Documentation :

Création d’un modèle de Prédiction à l’aide de Qlik Cloud AutoML

Introduction :

Ce rapport détaille les étapes pour créer un modèle de prédiction impliquant SSIS (SQL Server Intégration Services), Qlik Cloud, et AutoML dans le but de prédire les ventes pour l'année 2017 et les comparer avec les données qu’on a déjà à propos des ventes dans 2017.

1. **Qu'est-ce qu'AutoML ?**

L'apprentissage automatique automatisé, également connu sous le nom de ML automatisé ou AutoML, est une technologie émergente permettant d'automatiser les tâches d'apprentissage automatique, d'accélérer le processus de création de modèles, d'aider les data scientiste à se concentrer sur des tâches de plus grande valeur et d'améliorer la précision des modèles Machine Learning. AutoML tente d'automatiser certaines parties du flux de travail de la science des données et d'orienter la prise de décision basée sur les données.

1. **QlikCloud AutoML**

Qlik AutoML est un outil d'apprentissage automatique (Machine Learning) intégré dans la plateforme cloud de Qlik. Il a pour but d’aider les entreprises à prendre des décisions éclairées en prédisant les futurs comportements à partir des données passées et des modèles statistiques intégrés. Avec AutoML, même les utilisateurs sans expertise en science des données peuvent créer des modèles prédictifs puissants en quelques étapes simples.

1. **Taches réalisés pour effectuer la prédiction au niveau de qliksense :**
2. Fusion de l'ensemble des tables de la datawarehouse :

La phase initiale de notre travail impliquait le rassemblement de toutes les tables de données dispersées en utilisant SSIS dans Visual Studio. L'objectif était de créer un fichier Excel « full\_data » contenant l'ensemble des données disponibles. Cette démarche nous a permis d'effectuer une meilleure sélection et classification des données pertinentes pour notre modèle.



1. Importation du Résultat dans Qlik Cloud :

L'importation du fichier Excel "full data" dans Qlik Cloud vise à faciliter une analyse plus poussée en utilisant les fonctionnalités avancées de Qlik. À partir de ce fichier, nous allons créer le dataset correspondant aux données de ventes de 2015 et 2016, ainsi que les données d'entraînement pour l'année 2017, en utilisant un script QlikView.

Une image contenant texte, capture d’écran, affichage, Système d’exploitation

Description générée automatiquement

1. Générer le dataset et data training à partir de QlikView

L'objectif de notre modèle est de prédire la moyenne des ventes pour chaque produit. Dans notre exemple, nous avons sélectionné les données suivantes que nous considérons pertinentes pour superviser et entraîner notre modèle :

* **Sales :** Représente la moyenne des ventes pour chaque produit pour chaque année.
* **GrossSales** : Correspond à la moyenne des ventes brutes, soit le montant total des ventes d'un produit sans déduire les coûts, les remises, les retours ou autres déductions.
* **Unit Price** : Indique la moyenne du coût ou du prix associé à l'achat d'une seule unité du produit en question.
* **Sales Qty** : Représente la moyenne du nombre d'unités d'un produit vendues.
* **Margin** : Renvoie la moyenne de la marge bénéficiaire des produits vendus.
* **QualRating** : Mesure la qualité du produit.
* **Catalog Price** : Correspond au prix de vente du produit.
* **Item Number** : Est un identifiant unique associé à chaque produit.
* **Year** : Indique l'année à laquelle les données de vente se rapportent.

En adoptant la moyenne comme méthode d'agrégation sur une période spécifique, le modèle de prévision gagne en stabilité et en précision, car il atténue les fluctuations brusques des données. Cette approche peut mettre davantage en évidence les tendances, surtout lorsque les ventes connaissent des variations importantes d'un jour à l'autre.

Le script ci-dessous permet de créer à partir de l’application "full\_data" deux ensembles de données distincts : "DatasetFinal2.qvd" comprenant les données de ventes pour chaque produit pour les années 2015 et 2016 utilisé comme donnée d’apprentissage dans nôtre model, ainsi que "DatasetTest2.qvd" qui contient les données de ventes pour chaque produit pour l'année 2017. Ce dernier servira de point de référence pour la comparaison avec les résultats de prédiction de notre modèle et ainsi entrainer nôtre model et vérifier sa fiabilité.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran, affichage, Système d’exploitation

Description générée automatiquementUne image contenant texte, capture d’écran, affichage, Système d’exploitation

Description générée automatiquement

1. Création du model et importation du dataset

Au cours de cette étape, nous avons mis en place un modèle et avons importé notre ensemble de données de ventes des années 2015 et 2016. Nous avons sélectionné la feature "Sales" comme objectif principal du modèle, car notre intérêt porte sur la prévision de la moyenne des ventes pour l'année suivante.

Qlik AutoML s'est ensuite chargé de rechercher l'algorithme le mieux adapté à notre problème, en mettant l'accent sur l'obtention d'une précision et fiabilité maximale, mesurée par l'indicateur R2.

De plus, Qlik AutoML offre diverses fonctionnalités lors de la définition du modèle, notamment l'optimisation des hyperparamètres. Cette fonction permet d'ajuster les paramètres de l'algorithme sélectionné par Qlik AutoML, ce qui permet d'optimiser davantage le modèle et d'améliorer sa précision.

N.B : Nous n’avons pas sélectionné les features « ItemNumber » et « Year » car ce sont des données catégoriques utilisé pour le regroupement des données de la dataset.

Une image contenant texte, nombre, ligne, Police

Description générée automatiquement

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

Description générée automatiquement

On peut aussi voir l’importance et le poids pour chaque feature par rapport à l’algorithme sélectionné.

1. Déploiement et training du model

Une fois le modèle prêt, nous avons déployé le model afin de le rendre utilisable et vérifier son accuracy et fiabilité. La première étape consiste à importer la data training dans le modèle, notre data training comporte les données de ventes pour l’année 2017 en respectant le même format et features de notre dataset, Qlik AutoML comprend automatiquement que la feature « ItemNumber » est notre clé primaire qu’on va détailler par la suite son importance. Qlik AutoML généré par défaut trois ensemble de donnée QlikView :

Après avoir préparé le modèle, nous avons procédé au déploiement de celui-ci pour le rendre opérationnel, tout en évaluant sa précision et sa fiabilité. La première étape de ce processus consiste à importer les données d'entraînement dans le modèle. Notre jeu de données d'entraînement comprend les données de ventes pour l'année 2017, respectant le même format et les mêmes caractéristiques que notre ensemble de données initial. Il est important de noter que Qlik AutoML reconnaît automatiquement que la caractéristique "ItemNumber" comme clé primaire, et nous expliquerons par la suite son importance.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Par défaut, Qlik AutoML génère trois ensembles de données QlikView, mais parmi eux, les deux les plus pertinents sont :

🡪DatasetTest2\_Prediction\_SHAP.qvd : Cet ensemble contient les valeurs associées à l'axe des abscisses pour chaque caractéristique (feature) et pour chaque "ItemNumber".

🡪DataTest2\_Prediction.qvd : Cet ensemble comprend les résultats du modèle et les prédictions des ventes pour l'année 2017 pour chaque produit c’est-à-dire "ItemNumber".Une image contenant texte, diagramme, ligne, capture d’écran

Description générée automatiquement

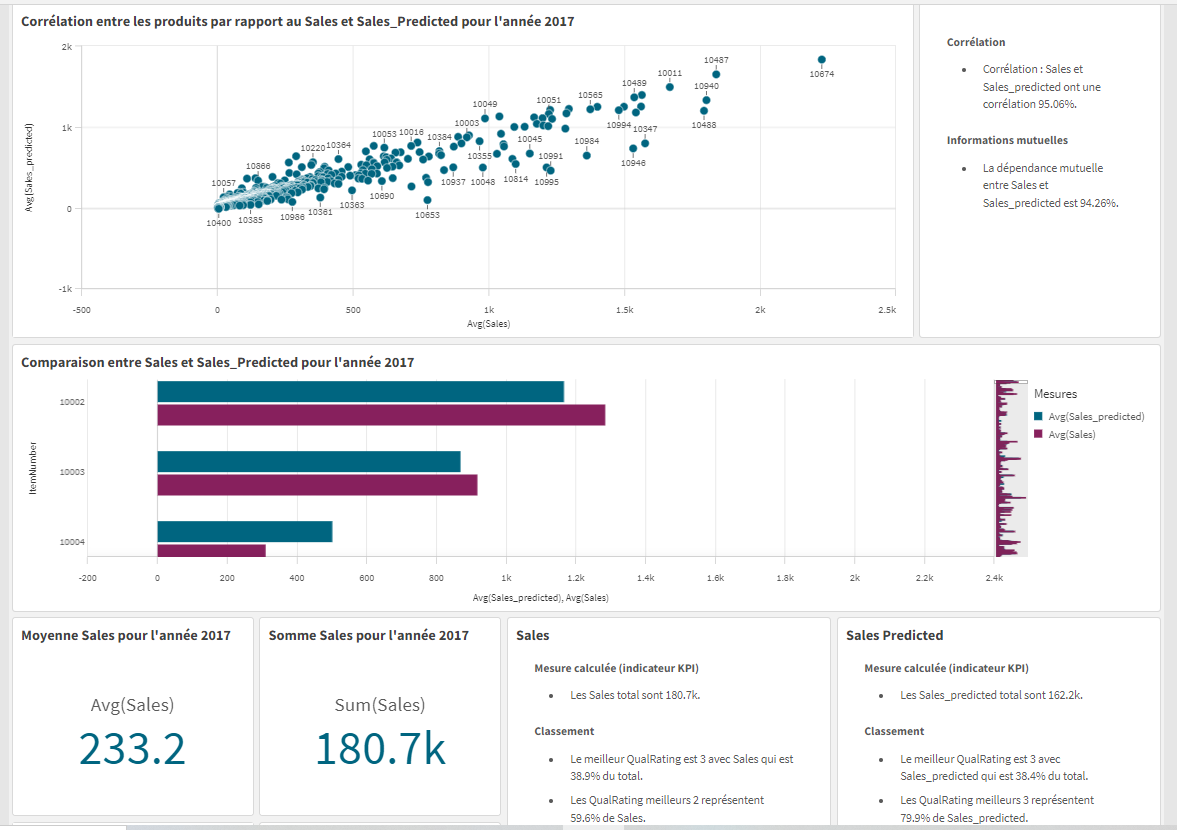
6-Visualisation des données :

Une image contenant cercle, verre à main, conception

Description générée automatiquement

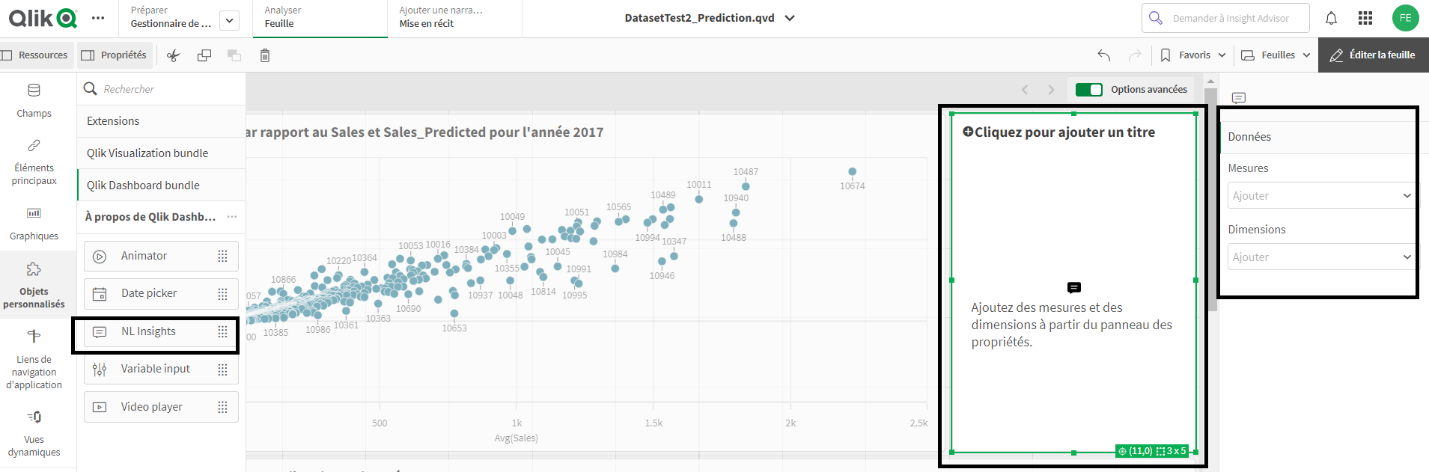
Nous avons créé une application d'analyse pour l'ensemble de données "DataTest2\_Prediction.qvd". Dans le gestionnaire de données, nous avons effectué une opération de jointure en associant l'ensemble de données initial d'entraînement, "DatasetTest2.qvd", avec l'ensemble de données de prédiction, "DatasetTest2\_Prediction.qvd" et « DatasetTest2\_Prediction\_SHAP», en utilisant la caractéristique "ItemNumber". Cette caractéristique revêt une grande importance dans ce contexte car elle nous permet de fusionner les deux ensembles et nous faciliter la visualisation des données.

Une fois que les ensembles de données QlikView ont été fusionnés, nous pouvons élaborer une feuille d'analyse pour visualiser, au moyen d'un Dashboard, l'écart entre la moyenne des ventes réelles de 2017 et la moyenne des ventes prédites par notre modèle. Cette démarche nous permet de vérifier la précision du modèle.



D’après ce Dashboard, on peut déduire que notre modèle concorde étroitement avec le degré de fiabilité calculé par l'indicateur R2, qui atteint 85%. L'adoption de la moyenne en tant que méthode d'agrégation a véritablement permis de stabiliser les données et de réduire les variations entre les ventes pour chaque produit pour l’année 2017.

Comme vous pouvez le constater, QlikView nous offre la possibilité de générer automatiquement des commentaires pertinents en relation avec nos graphiques :



Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Sélectionnez la section « Objets personnalisés » et faites glisser un objet « NL insights » dans votre Dashboard. Ensuite, choisissez les mesures et dimensions que vous souhaitez analyser, et QlikView se chargera d'afficher les conclusions sous forme de commentaires.

Conclusion :

Cette tâche complexe de mise en place d'un flux de données impliquant SSIS, Qlik Cloud et Auto-ML a été menée à bien avec succès. Nous avons réussi à consolider les données, à les importer dans Qlik Cloud pour l'analyse, à créer des datasets correspondantes des dates 2015-2016 et d’une data d’entrainement de 2017, à entraîner des algorithmes et à obtenir des prédictions de ventes significatives pour l'année à venir. Cette expérience a été une opportunité exceptionnelle pour enrichir notre connaissance en matière de data science et de machine Learning. Elle nous a permis de comprendre comment ces technologies peuvent être intégrées pour résoudre des problèmes commerciaux concrets.