Final Project Report - Optimizing Merage’s Online Advertising - SMP Data

Team 9: Sujaya Darke, Sammi Chang, Isha Chaudhari, Song Han, Poonam Ahuja

**Executive Summary:**

University graduate divisions fiercely compete against each other to maximize enrollment for their respective programs. In order to maximize enrollment and overall market penetration, UCI Merage graduate school is looking for ways to improve the efficiency and effectiveness of their online advertising. The aim of this report is to assess UCI’s digital analytics data like website data, social media, digital marketing strategy with the focus of increasing the number of student enrollments at UCI’s Specialty Masters Program (SMP). By the end of the project, we hope to provide powerful recommendations to Merage to optimize their advertising efforts and aim for a higher return on their marketing investment.

**Introduction and Purpose:**

Through this analysis, we would like to know which types of applicants are more likely to apply using applicant profile data, which applicant lead sources and platforms are optimal to advertise with in order to increase our conversion rates, and whether we can use application submission data to derive a meaningful conclusion about applicant behavior. We would also like to know whether we can improve our conversions, namely number of clicks and number of application submissions.

Our overall goal is to help Merage improve online advertising strategy on social media platforms for their Specialty Masters Programs and gain more insight on the effectiveness of different lead sources. We would also like to use applicant data to help Merage further develop their campaigns to target a more effective audience segment. Additionally, we would like to provide additional data collection strategies to perform more powerful and effective campaign optimization in the future.

**Dataset Brief:**

We were provided three types of data: campaign data, leads data, and applicant data. The campaign data included engagement and performance data. The leads data included lead source data, applicant stage data (how far into the application process an applicant reached), limited gender data, and date they first visited the site. The applicant data included alternate lead sources, application submit date data, and program and residency data.

**Data Description, Exploratory Data Analysis, and Modeling**

A. Applicants Data

*Dataset Description.*

The applicant data that was given to us consisted of 4354 observations and 7 variables. This dataset includes key information about the applicants and the lead objects that led them to applying. The difference between this dataset and the Leads dataset is that all the students here applied. The variables consisted of:

1. Programs: The Masters programs that Merage offers. This was filtered to only include those in SMP: {FIN, MIE, MPA, MSBA}
2. Application Submit Date- The date that the application submitted their application.
3. First Activity Date – The date that the applicant first visited the site.
4. Lead: {Gmass, GMAT, GRE, Grad Fair, Hobsons, Internal Inquiry Form, Online App, Event Registration qualified List Upload, Unqualified List Upload, Web Inquiry Form, World MBA Tour, You Visit}. The lead object that brought the applicant to the site.
5. Start.Term.And.Year - The start term and year that the applicant applied for.
6. Residency.Status - {Citizen or Permanent Resident, International Applicant} Indicates the residency status of the applicant at time of application.
7. Pardot.First.Referrer - This attribute refers to webpage that the student came from. Most applicants come from the Merage website, so this column was largely useless and so removed from the dataset.

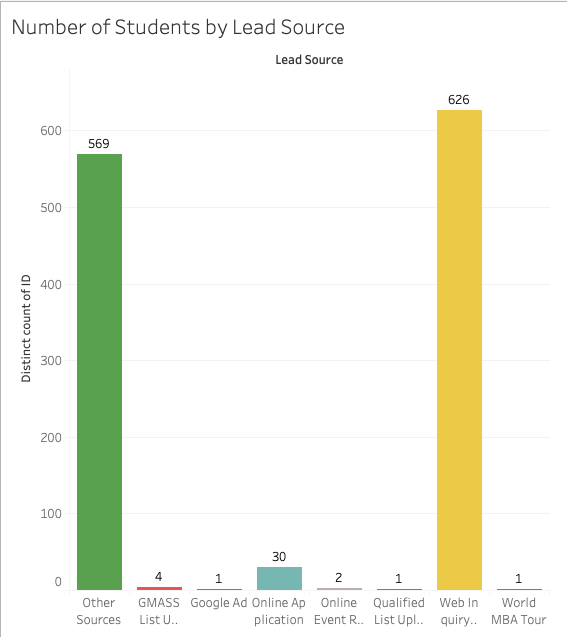
*Data Cleaning and Preparation.*

The data was cleaned and prepared by creating dummy variables and engineering new ones. Firstly, new variables were created for program – one dummy variable for each program in SMP. Then, several dummy variables were included to incorporate Residency status and Lead in our analysis. The term and year was not used.

Two variables were engineered from the two dates given in the original dataset. These are time\_elapsed, which is equal to application submit date – first activity date, and time\_elapsed\_since\_app\_window\_opens, which is equal to application submit date – date application window opened. In other words, time elapsed measures the time from the applicant’s first website visit to their submission date. The latter variable measures the time from the application window opening to the applicant’s submission date.

*Exploratory Data Analysis.*

The below graph shows that amongst all the leads, web inquiry form has the highest numbers, so it is the most useful lead in getting students to apply. Online application only has 30 counts, but it still has a greater count than the others. It is important to note that there were almost half of the students in the data that did not come from any lead source (569 students).

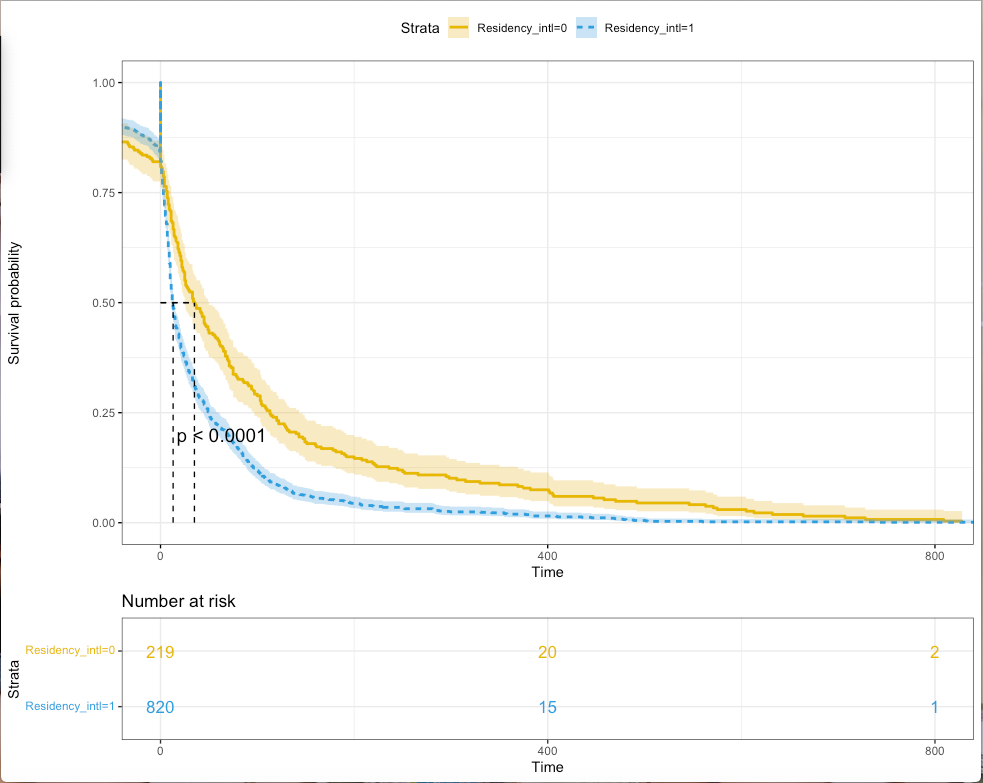


*Modeling.*

The most interesting aspect of this dataset is the calculated variable: time elapsed. We can compare the time elapsed between first activity and application submit date against many variables to bring light to information that will help Merage School increase total application numbers for SMP.

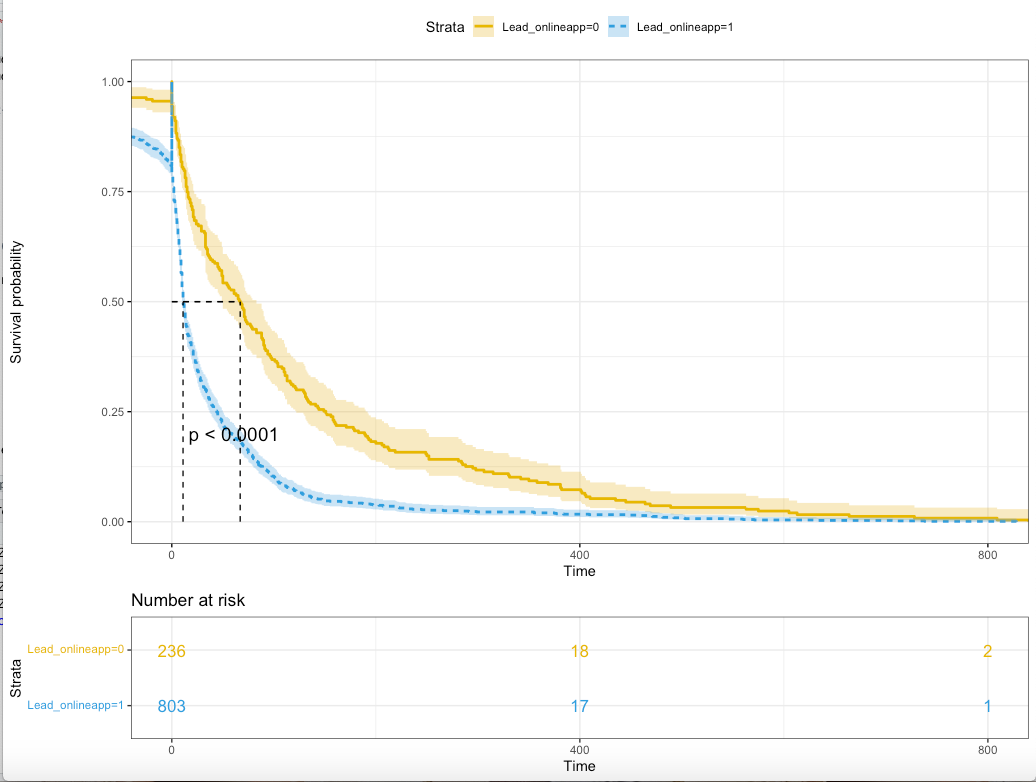
Survival Analysis.

~by Residency Status (International = 1, Citizen = 0)



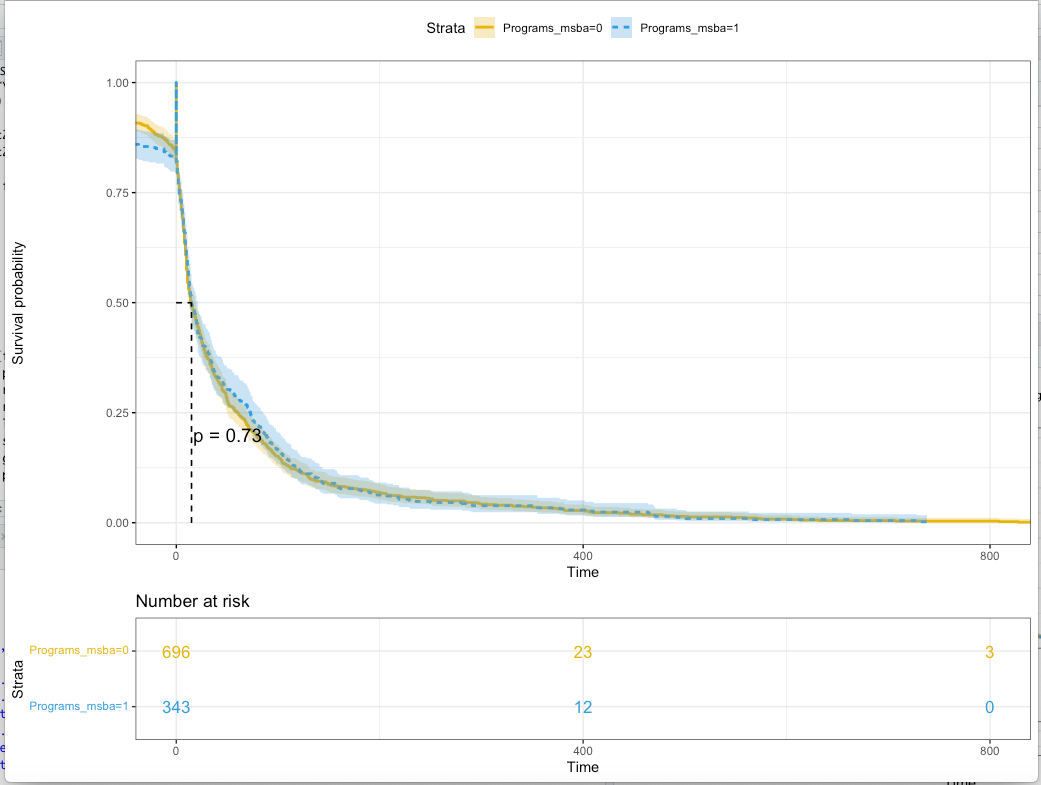
International Students apply to SMP earlier than Residents. Several factors could factor into this: application submission dates are earlier for international students, and they need more time to make decisions due to reasons like visa issues and extensive paperwork. What we can derive from this graph is that domestic students do not follow the same pattern as international students. Residents tend to take their time with decision making or writing their application. With this in mind, we encourage Merage to focus targeted marketing efforts to domestic students 1-2 months later than international students.

~by Lead is Online App (1) or other types of Leads (0)

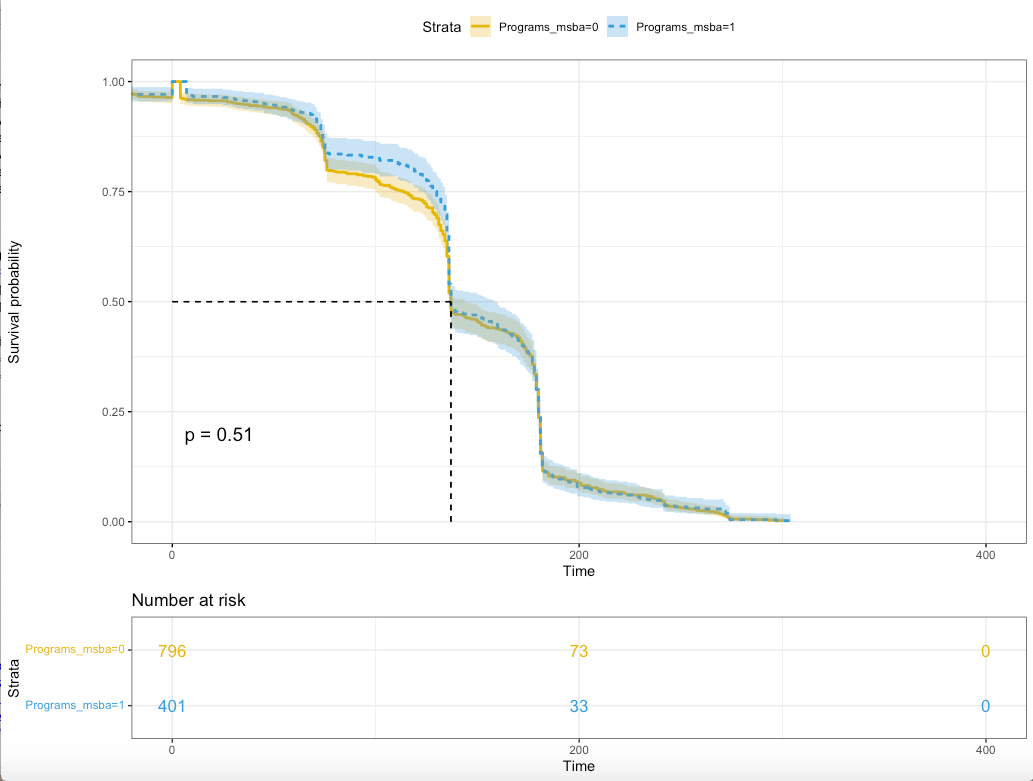


Those who submit online applications directly without coming through other lead sources tend to submit their application earlier. This is probably true because those coming through other sources need to be reminded rather than those who start the online application right away. Furthermore, we are not sure how lead source data is collected. If it is collected through the application as a question on the form, the results could be biased (applicants could accidentally glaze over the question or select a random answer). Moreover, the default answer could be “lead: online application.” However, since there were only 30 datapoints for online application, the graph might be biased.

~by Programs: MSBA (0) and all others (1)



In the above graph, it is evident to see, along with the p-value, that those who apply for MSBA and those who apply for all other programs have similar timelines. There is no difference in survival between students of different programs. All in all, most students tend to submit their applications in under 200 days from first submitting a web inquiry or first visiting the website. After 20 days, fifty percent of students will have already submitted their application.

Similarly, we run a survival model on Programs\_msba, but this time with a different dependent variable: time\_elapsed\_since\_app\_window\_opens. Thus, here we are setting the timer at precisely the date that the window to submit the application opens. The dependent variable will try to explain if there is any difference in timing between students of different programs. 

Here, our results agree with that of the above – there is no significant difference. More interestingly, however, the survival curve tends to plateau, fall sharply, and plateau again, almost in a cycle. This can be explained by the different application deadlines. For example, the first-round deadline to submit might be January 15th , while the second-round deadline may be March 1st , the third-round deadline may be April 15th, and so on. So, as the curve approaches the deadline, more and more students fail to survive, resulting in a sharp downward trend; that is, a good number of them rush to submit their applications as the date approaches a deadline. And then when one deadline passes, a new round starts, resulting in a plateau for the first dozen days. This plateau gets “weaker” as time approaches infinity. Most notably, the third plateau is shorter and even has a negative slope, indicating that students that have not submitted by then are rushing to submit for the last round of deadlines. The slope of the curve then approaches zero as the curve finishes off.

*Recommendations to Merage*

To summarize, we ran different models to investigate the relationship between the time elapsed variable and residency and lead. We also wanted to see if there was a difference in our time elapsed variable between the MSBA program and all other programs in SMP. There is no significant difference—we can see this from the above survival graphs that the p-values are insignificant and that the lines basically overlap. From this, we recommend that Merage does not put too much effort in advertising for different programs. Students behave differently based on how they find the application and if they are a citizen or not, but the time it takes to apply is not affected by the candidate’s preferred program.

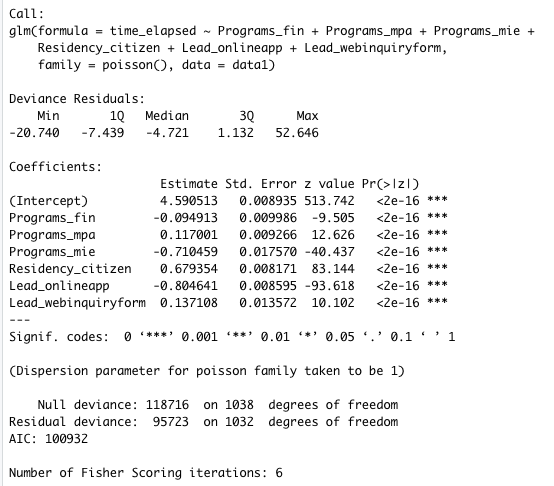
Our recommendations to Merage are to put more effort into continuing to remind and nudge those coming from other sources. Overall, increasing the number of applications will allow Merage to be more selective or increase student capacity and grow the program. Having a more selective vetting process will allow Merage to increase in rankings for SMP. We recommend having automated phone calls or email marketing to nudge Leads from other sources. Giving discounts for application fees or offering conversations with successful alumni could be other ways of catching interest of other leads.

After seeing these results from survival analysis, we wanted to confirm our claims by running a regression. We chose Poisson regression since the time elapsed field is a count variable. We chose to use time elapsed (in days) as the dependent variable.

An Aside: A Model for Confirmation -- Poisson Regression.

After running a few regressions, we realized we should not include the first activity date, or the application submit date in the regression since the time elapsed variable is calculated off those two variables and so will cause multicollinearity. Also, since we had two dummy variables for Residency: international and citizen, we chose to only include one in our analysis to avoid over-explaining the data; just one of those two gives us the desired results. We made similar decisions when looking at programs and leads variables.

Our regression supports the conclusions that were found in the survival analysis.



There is less than a day’s difference between the different programs’ beta values. Those who submit their application from the online application have a negative coefficient and web inquiry form has a positive coefficient. The regression summary supports that citizens take longer time to submit the application rather than non-citizens. Lastly, it is important to note that all the variables are greatly significant likely due to all the variables being binary. However, it is still an interesting investigation and does confirm and shed insight on the various effects on the time of applying.

B. Leads Data.

*Data Description.*

The leads data that we were given consisted 3158 observations of student data and 11 variables. It contains key information about the lead and conversion objects for each student for the application years 2018 and 2019. These variables included the of the salutation of the web inquirer, the date of the inquiry (taken from Sales Force Pardot), the lead source, the conversion object (taken from Pardot), the Pardot first referrer, the program that the web inquiry form was submitted for, and the program that the student was interested in. Note that Pardot is Salesforce’s automated marketing solution. We describe these in detail further.

1. *Salutation*: {Mr., Mrs., Ms. M., Dr., M, Prof.} This column was used to identify the gender of the student. There were too few counts for Dr. And Prof., so these were removed from the dataset, although it would have been insightful to include a dummy variable indicating a degree level of Doctor. Additionally, the meaning of “M.” was ambiguous, and may have represented a missing value, so this was removed from the dataset.
2. *Created.Date* - The created date of the web inquiry. Some of these values were missing, and because the values were same as those in Pardot.Created.Date, the latter was used.
3. *Lead.Source* - {GMASS List Upload, Google Ad, Online Application, Online Event Registration, Qualified List Upload, Web Inquiry Form, and World MBA Tour}. This describes the source that the student came from. Only web inquiry form and online event registration were feasible to include; the others had too few counts (<10). Many values were missing, indicating that the applicant did not come from a lead source.
4. *Student.Stage* - {applicant, inquiry, archive}. The stage of the student. Inquiry denotes a student who had only inquired about the program but did not apply; Applicant refers to a student who started an application but did not submit; and Archive denotes a student who applied to the program.
5. *Pardot.Created.Date* - The date that the student’s information went into the Salesforce system.
6. *Pardot.Conversion.Object.Name* - This column is more granular than lead source and provides some details. The values are: SMP Major Event Registration, Web Inquiry Form AIIMBA, Web Inquiry Form CTA, Web Inquiry Form FTMA, Web Inquiry Form Masters, Web Inquiry Form MFin, Web Inquiry Form MIE, Web Inquiry Form MPAC, Web Inquiry Form MSBA, Web Inquiry Form MSBTM, Web Inquiry Form MSEM, Youvisit Form.
7. *Pardot.First.Referrer.Type* - This column indicates the type first referrer of the student. For example, some came from Google Ads, and some from Google Natural Search. However, since 99% of values were null, this column was removed from the dataset as well.
8. *Pardot.First.Referrer* - This attribute refers to webpage that the student came from. Most students come from the Merage website, so this column was largely useless and so removed from the dataset.
9. *Program* – the program that the student is interested in or ultimately applied for. These were filtered through to include only the programs in SMP.

*Data Preparation and Cleaning.*

The dependent variable that we are interested is Archive (meaning that the student submitted an application), and we want to analyze the effect of different variables, namely, date that the student visited the website, the gender, whether a web inquiry form was submitted, and whether the student registered for an online event, on archiving. Secondarily, we want to explore the effect of these same variables on Archive *or* Applicant; in other words, we want to explore the effect of these variables on starting an application. A dependent dummy variable, arch\_or\_appl, was made to indicate 1 if an application was started, and 0 otherwise.

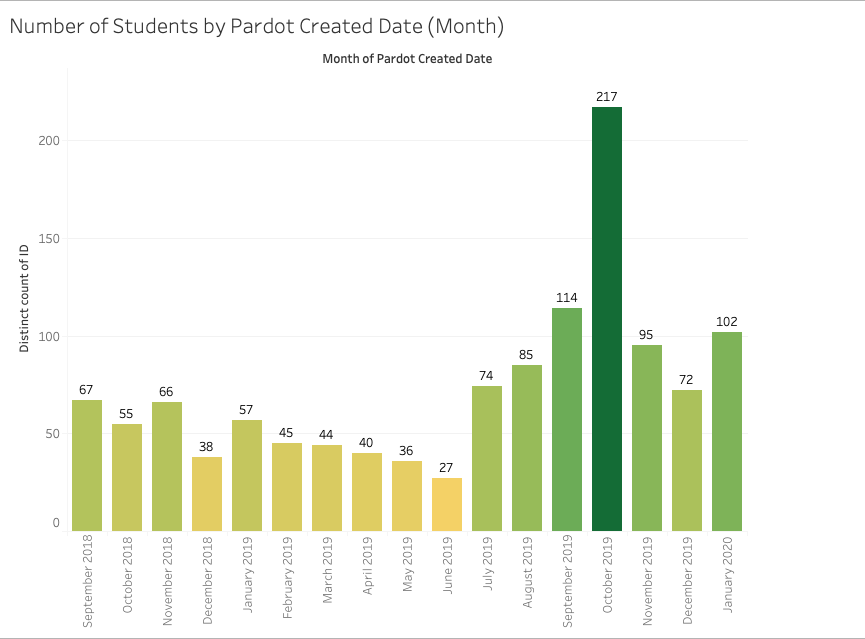
Dummy variables were created for the different leads and conversion objects. After one-hot encoding, we had many more columns, but because of low counts and high p-values, only one lead and one conversion object was kept: web inquiry form and online event registration. Dummy variables were also made for the different student stages, and the different programs. A dummy variable, Male, was created from Salutation, which was 1 where the column reads “Mr.” and 0 otherwise. The date was taken from Pardot.Created.Date -- the package stringr was used to extract only the month and date of the strings in this date column, and then they were converted to POSIXCT objects. Lastly, because we don’t have any 2020 data past January, any 2020 data was removed.

Some variables were engineered to add to the analysis. We had information for dates spanning from 2018 to 2020. Note that all dates were mapped to the same year. Then, to include date in the regression analysis, another variable, time\_til\_oct was created, which calculated the absolute value difference of the date from October. This was done because the application window opens in October; thus this variable explains the relationship between the difference in days the student first visited the site *from* the application window opening date, and applying to the program.

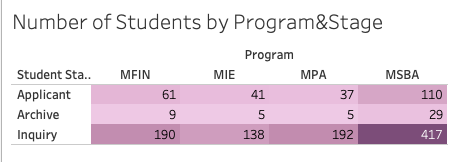
After data cleaning, the dataset consisted of 1135 observations and 24 variables, although only 5 of these variables were used, which consists of the two dependent variables: arch\_or\_appl, stage\_archive and the three independent variables: web\_inq, event\_reg, and time\_til\_oct, male. There were only 49 rows for which the student applied to the program, so an oversampling was done for the minority class to balance the dataset.

*Exploratory Data Analysis.*

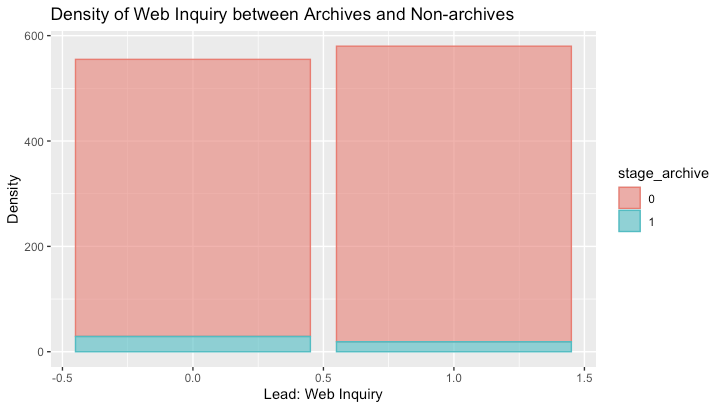
The distribution of dates that students had visited the Merage website is shown below. Again, we see that there is a spike in October when application windows open.



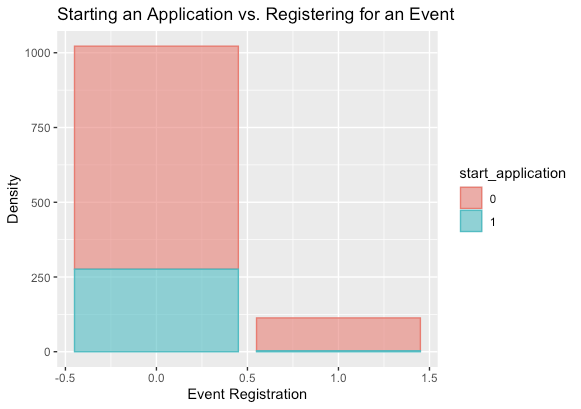
The below figure shows that students are either Applicants, Archive, or Inquiry (with archive being our focus), and MSBA has the most students interested in the program.



The below graph shows the distribution of submitting a web inquiry for two groups: those that submitted an application, and those that did not. Interestingly, there are many more who did not submit a web inquiry (indicated by 0 on the x-axis) who eventually submitted an application (indicated by the blue bars). On the other hand, among the red bars (those who did not submit an application), there were many more students who submitted a web inquiry. This may be because the students who submit a web inquiry are still in the process of researching the program that they want to apply to; people who did not submit one are perhaps more certain that they will apply to Merage’s program.



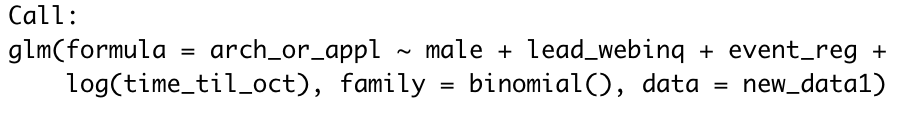
The below figure shows the distribution of event registration vs. starting an application. Among those who did not submit an application, most also did not register for an event. However, among those who did submit an application (blue bar), much more did not register for an event than register for an event. It also seems from this graph that a comparison of the respective proportions will show that event registration may not be that useful for conversion.

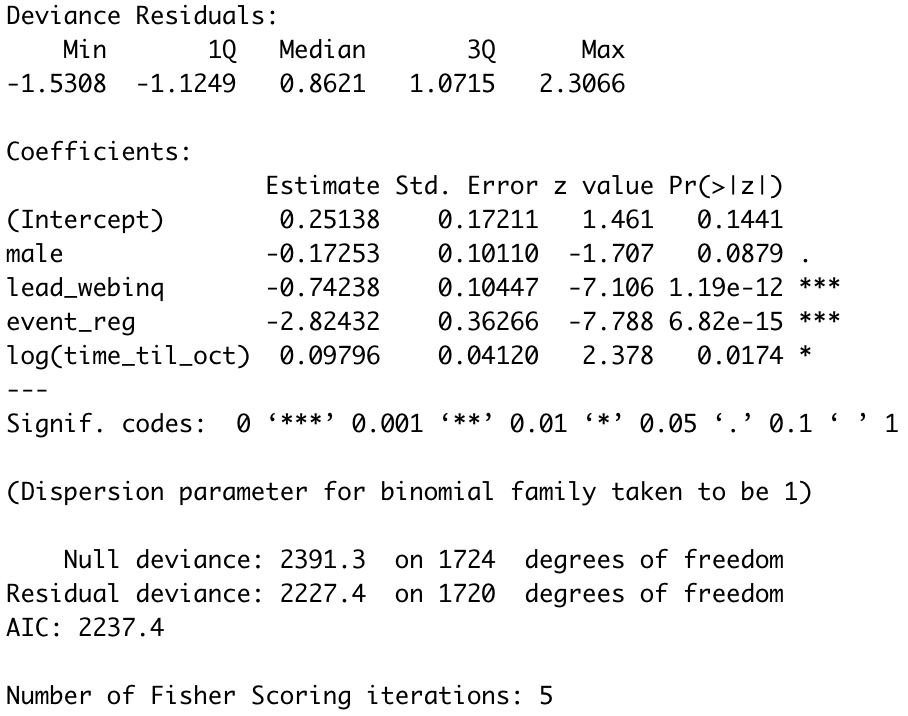


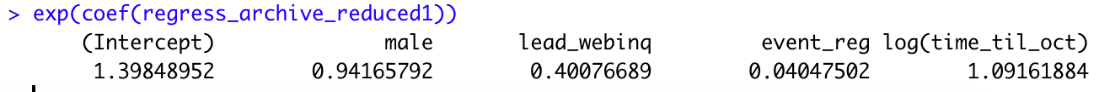
*Modeling.*

**A logistic regression was conducted to explore the effect of gender, web inquiry, event registration, and date on starting an application:**

At a .1 alpha level, the gender is significant. And then, at a .05 alpha level, lead\_webinq, event\_reg, and date are significant. The coefficients of web inquiry and event registration also show our findings agree with the preliminary analysis from the data visualizations. In particular, web inquiry and event registration have negative coefficients, which show that they have a negative effect on starting an application. The date has a positive coefficient, which shows that as the first visit date gets further away from the application window’s opening date, students are more likely to apply. This may be because students whose first visits are earlier in the year are more likely to apply. The day variable was logged because it was on a scale that would overshadow the other binary variables in the analysis.







+Interpretation of model:

Effect of gender: For every unit increase in Male, the odds of submitting an application decrease by 5.8%=( (.942-1)\*100%). More specifically, for every unit increase in male, y is expected to decrease by 17.3%.

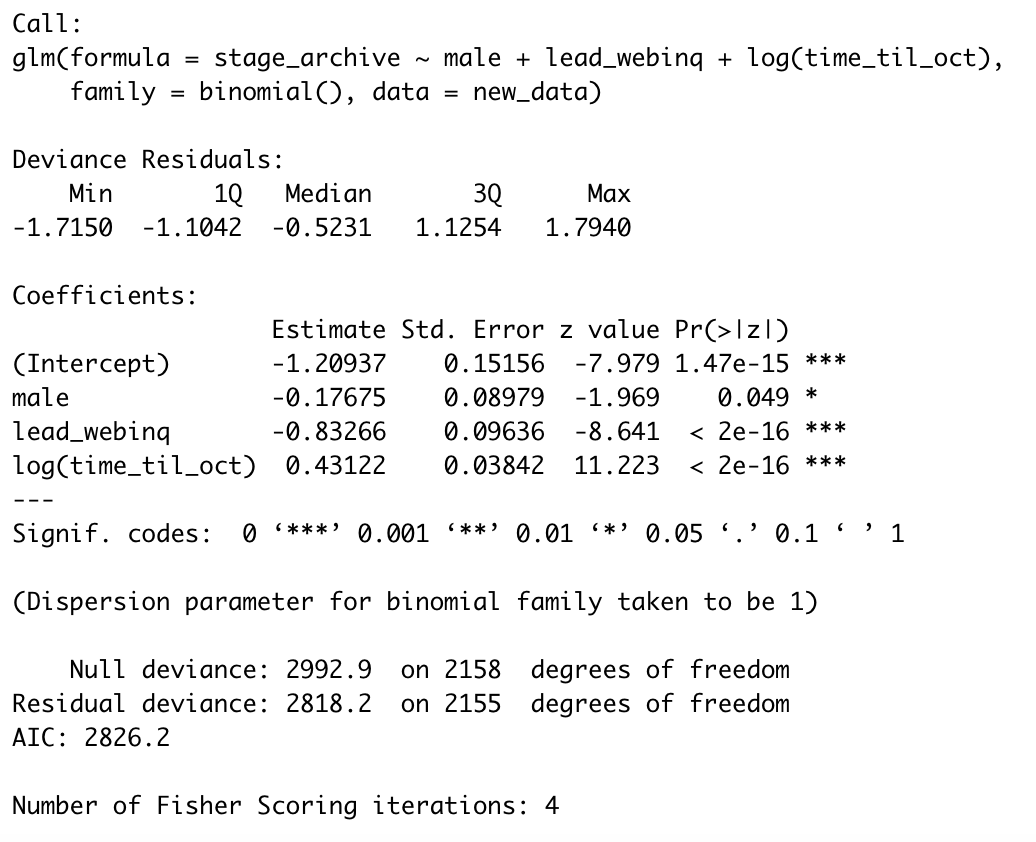
Effect of submitting a web inquiry or not: For every unit increase in lead\_webinq, the odds of submitting an application decrease by 59.9%=( (.40076689-1)\*100%). More specifically, for every unit increase in lead\_webinq, y is expected to decrease by 74.238%.

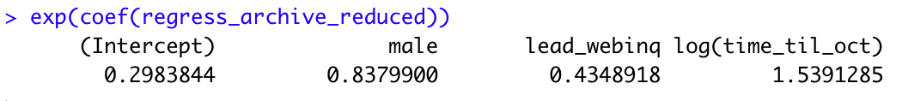
Effect of registering for an event or not: For every unit increase in event\_reg, the odds of submitting an application decrease by 95.9%=( (.04047502-1)\*100%). More specifically, for every unit increase in event\_reg, y is expected to decrease by 282.24%.

Effect of day-distance from start of application submit window: For every unit increase in time\_til\_oct, the odds of submitting an application increase by 9% ((1.09-1)\*100). More specifically, for every 1% increase in time\_til\_oct, y is expected to increase by 0.09796%.

**Similarly, now with a different dependent variable, we conducted a logistic regression to explore the effect of gender, web inquiry, event registration, and date on submitting an application:**

At a .05 alpha level, male, lead\_webinq, and the date are significant. This result also agrees with the first bar graph shown in this section; a student who submits a web inquiry is less likely to apply. Also, as the first visit time gets further away from October, students are more likely to apply. This time, the date variable has a smaller p-value and standard error than the previous regression, indicating that date is a more important factor in predicting submitting an application than predicting starting an application. Furthermore, a regression was run with event\_reg also, but the variable was insignificant. This may be because there is not enough data to test this, or also because event registrations prompt many people to apply for the program, but not necessarily to submit an application.





+Interpretation of model:

Effect of gender: For every unit increase in Male, the odds of submitting an application decrease by 16.2%=((.83799-1)\*100%). More specifically, for every unit increase in male, y is expected to decrease by 17.8%.

Effect of submitting a web inquiry or not: For every unit increase in lead\_webinq, the odds of submitting an application decrease by 56.51%=((.4348918-1)\*100%). More specifically, for every unit increase in lead\_webinq, y is expected to decrease by 83.27%.

Effect of day-distance from start of application submit window: For every unit increase in time\_til\_oct, the odds of submitting an application increase by 53.91% (=(1.5391-1)\*100). More specifically, for every 1% increase in time\_til\_oct, y is expected to increase by 0.43122%.

*Recommendations to Merage*

Surprisingly, if a student fills out a web inquiry form, the student is less likely to apply. And if a student registers for an event or fills out a web inquiry form, the student becomes less likely to start an application. Our recommendation is to ask applicants or inquirers for feedback. Merage may send surveys to gauge satisfaction levels on customer service interactions that include email conversations or meetings with Merage representatives. The same can be done for events – perhaps an event was not insightful or engaging enough to persuade the student to apply. Merage must feedback receive feedback to find out why web inquiry forms or events do not boost application levels. Another hypothesis may be that the students who submit web inquiry forms or register for events are simply those who are still in the research phase and are trying to discover new universities to apply to; thus they are perhaps less certain about applying to UCI. On the other hand, those who do not submit inquiries or register for events may be more “set” on applying – likely due to their trust for the already well-known University of California brand. Nevertheless, as web inquiries and events are originally purposed to persuade those who are uncertain of applying, it is beneficial to investigate why events and web inquiry systems are having a negative effect on applying.

C. Campaign Data.

The campaign data that was given to us consisted of 40 observations of campaign performance data and 12 variables after prepping and cleaning. It contains important information about the performance of a campaign as well as the various engagement metrics of the displayed ad. Unique clicks was used as our dependent variable, since that is the industry standard for measuring conversion. Additionally, post, page, and ad are used interchangeably since an ad on the two platforms may be displayed in various formats.

The following variables were used to generate our analysis:

1. *IsFacebook* – This variable was derived from the campaign name column. We determined whether the campaign ran on Facebook or Instagram, and then assigned 0 or 1 depending on if it ran on Facebook or not.
2. *Reach* – the total number of different people exposed to the ad at least once during a campaign.
3. *Impressions* – the total amount of times the ad was displayed.
4. *AmountSpent* – the total amount of money spent on the campaign.
5. *PageEngagement* – total engagement with the ad post.
6. *PageLikes* – the total number of likes the ad received.
7. *PostComments* – the total number of comments on the ads displayed.
8. *PostEngagement* – the total amount of engagement with the post
9. *PostReactions* – the total number of reactions on the ad, such as like, love, anger, sad, etc.
10. *PostShares* – the number of times the post or ad was shared.
11. *CampaignDuration* – the total duration of the campaign in number of days, derived by taking the difference between the start date and the end date of a campaign.
12. *UniqueClicks* – the number of times an ad was clicked by a unique user.

*Data Cleaning and Preparation.*

The dependent variable that we are interested in is UniqueClicks (meaning the number of times an ad was clicked by a unique user), and we want to analyze the effect of different variables such as ad engagement metrics like post shares and reactions and performance metrics like clicks and impressions.

First, we removed any of the variables that were calculated fields to reduce highly correlated variables and any potential multicollinearity issues since many of the columns were calculated fields and proportions created using other columns in the data. Next, we removed any columns with many missing or 0 values since those columns would not be useful to us.

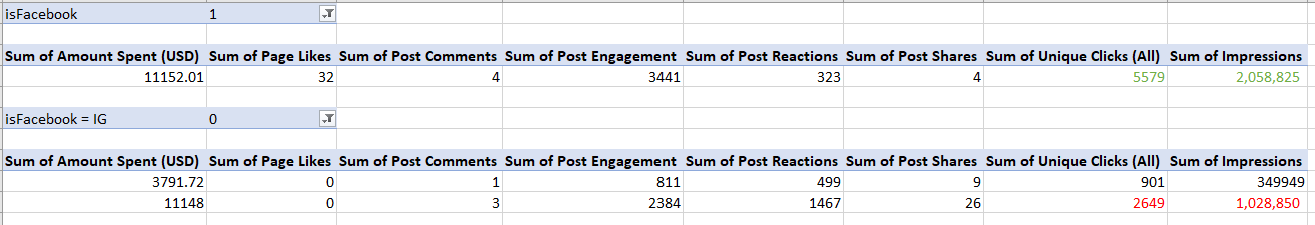
Then, we created dummy variables for any categorical variables in our data. We used the campaign name column to determine which platform the ad was displayed on using the grep text analysis function in r, which detects the presence of certain words or combinations of words. Since all of the campaign names contained either Facebook or Instagram in the title based on platform, we used this to create a dummy variable, isFacebook, which was 1 if the campaign ran on Facebook or 0 if it ran on Instagram.

Additionally, we also calculated campaign duration by taking the difference between the campaign start date and end date which resulted in values between 3 days and 7 days, depending on campaign length.

*Exploratory Data Analysis*

We then ran some basic exploratory data analysis to see the differences in advertising performance across the two platforms to see if we could see any trends prior to modeling.

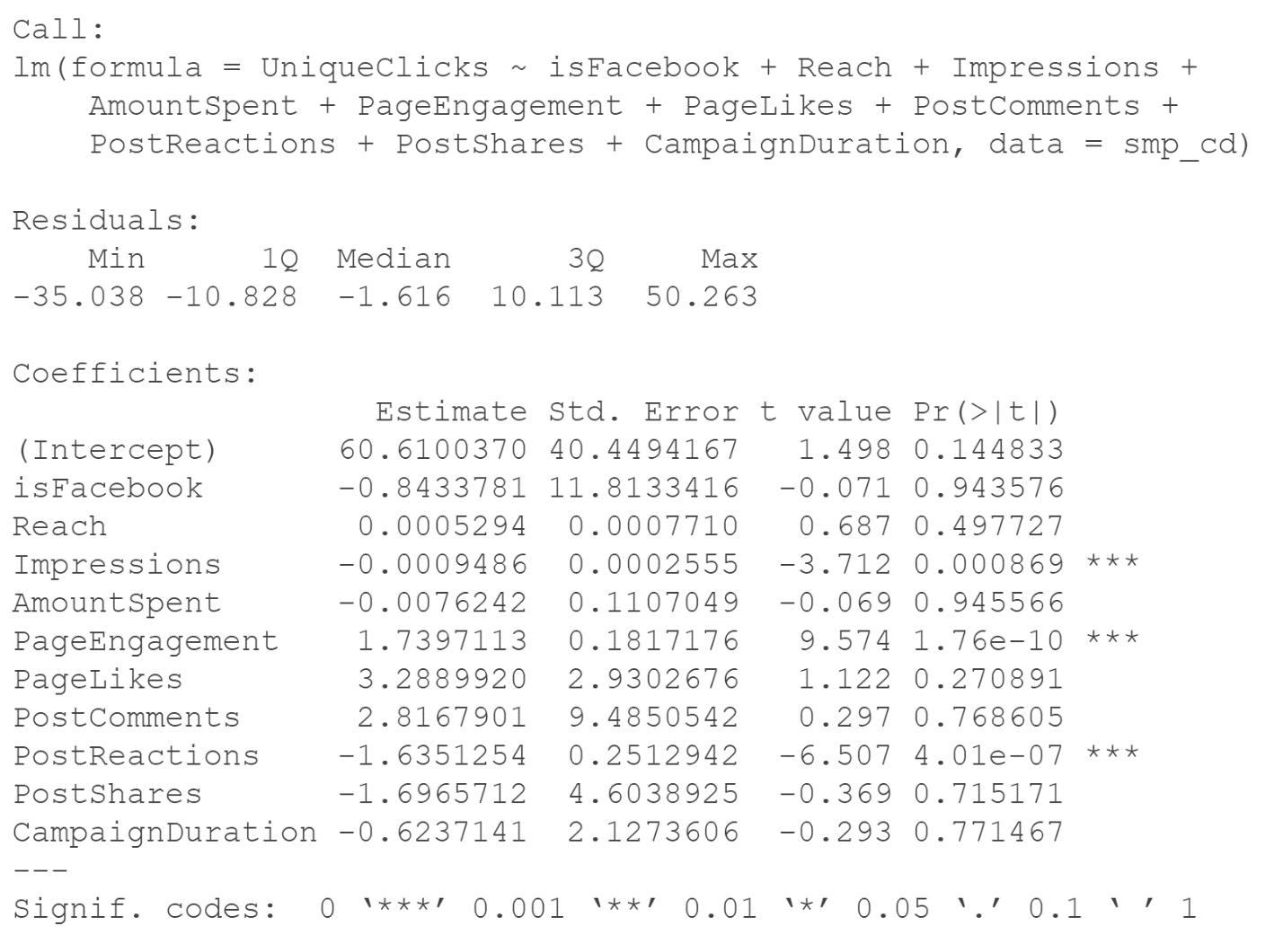
First, we took the sums of appropriate variables separated by platform. We noticed that the total spend between the two platforms was split 75%-25%, so we tried to replicate equal spend by multiplying the spend on Instagram to match that on Facebook. Since we had to multiply spend by a factor of 2.94 to achieve these results, we multiplied the rest of the aggregates for Facebook by 2.94 as well so that we could finally compare the performance on the platforms side by side.



When we compare the performance data for each platform side by side, we can see that for the same amount of money, Merage receives only half the clicks and impressions than it does if it ran on Facebook, which shows that Instagram advertising is almost 2x as expensive as Facebook.

*Modeling*

We decided to run a simple linear regression to see the effect of different types of engagement and performance on unique clicks.



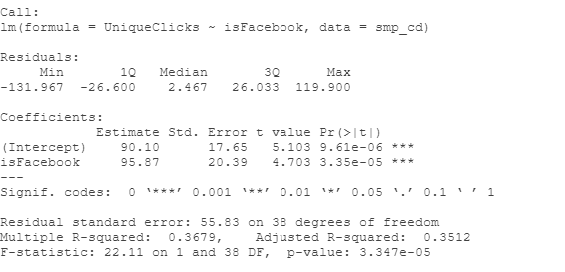
From the regression above, we can see that the effect of Impressions, Page Engagement, and Post reactions are significant on unique clicks. We can interpret the above results as:

Each additional impression will decrease unique clicks by –0.0009. This implies that Merage is showing the ads too frequently to the same users. An addition of a frequency cap (a maximum amount of times a user can see the ad in a given time period) may help to alleviate this issue and in turn increase ROI.

Each additional unit of PageEngagement leads to an increase of 1.74 clicks. Merage should try to increase their engagement with the ad by including interactive captions (such as broad open-ended questions, for example, “What are you most excited about learning about Data Science?”) to try to increase unique clicks.

Each additional post reaction decreases the number of unique clicks by 1.64. This is generally because a user either reacts to a post or clicks, but not both. Merage can try to increase the likelihood of a user clicking by making their ads more attractive by providing an incentive to click such as the chance to receive a fee waiver code by clicking on the ad. This in conjunction with a frequency cap on the ad will incentivize clicks over reactions, since the user may not see the ad again if they don’t click.

Additionally, we wanted to see the impact of advertising on Facebook versus Instagram on our dependent variable clicks to see which platform gives us a better result. We ran a linear regression and found the following:

If a campaign runs on Facebook, it is likely to get about 95 more clicks than if it were to run on Instagram. The above model shows that Merage should advertise on Facebook more than Instagram because Facebook gives us a higher number of clicks (our conversion metric) than Instagram.

**Limitations**

The most crucial limitation concerned the overall sample size of our data sets. In the Leads dataset, after filtering for only SMP programs, the number of observations for “archive” were reduced. Similar situations occurred in the Applicants dataset as well as in the campaign dataset.

For SMP programs, most observations were Facebook campaign data, so we didn’t have many observations for other advertising platforms (such as LinkedIn) and could not analyze that data.

Another problem we ran into was the absence of a common variable that tied together all our data sets. Because of this, we couldn’t effectively use all the data available to us. For example, we had campaign data and click data, but we didn’t have specialized click tracker data that would have showed us where the individuals who clicked on the ads later went. If we had this data, we may have been able to track the complete progress from viewing of an impression to submitting an application.

**Recommendation**

Our recommendation to Merage School is to find a unique identifier to tie all datasets together. This will allow for more robust insights, modeling and the use of several different techniques. We felt limited in our abilities and tried to get as creative as we could within the given constraints.

Looking into the actionable insights from each of the datasets viz. Campaign, Leads and Applicants, a summary of our recommendations for Merage is:

1. Focus marketing efforts on Facebook (FB), as a preferred social media channel and explore ways to engage users to like the Merage FB page. Seeing the success of Facebook over other social media channels, it would be better to focus on this channel rather than spreading thin resources over many social media channels.
2. Enhance the conversion rate of applicants by exploring automating drip mailer communication strategy (sending out emails to applicants who are stuck at a non-archive stage), targeted one-on-one counseling for various lead stages from the conversion funnel. Take a customized effort to improve overall male conversion rate and to increase applicant numbers for female while retaining conversion rate .
3. Target marketing toward international students for first few submission deadlines and residents for later deadlines.
4. Minimize drop off from lead sources other than “online form” by improving engagement efforts.
5. Send satisfaction surveys to gauge satisfaction levels and receive feedback on web inquiry form systems and Merage events.