Stripe Merchant clustering and churn predictions

[Zhaox334@umn.edu](mailto:Zhaox334@umn.edu)

Introduction

In this project, a two-year transaction history data from Stripe’s merchant was given to achieve two goals. The first is merchant segmentation and the second is customer churn prediction.

For question 1, I created features for individual merchants and trained two clustering models with unsupervised learning algorithms after deciding the best number of clusters and hyper-parameter tuning. I was able to visualize my clustering result and have an interpretation of individual clusters from the center of the clusters.

For question 2, I first defined churn and created churn labels for the merchants in both empirical and theoretical ways. I then compared, trained and hyper-tuned a classification model with the label and feature from part 1 and other time series features from the original data to predict future churns of individual merchants. I tried to determine the reasons for the churn and provided some insights on how to proceed with the churning merchants.

The report will describe my approach to solving the two questions. Starting with my understanding of the data and feature engineering, followed with my choice of models, hyper-parameters. And lastly some ideas on how to approach churn merchants.

Keywords: customer segmentation, unsupervised learning, churn prediction

1. Pre-assumptions

There are some assumptions I made when I analyzed the dataset and trained the models, which could cause huge difference in data interpretation and model performance. It’s important to state them in the first place.

* If one merchant does not have transaction at the last day of 2034, I won’t automatically take them as stop processing with Stripe and they are not taken as churn. My definition of churn is defined in section 4. This could cause a big difference in terms of model performance. One would have far more positive (churn) labels than another, hence, change the underlying data distribution.
* If merchants churn, they will not be back. Whether with the same id or a different id.
* Different merchant id means different merchant, there will not be two ids associate with same store or merchant.

2. Exploratory Data Analysis and Feature Engineering and Data Preparation

This section introduces my understanding of the dataset and how I create new features for later problems.

2.1 EDA

The data has 1 million rows and 3 columns, and it is a time series of individual merchant’s transaction with 2-year range. Data is clean but needs more features in order to apply machine learning models.

Since we are focusing on clustering and churn prediction instead of predict future sales, my focus will be on merchant level features, which are retrieved from time series data.

2.2 Feature Engineering

The new features come with the following categories.

Transaction volume, transaction total amount, frequency, tenor with Strip and different time of the day (morning, night etc.).

* sales volume, amounts(mean, variance).
* Frequency of the transitions (daily, weekly etc.)
* when did sales happens when(sales per day/week/month, time of the day morning, noon, night.
* start and date on strip, how long they have been as a customer.

Then, there is a divide between features to get per day features.

This is based on my experience working with retail time series. The idea is to use existing time series to come up with content or feature represent to separate merchants from different clusters.

I calculated VIF score to detect multicollinearity. There is only one feature that has about 30 VIF score. Since VIF does not necessarily affect model performance, unless they are perfectly correlated, I decided to keep all of them.

After working with the data more, I came up with more detailed information like below.

* booming business more likely to stay, what’s the metric to identify them? trend? How to identify stable business, unhealthy business.
* instead of using mean, take the most recent average, with weights
* small business to large business in long term, think about long term change underlying model distribution.

To find those information, I used TSFresh module to add additional features. TSFresh uses p value to identify important features from time series. Example would be trend, means, variance and categorical data like has\_duplicate or not. It is massive and end up with 700+ features.

When dealing with supervised learning, the features could be condensed or purified to more important features. In my application here, I tried to use PCA to reduce dimension and integrate later models. The result was not as impressive as I thought, presumably due to the existing features catch good portion of the clustering problem. I can see it works with other harder prediction problems.

3. Clustering and model training

This part talks about

3.0 Metric consideration

AIC, Sihouette Score and Inertia

3.1 Model comparisons

* model wise, can use kmeans or EM, elbow to decide number of clusters

Kmeans vs EM vs something else

3.2 Hyper-parameter turning

Number of clusters K with elbow methods.

* To discuss: K could have multiple correct answers. (growing, stable, declining business; large medium, small, tiny business; subscription business, market place business)
* With more features, clusters actually change with 4 as the best.
* This is where domain knowledge or bias come in.

EM tuning with covariance type

3.25 what does individual cluster represent.

3.3 Sihouette visualization

3.4 Clustering ensemble

Sihouette A large variety of clustering algorithms has been proposed: k-Means, EM (Expectation Maximization), based on spectral graph theory, hierarchical clustering algorithms like Single-Link, Fuzzy c-Means, etc. (see Refs. 47 and 94). However, as it is known, there is no clustering method capable of correctly the underlying structure for all data sets. When we apply a clustering algorithm to a set of objects, it imposes an organization to the data following an internal criterion, the characteristics of the used (dis)similarity function and the dataset. Hence, if we have two di®erent clustering algorithms and we apply them to the same dataset, we can obtain very di®erent results. But, which is the correct one? How can we evaluate the results? In clustering International Journal of Pattern Recognition and Arti¯cial Intelligence Vol. 25, No. 3 (2011) 337372 #.c World Scienti¯c Publishing Company DOI: 10.1142/S0218001411008683 337 analysis, the evaluation of results is associated to the use of cluster validity indexes (CVI),11 which are used to measure the quality of clustering results. Nevertheless, the use of the CVIs is not the denite solution. There is no CVI that impartially evaluates the results of any clustering algorithm. Thus, we can say that di®erent solutions obtained by di®erent clustering algorithms can be equally plausible, if there is no previous knowledge about the best way to evaluate the results. Roughly, we can assure that for any clustering algorithm there is a CVI that will evaluate satisfactorily its results. The idea of combining di®erent clustering results (cluster ensemble or clustering aggregation) emerged as an alternative approach for improving the quality of the results of clustering algorithms. It is based on the success of the combination of supervised classi¯ers

No longer appear if they choose to stop, they will no longer be back.

“*If the merchant stops processing with Stripe, then they would no longer appear.  “*  does this mean once gone, never back?

Question 2

4. Churn clarification/ assumption

Currently churn is assumed as 90%. This will need to be discussed and consulted with domain expert.

This also affects model performance. A less strict threshold will have more positive label and could increase the complexity or change the distribution of the positive data, thus causing false negative or false positive.

No longer appear if they choose to stop, they will no longer be back.

“*If the merchant stops processing with Stripe, then they would no longer appear.  “*  does this mean once gone, never back?

5. Churn Identifier

Empirical vs theoretical

6. Churn prediction

6.1 Model comparison

6.2 Result

6.3 Analysis? Feature importance

6.4 Find the churn and work on it.

Prioritize

Assumption is power user is important or revenue? Maybe both, first is general population, maybe smaller transaction amount, but could be exponentially expanded better for long run.

Incentive for more transactions and customer retention.

Business cost vs life time value.

Summary

Future work

Time series data, After using pca, mention ica, random projection etc.

Clustering ensemble

Reference

1. Christ, Maximilian, Andreas W. Kempa-Liehr, and Michael Feindt. "Distributed and parallel time series feature extraction for industrial big data applications." *arXiv preprint arXiv:1610.07717* (2016).

2. Vega-Pons, Sandro, and José Ruiz-Shulcloper. "A survey of clustering ensemble algorithms." *International Journal of Pattern Recognition and Artificial Intelligence* 25.03 (2011): 337-372.

3. On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled\* SDM’2010 Columbus, OH

4. Clustering Ensemble <https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_ensemble.pdf>

5. ClusterEnsemble Pypi page https://pypi.org/project/ClusterEnsembles/

6. TSFresh https://tsfresh.readthedocs.io/en/latest/