Machine Learning Engineer Nanodegree
Capstone Proposal

Mateus Medeiros Furquim Mendonça - mateus@mfurquim.dev

1 Domain Background

One of the principal steps in the machine learning process is the feature engineering. It falls into the Data Preparation step in the CRISP-DM, which is usually the most time consuming (60%~70% of time in overall project)[1]. In feature engineering there are two different approaches: **feature selection** and **feature extraction**. Principal Component Analysis (PCA) is a process of feature extraction to reduce the dimensionality of the data by understanding the individual significance for a feature on the result of the outcome[2]. If two features are directly related and change at the same rate, then there is no need to have both of them to predict the output. For instance, if there are two features which measure a phenomenon, one in meters and another one in inches, they are measuring the same thing but in different scales. So there would only be need to keep one of them.

2 Problem Statement

It is expected that the Principal Component Analysis technique will improve the time of training and accuracy of prediction. This project aims to compare the performance of the model with and without applying the PCA techinque for reduction of dimensionality. By the end, there should be a better understanding on the relationship between the percentage of variance retained and the performance of training and prediction.

Furthermore, this project will also compete in the *Udacity+Arvato: Identify Customer Segments Kaggle Competition*[3], The goal of the competition is to help a mail-order company, which sells organic products, to increase efficiency in the customer acquisition process. To figure out which people in Germany are most likely to be new customers, the attributes of existing clients are analyzed and matched against a bigger data set that includes attributes for people in the country.

3 Datasets and Inputs

The data set used in this project will be the Arvato's data set provided by the course in the Kaggle Competition[3], The Arvato's data set is a compilation of financial data from their customers and the Germany demographic data. The customers data set has 369 features (columns) and almost 200 thousand observations (rows), whereas the Germany data set contains 366 features (columns) and almost 900 thousand observations (rows).

Those features are separated in the following different levels of information: (**person**) description of the person and his/her habits; (**household**) number of people, estimated net income, and the structure of the house; (**building**) number of households and type of building; (**microcell**) CAMEO[1] typology, number of family houses in the cell, and share of car owners; (**transactional**) unique activity data regarding the mail-order activity of consumers; (**postcode**) distance to next metropole and density of inhabitants; (**automobile**) share of cars differentiated; and (**community**) share of unemployed person and number of inhabitants in the community.

4 Solution Statement

The proposed solution is to understand the relationship between all 369 features and the likelihood of a German citizen being a customer. Because of the nature of the problem - i.e., having labeled data - we will train a supervised model to predict a potential customer and help the marketing team to focus their effort on them.

Furthermore, a Principal Component Analysis will be applied on the training set. A total of K dimensions will be kept such that at least 85% of the variance is retained. The model will be trained using $k \in [K, m]$ features, where m is the total number of features (366). For each dimension reduction, there will be a corresponding plot for the model's training and prediction performance.

At the end, there should be a clear visualization on how the dimensionality reduction affects the training speed and prediction accuracy.

5 Benchmark Model

For this problem, the model without PCA will be used as benchmark to compare the solution with different number of dimensions and percentage of variance retained. The best model will be submited to the aforementioned kaggle competition and compared with the current leaderboard.

6 Evaluation Metrics

Training performance will be measured by speed and prediction results are going to be evaluated using the area

1. The CAMEO profiles contain socio-demographic and lifestyle data at microcell level based on parameters such as age, education, income, and general interests.

under the ROC curve[2]. Another secondary evaluation metric will also be used just for informative purpose: the F1 Score - which is defined as the harmonic mean of precision and recall. It is specially useful when having imbalanced data because it takes into account both errors (false positives or false negatives)[4].

7 Project Design

7.1 Explore and Process Data

After downloading the data from the Arvato kaggle competition[3], it comes the Exploratory Data Analysis (EDA) step. In the EDA, the data will be cleaned, explored, and pre-processed. The NaN (Not a Number) and missing values are going to be treated accordingly to the feature type (binary, categorical, continuous etc.) and context. Further features might be engineered in order to consolidate one or more features.

Most machine learning models expect standardized data values. Therefore, this step might also involve normalizing and converting the format of the data. In addition, the data will be split into training[3], validation[4] and test[5] data sets.

Visualize data.

7.2 Modeling

This step will focus on developing the model. A model will be selected and trained using the training dataset. After creating the model, the time of training is going to be recorded and the model will be evaluated and validated by the chosen metric - **Area Under the ROC Curve** (AUC) - against the validation data set.

7.3 Principal Component Analysis

When the first model (benchmark model) is developed and evaluated, the next step will be to apply the PCA technique to reduce the number of dimensions such that at least 85% of the variance is retained. For each dimension reduction, a new model will be trained and evaluated in order to get the corresponding computation time and the AUC.

References

[1] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems*. Apress, 2017. [Online]. Available: https://books.google.com.br/books?id=9ClEDwAAQBAJ

[2] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[3] Arvato. (2018, Nov) Identify customers from a mailout campaign. [Online]. Available: https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview

[4] J. Korstanje, "The f1 score," 2021. [Online]. Available: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

2. The Receiver Operating Characteristic curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

3. Training dataset is used to train the model.

4. Validation dataset is used for model tuning and selection.

5. Test dataset is used after training for evaluation of the model