Dossier d’architecture technique

Version 1.0

30/02/2022

Historique

Création du document : 30/02/2022

Groupe 5

**El Harrar Ethan**

**2024 – 2025**

# Sommaire

Table des matières

[Sommaire 2](#_Toc118904261)

[Présentation générale 3](#_Toc118904262)

[1 •  Présentation du client 3](#_Toc118904263)

[2 •  Présentation de l'équipe 3](#_Toc118904264)

[3 •  Contexte de développement 3](#_Toc118904265)

[Architecture applicative 4](#_Toc118904266)

[1 •  Liste des composants applicatifs 4](#_Toc118904267)

[2 •  Schéma détaillé de l'architecture applicative 4](#_Toc118904268)

[Architecture Technique 5](#_Toc118904269)

[1 •  Liste ressources 5](#_Toc118904270)

[2 •  Listes détaillé des blobs 5](#_Toc118904271)

[3 •  Liste des bases de données 5](#_Toc118904272)

[4 •  Listes des tables 6](#_Toc118904273)

[5 •  UML 6](#_Toc118904274)

# 

# Présentation générale

## 

## • Présentation du client

Dans ce projet, je m’intéresse aux entreprises qui proposent des assurances automobiles. Cela inclut :

* **Les banques qui font de l’assurance auto** : Ce sont des banques qui ne se limitent pas aux comptes et crédits, mais qui proposent aussi des assurances pour les voitures. Leur objectif est d’attirer plus de clients et de fidéliser ceux qu’elles ont déjà en leur offrant des services diversifiés. Parmi ces banques, je me concentre pour l'instant uniquement sur les banques locales telles que la Bred, BNP Paribas, Société Générale, LCL, CIC, ainsi que plusieurs autres.
* **Les assureurs spécialisés en automobile** : Ces entreprises sont expertes dans le domaine des assurances pour les véhicules. Elles cherchent à se différencier sur le marché pour garder leurs clients, car il y a beaucoup de concurrence et les clients peuvent facilement aller voir ailleurs. Parmi les assurances auto, je me focalise uniquement sur celles spécialisées dans ce domaine, comme Direct Assurances, Eurofil, Assurpeople, et bien d'autres.

Ces deux types d’entreprises doivent faire face à un problème commun : le churn, ou la perte de clients. Comme le marché est très compétitif et que les consommateurs peuvent facilement changer d’assureur pour une meilleure offre, il est super important pour elles de comprendre pourquoi leurs clients partent et comment elles peuvent les garder.

## • Présentation de l'équipe

Pour ce projet, je serai seul à le réaliser. Je m’appelle Ethan, j’ai 24 ans et je suis étudiant spécialisé en data science et data analytics. J’ai choisi de travailler seul sur ce projet parce que j’ai une vision claire de ce que je veux faire et de comment l’implémenter. Cela me permet d’aller directement à l’essentiel et de gagner du temps en évitant les allers-retours qu’il peut y avoir dans un travail d’équipe. Je suis motivé à relever ce défi et à mener ce projet à bien du début à la fin.

## 3 • Contexte de développement

Dans le domaine de l’assurance automobile, garder ses clients est devenu un vrai défi. Avec la concurrence qui ne fait qu’augmenter et les clients qui peuvent facilement comparer les offres sur Internet, beaucoup d’assureurs voient leurs clients partir pour d’autres entreprises. Ce phénomène, qu’on appelle le "churn", coûte cher, car recruter de nouveaux clients est souvent plus compliqué et plus cher que de garder ceux qu’on a déjà.

C’est dans ce contexte que j’ai décidé de faire ce projet. Mon objectif est de mieux comprendre pourquoi les clients partent et de développer un modèle qui pourrait aider à prédire ces départs. L’idée, c’est d’apporter des solutions concrètes aux banques et aux assureurs automobiles pour les aider à fidéliser leurs clients et rester compétitifs.

# 

# Architecture applicative

## 

## 1 • Liste des composants applicatifs

Le projet s’appuie sur les outils et composants suivants pour collecter, traiter et analyser les données ainsi que pour développer le modèle de classification du churn :

**Sources de données**

Fichiers CSV provenant de Kaggle : address, autoinsurance\_churn, customer, demographic, et termination.

**Outils utilisés**

**Databricks File Store (DBFS)** : Pour stocker les fichiers CSV et les rendre accessibles pour les analyses.

**PySpark** : Chargement des fichiers CSV dans Databricks pour une gestion optimisée des données volumineuses.

**Préparation et traitement des données**

**Nettoyage et exploration**

* Suppression des valeurs manquantes et gestion des anomalies.
* Analyse exploratoire des relations entre les données des différents fichiers.

**Transformation des données**

Fusion des fichiers CSV pour constituer une table unique (ou des tables) intégrant les informations clés.

**Outils utilisés**

* Databricks Notebooks : Utilisation de bibliothèque python.

**Développement et entraînement des modèles**

**Méthodologie**

* Modèles supervisés (régression logistique, arbres de décision, etc.) pour prédire la probabilité de churn.
* Tests et validation croisée pour optimiser les performances.
* **Outils utilisés** :
  + **Databricks MLflow** : Suivi des expérimentations et gestion des modèles.
  + **Scikit-learn** et **PySpark MLlib** : Entraînement des modèles.

**Visualisation et documentation des résultats**

* **Résultats** :
  + Génération de tableaux récapitulatifs (CSV) avec les prédictions de churn.
  + Graphiques et visualisations des insights clés dans les notebooks Databricks.
* **Outils utilisés** :
  + **Matplotlib** : Création de graphiques directement dans les notebooks.

**Exécution**

**Approche d’exécution** :

* Pas d’interface utilisateur graphique. Toutes les étapes se font via des scripts Python exécutés dans les notebooks Databricks.

## 2 • Schéma détaillé de l'architecture applicative

Une image contenant diagramme, texte, capture d’écran, conception

Description générée automatiquement

# Architecture Technique

## 1 • Liste ressources (exemple)

## 2 • Listes détaillé des blobs

## 3 • Liste des bases de données

## • Listes des tables

**Termination**

* Contient les informations liées à la suspension des comptes (ACCT\_SUSPD\_DATE).
* Permet de suivre les comptes suspendus et de fournir des informations pour l'analyse du churn.

**Demographic**

* Stocke les données démographiques des individus, comme le revenu, l'état civil, et le nombre d'enfants.
* Sert à enrichir les analyses clients et les comportements dans d'autres tables.

**Customer**

* Regroupe les informations personnelles des clients (âge, durée de fidélité, date de naissance, etc.).
* Est le point central pour lier les données démographiques, géographiques, et comportementales.

**AutoInsuranceChurn**

* Table centrale qui rassemble les informations liées aux clients, leurs comportements, et leurs chances de churn.
* Intègre des données enrichies des autres tables pour des analyses avancées.

**Address**

* Contient les informations de localisation géographique (adresse, latitude, longitude, ville, etc.).
* Permet de localiser les clients et d'analyser les tendances géographiques.

## 

## 5 • UML

