Project Report: Customer Lifetime Value Prediction Model

Introduction

In the current digital era, the insurance sector is leveraging data-driven solutions to assess customer risk, predict claims, and improve decision-making. This project aims to build a predictive model that classifies and evaluates insurance claim-related data. By applying machine learning techniques, the project helps in identifying potential risk patterns and provides insights that can assist organizations in optimizing their services.

Abstract

This project focuses on developing a predictive model using an insurance dataset containing demographic and policy-related information. The workflow involves data preprocessing, feature engineering, handling missing values, and applying machine learning models to predict outcomes. Models such as Logistic Regression, Random Forest, and XGBoost were trained and evaluated. The results highlight the model's effectiveness in capturing important features like income, policy type, and claim amount, enabling accurate predictions. This project demonstrates the importance of structured data preparation and ensemble learning in building robust risk prediction systems.

Tools Used

- **Python** – Programming language for development
- **Pandas & NumPy** – Data preprocessing and manipulation
- **Scikit-learn** – Model building and evaluation
- **XGBoost** – Advanced gradient boosting algorithm
- **Matplotlib** – Data visualization
- **Joblib** – Saving and loading trained models

Steps Involved in Building the Project

1. **Data Collection & Loading**
   - Imported training and testing datasets in CSV format.
   - Inspected dataset structure, size, and key variables.
2. **Exploratory Data Analysis (EDA)**
   - Checked for missing values and handled them using median imputation.

- Analyzed unique values in categorical columns like gender, income, and policy type.

3. **Data Preprocessing & Feature Engineering**
   - Encoded ordinal features such as income (≤2L, 2L–5L, etc.) and policy type (Silver, Gold, Platinum).
   - Transformed categorical variables into numeric format for machine learning models.
4. **Model Building**
   - Applied different algorithms including Logistic Regression, Random Forest, and XGBoost.
   - Tuned hyperparameters for improved model performance.
5. **Model Evaluation**
   - Assessed models based on accuracy and predictive performance.
   - Compared results to select the most suitable model.
6. **Model Deployment Preparation**
   - Exported the trained model using **joblib** for future use.

Conclusion

The project successfully demonstrated the end-to-end process of building a predictive model for insurance data. Data preprocessing and feature engineering significantly improved model performance, while ensemble methods such as Random Forest and XGBoost achieved strong predictive accuracy. This project highlights the role of machine learning in insurance analytics, helping businesses minimize risks and make informed decisions. Future work may include incorporating more advanced techniques like deep learning or integrating external datasets to enhance model robustness.