**Propensify**

**Prediction of Marketing Campaign Response**

**Author**: Supraja Singu  
**Course**: Data Science Bootcamp, upGrad

**Introduction**

Insurance is a critical component of financial security, offering protection against unforeseen events. However, identifying the right customers to target with marketing efforts can be a complex and resource-intensive task. Without an effective strategy, insurance companies may end up spending valuable resources on individuals who are unlikely to engage with their services, leading to inefficient campaigns and missed opportunities.

The primary goal of this project is to build a propensity model that predicts the likelihood of potential customers engaging with marketing efforts. By analyzing a historical dataset (train.csv), the model will identify patterns and behaviors that can be used to predict which individuals from a list of potential customers (test.csv) are most likely to respond to marketing. This solution aims to help the insurance company optimize its marketing strategy, ensuring that resources are focused on the most promising leads.

**Dataset Overview**

This project focuses on developing a propensity model to predict customer engagement with marketing campaigns for an insurance company. The model leverages historical customer data to identify which leads are likely to respond positively, ultimately optimizing marketing efforts.

The dataset consists of two primary components:

1. Training Data (train.csv):
   * Contains detailed information on past customer interactions and responses to previous marketing campaigns.
   * Features include customer demographics, contact history, and past response behaviour.
   * Imbalanced Dataset: The response variable is highly skewed, with approximately 89% of customers who did not respond to campaigns and only 11% who did. This presents a challenge in model development, requiring techniques such as resampling or class weighting to address the imbalance.
2. Test Data (test.csv):
   * Contains similar features as the training data but without the response variable.
   * This data will be used to predict future customer responses, providing insights for targeted marketing campaigns.

**Methodology**

**3.1 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was performed to understand the dataset better. This involved:

* **Descriptive Statistics**: Generated summary statistics to understand the central tendencies and spread of numerical features.
* **Check for Duplicate Values:** Identified and removed any duplicate entries to maintain data integrity.
* **Check for Missing Values:** Assessed missing data to determine appropriate strategies for imputation or removal.
* **Target Variable Analysis:** Analyzed the target variable distribution, noting the significant imbalance between responders and non-responders.
* **Numerical Feature Analysis:** Examined the distribution and spread of numerical features like age and income.
* **Categorical Feature Analysis:** Analyzed categorical features to understand the frequency distribution.
* **Visualizations**: Using graphs and charts to visualize the numerical and categorical features.

**3.2 Data Cleaning**

Data cleaning involved several steps:

* **Data Aggregation:** Performed data aggregation based on domain knowledge to create more meaningful insights.
* **Handle Missing Data:** Applied various imputation methods (e.g., mean, median, mode) to handle missing values.
* **Removed Unnecessary Features:** Removed irrelevant features based on correlation analysis to reduce noise and improve model performance.

**3.3 Feature Engineering**

Feature engineering was conducted to improve the model's performance:

* **Transformation of Existing Features**: Modified existing features for better interpretability and modeling.

**3.4 Dealing with Imbalanced Data and Feature Selection**

Given the highly imbalanced nature of the dataset, various techniques were implemented to address this issue, such as:

* **Dealing with Skewed Data**: Applied the Box-Cox transformation to normalize skewed numerical features and make them more suitable for modeling.
* **Encoding and Standardization**: Encoded categorical features using one-hot encoding and standardized numerical features to ensure uniform scaling.
* **Feature Selection**: Applied Principal Component Analysis (PCA) for dimensionality reduction, selecting the most important features to improve model efficiency.
* **Handling Imbalanced Data**: Used the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class, balancing the dataset for better model performance.

**3.5 Model Selection**

Different classification models were considered for this project, including:

* Logistic Regression
* Random Forest
* Ridge Classifier
* Adaboost Classifier

The choice of model was based on their performance on the validation dataset and their ability to handle class imbalance.

**3.6 Model Training**

The dataset was split into training and testing sets, typically with a ratio of 70/30. The training set was used to fit the model, tuning hyperparameters as necessary to improve performance.

**3.7 Model Validation**

Model performance was evaluated using various metrics, such as:

* Accuracy
* Precision
* Recall
* F1 Score
* ROC-AUC Score

This evaluation helped ensure that the model generalizes well to unseen data.

**Results and Discussion**

Upon evaluating the model's performance, it was found that the accuracy exceeded 75%, meeting the project's success criteria. Key performance metrics demonstrated that the model was capable of detecting fraudulent transactions effectively, balancing both precision and recall.

**Key Findings**:

* The Random Forest model outperformed others in terms of accuracy and F1 Score.

**Future Work**

Several avenues for future improvement were identified:

* **Exploring Advanced Algorithms**: Investigating more sophisticated models like neural networks or ensemble methods could potentially enhance prediction accuracy.
* **Real-Time Monitoring**: Developing a system for continuous learning, where the model adapts to new data over time.

**Conclusion**

In this analysis, we performed a comprehensive exploration and modeling process for credit card fraud detection using an imbalanced dataset. The steps included:

* **Data Exploration and Preprocessing**: Loading and exploring the dataset, visualizing class imbalance, and performing necessary preprocessing steps.
* **Handling Imbalanced Data**: Applying techniques such as undersampling and oversampling to address class imbalance, ensuring that our models could better learn from the minority class.
* **Model Training and Evaluation**: Training and evaluating various classifiers, including Logistic Regression, Decision Tree, and Random Forest, using metrics such as confusion matrices, classification reports, ROC-AUC scores, and ROC curves.
* **Model Saving**: Saving the trained models for future use, ensuring easy loading and application for predictions on new data.

**Key Findings**:

* **Model Performance**: The Random Forest Classifier demonstrated high accuracy in detecting fraud, showing that it is a strong candidate for deployment. The ROC-AUC scores and ROC curves provided insights into each model’s performance, particularly in distinguishing between fraudulent and non-fraudulent transactions.
* **Impact of Imbalance Handling**: Techniques for balancing the dataset were essential in improving model performance and ensuring that the minority class was adequately represented in the training process.

Overall, this analysis has provided a robust framework. The insights gained from PCA and model evaluations will be instrumental in refining our approach and improving detection capabilities. Future work could involve fine-tuning models further, experimenting with additional features, and exploring other advanced techniques for handling imbalanced data.