Review link:
https://review.udacity.com/#!/reviews/3317096

Machine Learning Engineer Nanodegree

Capstone Proposal

Basmah Alabdullatif

12/10/2021

Predict Responses to Marketing Campaign

Domain background:

Marketing campaigns have become an important tool for companies to attract new customers. A well-planned campaign is essential to reduce expenses and get a good marketing ROI at the same time. Therefore, companies need a successful and cost-efficient campaign strategy to help companies stay competitive and gain more market shares.

Problem statement:

Spending the marketing budget on the wrong customers (customers less likely to purchase products) will waste the companies money.

Having previous campaigns data, and how customers act to them, we are going to classify whether a customer is likely to engage in our current campaign or not, to head the marketing effort in the right direction.

Dataset:

This data-set from Kaggle contains information of 2240 customers with 29 columns including demographics, buying habits, and how they acted to previous campaigns, here are the features:

1- ID: Unique identification code for each customer
2- Year_Birth: The DOB Year of the customer
3- Education: Customer's level of education.
4- Marital_Status: Customer's status of Marriage
5- Income: Customer's annual Income
6- Kidhome: Number of children under 13 in Customer's house
7- Teenhome: Number of children between 13-19 in Customer's house
8- Dt_Customer: Date of customers enrollment
9- Recency: Number of days since last purchase
10- MntWines: the amount of money spent on Wines in the last 2 years.
11- MntFruits: the amount of money spent on Fruits in the last 2 years.
12- MntMeatProducts: the amount of money spent on Meat products in the last 2 years.
13- MntFishProducts: the amount of money spent on Fish products in the last 2 years.
14- MntSweetProducts: the amount of money spent on Sweet products in the last 2 years.
15- MntGoldProds: Dollar the amount of money spent on Gold products in the last 2 years.
16- NumDealsPurchases: Number of purchases made with a discount.
17- NumWebPurchases: Number of purchases made through the company's website.
18- NumCatalogPurchases: Number of purchases made using the catalog.
19- NumStorePurchases: Number of purchases made directly in-store.
20- NumWebVisitsMonth: Number of visits made through the company's website.
21- AcceptedCmp1: 1 if the customer accepted the offer in the 1st campaign, 0 otherwise.
22- AcceptedCmp2: 1 if the customer accepted the offer in the 2nd campaign, 0 otherwise.
23- AcceptedCmp3: 1 if the customer accepted the offer in the 3rd campaign, 0 otherwise.
24- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise.
25- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise.
26- Complain: 1 if the customer complained in the last 2 years, 0 otherwise.
27- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise.
28- Z_CostContact cost spent on this customer (for marketing).
29- Z_Revenue revenue gained from this customer.

Data-set link: https://www.kaggle.com/rodsaldanha/arketing-campaign

Solution statement:

We are going to train a machine learning model using previous data to predict which customer is worth spending the marketing effort on. This is a classification problem that aims to classify each customer as (will respond, will not respond) based on their personal data, demographics, buying habits, and how they acted to previous campaigns.

Benchmark model:

To assess how our final model is performing we are planning to compare the results of the final model with the following benchmark: Data Analysis Campaign project by Rodolfo Saldanha using deep learning (Keras) which gave a 0.867 accuracy score. More details can be found here:

https://www.kaggle.com/rodsaldanha/data-analysis-campaign

We will use different algorithms and try to beat this result

Evaluation metrics:

As this is a classification problem, we will calculate accuracy, precision, recall, and generate a confusion matrix to evaluate our model's performance.

Project design:

We will use Python 3.6 in this project. The following libraries might be used: NumPy, Pandas, SciKit-Learn, SciPy, sklearn, xgboost, Matplotlib. Our project will go through the following steps:

Loading Data:
This data is available on Kaggle as a .csv file, we will simply load it to our workspace

Data exploration:
After loading the data, EDA will be done to understand the data, find data patterns, trends and extract useful insights to select features.

Pre-processing:
In this step, we will perform data cleansing, deal with missing data, and tackle outliers. We may also need scaling for numerical data.

Feature engineering
If needed, we may add new useful columns extracted from available ones. This may include but not limited to:

- Age from Year_Birth
- Number of months the customer is enrolled with the company from Dt_Customer
- Number of offers accepted from the past campaigns.

Training the model

After splitting the data to train and test, we will train the model using several algorithms from sklearn library (after tuning their parameters using GridSearchCV) such as:

- SVM – (hyperparameters like Kernels, penalty parameter, Gamma, ..)
- Decision trees – (criterion, max_depth, max_features)
- Random forest – (hyperparameters like max_depth, n_estimators n_estimators, max_features, ..)
- KNN – (hyperparameters like: leaf size, # of neighbors , ..)
- Logistic regression - (hyperparameters like: penalty, C, ..)

Evaluate the model

Using different evaluation measures as explained in the previous section, we will assess the performance of the model and we may need to go back to the previous step and work on parameter tuning until good results are reached, then we select the best algorithm.

Deployment

In this step, we will convert our work into a product.