Machine Learning Engineer Nanodegree

Capstone Proposal

Pawel Wisniewicz
05-06-2020

Proposal

Domain Background

Nowadays by paying by card, by online shopping, by using our smartphones, by surfing the internet we are leaving a digital footprint. More and more data is collected, due to that, there are new possibilities for businesses to explore that data and find some insights which can boost their effectiveness. Marketing is one of the fastest-growing fields utilizing big data opportunities. Forbes says that Machine learning is using for variety of applications in the marketing. User personalization, Customer segmentation are just a few processes supported by algorithms. Machine learning can eliminate the guesswork and improve targeted advertising effectiveness.

Problem Statement

The main challenge is to improve the selection of people for targets of a marketing campaign. There is a lot of data gathered about customers, that data can provide some useful insights to perform the population segmentation where the new demographic data can be analysed. Similar attributes can be found to help identify prospective customers. There is also a classification problem to be solved. To predict if any person is likely to become a new customer.

Dataset and Inputs

This dataset in a real-life customer's data delivered by Alvato Financial Solutions. They provide 4 datasets:

- Demographics information about the general population in Germany. 891,211 samples with 366 features.
- Demographics information about customers of the sales company in Germany. 191,652 samples with 369 features.
- Demographics information about people targeted with the campaign. 42,982 samples with 367 features.

- Demographics information about people targeted with the campaign. 42,833 samples with 366 features.

The general population dataset includes different information such as age, gender, personal profile, purchase information, state of possession, and many other features provided for all individuals.

In the customers' dataset, there are the same features plus 3 additional 'CUSTOMERS_GROUP', PRODUCT_GROUP', and 'ONLINE_PURCHASE'. These extra columns provide more precise information about every customer profile.
The third dataset contains information about targeted people along with additional column 'RESPONSE' indicating whether each person was attracted by the campaign or not. The remaining features are the same.
The last dataset includes test data, the same columns like in the third dataset except for 'RESPONSE' column, for this portion of data we need to predict how likely each person is to become a new customer.

There are also two excel files:

- List of attributes and descriptions segmenting them by the type of attribute e.g. Person, Household, etc.
- A detailed list of attributes with their values described.

Solution Statement

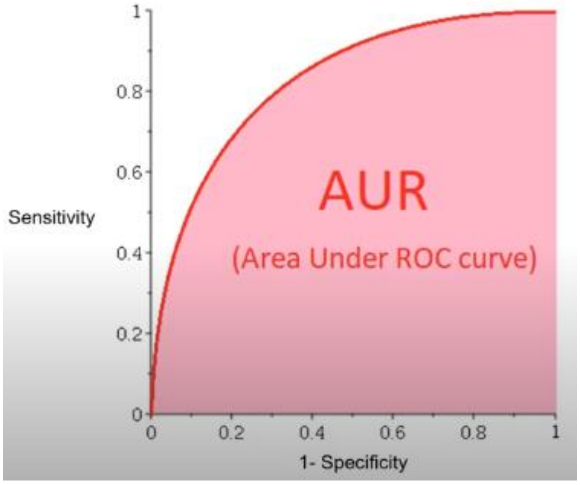
Utilizing available data about each person will help to identify prospective customers more efficiently, reducing the number of failures. The first model will perform population segmentation to identify parts of the population dataset that can be matched with the typical customer profile. K-Means Clustering can be used to find similarities between samples. The second model will perform the classification to determine the probability of each individual becoming a new customer. Supervised classification technique such as Random Forest can be used.

Benchmark Model

Because this a Kaggle competition a benchmark model can be the best AUC score for the test set. The best model got AUC value of 0.81 and my goal will be to get more then 0.79 score.

Evaluation Metrics

Model performance is measured using AUC which results in the ratio between True positive and False positive rates. The data is unbalanced because there are many more individuals who were targeted and did not respond.



Project Design

First, I will start with data exploration, I will use lists of attributes to understand demographics information. Then I will analyze each column to explore the scale of measurements whether this is a nominal, ordinal, or numeric (interval, ratio) scale. It is also worth to check how much missing values are there.

Then using unsupervised machine learning techniques. I will perform clustering on the general population dataset before that I need to use various techniques for dimensionality reductions such as PCA. After segmentation, I will select the most important features which can help to determine which individuals of the general population are more likely to become a new customer.

Then I will apply supervised machine learning techniques to find the best model which will be used to predict the probability of becoming a new customer for each individual in the test dataset.
Before selecting the best model I will perform features engineering to reduce noise. Then checking the correlation between features and using recursive feature elimination I will select the most important features to input into models. Too many features may lead to the model being less accurate, particularly on unseen data. This is because the model can be overfitted. Next, I will perform k-fold cross-validation to select the best performing model. I will test a couple of classification models like Random Forest, Support Vector Classification and XGBoost. Then using grid search I will train the best models across different hyperparameters and select the model with the highest AUC.

References

1. http://scott.fortmann-roe.com/docs/BiasVariance.html
2. https://www.forbes.com/sites/mariyayao/2018/04/10/14-ways-machine-learning-can-boost-your-marketing/#2c7d992911b6
3. https://www.youtube.com/watch?v=egTNM8NSa2k