Solution BI Afrique - Casablanca

Rapport du projet du recrutement de stage PFE

Prédiction des Ventes et Analyse Client : Une Solution Alimentée par l'IA pour Optimiser les Performances Commerciales

Filière : Business Intelligence & Analytics

Réalisé par :

Amin Benali

Année universitaire 2024/2025

Table des matières

Table des figures

Chapitre 1

Contexte général du projet

Introdution

Ce chapitre aborde le contexte général du projet. Il présente le contexte et la description ainsi que les objectifs du projets.

1.1 Le contexte du projet

Dans le cadre du processus de recrutement de l'entreprise Solution BI Afrique, ce projet me permet de mettre en valeur mes compétences en tant qu'élève ingénieur en dernière année à l'École Nationale Supérieure de l'Informatique et d'Analyse des Systèmes (ENSIAS), spécialisé en Business Intelligence et Analytics. Il représente également une occasion de participer au processus de recrutement pour un stage de fin d'études (PFE) au sein de l'entreprise.

1.2 Description du projet

Ce projet vise à développer des dashboards interactifs qui présenteront les indicateurs clés de performance (KPI), les prédictions de ventes, ainsi qu'une segmentation des clients. Pour atteindre cet objectif, il est nécessaire de concevoir un pipeline de données capable de collecter, traiter, analyser et exploiter les données pour alimenter ces dashboards. L'objectif ultime est de fournir une solution basée sur l'intelligence artificielle (IA) pour aider à optimiser les performances commerciales et soutenir la prise de décision stratégique.

1.2.1 Les Objectifs du Projet

Dans le cadre de ce projet, l'objectif est de concevoir une solution complète et efficace qui exploite les données pour optimiser les performances commerciales. Les principaux objectifs à

atteindre sont les suivants :

— Développer des dashboards interactifs pour visualiser les indicateurs clés de performance (KPI), les prédictions de ventes et la segmentation des clients.

— Construire un pipeline de données pour la collecte, le traitement, l'analyse et l'exploitation des données.

— Implémenter un algorithme de prédiction basés sur le machine learning pour estimer les tendances et anticiper les besoins du marché.

— Mettre en œuvre une segmentation client avancée pour mieux comprendre les comportements et personnaliser les stratégies commerciales.

Conclusion

En conclusion, ce chapitre a exposé le contexte et les objectifs du projet, qui visent à développer une solution basée sur l'intelligence artificielle pour améliorer les performances commerciales, en mettant l'accent sur la création de dashboards interactifs et la mise en place d'un pipeline de données.

Chapitre 2

Analyse et Conception

Introdution

Ce chapitre de l'analyse et la conception commence par une analyse fonctionnelle qui permet de définir les exigences fonctionnelles du projet. Ensuite, l'architecture globale du pipline qui servira comme un guide dans la réalisation. Et se termine par les choix effectués pour l'implimentation.

2.1 Analyse Fonctionnelle

L'objectif est de concevoir un pipeline de données intégrant un ensemble d'étapes clés pour une exploitation des données dans Power BI. Ce devra satisfaire à des exigences fonctionnelles qui seront détaillées dans cette section.

2.1.1 Exigences Fonctionnelles

1. **Collecte et prétraitement des données :** Mettre en place le processus de collecte et de traitement des données.
2. **Analyses et prédictions :** Développer un modèle de prédiction afin d'anticiper les ventes.
3. **Segmentation des clients :** Implémenter un algorithme de machine learning pour créer une segmentation avancée des clients en fonction de leurs comportements et caractéristiques,
4. **Construction des APIs :** Créer des interfaces de programmation applicative (APIs) pour la récupération des données .
5. **Construction des dashboards :** Concevoir des dashboards dynamiques et interactifs dans Power BI pour visualiser les indicateurs clés de performance, les prédictions de ventes

et les analyses clients.

2.2 Architecture du Systeme

Le système doit répondre aux exigences fonctionnelles. Le produit final sera constitué de dashboards contenant des indicateurs clés de performance (KPI) relatifs aux ventes actuelles, aux prévisions des ventes pour un trimestre, ainsi qu'à la segmentation des clients. Les données utilisées devront passer par une phase de traitement pour la préparation, suivie d'une analyse à l'aide d'algorithmes de machine learning pour les prévisions de ventes et la segmentation des clients. Une fois ce processus est terminé, un backend sera conçu pour récupérer les résultats, et les exploiter dans Power BI.

Le schéma suivant illustre ces étapes :

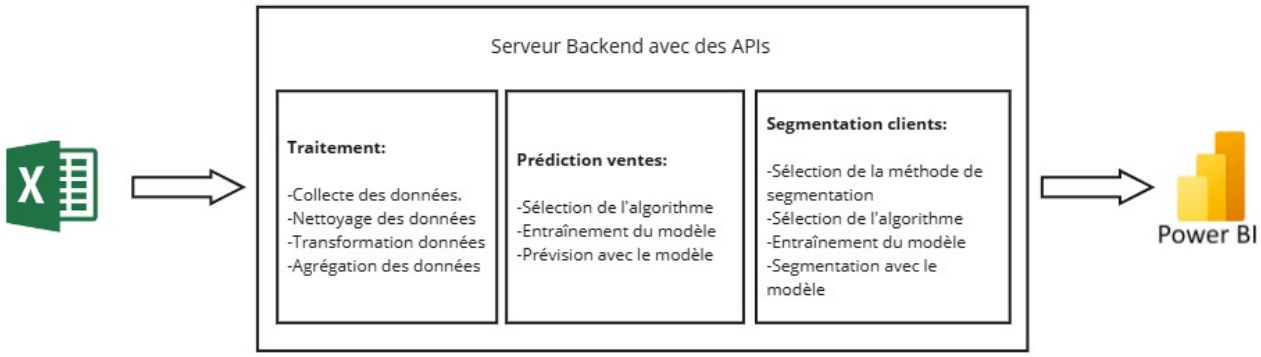

FIGURE 2.1 – Image illustrant l'architecture de l'application

2.3 Choix d'implémentation

2.3.1 Présentation du Dataset

Le dataset **Online Retail II**, disponible sur Kaggle : `https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset`, contient les transactions réalisées par une entreprise britannique spécialisée dans la vente en ligne de cadeaux uniques pour diverses occasions. Les données couvrent une période de deux ans, allant du **1er décembre 2009** au **9 décembre 2011**. L'entreprise,.

Champs du dataset :

— **InvoiceNo** : Numéro de facture unique (6 chiffres). Les factures commençant par "C" indiquent des annulations.

— **StockCode** : Code produit unique (5 chiffres).

— **Description** : Nom du produit.

— **Quantity** : Quantité de produits par transaction.

— **InvoiceDate** : Date et heure de la transaction.

— **UnitPrice** : Prix unitaire (en livres sterling).

— **CustomerID** : Identifiant client unique (5 chiffres).

— **Country** : Pays de résidence des clients.

2.4 Algorithme de prédiction des ventes : LSTM

Les algorithmes de prévision des ventes sont essentiels pour aider les entreprises à anticiper la demande, optimiser les stocks et maximiser les profits. Plusieurs approches couramment utilisées incluent la moyenne mobile, le lissage exponentiel, la régression, ARIMA, SARIMA et les réseaux de neurones. Les réseaux de neurones sont particulièrement populaires, notamment pour leur capacité à traiter des relations non linéaires.

Dans ce cadre, j'ai choisi d'utiliser l'algorithme LSTM (Long Short-Term Memory), un type de réseau de neurones récurrent, car il est particulièrement adapté aux données de séries chronologiques et peut gérer des dépendances temporelles complexes. Les LSTM sont capables de capturer des tendances à long terme tout en tenant compte des variations saisonnières. Cette solution est efficace, puisque je travaille avec un ensemble de transactions commerciales couvrant une période de deux ans.

2.5 Segmentation clients

2.5.1 Méthodes RFM

J'ai choisi la méthode RFM (Récence, Fréquence, Montant) pour segmenter les clients car elle offre une approche simple et efficace pour analyser le comportement des clients en fonction

de trois critères clés :

Récence : la date du dernier achat, car les clients ayant acheté récemment sont plus enclins à revenir. Fréquence : la régularité des achats, ce qui permet de repérer les clients fidèles et réguliers. Montant : la somme dépensée, qui donne une indication sur la valeur des clients et leur propension à effectuer des achats plus importants.

Cette méthode permet de mieux comprendre quels clients sont les plus susceptibles d'effectuer des achats répétés et lesquels génèrent le plus de revenus, ce qui est essentiel pour cibler efficacement les campagnes marketing. Cette méthode est particulièrement adaptée car elle s'appuie directement sur les champs présents dans la dataset, tels que InvoiceDate (pour calculer la récence), Quantity et UnitPrice (pour déterminer le montant dépensé),ainsi que le nombre de transactions effectuées par chaque CustomerID (pour calculer la fréquence des achats).

2.5.2 Algorithme de clustering : K-Means

Il existe plusieurs algorithmes de clustering populaires, tels que K-Means, DBSCAN, Hierarchical Clustering et Gaussian Mixture Models (GMM). Chaque méthode a ses avantages selon les caractéristiques des données. Dans mon cas, j'ai choisi d'utiliser l'algorithme K-Means car il est l'une des méthodes de clustering les plus courantes et performantes pour la ségmentation des clients. Cette méthode est simple à implémenter et efficace pour segmenter des données en groupes distincts.

2.6 Framework backend : Flask

Flask est un framework léger et flexible pour le développement d'applications web en Python. Il permet de créer rapidement des API RESTful. Flask est particulièrement adapté pour des projets où la simplicité, la rapidité de développement sont essentielles.

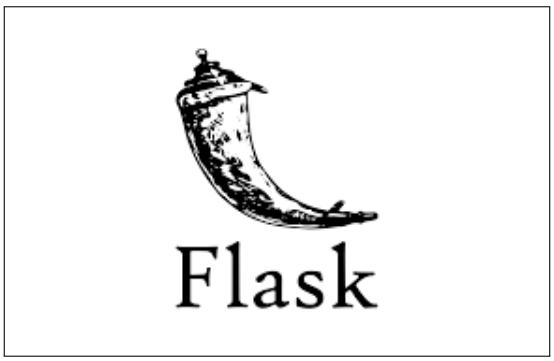

FIGURE 2.2 – Flask logo

Conclusion

En conclusion, ce chapitre a permis de détailler les fondations conceptuelles et techniques sur lesquelles repose le projet.

Chapitre 3

Réalisation

Introdution

Ce chapitre présente la phase de réalisation du projet, dans laquelle les différentes étapes de développement du pipeline sont détaillées. Finalement, il présente les tableaux de bord, qui constituent l'objectif final du projet.

3.1 Exploration des données et nettoyage des données

Voilà une description de notre datset :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
```

FIGURE 3.1 – Description de la dataset

Cette description montre que nous avons 1.067.371 instances dans le dataset. Cependant, les champs Customer ID et Description contiennent des valeurs manquantes. Les valeurs manquantes dans le champ Description ne sont pas assez importantes, car elles n'impactent pas directement les calculs ou les analyses prévues. En revanche, les valeurs manquantes dans Customer ID sont essentielles, car ce champ est indispensable pour effectuer correctement les calculs du modèle RFM utilisé pour la segmentation des clients. D'où le besoin pour éliminer les instances où les valeurs de Customer ID sont manquantes. Ainsi, nous devons supprimer les

instances dupliquées afin que le dataset sera pret pour l'exploitation.

```
<class 'pandas.core.frame.DataFrame'>
Index: 797885 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Invoice      797885 non-null  object
 1   StockCode    797885 non-null  object
 2   Description  797885 non-null  object
 3   Quantity     797885 non-null  int64
 4   InvoiceDate  797885 non-null  datetime64[ns]
 5   Price        797885 non-null  float64
 6   Customer ID  797885 non-null  float64
 7   Country      797885 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 54.8+ MB
```

FIGURE 3.2 – Description de la dataset aprés le nettoyage

3.2 l'entrainement du modèle LSTM

Le modèle LSTM est entraîné avec des paramètres choisis après une considération pour assurer une bonne performance :

- **window_size = 30** : La taille de la fenêtre détermine combien de jours passés qui seront utilisées pour prédire les ventes du jour suivant.
- **50 unités dans la couche LSTM** : Le nombre de neurones dans la couche LSTM. 50 unités est un compromis souvent utilisé pour des séries temporelles de taille moyenne afin d'éviter le surajustement.
- **Fonction d'activation ReLU** : La fonction d'activation ReLU (Rectified Linear Unit) est souvent utilisée dans les réseaux neuronaux pour introduire de la non-linéarité. Elle est largement utilisée pour les tâches de séries temporelles et les réseaux LSTM.
- **Optimiseur Adam avec taux d'apprentissage de 0.001** : Des paramètres très populaires souvent utilisés avec LSTM.
- **Entraînement sur 50 époques** : Le nombre d'époques détermine le nombre de fois que le modèle passe sur l'ensemble des données.

Après l'entraînement, le modèle est sauvegardé. Lors de la prévision future, cette sauvegarde du modèle permettra de le réutiliser pour faire des prévisions lors de l'utilisation de l'API.

3.3 Utilisation de RFM et de K-means pour la segmentation

La RFM est utilisée pour la segmentation. Elle repose sur trois critères clés :

— **Récence** : Le nombre de jours écoulés depuis la dernière transaction du client.

— **Fréquence** : Le nombre total de transactions effectuées par le client.

— **Montant** : Le montant total dépensé par le client.

Ensuite, pour effectuer la segmentation des clients, l'algorithme KMeans est utilisé. KMeans est un algorithme de clustering non supervisé qui regroupe les données en un nombre défini de clusters en fonction de la similarité entre les données. Dans notre cas, les données sont les scores RFM des clients.

Voici les paramètres utilisés dans le KMeans :

— `n_clusters=5` : Le nombre de clusters dans lesquels les clients doivent être regroupés.

— `random_state=42` : Le paramètre de graine aléatoire permet d'assurer que les résultats du clustering sont reproductibles.

3.4 Restful APIs

Pour les API, trois adresses ont été créées pour retourner les résultats :

— `/OriginalData` : L'utilisation de cette API permet d'obtenir les données sous format JSON après le nettoyage.

— `/ForecastNextQuarter` : Cette API permet de retourner les prévisions des ventes pour les 30 jours suivants le dernier jour enregistré. Les données sont au format JSON et contiennent deux champs : `date` et `prévision`.

— `/RFM_Clustering` : Cette API permet de retourner la segmentation des clients. Elle contient 5 champs : `Customer ID`, `Recency`, `Frequency`, `Monetary` et `Cluster`.

L'utilisation de ces API permettra d'exploiter ces données sur Power BI en facilitant l'intégration directe des résultats dans des tableaux de bord. Grâce à ces API, on peut obtenir les données les plus récentes d'une manière directe et les exploiter facilement avec Power BI.

3.5 Résultats finaux

Ce processus a permis de créer des tableaux de bord pour visualiser les les performances et les tendances des ventes et analyser la segmentation des clients.

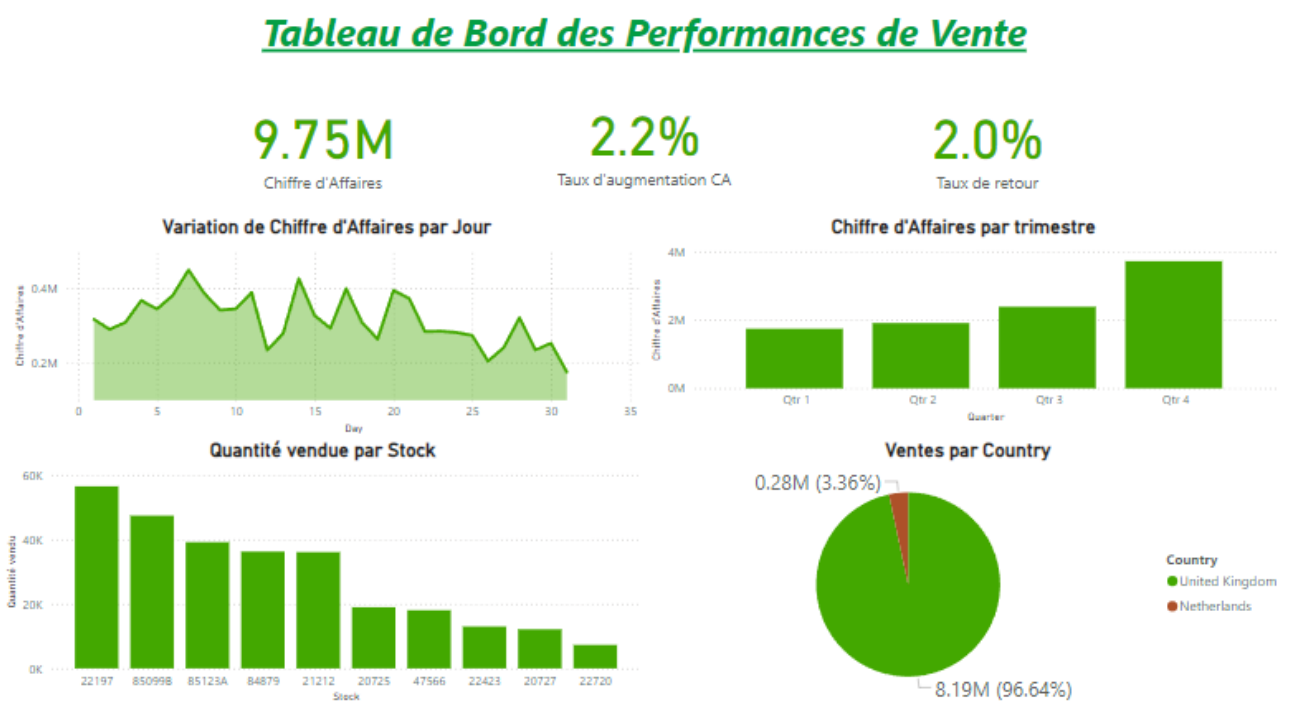

FIGURE 3.3 – Tableau de bord des performances de ventes

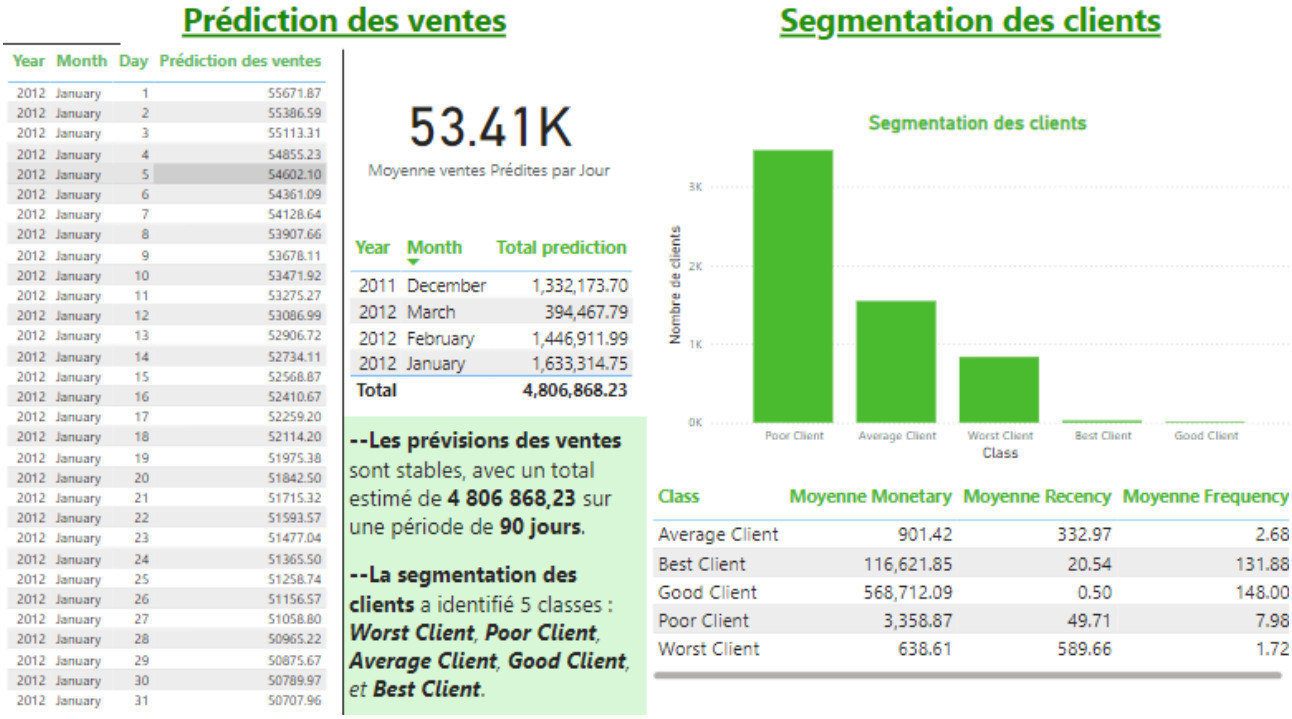

FIGURE 3.4 – Tableau du bord des Prévions de Ventes et Segmentation de Cliens

Conclusion

Ce chapitre décrit les étapes concrètes de la réalisation du projet. Il couvre la mise en place du pipeline et présente, à la fin, les tableaux de bord créés à l'aide de ce pipeline.