Smart Premium: Insurance Premium Prediction App

Project Overview

SmartPremium is a machine learning project designed to predict insurance premiums for customers based on personal, financial, and policy-related data. The project leverages **XGBoost** and other regression models for accurate prediction and provides a user-friendly **Streamlit web application** for real-time premium estimation.

Problem Statement

Insurance premiums depend on factors such as age, income, health status, occupation, location, and claim history. The goal of this project is to **build an ML model that predicts insurance premiums** accurately based on these customer and policy attributes.

Project Approach

Step 1: Understanding the Data

- Load and inspect the dataset.
- Identify **numerical, categorical, and text features**.
- Explore missing values, inconsistencies, and skewed distributions.
- Perform **EDA** using visualizations to understand relationships between features.

Step 2: Data Preprocessing

- Handle missing values (median for numerical, mode for categorical).
- Encode categorical variables using **One-Hot Encoding**.
- Split dataset into **training (80%)** and **evaluation (20%)** sets.
- Apply **scaling** for numerical features to standardize ranges.

Step 3: Model Development

- **Regression models used:**
    - Linear Regression

    - Decision Tree Regressor
    - Random Forest Regressor
    - **XGBoost Regressor** (Best model)
- Evaluate models using **RMSE, MAE, R^2, RMSLE**.
- Select **XGBoost** as the final model for deployment.

Step 4: ML Pipeline & MLflow Integration

- Build a **pipeline**: preprocessing → training → evaluation.
- Track experiments with **MLflow**:
    - Log model parameters, metrics, and versions.
    - Store trained model for deployment.

Step 5: Model Deployment with Streamlit

- Develop **Streamlit web app** with input fields for customer data.
- Integrate **trained XGBoost pipeline** for real-time prediction.
- Deploy on **Streamlit Cloud, Heroku, or AWS** for accessibility.

Dataset

- **Source:** Google Drive Link
- **Format:** CSV
- **Size:** 2L+ records, 20+ features
- **Target Variable:** `Premium Amount` (insurance premium)
- **Feature Types:** Numerical, Categorical, Text
- **Key Features:**
    - Age, Gender, Annual Income, Health Score, Previous Claims, Vehicle Age, Credit Score, Insurance Duration
    - Marital Status, Education Level, Occupation, Location, Policy Type, Smoking Status, Exercise Frequency, Property Type

- **Data Characteristics:** Missing values, skewed distributions, incorrect data types (simulating real-world complexity)

Modeling & Pipeline

- Preprocessing handled using **ColumnTransformer**:
    - **Numerical:** StandardScaler
    - **Categorical:** OneHotEncoder with `handle_unknown='ignore'`
- ML Pipeline integrated **XGBoost** as the final model.
- Model saved using **joblib** (`best_model.pkl`).

Streamlit App

- Input features via sliders, number inputs, and dropdowns.
- Real-time prediction output:
- `Predicted Insurance Premium: 6.60`

Values are scaled based on dataset distribution.

Project Deliverables

- Jupyter Notebook with code, EDA, and results
- ML Pipeline integrated with MLflow
- Trained Model for deployment
- Streamlit Web App code and link

Evaluation Metrics

- **RMSE:** Root Mean Squared Error
- **R^2 Score:** Variance explanation
- **MAE:** Average prediction error
- **RMSLE:** Logarithmic error metric