Health Insurance Cost Prediction

Abstract

Insurance is a policy that helps to cover up all loss or decrease loss in terms of expenses incurred by various risks. A number of variables affect how much insurance costs. These considerations of different factors contribute to the insurance policy cost expression. Machine Learning(ML) in the insurance sector can make insurance more effective. In the domains of computational and applied mathematics the machine learning (ML) is a well-known research area. ML is one of the computational intelligence aspects when it comes to exploitation of historical data that may be addressed in a wide range of applications and systems. There are some limitations in ML so; Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry and thus it requires few more investigation and improvement. Using the machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Decision Tree Regression and Gradient Boosting Regression and also streamlit as a framework. We had used a medical insurance cost dataset for the cost prediction purpose, and machine learning methods are used to show the forecasting of insurance costs by regression model comparing their accuracies.

Introduction

We live on a planet full of threats and uncertainty. People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets. But, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to shield them. Therefore, Insurance is one of the policies that either decreases or removes loss costs incurred by various risks. The value of insurance in the lives of individuals. That's why it becomes important for insurance companies to be sufficiently precise to measure the amount covered by this specific policy and the insurance charges which must be paid for it. Various parameters or factors play an important role in estimating the insurance charges and Each of these is important. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore absolutely critical to carry out these tasks with high accuracy. So, the possibility of human mistakes is high, so insurance agents also use different tools to calculate the insurance premium. And thus, ML is beneficial here. ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on insurance data from the past. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict insurance costs. The value of insurance fees is based on different variables. As a result, insurance fees are continuous. Regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used. Preprocessing of the dataset is done first. Then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, Decision Tree Regression and Gradient Boosting Regression. It is found that the gradient boosting provides the highest accuracy with an r-squared value of 86.7853. The inclusion of a novel method of insurance cost estimation is the main goal of this work.

Need for Insurance:

Health insurance is crucial for several reasons:

- Financial Protection: Health insurance helps protect you from high medical costs. Without insurance, you would be responsible for paying the entire cost of medical services out of pocket, which can be very expensive, especially for major illnesses, surgeries, or long-term treatments. Health insurance provides coverage for these costs, reducing your financial burden.

- Access to Healthcare: Having health insurance ensures that you have access to a wide range of healthcare services and providers. It allows you to visit doctors, specialists, and hospitals, receive necessary treatments, and access preventive care such as vaccinations and screenings. Insurance coverage gives you the freedom to seek medical help when needed, without worrying about the cost.

- Preventive Care: Many health insurance plans offer coverage for preventive services, including annual check-ups, vaccinations, screenings, and counseling. Preventive care is essential for early detection and prevention of diseases, improving overall health outcomes. With insurance, you are more likely to receive timely preventive care, reducing the risk of developing severe health issues.

- Chronic Disease Management: If you have a chronic condition like diabetes, asthma, or heart disease, health insurance is vital for managing your ongoing healthcare needs. It covers regular doctor visits, medications, and treatments required to keep your condition under control. Health insurance ensures you can afford the necessary care to maintain your health and manage chronic diseases effectively.

- Emergency Medical Services: Accidents and emergencies can happen unexpectedly, and they often result in high medical bills. Health insurance provides coverage for emergency room visits, ambulance services, surgeries, and hospital stays. It offers peace of mind, knowing that you won't face exorbitant costs if you require urgent medical attention.

- Prescription Medications: Health insurance often includes coverage for prescription medications, which can be costly, particularly for long-term treatments. With insurance, you pay a reduced price for medications, making them more affordable and accessible.

- Mental Health Services: Mental health is as crucial as physical health, and many health insurance plans provide coverage for mental health services, including therapy, counseling, and psychiatric medications. Insurance coverage ensures that individuals can receive the mental healthcare they need without financial strain.

Need of Health Insurance Cost Prediction:

Health cost prediction plays a significant role in various aspects of healthcare management and planning. Here are some ways in which health cost prediction can be beneficial:

Budgeting and Financial Planning: Predicting health costs helps individuals, families, and businesses plan their budgets and allocate resources accordingly. By having an estimate of future healthcare expenses, individuals can save or purchase appropriate health insurance coverage to meet their anticipated needs. Similarly, businesses can plan for employee benefits and allocate funds for healthcare expenses in their financial plans.

Insurance Premium Calculation: Health cost prediction helps insurance companies determine premium rates for their health insurance plans. By analyzing historical data and projecting future costs, insurance providers can calculate premiums that align with expected expenses. Accurate cost prediction ensures that premiums are set at a level that covers anticipated healthcare costs, allowing insurance companies to remain financially stable while providing coverage to policyholders.

Resource Allocation: Health cost prediction is valuable for healthcare providers and organizations in managing their resources effectively. By forecasting future healthcare costs, providers can allocate resources such as staff, medical supplies, and infrastructure appropriately. For example, if there is a projected increase in demand for specific healthcare services, providers can plan to hire additional staff or invest in equipment and facilities accordingly.

Policy Planning and Decision-Making: Health cost prediction assists policymakers and government agencies in making informed decisions regarding healthcare policies and regulations. By understanding future cost trends, policymakers can design programs and initiatives to address potential cost drivers, improve cost-efficiency, and ensure sustainable healthcare financing. This information helps in making decisions related to healthcare funding, reimbursement policies, and resource allocation at a broader level.

Fraud Detection and Prevention: Health cost prediction models can be utilized to identify anomalies and patterns that indicate potential fraud or abuse in healthcare billing and claims. By analyzing historical data and comparing it with predicted costs, insurers and regulatory bodies can identify suspicious billing practices and investigate further to prevent fraud. This can help in reducing healthcare costs by eliminating fraudulent claims and ensuring that resources are used appropriately.

Population Health Management: Health cost prediction models can aid in population health management strategies. By analyzing health cost trends and patterns within a specific population, healthcare organizations can identify high-risk groups, target interventions, and allocate resources for preventive care and disease management programs. This proactive approach can help reduce future healthcare costs by focusing on preventive measures and early interventions.

Data Description

We had used a dataset from Kaggle Site for creating our prediction model. This data set includes nine attributes and the data set has splitted into two-parts : training data and testing data.For training the model, 80% of total data is used and the rest for testing.To build a predictor model of medical insurance cost the training dataset is applied and to evaluate the regression model, test set is used. The following table shows the Description of the Dataset.

Age	Age of client
Sex	Male/Female
BMI	Body Mass Index
Children	Number of children/kids the client has
Smoker	Whether a client is a smoker or not
Region	Whether the client lives in southwest, northwest, southeast or northeast
Charges	Medical Cost the client pay

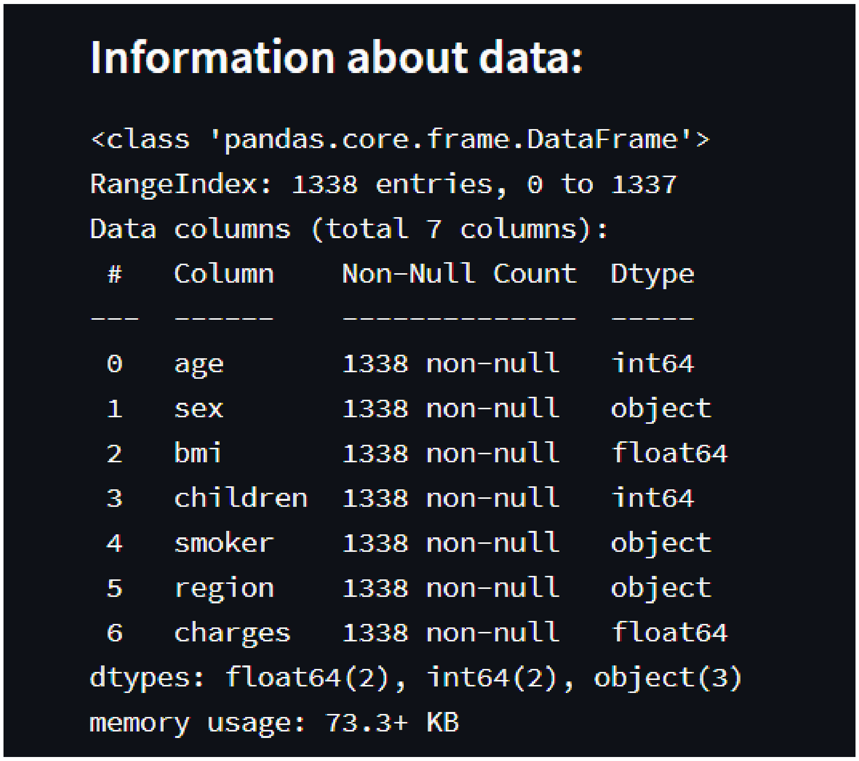
- age: age of the primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9

- children: number of children covered by health insurance, number of dependents
- smoker: smoking or not
- region: the beneficiary’s residential area in the US, northeast, southeast, southwest, northwest.
- charges: individual medical costs billed by health insurance

Since we are predicting insurance costs, charges will be our target feature.

Data Ceaning:

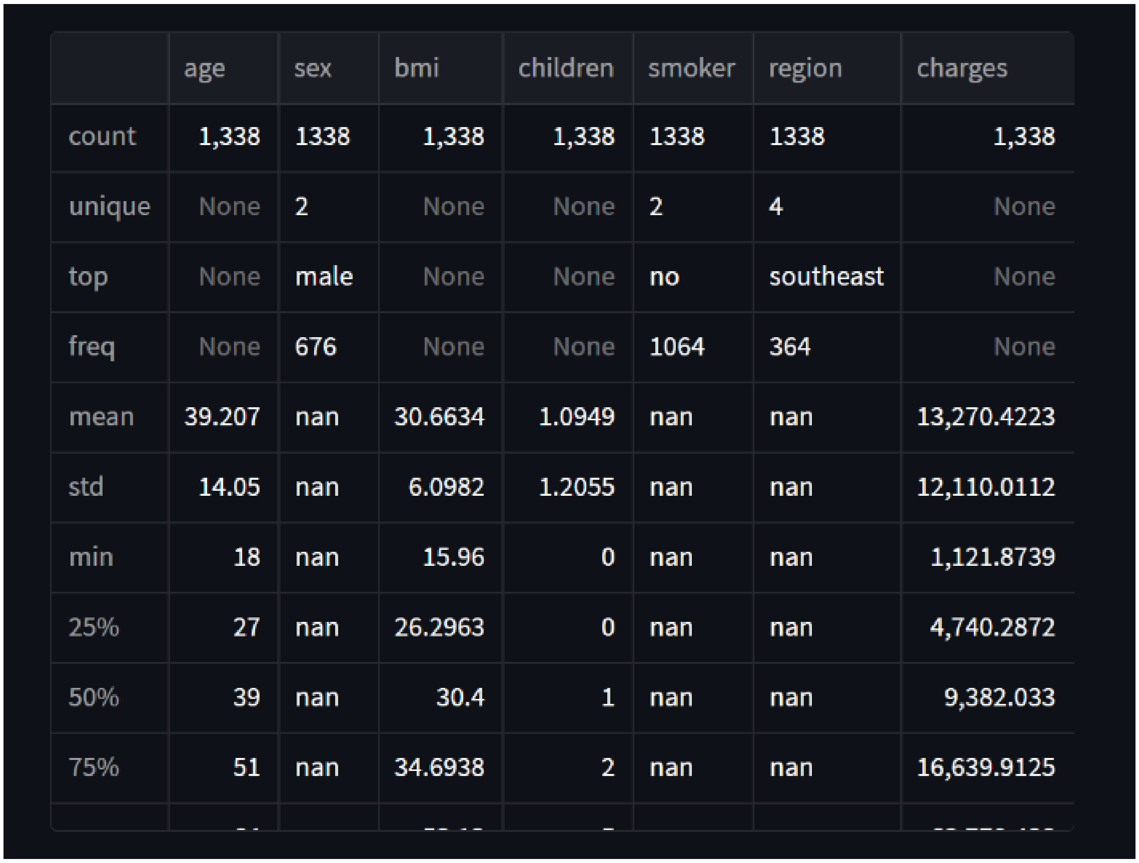
Check the info:



Information about data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Statistics of data:

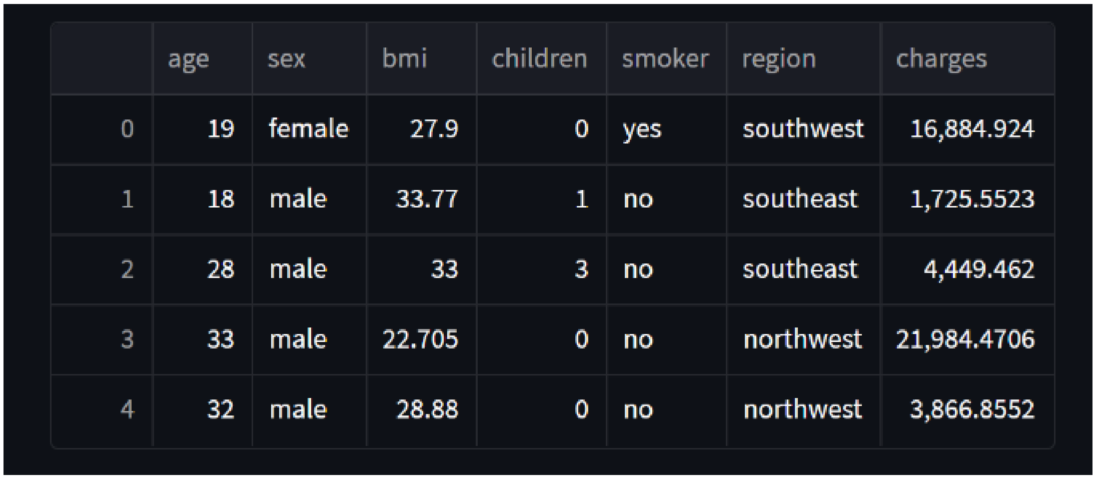


	age	sex	bmi	children	smoker	region	charges
count	1,338	1338	1,338	1,338	1338	1338	1,338
unique	None	2	None	None	2	4	None
top	None	male	None	None	no	southeast	None
freq	None	676	None	None	1064	364	None
mean	39.207	nan	30.6634	1.0949	nan	nan	13,270.4223
std	14.05	nan	6.0982	1.2055	nan	nan	12,110.0112
min	18	nan	15.96	0	nan	nan	1,121.8739
25%	27	nan	26.2963	0	nan	nan	4,740.2872
50%	39	nan	30.4	1	nan	nan	9,382.033
75%	51	nan	34.6938	2	nan	nan	16,639.9125

Here is some descriptive statistic

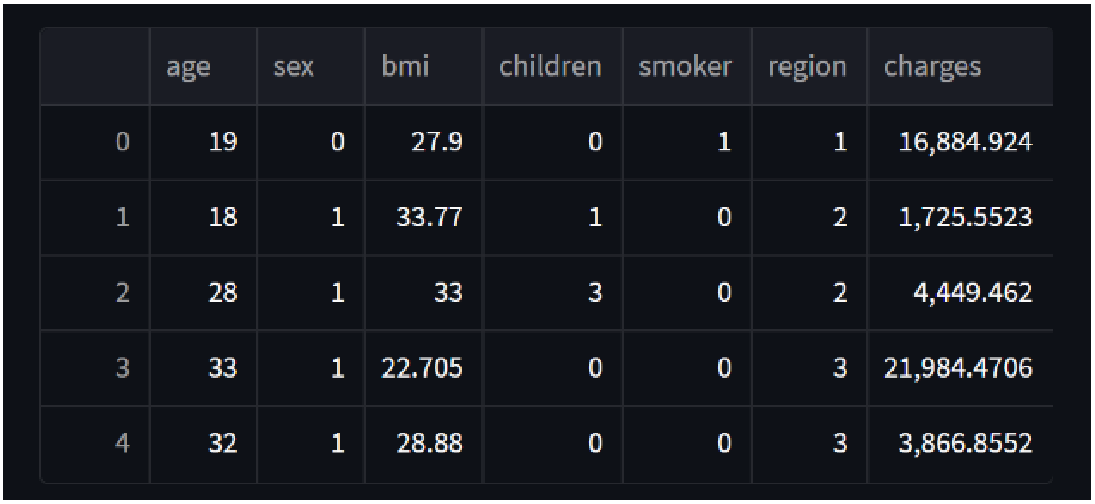
Converting string values of Columns to numerical values:

Before converting:



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16,884.924
1	18	male	33.77	1	no	southeast	1,725.5523
2	28	male	33	3	no	southeast	4,449.462
3	33	male	22.705	0	no	northwest	21,984.4706
4	32	male	28.88	0	no	northwest	3,866.8552

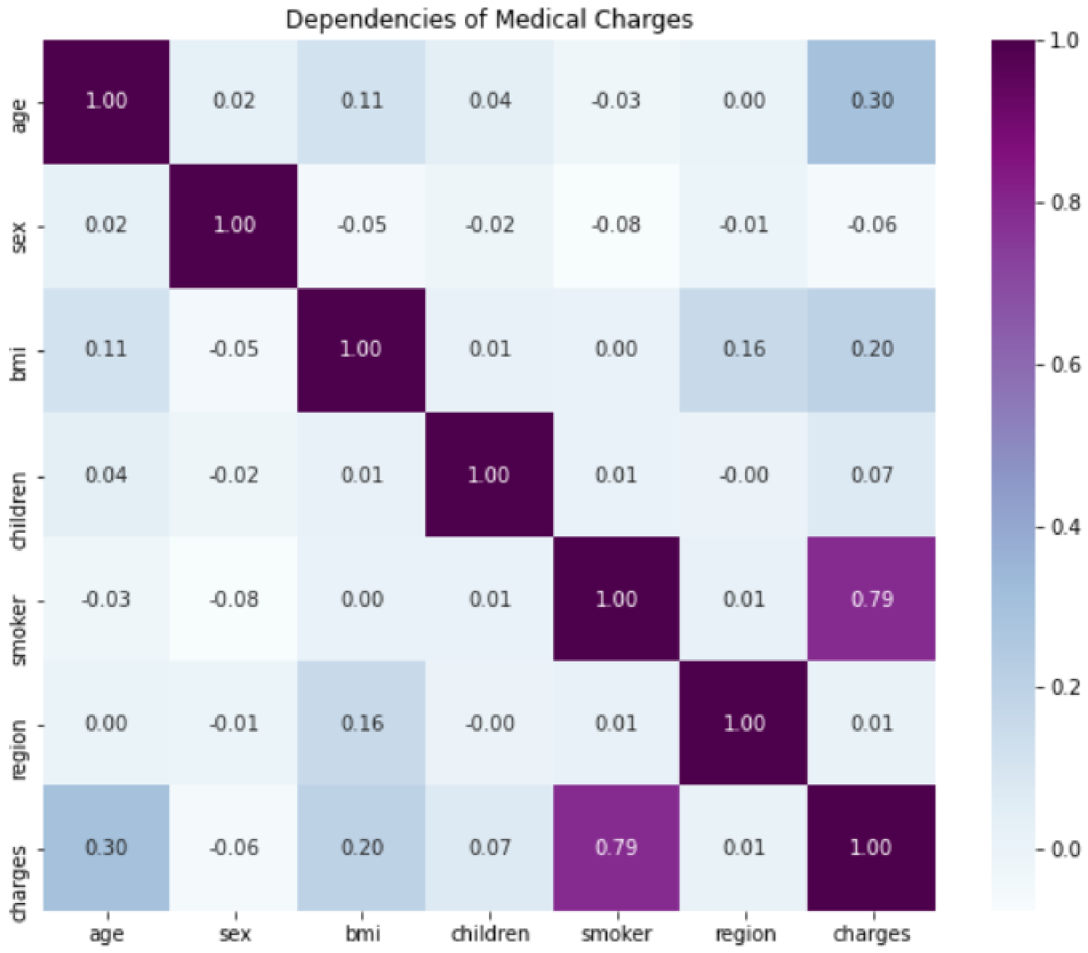
After converting:



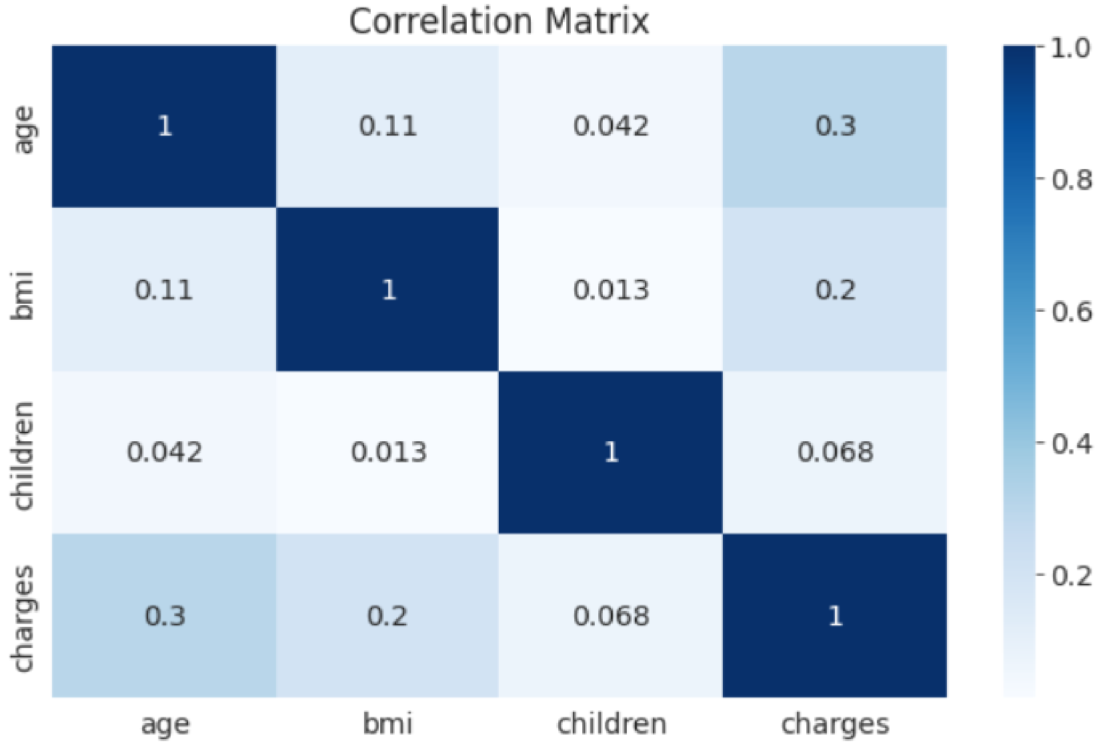
	age	sex	bmi	children	smoker	region	charges
0	19	0	27.9	0	1	1	16,884.924
1	18	1	33.77	1	0	2	1,725.5523
2	28	1	33	3	0	2	4,449.462
3	33	1	22.705	0	0	3	21,984.4706
4	32	1	28.88	0	0	3	3,866.8552

In terms of categorical features, the dataset has a similar number of people for each category, except for smoker. We have more non-smokers than smokers, which makes sense. The charges itself varies greatly from around $1,000 to $64,000.

Plotting distribution of every column and allowing user to plot their selected plots.
We will use Mean Absolute Error (MAE) as our metrics. These three metrics can be used depends on the business point of view. To see what we mean, consider the true value of one observation of charges be $10,000. Assume the model predictions are exactly the same as true values, except for this particular observation which the model predicts as x. We will vary x from $1,000 to $19,000 and see the resulted error.

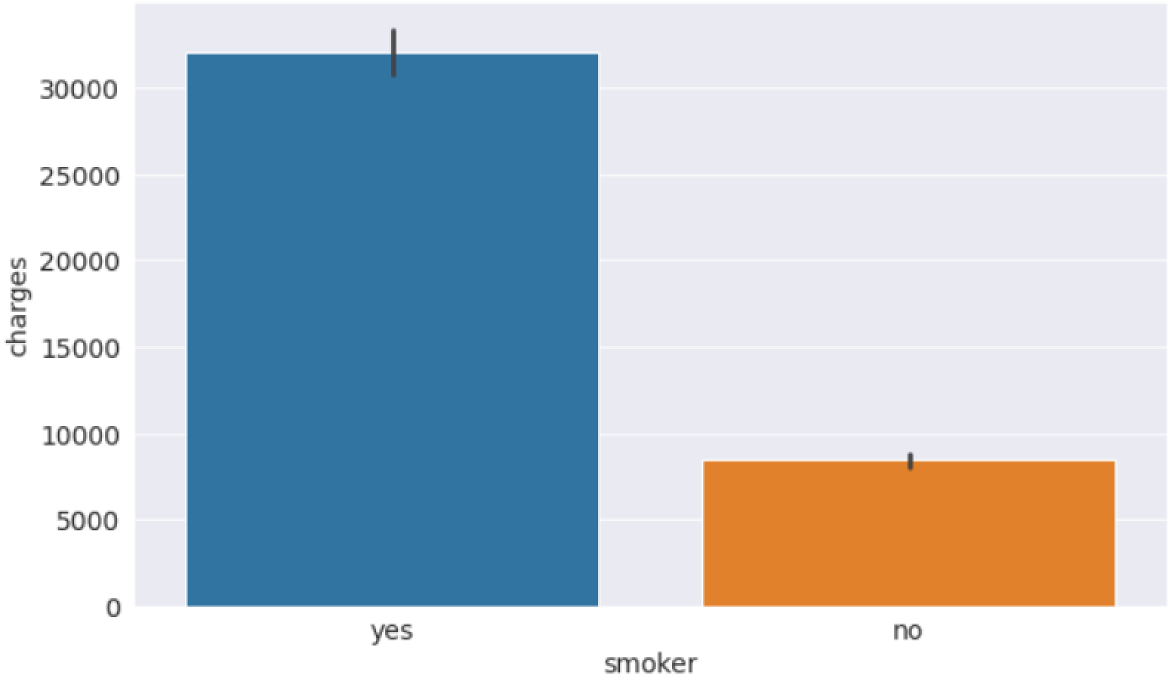


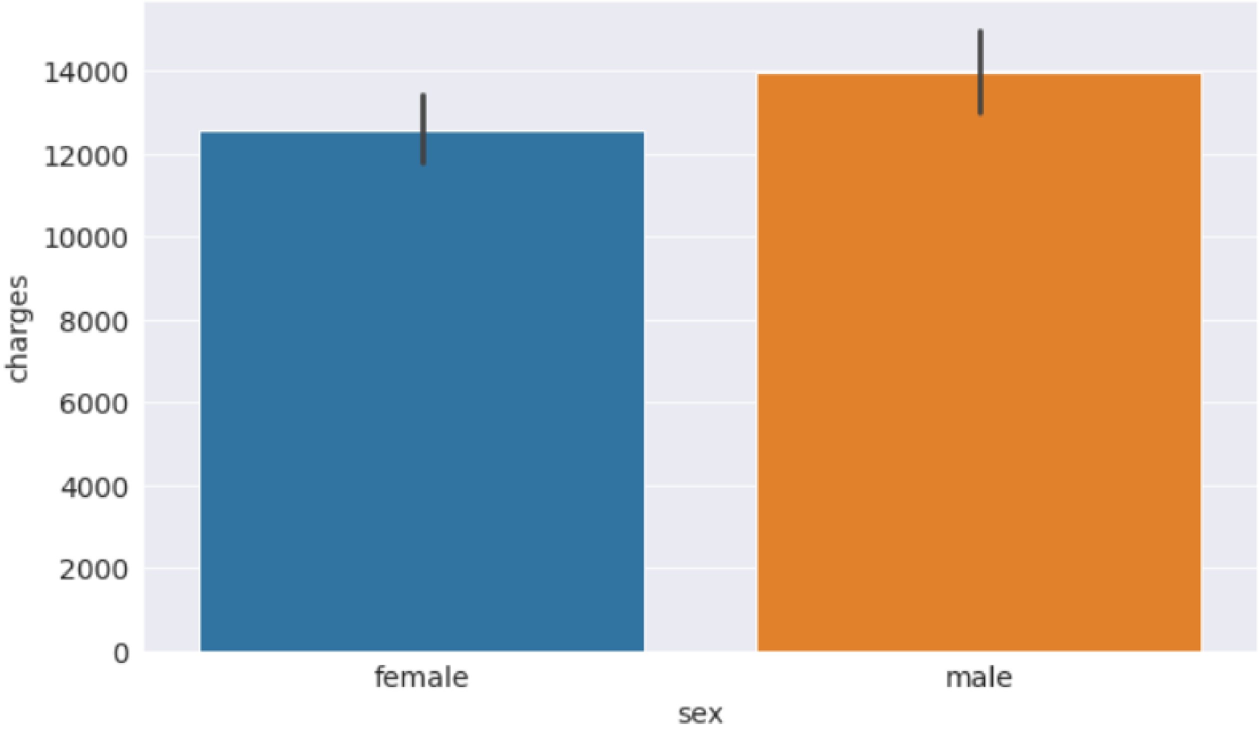
Dependencies of Medical Charges

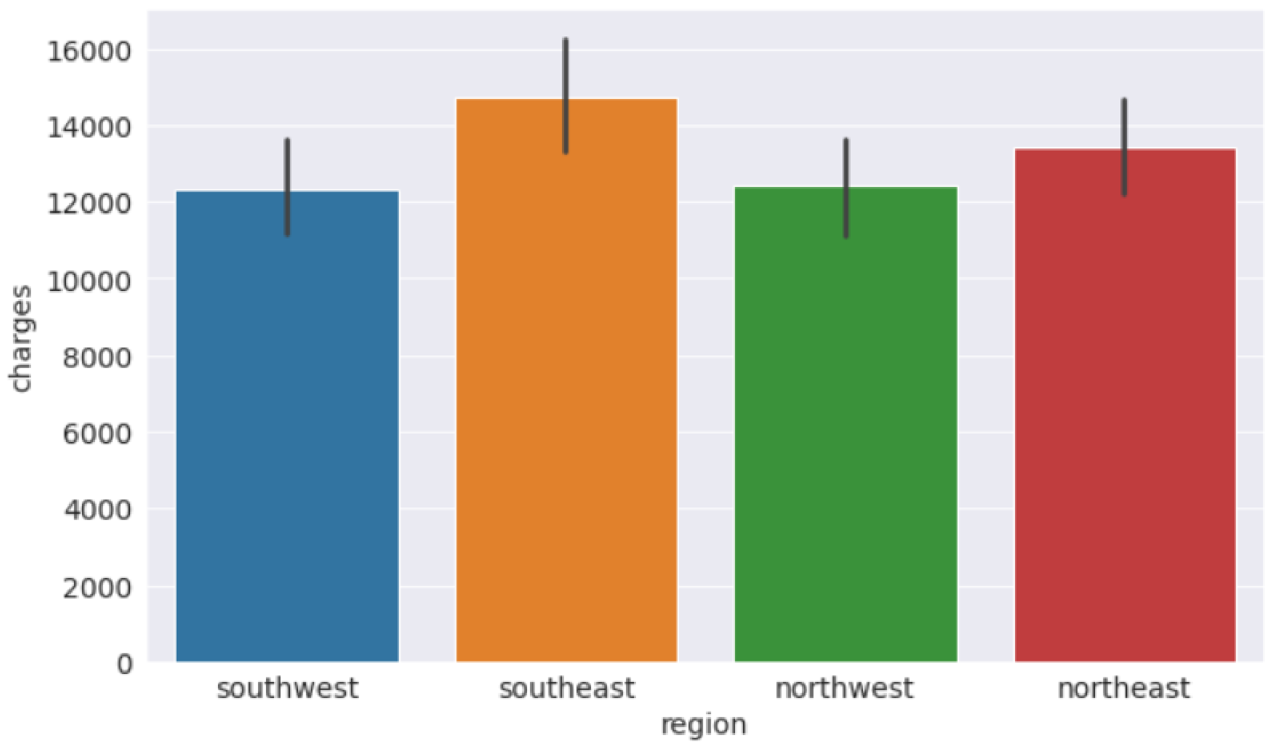


Correlation Matrix

How other factors affects charges:






16000
14000
12000
10000
8000
6000
4000
2000
0
charges
southwest
southeast
northwest
northeast
region

Gradient Boosting

Gradient boosting algorithmic program is one among the foremost powerful algorithms within the field of machine learning. As we all know that the errors in machine learning algorithms are generally classified into 2 classes i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it's accustomed to minimize bias error of the model.

Gradient boosting algorithms are often used for predicting not solely continuous target variables as a regression however additionally categorical target variables (as a Classifier). Once it is used as a regressor, the price operates as Mean square. Error (MSE) and when it is used as a classifier then the price operates as Log loss.

Implementation:

The objective of the study is to prophetic the insurance cost supported age, BMI, kid number, the region of the person living, sex, and whether or not a shopper is smoking or not, drinks alcohol or not, having diabeties or not . These options contribute to our target variable prediction of insurance costs.

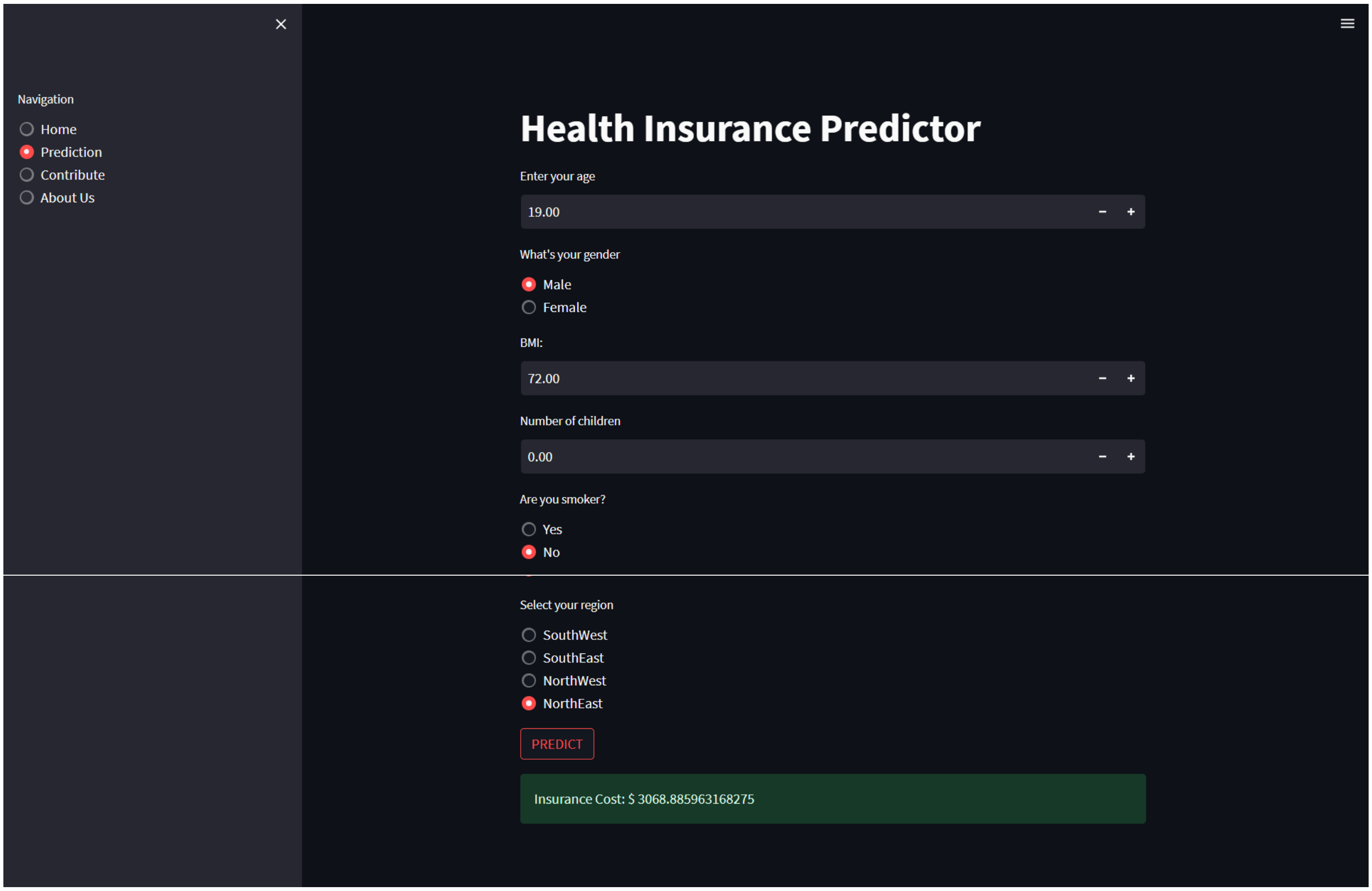
For the measuring of the value of insurance, many regression models are applied during this study. The dataset is split into 2 sections.

One half for model training and also the other part for model analysis or testing. During this study, the info set is separated into two-part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data. The accuracy is checked with the assistance of r2 score.

Model Performance:

Algorithm Used	R2 Score
Linear Regression	74.4738141
Decision Tree Regression	69.0465611
Gradient Boosting Regression	86.8600199

Outputs:



Health Insurance Predictor

Enter your age

19.00

What's your gender

● Male
○ Female

BMI:

72.00

Number of children

0.00

Are you smoker?

○ Yes
● No

Select your region

○ SouthWest
○ SouthEast
○ NorthWest
● NorthEast

Insurance Cost: $ 3068.885963168275

Health cost prediction can provide several benefits to a business:

Financial Planning: By accurately predicting health costs, businesses can plan and allocate their financial resources more effectively. This allows them to budget for healthcare expenses, set aside appropriate funds, and make informed decisions about insurance plans and employee benefits.

Cost Control: Health cost prediction helps businesses identify cost drivers and take proactive measures to contain expenses. By analyzing historical data and trends, businesses can anticipate future healthcare costs and implement strategies to mitigate them. This may involve negotiating better contracts with healthcare providers, implementing wellness programs to improve employee health, or exploring cost-effective insurance options.

Employee Benefits Management: Predicting health costs allows businesses to design and manage employee benefits packages more efficiently. It helps them determine appropriate coverage levels and select the most suitable insurance options based on projected costs. This enables businesses to provide comprehensive healthcare benefits to employees while managing the financial impact on the organization.

Risk Management: Health cost prediction helps businesses identify potential risks associated with employee health and wellness. By analyzing data and trends, businesses can assess the impact of certain health conditions or behaviors on their workforce and take preventive measures. This might involve implementing wellness initiatives, promoting healthy lifestyles, or targeting specific health conditions to reduce their prevalence and associated costs.

Employee Productivity and Engagement: Effective management of healthcare costs can lead to a healthier and more engaged workforce. Health cost prediction helps businesses identify areas where investment in preventive care and wellness programs can improve employee health and productivity. By promoting employee well-being, businesses can enhance overall productivity and reduce absenteeism, ultimately improving business performance.

Competitive Advantage: Businesses that can accurately predict and manage health costs gain a competitive edge. By optimizing healthcare expenditures, they can offer more attractive employee benefits packages, which can help attract and retain top talent. Additionally, effective cost management can result in lower operating expenses, potentially allowing businesses to offer more competitive pricing to customers.