Classification Project Write-up

Predicting the Functionality of Tanzania Waterpoints

Abstract

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4, and P5).In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). They plan to use the same strategy for the new markets and have identified 2627 new potential customers. We are required to help the manager to predict the right group of the new customers.

I tried several different ml models and finally choose random forest to get a highest accuracy model, I also built a model for the hard-to-tell part(Segment B)

Design

Data comes from a Kaggle program, links are as follows: https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation.

Classifying statuses accurately via machine learning models would enable the Boyota company take action to successfully assign their 2500+ potential customer into correct Segment groups, improving their customer management and open new US market

Data

Data columns are attached in the final page.

Data comes from a Kaggle program, links are as follows: https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation

Individual sample/unit of analysis in this project are customers, every single row represent one customer.

I will work on all the features to evaluate the best model and help the company group customers.

The dataset are separated into two parts, one parts for training(8000+rows) and the other part for test(2500+ rows)

Algorithms

Feature Engineering

1. Normalizing numerical data for KNN model
2. Converting categorical features to binary dummy variables
3. Selecting subsets of the total unique values for categorical features that were converted to dummies, according to the number of samples they were associated with and their contribution to certain statuses

Models

Naïve bayes, k-nearest neighbors, random forest classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

Model Evaluation and Selection

The entire training dataset of 8068 records was split into 70/30 train-test-split data, we also have 2554 holdout data, and all scores reported below were calculated with Random forest on the training and testing portions.

A special prediction for segment B is also included(by random forest)

Final random forest train scores: 10 features with class weights

- F1 0.72
- precision 0.71
- recall 0.71

test

- F1: 0.47
- Precision: 0.46
- Recall: 0.46
-

Final random forest train scores for segment B: 10 features with class weights

- F1 0.85
- precision 0.84
- recall 0.70

test

- F1: 0.74

- Precision: 0.58
- Recall: 0.54

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

In addition to the slides and visuals presented, will be embedded on my personal website and blog.