Logo

Description automatically generated

Week 4 Practice

Craig Perkins – [perkins.cr@northeastern.edu](mailto:perkins.cr@northeastern.edu)

Northeastern University

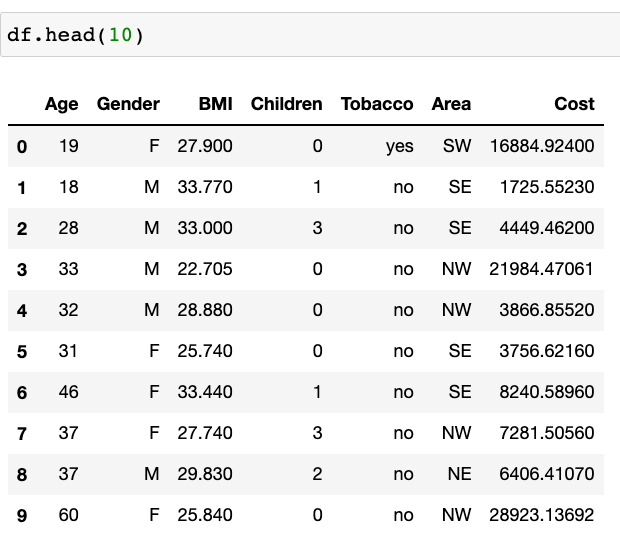
December 2022 – ALY 6150

**Week 4 Practice  
Introduction**

In this assignment I analyze data about costs of healthcare. The data I analyze contains 1338 records with 7 features include the cost of the service rendered. The first section is the exploratory analysis portion of this assignment that examines the data in the dataset visualize the distribution of the data and obtain descriptive statistics. The second section discussed the cleaning and preparation of the data for predictive analysis and the third section discusses the methods used to build a regressor to predict cost of a service. The regressor is analyzed to determine important features in determining the cost.

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| **Age** | Age of the patient |
| **Gender** | Gender of the patient |
| **BMI** | Body Mass Index |
| **Children** | Number of Children |
| **Tobacco** | Smoking (Y/N) |
| **Area** | NW,NE,SW,SE |
| **Cost** | Cost of service rendered |

Samples of rows within the dataset:



Descriptive Statistics of the dataset:

Table

Description automatically generated

**Exploratory Data Analysis**

For the exploratory data analysis, I wanted to get a sense of the demographic data and its distribution contained in the dataset. I chose a variety of different methods to visualize the data in the dataset including a pie chart, bar chart, box plot and histogram. In this section I also analyze the distribution of cost as cost is the variable that is being predicted based on the other features of the dataset. Before continuing with the rest of the analysis, I hypothesize that higher BMI and Tobacco use will contribute to higher costs associated with the service rendered. 18.5 – 24.9 is considered healthy BMI for reference. The histogram on the bottom right shows the distribution of BMI in the dataset and it appears this dataset contains people whose BMI is higher than healthy on average.

Chart

Description automatically generated

The next chart is a histogram of the costs of all 1338 records in the dataset.

Chart, histogram

Description automatically generated

From this chart you can see that the majority of records have a cost below $15,000. Unfortunately, this dataset does not contain information pertaining to what the cost is for which would could add another interesting dimension to this analysis. The purpose of this exercise is to correlate, Age, BMI, Gender, Smoking Status and Children with cost for healthcare.

**Data Preparation**

To prepare the dataset for regression I performed a couple of steps:

1. Find missing values and impute if needed – imputing will be done by the median if its needed
2. One hot encode categorical data such as Gender, Region and Tobacco. One hot encoding the Boolean values is equivalent to transforming their text based values of (yes/no) to the binary output of (1/0)
3. Use MinMax scaling to scale the numerical features between 0 and 1

The dataset contains no missing data:

Table

Description automatically generated

The output of one-hot encoding transformed the categorical data into many columns containing binary (1 / 0) output:

Table

Description automatically generated

The output of MinMaxScaling made the numerical columns between 0 and 1 which is beneficial for Machine Learning models that rely on Euclidean distances.

Table

Description automatically generated

**Regression Analysis**

Similar to last week I first chose to perform a DecisionTreeRegression which is a good model for showing the most important features when making a decision. During this analysis, I found that my intuition of smoking, high BMI and high age were important features in predicting whether the cost is higher for healthcare.

The data was split in 25% train and 75% test data. I created a tree with max depth of 4 to be able to visualize the output in a concise way.

A picture containing text, transport, lawn mower

Description automatically generated

In order of importance, Tobacco\_yes is the most important feature, followed by Age and then by BMI.

The next model I performed is a RandomForestRegressor which itself is composed of tiny decision trees organized in an ensemble to predict the final output. Each decision tree grabs a subset of features (bagging) to produce its tiny decision tree for the ensemble.

To compare the 2 models produced I chose the Mean Squared Error to compare which model performed better on the test dataset. See the output below for the Decision Tree vs. the Random Forest model. The Random Forest model has the lower of the MSEs.

Graphical user interface, text, application, email

Description automatically generated

**Conclusion**

This was a good continuation from last week’s analysis and another good example of how analytics is applied in healthcare. Cost is a big problem in the United States and this assignment showed how unhealthy lifestyles can be a contributing factor in the cost of healthcare. Higher BMI and smoking correlate with higher costs and one can draw conclusions that cessation programs and regular exercise and healthy dieting can be used to combat high healthcare costs. The best way to rein in healthcare spending is to prevent a patient from getting a chronic illness in the first place. These are interesting analyses that shed light on interesting problems and I hope to one day transform these analyses into concrete action to try to push the needle forward and help the US healthcare system deliver care more effectively.