## **Predicting Insurance Claim Frequency: A Case Study in Classification on Imbalanced Data**

This project focuses on predicting the frequency of insurance claims using a dataset containing information about insurance policies, vehicles, and customers. The primary goal is to build a robust classification model that can accurately identify instances where claims are likely to occur, even in the face of imbalanced data. This project demonstrates the process of handling class imbalance in classification, highlighting the challenges and solutions involved.

### **Project Goal**

The main goal is to develop a machine learning model capable of predicting whether an insurance policy will result in a claim. This prediction is crucial for insurance companies to manage risk effectively and set premiums appropriately.

**Project Process**

1. **Data Loading and Exploration:** Loaded the insurance claims dataset using pandas and explored its structure, data types, and missing values.
2. **Data Visualization:** Created visualizations to understand the distribution of claim status, numerical features (customer age, vehicle age, subscription length), and categorical features (region code, segment, fuel type).
3. **Data Preprocessing:** Handled class imbalance by oversampling the minority class (claims with status 1) using the resample method from scikit-learn.
4. **Feature Engineering:** Encoded categorical features using LabelEncoder for model training.
5. **Model Training:** Trained a RandomForestClassifier model on the oversampled data to predict claim status.
6. **Model Evaluation:** Evaluated the model's performance using metrics like classification report and accuracy score.
7. **Comparison with Original Data:** Applied the trained model to the original dataset and compared predictions with actual claim statuses.
8. **Visualization of Results:** Created a bar chart to visualize the classification accuracy.

### **Methods**

1. Data Exploration and Preprocessing: The project starts with an in-depth analysis of the dataset, including:
   * Examining the distribution of classes in the target variable (claim\_status) to understand the level of imbalance.
   * Visualizing the distributions of key numerical features like subscription length, vehicle age, and customer age.
   * Analyzing the distribution and relationships of categorical features such as region code, segment, and fuel type.
2. Handling Class Imbalance: Since the dataset exhibits a significant imbalance between the claim\_status classes (more instances of no claim than claims), the project utilizes oversampling to balance the classes. This involves replicating instances from the minority class to achieve an equal representation of both classes.
3. Feature Selection: To identify the most impactful features for prediction, the project employs feature importance techniques. This involves training a Random Forest model and analyzing the feature importance scores to identify the top variables influencing the claim\_status prediction.
4. Model Training: Using the balanced dataset and selected features, a Random Forest classifier is trained to predict the claim\_status.
5. Model Evaluation: The trained model is evaluated on the test dataset using metrics suitable for imbalanced datasets, such as:
   * Precision: Proportion of correctly predicted positive instances out of all predicted positives.
   * Recall: Proportion of correctly predicted positive instances out of all actual positives.
   * F1-Score: Harmonic mean of precision and recall.
   * Accuracy: Overall proportion of correctly classified instances.
   * AUC (Area Under the ROC Curve): Measures the overall ability of the model to distinguish between the two classes.

### **Results**

The developed Random Forest model achieved impressive performance on the test dataset, demonstrating high precision, recall, and F1-score for both classes. The high recall for the "claim" class is particularly noteworthy, indicating that the model is effective at identifying instances where claims are likely to occur.

### **Features**

* Key Features: The project identified features such as subscription length, vehicle age, customer age, region code, model, and engine type as having a significant influence on predicting claim frequency.
* Unique Identifier: While policy\_id was initially identified as a highly important feature, it was removed during model training as it is not directly relevant for prediction.

### **Tools**

* Python: The core programming language for data processing, analysis, and model building.
* Pandas: A powerful library for data manipulation and analysis.
* Scikit-learn (sklearn): A widely used library for machine learning algorithms, including Random Forest classifiers.
* imblearn: A library specifically designed for handling imbalanced datasets, including oversampling techniques like SMOTE.
* Matplotlib and Seaborn: Libraries for data visualization.

### **Importance**

This project has significant implications for the insurance industry:

* Risk Management: The model can help insurance companies better assess risk associated with individual policies, leading to more accurate premium calculations.
* Fraud Detection: By identifying patterns associated with fraudulent claims, the model can assist in detecting and preventing fraudulent activities.
* Customer Segmentation: The model can be used to identify customer segments with higher claim frequency, allowing insurance companies to develop targeted risk management and customer service strategies.

### **Conclusion**

This project showcases the process of building a robust classification model for predicting insurance claim frequency, addressing the challenges associated with imbalanced data. By utilizing appropriate methods for handling class imbalance, feature selection, and model evaluation, the project demonstrates the potential of machine learning to improve risk management and decision-making in the insurance industry.

**References:**Machine Learning for Insurance Claim Prediction<https://www.xenonstack.com/use-cases/insurance-claim-prediction/>