Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Critères d'évaluation des classificateurs par apprentissage automatique : application aux statistiques de prix

par Serge Goussev et William Spackman

Date de diffusion : le 5 novembre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Critères d'évaluation des classificateurs par apprentissage automatique : application aux statistiques de prix

Serge Goussev et William Spackman[1]

Résumé

La modernisation des statistiques sur les prix par les organismes nationaux de statistique (ONS), comme Statistique Canada, met l'accent sur l'adoption d'autres sources de données qui comprennent presque la totalité de l'univers des produits vendus dans le pays, une échelle qui nécessite la classification des données par apprentissage automatique. Le processus d'évaluation des classificateurs permettant de sélectionner ceux qui conviennent à la production ainsi que de surveiller les classificateurs une fois qu'ils servent à la production doit être fondé sur des paramètres robustes pour que soit mesuré le taux de classification erronée. Étant donné que les mesures couramment utilisées, comme le score Fβ, peuvent ne pas tenir compte des principaux aspects applicables aux statistiques de prix dans tous les cas, comme l'importance inégale des catégories, il faut examiner attentivement l'espace métrique pour choisir les méthodes appropriées d'évaluation des classificateurs. Le présent document de travail présente l'espace métrique applicable aux statistiques de prix et propose un cadre opérationnel d'évaluation et de surveillance des classificateurs, en portant un intérêt particulier aux besoins de l'Indice des prix à la consommation du Canada et en démontrant les paramètres étudiés au moyen d'un ensemble de données accessibles au public.

Mots clés : indice des prix à la consommation; classification supervisée; mesures d'évaluation; taxonomie

1. Introduction

Les nouvelles sources de données sont aujourd'hui essentielles pour les organismes nationaux de statistique (ONS) parce qu'elles complètent les sources traditionnelles comme les données recueillies sur le terrain. Pour être employées dans le cadre des statistiques officielles sur le prix, comme l'Indice des prix à la consommation (IPC), les nouvelles sources à volume élevé comme les données transactionnelles (données de lecteurs optiques ou technologie de point de vente) et les données en ligne (par moissonnage des sites Web de détaillants) doivent d'abord être correctement classées aux fins de correction des catégories de produits au moyen de classificateurs d'apprentissage automatique.[2] Lors de l'élaboration de classificateurs, les ONS doivent évaluer et sélectionner un classificateur approprié aux fins de production, ainsi que surveiller son rendement sur une base mensuelle (Eurostat, 2017; Martindale et coll., 2020, 19). Une des questions à laquelle doivent répondre les ONS consiste à savoir quelles mesures s'appliquent à l'évaluation et à la surveillance des classificateurs dans l'espace des statistiques de prix. En effet, le score Fβ et d'autres mesures couramment utilisées ne tiennent pas compte de plusieurs aspects clés applicables aux statistiques de prix dans tous les cas, notamment de la nature hiérarchique de la taxonomie selon laquelle tous les enregistrements doivent être classifiés ainsi que de l'importance inégale de certaines catégories dans la taxonomie par rapport à d'autres. Pour mieux faire comprendre l'espace métrique disponible et contribuer aux considérations théoriques antérieures sur la sélection de mesures dans les statistiques de prix (UK Statistics Authority, 2019), le présent document de travail examine les mesures d'évaluation hiérarchique ainsi que d'autres critères applicables aux statistiques de prix de façon à appuyer le travail d'évaluation, de sélection et de surveillance des classificateurs par les ONS. Le document met particulièrement l'accent sur les mesures applicables à la classification à classes multiples, puisque les efforts des ONS sont axés principalement sur la classification d'ensembles de données contenant plusieurs segments de consommation.

[1] Serge Goussev, Statistique Canada, 150, promenade Tunney's Pasture, Canada, K1A 0T6 (serge.goussev@statcan.gc.ca); William Spackman, Statistique Canada, 150, promenade Tunney's Pasture, Canada, K1A 0T6 (william.spackman@statcan.gc.ca);

[2] Voir le chapitre « 10 : Données numérisées » du *Manuel de l'indice des prix à la consommation : concepts et méthodes* (International Labour Office 2020), pour en savoir plus sur le processus d'intégration de données provenant d'autres sources – comme les données de moissonnage – à l'indice des prix à la consommation, par exemple.

L'article est structuré de la façon suivante. La première section donne un aperçu des objectifs opérationnels et d'apprentissage automatique ainsi que des mesures applicables à l'évaluation de modèles. Ensuite, on propose un cadre opérationnel qui comporte les mesures applicables à une analyse et une surveillance efficaces des classificateurs mis en œuvre. Enfin, on propose une démonstration empirique à partir d'un ensemble de données ouvertes pour démontrer l'application des mesures étudiées dans un cas d'utilisation réaliste.

2. Objectifs et mesures proposées

2.1 Mesures de la qualité de la classification

La classification à classes multiples des enregistrements dans les statistiques de prix consiste à attribuer chaque enregistrement d'un ensemble de données au niveau le plus bas ou au niveau de la feuille d'une taxonomie hiérarchique à arborescence.[3] En général, dans l'IPC, la classification se situe au niveau de la catégorie des produits élémentaires (PE), ou d'un ensemble de biens ou de services relativement homogènes ayant des utilisations et des mouvements de prix semblables, soit le niveau le plus bas pour lequel des pondérations des dépenses sont disponibles (Bureau international du Travail, 2020). Les catégories de classification pourraient également être plus basses selon les besoins et les données des ONS, puisque certaines catégories de PE ne sont pas parfaitement homogènes et que plusieurs solutions de rechange sont à l'étude (Office for National Statistics [ONS du Royaume-Uni], 2020). Après le classement de tous les enregistrements, les indices de prix sont calculés au moyen d'une formule d'indice de prix différente.

Deux types de méthode de classification peuvent s'appliquer à ce type de problème : soit celles qui tiennent compte de la taxonomie et celles qui ne le font pas (Silla et Freitas, 2011). Si une classification erronée doit être pénalisée de façon égale, un cas où chaque catégorie de PE est supposée avoir un mouvement de prix indépendant des autres catégories, le cas le plus simple s'applique. Dans ce cas, les chercheurs peuvent ignorer la hiérarchie des classes et prédire au niveau le plus bas ou au niveau de la feuille; cette méthode utilise une classification plate conforme aux méthodes classiques d'apprentissage automatique supervisé (Silla et Freitas, 2011; Costa, et coll., 2007). Les mesures d'évaluation traditionnelles, comme la précision, le rappel et le score Fβ (avec une sélection de β appropriée), s'appliquent à ce cas et ont été largement adoptées par les ONS (Office for National Statistics, 2021; Office for National Statistics, 2020; Martindale, et coll., 2020).[4]

Plusieurs approches des scores Fβ doivent être prises en compte dans l'application de la mesure aux statistiques de prix. Premièrement, il est essentiel de tenir compte du score Fβ par catégorie de chaque classificateur pour mieux distinguer les classificateurs qui peuvent avoir de bonnes performances globales, mais avoir tendance à prédire mal un certain nombre de catégories. Dans les statistiques de prix, cela est d'autant plus important que certaines catégories sont plus importantes que d'autres en raison de leur importance relative plus élevée dans les dépenses de consommation. Deuxièmement, les scores Fβ par catégorie peuvent être agrégés en un score Fβ par modèle pour fournir une vue d'ensemble des performances du modèle. Il est possible d'appliquer trois techniques d'agrégation : une pondération d'échantillonnage, une pondération égale et une pondération des dépenses. Un score Fβ pondéré par l'échantillon est une moyenne du score Fβ par classe, pondérée par le nombre d'enregistrements présents dans chaque classe pendant l'évaluation, ce qui biaise de façon inhérente la mesure vers un résultat global plus élevé dans les modèles qui ont de bonnes performances sur les classes plus grandes. Il est aussi possible d'utiliser un score Fβ pondéré également ou par macro, qui applique une moyenne arithmétique simple de tous les scores Fβ par catégorie sans pondération. Cela garantit l'égalité entre toutes les classes, ce qui favorise les modèles ayant de bonnes performances dans tous les cas, même dans les classes avec peu de points de données. Enfin, on peut utiliser un score Fβ pondéré par les dépenses, basé sur une moyenne pondérée de tous les scores Fβ par catégorie, avec des poids tirés des poids du panier de dépenses de l'IPC représentant l'importance relative de chaque catégorie. Cette méthode

[3] Les taxonomies courantes des statistiques sur les prix comprennent des classifications internationales des biens et des services comme la Classification des fonctions de consommation des ménages (COICOP), le Système de classification des produits de l'Amérique du Nord (SCPAN), ou des taxonomies propres à des ONS, comme la classification de l'Indice des prix à la consommation utilisée par Statistique Canada dans le calcul de l'IPC.
[4] Voir UK Statistics Authority 2019 pour une présentation détaillée de ces paramètres.

favorise les modèles qui donnent de bons résultats pour les classes relativement plus importantes en ce qui a trait à l'utilisation finale des données classifiées, un cas particulièrement applicable à des statistiques sur les prix comme l'IPC.

2.2 Mesures de qualité de la classification hiérarchique

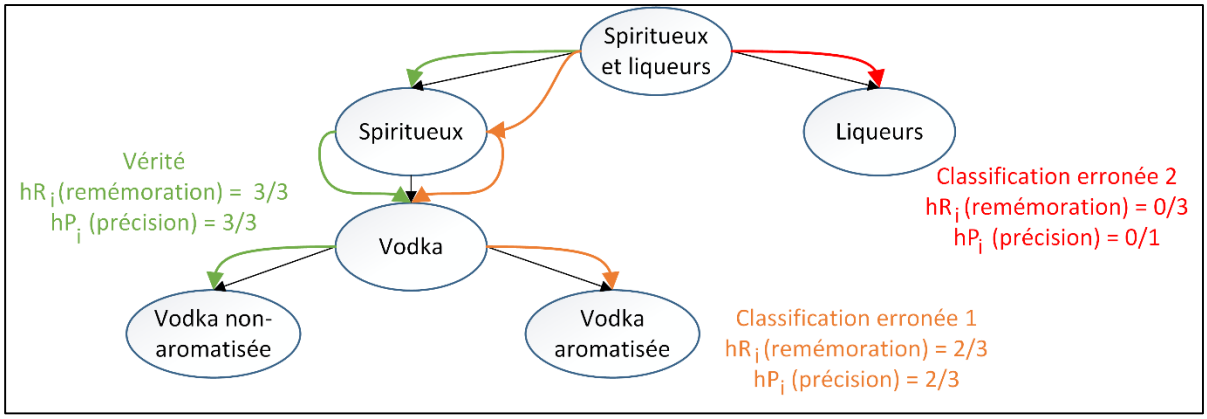
La supposition d'une indépendance entre les catégories et la pénalisation égale des classifications erronées sont inhérentes aux mesures de rendement normalisées. Cette méthode ne peut pas entièrement être appliquée à la classification dans le contexte des statistiques sur les prix. La classification erronée d'un enregistrement dans une catégorie semblable peut être moins problématique que sa classification dans une catégorie sans lien. En effet, de cette façon, le classificateur reproduirait de plus près l'action d'humains qui feraient plus d'erreurs dans des catégories étroitement liées. Cela refléterait l'application dans le monde réel, car les efforts des ONS en matière d'étiquetage manuel des données ont montré que le maintien de l'uniformité de la classification n'est pas négligeable pour les catégories étroitement liées, surtout quand un produit se situe à la limite de deux classes possibles (Office for National Statistics 2020, 7). La méthode hiérarchique est également applicable si les chercheurs des ONS utilisent des méthodes de classification personnalisées, propres à la taxonomie, non plates et plus complexes, comme la méthode à classificateur local ou une méthode à classificateur global (big bang). Dans les méthodes à classificateur local, un pipeline de classificateurs séquentiels soutient un arbre décisionnel de choix, où chaque enregistrement est classé dans l'ensemble pour atteindre le dernier niveau le plus bas, alors que dans une approche globale, un modèle de classification unique et relativement complexe est entraîné sur les données et toute la hiérarchie de classe est prise en compte (Silla et Freitas, 2011). Les ONS peuvent trouver ces approches efficaces et applicables dans les cas où un autre fournisseur de données couvre une multitude de segments de consommation disparates.[5]

Au sein de la classification hiérarchique, une considération clé est la « proximité » des catégories dans la pénalisation des classifications erronées (Sun et Lim, 2001), les erreurs dans les catégories étroitement liées étant considérées comme moins problématiques que celles dans des catégories éloignées. Dans les statistiques sur les prix, les taxonomies sont conçues dans la plupart des cas pour classer des catégories semblables sous le même parent, si bien que la classification erronée des enregistrements « proches » les éloigne habituellement d'un seul lien par rapport à la classification correcte. Les mesures de profondeur ou le fait de se concentrer sur le nombre de liens correctement prédits, du nœud racine au nœud final de la feuille pour chaque enregistrement, sont un moyen applicable et interprétable de quantifier la proximité aux fins des statistiques de prix. Les mesures de profondeur peuvent comporter une limite, en cas de classification qui peut être arrêtée au nœud parent sans continuer au nœud d'enfant le plus bas. Or ce cas ne s'applique pas aux statistiques sur les prix, car tous les enregistrements doivent être classés jusqu'au niveau de feuille le plus bas.

Comme pour les mesures de classification classiques, trois mesures hiérarchiques peuvent être calculées : hP_i représentant la précision hiérarchique, ou le rapport des liens de taxonomie correctement prédits sur les liens totaux précités; hR_i représentant le rappel hiérarchique, ou le rapport des liens de taxonomie correctement prédits sur le nombre réel de liens; et $hF\beta$ représentant le score Fβ hiérarchique. La figure 1 montre une application visuelle de ces mesures, avec trois exemples. Plus précisément, la classification correcte d'un enregistrement donnerait un rappel et une précision de 1, tandis que la pénalité pour classification erronée dépend de la distance par rapport à la vérité. La classification de la Vodka non aromatisée dans la Vodka aromatisée utilise deux liens corrects déplacés dans la hiérarchie, ce qui donne un rappel de 2/3 et une précision de 2/3. La classification de l'enregistrement dans Boissons alcoolisées donne une précision et un rappel de 0. Une fois qu'un score Fβ par catégorie est calculé au moyen de ces mesures hiérarchiques, les scores Fβ propres au modèle agrégé pondéré par panier, échantillonnage et macro peuvent également être déterminés de la même manière qu'avec les mesures de classification classiques.

[5] Par exemple, de nombreuses tâches de classification des vêtements traitent des données de détaillants, où l'entreprise vend des produits qui tomberaient à la fois dans la catégorie COICOP 03 (Vêtements et chaussures) et COICOP 09 (Loisirs et culture), plus précisément 09.3.2 (Articles de sport, matériel de camping et matériel pour activités de plein air).

Figure 1
Exemple de mesures de classification hiérarchique avec une classification correcte et deux classifications erronées



2.3 Considérations relatives aux opérations et aux processus

Pour appuyer la sélection initiale d'un classificateur performant à des fins de production, il faut tenir compte d'autres contraintes relatives aux ressources et aux activités. Premièrement, alors que la plupart des statistiques sur les prix sont calculées de façon mensuelle ou trimestrielle, les données sont reçues régulièrement, et leur classement, la vérification de leur qualité, et leur agrégation sont réalisés selon un calendrier strict. Bien que cela soit peu probable, les modèles qui nécessitent un long délai de prédiction peuvent limiter le temps consacré à d'autres tâches des ONS. De même, si l'infrastructure de TI est limitée, il faut peut-être aussi prendre en compte la taille de stockage du modèle, car certains modèles d'apprentissage automatique occupent beaucoup d'espace sur le disque. Enfin, il faut parfois aussi considérer la complexité et l'explicabilité du modèle, ainsi que la capacité d'appuyer le processus régulier d'assurance de la qualité, si un modèle est utilisé avant que les données prédites soient incorporées dans un indice publié.

3. Critères d'évaluation proposés

En conjuguant les mesures ci-dessus, on propose une méthode en trois étapes permettant d'évaluer un modèle de classification et de le surveiller après son déploiement.

3.1 Étape 1 – Analyse des mesures agrégées

Premièrement, il faut sélectionner le β dans le score Fβ. Deuxièmement, les mesures classiques de la qualité de la classification applicables (les scores Fβ étant $F\beta_{mw}$ avec pondération égale ou par macro, $F\beta_{sw}$ avec pondération d'échantillonnage, et $F\beta_{bw}$ avec pondération du panier) et les mesures de la qualité de la classification hiérarchique (les scores Fβ hiérarchiques étant $hF\beta_{mw}$ avec pondération égale ou par macro, $hF\beta_{sw}$ avec pondération d'échantillonnage; et $hF\beta_{bw}$ avec pondération du panier) des sections 2.1 et 2.2 peuvent être comparées isolément pour chaque modèle ou conjuguées dans une moyenne permettant de générer un score global par modèle pour en faciliter l'interprétabilité (1). Il faudrait mener des recherches sur les mesures qu'il faut inclure dans le score total pour savoir si certaines peuvent être exclues et déterminer le type de moyenne à utiliser. Le score final doit être interprétable, ce qui est accompli grâce à une moyenne qui limite le score à une valeur entre 0 et 1, et sensible aux valeurs aberrantes, ce qui donne un score total inférieur pour le modèle si un type spécifique de score Fβ est faible. La formule 1 utilise une moyenne arithmétique comme exemple de score du modèle final.

$$Model\ score\ = \frac{1}{6}[F\beta_{mw}\ + F\beta_{sw} + \ F\beta_{bw} + hF\beta_{mw} + hF\beta_{sw} + hF\beta_{bw}] \quad (1)$$

3.2 Étape 2 – Analyse par classe

Parallèlement à l'étape 1 ci-dessus (section 3.1), un score Fβ pour chaque classe du niveau le plus bas est calculé pour chaque modèle de classification, ce qui permet la comparaison à un niveau inférieur. Il faut utiliser un seuil contextuel pour la statistique des prix calculée et les processus opérationnels pour éliminer les modèles dont les performances sont mauvaises dans un trop grand nombre de catégories. La sélection des modèles devrait être envisagée dans le contexte de l'effort requis aux fins d'assurance de la qualité après la classification, et de l'importance des catégories ayant un faible rendement, comme leur importance dans le panier final de l'IPC.

3.3 Étape 3 – Considérations relatives aux opérations et aux processus

Le cas échéant, selon les exigences des ONS, des considérations opérationnelles et de processus spécifiques doivent être prises en compte.

3.4 Tout mettre ensemble

Les étapes ci-dessus se conjuguent pour créer un cadre opérationnel que les ONS peuvent utiliser pour l'appliquer à la fois à l'évaluation initiale de la sélection d'un classificateur performant à des fins de production ainsi qu'à la surveillance des performances du classificateur une fois qu'il est déployé. Aux fins de l'évaluation initiale, les trois étapes doivent être réalisées. On effectue les deux premières en parallèle sur des données étiquetées de manière robuste, en sélectionnant des modèles qui donnent de bons résultats dans plusieurs catégories et du classement de modèles et en classant les modèles en résultant dans le but de choisir un modèle optimal. On peut tenir compte des considérations relatives aux opérations et aux processus, le cas échéant, pour évaluer si le modèle le mieux coté est idéal dans un processus de production, et si le déploiement du modèle arrivant en deuxième position est préférable. Une fois qu'un modèle est déployé à l'appui d'un processus de production, il faut en surveiller les performances, notamment pour évaluer quand le modèle doit être entraîné de nouveau (UK Statistics Authority, 2019, 9; Eurostat, 2017, 23; Martindale et coll., 2020, 19). Le cadre proposé, en particulier les étapes 1 et 2, peut être appliqué à un échantillon de nouveaux produits pour chaque mois ayant fait l'objet d'une assurance de la qualité ou d'une validation de l'exactitude des catégories prédites. Si l'ONS déploie un nouveau modèle pour remplacer un modèle existant, il faut soumettre le nouveau modèle à un test A/B ou le déployer en parallèle du modèle existant afin d'évaluer qu'il fonctionne aussi efficacement que prévu hors échantillon.

4. Test empirique

4.1 Données et prétraitement

Pour démontrer le rendement de la méthode, nous avons choisi un ensemble de données publiques mises à disposition par l'État de l'Iowa concernant les données sur les ventes d'alcool (État de l'Iowa, 2021). Les données contiennent des variables structurées et ressemblent aux données de lecteur optique souvent utilisées par les ONS. Plus précisément, les données sont des données sur les transactions électroniques par produit unique vendu à une date précise chez un titulaire de permis de vente d'alcool de catégorie « E ». De plus, l'ensemble de données contient les variables de vente et de volume nécessaires au calcul de l'indice de prix (« Bouteilles vendues », « Volume vendu (en litres) », « Vente (en dollars) »), les variables catégoriques (« Catégorie » et « Nom de catégorie ») qui peuvent servir à cartographier les produits de l'ensemble de données selon une taxonomie hiérarchique, et les variables de définition du produit (« Description de l'article », « Fournisseur ») qui peuvent servir pour les caractéristiques de prédiction d'un modèle de classification.

En effectuant en parallèle les méthodes courantes de mise en œuvre des données de lecteur optique (Eurostat, 2017, 24), nous cartographions les catégories d'ensembles de données selon 14 codes du niveau le plus bas dans une taxonomie hiérarchique personnalisée, en parallèle des codes de taxonomie les plus bas utilisés dans les statistiques de prix comme les codes de taxonomie des produits élémentaires. Pour ce qui est des catégories de spécialité dans l'ensemble de données, nous établissons manuellement la correspondance directe entre les produits individuels et nos codes de taxonomie, car ils sont interprétés comme des catégories fourre-tout pour les types hétérogènes de produits qui devraient appartenir à la catégorie de produits semblables. Après l'avoir cartographié, nous traitons le code

cartographié comme la véritable étiquette de chaque produit et poursuivons notre démonstration en ignorant la variable de catégorie de l'ensemble de données. On a créé une hiérarchie artificielle à quatre niveaux pour évaluer les mesures hiérarchiques, avec une seule racine divisée en trois types de boissons alcoolisées (alcools, cocktails/prêts-à-boire et spiritueux), les spiritueux étant divisés en huit types (brandy, gin, mezcal, rhum, tequila, vodka, whisky et autres spiritueux), avec vodka et whisky encore divisé respectivement en deux et quatre sous-types (vodka aromatisée et non aromatisée; et bourbon, whisky irlandais, scotch et autres whiskys).

4.2 Classification et réglage de précision

Nous utilisons deux années de l'ensemble de données pour le test empirique : les données de 2019 servent à l'entraînement initial, le réglage et l'évaluation du modèle, tandis que les données de 2020 servent à la démonstration du processus de surveillance mensuelle. Les données de 2019 contiennent 3 208 produits uniques; celles de 2020 contiennent 941 nouveaux produits, non observés en 2019.

Le pipeline de classification comprend deux étapes : prétraitement et classification. Aux fins du prétraitement, on a effectué un prétraitement par traitement du langage naturel (TLN) courant, comprenant la suppression de caractères spéciaux, la segmentation en unigrammes de mots, la suppression de mots vides et la vectorisation TF-IDF (fréquence du terme-fréquence inverse de document). La TD-IDF est courante, facile à mettre en œuvre, relativement peu coûteuse du point de vue du calcul, et elle est plus robuste qu'un vectoriseur de dénombrement simple type. Pour l'étape de la classification, on a mis à l'essai cinq modèles de classification classiques largement adoptés en ignorant la hiérarchie et en utilisant la méthode de classification plate (voir le tableau 1). Les données de 2019 ont été divisées en un ensemble d'entraînement (80 % des données de 2019) et de validation (20 %), au moyen d'un échantillonnage aléatoire stratifié. Pour chacun des pipelines de classification candidats, les hyperparamètres du modèle ont été sélectionnés au moyen d'une validation croisée en trois parties de l'ensemble de données d'entraînement. Chaque pipeline a ensuite été entraîné, au moyen des hyperparamètres sélectionnés, sur tout l'ensemble de données d'entraînement. Les mesures des critères d'évaluation ont ensuite été calculées à partir de l'ensemble de données de validation; les résultats sont résumés dans le tableau 1. On réajuste ensuite le modèle sélectionné final au moyen de toutes les données de 2019 (entraînement et validation) pour le déployer en production. Le modèle de production sert à prédire les nouveaux produits trimestriellement, pour 2020. Les performances du modèle pour les nouveaux produits sont évaluées chaque trimestre au moyen des mêmes critères d'évaluation; le score du modèle pour chaque trimestre de 2020 est présenté dans le tableau 3.

Un β de 1 pour tous les scores Fβ a été sélectionné, car des faux négatifs et des faux positifs ont été observés dans les modèles d'entraînement. Pour le score combiné, nous avons inclus tous les scores F1 applicables à des fins de démonstration et adopté une moyenne arithmétique non pondérée pour rendre le score final sensible aux valeurs aberrantes. Nous avons choisi un seuil de 0,8 en nous basant sur les constatations selon lesquelles les classificateurs aux performances élevées au-delà de 0,8 donnent des indices de prix supérieurs (Office for National Statistics 2020).

4.3. Résultats

L'ensemble de données de démonstration montre que la machine à vecteur de support (SVM pour l'anglais *Support Vector Machine*) et le réseau neuronal peu profond (RN) obtiennent généralement les meilleurs résultats pour toutes les mesures d'évaluation et ont obtenu les deux meilleurs scores de modèle agrégés (tableau 1). Le test de seuil a montré que dans les trois meilleurs modèles, le modèle SVM avait une catégorie de niveau inférieur (whiskys irlandais) avec un score F1 inférieur à 0,8, le modèle RN en avait deux, et AdaBoost en avait trois (tableau 2). En raison de ses performances élevées, la machine à vecteur de support (SVM) a été choisie comme modèle de sélection sur les données de 2020; les résultats trimestriels de toutes les mesures sont résumés dans le tableau 3.

Tableau 1
Statistiques d'entraînement et de test pour les données de 2019 fondées sur les divers scores Fβ décrits à la section 3.1.

Modèle	$F1_{sw}$	$F1_{mw}$	$F1_{bw}$	$hF1_{sw}$	$hF1_{mw}$	$hF1_{bw}$	Score du modèle
SVM	0,8795	0,857371	0,876771	0,916849	0,908075	0,918435	0,892834
RN peu profond	0,871776	0,86666	0,860297	0,919695	0,918763	0,916391	0,892264
Bayésien naïf	0,841622	0,832533	0,818937	0,901617	0,900732	0,891479	0,864487
AdaBoost	0,844433	0,837334	0,846729	0,88011	0,874676	0,886112	0,861566
Forêt aléatoire	0,839274	0,829134	0,843871	0,877714	0,871871	0,884811	0,857779

Tableau 2
Statistiques d'entraînement et de test pour les trois meilleurs modèles fondées sur les données de 2019 et divers scores Fβ décrits à la section 3.1.

Modèle	Agrégat élémentaire (PE)	Score Fβ
SVM	Whiskys irlandais	0,556
RN peu profond	Autres spiritueux	0,714
	Whiskys irlandais	0,762
AdaBoost	Autres spiritueux	0,632
	Whiskys irlandais	0,632
	Alcools	0,746

Tableau 3
Score du modèle appliqué par trimestre aux données de 2020

Trimestre	$F1_{sw}$	$F1_{mw}$	$F1_{bw}$	$hF1_{sw}$	$hF1_{mw}$	$hF1_{bw}$	$Model\ score$
2020T1	0,791704	0,761984	0,779637	0,861476	0,85468	0,862852	0,818722
2020T2	0,81019	0,788765	0,813265	0,854453	0,826787	0,859456	0,825486
2020T3	0,800554	0,700787	0,780164	0,886857	0,827672	0,877761	0,812299
2020T4	0,782575	0,784638	0,763883	0,869312	0,869071	0,849384	0,819811

5. Conclusion

Cette recherche a permis d'établir la définition et la démonstration de mesures clés permettant, d'une part, d'évaluer et sélectionner des classificateurs à des fins de production et, d'autre part, de surveiller leurs performances dans le calcul de statistiques sur les prix. Notre démarche a introduit l'ensemble des mesures qui peuvent être prises en considération dans l'évaluation des performances des modèles, mais il faudrait mener des recherches complémentaires pour déterminer comment sélectionner les mesures les plus appropriées à partir de cet ensemble de mesures. Plus précisément, ces autres recherches doivent premièrement modéliser explicitement l'effet d'une classification erronée sur un indice de prix, et envisager ce risque conjointement avec des méthodes permettant de marquer les enregistrements et d'assurer leur qualité avant le calcul de l'indice de prix de façon à réduire le plus possible tout biais de classification erronée dans l'indice de prix final. Deuxièmement, les futures recherches doivent évaluer empiriquement l'effet des diverses mesures proposées à la section 3.1 sur le suivi des performances en matière de classification erronée, ainsi que la pondération idéale de diverses mesures en vue d'obtenir le score total le plus optimal. Enfin, ces recherches doivent déterminer si le choix de la méthode d'indice de prix appliqué ainsi que la disponibilité des données de pondération ont une incidence sur les mesures et leur pondération dans le score d'évaluation final.

Remerciements

Les auteurs aimeraient remercier Junxiao Ma pour son appui à une version antérieure du présent article.

Bibliographie

Costa, E. P., A. C. Lorena, A.C.P.L.F. Carvalho et A. A. Freitas (2007), « A review of performance evaluation measures for hierarchical classifiers. » *Evaluation methods for machine learning II: Papers from the AAAI-2007 workshop.* p. 1-6

Eurostat (2017), « Practical Guide for Processing Supermarket Scanner Data. » Indice des prix à la consommation harmonisé.

Bureau international du travail (2020), *Manuel de l'indice des prix à la consommation : Théorie et pratique*. https://www.ilo.org/global/statistics-and-databases/WCMS_331155/lang--fr/index.htm

Martindale, H., E. Rowland, T. Flower et G. Clews (2020), « Semi-supervised machine learning with word embedding for classification in price statistics. » *Data & Polic* 2.

Office for National Statistics (2020), « Automated classification of web-scraped clothing data in consumer price statistics. » Office for National Statistics. https://www.ons.gov.uk/economy/inflationandpriceindices/articles/automatedclassificationofwebscrapedclothingdatainconsumerpricestatistics/2020-09-01.

Office for National Statistics (2021), « Classification of new data in UK consumer price statistics ». Office for NationalStatistiscs.https://www.ons.gov.uk/economy/inflationandpriceindices/articles/classificationofnewdatainukconsumerpricestatistics/2021-04-06.

Silla, C. N. et A. A. Freitas (2011), « A survey of hierarchical classification across different application domains. » *Data Mining and Knowledge Discovery 22.1* p 31-72.

État de l'Iowa (2021), « 2019 Iowa Liquor Sales. » *data.iowa.gov.* 10 01. https://data.iowa.gov/Sales-Distribution/2019-Iowa-Liquor-Sales/38x4-vs5h.

Sun, Aixin et Ee-Peng Lim (2001), « Hierarchical text classification and evaluation. » *Proceedings 2001 IEEE International Conference on Data Mining.* p. 521-528.

UK Statistics Authority (2019), « Guidelines for selecting metrics to evaluate classification in price statistics production pipelines. » Advisory Panel on Consumer Prices – Technical, UK Statistics Authority. https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2019/08/APCP-T1910-Classification-metrics-guidelines.pdf.