statProject

2022-05-20

Contents

List of Figures

List of Tables

1 Introduction

A health insurance company can only make money if it collects more than what it spends on the medical care of its beneficiaries. On the other hand, even though some conditions are more prevalent for certain segments of the population, medical costs are difficult to predict since most money comes from rare conditions of the patients. The aim of this project is to first analyze the factors that influence medical costs by explorating the dataset and all the components in order to discover correlations between datas, and secondly try to build an adequate model that can accurately predict insurance costs bsed on the data and optimize its performance

2 Data Collection

2.1 Variables description

- GENDER: gender of the student (Boy/Girl);
- AGE: age of the student;
- EDUCATION LEVEL: educational institution level (School/University/College);
- INSTITUTION TYPE: type of the educational institution (Government/Non Government);
- IT STUDENT: whether the the student is an IT (Information Technology) student or not;
- LOCATION: whether the student is living in town or not;
- LOAD SHEDDING: level of load shedding, which is a way to distribute demand for electrical power across multiple power sources (High/Low);
- FINANCIAL CONDITION: financial condition of the student's family;
- INTERNET TYPE: internet type used mostly in the device;

- NETWORK TYPE: network connectivity type (3G/4G);
- CLASS DURATION: daily class duration in hours;
- SELF LMS: institution's own availability of a LMS (Learning Management System), which is a software used manage a specific learning process;
- DEVICE: device used mostly in class (Mobile, Computer, Tablet);
- ADAPTABILITY LEVEL: adaptability level of the student (High, Moderate, Low).

2.2 Descriptive statistics

2.3 Data preparation

```
data <- read.csv(file = "students_adaptability.csv")
glimpse(data)
```

```
## Rows: 1,205
## Columns: 14
## $ Gender              <chr> "Boy", "Girl", "Girl", "Girl", "Girl", "Boy", "Boy~
## $ Age                 <chr> "21-25", "21-25", "16-20", "11-15", "16-20", "11-1~
## $ Education.Level     <chr> "University", "University", "College", "School", "~
## $ Institution.Type    <chr> "Non Government", "Non Government", "Government", ~
## $ IT.Student          <chr> "No", "No", "No", "No", "No", "No", "No", "No", "N~
## $ Location            <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "~
## $ Load.shedding       <chr> "Low", "High", "Low", "Low", "Low", "Low", "Low", ~
## $ Financial.Condition <chr> "Mid", "Mid", "Mid", "Mid", "Poor", "Poor", "Mid",~
## $ Internet.Type       <chr> "Wifi", "Mobile Data", "Wifi", "Mobile Data", "Mob~
## $ Network.Type        <chr> "4G", "4G", "4G", "4G", "3G", "3G", "4G", "4G", "4~
## $ Class.Duration      <chr> "3-6", "1-3", "1-3", "1-3", "0", "1-3", "0", "1-3"~
## $ Self.Lms            <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "~
## $ Device              <chr> "Tab", "Mobile", "Mobile", "Mobile", "Mobile", "Mo~
## $ Adaptivity.Level    <chr> "Moderate", "Moderate", "Moderate", "Moderate", "L~
```

```
dim(data)
```

```
## [1] 1205   14
```

```
zero <- data[data$Class.Duration== 0,]
dim(zero)
```

```
## [1] 154  14
```

3 Exploratory Data Analysis (EDA)

3.1 Categorical variable statistics

3.1.1 Inspiration

3.1.2 Hypothesis 1

3.1.3 Hypothesis 2

3.2 Numerical variable statistics

3.2.1 Inspiration

3.2.2 Hypothesis 1

3.2.3 Hypothesis 2

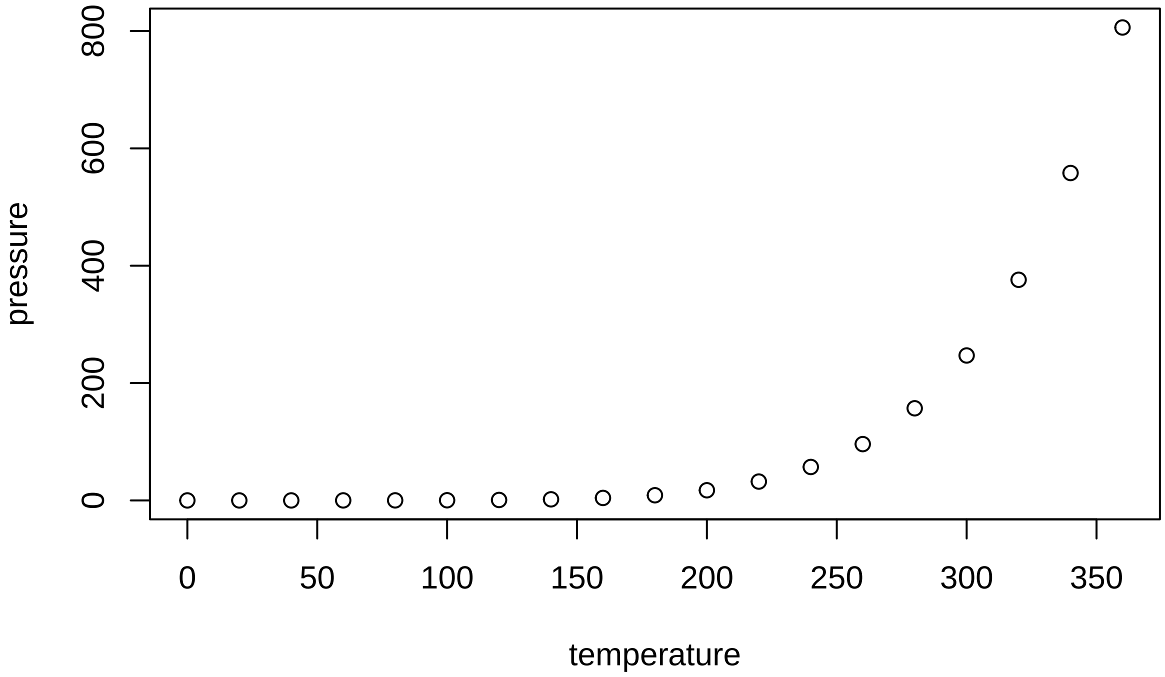
4 Model Analysis

4.1 Accuracy metrics

5 Model Evaluation

6 Conclusion

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.