

**IST 687: Introduction to Data Science**

**Final Project Report**

**Analyzing healthcare cost information from an HMO (Health Management Organization)**

Shruti Rao

Shubham Koshti

Kishan Polekar

Janhavi Ghuge

Contents

[Introduction 3](#_Toc134556620)

[Objective 3](#_Toc134556621)

[Background 3](#_Toc134556622)

[Context 4](#_Toc134556623)

[Business Questions 5](#_Toc134556624)

[Initial Business Questions 5](#_Toc134556625)

[Final Business Questions 5](#_Toc134556626)

[Data Analysis 6](#_Toc134556627)

[Data Acquisition 6](#_Toc134556628)

[Data Cleaning and Variable Analysis 7](#_Toc134556629)

[Descriptive Statistics and Visualizations 10](#_Toc134556630)

[Independent variables analysis 10](#_Toc134556631)

[Graph Visualizations 11](#_Toc134556632)

[Modelling Techniques and Prediction 12](#_Toc134556633)

[Conclusion and Recommendations 14](#_Toc134556634)

Introduction

## Objective

We act as a consultancy company for HMOs (Health Management Organizations), which are associations of health insurance companies that provide medical services in return for a predetermined yearly fee.

To lower the HMO's overall health care costs, we specifically advise the HMO on how to cut costs. Our goal is to identify the key factors that determine why some people need more medical attention than others, identify those who will spend a lot of money on healthcare in the upcoming year, and identify those who will be expensive to the insurance companies.

## Background

Health Management Organizations (HMOs) are medical insurance groups that offer healthcare in exchange for a set annual charge. The dataset that we were provided with had 14 columns (attributes) and 7582 rows (observations). The attributes focus on several categories, such as patient’s unique identifier, age, bmi, number of children, marital status, physical activity, smoking status, etc.

Based on the given information, we were supposed to deliver actionable insights via our research and try to anticipate the factors that affect the spending cost of patients on healthcare. Moreover, we had to identify an ideal value above which the healthcare costs of patients would be termed as expensive. Depending on our analysis, we were assigned to offer some suggestions on how patients can reduce their healthcare expenditures.

## Context

To analyze the dataset, we have divided the columns into three categories:

|  |  |
| --- | --- |
| **Patient’s Basic Information** | |
| **Variable** | **Description** |
| x | Unique Identifier |
| age | Age of the person at the end of the year |
| Gender | Gender of the person |
| education\_level | The amount of College Education |
| married | Marital Status of the individual |
| num\_children | Number of children |

|  |  |
| --- | --- |
| **Patient’s Geographical Information** | |
| **Variable** | **Description** |
| location | Location according to states in US |
| location\_type | Type of Location – Urban/Country |

|  |  |
| --- | --- |
| **Patient’s Health Information** | |
| **Variable** | **Description** |
| exercise | Actively exercising or not |
| smoker | If the person is an active smoker or not |
| hypertension | If the person has hypertension or not |
| bmi | Body Mass Index of the person |
| yearly\_physical | If the person has visited their doctor for frequent check-ups or not |
| cost | Total healthcare cost for a particular person, within the past year |

# Business Questions

We have figured out some of the business questions to come up with analysis techniques to achieve the project deliverables. Few of them are as follows:

## Initial Business Questions:

1. What are the factors that are affecting the healthcare costs of the patients in the database? And predict potential expensive healthcare patients in the coming year.
2. What are the precautions or actions that need to be taken to lower the patient’s healthcare costs? Enlist recommendations for the same.

## Final Business Questions:

1. What are the significant factors that are affecting the cost of patients in the USA?
2. What are the most expensive and least expensive states in terms of healthcare patients’ costs?
3. Which group of people are spending more on healthcare compared to other groups?
4. How is the distribution of cost variable in the dataset? What would be the range in which most of the patients spend on healthcare?
5. Does smoking affect the cost of insurance premiums for patients?
6. How are geographical attributes of the patients affecting their healthcare costs?
7. Are exercise or yearly checkups significant factors in reduced healthcare costs of patients?
8. What steps can be taken to lower the companies’ as well as the customer’s healthcare costs?
9. What measures can be instructed to build a healthier community?

# Data Analysis

## Data Acquisition

We were provided with a data frame of 14 columns and around 7582 observations. We first downloaded the .csv file from the desired link and imported it into R Studio, using read\_csv() function into a new data frame named HMO\_data. Then, displayed the first 6 observations using head() function to get a basic idea of the dataset.

Text

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated with medium confidence

## Data Cleaning and Variable Analysis

Statistical Analysis of variables – We identified 6 attributes (columns) that were of numerical type. We ran functions such as summary() to analyze the statistical characteristics of each variable.

Table

Description automatically generated with medium confidence

**Identifying and removing missing values –** Using statistical methods, we identified that there are NA values in variables such as bmi and hypertension. This might be due to manual errors while entering patient’s information in the database.

We used na\_interpolation() function from the imputeTS package to remove the N/A values from the variables mentioned above.

**I**Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Detecting possible outliers –** We analyzed the cost variable and identified that the outliers lie within the top and bottom 0.5% of the data. Using quantile() function, removed the possible outliers.

Graphical user interface, text, application, Teams

Description automatically generated

# Descriptive Statistics and Visualizations

## Independent variables analysis

We plotted bar graphs for each of the potentially significant variable against the cost variable. We analyzed that the smoker attribute has the most significance on the cost variable.

Chart, box and whisker chart

Description automatically generated

Moreover, we plotted a scatterplot for age and bmi variables against the cost variable and observed that there exists a slightly linear relationship between them.

Chart

Description automatically generated

## Graph Visualizations

To visualize the impact of cost variable geographically, we plotted maps which explained the distribution of cost variable in the states included in the dataset. We found that the state of Maryland has a higher average cost as compared to other states.

A map of the united states

Description automatically generated

Moreover, after creating the expensive variable, we visualized the data geographically by plotting maps to demonstrate distribution of population of expensive and non-expensive patients in various states.

A map of the united states

Description automatically generated

# Modelling Techniques and Prediction

Before running models on the data, we divided the dataset into training and testing datasets. The proportion that we selected was 70:30 i.e., 70% training data and 30% testing data.

Firstly, we applied a linear regression model to the dataset. Using this model, we were able to predict the cost variable with almost 59% accuracy using predictors such as age, smoker, exercise, bmi and hypertension.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, font, screenshot, algebra

Description automatically generated

Secondly, we used SVM model using the ‘svmRadial’ method. This is the most popular and most used method as it is very similar to a Gaussian distribution.

We achieved an accuracy of around 93% using the SVM model. To get this accuracy number, we used the confusion matrix function from the caret package.

A screenshot of a computer

Description automatically generated with medium confidence

Thirdly, we used the K-SVM model. This is a Kernel-SVM approach. The difference here as compared to the normal SVM approach is that we specify the number of Kernels (points) that we will use closest to the current point to determine if they are similar or not for classification.

This method uses the efficiency of SVM along with the accuracy of KNN (nearest neighbor) method. The accuracy that we achieved using this method was better as compared to SVM model, this one being around 98.36%. We used the confusion matrix function to observe the model accuracy.

A screenshot of a computer

Description automatically generated with medium confidence

# Conclusion and Recommendations

* People that are physically active on a regular basis are observed to have lower medical costs than those that do not.
* Regular exercise is also beneficial for mental health, as it helps to reduce stress and anxiety, improves mood and cognitive function, and promotes better sleep quality, all of which can lead to better overall health outcomes and potentially lower medical costs.
* Our analysis has highlighted that smoking has a significant impact on medical expenses. To address this issue, it is important to create awareness about the dangers of smoking and encourage individuals to quit smoking.
* In addition to social media campaigns, other awareness strategies can include organizing health fairs and community events, distributing educational materials, and partnering with local schools and businesses to promote smoking cessation programs.
* Promoting regular health checkups for individuals over the age of 40 is crucial to reduce the risk of health issues related to smoking.
* These checkups can help to detect potential health problems early, offer support and guidance on quitting smoking, and monitor the progress of individuals who have quit smoking. By taking these steps, we can help individuals lead healthier lives and reduce the burden of medical costs associated with smoking-related illnesses.