Étude sur les prix de produits électroniques vendus entre particuliers

Mohamed Maftah
mohamed.maftah@polymtl.ca

Mehdi Miah
mehdi.miah@gmail.com

19 avril 2018

1 Introduction

1.1 Contexte et enjeux

Es-tu désireux d'arrondir tes fins de mois simplement et rapidement ? La solution : revends tes anciens vêtements ou jeux-vidéo qui prennent la poussière dans tes placards. En effet, les sites de ventes de particulier à particulier sont légions sur le web depuis le mythique eBay, en passant par leboncoin ou encore le fameux Kijiji. Ces sites rencontrent une forte affluence et doivent traiter des articles de nature très variée : de la console de jeux aux livres en passant par les vêtements.

Malgré une apparente simplicité, il peut être difficile pour le vendeur de compléter certaines informations, sans la supervision d'un opérateur, telle que le prix de vente d'un produit. Ce choix est d'autant plus critique qu'il est aussi essentiel pour le futur acheteur que pour le vendeur. Pour ce dernier, le sous-estimer signifierait des revenus moindres et le sur-estimer reviendrait à perdre en compétitivité.

Les utilisateurs de Mercari, un site communautaire de ventes d'articles basé au Japon, connaissent exactement ce problème. En partenariat avec Kaggle, un site de compétition en data science, Mercari souhaite développer une méthode permettant de prédire le prix d'un article une fois que l'utilisateur a indiqué la nature du produit et son état. Ces prédictions permettront alors au site web de recommander des prix au futur vendeur, de sorte à ce qu'il puisse vendre son produit au prix du marché, et donc aussi faciliter la recherche des produits par les futurs acheteurs lors de leurs requêtes.

Les données sont disponibles sur `www.kaggle.com/c/mercari-price-suggestion-challenge/data`.

1.2 Objectifs

L'objectif de ce projet est de déterminer le prix d'un article à partir de sa nature, son état (neuf et/ou degré d'usure) et d'une description textuelle du produit en anglais.

1.3 Méthodologie

L'objectif principal de ce projet est de prédire le prix d'un article à partir des méthodes traitées en IND6212. De plus, il sera également pertinent de comprendre les données et d'en extraire de l'information pour évaluer la pertinence du modèle et identifier les principales difficultés.

Pour mesurer la qualité de prédiction, le challenge sur Kaggle reposait sur la métrique RMSLE (Root Mean Squared Logarithmic Error) : pour N produits, si les véritables prix sont $y = (y_i)_{i\in[\![1,N]\!]}$ et les prix estimés $\hat{y} = (\hat{y}_i)_{i\in[\![1,N]\!]}$, l'erreur mesurée est alors

$$\mathrm{RMSLE}(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\ln(1+y_i) - \ln(1+\hat{y}_i)]^2}$$

Pour contrôler le sur-apprentissage, la base de données sera segmentée aléatoirement : une partie servira à l'apprentissage (80% des données) et l'autre partie à tester le modèle (20% des données).

2 Présentation et traitement des données initiales

Les données initiales de Kaggle contiennent près de 1.5 millions de produits répartis en 10 familles, dont 'Electronics' qui regroupe 8.3% des données de départ soit 122632 articles.

Il a été décidé de centrer l'analyse sur cette famille de produits car elle nous est plus familière et présente la plus grande variabilité monétaire ce qui la rend particulièrement intéressante à étudier.

La table 1 indique le format des données de la base.

Variable	Description	Format	Modalité/étendue
train_id	Identifiant du produit	Numérique	1 :999988
name	Nom du produit	Texte	
item_condition_id	Etat du produit	Catégorielle	1 :5
cat2	Catégorie du produit	Catégorielle	8 modalités
cat3	Sous-catégorie du produit	Catégorielle	51 modalités
brand_name	Marque du produit	Catégorielle	307 modalités
price	Prix en dollars US auquel l'article a été vendu	Numérique	de 0 à 1909
shipping	1 si les frais d'envoi ont été payés par le vendeur, 0 sinon	Numérique	0 ou 1
item_description	Description du produit	Texte	

TABLE 1 – Format des variables de la base de données

2.1 Pré-traitement des données

S'agissant de données renseignées par le vendeur lui-même, et non par un "opérateur expert", la première étape est de les valider en étudiant ses valeurs manquantes, ses granularités et sa cohérence.

Voici une liste de toutes les particularités observées au sein des articles électroniques :

- 1 observation présente un nom manquant : il a été décidé de la laisser telle quelle ;
- 7507 observations présentent une description manquante : il a été décidé de mettre un marqueur précis ("NoDescriptionYet") pour évaluer l'impact du non renseignement de description sur le prix de vente ;
- 62629 observations n'ont pas de marque : il a été décidé de combler les valeurs manquantes par un marqueur précis ("brandMissing") pour évaluer l'impact de l'absence de marque ;
- il y a 308 marques différentes. Après avoir obtenu leur fréquence, il a été décidé de conserver les six plus fréquentes (dans l'ordre décroissant en nombre d'articles : "brandMissing" à 51%, Apple, Nintendo, Sony, XBox et Samsung à 3.3%) et de regrouper les autres marques sous un nom commun ("otherBrand"), qui représentent au final 9.9% des articles ;
- 58 observations ont un prix de vente nul (don ?) : il a été décidé de les retirer de l'étude car ce sont des valeurs aberrantes ;
- en accord avec le RMSLE, il a été décidé de prédire $\texttt{log\_price} = \ln(\texttt{price} + 1)$ et de mesurer la qualité de la prédiction par un traditionnel RMSE [1]. Par ailleurs, le rapport utilisera le terme "prix" pour désigner autant la transformée logarithmique que le véritable prix. L'ambiguïté sera levée par la présence de l'unité (en dollar américain) du véritable prix et par les ordres de grandeur ;
- certains articles ne sont pas classifiés dans les bonnes sous-catégories. Il est ainsi possible que près de 10% des articles dans "Cell Phones & Smartphones" correspondent à des accessoires (estimation faite à partir de la présence de mots relatifs aux accessoires dans le nom du produit).

Une fois ces pré-traitements effectués, la base de données est enregistrée et un extrait a été déposé sous Moodle.

1. cela permet également d'exploiter les propriétés mathématiques de l'estimateur des moindres carrés

2.2 Analyse descriptive des données numériques et catégorielles

Avant la phase de prédiction, il convient d'évaluer l'effet (global) de chaque paramètre sur le prix. En premier lieu, en terme de volumétrie, 70% des produits proviennent soit de la catégorie "Cell Phones & Accessories", soit de "Video Games & Consoles" et plus précisemment des sous-catégories "Games" (21.6% des articles) et "Cases, Covers & Skins" (20.1%).

En second lieu, en terme de lien avec le prix de vente, les catégories des produits affectent fortement le prix : en effet, le prix médian d'un article en "Media" est de 10$ contre 40$ pour des produits de "Computers & Tablets".

Puis, la marque affecte également le prix : les produits sans marque ont un prix médian de 11$, contre environ 17$-22$ pour les produits de marque Apple, Nintendo, Samsung, Sony ou XBox. Étrangement, les produits les plus chers sont ceux des autres marques, dont le prix médian est de 36$.

Ensuite, l'effet de l'état du produit (neuf, comme neuf, usé, etc) sur le prix est contre-intuitif : en effet, à partir des données, il semblerait à première vue que plus un produit est usé, plus il est cher. Cela est dû à un biais de sélection : les articles vendus en occasion sont en général des produits qui avaient une certaine valeur à l'achat, tels que les ordinateurs ou des smartphones. A l'opposé, les accessoires, peu chers, ne sont pas revendus sur le marché de l'occasion mais plutôt jetés. A vrai dire, il semblerait qu'il existe, pour chaque sous-catégorie, une corrélation entre le prix médian et la part de produits d'occasion : pour une sous-catégorie, plus le prix médian est élevé, plus la part de produits d'occasion est forte. Cet effet est illustré par la figure 1.

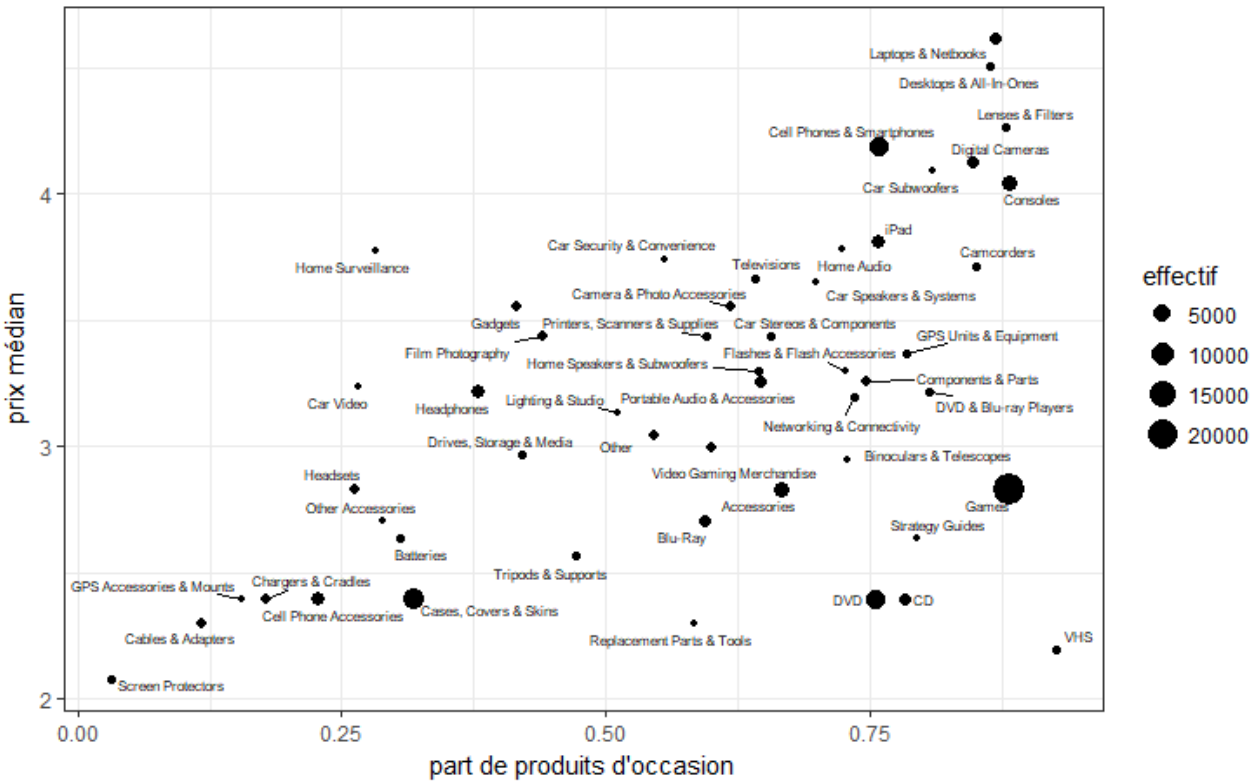

FIGURE 1 – Part de produits d'occasion et prix médian par sous-catégorie d'articles

Et enfin, des résultats similaires peuvent être trouvés avec les frais d'expédition : plus la sous-catégorie a un prix médian faible, plus la probabilité que les frais d'envoi soit prise en charge par le vendeur est élevée. Ainsi, 87% des protections des écrans de smartphone ont des frais de port inclus contre 40% des ordinateurs portables. Cela s'explique par un effet psychologique : plus un produit est cher, plus la part des frais de port est négligeable dans la somme totale et plus l'acheteur est consentant à les payer.

2.3 Analyse descriptive des données textuelles

Après avoir étudié les variables numériques et catégorielles, il est temps d'étudier le cas des données textuelles, qu'il faudra tout d'abord formater.

2.3.1 Formatage des données textuelles

S'agissant de données non structurées, il convient tout d'abord de formater les variables textuelles : les noms des produits et leurs descriptions. La méthodologie suivante a été adoptée :

- mise en minuscule : l'effet des majuscules est ignoré ;
- suppression des signes de ponctuation ;
- conservation des nombres : parfois, les nombres désignent des modèles (iPhone 6, PS4) ou bien des caractéristiques (64 Gb, 5 inches) qui peuvent influer sur le prix ;
- tokenisation : seuls les mots (1-gram) sont considérés ;
- pas de correction orthographique ;
- suppression du marqueur "[rm]" introduit par Kaggle dans les données textuelles pour censurer les indications sur le prix du produit ;
- conservation des stopwords classiques pour les noms car les prépositions "for" ou encore "with" permettent de spécifier le modèle ou des suppléments ;
- pas de lemmatisation : on souhaite différencier les pluriels des singuliers. En effet, "games" a une autre signification que "game" en terme monétaire.

2.3.2 Résultats des analyses de text mining

A partir du formatage effectué dans la partie 2.3.1, il est possible de mettre en valeur les mots qui apparaissent le plus fréquemment dans les noms des produits, et qui donc sont susceptibles d'avoir une plus grande importance dans un modèle de prédiction.

En considérant les noms des articles comme corpus de texte, il est possible d'obtenir le nuage de mots illustré par la figure 2 indiquant les mots les plus souvent utilisés. Les résultats sont cohérents avec ceux obtenus précédemment montrant la forte présence des produits relatifs aux smartphones et aux jeux vidéo.

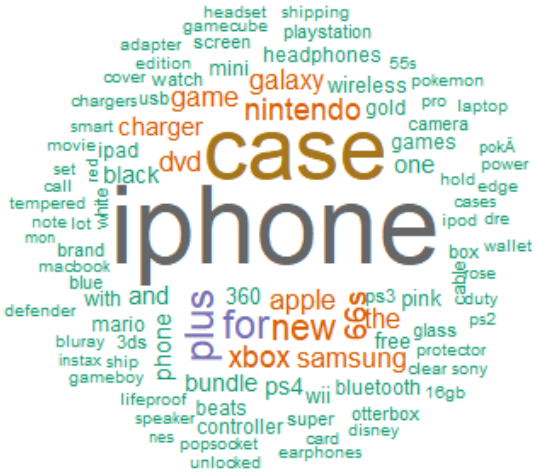

FIGURE 2 – Nuage des mots utilisés dans les noms des produits

Le nuage de mots ne se base uniquement que sur la fréquence d'apparition des mots. Cette approche ne permet pas de quantifier la valeur monétaire de chacun des substantifs du corpus. Pour cela, il est préférable d'utiliser les règles d'association : il sera possible de mesurer l'effet de l'apparition d'un mot particulier dans un produit sur son prix.

3 Règles d'association

Pour un produit donné, il serait par exemple souhaitable d'associer l'apparition du mot "iphone" à un produit cher, et inversement, à un produit peu cher si le mot "case" apparaît.

3.1 Mise en matrice et application

A partir du corpus sur les noms des produits, également utilisé pour le nuage de mots de la figure 2, et en ne conservant que les mots qui apparaissent dans au moins 0.1% des produits (cela permet de réduire la dimension de la matrice par un facteur 45), une matrice avec 122632 observations et 505 colonnes est obtenue. Chacune des colonnes indique la présence ou nom du mot associé par une valeur binaire.

De plus, afin d'associer une valeur monétaire aux mots, une colonne est rajoutée en catégorisant les prix des produits en trois groupes : "Cheap" (pour les 25% des produits les moins chers), "Expensive" (pour les 25% des produits les plus chers) et "Moderate" pour le reste. Ces seuils ont été choisis empiriquement.

Puisque les algorithmes descriptifs vus en cours donnent les mêmes règles et que le temps de calcul n'est pas un problème dans ce contexte, l'algorithme A-Priori est utilisé avec les paramètres suivants : support à 0.1%, confiance à 80% et des règles de longueur 4 (jusqu'à 3 mots dans la partie des antécédents). Ces paramètres on été choisis après plusieurs essais.

3.2 Résultats

Afin de déterminer l'impact des mots sur le prix, il est plus pertinent d'étudier ceux impliquant avec une forte probabilité les classes "Cheap" ou "Expensive", pour obtenir les règles les plus intéressantes.

Ainsi, nous obtenons 82 règles dont la table 2 en résume cinq.

Antécédents	Conclusion	Support	Confidence	Conviction
{wireless,beats,solo}	{Expensive}	0.0017	0.9952	4.0842
{edition,classic}	{Expensive}	0.0010	0.9776	4.011
{unlocked,iphone}	{Expensive}	0.0042	0.9735	3.9952
{iphone,case,slim}	{Cheap}	0.0013	0.9060	3.4562
{protectors}	{Cheap}	0.0015	0.8596	3.2791

TABLE 2 – Quelques règles d'association

Contrairement à l'approche fréquentiste (nuage de mots), il est possible ici de déterminer les mots qui impliquent un prix élevé ou non. Ainsi, certains mots comme "iphone" et "case", n'ont pas d'influence à eux seuls sur le prix mais que c'est leur combinaison avec d'autres mots qui affecte la valeur monétaire.

4 Prédiction du prix des produits

Une fois que les données ont été comprises, il est temps de prédire le prix d'un produit.

4.1 Méthode baseline

La première méthode consiste à évaluer le prix moyen par groupe partageant les mêmes critères : sous-catégorie (51 modalités), frais d'envoi (2), état (5) et présence de marque (2). A partir de ce modèle de base, qui servira à comparer les performances d'un modèle plus sophistiqué, l'erreur atteint 0.67.

Cette méthode ne réside que sur une partition de l'espace des articles en 1020 groupes. Cependant elle ne permet pas de détecter les produits mal classés, de gérer les biais de sélection et les données textuelles.

4.2 Random forest

La méthode complexe est un random forest utilisant la matrice créée à partir des noms de produits construite avec la méthodologie des parties 2.3.1 et 3.1. Cela permet l'obtention d'une matrice binaire de taille 122632×505. A cela, sont rajoutées six variables numériques ou catégorielles : la catégorie, la sous-catégorie, la marque, la gestion des frais d'envoi, l'état et la transformée logarithmique du prix. Cette dernière sera ainsi prédite à partir de 510 variables (dont 506 binaires).

Après plusieurs essais en variant le nombre d'arbres et le nombre de variables par arbre, la forêt a été construite avec 40 arbres à 40 variables chacune [2]. En accord avec la méthodologie décrite en 1.3, l'apprentissage s'effectue sur 80% des données choisies aléatoirement.

Les résultats de ce modèle sont très satisfaisants :
— le score obtenu sur la base de test est de 0.53 ;
— lors de l'apprentissage, la forêt aléatoire calcule un score de test (à partir des observations non choisies) qui vaut 0.53 : cela montre que le sur-apprentissage est évité ;
— l'importance des variables est illustrée dans la figure 3 : les cinq variables supplémentaires sont les plus importantes. Puis, viennent les variables binaires classées par puissance de discrimination. De plus, les prépositions "for" et "with" apparaissent dans les 30 variables les plus importantes : cela confirme notre intuition lors de la conservation des stopwords dans la partie 2.3.1.

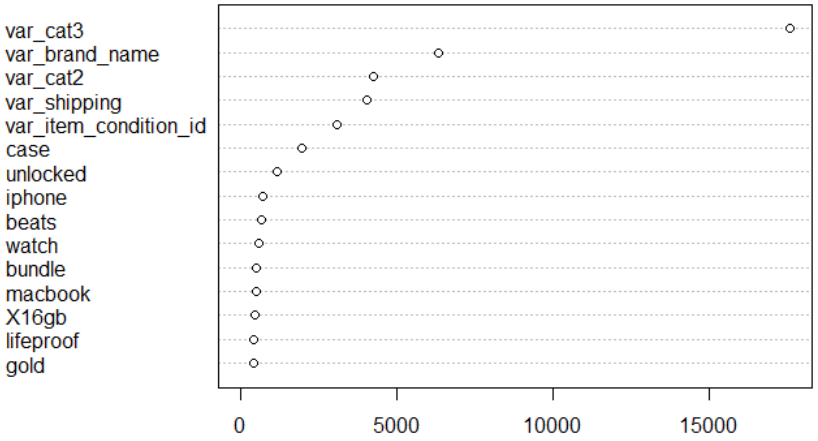

FIGURE 3 – Importance des 15 premiers variables dans le random forest

5 Conclusion

Le modèle de forêt aléatoire réduit l'erreur de 20% par rapport à celle de la méthode baseline. Toutefois, ce modèle n'indique pas l'effet sur le prix : pour cela, il faut se référer aux règles d'associations pour interpréter. De plus, il ne permet pas de gérer les effets de composition : en créant de faux produits où seules les variables de frais d'envoi et d'état changent, il n'est pas clair que l'état, toute chose égale par ailleurs, réduit le prix ; de même avec les frais d'envoi. Ainsi, si cette méthode devait être implémentée sur le site de Mercari, il faudrait mieux proposer à l'utilisateur des intervalles de prix, et non pas un prix fixe.

Le modèle du random forest peut-être amélioré en intégrant les bigrams ou trigrams (cf les règles d'association) et les descriptions du produit. Par ailleurs, les meilleurs modèles utilisés sur Kaggle résident pour la très grande majorité sur les réseaux de neurones profonds, qui atteignent un score de 0.38 sur l'ensemble des familles de produits (et non pas que sur 'Electronics').

L'ensemble du code est disponible sur Github : `www.github.com/Guepardow/Mercari`

2. cette forêt nécessite 50mn de calcul, et est plus performante qu'une forêt de 60 arbres à 170 variables (le tiers du nombre de paramètres qui est la valeur par défaut du package randomForest sous R) et qui nécessite 10h de calcul