Capstone Project proposal

Boukary

2020-05-08

Machine Learning Engineer Nanodegree

Capstone Project

Boukary Ouedraogo
Mai 8th, 2020

Project Overview

The last decade has been the decade of data, as data storage capacity has grown and the need to process data for added value has become apparent to all organizations. This need of exploiting data has led to the need of for innovative tools to extract knowledge from data . Machine learning is a set of methods at the intersection of several disciplines (mathematics, statistics, computer science...) allowing to extract insights from data to allow companies to increase their profits. In this nanodogree, dedicated to machine learning, we have learned many of these machine learning techniques over the weeks. From predicting the price of housing in Boston area to building a plagiarism detection model to deploying a sentiment analysis model with sagemaker, we have learned and applied a variety of techniques from end to end. It's time to apply all the knowledge we've acquired to solve a real-life project. The project to be discussed here consist of helping Arvato Financial Solutions to get to better know its customers and increase the effectiveness of it mail-ordering in order to attract new customers through client segmentation and a supervised learning model . The

datasets are provided by Arvato Financial Solutions, a subsidiary of the Bertelsman Group

Problem Statement

“How can a client – mail order company selling organic products – acquire new clients in a more efficient way?” This is the problem of this project that we must deal with.

The project will be carried out in four stages.

The first stage will consist of pre-processing and data preparation which is a prerequisite in any machine learning project. In this part we will use imputation methods to process missing data, we will convert qualitative variables into quantitative variables... For example for variables with more than 50% missing values we will combine the imputation of missing values with the creation of indicator variables that show the location of the imputed entries.

The second step is to use an unsupervised learning method to segment the clientele. The purpose of this customer segmentation step is to learn about the company’s customer profiles from data on demographic characteristics. The segmentation will identify the value segments and will be used to construct the target variable that will be used in the supervised learning model.

The third step is to apply supervised learning methods on a new data set to predict the clients that could belong to the value segments we had previously identified.

The last step is to submit the results of the supervised learning model on the kaggle platform to evaluate its performance compare with models built by others.

Data and inputs

In order to carry out this project, Arvato has provided us with 4 datasets:

- `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).

- `DIAS Information Levels - Attributes 2017.xlsx:` a top-level list of attributes and descriptions, organized by informational category

- `DIAS Attributes - Values 2017.xlsx:` a detailed mapping of data values for each feature in alphabetical order

Solution statement

Given the nature of the problem and the data available to us, I have decided to adopt the following resolution plan: For the part concerning unsupervised learning, I will use PCA to calculate the principal components. Then I will retrieve the first most important principal components to cluster the clients using the Kmeans method. As we have more than 300 variables, before doing a Kmeans it is more than necessary to perform a dimension reduction otherwise we will get noisier clustering.

In the supervised learning part I will test several classification models to select the one that gives the best results. Among the models best suited to approach this type of problem we have mainly:

- Logistic regression
- randomforest,
- xgboost classifier,
- H_2O GBM
- Support Vector Machine Classifier (SVC)

The list is not yet exhaustive because this is only a proposal for now and I am open to test other models. I will also tune the hyperparameters of certain models like the Xgboost to retain the best

Benchmark

For this type of problem, the xgboost is known to give better results. XGBoost is an ensamble method that uses many trees to take a decision so it gains power by repeating itself. This method dominates many Kaggle competitions and achieves state-of-the-art results on a variety of datasets. My benchmark model will therefore be an xgboost classifier

Metrics

For the PCA, I will consider the variances explained by each principal component and then I will choose the first n-largest principal components that capture at least 80% of the total variance.

To determine the optimal number of class in the Kmeans, I will use elbow graph.

The expected metric in Kaggle is the area under the ROC curve (AUC). I will therefore use mainly the AUC as the metric. This does not prevent me from looking at other metrics to compare the performance of my models.

Project Design

1. Data cleansing: Raw data requires a thorough data clean-up before building a machine learning model. I will pay particular attention to missing data and outliers. For missing data I will calculate the percentages of missing values for variables and for observations. Observations with more than 50% missing values will be removed and variables with more than 50% missing values will be replaced by dummy variables before being imputed. The remaining variables with missing values will be imputed. Outliers will also be processed.

2. Visualization of the data : Visualization allows a first overview of all the variables in the dataset. It can help to take directions in the

modeling. The seaborn and matplotlib libraries will be used in this step.

3. Feature engineering: PCA implementation, and selection of the principal components that capture at least 80% of the total variance of the dataset.

4. Model selection :

- Selection of the optimal number of classes using the elbow graph method.
- Implementation of the Kmeans algorithm for clustering clients.
- Implementation of the above mentioned models (SVM, random forest, logistic regression, Xgboost) then selection of the most appropriate model to forecast customer acquisition by targeted campaigns.

5. Setting up the model : Find the hyperparameters that best suit our model using GridSearchCV or RandomizedSearchCV.

6. Test and Predict : After building our model on the training dataset we will test its performance on the test dataset using the appropriate metrics especially the AUC.

---

references

1. Udacity machine learning project github repository
2. Udacity machine learning engineer nanodegree courses
3. Scikit-learn website
4. Xgboost course by kaggle
5. Handling missing values by kaggle
6. H_2O.ia website