Machine Learning Engineer Nanodegree

Capstone Proposal

Jonathan Hasan

1/1/2020

Customer Segmentation for Arvato Financial Services

Background:

The road to AI utilization in the financial industry took a long time to develop. In the 1950's and 1960's AI was not viewed favorably due to poor performance. At the time, there was little electronic data and compute power and thus AI could not achieve anything meaningful. This kickstarted the so called "AI Winter" where there was a notable lack of interest and investment in artificial intelligence. The 1980s showed a revival of interest in AI thanks to pioneers like Geoffrey Hinton who developed a lot of the underlying neural network theory that allowed computers to perform extremely well on image classification tasks. Countries all over the world began to take note of these improvements and began to pour massive funding into AI initiatives. Japan, the UK, and USA have had programs started through the Fifth Generation Computer Project, Alvey Program and the Strategic Computing Initiative respectively. It is also interesting to note that Rennaissance technologies founded by Jim Simons also started during this time and was one of the first hedge funds to utilize machine learning and AI in their everyday trading operations. The medallion fund is one of the most successful funds to ever exist, it only had one year where the returns were negative. The financial industry now benefits from AI thanks to all the research conducted by scientists in the 1980s.

The financial industry has an abundant amount of opportunities for applying machine learning to increase profits. It has managed to find its way into hedge funds, international banks insurance companies, regulatory agencies and Fintech. Supervised and unsupervised methods are successfully used every day for financial service tasks. According to the report titled "Artificial Intelligence in Finance" published by Bonnie G. Buchanan in April 2019, shows that there are three areas where AI is playing an instrumental role. These roles are fraud detection and compliance, chatbots and robot-advisory services and algorithmic trading. There are many more financial applications besides these and one of those will be shown in this project which is customer segmentation and binary classification.

Problem Statement:

In this project, the focus will be on customer segmentation and creating a supervised learning model for determining who will respond to a mail order campaign. Customer segmentation is the process of dividing a customer base into groups based on similar characteristics. Using feature engineering, features will be generated to try and devise a model for finding which people are most likely to be a customer.

Datasets and inputs:

Two datasets are provided called azdias and customers. The azdias dataset has demographics data for the general population of Germany and has the data of 891,211 people with 366 data points on each one. The customers dataset has the data of 191,652 people and has 369 features for each person. Each row of these files is a person and includes information not only about the individuals themselves but information about their surroundings, such as household, building, and neighborhood. These two files are the staging ground for determining the best features for the model. The customers dataset needs to be compared to the azdias dataset to see who would be most likely to become a customer. Once the optimal features are chosen then it is time to move on to the train dataset for the mailout dataset.

Solution statement:

The solution is a model that will be able to accurately determine whether an individual will be a customer given some attribute information. There are 369 features for each person to choose from. The model will make use of the most important features while discarding unnecessary ones. The model needs to be better than a random guess which would mean 50% accuracy. A good model in this case will achieve an accuracy of 90% or greater.

Benchmark Model:

I will try two different approaches to this problem. I will be creating a neural network with pytorch as well as using scikit learn models such as linearSVC to correctly classify whether a given individual will be a customer. The scikit learn linear models only work well on data that can be separated by a linear decision boundary. I need to reduce the feature space to allow this to happen. Two scripts will be provided that define training and prediction for each one as well as an extra model script for the pytorch model.

Evaluation Metric:

For this project, an accuracy metric will be used. This will determine how many correct classifications the model made. In addition, a confusion matrix will be used to see how many false positives and false negatives there are. Using the results from the confusion matrix, I will also try and focus on the recall metric as well. After all, it is desired to identify as many potential customers as possible.

Project Design:

Part one is Exploratory Data Analysis. We want to see how the data breaks down. We want to answer questions pertaining to how data is distributed, if there are any patterns, and summary statistics to name a few. It's also important to determine if there are any class imbalances. The number of people who did end up being customers have to be compared to the number of people who didn't become a customer. If we get a high accuracy but the data is not evenly distributed between positive and negative classifications, then it may not generalize well. Afterwards, it will be time to move to step 2 which is feature selection and engineering. There are many features that are available to choose from. Some of these features will play a larger role than others. It will be necessary to choose those features and remove all the rest. This will be done with statistical analysis techniques such as Anova. Finally, the chosen features will then be used as a basis for training a model with Scikit learn and a Pytorch neural network. These models will be trained on a training dataset provided by Arvato Financial Services and then used on a test dataset on Kaggle.

Sources:

1. https://machinelearningmastery.com/an-introduction-to-feature-selection/
2. https://www.deeplearning-academy.com/p/ai-wiki-evaluation-metrics-in-data-science
3. https://www.turing.ac.uk/sites/default/files/2019-04/artificial_intelligence_in_finance_-_turing_report_0.pdf