

Capstone Proposal: Bertelsmann/Arvato Project, A segmentation approach

John Zhao

yzh2013@yahoo.com

Introduction

This is the capstone proposal for Udacity Machine Learning Engineer Nano Degree. The proposal starts with some knowledge and background for the problem. Then I'll talk about the data and features of the projects. Next section is about model selection, training, and hyper parameter training. Last part is about metrics and overall design.

1. Domain Background

This project is about using customers' information for customer segmentation which is unsupervised learning, with feature deduction techniques and customer recommendation and classification (whether to respond to the campaign or not), which is a supervised learning.

Historically, financial company will use heuristic analytical techniques to analysis certain customers and make product promotion. It is a combination of art and science.

With the modern compute power increase, we can now use machine learning to process data and use statistical method to build machine learning models to make inference, whether for classification or regression.

2. Problem Statement

as mentioned earlier, this project is helping a financial firm to decide which customers they should send promotion to based on their demographic data.

There are 3 tasks. First is to make segmentation based on demographic data. Second is to train a supervised learning model to predict whether they should send promotion. Last part is to predict a series of data for Kaggle competition.

3. Datasets and Inputs

The input data has about 1 million rows and 370 columns. Each row represents one client. Each column represents some demographic features. Like age, geo-location, wealth level etc. the customer data has labels of whether promotions are sent to them.

More details of features are saved two explanation files.

The data is relatively clean, only about 6 columns has more than 30% missing, which needs to be resolved before further procedure.

4. Solution Statement

I will use similar approach like the population segmentation model in the first part of machine learning project.

I will start with data cleaning and exploratory data analysis. Next step is to use PCA method to reduce dimensionality. Then I will use K means unsupervised learning method to separate all the population. I will compare the difference/similarity between whole population and customers of the company.

Next session is using supervised learning to predict whether to respond to the campaign with sklearn models.

Last part is to use the model to make predictions.

5. Benchmark Model

In this project, there are not much benchmark models to compare with. If available, I might use the historical data from the company to compare my result with.

Alternatively, I am thinking to build some simple models with default parameters and compare with a fine tuned and more complexed models to see the difference and improvement.

6. Evaluation Metrics

Due to the nature of the imbalance data, I am using AUC-ROC score to measure the model.

I am also using a lot of percentage and percentage difference in the data analysis part to checkout the difference between whole population and customers.

7. project design

This is pretty much covered above. I will start with data cleaning, then I'll do some exploratory data analysis, some feature engineering and then I'll use PCA to change data dimension and feed the result to K means model for segmentation.

Next step is to build models for classification. I will try some easy model and complex models. And decide which is the best candidate. Lastly use grid search cv to fine tune the hyper parameters.