Project Proposal: Insurance Claim Prediction Using Machine Learning

Problem Statement

Insurance companies manage large volumes of claims daily, and accurately predicting whether a policyholder will file a claim is essential for effective risk management and premium optimization. Traditional approaches often rely on limited data and expert judgment, which can be inefficient or biased. This project aims to build a data-driven machine learning model to predict the likelihood of a policyholder filing an insurance claim using demographic, policy, and historical data.

Context

Predictive analytics is transforming the insurance industry by enabling smarter underwriting, better pricing strategies, and improved fraud detection. By identifying customers more likely to file claims, insurers can tailor their services, optimize premium structures, and proactively manage financial risk. This project fits within the data science for business decision-making domain, applying machine learning for classification and predictive modeling.

Criteria for Success

The project will be considered successful if it:
1. Develops a model with strong predictive performance (high accuracy, precision, recall, and ROC-AUC).
2. Identifies the most important features influencing claim likelihood.
3. Produces clear visual insights that can inform decision-making for insurers.
4. Presents results in a reproducible, interpretable, and ethical manner.

Scope of Solution Space

This project will focus on structured tabular data analysis, feature engineering and data preprocessing, machine learning model training and comparison (Logistic Regression, Random Forest, XGBoost), model interpretability through feature importance and SHAP analysis, and visualization and reporting of key results.

Constraints

Data limitations: The dataset does not include textual or temporal data often used in production systems.
Ethical constraints: The model must avoid bias related to sensitive features like gender or region.
Computational constraints: Limited to local or standard notebook environments.
Time constraint: Approximately 2-4 hours for this phase (problem identification and proposal).