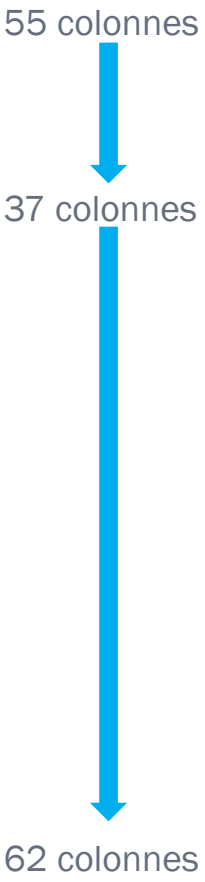
ayming

Projet debrief

Draft au 6 septembre 2019

Feature engineering

- **55 colonnes dont plusieurs ne portent aucune info (valeur unique)**
  - browserSize, browserVersion, cityId, criteriaParameters, flashVersion, language, latitude, longitude, mobileDeviceBranding, mobileDeviceInfo, mobileDeviceMarketingName, mobileDeviceModel, mobileInputSelector, networkLocation, operatingSystemVersion, screenColors, screenResolution, socialEngagementType, visits
  - Reste 37 colonnes
- **Feature engineering**
  - Passage en numéric : fullVisitorI, pageviews, hits, visitNumber, transactionRevenue, visitId
  - Dates : champs date et visitStartTime transformé en date et heure (fonction (as.POSIXct
    - Extration de nouvelles colonnes : mois, jour de semaine, ..
    - Passage des dates elles memes en numéric
- **Nouvelles features**
  - Nb de session par joueur
  - Délai entre session
  - **sessionId** : Split en 2 colonnes sumériques (de part et d'autre du « - »
  - **gclId** : Extraction des 3 premiers et 3 derniers caracteres (présente des régularités. On pourrait pousser plus loin le parsing
  - **referralPath** : détection des paths avec google (on pourrait pousser plus loin le parsing)
  - **Keywords** : features indicatrisce de « youtube, google, matching, Content targeting, vertical targeting,, provided, store, shop ….
- **Traitement des NA**
  - « ex na » pour les strings
  - Médian pour les mumérics
  - 0 pour la target transaction revenue
- **Factorisation des strings**, réduction du nombre de modalités à 1000 (information mutuelle)
- **Passage de la target en log(x+1)**



2 approches possibles : prédiction de montant par session puis agrégartion, ou construire un dataset par joueur puis prédiction directe du montant par joueur

Scénario A

62 colonnes

Point d'attention : Séparation train test

Attention à splitter au hasard par joueur et non au hasard par session dans le jeu d'évaluation les nouveaux joueurs n'ont aucun historique connu

?

Scénario B

P colonnes

Point d'attention : feature enginnering

Possibilités infinies de création de feautures agrégées par joueur

Scénario A1

- directement une régression ?

Scénario A2

- d'abord une classif pour mettre de coté les 0 puis une régresssion ?

Quel point d'attention si stratégie de modèles successifs ?

Une question importante avant de démarrer…
→ laquelle ?

- Quelle serait la performance d'une prédiction naïve ?

- Exemple : prédiction aucun achat pour tout le monde
  - RMSE : 2,02

- Cette valeur est essentielle pour porter un jugement de valeur sur le résultat …

Approche A2 :
→ modalité d'exploitation du jeu de données

Modéliser régression et classification sur des jeux indépendants est essentiel !

714K joueurs

classif — 25%

régression — 50%

Test final — 25%

Faire le sampling sur les joueurs pas sur les sessions
....
Au hasard bien sûr. Jamais les premières lignes

train1

valid1

train2

valid2

Test final

Idem pour le découpage train test

train1

valid1

train2

valid2

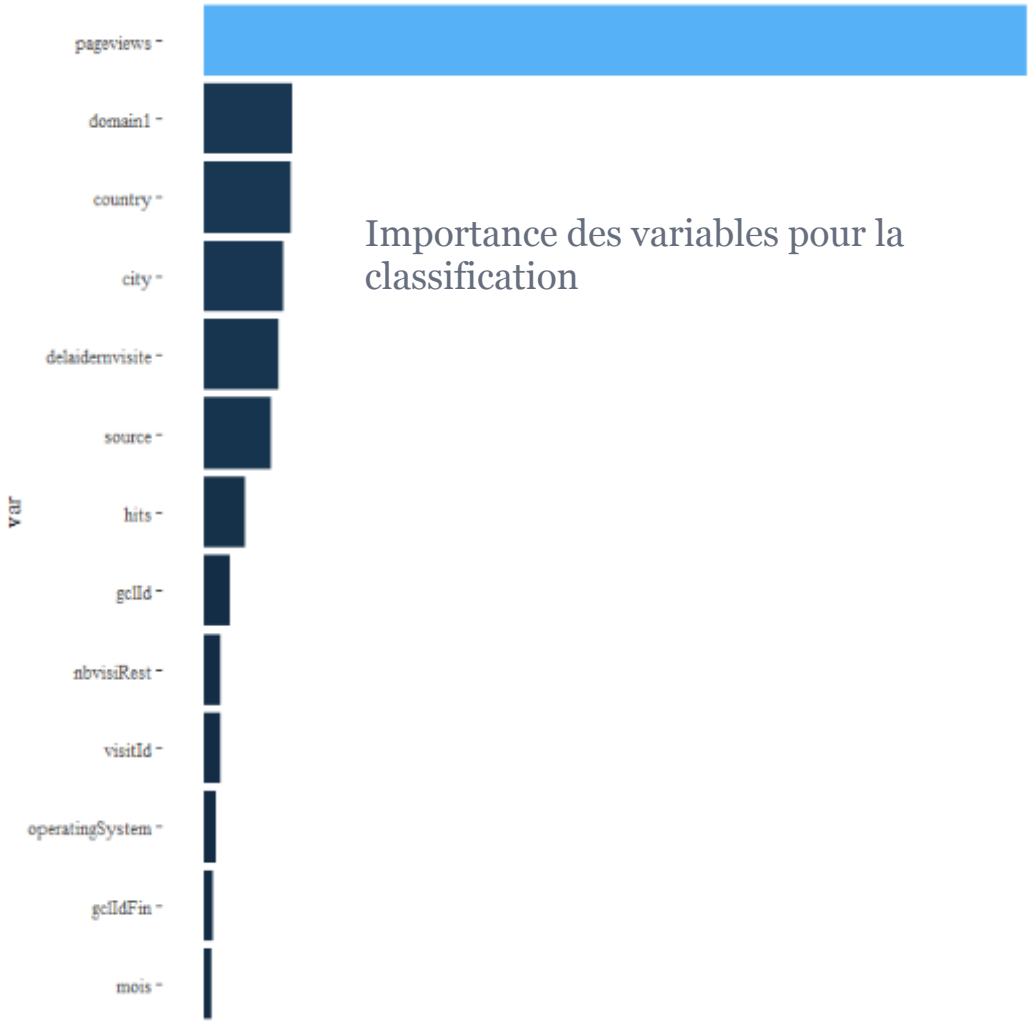
Test final

900 sessions

Au final toutes les données ont été utilisées,
Au max 150K lignes à la fois (classification)

Approche A2 :
→ 1 : Phase de classification

- 98% d'auc : Encouragement à poursuivre dans cette voie



Importance des variables pour la classification

Observations par ordre croissant de score

0.9999

Valeur sûrement non nulle : régression nécessaire

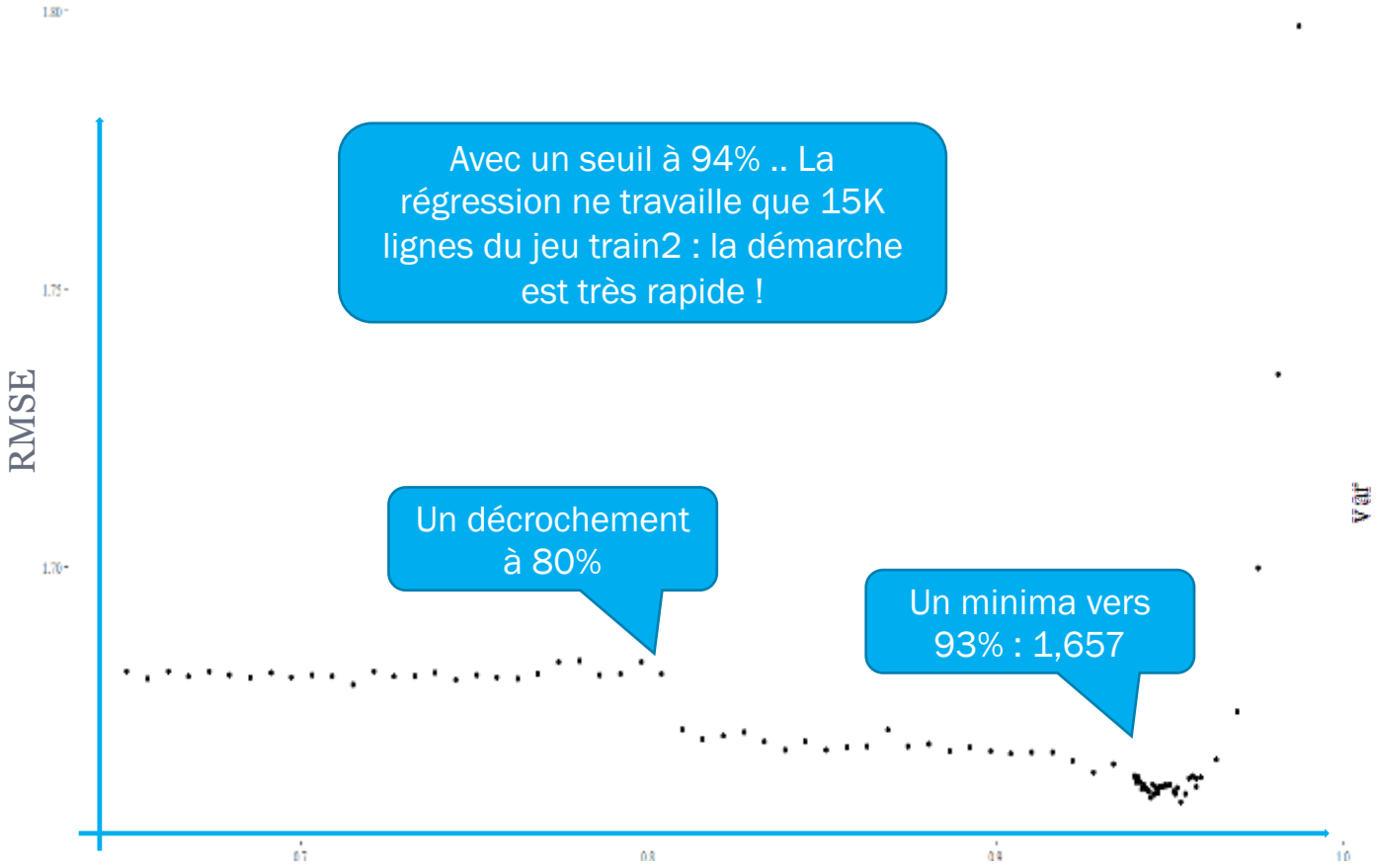
Ou placer le seuil séparant les prédictions « 0 » retenues dès la phase de classification et celles nécessitant une régression pour évaluer la valeur ?

Valeur sûrement 0

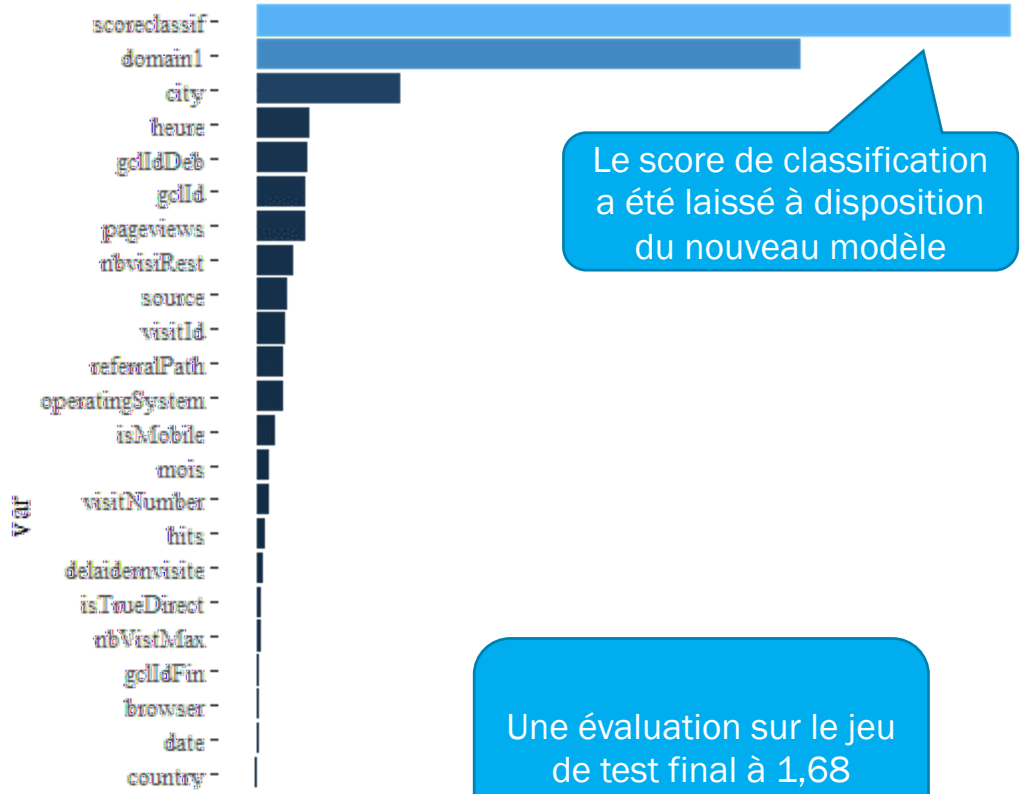
0.0001

Approche A2 :
→ 2 : Phase de régression :

Pour déterminer le seuil .. La seule démarche est de tester !

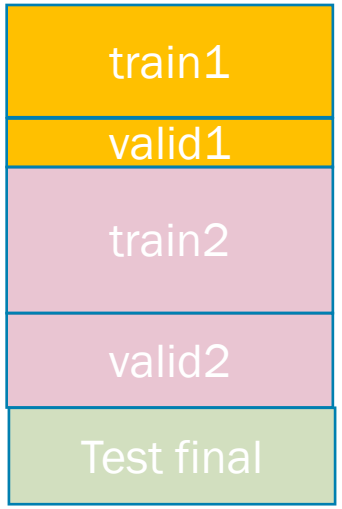


Importance des variables pour la régression

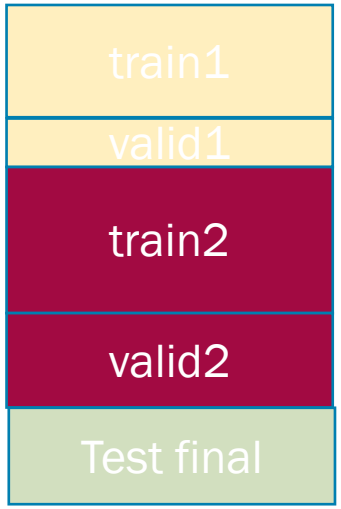


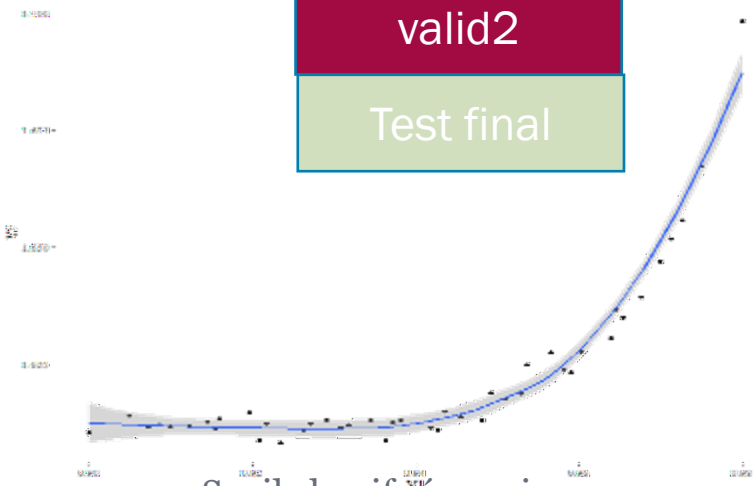
Approche A2 :
→ 2 : synthese de l'approche

Classification



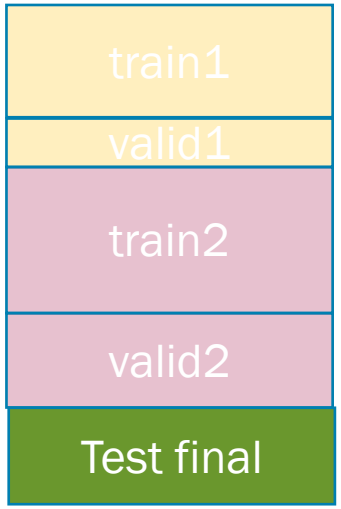
Régression au-delà d'un certain score
En dessous : 0
Au dessus : prédiction de la classification





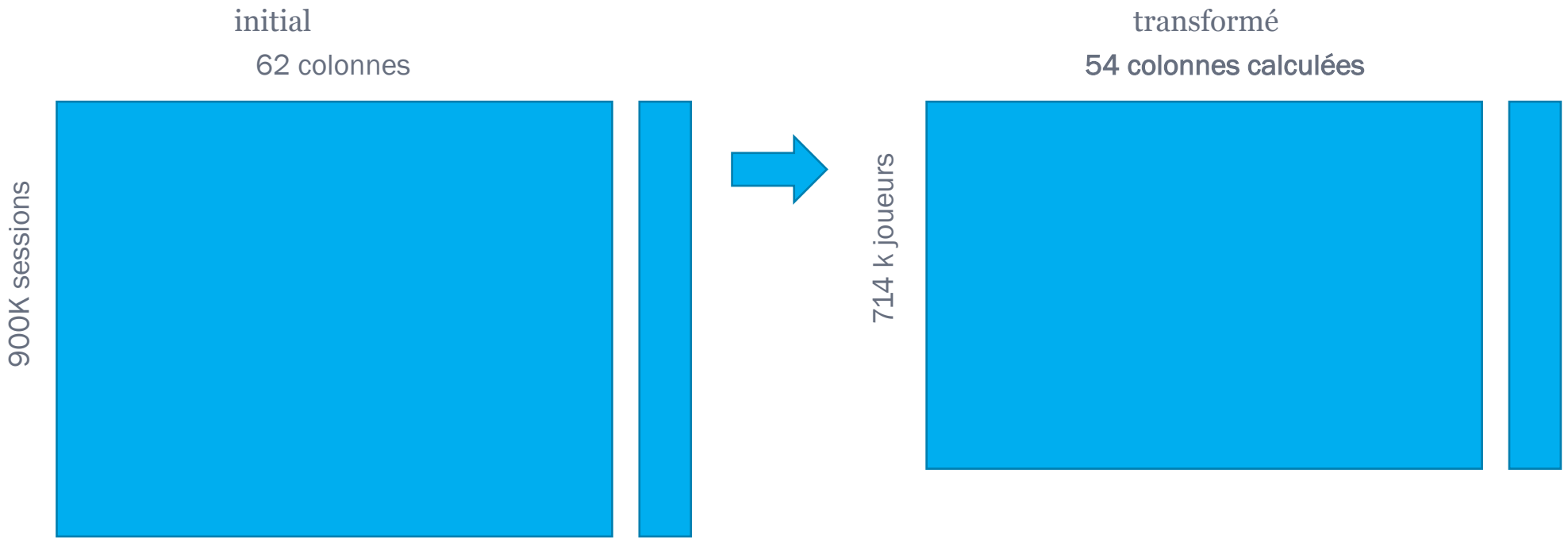
Seuil classif régression

Jeu de test



AUC 0.985 **Régression** **1.605** **Final** **1.61**

ScenarioB



RMSE : 1,67 … approche mise de côté

Pour aller plus loin

- Réapprendre avec toutes les données
- Intégrer le score par session dans la régression par visiteur (calculer des moyennes et des sonnes de score ?
- Combiner des modèles : comment ?