University Enrollment Lead Scoring Model

Mark Brandes

February 14, 2017

## The Opportunity

My client is a university system whose number one goal is getting qualified students enrolled in one of their university locations. In order to increase the number of enrollments the University is planning on investing in a lead generation marketing campaign. Some of the previous lead generation campaigns were so successful that they ended up not my the resources required to follow up on all the leads.

The solution I am proposing is to create a lead scoring model that would give each lead a certain probability of enrolling. Using this information they will then be able to prioritize which leads they follow up with first. This is will help them keeps costs and staffing in place while lowering the chance they miss out on a great candidate.

An added benefit of this model will be the ability to learn which of their marketing channel brings in the most leads with a high probability of enrolling.

## Adding needed libraries

We'll get started by adding the libraries that will be needed to create this model.

library(dplyr)  
library(tidyr)  
library(readr)  
library(caTools)  
library(ROCR)  
library(lubridate)

## The Data

Now that the libraries are in we can go about importing the data.

original <- read\_csv("Prospectdata.csv")  
allrecords <- original  
str(allrecords)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 17980 obs. of 19 variables:  
## $ CreatedOn : chr "12/5/2016" "2/12/2017" "7/15/2016" "7/15/2016" ...  
## $ Location : chr "St. Louis Home Campus" "St. Louis Home Campus" "St. Louis Home Campus" "St. Louis Home Campus" ...  
## $ InquiryDate : chr "12/5/2016" "2/12/2017" "7/15/2016" "7/15/2016" ...  
## $ InquirySource : chr "Web, Create Account" "Online Application" "Web, Create Account" "Web, Inquiry" ...  
## $ ProspectStatus : chr "Application Submitted" "Application Submitted" "Inquiry" "Inquiry" ...  
## $ ProspectDate : chr "12/5/2016" "2/12/2017" "7/15/2016" "7/15/2016" ...  
## $ ApplicationStatus : chr "Submitted" "Submitted" NA NA ...  
## $ ApplicationStartedDate : chr "12/25/2016" "2/12/2017" NA NA ...  
## $ ApplicationCompletedDate : chr NA NA NA NA ...  
## $ ApplicationSubmittedDate : chr "12/25/2016" "2/12/2017" NA NA ...  
## $ ApplicationMovedtoERPDate: chr NA NA NA NA ...  
## $ AdmitDate : chr NA NA NA NA ...  
## $ EnrolledDate : chr NA NA NA NA ...  
## $ AcademicLevelofInterest : chr "Graduate" "Graduate" "Graduate" "Graduate" ...  
## $ academicProgramofInterest: chr "art History and Criticism" "art History and Criticism" "art History and Criticism" "art History and Criticism" ...  
## $ AnticipatedEntryTerm : chr "Fall I, 2017" "Fall I, 2017" "Fall I, 2017" "Fall I, 2017" ...  
## $ MarketingCampaign : chr "(not%20set)" "(referral)" "(direct)" "(referral)" ...  
## $ MarketingMedium : chr "(not%20set)" "referral" "(none)" "referral" ...  
## $ MarketingSource : chr NA "webster.edu" "(direct)" "webster.edu" ...  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 19  
## .. ..$ CreatedOn : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Location : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ InquiryDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ InquirySource : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ProspectStatus : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ProspectDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ApplicationStatus : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ApplicationStartedDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ApplicationCompletedDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ApplicationSubmittedDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ ApplicationMovedtoERPDate: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ AdmitDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ EnrolledDate : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ AcademicLevelofInterest : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ academicProgramofInterest: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ AnticipatedEntryTerm : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ MarketingCampaign : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ MarketingMedium : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ MarketingSource : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

From the structure of the data, this first thing you notice is that every field was imported as a character string. Clearly that is not going to work for creating our model so we will need to clean up the data.

We will start by looking at each of the variables individually in order to determine its value to the dataset.

* CreatedOn - The date when the record was created in the database
  + Need to change this into the proper date format
* Location - Which location the person is interested in attending
  + This variables has too many values so we will engineer new feature that will be more helpful to the model
* Inquiry Date- The date when the person placed an inquiry
  + This date in and of itself will not help but creating a binary value for whether or not the person placed an inquiry may help.
* Inquiry Source - This will tell us if the inquiry was done online, offline, in person, on a website, etc.
  + Need to change this field into a factor
* Prospect Status - This is an internal tracking field showing where the person is in the enrollment funnel
  + Not helpful to the model so this will be removed
* Prospect Date - The date that the most recent prospect status was set.
  + Not helpful to the model so this will be removed
* Application Status - This is the current status of the person's application.
  + Not helpful to the model so this will be removed
* Application Started Date - Date the application was started
  + Everyone interested in enrolling must fill out an application so I don't believe a binary variable will help here. However, using this field along with the Application Submitted Date may give us a useful time period value that we can use
* Application Completed Date - Date the application was completed
  + Intermediary step between started and submitted that is not consistently filled in. Will be removed from the dataset.
* Application Submitted Date - Date the application was submitted
  + Will be used with the Started Date to determine a time period between
* Application Moved to ERP Date - Date the application was moved to ERP
  + Seems to be rarely used. Will be removed form the dataset
* Admit Date - Date the person was admitted to school
  + Everyone needs to be admitted in order to be enrolled so this will not help the model
* Enrolled Date - Date when the person was enrolled in classes
  + Dependent variable for our model
* Academic Level of Interest - The variable contains the same value of "Graduate" for all observations
  + This variable will be deleted
* Academic Program of Interest - This is the program the person is interested in
  + Helpful variable but there are too many values for a factor so some feature engineering will need to take place
* Anticipated Entry Term - The term when the person is interested in enrolling
  + This field will not be helpful to our model so it will be removed
* Marketing Campaign - This is the campaign used to access the website if they submitted their records online
  + Need to clean up some values, replace the N/A's with (not set), and then change into a factor
* Marketing Medium - This is the medium used to access the website if they submitted their records online
  + Need to clean up some values, replace the N/A's with (not set), and then change into a factor
* Marketing Source - This is the source used to access the website if they submitted their records online
  + This field has too many values with no obvious way of grouping them. It could be helpful to create a new binary variable from this as to whether or not the record has a source

## Cleaning the data

Now that we have created a plan for our data, we will start the cleaning process by removing the variables that will not be helpful to our model or analysis.

allrecords <- allrecords %>%   
 select(-ProspectStatus) %>%  
 select(-ProspectDate) %>%  
 select(-ApplicationStatus) %>%  
 select(-ApplicationCompletedDate) %>%  
 select(-ApplicationMovedtoERPDate) %>%  
 select(-AdmitDate) %>%  
 select(-AcademicLevelofInterest) %>%  
 select(-AnticipatedEntryTerm)

Next we need to replace the NA values in our independent variables to something useful. For InquirySource, Location, MarketingMedium, and MarketingCampaign, I choose the value of "not set" to replace the NA values. Two of those fields use the value of "none" for some records so I wanted to use something other than that.

For academicProgramofInterest I decided to use "Undecided" as my NA replacement value. My reasoning was that since the prospects did not pick a program they could be assumed to be undecided about there program choice. This is a value that was already used in the data for other sources so it also kept consistency.

Two other changes I made were to the MarketingMedium variable. There were a couple values that should have been labeled something else so I fixed those.

allrecords <- allrecords %>%   
 replace\_na(list(InquirySource = "not set")) %>%  
 replace\_na(list(Location = "not set")) %>%  
 replace\_na(list(academicProgramofInterest = "Undecided")) %>%  
 replace\_na(list(MarketingMedium = "not set")) %>%  
 replace\_na(list(MarketingCampaign = "not set"))  
  
allrecords$MarketingCampaign <- gsub("%20"," ", allrecords$MarketingCampaign)  
allrecords$MarketingCampaign <- gsub("(not set)","not set", allrecords$MarketingCampaign, fixed = TRUE)  
allrecords$MarketingMedium <- gsub("%20"," ", allrecords$MarketingMedium)  
allrecords$MarketingMedium <- gsub("(not set)","not set", allrecords$MarketingMedium, fixed = TRUE)  
allrecords$MarketingMedium <- gsub("FacebookAd","PaidSocial", allrecords$MarketingMedium)  
allrecords$MarketingMedium <- gsub("OnlineMarketing","PaidDisplay", allrecords$MarketingMedium)

At this point the last thing to do was to change the formatting of the variables. I changed all the dates to the mm/dd/yyyy format. I also changed InquirySource, MarketingCampaign, and Marketing Medium to factors.

allrecords$CreatedOn <- as.Date(allrecords$CreatedOn, format="%m/%d/%Y")  
allrecords$InquiryDate <- as.Date(allrecords$InquiryDate, format="%m/%d/%Y")  
allrecords$ApplicationStartedDate <- as.Date(allrecords$ApplicationStartedDate, format="%m/%d/%Y")  
allrecords$ApplicationSubmittedDate <- as.Date(allrecords$ApplicationSubmittedDate, format="%m/%d/%Y")  
allrecords$EnrolledDate <- as.Date(allrecords$EnrolledDate, format="%m/%d/%Y")  
  
  
allrecords$InquirySource <- as.factor(allrecords$InquirySource)  
allrecords$MarketingCampaign <- as.factor(allrecords$MarketingCampaign)  
allrecords$MarketingMedium <- as.factor(allrecords$MarketingMedium)

## Feature Engineering

I had mentioned previously when looking at the individual variables that we would need to do some feature engineering. Now that the data is clean we will start that process. The first thing I will do is read in the reference files for grouping Location and academicProgramofInterest. These files were created using domain knowledge from the client.

progref <- read\_csv("ProgramGroupReference.csv")  
locref <- read\_csv("LocationReference.csv")

### New Variables

This first step I will take is creating three new binary variables that will either have a value of 1 or 0.

The most important new variable is "enrolled"" which will have a value of 1 if the EnrolledDate field has a date and a value of 0 if there is no enrolled date. This variable will serve as the dependent variable for our model.

The next variable I will create is called "inquired which will be similar to how I created enrolled. If there is an InquiryDate then it gets a value of 1 and if there is no date it gets a value of 0. The idea here is that the actual date of the inquiry doesn't seem to hold much value, but the fact that they did or did not inquire before they applied may have some.

The third variable I will create is for MarketingSource. MarketingSource has far too many values to be of use in our model. However, as with InquiryDate, there may still be value in whether or not this record has a source. So this variable will be called "HasSource" and will have a value of 0 if the MarketingSource is "NA" and a value of 1 if not.

The final step for these three new variables will be to set their format to factor.

allrecords <- allrecords %>%  
 mutate(enrolled=if\_else(is.na(allrecords$EnrolledDate),0,1)) %>%  
 mutate(inquired=if\_else(is.na(allrecords$InquiryDate),0,1)) %>%  
 mutate(HasSource=if\_else(is.na(allrecords$MarketingSource),0,1))  
  
allrecords$enrolled <- as.factor(allrecords$enrolled)  
allrecords$inquired <- as.factor(allrecords$inquired)  
allrecords$HasSource <- as.factor(allrecords$HasSource)

### Grouping Location and AcademicProgramofInterest

My next step in feature engineering will using the translation file we added to join in smaller groupings on those values. As mentioned earlier these files were made with domain knowledge from the client.

The join with the ProgramReference file will add a "Group" variable to the dataset. This group variable will turn out to be a factor with 9 levels which will be much more helpful than our character values in AcademicProgramofInterest.

The join with the LocationReference file will add "Metro", "Region", "Characteristic", and "Characteristic2" variables to our dataset. The "Metro" variable will group the locations by their placement in a specific Metro in the country. The "Region" variable will group those regions even more broadly. The Characteristic variables use a different type of grouping based on the characteristics of those location instead of their geography.

Once again we turn all these variables into factors.

allrecords <- left\_join(allrecords,progref,by="academicProgramofInterest")  
allrecords <- left\_join(allrecords,locref,by="Location")  
  
allrecords$Group <- as.factor(allrecords$Group)  
  
allrecords$Metro <- as.factor(allrecords$Metro)  
allrecords$Region <- as.factor(allrecords$Region)  
allrecords$Characteristic <- as.factor(allrecords$Characteristic)  
allrecords$Characteristic2 <- as.factor(allrecords$Characteristic2)  
  
str(allrecords)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 17980 obs. of 19 variables:  
## $ CreatedOn : Date, format: "2016-12-05" "2017-02-12" ...  
## $ Location : chr "St. Louis Home Campus" "St. Louis Home Campus" "St. Louis Home Campus" "St. Louis Home Campus" ...  
## $ InquiryDate : Date, format: "2016-12-05" "2017-02-12" ...  
## $ InquirySource : Factor w/ 8 levels "Internal Application",..: 7 3 7 8 2 3 3 3 3 3 ...  
## $ ApplicationStartedDate : Date, format: "2016-12-25" "2017-02-12" ...  
## $ ApplicationSubmittedDate : Date, format: "2016-12-25" "2017-02-12" ...  
## $ EnrolledDate : Date, format: NA NA ...  
## $ academicProgramofInterest: chr "art History and Criticism" "art History and Criticism" "art History and Criticism" "art History and Criticism" ...  
## $ MarketingCampaign : Factor w/ 23 levels "(direct)","(organic)",..: 22 3 1 3 22 2 1 22 18 3 ...  
## $ MarketingMedium : Factor w/ 9 levels "(none)","email",..: 4 9 1 9 4 5 1 4 3 9 ...  
## $ MarketingSource : chr NA "webster.edu" "(direct)" "webster.edu" ...  
## $ enrolled : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 1 ...  
## $ inquired : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 2 ...  
## $ HasSource : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 1 2 2 ...  
## $ Group : Factor w/ 9 levels "ArtsMusic","Business",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Metro : Factor w/ 48 levels "Augusta","Austria",..: 43 43 43 43 16 43 43 43 43 43 ...  
## $ Region : Factor w/ 21 levels "Arizona","Arkansas",..: 11 11 11 11 17 11 11 11 11 11 ...  
## $ Characteristic : Factor w/ 8 levels "Home","Hospital",..: 1 1 1 1 3 1 1 1 1 1 ...  
## $ Characteristic2 : Factor w/ 11 levels "AirForce","Army",..: 3 3 3 3 6 3 3 3 3 3 ...

submitted <- allrecords %>% filter(!is.na(ApplicationSubmittedDate))  
submitted <- mutate(submitted,appsubmitdays=difftime(CreatedOn,ApplicationSubmittedDate,units="days"))  
submitted <- filter(submitted, appsubmitdays >= 0)

## Creating the Model

Now that our data is cleaning and feature engineering is completed we can being to create our model. In this case I decided to use a logistic regression model since my independent variable is binary. I also like the fact that my output from the model will be the probability that a person will enroll which is exactly what I want to base my leading scoring on.

First we are going to take a look at a table of our dependent variable to get a feel for what our baseline model would be.

table(allrecords$enrolled)

##   
## 0 1   
## 15473 2507

15473/(15473+2507)

## [1] 0.8605673

So it looks like if we had our model always choose "not enrolled" then we would have an 86% accuracy rate. This is the number we are hoping our model will beat.

Now we are going to break up our dataset into a training set and testing set.

set.seed(9)  
split <- sample.split(allrecords$enrolled,SplitRatio = 0.75)  
allrecTrainSet <- subset(allrecords, split==TRUE)  
allrecTestSet <- subset(allrecords, split==FALSE)

We are going to use the training set in the glm function to make our model.

allrecLogModel <- glm(enrolled ~ InquirySource + inquired + MarketingCampaign + MarketingMedium + HasSource + Group + Metro + Region + Characteristic + Characteristic2, data = allrecTrainSet, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(allrecLogModel)

##   
## Call:  
## glm(formula = enrolled ~ InquirySource + inquired + MarketingCampaign +   
## MarketingMedium + HasSource + Group + Metro + Region + Characteristic +   
## Characteristic2, family = binomial, data = allrecTrainSet)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5424 -0.2250 -0.0205 0.0000 3.9723   
##   
## Coefficients: (39 not defined because of singularities)  
## Estimate  
## (Intercept) -4.170e-01  
## InquirySourcenot set -1.993e+01  
## InquirySourceOnline Application -1.754e+00  
## InquirySourceOnline Event Registration -4.288e+00  
## InquirySourceOpen Information Session -2.245e+00  
## InquirySourceWeb Create Account - Default Create Account Form 1.538e+01  
## InquirySourceWeb, Create Account -2.627e+00  
## InquirySourceWeb, Inquiry -8.870e+00  
## inquired1 NA  
## MarketingCampaign(organic) -1.220e-01  
## MarketingCampaign(referral) -2.111e-01  
## MarketingCampaign1MBA students%2C 2015 -1.926e+01  
## MarketingCampaignFY1516Grad\_Columbia -3.556e+01  
## MarketingCampaignFY1516Grad\_LittleRock -1.975e+01  
## MarketingCampaignFY1516Grad\_Orlando 4.936e-03  
## MarketingCampaignFY1516Grad\_SanAntonio -3.735e+01  
## MarketingCampaignFY1516Grad\_Webster -2.153e+01  
## MarketingCampaignFY1516STLUG\_Webster -3.843e+01  
## MarketingCampaignFY1617Grad\_Webster -1.956e+01  
## MarketingCampaignFY1617Grad\_Webster\_COL -2.535e+00  
## MarketingCampaignFY1617Grad\_Webster\_LIT -3.650e+01  
## MarketingCampaignFY1617Grad\_Webster\_ORL -2.070e+01  
## MarketingCampaignFY1617Grad\_Webster\_SAN -1.172e+00  
## MarketingCampaignFY1617Grad\_Webster\_STL -1.997e+01  
## MarketingCampaignglobalmaemails -1.857e+01  
## MarketingCampaignGraduate\_ALL -4.431e-01  
## MarketingCampaignGraduate\_SANA -3.120e+01  
## MarketingCampaignGraduate\_STL -1.004e-02  
## MarketingCampaignnot set -2.435e-01  
## MarketingCampaignUS\_Military\_endofyear -6.554e-01  
## MarketingMediumemail NA  
## MarketingMediumMicrosite NA  
## MarketingMediumnot set NA  
## MarketingMediumorganic NA  
## MarketingMediumPaidDisplay 1.931e+01  
## MarketingMediumPaidSocial NA  
## MarketingMediumRadioMarketing NA  
## MarketingMediumreferral NA  
## HasSource1 NA  
## GroupBusiness 1.293e+00  
## GroupCounseling 5.026e-01  
## GroupLeadership 1.635e+00  
## GroupMarketing and Communication 1.155e+00  
## GroupNursing and Health Care 7.480e-01  
## GroupTeaching and Education 1.226e+00  
## GroupTechnology 1.325e+00  
## GroupUndecided 6.394e-01  
## MetroAustria -1.486e+00  
## MetroCalifornia 1.764e+00  
## MetroCharleston 5.162e-01  
## MetroChicago 1.298e+00  
## MetroChina -2.757e+00  
## MetroColorado Springs -3.583e-01  
## MetroColumbia -2.863e-03  
## MetroDC 1.316e-01  
## MetroDenver -1.159e+00  
## MetroEl Paso -6.341e-01  
## MetroFayetteville -8.709e-01  
## MetroFort Leonard Wood 7.559e-01  
## MetroFort Smith Metro 1.288e-01  
## MetroGreece -1.751e+01  
## MetroGreenville 7.505e-03  
## MetroHinesville -2.242e-02  
## MetroHospital 1.195e+00  
## MetroIrvine 8.467e-01  
## MetroJacksonville -2.576e-01  
## MetroKansas City 5.164e-01  
## MetroLakeland 3.989e-01  
## MetroLawton -1.861e+00  
## MetroLittle Rock -5.278e-01  
## MetroLos Angeles -2.785e-01  
## MetroLouisville -7.489e-01  
## MetroMelbourne -1.354e-01  
## MetroMemphis 1.526e+00  
## MetroMerritt Island 7.401e-01  
## MetroMyrtle Beach 3.187e-01  
## MetroNetherlands -1.791e+00  
## Metronot set -3.440e+01  
## MetroOcala -1.126e-01  
## MetroOklahoma City -7.722e-01  
## MetroOnline 2.834e-01  
## MetroOrlando 1.220e+00  
## MetroPhoenix 5.489e-01  
## MetroRolla 3.177e-01  
## MetroSalt Lake City -1.253e-01  
## MetroSan Antonio 2.092e-01  
## MetroSavannah -1.830e+00  
## MetroSpokane -1.306e+01  
## MetroSt Louis 1.193e-01  
## MetroSt Petersburg 1.937e-01  
## MetroStudy Abroad -1.779e+01  
## MetroSumter 7.437e-02  
## MetroSwitzerland -1.367e+00  
## MetroWichita 2.413e+00  
## RegionArkansas NA  
## RegionCalifornia NA  
## RegionColorado NA  
## RegionDC NA  
## RegionFlorida NA  
## RegionGeorgia NA  
## RegionIllinois NA  
## RegionKansas NA  
## RegionMaryland NA  
## RegionMissouri NA  
## RegionNonUS NA  
## RegionNorth Carolina NA  
## Regionnot set NA  
## RegionOklahoma NA  
## RegionOnline NA  
## RegionSouth Carolina NA  
## RegionTennessee NA  
## RegionTexas NA  
## RegionUtah NA  
## RegionWashington NA  
## CharacteristicHospital 5.678e-01  
## CharacteristicMetro 4.529e-01  
## CharacteristicMilitary 1.608e+00  
## CharacteristicNonUS NA  
## Characteristicnot set NA  
## CharacteristicOnline NA  
## CharacteristicSchool -1.302e+00  
## Characteristic2Army 1.957e-01  
## Characteristic2Home NA  
## Characteristic2Hospital NA  
## Characteristic2Marines 5.051e-01  
## Characteristic2Metro NA  
## Characteristic2Navy -1.120e+00  
## Characteristic2NonUS NA  
## Characteristic2not set NA  
## Characteristic2Online NA  
## Characteristic2School NA  
## Std. Error  
## (Intercept) 7.871e-01  
## InquirySourcenot set 1.522e+02  
## InquirySourceOnline Application 2.371e-01  
## InquirySourceOnline Event Registration 7.577e-01  
## InquirySourceOpen Information Session 1.150e+00  
## InquirySourceWeb Create Account - Default Create Account Form 1.563e+03  
## InquirySourceWeb, Create Account 2.615e-01  
## InquirySourceWeb, Inquiry 7.523e-01  
## inquired1 NA  
## MarketingCampaign(organic) 9.258e-02  
## MarketingCampaign(referral) 1.077e-01  
## MarketingCampaign1MBA students%2C 2015 1.075e+04  
## MarketingCampaignFY1516Grad\_Columbia 1.563e+04  
## MarketingCampaignFY1516Grad\_LittleRock 1.863e+04  
## MarketingCampaignFY1516Grad\_Orlando 1.863e+04  
## MarketingCampaignFY1516Grad\_SanAntonio 1.664e+04  
## MarketingCampaignFY1516Grad\_Webster 1.863e+04  
## MarketingCampaignFY1516STLUG\_Webster 1.863e+04  
## MarketingCampaignFY1617Grad\_Webster 1.585e+04  
## MarketingCampaignFY1617Grad\_Webster\_COL 1.076e+04  
## MarketingCampaignFY1617Grad\_Webster\_LIT 1.668e+04  
## MarketingCampaignFY1617Grad\_Webster\_ORL 1.075e+04  
## MarketingCampaignFY1617Grad\_Webster\_SAN 1.076e+04  
## MarketingCampaignFY1617Grad\_Webster\_STL 1.863e+04  
## MarketingCampaignglobalmaemails 5.357e+03  
## MarketingCampaignGraduate\_ALL 3.222e-01  
## MarketingCampaignGraduate\_SANA 1.863e+04  
## MarketingCampaignGraduate\_STL 1.261e+00  
## MarketingCampaignnot set 1.128e-01  
## MarketingCampaignUS\_Military\_endofyear 1.076e+04  
## MarketingMediumemail NA  
## MarketingMediumMicrosite NA  
## MarketingMediumnot set NA  
## MarketingMediumorganic NA  
## MarketingMediumPaidDisplay 1.521e+04  
## MarketingMediumPaidSocial NA  
## MarketingMediumRadioMarketing NA  
## MarketingMediumreferral NA  
## HasSource1 NA  
## GroupBusiness 6.139e-01  
## GroupCounseling 6.287e-01  
## GroupLeadership 6.180e-01  
## GroupMarketing and Communication 6.323e-01  
## GroupNursing and Health Care 6.251e-01  
## GroupTeaching and Education 6.250e-01  
## GroupTechnology 6.226e-01  
## GroupUndecided 6.407e-01  
## MetroAustria 5.172e-01  
## MetroCalifornia 1.218e+00  
## MetroCharleston 4.974e-01  
## MetroChicago 1.117e+00  
## MetroChina 1.107e+00  
## MetroColorado Springs 4.749e-01  
## MetroColumbia 4.255e-01  
## MetroDC 4.038e-01  
## MetroDenver 8.173e-01  
## MetroEl Paso 4.197e-01  
## MetroFayetteville 4.141e-01  
## MetroFort Leonard Wood 4.467e-01  
## MetroFort Smith Metro 1.034e+00  
## MetroGreece 3.541e+03  
## MetroGreenville 5.326e-01  
## MetroHinesville 5.962e-01  
## MetroHospital 5.677e-01  
## MetroIrvine 6.379e-01  
## MetroJacksonville 5.452e-01  
## MetroKansas City 3.994e-01  
## MetroLakeland 6.589e-01  
## MetroLawton 5.956e-01  
## MetroLittle Rock 4.974e-01  
## MetroLos Angeles 5.818e-01  
## MetroLouisville 6.669e-01  
## MetroMelbourne 5.218e-01  
## MetroMemphis 7.937e-01  
## MetroMerritt Island 6.771e-01  
## MetroMyrtle Beach 4.927e-01  
## MetroNetherlands 5.673e-01  
## Metronot set 1.949e+03  
## MetroOcala 5.979e-01  
## MetroOklahoma City 8.259e-01  
## MetroOnline 4.385e-01  
## MetroOrlando 4.746e-01  
## MetroPhoenix 7.076e-01  
## MetroRolla 7.445e-01  
## MetroSalt Lake City 5.408e-01  
## MetroSan Antonio 4.304e-01  
## MetroSavannah 7.635e-01  
## MetroSpokane 1.075e+04  
## MetroSt Louis 4.263e-01  
## MetroSt Petersburg 5.541e-01  
## MetroStudy Abroad 1.075e+04  
## MetroSumter 8.457e-01  
## MetroSwitzerland 5.475e-01  
## MetroWichita 2.814e+00  
## RegionArkansas NA  
## RegionCalifornia NA  
## RegionColorado NA  
## RegionDC NA  
## RegionFlorida NA  
## RegionGeorgia NA  
## RegionIllinois NA  
## RegionKansas NA  
## RegionMaryland NA  
## RegionMissouri NA  
## RegionNonUS NA  
## RegionNorth Carolina NA  
## Regionnot set NA  
## RegionOklahoma NA  
## RegionOnline NA  
## RegionSouth Carolina NA  
## RegionTennessee NA  
## RegionTexas NA  
## RegionUtah NA  
## RegionWashington NA  
## CharacteristicHospital 2.528e-01  
## CharacteristicMetro 1.742e-01  
## CharacteristicMilitary 2.731e-01  
## CharacteristicNonUS NA  
## Characteristicnot set NA  
## CharacteristicOnline NA  
## CharacteristicSchool 1.060e+00  
## Characteristic2Army 2.324e-01  
## Characteristic2Home NA  
## Characteristic2Hospital NA  
## Characteristic2Marines 6.138e-01  
## Characteristic2Metro NA  
## Characteristic2Navy 6.605e-01  
## Characteristic2NonUS NA  
## Characteristic2not set NA  
## Characteristic2Online NA  
## Characteristic2School NA  
## z value  
## (Intercept) -0.530  
## InquirySourcenot set -0.131  
## InquirySourceOnline Application -7.395  
## InquirySourceOnline Event Registration -5.659  
## InquirySourceOpen Information Session -1.953  
## InquirySourceWeb Create Account - Default Create Account Form 0.010  
## InquirySourceWeb, Create Account -10.045  
## InquirySourceWeb, Inquiry -11.791  
## inquired1 NA  
## MarketingCampaign(organic) -1.318  
## MarketingCampaign(referral) -1.960  
## MarketingCampaign1MBA students%2C 2015 -0.002  
## MarketingCampaignFY1516Grad\_Columbia -0.002  
## MarketingCampaignFY1516Grad\_LittleRock -0.001  
## MarketingCampaignFY1516Grad\_Orlando 0.000  
## MarketingCampaignFY1516Grad\_SanAntonio -0.002  
## MarketingCampaignFY1516Grad\_Webster -0.001  
## MarketingCampaignFY1516STLUG\_Webster -0.002  
## MarketingCampaignFY1617Grad\_Webster -0.001  
## MarketingCampaignFY1617Grad\_Webster\_COL 0.000  
## MarketingCampaignFY1617Grad\_Webster\_LIT -0.002  
## MarketingCampaignFY1617Grad\_Webster\_ORL -0.002  
## MarketingCampaignFY1617Grad\_Webster\_SAN 0.000  
## MarketingCampaignFY1617Grad\_Webster\_STL -0.001  
## MarketingCampaignglobalmaemails -0.003  
## MarketingCampaignGraduate\_ALL -1.375  
## MarketingCampaignGraduate\_SANA -0.002  
## MarketingCampaignGraduate\_STL -0.008  
## MarketingCampaignnot set -2.158  
## MarketingCampaignUS\_Military\_endofyear 0.000  
## MarketingMediumemail NA  
## MarketingMediumMicrosite NA  
## MarketingMediumnot set NA  
## MarketingMediumorganic NA  
## MarketingMediumPaidDisplay 0.001  
## MarketingMediumPaidSocial NA  
## MarketingMediumRadioMarketing NA  
## MarketingMediumreferral NA  
## HasSource1 NA  
## GroupBusiness 2.106  
## GroupCounseling 0.799  
## GroupLeadership 2.646  
## GroupMarketing and Communication 1.826  
## GroupNursing and Health Care 1.197  
## GroupTeaching and Education 1.962  
## GroupTechnology 2.128  
## GroupUndecided 0.998  
## MetroAustria -2.873  
## MetroCalifornia 1.448  
## MetroCharleston 1.038  
## MetroChicago 1.162  
## MetroChina -2.491  
## MetroColorado Springs -0.754  
## MetroColumbia -0.007  
## MetroDC 0.326  
## MetroDenver -1.418  
## MetroEl Paso -1.511  
## MetroFayetteville -2.103  
## MetroFort Leonard Wood 1.692  
## MetroFort Smith Metro 0.125  
## MetroGreece -0.005  
## MetroGreenville 0.014  
## MetroHinesville -0.038  
## MetroHospital 2.104  
## MetroIrvine 1.327  
## MetroJacksonville -0.473  
## MetroKansas City 1.293  
## MetroLakeland 0.605  
## MetroLawton -3.125  
## MetroLittle Rock -1.061  
## MetroLos Angeles -0.479  
## MetroLouisville -1.123  
## MetroMelbourne -0.260  
## MetroMemphis 1.922  
## MetroMerritt Island 1.093  
## MetroMyrtle Beach 0.647  
## MetroNetherlands -3.157  
## Metronot set -0.018  
## MetroOcala -0.188  
## MetroOklahoma City -0.935  
## MetroOnline 0.646  
## MetroOrlando 2.571  
## MetroPhoenix 0.776  
## MetroRolla 0.427  
## MetroSalt Lake City -0.232  
## MetroSan Antonio 0.486  
## MetroSavannah -2.397  
## MetroSpokane -0.001  
## MetroSt Louis 0.280  
## MetroSt Petersburg 0.350  
## MetroStudy Abroad -0.002  
## MetroSumter 0.088  
## MetroSwitzerland -2.498  
## MetroWichita 0.857  
## RegionArkansas NA  
## RegionCalifornia NA  
## RegionColorado NA  
## RegionDC NA  
## RegionFlorida NA  
## RegionGeorgia NA  
## RegionIllinois NA  
## RegionKansas NA  
## RegionMaryland NA  
## RegionMissouri NA  
## RegionNonUS NA  
## RegionNorth Carolina NA  
## Regionnot set NA  
## RegionOklahoma NA  
## RegionOnline NA  
## RegionSouth Carolina NA  
## RegionTennessee NA  
## RegionTexas NA  
## RegionUtah NA  
## RegionWashington NA  
## CharacteristicHospital 2.246  
## CharacteristicMetro 2.600  
## CharacteristicMilitary 5.887  
## CharacteristicNonUS NA  
## Characteristicnot set NA  
## CharacteristicOnline NA  
## CharacteristicSchool -1.228  
## Characteristic2Army 0.842  
## Characteristic2Home NA  
## Characteristic2Hospital NA  
## Characteristic2Marines 0.823  
## Characteristic2Metro NA  
## Characteristic2Navy -1.695  
## Characteristic2NonUS NA  
## Characteristic2not set NA  
## Characteristic2Online NA  
## Characteristic2School NA  
## Pr(>|z|)   
## (Intercept) 0.59629   
## InquirySourcenot set 0.89578   
## InquirySourceOnline Application 1.41e-13 \*\*\*  
## InquirySourceOnline Event Registration 1.52e-08 \*\*\*  
## InquirySourceOpen Information Session 0.05085 .   
## InquirySourceWeb Create Account - Default Create Account Form 0.99215   
## InquirySourceWeb, Create Account < 2e-16 \*\*\*  
## InquirySourceWeb, Inquiry < 2e-16 \*\*\*  
## inquired1 NA   
## MarketingCampaign(organic) 0.18761   
## MarketingCampaign(referral) 0.05003 .   
## MarketingCampaign1MBA students%2C 2015 0.99857   
## MarketingCampaignFY1516Grad\_Columbia 0.99819   
## MarketingCampaignFY1516Grad\_LittleRock 0.99915   
## MarketingCampaignFY1516Grad\_Orlando 1.00000   
## MarketingCampaignFY1516Grad\_SanAntonio 0.99821   
## MarketingCampaignFY1516Grad\_Webster 0.99908   
## MarketingCampaignFY1516STLUG\_Webster 0.99835   
## MarketingCampaignFY1617Grad\_Webster 0.99902   
## MarketingCampaignFY1617Grad\_Webster\_COL 0.99981   
## MarketingCampaignFY1617Grad\_Webster\_LIT 0.99825   
## MarketingCampaignFY1617Grad\_Webster\_ORL 0.99846   
## MarketingCampaignFY1617Grad\_Webster\_SAN 0.99991   
## MarketingCampaignFY1617Grad\_Webster\_STL 0.99914   
## MarketingCampaignglobalmaemails 0.99723   
## MarketingCampaignGraduate\_ALL 0.16902   
## MarketingCampaignGraduate\_SANA 0.99866   
## MarketingCampaignGraduate\_STL 0.99365   
## MarketingCampaignnot set 0.03093 \*   
## MarketingCampaignUS\_Military\_endofyear 0.99995   
## MarketingMediumemail NA   
## MarketingMediumMicrosite NA   
## MarketingMediumnot set NA   
## MarketingMediumorganic NA   
## MarketingMediumPaidDisplay 0.99899   
## MarketingMediumPaidSocial NA   
## MarketingMediumRadioMarketing NA   
## MarketingMediumreferral NA   
## HasSource1 NA   
## GroupBusiness 0.03524 \*   
## GroupCounseling 0.42405   
## GroupLeadership 0.00815 \*\*   
## GroupMarketing and Communication 0.06787 .   
## GroupNursing and Health Care 0.23146   
## GroupTeaching and Education 0.04976 \*   
## GroupTechnology 0.03335 \*   
## GroupUndecided 0.31828   
## MetroAustria 0.00407 \*\*   
## MetroCalifornia 0.14763   
## MetroCharleston 0.29934   
## MetroChicago 0.24510   
## MetroChina 0.01274 \*   
## MetroColorado Springs 0.45064   
## MetroColumbia 0.99463   
## MetroDC 0.74438   
## MetroDenver 0.15607   
## MetroEl Paso 0.13087   
## MetroFayetteville 0.03546 \*   
## MetroFort Leonard Wood 0.09063 .   
## MetroFort Smith Metro 0.90090   
## MetroGreece 0.99606   
## MetroGreenville 0.98876   
## MetroHinesville 0.97000   
## MetroHospital 0.03536 \*   
## MetroIrvine 0.18441   
## MetroJacksonville 0.63653   
## MetroKansas City 0.19605   
## MetroLakeland 0.54488   
## MetroLawton 0.00178 \*\*   
## MetroLittle Rock 0.28855   
## MetroLos Angeles 0.63222   
## MetroLouisville 0.26141   
## MetroMelbourne 0.79522   
## MetroMemphis 0.05457 .   
## MetroMerritt Island 0.27442   
## MetroMyrtle Beach 0.51776   
## MetroNetherlands 0.00160 \*\*   
## Metronot set 0.98591   
## MetroOcala 0.85063   
## MetroOklahoma City 0.34977   
## MetroOnline 0.51810   
## MetroOrlando 0.01014 \*   
## MetroPhoenix 0.43788   
## MetroRolla 0.66953   
## MetroSalt Lake City 0.81677   
## MetroSan Antonio 0.62685   
## MetroSavannah 0.01653 \*   
## MetroSpokane 0.99903   
## MetroSt Louis 0.77963   
## MetroSt Petersburg 0.72663   
## MetroStudy Abroad 0.99868   
## MetroSumter 0.92992   
## MetroSwitzerland 0.01250 \*   
## MetroWichita 0.39124   
## RegionArkansas NA   
## RegionCalifornia NA   
## RegionColorado NA   
## RegionDC NA   
## RegionFlorida NA   
## RegionGeorgia NA   
## RegionIllinois NA   
## RegionKansas NA   
## RegionMaryland NA   
## RegionMissouri NA   
## RegionNonUS NA   
## RegionNorth Carolina NA   
## Regionnot set NA   
## RegionOklahoma NA   
## RegionOnline NA   
## RegionSouth Carolina NA   
## RegionTennessee NA   
## RegionTexas NA   
## RegionUtah NA   
## RegionWashington NA   
## CharacteristicHospital 0.02470 \*   
## CharacteristicMetro 0.00933 \*\*   
## CharacteristicMilitary 3.94e-09 \*\*\*  
## CharacteristicNonUS NA   
## Characteristicnot set NA   
## CharacteristicOnline NA   
## CharacteristicSchool 0.21956   
## Characteristic2Army 0.39982   
## Characteristic2Home NA   
## Characteristic2Hospital NA   
## Characteristic2Marines 0.41056   
## Characteristic2Metro NA   
## Characteristic2Navy 0.09001 .   
## Characteristic2NonUS NA   
## Characteristic2not set NA   
## Characteristic2Online NA   
## Characteristic2School NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10893.1 on 13484 degrees of freedom  
## Residual deviance: 5792.9 on 13393 degrees of freedom  
## AIC: 5976.9  
##   
## Number of Fisher Scoring iterations: 18

Now that our model is made we want to look at some statistics on the predictions its making. We will also look at the average probility for each of our dependent variable outcomes.

predictallrecTrain <- predict(allrecLogModel, type = "response")  
summary(predictallrecTrain)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000000 0.0000000 0.0004092 0.1394000 0.2805000 1.0000000

tapply(predictallrecTrain, allrecTrainSet$enrolled, mean)

## 0 1   
## 0.08374953 0.48302484

So we are getting a much higher average probibility for enrolled which is a good sign. Now let's look at our confusion matrix from these predicted values. We are going to start by using a threshold value of 0.5.

table(allrecTrainSet$enrolled,predictallrecTrain >0.5)

##   
## FALSE TRUE  
## 0 11275 330  
## 1 1085 795

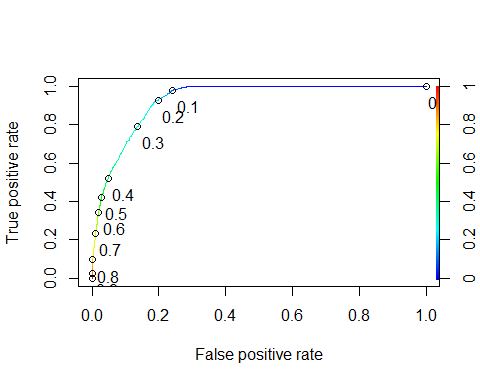
(11275+795)/(11275+795+1085+330)

## [1] 0.8950686

So our accuracy rate with a threshold of 0.5 is 89.5% which is already better than our baseline model. Now let's see if we can improve it by picking an optimized threshold value. We will do this by looking at the ROC curve.

## ROC curve and AUC

ROCRpred = prediction(predictallrecTrain, allrecTrainSet$enrolled)  
ROCRperf = performance(ROCRpred, "tpr","fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))



auc.tmp <- performance(ROCRpred,"auc"); auc <- as.numeric(auc.tmp@y.values)  
auc.tmp

## An object of class "performance"  
## Slot "x.name":  
## [1] "None"  
##   
## Slot "y.name":  
## [1] "Area under the ROC curve"  
##   
## Slot "alpha.name":  
## [1] "none"  
##   
## Slot "x.values":  
## list()  
##   
## Slot "y.values":  
## [[1]]  
## [1] 0.9283588  
##   
##   
## Slot "alpha.values":  
## list()