Insurance Policy Response Prediction

1. Executive summary

This project developed a high-performance machine learning pipeline to predict customer interest in vehicle insurance. By processing over 380,000 records and addressing a massive class imbalance, we successfully trained an **AdaBoost** model that achieves a high recall for the "at-risk" (interested) class, providing a strategic advantage for targeted marketing and resource optimization.

---

2. Exploratory Data Analysis (EDA) & Inferences

The EDA phase was crucial in uncovering the underlying patterns and behavior of the customer base.

2.1 Key Statistical Findings

- **Class Imbalance**: Only about **12%** of customers responded positively to the insurance offers. This confirmed that standard accuracy would be a misleading metric.
- **Feature Associations (Cramér's V)**: We implemented Cramér's V to test the strength of association between categorical features and the target.
    - **Inference**: Vehicle_Damage and Previously_Insured were the strongest predictors. Customers with previous vehicle damage were significantly more likely to show interest.
- **Redundancy**: The id column was found to have zero predictive power and was removed to streamline the model.

2.2 Demographic Inferences

- Younger drivers (under 25) rarely had prior insurance and showed the lowest response rates.
- The Annual_Premium feature showed a significant right-skewed distribution, requiring robust handling during the transformation phase.

---

3. Data Preprocessing & Feature Engineering

To prepare the data for modeling, the following pipeline was established:

- **Categorical Encoding**: Used binary encoding for Gender, frequency encoding for Region_Code and Policy_Sales_Channel (to manage high cardinality), and One-Hot Encoding for Vehicle_Age.
- **Resampling (SMOTETomek)**: To combat the 12% minority class issue, we used a hybrid approach. **SMOTE** oversampled the minority class, while **Tomek Links** removed noisy examples from the majority class boundary to create a cleaner decision path.
- **Scaling**: Standard scaling was applied to Age and Vintage to ensure numerical features were on a uniform scale.

---

4. The Transition From Linear To Tree Models

A critical turning point in the project was the evaluation of model architectures.

4.1 Why Linear Models (e.g., Logistic Regression) Failed

- **Non-Linearity**: The relationship between age, premium, and interest is not a straight line. Linear models struggled to capture "pockets" of interested demographics.
- **Complex Interactions**: Insurance interest depends on interactions (e.g., Age + Vehicle_Damage). Linear models require manual feature engineering to see these, whereas Trees capture them automatically.
- **Outlier Sensitivity**: Linear models were disproportionately influenced by extreme values in Annual_Premium.

4.2 Why Tree-Based Models (AdaBoost) Succeeded

- **Sequential Learning**: AdaBoost focuses on "hard-to-classify" instances by updating weights, making it perfect for imbalanced data.
- **Non-Parametric**: Trees do not assume a normal distribution, allowing them to handle the skewed data discovered during EDA.
- **Feature Importance**: They provided clear insights into the hierarchy of drivers for customer interest.

---

5. Final Model Performance & Tuning

The final model was fine-tuned using RandomizedSearchCV and tracked via **MLflow**.

5.1 Final Performance Metrics (AdaBoost)

Metric	Score
Accuracy	**0.7929 (79.29%)**
Recall (Risk Class)	**0.7015 (70.15%)**
F1 Macro	**0.6604**

5.2 Detailed Classification Report

- **Class 0 (Not Interested)**: Precision: 0.95 | Recall: 0.81
- **Class 1 (Interested)**: Precision: 0.33 | **Recall: 0.70**

Strategic Inference: While the precision for "Interested" is 0.33, the **Recall of 70%** is the critical success factor. It ensures the sales team captures 70% of all potential leads, which is significantly higher than the baseline 12% hit rate.

---

6. Optimization & Experiment Tracking

To ensure reproducibility and quality, we utilized:

- **MLflow**: Logged parameters (learning rate, estimators) and metrics for every run.
- **Joblib**: The final fine-tuned model was serialized as best_model_AdaBoost_SMOTETomek.pkl for immediate deployment.

---

7. Conclusion

The project successfully transitioned from raw, imbalanced data to a robust predictive tool. By prioritizing **Recall** over simple **Accuracy**, the model provides the insurance company with an actionable list of high-probability leads, optimizing the sales funnel and reducing wasted marketing spend.