



# 「以健康為主的動態保費調整」方案

目前保費多是以年齡及性別做為參考依據，本企劃書的目的是為了提高保險公司預測客戶健康狀態的能力，擴大預測的參考變數，找出是否有其他的重要指標會反應投保人健康因素，以縮減保險公司與客戶間的資訊不對稱的風險困境，更探討是否有機會達到「以健康為主的動態保費調整」，讓要保人能夠以更少的費用購買保險，使購買意願增加。

城鎮科技場  
徐英愷 陳怡涵 劉緣梵 徐佳筠

# 目錄

壹、摘要 .....	2
貳、問題陳述與重要性 .....	3
參、資料處理及資料來源說明 .....	4
肆、資料分析及模型建構 .....	8
(一) 資料前處理 .....	8
(二) 資料分析 - 決策樹 .....	9
(三) 資料評估 - 決策樹 .....	14
(四) 資料分析與評估 - 關聯規則 .....	15
伍、結果說明及總結 .....	17
附錄 .....	19

# 「以健康為主的動態保費調整」方案

## 壹、摘要

目前保險費率大多參考以性別及年齡而組成的生命統計表，並要求投保人需告知自身的身體狀況。此舉除了讓保險公司承受更巨大的風險外，也容易使日後產生更多的糾紛，造成公司負面的形象。我們便希望能夠擴大保險公司能夠參考的變數，以提升預測客戶健康狀態的準確度。

我們首先採用 Tableau 進行資料概況分析，再採用 Azure Machine Learning Studio 做初步的機器學習分析。最後，以 R 運用決策樹演算法找出和健康狀況有密切關係的變數，把資料分成訓練集及測試集，並利用權數加以微調模型，並評估「以健康為主的動態保費調整」方案之可行性。

## 貳、問題陳述與重要性

往往，保險公司在核保人壽型保險時，主要依據性別與年齡作為應繳保費的依據，並在填寫健康狀態部分，需仰賴投保人「如實告知」的義務，一旦投保時有任何隱瞞的行為，都很有可能失去日後索賠的權利，因此常常有保險公司「拒賠」的糾紛案例。在信息不完全對稱的情況下，不光是保險公司所承受的風險最大，投保人也很有可能蒙受一定的風險，保險營銷人員為了自身的經濟利益，在推銷保險時隨意誇大保險責任，而對除外責任則輕描淡寫或避而不談，誘導客戶盲目投保。不管是客戶故意隱匿，又或者是無心之過，在保險公司承保之後卻因為「拒賠」的糾紛對公司所帶來的影響，小則賠錢了事，大則傷害公司名聲，對公司都是相當程度的傷害。

因此，本企劃書的目的就是為了提高保險公司預測客戶健康狀態的能力，擴大預測的參考變數，找出是否有其他的重要指標會反應投保人健康因素，例如，經濟活動、社會階層、居住地區等，使其不單單只侷限在以性別與年齡為主，以縮減保險公司與客戶之間的資訊不對稱的風險困境，更探討是否有機會達到「以健康為主的動態保費調整」，讓要保人能夠以更少的費用購買保險，使購買意願增加。

以「提高保險公司預測客戶健康狀態的能力」為目的，我們使用主資料「2011 年 England & Wales 的人口調查」為分析主體，運用決策樹演算法找出和健康狀況有密切關係的變數，把資料分成訓練集及測試集，並利用權數加以微調模型。最後，依據我們的分析結果將得到「那些要素與健康狀態具有高度相關性？」的答案，進而使我們有能力來評估「以健康為主的動態保費調整」方案之可行性。

# 參、資料處理及資料來源說明

## (一) 資料來源說明

僅使用規定之主表單資料，來源為英國統計局之開放資料，是從 2011 年 England & Wales 人口中，取樣 1% 人口調查之結構化資料 (Microdata)。

## (二) 資料處理說明

### 步驟一：檢視資料概況

首先，我們使用軟體 Tableau 進行資料的概況檢視，來瞭解資料的分布狀態。發現在地區 (Region)、性別 (Sex)、職業 (Occupation)、產業 (Industry)、社會階層 (Approximated social grade) 等變數都是分布較平均，而在家庭組成 (Family composition)、婚姻狀態 (Marital status)、健康狀態 (Health)、經濟活動 (Economic activity) 等是分布較不均勻的變數型態。分布最不均勻的變數為居住類型 (Residence type)、人口基數 (Population base)，單一類別比例都在 99% 以上，因此可以直接說，在這筆資料中幾乎所有受訪者都是「常住者 (Usual resident)」且居住型態都非受管理監控的 (communal establishment)。以下以表格檢視數據概況：

註：*communal establishment : A Communal Establishment is defined as an establishment providing managed residential accommodation. Managed means full-time or part-time supervision of the accommodation. (For example, prisons, large hospitals, hotels)*

資料名稱	資料概況
地區	分布均匀，以東南 (15.46%)、倫敦 (14.67%)、西北 (12.54%) 最多人
居住型態	「非」受管理監控的 (communal establishment) 居住者占大多數 (98.13%)
家庭組成	已婚/同性伴侶家庭 (Married/same-sex civil partnership couple family) 占比一半以上 (52.82%)、無家庭關係者 (16.97%)
人口基數	常住者 (Usual resident) (98.47%)
性別	男女參半(男：49.245%，女：50.755%)
年齡	0-15 歲 (18.75%) 最多，其次為 35-44 歲 (13.8%)、45-54 歲青壯年 (13.58%)，分布較為平均之變數
婚姻狀態	單身 (47.57%)、已婚或以註冊為同性伴侶 (37.59%)
學生比	非學生 (79.79%)、學生 (22.21%)
出生國家	大部分為英國人 (85.24%)
健康狀態	非常健康 (46.51%)、健康 (33.65%)、中等 (13.07%)、不健康 (4.31%)、非常不健康 (1.26%)
種族	白人 (84.46%)、亞洲/亞裔 (7.5%)
宗教	基督教占大宗 (58.53%)、無信仰 (24.86%)
經濟活動	受雇者 (37.92%)、退休 (17.11%)、創業者 (7.13%)
職業	專業人員 (11.25%)、非技術人員 (10.26%)、行政及秘書助理人員 (9.35%)，分布較均勻之變數
產業	零售業 / 汽機車修理業 (12.09%)、礦業 / 採石業 / 製造業 / 電力及瓦斯業 / 冷暖氣業 / 自來水業 (9.38%)、房仲業 / 專業科學科技業 / 行政業 (8.77%)，分布較均勻之變數
工時	無代碼 (53.06%)、全職 (27.02%)、兼職 (9.15%)，由於無代碼比例偏高，後分析將不參考此變數
社會階層	C1 (28.02%)、DE (21.72%)、AB (14.45%)、C2 (14.03%)

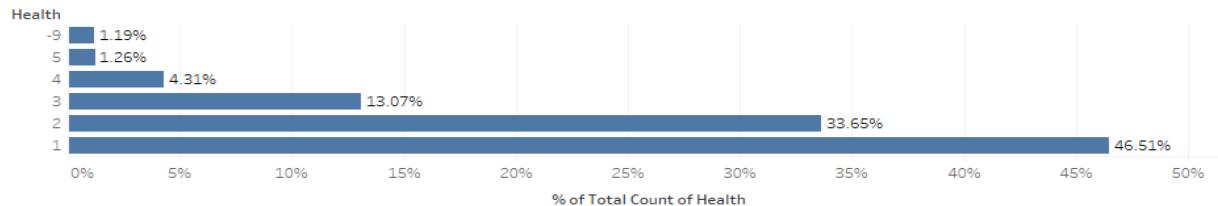
圖表 1：原始資料概況

其中，我們因此能將涵蓋整個數據的變數與未知因素過多的變數去除，例如：居住型態、人口基數、工時，而對其他變數進行下一階段的分析。下一步驟的分析我們將從分布較均勻的變數優先測試，例如：地區、性別、年齡、產業。

## 步驟二：深入探討健康狀態議題

從第一步驟的分析結果，最健康的人占 46.51%，保險公司所關心「不健康與最不健康」比例分別為 4.31% 及 1.26%。以下我們將以各變數與健康狀態分別探討，其中以地區、性別、年齡加以比較。

健康狀態分布



圖表 2：健康狀況分布

**地區：**從資料分布來看，最健康人口的區域主要分布在倫敦及東南地帶，而因為不健康人數比例本來就相對較低，因此在此表上分布情況較不明顯。但當我們將其單獨拉出來看時不健康人數相對較多的比例是落在西北地帶，其次才是倫敦與東南。而當中我們知道，因為東南 (15.46%)、倫敦 (14.67%)、西北 (12.54%) 地區分布的人數比例本來就較多，因此這個部份我們只能對數據有初步的認識，知道地區可能是影響健康的因素之一，但是影響的程度多大，重要性多大？我們將於第三部分做更進一步的資料分析。

地區與健康狀態分布

Health	Region									
	E12000001	E12000002	E12000003	E12000004	E12000005	E12000006	E12000007	E12000008	E12000009	W920000..
1	11,425	32,413	24,086	20,595	25,594	27,739	41,715	42,517	24,593	14,294
2	8,795	23,502	18,302	15,665	19,230	20,526	27,627	29,929	18,631	9,537
3	3,977	9,847	7,476	6,402	7,907	7,637	9,211	10,539	7,036	4,448
4	1,476	3,842	2,376	2,043	2,723	2,120	3,052	2,934	2,167	1,825
5	428	1,083	703	554	810	624	980	839	625	538

圖表 3：地區與健康狀態分布

**性別與年齡**：從性別與健康狀況分布來看，我們發現性別所影響健康狀態的效果較不顯著，而如果加入年齡的因素呢？最健康的人分布在 0-34 歲左右，而健康的人分布在 35-64 歲左右，性別在當中的影響因素依然不顯著。單獨將不健康的人拉出來比較時，我們發現性別在 75 歲以前影響因素不大，但到了 75 歲以後女性不健康與最不健康人數大幅上升。

### 性別與健康狀態分布

Age	Sex / Health									
	1					2				
	1	2	3	4	5	1	2	3	4	5
1	42,538	10,073	1,359	297	109	41,042	9,094	1,001	216	85
2	23,116	9,163	1,368	257	88	20,585	10,658	1,618	278	82
3	22,038	13,177	2,192	564	159	20,553	14,016	2,362	573	129
4	18,748	15,247	3,552	1,118	296	18,669	15,655	3,821	1,202	319
5	14,391	16,325	5,259	1,803	556	14,572	16,248	5,559	2,097	570
6	8,382	13,755	6,751	2,600	718	8,879	14,508	6,965	2,435	669
7	3,961	9,867	6,754	2,262	662	4,224	10,718	7,471	2,223	635
8	1,486	5,754	7,124	2,466	801	1,787	7,486	11,324	4,167	1,306

圖表 4：性別與健康狀態分布

### 小結：

從以上初步的結果來看，我們思考的是，是不是有更多的因素會影響健康狀態，而是現階段保險公司並未採納的衡量指標。又或者從性別資料分析結果，性別對影響健康狀態的效果真的顯著嗎？如果不顯著，是不是能找到比性別更具有參考性的指標因素，使保險公司預測投保人健康狀態的第一步驟變得更精細呢？又或者現在一般以性別與年齡作為保費的依據，彈性較低，如果加入其他要素的衡量指標，能使得在同樣性別與年齡下健康狀態相對健康的人能獲得較低的保費核保，增加保費調整的彈性，是否可以增加健康狀態優等生購買保險的意願？

因此，在第三部分，我們將採用更專業的模型建構方法來探討，究竟那些要素與健康狀態更具相關性？以及影響效果的重要程度又有多大？

## 肆、資料分析及模型建構

### (一) 資料前處理

#### 步驟 0

下載所需的套件以及載入資料。

```
#install.packages("readr")      #Read Rectangular Text Data
#install.packages("dplyr")        #A Grammar of Data Manipulation
library(readr)
library(dplyr)

insurance <- read_csv("01.2011_Census_Microdata.csv")
#str(insurance)
```

#### 步驟 1

重新命名每個欄位的變數名稱，將其空格刪除。

```
names = c("PersonID", "Region", "ResidenceType", "FamilyComposition",
"PopulationBase", "Sex", "Age", "MaritalStatus", "Student",
"CountryOfBirth", "Health", "EthnicGroup", "Religion", "EconomicActivity",
"Occupation", "Industry", "HoursWorked", "SocialGrade")
colnames(insurance) = names
```

#### 步驟 2

刪除健康狀態欄位結果為未知 (-9) 的資料。

```
insurance <- insurance %>% filter(Health != -9)
#str(insurance)
```

## 步驟 3

根據前述資料處理所作之分析，我們發現居住型態 (ResidenceType) 和人口基數 (PopulationBase) 在資料中的類別幾乎完全相同，因此我們將這兩項變數予以刪除。同時，我們也將資料的 Key (PersonID) 紿去除，並把剩下的變數轉為類別形式。

```
insurance <- insurance %>% select(-PersonID, -ResidenceType,  
-PopulationBase) %>%  
mutate_all(funs(as.factor(.)))  
#str(insurance)
```

## (二) 資料分析 - 決策樹

### 步驟 0

安裝決策樹分析所需之套件。

```
#install.packages("rpart")      #An algorithm of Decision Tree  
#install.packages("rpart.plot") #The visualization of "rpart" package  
#install.packages("partykit")   #Another algorithm of Decision Tree  
library(rpart)  
library(rpart.plot)  
library(partykit)
```

### 步驟 1

站在保險公司的立場來看，如果我們把病人預測為健康的狀態，會比將健康的人誤認為不健康的狀態來得嚴重許多，因此我們建立以下的損失矩陣一 (lossmatrix\_first)，對非常不健康至非常健康依序的比重為 120 : 24 : 6 : 2 : 1。

```
lossmatrix_first = matrix(c( 0,  1,  1,  1,  1,  
                           2,  0,  2,  2,  2,  
                           6,  6,  0,  6,  6,  
                          24, 24, 24,  0, 24,  
                         120,120,120,120,  0), byrow = T, nrow = 5)
```

此外，若將健康狀況誤判一個層級是在可接受的範圍，但如果誤判的程度越大，例如將健康狀況非常健康的人判斷為一般或是不健康 (1 -> 3 或 4)，這樣的誤判程度 (兩到三個層級) 就越不能接受，更何況是完全判斷錯誤 (四個層級) 的情況，如此一來會讓公司承受較大損失風險。因此，我們依此原因建立出了以下的損失矩陣二

(lossmatrix\_second) ，其對判斷錯誤的層級由大至小依序的比重為 24 : 6 : 2 : 1。

```
lossmatrix_second = matrix(c( 0, 1, 2, 6, 24,
                             1, 0, 1, 2, 6,
                             2, 1, 0, 1, 2,
                             6, 2, 1, 0, 1,
                             24, 6, 2, 1, 0), byrow = T, nrow = 5)
```

最後，我們將第一個矩陣 (lossmatrix\_first) 與轉置後的第二個矩陣 (lossmatrix\_second) 相乘後形成最終的損失矩陣。

```
lossmatrix = lossmatrix_first * t(lossmatrix_second)
```

## 步驟 2

我們將本次的分析資料以 80% : 20% 的比例分成訓練集及測試集。同時，依據以上計算出的損失矩陣，我們用訓練資料建立出一個完全生長的決策樹。

```
set.seed(1234)
idx_fold <- sample(1:5, nrow(insurance), replace = T)
idc_train <- idx_fold != 5

rpart_fit <- rpart(Health ~ ., data = insurance, subset = idc_train, method
= "class", parms = list(loss = lossmatrix), cp = -1)
```

## 步驟 3

利用步驟 2 產生出的模型來預測訓練及測試資料級的結果（健康狀態），並列印出混淆矩陣和分類錯誤率。

```
# Find the predicted class
pred_train <- predict(rpart_fit, insurance[idc_train, ], type = "class")
pred_test <- predict(rpart_fit, insurance[!idc_train, ], type = "class")

# See the prediction
table_train = table(real = insurance[idc_train, ]$Health, predict =
pred_train)
table_train
#   predict
#real      1      2      3      4      5
#1    74480  86745  40252  8218  2294
#2    15385  79885  34922  17068  5994
#3      973  10622  26458  14624  6995
#4      19   292   3258  12751  3438
#5       0    12    131   1045  4538

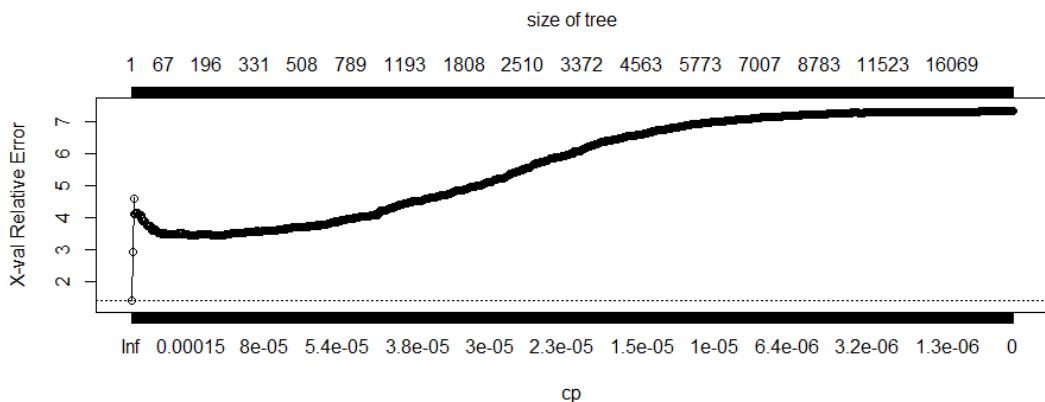
table_test = table(real = insurance[!idc_train, ]$Health, predict =
pred_test)
table_test
#   predict
#real      1      2      3      4      5
#1    15606  23662  10529  2485   700
#2    6342   16355  9663  4539  1591
#3    868    3535  4399  4062  1944
#4   119    594   1160  1838  1089
#5    41    125   294   592   406
# Calculate the classification error
error_train = 1 - sum(diag(table_train)) / sum(table_train)
error_train
#[1] 0.5601411
error_test = 1 - sum(diag(table_test)) / sum(table_test)
error_test
#[1] 0.6569692
```

## 步驟 4

因為完全生長且無限制的決策樹可能會造成模型對訓練資料集的過度擬合，因此我們透過對訓練資料進行 10-fold 交叉驗證的方式來修剪這棵樹，找出最佳的修剪程度 (cp\_best)，以降低樹的複雜度。

```
cp_matrix = printcp(rpart_fit)
cp_matrix = cp_matrix[c(-1,-2),]
cp_best = cp_matrix[which.min(cp_matrix[, "xerror"])], "CP"]

prune_fit <- prune(rpart_fit, cp = cp_best)
```



圖表 5：不同 cp 值底下的相對錯誤率比較

## 步驟 5

利用步驟 4 產生出的模型來預測訓練及測試資料級的結果（健康狀態），並列印出混淆矩陣和分類錯誤率。

```
# Find the predicted class
pred_prune_test <- predict(prune_fit, insurance[!idc_train, ], type =
"class")

# See the prediction
table_prune_test = table(real = insurance[!idc_train, ]$Health, predict =
pred_prune_test)
table_prune_test
#     predict
#real      1     2     3     4     5
#   1    415 30853 19306  2307   101
#   2     84 14535 17011  6442   418
#   3    10  1949  5211  6529  1109
#   4     2   251   936  2529  1082
#   5     0    59   198   786   415

# Calculate the classification error
error_prune_test = 1 - sum(diag(table_prune_test)) / sum(table_prune_test)
error_prune_test
#[1] 0.7946916
```

### (三) 資料評估 - 決策樹

我們對完整生長的決策樹和經由修剪後的決策樹進行比較，得出以下結果：

```
table_test;table_prune_test
#   predict
#real    1    2    3    4    5
# 1 15606 23662 10529 2485 700
# 2 6342 16355 9663 4539 1591
# 3 868 3535 4399 4062 1944
# 4 119 594 1160 1838 1089
# 5 41 125 294 592 406
#   predict
#real    1    2    3    4    5
# 1 415 30853 19306 2307 101
# 2 84 14535 17011 6442 418
# 3 10 1949 5211 6529 1109
# 4 2 251 936 2529 1082
# 5 0 59 198 786 415

error_test;error_prune_test
#[1] 0.6569692
#[1] 0.7946916
```

同時，根據修剪之後的決策樹，我們找出變數的相對重要性。

```
prune_fit$variable.importance
# EconomicActivity          Age      MaritalStatus     Occupation       Industry      SocialGrade
# 55125.1060        37957.9889     11544.6842      3325.5436      3245.3986     2575.8693
#FamilyComposition        Religion      Region      HoursWorked      EthnicGroup      Student
# 1853.0231        1777.5423     1757.8407      1338.4296      650.0216     255.0316
#           Sex      CountryOfBirth
# 235.3054        197.4269
```

圖表 6：變數相對重要性

## (四) 資料分析與評估 - 關聯規則

### 步驟 0

安裝關聯規則所需之套件。

```
#install.packages("arules")      #An algorithm of Association Rules  
#install.packages("arulesViz")    #The visualization of "arules" package  
  
library(arules)  
library(arulesViz)
```

### 步驟 1

依照關聯規則的需求，將資料轉換成交易資料的型態。

```
insurance_trans <- as(insurance, "transactions")
```

接著，我們找出規則筆數超過 100 則，信賴度高於 5 % 的規則 (rules)。

```
rules <- apriori(insurance_trans,  
                   parameter = list(maxlen = 5,  
                                     support = 100/562937,  
                                     confidence = 0.05))  
  
#summary(rules)
```

### 步驟 2

根據保險產業的特性來說，保險經紀人通常會在乎什麼樣的人容易有身體非常不健康的  
情況產生。因此，我們篩選出健康狀態為非常不健康且增益大於 1 的規則 (rulesOwn)。

```
rulesOwn <- subset(rules, subset = rhs %pin% "Health=5" & lift > 1)  
#summary(rulesOwn)
```

## 步驟 3

利用步驟 2 篩選出的規則 (rulesOwn) , 我們將前 10 大增益的規則列印出來。

```
rulesOwn_sort = sort(rulesOwn, by = "lift")
inspect(rulesOwn_sort[1:10])
#      lhs                                rhs
#[1] {Sex=1,MaritalStatus=2,EconomicActivity=8,SocialGrade=3} => {Health=5}
#[2] {Age=7,EconomicActivity=8}                      => {Health=5}
#[3] {Age=7,EconomicActivity=8,HoursWorked=-9}        => {Health=5}
#[4] {Age=7,Student=2,EconomicActivity=8}              => {Health=5}
#[5] {Age=7,Student=2,EconomicActivity=8,HoursWorked=-9} => {Health=5}
#[6] {Age=7,EthnicGroup=1,EconomicActivity=8}          => {Health=5}
#[7] {Age=7,EthnicGroup=1,EconomicActivity=8,HoursWorked=-9} => {Health=5}
#[8] {Age=7,Student=2,EthnicGroup=1,EconomicActivity=8}    => {Health=5}
#[9] {Age=7,CountryOfBirth=1,EthnicGroup=1,EconomicActivity=8} => {Health=5}
#[10] {Age=7,CountryOfBirth=1,EconomicActivity=8}           => {Health=5}
# support   confidence   lift   count
#[1] 0.0001794162 0.2121849 16.62677 101
#[2] 0.0002273789 0.2067851 16.20365 128
#[3] 0.0002273789 0.2067851 16.20365 128
#[4] 0.0002273789 0.2067851 16.20365 128
#[5] 0.0002273789 0.2067851 16.20365 128
#[6] 0.0001989565 0.2036364 15.95691 112
#[7] 0.0001989565 0.2036364 15.95691 112
#[8] 0.0001989565 0.2036364 15.95691 112
#[9] 0.0001811926 0.2000000 15.67197 102
#[10] 0.0001811926 0.1976744 15.48973 102
```

## 伍、結果說明及總結

### (一) 決策樹之結果說明

1. 雖然修剪之後的決策樹分類錯誤率較高，但我們相較在乎的資料（不健康和非常不健康的人），預測水準則能大幅提升，因此我們認為剪枝後的決策樹模型較適合拿來作保險資料的預測。
2. 經濟狀況 (Economic Activity)、年齡 (Age)、婚姻狀態 (Marital Status) 對健康狀態看起來影響的程度最大。
3. 性別在我們的模型中沒有像現實生活中的情況一樣是保險資費判斷的一大基準。

### (二) 關聯規則之結果說明

1. 我們發現年齡介於 65-74 歲間、因長期生病或是殘障而導致無經濟活動能力、住在英國地區、不是學生、或屬於白種人的特質，是最有可能造成是不健康規則的因子。
2. 長期生病或是殘障而導致無經濟活動能力其實和健康狀態有著非常大的關聯性，因此我們可以考慮將此變數刪除或將類別縮減。

### (三) 總結

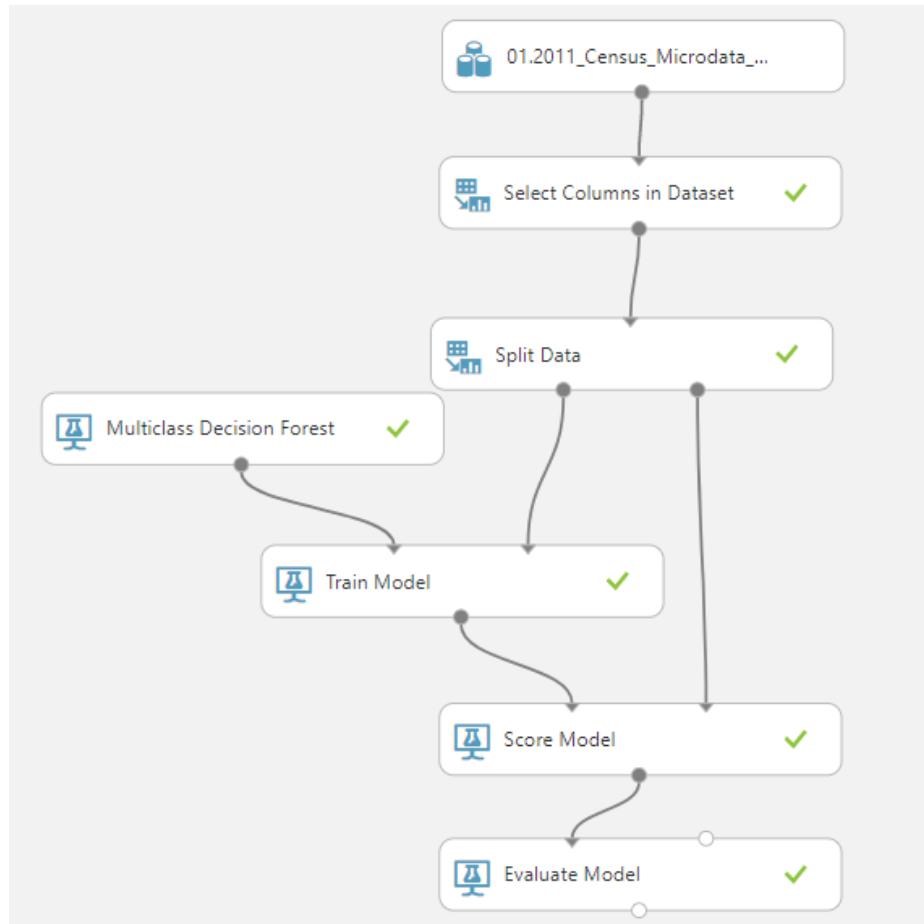
保險是金融商品之一，隨著科技的進步及金融市場的擴張，人人皆較容易取得相關的保險資訊及知識，市場需求逐年增加。以消費者的角度來看，最關心的是能不能買到最適合自己的商品，以及保險費用是否合宜、是否能負擔。以保險公司來說，則是有沒有足夠多的客戶作為保險基數。

在人壽保險中，健康狀態是理賠中主要的關鍵因素，因此我們想透過主資料，去分析群眾的健康程度，會不會依客戶的社會階層、性別、年齡...等特徵，有一定的影響權重比例。進而打破目前僅依性別及年齡調整保費的現況，同時站在要保人及保險公司的立場，探討「以健康為主的動態保費調整」是否可行，讓具有「高健康程度」跡象的要保人能夠以更少的費用購買保險，使購買意願增加、保險公司承保人數增加，而降低保險公司整體的承保風險，達到雙贏局面。

由以上的資料分析結果，發現目前對於要保人的性別和年齡的分類方法並不是最好的，其中我們得到，健康與否與性別的關係不顯著，與經濟狀況 (Economic Activity) 、年齡 (Age)、婚姻狀態 (Marital Status) 等項目有密切相關。結果也讓我們確信，現行保險費用的分類方式是可以嘗試加入其他評估要素讓保險公司更精確的衡量其健康可能的狀態，以降低整體承保的風險，使保費的調整能更加多元、彈性。而我們也深信，保費的彈性一旦增加，對投保人來說其實具有一定的吸引力存在，保險基數因此上升，如此一來將更能彰顯保險所帶來的價值。

# 附錄

## 無特殊損失矩陣之 Azure Machine Learning Studio 分析



圖表 7 : Azure Machine Learning 流程圖

## 步驟 0 資料前處理

刪除健康狀態欄位結果為未知 (-9) 的資料。

## 步驟 1 選定特徵值

除了人口基數、居住型態、工時三項，其他都是我們所選定的特徵值。



圖表 8：所選定的特徵值列表

## 步驟 2 資料分割

將資料隨機分割成兩部分，70%的資料用以建立模型，30%的資料用以評估模型的準確度。

## 步驟 3 選定演算法

我們選擇了多類別決策森林 (multiclass decision forest) 來實作，並使用 8 顆決策樹，並使每棵樹的深度最高為 32。

## 步驟 4 模型評估

由於健康狀況為 1 (非常健康) 的資料比較多，因此在分析上會盡可能地將其分析得最準確，卻也因此犧牲了健康狀況非常糟糕的準確度。然而，我們比較希望能夠準確地預測出健康狀況不好的人，因而改採用 R 實作並且自己訂定損失矩陣，以達到我們預期的分析結果。

		Predicted Class					
		1	2	3	4	5	
		Actual Class	1	2	3	4	5
	1		69.5%	26.0%	3.8%	0.6%	0.1%
	2		49.1%	38.0%	10.6%	2.0%	0.4%
	3		25.9%	39.2%	26.1%	7.3%	1.5%
	4		14.8%	31.2%	33.0%	17.6%	3.4%
	5		12.9%	27.3%	35.7%	19.8%	4.4%

圖表 9：實際及預測之健康狀況表