RAPPORT

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

ANTOINE MEYER
ALEKSANDRINA STOYANOVA-CHRISTEN

CLIENT: THE CLIENT

VERSION 1

DATE : 26 JANVIER 2021

Contexte

Une nouvelle compagnie d'assurance maladie souhaite proposer une formule personnalisée à ses futurs.es clients.es. 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. La compagnie d'assurance a fourni à notre start-up un historique des dépenses en frais de santé.

Objectifs

L'objectif est de développer un modèle de machine learning capable de prédire les frais médicaux de ses prospects.

ANTOINE MEYER
ALEKSANDRINA STOYANOVA-CHRISTEN

CLIENT: THE CLIENT

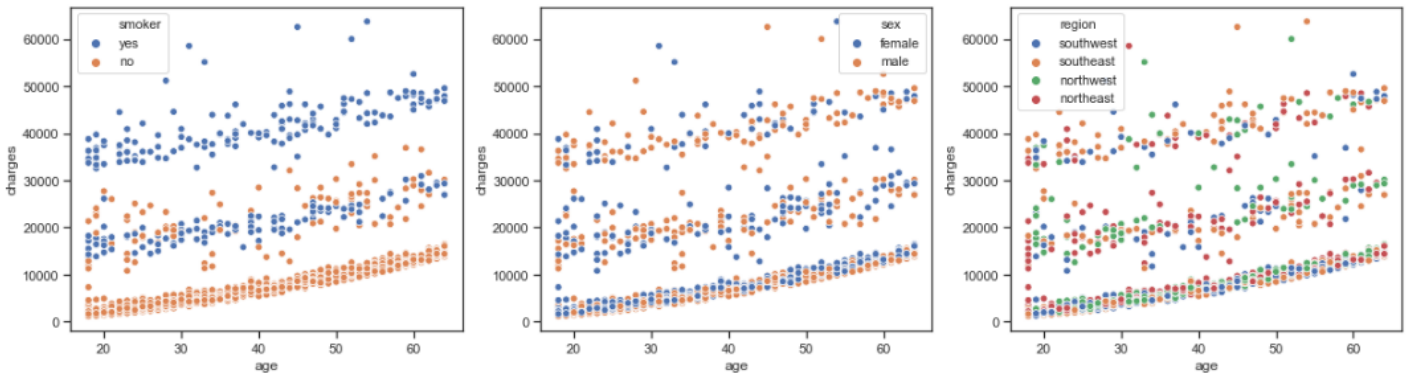
VERSION 1

DATE : 26 JANVIER 2021

Analyse des données

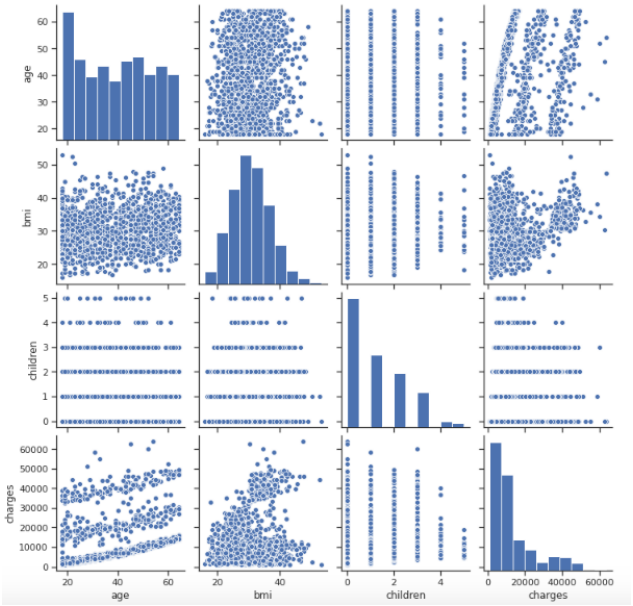
Le jeu de donnée contenu dans le lien **https://simplonline-v3-prod.s3.eu-west-3.amazonaws.com/media/file/csv/d51b9368-7437-420c-975c-81cac8790b68.csv** est un tableau comprenant **1338 lignes** disposées en **7 colonnes.** Ce jeu de donnée est **complet** (il ne possède aucune valeur non renseignée), et est composés de **4 variables quantitatives** (age/bmi/ children/charges) ainsi que de **3 variables qualitatives** (sex/smoker/region).

Le but sera donc de créer un modèle capable de déterminer la valeur de la variable **charges** à partir des 6 autres.



Distribution des frais en fonction de l'âge, pour chaque modalité des 3 variables qualitatives

Notre lecture graphique des différentes mise en relation entre variable nous a induit l'hypothèse que seul quelques variables avaient un impact significatif sur les frais. Nous avons ainsi procédé à l'établissement d'un premier modèle dans l'espoir de les déterminer, encodant au préalable les variables qualitatives en variable quantitative afin de pouvoir procéder, premièrement,
à une régression linéaire multiple.



Corrélation des différentes variables quantitatives

Construction des modèles

Le premier modèle que nous considérons est basé sur l'algorithme de la régression linéaire multiple. Après l'encodage des variables qualitatives **sex/smoker/region** (OneHotEncoding), le jeu de donnée est séparé au jeu d'entraînement et au jeu de test dans la proportion (70/30). Un premier modèle a été établit en utilisant toutes les variables explicatives **age/bmi/children/sex/smoker/region** dans la description de la variable à expliquée, **charges.** Ce modèle est caractérisé par une valeur du coefficient de la détermination ajusté (R^2-ajusté) égale à **0.762** et une précision de **53 %**.

L'analyse statistique descriptif a montré que certaines variables explicatives ne sont pas significatives statistiquement dans la description des **charges.** Afin d'améliorer notre modèle nous effectuons une sélection des variables explicatives en utilisant la technique de Backward Elimination qui nous permet d'écarter à chaque étape la variable explicative avec la plus grande p-valeur au-dessus un seuil de 0.05. À chaque étape cette variable est enlevée du jeu de donnée et le processus est répété en inspectant le R^2-ajusté ainsi que le critère d'information d'Akaike. Les itérations s'arrêtent quand les p-valeurs de toutes les variables explicatives restantes sont au-dessous de 0.05. En déroulant ce technique nous obtenons un modèle de régression linaire multiples exprimé par l'équation :

Charges = -12466.156 + 23347.627*__smoker_yes__ + 880.117*__region_northeast__ + +244.7259*__age__ + 340.0924*__bmi__ + 619.166*__children__.

Ce deuxième modèle est caractérisé par une valeur du coefficient de la détermination ajusté (R^2-ajusté) égale à **0.763** et une précision de **54 %** qui représentent une très faible amélioration en comparaison avec le premier modèle.

Modèle n°2 : Random Forest

La présence de variable à la fois quantitative et qualitative nous incite à utiliser un algorithme d'arbre de décision successifs. Nous avons ainsi séparés les données en deux échantillons **test/entrainement (30/70)** qui **conserve tout deux la fréquence de classe** au sein des valeurs qualitatives. Après entrainement, notre algorithme a déterminé avoir **une précision de 0,831** pour **une erreur d'en moyenne 11%** entre l'échantillon prédit et celui de test. Une fois le modèle créée, nous avons pu également observé l'impact des différentes variables sur la prédiction des frais. Il semblerait que seul 3 variables possèdent une influence majeur :

```
{'smoker_no': 0.7085556582840582,
 'bmi': 0.16329462128538194,
 'age': 0.11869483796720891,
 'children': 0.007895290227217017,
 'region_northeast': 0.000919772933792
 'sex_female': 0.00027696915058516176,
 'region_northwest': 0.000231715192166
 'region_southeast': 0.000131134959590
```

-"smoker" influe sur 70% du prix des frais
-"bmi" influe sur 16% du prix des frais
-"age" influe sur 11% du prix des frais
-------> 97% des frais sont expliqués

Optimisation

En passant par l'outil GridSearchCV nous avons pu optimiser notre modèle jusqu'à **une précision de 0.854** en sélectionnant des paramètres plus adaptés pour notre modèle. GridSearchCV nécessitant une grande puissance de calcul, nous nous sommes limités à deux hyper-paramètres : **n_estimators** et **max_depth.**

Ce gain de précision nous a permis d'éliminer les variables les moins influentes tout en revenant à une précision semblable à notre précision de départ (0.852), ce qui permet d'alléger le temps de calcul pour une qualité similaire.

Linear Regression vs Random Forest

Nous pouvons ainsi comparer le score de nos deux modèles et leur score respectif au gré des traitements que nous avons effectués au cours de cette étude.

La régression linéaire multiple nous suggère de conserver 5 variables (l, indiquée dans l'équation :

Charges = - 1.247e+04
+ 2.335e+04 * smoker (encodé)
+ 880.1172 * region (encodé)
+244.7259 * age
+340.0924 * bmi
+619.1660 * children.

Le score R^2- ajusté oscille autours de 0.762 et la précision en %, définit comme (100 - la moyenne des mean absolute percent error (mape)) est faible, 54%.

La RandomForest nous suggère de conserver 3 variables, indiquée comme suit :

- Smoker
- BMI
- age

Le score de précision oscille autours de 0.850.

Conclusion

La RandomForest semble être ainsi le modèle le plus performant pour prédire les frais.