Rapport Business — Analyse de Churn Client & Stratégie de Rétention

1. Contexte et objectif

La fidélisation des clients est un enjeu majeur pour toute entreprise, car le coût d'acquisition d'un nouveau client est souvent supérieur au coût de rétention d'un client existant.

Ce projet s'inscrit dans cette perspective et repose sur le jeu de données IBM HR Analytics, souvent utilisé pour simuler une situation de départ de clients ou d'employés (churn). L'objectif principal est d'identifier les individus présentant un risque élevé de départ, de comprendre les facteurs explicatifs de ce phénomène et de proposer des stratégies de rétention pertinentes et rentables.

Plus précisément, ce projet vise à :

- Déterminer les variables ayant le plus d'influence sur le départ des clients.
- Construire un modèle prédictif capable d'estimer la probabilité de churn.
- Traduire les résultats analytiques en recommandations opérationnelles pour améliorer la rétention et la rentabilité.

2. Jeu de données

Le jeu de données utilisé provient de la base publique IBM HR Analytics Employee Attrition Dataset disponible sur Kaggle via le lien : https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

Il comprend 1 470 observations et environ 35 variables, incluant des caractéristiques démographiques, professionnelles et comportementales (âge, niveau de satisfaction, salaire, ancienneté, fréquence des déplacements, etc.).

Ces variables permettent d'établir des corrélations entre les profils des employés et leur probabilité de départ.

Une analyse préliminaire a permis de constater un déséquilibre des classes : une majorité d'employés restent dans l'entreprise, tandis qu'une minorité quitte l'organisation. Ce déséquilibre a été pris en compte dans la phase de modélisation.

3. Analyse exploratoire des données

Une exploration descriptive approfondie a été réalisée afin de comprendre la structure du jeu de données et les tendances majeures.

Cette phase a permis de dégager plusieurs constats :

- Les employés les plus jeunes et ceux ayant une faible ancienneté présentent une probabilité de départ plus élevée.
- Les individus percevant un faible niveau de satisfaction au travail ou un salaire relativement bas sont plus susceptibles de quitter l'entreprise.
- Certaines variables, comme la fréquence des déplacements professionnels ou le déséquilibre entre vie professionnelle et personnelle, se sont révélées particulièrement explicatives.

L'analyse inclut également une segmentation des profils à l'aide de la méthode RFM adaptée au contexte RH, ainsi qu'une étude comparative entre les employés restés et ceux ayant quitté l'organisation.

4. Modélisation

Après la phase d'exploration, plusieurs modèles de machine learning ont été développés et comparés pour prédire le risque de churn.

Les principales étapes de la modélisation sont les suivantes :

1. **Préparation des données** : encodage des variables catégorielles, normalisation, et équilibrage des classes.
2. **Feature engineering** : création de nouvelles variables explicatives pertinentes (ex. ratio satisfaction/salaire, ancienneté ajustée).
3. **Entraînement et comparaison de modèles** :
    - Régression Logistique
    - Random Forest
    - XGBoost

- LightGBM

4. **Évaluation des performances** : utilisation des métriques standards (accuracy, précision, rappel, F1-score et ROC AUC).

Le modèle XGBoost a offert les meilleures performances globales.

Les résultats du rapport de classification sont présentés ci-dessous :

```
precision    recall  f1-score   support

           0       0.94      0.75      0.83       247
           1       0.36      0.74      0.49        47

    accuracy                           0.75       294
   macro avg       0.65      0.75      0.66       294
weighted avg       0.85      0.75      0.78       294
```

Le modèle atteint une précision globale de 75 %, avec un rappel de 74 % sur la classe minoritaire (les churners), ce qui constitue un bon compromis entre détection et stabilité.

Le ROC AUC atteint 0.84, confirmant la capacité du modèle à bien discriminer les individus à risque.

L'interprétabilité a été renforcée grâce à l'utilisation des valeurs SHAP, permettant d'identifier les variables ayant le plus d'impact sur la probabilité de départ. Les principales variables influentes incluent la satisfaction au travail, le niveau de salaire, la fréquence des déplacements et l'ancienneté.

5. Impact business

L'évaluation de l'impact économique du churn repose sur la notion de Customer Lifetime Value (CLV), représentant la valeur totale qu'un client (ou employé) apporte à l'entreprise au cours de sa relation.

Le calcul du CLV estimé s'appuie sur la formule :

CLV = revenu moyen × marge moyenne × durée moyenne de rétention

En se basant sur des hypothèses réalistes (revenu annuel moyen de 3 000 €, marge de 30 %, durée moyenne de 3 ans), le CLV moyen estimé est d'environ 2 700 €.

Une matrice coût-bénéfice a ensuite été construite pour évaluer la rentabilité des actions de rétention.

Elle permet de déterminer le seuil optimal de probabilité à partir duquel une intervention devient économiquement justifiée.

Ainsi, cibler un client avec un risque de churn supérieur à 0.7 et un CLV élevé maximise le retour sur investissement des actions de fidélisation.

6. Segmentation et priorisation

Sur la base du score de churn et du CLV, une segmentation opérationnelle a été mise en place afin d'orienter les actions de rétention selon la valeur stratégique des individus.

- **Priorité 1** : Clients à haut CLV et risque de churn élevé → action immédiate (campagnes personnalisées, entretiens RH, avantages ciblés).
- **Priorité 2** : Clients à CLV moyen et risque de churn élevé → actions automatisées ou semi-personnalisées.
- **Priorité 3** : Clients à faible CLV et risque de churn élevé → suivi ponctuel ou actions à coût limité.
- **Priorité 4** : Clients à faible risque → monitoring passif et maintien des bonnes pratiques.

Cette segmentation permet d'optimiser les ressources marketing ou RH et de maximiser l'efficacité des campagnes de rétention.

7. Recommandations opérationnelles

Les résultats du modèle et les analyses business conduisent à plusieurs recommandations :

1. **Cibler les individus à fort potentiel économique et risque élevé** à l'aide de campagnes de rétention prioritaires.
2. **Mettre en place des tests A/B** pour comparer différentes stratégies de fidélisation (offres, primes, communication personnalisée).
3. **Intégrer le scoring de churn au sein du CRM** afin d'automatiser la détection et la priorisation des individus à risque.
4. **Mesurer l'efficacité des actions dans le temps** (30, 60 et 90 jours après intervention) pour ajuster les stratégies selon le retour observé.

5. **Déployer un tableau de bord de suivi dynamique** pour visualiser les taux de churn, les scores prédictifs et les indicateurs de performance.

8. Limites et Points d'Attention

Malgré des résultats encourageants, plusieurs limites doivent être prises en compte :

- Le jeu de données est de taille modeste (1 470 observations), ce qui limite la généralisation des résultats.
- Le contexte simulé (jeu de données IBM) ne reflète pas nécessairement la complexité d'un environnement réel.
- Certaines valeurs chiffrées, notamment celles relatives au CLV, reposent sur des hypothèses théoriques et mériteraient d'être validées avec des données économiques réelles.
- Une validation croisée plus robuste (par exemple avec un jeu de données temporel) permettrait de confirmer la stabilité du modèle.

Pour aller plus loin, il serait pertinent d'enrichir le modèle par :

- L'ajout de données temporelles pour suivre l'évolution du comportement des individus.
- L'utilisation de modèles de survie (Cox ou Kaplan-Meier) afin d'estimer la durée avant le churn.
- L'intégration du modèle dans un système décisionnel en production (API ou CRM).

9. Conclusion

Ce projet démontre comment une approche de data science peut être mobilisée pour comprendre et anticiper le phénomène de churn.

Grâce à une combinaison d'analyse exploratoire, de modélisation prédictive et d'interprétation des résultats, il a été possible de proposer une stratégie de rétention ciblée et mesurable.

Le modèle final, basé sur XGBoost, présente une performance solide (ROC AUC de 0.84) et permet de détecter efficacement les individus à risque. Les recommandations opérationnelles issues de cette étude constituent une base pour des actions concrètes et orientées vers la rentabilité.

En résumé, cette démarche met en évidence la valeur ajoutée de l’analyse prédictive pour la prise de décision stratégique et la gestion proactive de la fidélisation client ou employé.

Script Voix-Off - L'Algorithme d'Apriori