Machine Learning Engineer Nanodegree

Capstone Proposal

Keyvan Tajbakhsh
16st, July 2019

Customer Segmentation for Arvato Financial Services

Domain Background

Nowadays, the importance of treating customers as the principal asset of an organization is increasing in value. Companies are rapidly investing in developing strategies for targeting customers in a specific manner .

The concept of business intelligence has a crucial role to play in making it possible for organizations to use technical expertise for acquiring better customer insight for outreach programs. In this scenario, we will discuss of how machine learning techniques can impact marketing decisions and how to build a application (predictor) for this use (1).

Problem Statement

Despite the extensive literature on market segmentation, there is no overall consensus about the optimal segmentation methodology. One reason for this diversity is the fact that segmentation can be observed from different perspectives (2).

The challenge in the first part of this project is aimed at making use of machine learning techniques in grouping customers regarding their demographics characteristics (3) and secondly try to group them regarding their geographical location in Germany.

The second part is focused on supervised learning using different machine learning techniques for binary classification. The labeled dataset provided by Udacity contains demographics data for individuals who were targets of a marketing campaign. This campaign concern a mail list as labeled column to see if a customer is in the mail list marketing campaign (labeled 1) or not (labeled 0). Logistic regression as a benchmark and neural network would be implemented to achieve this section.

Datasets and Inputs

Customers datasets (Unsupervised dataset) are available on Udacity platform containing demographics characteristics of each individual. The first demographics dataset will be used to create our clustering model. However, due to high number of features, reducing dimensionality before creating the clustering model has to be considered.

Here after, the train supervised dataset containing almost the same features as demographics dataset with an additional labeled binary column (0 and 1); representing if a customer is in the mail list marketing campaign (1) or not (0). This dataset will be used to implement our supervised model. Then we will be able to deploy our model as a predictor which it should be able to predict test data if a customer will be in mail list of the company marketing campaign or not.

It is important to note that the supervised data contains 42962 data points. Moreover, we observe that the labeled column contains 42430 labeled 0 and 532 labeled 1. We should take this into account when we consider to set our hyperparameters and also during evaluation of our model.

Solution Statement

In this project, the key concepts discussed are as follows:

1. Preprocessing and visualiztion
2. Dimensionality reduction with Principal Components Analysis (PCA)
3. Customer segmentation with K-means clustering
4. Binary classification with Convolution Neural Network (CNN)

The dataset provided needs to be cleaned and preprocessed before building any machinelearning model. We observe a large amount of NaN values (NaN), Unknown (-1) and None (0) values all interpreted as missing values. To understand the process step by step we would procees as follows:

1. Identifying and dropping non values columns
2. Convert remaining non values data to numeric values
3. Replacing unknown and none data with NaN
4. Analyzing NaN values in dataset
5. Dropping columns with more than 10% of NaN values
6. Dropping the rows only if all of the values in the row are missing
7. Replacing NaNs with -1 value
8. Checking data type and cleaned values
9. Normalizing the data

Once our data is preprocessed, we will use PCA to reduce the dataset dimensionality. The PCA is a common tool used in data analysis and machine learning projects. Our dataset is composed of 366 features which in relatively large. In order to prevent facing difficulties with K-means clustering we should reduce the number of dimensions using this technique. Hereafter we will be ready to feed our clustering model with our reduced dataset.

K-means is one of the most widely used clustering algorithms since it is simple and efficient. The aim of K-means algorithm is to divide M points in N dimensions into K clusters (assume k centroids). These centroids should be placed in a wise fashion so that the results are optimal which otherwise can differ if locations of the centroids change.

For the final step we will build our CNN model using the labeled dataset (Udacity_MAILOUT_052018_TRAIN.csv). Convolutional neural networks (CNN) utilize layers with convolving filters that are applied to the features. Originally invented for computer vision, CNN models have subsequently been shown to be effective for supervised classification. Note that we should apply the techniques discussed previously (PCA and K-means) on the training and testing dataset prior to build our CNN model.

Benchmark Model

To evaluate our method for this project we will use logistic regression to create our benchmark model. This binary classifier is implemented, then compared to our

CNN classifier performance. The goal would be to outperform the benchmark in order to prove how our model is efficient.

Evaluation Metrics

Once our supervised model is trained and deployed, we evaluate the predicted output in comparison to the training dataset. To evaluate the predictor we'll calculate false negatives and positives as well as recall and precision.

True Positive = Correctly predicted part of mail list (1s are 1s)
True Negative = Correctly predicted not part of mail list (0s are 0s)
False Positive = Uncorrectly predicted part of mail list (0s are 1s)
False Negative = Uncorrectly predicted not part of mail list (1s are 0s)

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

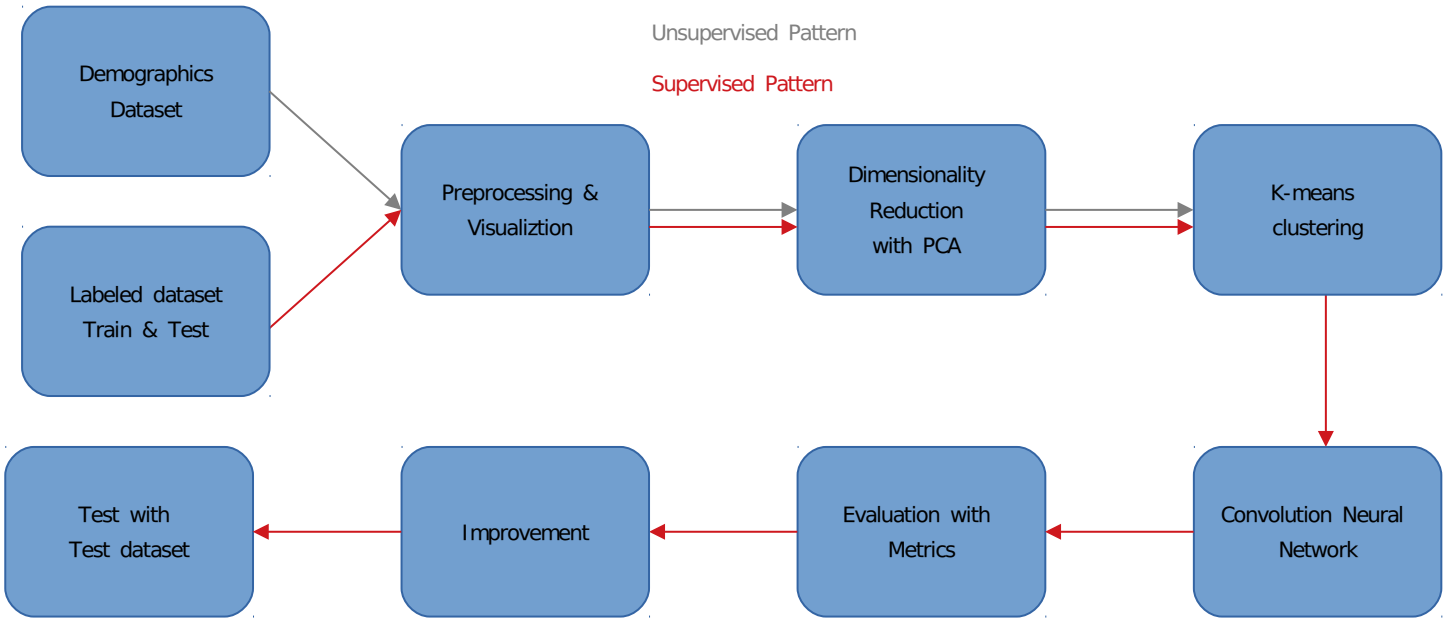
The figure above illustrate how different metrics are calculated.

Project Design

In this section the following chart summarizes the workflow of this project. We distinguish two patterns in gray and red color respectively, representing unsupervised flow and supervised modeling flow. The question we must asking is why we have different dataset for each pattern and not using the supervised data set for both tasks (unsupervised and supervised learning)?

In my opinion there is a need of large dataset for K-means clustering to potentially get an meaningful output while supervised learning needs smaller dataset to train an efficient model.

One of the most important aspect of the workflow described below is the improvement section. The improvement can be done through tuning hyperparameters of our supervised model and refining evaluation metrics such as recall and precision to compensate the imbalance of the dataset.



Project work flow differentiated with two pattern
gray as unsupervised pattern and red supervised pattern

Publications

https://www.researchgate.net/publication/326403359_Market_Segmentation_Analysis_and_Visualization_Using_K-Mode_Clustering_Algorithm_for_E-Commerce_Business

https://www.researchgate.net/publication/326706602_Approaches_to_Clustering_in_Customer_Segmentation

http://spider.sci.brooklyn.cuny.edu/~kopec/cis718/fall_2005/sdarticle2.pdf