Private Health Insurance Analysis

Minali, Gulcan, Irina
29 Aug 2023

Project Goal

Our goal is to employ Machine Learning to forecast Health Insurance charges for a business named "Happy Life Health Cover" in the US by analyzing a dataset that contains information on several factors that impact medical expenses and insurance premiums in the country.

Introduction

What is Health Insurance?

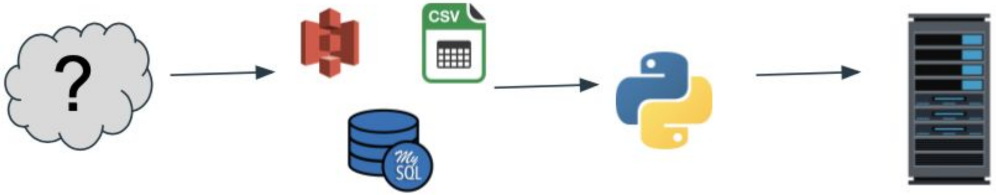
- Health insurance is a type of coverage that helps individuals pay for medical expenses and services.
- Premiums are the regular payments individuals or employers make to maintain health insurance coverage.
- Health insurance plans differ in coverage levels, benefits, limits, and exclusions.

Relevance of Machine Learning in predicting Health Insurance costs

- Able to analyze large amounts of data and identify patterns, correlations, and trends to estimate or forecast future costs.
- By using historical data and relevant variables, Machine Learning algorithms can learn from patterns and make predictions about future costs.
- Machine Learning models can adapt and improve over time as they receive more data and feedback, leading to more accurate cost predictions.

Technologies Used

- Pandas
- Matplotlib
- Seaborn
- Numpy
- Scikit-learn
- Sklearn
- Tableau



Data Collection and Cleaning

1.Data Source

https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning

2. Data Cleaning

The data from Kaggle contained 1,000,000 lines of entries.

The data was checked for empty and duplicate entries.
Attributes with categorical data were checked for unique values.
Each attribute contained acceptable number of categories for our analysis.

Machine Learning Algorithms Used

❖ **The following models were used:**

- Linear Regression
- Decision Tree Regressor

Decision Tree Regressor Data preparation
Data was separated as noted below and OneHotEncoder was used on categorical data.

Features	Target
•Numerical Data:	•Charges
•Age	
•bmi	
•children	
•Categorical data	
•gender	
•smoker	
•region	
•medical_history	
•family_medical_history	
•exercise_frequency	
•occupation	
•coverage_level	

Machine Learning - Linear Regression

As seen above,the difference between Root Mean Squared Error and Mean Squared Error for Training and testing values is low ,it can be concluded that the data is slightly overfitting.

```
Training Data Score for Linear Regression: 0.9957266917943283
Testing Data Score for Linear Regression: 0.9957191270914253
Mean Squared Error for Linear Regression for Testing Values: 83546.37958252191
Root Mean Squared Error for Linear Regression for Testing Values: 289.0439059771403
Mean Squared Error for Linear Regression for Training Values: 83308.53864396818
Root Mean Squared Error for Linear Regression for Training Values: 288.63218573812617
R2 score for Linear Regression: 0.9957191270914253
```

	Prediction for Linear Regression	Actual for Linear Regression
0	12379.274170	12481.068956
1	18784.150635	18299.071994
2	18862.621338	18846.795608
3	21283.642822	21597.663069
4	25182.140869	25596.721389

Machine Learning - Tree Regression

As seen above,the difference between Root Mean Squared Error and Mean Squared Error for Training and testing values is very high ,it can be concluded that the data is overfitting.

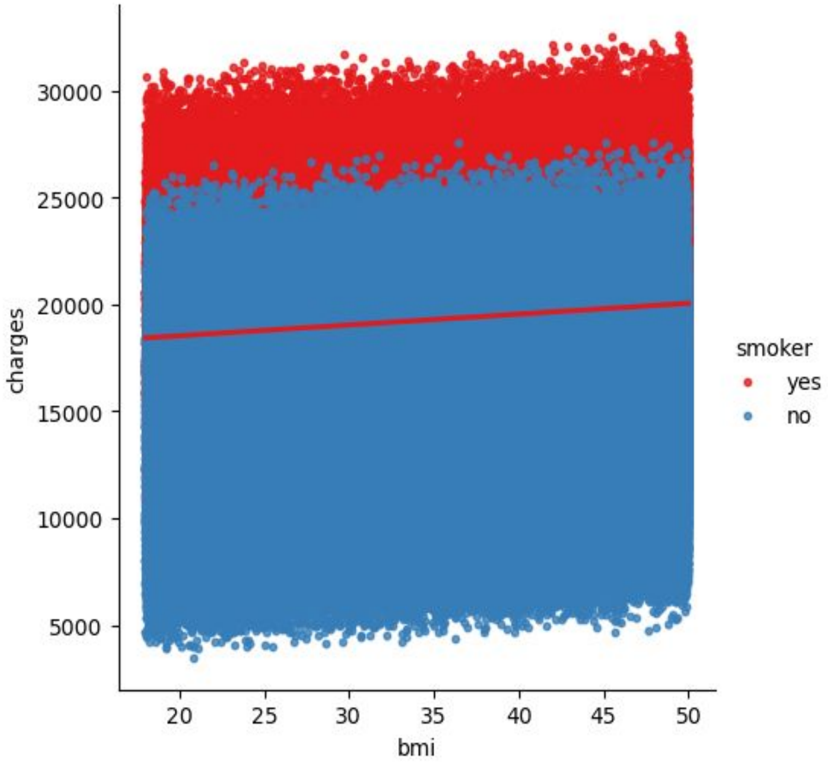
```
Training Data Score for Decision Tree Regression: 0.999999805257673
Testing Data Score for Decision Tree Regression: 0.9868414139129958
Mean Squared Error for Decision Tree Regression: 256805.6215338956
Root Mean Squared Error for Decision Tree Regression: 506.7599249485851
Mean Squared Error for Decision Tree Regression for Training Values: 83308.53864396818
Root Mean Squared Error for Decision Tree Regression for Training Values: 288.63218573812617
R2 score for Decision Tree Regression: 0.9868414139129958
```

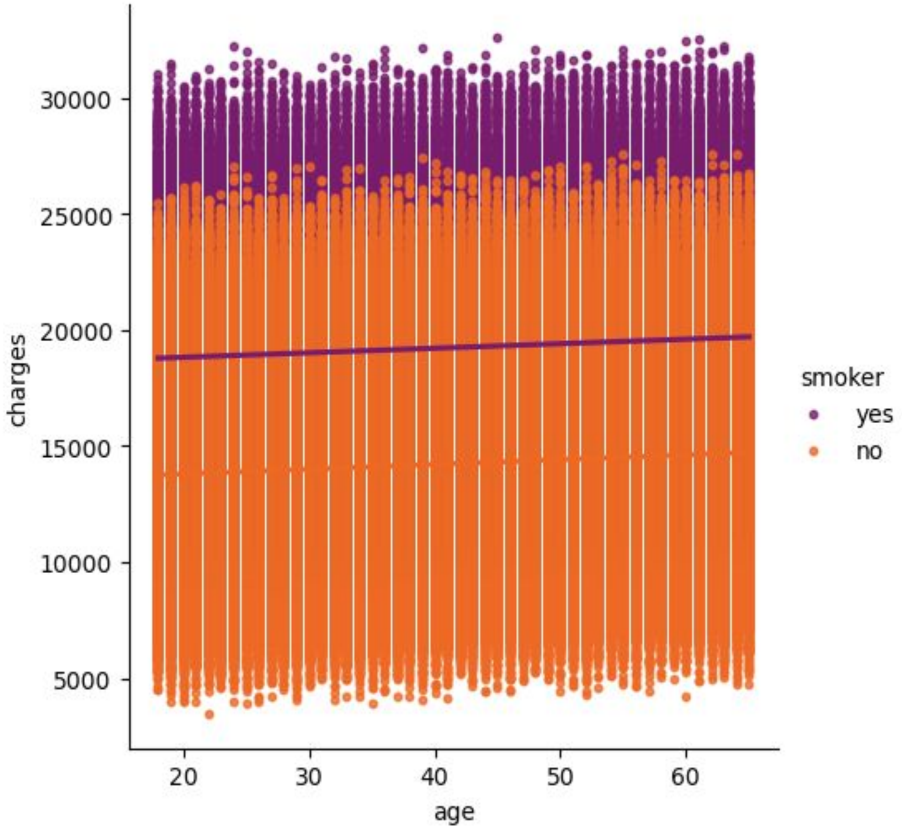
	Prediction for Decision Tree Regression	Actual for Decision Tree Regression
0	12608.414913	12481.068956
1	19359.470966	18299.071994
2	19172.322317	18846.795608
3	21409.863225	21597.663069
4	25362.394352	25596.721389

Conclusion

A Machine Learning model offers swift and precise predictions for specific scenarios. In our scenario, it generated insurance premiums not just for the customer's selected plan option, but also for alternative choices.

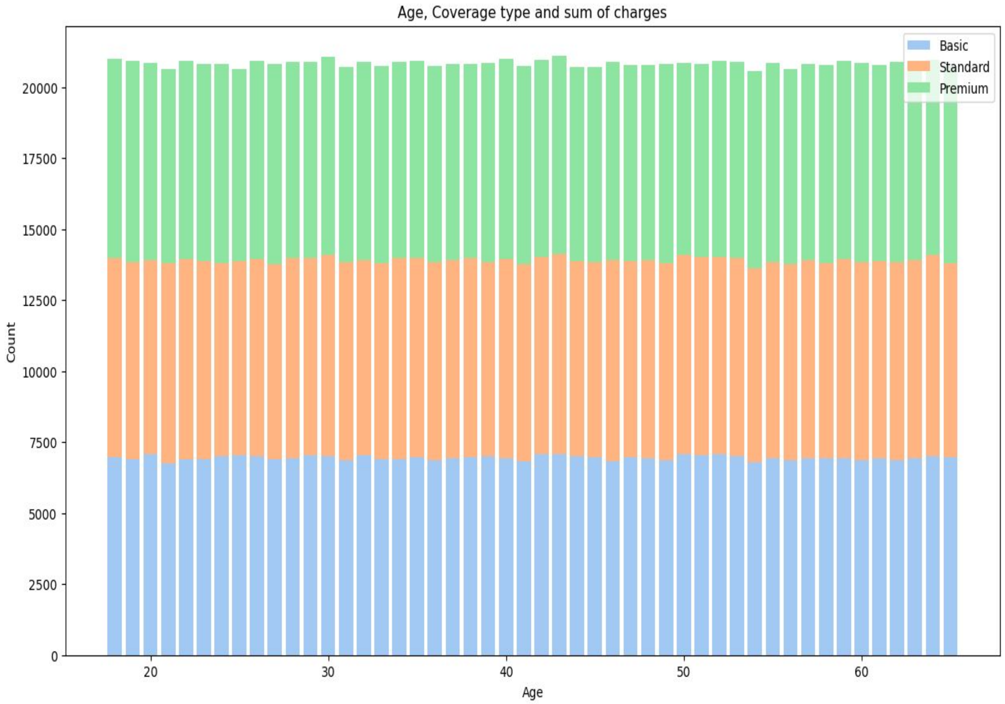
Charges by BMI, Age and Smoker Status



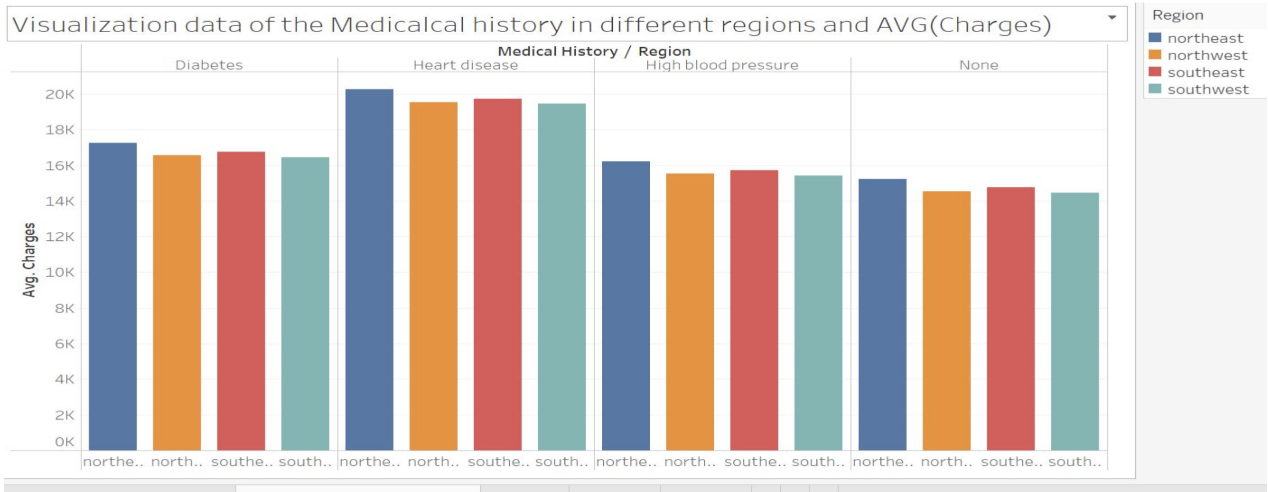


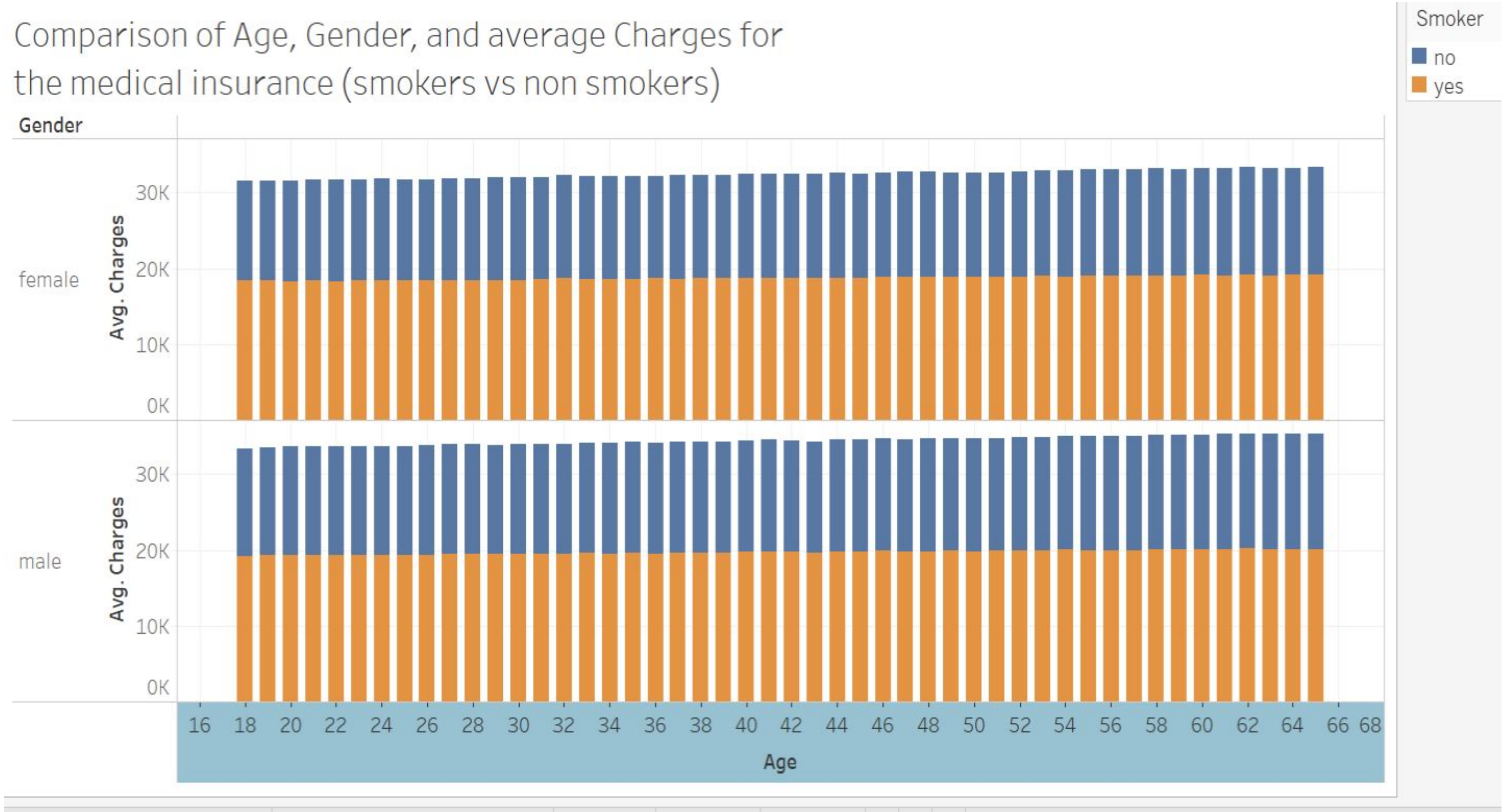
Coverage Level by Age Group

The chart provides an overview of three distinct coverage types: Basic, Standard, and Premium. Among these, the Basic coverage stands as the most cost-effective option, while the Premium coverage commands a higher charge compared to the others. Although there is a marginal fluctuation in charges across different age groups, this variation demonstrates minimal impact on the overall pricing structure.



Medical History in Regions





Comparison of Age, Gender, and average Charges for the medical insurance (smokers vs non smokers)
Smoker
no
yes
Gender
female
male
Avg. Charges
0K
10K
20K
30K
Age
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68

Thank You!