**Customer Conversion Prediction**

**Project Overview**

**Domain**

Data Science / Machine Learning

**Problem Statement**

You are working for a new-age insurance company and employ multiple outreach plans to sell term insurance to your customers. Telephonic marketing campaigns remain one of the most effective ways to reach people, but they incur significant costs. Therefore, it is essential to identify the customers who are most likely to convert beforehand, so they can be specifically targeted via call. Given the historical marketing data of the insurance company, we are required to build an ML model to predict if a client will subscribe to the insurance.

**Approach**

The project follows a structured approach to achieve the objective:

1. **Clean**: Handle missing values, outliers, duplicates, and ensure data is in the right format.
2. **Explore**: Perform exploratory data analysis to understand the data and extract insights.
3. **Pre-process**: Transform the data into a suitable format for modeling.
4. **Model**: Train various machine learning models.
5. **Validate**: Evaluate the models using appropriate metrics.
6. **Enhance**: Improve the model performance using techniques like feature selection, balancing the dataset, etc.
7. **Explain**: Analyze the important factors contributing to the conversion rate.
8. **Deploy**: Prepare the model for deployment.

**Data Set Explanation**

The historical sales data includes the following features:

* **age**: Age of the client (numeric)
* **job**: Type of job
* **marital**: Marital status
* **educational\_qual**: Education status
* **call\_type**: Contact communication type
* **day**: Last contact day of the month (numeric)
* **mon**: Last contact month of the year
* **dur**: Last contact duration in seconds (numeric)
* **num\_calls**: Number of contacts performed during this campaign for this client
* **prev\_outcome**: Outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")
* **y**: Target variable indicating if the client subscribed to the insurance (yes/no)

**Data Preprocessing**

1. **Loading and Initial Analysis**:
   * Data is loaded using pandas.
   * Initial shape and column names are checked.
2. **Cleaning**:
   * Missing values are checked.
   * Duplicates are removed.
   * Outliers in numeric columns are clipped.
   * Data types are ensured to be appropriate.
3. **Feature Engineering**:
   * Categorical variables are encoded.
   * Target variable is mapped to binary (0: yes, 1: no).

**Exploratory Data Analysis (EDA)**

* The average age for each target class was analyzed.
* Job types, marital status, and educational qualifications were examined for their impact on the target variable.

**Modeling**

1. **Data Splitting**:
   * Data is split into training and testing sets using stratified sampling to maintain the target variable distribution.
2. **Model Training and Validation**:
   * Models like Decision Trees, Random Forest, and XGBoost were trained.
   * Techniques like undersampling (Cluster Centroids), oversampling (SMOTE), and a combination of both (SMOTEENN) were used to handle class imbalance.
3. **Evaluation**:
   * The models were evaluated using the F1-Score as the primary metric to balance precision and recall.

**Results and Model Selection**

* The RandomForestClassifier and XGBClassifier showed promising results with balanced data using SMOTE.
* Feature importance was analyzed to identify the key factors influencing conversion.

**Conclusion and Future Work**

The project successfully developed a machine learning model to predict customer conversion for telemarketing campaigns. The model's performance, as measured by the F1-Score, indicates it can effectively target potential customers, thus optimizing campaign costs. Future work could involve further tuning the models, exploring additional features, and implementing the model in a real-time deployment scenario.