Le minage des données

Objectifs

Présentation

- Définir le minage de données
- Décrire le procès de minage
- Découvrir les minage sous SQL Server

11.

Le minage des données

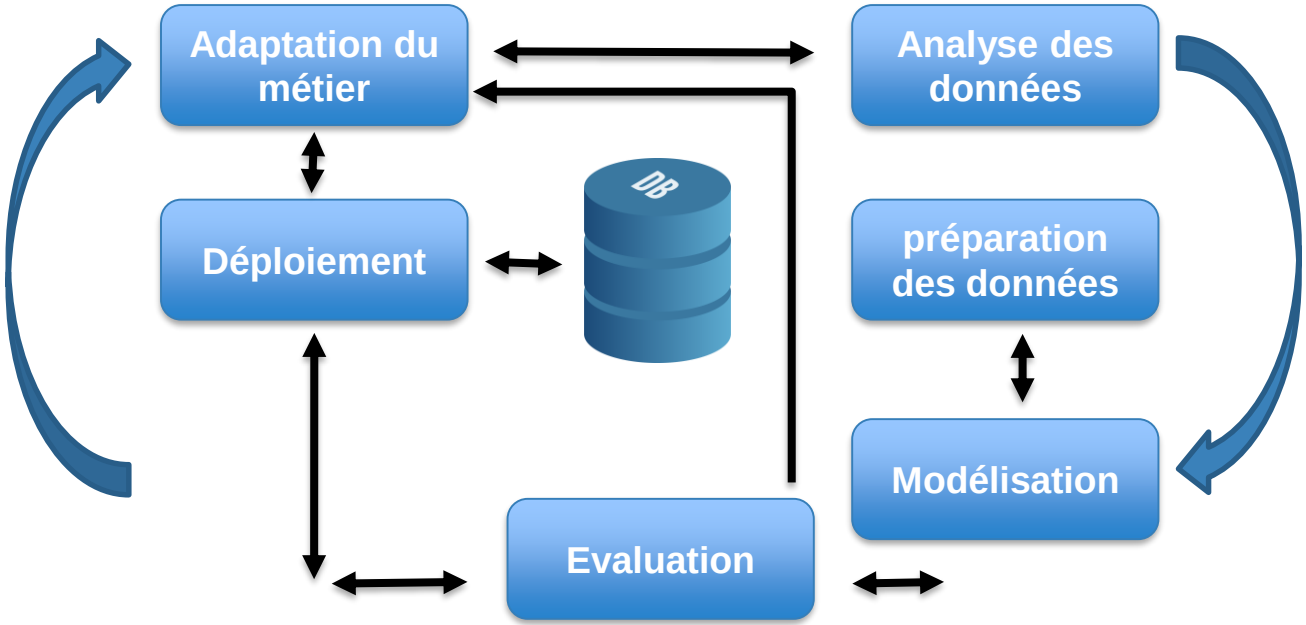
Découvrons ensemble le minage des données!!!!

Définir le minage

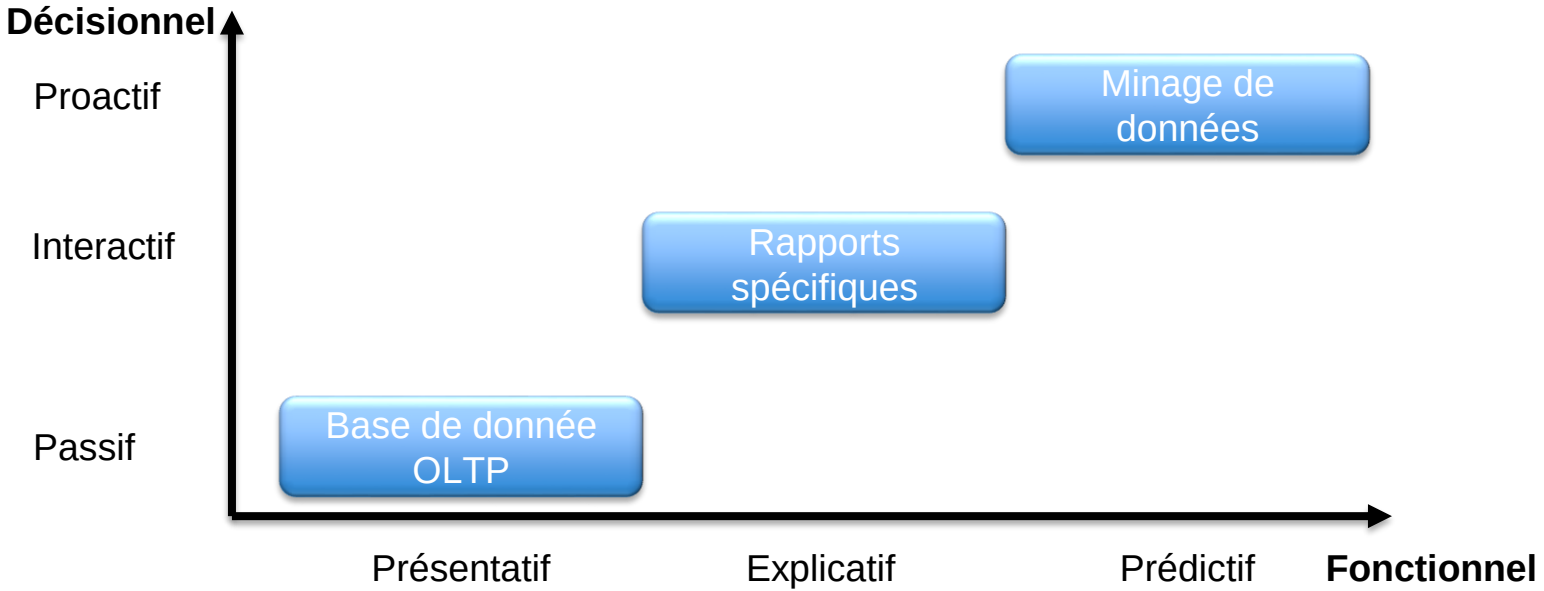
C’est quoi?

- Le minage des données répond au dilemme beaucoup de données et peut d’informations
- La principale idée derrière le minage c’est la prévision qui est la clé principale du BI
- Le minage propose un ensemble d’algorithmes prédéfinis sous SQL Server qui permettent de créer des prévisions selon le contexte

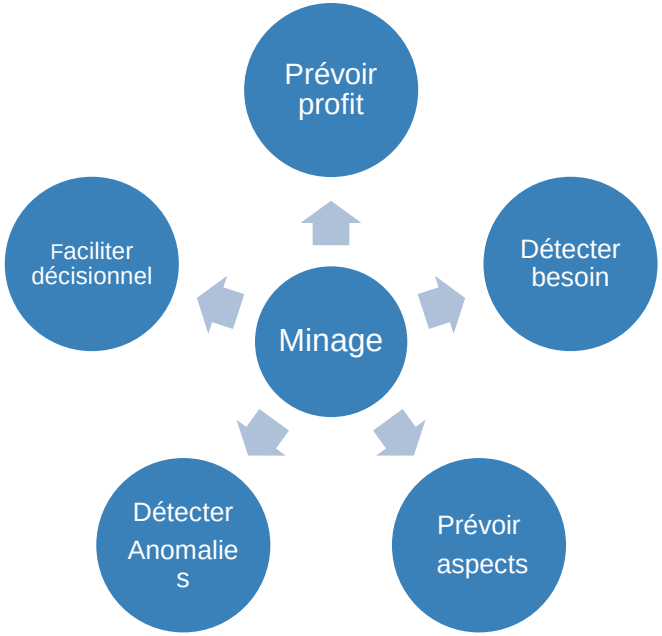
Le procès du minage



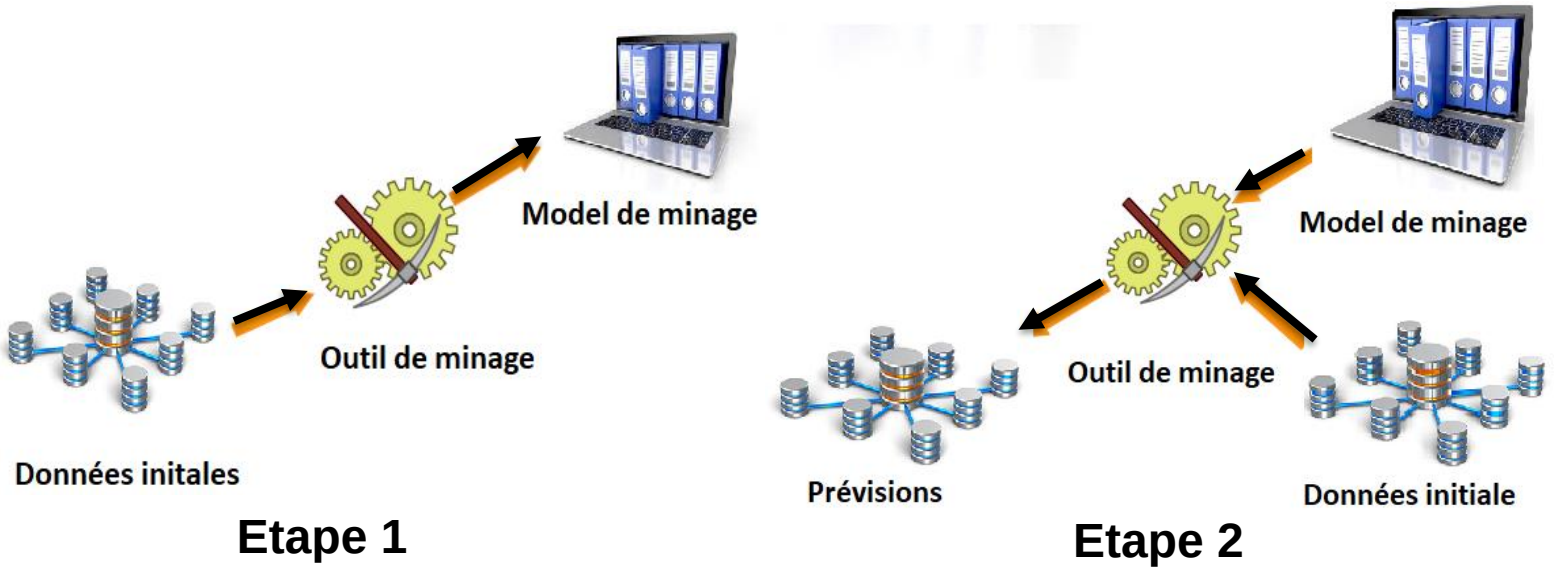
Le procès du minage



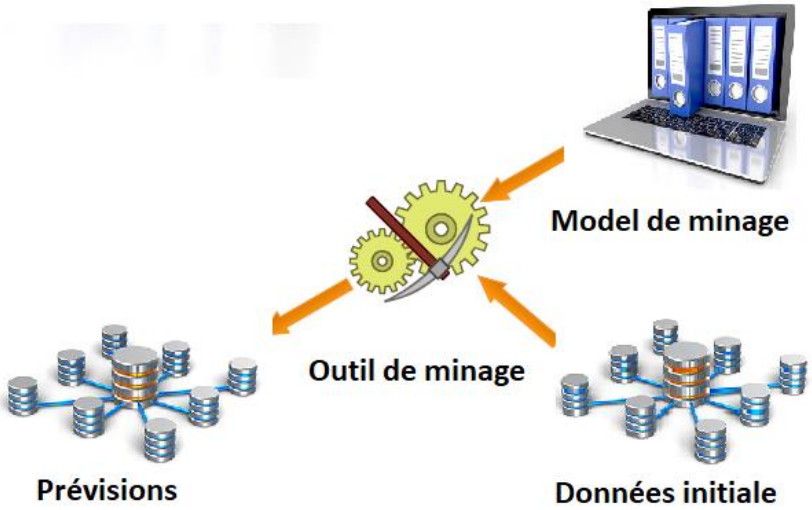
Le procès du minage



Le procès du minage



Le procès du minage



Le procès du minage

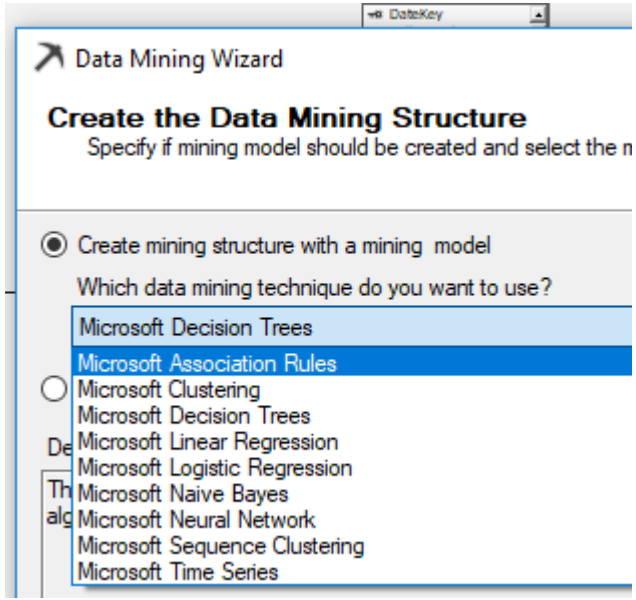
Quel sont les domaines de minage?

- **Classification:** La ségrégation des cas sur la base d'un critère donné
- **Régression:** La recherche de corrélation entre plusieurs facteurs
- **Segmentation:** La détection de bord ou de contour
- **Association:** La recherche de traits en commun entre les éléments
- **Prévision:** L'estimation des résultats futur sur la base des études empiriques
- **Exploration:** L'obtention des informations à caractères explicatives des phénomènes étudiés

Le miange sous SQL Server

Qu’est ce qui nous offre?

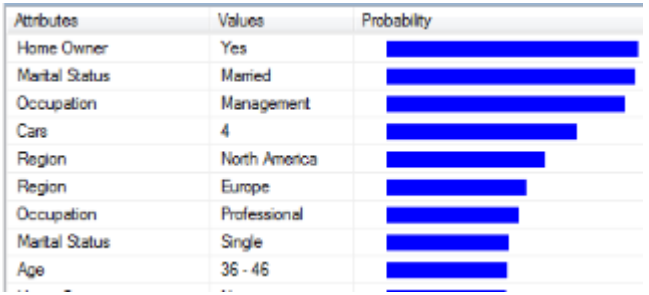
- Réduit la complexité de développement des procédées des minages
- Offre 9 algorithmes de minage parmi les plus utilisés
- Une interface simple avec outils visuel intuitifs
- Un Framework pour augmenter les capacités de minage et une parfaite intégration avec les autres outils Microsoft exemple Excel



Le miange sous SQL Server

Microsoft Naîves Bayes

- C’est un algorithme de classification basé sur le théorème de Bayes
- C’est un algorithme simple et rapide mais qui ne prend pas en compte l’interdépendance des faits représentés par les divers colonnes
- Utilisé pour une exploration initiale

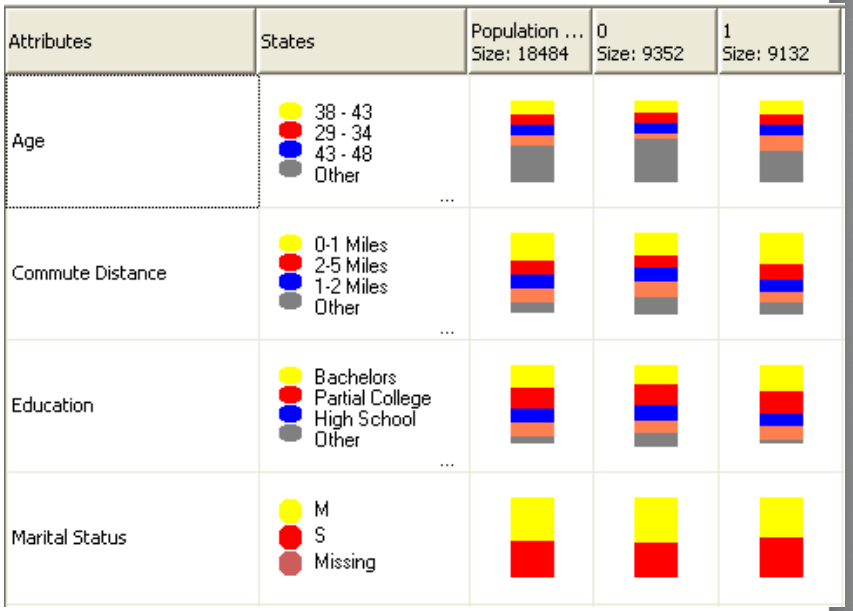


Attributes	Values	Probability
Home Owner	Yes	
Marital Status	Married	
Occupation	Management	
Cars	4	
Region	North America	
Region	Europe	
Occupation	Professional	
Marital Status	Single	
Age	36 - 46	

Le miange sous SQL Server

Comment sa fonctionne?

- Prenons l'exemple d'une compagnie qui veut lancer une compagnie de marketing
- Elle veut envoyer des mails seulement vers les clients potentiels
- L'algorithme de Bayes calcule la probabilité de chaque état de chaque colonne d'entrée



Le miange sous SQL Server

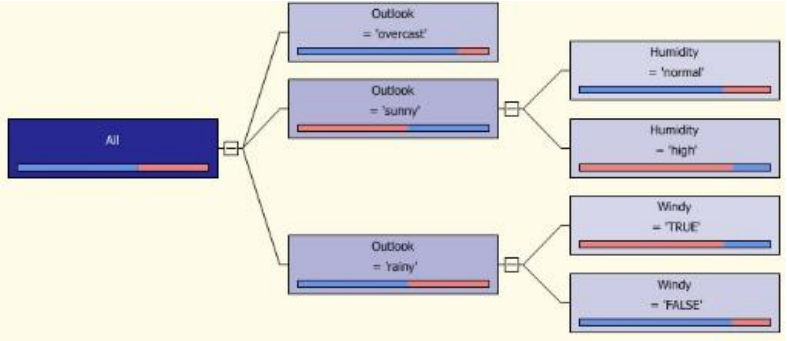
Données entrées?

- **La clé:** L'identifiant au niveau d'une entité sa pourra être numérique ou texte
- **Les noms de colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir
- **La colonne de prévision:** Au moins une colonne doit être préservée à la prévision
- Les données d'entré doivent être absolument discrètes sinon nous les discrétisons
- Les attributs relatifs aux données d'entrée doivent être indépendants

Le miange sous SQL Server

Microsoft Descision Trees

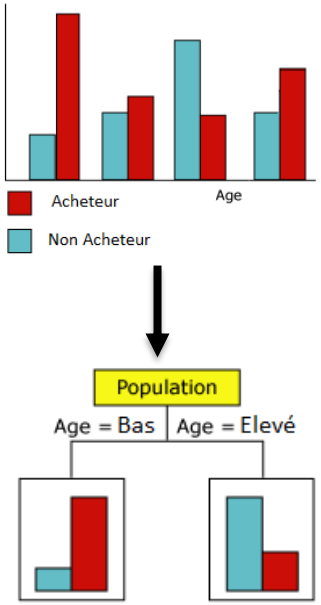
- C’est un algorithme de classification et de régression en mode discret et continu
- Dans un contexte discret, les prévisions se basent sur la relation entre les colonnes celles d’entrée et celle de prévision, c’est la tendance
- Dans contexte continu, il utilise une simple régression linéaire pour trancher
- L’algorithme crée un arbre décisionnel relatif à chaque colonne de prévision



Le miange sous SQL Server

Comment sa fonctionne?

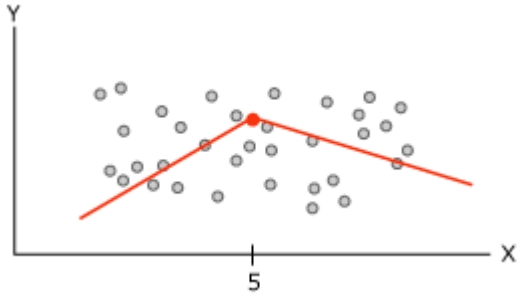
- Dans le contexte discret, prenons l'exemple d'une compagnie qui veut lancer une compagnie de marketing
- Elle veut hiérarchiser les clients envers l'achat d'un produit particulier selon leurs caractéristiques
- Après une corrélation, il présente la prévision au niveau d'un nœud et il passe vers la caractéristique suivante



Le miange sous SQL Server

Comment sa fonctionne?

- Dans le contexte continu, c'est la régression linéaire qui sera utilisée
- La segmentation se base sur le point de non linéarité représenté par le model
- Plusieurs model sont utilisés à la fois pour améliorer le résultat de segmentation



Le miange sous SQL Server

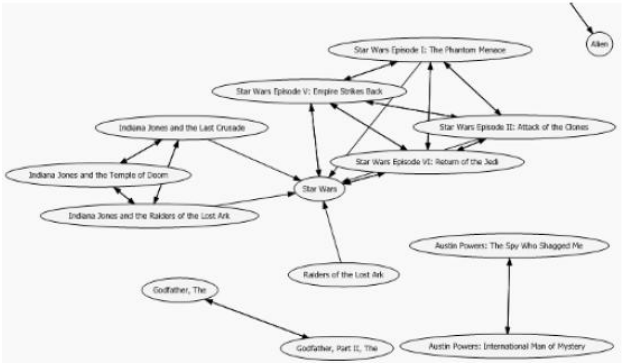
Données entrées?

- **La clé:** L'identifiant au niveau d'une entité numérique ou texte, elle ne doit pas être composée
- **Les noms de colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir qui peuvent être continues ou discrètes
- **La colonne de prévision:** Au moins une colonne doit être préservée à la prévision qui peuvent être continues ou discrètes

Le miange sous SQL Server

Microsoft Associations

- C’est un algorithme de recommandation qui se base sur des études empiriques
- Une association est composé d’un ensemble dit de **Itemsets**
- Exemple, nous avons deux produits X et Y, en compte le nombre de fois où les X et Y sont au panier
- La prévision sera basée sur le nombre de fois d’associations X,Y effectuées par le client

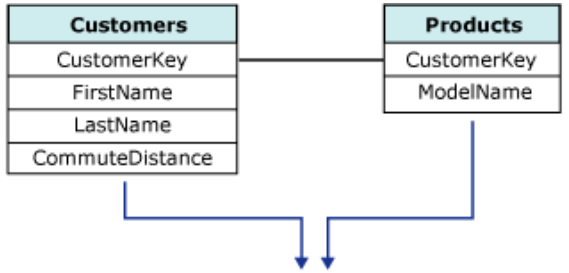


Le miange sous SQL Server

Données entrées?

- **La clé :** L'identifiant au niveau d'une entité numérique ou texte, il ne doit pas être composé
- **Les noms de colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir doivent être discrètes
- **La colonne de prévision:** Une seule colonne doit être préservée à la prévision, les données doivent être discrètes

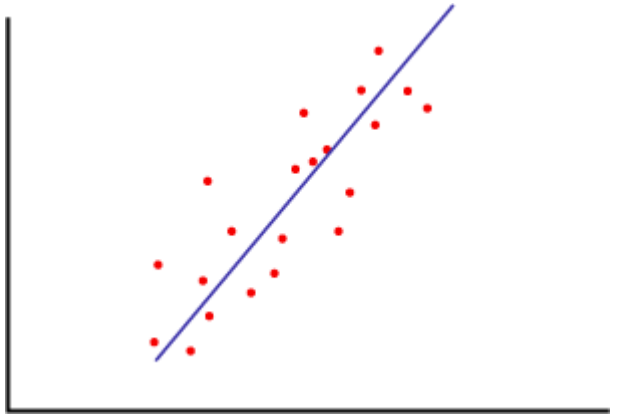
Remarque: La colonne de prévision est généralement au niveau d'une table imbriquée



Le miange sous SQL Server

Microsoft Linear regressions

- C'est un algorithme considéré comme une variance de Binary Tree utilisé dans un contexte continu
- Il permet de prévoir une corrélation entre deux phénomènes différents
- Il est possible de prévoir la relation entre la hausse des ventes d'un produit X par apport à la baise d'un produit Y



Le miange sous SQL Server

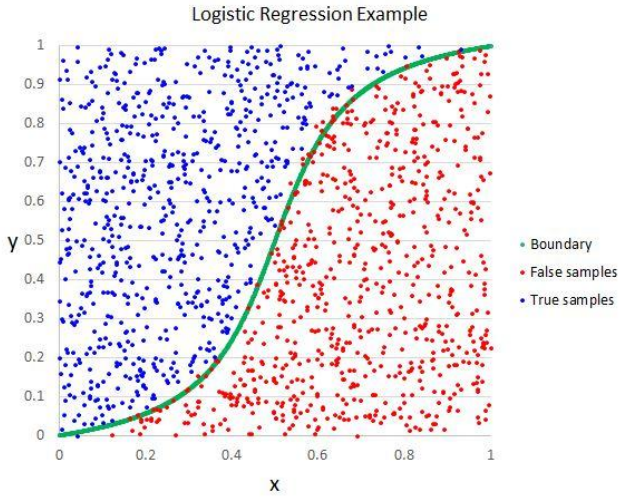
Données entrées?

- **La clé :** L'identifiant au niveau d'une entité, il ne doit pas être composé
- **Les noms de colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir, elles doivent être continues
- **La colonne de prévision:** Au moins une colonne doit être préservée à la prévision, les données doivent être continues et non de type date

Le miange sous SQL Server

Microsoft logistic regressions

- C'est un algorithme considéré comme un algorithme de régression non linéaire
- Il permet de prévoir une corrélation entre deux phénomènes différents comme le modèle linéaire mais pour prévoir plusieurs facteurs à la fois
- Prenons un groupe de personnes homogènes et leur attitude à acheter un produit X, l'algorithme peut déterminer la probabilité d'achat de se produit de l'une de ces personnes



Le miange sous SQL Server

Données entrées?

- **La clé :** L'identifiant au niveau d'une entité non composé
- **Les colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir
- **La colonne de prévision:** Au moins une colonne doit être préservée à la prévision

Remarque: Les colonnes imbriquées ne sont pas prises en considération comme colonnes de prévision dans ce cas

Le miange sous SQL Server

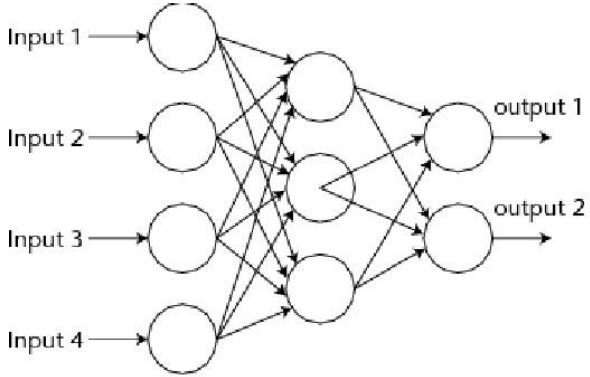
Microsoft Neutral Network

- L'algorithme Microsoft Neural Network est une implémentation du réseau neuronal adaptable pour Machine Learning
- L'algorithme fonctionne en testant chaque état possible de l'attribut d'entrée sur chaque état possible de l'attribut prévisible
- Les probabilités sont calculées pour chaque combinaison en fonction des données d'entraînement
- La sortie de cet algorithme est multiple, il calcule la probabilité cas par cas
- Ces probabilités peuvent être ensuite réutilisées dans des opérations de classification ou régression
- Il est possible de le combiner avec une analyse des associations

Le miange sous SQL Server

Microsoft Neutral Network

- Il est possible de l'utiliser pour mesurer le degré de succès d'une compagne de marketing
- Prévoir les comportements boursiers
- Minage du texte
- En général, toute prédiction qui abouti à plusieurs alternatives à la fois



Le miange sous SQL Server

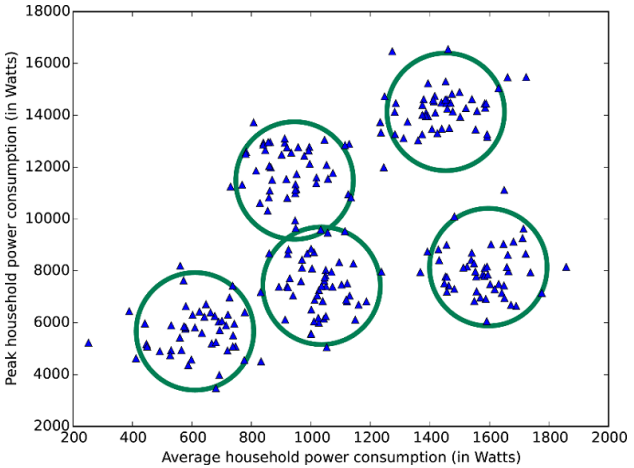
Données entrées?

- **Couche d'entrée:** les nœuds d'entrée comprennent des données et leurs probabilités.
- **Couche cachée:** c'est l'endroit où les différentes probabilités des entrées sont attribuées aux poids
- **Couche de sortie:** Les nœuds de sortie représentent des valeurs d'attribut prévisibles pour le modèle d'exploration de données

Le miange sous SQL Server

Microsoft Clustering

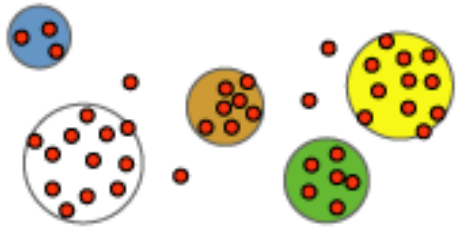
- C'est un algorithme de segmentation ou de cluster et d'exploration
- Itère sur les cas dans un ensemble de données pour les regrouper en clusters qui contiennent des caractéristiques similaires
- Ces groupements sont utiles pour
  - ✓ Explorer les données
  - ✓ Identifier les anomalies dans les données
  - ✓ Créer des prédictions



Le miange sous SQL Server

Comment sa fonctionne?

- L'algorithme identifie d'abord les relations dans un ensemble de données et génère une série de clusters basés sur ces relations
- Un diagramme de dispersion est un moyen utile de représenter visuellement les données
- Après avoir défini les clusters, l'algorithme calcule à quel point les grappes représentent les groupements des points et l'opération à nouveau jusqu'à arriver au seuil limite



Le miange sous SQL Server

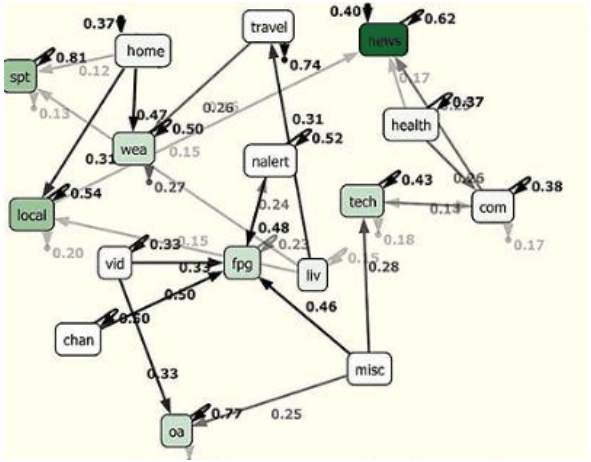
Données entrées?

- **La clé :** L'identifiant au niveau d'une entité non composée qui peut être numérique ou texte
- **Les colonnes d'entrée:** Les colonnes qui contiennent les données à prévoir
- **Les colonnes de prévision:** Elle est optionnelle cette fois ci car l'algorithme n'a pas besoin d'une colonne prévisible pour construire le modèle, la colonne de prévision doit être marquée comme **PredictOnly**

Le miange sous SQL Server

Microsoft Sequence Clustering

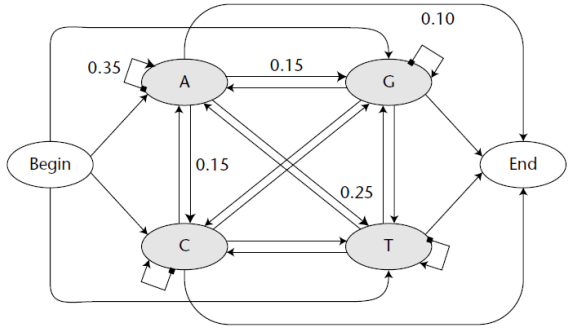
- Vous pouvez utiliser cet algorithme pour explorer des données contenant des événements pouvant être liés dans une séquence
- l'algorithme trouve les séquences les plus courantes, et effectue le regroupement pour trouver des séquences similaires
- Exemple les zones de click au niveau des pages d'un site web



Le miange sous SQL Server

Comment sa fonctionne?

- L'algorithme est hybride il combine les techniques de regroupement avec la chaîne de Markov
- Les données représentent généralement une série d'événements ou d'états transitions dans un ensemble de données
- L'algorithme examine toutes les probabilités de transition et mesure les différences entre toutes les séquences possibles



Transition probabilities:
$P(x_i = G|x_{i-1}=A) = 0.15$
$P(x_i = C|x_{i-1}=A) = 0.15$
$P(x_i = T|x_{i-1}=A) = 0.25$
$P(x_i = A|x_{i-1}=A) = 0.35$
$P(x_i = End|x_{i-1}=A) = 0.10$

Le miange sous SQL Server

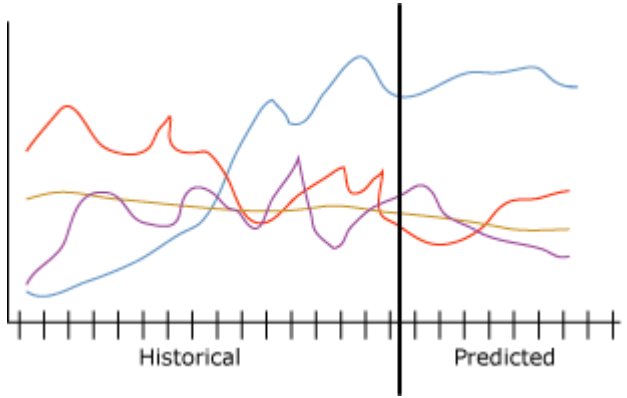
Données entrées?

- **La clé :** L'identifiant au niveau d'une entité non composée
- **Une colonne de séquence :** Cette colonne doit exister dans une table imbriquée, elle contient les identifiants des séquences triables
- **Des colonnes non séquentielles :** Elles sont optionnelles

Le miange sous SQL Server

Microsoft Time Series

- L'algorithme fournit plusieurs sous algorithmes optimisés pour la prévision de valeurs continues
- Tel que les ventes de produits, au fil du temps
- Une caractéristique importante de l'algorithme c’est qu'il peut effectuer une prédiction croisée.
- Les ventes observées d'un produit peuvent influer les ventes prévues d'un autre produit



Le miange sous SQL Server

Comment sa fonctionne?

- L'algorithme se base sur deux principaux algorithmes ARTXP pour l'analyse des flux de données rapides et ARIMA pour l'analyse des flux de données moins rapides
- Il est possible d'utiliser une combinaison des deux algorithmes
- Un seul modèle peut contenir des ventes de plusieurs produits, à condition qu'il n'y ait qu'un enregistrement de nom de produit unique pour chaque tranche de temps

Le miange sous SQL Server

Données entrées?

- **La clé :** L'identifiant au niveau d'une entité numérique ou date seulement
- **Une colonne de prévision:** Cette colonne doit contenir des valeurs continues , la colonne doit contenir des valeurs continues et les valeurs doivent être uniques pour chaque série
- **Des colonnes clés d'indentification optionnelles :** Elles sont optionnelles, elles doivent contenir des données uniques

Le miange sous SQL Server

Classification	Régression	Ségmentation	Prévision	Association
• Decision Trees • Logistic Regression • Naïve Bayes • Neural Networks	• Decision Trees • Linear Regression • Logistic Regression • Neural Networks	• Clustering • Sequencial clustering	• Time Series	• Association Rules • Decision Trees

Le miange sous SQL Server

Les méthodes d'applications

- **SSDT:** L'outil offre déjà un environnement visuel confortable pour créer des structure de minage
- **DMX:** Il est possible de créer des model de minage à travers un langage très proche du SQL
- **ADOMD.Net:** Cet API permet d'intégrer les minage dans les applications personnalisées
- **Les plugins:** Comme les plugin Excel

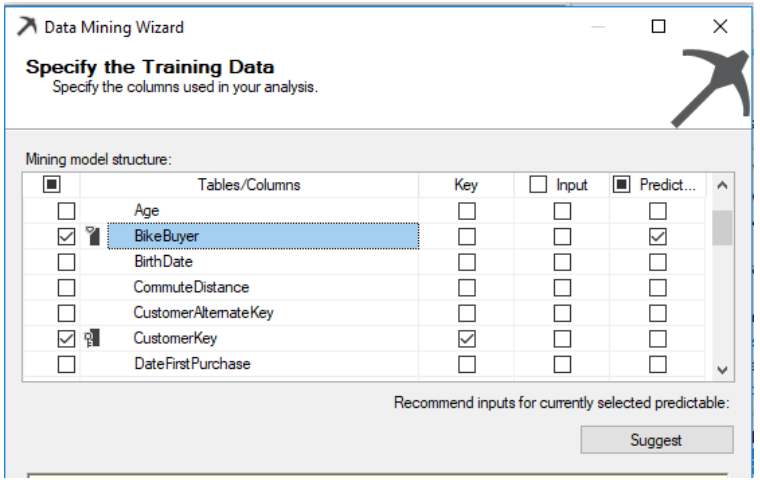
```
/*Exemple de structure de requête
DMX*/
SELECT FLATTENED TOP <colonnes>
FROM <model>
PREDICTION JOIN <table>
ON <mappage>
WHERE <filter>
ORDER BY <expression>
```

Lab:Structure de minage de classification sous SSDT

Plusieurs dimensions et une seule meusure

Création de la structure de minge

- Création de structure de minage
- Nous utilisons [AdventureWorkDW2012][vTargetMail] comme table d'entrée
- Nous spécifions CustomerKet comme clé et BikeBuyer comme colone de prévision

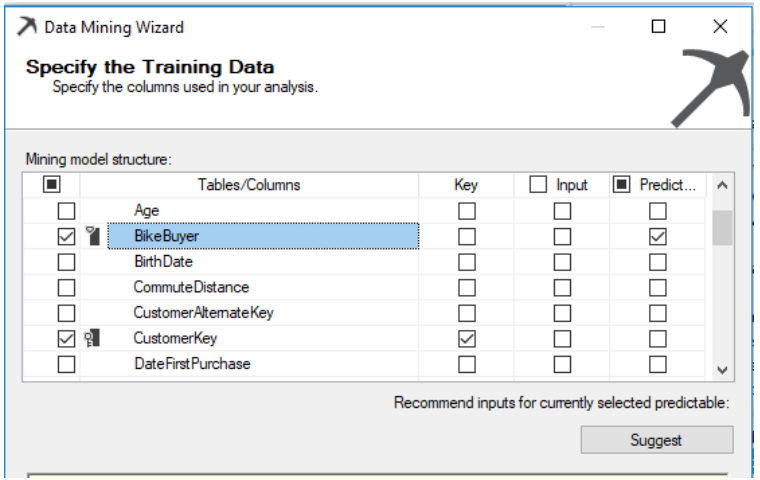


Plusieurs dimensions et une seule meusure

Création de la structure de minge

- ❑ Nous cliquons sur le bouton suggest pour voir les suggestions d'entrées

- ❑ Nous choisissons comme entrées
- ✓ Age
- ✓ Nombre de voitures
- ✓ Nombre d'enfants
- ✓ Niveau
- ✓ Salaire annuel
- ✓ Région

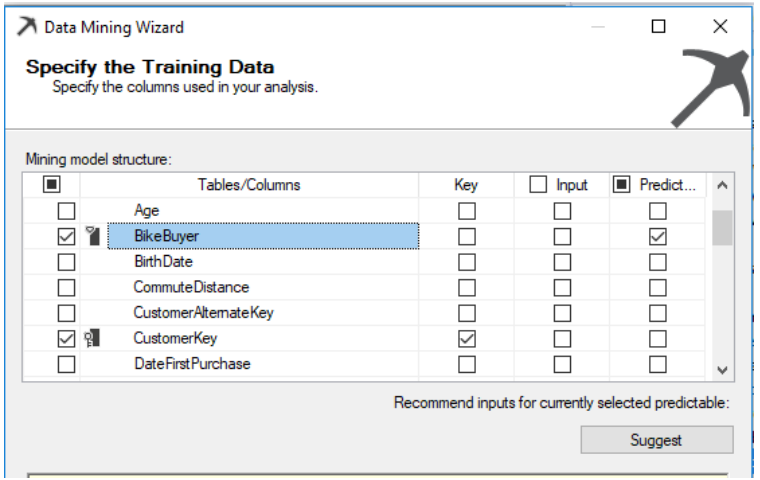


Plusieurs dimensions et une seule meusure

Création de la structure de minge

- ❑ Nous cliquons sur le bouton suggest pour voir les suggestions d'entrées

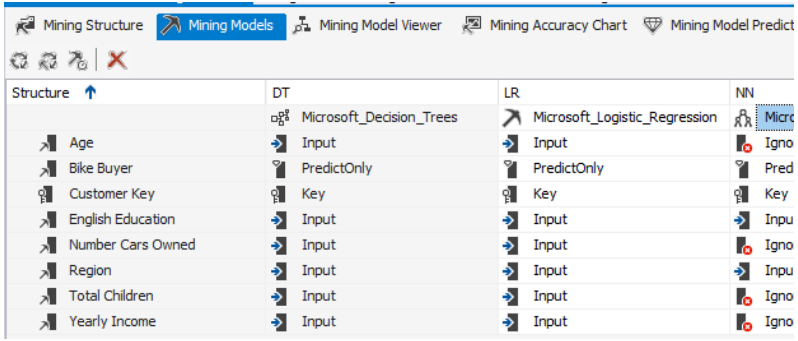
- ❑ Nous choisissons comme entrées
- ✓ Age
- ✓ Nombre de voitures
- ✓ Nombre d'enfants
- ✓ Niveau
- ✓ Salaire annuel
- ✓ Région



Plusieurs dimensions et une seule meusure

Création de la structure de minge

- ❑ Nous explorons les données un fois le model est crée
- ❑ Nous ajoutons les models de minage à partir de l'onglet **Mining Models**
  - ✓ Decision Tree
  - ✓ Logistic Regression
  - ✓ Neutral Network

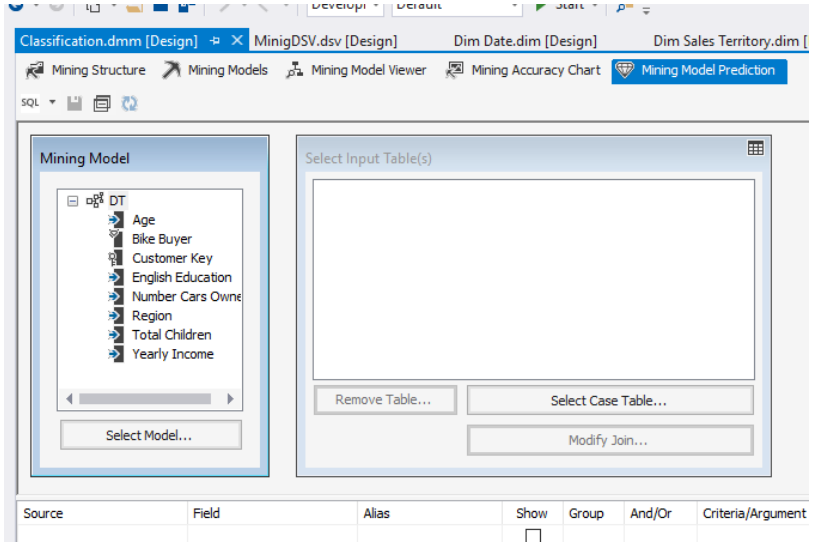


Plusieurs dimensions et une seule meusure

Création de la structure de minge

- Au niveau de l'onglet **Minig Model Viewer** nous essayons d'explorer les divers model ajoutés
- Essayons l'onglet **Dependency Network** sous **Minig Model Viewer**

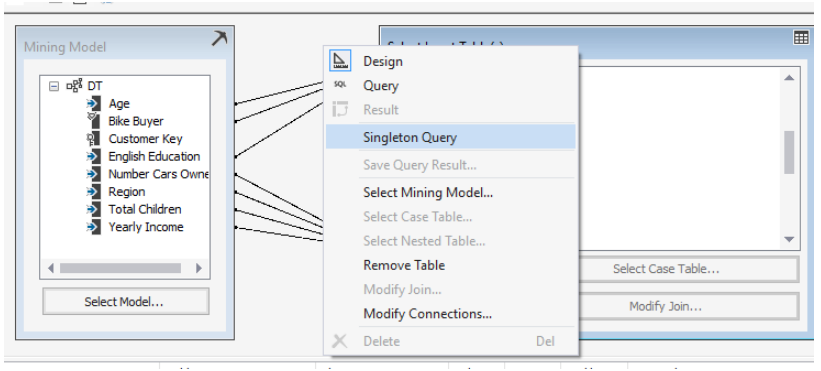
- Allons vers L'onglet **Mining Model Prediction**



Plusieurs dimensions et une seule meusure

Création de la structure de minge

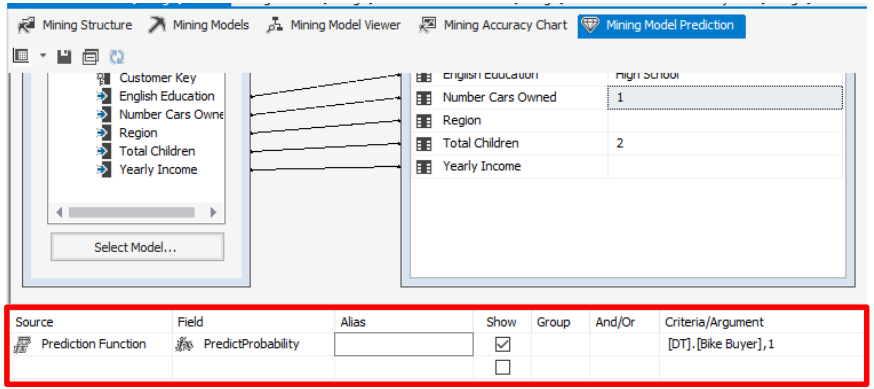
- Allons vers L'onglet **Mining Model Prediction** essayons
- Essyons de selectionner le model decision tree
- Choisissons la table **vTargetMail** à droite
- Il est possible de raisonner niveau table ou niveau cas singulier



Plusieurs dimensions et une seule meusure

Création de la structure de minge

- Choisissons le cas de prévision singulier
- Avec les paramètres suivants

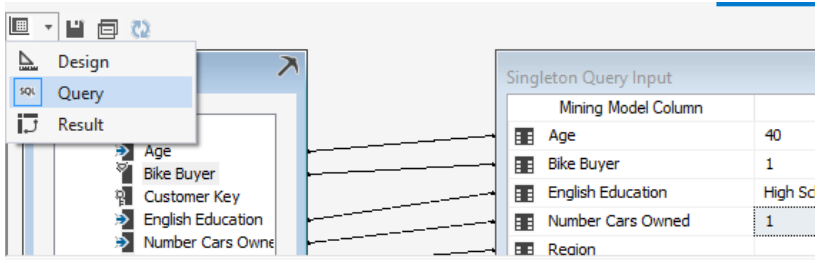


Plusieurs dimensions et une seule meusure

Création de la structure de minge

- Choisissons l'option Query pour voir la requête en DMX
- Choisissons le résultat pour voir la prévision en pourcentage

- Revenons vers le raisonnement niveau table et répetons les configurations



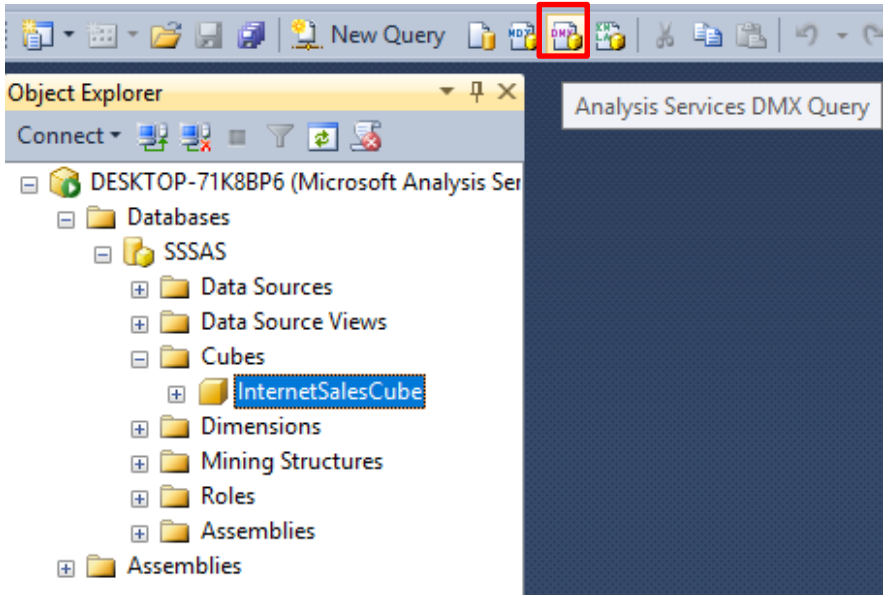

```
SELECT
  (PredictProbability([DT].[Bike Buyer],1)) as [Achat]
From
  [DT]
NATURAL PREDICTION JOIN
(SELECT 40 AS [Age],
  1 AS [Bike Buyer],
  'High School' AS [English Education],
  1 AS [Number Cars Owned],
  2 AS [Total Children]) AS t
```

Lab:Le minage sous SSMS

Les dimensions usage

Le mode d'emploi

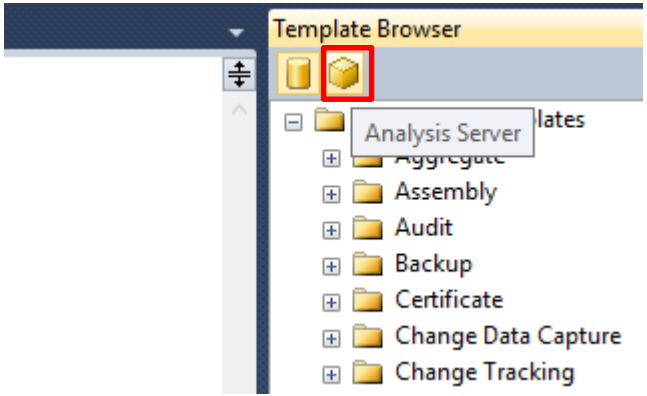
- Nous créons une nouvelle requête DMX sous SSMS
- Cliquons **Template Explorer** sous le menu **View**



Les dimensions usage

Le mode d'emploi

- Cliquons **Analysis Server** pour avoir les divers templates DMX
- Choisissons **Base prediction** comme template
- Essayons de reconstruire la requête précédment faite au niveau de **SSDT**



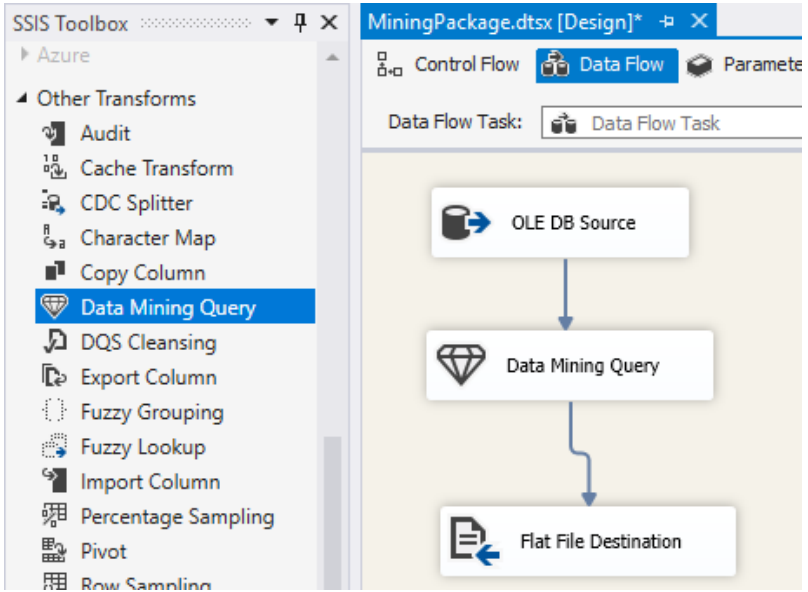
“

Lab:Le minage avec SSIS

Les dimensions usage

Le mode d'emploi

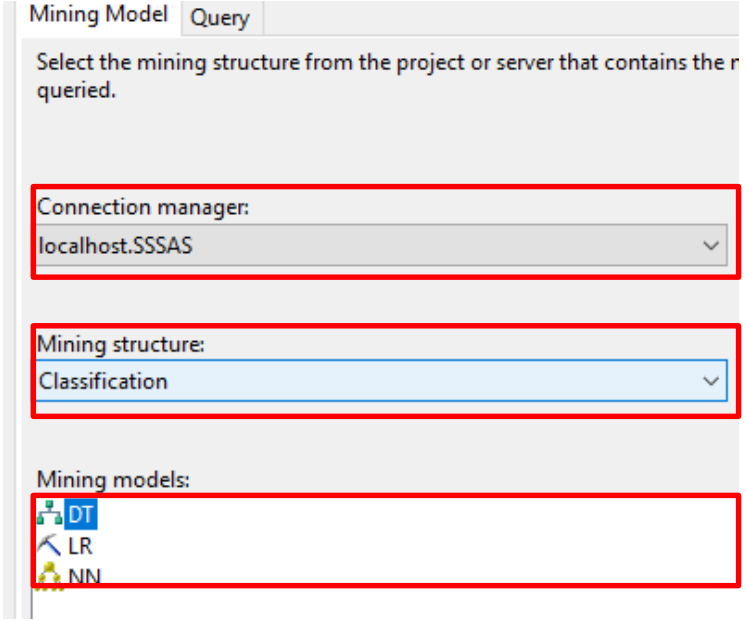
- Il est possible de miner sous SSIS avec le Task **Data Mining Query**



Les dimensions usage

Le mode d'emploi

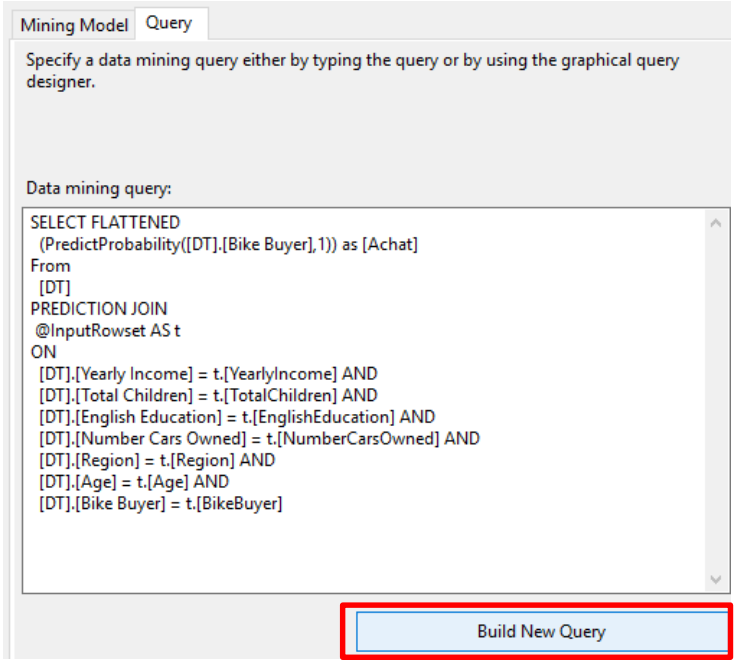
- Au niveau de l'onglet **Mining Model** parametrons le connection manager
- Spécifions la structure de minage
- Choisissons le model de minage



Les dimensions usage

Le mode d'emploi

- Construisons la requête de minage avec DMX ou avec l'assistant en cliquant le bouton **Build New Query**



Les dimensions usage

Le mode d'emploi

- Plaçons un Data Viewer pour visualiser les données au niveau du package SSIS