UDACITY

MACHINE LEARNING ENGINEER NANODEGREE

Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Ozayr Moorad

April 22nd 2020

DOMAIN BACKGROUND

Arvato Bertelsmann is a global services company founded in 1996 situated in Gutersloh Germany (Ostrowski, 2016, p. 13). It provides services in customer support, logistics, finance and IT. The domain of interest for this project is the financial solutions services that Arvato provides, more specifically data driven targeted marketing. The client, a mail order company, requires information as to what type of demographic group their customers fall into. They wish to use this information to concentrate their marketing campaigns to ensure maximum customer acquisition. Hence the goal of this project will be to utilize the provided data augmented with machine learning techniques to uncover patterns that a human analyst might miss.

THE CLIENT

Not much is known about the client except that it is a mail order company that's wants to ensure datapoints are being used to facilitate the decision-making process. I thought it apt to add a few points from the definition of a mail order company that I had come across during my research (Mail-Order-Business, n.d.)

- Mail-order businesses date back to pre-Revolutionary War days, when gardeners and farmers ordered seeds through catalogues.
- Historically, mail-order businesses became successful because they offered a wider variety of goods than could be found in local retail outlets.
- goods purchased through the mail were often less expensive than those available locally.
- mail-order houses, blessed with the capacity to maintain far larger inventories than many of their retail competitors, could afford to offer more sizable discounts.
- Indeed, for consumers in remote rural sections of the country, their isolation from commercial centres made catalogue or mail-order shopping a necessity.
- Finally, individuals pursuing a hobby or special interest were more likely to locate those hard-to-find items in a specialty catalogue than in a store.

One of my motivations for choosing this project was to test the hours of data science reading that I have done over the past few months ever since the beginning of my journey into Machine learning, I believe in order to be a good machine learning engineer one has to have good data science skills. This project would also give me the ability to put my insight extraction and philosophical skills to the test on something that has real world implications.

PROBLEM STATEMENT

The question we seek to answer as put forth by Timo Reis from Arvato is "How can the mail-order company acquire new clients more efficiently?".

For the first part of the project, one set of solutions we may investigate or implement is the utilization of machine learning models that will allow us to uncover patterns and insights from the data provided. We may use unsupervised learning techniques such as KMeans clustering to cluster the population and customers into segments giving us an idea of who or what type of people our customers are. The second part is well suited for supervised machine learning, where we are trying

to predict who will respond to a mail order campaign given historical data of who had responded to a previous mail order campaign.

DATASETS AND INPUTS

The following information on the datasets to be used are taken as if from the project workspace.

There are four data files associated with this project:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Provided along with the data set are 2 additional files. Viz:

- DIAS Information Levels - Attributes 2017.xlsx

And is described as a top-level list of attributes and descriptions, organized by informational category.

- DIAS Attributes - Values 2017.xlsx

is a detailed mapping of data values for each feature in alphabetical order.

These files will be used to gain an understanding of the features which will help with the cleaning as well as the interpretation process.

SOLUTION STATEMENT

My proposal deviates a little form the norm, I propose using an unsupervised method to detect and understand segments that customers fall into, I then will formulate the same problem from a supervised perspective to further enhance the Interpretability of the insights gained. This will be done by labelling the general population and customers as separate classes. Thereafter optimizing a machine learning model to predict, given the data, if the individual belongs to the customer class or not. I will then use shapely values which are both accurate and consistent (Lundberg, 2018) to explore the effects of perceived important features and their interactions. (A Unified Approach to Interpreting Model Predictions, 2017). In my opinion this better allows me to quantify and measure the results obtained from the unsupervised section. I will then use the insights gained to produce a supervised model that can predict the probability of a customer responding to a mail order campaign.

BENCHMARK MODEL

The benchmark for the customer segmentation part of the problem is formulated from a supervised learning perspective, to facilitate a quick method of getting some kind of benchmark I fused the customer and population datasets labelling them respectively. After dropping columns with more than 80% missing values, I fitted an XGBoost model which achieved a 0.93 AUC score to predict the customer class. I then used Shapely values (A Unified Approach to Interpreting Model Predictions, 2017) to get some quick insights into important features and their interactions, this pre analysis will be detailed in the final report.

For the mail order response prediction again an XGBoost model was fitted on the raw data with minor clean up which attained an AUC score of 0.72

EVALUATION METRICS

I plan on using the formulation of the customer segmentation part from a supervised perspective as sort of a metric to compare the unsupervised model against.

For the supervised mail-out response problem, due to the stark imbalance in the dataset AUC score will be used. Also due to the fact that this is what the Kaggle competition scores against.

PROJECT DESIGN

1. Data Dictionary analysis: The relevant information files need to be investigated to get an understanding of the data we are being presented with.
2. Pre-Analysis: With minimal clean-up i.e. Dropping features with missing values, dropping object features and fitting an XGBoost model to the data then using shapely values to explore feature interactions and further our understanding of the data.
3. Data memory optimizations: Seeing that the data frame is quite large, it will be worth it to look at ways that it can be processed in order to be sorted more efficiently. This may also result in speedups down the processing pipeline.
4. Exploratory data analysis:
    a. Feature analysis: Analyse features, check against data dictionary. Do we have documentation for all features? what are the feature data types? if there are groupings, what groups do our features fall into?
    b. Missing value analysis: Do missing values have a pattern or structure or is it random? Can missing values represent some underlying information about data points or are they just missing values? How can we fill missing values intelligently without introducing bias into the dataset?
    c. Feature Engineering: How can we engineer new features to better represent the data or interactions within the data? which features are redundant? seem incorrectly documented?
    d. Correlation analysis: Investigate correlation between features to gain further insight into what features mean and how they interact. Will get a better understanding of the data set in general.

5. Pre modelling processing: Before fitting a model to the data we need to pre process the data such as scaling and PCA for dimensionality reduction as this will enable our unsupervised and supervised models perform better.
6. Model Selection: We can then fit multiple models to the data to investigate the best one
7. Tuning and enhancement: The chosen model will then be tuned to further increase performance and generalization.
8. Predict and evaluate: Apply our final model to the data and use the evaluation metrics to evaluate our results.

WORKS CITED

A Unified Approach to Interpreting Model Predictions. (2017). *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 10.

Lundberg, S. (2018, April 17). *interpretable-machine-learning-with-xgboost*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Mail-Order-Business. (n.d.). Retrieved from www.referenceforbusiness.com: https://www.referenceforbusiness.com/small/Inc-Mail/Mail-Order-Business.html

Ostrowski, J. (2016). *"Arvato baut aus".* Neue Westfälische (in German).