Machine Learning Engineer Nanodegree

Rick Wuebker
October 30th, 2020

Capstone Proposal

Domain Background

Arvato is an internationally active services company that develops and implements innovative solutions for business customers from around the world **[1]**. They have teamed up Udacity to offer a capstone machine learning project. Their client, a mail-order organic products company, would like to use machine learning to determine which members of the Germany population would be good candidates for their advertising campaign.

I'm particularly interested in this product because there are two different problems to solve. The first is to use unsupervised learning to group their existing customers into different segments and to determine which features of the population would be useful to determine if members of the population would be good candidates for the organic products. The second problem would be to combine those features into a supervised learning model to predict which members of the German population to target.

Problem Statement

Using data provided by Arvato that includes characteristics of current customers and for characteristics of the German population, create a machine learning model to predict which people in Germany would be good candidates to target for their advertising campaign.

The expected strategy is to use unsupervised learning to determine features relevant from the current customers and population to build a supervised learning model that predicts which members of the German population are good targets for a marketing campaign.

Datasets and Inputs

There are four datasets provided:

1. Udacity_AZDIAS_052018.csv; Rows: 891,221; Columns: 366. This dataset gives information on the German population.
2. Udacity_CUSTOMERS_052018.csv; Rows: 191,652; Columns: 369 This datasets gives information on the existing customers of the client.
3. Udacity_MAILOUT_052018_TRAIN.csv; Rows: 42,962; Columns: 367; A training dataset for the supervised model which gives features of past marketing campaign attempts to acquire customers as well as the result.
4. Udacity_MAILOUT_052018_TEST.csv; Rows: 42,833; Columns: 366; A test dataset to accompany (3)

Solution Statement

My solution strategy consists of the following:

1. Data Processing. This will be necessary to deal with missing values in the data. Also oversampling will be used on the target dataset if it is unbalanced.
2. Unsupervised Learning to Segment the population: This step will involve using PCA (Principal Component Analysis) and K-Means Clustering to get a better understanding of what sets their customers apart.
3. Develop a supervised machine learning algorithm and test these strategies:
    a. Logistic Regression
    b. XGBoost Classifier
    c. BalancedBaggingClassifier
    d. BalancedRandomForestClassifier
    e. HistGradientBoostingClassifier

Benchmark Model

The benchmark model for this highly imbalance dataset will be a model where we guess all are negative. This model would have an accuracy of 0.98 and an AUC of 0.50. The linear model should be better than 50/50 but this does not seem to be a linear dataset so it will probably not generalize. We will see if we can use a non-linear model to develop a better model.

Evaluation Metrics

The main metric we will use to evaluate the models will be auc. But since we are also interested in obtaining as many members of the German population as we can, we will also focus on recall, which is to minimize false negatives.

Project Design

The project will progress in the following order:

1. Data Processing: Cleaning of the data for missing values and other modeling issues. I will also be splitting up the test data set into two different sets for validation and testing.
2. Customer Segmentation: This portion will involve using PCA and K-means to learn features of their existing customers to determine what will be the best predictors of which members of the German population will become new customers.
3. Develop the Benchmark Model: Here we will apply the benchmark model to begin, which will be a Logistic Regression model. We will use regularization to keep our model from overfitting and to generalize well and to find the best accuracy and recall.
4. Develop the XGBoost Model: Here we will apply XGBoost to the features and also tune the hyperparameters to find the model with the best accuracy and recall.
5. Develop a Deep Learning Model: We will use different deep learning architectures to try to find the best model to fit the data.
6. Compare the models from (3), (4) and (5) and determine which is the best for prediction and generalization and submit this model to Kaggle **[2]**.

References

[1] Arvato-Bertelsmann [Online]

Available: https://www.bertelsmann.com/divisions/arvato/#st-1 [Assessed October 30th, 2020]

[2] Kaggle [Online]
Available: https://www.kaggle.com/c/udacity-arvato-identify-customers [Assessed October 30th, 2020]