**Insurance Purchase Prediction Using Customer Data**

**Top-Level Summary:**

A new-age insurance company employs multiple outreach plans to sell term insurance to its customers. Telephonic marketing campaigns remain one of the most effective ways to reach potential clients, however these campaigns incur significant costs. To optimize resources and reduce expenses, it is crucial to identify the customers who are most likely to convert beforehand so that they can be specifically targeted via calls. Given historical marketing data, the objective is to build a machine learning model that will predict if a client will subscribe to the insurance.

**1. Project Introduction**

In the competitive insurance sales industry, the efficient allocation of marketing resources is essential. Telephonic marketing, despite its effectiveness, can be expensive. Thus, predicting potential subscribers beforehand can greatly enhance the cost-efficiency of marketing initiatives. This report describes the development and assessment of a predictive model to achieve this aim.

**2. Data Description**

The dataset comprises 11 columns, including 10 features and 1 target variable. Below is a detailed explanation of each column:

* **Features:**
* **Age: Type: Numeric**
  1. **Description**: Represents the age of the client. This is an important demographic feature that can influence the likelihood of a client subscribing to insurance. For instance, certain age groups may be more inclined to purchase insurance due to their stage in life or financial planning needs.
* **Job: Type: Categorical**
  1. **Description**: Indicates the type of job the client has. Examples include "admin", "blue-collar", "entrepreneur", etc. The type of occupation can reflect the client's income level, job stability, and potential interest in insurance products.
* **Marital: Type: Categorical**
  1. **Description**: Shows the marital status of the client. Common categories are "single", "married", and "divorced". Marital status can influence financial decisions and priorities, impacting the likelihood of subscribing to insurance.
* **Educational\_qual: Type: Categorical**
  1. **Description**: Reflects the education level of the client, such as "primary", "secondary", "tertiary", etc. Education can affect a client's understanding of insurance benefits and their financial literacy.
* **Call\_type: Type: Categorical**
  1. **Description**: Type of communication used to contact the client. Categories may include "cellular", "telephone", etc. The mode of communication can impact the effectiveness of the outreach and the client's responsiveness.
* **Day: Type: Numeric**
  1. **Description**: The day of the month when the last contact was made. This can capture time-specific patterns in client behavior or responsiveness.
* **Mon: Type: Categorical**
  1. **Description**: The month of the year when the last contact was made. Examples are "Jan", "feb", "mar", etc. Seasonality can affect client decisions and their availability.
* **dur: Type: Numeric**
  1. **Description**: Duration of the last contact in seconds. Longer call durations may indicate more engaged conversations, which could correlate with a higher likelihood of conversion.
* **Num\_calls: Type: Numeric**
  1. **Description**: Number of contacts made with the client during the current campaign. This feature captures the persistence of the outreach efforts and the client's responsiveness.
* **Prev\_outcome: Type: Categorical**
  1. **Description**: Outcome of the previous marketing campaign for this client. Categories include "unknown", "other", "failure", and "success". Past behavior can be a strong predictor of future actions, with previous success indicating a higher likelihood of future conversion.

**Output Variable (Desired Target):**

* **Y: Type: Binary**
  1. **Description**: Indicates whether the client has subscribed to the insurance (1 for yes, 0 for no). This is the target variable that the model will predict, based on the provided features.

**3. Data Analysis and Preprocessing**

**3.1. Data Import and Initial Assessment**

The dataset was imported and inspected to assess its structure and completeness. We determined the data types, identified any missing values, and verified the absence of anomalies.

**3.2. Data Cleaning**

* **Missing Values:** Addressed through appropriate imputation or deletion strategies based on their impact on model performance and context.
* **Outliers:** Identified and managed to mitigate their influence on the results.
* **Duplicates:** Detected and resolved to ensure data integrity and accuracy.
* **Data Structure:** Verified to ensure alignment with expected formats and structures.
* **Data Types:** Examined to confirm compatibility with analysis and modelling requirements.

**3.3. Data Exploration and Insights (EDA):**

EDA revealed the following key insights:

* **Age:** Age strongly influences customer conversion: around 40.8 years old, lower likelihood; around 41.8 years old, higher likelihood.
* **Job:** Students, retirees, and the unemployed are more likely to purchase insurance based on the provided data.
* **Marital Status**: Singles are more likely to buy insurance
* **Education Level:** Individuals with tertiary education are more likely to purchase insurance.
* **Call Type and Timing**: From the above plot, Call\_type has a significant impact. Moreover, there is a positive correlation between call duration and the likelihood of customers buying insurance, indicating that longer calls tend to result in higher conversion rates.
* **Month and Day**: During March, December, and September, there is a notable increase in insurance purchases, as observed from the bar graph. Additionally, there is a higher likelihood of buying insurance at the beginning of each month, as indicated by the graph.
* **Previous Campaign Outcome**: Having a positive previous outcome increases the likelihood of obtaining insurance.

**4. Feature Engineering and Selection**

* Categorical features were transformed using label encoding instead of one-hot encoding.
* Categories were aggregated to reduce complexity and streamline the model.
* Numerical features were scaled to maintain consistency.
* Additional engineered features, such as interactions between call duration and previous campaign outcomes, were incorporated to enhance model effectiveness.

**5. Model Building**

**5.1. Model Selection**

Various models were evaluated, including

* Logistic Regression
* Decision Trees Classifier
* Random Forests Classifier
* XGB Classifier

**5.2. Model Training**

The models were trained using an 80% training data and 20% testing data split, with cross-validation employed to ensure reliable performance estimates.

**6. Model Evaluation**

**6.1. Performance Metrics**

* **F1-Score:** The primary metric used to balance precision and recall.
* **Accuracy:** Offers a general sense of the overall model performance.

The best model achieved an F1-Score of 0.92, demonstrating a good balance between precision and recall.

**6.2. Feature Importance**

Analysed the importance of each feature in the model’s decision-making process:

* **Call Duration**: Most influential factor.
* **Previous Campaign Outcome**: Strong indicator of future behaviour.
* **Number of Calls**: More attempts generally correlated with higher conversion rates.

**7. Model Tuning and Enhancement**

**7.1. Hyperparameter Tuning**

Hyperparameters of the selected models were fine-tuned using Cross-val-score, optimizing their performance.

**7.2. Feature Selection**

Techniques such as model. feature\_impotances were applied to select the most impactful features, thereby reducing model complexity and enhancing interpretability.

**8. Model Interpretation**

* Through the use of various machine learning models such as Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, the company can predict the likelihood of a client subscribing to the insurance. These predictions will enable the company to focus its marketing efforts on high-probability customers, thereby reducing costs and increasing conversion rates.

**9. Results and Business Impact**

**9.1. Summary of Findings**

* Key factors influencing client subscriptions were identified, with call duration and previous campaign outcomes emerging as the most significant predictors.
* The developed model achieved an F1-Score of 0.78, indicating robust predictive capability.

**9.2. Business Impact**

* By targeting the top 30% of the most likely subscribers identified by the model, the company stands to potentially reduce telephonic marketing costs by 50%.
* Insights from the model can inform strategic adjustments in marketing efforts, enabling focused and efficient resource allocation in high-impact areas.

**10. Conclusion**

This project has achieved a predictive model aimed at improving the efficiency of telephonic marketing campaigns for the insurance company. The deployment of this model is poised to deliver significant cost savings and higher conversion rates. Ongoing monitoring and fine-tuning will be essential to ensure the model's continued effectiveness in adapting to changing market conditions.