



12 March 2019

Alexandre ALLANI
Adrien CLOTTEAU

Rapport projet S5

Challenge Total



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Contents

1	Introduction	3
2	Contexte	3
3	Visualisation des données	4
4	Préparation des données (Data Engineering)	5
	Modification du jeu de données	5
4.1	Segmentation temporelle	5
4.1.1	Distinction des vacances	5
4.1.2	Distinction des dates de départs	6
4.1.3	Segmentation de la période selon les journées et mois	6
4.1.4	Suppression de la colonne date	6
4.2	Influence des catégories	7
4.3	Transformation des données	8
4.3.1	One Hot Encoding	8
4.3.2	Normalisation des données	9
5	Modélisation	10
5.1	Modélisation basée sur le Data Engineering	10
5.1.1	Méthodologie d'évaluation des modèles	10
5.1.2	Méthodologie d'amélioration des modèles	10
5.1.3	Régression linéaire	11
5.1.4	Decision Tree Regressor et Extra Tree Regressor	11
5.1.5	Random Forest Regressor	11
5.1.6	Gradient Boosting et XGBOOST Regressor	12
5.1.7	Multi-layer Perceptron (MLP)	12
5.1.8	Stacking	12
5.1.9	Post-traitement	13
5.2	Modélisation basée sur les Time Series	13
6	Analyse des résultats	14
7	Conclusion	15
8	Annexe	16
8.1	Vente par Catégorie de produit	16

1 Introduction



Vous trouverez l'ensemble de ce qui a été fait en suivant le lien ci-dessous.
<https://github.com/Syndorik/ChallengeTotal>.

En suivant le lien Github, vous trouverez:

- Challenge-Total-Overlook-of-the-data.ipynb : Jupyter Notebook sur la visualisation des données
- Data Engineering and Modeling.ipynb : Jupyter Notebook sur la préparation des données et la modélisation
- Time_series.ipynb : Jupyter Notebook sur la modélisation avec les Time Series.

2 Contexte

3 Visualisation des données



Le code lié à cette partie est disponible sur le lien suivant
[https://github.com/Syndorik/ChallengeTotal/blob/master/
Challenge-Total-Overlook-of-the-data.ipynb](https://github.com/Syndorik/ChallengeTotal/blob/master/Challenge-Total-Overlook-of-the-data.ipynb).

4 Préparation des données (Data Engineering)



Le code lié à cette partie est disponible sur le lien suivant https://github.com/Syndorik/ChallengeTotal/blob/master/Data-Engineering-and_Modeling.ipynb.

Dans cette partie, nous allons présenter la préparation des données faites avant les dernières modélisation. En suivant la méthodologie CRISP, cette préparation des données est le fruit de plusieurs tests/modélisation. Nous ne présenterons que les résultats fructueux, qui ont mené à une amélioration du modèle.

Modification du jeu de données Le jeu de données étant Français et l'interpréteur utilisé étant sous le format anglo-saxon, les valeurs numériques (tels que la température, la nébulosité etc.) étaient considérés comme des chaînes de caractère à cause de la virgule. Nous avons donc remplacé les "," par des "." pour pallier à ce problème.

Par ailleurs, le jeu de données ne contenant aucune valeur non acquises (NA values), il n'y a pas eu de pré-traitement à faire.

4.1 Segmentation temporelle

4.1.1 Distinction des vacances

L'une des premières améliorations effectuée au niveau des données est la segmentation du jeu de données en période de vacances et période de "travail". En effet, il est important de modéliser cette saisonnalité car les ventes sont très différentes pendant et hors des périodes de vacances. Le graphique ci-dessus, **Figure 1**

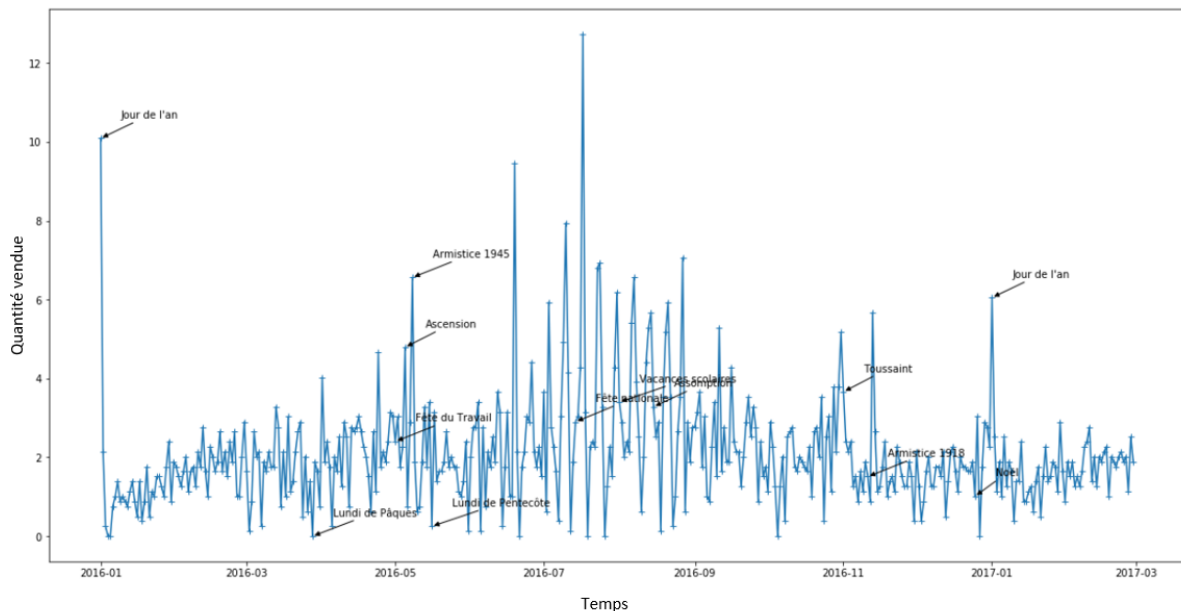


Figure 1: Vente de baguette au cours de la période

représente les ventes des baguettes sur les deux stations durant la période couverte par le jeu de données. Les baguettes sont le produit le plus vendu, tous produits confondus. Nous pouvons observer que les ventes lors des jours fériés et périodes de vacances sont beaucoup plus grandes. En particulier, les grandes vacances sont plus sujettes à de fortes variations.

Ces observations sont logiques, les périodes de vacances correspondent aux périodes de départs. Les familles vont en général plus s'arrêter dans les stations services, ainsi la quantité de produit vendu est beaucoup plus grande.

Segmenter l'année permet de mettre en évidence les périodes où les ventes sont très variables par rapport à la moyenne et donc permet d'affiner notre modèle. En se basant sur les [vacances 2015-2016](#) et les [vacances 2016-2017](#), nous avons pu segmenter la période selon les vacances scolaires et les zones de départ en créant une nouvelle variable "holidays".

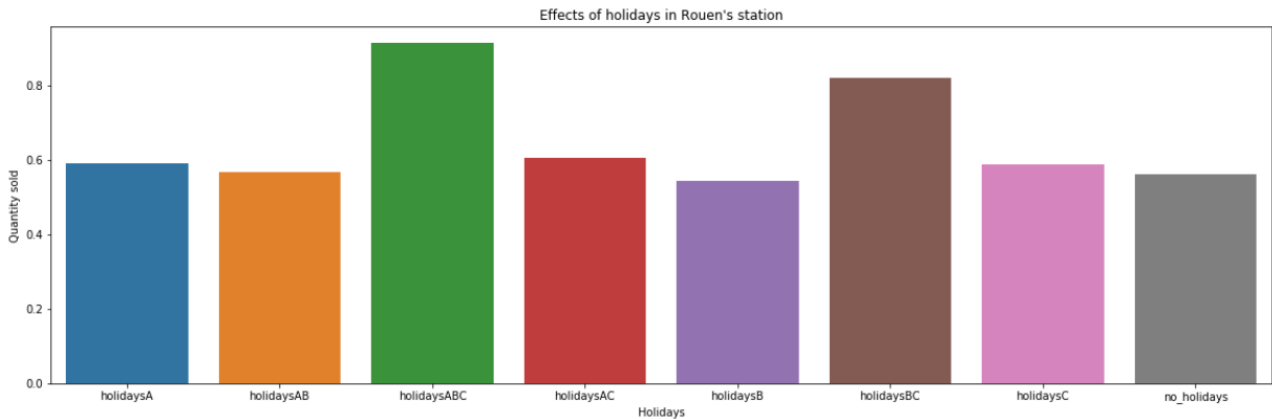


Figure 2: Vente cumulé par semaine selon la journée

Après création de ces variables, nous pouvons comparer les ventes moyennes entre les différentes périodes de vacances grâce à la **Figure 2**. Il apparaît qu'il y a une réelle différence lorsque toutes les zones (A, B, C) sont en vacances ou juste les zones (B,C). Dans les autres cas, il n'y a pas beaucoup de différence avec les périodes de travaux. Cela s'explique par le fait que les stations situés à Rouen sont en zone B, et la plupart des voyageurs se dirigeant vers Rouen doivent passer par Paris (zone C). Ainsi lorsque les deux zones sont en vacances en même temps, il y a bien plus d'aller-retour qu'en période normale. Nous choisissons de garder la discrimination entre les vacances pour les trois zones (A,B,C), les vacances pour les zones B et C, et finalement les périodes de "travail".

4.1.2 Distinction des dates de départs

Nous pouvons voir sur la **Figure 1** que les premières journées et dernières journées de vacances sont particulièrement haut en terme de quantité vendue. Cela n'est pas surprenant, ce sont des journées où beaucoup de famille sont sur les routes, ainsi il y a mécaniquement plus de ventes qui sont effectuées. Il est donc intéressant de discriminer ces journées afin de mieux les modéliser. Pour cela nous avons créé une nouvelle variable "departure" qui segmente le jeu de donnée selon les dates de départs.

4.1.3 Segmentation de la période selon les journées et mois

Dans l'optique d'exploiter au maximum les saisonnalités des ventes, nous avons continué à travailler sur les dates. Comme nous pouvons le remarquer sur la **Figure 1**, les ventes atteignent un maximum local de manière périodique environ tout les 7 jours. Il y a une périodicité remarquable au niveau de la semaine. En effet, le trajet Paris Rouen ne dure que 2 heures, il est envisageable que certaines personnes font l'aller-retour chaque semaine pour rentrer le week-end, et donc s'arrêtent aux deux stations services.

La **Figure 3** représente le nombre de vente cumulé sur la période du jeu de donnée en fonction des jours de la semaine. Elle met en évidence cette périodicité. Les ventes des journées du Vendredi et Samedi sont bien au dessus des autres jours de la semaine. De manière générale, les ventes sont faibles pendant la semaine et haute pendant le week-end. C'est pourquoi nous avons différencié chaque jour de la semaine dans notre jeu de données.

4.1.4 Suppression de la colonne date

Une fois ces transformations réalisés, nous supprimons la colonne date. En effet pour les modèles que nous allons utiliser (excepté pour les Time Series), garder les dates va entraîner du sur-apprentissage. Les modèles

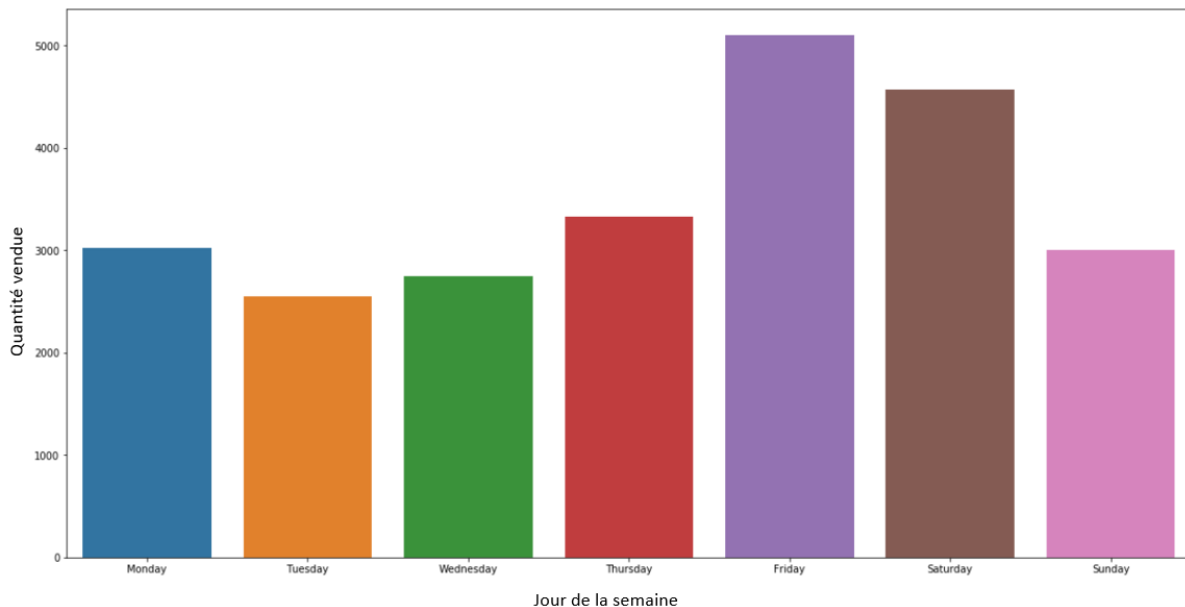


Figure 3: Vente cumulé par semaine selon la journée

vont apprendre sur la date exacte (jour-mois-année), sachant que ces dates sont passés, cela va entraîner une grande erreur.

Il est à noter que toutes ces segmentations temporelles ont pour but d'affiner le modèle, et retirer le plus d'information de la variable *date*. Cependant pour le modèle basé sur les Time Series l'ensemble de ces segmentations ne seront pas conservés, car incompatible avec le modèle. Nous expliciterons plus ce sujet dans la **section 5.2**

4.2 Influence des catégories

Nous voulons savoir si la distinctions par catégorie était utile, sachant qu'il y a inclusion des catégories : $cat6 \subset cat5 \subset cat4$. La **Figure ?? et ??** représente les ventes en fonction de la catégorie 5 pendant la période.

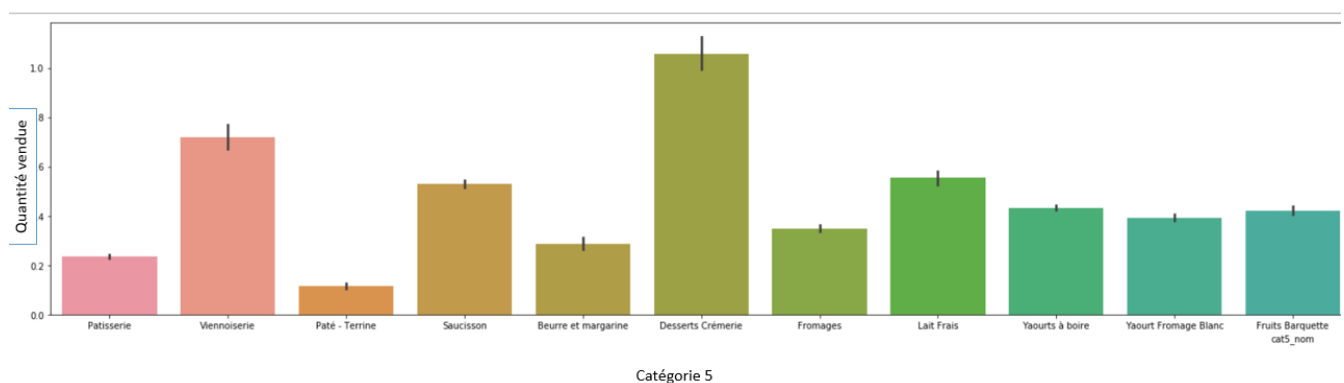


Figure 4: Vente cumulé en fonction de la catégorie 5 (partie 1)

Il est clair qu'il y a une différence de vente en fonction de la catégorie. La catégorie *Sandwich* ou encore *Dessert Crèmerie* domine, tandis que la catégorie *pâté* et *autre charcuterie* font de faibles ventes. Cependant comme on peut le voir sur la **Figure 10** (Annexe 8.1), il y a peu de différence entre les sous catégories, en particulier entre la catégorie 5 et 6 qui paraissent être presque les mêmes. Cela s'explique par l'inclusion des catégories.

En réalisant les tests de modélisation, nous avons supprimé la catégorie 5 pour éviter la redondance des

données, ce qui n'a pas affecté nos modèles. Ainsi on conservait le regroupement par catégorie qui permet d'avoir une version plus grossière de nos produit, et un certain clustering, tout en limitant le nombre de variables pour éviter le *Fléau de la dimension* ("*curse of dimensionality*").

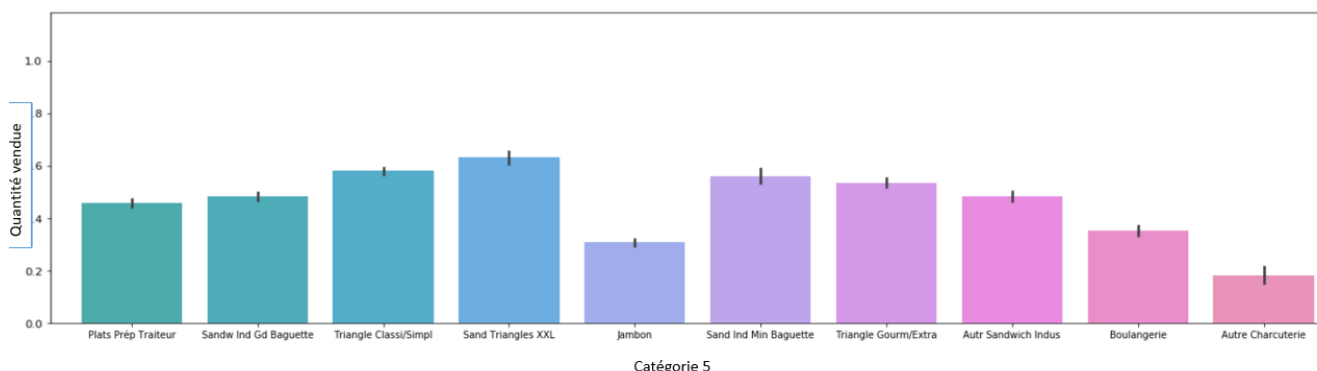


Figure 5: Vente cumulé en fonction de la catégorie 5 (partie 2)

4.3 Transformation des données

4.3.1 One Hot Encoding

Le One Hot Encoding est une transformation des données catégoriel en vecteur binaire. La **Figure 6** représente une telle transformation. Le principe est de créer un ensemble de nouvelles variables représentant

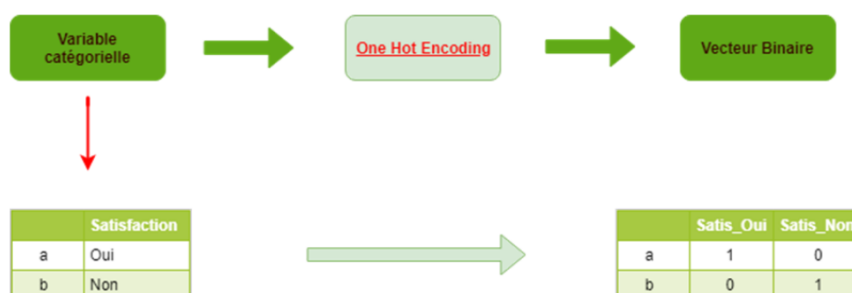


Figure 6: Principe du One Hot Encoding

une variables catégorielle (une variable par possibilité différente) prenant des valeurs binaires (0 ou 1). Il est important de faire cette transformation car plusieurs algorithme de machine learning n'arrivent pas à travailler avec les variables catégorielles. En l'occurrence, concernant notre projet les modèles suivants ne supportent pas les variables catégorielles:

- Modèle linéaire
- Réseau de Neurone

Cependant il y a plusieurs désavantage à cette technique:

- L'augmentation du nombre de dimensions qui est la cause de la "*curse of dimensionality*". Le traitement de donnée devient plus difficile et long, car les données prennent plus de place en mémoire
- Les éléments du vecteurs sont équidistant, impliquant que chaque modalité d'une variable catégorielle ait la même importance ce qui n'est pas le cas. (Par exemple, le fait qu'il y a des vacances en zone A, B et C est plus important que le fait qu'il n'y ait pas de vacances)
- Le vecteur binaire ne contient une information sémantique encodée, ce qui rend plus difficile la compréhension.

Cette méthode étant simple à appliquer et permettant l'utilisation direct de la majorité des algorithme, nous allons tout de même l'utiliser dans la suite du projet.

4.3.2 Normalisation des données

De même que pour le One Hot Encoding, il est nécessaire (ou préférable) pour certains algorithmes, en particulier le Réseau de Neurone, d'avoir des données normalisées. En effet le Réseau de Neurone ne prend en entrée que des données normalisées. Nous avons donc normalisé par rapport à la différence minimum/ maximum pour chaque variable numérique.

La normalisation permet aussi d'avoir une convergence plus rapide pour les algorithmes basés sur l'optimisation par gradient (Gradient boosting). Ce dernier se base sur une step size (le gradient de la fonction d'erreur multiplié par le taux d'apprentissage), cependant si les variables ne sont pas sur la même échelle, les step size sont différentes (car le gradient est différent) rendant la méthode de gradient boosting plus longue.

De plus elle permet d'avoir de meilleurs résultats pour le modèle linéaire, car celui-ci est moins affecté par les grandes variations. Les variables ayant un grand intervalle de définition vont induire une grande variance au modèle, ce qui leur donne trop d'importance.

Pour toutes ces raisons, nous appliquons une normalisation des données.

5 Modélisation



Le code lié à cette partie est disponible sur le lien suivant <https://github.com/Syndorik/ChallengeTotal/blob/master/Data-Engineering-and-Modeling.ipynb>.

Dans cette section, nous présentons l'ensemble des modèles qui ont été testés pour modéliser le problème posé par Argedis.

5.1 Modélisation basée sur le Data Engineering

Le modèle doit répondre à un problème de régression de machine learning supervisé. Ici, on considère la variable comme continue (classement supervisé), bien qu'elle soit en réalité discontinue. En effet, comme précisé en Introduction, la quantité vendue d'un produit est le multiple d'un certain nombre propre au produit. Ainsi dans l'idéal, un modèle aurait été développé par produit. Dans ce cas là, le problème aurait été un problème de classification. Cependant, au vu de la taille faible de notre jeu de données, il peut être compliqué de faire un modèle par produit. On a choisi de garder comme base d'entraînement l'ensemble des produits, cette

Une autre possibilité aurait été de faire un modèle apprenant sur les périodes de travail (hors vacances) et un modèle apprenant sur les vacances. Comme nous l'avons vu, les périodes de vacances sont anormales : période de haute variance de quantité d'articles vendues. Donc cela va avoir tendance à perturber les modèles. Cependant nous n'avons qu'une année et demi de données. La période trop courte pour pouvoir faire cela. Il est à noter que les Time Series vont nous aider sur cette problématique. Nous allons détailler l'ensemble des modèles suivant :

- Régression linéaire
- Decision Tree Regressor
- Extra Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBOOST
- Multi-layer Perceptron (MLP)

5.1.1 Méthodologie d'évaluation des modèles

Comme nous l'avons énoncé en Introduction, notre modèle est évalué selon le RMSE : $\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$. La méthode que nous avons utilisée pour s'assurer de la validité d'un modèle est la validation croisée (cross-validation). Le principe est d'entraîner le modèle sur une partie du jeu de données, 80% par exemple et d'utiliser les 20% restant pour le tester. Le processus est répété un nombre n de fois, et la moyenne des scores RMSE obtenus nous permet d'évaluer le modèle

5.1.2 Méthodologie d'amélioration des modèles

Chaque modèle, que nous allons expliciter dans les parties suivantes, possède des paramètres qui lui sont propres. Pour chaque modèle nous allons énoncer les paramètres les plus importants. Cependant la méthodologie pour optimiser ces derniers est la même. Pour chaque combinaison de paramètres, nous allons lancer l'apprentissage d'un modèle et son évaluation par cross-validation (en RMSE).

Les paramètres du modèle ayant la meilleure évaluation seront considérés comme optimum. Dans la pratique, il existe une multitude de paramètres pour chaque algorithme, en général nous n'appliquons cette méthode que pour certains d'entre eux.

5.1.3 Régression linéaire

Ce modèle essaie de décrire les relations entre la variable à prédire et les variables d'entraînement comme linéaire (du type : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots$ avec Y la variable à prédire et les X_i les variables d'entraînements). La méthode utilisée pour créer la courbe de régression est la méthode des moindres carrés. On obtient après entraînement un score R^2 de 35.4%, ce qui veut dire que ce modèle explique 35.4% de la variance. Le score RMSE est le suivant



RMSE = 0.9568

Ce score est assez faible surtout en comparaison des autres modèles. Cela est assez logique, car les relations au sein de notre jeu de données ne sont clairement pas linéaire, et il est assez difficile de modéliser des relations non linéaire avec ce modèle. C'est possible via l'introduction de termes d'interaction, mais cela reste assez primaire comme modélisation.

5.1.4 Decision Tree Regressor et Extra Tree Regressor

Ces modèles se basent sur la création d'arbre afin de fournir une prédiction. Le principe est que chaque feuille de l'arbre de décision va utiliser une variable pour effectuer une partition des données. Les arbres de décision permettent d'avoir un modèle qui est compréhensible en sortie d'apprentissage. Nous obtenons les résultats suivant :



$RMSE_{Decision-tree-Regressor} = 0.669$ et $RMSE_{Extra-tree-regressor} = 0.646$

La différence entre le Decision Tree Regressor et l'Extra Tree Regressor repose sur la manière séparer la population au sein d'un noeud afin de créer une branche. L'Extra Tree Regressor va prendre un certain nombre de variables au sein d'un noeud, regarder quelle est la meilleure manière de séparer le noeud à partir de ces variables et recommencer le processus plusieurs fois. Le meilleur *split* est alors choisi.

Cela explique pourquoi l'Extra Tree Regressor possède un meilleur score RMSE. Cependant, le grand inconvénient de ces méthodes est qu'elles ont tendance à sur-apprendre ou au contraire pas assez apprendre en fonction de la profondeur de l'arbre. C'est pourquoi nous nous sommes concentrés sur les modèles ensemblistes d'arbre de décision.

Paramètres principaux à optimiser : la profondeur de l'arbre, le nombre minimum de feuilles

5.1.5 Random Forest Regressor

Les forêts d'arbre de décision ou random forest, se base sur la création de plusieurs arbres de décision. Le principe est de créer plusieurs arbres de décision de manière aléatoire : sélection aléatoire des variable et du jeu d'apprentissage (bootstrap sample). Une fois l'apprentissage des arbres fini, il suffit de moyenner la sortie de chaque arbre de décision pour obtenir la variable à prédire. C'est le principe du bagging. Nous obtenons les résultats suivant :



RMSE = 0.57997

Le score obtenu est meilleur que pour les précédents modèles utilisés. Cela s'explique par l'aspect général du modèle qui permet de compenser les problèmes de sur-apprentissage que l'on peut rencontrer avec un arbre de décision unique. L'inconvénient de cette méthode est le temps d'exécution qui peut être assez long. Paramètres principaux à optimiser : la profondeur de l'arbre, le nombre minimum de feuilles, le nombre d'arbre maximum

5.1.6 Gradient Boosting et XGBOOST Regressor

Ces deux modèles sont des versions modifiés des forêts d'arbre de décision. Au lieu de se baser sur du bagging pour créer les arbres de décisions, ils se basent sur des méthodes de gradient boosting. Le principe est de créer des arbres de décisions à la chaîne. Cependant entre chaque création d'arbre, l'erreur est minimisée en utilisant la méthode de descente du gradient. Nous obtenons les résultats suivant :



$$RMSE = 0.57909$$

Le score obtenu est meilleur que pour les forêts d'arbres aléatoires. Cependant, la différence se fait au millième du RMSE. Sachant que notre prédiction est continue alors que la quantité vendue par produit prend des valeurs discontinue, la différence obtenue n'est pas significative. En effet l'erreur dû à l'échantillonnage est plus grande.

On peut trouver une explication à ces résultats par le fait que les deux algorithmes reposent sur la création d'arbre de décision. Les forêts d'arbre aléatoires vont en général essayer de réduire la variance tandis que les méthodes de boosting vont essayer de réduire le biais. Ici, il apparaît que dans les deux cas nous obtenons une erreur similaire.

Paramètres principaux à optimiser : le taux d'apprentissage, la profondeur de l'arbre, le nombre minimum de feuilles, le nombre d'arbre maximum.

5.1.7 Multi-layer Perceptron (MLP)

Cet algorithme repose sur la création d'un réseau de neurone. Il va apprendre une fonction à partir du jeu de données. L'avantage est la capacité de modéliser les non linéarités. Le réseau de neurone comporte 4 couches dont deux caché. La première est l'input layer, la seconde est composée de 32 neurones, la deuxième de 16 neurones, et la dernière de une neurone (correspondant à la variable à prédire). La fonction d'activation utilisé est la fonction "relu".



$$RMSE = 0.685$$

Le score obtenu est moins bon que pour les autres modèles testés auparavant. Cela est probablement du au fait que notre réseau de neurone n'est pas adapté à notre problème et qu'il pourrait être optimisé. Cependant le temps d'apprentissage du réseau étant relativement long, nous n'avons pas insisté sur cette voie.

5.1.8 Stacking

Cette méthode ensembliste permet de combiner les meilleurs modèles précédemment créé afin d'obtenir un modèle plus performant. Dans notre cas, nous avons pris 3 modèles différents, le modèle linéaire, le modèle de forêt aléatoire et le modèle Gradient Boosting. Le principe est décrit en **Figure 7**. Le jeu de données est divisé en différente base d'apprentissage pour différent modèle. Les prédictions faites à partir des modèles vont ensuite servir de base d'apprentissage au meta-modèle. Dans notre cas ce sera un XGBOOST. Nous obtenons alors le résultat suivant :



$$RMSE = 0.5735$$

Le score obtenu est meilleure mais ne varie que de quelques centièmes. En effet, le stacking est utile pour combiner des modèles qui ont des zones d'erreurs assez différentes. Les modèles basés sur les forêts d'arbres de décision et les modèles basés sur XGBOOST ont des zones d'erreurs assez similaires. Ainsi, nos modèles sont trop similaire pour bénéficier du stacking.

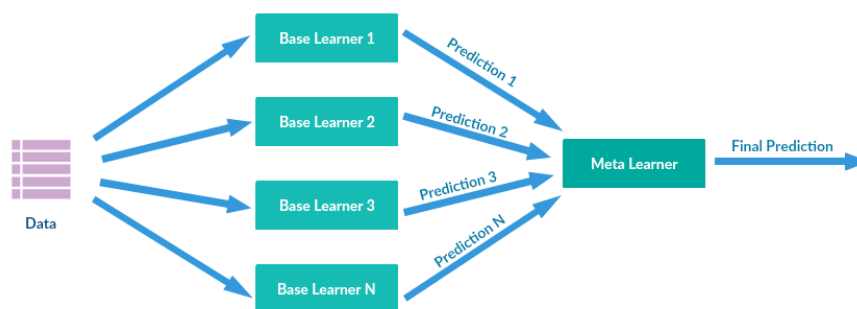


Figure 7: Principe du stacking

5.1.9 Post-traitement

Nous avons implémenté deux post-traitement en vue d'améliorer la qualité de notre modèle.

- Il existe des produits qui étaient disponible dans une station mais pas dans l'autre, et vice-versa. Nous savons dès lors, que s'il est demandé de prédire le nombre de vente d'un produit qui n'existe pas dans la station, il sera prédit automatiquement à 0.
- Nous savons qu'un produit peut prendre des valeurs de quantité vendue qui ne sont que des multiples d'une quantité de base. Cette quantité de base est différente pour chaque produit. Ainsi, nous avons tenter d'approximer chaque quantité vendu soit par sa valeur de multiple supérieur soit par sa valeur de multiple inférieure.

Le premier point du post-traitement est assez efficace et améliore effectivement les prédictions des modèles. Le deuxième point est bien plus problématique. Nous l'avons testé sur les résultats du modèle stacké et nous sommes passés d'un RMSE de 0.5735 à un RMSE de 0.72. Nous nous attendions à une régression légère du score RMSE, cependant ici cela est bien trop grand. Il y a plusieurs explications possibles. Tout d'abord, notre post-traitement ce trompe pour chaque produit et retourne toujours la mauvaise valeur (ie il retourne la valeur au-dessus quand la valeur réelle est en dessous et inversement). Cette première explication semble assez improbable. Une seconde explication serait une mauvaise compréhension de la variable quantité vendue. En effet, aucune information dans le jeu de donnée ne nous dis ce qu'elle représente exactement.

5.2 Modélisation basée sur les Time Series



Le code lié à cette partie est disponible sur le lien suivant https://github.com/Syndorik/ChallengeTotal/blob/master/Time_series.ipynb.

6 Analyse des résultats

Etudes par produit aurait pu être plus fructueuse. Cela n'a pas été fait, mais peut donner des meilleurs résultats, surtout que les ventes de produits sont très différentes et varie différemment. Plusieurs analyses pour l'échec du post-processing : Pas assez précis, une erreur de compréhension de la variable à prédire, toujours une mauvaise prédiction (en haut quand c'est en bas et vice versa)

7 Conclusion

8 Annexe

8.1 Vente par Catégorie de produit

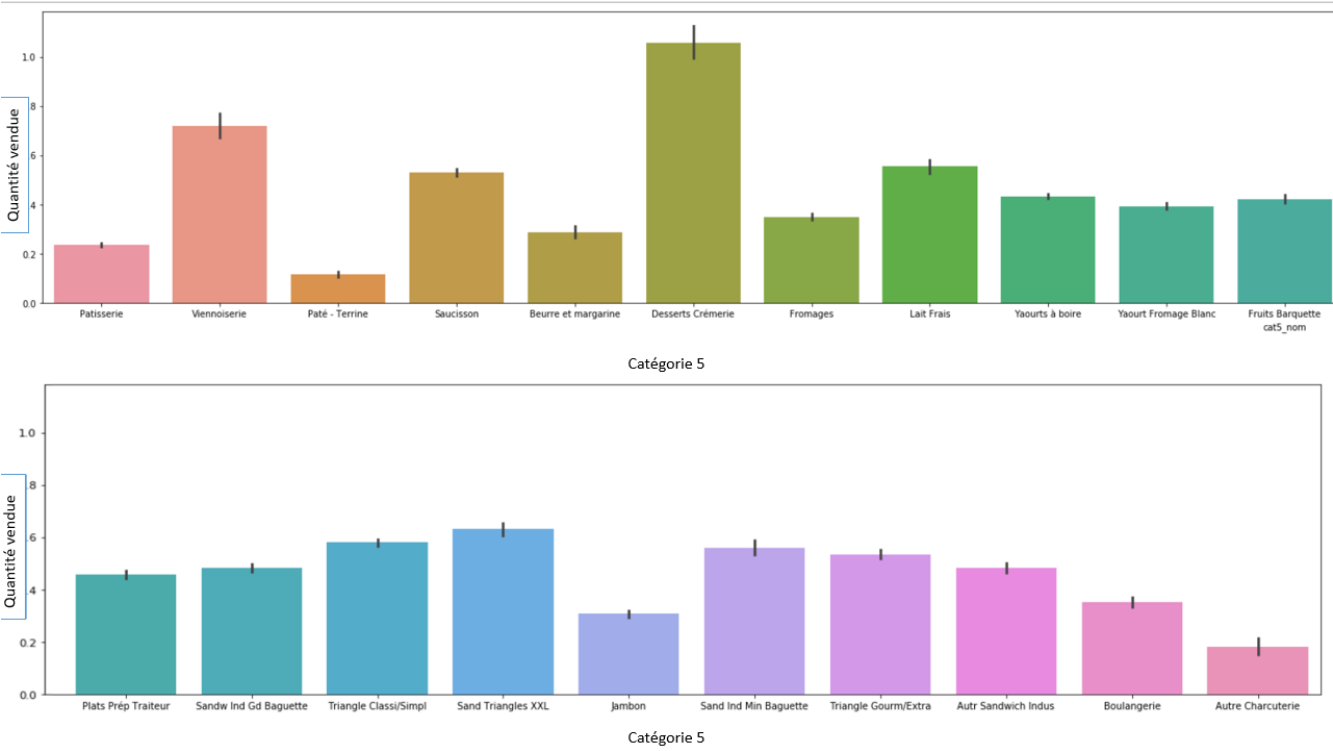


Figure 8: Vente cumulé en fonction de la catégorie 5

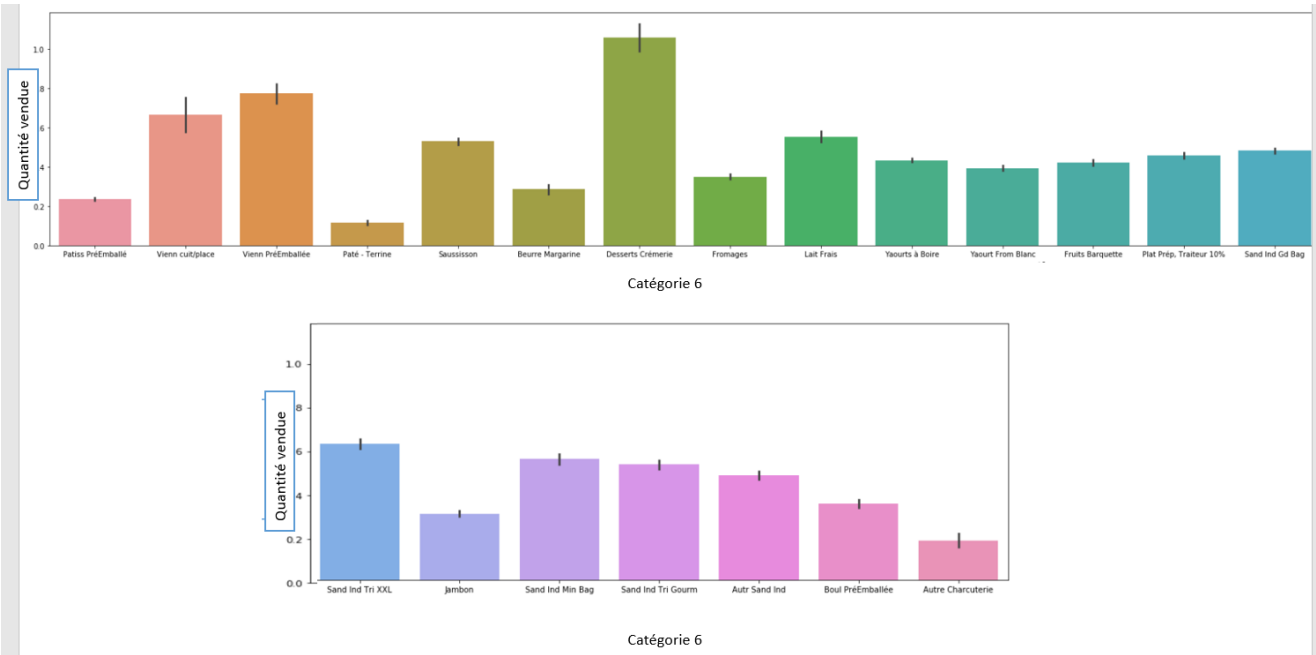


Figure 9: Vente cumulé en fonction de la catégorie 6

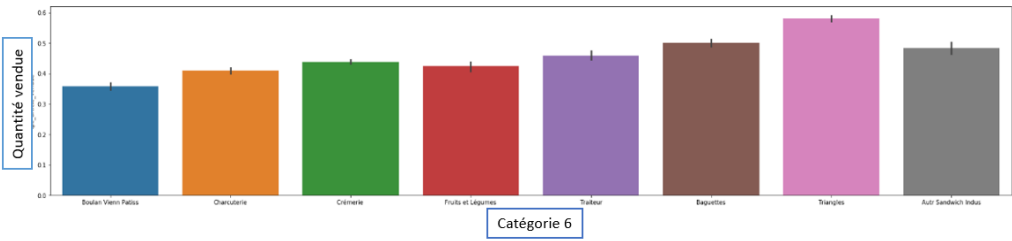


Figure 10: Vente cumulé en fonction de la catégorie 4