eHealth

eHealth |Cohort Analysis

Test Report

Date: 22nd March 2023
Created by: Varun Bhavnani
Created for: eHealth
No. of Pages: 9

Contents

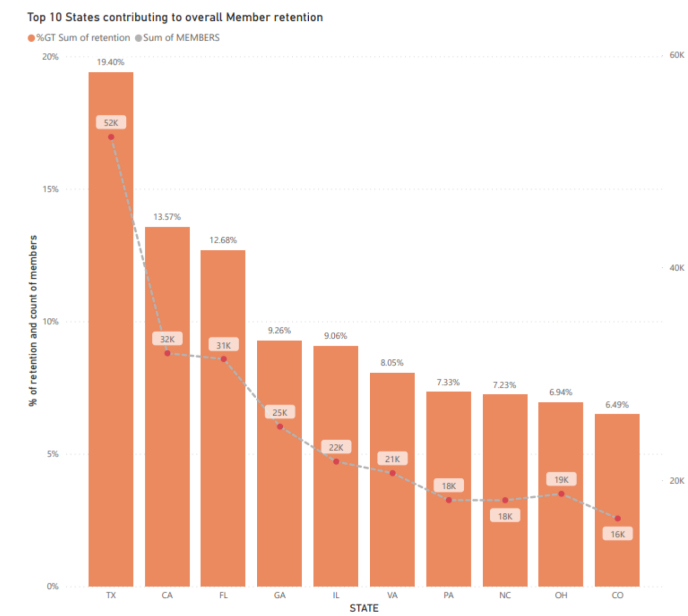
Problem Statement 3
Descriptive Statistics 3
Part 1: Lifetime Value for a cohorts 4
Part 2: Cohort Analysis 4
2.1. Factors that impact the length of member Lifetime 4
2.2. Monthly & Cumulative Retention Curve 6
2.3. Cumulative Retention Curve 2019 prediction 7

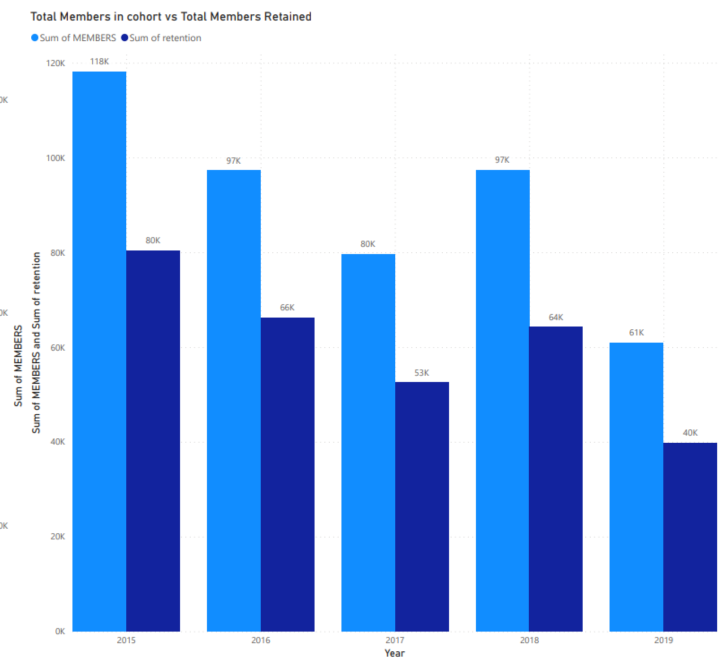
Problem Statement

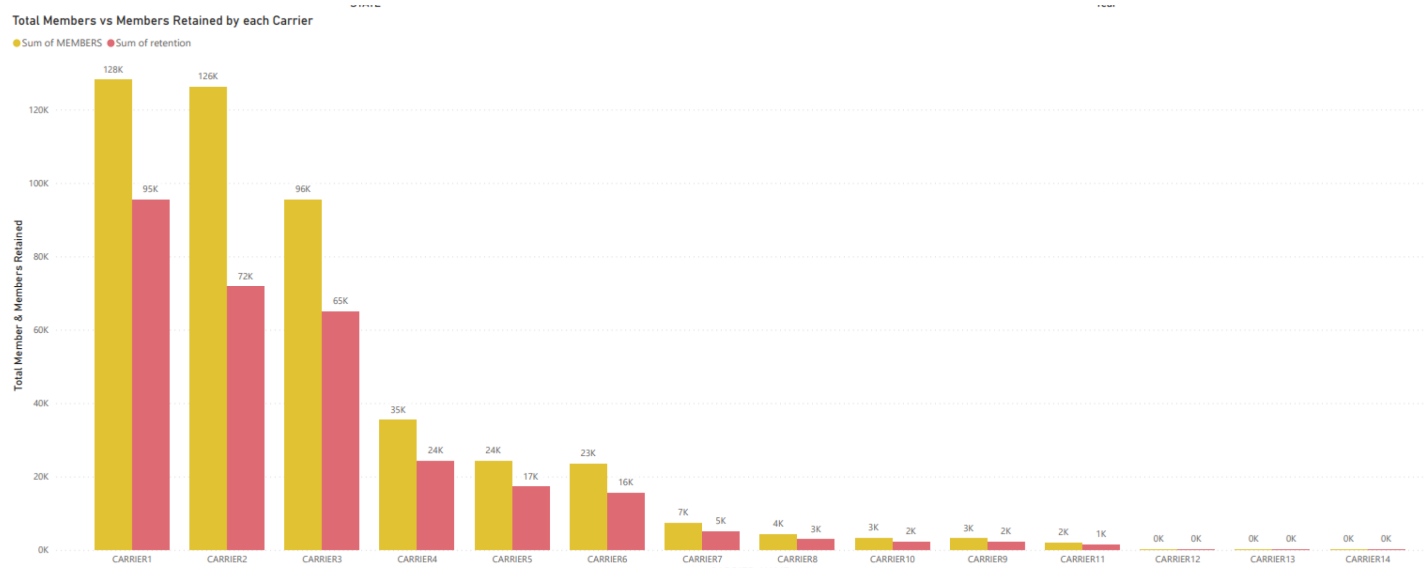
Cohort Analysis is a form of behavioral analytics that takes data from a given subset and groups it into related groups rather than looking at the data as one unit. The groupings are referred to as cohorts. They share similar characteristics such as time and size. In this report, we perform cohort analysis on paying member historical data to understand how each cohort performs, factors affecting member's lifetime & Cumulative Retention Rate forecasting.

Descriptive Statistics

Below are some statistics for the given data:







Part 1: Lifetime Value for a cohorts

The formula for Customer Lifetime Value (LTV) for a particular cohort, using the given metrics, can be calculated as follows:

LTV = (CMPM * MRR * IR) / (1 - MRR)

Where:

CMPM = Commission Per Member Per Month

MRR = Monthly Retention Rates across the cohort lifetime

IR = Interest Rate

The numerator in the formula represents the expected revenue generated from each member in the cohort over their lifetime, while the denominator represents the expected percentage of members who will churn or leave during the cohort lifetime.

To calculate LTV for a specific cohort, we need to calculate the average values of CMPM, MRR, and IR across the cohort lifetime. It is important to note that LTV is a forward-looking metric that is based on assumptions about future customer behavior. As such, it is subject to variability and uncertainty and should be used in conjunction with other metrics and analysis to inform business decisions

Part 2: Cohort Analysis

2.1. Factors that impact the length of member Lifetime

To understand the factors that affect the member Lifetime, we first work on defining cohorts, calculating the lifetime of members & **regrouping different states into US regions** in order generalize our findings

- Define cohorts: We can define cohorts based on the start date or start month of the policy. For example, we can create cohorts based on the month in which the policy was started, such as the January 2015 cohort, February 2015 cohort, and so on
- Calculate the lifetime of members: We can calculate the lifetime of members by subtracting the churn date from the start date. If the policy is still active, we can use the end of the dataset, December 2019, as the churn date

Analyze the lifetime of members by cohort: We can calculate the average lifetime of members in each cohort and compare them to see if there are any trends or differences

Identify factors that impact the length of member lifetime: Utilized OLS mode to understand the impact of each variable

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Member Lifetime   R-squared:                       0.161
Model:                            OLS   Adj. R-squared:                  0.161
Method:                 Least Squares   F-statistic:                     3234.
Date:                Wed, 22 Mar 2023   Prob (F-statistic):               0.00
Time:                        15:25:29   Log-Likelihood:            -1.7922e+06
No. Observations:              303235   AIC:                         3.584e+06
Df Residuals:                  303216   BIC:                         3.585e+06
Df Model:                          18
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          87.0724      0.162    537.256      0.000      86.755      87.390
MEMBERS        -3.9924      0.164    -24.329      0.000      -4.314      -3.671
MAX_DURATION   21.9542      0.200    109.715      0.000      21.562      22.346
Midwest        -1.6225      0.177     -9.167      0.000      -1.969      -1.276
Northeast      -1.5678      0.168     -9.326      0.000      -1.897      -1.238
West           -5.6777      0.185    -30.760      0.000      -6.040      -5.316
CARRIER1      -23.3397      6.559     -3.559      0.000     -36.195     -10.485
CARRIER10      -2.3634      1.196     -1.976      0.048      -4.707      -0.020
CARRIER11      -4.5338      0.987     -4.595      0.000      -6.468      -2.600
CARRIER12       0.3307      0.285      1.159      0.247      -0.229       0.890
CARRIER13      -0.6795      0.278     -2.440      0.015      -1.225      -0.134
CARRIER2      -11.0508      6.001     -1.841      0.066     -22.813       0.711
CARRIER3      -12.5868      5.795     -2.172      0.030     -23.946      -1.228
CARRIER4       16.1674      3.837      4.214      0.000       8.647      23.687
CARRIER5       -9.0450      3.269     -2.767      0.006     -15.452      -2.638
CARRIER6        4.8656      3.116      1.562      0.118      -1.241      10.973
CARRIER7       -8.4432      1.814     -4.655      0.000     -11.998      -4.888
CARRIER8       -0.5957      1.380     -0.432      0.666      -3.301       2.110
CARRIER9       -4.0129      1.207     -3.325      0.001      -6.379      -1.647
==============================================================================
Omnibus:                   162804.272   Durbin-Watson:                   1.593
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2707652.285
Skew:                           2.217   Prob(JB):                         0.00
Kurtosis:                      16.951   Cond. No.                         102.
==============================================================================
```

Note:

- *Our statistical significant level (alpha) is 0.05*
- *The Ordinary Least Square (OLS) model is the appropriate approach for analyzing the impact of variables such as US regions, carriers, max policy duration,etc. Grouping observations by cohorts may obscure the effects of these variables and hinder accurate interpretation, particularly for non-averaging variables like carriers and US regions*
- *Our dependent variable (Y) is Member's lifetime that is calculated in number of days*
- *Categorical variables like 'South' and 'CARRIER14' is skipped to avoid multicollinearity issue*

Interpretation:

- Carriers: 12,2,6, & 8 are insignificant in terms of impact on member lifetime
- Some statistical significant factors that affect member's lifetime value:
  - Members in a policy: One addition of a member in a policy can decrease the lifetime by 3 days
  - Customers from West U.S region: If a policy is initiated from west US region, it's likely possible that it will decrease the member's lifetime by 3 days as compared to those policies initiated from south region
  - Maximum duration of the policy: If maximum duration of a policy is increase by 1 day, member's lifetime increases by 22 days

- ○ Insurance carrier 1: If Insurance carrier 1 is selected against carrier 14, the lifetime decreases by 23 days
- Rest of the variables do have impact on lifetime of members which can be viewed from the OLS output above

2.2. Monthly & Cumulative Retention Curve

Following are the steps taken in-order to calculate the retained members for each Cohort:

- Cohort start date: Identified the date when the cohort of users started their subscription. For example, if all users started their subscription in January 2022, then January 2022 would be the cohort start date.
- Cohort end date: Identified the date when the cohort of users completed their first month of subscription. For example, if all users started their subscription in January 2022, then February 2022 would be the cohort end date.
- Number of retained users: Counted the number of users who are still active at the end of the current month and who were also active in the previous month. Identified the users whose CHURN_DATE is greater than or equal to the last day of the previous month and whose START_DATE is less than or equal to the last day of the current month

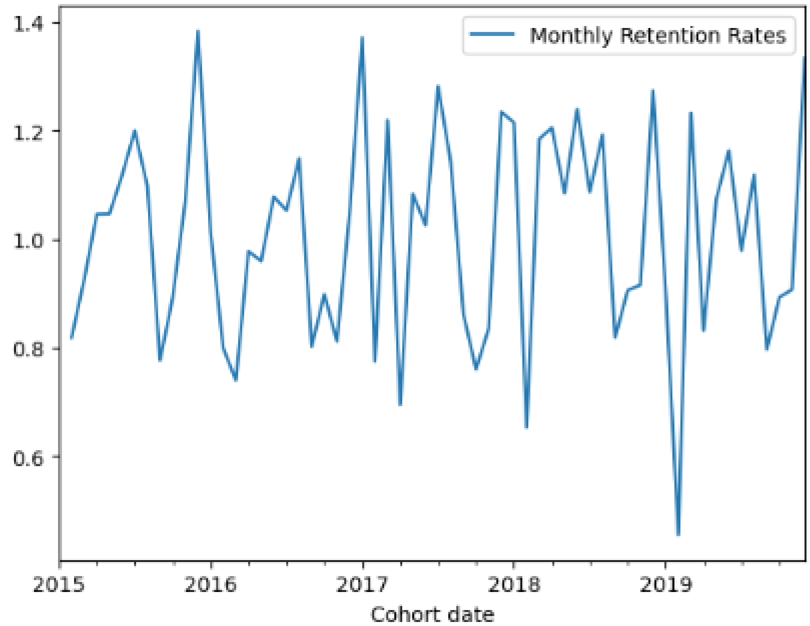
```
# Define the conditions to assign retention values
conditions = [
   (df['CHURN_DATE'] >= pd.to_datetime(df['Cohort']) - pd.DateOffset(months=1)) &
(pd.to_datetime(df['START_DATE']) <= pd.to_datetime(df['Cohort'])+ pd.DateOffset(months=1)),
   (df['CHURN_DATE'] < pd.to_datetime(df['Cohort']) - pd.DateOffset(months=1)) & (pd.to_datetime(df['START_DATE'])
> pd.to_datetime(df['Cohort'])+ pd.DateOffset(months=1))
]

# Define the values to assign for each condition
values = [1, 0]
```

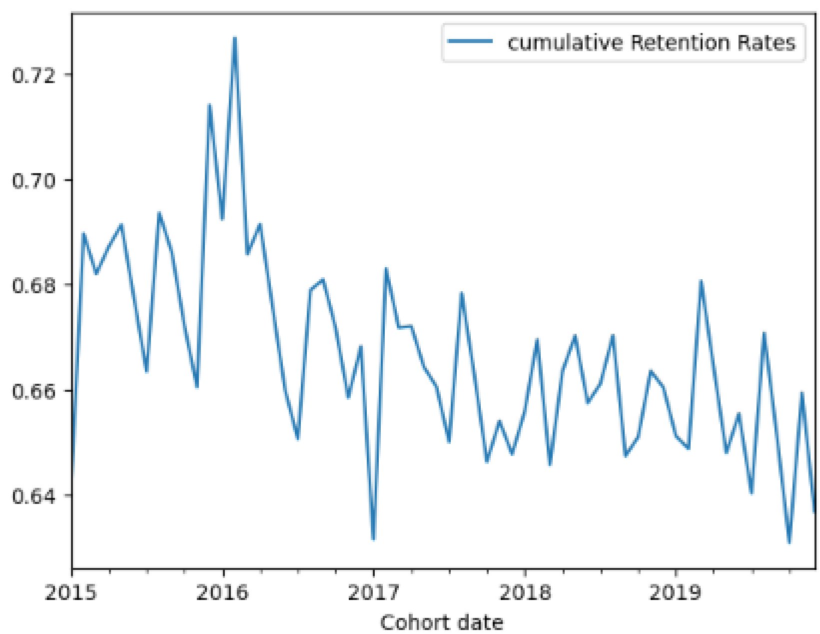
- Monthly Retention Rate: Retained members at month t/Retained members at month t-1
- Cumulative Retention Rate: Retained members at month t/Starting number of members at the beginning of a cohort's life
- Final Data:

	Cohort	total active members	total retention	Cohort date	Monthly Retention Rates	cumulative Retention Rates
20	January 2015	11011	7065	2015-01-01	NaN	0.641631
15	February 2015	8385	5782	2015-02-01	0.818401	0.689565
35	March 2015	7838	5345	2015-03-01	0.924421	0.681934
0	April 2015	8131	5587	2015-04-01	1.045276	0.687123
40	May 2015	8454	5844	2015-05-01	1.046000	0.691270
30	June 2015	9638	6528	2015-06-01	1.117043	0.677319
25	July 2015	11798	7828	2015-07-01	1.199142	0.663502
5	August 2015	12381	8587	2015-08-01	1.096960	0.693563
55	September 2015	9726	6672	2015-09-01	0.776988	0.685996
50	October 2015	8846	5946	2015-10-01	0.891187	0.672168
45	November 2015	9626	6358	2015-11-01	1.069290	0.660503
10	December 2015	12318	8795	2015-12-01	1.383297	0.713996

Monthly Retention Curve:



Cumulative Retention Curve:



2.3. Cumulative Retention Curve 2019 prediction

AutoRegressive Integrated Moving Average (ARIMA) model is a right model to forecast the cumulative retention rate for 2019 using previous year's data. Following are the steps taken:

- Stationary Trend test: to use ARIMA time series forecasting model, we need to first check if the cumulative retention rate trend is stationary or not. To perform any form of forecasting, we need to first remove any non-stationary trend. The Augmented Dickey-Fuller test is one such measure that statsmodel readily provides. The ADF test aims to reject the null hypothesis that the given time-series data is non-stationary. It calculates the p-value and compares it with a threshold value or significance level of 0.05. If the p-value is less than this level, then the data is stationary; else, the differencing order is incremented by one

```
ADF Statistic: -2.8751159621306908
p-value: 0.04831913759912574
Critical Values:
	1%: -3.548493559596539
	5%: -2.912836594776334
	10%: -2.594129155766944
```

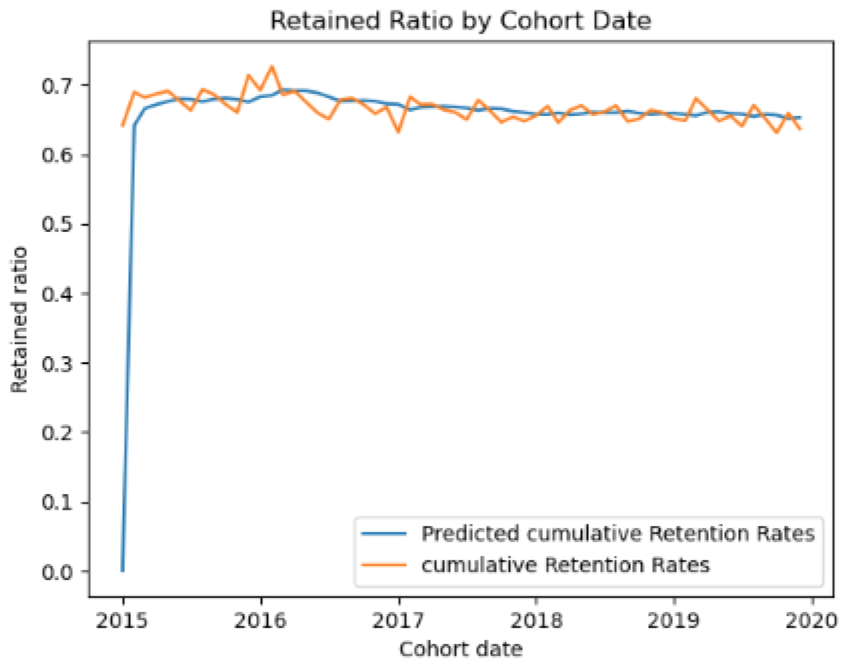
The p-value is less than alpha 0.5, so the data is stationary

- Then we fit the ARIMA model and calculated the **Root Mean Squared Error (RMSE) = 0.0145**. Lesser RMSE value, better the model
- Here is the actual Cumulative Retention rate vs its Predicted value

Actual Cumulative Retention Rate 2019	Predicted Cumulative Retention Rate 2019
0.6512136453	0.6591713205
0.6487909983	0.657610543
0.6805300714	0.6558706597
0.6641951686	0.6607032461
0.6480313248	0.6614143687
0.6554399243	0.6587902837
0.6404006047	0.6581186227
0.6707964602	0.654636373
0.6514889944	0.6577909722
0.6309133489	0.6565703578
0.6594652984	0.6515262541
0.6368616709	0.6530585068

Accuracy in terms of RMSE: 0.0145

Cumulative Retention rate vs Predicted Cumulative Retention rate



- End of Document -