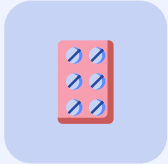
Data Mining for Healthcare

Understanding the Factors that Influence Healthcare Plan Enrollment

Presentation Overview

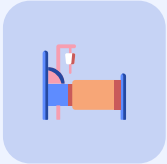


Topic Introduction

What we hope to gain insight on for our company

Data Mining Methods

Logistic regression, decision trees, and clustering



Implications

The impact on our future business strategies

Scenario

We are a **health insurance company** looking to gain more customers and have existing customers upgrade their health plans. We have **survey data for over 12,000 people** and are looking at different data mining techniques where we can generate useful insights into **how we can get more customers**.

Data in the survey includes:

- Enrolled in Health plan
- Income
- Marital status
- Education
- Gender
- Age
- Self Esteem

Figure Out Significant Factors – Logistic Regression

- HealthPlan: Binary
- Conducted **logistic regression** on HealthPlan by some **observable features of customers**
- HealthPlan= β0 + β1Age + β2Urban + β3Black + β4Hispanic + β5White + β6Christian + β7Male + β8FamilySize + β9Height + β10Weight + β11Income + β12 Marital_Status + β13Education + β14WeeksEmployed

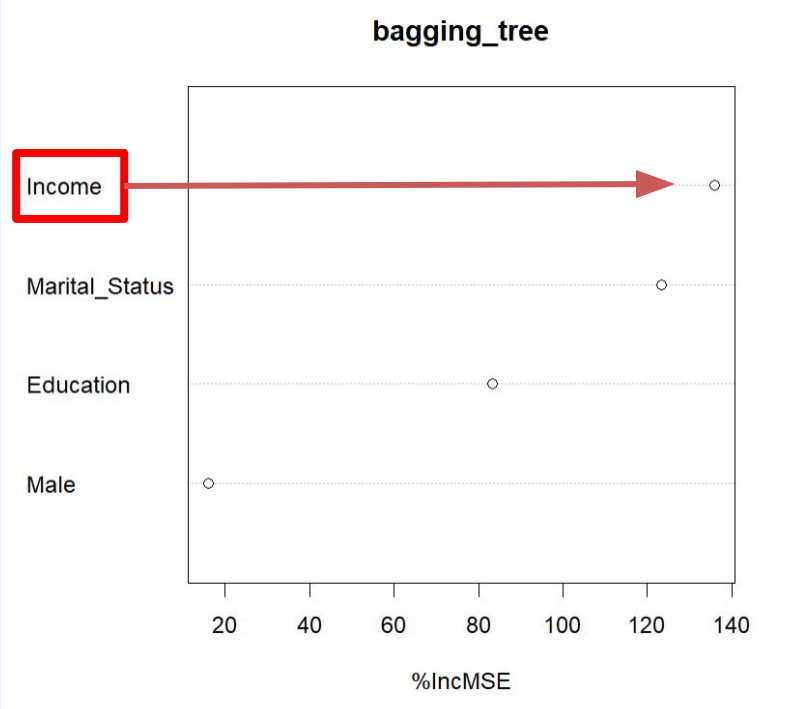
Figure Out Significant Factors – Logistic Regression

Variable	Estimate
Gender (Male=1)	-8.438e-01***
Income	3.897e-05***
Marital_Status	-1.204e-01***
Education	1.962e-01***

- **Women** are more likely to purchase the health plan
- **Income and education** are positively associated with health plan
- **Marital status** also has significant effect on health plan
- Finding these significant factors can further **determine our decision tree** model

Preparation for Decision Trees

Next, we build multiple **decision tree models** by splitting the case **based on 4 significant factors** found in logistic regression.



- Construct the **ensemble tree**
  - Bagging strategy
- **Rank** 4 factors
  - Income is the most important predictor variable

```
> importance_df
                %IncMSE IncNodePurity
Male           16.03696      25.79385
Income        135.95204     470.46125
Marital_Status 123.27056     924.83693
Education      83.22614      169.51380
```

- **Specify splits** in percentage and number
  - Regression tree
  - Classification tree

Regression and Classification Trees

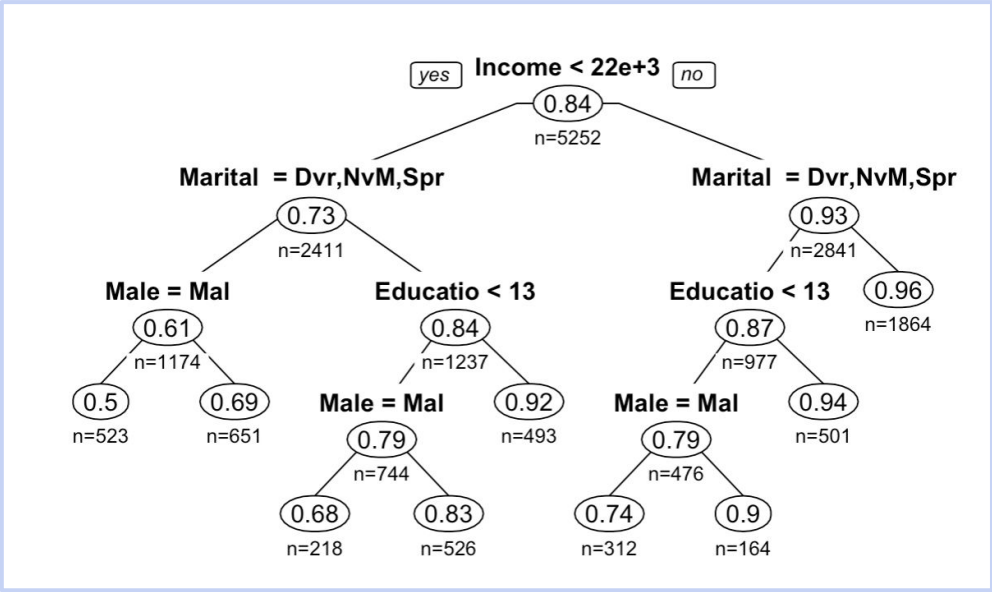


Figure 1 : Regression Tree

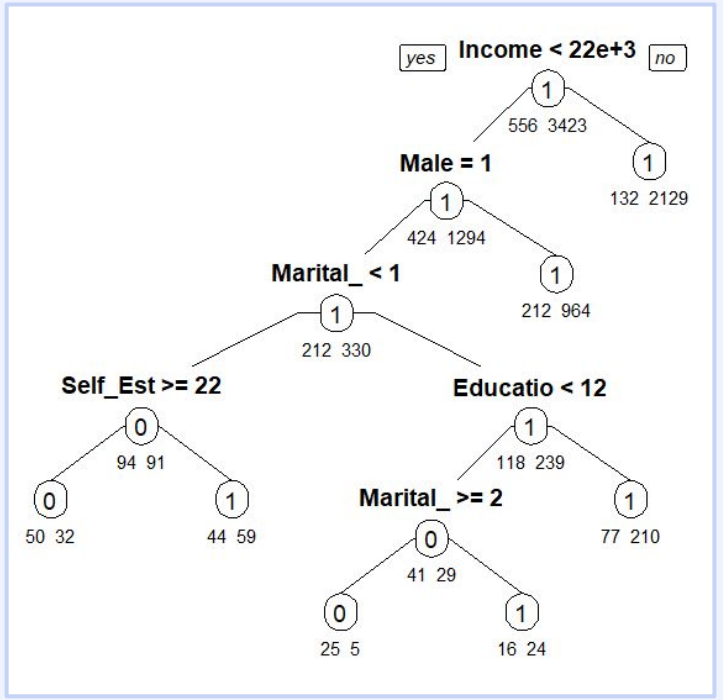


Figure 2 : Classification Tree

Decision Trees Impact

- We can use these decision trees as **the basis of a script** insurance agents can use when they sell insurance. They can **use the splits in the tree (Figure 1) as questions** to ask prospective customers and go down the tree until they reach a leaf node which gives them the **probability of them buying a health plan**.
- Depending on how high or low the probability is, they can **offer discounts** as needed as well the **amount of selling pressure** they can use.
- The insurance agent can use the classification tree (Figure 2) to understand from their questions at each split, whether or not the customer will buy a health plan.
- Overall, these decision trees and each customer's response to their questions will provide the insurance agent with a better understanding of the **factors affecting their customers decisions**.

Additional Research - Unsupervised Data Mining

Marketing budgets are always limited, how can we effectively market our product to the right group of customers?

Target Marketing

How?
→ Perform K-means clustering
→ Find different groups of customer with similar traits
→ Apply the correct marketing strategies/call script to the correct group

Data Preparation for K-mean Clustering

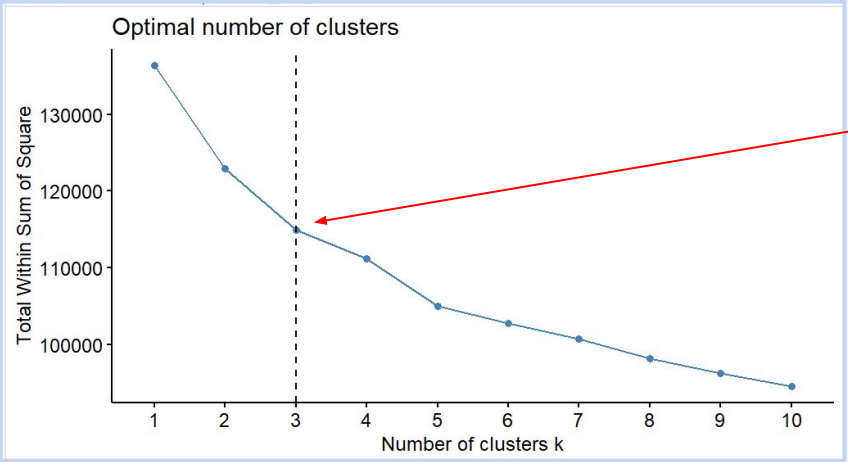
- Loading the Survey dataset
- Remove NA values
- Scale the dataset

```
library(cluster)
library(readxl)
library(factoextra)
myData <- read_excel("Survey.xlsx")
myData <- na.omit(myData)
View(myData)

# Load and prepare the data
myData <- scale(myData) # Scale the data to have mean=0 and variance=1
```

Finding the optimal K - WSS Elbow Method

```r
# Perform k-means clustering with different k values
set.seed(1) # Set the seed for reproducibility
wss <- c()
for (i in 1:10) {
  kmeans_fit <- kmeans(myData, centers = i, nstart = 25)
  wss[i] <- kmeans_fit$tot.withinss
}

# Plot the within-cluster sum of squares (WSS) against the number of clusters (k)
fviz_nbclust(myData, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = "dashed") # Add a vertical line at the optimal k value
```

The within-cluster sum of squares (WSS): A measure of the compactness of the clusters in k-means clustering.

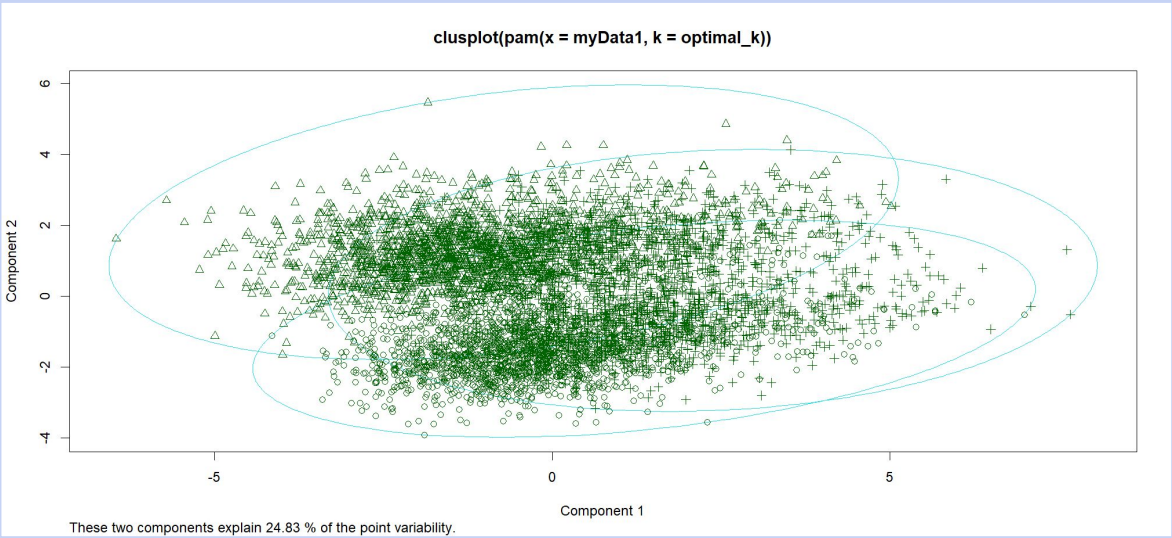


Based on the **WSS Elbow Method**, the optimal K is 3

K-means Clustering - Cluster Map

```
Numerical information per cluster:
     size  max_diss  av_diss diameter separation
[1,] 2026 11.387406 4.616274 13.86158   1.424277
[2,] 1866  9.452062 4.768712 13.80123   1.401355
[3,] 1791 10.073585 5.185911 15.16824   1.401355
```

```
40  optimal_k <- 3
41  set.seed(1)
42  kResult <- pam(myData, k=optimal_k)
43  summary(kResult)
44
45  plot(kResult)
```



Results & Solutions

```
myData <- data.frame(myData, kResult$clustering)
summary(subset(myData, myData$kResult.clustering == 1))
summary(subset(myData, myData$kResult.clustering == 2))
summary(subset(myData, myData$kResult.clustering == 3))
```

Cluster 1

```
  HealthPlan
Min.   :-2.48302
1st Qu.: 0.40267
Median : 0.40267
Mean   : 0.05228
3rd Qu.: 0.40267
Max.   : 0.40267
```

Majority enrolled in a health plan

Cluster 2

```
  HealthPlan
Min.   :-2.4830
1st Qu.: 0.4027
Median : 0.4027
Mean   : 0.1135
3rd Qu.: 0.4027
Max.   : 0.4027
```

Few are enrolled in a health plan

Cluster 3

```
  HealthPlan
Min.   :-2.4830
1st Qu.: 0.4027
Median : 0.4027
Mean   :-0.1774
3rd Qu.: 0.4027
Max.   : 0.4027
```

Possible Marketing Solutions

- Solution 1 (Retargeting) - Cluster 2 is the best group for policy renewal or upgrade.
- Solution 2 (Cold Calls) - Cluster 3 is the best group for cold calling.

Impact on Future Business

- Logistic regression showed us the **variables that impacted enrollment** the most
  - Potential for follow-up surveys
- Decision trees provide a **level-by-level breakdown** of the significant variables
  - Insurance agents can follow as a **"script"** when looking to sell policies
  - **Customize policies** based on decision tree probabilities
- Cluster analysis broke up respondents with varying rates of enrollment
  - Tailor **marketing** efforts to each specific cluster

Thank you for listening!