esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

Projet Machine Learning

Réalisé par:
Racha Maghrebi
Nada Belaidi
Khaled Charaabi
Abir Lakhal
Fares Khechana.

Plan:

01

Introduction et problématique étudiée

Le terme « churn », très utilisé en marketing, désigne une tendance à la baisse de la clientèle. Cette notion, aussi connue en français sous le terme « d'attrition », permet d'analyser la fidélité d'une clientèle et l'impact d'actions marketing sur celle-ci.
En prévoyant le taux de désabonnement des clients, les entreprises peuvent immédiatement prendre des mesures.
La prédiction peut être effectuée en analysant les données des clients à l'aide de techniques d'exploration.

On propose dans notre projet une prédiction du taux de désabonnement des clients en utilisant:

Algorithmes	**Méthodes de sélection de caractéristiques**
-KNN. -CART. -SVM. -Régression logistique. -Naive Bayes. -Random Forest.	-SFS. -SBS. -SFFS. -SFBS. -Sélection par corrélation de Pearson. -Manipulation des données:fonction get_dummies.

02

Préparation des données

I. Compréhension des données :

-Importation et visualisation de la dataset.

(dataset : 22 caractères ,7043 valeurs)

-Test sur l’existence des valeurs nulles ou bien manquantes.

(pas des valeurs nulles)

-Représentation de quelques caractères.

II. Nettoyage des données:

-Imputation en cas de valeurs manquantes ou nulles.

→ Absence de valeurs manquantes ⇒ Nous n'avons pas effectué une imputation

-conversion numérique de tous les caractères pour faciliter le traitement des algorithmes .

→ On a converti des colonnes comme "Partner", "Monthly Charges", "Contract", etc..

-Remplacer les cases vides (espace) par des valeurs nulles.

→ L'une des valeurs de la colonne "Total Charges" est un espace.

III. Feature selection :

Feature Selection est un processus dans lequel vous sélectionnez automatiquement les caractéristiques de nos données qui contribuent le plus à la variable de prédiction ou à la sortie qui nous intéresse.

Minimiser la perte d'information venant de la suppression de toutes les autres variables.

Avantages:

- Réduit le surajustement
- Améliore la précision
- Réduit le temps d'entraînement (Training)

méthodes proposés par l'article	amélioration
-SFS -SBS -SFFS -SFBS	-Sélection par corrélation de Pearson -Manipulation des données:fonction get_dummies

VI. transformation de données :

- NORMALISATION:

  -MinMaxScaler:

  Transformer les caractéristiques en mettant chaque caractéristique à l'échelle d'une plage donnée.

  Cet estimateur met à l'échelle et traduit chaque caractéristique individuellement de manière à ce qu'elle se trouve dans la plage donnée sur l'ensemble de formation, par exemple entre zéro et un.

V. Réduction de dimensionnalité :

--> On a appliqué l'ACP (Analyse en composantes principales)

-principe de l'ACP:

L'ACP consiste à remplacer une famille de variables par de nouvelles variables de variance maximale, non corrélées deux à deux et qui sont des combinaisons linéaires des variables d'origine.

03

Travail effectué et Résultats Obtenus

Corrélation Pearson:

Une corrélation est une relation qu'il y a entre différentes variables. Il est important d'identifier les interdépendances entre variables avant l'élaboration du modèle. Car cela peut vous guider dans le choix du modèle . Voici ce que cette méthode a donné avec nos 6 modèles.

algorithme	Training score	Testing score	score
SVM	0.74	0.73	0.73
Naive Bayes	0.90	0.49	0.72

SFS :

"Sequential Feature Selector" trouve le meilleur sous-ensemble de caractéristiques en ajoutant une caractéristique qui améliore le mieux le modèle à chaque itération.
La mesure à utiliser pour évaluer le classificateur: Precision(Accuracy)/score(roc_auc).

SFS (AUC)

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM	0.80	**0.80**	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9494929006085193	0.7723615712257453	0.7723615712257453
regression logistique	0.7943204868154158	0.8064363464268812	0.8069096071935636

SFS(Accuracy):

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM	0.80	0.80	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9466531440162271	0.7742546142924751	0.7665314401622718
regression logistique	0.7943204868154158	0.8064363464268812	0.8069096071935636
Naive Bayes	0.7180527383 36714	0.7425461429247515	0.7425461429247515

SBS:

SBS (Sequential Backward Selection)est le contraire de SFS. SBS commence avec toutes les caractéristiques et supprime la caractéristique qui a le moins d'importance pour le modèle à chaque itération.

SBS (AUC)

algorithme	training	testing	score
KNN	0.79	0.81	0.8078561287269286
SVM	0.78	0.80	0.8035967818267865
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.949290060851927	0.7723615712257453	0.7723615712257453
regression logistique	0.7945233265720081	0.8078561287269286	0.8078561287269286

SBS(Accuracy)

algorithme	training	testing	score
KNN	0.79	0.81	0.8054898248935163
SVM	0.78	0.80	0.8035967818267865
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9501014198782961	0.780407004259347	0.780407004259347
regression logistique	0.7945233265720081	0.8078561287269286	0.8078561287269286
Naive Bayes	0.718052738336714	0.7425461429247515	0.7425461429247515

SFFS :

La sélection directe séquentielle flottante (SFFS) démarre à partir de l'ensemble vide. Après chaque pas en avant, SFFS effectue des pas en arrière tant que la fonction objective augmente.

SFFS(Auc)

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM(rbf)	0.80	0.80	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9486815415821501	0.7756743965925225	0.7756743965925225
régression logistique	0.7943204868154158	0.8064363464268812	0.8064363464268812

SFFS(Accuracy)

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM (rbfl)	0.80	0.80	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.945841784989858	0.7714150496923805	0.7714150496923805
régression logistique(l1 penalty)	0.7949290060851927	0.8069096071935636	0.8069096071935636
Naive Bayes	0.90	0.50	0.7425461429247515

SBFS:

La sélection séquentielle flottante vers l'arrière (SFBS) commence à partir de l'ensemble complet.

SBFS(Auc)

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM	0.80	0.70	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9466531440162271	0.7756743965925225	0.7756743965925225
regression logistique	0.7949290060851927	0.8064363464268812	0.8017037387600567
Naive Bayes	0.90	0.50	0.7425461429247515

SBFS(Accuracy)

algorithme	training	testing	score
KNN	0.80	0.81	0.8054898248935163
SVM(rbf)	0.80	0.80	0.8017037387600567
CART	0.7845841784989858	0.8035967818267865	0.8035967818267865
Random forest	0.9498985801217038	0.77520113582584	0.77520113582584
regression logistique	0.7949290060851927	0.8064363464268812	0.8064363464268812
Naive Bayes	0.90	0.50	0.7425461429247515

Dummies:

est utilisé pour la manipulation des données. Elle convertit les données catégorielles en variables factices ou en variables indicatrices.

algorithme	training	testing	score
KNN	0.79	0.79	0.78
SVM(Gaussian Kernel)	0.81	0.81	0.80
CART	0.78	0.79	0.80
Random forest	Accuracy score : 0.80	ROC-AUC score : 0.67	0.80
régression logistique	0.8006085192697768	0.8064363464268812	0.8102224325603408
Naive Bayes	0.50	0.80	0.73

04

Perspectives sur le travail effectué

Basé sur l’expérience, le résultat montre que l’algorithme KNN fonctionne bien avec le modèle SBS par rapport aux autres avec la précision de l’entraînement 0.81 et des tests 0.80 et un score général égale à 0.81.

05

Apports académiques

Ce projet a été très enrichissant pour nous, car il nous a permis de découvrir le domaine de l'analyse décisive , ses acteurs et ses contraintes. Il nous a permis de participer concrètement à ses enjeux au travers une missions réaliste et nous avons ainsi pu mettre en pratique tout ce que nous avons appris en classe et comprendre l'importance de la data science dans le domaine du marketing et du business.

Merci pour votre attention.