Insurance Claim Prediction

1.Executive Summary

This project focuses on developing a machine learning model to predict insurance claim frequency based on policyholder and vehicle attributes. Accurate prediction of insurance claims can assist insurers in risk assessment, fraud detection, and premium pricing. The study leverages advanced modeling techniques, feature importance analysis, and class imbalance handling to build a reliable prediction framework.

2. Problem Statement

- **Background:** The insurance industry faces challenges in predicting claim frequency due to the complexity and variability in customer demographics, vehicle attributes, and other factors. Manual evaluation can lead to inconsistencies and biases.
- **Objective:** Develop a machine learning model to predict the likelihood of insurance claims using structured data.
- **Scope:** Utilize a publicly available dataset containing detailed insurance-related features and assess multiple models to optimize prediction accuracy.

3. Data Source

- **Dataset:** Contains 58,592 entries and 41 columns, including the target variable **claim_status.**

  Key Features:

  - Numerical: subscription_length, vehicle_age, customer_age.
  - Categorical: region_code, segment, fuel_type, model.
  - Target Variable: claim_status (0: No Claim, 1: Claim).

4. Methodology

- ### Data Preparation:

  Performed Exploratory Data Analysis (EDA) to identify patterns and trends in both numerical and categorical features. Balanced the dataset using oversampling techniques to address the significant class imbalance in claim_status. Dropped irrelevant features like policy_id and performed feature encoding for categorical variables.

- ### Feature Selection

  Identified the top 10 influential variables for claim prediction using Random Forest feature importance.

- **Model Development**

Built and evaluated multiple models, including Logistic Regression, Random Forest, Decision Tree, KNN, AdaBoost, and XGBoost model

Testing on Original Imbalanced Data

Applied the trained model on the original imbalanced dataset to evaluate its real-world applicability. The model demonstrated exceptional precision and recall, effectively identifying claims.

5. Key Deliverables

- Model: A machine learning model capable of accurately predicting insurance claims.
- Insights: Identification of the most critical features impacting claim prediction.

6. Challenges

- **Class Imbalance:** Addressed through oversampling to ensure unbiased predictions for minority classes.
- **Feature Encoding:** Categorical variables required careful encoding to balance interpretability and model performance.
- **Outliers:** Retained outliers to reflect valid variability, especially in vehicle and customer attributes.

7. Results

- **Best Model Performance (Random Forest)**
- **Precision:** 1.00 (No Claims), 0.95 (Claims)
- **Recall:** 0.94 (No Claims), 1.00 (Claims)
- **F1-Score:** 0.97 for both classes.
- **AUC-ROC Curve:** Above 0.97.
- **Classification Accuracy on Original Imbalanced Data:** 97.4% correctly classified and 2.6% misclassified.

8. Practical Applications

- **Risk Assessment:** Identify high-risk policies for proactive measures.
- **Fraud Detection:** Flag suspicious claims based on feature discrepancies.
- **Premium Optimization:** Offer personalized pricing based on claim likelihood.
- **Customer Support:** Provide insights to policyholders on factors influencing their premiums.

9. Conclusion

This project demonstrates the power of machine learning in insurance claim prediction. By addressing class imbalance, optimizing model performance, and providing actionable insights, the study highlights a practical solution for challenges faced by the insurance industry. The Random Forest model's exceptional performance, with a high AUC-ROC score and accurate predictions on the imbalanced dataset, reinforces its utility in real-world applications.