STUDENT: Damir Valput

Nanodegree program: Machine Learning Engineer

CAPSTONE PROJECT PROPOSAL: Starbucks Project

Domain Background

The problem I will solve in this capstone project is one of the three proposed ones dealing with Starbucks simulated data set.

The problem domain is in the field of targeted marketing (i.e. providing targeted offers), an area in which through data analysis and predictive modelling we wish to better understand different customer profiles and what types of offers they respond to. Many companies that offer specific products or services benefit from using targeted marketing strategies. Presenting customers with targeted offers is very beneficial to businesses as it encourages customers to choose that company over a competitor, increases chances that people spend money on a certain product, and yields more focused marketing efforts in terms of a higher ROI (Return on Investment) [1]. Customers should be targeted in the appropriate and relevant manner: too intensive targeting can result in their fatigue, whereas too weak advertising can result in profit loss or missed opportunities.

Data science, and artificial intelligence techniques in general, have been successfully applied in marketing for some time now, ranging through a variety of applications: lead targeting, customer segmentation, sentiment analysis, pricing strategy, product development, etc. [2] Data science and machine learning in targeted marketing are particularly useful as they can provide insights about customers' behaviour and decision making beyond what is intuitive to a human brain, which consequently allows business expansion into new markets or places. [2]

I find this problem particularly interesting as it focuses on applying machine learning to a problem in an area that is novel for me, and thus I believe it will enable me to learn a great deal of new concepts and techniques.

Problem Statement

In this problem, I am presented with three different datasets on Starbucks customers and the transactions they made in the past. My task is to combine transactional, demographic and offer data to determine which demographic groups respond best to which offer type.

I intend to perform market segmentation of the customers in the database, i.e. customers will be categorised according to their characteristics into similar groups (clusters) with respect to

their historical purchasing behaviour and demographic characteristics. This will be done taking into account the information whether a purchase was done as a response to an offer since I wish to understand what a certain customer group buys when not receiving any offers. Furthermore, upon the market segmentation, I plan to design a predictive model that answers the question whether a particular customer will respond to a certain type of offer and therefore whether the offer should be sent to them.

Datasets and Inputs

The dataset for this project was provided by Starbucks and it consists of simulated data that mimics customer behavior on the Starbucks rewards mobile app. This dataset is a simplified version of the real Starbucks app data because the underlying simulator only has one product, whereas Starbucks actually sells dozens of products. The dataset is structured (tabular data) and contains these three tables:

- “portfolio” - contains metadata on different offer types provided by Starbucks
- “profile” - contains demographic data for each customer registered in the app, e.g. age, gender, income, etc.
- “transcript” - consists of records of transactions, offers received, offers viewed, and offers completed for each customer in the “profile” table

Each table is in .json format. The data will be loaded, cleaned and analysed via a standard exploratory data analysis process, with the objective of having them ready for their usage in the feature engineering process.

Solution Statement

The project will follow a standard machine learning workflow: starting with the loading and explorative analysis of the datasets, cleaning the data and feature engineering. Two machine learning applications will be developed and tested.

Firstly, an unsupervised clustering algorithm (most likely k-means or HDBSCAN) will be developed to perform market segmentation, i.e. to cluster the customer into similar profile groups based on their demographic and purchasing characteristics. In the second machine learning task, a supervised learning algorithm will be developed and validated with the objective of predicting whether a particular customer will respond to a certain offer type, i.e. whether that offer should be sent to the customer. This is a binary classification problem. For this ML task adequate features will be selected based on correlation analysis, and the results of the clustering algorithm could be used as features in this task if deemed useful and appropriate. Moreover, to prevent any data leakage, the dataset will be split into train and validation sets before any data preparation. [3]

Benchmark Model

A benchmark model for the supervised learning problem will be designed to compare the developed binary classification model to. The benchmark model will be a random generator that will decide whether a client should be presented with an offer or not by flipping a fair coin (50% chance for each outcome). The performance of the benchmark model will be measured using the same set of evaluation metrics as for the defined solution, and I expect to see a significantly better performance of my developed binary classifier compared to the random decision maker.

Evaluation Metrics

With clustering algorithms there is nothing to test and evaluate since there is no ground truth. However, certain metrics can be used to examine the quality of clustering and they might be considered in this project as well. For example, to determine an optimal number of clusters I might rely on the elbow method.

In the problem of binary classification, a standard set of evaluation metrics for classification algorithms will be used. Concretely, I will rely on accuracy, confusion matrix, F1 score, and any other metric deemed appropriate for the case study in question (precision, recall, ROC AUC score, etc.).

Project Design

As mentioned already, the project will follow a typical machine learning workflow. I will start with an explorative descriptive analysis (EDA) to better understand the datasets at hand and detect any outliers, anomalies, or missing or erroneous values. Afterwards, the data are going to be cleaned and prepared for the feature engineering process. Before the feature engineering, the dataset will be split into training and validation dataset (a test data set will be considered as well depending on the final dataset size).

For the clustering task, a k-means or HDBSCAN algorithm will be considered. PCA will be applied to reduce the dimensionality of the data set.

To solve the classification problem, I will consider and test several different machine learning models. Among the potential candidates are logistic regression, decision trees based models (e.g. random forest), SVM, Naive Bayes, XGBOOST classifier, or a customer neural network. I will make the final decision on the models to try out upon performing the full exploratory and correlation data analysis. I will also consider ensemble learning, i.e. ensembles of the above proposed learners.

References

1. Targeted Marketing: Explore the Strategy of Targeted Marketing. Last updated: November 28, 2020. Available at: https://www.marketing-schools.org/types-of-marketing/targeted-marketing/#:~:text=Targeted

%20marketing%20identifies%20an%20audience.for%20those%20preferred%20market%20segments.
2. Data Science in Marketing: A Comprehensive Guide (with examples). Published on: May 26, 2020. Available at: https://nogood.io/2020/05/26/data-science-marketing-guide/
3. How to Avoid Data Leakage When Performing Data Preparation. Updated on: August 17, 2020. Available at: https://machinelearningmastery.com/data-preparation-without-data-leakage/