Analyse et prédiction des annulations dans les réservations hôtelières

WAKRIM YOUSRA

Youssra.wakrim@etu.uae.ac.ma

Ecole National Des Sciences Appliquées D'Al Hoceima

Résumer :

Cet article de données décrit deux ensembles de données contenant des informations sur la demande hôtelière. L'un des hôtels (H1) est un hôtel de villégiature, tandis que l'autre (H2) est un hôtel en ville. Les deux ensembles de données partagent la même structure, avec 31 variables décrivant 40 060 observations pour H1 et 79 330 observations pour H2. Chaque observation représente une réservation d'hôtel. Les deux ensembles de données comprennent des réservations prévues entre le 1er juillet 2015 et le 31 août 2017, incluant à la fois les réservations effectivement arrivées et celles annulées. Étant donné qu'il s'agit de données réelles d'hôtels, tous les éléments relatifs à l'identification des hôtels ou des clients ont été supprimés. En raison de la rareté des données commerciales réelles disponibles à des fins scientifiques et éducatives, ces ensembles de données peuvent jouer un rôle important dans la recherche et l'enseignement en gestion des revenus, apprentissage automatique ou fouille de données, ainsi que dans d'autres domaines.

Tableau des spécifications :

- **Domaine** : Gestion hôtelière
- **Sous-domaine** : Gestion des revenus
- **Type de données** : Fichiers texte et objets R
- **Acquisition** : Extraction via requêtes TSQL depuis les bases SQL des systèmes PMS
- **Format** : Mixte (brutes et prétraitées)
- **Facteurs expérimentaux** : Création de variables à partir de différentes tables, point temporel défini la veille de l'arrivée.
- **Caractéristiques** : Extraction via TSQL et analyse avec R.
  Lieu : H1 (Algarve, Portugal) et H2 (Lisbonne, Portugal).

Valeur des données :

- L'analyse descriptive peut être utilisée pour mieux comprendre les tendances, les modèles et les anomalies des données.
- Les données permettent de mener des recherches sur divers problèmes tels que la prédiction des annulations de réservations, la segmentation des clients, la saturation de la clientèle et la saisonnalité.
- Les chercheurs peuvent utiliser ces ensembles de données pour comparer des modèles de prédiction d'annulations de réservations avec des résultats déjà connus (par exemple [1]).
- Les chercheurs en apprentissage automatique peuvent évaluer les performances de différents algorithmes pour résoudre des problèmes similaires (classification, segmentation ou autres).
- Les éducateurs peuvent utiliser ces données pour enseigner des problèmes de classification ou de segmentation en apprentissage automatique.
- Les éducateurs peuvent également les utiliser pour des formations en statistiques ou en fouille de données.

Dataset :

Dans les industries du tourisme et des voyages, les recherches sur la prévision de la demande et la gestion des revenus utilisent principalement des données de l'aviation (format PNR). Cependant, des secteurs comme l'hôtellerie, les croisières, et les parcs à thème ont des besoins spécifiques nécessitant des données propres. Pour combler cette lacune, deux ensembles de données hôtelières ont été partagés pour développer des modèles de prédiction d'annulation de réservations. Afin d'éviter toute fuite d'informations futures, les données ont été extraites du journal des modifications des réservations, avec un horodatage relatif à la veille de la date d'arrivée, garantissant ainsi que les informations futures ne soient pas utilisées prématurément.

I. INTRODUCTION :

Dans l'industrie hôtelière, la prévision des réservations est essentielle pour optimiser les opérations et améliorer l'expérience client. La capacité à anticiper la demande en fonction des saisons et des types d'hôtels est cruciale pour la gestion des ressources, des revenus et des services personnalisés.

Ce travail présente une étude visant à développer un système de prédiction des réservations d'hôtel en utilisant des techniques de nettoyage des données, d'analyse des relations entre attributs, d'ingénierie des fonctionnalités, d'analyse spatiale et l'application de six algorithmes d'apprentissage automatique : **régression logistique**, **KNN**, **forêt aléatoire**, **arbre de décision**, **CatBoost, XGBoost** et **LightGBM.**

L'étude commence par un nettoyage des données pour éliminer les incohérences et améliorer leur qualité. Ensuite, une analyse des relations entre les variables permet d'identifier les corrélations influençant les comportements de réservation. Une ingénierie des fonctionnalités est réalisée pour créer de nouvelles variables et améliorer les prédictions, complétée par une analyse spatiale pour évaluer l'impact géographique sur la demande.

Bien que de nombreuses études se concentrent sur l'aviation, peu ont appliqué ces techniques au secteur hôtelier. Cette recherche se distingue par l'utilisation de six algorithmes avancés pour améliorer la précision des prédictions. Les résultats obtenus fournissent des insights précieux pour les gestionnaires d'hôtels, les acteurs du secteur du voyage et les chercheurs, permettant d'optimiser les stratégies de réservation et d'améliorer la satisfaction client.

II. CONTEXTE :

Cette étude s'appuie sur des travaux antérieurs en apprentissage automatique et prédiction des réservations hôtelières, abordant des défis comme les annulations et la surréservation. En intégrant des algorithmes divers tels que la régression logistique et les forêts aléatoires, nous visons à améliorer la précision des prévisions de la demande hôtelière. Nous mettons l'accent sur le nettoyage des données, l'analyse des fonctionnalités et l'analyse géographique pour optimiser les performances des modèles. L'objectif est de proposer un système fiable qui facilite la prise de décision et améliore la satisfaction client dans l'industrie hôtelière.

III. Méthodologie :

Le jeu de données "hotel_bookings.csv" de ce projet contenant **119 390 échantillons** et **32 colonnes.** Il inclut des informations essentielles telles que le type d'hôtel, le statut de la réservation (Canceled or no), le lead time, la date d'arrivée, les données démographiques des clients, le segment de marché, et les comportements précédents des invités. Le jeu de données comprend aussi des informations financières et logistiques, comme le tarif moyen quotidien (ADR) et les demandes spéciales.

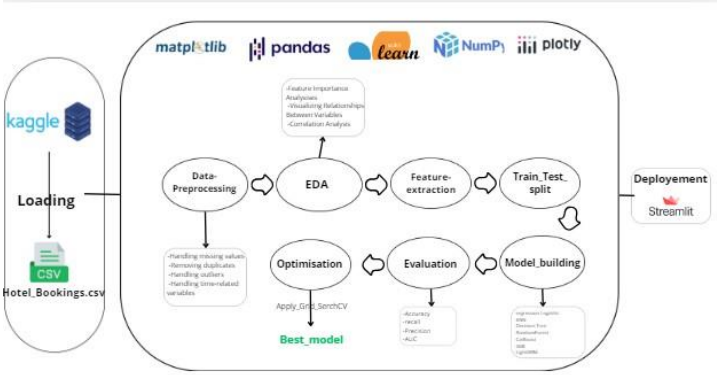


Fig. 1. La méthodologie de projet

1) *Collecte des données :*

Les données ont été collectées depuis Kaggle [1].

2) *Analyse Exploratoire des Données (EDA) :*

- **Univariate Analysis** : Analyse des distributions (**lead time**, **ADR**, types d'hôtels).
- **Bivariate Analysis** : Étude des relations entre **is_canceled** et des variables comme le segment de marché.
- **Multivariate Analysis** : Exploration des interactions via des heatmaps et matrices de corrélation.
- **Geographical Distribution** : Visualisation des origines géographiques des visiteurs.

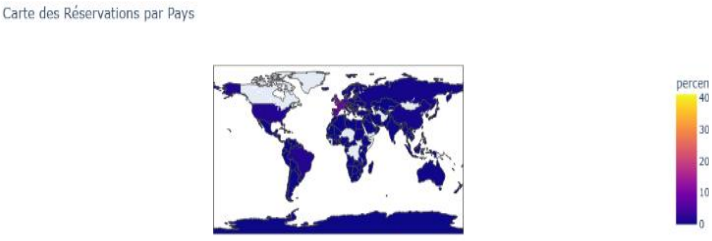


Fig. 2. Carte des réservations par Pays

3) *Nettoyage et prétraitement des données :*

Tout d'abord, les valeurs manquantes ont été identifiées et traitées soit par suppression, soit par imputation appropriée. Ensuite, les doublons ont été supprimés afin d'assurer l'unicité des observations. Les anomalies et valeurs aberrantes, notamment dans les variables numériques comme **`adr`**, ont été détectées grâce aux visualisations en **boxplot** et corrigées pour éviter les biais. Enfin, les types de données ont été standardisés et les variables catégorielles nettoyées pour garantir une analyse et un entraînement des modèles de machine learning optimaux.

4) *Division des Données :*

Pour la division des données, On a séparé l'ensemble en données d'**entraînement** et de t**est** en utilisant un ratio classique de **80/20**, afin de garantir une évaluation indépendante des modèles. Nous n'avons pas appliqué la technique **SMOTE** (Synthetic Minority Oversampling Technique), car bien que les classes soient légèrement déséquilibrées, cet écart n'était pas suffisamment extrême pour nécessiter un suréchantillonnage artificiel. L'équilibre relatif des classes **Not Canceled** et **Canceled** a permis de préserver la distribution naturelle des données et d'évaluer les performances des modèles sans introduire de biais potentiels liés à la génération de données synthétiques. Cette approche vise à refléter une situation réelle et pratique dans le domaine hôtelier.

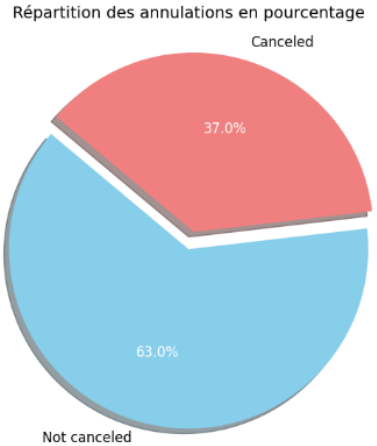


Fig. 3. Répartition des annulations en pourcentage.

5) *Modèles Utilisés :*

Dans ce projet, un total de **sept algorithmes de machine learning** incluant des modèles de classification variés tels que **Logistic Regression**, **K-Nearest Neighbors (KNN)**, **Decision Tree**, **Random Forest**, ainsi que des modèles de boosting avancés comme **XGBoost**, **CatBoost** et **LightGBM**, afin de comparer leurs performances sur la prédiction des annulations de réservations.

Regression Logistic :

La régression logistique est une technique d'apprentissage supervisé utilisée pour prédire la probabilité qu'une réservation soit annulée ou non dans le contexte des hôtels. Elle repose sur la fonction sigmoïde :

$$h_\theta(X) = \frac{1}{1 + e^{-\theta X}}$$

Où hθ(X) donne la probabilité estimée que l'observation X. La fonction de coût est définie par :

$$J(\theta) = -\frac{1}{m}\sum^{m}[\, y^{(i)} \log\left(h_\theta(X^{(i)})\right) + (1 - y^{(i)}) \log\,(1 - h_\theta(X^{(i}$$

La descente de gradient est utilisée pour ajuster les paramètres θ, minimisant ainsi l'erreur et optimisant la prédiction des annulations.

KNN :

(K-Nearest Neighbors) est une méthode d'apprentissage supervisé utilisée pour prédire si une réservation sera annulée en fonction de ses voisins les plus proches dans l'espace des caractéristiques. L'algorithme fonctionne en trouvant les **K voisins les plus proches** d'une observation, puis en déterminant la classe (annulé ou non) en fonction de la majorité des voisins.

Le principe est basé sur la distance (souvent Euclidienne) entre les points de données :

$$d\,(a, b) = \sqrt{(\sum(\mathrm{ai} - \mathrm{bi})^2)},$$

où *a* et *b* sont deux points dans l'espace des caractéristiques. KNN ne nécessite pas de phase d'entraînement explicite, mais la valeur de **K** et la métrique de distance doivent être choisies pour obtenir les meilleures prédictions. Une fois les voisins identifiés, la classe de l'observation est prédite en fonction de la classe la plus fréquente parmi les voisins.

- Decision Tree Classifier :

Le **Decision Tree Classifier** est un algorithme supervisé qui divise les données en segments basés sur des critères décisionnels. Dans le cadre de la prédiction des annulations de réservations d'hôtel, l'arbre de décision identifie les attributs clés (par exemple, le délai de réservation ou le type de client) pour classer les observations. Chaque nœud représente une question ou une règle, et les branches mènent à des résultats possibles. Le modèle utilise des mesures comme l'entropie ou l'indice de Gini pour optimiser les divisions, visant à minimiser l'impureté à chaque étape. Sa simplicité et son interprétabilité en font un outil puissant pour la prédiction et l'analyse des décisions.

- Random Forest classifier :

Le Random Forest Classifier combine plusieurs arbres de décision pour améliorer la précision et réduire le surapprentissage. Chaque arbre est entraîné sur un sous-échantillon des données, et la prédiction finale est obtenue par vote majoritaire des arbres. L'algorithme utilise des techniques comme le bootstrap et la sélection aléatoire d'attributs pour diversifier les arbres, ce qui renforce sa robustesse. La réduction de l'impureté dans chaque arbre est généralement mesurée via l'indice de Gini, calculé par :

$$G = 1 - \sum_{i-1}^{n} p_i^2$$

Où pi est la proportion d'observations dans la classe (i).

Ce modèle est particulièrement adapté à des données complexes et améliore la précision des prédictions pour les annulations de réservation d'hôtel.

- CatBoost Classifier :

Le **CatBoost Classifier** est un modèle de gradient boosting conçu pour gérer efficacement les variables catégoriques sans nécessiter d'encodage préalable. Il utilise un calcul basé sur des permutations pour réduire le surapprentissage et améliore la précision tout en maintenant une rapidité d'exécution. Ce modèle est particulièrement performant pour les données avec de nombreuses caractéristiques catégoriques.

- XGBoost :

Le **XGBoost Classifier** est une implémentation optimisée du gradient boosting qui utilise une régularisation L1L_1L1 et L2L_2L2 pour contrôler le surapprentissage. Sa fonction de coût est ajustée à chaque itération pour minimiser l'erreur :

$$\text{Obj} = \sum_{i-1}^{n} l(y_i, \hat{y}_i) + \Omega(f)$$

Où Ω(f) contrôle la complexité du modèle. Il excelle dans la gestion des données déséquilibrées, comme les annulations d'hôtel.

- LightGBM :

Le **LightGBM Classifier** est un modèle de gradient boosting optimisé pour la vitesse et la mémoire. Il utilise une stratégie de croissance en feuilles (leaf-wise) et la réduction des erreurs de variance pour construire des arbres plus profonds et précis. Sa capacité à gérer de grands ensembles de données en fait un choix efficace pour les prédictions complexes.

IV. Résultats et discussion :

1)*Les métriques de performances :*

Pour évaluer la performance d'un modèle Machine Learning, diverses métriques peuvent être utilisées dans le cadre de l'analyse du modèle. Pour ce projet d'analyse des réservations, nous avons utilisé des métriques telles que :

- Exactitude (Accuracy)
- Précision (Precision)
- Rappel (Recall)
- AUC

Avant de définir chaque métrique de performance, examinons les termes suivants que nous utilisons lors de la définition de la performance d'un modèle.

- **True Positive (TP)** - Le résultat lorsque le modèle prédit correctement une classe positive.
- **True Negative (TN)** -Le résultat lorsque le modèle prédit correctement une classe négative.
- **False Positive (FP)** - Le résultat lorsque le modèle prédit à tort une classe positive comme une classe négative.
- **False Negative (FN)** - Le résultat lorsque le modèle prédit à tort une classe négative comme une classe positive.

4.1) Accuracy :

L'exactitude d'un modèle de classification peut être définie comme le rapport entre le nombre total de prédictions Correctes et le nombre total de prédictions. L'équation pour l'exactitude peut être formulée comme suit :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

4.2) Precision :

La précision est définie comme le rapport entre le nombre d'étiquettes positives correctement classifiées et le nombre total d'étiquettes positives. L'équation de la précision peut être formulée comme suit :

$$Precision = \frac{TP}{TP + FP}$$

4.3) Recall :

Le rappel est le rapport des vrais positifs aux vrais positifs et aux faux négatifs. Cela revient à identifier le nombre de positifs correctement étiquetés. L'équation pour le rappel est la suivante :

$$Recall = \frac{TP}{TP + FN}$$

4.4) F1 Score :

Une autre métrique de performance utilisée est le score F1, cette métrique est utilisée pour résumer la précision et le rappel afin de fournir de meilleurs résultats. Le score F1 peut être défini comme la moyenne harmonique de la précision et du rappel, se situant entre 0 et 1. L'équation pour le score F1 est la suivante :

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.4) AUC :

L'AUC (Area Under the Curve) mesure la capacité d'un modèle à distinguer entre les classes en évaluant l'aire sous la courbe ROC, définie par :

$$\mathrm{AUC} = \int_0^1 TPR(FPR^{-1}(x))\,dx$$

Où TVP (taux de vrais positifs) et TFP (taux de faux positifs) quantifient la performance du modèle à différents seuils.

De plus, l'évaluation de la performance du modèle d'analyse des réservations d'hôtel est enrichie en tenant compte d'une matrice de confusion. Une matrice de confusion offre une ventilation détaillée des prédictions du modèle, indiquant le nombre d'instances de vrais positifs (TP), de vrais négatifs (TN), de faux positifs (FP) et de faux négatifs (FN). Cette matrice est particulièrement précieuse pour comprendre la capacité du modèle à classifier correctement les annulations positifs et négatifs.

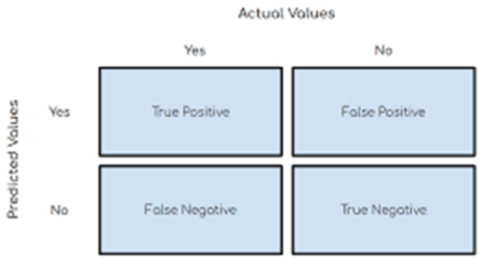


Fig. 4. Matrice de confusion

2) Comparaison algorithmique et évaluation des performances :

Les résultats montrent que les algorithmes d'apprentissage automatique utilisés dans ce projet, notamment la Régression Logistique, le K-Nearest Neighbors (KNN), le Decision Tree Classifier, le Random Forest, CatBoost, LightGBM et XGBoost, ont tous démontré une performance élevée en termes de précision lorsqu'ils ont été appliqués au dataset des réservations hôtelières.

Classification Report (Test):

	precision	recall	f1-score	support
0	0.79	0.92	0.85	14921
1	0.81	0.59	0.68	8876
accuracy			0.79	23797
macro avg	0.80	0.75	0.76	23797
weighted avg	0.80	0.79	0.79	23797

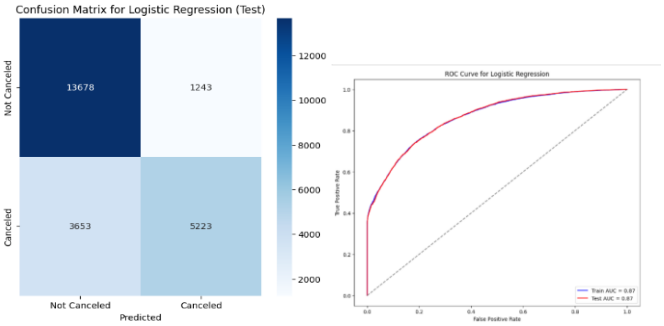


Fig.5. Rapport de classification et matrice de confusion pour la Regression Logistique

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.89	0.87	14921
1	0.80	0.73	0.76	8876
accuracy			0.83	23797
macro avg	0.83	0.81	0.82	23797
weighted avg	0.83	0.83	0.83	23797

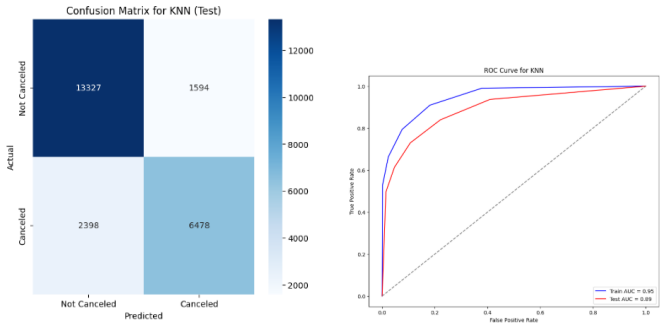


Fig. 6. Rapport de classification et matrice de confusion pour KNN.

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.82	0.84	14921
1	0.72	0.76	0.74	8876
accuracy			0.80	23797
macro avg	0.79	0.79	0.79	23797
weighted avg	0.80	0.80	0.80	23797

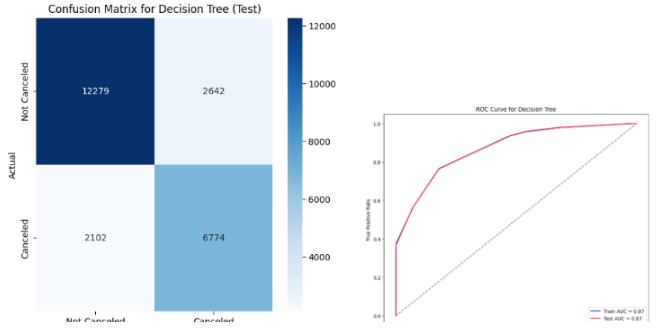


Fig. 7. Rapport de classification et matrice de confusion pour Décision Tree

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.89	0.87	14921
1	0.79	0.74	0.77	8876
accuracy			0.83	23797
macro avg	0.82	0.81	0.82	23797
weighted avg	0.83	0.83	0.83	23797

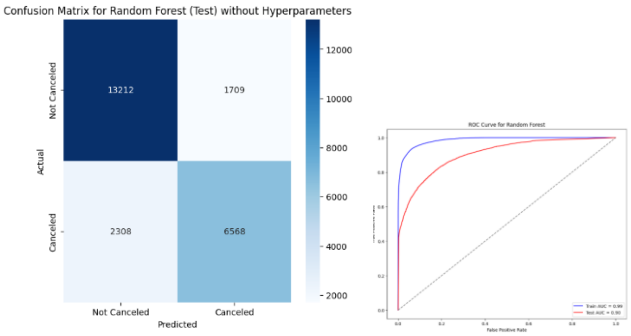


Fig. 8. Rapport de classification et matrice de confusion pour Random Forest Classifier

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.91	0.88	14921
1	0.82	0.73	0.78	8876
accuracy			0.84	23797
macro avg	0.84	0.82	0.83	23797
weighted avg	0.84	0.84	0.84	23797

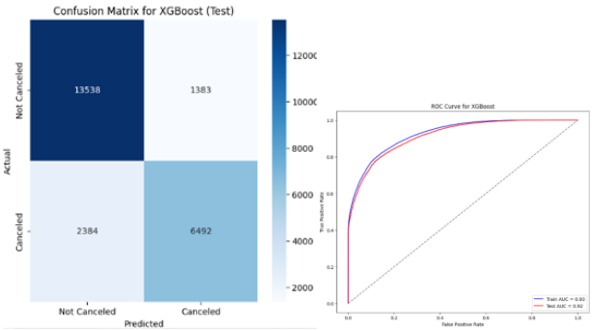


Fig. 10. Rapport de classification et matrice de confusion pour XGBoost.

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.91	0.88	14921
1	0.83	0.73	0.78	8876
accuracy			0.84	23797
macro avg	0.84	0.82	0.83	23797
weighted avg	0.84	0.84	0.84	23797

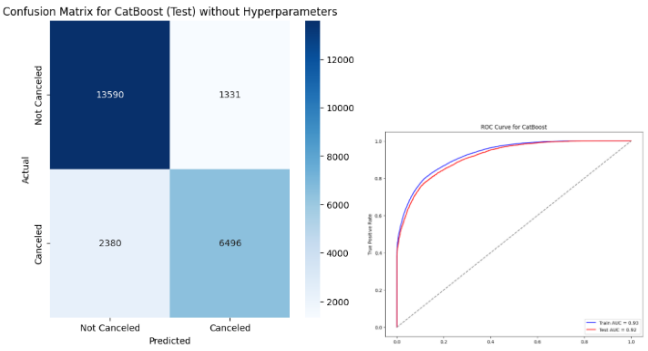


Fig.11. Rapport de classification et matrice de confusion pour CatBoost.

Classification Report (Test):

	precision	recall	f1-score	support
0	0.85	0.90	0.87	14921
1	0.81	0.73	0.77	8876
accuracy			0.83	23797
macro avg	0.83	0.81	0.82	23797
weighted avg	0.83	0.83	0.83	23797

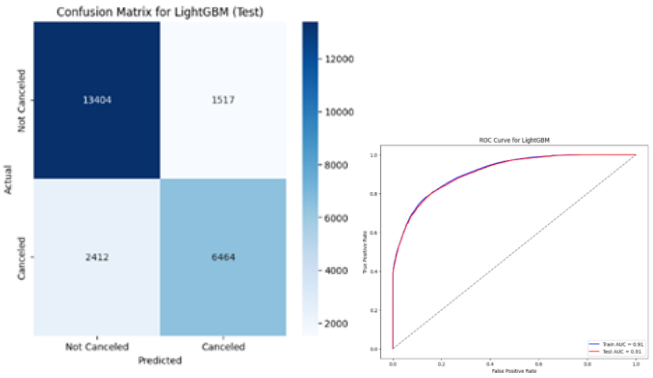


Fig. 12. Rapport de classification et matrice de confusion pour LightGBM

Les résultats des différents modèles appliqués sur le dataset des réservations hôtelières montrent des performances variables selon les métriques d'évaluation telles que l'Accuracy, la Précision, le Recall, le F1-Score et l'AUC. Avant l'optimisation des hyperparamètres, **le modèle Random Forest** se distingue avec une **Accuracy de 93.98% sur les données d'entraînement** et une **Test Accuracy de 83.11%**, accompagné d'une **AUC de 0.9681 sur les données de test**, ce qui en fait l'un des modèles les plus performants pour cette tâche. Toutefois, les modèles **CatBoost** et **XGBoost** montrent également des résultats prometteurs avec des **AUC test respectifs de 0.9177 et 0.9166**, soulignant leur capacité à bien distinguer les classes positives et négatives. En revanche, bien que les performances du **Decision Tree** et du **KNN** soient respectables, ils affichent un certain écart entre les métriques d'entraînement et de test, suggérant un risque de surapprentissage. La **Régression Logistique** offre des résultats plus équilibrés avec une bonne généralisation, atteignant une **Test Accuracy de 79.42%**. Enfin, le **LightGBM** affiche une **AUC de 0.9100**, confirmant son efficacité parmi les modèles testés

Le tableau 1 présente tous les résultats de chaque modèle.

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1-Score	Test F1-Score	Train AUC	Test AUC
0	Logistic Regression	0.795899	0.794260	0.797611	0.796138	0.795899	0.794260	0.787373	0.785787	0.867041	0.867723
1	KNN	0.875940	0.832248	0.875259	0.830728	0.875940	0.832248	0.874827	0.830469	0.947228	0.890793
2	Decision Tree	0.800500	0.800647	0.804241	0.803697	0.800500	0.800647	0.801801	0.801753	0.871777	0.870842
3	Random Forest	0.938984	0.831197	0.938904	0.829743	0.938984	0.831197	0.938706	0.829911	0.986189	0.904636
4	XGBoost	0.851757	0.841703	0.850777	0.840614	0.851757	0.841703	0.849872	0.839542	0.926433	0.916695
5	CatBoost	0.853690	0.844056	0.852872	0.843129	0.853690	0.844056	0.851687	0.841807	0.928214	0.917730
6	LightGBM	0.838362	0.834895	0.836918	0.833482	0.838362	0.834895	0.836595	0.832916	0.912940	0.910066

Tableau1. Evaluation les models avec les metrics.

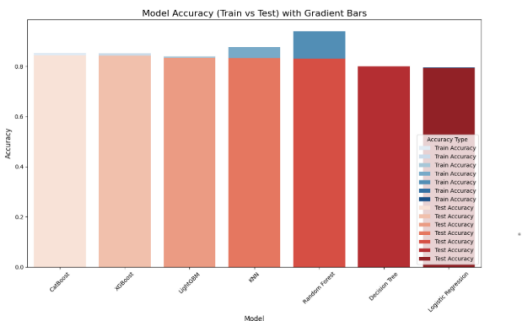


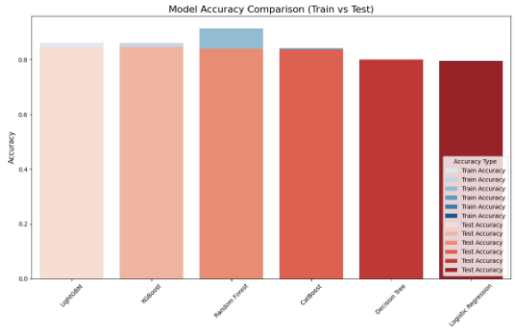
Fig. 13. Performance des models.

3) Amélioration des modèles par ajustement D'hyperparamètres :

L'ajustement des hyperparamètres est une étape essentielle dans l'optimisation des modèles d'apprentissage automatique. Il consiste à ajuster les paramètres externes des algorithmes afin d'améliorer leurs performances sur un jeu de données spécifique. Dans notre projet, nous avons appliqué cette méthodologie d'ajustement des hyperparamètres aux modèles. L'objectif principal était d'étudier l'impact de ces ajustements sur la performance de chaque modèle. Pour ce faire, nous avons utilisé la technique de validation croisée avec **GridSearchCV**, qui permet d'explorer différentes combinaisons de paramètres et de sélectionner les meilleurs pour chaque modèle. Les hyperparamètres clés ajustés pour chaque modèle incluent la profondeur maximale **(max_depth),** le nombre d'estimateurs **(n_estimators),** et les taux d'apprentissage **(learning_rate)** pour les modèles basés sur l'arbres de décision et le boosting. Cette optimisation a montré des améliorations notables des performances des modèles, comme en témoignent les résultats obtenus après validation croisée et ajustement des hyperparamètres.

	Model	CV Accuracy	CV Precision	CV Recall
3	XGBoost	0.845695	0.826759	0.737783
5	LightGBM	0.845579	0.826114	0.738265
2	Random Forest	0.839927	0.814633	0.734832
4	CatBoost	0.837522	0.816088	0.724360
1	Decision Tree	0.800164	0.714646	0.766076
0	Logistic Regression	0.795836	0.807490	0.588938

Tableau.2. Validation Croisée.



Model	L'exactitude (Avant L'ajustement des Hyperparamètres) (%)	L'exactitude (Après L'ajustement des Hyperparamètres) (%)
Random Forest	83%	84.02%
Catboost	84.40%	83.68%
XGBoost	84.17%	84.62%
LightGBM	83.48%	84.66%

Tableau.3. Comparaison des models avant et après L'hyperparamètre

V. Perspectives:

Cette étude a démontré l'efficacité des algorithmes d'apprentissage automatique pour prédire les réservations d'hôtels. Pour aller plus loin, des pistes d'amélioration incluent l'utilisation de méthodes d'ensemble (comme le Voting Classifier) pour combiner les forces des modèles, l'analyse temporelle pour capturer les tendances saisonnières, et la sélection des variables clés pour optimiser les performances. L'intégration de données externes (météo, événements locaux) et l'apprentissage en ligne permettraient également d'adapter les modèles en temps réel, offrant des prédictions plus précises et une meilleure gestion des réservations.

VI. Conclusion :

Cette étude a exploré l'efficacité des algorithmes d'apprentissage automatique dans la prédiction des réservations d'hôtels, en mettant en œuvre des modèles performants tels que les arbres de décision, les forêts aléatoires, CatBoost, XGBoost et LightGBM. L'ajustement des hyperparamètres via GridSearchCV a permis d'optimiser les performances de ces modèles, mettant en évidence l'importance d'une configuration précise pour des résultats fiables.

Les résultats obtenus soulignent que les modèles basés sur des techniques de boosting, notamment XGBoost et LightGBM, surpassent les autres en termes de précision et de robustesse, tout en s'adaptant efficacement à des données complexes. Cependant, cette recherche a également révélé des défis, notamment liés à l'analyse temporelle et à l'impact potentiel de facteurs externes, qui pourraient améliorer davantage les prédictions.

Dans une perspective plus large, cette étude met en lumière l'importance de combiner des approches algorithmiques avancées, des techniques d'optimisation et l'intégration de données contextuelles pour répondre aux besoins spécifiques du secteur hôtelier. Ces travaux offrent une base solide pour des recherches futures visant à affiner les modèles prédictifs, en intégrant des données en temps réel, des analyses temporelles et des méthodes de segmentation client. Par conséquent, cette étude contribue à démontrer comment l'apprentissage automatique peut transformer la gestion des réservations d'hôtels en un processus plus intelligent, précis et orienté vers les données.

VII. Références :

[1] N. Antonio, A. Almeida, L. Nunes, Predicting hotel bookings cancellation with a machine learning classification model, in: Proceedings of the 16th IEEE International Conference Machine Learning Application, IEEE, Cancun, Mexicopp. 1049–1054.doi:10.1109/ICMLA.2017.00-11, 2017.

[2] https://www.kaggle.com/code/swetarajsinha/hotel-bookings

[3] P. H. Saputro and H. Nanang, 'Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets. Journal of Applied Data Sciences,2(2021)40-46