Insurance Pricing Forecast Using XGBoost Regressor
Project Overview

Overview

Insurance companies cover expenses the policyholder incurs from damages to health or property policies commonly offered: medical bills, house, motor vehicle, and fire insurance, and financial losses such as a loss of income against a fee or premium paid by the client. Traditional approaches to premium calculation require a lot of time-consuming human labor and are getting more complicated daily to capture the increasingly complex interactions in the data.

Insurance firms should normally collect a higher premium than the amount given to the insured individual if that person files a valid claim to generate a profit. Since profitability is the fundamental factor that helps the insurance firm survive, they need a mechanism for reliably forecasting healthcare expenses.

Hence, our goal is to build a machine learning model that helps establish the rates by predicting the charges or payouts done by the health insurance firm to maintain profitability.

In this project, we will primarily focus on building an XGBoost Regressor to determine healthcare expenses based on features such as age, BMI, smoking, etc. We will also learn about categorical correlation, build a linear regression model as a baseline and compare it with the results of the XGBoost Regressor. We will eventually learn how to communicate technical results to stakeholders who are not technical.

Aim

This data science project aims to build and evaluate linear and xgboost regression models and determine the healthcare charges of each customer. This analysis will help the insurance firm to strategize a premium plan that will help maximize the profits.

Data Description

The insurance price forecast dataset contains historical records for 1338 insured customers. The column definitions are below

- age: Age of the primary beneficiary.
- sex: Gender of the primary beneficiary.
- BMI: Body mass index of primary beneficiary
- children: Number of children the primary beneficiary has.
- smoker: Whether the primary beneficiary smokes.
- region: The primary beneficiary's residential area in the US.

- charges: Individual medical costs billed by health insurance.

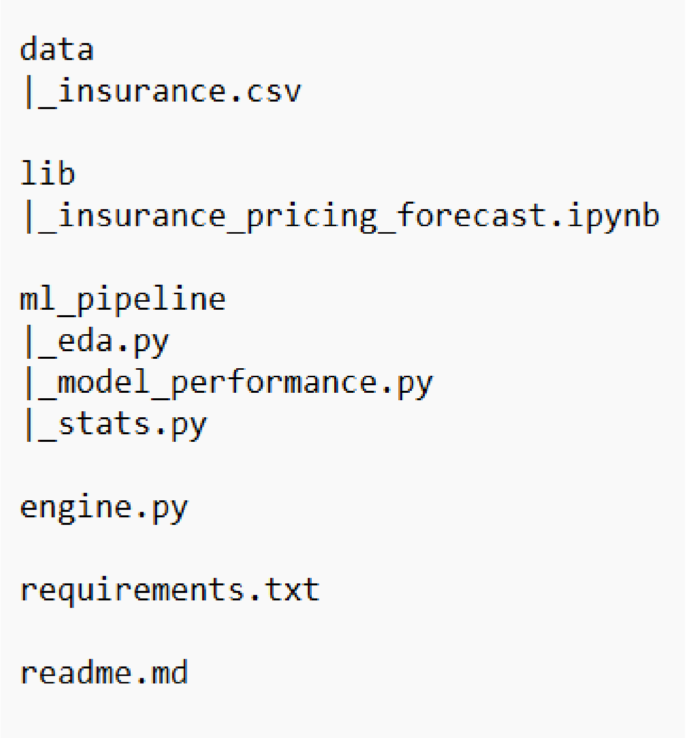
Tech Stack

- ➔ Language: Python
- ➔ Libraries: pandas, numpy, matplotlib, plotly, statsmodels, sklearn, xgboost, skopt

Approach

- Exploratory Data Analysis (EDA)
    - Distributions
    - Univariate Analysis
    - Bivariate Analysis
    - Correlation
        - Pearson Correlation
        - Chi-squared Tests
        - ANOVA
- Build and evaluate a baseline linear model
    - Linear regression assumptions
    - Data preprocessing
    - Model training
    - Model evaluation
        - RMSE
- Improve on the baseline linear model
    - Introduction to a non-linear model - XGBoost
    - Data preprocessing
    - Using Sklearn's `Pipeline` to optimize the model training process
    - Model evaluation
        - RMSE
    - Comparison to the baseline model
- Presenting the results to non-technical stakeholders

Modular code overview:



```
data
|_insurance.csv

lib
|_insurance_pricing_forecast.ipynb

ml_pipeline
|_eda.py
|_model_performance.py
|_stats.py

engine.py

requirements.txt

readme.md
```

Once you unzip the modular_code.zip file, you can find the following folders.

1. data
2. lib
3. ml_pipeline
4. engine.py
5. requirements.txt
6. readme.md

1. The lib folder is a reference folder and contains the original ipython notebook as in the lectures.

2. The ml_pipeline folder contains all the functions put into different python files, which are appropriately named. The engine.py script then calls these python functions to run the steps in one go to train the model and print the results.

3. The requirements.txt file has all the required libraries with respective versions. Kindly install the file using the command **pip install -r requirements.txt**

4. **All the instructions for running the code are present in readme.md file**

Project Takeaways

1. Understanding the insurance pricing problem statement
2. Exploratory Data Analysis on Categorical and Continuous Data
3. Univariate Data Analysis
4. Bivariate Data Analysis
5. Understand Correlation Analysis
6. Categorical Correlation with Chi-squared
7. Correlation between Categorical and Target Variables with ANOVA
8. Label Encoding for Categorical Variables
9. Understanding Linear Regression Assumptions
10. Implementing Linear Regression
11. Validating Linear Regression Assumptions
12. Understanding XGBoost Regressor
13. Implementing XGBoost Regressor
14. Building pipelines with Sklearn's Pipeline operator
15. Implementing BayesSearchCV for XGBoost Hyperparameter Optimization
16. Evaluating Models with Regression Metrics - RMSE
17. Presenting Non-Technical Metrics for Stakeholders