Problem to solve:

Increase the number of enrollments for the university. One way to do this is to spend money on a lead acquisition campaign. However, the university doesn’t have the resources to go through the large number of leads they receive. They need a way to prioritize the leads that have the best chance of enrolling.

Client:

Mark University – If I can prioritize their leads they will use that prioritization to decide which inquires they will spend the majority of their time communicating with. This is will help them keeps costs and staffing in place while lowering the chance they miss out on a great candidate. A byproduct of this process could also be that they learn which marketing channels they should focus on which again will help them keeps costs under budget while increasing effectiveness.

Data:

* Full Name
  + Some Duplicates. How to handle? Really duplicates or just people with the same name?
* First Name
* Last Name
* Created On
  + Date the record was created on
* Location
  + Location where the prospect would be interested in attending. What if prospect places a request for multiple locations? So you would have duplicates in name but the location would make the entire observation no longer a duplicate.
* Inquiry Date
* Inquiry Source
  + How did this record get placed in the database
* Prospect Status
  + Shows the prospects current status in the enrollment funnel
* Prospect Date
  + Shows the date they gained their current status
* Application Status
  + The current status of their application. Blank should mean they have not applied
* Application Started Date
* Application Completed Date
* Application Submitted Date
* Application Moved to ERP Date
* Admit Date
* Enrolled Date
  + This will be the dependent variable. I will need to change the value to a factor which has values or either 1 or 0. 1 will mean they enrolled and 0 will mean they did not
* Academic Level of Interest – All values are “Graduate” - TO BE DELETED
* Academic Program Level of Interest – Program they are interested in. Character Field - There are a lot of possible values or this variable
* Anticipated Entry Term – when they are thinking they will start school
* Marketing Campaign
  + the campaign used to access website prior to becoming prospect
* Marketing Medium
  + the medium used to access website prior to becoming prospect
* Marketing Source
  + the source used to access website prior to becoming prospect
* Reference Number
  + Assigned in the system which contact record is being referenced (good criteria for keeping this version of duplicate?)

Data Cleaning:

* Explore name duplicates to determine how many variables they are matching on
* Decide whether or not to remove duplicates
* Missing Values
  + Name fields – remove any observations with missing data
  + Location – change missing values to “None”
  + Marketing Campaign, Source, and Medium - all n/a should be “(not set)
* Data to Remove
  + Delete all records with any date prior to 6/1/2016
  + Delete Academic Level of Interest column
  + Delete Anticipated Enter Term column
  + Delete Reference Number Column
  + Delete name fields
* New to Create
  + Days Between prospect date and app submitted date
  + Change Enrolled Date to binary
* Data to Mutate – need to group in a meaningful way – too many possible values
  + Academic Program of Interest
  + Marketing campaign, source and medium
  + Location

Initial findings:

Summary of the logistic regression model returned with too much information it filled up the console. I will need to find a way to limit options in the “Data to Mutate” variables above.

Approach:

Initial approach would be using logistic regression on the enrollment variable in order to create a model which will predict the probability that a particular prospect enrolls.

Devilerables:

Code

Paper on Methodology along with the results on test data