

Projet Data Mining

Contexte du Projet

Vous avez été contacté par un concessionnaire automobile afin de l'aider à mieux cibler les véhicules susceptibles d'intéresser ses clients. Pour cela il met à votre disposition :

- Son catalogue de véhicules
- Son fichier clients concernant les achats de l'année en cours
- Un accès à toutes les informations sur les immatriculations effectuées cette année
- Une brève documentation des données
- Un vendeur (voir son interview ci-dessous)

Votre client sera satisfait si vous lui proposez un moyen afin :

- Qu'un vendeur puisse en quelques secondes évaluer le type de véhicule le plus susceptible d'intéresser des clients qui se présentent dans la concession
- Qu'il puisse envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sélectionnés par son service marketing.

Documentation des Données

Les fichiers de données à votre disposition vous sont décrits dans les tables ci-dessous. Pour chaque attribut du fichier, vous sont donnés son nom, son type (numérique, caractères, catégoriel ou booléen) sa description et son domaine de valeurs.

Certains attributs peuvent comporter des valeurs manquantes ou incorrectes (erreur de saisie par exemple). Celles-ci sont représentées par une cellule vide ou bien contenant une valeur hors du domaine de valeurs de la variable (valeurs « ? », « » ou « N/D » par exemple).

Informations Données par le Concessionnaire

L'interview du gestionnaire de la concession automobile nous a permis de définir le contexte et les objectifs suivants :

« Les différents véhicules de notre catalogue répondent à des besoins différents. Certains sont petits afin de mieux circuler en ville, d'autres ont de l'espace pour transporter toute une famille tandis que certains sont plus puissants et destinés à une clientèle plus fortunée. Nous souhaitons définir différentes catégories de véhicules afin de mieux comprendre les désirs des clients et proposer aux nouveaux clients le véhicule le plus adapté à leurs besoins. »

Immatriculations.csv : informations sur les immatriculations effectuées cette année

Attribut	Type	Description	Domaine de valeurs
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA 99 »
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundaï, Jaguar, Kia, Lancia, Mercedes, Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beetle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance		Puissance en chevaux Din	[55, 507]
Longueur		Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d'occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

Catalogue.csv : catalogue de véhicules

Attribut	Type	Description	Domaine de valeurs
Marque	caractères	Nom de la marque du véhicule	Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundai, Jaguar, Kia, Lancia, Mercedes, Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo
Nom	caractères	Nom du modèle de véhicule	S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beetle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4
Puissance	numérique	Puissance en chevaux Din	[55, 507]
Longueur	catégoriel	Catégorie de longueur	courte, moyenne, longue, très longue
NbPlaces	numérique	Nombre de places	[5, 7]
NbPortes	numérique	Nombre de portes	[3, 5]
Couleur	catégoriel	Couleur	blanc, bleu, gris, noir, rouge
Occasion	booléen	Véhicule d'occasion ?	true, false
Prix	numérique	Prix de vente en euros	[7500, 101300]

Clients.csv : fichier clients concernant les achats de l'année en cours

Attribut	Type	Description	Domaine de valeurs
Age	numérique	Age en années du clients	[18, 84]
Sexe	catégoriel	Genre de la personne	M, F
Taux	numérique	Capacité d'endettement du client en euros (30% du salaire)	[544, 74185]
SituationFamiliale	catégoriel	Situation familiale du client	Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule
NbEnfantsAcharge	numérique	Nombre d'enfants à charge	[0, 4]
2eme voiture	booléen	Le client possède déjà un véhicule principal ?	true, false
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA99 »

Marketing.csv : clients sélectionnés par le service marketing

Attribut	Type	Description	Domaine de valeurs
Age	numérique	Age en années du clients	[18, 84]
Sexe	catégoriel	Genre de la personne	M, F
Taux	numérique	Capacité d'endettement du client en euros (30% du salaire)	[544, 74185]
SituationFamiliale	catégoriel	Situation familiale du client	Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule
NbEnfantsAcharge	numérique	Nombre d'enfants à charge	[0, 4]
2eme voiture	booléen	Le client possède déjà un véhicule principal ?	true, false

Voici quelques étapes à suivre :

1. Analyse exploratoire des données :

L'analyse exploratoire des données vous permettra d'identifier d'éventuels problèmes dans les données (valeurs incohérentes, valeurs manquantes, etc.) et découvrir d'éventuelles propriétés de l'espace des données (valeurs doublons, variables liées, variables d'importance particulière ou bien inutiles, etc.).

Appliquez pour cela différentes méthodes d'analyse des données (statistiques descriptives, histogrammes, nuages de points, etc.).

2. Identification des catégories de véhicules :

Vous devez à partir des informations du Catalogue identifier des catégories de véhicules (citadine, routière, sportive, etc.) en fonction de leur taille, puissance, prix, etc. Ces catégories doivent correspondre à divers besoins de la part des clients (une grande voiture pour les familles nombreuses, une petite voiture pour circuler en ville, etc.).

Ces catégories de véhicules constitueront les classes à prédire durant les étapes suivantes du processus.

3. Application des catégories de véhicules définies aux données des Immatriculations :

Les données d'Immatriculations contiennent les informations sur les véhicules vendus cette année. L'objectif est d'attribuer à chacun de ces véhicules la catégorie qui lui correspond en utilisant le modèle définissant les catégories de véhicules généré précédemment.

4. Fusion des données Clients et Immatriculations :

Les données Clients contiennent les informations sur les clients ayant les véhicules vendus cette année. L'objectif est de faire la fusion entre les données des Clients et des Immatriculations afin d'obtenir sur une même ligne l'ensemble des informations sur le client (âge, sexe, etc.) et sur le véhicule qu'il a acheté (avec sa catégorie).

Cet ensemble de données servira lors des étapes suivantes pour l'apprentissage de la catégorie de véhicules (variable cible) la plus adaptée à un client selon ses caractéristiques (variables prédictives).

5. Création d'un modèle de classification supervisée pour la prédiction de la catégorie de véhicules :

L'objectif de cette étape est de créer à partir du résultat de la fusion précédente un classifieur (modèle de classification supervisée) permettant d'associer aux caractéristiques des clients (âge, sexe, etc.) une catégorie de véhicules.

Testez les différentes approches et algorithmes (arbres de décision, random forests, support vector machines, réseaux de neurones, deep learning, etc.), avec pour chaque algorithme plusieurs paramétrages testés, afin d'obtenir un classifieur aussi performant que possible.

L'évaluation et la comparaison des performances de chaque configuration algorithmique (un algorithme et un paramétrage spécifiques) testée sera réalisée grâce aux matrices de confusion et mesures d'évaluation calculées à partir des résultats des tests des classifieurs.

6. Application du modèle de prédiction aux données Marketing :

Les données Marketing contiennent les informations sur les clients pour lesquels on souhaite prédire une catégorie de véhicules.

L'objectif est de prédire pour chacun de ces clients la catégorie de véhicules qui lui correspond le mieux en utilisant le classifieur généré durant l'étape précédente.