Capstone Proposal Udacity Machine Learning Nanodegree

Arno Gils

January 2020

Abstract

This capstone proposal is intended to give a clear overview of the way in which applying machine learning algorithms and techniques can help you solve a real world problem. In this case we focus on how a company which is selling organic products can acquire clients more effectively. Specifically it will give a general overview of the steps which need to be taken to separate customers into meaningful segments. Furthermore it will highlight which steps actually need to be taken to predict if customers in these segments are likely to buy something.

1 Domain background

The retail environment has witnessed a drastic transformation, especially in the last two decades, with the rapid growth of electronic commerce (e-commerce). Although the application of technology-based online retail services has grown rapidly in recent years, an understanding to attract, retain, and satisfy customers in such contexts remains limited (Sahney, 2008). Deep insights in these areas therefore could greatly benefit a company, e.g. by targeting potential new customers more effectively. This capstone is about how these insights can be generated by utilizing different machine learning techniques and using data from a mail order company which is selling organic products.

2 Problem statement

How can the mail order company which is selling organic products acquire more client effectively?

3 Datasets and inputs

The data is provided to us by Arvato Financial Services with the sole purpose of being used in this project. It consists of four separate datasets:

- **Udacity_AZDIAS_052018.csv**: Contains demographics data for the general population of Germany.
- **Udacity_CUSTOMERS_052018.csv**: Contains demographics data for a mail-order company in Germany.
- **Udacity_MAILOUT_052018_TRAIN.csv**: Contains demographics data for individuals who were targets of a marketing campaign. Will be used to train a supervised learning algorithm.
- **Udacity_MAILOUT_052018_TEST.csv**: Contains demographics data for individuals who were targets of a marketing campaign. Will be used to evaluate the performance of a supervised learning algorithm.

4 Solution statement

Increasing efficiency in targeting potential new customers can be achieved by combining both unsupervised and supervised machine learning techniques and algorithms. Unsupervised learning focuses on generating insights and patterns from the available data without having available an output label. In the first part of this capstone project we will implement a K-means algorithm to cluster customers into separate segments.

In the second part of the capstone project we will use supervised machine learning. Supervised machine learning focuses on generating an output based on data which consists of both an input and output label. To create a probability about which customers are likely to buy something we will implement in this capstone project a two binary classifiers. Specifically we will use a Decision Tree and a Random Forest.

5 Benchmark model

To evaluate the performance of our algorithms we first implement a Decision Tree. This will serve as a benchmark for the Random Forest model. We will use a Decision Tree as a benchmark because in general they perform well out of the box for tabular data.

A second benchmark which can be used during model development is the Kaggle leaderboard. We evaluate the performance of our model against this leaderboard to see how we rank up against the implementation of other students. This way we have a general sense about if we can implement further improvements.

6 Evaluation metrics

While looking at the data, the distribution of class labels seems to be a bit unbalanced. Therefore precision and recall will be used as evaluation metrics for the supervised learning part of this capstone project.

Precision = tp/tp+fp. Precision measures which proportion of positive identifications was actually correct, where TP = true positives and FP = false positives.

Recall = tp /tp+fn. Recall measures which proportion of actual positive class labels was identified correctly, where TP = true positives and FN = false negatives.

7 Project design

To complete this capstone project the following steps are roughly needed.

- Data prepararation. This comprises of exploratory data analysis (EDA) and data cleaning. EDA will be used to gather insights about the data. As a second step we will clean the data to get it in the most suitable format to train a machine learning model (both unsupervised supervised).
- Model development. The algorithms we use in this project all come from the Scikit-learn Python library. Specifically the K-means, Decision Tree and Random Forest implementations from this package will be used.
- Model training. Models will be trained on the available training datasets.
- Model evaluation. The performance of the models will be evaluated by using the available testing datasets and using the precision and recall metrics described above.
- Hyperparameter tuning. To improve the performance of the models we will use hyperparameter tuning to find the optimal set of parameters of the model for the given problem at hand.
- Model deployment. The model will be exposed as an endpoint. The idea behind this is that the model then can easily be reused for new cases / or be integrated in a technical architecture.

Literature

Shaney, S. 2008. Critical Success Factors in Online Retail – An Application of Quality Function Deployment And Interpretive Structural Modeling.