Analysis and Predictive Modelling for Health Insurance Charges

Project proposal by Group 9

BUAN 6337.006 Predictive analytics for Data Science – Spring 23

Objective

The aim of the project is to help provide the product owner and/or pricing manager of an insurance company with data backed insights on key characteristics of the customer to consider while segmenting and targeting their customer base. We will also try to predict the appropriate insurance charges based on the available data for the prospective customer. This could help the insurance firm find the right premium rate for customers helping the company to control risk undertaken.

For this project, we are using the US Health Insurance Dataset taken from Kaggle. In the research report we will be analyzing an insurance dataset that includes details on the age, BMI, number of children, and smoking habits of a sample of Americans. This study's main goal is to investigate the connections between these factors and how they affected health results. To find significant associations and investigate the predictive potential of each variable, we carried out a series of statistical analyses, including correlation analysis and linear regression models. According to our research, smoking and BMI are strongly linked to several negative health outcomes, including a higher chance of developing long-term conditions like diabetes and heart disease. Age and the number of offspring were also discovered to be important predictors of health outcomes, even though the relationships were less consistent across various health indicators. Overall, our findings emphasize the significance of considering a variety of variables, such as lifestyle and demographic characteristics, when determining health risks and creating interventions to enhance health outcomes. The importance of reliable, high-quality datasets in advancing our comprehension of complex health issues and guiding public health policy is underlined by this research.

Data description and preparation

Data source https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

Data format

The insurance costs are provided against the insured's age, sex, BMI, number of children, smoking status, and region in the 1338 rows of insured data in this dataset. We will be using all the data to perform analysis as all the columns are relevant and the size of the data can be easily handled.

Data description

The following factors are included in the dataset:
age: The main beneficiary's age, expressed in years (numeric)
sex: The main beneficiary's gender (categorical: male, female)
BMI: The main beneficiary's body mass index (numeric)
children: number of toddlers whose insurance is provided (numeric)
smoker: Regardless of whether the main benefactor smokes (categorical: yes, no)
region: the US neighborhood where the recipient resides (categorical: NE, SE, SW, NW)
charges: Health insurance charges for individual medical insurance (numeric)

Data cleaning

The dataset contains no missing or undefined numbers. We would be using the entire dataset with some modifications as needed in the data preparation stage.

Data preparation

Multiple categorical variables, including "sex," "smoker," and "region," are present in the US Premium dataset. Two common techniques for working with categorical data are One hot encoding and label encoding. We will be using both based on the need.

Description of the summary statistics for the data

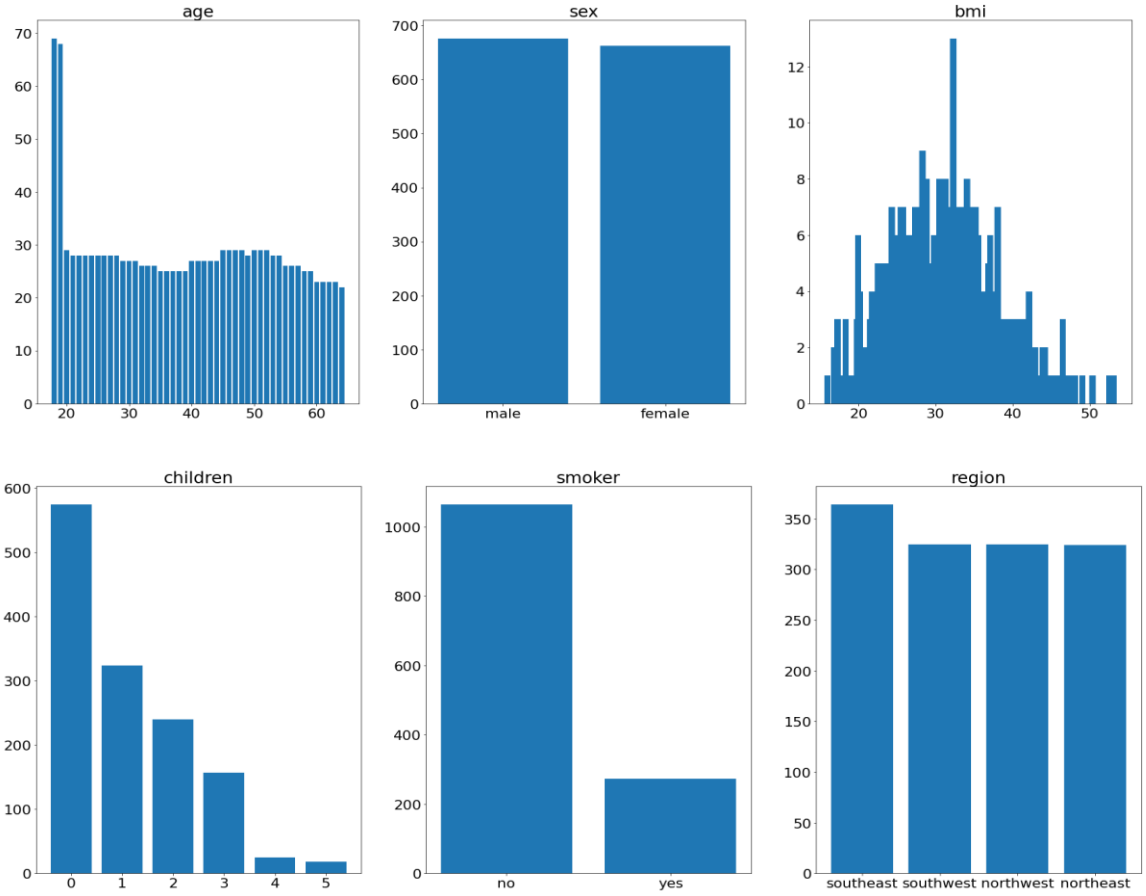
```
In [83]: df.describe()
```

Out[83]:

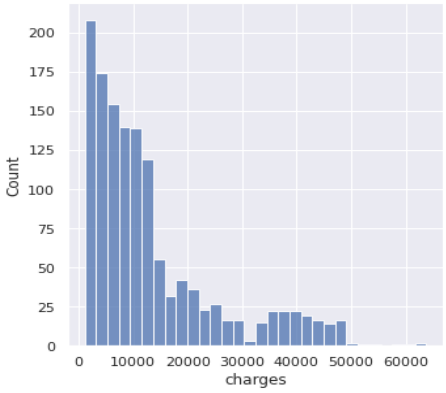
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Data visualization and exploratory analysis

Histograms were used as the first step of EDA to understand the distribution of data. Below are two subplots for the frequency distributions of explanatory variables.

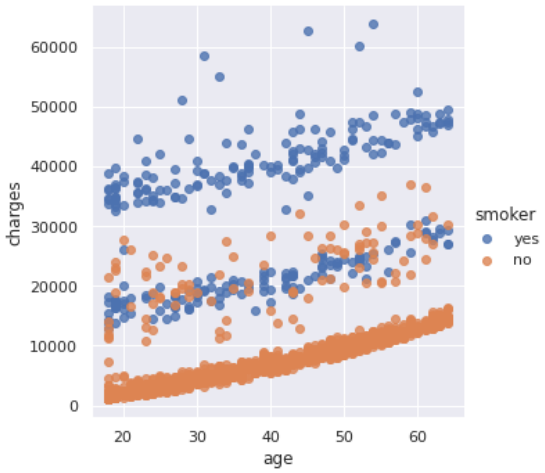


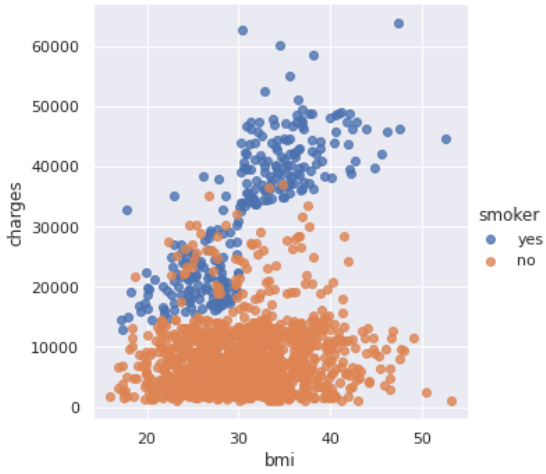
And the distribution of the explained variable – insurance charges is as follows:

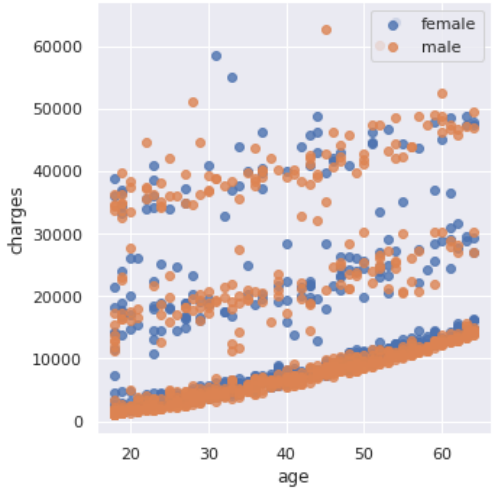


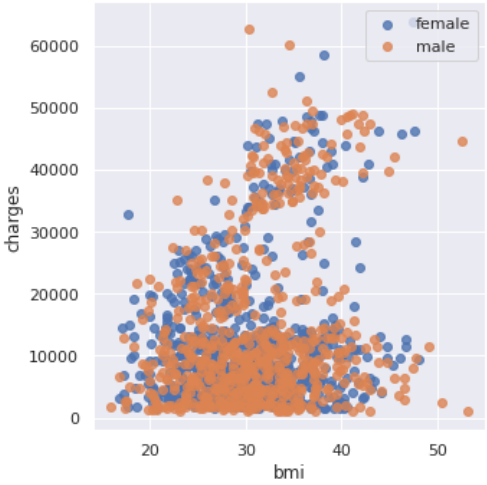
We then plotted scatter plots to understand how the variables ‘age’ and ‘bmi’ vary with charges. General assumption is that chargers are high for old aged people and people with high bmi. We wanted to check if the general assumption holds true for this dataset and wanted to check how the trends are with are varying especially amongst smokers and non-smokers and males and females.

Here are plots:

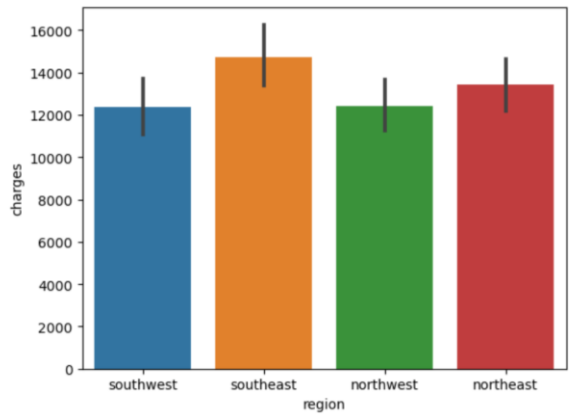


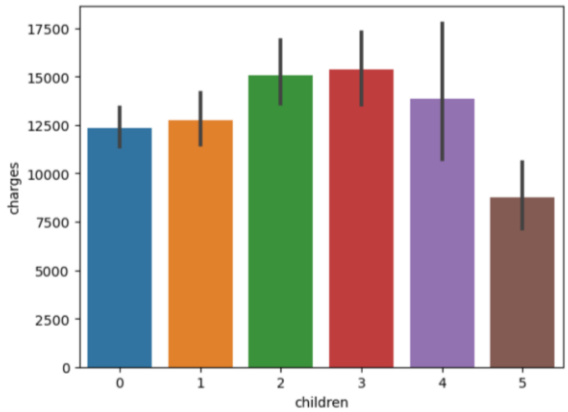






We also wanted to see if region or children plays any role on the insurance premium





Conclusion

In conclusion, our analysis of the US Health Insurance Dataset has provided insights into the key characteristics of customers that insurance companies can consider while segmenting and targeting their customer base. Our findings suggest that smoking and BMI are strongly linked to negative health outcomes and significantly impact insurance charges, while age and the number of children in the household are also important predictors of health outcomes. However, we also found that the impact of gender on insurance charges was not significant. Additionally, we observed that there is some variation in insurance charges based on region and the number of children, but further analysis is needed to determine if these differences are statistically significant.

Based on these preliminary observations we would want to build a multiple linear regression predictive model to be able to quantify the impact of each of these variables on insurance charges/outcome variable. We would evaluate our model using key performance metrics, i.e., RMSE, MSE, MAE, r2, and adjusted r2. We also want to conduct statistical tests to determine the significance of the differences in insurance charges based on region and the number of children. By doing so, we aim to provide insurance companies with data-backed insights to help them determine the right premium rate for customers and control the risks associated with providing insurance.