

Final Capstone Project Report

Insurance Premium Prediction using Machine Learning

1. Project Overview

The objective of this project is to build a machine learning model that predicts health insurance premiums based on demographic, lifestyle, and medical attributes. Accurate prediction helps insurance companies assess risk and automate premium calculations at scale.

2. Dataset Description

Source: SQLite database (regression.db)

Table Used: insurance_prediction

Records: ~1,000,000

Features:

Age, Gender, BMI, Children, Smoker, Region, Medical History, Family Medical History, Exercise Frequency, Occupation, Coverage Level

Target:

Charges – insurance premium amount

3. Data Splitting Strategy

Training: First 700,000 records

Evaluation: Next 200,000 records

Production: Remaining records

4. Exploratory Data Analysis

EDA was performed on a random sample of 100,000 records. Insurance charges were right-skewed. Smoking status, BMI, and age showed strong influence on premium values.

5. Data Preprocessing & Feature Engineering

Numerical features were scaled using StandardScaler. Categorical features were one-hot encoded. Missing numerical values were imputed using median, while categorical missing values were treated as 'Unknown'.

6. Models Evaluated

Linear models (Ridge, SGDRegressor) were used as baselines. Random Forest Regressor was selected as the final model due to superior performance and robustness.

7. Model Evaluation

Random Forest achieved an R^2 of approximately 0.99 with low MAE and RMSE on the evaluation dataset.

8. Feature Importance

Key features included smoking status, BMI, age, coverage level, and medical history indicators.

9. Production Implementation

Separate Python scripts were created for training (train.py) and prediction (predict.py). Preprocessing and model artifacts were saved for consistent inference.

10. Conclusion

This project demonstrates an end-to-end, industry-grade machine learning pipeline, from data ingestion to production-ready predictions.

Status: Capstone Project Completed Successfully