Kutay & Elena

RAPPORT

Étude - Estimer le coût de la couverture médicale d'un.e américain.e

---

Simplon Microsoft IA

STRASBOURG

26 janvier 2021

Sommaire

Contexte et objectif

Une nouvelle compagnie d'assurance maladie souhaite proposer une formule personnalisée à ses futurs.es clients.es. Un fichier “csv” contenant des données nécessaire à l’étude a été mis à notre disposition. Afin d'établir son business model, la compagnie doit être en mesure d'estimer les frais médicaux facturés par l'assurance santé pour ses prospects.

Elle fait appel à votre start-up qui développe des solutions en IA pour développer un modèle de machine learning capable de prédire les frais médicaux de ses prospects. La compagnie d'assurance fournit à votre start-up un historique des dépenses en frais de santé.

L’bjectif c’est de prédire les frais médicaux individuels des américains en se basant sur des critères de base (âge, sexe, bmi, nombre d'enfants, fumeur, région)

Exploration des données

Nous disposons d’un jeu de données comportant des données catégorique : sex, smoker et régions. Nous connaissons les âges des personnes, le bmi (l’indice corporel), les smokers (les fumeurs et les non-fumeurs) , la région ainsi que les charges.

Pour notre analyse nous utiliserons **Google Colaboratory** comme outil pour le traitement des données, mais aussi pour faciliter le travail collaboratif.

Nous connaissons la sortie (les charges) et les variables indépendantes à prendre en compte (le reste des colonnes). Les valeurs sont catégorielles et quantitatives . Nous devons appliquer l'encodage sur l'ensemble de données, car il y a des mots présents : pour les colonnes "sexe", "fumeur" et "region".

Construction du Modèle

Au vue de toutes ses informations, nous choisissons d'utiliser un modèle de machine learning supervisé : la régression linéaire multiple, le rendom forest, l'arbre de décision. Le jeux de données ne comportait pas des valeurs manquantes.

Comme le temps alloué à l'analyse n'était pas suffisant, on a créé des modèles basiques pour avoir un aperçu sur la prédiction.

Avant de créer le modèle, on analyse quelles sont les variables explicatives dans notre jeux de données et qui est notre variable "expliquée" ou "target" , quelles sont les variables à utiliser pour le modèle et qui sont celles qui ont une forte relation linéaire avec la variable " charges".

- La variable y ou "target": **charges**; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante
- Les variables "sex", "smoker" et "région" sont des variables explicatives et catégorielles.
- Les variables X ou "features": **age**, **sex**, **bmi**, **children**, **smoker**, **region**. Ce sont des variables explicatives, appelées également "régresseurs" ou "variables indépendantes".

Matrice de corrélation

En premier lieu, afin de trouver quelles sont les variables qui ont une forte relation linéaire avec la variable "target", on fait une matrice de corrélation. Le rôle de la matrice de corrélation est d'indiquer les valeurs de corrélation en mesurant le degré de relation entre les variables.

Les coefficients de corrélation se situent dans l'intervalle [-1,1].

- si le coefficient est proche de 1 c'est qu'il y a une forte corrélation positive
- si le coefficient est proche de -1 c'est qu'il y a une forte corrélation négative
- si le coefficient est proche de 0 en valeur absolue c'est qu'il y a une

Boîte à moustache

On a utilisé le box plot (appelé aussi une boîte à moustaches). C'est un graphique utilisé fréquemment pour l'exploration des données. Il permet de visualiser, pour une variable ou pour un groupe d'individus, le comportement global des individus. Il reflète à la fois la tendance centrale et une idée de la distribution des données. Ce graphique est surtout intéressant lorsqu'on essaye de comparer des variables sur des échelles similaires ou lorsqu'on essaye de comparer des groupes d'individus sur la même variable.

1. Multiple Linear Regression

- ❏ Entraînement de la régression linéaire multiple
- ❏ Prédire les résultats de l'ensemble de test
- ❏ Évaluation des performances du modèle
- ❏ Obtenir l'équation de régression linéaire finale avec les valeurs des coefficients

L'analyse par **régression linéaire multiple** est une des solutions qui existe **pour** observer les liens entre une variable quantitative dépendante et **n** variables (quantitatives) indépendantes. Dans notre cas, la variable indépendante ou " target" sera les "charges", les autres valeurs seront les variables dépendantes. Traitement sur les variables catégorielles afin qu'elles puissent être prises en compte dans l'analyse.

1. Decision Tree Regression

- ❏ Splitting the dataset into the Training set and Test set
- ❏ Training the Decision Tree Regression model on the Training set
- ❏ Predicting the Test set results
- ❏ Evaluating the Model Performance

2. Random Forest Regression

- ❏ Splitting the dataset into the Training set and Test set
- ❏ Training the Random Forest Regression model on the whole dataset
- ❏ Predicting the Test set results
- ❏ Evaluating the Model Performance

Conclusion

Dans notre étude, on a utilisé plusieurs modèles et méthodes pour augmenter le résultat. On a utilisé des arbres de décision

On s'est confronté à un manque de temps dans notre démarche, ce qui explique l'absence de certains aspects importants dans le projet : on n'a pas eu le temps pour examiner le Random Forest

Conformément aux résultats, on a pu constater que le modèle random-forest présentait les meilleurs résultats - le plus optimal pour prédire le modèle de ML. Pour les features engineering sur l'arbre de décision, on a observé une meilleure augmentation au niveau du résultat.