Smart Premium – Insurance Premium Prediction App

1. Project Overview

The **Smart Premium App** is a predictive analytics solution designed to estimate insurance premiums for customers based on their personal, financial, and health-related information. By leveraging machine learning techniques, this application provides a fast and accurate premium prediction, assisting insurance companies and customers in decision-making.

---

2. Objective

The main objectives of this project are:

- To build machine learning models that predict insurance premiums based on historical customer data.
- To provide a user-friendly **Streamlit interface** for real-time premium prediction.
- To track and compare models using **MLflow** for reproducibility and model selection.

---

3. Dataset

- **Source**: Synthetic / collected historical insurance customer data.
- **Features**:
    - Demographics: Age, Gender, Marital Status, Education Level, Location
    - Financial: Annual Income, Credit Score, Number of Dependents
    - Health: Health Score, Smoking Status, Exercise Frequency
    - Insurance: Policy Type, Insurance Duration
    - Vehicle: Vehicle Age

- **Target**: Premium Amount

The dataset was preprocessed to handle missing values, encode categorical variables, and normalize numerical features.

---

4. Data Preprocessing

- Dropped irrelevant columns such as `id`.
- Handled missing values using mean/median imputation.
- Outlier handling was done using IQR for numeric features.
- Categorical variables were encoded using a combination of **label encoding** and **one-hot encoding**.
- Train-test split: 80% training, 20% evaluation.

---

5. Machine Learning Models

The following models were tested:

- **Linear Regression (LR)**
- **Decision Tree Regressor (DT)**
- **Random Forest Regressor (RF)** *(best manual model)*
- **XGBoost Regressor** *(best MLflow model)*

Evaluation Metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Root Mean Squared Log Error (RMSLE)

- R^2 Score

Random Forest performed best in manual evaluation, while XGBoost was the best-performing model in MLflow experiments.

---

6. Streamlit App

A **user-friendly interface** was developed using **Streamlit** for real-time predictions.

Features of the app:

- Users can enter personal, financial, and health details.
- Real-time calculation and display of predicted insurance premium.
- Handles both numerical and categorical inputs.
- Includes proper validation and error handling.

App Workflow:

1. User inputs personal, financial, health, and insurance information.
2. Inputs are preprocessed to match model expectations.
3. Categorical variables are encoded, numeric variables are scaled.
4. Model predicts premium and results are displayed instantly.

---

7. Tools & Technologies

- **Programming Language:** Python
- **Libraries:** pandas, numpy, scikit-learn, xgboost, seaborn, matplotlib, Streamlit, joblib, MLflow
- **Version Control:** Git & GitHub
- **Environment:** Anaconda / virtual environment

---

8. Challenges & Solutions

- **Handling large datasets:** The dataset was too large to push entirely to GitHub; `.gitignore` was used to exclude unnecessary data files.
- **Model compatibility issues:** Fixed **NumPy and PyArrow version conflicts** to ensure smooth library functionality.
- **Categorical encoding mismatch:** Ensured Streamlit input dataframe columns matched the trained model.
- **Tracking models:** Implemented **MLflow** for organized model tracking and selection.

---

9. Conclusion

The Smart Premium app provides an **efficient, accurate, and easy-to-use solution** for predicting insurance premiums.

- **Manual model:** Random Forest performed best in manual evaluation.
- **MLflow model:** XGBoost was the best-performing model in logged experiments.

The combination of machine learning with a Streamlit interface allows non-technical users to estimate premiums quickly. The system can be further enhanced with larger datasets, real-time data integration, and advanced models for improved accuracy.

---

10. Future Enhancements

- Integration with real insurance company databases.
- Adding feature importance visualization for explainable AI.
- Implementing API endpoints for mobile app integration.
- Optimizing for very large datasets using cloud solutions.

---

11. Project Repository

GitHub: https://github.com/Divzdj/SmartPremium