Rapport Scientifique

Analyse de l'Intention d'Achat en Ligne

(Online Shoppers Purchasing Intention Dataset)

BOUSBOULA MOUAD

3 décembre 2025

Résumé

Ce rapport étudie la capacité à prédire, à partir du comportement de navigation en ligne, si un visiteur effectuera un achat. Le dataset utilisé, composé de plus de 12 000 sessions, contient des variables décrivant les interactions des utilisateurs avec un site e-commerce. Une analyse exploratoire, la mise en place de modèles de classification (régression logistique, Random Forest, etc.) ainsi qu'une évaluation à l'aide de métriques pertinentes (Accuracy, F1-Score, ROC-AUC) permettent d'examiner les facteurs qui influencent l'intention d'achat et d'évaluer la performance des modèles prédictifs.

1 Introduction

Avec l'essor du commerce en ligne, comprendre les comportements de navigation des visiteurs est essentiel pour optimiser les stratégies marketing. Le dataset *Online Shoppers Purchasing Intention* fournit des informations détaillées sur les sessions utilisateurs : durée de consultation des pages, catégories de pages visitées, taux de rebond, valeurs de pages, ainsi que des variables techniques et temporelles.

La problématique principale de ce projet est la suivante :

> **Peut-on prédire si une session utilisateur aboutira à un achat en se basant sur les comportements de navigation et les caractéristiques de la session ?**

Les objectifs de cette étude sont :

— analyser les déterminants du comportement d'achat ;

— mettre en place un modèle de machine learning pour prédire la variable cible `Revenue` ;
— interpréter les variables influentes et les erreurs du modèle ;
— proposer des pistes d'amélioration pour une application marketing.

2 Méthodologie

2.1 Description du dataset

Le dataset comporte 12 330 sessions et 18 variables. Il contient :
— des variables numériques : durées de visite, taux de rebond, taux de sortie, valeur de page ;
— des variables catégorielles : *VisitorType*, navigateur, région, TrafficType, mois, Weekend ;
— une variable cible **Revenue** indiquant si la session a abouti à un achat.

Le dataset présente un léger déséquilibre de classes, car la majorité des utilisateurs ne réalisent pas d'achat.

2.2 Prétraitement

Les étapes de préparation incluent :
— conversion de `Revenue` en classe binaire 0/1 ;
— encodage One-Hot des variables catégorielles ;
— standardisation des variables numériques ;
— séparation en ensemble d'entraînement (70%) et de test (30%).

2.3 Analyse exploratoire (EDA)

L'EDA a permis d'identifier plusieurs tendances importantes :
— corrélation positive entre `ProductRelated_Duration` et la probabilité d'achat ;
— corrélation négative entre `BounceRates` et `Revenue` ;
— importance de `PageValues` ;
— contribution du type de visiteur (*Returning Visitor*).

Des visualisations ont été utilisées : heatmap de corrélations, scatter plots, boxplots et clustering K-Means.

2.4 Modèles utilisés

Trois modèles principaux ont été testés :

— **Régression logistique** : simple, rapide, interprétable ;

— **Random Forest** : gère bien les non-linéarités ;

— **XGBoost** : très performant sur données tabulaires.

Les métriques d'évaluation retenues sont :

— *Accuracy* ;

— *Precision*, *Recall*, *F1-Score* ;

— *ROC-AUC*.

3 Résultats et Discussion

3.1 Performances des modèles

Les performances exactes dépendent des résultats obtenus lors de l'exécution du notebook. Voici un tableau à remplir :

Modèle	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Régression Logistique	[..]	[..]	[..]	[..]	[..]
Random Forest	[..]	[..]	[..]	[..]	[..]
XGBoost	[..]	[..]	[..]	[..]	[..]

TABLE 1 – Comparaison des performances obtenues par les modèles.

3.2 Analyse des erreurs

L'analyse de la matrice de confusion révèle :

— un grand nombre de vrais négatifs, cohérent avec le déséquilibre du dataset ;

— des faux négatifs problématiques, car ils correspondent à des acheteurs non détectés ;

— des faux positifs, pouvant entraîner des actions marketing inutiles.

L'étude des importances de variables (feature importance) montre :

— un fort impact de `PageValues` ;

— l'importance de `ProductRelated_Duration` ;

— l'effet négatif de `BounceRates`.

4 Conclusion

Cette étude montre qu'il est possible de prédire efficacement l'intention d'achat via les comportements de navigation. Les modèles Random Forest et XGBoost obtiennent généralement de meilleures performances que la régression logistique.

Limites

— déséquilibre des classes influençant les métriques ;
— absence de signaux comportementaux plus fins (clics, scrolls) ;
— variabilité temporelle non prise en compte.

Pistes d'amélioration

— utiliser SMOTE pour équilibrer la classe positive ;
— tester des modèles plus avancés (LightGBM, CatBoost) ;
— intégrer des données temporelles plus détaillées ;
— utiliser l'explicabilité par SHAP pour comprendre les prédictions individuelles.