Homework 3 - Regression Challenge

Guidelines and Instructions

Please read the following instructions carefully before starting your assignment.

- Download the following datasets from Blackboard:
  - `health_train.csv` (Training Data with target)
  - `health_test_features.csv` (Test Data features only)
- You must submit two separate files to Blackboard before Monday, December 15th at 11:59pm (Midnight)
  1. Python Script `.py`: A clean and runnable Python script containing the code that generates your final prediction file.
  2. Prediction File (`.csv`): A CSV file named predictions.csv containing your predicted premiums.
- Policy regarding the use of AI tools (e.g., Copilot, ChatGPT).
  - AI tools cannot be used for generating the entire content of your work. You are permitted to use them as a tool to help you debug code or understand concepts.
  - The primary goal of this assignment is to develop your own critical thinking and analytical process. Relying solely on an LLM to provide a solution will deprive you of this experience.
  - A randomly-selected student will be invited to give a brief presentation of his/her methodology in the next session. A failure to show a full understanding of your own submission will result in a big penalty.

How to create your `predictions.csv` file?

Your final CSV file must contain a single column with the same number of rows as the test dataset (`health_test_features.csv`). The file should contain your predicted continuous values (floats) with no header.

You can use the following Python snippet (using `pandas`) to save your predictions. Assume `my_predictions` is a NumPy array or list containing your predicted premiums.

```
import pandas as pd
import numpy as np

# Example where my_predictions is your predictions of the premiums
# Ensure it has the same length as the test set

df_to_submit = pd.DataFrame(my_predictions)

# Save without header and without index
df_to_submit.to_csv('predictions.csv', header=False, index=False)
```

Introduction

You are a data scientist working for a private health insurance provider. The team wants to update their pricing model to better reflect the risk profile of new customers. Your task is to analyze historical data and build a predictive model that estimates the annual premium for a client based on their health metrics and demographics.

Data Description

The dataset contains 6 features describing the beneficiary.

Explanatory variables:

- `Age_Years`: Age of the primary beneficiary.
- `Gender`: Gender of the beneficiary (*F*, *M*).
- `Body_Mass_Index`: BMI (kg/m^2).
- `Dependents`: Number of children or dependents covered by the insurance plan.
- `Tobacco_User`: Whether the beneficiary is a smoker (*Y*, *N*).
- `Residential_Area`: The anonymized geographic zone of the beneficiary (*Metro_A*, *Metro_B*, *Rural_A*, *Rural_B*).

Dependent variable (Target):

- `Annual_Premium`: The annual insurance premium charged to the customer.

Your Task

Your task is to build a regression model to predict the `Annual_Premium` for the customers in the test set.

Grading

Your grade will be determined by comparing your predicted premiums against the true, hidden values using the Root Mean Squared Error (RMSE). The RMSE measures the standard deviation of the prediction errors.