CAL & UMAA Knowledge Transfer Documentation

Sameeksha Aithal | Xiangke Chen | Sam Musch | Pardha Pitchikala | Patrick Seng

University of Minnesota
Driven to Discover℠

0 Document Purpose

This document was designed to provide a detailed description of our analytics process including but not limited to: decisions made, tools and techniques used, deliverables produced and estimated value added.

For exhaustive documentation of each script produced, please see the “6 - Knowledge Transfer Documentation” folder which is included in the folder structure that has been delivered to UMAA via Box.

1 Project Definition and Introduction

1.1 Background & Context

Even with significant investment, UMAA's new member acquisition and member retention revenue have not met expectations in recent years. UMAA has experienced relatively stable Lifetime membership dollars over the past 5 years whereas Annual membership dollars have experienced yearly decline in the same time period.

1.2 Key Question

UMAA is hoping to better understand their data and how it can be used to increase solicitation efficiency. UMAA hopes that by understanding members they can more efficiently allocate their solicitation budget to drive new member enrollment and member retention.

1.3 Data Sources

The primary data sources were 5 point-in-time data extracts pulled from UMAA's DMS system and are listed below. These data files were pulled in January of 2020. Our provided solutions (dashboards and predictive models) rely on these files in their current form. We do however recommend that the data is refreshed regularly to ensure most up to date insights and accurate analysis.

Individual_info.csv - Demographic information for all UMN alumni. Also includes known information about family of alumni if they have been included in joint memberships.

Academics - Academic information for all alumni including college, degree level etc.

Membership.csv - Transactional style membership table capturing every membership purchase throughout the history of UMAA. This table includes $0 transactions representing complimentary memberships.

Engagement.csv - Yearly engagement scores for all alumni beginning in 2015. Contains Annual, Life engagement for every alumni on a yearly basis. UMAA specific engagement scores are designated with the years 1919, 1920 representing 2019 and 2020 respectively.

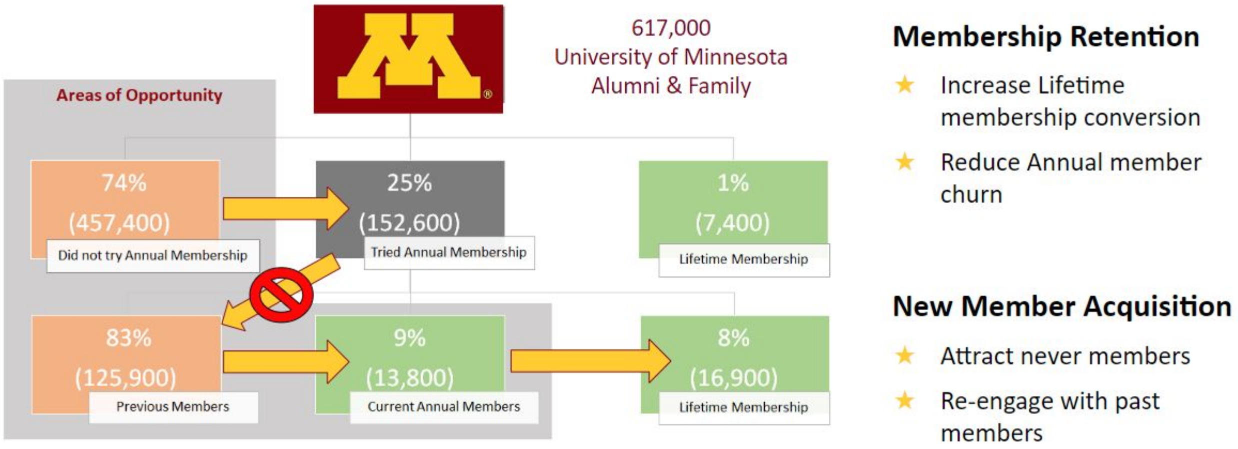
Events.csv - Event attendance for all alumni since 2015.

Emails.csv- Email records from all alumni. Includes information on if email was sent, bounced, opened, clicked etc.

2 Solution Overview

2.1 Solution Summary

The analytics solution provided by the CAL team was structured into several key components. After a preliminary membership breakdown we identified key areas of opportunity for UMAA to drive membership through increased solicitation efficiency.



617,000
University of Minnesota
Alumni & Family

Areas of Opportunity

74% (457,400) Did not try Annual Membership

25% (152,600) Tried Annual Membership

1% (7,400) Lifetime Membership

83% (125,900) Previous Members

9% (13,800) Current Annual Members

8% (16,900) Lifetime Membership

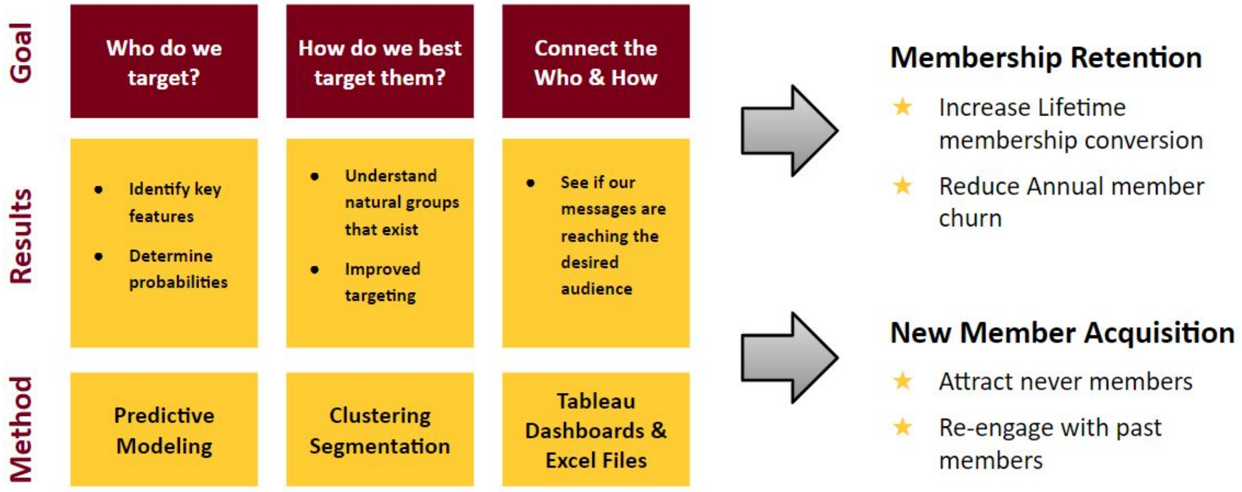
Membership Retention

- ★ Increase Lifetime membership conversion
- ★ Reduce Annual member churn

New Member Acquisition

- ★ Attract never members
- ★ Re-engage with past members

For both membership retention and new member acquisition objectives we utilized a sequential analytics structure centered around understanding who to target, how to target them and finally how to connect the two and measure success. Each of these pieces is discussed in further detail throughout this knowledge transfer document.



Goal	Who do we target?	How do we best target them?	Connect the Who & How
Results	• Identify key features • Determine probabilities	• Understand natural groups that exist • Improved targeting	• See if our messages are reaching the desired audience
Method	Predictive Modeling	Clustering Segmentation	Tableau Dashboards & Excel Files

Membership Retention

- ★ Increase Lifetime membership conversion
- ★ Reduce Annual member churn

New Member Acquisition

- ★ Attract never members
- ★ Re-engage with past members

3 Methodology

3.1 Description of Methods

3.1.1 Random Forest Predictive Modeling

Predictive modeling was used to determine who UMAA should be targeting when reaching out to alumni for membership upgrades, renewals and new membership purchases.

Random Forest is an improved version of the decision tree predictive model. To make the content easier to understand, let’s start with the decision tree. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:

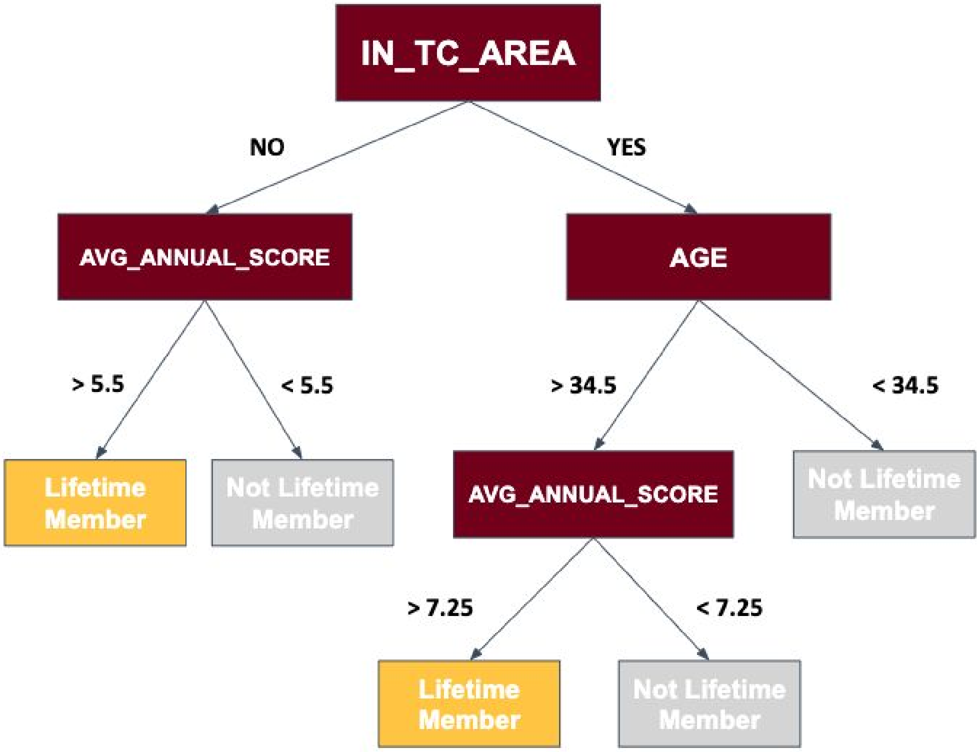
- Decision nodes – typically represented by squares
- Chance nodes – typically represented by circles
- End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

In terms of our project, the decision tree will help us identify who are more likely to become life-time members, fail to renew their membership or become new UMAA members.

The model will return a flow-chart like this:



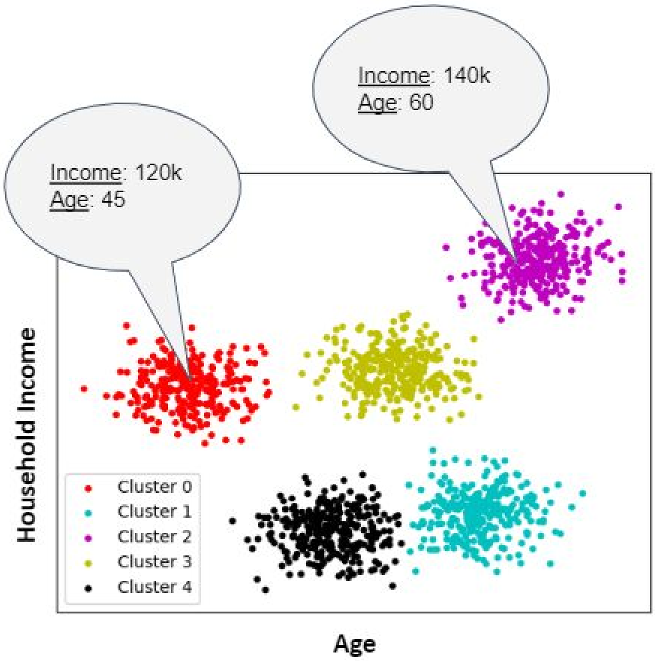
In our predictive modeling we are trying to understand key characteristics of those who became life members, not renew membership and purchase a new membership. In addition, we are calculating corresponding probabilities for each of these events occuring.

3.1.2 Clustering

In this analysis, we used clustering to identify the different demographic and behavioral patterns among current members. This will allow UMAA to understand how to best contact UMN alumni for membership upgrades, renewals and new membership purchases. Contacting alumni with content that best fits their consumption habits will maximize engagement and likelihood of purchase.

Clustering is a data driven segmentation technique that groups a population in such a way that the data points in the same groups are more similar to other data points in the same group than those in other groups. The aim is to segregate groups with similar traits and assign them into clusters.(https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/) This technique identifies natural groups present in the data and helps explore different behavioral patterns present in the data. This information can be used by businesses for better marketing decisions based on the preferences of each cluster group.

Below is an example of the output of clustering for household income and age.



We can see that the group in red has an average age of 45 and average household income value of 120,000 whereas the one in pink has an average age of 60 and an average household income value of 140,000.

While clustering based on demographics (above) was an important starting point. We found that behavioral clustering, based on alumni’s consumption behavior yielded results that best differentiated different alumni groups. Furthermore, we believe that these clusters will best enable UMAA to leverage personalized content to increase engagement and solicitation efficiency.

4 Data Engineering

4.0 Script to transform data files

The tables Academic, Events, Individual Info and Membership had double pipe delimiter. This was replaced with a single pipe using a python script. (read_academic.py, read_events.py, read_individual_info.py, read_membership.py).

If single pipe files are produced instead, the above python files can be disregarded.

4.1. Individual Info File

This file consists of demographic features of all alumni. There were a total of 617,372 records of all current,past, and non-members.

4.1.1 Data Cleaning (code file - Cleaning.R)

Out of 36 fields, 19 fields are chosen as they best represent the alumni information. The fields chosen are ID_DEMO, SPOUSE_ID, MARITAL_STATUS, GENDER, AGE, BIRTH_YEAR, DEATH_YEAR, ZIP_CODE, IN_TC_METRO_AREA, HOUSEHOLD_INCOME, IS_TC_GRAD, MOST_RECENT COLLEGE, MOST_RECENT_GRAD_YEAR, ATHLETIC_INTEREST, TRAVEL_INTEREST, MEMBERSHIP_STATUS_CODE, MEMBERSHIP_TYPE_CODE, FIRST_EVER_START_DATE, EXP_DATE_LAST_MEMBERSHIP.

4.1.2 Data Transformation

Blank values in the table are replaced with NA. Records of 39, 615 non-TC graduates are filtered out of the table. Thus, 577,757 records remain.

All records with a DEATH_YEAR value of either 9999 or NA are retained. 57,961 records are eliminated. MOST_RECENT_GRAD_YEAR is used to determine the BIRTH_YEAR for all those alumni with BIRTH_YEAR value of NA. AGE is ascertained using BIRTH_YEAR and records with AGE value less than 0 or greater than 100 are removed from the table. 504,573 records are retained. Average of AGE is used to impute values for records with NA values for AGE.

Zip code and household income information was taken from the United States Census Bureau(https://www.census.gov/search-results.html?q=minnesota+median+income&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8), and a github site (https://gist.github.com/Radcliffe/5ba82ed06ba92fa5e27e), and merged with the individual info table to determine the TC metro area residents. The records of 19,623 non-TC residents are filtered out of the dataset. The household income is also imputed using this third party data based on ZIP CODE and AGE. 484,950 records are retained. Missing values for HOUSEHOLD_INCOME are imputed using the mean HOUSEHOLD_INCOME value for AGE.

GENDER with M or F are preserved. Thus, 484,712 records of all alumni are conserved out of which 27,523 are current members.

4.2 Academic File

This file contains all students' academic information including college, major and graduate year. There are 1,699,908 records. The earliest student graduated from 1888, and the lastest student graduated from 2019. There are some duplicates (one person, multiple records) due to system change, or major code change. Hence, we only want to keep one record for each person. Per discussion with UMAA members, we noticed that students whose college code (column CODE_CLG_QUERY) length greater than 3 was not qualified, so we removed those records. For other duplicates, we leave the most updated one.

4.2.1 Data Cleaning (code file - Cleaning.R)

1) There are Null values in YEAR_GRAD columns. Hence, they were removed.
2) One student might have attended the school multiple times, ended up with several records the same except for YEAR_GRAD. Hence, they were removed.
3) The length of column CODE_CLG_QUERY greater than 3 was false value per discussion. Hence, they were removed.

Other than the above issues, the file is clean and no more data cleaning was required.

4.2.2 Data Transformation

The output file is aggregated level, containing academic information for each person and they only appear once in the file.

4.3 Membership file

This file is a transactional style table that provides information on every membership purchase that has occurred with UMAA. It includes complimentary membership transactions, reflecting a $0 balance for these line items.

4.3.1 Data Cleaning (code files: 1 - Cleaning.R)

Key data cleaning steps for the membership file included removing records that were of higher value than $1000 as they indicated a donation and did not belong in this table. Additionally, records below $0 were erroneous and removed. There was also an ID that had more than 50 transactions attributed to it. This record was deemed a “holding” record and was removed. Finally, we included only records that had “Annual” or “Life” in the MEMBERHSIP_LEVEL field to ensure we were only capturing true membership transactions in our analysis.

4.3.2 Data Transformation

Membership file's data transformation was a relatively complex process that has been explained in detail in files "4.1 - Predictive Model Prep Existing Member" and "4.2 - Predictive Model Prep New-Member". The main objective was to identify the member status for all alumni in each of the past five years. We selected a five year window to track alumni as we deemed this was an appropriate length of time to capture relevant historical membership information and also due to the engagement score tracking beginning in 2015. In R scripts 4.1 and 4.2, membership statuses for all alumni were captured over the past five years and then rolled up to produce a file that has one row per alumni ID with columns indicating the number of years each alumni was an annual member, a life member or a non-member. This was then joined with 5 years of engagement, events and email data to produce the files that are read into the predictive model.

4.4 Engagement score file

This file contains Annual and Lifetime engagement scores along with sub engagement scores like Loyalty, Stay Informed, Volunteering, Donation, UMAA membership and Events attended scores. The data is at Alumni and year level. We have data from Fiscal year 2015 to 2019. Until 2018, scores are available only at UMN level. However, UMAA has started calculating UMAA specific scores starting 2019. Hence, UMAA scores are available only for 2019 and we can identify them by selecting YEAR_FISCAL = 1919.

4.4.1 Data Cleaning (code file: 1 - Cleaning.R)

- There are Null values in Year_Fiscal columns. Hence, they were removed
- Values for Year_Fiscal = 2014 are 0's. Hence, they were removed

Other than the above issues, the file is clean and no more data cleaning was required

4.4.2 Data Transformation

So far, we have used only the UMN Total Annual score column (ENGAGEMENT_TOTAL_ANNUAL) since it is most frequently used by the UMAA team. Based on the analysis, we used the average value of this column for each Alumni and will be indicated in respective sections.

4.5 Emails file

This file contains every email that has been sent to every person along with information like a pseudo-subject line, date sent, and email status (eg clicked, opened, etc).

4.5.1 Data Cleaning (code file - Cleaning.R)

- We added in manual categories for each email to better understand the content a person responds to.
- We added in a fiscal year by using the year + month an email was sent.
- We added a column called “status_clicked”, because a “click” is a subset of an “open” but they are categorized separately in the original files.

Other than the above issues, the file is clean and no more data cleaning was required

4.6 Events file

This file contains every event that has been held along with information like a pseudo-subject line, date of event, and which people attended.

4.5.1 Data Cleaning (code file - Cleaning.R)

- We added in manual categories for each event to better understand the content a person responds to.
- We added in a fiscal year by using the year + month an event was held.

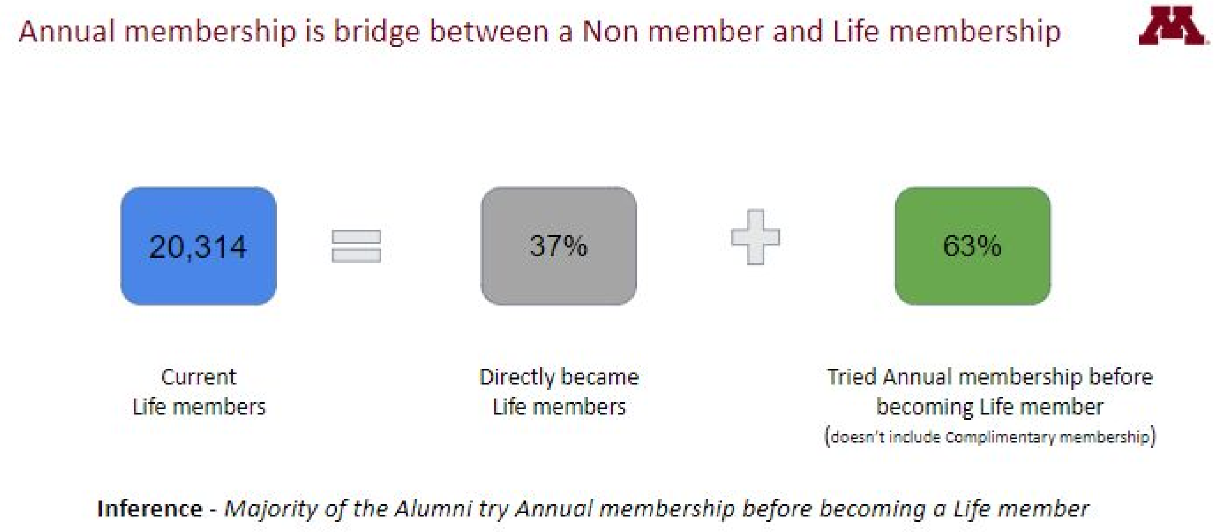
Other than the above issues, the file is clean and no more data cleaning was required

5 Insights and Trends

After exploring the **Membership** and **Individual info** data we found few insights that helped us identify the areas of opportunities to increase membership retention and renewals.

5.1. Life membership Analysis (code file - Insights and Trends.R)

Figure 5.1 - Past behaviour of Lifetime members



Annual membership is bridge between a Non member and Life membership

20,314 = 37% + 63%

Current Life members

Directly became Life members

Tried Annual membership before becoming Life member (doesn't include Complimentary membership)

Inference - *Majority of the Alumni try Annual membership before becoming a Life member*

There are 20,314 Life members in the fiscal year 2020 and looking at their past behaviour, we found that 37% of them directly bought Lifetime members whereas the rest of 63% first tried Annual membership at least once, before buying a Life membership.

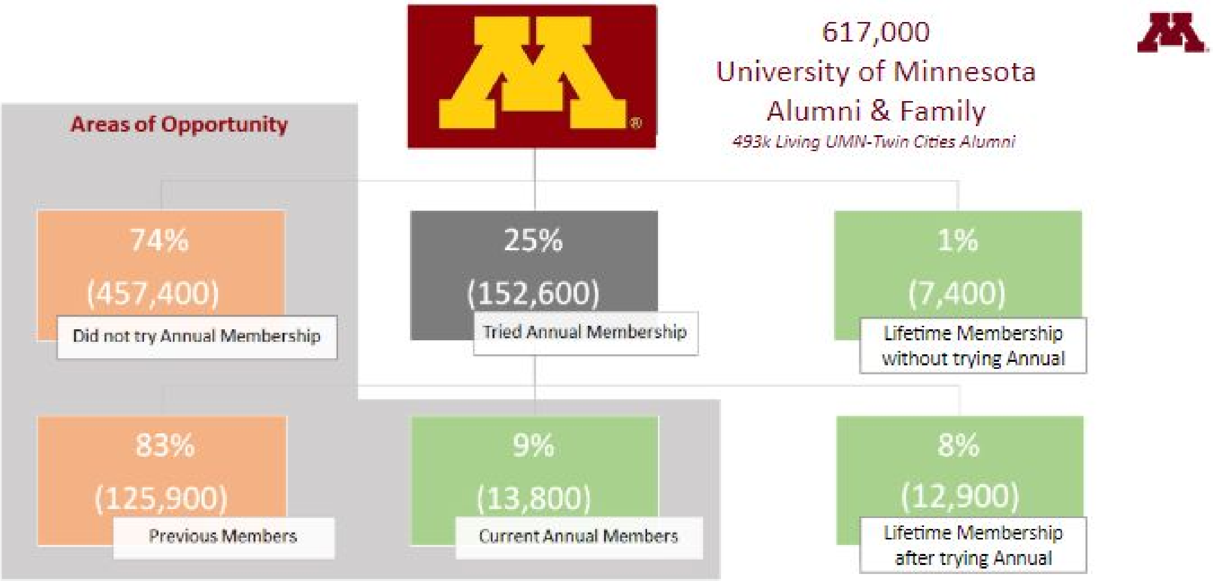
From this, we infer that Annual membership is a **bridge** for a Non member to buy a Life membership.

Please note that Alumni who tried complementary membership and purchased Life membership without trying Annual membership are included in Directly became Life members bucket.

5.2. Grouping Alumni based on membership status (code file - Insights and Trends.R)

An Alumni can be a Non member or an Annual member or a Life member. We wanted to group all the Alumni based on their membership status.

Figure 5.2 - Alumni groups based on membership status



Currently UMAA has data for approximately 617,000 Alumni including their family members. Out of these Alumni, 74% of them have not tried Annual membership so far. Next, 25% of them tried Annual membership at least once. The rest 1% directly bought Life membership.

Further, drilling down the group of Alumni who bought Annual membership at least once, we observed that 83% of them discontinued their Annual membership and became non members. Next, 9% of them are current Annual members and the remaining 8% converted from Annual to Life membership.

From this Big picture we identified 3 groups as areas of opportunities.

1) Alumni who have not tried Annual membership so far
2) Alumni who tried Annual membership at least once but discontinued
3) Alumni who are current Annual members

As explained previously, using this understanding UMAA can increase Membership retention and New member Acquisition.

In order to achieve this, we used 2 techniques.
Technique 1 - Prediction: Using predictive models we want to identify the probability of each Annual member becoming a Life member. This helps UMAA to select the Alumni with high probability for targeting. That is we know **WHO** to target.

Technique 2 - Segmentation: Each Alumni has unique interests. Using Segmentation we want to identify the interests of each Alumni. It can be sports emails or social events or Networking events etc. This part is identifying **HOW** to target.

Combining both the techniques, UMAA can efficiently target Annual members who are more likely to become Life members.

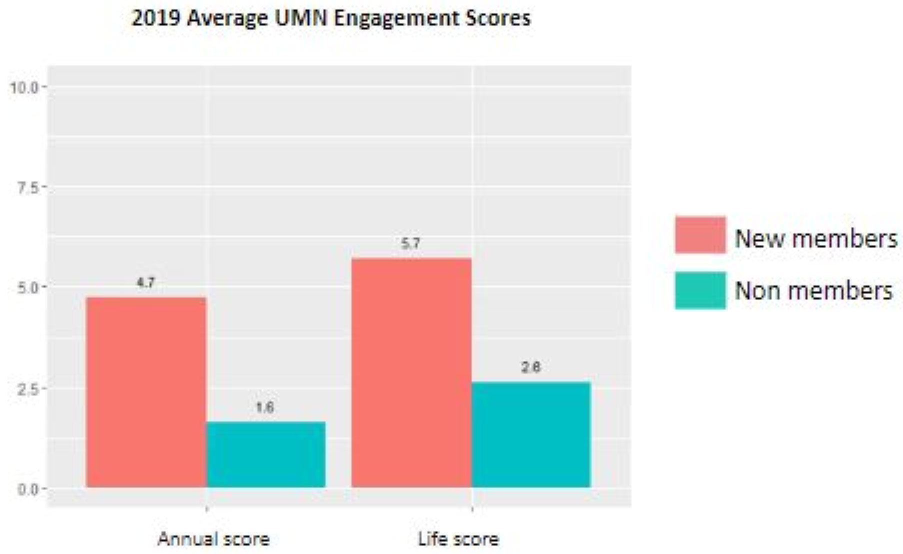
5.3. New members behaviour analysis

Up on doing analysis on New members of 2019, we found few notable insights. The below sections help UMAA understand in which sub engagement scores New members differ from Non members and also how the New members' behaviour changes in their membership year.

5.3.1 Engagement Scores comparison

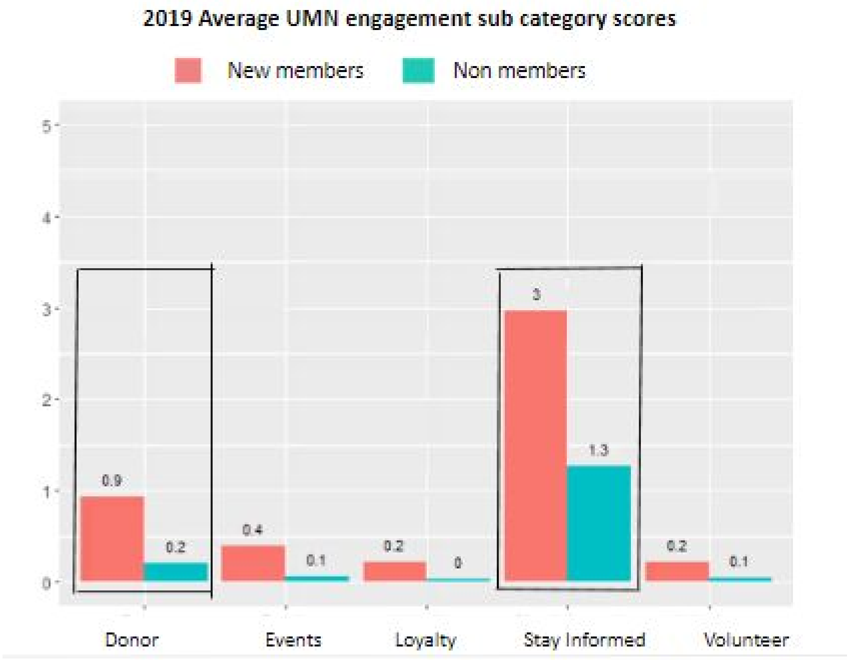
For this analysis, **New Members** are the Alumni who did not have any membership in 2018 and purchased a membership in 2019. Non members are the Alumni who did not have any membership in 2018.

Plotting UMN Annual scores for these 2 categories shows that New Members have 3x higher scores than Non-Members.



Please note that Membership subcategory scores have been removed from total Annual scores for this analysis.

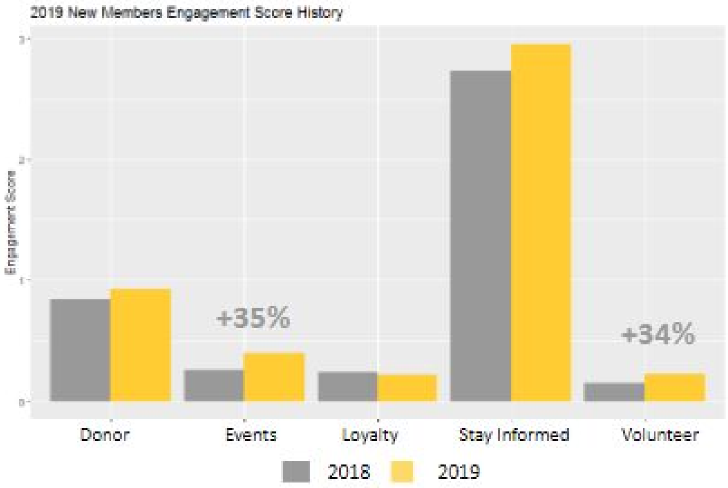
Next, we wanted to dig deeper and identify which subcategories are causing this huge difference. Hence,, when we drilled down into the subcategories we found that Donation score and Stay Informed score are differentiating New members from Non members.



Hence, it is worth contacting a few alumni who have a good Donation and Stay Informed score but do not have any membership and identify reasons for not taking any membership. Also, It is an indication that they are good candidates when UMAA is targeting for New member acquisition.

5.3.2 Behavior change in New members

We considered New members in 2019 and compared their behavior from 2018 to 2019 to check in which engagement categories their behaviour has changed significantly. We found that New members have attended 35% more events in 2019 which is their membership year compared to 2018 and it is the same with volunteering activities.



This data exploration on new members gave us the idea to use the subcategory scores in our predictive model to identify the likelihood of becoming a new member

6 Predictive Modeling

6.1 Model Selection

We used Decision Tree and Random Forest for prediction. Decision trees are easy to use and interpret. Random Forest could be understood as an improved version of the decision tree with higher model performance.

Here we used our lifetime conversion model for illustration purposes.

The model is predicting the likelihood of annual members to become life members, the likelihood of annual members to leave.

Model is built on 12K annual members in 2018. We considered demographic and members' behaviors including but not limited to age, household income, gender, annual engagement score, events attended, click through rate

We used their membership status in 2019 for validation to evaluate the model performance. And the model is ready and well-trained, the model could be used for predicting the membership status in 2020. This would help UMAA customize and better design their solicitation strategy next year.

6.2 Evaluation Methodology

We mainly use three metrics to evaluate this model, and the importance is ordered as below:

- Recall
- Precision
- Accuracy

Recall:
Recall means that among all Lifetime members in 2019, what percent of them did we identify? And how many life members could we retrieve in this model. The higher the recall, the more true life members could identify.

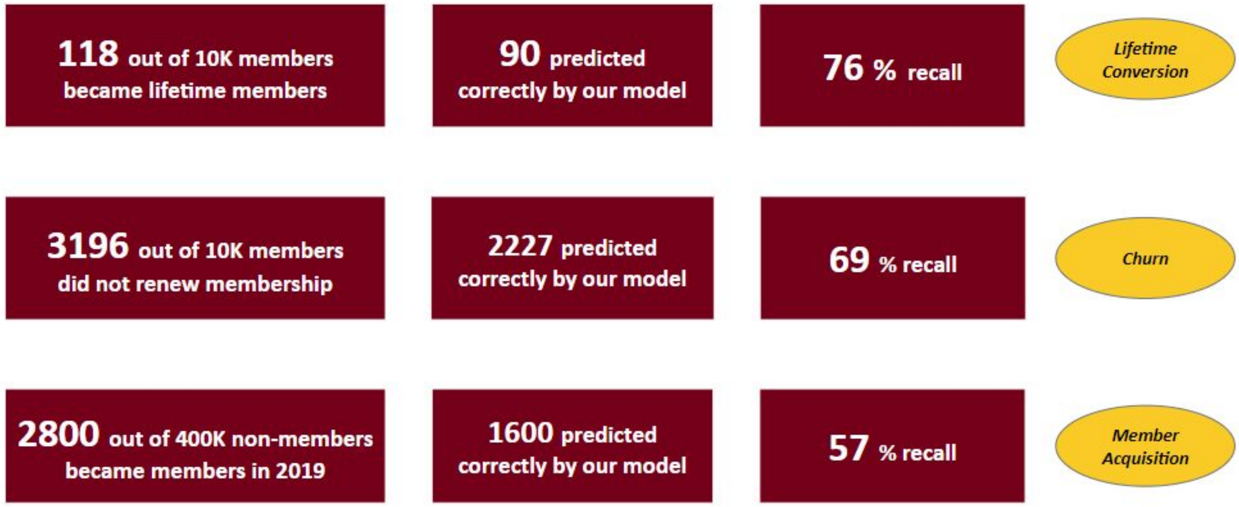
Precision:
Precision means that among all alumni that we predicted to become Lifetime members, how many actually are Lifetime members? The higher the precision, the less mistake we made in the prediction.
When we predict people to become life members, we want the prediction to be more useful and accurate. This measure basically reflects it.

Accuracy:

Accuracy means that the percentage of predictions are correct? We could have two prediction results, one for lifetime, one for not becoming lifetime. This measure basically considers the both sides and makes sure that the balance of two prediction outcomes.

Business Value
By having those probabilities, instead of sending emails to all the annual members, UMAA can know who to target first and avoid the risk of over-sending emails to them. Meanwhile, UMAA can also improve the solicitation efficiency.

Below are the final performance results of our predictive models.

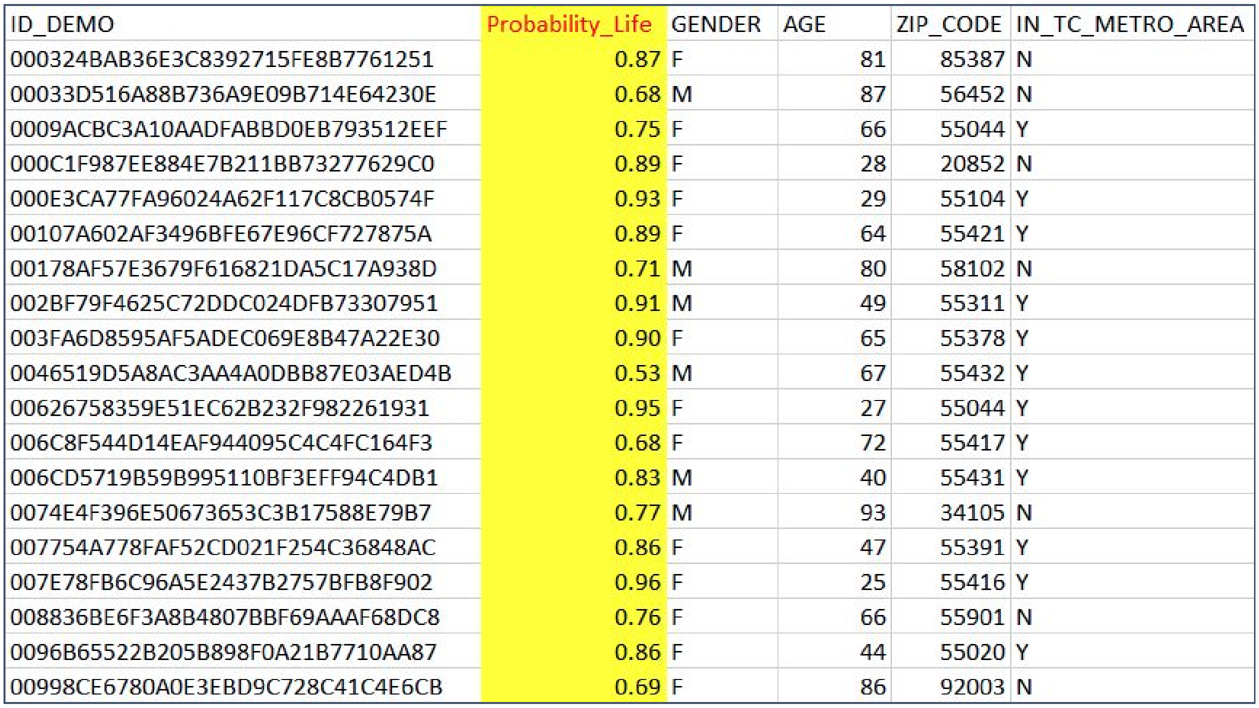


118 out of 10K members became lifetime members	90 predicted correctly by our model	76 % recall	Lifetime Conversion
3196 out of 10K members did not renew membership	2227 predicted correctly by our model	69 % recall	Churn
2800 out of 400K non-members became members in 2019	1600 predicted correctly by our model	57 % recall	Member Acquisition

6.3 Lifetime Conversion Analysis

The objective of lifetime conversion model is to identify which annual members are more likely to become lifetime members. There are around 12K annual members in 2018, we are predicting their membership status next year, to see if they would convert or not. Out of 12K annual members, 118 of them became lifetime members. Our model predicts 90 of them correctly, which gives us 76% recall.
From the lifetime conversion model, we have a probability spreadsheet where you could find the corresponding probability of converting to lifetime member for each annual member. And we also identify some leading indicators that associate with the membership status change.

Probability of Lifetime Conversion
The model will return a probability for each annual member converting to life member.

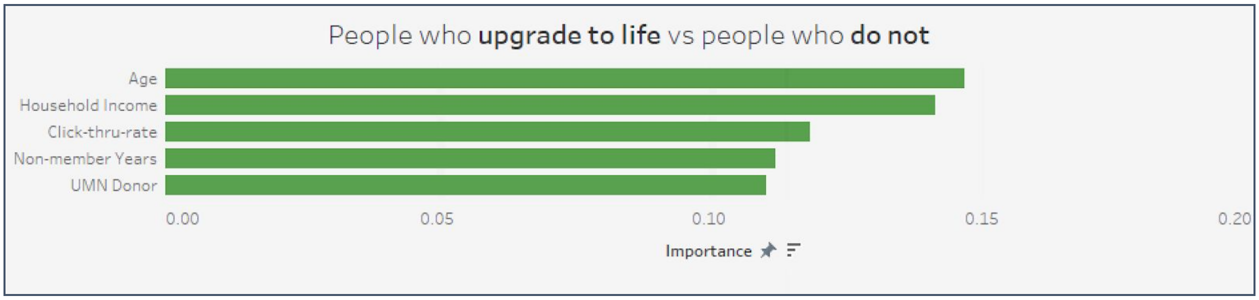


ID_DEMO	Probability_Life	GENDER	AGE	ZIP_CODE	IN_TC_METRO_AREA
000324BAB36E3C8392715FE8B7761251	0.87	F	81	85387	N
00033D516A88B736A9E09B714E64230E	0.68	M	87	56452	N
0009ACBC3A10AADFABBD0EB793512EEF	0.75	F	66	55044	Y
000C1F987EE884E7B211BB73277629C0	0.89	F	28	20852	N
000E3CA77FA96024A62F117C8CB0574F	0.93	F	29	55104	Y
00107A602AF3496BFE67E96CF727875A	0.89	F	64	55421	Y
00178AF57E3679F616821DA5C17A938D	0.71	M	80	58102	N
002BF79F4625C72DDC024DFB73307951	0.91	M	49	55311	Y
003FA6D8595AF5ADEC069E8B47A22E30	0.90	F	65	55378	Y
0046519D5A8AC3AA4A0DBB87E03AED4B	0.53	M	67	55432	Y
00626758359E51EC62B232F982261931	0.95	F	27	55044	Y
006C8F544D14EAF944095C4C4FC164F3	0.68	F	72	55417	Y
006CD5719B59B995110BF3EFF94C4DB1	0.83	M	40	55431	Y
0074E4F396E50673653C3B17588E79B7	0.77	M	93	34105	N
007754A778FAF52CD021F254C36848AC	0.86	F	47	55391	Y
007E78FB6C96A5E2437B2757BFB8F902	0.96	F	25	55416	Y
008836BE6F3A8B4807BBF69AAAF68DC8	0.76	F	66	55901	N
0096B65522B205B898F0A21B7710AA87	0.86	F	44	55020	Y
00998CE6780A0E3EBD9C728C41C4E6CB	0.69	F	86	92003	N

According to the probability, UMAA could prioritize members with high likelihood of conversion. In addition, UMAA can combine these insights with member behaviors to better solicit Lifetime memberships to these Annual Members.

Leading Indicators

In addition, a relatively important feature is concluded from the model that drives annual members to convert to life members. These features importance can guide UMAA to change the existing slicing strategy and add new important features for identifying potential members in the future analysis.



However, when you run the model multiple times, the order of the leading indicators might change a bit since some randomness exists in the models. So you might notice the difference. Overall, you can find some features always stand out how many times you run the model, they are very important and informative.

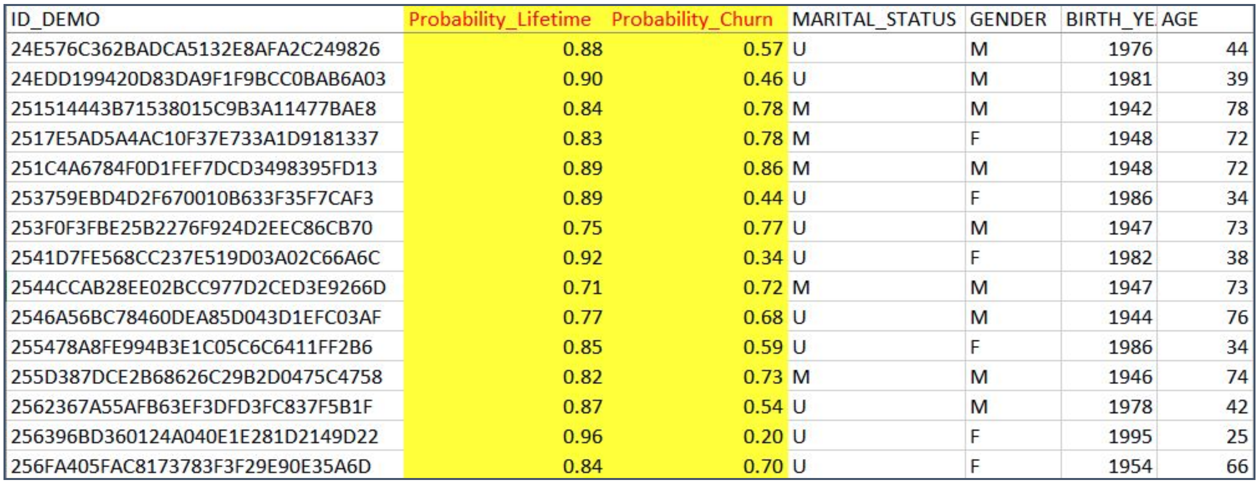
6.4 Churn Analysis

The objective of the churn model is to identify which annual members are unlikely to leave. This training data is the same as what we have for the lifetime conversion model. There are around 12K annual members in 2018, we are predicting their membership status next year, to see if they would convert or not. Out of 12K annual members, 3196 of them became lifetime members. Our model predicts 2227 of them correctly, which gives us 69% recall.

From the churn model, we have a probability spreadsheet where you could find the corresponding probability of not renewing a member for annual members. And we also identify some leading indicators that associate with the membership status change.

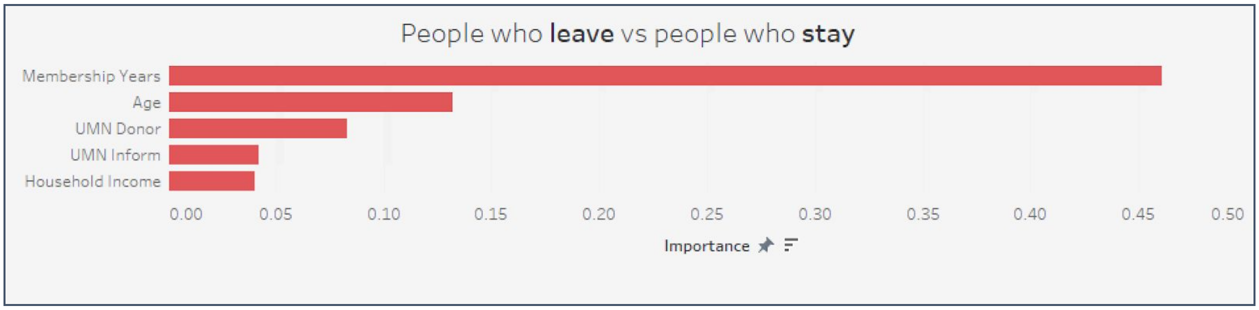
Probability of Churn

The model will return a probability for each annual member not renewing membership.



ID_DEMO	Probability_Lifetime	Probability_Churn	MARITAL_STATUS	GENDER	BIRTH_YE	AGE
24E576C362BADCA5132E8AFA2C249826	0.88	0.57	U	M	1976	44
24EDD199420D83DA9F1F9BCC0BAB6A03	0.90	0.46	U	M	1981	39
251514443B71538015C9B3A11477BAE8	0.84	0.78	M	M	1942	78
2517E5AD5A4AC10F37E733A1D9181337	0.83	0.78	M	F	1948	72
251C4A6784F0D1FEF7DCD3498395FD13	0.89	0.86	M	M	1948	72
253759EBD4D2F670010B633F35F7CAF3	0.89	0.44	U	F	1986	34
253F0F3FBE25B2276F924D2EEC86CB70	0.75	0.77	U	M	1947	73
2541D7FE568CC237E519D03A02C66A6C	0.92	0.34	U	F	1982	38
2544CCAB28EE02BCC977D2CED3E9266D	0.71	0.72	M	M	1947	73
2546A56BC78460DEA85D043D1EFC03AF	0.77	0.68	U	M	1944	76
255478A8FE994B3E1C05C6C6411FF2B6	0.85	0.59	U	F	1986	34
255D387DCE2B68626C29B2D0475C4758	0.82	0.73	M	M	1946	74
2562367A55AFB63EF3DFD3FC837F5B1F	0.87	0.54	U	M	1978	42
256396BD360124A040E1E281D2149D22	0.96	0.20	U	F	1995	25
256FA405FAC8173783F3F29E90E35A6D	0.84	0.70	U	F	1954	66

Leading Indicators



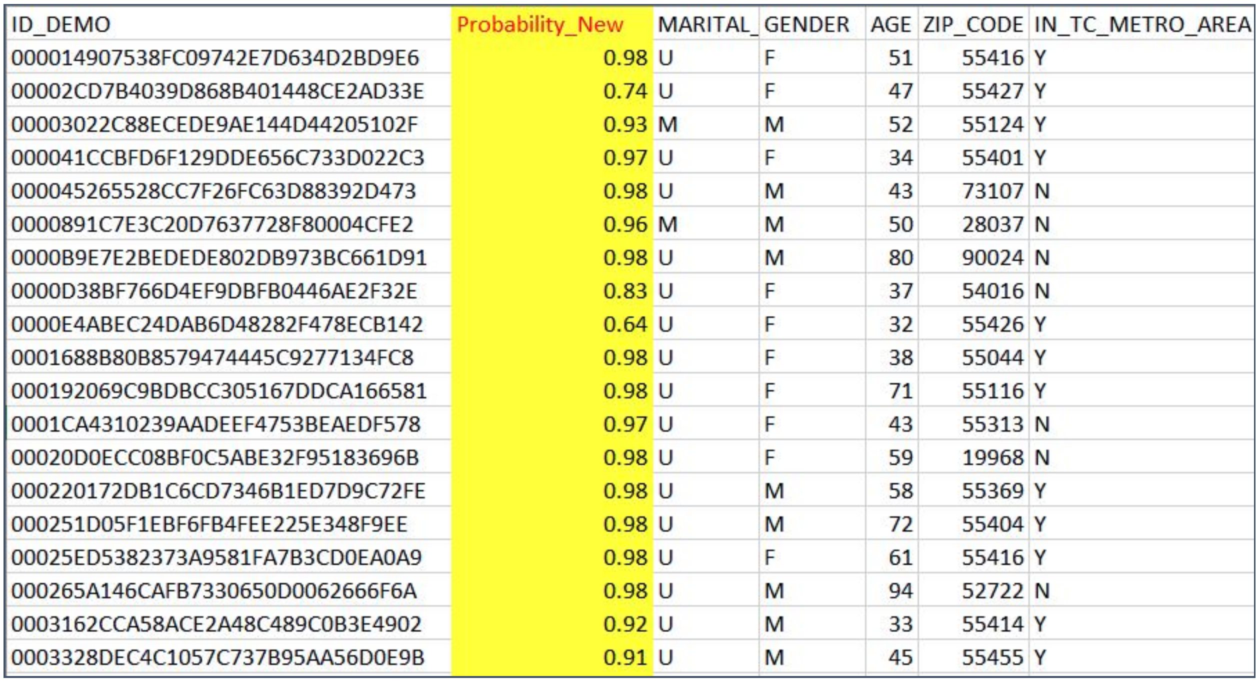
6.5 New Member Acquisition

The objective of the new member acquisition model is to identify which non-members are likely to become members. There are around 400K annual members in 2018, we are predicting their membership status next year, to see if they would convert or not. Out of 400K annual members, 2800 of them became lifetime members. Our model predicts 1600 of them correctly, which gives us 57% recall.

From the new member model, we have a probability spreadsheet where you could find the corresponding probability of becoming a member for non-members. And we also identify some leading indicators that associate with the membership status change.

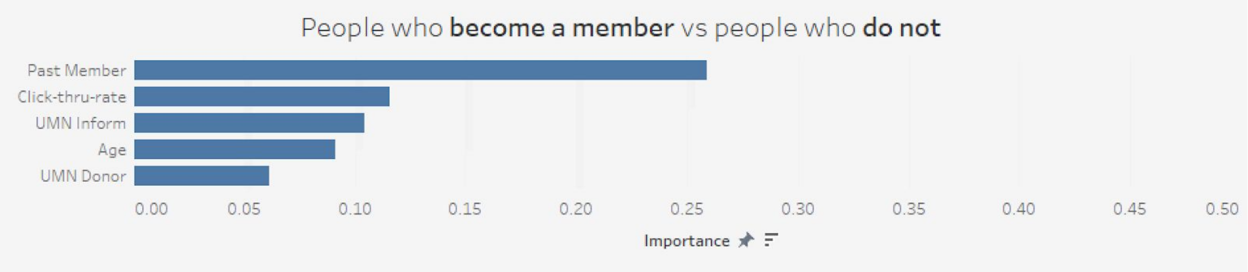
Probability of new member acquisition

The model will return a probability for each annual member becoming a new member.



ID_DEMO	Probability_New	MARITAL_	GENDER	AGE	ZIP_CODE	IN_TC_METRO_AREA
000014907538FC09742E7D634D2BD9E6	0.98	U	F	51	55416	Y
00002CD7B4039D868B401448CE2AD33E	0.74	U	F	47	55427	Y
00003022C88ECEDE9AE144D44205102F	0.93	M	M	52	55124	Y
000041CCBFD6F129DDE656C733D022C3	0.97	U	F	34	55401	Y
000045265528CC7F26FC63D88392D473	0.98	U	M	43	73107	N
0000891C7E3C20D7637728F80004CFE2	0.96	M	M	50	28037	N
0000B9E7E2BEDEDE802DB973BC661D91	0.98	U	M	80	90024	N
0000D38BF766D4EF9DBFB0446AE2F32E	0.83	U	F	37	54016	N
0000E4ABEC24DAB6D48282F478ECB142	0.64	U	F	32	55426	Y
0001688B80B8579474445C9277134FC8	0.98	U	F	38	55044	Y
000192069C9BDBCC305167DDCA166581	0.98	U	F	71	55116	Y
0001CA4310239AADEEF4753BEAEDF578	0.97	U	F	43	55313	N
00020D0ECC08BF0C5ABE32F95183696B	0.98	U	F	59	19968	N
000220172DB1C6CD7346B1ED7D9C72FE	0.98	U	M	58	55369	Y
000251D05F1EBF6FB4FEE225E348F9EE	0.98	U	M	72	55404	Y
00025ED5382373A9581FA7B3CD0EA0A9	0.98	U	F	61	55416	Y
000265A146CAFB7330650D0062666F6A	0.98	U	M	94	52722	N
0003162CCA58ACE2A48C489C0B3E4902	0.92	U	M	33	55414	Y
0003328DEC4C1057C737B95AA56D0E9B	0.91	U	M	45	55455	Y

Leading Indicators



7 Clustering

In this analysis, we used clustering to identify the different demographic and behavioral patterns among current members. This will allow UMAA to understand how to best contact UMN alumni for membership upgrades, renewals and new membership purchases. Contacting alumni with content that best fits their consumption habits will maximize engagement and likelihood of purchase.

7.1 Demographic

7.1.1 Data Preparation

Selected demographic features of 25,690 current members were used to determine the different patterns within the current members. GENDER, AGE, HOUSEHOLD_INCOME, IN_TC_METRO_AREA were the demographic information that was used in the cluster. In addition to this, the average annual engagement score was used from the ENGAGEMENT_CLEANED file. The score with NA values was replaced with 0. Categorical data was encoded with numeric values to convert demographic information into numeric data. This numeric data was then subjected to standardization to convert all variables into a comparable scale.

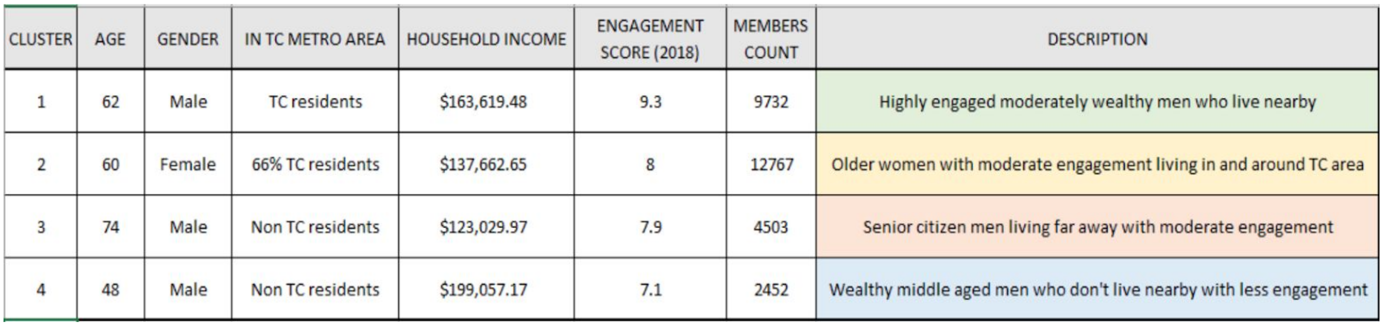
7.1.2 Analysis

Once the features were transformed, correlation plot was plotted to ensure that no two variables are strongly correlated. The clustering method used for this analysis is k-means clustering for its simplistic nature. Elbow plot was used to validate the optimal number of clusters. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE). (https://bl.ocks.org/rpgove/0060ff3b656618e9136b) When an elbow curve was plotted for the dataset, the plot indicated 4 as the optimal number of clusters.

7.1.3 Results

When the k-means clustering algorithm was executed, we obtained 4 groups and defined them as follows:

1. Highly engaged moderately wealthy men who are TC residents
2. Moderately engaged older women who live in and around TC metro area
3. Moderately engaged Senior citizen men who live away from TC metro area
4. Less engaged wealthy middle aged men who don’t live in the TC metro area

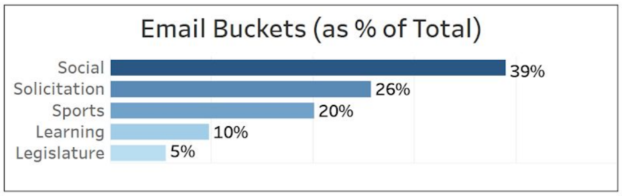


CLUSTER	AGE	GENDER	IN TC METRO AREA	HOUSEHOLD INCOME	ENGAGEMENT SCORE (2018)	MEMBERS COUNT	DESCRIPTION
1	62	Male	TC residents	$163,619.48	9.3	9732	Highly engaged moderately wealthy men who live nearby
2	60	Female	66% TC residents	$137,662.65	8	12767	Older women with moderate engagement living in and around TC area
3	74	Male	Non TC residents	$123,029.97	7.9	4503	Senior citizen men living far away with moderate engagement
4	48	Male	Non TC residents	$199,057.17	7.1	2452	Wealthy middle aged men who don't live nearby with less engagement

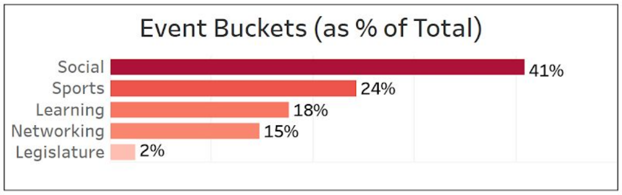
7.2 Behavioral

7.2.1 Data Preparation (code file - Prepping for Clustering.R)

Note that we first created manual categories for each email / event. Thorough documentation has been created for these categories and sent to UMAA already. Here, we are showing the 5 categories of emails & events that we created (not including “Solicitation” from emails).



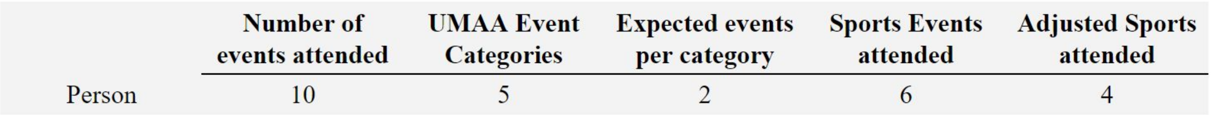
Email Buckets (as % of Total)



Event Buckets (as % of Total)

The following image walks through an example of how we calculated adjusted scores to go into our clusters. We are looking to identify the type of content that someone likes above and beyond the others. We do this by comparing someone’s behavior on average vs their behavior when it comes to each specific type of content.

In this example, we would have expected the person to attend 2 events per category if we know that they attended 10 events in total. They actually attended 6 sports events, which gives them a positive score of 4 for sports.



	Number of events attended	UMAA Event Categories	Expected events per category	Sports Events attended	Adjusted Sports attended
Person	10	5	2	6	4

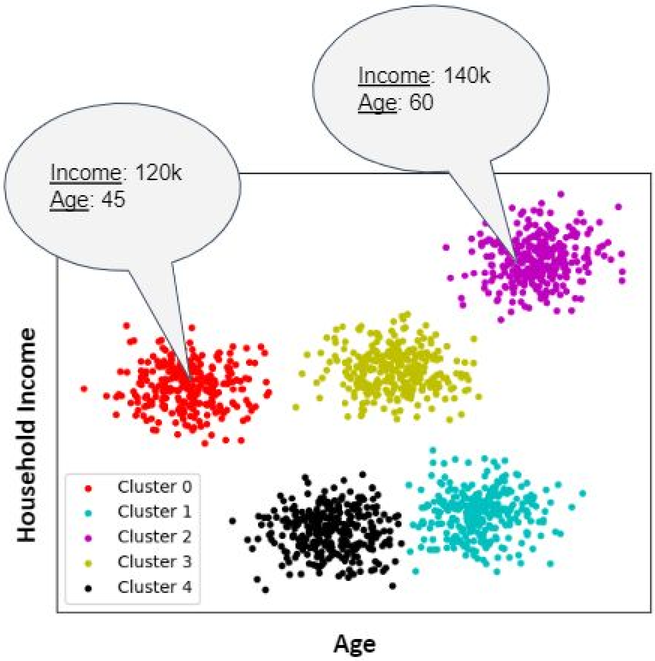
7.2.2 Analysis (code file - Clustering (Behavioral).R)

After obtaining an events score and an email score for each category, we added the two together to create one category per type of content. The resulting 5 columns were

- Sports

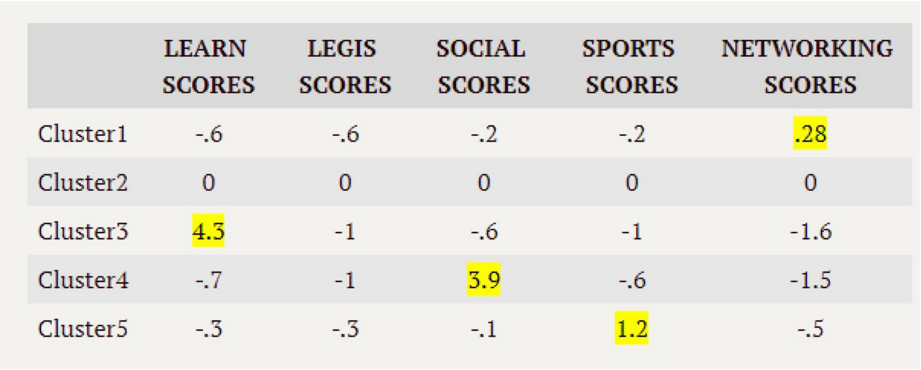
- Social
- Networking
- Learning
- Legislature

We used an algorithm called “k-means” in order to separate our members into 4 distinct groups. An intuitive plot of how this works is shown below.



7.2.3 Results

We ended up with 5 clusters. 4 of them were meaningful, and the 5th ended up capturing the rest of the people.



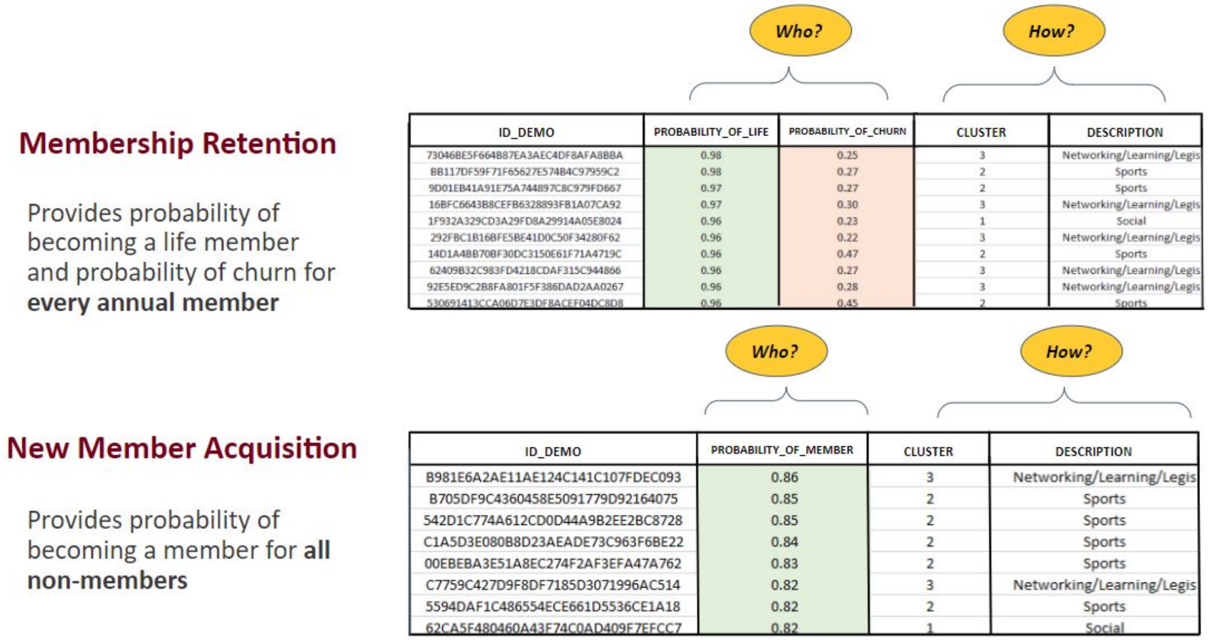
	LEARN SCORES	LEGIS SCORES	SOCIAL SCORES	SPORTS SCORES	NETWORKING SCORES
Cluster1	-.6	-.6	-.2	-.2	.28
Cluster2	0	0	0	0	0
Cluster3	4.3	-1	-.6	-1	-1.6
Cluster4	-.7	-1	3.9	-.6	-1.5
Cluster5	-.3	-.3	-.1	1.2	-.5

The people from the top 4 clusters pose a great opportunity for personalized messaging / incentives directed towards what they actually care about.

8 Business Value

Predictive modeling (*Who*) was used to determine who UMAA should be targeting when reaching out to alumni for membership upgrades, renewals and new membership purchases. Clustering analysis (*How*) will assist UMAA in understanding how to best contact UMN alumni for membership upgrades, renewals and new membership purchases. Contacting alumni with content that best fits their consumption habits will maximize engagement and likelihood of purchase.

Connecting these two pieces, UMAA will be able filter members based on their probability of lifetime membership purchase, churn or new member purchase and connect with them using personalized content to drive membership.



Membership Retention

Provides probability of becoming a life member and probability of churn for **every annual member**

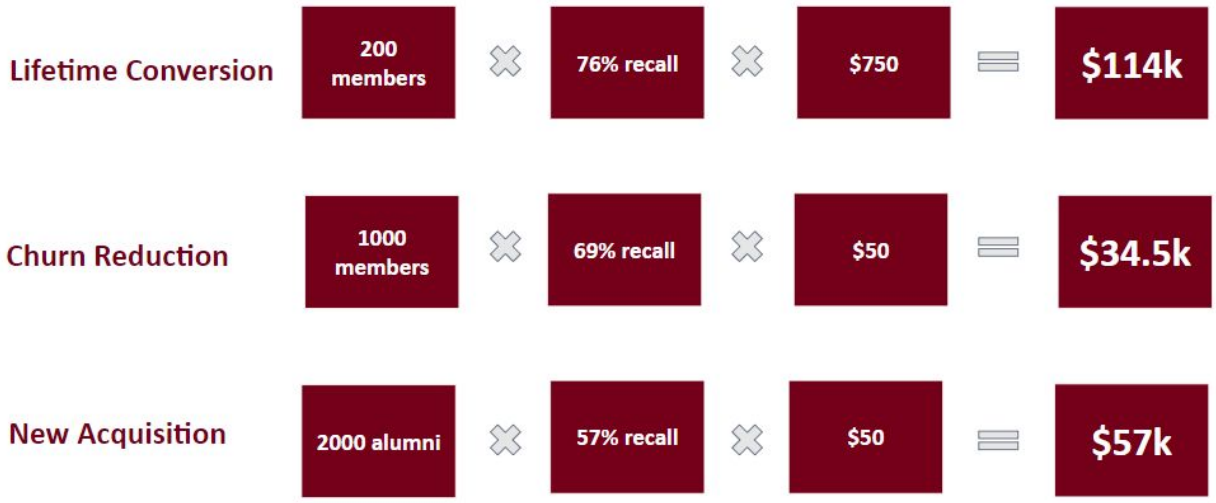
ID_DEMO	PROBABILITY_OF_LIFE	PROBABILITY_OF_CHURN	CLUSTER	DESCRIPTION
73046BE5F664B87EA3AEC4DF8AFA8BBA	0.98	0.25	3	Networking/Learning/Legis
BB117DF59F71F65627E574B4C97959C2	0.98	0.27	2	Sports
9D01EB41A91E75A744897C8C979FD667	0.97	0.27	2	Sports
16BFC6643B8CEFB6328893FB1A07CA92	0.97	0.30	3	Networking/Learning/Legis
1F932A329CD3A29FD8A29914A05E8024	0.96	0.23	1	Social
292FBC1B16BFE5BE41D0C50F34280F62	0.96	0.22	3	Networking/Learning/Legis
14D1A4BB70BF30DC3150E61F71A4719C	0.96	0.47	2	Sports
62409B32C983FD4218CDAF315C944866	0.96	0.27	3	Networking/Learning/Legis
92E5ED9C2B8FA801F5F386DAD2AA0267	0.96	0.28	3	Networking/Learning/Legis
530691413CCA06D7E3DF8ACEF04DC8D8	0.96	0.45	2	Sports

New Member Acquisition

Provides probability of becoming a member for **all non-members**

ID_DEMO	PROBABILITY_OF_MEMBER	CLUSTER	DESCRIPTION
B981E6A2AE11AE124C141C107FDEC093	0.86	3	Networking/Learning/Legis
B705DF9C4360458E5091779D92164075	0.85	2	Sports
542D1C774A612CD0D44A9B2EE2BC8728	0.85	2	Sports
C1A5D3E080B8D23AEADE73C963F6BE22	0.84	2	Sports
00EBEBA3E51A8EC274F2AF3EFA47A762	0.83	2	Sports
C7759C427D9F8DF7185D3071996AC514	0.82	3	Networking/Learning/Legis
5594DAF1C486554ECE661D5536CE1A18	0.82	2	Sports
62CA5F480460A43F74C0AD409F7EFCC7	0.82	1	Social

The below graphic illustrates a scenario in which UMAA targets the top 10% most likely to convert to Lifetime membership, churn or become a new member. Accounting for our model performance, we estimate a savings of $205,000 by by leveraging our proposed solution.



Lifetime Conversion 200 members × 76% recall × $750 = $114k

Churn Reduction 1000 members × 69% recall × $50 = $34.5k

New Acquisition 2000 alumni × 57% recall × $50 = $57k

9 Tableau Dashboard

Tableau dashboards were created to help UMAA measure how each content cluster is responding to different types of messages over time.

Please see additional “Tableau Guide.pdf” for comprehensive documentation on these Tableau dashboards.

10 Limitations

The email and event categories were manually created, so UMAA will need to add in new categories manually where they see fit. Additionally, we created categories based on pseudo-subject lines of each email & event based on what we believed was correct (and checked with UMAA). UMAA has also been moving towards sending 1 single weekly email with all categories included, which our algorithm currently will not be able to implement on a per-category basis.

In addition, in our predictive analysis, we had a difficult time dealing with imbalanced data of members and non-members. Resampling methodology was used to maximize model performance despite this, however the severity of the imbalance may have affected results.

11 Future Analysis & Adjustments

11.1 Future Analysis Opportunities

During the project, the CAL team was made aware that emails were increasingly moving towards the “newsletter” format, including multiple content types in the same message. While this does limit the use of our content cluster analysis, we believe that the analysis is still valuable to UMAA and should be utilized when tracking emails that do relate to a single content category. Future analysis could be performed on how alumni interact with different sections of the “newsletter” style emails.

UMAA specific engagement scores were collected beginning in 2019. Due to the limited amount of data, we were not able to draw meaningful conclusions from these engagement scores and instead focused on UMN engagement and sub engagement scores. Given the nature of this data, and it’s direct relevance to UMAA, we suggest that analysis is performed on these UMAA engagement scores in the future.

Finally, the CAL team has recommended that UMAA begin connecting an appeal code to each membership solicitation email. This will enable UMAA to attribute specific emails to membership purchases. Future analysis can then be performed on these emails in conjunction with their appeal codes to better understand the efficacy of membership solicitations.

11.2 Database Adjustments

We suggest adding 2 new tables. One should be **Events_detail** and the other should be **Emails_detail**. This will allow us to reduce the size of the existing tables for **emails_individual** and **events_individual.**

1) **Emails_detail:** This table provides 1 line per email CODE_PATH and no information about ID_DEMO. Below columns need to be populated.
    a) CODE_PATH to link to full table
    b) NBR_GROUP to link events. This assumes that each email refers to 1 event at most though
    c) APPEAL_CODE to link conversions
    d) Major category - Sports, Learning, Solicitation, etc
    e) Minor category - Football, Webinar, etc
    f) DESC_PATH
    g) Group being targeted - UMAA members, everyone, etc
    h) Region/area sent to
    i) Date sent

Creating above table allows to reduce the size of **Emails_individual** table by providing 1 line per person who received the email and this table contains the below columns

a) ID_DEMO to track the person email was sent to
b) Status (clicked, opened, etc)

2) **Events_detail**: There should be a new table for events with 1 line per event and below columns need to be populated.

a) NBR_GROUP to link full table
b) Major category - Sports, Learning, Social, etc
c) Minor category - Football, Fundraiser, etc
d) NAME_GROUP for full detail
e) Group being targeted - UMAA members, everyone, etc
f) Location of event
g) Date being held

Creating the above table allows to reduce the size of the **Event_individual** table by providing 1 line per person who attended the event and this table contains ID_DEMO for the person it was sent to.

Membership table:

The "Membership_level" column should be broken up. There should be a column for

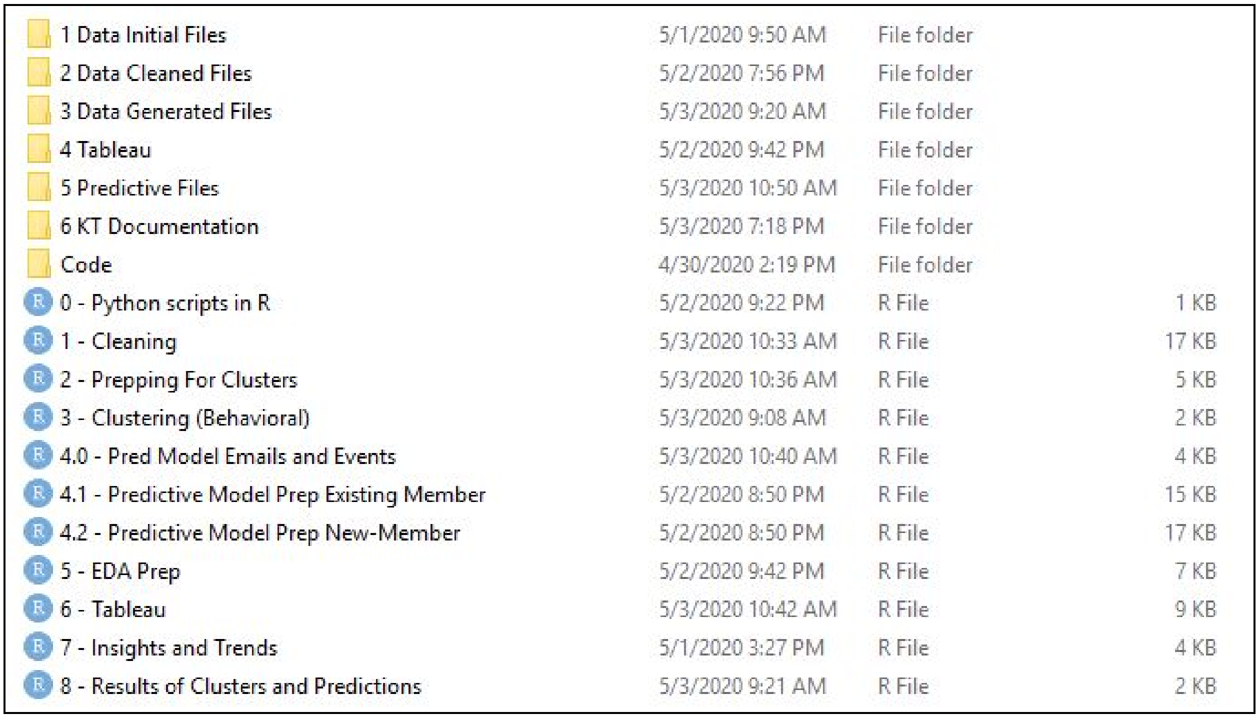
a) Purchase type - Annual or Lifetime
b) Comp or not
c) Person type - Student, Recent Grad, Employee, etc
d) Being Bought For - Joint, Single, Joint Secondary, etc
e) Number of Years - 1, 3, 5, etc

Engagement table:

The scores that are being generated for UMAA specific (ie the ones with FY1919) should either be additional columns (specifying UMN or UMAA score) or should be in a separate table altogether, but not additional rows.

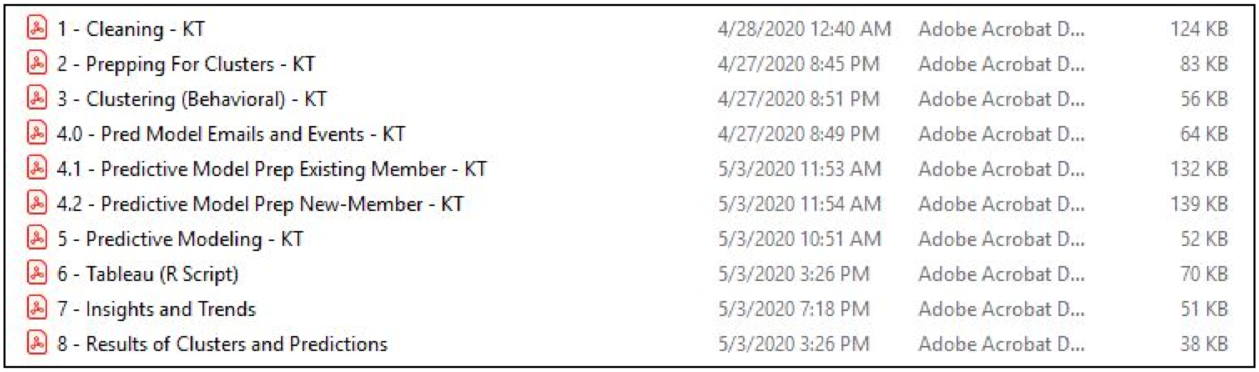
12 Additional Materials

A comprehensive folder structure, including scripts and detailed documentation has been provided to UMAA via BOX. Below is a screenshot of this folder structure, scripts and documentation. Additionally, video documentation of scripts and recordings of knowledge transfer sessions have been provided to UMAA via Google Drive.



1 Data Initial Files	5/1/2020 9:50 AM	File folder	
2 Data Cleaned Files	5/2/2020 7:56 PM	File folder	
3 Data Generated Files	5/3/2020 9:20 AM	File folder	
4 Tableau	5/2/2020 9:42 PM	File folder	
5 Predictive Files	5/3/2020 10:50 AM	File folder	
6 KT Documentation	5/3/2020 7:18 PM	File folder	
Code	4/30/2020 2:19 PM	File folder	
0 - Python scripts in R	5/2/2020 9:22 PM	R File	1 KB
1 - Cleaning	5/3/2020 10:33 AM	R File	17 KB
2 - Prepping For Clusters	5/3/2020 10:36 AM	R File	5 KB
3 - Clustering (Behavioral)	5/3/2020 9:08 AM	R File	2 KB
4.0 - Pred Model Emails and Events	5/3/2020 10:40 AM	R File	4 KB
4.1 - Predictive Model Prep Existing Member	5/2/2020 8:50 PM	R File	15 KB
4.2 - Predictive Model Prep New-Member	5/2/2020 8:50 PM	R File	17 KB
5 - EDA Prep	5/2/2020 9:42 PM	R File	7 KB
6 - Tableau	5/3/2020 10:42 AM	R File	9 KB
7 - Insights and Trends	5/1/2020 3:27 PM	R File	4 KB
8 - Results of Clusters and Predictions	5/3/2020 9:21 AM	R File	2 KB

6 - KT Documentation:



1 - Cleaning - KT	4/28/2020 12:40 AM	Adobe Acrobat D...	124 KB
2 - Prepping For Clusters - KT	4/27/2020 8:45 PM	Adobe Acrobat D...	83 KB
3 - Clustering (Behavioral) - KT	4/27/2020 8:51 PM	Adobe Acrobat D...	56 KB
4.0 - Pred Model Emails and Events - KT	4/27/2020 8:49 PM	Adobe Acrobat D...	64 KB
4.1 - Predictive Model Prep Existing Member - KT	5/3/2020 11:53 AM	Adobe Acrobat D...	132 KB
4.2 - Predictive Model Prep New-Member - KT	5/3/2020 11:54 AM	Adobe Acrobat D...	139 KB
5 - Predictive Modeling - KT	5/3/2020 10:51 AM	Adobe Acrobat D...	52 KB
6 - Tableau (R Script)	5/3/2020 3:26 PM	Adobe Acrobat D...	70 KB
7 - Insights and Trends	5/3/2020 7:18 PM	Adobe Acrobat D...	51 KB
8 - Results of Clusters and Predictions	5/3/2020 3:26 PM	Adobe Acrobat D...	38 KB