Private Health Insurance Analysis

Ray, Ina, Laszlo
16 May 2023

Project Aim:

Our goal is to employ Machine Learning to forecast Health Insurance prices 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.

Technologies Used:

- *Python modules and libraries*
  - *flask*
  - *flask-sqlalchemy*
  - *sqlalchemy*
  - *pandas*
  - *scikit-learn*
  - *joblib*
  - *json*
  - *numpy*
  - *Sklearn*
- *SQLite*
- *HTML / Css / JavaScript / Bootstrap*
- *D3.plotly*
- *Chart.js*

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.*

Data Collection, Pre-processing:

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*

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	

Findings:

```
LinearRegression() :
Training Data Score: 0.9957267068571238
Testing Data Score: 0.9957190785263942
      Prediction        Actual
0   12378.602826   12481.06896
1   18783.920686   18299.07199
2   18862.831011   18846.79561
3   21283.839914   21597.66307
4   25182.982849   25596.72139


DecisionTreeRegressor() :
Training Data Score: 0.9999998052576735
Testing Data Score: 0.9868460216601581
      Prediction        Actual
0   12608.41491   12481.06896
1   19359.47097   18299.07199
2   19172.32232   18846.79561
3   21409.86323   21597.66307
4   25362.39435   25596.72139
```

Prediction:

Model used: model_DecisionTreeRegressor.joblib

Data Input for Prediction
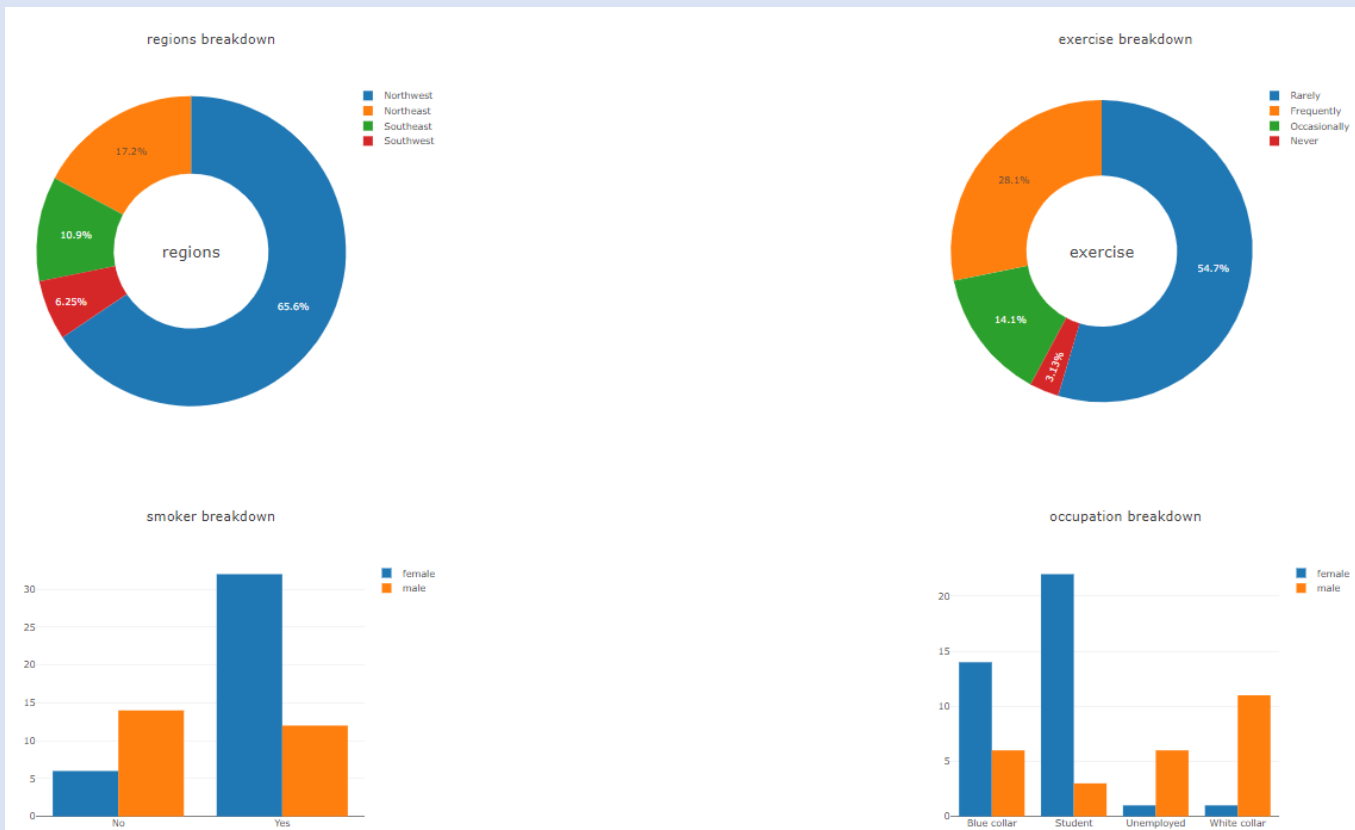- HTML form using flask
- Data submitted saved in SQLite

Analysis of Submitted Data:

SQLite data was used for business opportunity analysis

Limitations & Challenges:

- *Large dataset*
- *Dataset appeared to be fabricated to create uniform dataset for all categories.*
    - *Issue for data analysis, not for machine learning*

Conclusion:

- *A Machine Learning model provides fast and accurate prediction for a given situation.*

- *In our case it provided the insurance premium not only for the plan option chosen by the customer, but also for alternative options as well.*