Problèmes soumis au 11e atelier de résolution de problèmes industriels de Montréal

Air Canada
Relation et décalages entre les recherches et les tendances de réservation

Les recherches de voyage en ligne peuvent être considérées comme une intention de voyager et sont souvent une première étape nécessaire à l'action même de réservation. Des pics dans le volume de recherches peuvent parfois indiquer une augmentation soudaine de la demande des passagers, tandis qu'une augmentation du taux de « recherche-pour-réserver » (look-to-book en anglais) peut révéler un manque de compétitivité ou des problèmes sur le site Web qui présentent une opportunité pour l'analyste en gestion de revenu et peuvent aider à améliorer l'expérience d'achat et de réservation.

En utilisant les données de réservation et de recherche, il serait possible d'identifier la relation existante entre les deux signaux et de construire un modèle de détection des valeurs aberrantes qui identifiera les changements dans ces tendances. Air Canada pourrait ainsi corriger ses prévisions ou modifier sa politique d'inventaire.

Banque nationale du Canada
Anonymisation de données

La sécurité et la confidentialité des données est aujourd'hui au cœur des préoccupations des organisations. L'enjeu est de trouver un compromis entre la valorisation des données et le risque d'utiliser des données personnelles pour des projets analytiques et d'intelligence artificielle.

La gouvernance des données partage la responsabilité avec d'autres équipes, notamment les experts en cybersécurité, de mettre en place des mesures nécessaires pour garder un contrôle granulaire sur les droits d'accès de chaque employé. Par exemple, elle définit des niveaux d'accès par rôle directement sur les données. Cependant, la mise en place de ces mesures et des processus qui en découlent restreint l'accès aux données, et ce, quelle que soit l'utilisation qu'on veut en faire. Cette approche atteint aussi ses limites lorsqu'on souhaite donner des accès partiels à des bases de données, car les risques d'inférence et de re-identification augmentent.

Des options existent et constituent des domaines de recherche très actifs. Elles peuvent être regroupées en trois catégories (non disjointes) :

(a) Techniques d'anonymisation de données (p. ex. : "k-anonymisation") visant à rendre moins sensibles les données d'origine;
(b) Analyses conçues pour conserver la confidentialité des données (p. ex.: "Differential Privacy"). On utilise les données d'origine et on désensibilise les calculs sur les données.
(c) Modèles génératifs de données synthétiques.

Ces approches introduisent le concept de mesure de confidentialité et donc une quantité associable au risque des données.

Compte tenu de leurs avantages et inconvénients inhérents, quelle(s) approche(s) faudrait-il favoriser pour optimiser le compromis entre le risque de confidentialité et la valorisation des données?

Coveo
Analyse des données ClickStream de commerce électronique pour prédiction d'intention et recommandations

Coveo partage un riche ensemble de données anonymisées sur le comportement d'utilisateurs sur un site de commerce électronique à des fins de recherche. L'ensemble de données comprend plus de 30 millions d'interactions de produits uniques par de vrais acheteurs – les données de parcours de navigation sont enrichies par des métadonnées de catalogue (vectorisées) et par un comportement de recherche précis, montrant non seulement les produits cliqués après l'émission d'une requête, mais également les produits vus et non-cliqués (c'est-à-dire la rétroaction négative).

Dans la continuité des recherches précédentes (de Coveo et de la communauté) et avec le SIGIR Data Challenge 2021, nous accueillons des recherches portant sur deux défis :

(a) Une tâche de recommandation basée sur une session, où un modèle est invité à prédire les prochaines interactions entre les acheteurs et les produits, sur la base des interactions précédentes des produits et des requêtes de recherche au sein d'une session;
(b) Une tâche d'abandon de panier, où, étant donné une session contenant un événement d'ajout au panier pour un produit X, un modèle est invité à prédire si l'acheteur achètera X ou non dans cette session.

Les participants sont encouragés à lire les articles évalués par les pairs suivants https://arxiv.org/abs/1907.00400 et https://rdcu.be/b8oqN, pour mieux comprendre les problèmes théoriques et pratiques sous-jacents : déséquilibre de classe, taux de conversion, etc. L'article SIGIR Data Challenge contient une analyse documentaire approfondie, des détails approfondis sur l'ensemble de données et des discussions sur les cas d'utilisation cibles ; le Leaderboard SIGIR peut être consulté pour des benchmarks quantitatifs (des fichiers tests pour répliquer la phase de soumission peuvent être demandés à Coveo) ; le référentiel open source SIGIR contient des explications détaillées sur l'ensemble de données, des modèles de base prêts à l'emploi et des liens vers des implémentations réussies.

Environnement et changement climatique Canada
Développement d'un générateur de textes météo

Les prévisions météorologiques comprennent de nombreuses façons d'exprimer les prévisions possibles. Cependant, les textes des prévisions sont très structurés et très limités dans leur formulation. De plus, il n'existe qu'une bonne façon standard de « dire » la météo considérant un set de concepts donnés.

Afin de continuer à fournir les services de prévision et d'information météorologiques de qualité aux Canadiens, le SMC souhaite développer un générateur de texte de prévisions météorologiques, en français et en anglais, qui utilise des concepts météorologiques représentant, sous forme codée, la prévision météo.

Nous avons deux objectifs.
(a) Bâtir un prototype qui démontrerait la faisabilité d'une approche de Traitement naturel du langage (NLP en anglais) pour développer un générateur de textes de prévisions météorologiques.
(b) Obtenir une estimation des points forts et des points d'amélioration d'une telle approche en comparaison avec le système existant.

IATA
Analyse corrélative du partage de code des compagnies aériennes

Le programme IOSA est une norme de l'industrie qui est utilisée par les compagnies aériennes pour coopérer. Un mode de coopération fréquent est le partage de codes. En utilisant les données de trafic, il serait souhaitable de découvrir et comprendre le marché IOSA et ses besoins à travers l'optique des relations de partage de code. L'objectif est de comprendre pour quelle destination, à quelle ampleur et à

quelle profondeur les compagnies aériennes font du partage de code entre elles, et quelle est la corrélation avec la certification IOSA.

Accessoirement, IATA aimerait dans la mesure du possible explorer toutes autres corrélations potentielles avec les extrants produits par une compagnie aérienne, par exemple, explorer la relation entre les activités de partage de codes et certaines métriques cruciales des compagnies aériennes.

Les travaux d'IATA montrent que les compagnies aériennes certifiées IOSA ont généralement plus de relations de partage de codes avec d'autres compagnies aériennes que les compagnies aériennes non certifiées. Toutefois, cela varie d'une compagnie à l'autre. En outre, les réseaux de transporteurs traditionnels sont pour la plupart certifiés IOSA. Par conséquent, IATA ne peut pas établir de corrélation nette entre la certification IOSA et le partage de codes des compagnies aériennes. L'association aimerait avoir des conclusions sous forme d'informations commerciales et comprendre la méthodologie utilisée.

Société générale
Sélection de variables catégorielles dans la modélisation des risques

Dans les institutions financières, les caractéristiques catégorielles apparaissent assez souvent dans les ensembles de données de crédit et dans les modèles de conformité, par exemple, les caractéristiques liées au profil de risque des clients.

Les méthodes traditionnelles de sélection d'entités (par exemple, signification statistique, élimination d'entités récursives, LASSO) ne fonctionnent pas bien avec les entités catégorielles, car ces méthodes conservent certains niveaux en supprimant d'autres de la même entité. L'approche LASSO groupé s'est montrée plus stable en termes de sélection de variables mais présente des lacunes en termes de prévisibilité. Pour une caractéristique donnée, serait-il plus approprié de concevoir une méthode qui agrège certains niveaux avoisinants afin d'obtenir un espace de représentation des caractéristiques qui varie mieux avec la variable à prédire?

En raison des nombreuses façons de représenter les variables catégorielles et de sélectionner les variables importantes, nous nous demandons quelles sont les méthodes les plus appropriées pour améliorer la sélection des caractéristiques catégorielles.

TD Canada
Prévision du risque de portefeuilles

En pratique, les gestionnaires de portefeuille utilisent soit le modèle fondamental, soit le modèle statistique, soit les deux à la fois, pour modéliser la volatilité du rendement futur d'un portefeuille.

Le modèle fondamental spécifie des caractéristiques qui influencent les covariances des rendements des actions. Par exemple, les actions d'une même industrie auront des rendements plus corrélés entre elles que celles issues d'industries différentes. Le levier financier ou la taille des entreprises semblent aussi expliquer une partie de la variation commune des rendements. Moyennant une liste de ces caractéristiques et leur évolution historique ainsi qu'un historique de rendements des actions, il est possible d'estimer la matrice de covariance des rendements des facteurs de risque pour calculer l'exposition d'un titre à ces facteurs. De plus, en les agrégeant nous sommes capables de calculer le risque de tout portefeuille.

Le modèle statistique part d'une pleine matrice de covariance. Celle-ci a souvent été calculée avec des rendements filtrés par un processus de "winsorisation". Une analyse de composantes principales (PCA) identifie les principales dimensions de la matrice de covariance en les classant par importance mais sans pour autant identifier les facteurs auxquels elles se réfèrent. Les facteurs de risque sont inférés empiriquement.

À ce jour, GPTD a construit des modèles de risque selon ces deux types de modélisation. Les modèles statistiques (qui utilisent moins d'information) fonctionnent bien, surtout quand les marchés boursiers sont moins volatiles mais moins lorsque la volatilité des marchés augmente fortement. Peu de consensus existe sur la façon optimale de combiner le modèle fondamental et le modèle statistique ou encore sur la possibilité d'en construire un troisième. GPTD cherche à savoir quelle serait la meilleure manière de prévoir le risque d'un portefeuille qui utilise les rendements historiques sur les actions d'un certain univers ainsi que les caractéristiques fondamentales de ces actions.