Project Report

Project Title: Health Insurance Premium Prediction

Role: Data Scientist – Model Development, Governance & Deployment

Objective

To build a reliable system that can predict health insurance premium for any individual based on their health and personal details.
The system should be:

- Highly accurate (>97%)
- Fair and transparent
- Easy for insurance underwriters to use

Scope of Work

- Understanding the data and cleaning it
- Creating useful features
- Building and testing machine learning models
- Deploying the model as a Streamlit application
- Documenting the entire process for business use

Exploratory Data Analysis (EDA)

- Only a few missing values were found → safely removed
- Some extreme values were removed to avoid unusual errors
- Most customers:
  - Earn 0–40 Lakhs per year
  - Are male and living in the Southeast region
  - Are unmarried
  - Have normal BMI
  - Are non-smokers
- Age and genetic health factors showed a positive relationship with the insurance premium

→ Higher age or health risks → higher premium

Feature Engineering

- Converted medical history into a measurable risk score
- Converted categorical columns using encoding methods so the model can understand them
- Removed highly correlated features to avoid confusion for the model
- Data was scaled to ensure equal importance to all features

Model Building

We first tried a single model but found:

- Heavy errors for younger customers (age ≤ 25)

Solution: **Build two separate models**

Model	**Group**	**Algorithm**	**Accuracy (R^2 Score)**
Model A	Age > 25	XGBoost	99.7%
Model B	Age ≤ 25	Linear Regression	98.82%

After segmentation:

- Errors dropped significantly
- Predictions became fairer & more stable

Both models were saved and linked together for deployment.

Evaluation Metrics

- Accuracy measured using R^2 Score, MAE
- After improvement:
  - Model A: Only 0.5% high-error cases
  - Model B: About 2% high-error cases
    Over 97% of predictions now within 10% difference of actual premium

AI Governance

- Model was evaluated for its fairness & bias.
- Residual risk evaluation was done.
- SHAP was used to explain overall contribution of each feature to the premium prediction.

Deployment

- User Interface built using Streamlit
- Hosted on Streamlit Cloud
- Underwriters can:
  - Enter individual details
  - Instantly get premium predictions
  - See which model is used for transparency

Tech Stack Used

- Python
- Pandas, NumPy, Scikit-learn
- XGBoost

- SHAP for Explainability
- Streamlit for UI
- Joblib for model saving

Business Impact

- Faster premium estimation - reduces underwriter workload
- More consistent pricing decisions
- Improved customer experience
- Scalable and accessible from anywhere online

Conclusion

The model is:

- Highly accurate
- Fair across different age groups
- Ready to support real underwriting workflows
- This MVP proves that AI can greatly improve efficiency in pricing health insurance policies.

Recommendations to Stakeholders

- Expand dataset to include more health details → even better accuracy
- Plan for yearly retraining → adapt to new pricing rules and trends
- Introduce uncertainty estimates → show confidence levels to underwriters
- Gradually integrate this into the official underwriting process

Deployed Link: Streamlit

Video Explanation: Link

Project By:
Shijin Ramesh
Data Scientist
LinkedIn | Portfolio