Projet Data Science

Obésité et Couverture Santé aux Etats-Unis

Jedha Bootcamp

Khadija Ajimi

Août 2018

Introduction

Contrairement à la France, les Américains n'ont pas l'équivalent de notre Sécurité Sociale (française). L'immense majorité souscrit à des assurances privées qui couvrent les risques liés à leur santé. Ce qui représente pour chaque américain un budget conséquent au sein du ménage. C'est pourquoi, nous allons estimer le coût moyen en dépenses de santé par personne. De plus, l'état américain dépense chaque année 215 milliards de dollars dans le secteur de la santé uniquement en ce qui concerne l'obésité. Cependant, tout le monde n'en bénéficie pas forcément. Donc, nous allons vérifier si ce sont bien les américains obèses qui sollicitent le plus souvent une couverture santé.

1) Données statistiques observationnelles

Aux Etats-Unis, les factures médicales sont la première raison de banqueroute personnelle. A cela, s 'ajoutent de nouvelles données indiquant que la part d'Américains obèses a bondi de 6% en 10 ans. Selon une étude publiée le 23 mars 2018 dans le Journal of the American Medical Association (JAMA), près de 40% des Américains de plus de 20 ans étaient obèses en 2016 contre 34% en 2007.
Pour parvenir à ces résultats, les auteurs de l'étude - des chercheurs des Centres de contrôle et de prévention des maladies (CDC) - ont comparé les données de plus de 43.000 Américains recueillies via des questionnaires entre 2007 et 2008 par rapport à celles collectées entre 2015 et 2016.

2) Données économiques

Outre, l'étude (mentionnée ci-dessus) menée par les chercheurs, les autorités publiques se sont penchées sur le problème et ont constaté que l'obésité est associée à une sur-mortalité. Par exemple, la chirurgie bariatrique réduit pourtant de moitié le risque de décès chez les personnes obèses, et se montre plus efficace à long terme que le régime et l'exercice pour les personnes en situation d'obésité morbide. Dans le même temps, la pose d'un anneau gastrique coûte, sans assurance, entre 20.000 et 35.000 dollars (soit entre 16.800 et 29.300 euros). L'opération chirurgicale n'est pas couverte par le programme Medicaid, qui prend en charge les plus démunis, soit 760.000 personnes.

3) Mise en place d'outils décisionnels

Afin d'approfondir cette étude concernant le coût des dépenses de santé des Américains. Nous allons mettre en place des outils informatiques décisionnelles comme le Machine learning. Il est important de rappeler que nous travaillons avec le langage python dans un environnement Spyder, La distribution d'Anaconda Cloud est riche en bibliothèques utilisées pour exécuter le code informatique. Ces bibliothèques sont : Pandas, Numpy, Matplot lib, Sklearn et Seaborn.

D'abord, on se penchera sur l'estimation du coût moyen des charges supportées par les américains pour leur couverture santé. Ensuite, nous allons essayer de mettre en évidence la catégorie d'individus qui sollicitent le plus souvent une couverture santé.

a- Collecte des données

Les datasets ont été collectées à partir de la plateforme Kaggle dont l'adresse du site est https://www.kaggle.com/. Ces datasets sont des fichiers aux formats « .csv ». Avant d'utiliser ces 2 types de datasets : *« insurance.csv »* et *le doublet train set « insurance3r2.csv » et test set « insurance2.csv ».*

b - Nettoyage des données

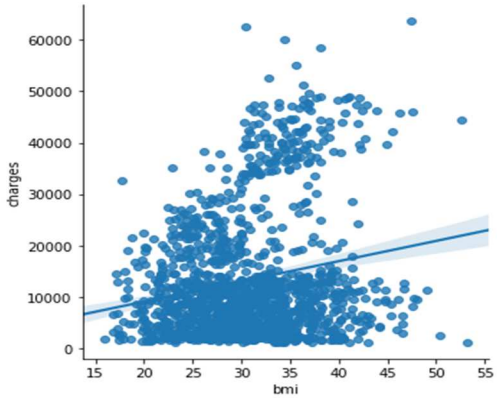
Pour l'ensemble des fichiers csv récupérés, nous avons procéder à l'encodage des données catégorielles suivantes : sexe, région et fumeur ou non (smoker) ou non. La catégorie « région » a été exclue de l'étude car non significative. De plus, une normalisation est nécessaire pour mettre à l'échelle l'ensemble des unités. Nous avons procéder de la même manière pour les 2 datasets de la régression logistique décrite plus loin.

c - Application des modèles d'études

Pour répondre à notre problématique, nous avons choisi la méthode de régression linéaire multiple afin de tester les variables explicatives et déterminer la variable dépendante qui correspond aux « charges » de santé. Avec cette méthode de Machine Learning, nous allons prédire le coût moyen des dépenses d 'assurance maladie.

d - Phase de Test

Avant d'effectuer le split (= séparer en 2 parties train set et test set) du dataset « insurance.csv » en train set et test set, l'aide de la visualisation exploratoire et après plusieurs tests (voir annexe 1 pour les codes), on obtient les résultats suivants :



Figures 1 : Test avec 1339 lignes

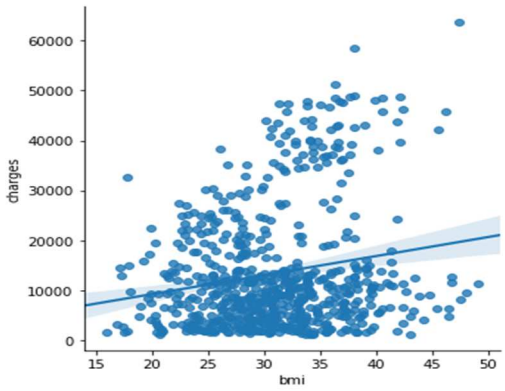


Figure 2 : Test avec 700 lignes

Résultats et interprétation :

Avant même d'estimer le coût moyen des dépenses de santé, la visualisation exploratoire nous indique une tendance. En effet, on s'aperçoit que ce sont les individus obèses, dont l'IMC est élevé qui dépensent le plus pour leur couverture santé. D'après ce graphique, la densité de population semble très importante0 au niveau de l'IMC 30. Autrement dit, les autres facteurs testés (sexe, âge, le fait d'être fumeur) n'influent pas de manière significative sur les dépenses de santé comme pour l'IMC. Après calculs, nous estimons que : l**e coût moyen** des dépenses de santé est de **4000 dollars.**

4) Définition et rappel de l'IMC OU BIM

L'obésité est définie par le fait d'avoir un indice de masse corporel (IMC ou BIM) supérieur ou égal à 30. Le calcul de l'indice de masse corporelle (IMC) vous permet d'estimer si vous êtes trop maigre, si votre poids est normal, si vous êtes en surpoids ou si vous êtes obèse. Pour calculer votre IMC, il vous faut diviser votre poids (en kilogrammes) par le carré de votre taille (en mètre). Par exemple, si vous pesez 65 kilos et que vous mesurez 1.70m, vous devez effectuer l'opération suivante. Par exemple, 65/ [1.70×1.70] = 65/2.89 = 22.49 Votre indice de masse corporelle est 22.49. L'obésité est un facteur de risque de maladies cardiovasculaires, de diabète ou encore de divers cancers.

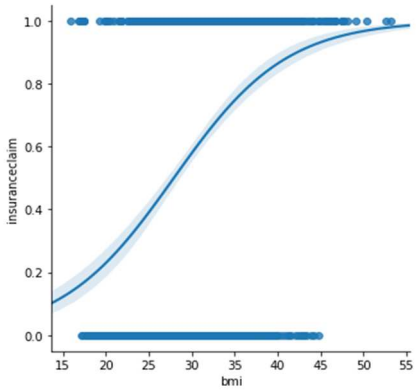
Comme nous l'avons rappelé précédemment, au niveau économique, les frais de chirurgies sont très

onéreux : entre 20.000 et 35.000 dollars. Malgré les sommes exorbitantes, nous allons voir si ce sont bien ces Américains obèses sont les plus nombreux à solliciter une couverture santé.

5) Catégorisation d'individus par la méthode de régression logistique

Rappelons notre hypothèse : Ce sont *les Américains touchés par l'obésité qui représentent la catégorie d'individus sollicitant le plus souvent une demande de couverture santé.*

Concernant le choix des variables explicatives, nous avons effectuer le même nettoyage de données comme décrit dans la rubrique **3) b-** expliquée plus haut. La visualisation exploratoire nous donne déjà une tendance proche de La courbe en « S » (ou fonction sigmoïde) caractéristique de la prédiction par la régression logistique (Voir figure 3).



Figures 3 ; Résultat exploratoire

Après le split, du dataset, en train set « insurance3r2.csv » et test set « insurance2.csv », on procède au calcul du % de chance pour que la réponse corresponde à notre hypothèse (voir annexe 2 pour le code).

Calculs et interprétation des résultats :

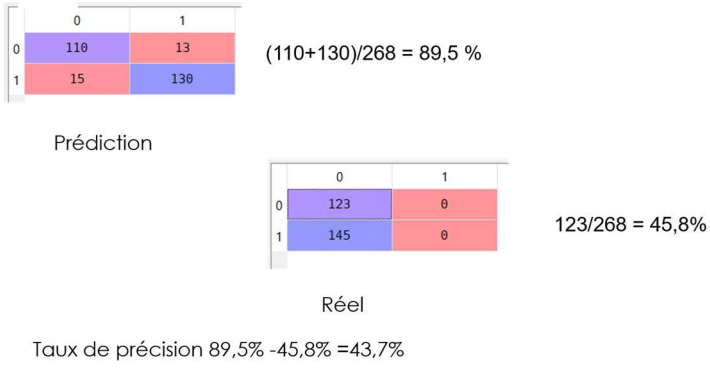


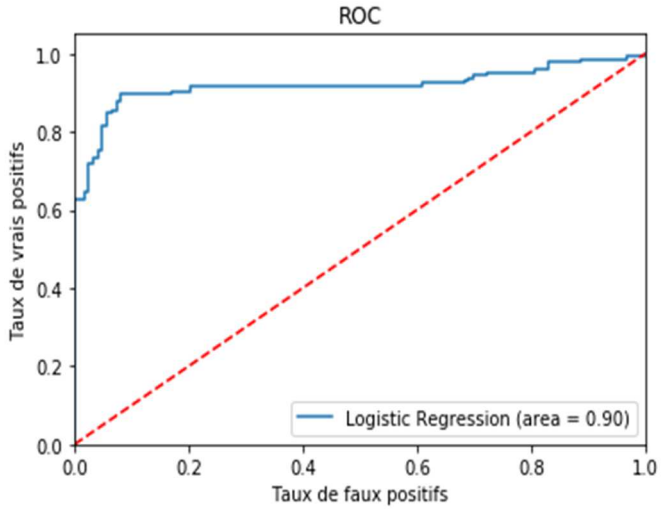
Figure 4 : Matrice de confusion

D'après la méthode de régression logistique, nous avons appliquer le principe de la matrice de confusion (voir figure 4):

(Hypothèse positive + Hypothèse négative) / (Hypothèse positive + Hypothèse négative + 15+ 13).

Avec le dataset d'entraînement le résultat de cette formule est : (110+130) /268 = 89,5 % de chance que ce soient les Américains dont l'IMC est élevée qui sollicitent le plus une couverture santé.

Avec le test set, le résultat est de : 123/268 = 45,8% de chance que ce soient les Américains dont l'IMC est élevée qui sollicitent le plus une couverture santé.
Le Taux de précision de notre modèle est bien de : **89,5-45,8 =43,7 %**



Figures 5: : Mesure de la Performance du test

Enfin d'après la figure 5 ci-dessus, le ROC (Receiver Operating Characteristic) nous conforte dans notre modèle de prédiction. La courbe en bleue est proche de 1 donc oui ce sont bien les Américains obèses qui entreprennent les démarches pour obtenir une couverture santé.

Conclusion et Perspectives

A l'aide d'outils décisionnels, nous avons pu estimer le coût moyen des dépenses de santé par personne aux Etats-Unis : 4000 dollars. Puis nous avons constaté que malgré les charges importantes dépensées, ce sont bien les Américains touchés par l'obésité qui sollicitent le plus souvent une couverture santé. Malgré le nombre de demandes croissantes en couvertures santé, variant en fonction de l'IMC, il semblerait que les pouvoirs publics ne s'investissent pas suffisamment. De plus, les assureurs privés ont tendance à rejeter les dossiers car considérés comme des opérations de rhinoplasties. Pourtant l'obésité est reconnue comme maladie chronique par l'OMS. Les assureurs comme Obamacare et Medicaid ne couvrent pas non plus les cas d'obésité.

Il serait préférable de mener une enquête publique pour interroger tous les Américains dont les dossiers ont été rejetés par les compagnies d'assurances. La fixation d'un tarif minimum par les pouvoirs publics pour garantir un minimum une couverture santé, serait une solution idéale.

Annexes : codes utilisés

Annexe 1 : Estimation du coût moyen ds dépenses de santé

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# import Dataset
dataset = pd.read_csv("insurance.csv")

# permet d'isoler les lignes et les colonnes, : pour tout prendre iloc index localisation
x= dataset.iloc[:,[0,1,2,4,5]]
y= dataset.iloc[:,6]

# Gérer les variables catégoriques
x = pd.get_dummies(x,drop_first= True)

# Visualisation exploratoire avec Seaborn (script lancé ligne par ligne afin de procéder par élimination)
sns.lmplot(x="bmi",y="charges", data = dataset)
sns.lmplot(x="âge",y="charges", data = dataset)
sns.lmplot(x="smoker",y="charges", data = dataset)
sns.lmplot(x="children",y="charges", data = dataset)

# Séparer un dataset en training set et test set
from sklearn.cross_validation import train_test_split
x_train, x_test,y_train, y_test = train_test_split(x,y, test_size = 0.2)
# si pas de random state on peut tester et l'écart

# Normalisation pour mettre à l'echelle
from sklearn.preprocessing import StandardScaler
sc_x = StandardScaler()
x_train = sc_x.fit_transform(x_train)
x_test = sc_x.transform(x_test)

from sklearn.linear_model import LinearRegression
regressor= LinearRegression()
regressor.fit(x_train, y_train)

# Prediction
y_pred = regressor.predict(x_test)

# Méthode des r2
ecart = ((y_pred - y_test)**2)**(1/2)
ecart.mean()
```

Annexe 2 : Méthode de régression logistique

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# import Dataset
dataset = pd.read_csv("insurance3r2.csv")
x = dataset.iloc[:,[0,1,2,4,5,7]]
y = dataset.iloc[:,8]

# Visualisation exploratoire
#sns.lmplot(x="children", y= "insuranceclaim", data = dataset, logistic = True)
sns.lmplot(x="bmi", y= "insuranceclaim", data = dataset, logistic = True)

x= pd.get_dummies(x, drop_first = True)

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test =  train_test_split(x,y, test_size = 0.2, random_state =0)

# Feature Scaling <=> normaization
from sklearn.preprocessing import StandardScaler
sc_x = StandardScaler()
x_train = sc_x.fit_transform(x_train)
x_test = sc_x.transform(x_test)

# Linear_model import LogisticRegression
classifier = LogisticRegression(random_state=0)
classifier.fit(x_train, y_train)

#Predire les resultatsY_pred = classifier.predict(X_test)
y_pred = classifier.predict(x_test)

#matrice de confusion
cm = confusion_matrix(y_test, y_pred)

#matrice de confusion de Comparaison environ 46%
cm_2 = confusion_matrix(y_test, y_pred_3)

#Courbe ROC,, on calcule le ROC (Receiving Operator Characteristic)
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, classifier.predict(x_test))
fpr, tpr, thresholds = roc_curve(y_test, classifier.predict_proba(x_test)[:,1])
plt.figure()

# on ajoute faux positif, faux negatif et un label
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
```

```
plt.xlabel('Taux de faux positifs')
plt.ylabel('Taux de vrais positifs')
plt.title('ROC')
plt.legend(loc="lower right")
plt.show()
```

Ressources

L’obésité : ce qu’il faut savoir
- https://www.lelynx.fr/mutuelle-sante/medecine/maladie/alimentaire/regime/obesite/

L’assurance santé aux Etats-Unis
- https://frenchdistrict.com/articles/systeme-sante-americain-securite-sociale-assurances-medicales/

- Aux États-Unis, deux adultes sur cinq sont obèses
http://sante.lefigaro.fr/article/aux-etats-unis-deux-adultes-sur-cinq-sont-obeses/

- L'impossible traitement de l'obésité aux États-Unis
http://www.slate.fr/story/161407/impossible-traitement-obesite-etats-unis

- L’Obésité aux états unis , enjeux économiques et défis politiques, Anne-Sophie Cérisola et Jacques Mistral Document de travail 2004.01 ? Agence financière, Albassade de France à Washington
http://doc.hubsante.org/doc_num.php?explnum_id=977

-USA: l'obésité coûte 215 Mds $ par an
http://www.lefigaro.fr/flash-actu/2010/09/14/97001-20100914FILWWW00626-usa-l-obesite-coute-200-mds-par-an.php