SKAFOS LLC ML TEST PROJECT

# PROPOSAL

Data set used:

Please find the URL below for the Data Set (Found in Google dataset, Kaggle)

<https://archive.ics.uci.edu/ml/datasets/online+retail>

About the dataset:

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion giftware.

**ML: Algorithms Used to check whether the customer is from UK or not (Binary Classification)**

The spot-check / traditional ML algorithms used for this binary classification are:

LR: Binary Logistic Regression, uses sigmoid function (linear algorithm), Poor performance (scaled version is better)

LDA: Linear algorithm using fisher’s linear discriminant, poor performance

Nearest Neighbors: Non – linear algorithm, ok performance

Decision Tree: Non – linear algorithm, ok performance

Naïve Bayes: Non – linear algorithm, ok performance

SVM: Non – linear algorithm, ok performance

Tradeoffs / Drawbacks:

1. This binary classification might not be useful. Might have picked the wrong target variable.
2. Did under sampling to make balanced data set for training, validation and testing.
3. Data was skewed, so merged all the countries other UK to be 0.

When the model degrades in the production (concept drift), we do have continuously monitor the metrics and retrain, redesign, and redeploy the model. This because the data changes continuously (can use unit tests for data validation) for this we may need incorporate a weighting schema where in more weight is given to recent data than the old one.

Need to build to productionized ML system such that it continuously evaluates and retrain the model.

(Personally, I have not done this).

Deployment options:

Via Rest APIs: Flask (which personally I do) or using Docker containers. Model streaming using Spark or in AWS, using Sage maker similarly for other cloud services such as Azure, google cloud.

Other Notes To increase drive conversion and increase customer management:

1. Customer segmentation should be done using RFM analysis [3] and K means to increase drive conversion by understanding who best customers are, who have the potential to increase revenue and who are likely respond to promotions campaign. (Done only K Means in the code to segment the customers).
2. Data Mining: Associate rule mining using Apriori Algorithm recommend items that are bought together.
3. To reduce the amount of time in training used a package called fuzzywuzzy [1] [2] to calculate similarity of items using Levenshtein distance.

Thanks

Sai Kasyap Kamaraju

References:

1. <https://github.com/seatgeek/fuzzywuzzy>
2. <https://towardsdatascience.com/fuzzywuzzy-find-similar-strings-within-one-column-in-a-pandas-data-frame-99f6c2a0c212>
3. <https://www.kaggle.com/umutboz/rfm-online-retail>