Merchant Clustering and Churn Analysis

David Tan

March 8, 2022

Problem 1: Merchant Segmentation

We want to understand merchants payment activity to try to infer their types of business.

Data:

- 2-year period (2033 - 2034)
- Random sample of future merchants using Stripe
- If the merchant stops processing with Stripe, they would no longer appear.

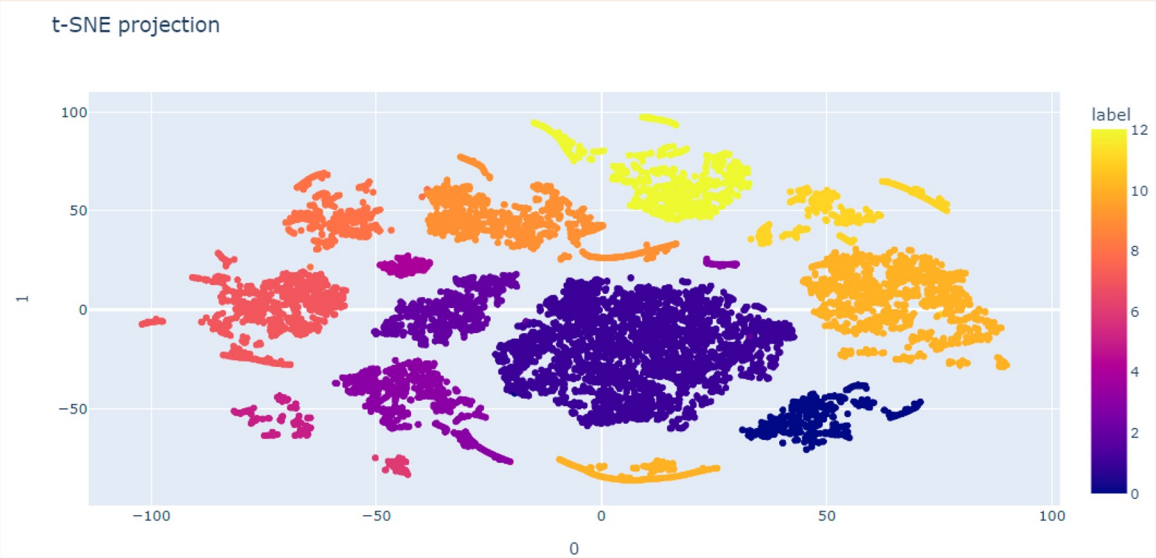
Goal:

- Identify different kinds of businesses in the sample
- Generate assignments for each merchant

Problem 1: Summary & Conclusion

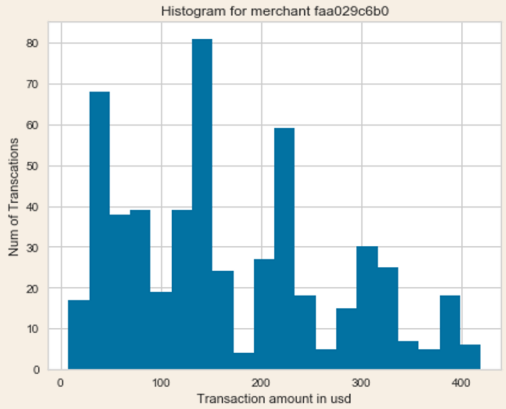
Based on K-means clustering, all merchants are assigned into 13 different groups (labeled from 0 to 12).

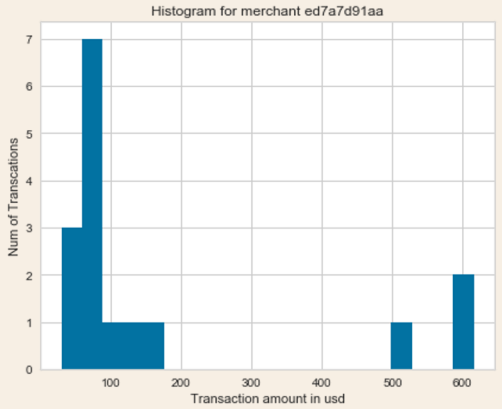
The t-SNE plot is a 2D visualization to visualize our cluster assignment. Merchants seem to be well segmented into 13 groups.

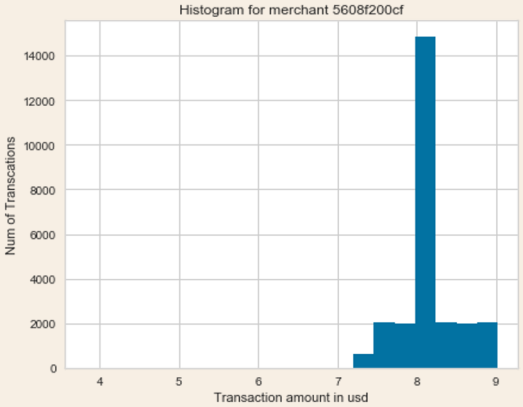


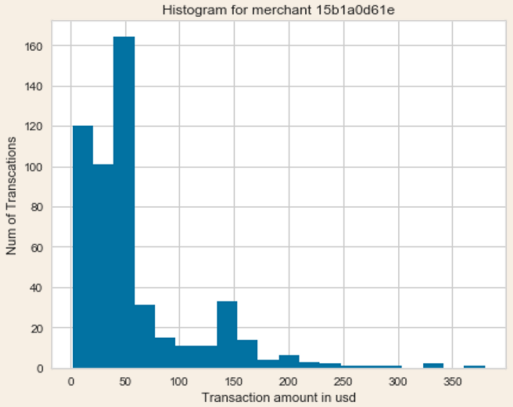
Problem 1: Exploratory Data Analysis

- The raw data has 1,513,719 rows and 3 columns
- No missing values
- Number of merchant: 14,351
- Histograms show that transaction distributions are skewed (not symmetrical).





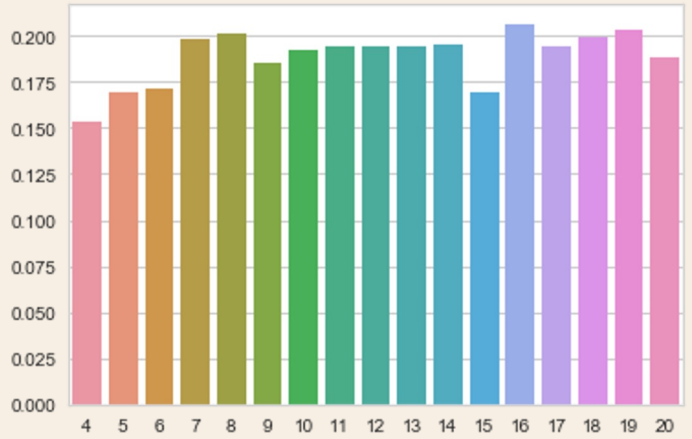




Problem 1: K-Means Clustering

- We choose K-means clustering to partition merchants into k clusters in which each cluster represents a type of business.

- Based on Elbow method and Sihouette scores, we choose k = 16.



Problem 1: K-Means Clustering

- After centroids quality check and K-means re-fit, all merchants were assigned into 13 groups based on their similarity.
- The final results in csv format. Column ‘label’ indicates each merchant’s assignment.

	merchant	daytime	evening	weekday	weekend	restaurant	quarter1	quarter2	quarter3	quarter4	median_spend	median_num_monthly_transactions	iqr	tenure_in_month	label
0	faa029c6b0	0.356618	0.643382	0.863971	0.136029	0.279412	0.180147	0.134191	0.273897	0.411765	145.990	30.0	150.4200	16.754855	1
1	ed7a7d91aa	0.187500	0.812500	0.875000	0.125000	0.375000	0.500000	0.062500	0.000000	0.437500	64.920	1.5	78.1025	12.918453	7
2	5608f200cf	0.287041	0.712959	0.858145	0.141855	0.244238	0.195829	0.250392	0.286728	0.267051	8.200	1318.0	0.0000	21.070087	6
3	15b1a0d61e	0.136276	0.863724	0.980806	0.019194	0.351248	0.201536	0.205374	0.238004	0.355086	44.660	36.0	37.4000	14.068934	1
4	4770051790	0.531359	0.468641	0.858885	0.141115	0.376307	0.000000	0.468641	0.456446	0.074913	288.545	34.0	327.5725	8.679853	1

Problem 2: Churn Prediction

Sometimes a merchant may stop processing with the online payment company, which we call churn. We are interested in identifying and predicting churn.

Goal:

a) come up with a concrete definition for churn

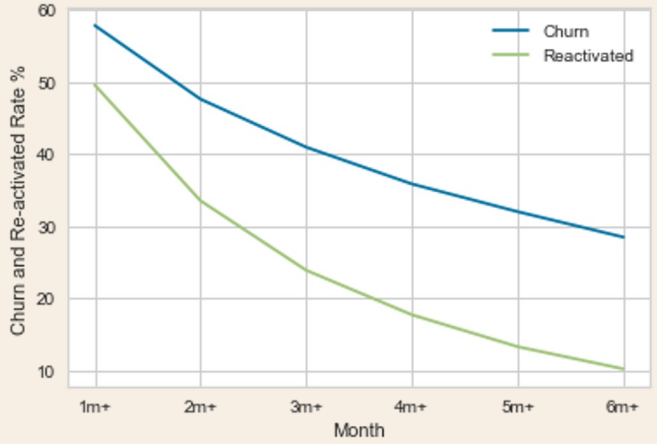
b) identify merchants that have already churned in the dataset

c) build a model to predict which active merchants are most likely to churn in the near future.

Problem 2: Summary & Conclusion

- We define churn as a merchant who has at least 4 months or more no activity after the last transaction.

- We labelled merchants with 4m+ no transaction since the last transaction as churn merchants, others no-churn merchants. Column “churn” is a binary variable indicating churn merchants.



merchant	daytime	evening	weekday	weekend	restaurant	quarter1	quarter2	quarter3	quarter4	tenure_in_month	churn	reactivated
faa029c6b0	0.356618	0.643382	0.863971	0.136029	0.279412	0.180147	0.134191	0.273897	0.411765	16.754855	0.0	0.0
ed7a7d91aa	0.187500	0.812500	0.875000	0.125000	0.375000	0.500000	0.062500	0.000000	0.437500	12.918453	0.0	1.0
5608f200cf	0.287041	0.712959	0.858145	0.141855	0.244238	0.195829	0.250392	0.286728	0.267051	21.070087	0.0	0.0
15b1a0d61e	0.136276	0.863724	0.980806	0.019194	0.351248	0.201536	0.205374	0.238004	0.355086	14.068934	0.0	0.0
4770051790	0.531359	0.468641	0.858885	0.141115	0.376307	0.000000	0.468641	0.456446	0.074913	8.679853	0.0	0.0

Problem 2: Churn Prediction Model

- Our final model is based on LightGBM and it is chosen by the maximum accuracy score and F1 Score among all candidate models.

- With exploratory data analysis, we inherited features developed from feature generation in part one. It was concluded that there were no missing observations in the dataset. However, we've seen outliers from the boxplots.

- In total we developed 8 ML models. Accuracy score, Precision, Recall, F1-score, classification report and confusion matrix were all generated. We also evaluated each model by 5-fold cross validation.

Problem 2: LightGBM Model Results

- The model based on LightGBM hyperparameter optimization became the winning model with the maximum AUC (0.898542). This means our model can predict 89.9% accurate churn cases.
- Features tenure_in_month, median_num_monthly_transactions, quarter 4 and median_spend are most relevant to churn. Our business team need to pay particular attention to those.