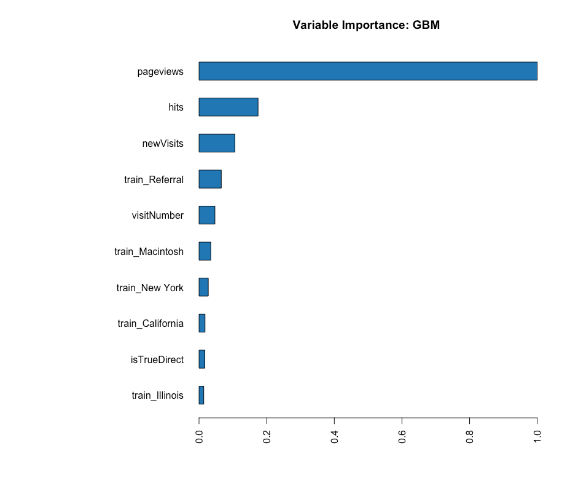
* + En classification
    - Régression Logistique



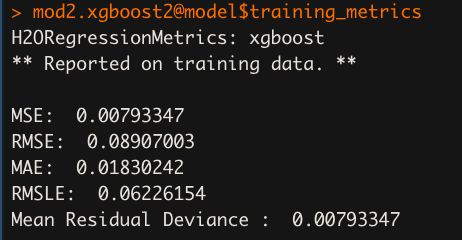
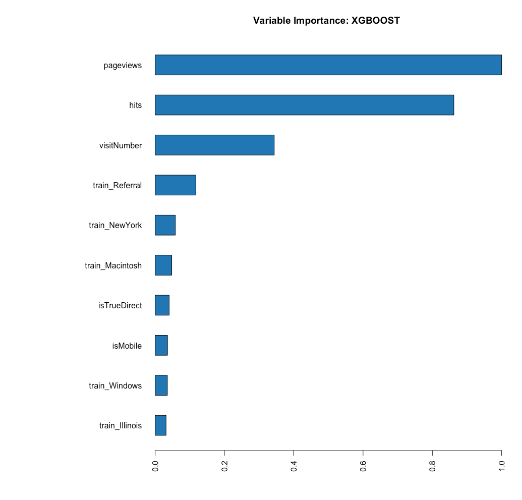
* + - Random forest



* + - GBM



* + - XGBoost



Conclusion : On observe que les variables les plus importantes sont les variables pageviews, hits et visitnumber. Nous avons également remarqué que la région et des informations qui peuvent avoir un lien avec la classe sociale (possession d’un macintosh, etc.). Nous avons eu l’idée d’ajouter une variable qui permet de savoir si la session a lieu dans une région riche ou non à partir des états des états-unis. On remarque que la classification est plus performante que la régression et qu’avec un peu plus de temps nous aurions pu affiner l’analyse et identifier un profil d’acheteur en creusant la piste de l’identification de classes sociale à partir des variables du dataset.