CAHIER DES CHARGES

Projet de Fin d'Études
Data Analyste

Analyse Prédictive et Segmentation Clientèle dans un Contexte Bancaire

Nom : Hassan Issil
Encadrant : Mr. Yassine Ammami

2025/2026

1 Introduction et Contexte du Projet

Dans un environnement bancaire de plus en plus concurrentiel, la rétention client est devenue une priorité stratégique. Le *churn* (désabonnement) entraîne non seulement une perte de revenus directs, mais aussi une augmentation des coûts d'acquisition. Les institutions financières disposent aujourd'hui de volumes massifs de données transactionnelles, démographiques et comportementales, qu'elles peuvent exploiter via des techniques de *Business Intelligence* (BI) et de *data science* pour anticiper les départs, personnaliser les offres et optimiser leur relation client.

Ce projet s'inscrit dans cette logique : il vise à transformer des données brutes en *insights* actionnables, en combinant visualisation interactive, modélisation statistique et bonnes pratiques ETL. Les données utilisées sont fictives mais réalistes (ex. : *Churn Modelling*, *Bank-Customers-Demo*), ce qui permet une expérimentation sans risque sur des cas métier pertinents.

2 Problématique Métier

Comment identifier, comprendre et prédire les clients à risque de churn afin de mettre en place des actions ciblées de fidélisation ?

Plus précisément :

— Quels facteurs (géographiques, comportementaux, démographiques) influencent le plus le départ d'un client ?
— Peut-on segmenter la clientèle de manière fine pour adapter la communication marketing ?
— Est-il possible de prédire la probabilité de churn avec un modèle robuste et interprétable ?

3 Objectifs du Projet

3.1 Objectifs Principaux

— Réaliser une analyse exploratoire approfondie des données clients (profils, comportements, tendances)
— Visualiser les disparités de churn par variables clés (genre, pays, âge, produits détenus, etc.) via Tableau
— Construire et évaluer un modèle de régression logistique pour prédire le churn
— Segmenter la clientèle selon des critères géodémographiques et comportementaux
— Automatiser un pipeline ETL (Extraction, Transformation, Chargement) pour assurer la reproductibilité

3.2 Objectifs Secondaires

— Appliquer des tests statistiques (ex. : χ^2) pour valider la significativité des liens observés
— Évaluer la performance du modèle via des métriques pertinentes (CAP Curve, AUC, précision, rappel)
— Documenter l'ensemble du processus dans un repository GitHub propre et structuré

4 Périmètre et Livrables Attendus

TABLE 1 – Livrables du projet 2secondarywhite

Livrable	Description	Format
1. Dataset nettoyé et documenté	Données après wrangling (gestion des doublons, valeurs manquantes, formats)	CSV + dictionnaire de données
2. Dashboard interactif (Tableau)	Visualisations clés : churn par genre/pays/âge, segmentation, KPI	Fichier .twb + export PDF
3. Modèle prédictif (Python)	Modèle de régression logistique avec coefficients interprétés	Notebook Jupyter + script .py
4. Rapport d'analyse	Interprétation métier des résultats, recommandations	PDF (10–15 pages)
5. Repository GitHub	Code, données, documentation, README clair	Lien public
6. Présentation orale	Synthèse visuelle des *insights* et démonstration du dashboard	PowerPoint / Google Slides

HORS PÉRIMÈTRE : Déploiement en production, intégration API temps réel, *machine learning* avancé (Random Forest, XGBoost).

5 Méthodologie et Approche Technique

Le projet suit une démarche CRISP-DM (*Cross-Industry Standard Process for Data Mining*) :

1. **Compréhension métier** → Définition des questions analytiques
2. **Compréhension des données** → Exploration, qualité, distribution
3. **Préparation des données** → ETL avec Python/SQL incluant :
   — Extraction depuis fichiers CSV
   — Transformation : gestion valeurs manquantes, encodage variables catégorielles (*One-Hot Encoding*), normalisation, création de *features*
   — Chargement : sauvegarde CSV pour Tableau et PostgreSQL (optionnel)
4. **Modélisation** → Construction d'un modèle de régression logistique (choisi pour son interprétabilité) comparé à un classifieur *baseline*. En parallèle, segmentation via clustering K-Means sur variables comportementales et démographiques
5. **Évaluation** → CAP Curve, matrice de confusion, validation croisée, métriques adaptées au déséquilibre des classes
6. **Déploiement** → Dashboard, rapport, documentation

Outils & Technologies

— **Langages :** Python (Pandas, Scikit-learn, Matplotlib, Seaborn), SQL
— **BI :** Tableau Desktop (version publique)
— **Base de données :** PostgreSQL (optionnel, pour exercices ETL)
— **Environnement :** Jupyter Notebook, VS Code, Git/GitHub
— **Statistiques :** Test du χ^2, analyse de variance, CAP Curve

6 Planning Prévisionnel (6 semaines)

Livraison finale : 10 février 2026

7 Contraintes et Hypothèses

Contraintes

— Données fictives → pas de contraintes RGPD, mais réalisme métier exigé
— Outils open source ou versions gratuites (Tableau Public, PostgreSQL)
— Temps limité (6 semaines) → focus sur la qualité, pas la complexité algorithmique

Hypothèses

— Les données reflètent fidèlement un contexte bancaire européen
— Le churn est binaire (parti/resté) et bien défini
— Toutes les variables nécessaires sont disponibles (âge, solde, produits, etc.)

Table 2 – Échéancier du projet

Semaine	Activités
S1	Audit des datasets, définition du scope, initialisation GitHub
S2	ETL Phase 1 & 2 : nettoyage, gestion erreurs, documentation
S3	Analyse exploratoire + visualisations Tableau (churn par variable)
S4	Modélisation : régression logistique, évaluation (CAP, AUC, validation croisée)
S5	Segmentation clientèle (K-Means), tests statistiques, amélioration dashboard
S6	Rédaction rapport, finalisation livrables, préparation soutenance

2secondarywhite

8 Indicateurs de Succès et Critères d'Évaluation

Le projet sera jugé sur :

— Pertinence métier des *insights* générés (réponses aux questions initiales)
— Qualité technique du code, du dashboard et du pipeline ETL
— Clarté de la communication (rapport, présentation, documentation GitHub)
— Rigueur analytique (validation statistique, interprétation des coefficients)
— Robustesse de la démarche : méthodologie CRISP-DM suivie et documentée avec choix techniques justifiés (gestion déséquilibre classes, validation modèle)

Succès atteint si :

— Le modèle prédit le churn avec une AUC > 0.80
— Le dashboard permet à un chef de produit de prendre une décision en < 2 minutes
— Le repository est clonable et reproductible par un pair
— Les principaux facteurs de churn sont identifiés et validés par l'analyse exploratoire
— Les segments identifiés présentent des profils et taux de churn distincts permettant des recommandations marketing différenciées

9 Conclusion

Ce projet constitue une démonstration concrète des compétences acquises en *Business Analytics* : de la préparation des données à la restitution stratégique. Il allie rigueur statistique, maîtrise des outils BI et sens du métier, trois piliers essentiels pour tout analyste moderne. Les résultats serviront non seulement de support de fin d'études, mais aussi de portfolio professionnel pour des postes en analyse de données, CRM ou intelligence client.

Annexe : Datasets Utilisés

— `Churn Modelling` (CSV)
— `Bank-Customers-Demo` (CSV)
— `FakeNamesUK/Canada` (CSV)
— `50_Startups` (CSV)
— `Email Offer` (CSV)