Tour Package predicting model

Abstract

A traveling agency just dropped new tour packages and dedicated hefty resources into marketing the tour packages. This however proved fruitless as proved by sale and pitch records.

Design

The goal of the project is build a classification model to predict whether a future buyer would buy a tour package or not. This will aid in building a customer profile which allows us to consider which customers to purseu

Data

The data set is built of 4888 instances (rows) and 20 features (columns). It is also worth noting that it is made of different datatypes; combining both categorical and numeric features.

Algorithms

I wanted to conduct a thorough testing of different models to see how it would affect the results and to help make the optimal choice.

Testing Scenario 1:

- Clean Data
- Test models (Decision Tree, Random Forest, XGBoost, AdaBoost, Logistic Regression)

Result: Decent Accuracy but very low recall. High number of false positives

Testing Scenario 2:

- Clean Data
- Oversample with SMOTE
- Test models (Decision Tree, Random Forest, XGBoost, AdaBoost, Logistic Regression)

Very High accuracy but little improvement on recall

Testing Scenario 3:

- Clean Data
- Oversample with RandomOverSampler
- Test models (Decision Tree, Random Forest, XGBoost, AdaBoost, Logistic Regression)

Best result. High accuracy, high recall.

Testing Scenario 4:

- Clean Data
- Feature Selection based only user profile
- Oversample
- Test models (Decision Tree, Random Forest, XGBoost, AdaBoost, Logistic Regression)

Low accuracy. Needs more tuning.

TOOLS

- Numpy and Pandas for data processing
- Scikit-learn for modeling
- Matplotlib and Seaborn for visualization