Merchant Segmentation and Churn Analysis

Rustin Shamloo

Problem Statements

Using dummy future data, how can we identify different kinds of businesses in the sample? Generate assignments for each merchant.

We want to identify and predict churn.
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.

- **Given:** Limited info on merchants
- **Deliverable:** Segment customers for deeper analysis
- **Process:**
  - Merchants are unlabeled
    - --> Unsupervised learning
      - Clustering is best approach here
  - Normalize data
  - Dummy encode any categorical into numerical type
  - Elbow method to find optimal # of clusters
  - K-means clustering
  - Analyze resulting clusters
  - Generate hypotheses about clusters
  - EDA to dive deeper and test hypotheses
  - Connect merchant findings to business recommendations

- **Given:** Churn = customers leaving, want to reduce churn
- **Deliverable:** Identify who has churned, predict who will churn
- **Process:**
  - Clearly define what churn is
  - Define time period and assumptions
  - Normalize data
  - Dummy encode categorical into numerical type
  - Generate a method for finding who churned month-to-month
  - Run classification model
  - Interpret accuracy
  - Dive deeper to determine important features
 - Connect churn findings to business recommendations

Customer Segmentation

Overview

Distribution:

- Transaction amounts ranged from **$2.01 - $260,000+**
- **1,513,719 transactions** made at **14,351 different merchants** during this 2-year period
  - **Less than one transaction per merchant per day**, on average
    -> how can we increase this metric?
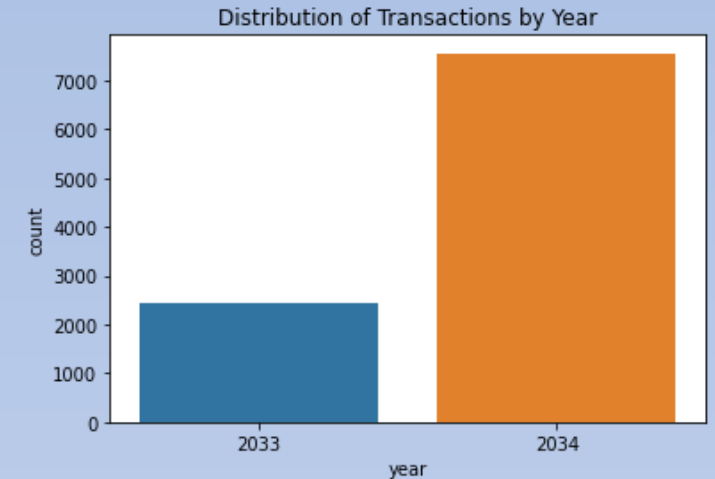- Few outliers, mostly transactions of less than **$100,000**

Exploratory Data Analysis
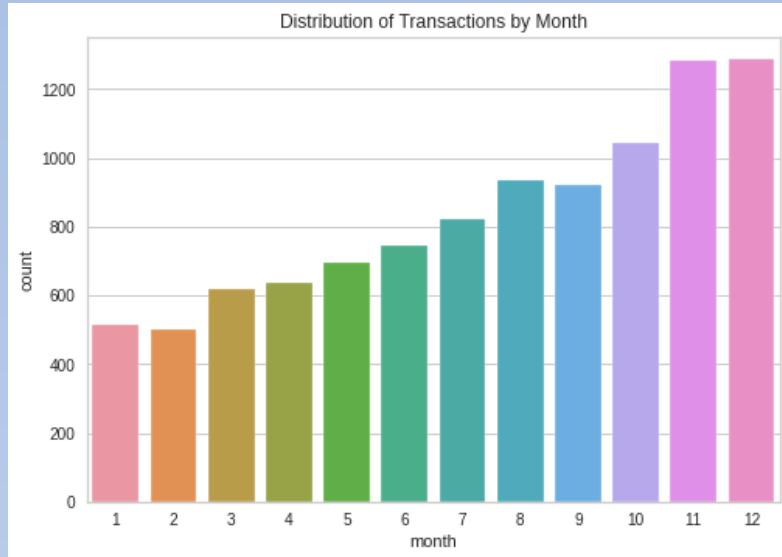
- More transactions in later months
  - Bias from large number of new merchants acquired in late 2033/2034?
  - Holidays?
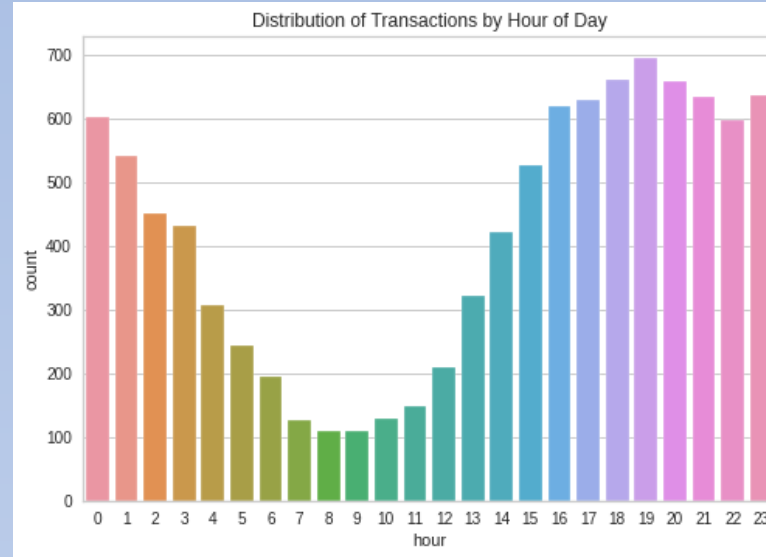
- Transactions ramp up during work hours, as expected
- Also more transactions after-hours
  - International merchants?
 - Restaurants, clubs, **e-commerce**, etc.?

- Business ramped up in 2034
  - **3x** the number of transactions, velocity ↑
- **Check customer churn, retention, and acquisition to investigate deeper**

Model Preparation

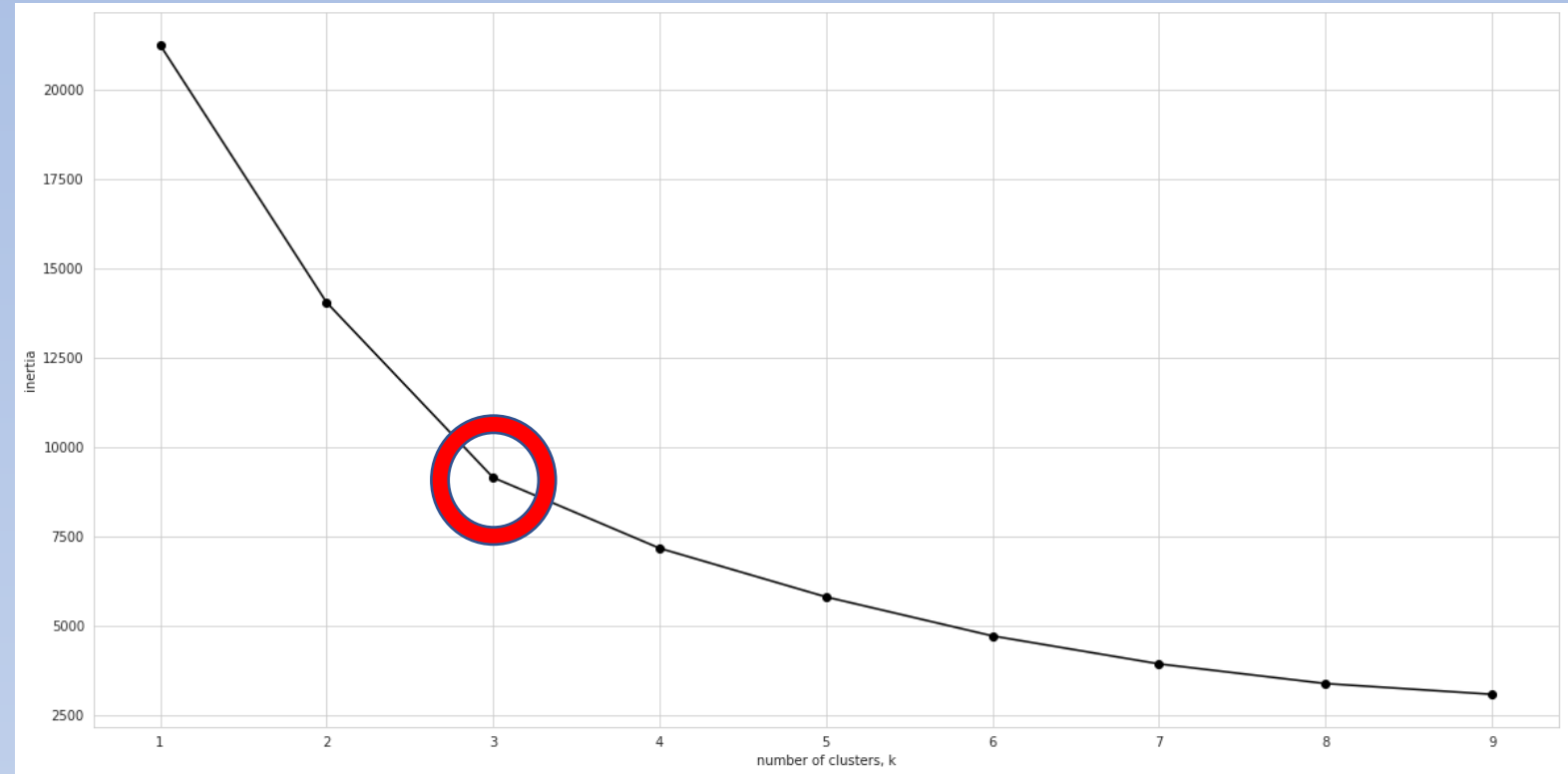| merchant	amount_dollars	year	month	day	hour
faa029c6b0	63.49	2034	6	17	23
ed7a7d91aa	38.54	2034	12	27	0
5608f200cf	7.89	2034	4	30	1
15b1a0d61e	44.52	2034	9	16	1
4770051790	202.03	2034	7	22	16
...	...	...	...	...	...
aaa631d7ca	105.18	2034	11	26	10
47b468a8c4	52.53	2034	7	31	2
e4c027a857	140.53	2034	4	23	17
2c328165ea	5.29	2034	10	26	2
6044a863a1	83.92	2034	11	2	3

- Split 'time' column to look across various granularities
  - But added more dimensionality
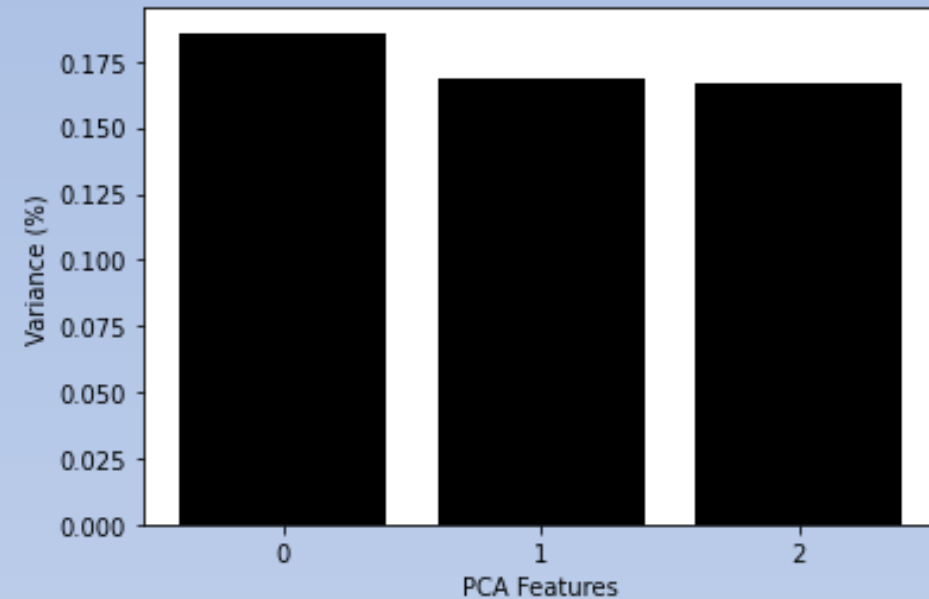
- Normalized numerical data before performing Principal Component Analysis (PCA)
- PCA to reduce dimensionality
- First 3 components explain roughly **50%** of variance

Model

- Merchants are not labeled or grouped
- Need to perform unsupervised learning
    - -> **K-means clustering**
    - Cluster merchants and investigate discerning attributes across clusters using their averages
- Used 'elbow method' to find optimal number of clusters
    - Segment merchants into **three** clusters
- Converted merchant names into numerical data for the model
- Mapped cluster labels back to data frame
    - Each merchant has an associated cluster now and is labeled

Model Results

| cluster	amount_dollars	year	month	day	hour
0	117.382975	2033.952357	6.740750	20.164470	18.144450
1	217.184228	2033.923565	6.702913	10.268532	7.696279
2	107.031430	2033.229610	9.861229	15.702744	13.208736

Cluster 0:
- Merchants process most transactions after regular business hours
- Relatively similar transaction sizes to cluster 2 merchants

Cluster 1:
- Merchants' average transactions are ~ **2x** larger
- Merchants process most transactions in the morning
 - see distribution of rest of hours

Cluster 2:
- Merchants process most transactions in the early afternoon
- Merchants process the smallest transactions, mostly in 2033
 - Have these merchants churned already in 2034?
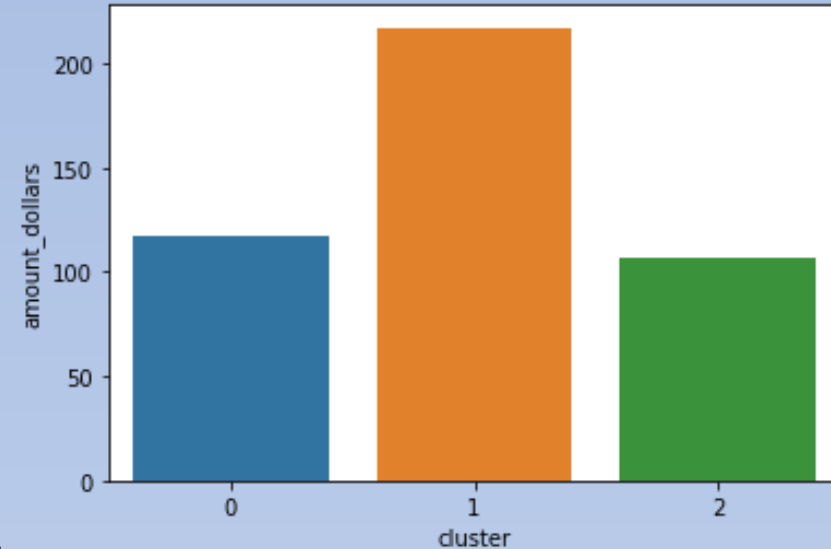
- Cluster 0 and 1 merchants processed many more transactions in 2034
- Are these merchants newer than those in cluster 2, or just doing better?
  - Cluster 0 is newer, cluster 1 is doing better

Cluster Analysis

Distribution of Transaction Size by Cluster

Transaction Sizes by Hour, Colored by Cluster
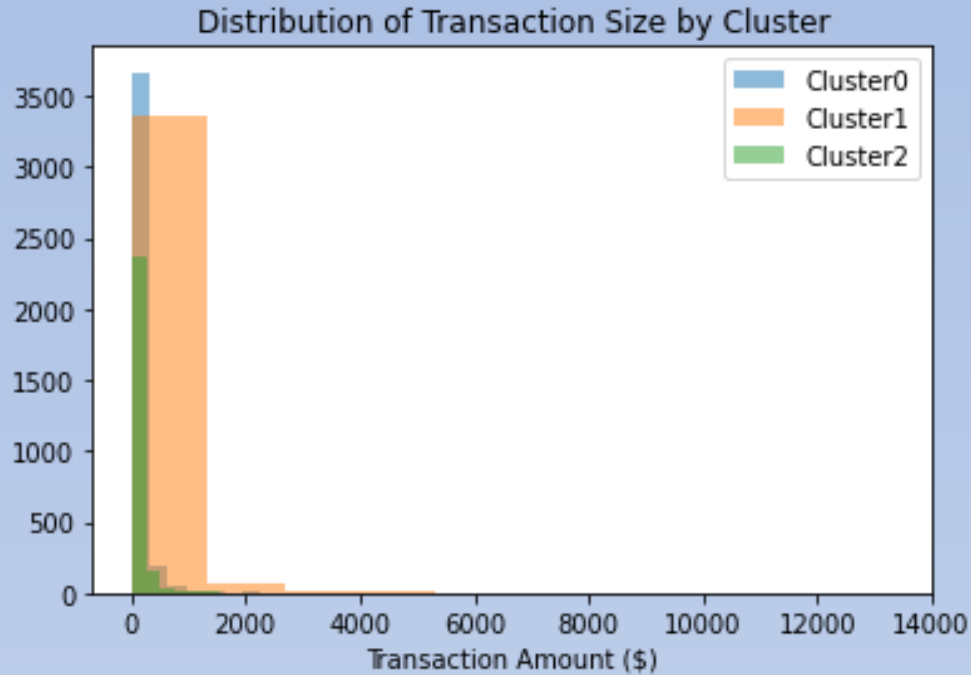
- Cluster 1 merchants have a large spread of transaction sizes, can process around the clock
  - Split amounts into buckets to see how much of each size?
- Cluster 2 merchants mainly process smaller transactions
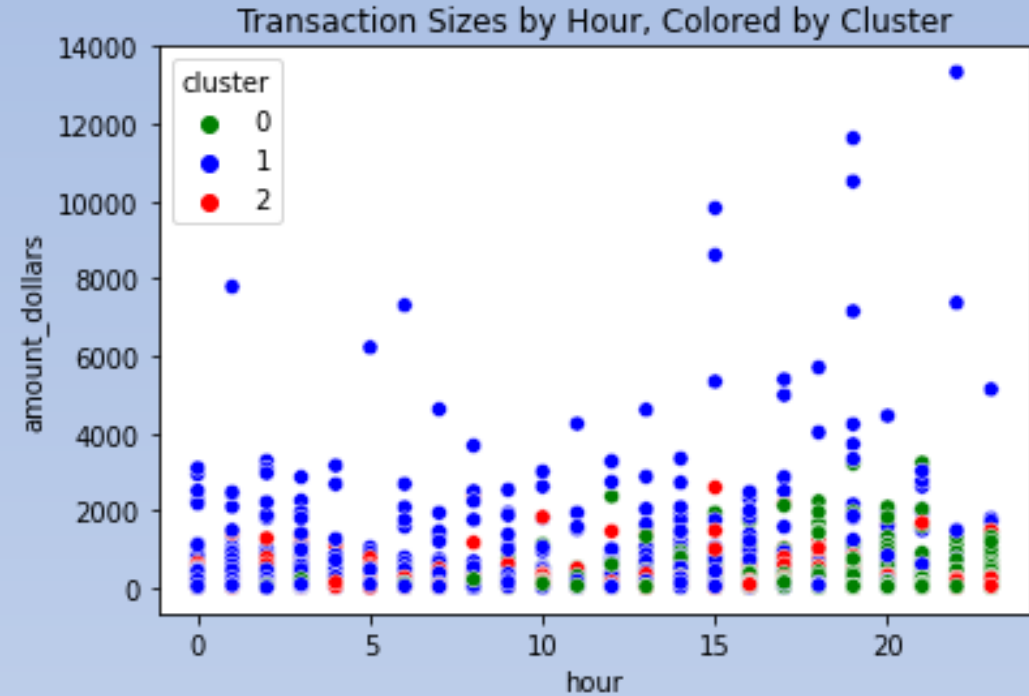
- Cluster 0 merchants process after-hours too, but much smaller transactions

- Cluster 1 merchants process the largest amount of big transactions

Recommendations

Cluster 0: Newer Merchants

- These are merchants who started in late 2033 / early 2034
- Accept transactions around the clock
- High volume merchants, although deal with relatively smaller transactions
- **They are viewed as a promising opportunity to grow** with them as they scale to handle larger transactions

Recommendation:
- Offer on-boarding to get them up to speed
- Good support to keep them feeling valued

Cluster 1: Big Merchants

- Very profitable merchants
- Big Merchants process the largest and the most amounts of transactions, around the clock
- Big Merchants also process, therefore sell, a wide range of items
- **Focus on retaining these merchants as best as possible**

Recommendation:
- Provide competitive exclusive crypto rewards programs
- Ensure customer support and experience is top-notch

Cluster 2: Unprofitable Merchants

- These are smaller, older merchants who process small transactions and are at high risk of churning
  - Some already did by 2034
- **If start-up, re-engage for potential future opportunities**

Recommendation:
- A/B tests with re-engagement, come back based on new features/promotion?
- Surveys about experience
- See which product they were using

Churn Analysis

What is Churn?

- Churn is essentially **the loss of a customer**
  - E.g., given merchant was active from start until March 2033, but not in April 2033 or beyond
- Up to **5x** more expensive to acquire new customers than to retain customers (Forbes)
- Stripe wants to keep churn below a certain threshold
  - **Keeping churn low increases customer retention and keeps revenue streams healthy**

- Churn rate = # of merchants who churned / # of active merchants at beginning of period
- Retention rate = 100% - churn rate

- Analyzing Subset: Q4 2034 (October – December 2034)
  - Representative of busiest months, with most recent merchant data
  - Assume findings from subset represents those from entire dataset

By The Numbers

Churn:

- **Churn Rate = 90%**
- High churn -> not always bad, but investigate
- **Is it 'positive' churn?**
  - Bad fit, unprofitable merchants, etc.
- 325 merchants churned between Oct and Nov 2034
- 350 merchants churned between Nov and Dec 2034

Acquisition:

- High churn but high acquisition too -> **leaky bucket effect**
- **Acquisition Rate = 99.3%**
  - **Net + 136 new merchants** in Q4
    - 749 -> 885 active merchants

merchant	churn
53b3fbeae2	0
524efc3692	1
0646c31f12	1
53b3fbeae2	0
e5393f7993	1
...	...
207cc1c4a7	0
eb772f6014	0
d087d4c321	0
a26684af60	1
062de5273b	0

Data frame showing each merchant has an associated churn label
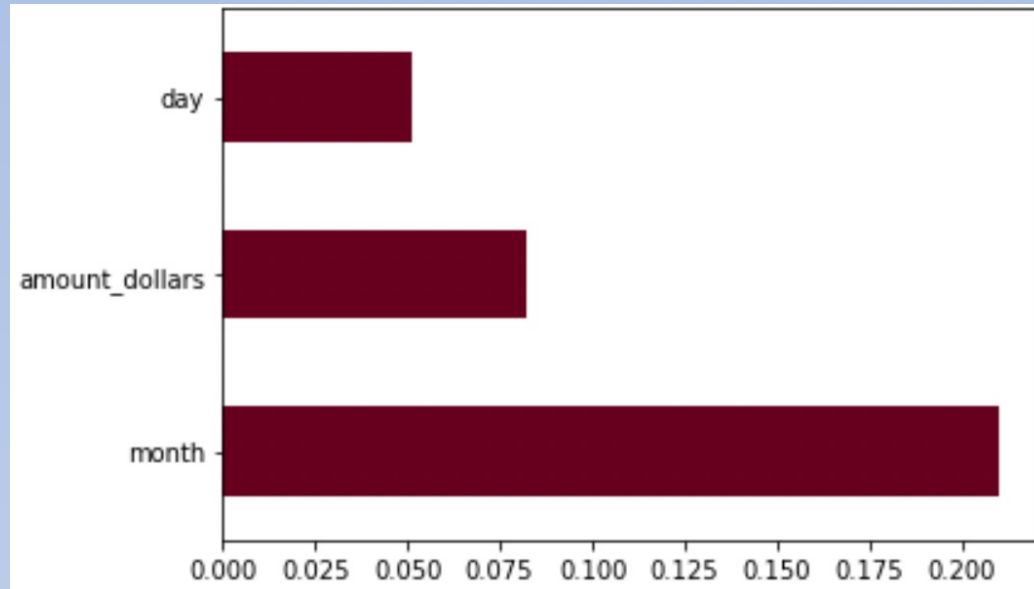
Churn Prediction Model

Predicted Result	0	1
Actual Result		
0	267	31
1	26	166

	precision	recall	f1-score	support
0	0.92	0.89	0.91	301
1	0.85	0.89	0.87	207
accuracy			0.89	508
macro avg	0.89	0.89	0.89	508
weighted avg	0.89	0.89	0.89	508

0.8917322834645669

- To predict whether a merchant will churn or not is a binary classification problem
  - Random Forest Classifier
  - Need to convert merchant to a numerical data type for model

Results:

- **89% accuracy**
  - High F_1 and recall while still maintaining high precision

What Impacts Churn?

- We see the importance (~ **20%**) that **month of transaction** demonstrates in predicting churn
- This model could be used to help predict customer churn based on different transaction attributes such as the month and transaction amount
  - May not be easy to implement, maintain, or interpret on a large scale, though…
    - **Model drift with more merchants scaling up?**

Recommendations

- **Follow-up surveys to discover if dissatisfaction is why churn is high**
  - Competitors, missing expectations, too expensive, sub-par customer service, etc.?
- Which clusters do churned merchants belong to?
  - **Cascade with customer segmentation analysis**
  - How many were profitable cluster 1 merchants?
- **Perform a conjoint study to find merchants' feature preferences**
- **Anomaly detection when churn is too high**
  - Live track and alert when a certain churn threshold has been reached in a specified timeframe

For retaining customers:
- Create marketing campaigns to upsell those currently using services

For expanding merchant base:
- Increase marketing budget
- Provide a cheaper subscription method to lower monthly costs for smaller merchants to be more inclusive
- A/B test with different re-engagement methods
- Adjust cost / profit margins potentially, if a big enough issue, to stay competitive

Thank You!

Questions