MISE AU POINT D'APPROCHES STATISTIQUES POUR UNE STARTUP

Sous la tutelle du Professeur Anne GEGOUT-PETIT

<u>Réalisé par :</u>
Julien Kossi KOWOU
Prince Foli ACOUETEY

<u>Membre du jury :</u>
Professeur Malika SMAIL

<u>Faculté des Sciences & Technologies de Nancy</u>
Département de Mathématiques
Université de Lorraine

Sommaire

Tables des figures

1 Introduction

1.1 Contexte du projet long

Eazee-Link, anciennement connue sous le nom de Vazee, est une startup française qui œuvre dans le secteur de la consommation hors domicile depuis 2014. L'objectif de l'entreprise est de faciliter la relation entre les marques, les établissements de restauration et les consommateurs (clients).

Eazee-Link propose un service de cartes et menus digitaux qui est facile à déployer et sans contraintes techniques pour les établissements de restauration. A ce jour, 2500 établissements en France sont en partenariat avec Eazee-Link.

Les clients peuvent scanner un QR code pour accéder au menu d'un établissement. Les établissements peuvent utiliser Eazee-Link pour modifier instantanément leur carte, mettre en place une ardoise, et consulter les avis en ligne.

Pour les marques, Eazee-Link est une solution qui permet de répondre à des enjeux marketing et commerciaux. Eazee-Link leur offre des solutions d'activation, de diffusion et d'analyse, leur permettant de suivre en temps réel leur présence dans les établissements équipés.

Dans le cadre de l'amélioration des solutions fournies aux marques, Eazee-Link a décidé de revoir sa méthodologie d'analyse des prix de boissons par marque en y intégrant un calcul de prix moyen par contenant (exemple de contenant : 12.5cl, …). Des contenants tel quel (12.5cl, …) seront qualifiés de contenants normalisés dans la suite du document. C'est dans ce contexte qu'a été effectué ce projet long.

1.2 Objectifs du projet long

Les objectifs de ce projet long peuvent être résumées en trois (3) :

- *Proposer une technique permettant de reconnaitre le contenant normalisé d'une boisson grâce aux informations disponibles (type d'établissement, prix, etc…)*
  Ce premier objectif concerne les boissons de toutes les marques partenaires d'Eazee-Link.

- *Déterminer dans quel contexte la moyenne pourrait être employée comme un indicateur de prix d'une boisson*
  Le second objectif vise spécialement la marque DESPERADOS.

- *Proposer une analyse de prix de boisson*
  Le troisième objectif vise également la marque DESPERADOS.

2 Comment reconnaitre un contenant normalisé ?

Les cartes et menus digitaux qu'Eazee-Link propose aux établissements de restauration, leur permet d'introduire des produits que pourront consulter des clients pour formuler leurs intentions d'achat.

A l'enregistrement de ces produits, spécialement des boissons, ils mentionnent les informations telles qu'un intitulé de contenant et le prix. Ils peuvent également supprimer le produit introduit à tout moment. Cependant, Eazee-Link n'impose pas de formule standard pour l'intitulé d'un contenant.

Ce libre choix que propose Eazee-Link aux établissements, a pour inconvénient d'avoir des boissons ayant des intitulés de contenants formulés comme « bouteille, bte, shoots, 2×10, galopin, grande, médium ».

Ces intitulées de contenants que nous pouvons qualifier de « personnalisés » car dépendant de l'établissement, ne sont pas pertinents lors d'une analyse de prix de boisson par marque. En effet, deux boissons identiques d'une même marque vendues dans deux établissements différents (localisation différente) ou non, sont susceptibles d'avoir deux intitulés de contenants personnalisés différents.

Le but de cette partie est de proposer une technique permettant de faire correspondre à deux boissons identiques d'une même marque vendues dans plusieurs établissements (différents ou non), un même contenant normalisé.

2.1 Notion d'observation

Les informations présentes dans la base de données d'Eazee-Link ont permis de définir une boisson comme une observation caractérisée par les variables suivantes :

- selene_product_id (identifiant numérique de la marque dans la base de données d'Eazee-Link)
- pricei_label, $i \in [\![1, 5]\!]$ (ces variables contiennent des contenants personnalisés)
- pricei, $i \in [\![1, 5]\!]$ (il s'agit du prix de vente en euros du pricei_label correspondant)
- updated_at (dernière date de mise à jour de la boisson)
- name (nom de l'établissement où la boisson est vendue)
- city (localisation de l'établissement où la boisson est vendue)
- postal_code (code postal de l'établissement où la boisson est vendue)
- type (type d'établissement qui vend la boisson ; ex : bar, restaurant, bar-restaurant, etc…)
- suppliers (fournisseurs de l'établissement)
- inside_capacity (capacité intérieure en nombre de clients de l'établissement)
- outside_capacity (capacité extérieure en nombre de clients de l'établissement)
- normalizedLabel (marque de la boisson)

Une observation désigne une boisson d'une marque avec une valeur donnée de contenant personnalisé, vendue dans un établissement, à un prix, et à une date donnée. De plus, des observations de marques différentes peuvent provenir d'un même établissement.

Au total, nous disposons de 300555 observations.

2.2 Approche proposée

Afin de faire correspondre à toute valeur des variables pricei_label ($i \in [\![1,5]\!]$) un contenant normalisé, les étapes suivies sont décrites dans le schéma ci-après :

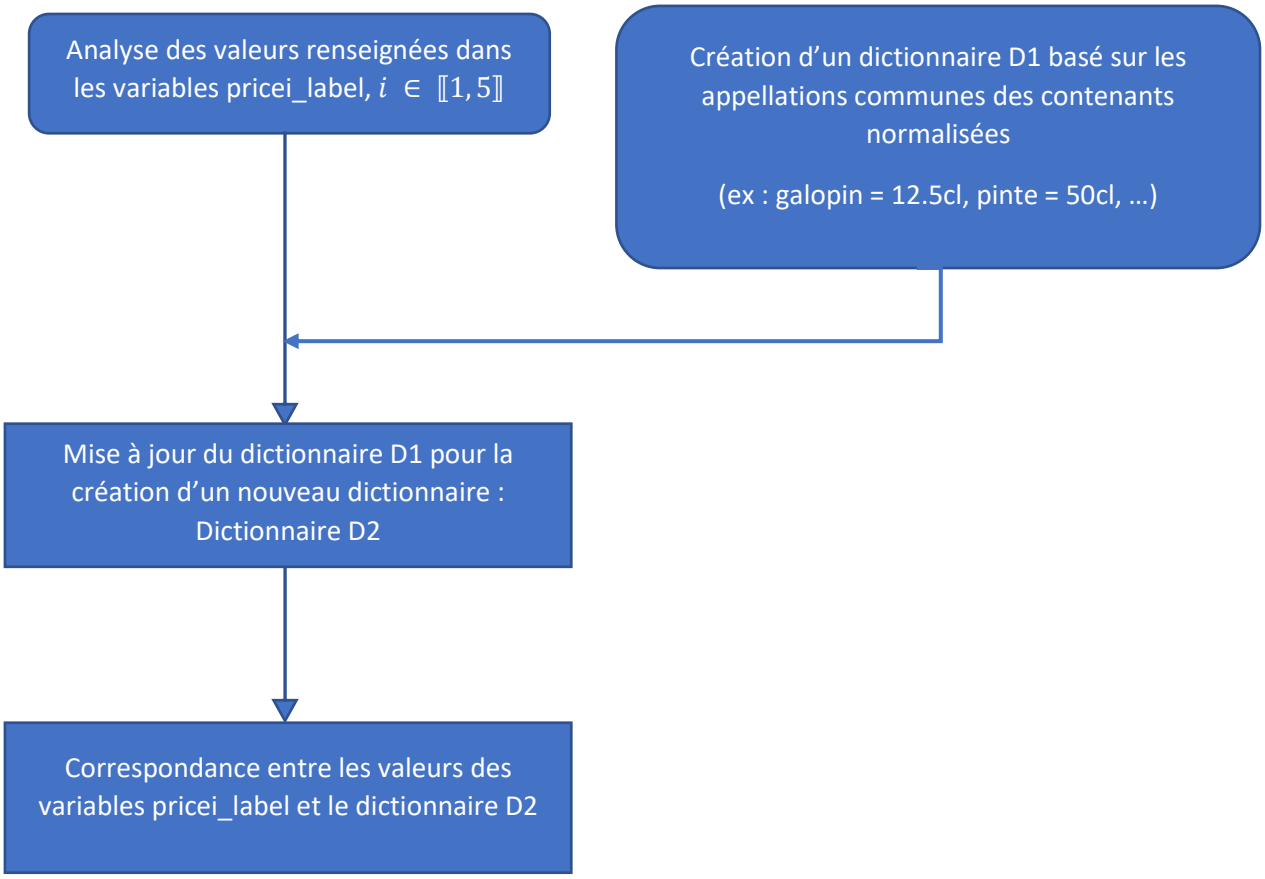

Figure 1 : Processus de normalisation d'un contenant personnalisé

Les variables pricei_label ($i \in [\![1,5]\!]$) ne sont généralement pas toutes renseignées ; certaines plus que d'autres :

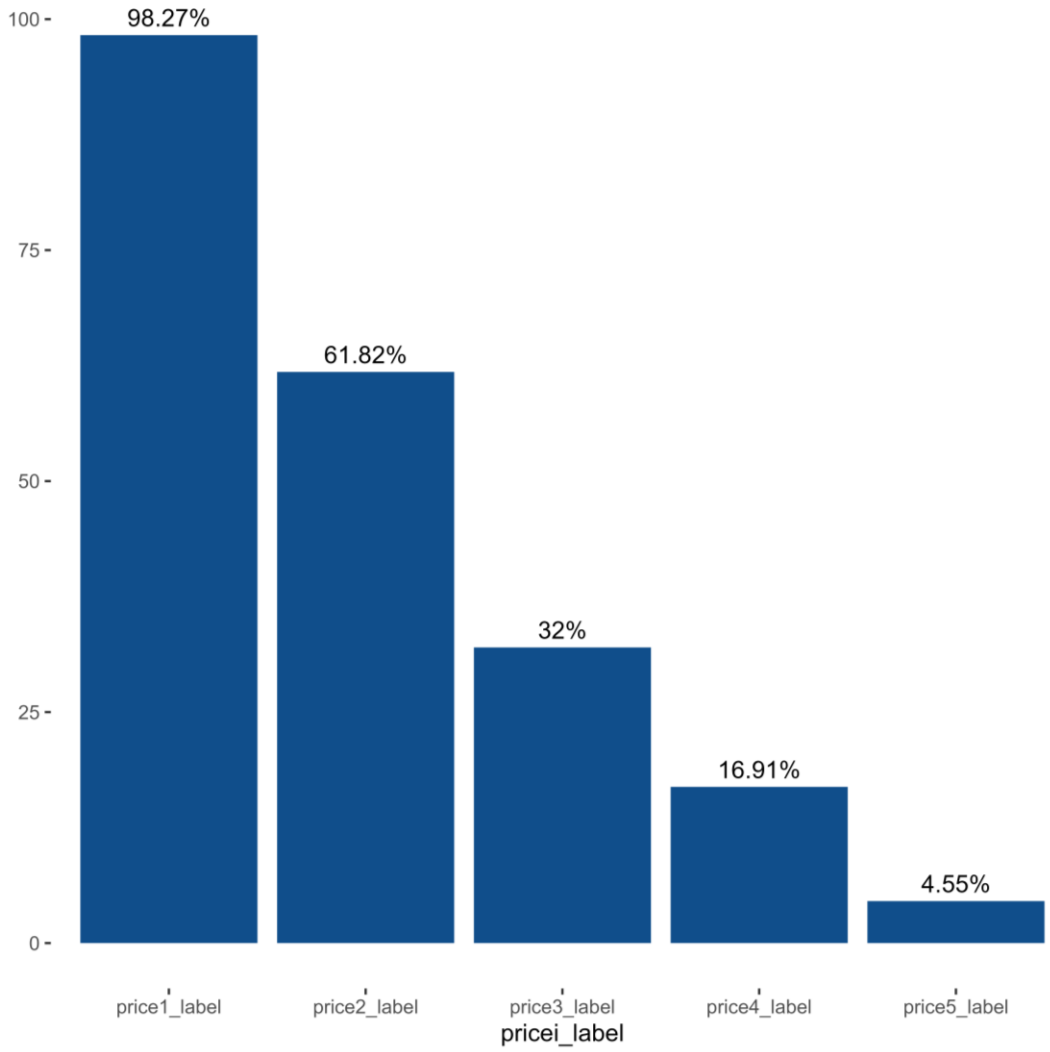

Figure 2 : Taux de pricei_label renseignés

L'approche utilisée n'a pas permis de trouver des correspondances à tous les contenants personnalisés des variables pricei_label ($i \in [\![1,5]\!]$) dans le dictionnaire D2 qui compte 300 correspondances.

Pour certaines observations, il y avait deux ou trois correspondances qui furent trouvées. Toutes ces observations n'ayant pas de correspondance ou ayant plus d'une correspondance furent écartées.

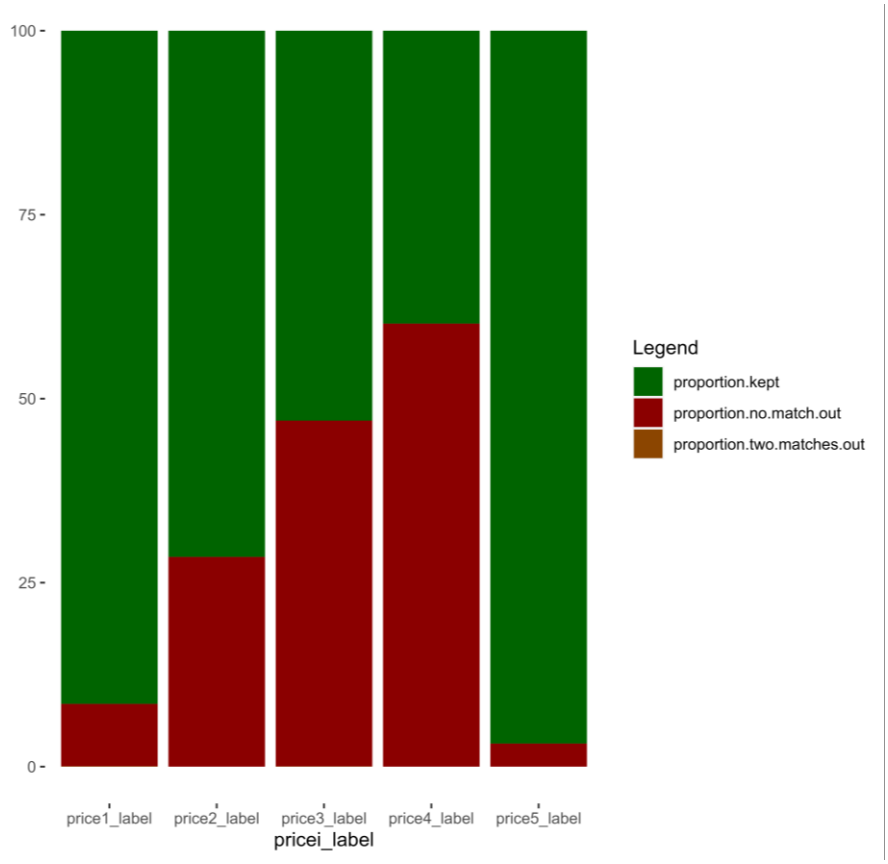

Figure 3 : Proportion d'observations gardées par pricei_label, i ∈ ⟦1,5⟧

Nous passons de 300555 observations à 248851 observations (soit 82.08% des contenants personnalisés qui ont pu être normalisés grâce au dictionnaire D2).

2.3 Conclusion

Nous passons de 300555 observations à 248851 observations (soit 82.08% des contenants personnalisés qui ont pu être traduits normalisés grâce au dictionnaire D2).

Par le suite, nous renommons toutes les variables pricei_label ($i \in [\![1,5]\!]$) par pricelabel ; de même toutes les variables pricei sont renommées price.

En effet, à l'issue de ces correspondances avec le dictionnaire D2, les variables pricei_label ($i \in [\![1,5]\!]$) représentent toutes le contenant normalisé d'une boisson. De plus, les variables pricei désignent toutes le prix d'une boisson.

3 Une moyenne est-elle adaptée comme indicateur de prix des boissons de marque DESPERADOS ?

Nous procédons à une étude de la distribution des prix de boissons de la marque DESPERADOS pour les contenants normalisés 25cl, 33cl et 50cl. Dans la base de données d'Eazee-Link cette marque est affiliée à trois (3) produits : DESPERADOS, DESPERADOSGINGER, DESPERADOSRED et DESPERADOSVIRGIN.

Ci-dessous le nombre d'observations par produit :

Produit	Nombre d'observations
DESPERADOS 25cl	172
DESPERADOS 33cl	2394
DESPERADOS 50cl	273
DESPERADOSRED 33cl	74
DESPERADOSVIRGIN 33cl	56
DESPERADOSGINGER 33cl	1

Le produit DESPERADOS est disponible en contenant 25cl, 33cl et 50 cl.

Les produits DESPERADOSRED et DESPERADOSVIRGIN ne sont disponibles qu'en contenant 33cl.

DESPERADOSGINGER ne compte qu'une seule observation, ce qui n'est pas pertinent pour une analyse. Il ne sera plus évoqué dans la suite de l'étude.

Ci-dessous un boxplot des prix de boissons selon ces produits :

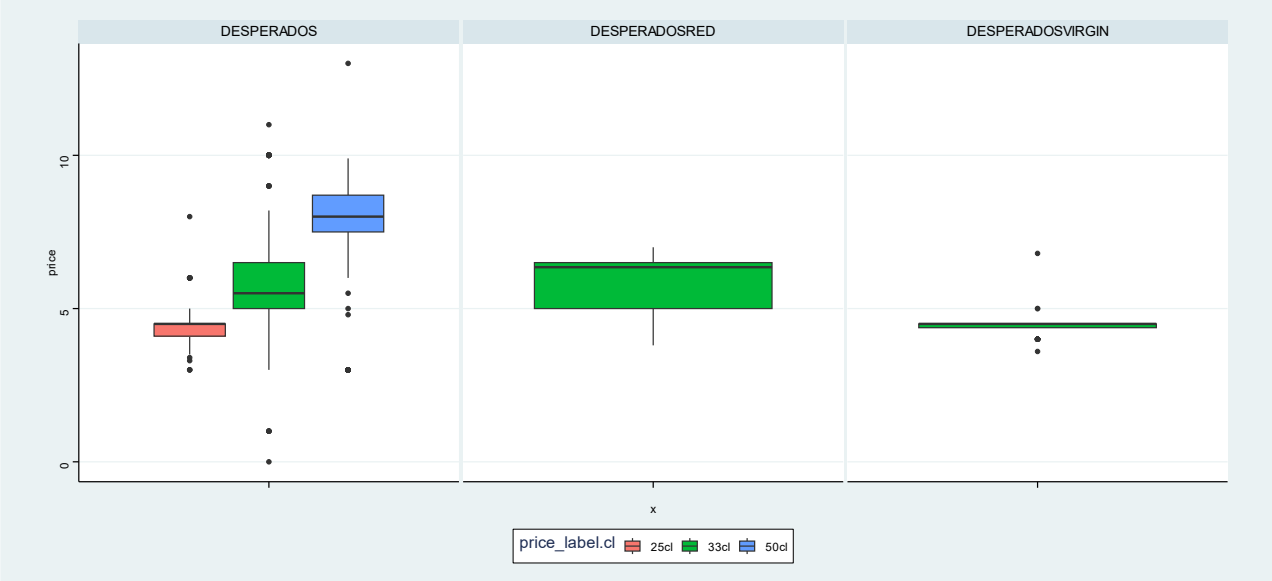

Figure 4 : Boxplot de prix des boissons pour la marque DESPERADOS en fonction du contenant

Pour le produit DESPERADOS, nous pouvons constater que le contenant 50cl coute généralement plus cher que les contenants 25cl et 33cl.

Nous procédons à un test de Kruskal-Wallis de niveau de confiance 95% pour comparer les distributions de prix des produits DESPERADOS 33cl, DESPERADOSRED 33cl et DESPERADOSVIRGIN 33cl :

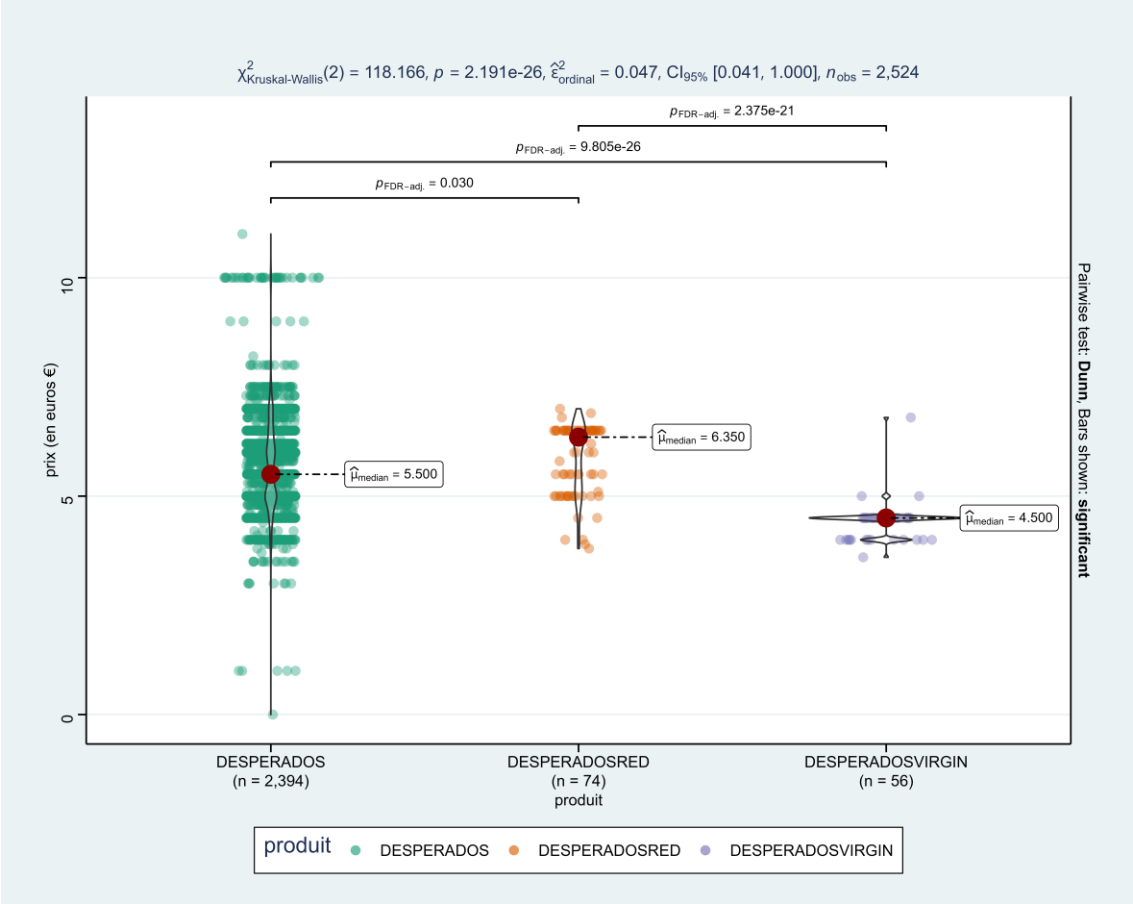

Figure 5 : Distribution des produits DESPERADOS 33cl, DESPERADOSRED 33cl et DESPERADOSVIRGIN 33cl

Les prix des produits DESPERADOS 33cl, DESPERAOSRED 33cl et DESPERADOSVIRGIN 33cl ont des distributions qui diffèrent significativement les unes des autres.

3.1 DESPERADOS 25cl

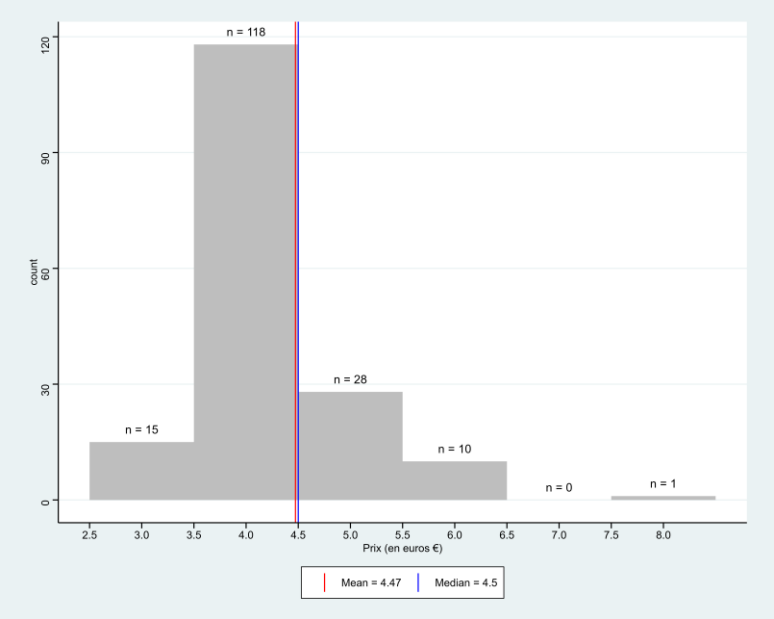

Figure 6 : Histogramme des prix du produit DESPERADOS 25cl

L'histogramme des prix montre qu'une moyenne n'est pas adaptée comme indicateur de prix du produit DESPERADOS 25cl. Nous pouvons affirmer que 50% des établissements vendent le produit DESPERADOS 25cl pour un prix inférieur ou égal à 4.5 euros.

3.2 DESPERADOS 50cl

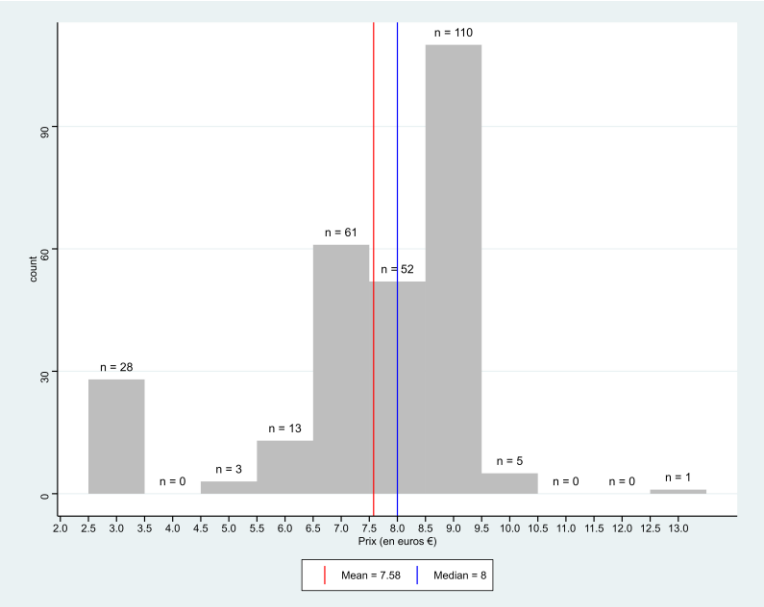

Figure 7 : Histogramme des prix du produit DESPERADOS 50cl

La moyenne n'est également pas adaptée dans ce cas comme indicateur de prix du produit DESPERADOS 50cl. Nous pouvons affirmer que 50% des établissements vendent le produit DESPERADOS 50cl pour un prix inférieur ou égal à 8 euros.

3.3 DESPERADOS 33cl

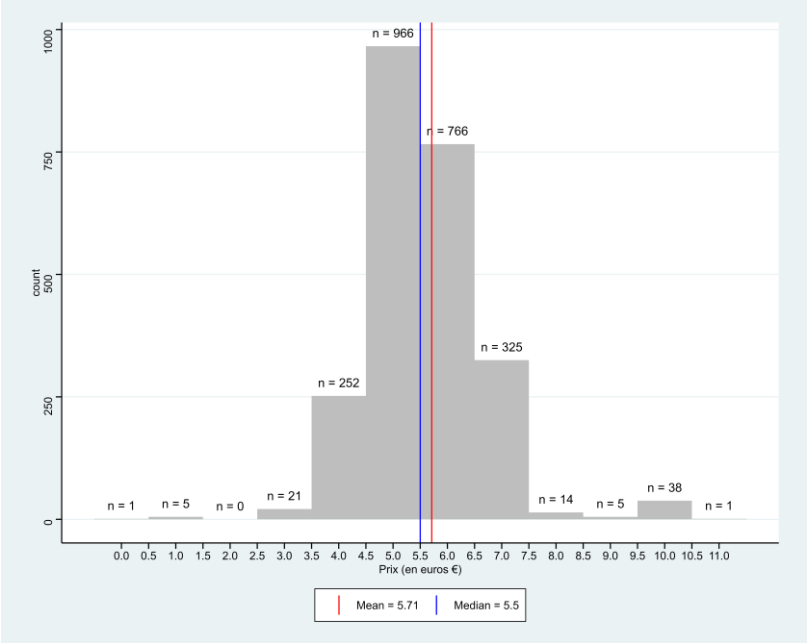

Figure 8 : Histogramme des prix du produit DESPERADOS 33cl

3.4 DESPERADOSRED 33cl

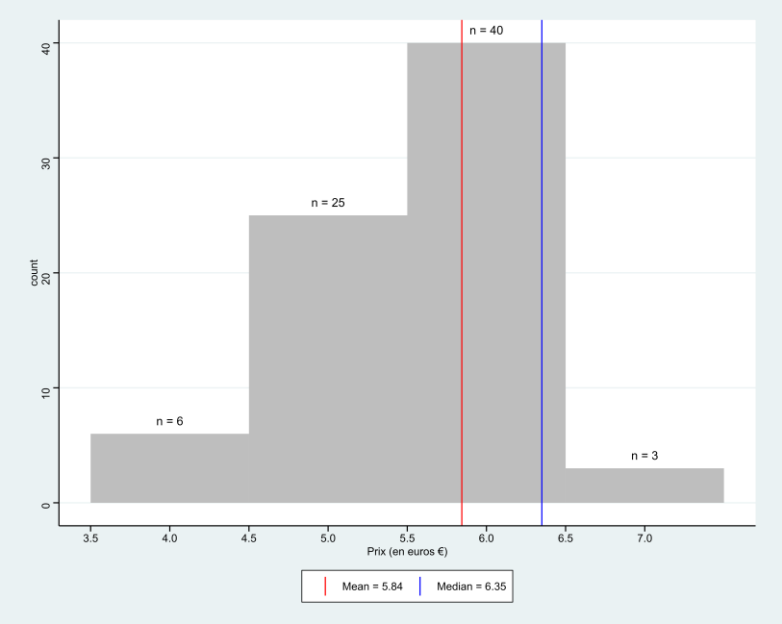

Figure 9 : Histogramme des prix du produit DESPERADOSRED 33cl

3.5 DESPERADOSVIRGIN 33cl

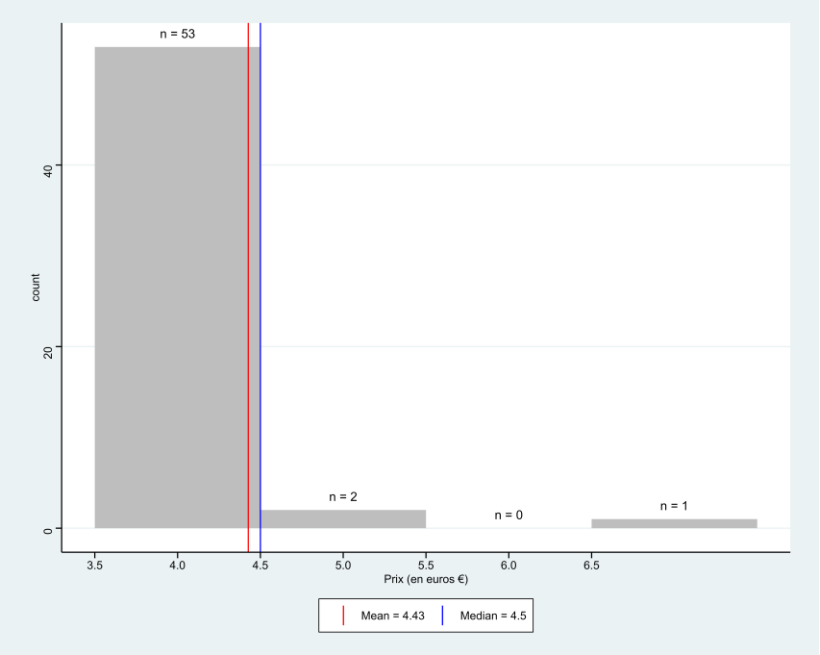

Figure 10 : Histogramme des prix du produit DESPERADOSVIRGIN 33cl

Dans l'immense majorité des cas, le produit DESPERADOSVIRGIN 33cl est vendu pour un prix inférieur ou égal à 4.5 euros.

3.6 Conclusion

Après analyse des histogrammes :

La moyenne ne peut être utilisée comme indicateur des prix pour aucun des produits de la marque DESPERADOS ayant pour contenant 25cl, 33cl ou 50cl. En effet, aucune des distributions de ces produits n'est symétrique, et d'autre part certaines de ces distributions présentent des valeurs extrêmes.

Utiliser la médiane (ou plus globalement les quartiles) donne des informations plus fiables sur les distributions des produits de la marque DESPERADOS.

Les produits de cette marque présentent des distributions différentes les unes des autres. Des statistiques telles que les quartiles ne peuvent être calculées sur la marque DESPERADOS sans prendre en compte le type de produit DESPERADOS, DESPERADOSVIRGIN ou DESPERADOSRED.

4 Analyse des valeurs manquantes

Parmi les variables d'intérêt utilisées dans la suite de l'étude, nous avons :

- city (localisation de l'établissement où la boisson est vendue)
- price (prix d'une boisson)
- type (type d'établissement qui vend la boisson ; ex : bar, restaurant, bar-restaurant, etc…)
- suppliers (fournisseurs de l'établissement)
- inside_capacity (capacité intérieure en nombre de clients de l'établissement)
- outside_capacity (capacité extérieure en nombre de clients de l'établissement)
- nb_days.

Seules les variables inside_capacity et outside_capacity présentent des valeurs manquantes.

La figure ci-dessous montre la proportion de valeurs manquantes présentes chez ces deux variables :

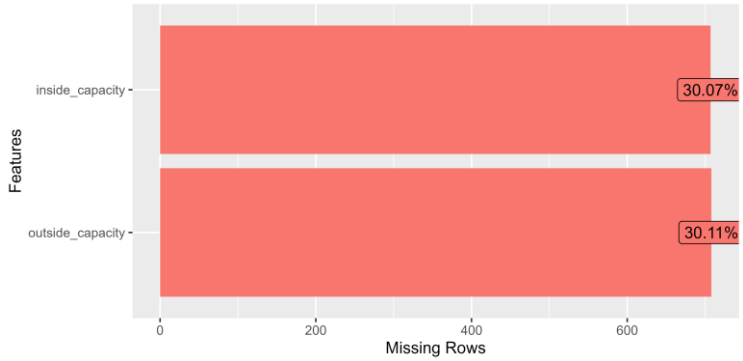

Figure 11 : Proportion de valeurs manquantes des variables inside_capacity et outside_capacity

Les valeurs prises par ces deux variables sont entières.

Nous ferons l'imputation de leurs valeurs manquantes par une méthode de forêt aléatoire en nous servant de leurs liens avec les autres variables d'intérêt.

La forêt aléatoire utilisée est composée de 10 arbres et le nombre d'itérations maximal est fixée à 1000.

5 Modèles de régression classiques

Dans un premier temps, nous transformons la variable updated_at (dernière date de mise à jour du prix d'une boisson) en variable quantitative comptant le nombre de jours par rapport à la date la plus ancienne présente. La date la plus ancienne est le 21/09/2021 et la plus récente le 20/12/2022. La nouvelle variable obtenue est appelée nb_days.

Dans un second temps, nous cherchons à expliquer le prix du produit DESPERADOS 33cl à l'aide des six (6) covariables :

- city (localisation de l'établissement où la boisson est vendue)
- type (type d'établissement qui vend la boisson ; ex : bar, restaurant, bar-restaurant, etc…)
- suppliers (fournisseurs de l'établissement)
- inside_capacity (capacité intérieure en nombre de clients de l'établissement)
- outside_capacity (capacité extérieure en nombre de clients de l'établissement)
- nb_days.

Nous disposons de 2351 observations. Elles proviennent de quatre types d'établissements présents dans 318 localisations différentes :

- *Part de chaque type d'établissement dans le nombre total des 2351 observations*

Type d'établissement	Part (%) du nombre total d'observations
Bar	17.74
Restaurant	4.55
Bar-Restaurant	27.17
Aucun des trois	50.53

- *Présence de chaque type d'établissement sur les différentes localisations*

Type d'établissement	Nombre de localisations
Bar	89
Restaurant	18
Bar-Restaurant	99
Aucun des trois	174

Nous constatons que toutes localisations ne contiennent pas tous les types d'établissements.

Deux modèles de régression seront étudiés : la forêt aléatoire et le XGBoost.

5.1 Forêt aléatoire

Nous procédons à une séparation des données en deux : un échantillon d'apprentissage (70% des 2351 observations) et un échantillon de test (30% des 2351 observations). L'échantillon d'apprentissage servira à entrainer le modèle ; quant à l'échantillon de test, il servira à évaluer la performance. Toutes les covariables quantitatives sont centrées et réduites dans les deux échantillons.

La forêt aléatoire construite est formée de 500 arbres. Chaque arbre procédera à une division jusqu'à ce qu'il y ait dans tous ses nœuds terminaux moins de 20 observations.

A l'issue de l'apprentissage, nous obtenons deux (2) statistiques permettant de juger de l'importance des variables :

- *Pourcentage d'augmentation de l'erreur quadratique moyenne (Model MSE Increase)* :

Il mesure l'importance d'une variable en calculant la réduction de la précision du modèle lorsque cette variable est exclue. Plus sa valeur est élevée pour une variable, plus cette variable est importante pour le modèle.

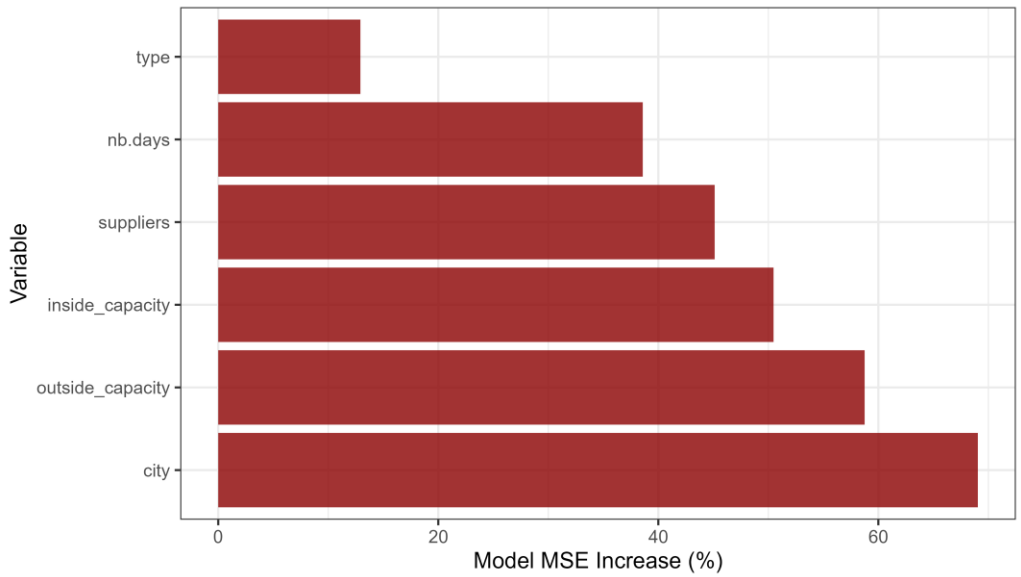

Figure 12 : Importance des variables selon le MSE

Toutes les variables utilisées permettent d'améliorer au moins de 10% le MSE du modèle.

- *Pureté totale des noeuds des arbres de décision :*

Cette mesure est calculée pour chaque variable en faisant la somme des améliorations de pureté des noeuds des arbres qui utilisent cette variable. Les variables qui ont une forte amélioration de la pureté (valeur élevée) sont considérées comme importantes.

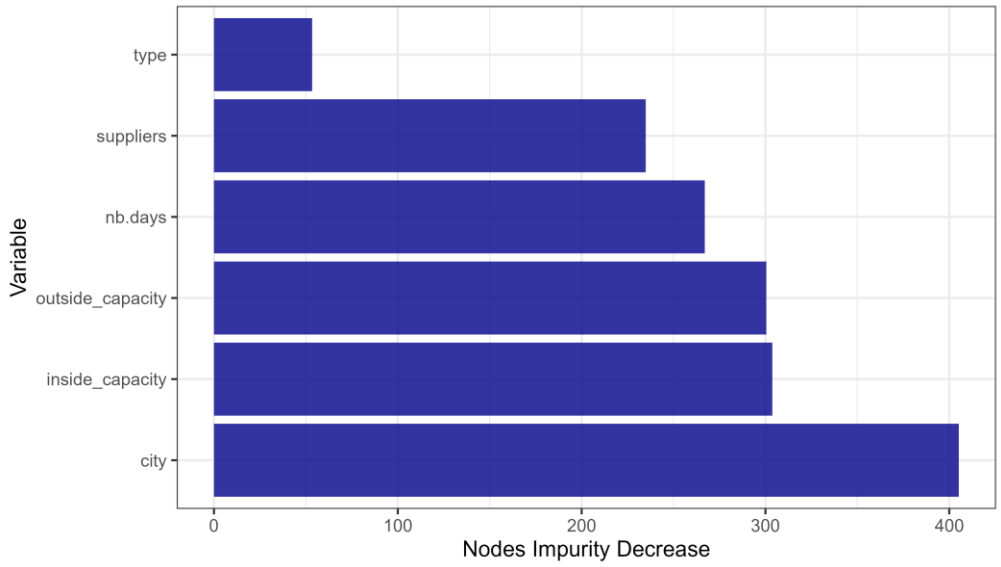

Figure 13 : Importance des variables selon la pureté des nœuds

Si nous considérons l'augmentation minimale de la pureté totale des nœuds de tous les 500 arbres (53.40 : celle de la variable type) et que toutes les variables ont été utilisées dans tous les arbres ; alors l'augmentation minimale de la pureté totale des nœuds par arbre est de 10.68%.

Sur l'échantillon test, les valeurs du MSE des 500 arbres varient entre 0.3 et 0.51 ; quant aux valeurs du R2, elles varient entre 56.1% et 74.31%.

En moyenne, le modèle de forêt arrive à expliquer 73.62% des informations du prix, et commet une erreur de 0.3 en termes de MSE.

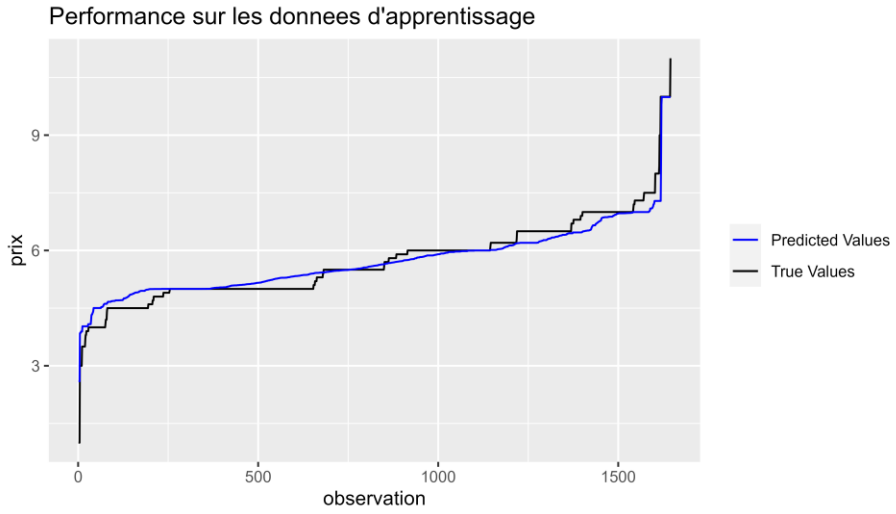

Figure 14 : Performances de la forêt aléatoire sur les données d'apprentissage

Afin de mieux évaluer la performance du modèle obtenu, nous réalisons une prédiction sur l'échantillon test. Nous calculons le MSE et nous obtenons 1.014 ; quant au R2, il vaut 19% :

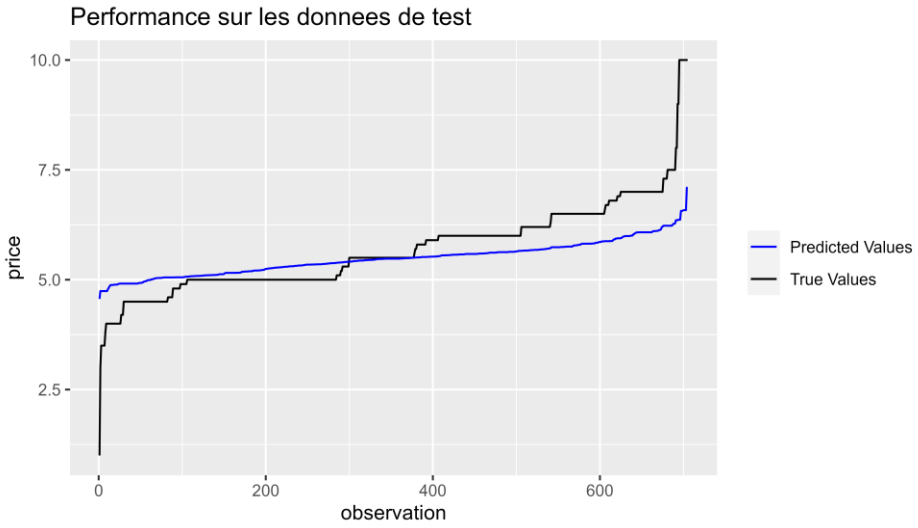

Figure 15 : Performances de la forêt aléatoire sur les données de test

Le modèle obtenu souffre de surapprentissage. Il est rejeté.

5.2 XGBoost

XGBoost (eXtreme Gradient Boosting) est une méthode d'apprentissage automatique basée sur des arbres de décision en cascade. L'un des avantages majeurs de XGBoost par rapport à la forêt aléatoire est sa capacité à gérer des données de grande dimension avec un nombre élevé de variables explicatives.

Il utilise un algorithme de descente de gradient pour optimiser une fonction de perte qui mesure l'erreur de prédiction, et il utilise également une régularisation L1 et L2 pour éviter le sur-apprentissage.

Cependant il n'accepte que des variables explicatives numériques. Nous transformons les variables qualitatives city, type et suppliers en variables numériques en utilisant la méthode du One-hot encoding.

- *Méthode du One-hot encoding*

Cette méthode consiste à créer une colonne binaire pour chaque modalité possible de chaque variable qualitative. Par exemple, si la variable "city" a trois modalités possibles : "Paris", "Lyon" et "Marseille", on créera trois nouvelles colonnes : "city_Paris", "city_Lyon" et "city_Marseille". Pour chaque ligne de données, la colonne correspondant à la modalité de la ville sera égale à 1 et toutes les autres seront égales à 0.

Après avoir codé toutes ces trois covariables, nous passons à de 6 à 402 covariables.

- Entrainement du modèle XGBoost

Le MSE est la fonction d'erreur à minimiser au cours de l'apprentissage. Le taux d'apprentissage ou la vitesse à laquelle le modèle apprend à chaque étape de l'entraînement qui est par défaut à 0.3 est réduit à 0.1. La profondeur maximale de chaque arbre de décision est fixée à 5. La fraction de l'échantillon d'apprentissage à utiliser pour l'entraînement de chaque arbre est 0.7. La fraction de variables à utiliser pour chaque arbre est de 0.8. La régularisation adoptée est une pénalisation de type Elastic-net. Le nombre d'itérations sur l'échantillon d'apprentissage est gardée à sa valeur de défaut 100.

Le MSE obtenu sur l'apprentissage est de 0.06, et le R2 vaut 95%. Sur l'échantillon test, le MSE vaut 0.26 et le R2 vaut 79%. Le modèle souffre d'un surapprentissage.

Nous décidons d'agir sur un autre paramètre du modèle XGBoost qui est la réduction minimale de la fonction de coût qui doit être réalisée avant qu'un nœud de l'arbre ne soit scindé pendant la construction de l'arbre. En effet, une valeur trop faible de ce paramètre peut être source de surapprentissage. Par défaut, il vaut 1. Nous l'augmentons à 2 : le MSE de l'apprentissage vaut 0.19 et celui de l'échantillon test vaut 0.34. Le problème de sur-apprentissage persiste toujours.

Nous gardons la réduction minimale de la fonction de coût à 2.

Nous agissons également sur le paramètre early_stopping_rounds. Le paramètre early_stopping_rounds permet d'activer l'arrêt précoce de l'apprentissage lorsque le MSE du modèle ne diminue plus après un certain nombre d'itérations. Il permet de trouver un point optimal où le modèle a une bonne capacité de généralisation sans trop s'adapter aux données d'entraînement.

Par défaut, il n'est pas activé. Nous testons trois de ces valeurs : 5, 10 et 20. Ci-dessous les résultats obtenus :

Valeur du paramètre early_stopping_rounds	MSE apprentissage	MSE test	R2 apprentissage	R2 test
5	0.21	0.33	82%	73%
10	0.22	0.35	81%	71%
20	0.2	0.32	83%	74%

Le meilleur parmi les modèles XGBoost testés, possède les paramètres :

- Taux d'apprentissage = 0.1.
- Profondeur maximale de chaque arbre de décision = 5
- Fraction de l'échantillon d'apprentissage pour l'entraînement de chaque arbre = 0.7
- Fraction de variables à utiliser pour chaque arbre = 0.8
- Pénalisation de type Elastic-net
- Le nombre d'itérations maximal sur l'échantillon d'apprentissage = 100
- Réduction minimale de la fonction de coût = 2
- early_stopping_rounds = 5

Son erreur de généralisation en termes de MSE vaut 0.33, tandis qu'à l'apprentissage le MSE vaut 0.21.

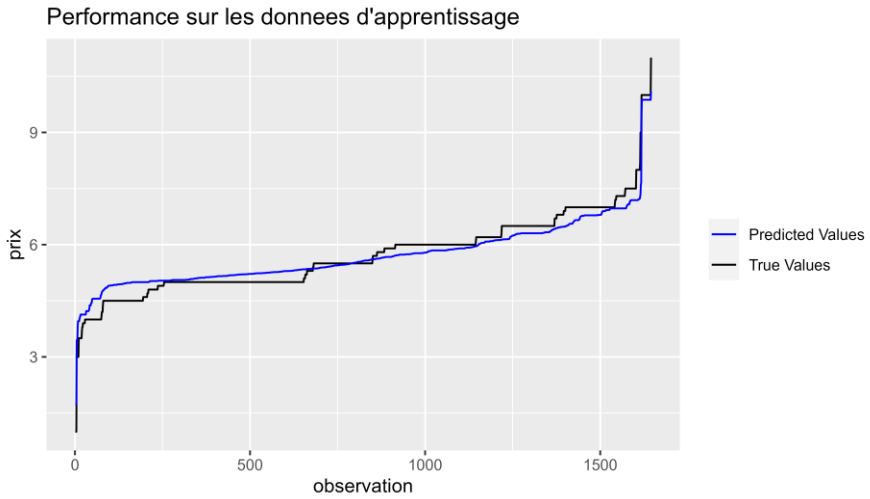

Figure 16 : Performance du XGBoost sur les données d'apprentissage

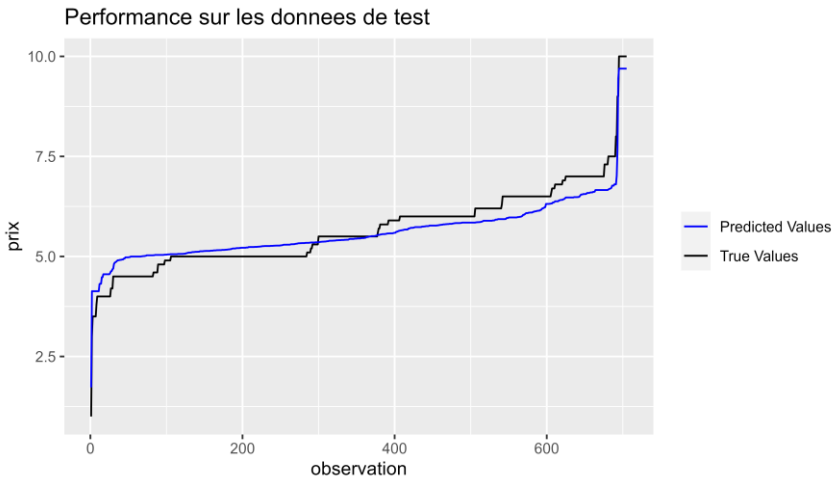

Figure 17 : Performance du XGBoost sur les données de test

5.3 Conclusion

Le surapprentissage est moins prononcé dans le cas du XGBoost que de la forêt aléatoire :

- *XGBoost : MSE apprentissage = 0.21, MSE test = 0.33*
- *Forêt aléatoire : MSE apprentissage = 0.3, MSE test = 1.014*

En effet, le XGBoost arrive à régler une part du surapprentissage dont souffre la forêt aléatoire. Cependant Il n'arrive pas à s'affranchir totalement du surapprentissage malgré tous les réglages que nous avons pu faire.

Ni le modèle de forêt aléatoire, ni le modèle de XGBoost ne sera également pas retenu.

6 Evolution du prix moyen de la marque DESPERADOS dans le temps

Dans cette partie, nous cherchons à modéliser l'évolution du prix moyen de la marque DESPERADOS dans le temps (indépendamment du type de contenant et du produit).

6.1 Construction de la série temporelle du prix moyen de la marque DESPERADOS

Avant de procéder à l'analyse de l'évolution du prix moyen de la marque DESPERADOS dans le temps, nous prenons soin de normaliser chacune des distributions de prix des produits DESPERADOS 25cl, DESPERADOS 33cl, DESPERADOS 50cl, DESPERADOSRED 33cl et DESPERADOS VIRGIN 33cl avant de les regrouper.

- *Prix moyen mensuel*

Dans un premier temps, nous procédons à une première agrégation mensuelle des prix afin d'avoir un prix mensuel. Après agrégation mensuelle, nous n'obtenons que 11 observations et de plus, nous constatons que de Mai 2022 à Novembre 2022, nous n'avons pas d'observations.

11 observations ne nous permettent pas de faire un bon ajustement de la série temporelle par un modèle. Nous choisissons de faire un regroupement bi-hebdomadaire des prix. De plus la période après Mai 2022 est écartée de la construction de la série temporelle du prix moyen bi-hebdomadaire.

- *Prix moyen bi-hebdomadaire*

Le nombre d'observations obtenues après construction de la série temporelle des prix moyens bi-hebdomadaire allant de Septembre 2021 à Mai 2022, est de 18. Ce nombre d'observations est toujours insuffisant pour l'ajustement d'un modèle.

Nous décidons de rester à un niveau de prix moyen hebdomadaire.

- *Prix moyen hebdomadaire*

Nous obtenons 30 observations de prix moyen hebdomadaire de Septembre 2021 à Mai 2022. Nous visualisons l'allure du prix hebdomadaire moyen ci-après :

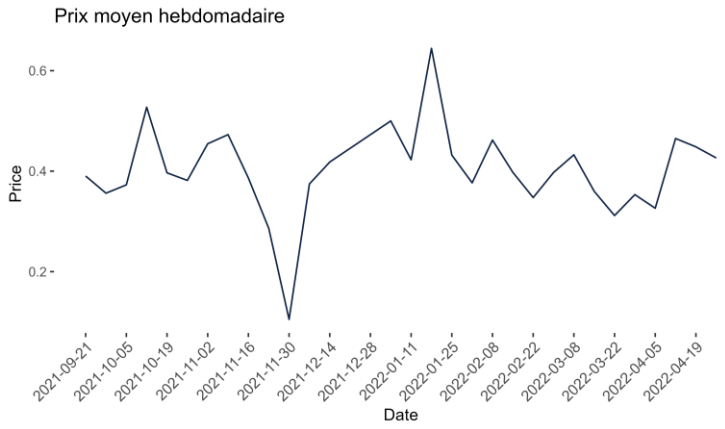

Figure 18 : Evolution du prix moyen hebdomadaire de Septembre 2021 à Mai 2022

Nous modélisons cette série dans la suite par un modèle SARIMA$(p, d, q)(P, D, Q)_s$.

6.2 Modélisation de la série temporelle du prix moyen hebdomadaire

- *Saisonnalité*

L'analyse de la Figure 18 permet à priori de déceler la présence d'une saisonnalité. Pour confirmer ce constat, nous visualisons l'autocorrélation de la série :

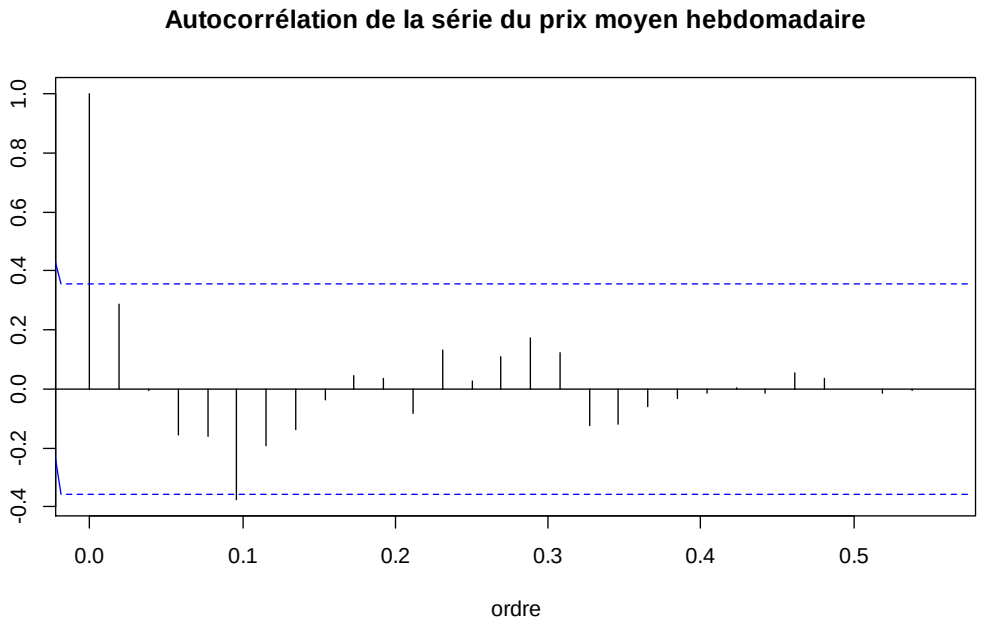

Figure 19 : Autocorrélation de la série du prix moyen hebdomadaire

L'analyse du graphe d'autocorrélation ne permet pas d'avoir une valeur exacte de la saisonnalité : les valeurs pouvant être choisies comme période sont de 3 semaines, 6 semaines et 9 semaines.

Nous choisissons la durée la plus faible constatée pour la période *s = 3*, et nous procédons à l'inférence des valeurs d, D, p, P, q et Q.

- *Tendance*

Après avoir désaisonnalisé la série de la Figure 18, nous analysons les graphique d'autocorrélation et d'autocorrélation partielle de la nouvelle série désaisonnalisée obtenue :

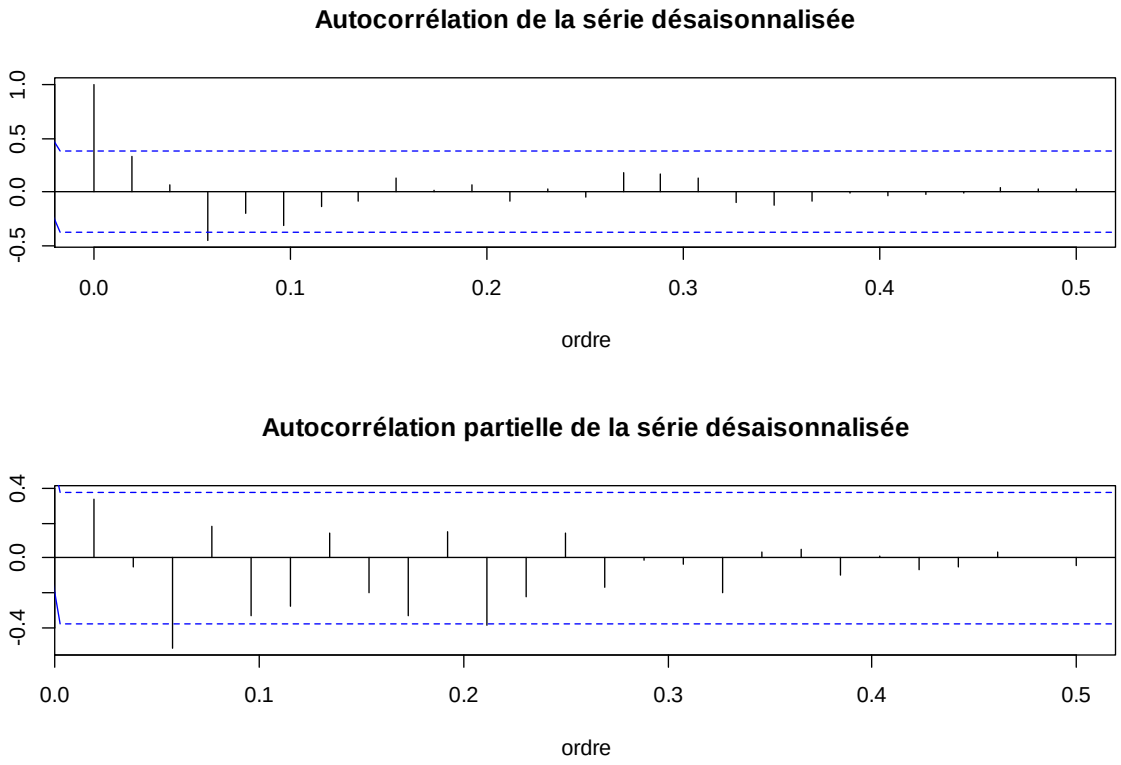

Figure 20 : Autocorrélation et autocorrélation partielle de la série du prix moyen hebdomadaire désaisonnalisée

La Figure 20 ne permet pas de se prononcer sur l'existence d'une tendance à priori. Cependant, par prudence nous estimons que Dmax = dmax = 1.

- *Ordres p, P, q et Q*

Quant aux ordres p, P, q et Q, nous estimons leurs valeurs maximales en analysant la Figure 20 : nous choisissons Pmax = 1, pmax = 3, Qmax = 1, qmax = 3.

- *Recherche du modèle optimal SARIMA(p, d, q)(P, D, Q)$_s$*

Nous faisons tourner tous les modèles SARIMA(p, d, q)(P, D, Q)$_s$ avec :

$3 \leq s \leq 9$, $0 \leq d \leq 1$, $0 \leq D \leq 1$, $0 \leq p \leq 3$, $0 \leq P \leq 1$, $0 \leq q \leq 3$, $0 \leq Q \leq 1$.

Nous testons 1792 modèles et nous ne retiendrons que celui avec l'AIC le plus faible.

Le modèle de plus faible AIC retenu est un SARIMA(0, 0, 1)(0, 0, 1)$_3$.

- *Validation du modèle optimal*

Afin de valider le modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$, nous devons étudier ses résidus.

Nous vérifions la normalité des résidus par un test de Shapiro de niveau de confiance 95% : la p-value obtenue vaut 0.01. Les résidus du modèle ne sont pas de loi normale.

Nous vérifions l'autocorrélation des résidus du modèle par un test de Box-Pierce de niveau de confiance 95% : la p-value obtenue vaut 0.62. Les résidus du modèle ne sont pas corrélés.

Nous validons le modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$. et nous visualisons la prévision du modèle sur les 30 observations de Septembre 2021 à Mai 2022.

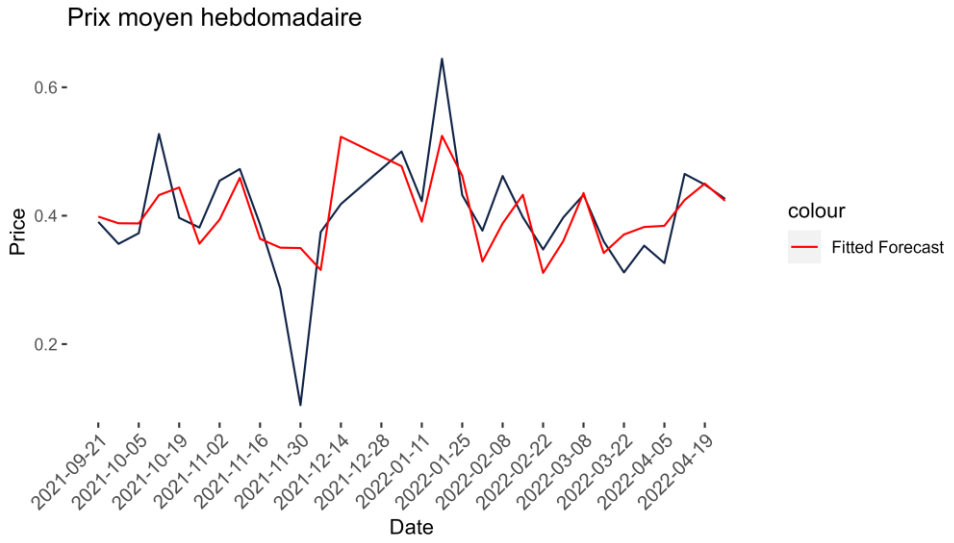

Figure 21 : Prévision du modèle SARIMA

Malgré la taille faible de l'échantillon (30 observations) dont nous disposons, le modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$ arrive à capturer l'évolution du prix moyen hebdomadaire de la marque DESPERADOS : d'où son MSE de 0.45%.

6.3 Conclusion

L'échantillon dont nous disposons permet de choisir un modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$ comme étant le plus adapté pour modéliser l'évolution du prix moyen hebdomadaire de la marque DESPERADOS.

Un taille plus conséquente d'observations serait la bienvenue afin de raffiner le modèle et ainsi augmenter ses performances.

7 Conclusion

Dans un premier temps, nous avons répondu à la question comment reconnaitre un contenant normalisé par la création d'un dictionnaire de mots. Ce dictionnaire de mots permet d'établir une correspondance entre un contenant personnalisé et sa valeur la plus probable de contenant normalisé pouvant lui être attribué.

Sur les 300555 observations initiales dont nous disposions, nous avons pu faire correspondre uniquement 82.08% d'entre elles à un contenant normalisé.

Dans un second temps, nous avons étudié la distribution des prix des produits de la marque DESPERADOS notamment DESPERADOS 33cl, DESPERADOS 25cl, DESPERADOS 50cl, DESPERADOSRED 33cl et DESPERADOSVIRGIN 33cl. L'analyse de leurs histogrammes a révélé que la moyenne est inadéquate comme indicateur de leurs prix. En effet, aucune des distributions de ces produits n'est symétrique, et certaines de ces distributions présentent des valeurs extrêmes.

Nous avons proposé d'autres statistiques dans ce contexte tels que la médiane (ou plus globalement les quartiles) qui communique des informations plus fiables sur les distributions des produits de la marque DESPERADOS.

Notre troisième fut de prédire le prix du produit DESPERADOS 33cl à l'aide des covariables city (localisation de l'établissement où la boisson est vendue), type (type d'établissement qui vend la boisson ; ex : bar, restaurant, bar-restaurant, etc…), suppliers (fournisseurs de l'établissement), inside_capacity (capacité intérieure en nombre de clients de l'établissement), outside_capacity (capacité extérieure en nombre de clients de l'établissement) et nb_days à l'aide de modèles de régression classiques telles que la forêt aléatoire et le XGBoost.

Ces deux modèles souffraient de surapprentissage, le surapprentissage étant très prononcé dans le cas de la forêt aléatoire. Aucun de ces deux modèles n'a été retenu.

Concernant ce dernier point, notre étude s'est terminée par l'analyse de l'évolution du prix moyen de la marque DESPERADOS. Vu les données dont nous disposions, l'analyse n'a pu se concentrer que sur l'évolution du prix moyen hebdomadaire.

Un modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$ a été retenu pour modéliser l'évolution de ce prix moyen. Il commet une erreur de 0.45% en termes de MSE. Cependant, nous ne pouvons nous prononcer à cet instant avec certitude sur le prix futur de la marque DESPERADOS sans entrainer le modèle SARIMA(0, 0, 1)(0, 0, 1)$_3$ sur plus d'observations. Mais, nous pouvons prédire d'une semaine à l'autre s'il y aura diminution ou baisse du prix hebdomadaire.