BAX 423 Big Data

What Happens in Vegas, Stays in Venmo

Text Analytics, Social Network Analytics, Predictive Analytics

Team Super Duper
Yuan Yuan Ge
Yining Hang
Yijun Huang
Joey Li
Yiying Wang

Table of Contents

Executive Summary

Venmo is one of the most popular peer-to-peer (P2P) mobile payment apps in U.S. It is popular for its social flavor that users are required to make transaction with a message describing what the transaction is about. In this report, we conducted text analytics, social network analytics, and predictive analytics to extract insights from Venmo transaction data. For text analytics, we extracted emojis and words from descriptions, then we matched them with corresponding topics using dictionaries. We found that different users have different spending behavior, which usually stabilized in about a month. For social network analytics, we counted the user's friends and friends of friends and calculated clustering coefficient and PageRank. For predictive analytics, we utilized the RFM model and analytics above to run regression model to predict the number of future transactions of users. We found that the social network metrics combined with the spending profile have the highest predictive power. The user with higher PageRank and more friends will have more transactions, which highlighted the unique feature of Venmo, sharing when transacting. As a result, we suggested Venmo to emphasize its social flavor and to encourage its user to expand their social networks.

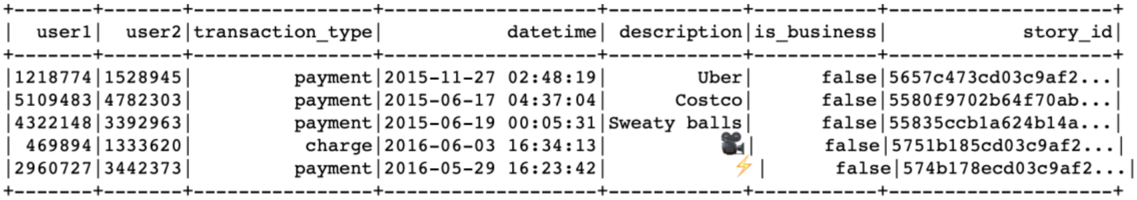
Introduction

Venmo is a peer-to-peer (P2P) mobile payment app owned by PayPal. It allows its users to exchange money with a click of a button. In the fourth quarter of 2019, Venmo has about 30 billion U.S. dollars volume with more than 40 million active accounts. What made Venmo so popular is that users are required to make their transaction with a message describing what the transaction is, which transforms transactions into sharing experiences. In this report, we will analyze the transaction data in three aspects. First, we applied text analytics to establish users' profiles. Then, we applied social network analytics to investigate users' networks. Finally, we used regression to predict the number of future transactions.

Data Characteristics

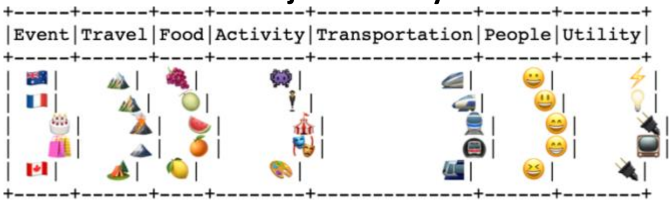
The transaction dataset has millions of transaction records from thousands of users. It includes the following features, user1, user2, datetime, transaction_type, description, is_business, and story_id. For text analytics, the focus will be description. We transformed descriptions into topics with emoji and word dictionary to better understand what the transaction is about. It is worth mentioning that we added some additional words to the default dictionary. For example, we added 'weee', a popular online food delivery service, into 'Food'. For social network analytics, the focus will be user1 and user2, we will use the combination of user1 and user2 to identify the user's friends as well as friends of friends and calculate social network metrics. For predictive analytics, we will use the RFM model as well as text and social network analytics to develop regression model to predict the number of future transactions.

Transaction dataset



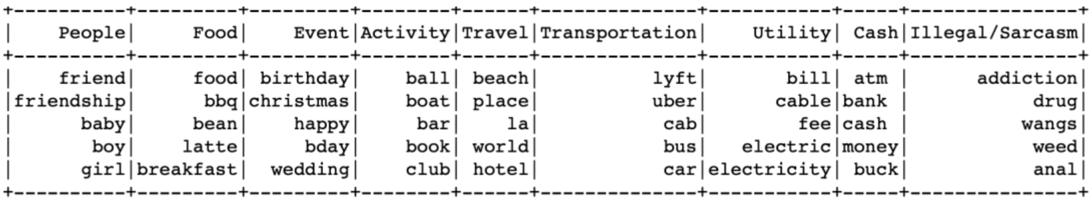
```
+-------+-------+----------------+-------------------+------------+-----------+--------------------+
|  user1|  user2|transaction_type|           datetime| description|is_business|            story_id|
+-------+-------+----------------+-------------------+------------+-----------+--------------------+
|1218774|1528945|         payment|2015-11-27 02:48:19|        Uber|      false|5657c473cd03c9af2...|
|5109483|4782303|         payment|2015-06-17 04:37:04|      Costco|      false|5580f9702b64f70ab...|
|4322148|3392963|         payment|2015-06-19 00:05:31|Sweaty balls|      false|55835ccb1a624b14a...|
| 469894|1333620|          charge|2016-06-03 16:34:13|          🎥|      false|5751b185cd03c9af2...|
|2960727|3442373|         payment|2016-05-29 16:23:42|          ⚡ |      false|574b178ecd03c9af2...|
+-------+-------+----------------+-------------------+------------+-----------+--------------------+
```

Emoji dictionary



```
+-----+------+----+--------+--------------+------+-------+
|Event|Travel|Food|Activity|Transportation|People|Utility|
+-----+------+----+--------+--------------+------+-------+
| 🇦🇺|    🏔|  🍇|      👾|            🚄|    😀|      ⚡|
| 🇫🇷|     ⛰|   🍈|       🕴 |            🚅|     😃|      💡|
|   🎂|    🌋|   🍉|       🎪|             🚆|     😄|      🔌|
|   🛍|    🗻|   🍊|       🎭|             🚇|     😁|      📺|
| 🇨🇦|    🏕|  🍋|      🎨|            🚈|    😆|     🔌|
+-----+------+----+--------+--------------+------+-------+
```

Word Dictionary



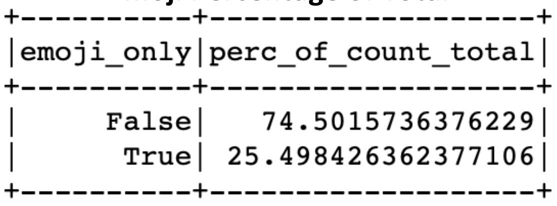
```
+----------+---------+---------+--------+------+--------------+-----------+-----+---------------+
|    People|     Food|    Event|Activity|Travel|Transportation|    Utility| Cash|Illegal/Sarcasm|
+----------+---------+---------+--------+------+--------------+-----------+-----+---------------+
|    friend|     food| birthday|    ball| beach|          lyft|       bill| atm |      addiction|
|friendship|      bbq|christmas|    boat| place|          uber|      cable|bank |           drug|
|      baby|     bean|    happy|     bar|    la|           cab|        fee|cash |          wangs|
|       boy|    latte|     bday|    book| world|           bus|   electric|money|           weed|
|      girl|breakfast|  wedding|    club| hotel|           car|electricity| buck|           anal|
+----------+---------+---------+--------+------+--------------+-----------+-----+---------------+
```

Text Analytics

The focus for text analytics is the description. As we mentioned, the description is a message shared by the user describing what the transaction is about. It is usually emoji, text, or a combination of both.

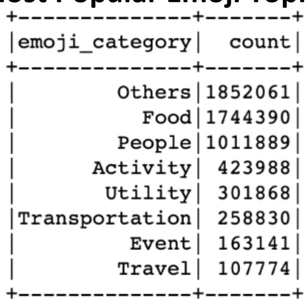
First, we extracted emojis and words from descriptions, then we matched them with corresponding topics using dictionaries. Sometimes, there are multiple emojis and words in one description, so we assigned a score of 0.7 for each word topics and a score of 0.3 for each emoji topics, and then we added up all the scores and chose the topic according to the highest score. Here are some interesting findings. Users like sending emojis when sharing experiences, and there are more than 25% of descriptions are emoji only. It seems that users are having a good time when sending emojis. The most popular emoji topics are Food, People, and Activity, and the most popular emojis are pizza 🍕, cheers 🍻, flying money💸, wine 🍷, and party popper 🎉.

Emoji Percentage of Total



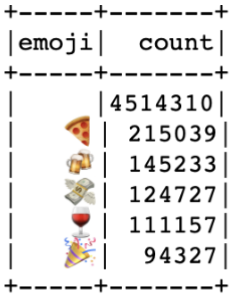
```
+----------+------------------+
|emoji_only|perc_of_count_total|
+----------+------------------+
|     False|  74.5015736376229|
|      True|25.498426362377106|
+----------+------------------+
```

Most Popular Emoji Topics



```
+--------------+-------+
|emoji_category|  count|
+--------------+-------+
|        Others|1852061|
|          Food|1744390|
|        People|1011889|
|      Activity| 423988|
|       Utility| 301868|
|Transportation| 258830|
|         Event| 163141|
|        Travel| 107774|
+--------------+-------+
```

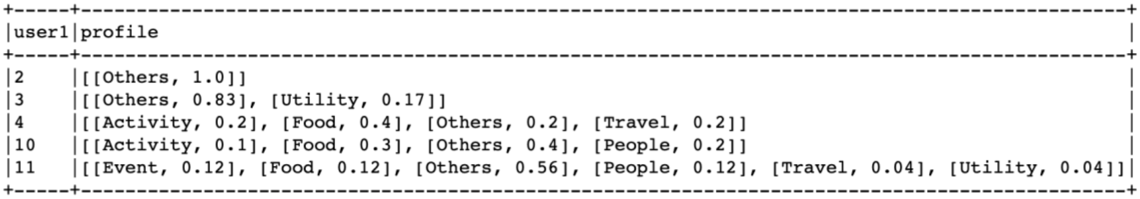
Most Popular Emojis



```
+-----+-------+
|emoji|  count|
+-----+-------+
|     |4514310|
|   🍕| 215039|
|   🍻| 145233|
|   💸| 124727|
|   🍷| 111157|
|   🎉|  94327|
+-----+-------+
```

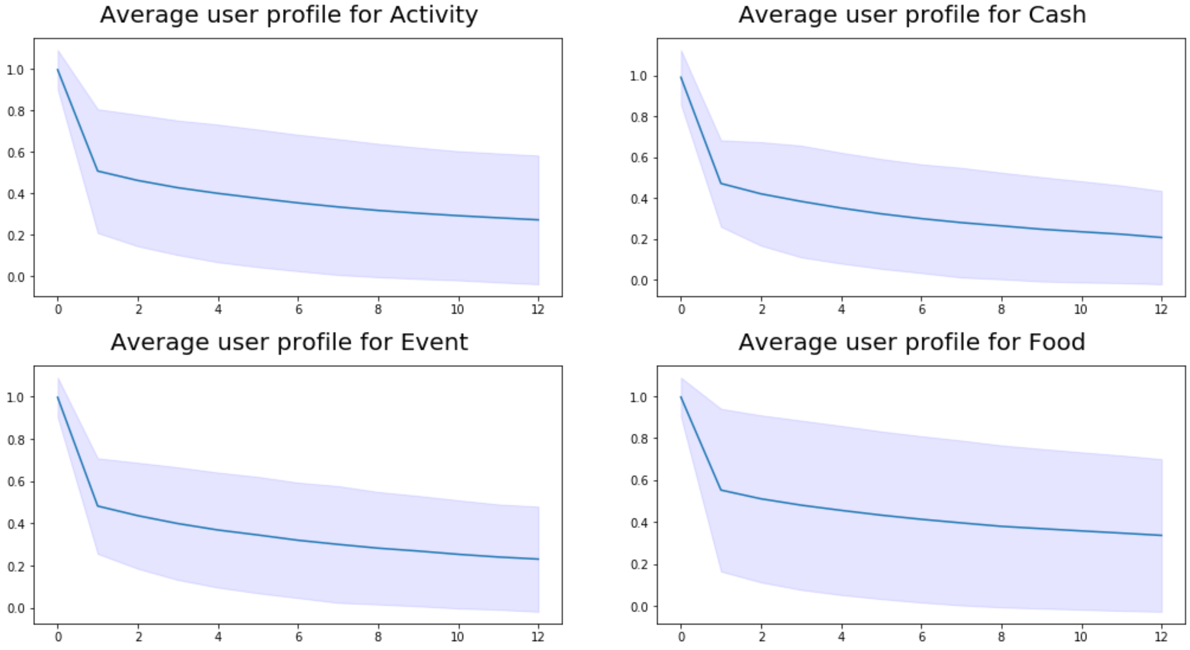
Then, we tried to establish user profiles. The static user profile identified the proportion of each type of spending for each user. For example, it is showed that user 2 spent most of the money on Food, and user 4 has a diverse set of different types of spending, Event, Food, People, Travel, and Utility. The dynamic user profile explored how the proportion of each type of spending change over time. We calculated the average and standard deviation across all users and excluded values that are zero. It is showed that the average percentage of different types of spending usually stabilized in around a month.

Static User Profile



```
+-----+---------------------------------------------------------------------------------------------+
|user1|profile                                                                                      |
+-----+---------------------------------------------------------------------------------------------+
|2    |[[Others, 1.0]]                                                                              |
|3    |[[Others, 0.83], [Utility, 0.17]]                                                            |
|4    |[[Activity, 0.2], [Food, 0.4], [Others, 0.2], [Travel, 0.2]]                                 |
|10   |[[Activity, 0.1], [Food, 0.3], [Others, 0.4], [People, 0.2]]                                 |
|11   |[[Event, 0.12], [Food, 0.12], [Others, 0.56], [People, 0.12], [Travel, 0.04], [Utility, 0.04]]|
+-----+---------------------------------------------------------------------------------------------+
```

Dynamic User Profile



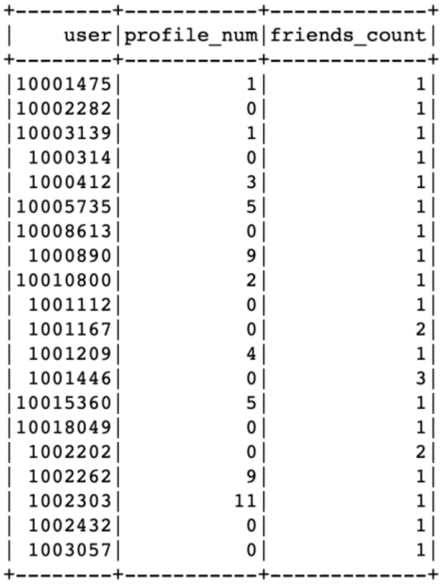
Social Network Analytics

We will use the user1 and user2 to identify the user’s friends as well as friends of friends and calculate social network metrics.

Friends are defined as other users who had transactions with the user. First, we union user1 and user2. Then, we selected distinct combinations. The computational complexity is 2n.

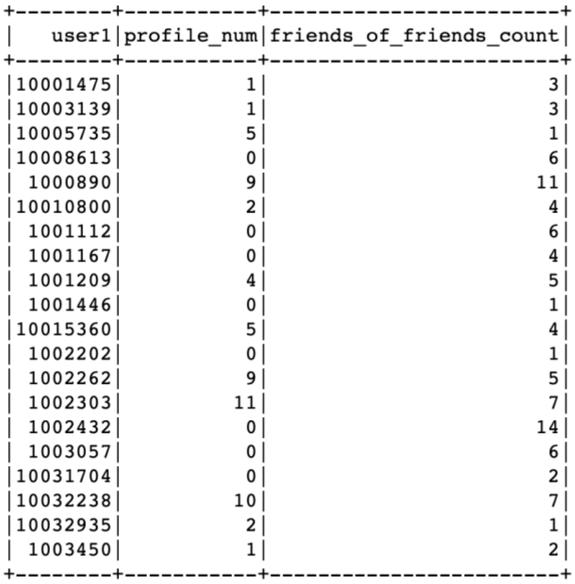
Friends of friends are defined as other users who had transactions with the user’s friends but not the user. First, we left join user table with user table, and user1 as the original user, user2 as friends, and user3 as friends of friends. Then, we excluded the rows where user3=user1 and user3 =user2. The computational complexity is O(M+N).

Friends Count



```
+--------+-----------+-------------+
|    user|profile_num|friends_count|
+--------+-----------+-------------+
|10001475|          1|            1|
|10002282|          0|            1|
|10003139|          1|            1|
| 1000314|          0|            1|
| 1000412|          3|            1|
|10005735|          5|            1|
|10008613|          0|            1|
| 1000890|          9|            1|
|10010800|          2|            1|
| 1001112|          0|            1|
| 1001167|          0|            2|
| 1001209|          4|            1|
| 1001446|          0|            3|
|10015360|          5|            1|
|10018049|          0|            1|
| 1002202|          0|            2|
| 1002262|          9|            1|
| 1002303|         11|            1|
| 1002432|          0|            1|
| 1003057|          0|            1|
+--------+-----------+-------------+
```

Friends of Friends Count

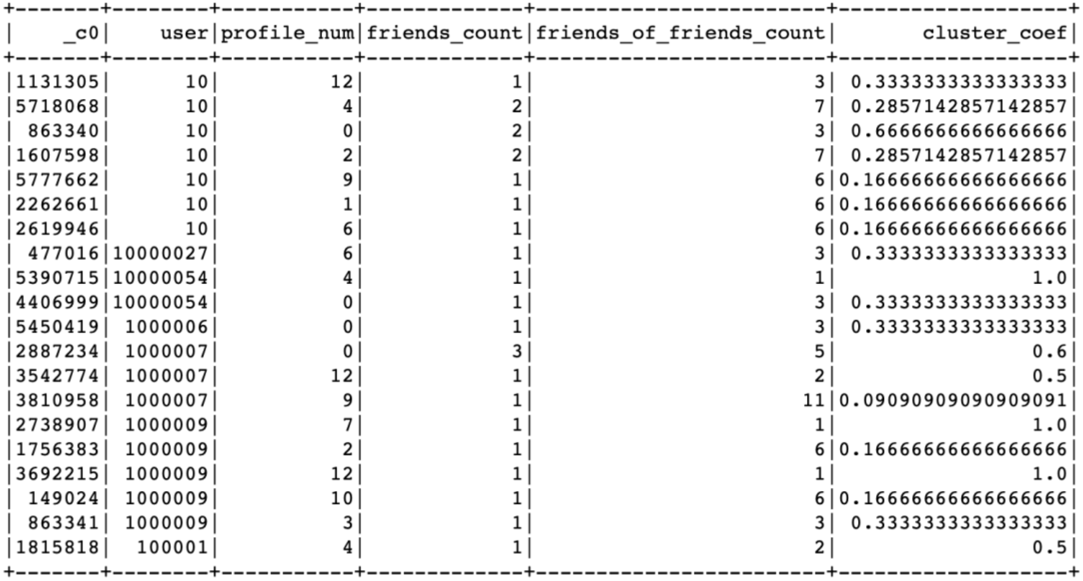


```
+--------+-----------+-----------------------+
|   user1|profile_num|friends_of_friends_count|
+--------+-----------+-----------------------+
|10001475|          1|                      3|
|10003139|          1|                      3|
|10005735|          5|                      1|
|10008613|          0|                      6|
| 1000890|          9|                     11|
|10010800|          2|                      4|
| 1001112|          0|                      6|
| 1001167|          0|                      4|
| 1001209|          4|                      5|
| 1001446|          0|                      1|
|10015360|          5|                      4|
| 1002202|          0|                      1|
| 1002262|          9|                      5|
| 1002303|         11|                      7|
| 1002432|          0|                     14|
| 1003057|          0|                      6|
|10031704|          0|                      2|
|10032238|         10|                      7|
|10032935|          2|                      1|
| 1003450|          1|                      2|
+--------+-----------+-----------------------+
```

We successfully identified users' friends as well as friends of friends. Then, we calculated the clustering coefficient and the PageRank to better capture their social network.

The Clustering Coefficient measures how connected a vertex's neighbor is to one another. In other words, if your friends all know each other, you have a high clustering coefficient. More specifically, it is calculated as the number of edges connecting a vertex's neighbors divided by the total number of possible edges between the vertex's neighbors. We calculated the clustering coefficient as friends count divided by the friends of friends count.

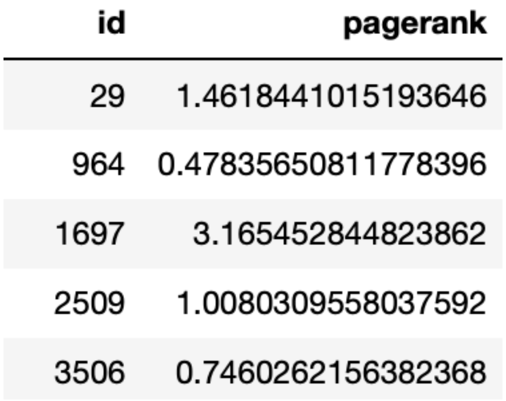
Clustering Coefficient



```
+-------+--------+-----------+-------------+------------------------+-------------------+
|    _c0|    user|profile_num|friends_count|friends_of_friends_count|        cluster_coef|
+-------+--------+-----------+-------------+------------------------+-------------------+
|1131305|      10|         12|            1|                       3| 0.3333333333333333|
|5718068|      10|          4|            2|                       7| 0.2857142857142857|
| 863340|      10|          0|            2|                       3| 0.6666666666666666|
|1607598|      10|          2|            2|                       7| 0.2857142857142857|
|5777662|      10|          9|            1|                       6|0.16666666666666666|
|2262661|      10|          1|            1|                       6|0.16666666666666666|
|2619946|      10|          6|            1|                       6|0.16666666666666666|
| 477016|10000027|          6|            1|                       3| 0.3333333333333333|
|5390715|10000054|          4|            1|                       1|                1.0|
|4406999|10000054|          0|            1|                       3| 0.3333333333333333|
|5450419| 1000006|          0|            1|                       3| 0.3333333333333333|
|2887234| 1000007|          0|            3|                       5|                0.6|
|3542774| 1000007|         12|            1|                       2|                0.5|
|3810958| 1000007|          9|            1|                      11|0.09090909090909091|
|2738907| 1000009|          7|            1|                       1|                1.0|
|1756383| 1000009|          2|            1|                       6|0.16666666666666666|
|3692215| 1000009|         12|            1|                       1|                1.0|
| 149024| 1000009|         10|            1|                       6|0.16666666666666666|
| 863341| 1000009|          3|            1|                       3| 0.3333333333333333|
|1815818|  100001|          4|            1|                       2|                0.5|
+-------+--------+-----------+-------------+------------------------+-------------------+
```

The PageRank is best known as the core metric behind Google's search engine. It includes three distinct factors, the number of vertices that link to the target, the PageRank centrality of the linking vertices, and the link propensity of the linking vertices. In other words, your PageRank will increase if you have more friends, if your friends have high PageRank, and your friends do not have many other friends. We used GraphX to calculate the PageRank.

PageRank



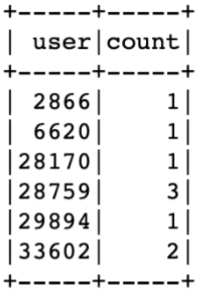
id	pagerank
29	1.4618441015193646
964	0.47835650811778396
1697	3.165452844823862
2509	1.0080309558037592
3506	0.7460262156382368

Predictive Analytics

One of the biggest questions in Customer Relationship Management (CRM) is to predict the number of future transactions. We predicted the total number of transactions a user will have by the end of their first year in Venmo based on transaction dataset as well as text and social network analytics above.

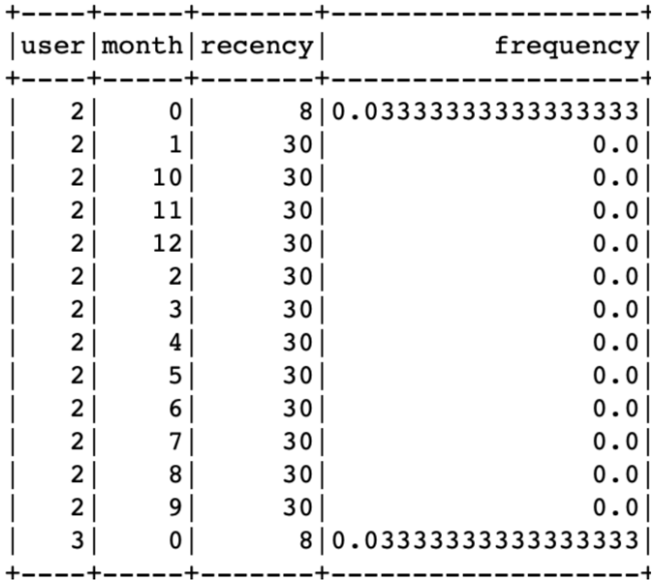
First, we created a dependent variable Y, which is the total number of transactions by the end of the first year. Then, we created the recency and frequency variables. Recency refers to the last time a user was active, and frequency refers to how often a user uses Venmo in a month. It is worth mentioning that if there is no transaction, the recency will be 30 and the frequency will be 0. Finally, we developed multiple regression models and plotted the accuracy of the model for each lifetime point.

Dependent Variable Y



```
+-----+-----+
| user|count|
+-----+-----+
| 2866|    1|
| 6620|    1|
|28170|    1|
|28759|    3|
|29894|    1|
|33602|    2|
+-----+-----+
```

Recency and Frequency



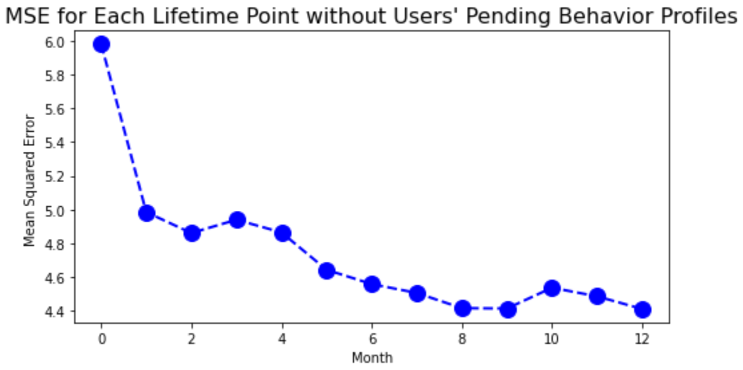
```
+----+-----+-------+-------------------+
|user|month|recency|          frequency|
+----+-----+-------+-------------------+
|   2|    0|      8|0.03333333333333333|
|   2|    1|     30|                0.0|
|   2|   10|     30|                0.0|
|   2|   11|     30|                0.0|
|   2|   12|     30|                0.0|
|   2|    2|     30|                0.0|
|   2|    3|     30|                0.0|
|   2|    4|     30|                0.0|
|   2|    5|     30|                0.0|
|   2|    6|     30|                0.0|
|   2|    7|     30|                0.0|
|   2|    8|     30|                0.0|
|   2|    9|     30|                0.0|
|   3|    0|      8|0.03333333333333333|
+----+-----+-------+-------------------+
```

Model

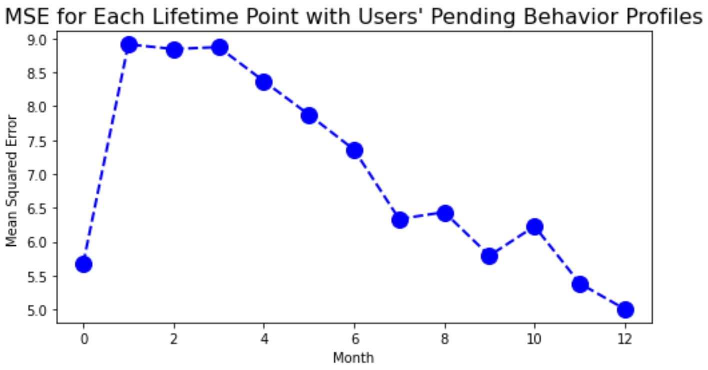
Model	**Independent Variables**
Model 1	recency and frequency
Model 2	recency, frequency, and spending profile
Model 3	social network metrics
Model 4	social networks metrics and spending profile

Model 1 regressed recency and frequency and model 2 regressed recency, frequency, and the spending profile. Both of the models have descending trends, which indicate that more recent month's data could better predict the outcome for the year. Surprisingly, the spending profile did not improve the performance of the model. Recency and frequency alone achieved better accuracy.

MSE for Model 1

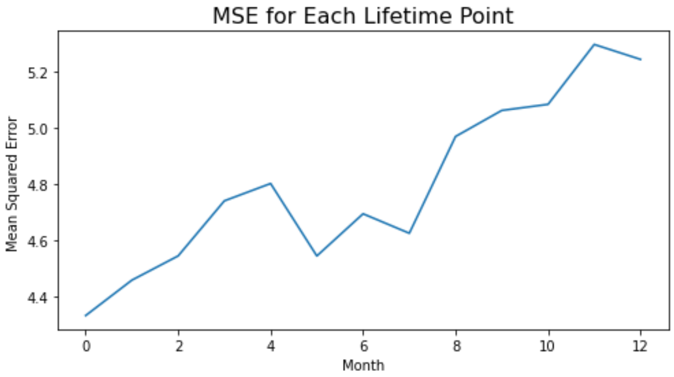


MSE for Model 2



Model 3 regressed social network metrics and model 4 regressed social network metrics and spending profile. Compared to recency and frequency, it seems that social network metrics performed better at predicting future transactions. Let's look into the coefficients, it seems that the user with higher PageRank and more friends will have more transactions. The result highlighted the unique feature of Venmo, sharing when transacting. Additionally, adding spending profile significantly improved the model performance, and all the types spending add to the total number of transactions, except for Illegal.

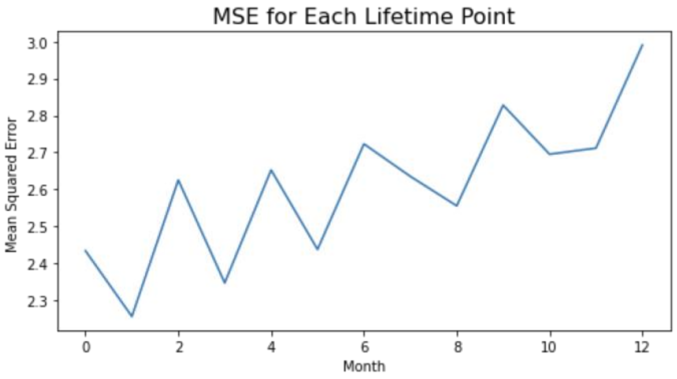
MSE for Model 3



Coefficients for Model 3

Independent Variables	Coefficients
pagerank	0.6266
friends_count	0.5798
friends_of_friends_count	0.0029
cluster_coef	-0.4104

MSE for Model 4



Coefficients for Model 4

Independent Variables	**Coefficients**
pagerank	0.2907
friends_count	0.2836
friends_of_friends_count	0.0014
cluster_coef	-0.2515
Activity	2.7021
People	2.8364
Transportation	2.7633
Event	2.5235
Utility	2.8739
Cash	2.6574
Travel	2.5434
Illegal_Sarcasm	0.0
Food	2.9636
Others	3.1651

Conclusion

We started with establishing the user's static and dynamic profile. It is showed that different users have different spending behavior, which usually stabilized in about a month. Additionally, user's enjoying sending emojis. Then, we tried investigating the user's social networks. We counted the user's friends and friends of friends and calculated clustering coefficient and PageRank. Finally, we utilized the RFM model and analytics above to run regression model to predict the number of future transactions of users. It is worth mentioning that we developed model for each lifetime point. We discovered that more recent month's data of recency and frequency could better predict the outcome for the year. We also discovered that the social network metrics combined with the spending profile have the highest predictive power. The coefficients indicate that the user with higher PageRank and more friends will have more transactions, which highlighted the unique feature of Venmo, sharing when transacting. Based on these analytics, we suggested Venmo to emphasize its social flavor and to encourage its user to expand their social networks. Additionally, Venmo could have more business partnerships, so that users could have a more diverse spending profile. With these methods, Venmo could enhance its user stickiness, have more active accounts, and achieve higher volume.