Analyse des causes des retards des trains marchandises de l'Office National des Chemins de Fer (ONCF) du Maroc

Sekou Kaba, Salah Eddine Khlail, Charles Bourbeau

Équipe 41

Introduction

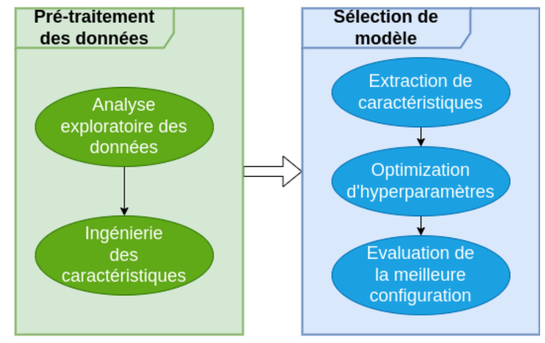
Pour toute organisation, privée ou public, oeuvrant dans le secteur du transport de marchandises, l'assurance de la qualité du transport ainsi que le respect des clauses contractuelles sont des facteurs de compétitivité de la plus haute importance. Au Maroc, l'ONCF est l'organisme public qui détient le monopole du transport ferroviaire des marchandises, et joue ainsi un rôle primordial dans le processus de développement industriel du pays. Conséquemment, les gestionnaires d'infrastructure ferroviaire se penchent assidûment sur la découverte d'avenues potentielles de diminution des causes de retard touchant les trains de marchandise, vu les coûts significatifs y étant associés. Dans ce contexte, l'ONCF vise en premier lieu à mener une étude des motifs de retards via l'agrégation d'un jeu de données décrivant les trajets victime de retard au courant de l'année 2021. La tâche cible vise le développement d'un système permettant l'inférence du motif de retard en fonction des caractéristiques représentant un trajet.

Dans ce travail, nous investiguons l'exploitation des avancées du domaine de l'apprentissage automatique afin de répondre au besoin criant de l'organisme partenaire. Notamment, nous démontrons qu'une approche qui apprend une transformation optimale de la nature qualitative des données en une représentation purement numérique maximise la performance de classification en comparaison avec des méthodes classiques tel que l'encodage one-hot.

Méthodes

La résolution du problème de classification proposé par l'ONCF a été accomplie selon une approche en deux temps dont la description haut-niveau est présentée à la Figure 1. D'abord, le pré-traitement des données a été réalisé afin d'acquérir une compréhension des variables représentant chaque trajet, et ainsi faciliter la sélection et l'ingénirie de caractéristiques présumablement liées la cible. Ensuite, nous avons appliqué une gamme de modèles appropriés au jeu de données afin d'en soutirer la solution optimale selon quatres métriques de performances en classification.

Figure 1 – Méthodologie de recherche



Les données utilisées dans ce projet ont été fournis par l'ONCF [1], et contiennent un total de 585 trajets de train. Chaque trajet est défini selon une combinaison de deux variables numériques et sept variables catégorielles, accompagné du motif de retard à classifier (Tableau 1).

Tableau 1 – Jeu de données ONCF

Nom de la variable	Échelle de mesure
Motif de retard	Nominale
Délai de départ	Continue (min.)
Délai d'arrivée	Continue (min.)
Identifiant de train	Nominale
Nature du train	Nominale
Machine	Nominale
Type de machine	Nominale
Chargé	Booléenne
Ligne	Nominale
Parcours	Nominale

Figure 2 – Distribution des classes

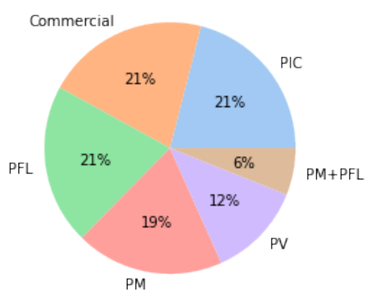
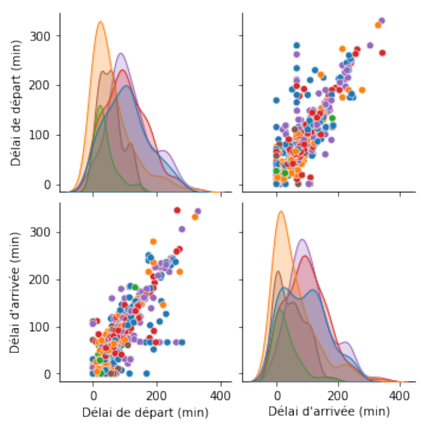
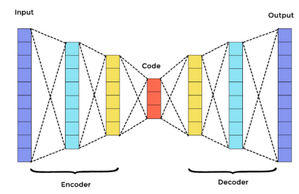


Figure 3 – Pairplot des variables continues



Puisque l'ensemble de caractéristiques comporte une majorité de variables catégorielles, une approche d'apprentissage de représentation basé auto-encodeur (Figure 4) a été développée afin de produire une représentation latente des trajets et permettre l'application valide d'une notion de distance entre ceux-ci.

Figure 4 – Architecture auto-encoder [2]



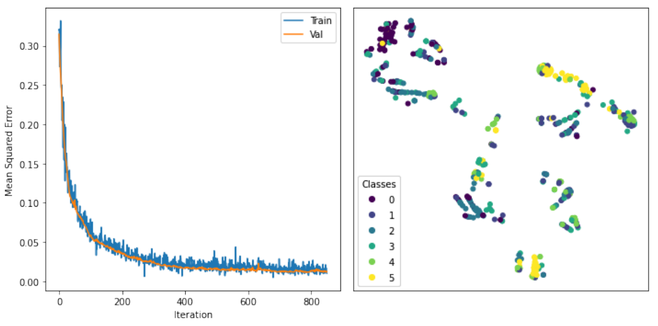
Résultats

Dans cette section, nous évaluons les différents modèles d'apprentissage automatique considérés en contrastant la méthode d'extraction de caractéristiques utilisée, soit identitaire ou basé auto-encodeur. Pour chaque modèle, nous performons l'optimization d'hyperparamètres selon une recherche en grille utilisant une validation à 10 plis. En plus des hyperparamètres du classifieur, on y inclut ceux spécifiant l'architecture de l'auto-encodeur. La performance de la meilleure configuration est rapportée sur un ensemble de test réservé totalisant 20% des données initiales.

Tableau 2 – Performance des différentes méthodes

Modèle	Métrique (%)			
	Exactitude	Précision	Rappel	Score F1
RFC	50.0	45.2	50.3	45.1
KNN	52.6	42.9	48.3	43.9
SVM	51.2	43.0	47.3	43.4
AE + RFC	55.1	56.7	54.3	55.47
AE + KNN	62.3	59.6	61.1	60.0
AE + SVM	57.3	55.6	53.2	54.4

Figure 5 – Entraînement d'un auto-encoder 16x16x12 (gauche) et visualisation T-SNE de l'espace (droite)



Discussion

L'expérimentation présentée dans la section précédente met en lumière le potentiel d'amélioration de performance provoqué par l'apprentissage d'une représentation complexe basé sur un réseau profond, dont le fonctionnement est impossible à répliquer par la conception de caractéristiques manuelles. L'augmentation de performance selon les quatres métriques présentées démontre qu'un apprentissage de repsésentation préliminaire facilite l'apprentissage subséquent sur la tâche cible de classification.

Conclusion

Comme la préservation de l'aspect concurrentiel du transport ferroviaire Marocain dépend de sa capacité à respecter ses engagements envers les entreprises marchandes, il est vital pour l'ONCF de comprendre et ainsi prévenir les diverses sources de retard. Ce travail établit l'aspect clé comme étant l'apprentissage profond d'une représentation qui permet aux stratégies d'IA d'opérer dans un espace numérique exploitant l'information essentielle des données. Cependant, la portée de ce projet se limite à l'évaluation de différentes architectures basées auto-encodeur. De ce fait, l'investigation d'approches similaires, telle que les machines Boltzmann restreintes, reste nécessaire afin de déterminer laquelle résulte en une représentation optimale pour cette tâche.

Références / References

[1] *Jeu de données de l'Office National des Chemins de Fer*. Portail National des données ouvertes du Royaume du Maroc. (2022). Retrieved December 18, 2022, from https ://data.gov.ma/fr

[2] *Autoencoders : Main components and architecture of autoencoder*. EDUCBA. (2022, May 3). Retrieved December 18, 2022, from https ://www.educba.com/autoencoders/