March 28th, 2022

Merchant Transaction Analysis

Table of Contents

1. Introduction

The dataset we use for this work includes a random sample of such transactions dated in the future. These transactions could be analyzed from multiple perspectives including but not limited to – transaction amounts, frequency, and the number of transactions. The goal of our analysis is to make use of such transactions aggregated at a merchant level and suggest offers and benefits using gamification to the customers to keep them engaged. Another goal is to predict the risk of merchants discontinuing software use in the future. This could help in targeting those merchants with specific schemes and questions for quantitative and qualitative feedback to help and improve the service.

1.1 Technical Problem Statement

In this section, we convert the business questions to technical problem statements that can be answered with given data.

For question 1,

Problem statement: Using unsupervised learning, segment the merchants in the data into different groups. Each group should have merchants with similar features clustered together.

For question 2,

Problem Statement: Define churn, annotate data and build a binary classification model for predicting the risk of a merchant churning in the near future

For clarity, the analysis for this problem statement can be broken down into a sequence of sub-problems:

- Market research about what and how is churn defined
- Validate or check findings with data
- Tag each merchant using the logic derived in the steps above
- Build a classification model, which takes available parameters and tags the merchants with a high risk of churning in the near future

2. Dataset

The dataset has randomly sampled merchant transaction activity, for merchants that start over a 2 year period (2033-2034). Each observation is a transaction amount in cents.

Topic	Value / Range
Total number of transactions	1513719

Number of unique merchants	14351
Transaction amount range	[201, 25920280]
Timeline of the data	1/1/33 – 12/31/34

Table 2.1: Dataset details

Figure 2.1, shows that the transaction amounts are fairly normally distributed.

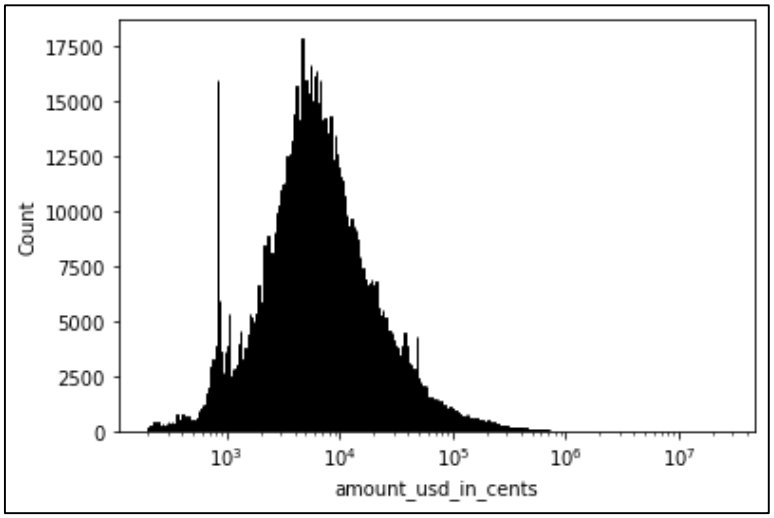

Figure 2.1: Transaction amount distribution

2.1 Dataset Limitations

The limitations of the data include, the data is rich in a number of data points but the features/columns are limited, no demographic information, purpose or type of transactions are given. The data is said to be of just 2 years, extrapolation beyond that could be deemed inappropriate. Some of the merchants are underrepresented in the dataset having only 1 transaction whereas the maximum could be upto 25512. All of them are equally weighted at this point.

2.2 Dataset Assumptions

We have assumed the following points for the data for our analyses:

1. The data is demographically diverse and does not depend on transactions in a specific region around the world.
2. All of these transactions are independent. If transactions happened between two merchants in the dataset, it is just associated with one of them, and the dataset has no duplicate entries.
3. The data has no outliers/fraudulent transactions since all the transaction amounts are within the current transaction limits.

4. The date 12-31-2034 is considered the end of time for analysis. This impacts tagging the churn prediction target variable for the current dataset
5. The value for money is considered for the years 2033 and 2034. The transaction amounts are inflation-adjusted for aiding comparison

3. Statistical Methods/Analysis

This section speaks about the statistical methods that we have used to answer the questions stated in section 1.1 and 1.2. To state them, we use K-means clustering to segment the merchants and Random Forest Classifier for churn prediction. The coding application used was Jupyter Notebook and the programming language was Python.

3.1 Question 1

The first question is as follows:

Using given data, how would you identify different kinds of businesses in the sample?

We address this question by aggregating the data at the merchant level and deriving 3 new fields. We then run the elbow method to figure out the optimal number of clusters and run K-means to segment the merchants.

3.1.1 Data Cleaning/Aggregation

Since the clustering is supposed to be performed at the merchant level, we aggregate the data at the merchant level. We derive a field named “total_amount” which adds all the transaction amounts associated with the merchant. We calculate the “number_of_transactions” and also calculate the average difference of days between two transactions for that merchant. The field is named “average_frequency”.

3.1.2 Dataset details and Descriptive Statistics

Following are the variables in the dataset used for analysis

- **Variables:** Total Amount, Average Frequency, Number of transactions
- **Derived fields**: All the variables above are derived fields

For analysis 1, table 3.1 shows the mean, and variance of all the features.

Group Name	Available Data	Sample Mean	Sample Variance
Total Amount	14351	1633295.61	41364215038244

Average Frequency	14351	19.1128	1544.8069
Number of transactions	14351	105.478	278466.846

Table 3.1: Descriptive Statistics for Analysis 1

3.1.3 Statistical method

K-means clustering is a method that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster[2].

The elbow method is used to determine the number of cluster segments, refer figure 3.1. Since the relationship between distortion and the number of clusters does not look significant after 4, we select 4 as the optimal number of clusters.

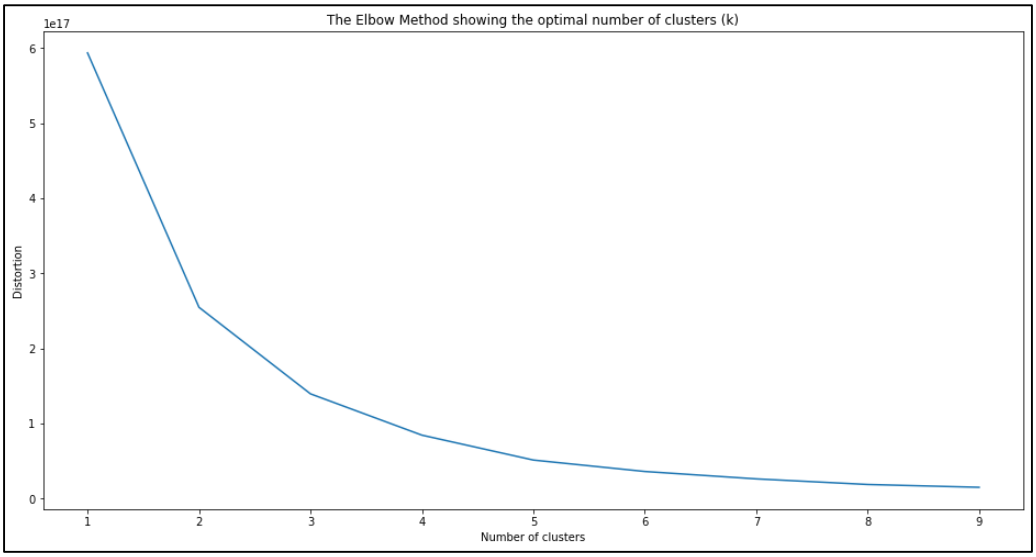

Figure 3.1: Elbow method output

3.1.4 Algorithm Assumptions

- Spherical Clusters: Kmeans assumes spherical shapes of clusters (with a radius equal to the distance between the centroid and the furthest data point)

3.2 Question 2

Please a) come up with a concrete definition for churn b) identify merchants that have already churned in the dataset, and c) build a model to predict which active merchants are most likely to churn in the near future.

We address this question by doing market research about churn, checking if the data shows similar findings, tagging each merchant with churn based on derived logic and then building a random forest binary classification model for predicting the risk of merchant churning or otherwise.

Churn Definition: A merchant is said to have churned if there have been no transactions from them in the last 90 days (here 10/3/2034-12/31/2024)

3.2.1 Deriving Churn definition

We used two methods to derive the churn definition,

1. Using a 95% confidence interval of the given data:

Using the average frequency below, we wanted to figure out the distribution of data for all merchants. 95% of values lay in 0 to 56.65. On the conservative side, we decide to round it to 60. We later ruled this out due to reliability issues without evidence.

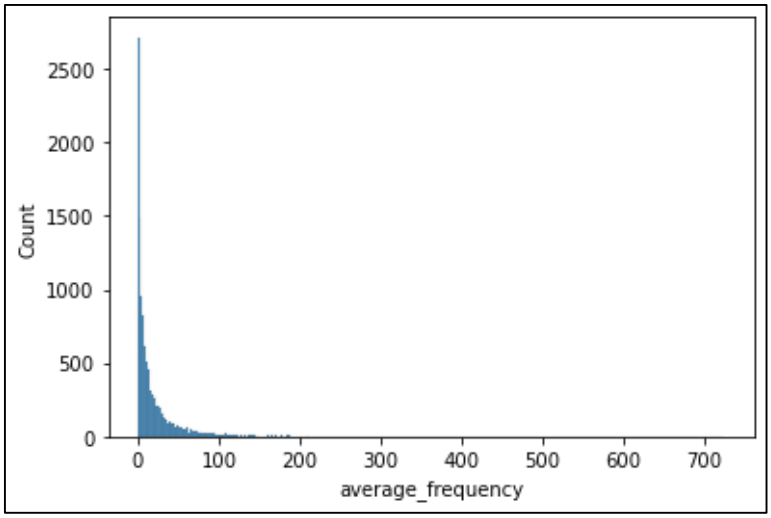

Figure 3.2: Average Frequency Distribution (Right-tail of a normal distribution roughly)

2. Market convention:

We finally used one of the market conventional values for churn (3 months). So a merchant who hasn't transacted in the last 90 days (10/03/2034-12/31/2034) was declared churned and others were not.

3.2.2 Data Cleaning and Aggregation

We use the original dataset and data from section 3.1 to get the last transaction date of each merchant and their last transaction amount. We use the former field to see if it fits in our churn definition and mark it as churned or otherwise. We then drop the field last_transaction_date from the data to avoid overfitting.

3.2.3 Dataset Details and Descriptive statistics

- **Variables:** Total Amount, Average Frequency, Number of transactions, last_transaction_amount, churn
- **Derived fields**: Total Amount, Average Frequency, Number of transactions, churn

For analysis 2, table 3.2 shows the size, mean, and variance of both the sample groups.

Group Name	Available Data	Sample Mean	Sample Variance
Total Amount	14351	1633295.61	41364215038244
Average Frequency	14351	19.1128	1544.8069
Number of transactions	14351	105.478	278466.846
Last Transaction Amount	14351	42410.855	71607469424.665

Table 3.2: Descriptive Statistics for Analysis 2

Churn Prevalence: 40.986%

We plotted scatterplots to see the correlation between variables, number_of_transactions and total_amount did show some positive correlation but not a strong one.

3.2.4 Statistical method

We build a random forest classifier(ensemble of decision trees) to predict the churn. We figured out the hyperparameters using a hyperparameter grid search running over 3-fold cross-validation. The benefit of doing that is we know if the model is not succeeding by chance on our splits.

We used random forests because they are developed on an underlying concept of bagging which provides the best ensemble (combination) of trees for prediction. Their chances of overfitting are comparatively less than decision trees. Parametric models that assume normal distributions don't handle outliers well, and the majority of popular statistical models assume normality. Random forests work in those scenarios as well. They are explainable and handle continuous and categorical variables well.

3.2.5 Model Assumptions

No formal assumptions

4. Results and Interpretation

This section details the results of each test, primarily focusing on their metrics, graphical outputs, and their final interpretation. This section also states how these tests could be used to make decisions.

4.1 Analysis 1: Merchant Cluster Assignment

Graph 4.1 shows the output from k-means. Table 4.1 shows the counts in each of the clusters:

Cluster ID	Number of merchants
0	13711
1	81
2	4
3	555

Table 4.1: Merchant distribution in clusters

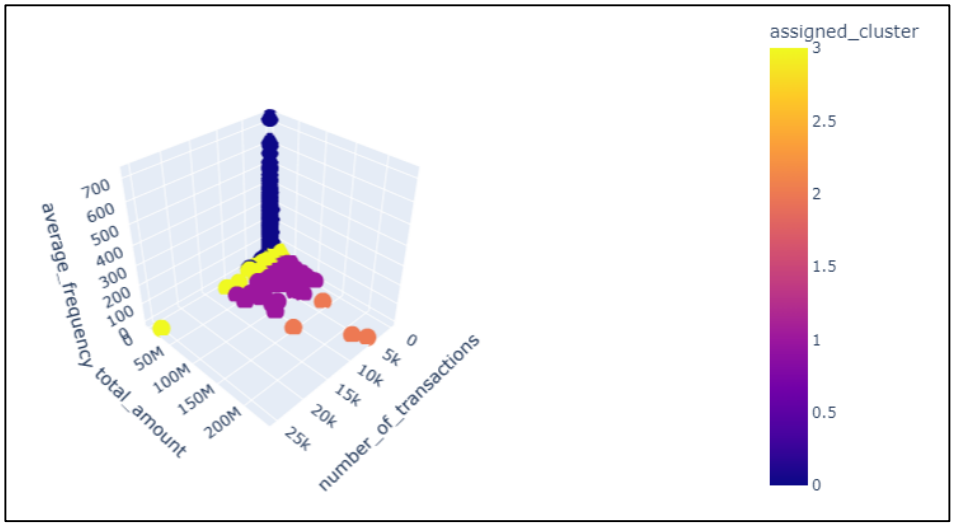

Figure 4.1: K-means output in 3-D space

The general insights from figure 4.1 include, the higher average frequency merchants are segmented in cluster 0. The ones with a higher total amount are in cluster 2. The ones with low average frequency and total amount irrespective of transactions are in cluster 3. The ones with all 3 features in the lower range are classified in 1.

Hence cluster 1 can be targeted with more schemes and offers to increase their engagement and value. The ones in cluster 0 can be targeted with offers for increasing their transaction frequency. Cluster 3 can be targeted for an increase in the amount and frequency offers. Similarly, cluster 2 can further be targeted for increasing the frequency of transactions.

4.2 Analysis 2: Churn Prediction

We get a recall sensitivity of predicting churn as 76% and a specificity of 64%. This means that out of 100 merchants that churn, our model is capable of picking 76 correctly. On the other hand, when 100 do not churn, our model calls 64 of them correctly as not churning but incorrectly calls 36 of them as churned.

Feature Name	Feature Importance (%)
Number of Transactions	33.79
Total amount	33.15
Average Frequency	27.53
Last transaction amount	5.53

Table 4.2: Feature importance in decision making of predicting churn

5. Conclusion

In summary, we use unsupervised machine learning to cluster the merchants in separate groups. These groups could be individually targeted as per the common features to help retain them. In analysis 2, we predict the risk of merchant churning using supervised machine learning. With more accurate labels and rich data features, this model can be further improved.

6. References

[1] https://en.wikipedia.org/wiki/K-means_clustering
[2] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

7. Appendices

This appendix was not used and was intentionally left blank.