Capstone Proposal

Mani Singh

6/01/2020

Machine Learning Engineer Nanodegree

Market Segmentation Capstone Proposal

Domain Background

Market segmentation is the process of dividing customers in a given market into groups based on common attributes. This is significant in marketing because it allows for personalization and more accurate targeting in advertising campaigns. Personalization is staritng to become necessary in marketing with 74% of customers expressing frustration when website content is not personalized. Moreover, 59% of customers say that personalization influences their purchasing decision. Instead of a generic, one size fits all offer, the statistics show that marketing should harness the power of personalization to provide targeted advertising to customers.

The idea of segmentation in marketing is not a new concept. In the 1600s, retailers would segment their customers by wealth. For example, shop owners would hold private showcases of goods in their home for wealthier customers. Also, in the late 1800s, German toy manufactuers were using geographic segmentation to produce toy models for specific countries, like American locomotives intended to be sold in the United States. Segmentation today is also done based on demographics and geography but has adapted to also include behavior and the customer's journey with respect to the buying process.

Problem Statement

Creating accurate and useful customer segments can be a daunting task. To create customer segments, a lot of reserach is required and a company may find that they need to conduct a study. This process can be time consuming and expensive. Often times, marketers are able to create simple segments, for example segmenting by country, age, gender, etc. But these segmentations are simple and do not contribute to understanding different segments of customers.

Datasets and Inputs

There are four data files that will be used in this project and they were all provided by Arvato Financial Solutions. The "Udacity_AZDIAS_052018.csv" file contains demographic information for the general population of Germany and has 891211 rows and 366 features. The "Udacity_CUSTOMERS_052018.csv" file contains data for the customers of a mail-order company and has 191652 rows and 369 features. This dataset has 3 extra features which provide details about the customers. The last two datasets, "Udacity_MAILOUT_052018_TRAIN.csv" and "Udacity_MAILOUT_052018_TEST.csv", contain demographic information about people who were targets of a marketing campaign. They have 42982 and 42833 rows respectively. The TRAIN file has 367 features while the TEST file has 366 due to the fact that the outcome variable, whether or not the recipient became a customer of the company, is left out of the TEST file so proper evaluation can be performed.

Collectively, the features of all four datasets are categorical. The files "Udacity_AZDIAS_052018.csv" and "Udacity_CUSTOMERS_052018.csv" will be used in unsupervised learning methods in order to create customer segments. These two datasets are very feature-rich which means the customer segments will give insight into the different types of groups that are customers of the company as well as their similarities with the general population. The last two files will be used to create a supervised learning model based on the customer segments. Specifically, the “Udacity_MAILOUT_052018_TRAIN.csv” will be used to train the model while the “Udacity_MAILOUT_052018_TEST.csv” will be used to test the model.

Solution Statement

In order to solve this problem, the first task will be to fit the CUSTOMER data to a clustering model, like K-means. This will yield customer grouping or segments. Clustering will also be used on the AZDIAS data, which fill find groupings for the general population of Germany. Then after analyzing the similarities between these two groupings, a supervised model, like logistic regression, can be used to make predictions. It will be trained on the TRAIN data and evaluated on the TEST data.

Benchmark Model

The benchmark model that will be used is logistic regression. Logistic regression is a fairly simplistic, supervised classifier that can be used to predict whether or not a person becomes a customer. Classification problems can be measured using various metrics like, accuracy, precision, and recall. However, for the benchmark model, just accuracy will be used. Logistic regression models can be prone to overfitting so it is expected that the final supervised model chosen will outperform the benchmark model.

Evaluation Metrics

As mentioned previously, the evaluation metric that will be used is accuracy. Accuracy will be a percentage and has the formula $\frac{\text{Number of Correct Predictions}}{\text{Number of Predictions}}$. The outcome variable of the supervised model is whether or not a person becomes a customer, which can be encoded with 1 or 0, where “1” is the person becomes a customer and “0” is they do not become a customer. Calculating the number of correct predictions would simply be getting the equality between the predicted vector and the actual vector. The number of predictions can be calculated by getting the length of the prediction vector.

Project Design

The first task of this project will be data cleaning. Data cleaning will consist mainly of two parts, deleting null values and removing unncessary columns. The features of the data are categorical, so methods like filling the null values with the median value of the column will not work. Given that each dataset has many observations, if there are only a small number of null values, they can be removed without significantly reducing our data. However, if there are many null values with proportion to the datasets, another method could be to use an machine learning algorithm, like K-NN, to predict what the missing values will be. In addition to null values, it will also be necessary to remove columns that have a significant amount of null values. For example, if more than 40% of column’s data is null, it may make more sense to remove the column instead of trying to impute the missing values. The imputed missing values are predicted and there is no way to verify that the prediction is accurate so imputation should be avoided if possible. In addition to maintaining the integrity of the data, removing columns typically results in a decrease in variance.

In many machine learning problems, it is usually a good idea to perform exploratory data analysis before cleaning the data. This is because exploratory data analysis allows for one to see which features need imputation. Also, if there are many null values, then the distribution and correlation graphs can be misleading. However, this approach is not perfect because if no exploratory data analysis is done, then it is difficult to know if a variable is worth cleaning. For example, during exploratory data analysis, one could find that unimportant features that will not help with model performance. So, these features will be removed but if no exploratory data analysis is done, this could never be known and time would be wasted on cleaning values for that feature. However, since the datasets have many features, performing exploratory data analysis is impractical. This is why exploratory data analysis and data cleaning will be done together to ensure that unneeded variables are removed while also practically analyzing the important features and their interactions.

After performing data cleaning and exploratory data analysis, the next step will be to reduce the dimensionality of the data. Reducing the dimensionality of the data will allow for faster training performance. This is necessary due to the large amount of features in the datasets. Principal Component Analysis, or PCA, will be used to perform dimensionality reduction. In order to find the optimal number of principal components, a scree plot will be used to see which principal components capture the most variance. After this nunber is found, PCA is performed and the resulting principal components can be used as the new "features" of the dataset. A new dataframe will be created where each column corresponds to a principal component. This new dataframe is what will be used in the clustering model.

The clustering model that will be used is K-means clustering. K-means clusteriung tries to create clusters where the points in the cluster are as similar as possible but while also making sure each cluster is as different as possible. In order to find the optimal number of clusters, the Silhouette Coefficient will be used. The Silhouette Coefficient for each point is, $\frac{b-a}{max(a,b)}$, where a is the distance between the point and its cluster and where b is the distance from the point to the nearest cluster that the point is not a part of. The mean Silhouette Coefficient across all observations can be calculated for different values of k and then plottedl Since the Silhouette Coefficient is a measure of how good the clusters are, the optimal number of clusters will be the poimt im the graph with the highest Silhouette Coefficient, which is just the peak of the graph. The optimal value of k can then be passed in as a parameter of the K-means model. After the model creates the clusters from the data, the clusters can be visualized using a heat map. Since there will likely be more than three principal components, the clusters cannot be visualized in space. From the visualization, an analysis of the influence that each principal component has on a given cluster can be done. Since each principal component is a combination of the features, a graph of the component makeup can be used to decide the most influential features of a principal component. After this analysis is done, the same process can be used to find clusters for the general population of Germany. Finally, the clusters from the CUSTOMERS data can be compared to and analyzed with the clusters found in the general population data to find the customer segments that are most likely to respond to a marketing campaign. In addition, more features can be added based on the PCA analysis of clusters which can improve the accuracy of the supervised model that will be trained.

The final task of this project will be to train and evaluate a supervised learning model that predicts whether or not a person will become a customer. The goal for this model will not just be accuracy, but also interpretability. Interpretability will allow the company to know which features are most influential when it comes to converting people to customers. However, accuracy is stll important because the interpretability of the model only matters if it is correct a significant amount of the time. This is why several different models will be considered. Logistic regressiom with lasso or ridge regularization, decision trees, random forests, boosted trees, XGB, SVM, and naive bayes will all be used and nested cross validation will be used to select the best model. However, it is not necessarily true that the model with the highest accuracy will be selected. A model that is balanced with regards to bias and variance will be selected. After a model is selected, the next task will be hyperparameter tuning to ensure that the model is being trained with the optimal hyperparamaters. Finally, the last step is to evaluate the model using the TEST data.