Insurance Cross-Selling Prediction

Introduction

In this project, the goal is to develop a predictive model for insurance cross-selling. The objective is to determine whether customers are likely to purchase additional insurance products when presented with cross-selling opportunities. Accurate predictions of customer behaviour can enhance marketing strategies and increase the success rate of cross-selling campaigns.

Task

The primary task involves creating a machine learning model capable of classifying customers into two categories: those inclined to purchase additional insurance (positive class) and those not inclined (negative class). This binary classification facilitates targeting the right customers and optimizing cross-selling efforts.

Prepare

To accomplish the task, several data preparation steps are undertaken:

1. **Data Loading:** Relevant libraries are imported, and the dataset containing customer information and historical cross-selling outcomes is loaded.
2. **Data Inspection:** The dataset is inspected to understand its structure, including dimensions, data types, and summary statistics. This step helps identify potential issues like missing values and duplicate records.
3. **Data Cleaning:** Data quality issues are addressed, including handling missing values and transforming data types as needed, ensuring dataset suitability for modelling.
4. **Feature Engineering:** Features that can improve model predictive power may be created or modified. This could involve encoding categorical variables, scaling numerical features, or deriving new variables.
5. **Data Splitting:** The dataset is divided into training and testing subsets to evaluate model performance. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) may be employed to balance class distribution, especially in the presence of class imbalance.

Steps

The key steps involved in this project include:

1. **Data Exploration:** The dataset is explored through visualizations and statistical analyses to gain insights into customer behaviour and potential patterns.

2. **Model Selection:** Appropriate machine learning algorithms for binary classification, such as K-Nearest Neighbors (KNN), Random Forest, or Logistic Regression, are chosen based on the nature of the problem and dataset.
3. **Model Training:** Selected models are trained using the training data, ensuring they learn patterns and relationships between features and the target variable.
4. **Model Evaluation:** Model performance is assessed using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. This helps identify the best-performing model.

Result

Model	**Accuracy Score (Train)**	**ROC-AUC Score (Train)**	**Accuracy Score (Test)**	**ROC-AUC Score (Test)**
KNN	0.874	0.954	0.753	0.730
Random Forest Classifier	0.990	0.999	0.807	0.765
Logistic Regression	0.657	0.706	0.621	0.705

These results provide an overview of the model performances on both the training and testing datasets. The metrics include accuracy score and ROC-AUC score, which are commonly used to evaluate binary classification models.

- The K-Nearest Neighbors (KNN) model achieves an accuracy of approximately 87.4% on the training data and 75.3% on the testing data. The ROC-AUC scores are 95.4% on the training data and 73.0% on the testing data.
- The Random Forest Classifier exhibits strong performance, with an accuracy of approximately 99.0% on the training data and 80.7% on the testing data. The ROC-AUC scores are exceptionally high, at 99.9% on the training data and 76.5% on the testing data.
- The Logistic Regression model has the lowest accuracy, with approximately 65.7% on the training data and 62.1% on the testing data. However, its ROC-AUC scores are relatively better at 70.6% on the training data and 70.5% on the testing data.

These results suggest that the Random Forest Classifier outperforms the other models in terms of accuracy and ROC-AUC score on the testing data, making it a strong candidate for predicting insurance cross-selling success.