**Objective**

The project aims to predict customer conversion rates for an insurance company by analyzing historical marketing campaign data and applying various machine learning models. The primary goal is to identify customers most likely to subscribe to insurance policies.

**Steps and Insights**

1. **Dataset**
   * The dataset includes demographic, contact, and campaign-related features such as age, job, marital status, education, contact type, campaign duration, and previous outcomes.
   * Target variable: y (whether the client subscribed to insurance).
2. **Data Cleaning and Preprocessing**
   * Duplicate rows were removed, reducing the dataset to 44,917 records.
   * Categorical columns with "unknown" values were analyzed:
     + Some "unknown" values were dropped (e.g., job, education\_qual) based on their low percentage.
     + Features with higher percentages of "unknown" (e.g., call\_type and prev\_outcome) were retained.
   * Outliers in numerical features (age, duration, num\_calls) were detected using the IQR method and removed.
   * Final dataset size: **36,940 rows**.
3. **Exploratory Data Analysis (EDA)**
   * **Univariate Analysis**:
     + Visualizations for target variable distribution showed class imbalance (88% "no" vs. 12% "yes").
     + Histograms and boxplots for numerical features highlighted their spread and central tendencies.
   * **Bivariate Analysis**:
     + Relationships between categorical and numerical features with the target variable were visualized.
   * Correlation analysis revealed weak correlations among numerical features.
4. **Data Encoding and Splitting**
   * Categorical features were encoded using one-hot and ordinal encoding.
   * Data was split into training and testing sets (80/20).
5. **Modeling**
   * Five machine learning models were applied:
     + **Logistic Regression**: Accuracy = 93%, but poor recall for the minority class.
     + **K-Nearest Neighbors (KNN)**: Best accuracy with k=13 = 92.73%.
     + **Decision Tree**: Accuracy = 90%, moderate precision and recall.
     + **Random Forest**: Best accuracy = 93.27%, decent precision and recall balance.
     + **XGBoost**: Accuracy = 92.85%, with strong feature importance visualization.
6. **Results**
   * **Best Model**: Random Forest achieved the highest accuracy (93.27%).
   * **Challenges**: Class imbalance affected minority class recall across all models, indicating room for improvement with techniques like oversampling or SMOTE.

**Key Takeaways**

* The project effectively prepares data for machine learning by handling missing values, outliers, and encoding features.
* Random Forest emerged as the most effective model for this dataset, but precision-recall imbalance suggests that further work is needed to optimize for minority class predictions.
* With the given models and preprocessing steps, the project provides actionable insights into targeting high-conversion customers.