UMDAA Analysis Report

Team member: Aditya Jadeja, Vishal Vindyala, Ching-Yun Chen, Shu-Ping Chen

1. Mission Statements

(1) Identify variables that are correlated to our desired outcomes of higher event attendance of first-time attendees and major gift prospect attendees.
(2) Hope to optimize both new and existing events in order to improve our strategy to engage alumni in these two groups.
(3) Understand what types of events attract the largest number of first- time attendees and major gift prospects so that we can use this information in planning future events.

2. Mission Objectives

(1) Based on the variables provided in the dataset, how can we get more first-time attendees to attend our events?
(2) Based on the variables provided in the dataset, how can we get more major gift prospects to attend our events?

3. Data Description

(1) Data resource: Alumni Association at University of Maryland
(2) Time period: 07/01/2013-11/30/2019 (fiscal year: July 1 - June 30)
(3) Description of variables:

In order to understand the data type and description of each variable, we created Table 1 to make a summary. Not only using the existing variables in the dataset, we also created a new variable called "year" after we combined the data in each year. In our data, there are nine categorical variables and six numerical variables in total.

Table1. Information of variables

Variable	Data Type	Description
Event Name	object	Name of event
Activity Code	object	5-digit code used to track alumni event attendance
Activity Description	object	Details of the activity
Location Code	object	4-digit code used with the activity code to denote the region and location an event takes place
Location Description	object	Details of the location
Group Code	object	3-digit code used with the activity code to denote the event's intended purpose and audience.
Group Description	object	Details of the group
Event Date	datetime64[ns]	Date of the event
Participated	int64	Number of people who participated the event
Average Age	int64	Average age of the people in the event
First Time Attendees	int64	Number of people who have never previously attended an event as an alumnus/a according to our database
Percentage First Time Attendees	float64	Percentage of the first-time attendees
Major Prospects	int64	Number of alumnus/a or friends of the university who has the ability to make a financial gift of $50,000 or more to the university
Percentage Major Prospects	float64	Percentage of the Major Prospects
year	object	Year period of the event

4. Exploratory Data Analysis

(1) Statistical information of the variables

For categorical variables, we could know some information about these variables, including the number of observations, the number of unique variables, the observation with the highest frequency, and its frequency through Table 2.

Table 2. Summary of categorical variables

	Event Name	Activity Code	Activity Description	Location Code	Location Description	Group Code	Group Description	year
count	622	622	622	622	622	622	622	622
unique	428	341	346	45	111	49	88	7
top	Baltimore Terps Game Watch	PEABG	CP AA-DC Terps Game Watch	PDON	CP DMV- On Campus	PS9	CP Social-General	2018
freq	14	14	12	116	52	176	72	219

For numerical variables, we could understand their number of observations, mean, standard deviation, minimum, maximum, and some percentiles through Table 3.

Table 3. Summary of numerical variables

	Participated	Average Age	First Time Attendees	Percentage First Time Attendees	Major Prospects	Percentage Major Prospect
count	622.000000	622.000000	622.000000	622.000000	622.000000	622.000000
mean	44.803859	40.117363	13.456592	0.276282	5.966238	0.102214
std	93.165049	9.741459	41.103936	0.242273	14.123466	0.131444
min	1.000000	19.000000	0.000000	0.000000	0.000000	0.000000
25%	10.000000	33.000000	1.000000	0.068523	0.000000	0.000000
50%	20.000000	40.000000	4.000000	0.237327	1.000000	0.058824
75%	44.750000	46.000000	11.000000	0.444444	5.000000	0.166667
max	1657.000000	75.000000	702.000000	1.000000	131.000000	0.818182

Hence, we concluded that not only realize the statistical information of each variable, but we could also know that there aren't any missing values in our data because each variable had the same number of observations.

(2) Changes of the percentage of first-time attendees and the major gift prospects over years

For the change of the first-time attendees, through Figure 1, we found that the percentage started from 0%, and went to the highest values in 2014 to 2015. However, in general, the percentage of first-time attendees had a downward trend as years passed.
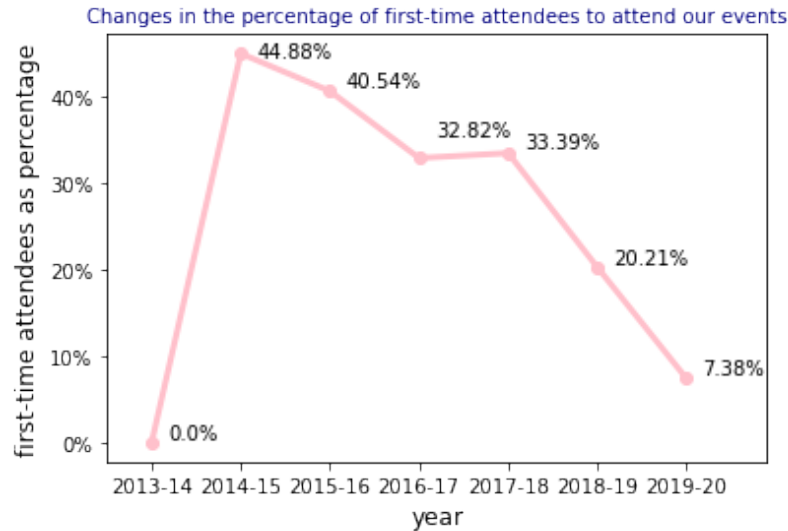
Figure 1. Changes in the percentage of first-time attendees to attend our events

For the change of the major gift prospects, although we found a little upwards from 2016 to 2017, in general, we could find a downwards trend year by year through Figure 2.
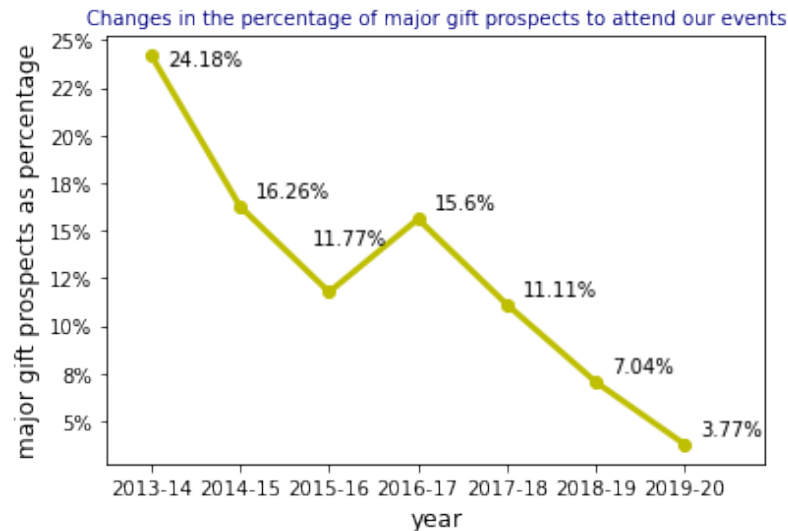

Figure 2. Changes in the percentage of major gifts prospects to attend our events

Consequently, we concluded that both percentages of the first-time attendees and the major gift prospects had a downwards trend in general when the year passes. Hence, we suggested that maybe there are some factors that would make these two percentages move in the same direction.

5. Correlation Plot

We made Figure 3 based on the numerical variables. We also could find that the variable, Participated, was highly correlated with First Time Attendees and Major Prospects. Consequently, we put more emphasis on "Participated" when we discussed the factors which had a great influence on First Time Attendees and Major Prospects.
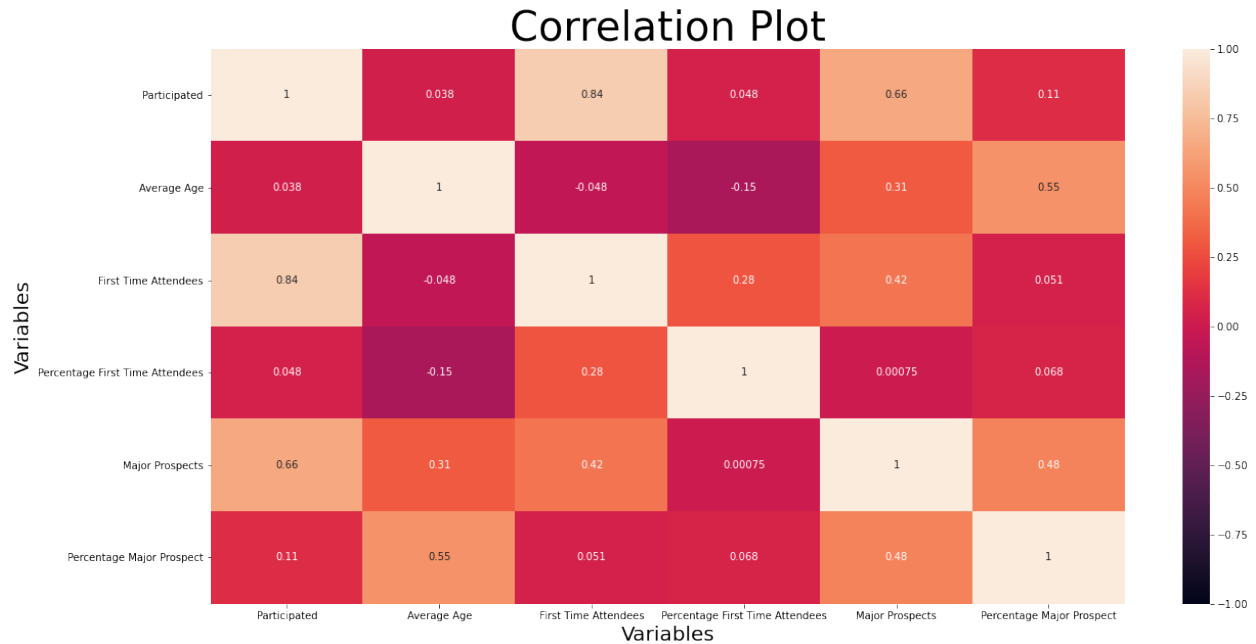
Figure 3. Correlation plot between numerical variables

6. Regression Model (First Time Attendees)

In this part, we conducted regression analysis in order to find out which variables or which combination of variables affects the dependent variable (First Time Attendees) the most. Thus, we developed seven models, each with a different combination of variables, so that we could compare the models based on their adjusted R-squared value and their mean squared error. The combination of variables for each of the seven models is as follows:

Model 1: Average Age, Participated
Model 2: Average Age, Participated, year, Major Prospects
Model 3: Participated, year, Major Prospects
Model 4: Participated, Major Prospects
Model 5: Participated, year, Major Prospects, Group Code, Location Code, Activity Code
Model 6: Participated, year, Major Prospects, Location Code, Activity Code
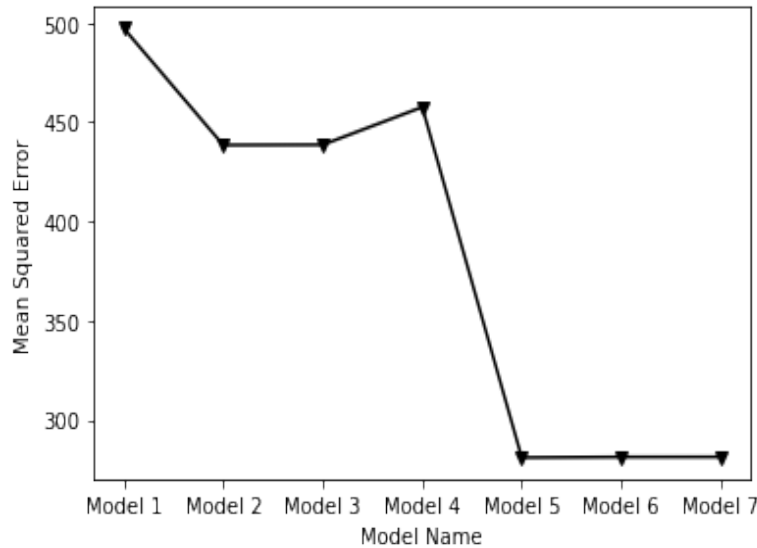Model 7: Participated, year, Major Prospects, Activity Code

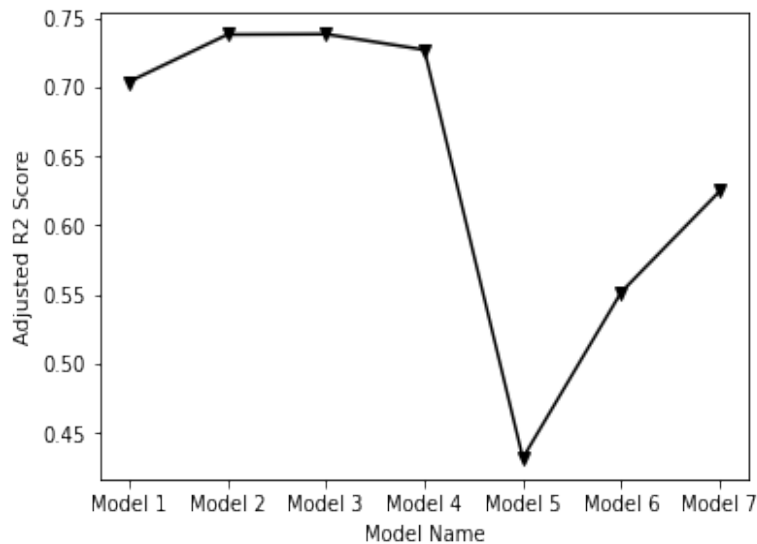**Figure 4.** Plot of MSE for different models

Figure 5. Plot of adjusted R-squared value for different models

Through Figure 4 and Figure 5, we could know the difference of the Mean Squared Error and Adjusted R-squared value between each model, respectively. Thus, we selected Model 3 as it has the best combination of the Mean Squared Error and Adjusted R-squared value.

7. Regression Model (Major Prospects)

In this part, we conducted regression analysis in order to find out which variables or which combination of variables affects the dependent variable (Major Prospects) the most. Thus, we developed seven models, each with a different combination of variables, so that we could compare the models based on their adjusted R-squared value and their mean squared error. The combination of variables for each of the seven models is as follows:

Model 1: Average Age, Participated, year, First Time Attendees

Model 2: Average Age, Participated, year, First Time Attendees, Group Code
Model 3: Average Age, Participated, year, First Time Attendees, Group Code, Location Code
Model 4: Average Age, Participated, year, First Time Attendees, Group Code, Location Code, Activity Code
Model 5: Average Age, Participated, year
Model 6: Average Age, Participated, First Time Attendees
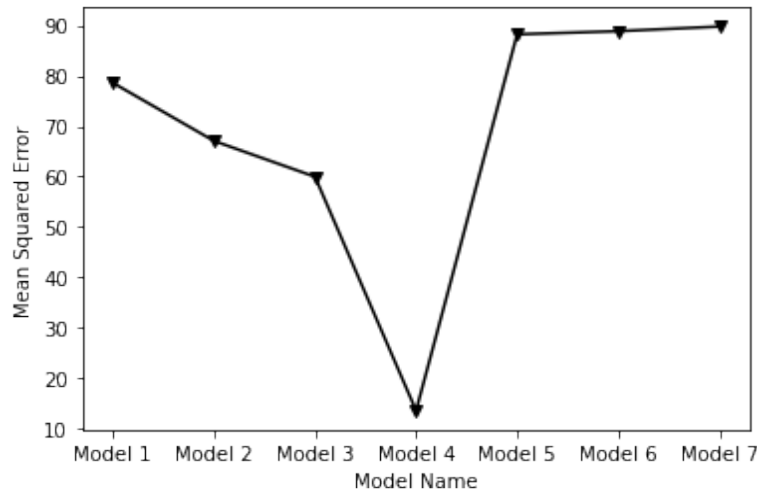Model 7: Participated, year, First Time Attendees

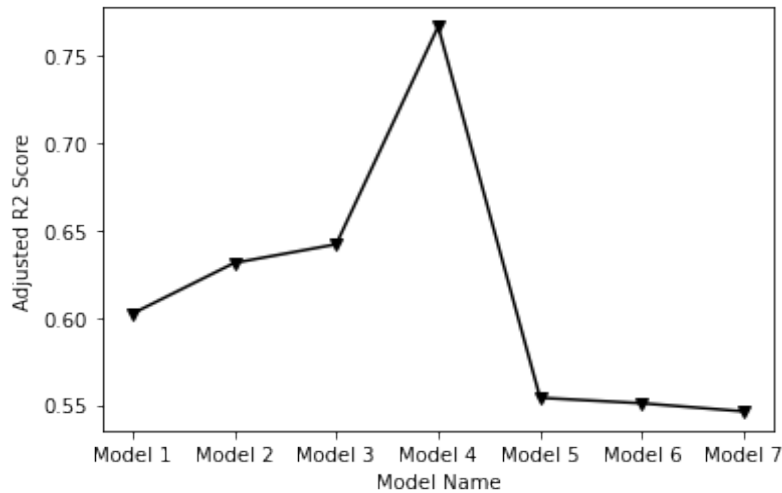**Figure 6.** Plot of MSE for different models

Figure 7. Plot of adjusted R-squared value for different models

Through Figure 6 and Figure 7, we could know the difference of the Mean Squared Error and Adjusted R-squared value between each model, respectively. Thus, we selected Model 4 as it had the lowest Squared Error and highest Adjusted R-squared value.

8. Attraction of events (First Time Attendees)

In this part, we analyzed the attraction of each Activity Code separately. For first-time attendees, we got the order of the average number of first-time attendees grouped by each Activity Code, to find the top Activity which contributed the largest number of first-time attendees. The results are shown below in Figure 8.

Figure 8. Difference of position between number and percentage for first time attendees

9. Attraction of events (Major Prospects)

In this part, we analyzed the attraction of each Activity Code separately. For Major Prospects, we got the order of the average number of Major Prospects grouped by each Activity Code to find the top Activity which contributed the largest number of Major Prospects. The results are shown below in Figure 9.

Figure 9. Difference of position between number and percentage for major prospects

10. Attraction of Group and Location

In this part, we analyzed the attraction of each Group Code and each Location Code separately. For first time attendees, we got the order of the average number of first-time attendees grouped by each group code and each location code, respectively, to find the top 1 group and top 1 location which contributed the largest number of first-time attendees. We also applied the same method to find the top 1 group and top 1 location for the major prospects. The results are shown below in Table 4.

Table 4. Summary of attraction of Group and Location

	Group Code	Group Description	Location Code	Location Description
First Time Attendees	PSS	CP Social-Students	PSAU	CP Southeast-Austin
Major Prospects	PH9	CP Stewardship-General	PNNA	CP Northeast-General

11. Conclusion and Recommendations

In conclusion, our goal is to maximize the number of first-time attendees and major prospects in planning future events. First, in order to increase the number of first-time attendees, our strategy is to hold more events like Virtual Book Club, for the group of Social-Students, or the location of Southeast-Austin. We can also optimize the overall participants in every event to get a higher number of first-time attendees. Second, in order to increase the number of major prospects, we consider holding more events like Virtual Book Club as well, for the group of Stewardship-General, or at the location of Northeast-General. Increasing the overall number of participants and first-time attendees is also beneficial for getting more major prospects. Moreover, the average age is also an important factor in this case, we could find that older participants are more likely to make financial donations. We use Table 5 to display the decision we made to increase the number of first-time attendees and major prospects.

Table 5. Decision making to achieve the goals

Goal	First time attendees ↑	Major prospects ↑
Decision	Activity: CP AA-Virtual Book Club Group: CP Social-Students Location: CP Southeast-Austin Participated	Activity: CP AA-Virtual Book Club Group: CP Stewardship-General Location: CP Northeast-General Participated First Time Attendees Average Age

12. Future Work

(1) We can try to explore the factors that motivate our alumni to participate in the events which could contribute to higher numbers of participants.

(2) In this report, we analyzed the factors that affect the number of First Time Attendees and Major Prospects separately. However, we hope to find a combination of all variables that can increase both First Time Attendees and Major Prospects.

(3) If we have more time to do the feature extraction, we can reduce the number of categorical variables by classifying them into groups. It helps us to simplify the model, even maintain the similar ability of interpretation of the original model.

13. Testing and evaluation

We used Jupyter Notebook to test and evaluate the methods we used in the report.