Ecole Internationale Supérieure Privée Polytechnique de Sousse

Machine Learning

Rapport d'Analyse sur la churn prédiction

Encadré par : *Mme. Marwa BEN JABRA*

Réalisé par : *Zeineb KHIARI*

Prédiction du départ des clients en utilisant des algorithmes de classification supervisés

Introduction :

Dans ce rapport, je présente une étude visant à prédire le départ des clients à l'aide d'algorithmes de classification supervisés. L'attrition des clients, également appelée churn, est un problème crucial pour de nombreuses entreprises. Nous avons exploré un ensemble de données disponible dans le jeu de données "Churn_Modelling.csv" afin de construire des modèles prédictifs.

Description des données :

Les données utilisées dans cette analyse proviennent du jeu de données "Churn_Modelling.csv". Elles contiennent des informations sur diverses caractéristiques des clients telles que le sexe, l'âge, le solde du compte, le score de crédit, le pays, etc. La variable cible est Exited, qui indique si le client a quitté l'entreprise 1 ou non 0.

Analyse exploratoire et visualisations :

J'ai commencé par explorer l'ensemble des données à l'aide de statistiques descriptives et de techniques de visualisation des données. La méthode describe() a fourni des statistiques sommaires, telles que la

moyenne, l'écart type et les quartiles, pour chaque variable numérique. Cela nous a permis d'avoir une compréhension initiale de la distribution des données.

Ensuite, j'ai utilisé un diagramme à barres pour visualiser la distribution des clients résiliés par rapport aux clients conservés, en les distinguant par genre. Cette visualisation a permis de mettre en évidence toute disparité entre les deux groupes et de mieux comprendre les caractéristiques des clients résiliés en fonction de leur genre.

De plus, pour traiter les variables catégorielles, j'ai procédé à l'encodage des étiquettes à l'aide du LabelEncoder de scikit-learn. Cela a permis de convertir les valeurs catégorielles en valeurs numériques, ce qui est nécessaire pour entraîner les modèles de machine Learning.

Résultats des modèles :

Nous avons entraîné trois modèles de classifications supervisés sur les données prétraitées : régression logistique, Random Forest Classifier et K-Nearest Neighbors (KNN). Nous avons divisé les données en ensembles d'entraînement et de test, puis évalué les performances des modèles à l'aide du score de précision.

❖ Les performances des modèles sont résumées comme suit :

	Régression Logistique	Random Forest Classifier	K-Nearest Neighbors (KNN)
Accuracy	81.5%	86.45%	79.5%
Precision	59.8%	—	70.2%
Recall	18.1%	—	79.5%
F1-score	27.7%	57.46%	72.1%

Discussion :

Le modèle Random Forest a affiché les meilleures performances en termes d'accuracy et de F1-score, bien qu'il reste de la marge pour l'amélioration. La régression logistique a montré des performances relativement faibles en raison d'une faible recall, indiquant qu'elle pourrait avoir du mal à détecter correctement les cas positifs de churn. Le modèle KNN a affiché une recall élevée mais a également montré des performances modérées dans d'autres métriques.

Conclusion :

En conclusion, cette analyse fournit un aperçu des performances des modèles dans la prédiction du churn des clients. Bien que le modèle Random Forest ait montré les meilleures performances, il reste des possibilités d'amélioration et des facteurs supplémentaires à explorer. Cette analyse sert de point de départ pour des études plus approfondies sur le churn des clients et ses implications pour l'entreprise.