PROMOTION GAIA, STRASBOURG
RODRIGUE ULUA & MORGANE HENTZEN-SOUCHET

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

Etude réalisée pour une nouvelle compagnie d'assurance maladie

—

27.01.2021

Contexte du projet

Une nouvelle compagnie d'assurance maladie souhaite proposer une formule personnalisée à ses futurs.es clients.es. Elle se situe aux **États-Unis**.

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 notre start-up pour développer **un modèle de machine learning capable de prédire les frais médicaux** de ses prospects.

La compagnie d'assurance nous fournit un jeu de données avec les informations suivantes sur des **individus américains** : l'**âge** (age), le **sexe** (sex), l'**indice de masse corporelle** (IMC en français, BMI en anglais), le **nombre d'enfants** (children), si **la personne fume ou non** (smoker), la **localisation aux États-Unis** avec un découpage en quatre grandes régions (region). La dernière donnée mise à notre disposition est le **coût de la couverture médicale** (charges) pour chaque individu.

Objectif du projet

Nous devons trouver **le meilleur modèle possible de machine learning** pour prédire **le coût de la couverture médicale** pour chaque américain.e en fonction d'une ou de plusieurs variables explicatives (les *features* : âge, sexe, IMC, nombre d'enfants, fumeur ou non et la localisation).

Outils

Nous avons travaillé avec le **langage Python** sur le notebook partagé de Jetbrains, Datalore. Nous avons également effectué quelques comparaisons de valeurs obtenues avec **Gretl**. En complément, nous avons aussi utilisé **Google Doc** pour le présent document et **Canvas** pour le powerpoint.

Limites de notre étude

Notre jeu de données est concerné par **quelques limites** que nous devons souligner :

- Les données concernent des individus américains et elles peuvent donc être sujettes à des biais de notre part.
- Nous pouvons trouver des valeurs aberrantes dans notre jeu de données dont il faut se méfier lors de l'interprétation.

Phase exploratoire du jeu de données

Pour nous aider à comprendre notre jeu de données, nous avons effectué des observations sous **une forme brute** : pour vérifier notamment l'absence de valeurs nulles - dites NaN - mais également pour obtenir un premier aperçu de la répartition des données avec des statistiques descriptives (*.describe()* par exemple).

Nous avons également généré des observations sous **la forme de graphiques** avec la visualisation des valeurs corrélées et une heatmap.

Cette première approche a été agrémentée d'un encodage des **variables catégorielles** (*sex*, *smoker*, *region*) en **variables numériques** pour pouvoir **les explorer** et **les exploite**r au même titre que les autres features.

La phase exploratoire nous a aidé à avoir une idée de l'ensemble de nos données et à développer une première hypothèse que nous gardons précieusement de côté : il se pourrait que **le fait d'être fumeur ou non** soit la variable qui génère le plus de coût pour la couverture médicale.

Nous aurons confirmation ou infirmation par la suite.

Quel(s) modèle(s) pour répondre aux attentes de la compagnie d'assurance maladie ?

Ce qu'il est important de souligner : nous cherchons **une variable continue en sortie de notre modèle de machine learning**. C'est la raison pour laquelle nous nous dirigeons en premier vers **le modèle de la régression linéaire multiple**, puis vers **le modèle de l'arbre de**

décision avec une régression et enfin **le modèle du random forest toujours avec une régression**.

Nous allons explorer les résultats obtenus avec ces trois grands modèles, leurs **performances**, leurs **marges d'erreurs** et les différentes **prédictions** générées.

La régression linéaire multiple

Pour pouvoir générer le modèle de la régression linéaire multiple (*LinearRegression*), nous avons décidé de reprendre la méthode de la **Backward Elimination** pour choisir les features qui pourraient nous permettre de construire un modèle avec **une prédiction fiable** (avec une *p-value* en-dessous de 0.05 pour les variables explicatives) et une **bonne performance** (*R^2 ajusté*).

La méthode Backward Elimination

- Étape 1 : on choisit un seuil de significativité (SL). Ici, il est de **5% (0.05)**. C'est un seuil suffisant pour prédire un coût.
- Étape 2 : on construit le modèle avec **tous les prédicteurs** (les features ou variables explicatives) possibles. Ici, ce sont les features age, sex, bmi, children, smoker et region.
- Étape 3 : on **compare** les modèles entre eux dans le notebook. On considère le prédicteur ayant la plus grande p-value. Si cette **p-value est supérieure au seuil de significativité**, on passe à l'**étape 4**. Sinon, on passe à la fin car le modèle est prêt.
- Étape 4 : on **enlève** le prédicteur (la feature) qui dépasse le **seuil de significativité de la p-value**.
- Étape 5 : on répète l'opération 3 et 4 autant de fois que nécessaire, jusqu'à ce que la p-value corresponde à ce que l'on cherche.

Conclusion : nous avons décidé de réaliser **quatre modèles différents**.

Le **premier modèle** a été réalisé à partir du **train_set avec toutes les features**, puisque nous avons assez de données pour nous permettre de le faire. Nous remarquons que le R^2 ajusté y est de 73% (0.732) et que la p-value la plus élevée est celle de la feature sex à 56% (0.562).

Le **deuxième modèle** généré nous permet de faire une comparaison avec le premier : nous prenons cette fois-ci l'ensemble du jeu de données à partir du **df_cmed avec toutes les**

features. Le R^2 ajusté y est de 75% (0.750) et la p-value la plus élevée est celle de la feature sex à 69% (0.69)

Nous décidons finalement de partir sur la dataframe de départ pour garder une performance de modèle plus élevée.

En suivant la méthode de la Backward Elimination, nous nous séparons de la feature sex pour générer le **troisième modèle** avec df_cmed. Le R^2 ajusté est toujours de 75% (0.750) et il n'y a plus de p-value au-dessus de notre seuil de significativité. Tout au plus, la p-value la plus élevée est celle de la variable region avec 2% (0.020). Nous pouvons nous arrêter à ce troisième modèle, puisqu'il est le plus performant et le plus fiable.

Nous avons néanmoins tenté un dernier modèle, par curiosité, en enlevant la feature region. Ce quatrième modèle a perdu en R^2 ajusté, puisqu'il est à 74%. **Cette dernière tentative nous a convaincu de conserver le troisième modèle et, de ce fait, de ne pas garder la feature sex pour la suite.**

Prédiction obtenue avec le modèle

Maintenant que nous avons choisi les features les plus explicatives (age, bmi, children, smoker, region), nous réalisons une prédiction du coût de la couverture médicale d'un.e américain.e.

Elle est accessible dans le notebook.

Performance du modèle

La performance maximale que le modèle nous a donnée est de 75%. Ce qui nous a poussé à essayer d'autres modèles pour voir si c'est possible d'obtenir une meilleure performance que celle de la régression linéaire multiple.

L'arbre de régression

Nous passons au second modèle de machine learning envisagé pour notre étude : celui de l'arbre de régression (*DecisionTreeRegressor*).

Les paramètres de l'arbre de régression

Pour les paramètres de l'arbre de décision, nous avons décidé d'opter pour : **random_state** que nous avons initialisé à 0, **max_depth** que nous avons initialisé à 3.

La visualisation de l'arbre de régression

Nous avons réalisé une visualisation de l'arbre de régression dans le notebook !

Prédiction obtenue avec ce modèle

Maintenant que nous avons choisi les paramètres de l'arbre de régression et que nous connaissons déjà les features qui nous intéressent, nous réalisons une prédiction du coût de la couverture médicale d'un.e américain.e.

Elle est accessible dans le notebook.

Performance du modèle

La performance de l'arbre de décision est de 89,7%. Ce qui est déjà mieux que celle de la régression linéaire multiple.

Malgré cette performance, nous nous sommes décidé de tester quand même le random forest pour voir si on aurait une meilleure performance.

Le random forest

Nous passons au troisième et dernier modèle de machine learning envisagé pour notre étude : celui des forêts aléatoires de régression (*RandomForestRegressor*).

Les paramètres optimaux du random forest

Pour les paramètres du random forest, nous avons décidé d'opter pour : random_state inialisé à 0, min_samples_leaf initialisé à 10 et min_samples_split initialisé à 30.

Prédiction obtenue avec ce modèle

Maintenant que nous avons choisi les paramètres optimaux du random forest en minimisant au possible l'overfitting et que nous connaissons déjà les features qui nous intéressent, nous réalisons une prédiction du coût de la couverture médicale d'un.e américain.e.

Elle est accessible dans le notebook.

Performance du modèle

La performance du random forest est de 91,2%. Elle est bonne !

Conclusion

Nous avons donc modélisé le coût de la couverture médicale d'un.e américain avec trois modèles différents de machine learning.

Nous retenons que le modèle de machine learning le plus performant est celui **du random forest**. Il s'agit du modèle avec **le R^2 ajusté le plus élevé (91%)** des trois modèles sélectionnés.

Quant aux variables qui peuvent expliquer le coût de la couverture médicale d'un.e américain.e : **le fait d'être fumeur ou non** (colonne smoker) est ce qui influence le plus le coût. **L'âge** et **l'indice de masse corporelle** (respectivement les colonnes age et bmi) jouent également un faible rôle dans le coût de ladite couverture médicale.