Capstone Proposal
Udacity Machine Learning Engineer Nanodegree

Felix Hauer

March 4, 2020

1 Domain Background

A business should know its customers. Not only to adapt the supply of the kind of goods the company produces or develops but also to use advertising right to get the attention of potential customers. Here it is important to speak out to the people who are most likely to buy a certain product and avoid those who are not interested at all. This will lead to a maximized return of each dollar that is put into the marketing budget. In this project, the business is a mail-order sales company in Germany.

My motivation for this project comes from my current profession. I work as a data scientist for risk models in the financial industry. This is a highly regulated business with strict limitations of the use of possible models and methods. This is the reason I want to try out some things I learned in the nano degree program here because I am not able to do this at my job.

2 Problem Statement

The first part of the project is to determine similarities of the demographics of the general population of Germany with the demographics of customers of the mail-order company. In the second part, the problem that was already mentioned in the first chapter of the proposal is to deliver the advertising to the right kind of customer. This is done with the help of a machine learning model that needs to predict a probability of how likely a customer will buy a product. The success of this model is quantified with the metrics that will be introduced in the *Evaluation Metrics* chapter.

3 Datasets and Inputs

The data that is used throughout this project consists of four datasets and was provided by Bertelsmann Arvato Analytics. The

- *AZDIAS* set consits of $891,211$ observations and 366 features and the

- *CUSTOMERS* set consits of 191, 652 observations and 369 features.

These two are datasets that are used for the first part of the project, which is a customer segmentation report. It is an unsupervised machine learning task, so there are no features that act as a target. The

- *MAILOUT* training set consits of 42, 982 observations and 367 features and the
- *MAILOUT* test set consits of 42, 833 observations and 366 features.

These two are datasets are used for the second part of the project, which is a supervised learning model. The 367th feature is the *response*, that needs to be predicted. In the test set this not available and it will be evaluated within the Kaggle competition. In addition to that, there are two Excel spreadsheets available that contain information about the data including attributes and the range of the data values.

4 Solution Statement

The solution will be a machine learning model. This can be interpreted as a mapping in the mathematical sense:

$$f \colon \mathbb{R}^p \mapsto [0, 1]$$

where $\mathbb{R}$ is the set of real numbers, p is the number of features after pre-processing the data and $[0, 1]$ is the interval that includes all real numbers from 0 to 1. f is an element of all machine learning models, for example, logistic regression, gradient boosters or neural networks. The value $f(x) \in [0, 1]$ will be interpreted as the probability that a recipient of the mailing will become a customer of the company. With the given datasets it is possible to train such a model and evaluate it.

5 Benchmark Model

As a baseline to check whether a model produces reasonable results, the first benchmark is a model that assigns every observation a positive label. This corresponds to the relative number of positive classes in the dataset. A developed model should at least beat this. The second benchmark will be a vanilla model. This is a very simple model. I will use a simple decision tree, which can be fitted which a small number of hyperparameters and is highly interpretable. The third benchmark will be the *Kaggle* leaderboard. Here I can compare my model with a community of the best data scientists of the world. I think a realistic expectation is to reach the top 25 %.

As already mentioned I will quantitatively compare the baseline model with the accuracy and the Kaggle competition with the AUC.

6 Evaluation Metrics

In this project, there will be two metrics to evaluate the performance of the model and one additional method to describe the predictive power of the model.

The first metric will be the accuracy which is defined as follows:

$$\text{acc} = \frac{TP + TN}{N}$$

where N is the number of predicted observations, TP (true positive) is the number of right predictions that are in the positive class and TN (false negative) is the number of the right predictions that are in the negative class. The accuracy is the metric in which one of the benchmark models is also evaluated, so it will be used to compare the two models. As the model will give us a score or a probability for each observation, one will need a *threshold* to assign a class. This problem leads to a metric that can be calculated without a threshold.

The second metric is the *area under the ROC curve* or short *AUC*. The ROC curve is generated by plotting the true positive rate against the false-positive rate. The AUC is usually a number between $\frac{1}{2}$ and 1 corresponds with the U-statistic from the Wilcoxon–Mann–Whitney test[1]. It is interpreted as the probability that a randomly drawn observation from the positive class has a higher score than a randomly drawn observation from the random class. The *AUC* is used in the Kaggle competition.

I also want to evaluate if the predicted probabilities are well-calibrated. This is important to do a cost calculation of a commercial campaign. The *AUC* is not suitable for this task, because is invariant to the absolute values of the class prediction. For example, you could add 0.1 to each prediction and the *AUC* would stay the same. I will use a so-called *calibration plot* to visualize this.

7 Project Design

The first part will be explorative data analysis to get an overview of the datasets. This will include calculation basic summary metrics, checking the datatypes of the predictors and finding and understanding missing values. Then I will pre-process the data. Numerical predictors are probably skewed and have outliers. As you can see in the Excel spreadsheets many categorial predictors have two values of missing data. They could probably be merged into one value.

The second part will be an unsupervised analysis of the *AZDIAS* and *CUSTOMERS*.

[1] $AUC = \frac{U}{n_1 n_2}$ where n_i, $i = 1, 2$ is the number of observations in class n_i.

Here I will use principal component analysis and k-means clustering to explain how these to datasets differ or correspond with each other.

Finally, the *MAILOUT* data will be examined with supervised machine learning techniques. I will first use a set of the most common algorithms, for example, support vector machines, decision trees, gradient boosters, and neural networks. Then I decide which of these performs best in terms of cross-validation AUC and will fine-tune it. With this model, I will enter the Kaggle challenge and finish this project.

8 References

- *The Elements of Statistical Learning, Hastie, Tibshirani and Friedman; Springer 2013*
- *Applied Predictive Modeling, Kuhn and Johnson; Springer 2013*
- `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`