SESUG Paper 179-2019

Prediction of College Admission Trend Predictive Analytics® 2019 Final Paper

Yash Prakash, Oklahoma State University

ABSTRACT

The number of applications received by the universities and higher educational institutions is increasing every year. Out of these huge number of applications, limited number of applicants are provided with offer of admission and out of it, only few applicants accept the offer. Therefore, it is very important for the universities and institutions to offer the admission only to the prospective applicants who are more likely to join their universities and institutions. A prospective student can be identified by finding interactions of applicants with the college over phone, direct mails, e-mails and other communication channels.

This paper uses the power of predictive analytics to identify the factors influencing application submission by the applicants and build operationalized models to predict the application submission based on students' communication with the university via e-mail.

The data used in this research has been provided by Marketing and Student Communication Center, Oklahoma State University. The dataset contains all the email communications date wise between the university and the students of the past three years starting from application submission to confirmed admission. The data set also contains the applicants' demographic information such as Active City, Region, Postal code, Gender, Race, First generation, etc. Also, the email communication between the student and the University are classified into categories such as Prospects, Inquirers, Applicants, Admitted, and Confirmed. The data is prepared and analyzed using different SAS tools like SAS Enterprise Guide®, SAS Viya® and SAS Enterprise Miner®.

Variables such as Gender, Race, Inquiry Date, Application Date, Percent Clicked, Percent Opened and Application Submission Date are found to be the most important factors affecting the admission decision of the candidate.

INTRODUCTION

Every year large number of applicants start application for the under-graduate course in universities. But, very few of them completes the application. A large percentage of the applicants either drop the application or miss dead line for the application. The method of approaching prospective students is in question, as sending application to prospective students incurs a cost and accounts for the reputation of the university. Interested students approach the college admission authorities, but a large set of students who are not very interested in the application damage the economic balance of universities. A very important problem that universities are facing is to choose right candidate who are the potential applicants and find the university right place to enhance their academic knowledge.

This paper aims to evaluate some key factors that may be important to identify the correct set of candidates to approach. The process should be in two stage, first is to select right set of candidate and second would be to setup a regular communication channel between university and applicants to increase the chances of converting the candidate into applicant. Different factors such as Academic Talent, Race, Age, Hispanic and many more along with Email that are send to applicants plays important role in enhancement of the number of applicants. The email content that are send to the prospective applicants are same. But, those emails should be sent to the right set of applicants.

Some key factors have been identified which plays a major role in the application process are race of the applicants, region from which applicants belong, gender, Academic talent, OSU legacy, first generation, percentage delivered and percentage clicked

The raw data provided by the Marketing and student relation department, Oklahoma State University for the analysis contains many missing values that need attention. Also, there are number of variables that are considered redundant such as active region, active city and active zip code. Students after receiving first email from the university starts enquiring about the program, campus life, housing facilities, sports facilities, health care facilities and so on. It is very important to know, what was the action taken by the student after enquiring, so calculating the number of days between enquiry and application date was important. If the number of days are increasing, then how the enrollment status is affected. A new variable named Enquiry to application has been created to calculate the number of days.

DATA DESCRIPTION

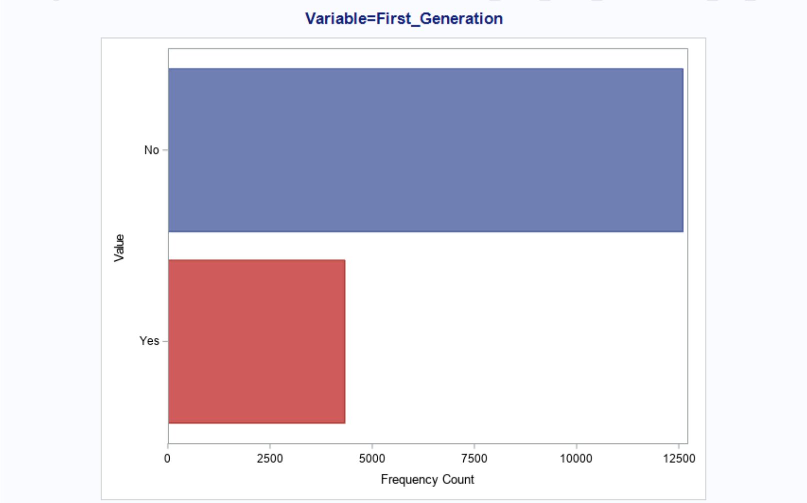
The data provided by Marketing department at Oklahoma State University contained approximately 130000 observations consists of three years from 2016 to 2018. There were total 38 variables available. Some of these variable is not of very much importance, as mentioned above, some were redundant. Also, some variables are irrelevant as they indicate the admission status, enrolment status etc. Some variables are the date wise first enquiry, second enquiry and so on. The text sent to individual students depends on the type of enquiries that student seeks. There were 10 demographic variables provided including Gender, Race, Hispanic etc. Data also provided the Delivered, clicked and opened rate. This information provides the effectiveness of email content. One variable was Active region, which indicates the region from which most of the applications are received. There were some missing values but very few that has been taken care off. Also based on the analysis demand some new variables were created.

DATA PREPERATION

The data contained considerable amount of missing value in some variable like date of enquiry, admission status etc. Also, those variables were not on research importance. Hence those variables were not considered in the analysis. Some of the important variables like gender also has missing values. Those missing values cannot be imputed based on any algorithms. Hence, left as it is. Out of 38 variables, 16 are selected as effective variables base on the result obtained from PROC MEANS, Stat Explore node, PROC SGPLOT and PROC FREQ. Defining level and converting binary into binary numeric values was done using SAS Enterprise Guide 7.1. Some missing values which are not in large number are removed and data set is partitioned into 70-30 ratio for training and validation purpose. Variable name student status is the target variable as it contains the information, whether the candidate has applied or not.

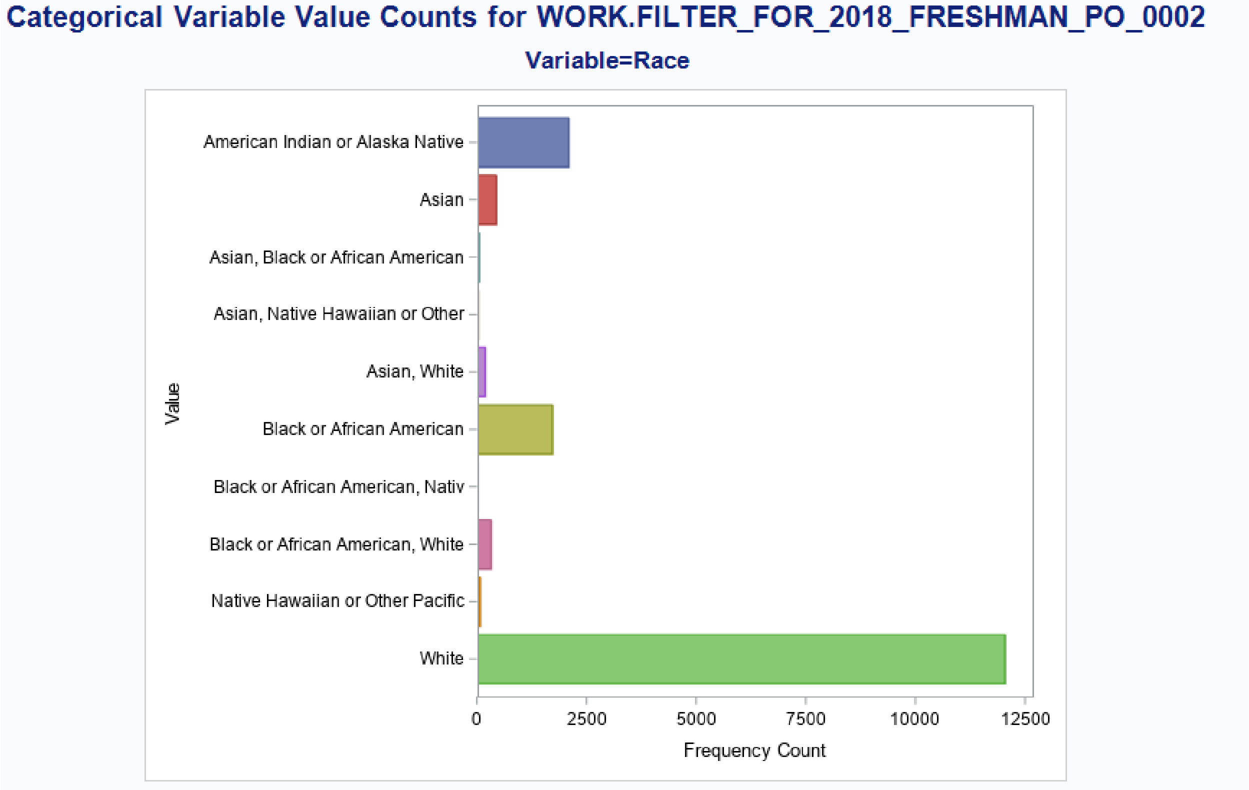
ANALYSIS

After data preparation, 16 variables kept in the data set for analysis. It was observed that those who were not the first generation were more inclined to application completion.



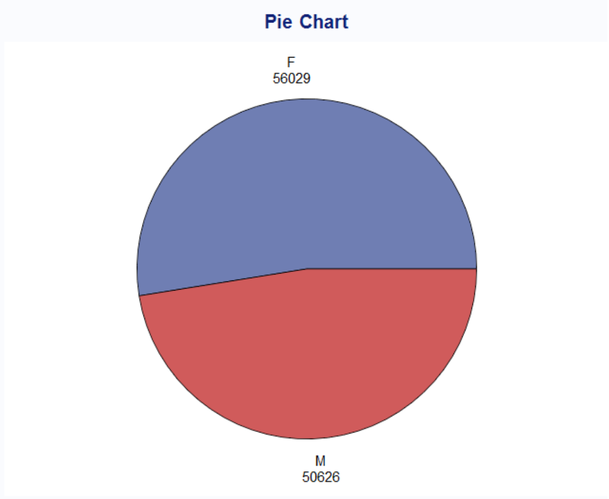
This result indicates that, the chances of applying those candidates whose parents have studied in Oklahoma State University is higher.

Also, the race distribution indicates that a large chunk of application has been received from white race candidate.

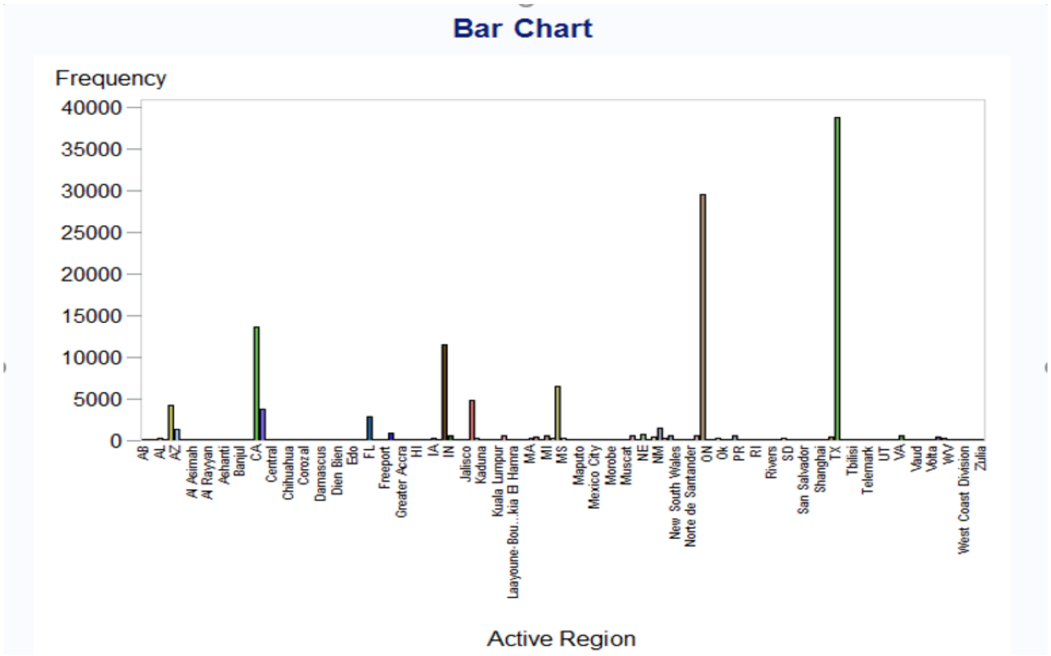


Hence, focusing on white people would be better. simultaneously other race should not be ignored. American Indian, Asian, Black or African American races should also be focused with different approach and offerings.

Gender wise application distribution indicates that female applicants are more than male applicants.

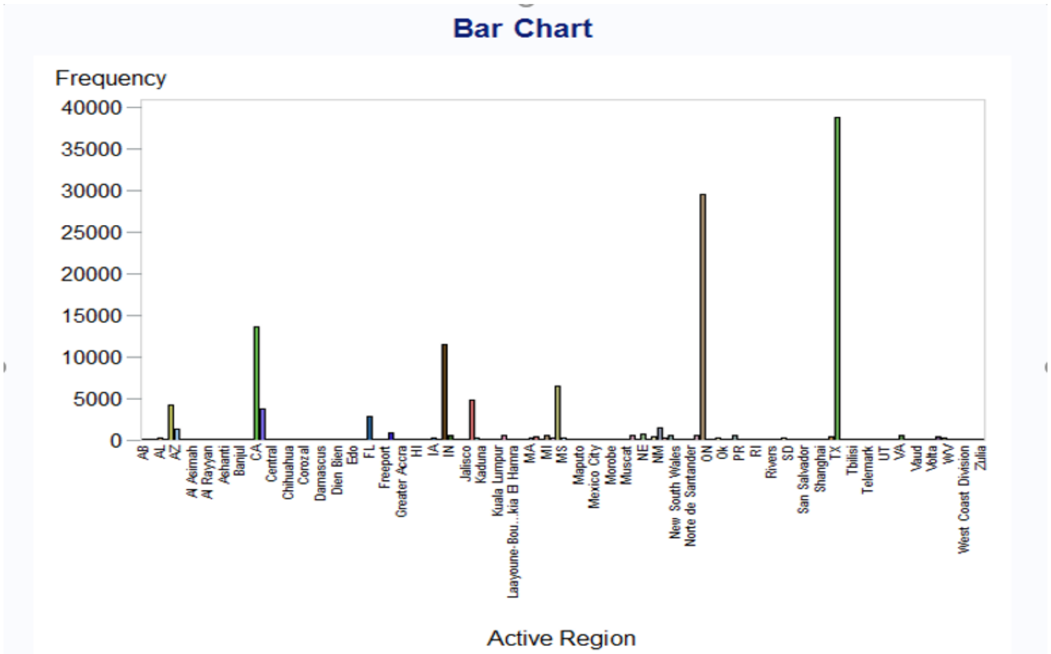


From descriptive analysis, it was observed that Texas and Oklahoma are the region from where maximum number of applications received. The data constitute year 2016, 2017 and year 2018. Similarly, there are states from where there is very less application have been received. So, region plays a major role and need to focused while approaching prospective applicants.

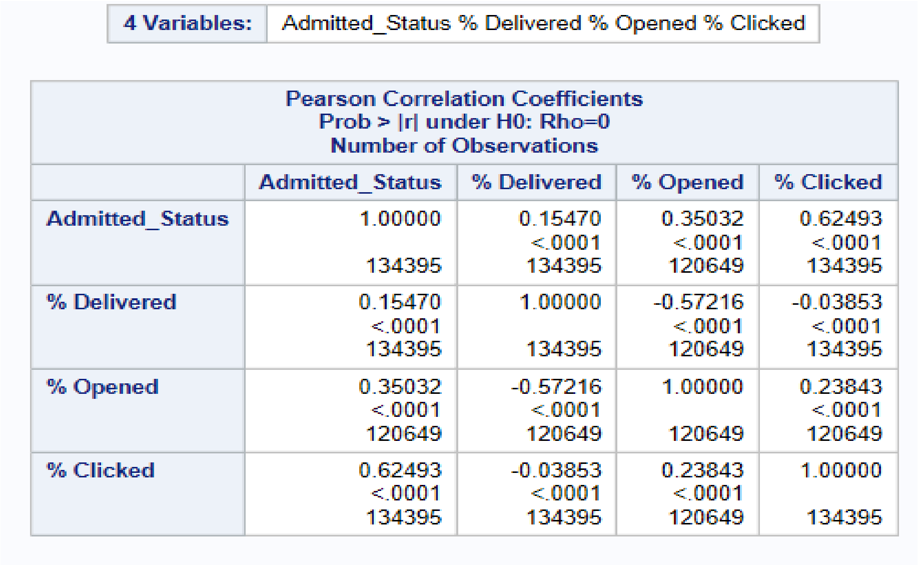


Enrolment status for other race has negligible significance. Next parameter that has been identified as important is the active region. If we look at the graph, Texas is the region from where maximum number of applications received. The data constitute year 2016, 2017 and year 2018. Similarly, there are states from where there is very less application have been received. So, region plays a major role and need to focused while approaching prospective applicants.

Region wise distribution of the applicants who successfully submitted application



Correlation among the selected student and the % Delivered emails, % clicked emails and % Opened email: The p-values for the correlation among each pair value is less than 0.05, so the correlation is significant. Only a certain percent of the promotional and informational email that are send to the applicants gets delivered. Some are by default gets delivered into spam. Out of those in boxed email, only a certain percent of the delivered emails is open and only very few emails get attention. The application completion and email delivered, clicked and open rate are highly correlated. If some action has been taken on any email, that indicates the interest of the application for the university. And the action increases the chances of the application completion.

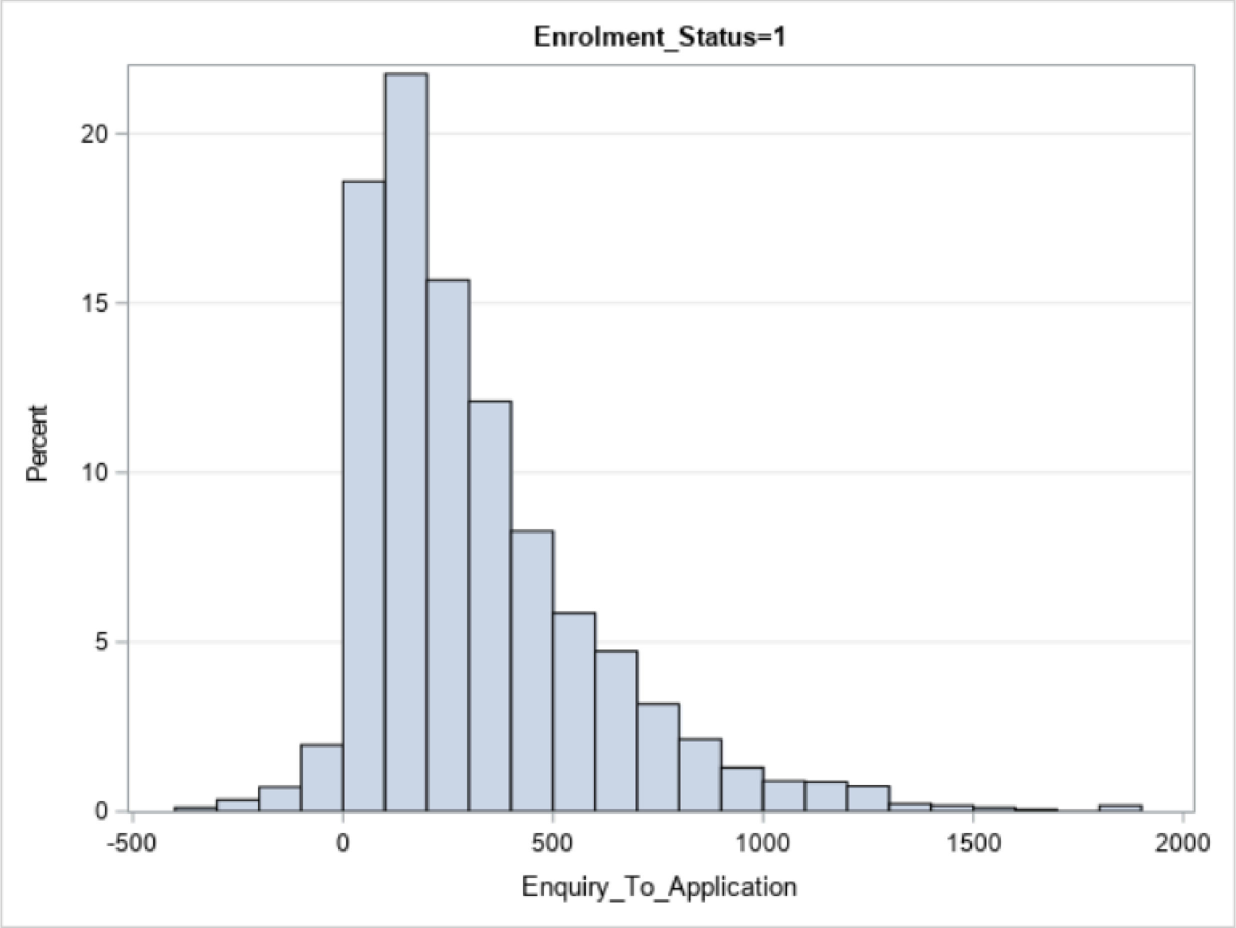


4 Variables:	Admitted_Status % Delivered % Opened % Clicked

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

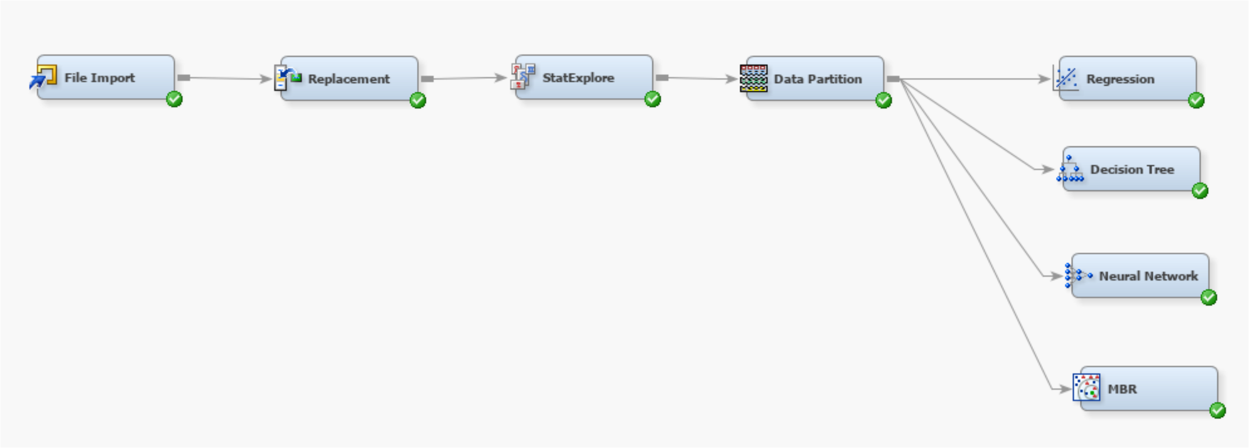
	Admitted_Status	**% Delivered**	**% Opened**	**% Clicked**
Admitted_Status	1.00000 134395	0.15470 <.0001 134395	0.35032 <.0001 120649	0.62493 <.0001 134395
% Delivered	0.15470 <.0001 134395	1.00000 134395	-0.57216 <.0001 120649	-0.03853 <.0001 134395
% Opened	0.35032 <.0001 120649	-0.57216 <.0001 120649	1.00000 120649	0.23843 <.0001 120649
% Clicked	0.62493 <.0001 134395	-0.03853 <.0001 134395	0.23843 <.0001 120649	1.00000 134395

Another observation indicates that if the duration between enquiry and application is not too large or not too small, i.e. if the duration of first enquiry and application submission is approximately six months, the applicant is more likely to join the school if offered admission.



This observation narrow-down our focus group. The descriptive analysis done above indicates further analysis for better prediction.

After descriptive analysis and data observation, different model were build on the training data set. The models were validated with the use of validation dataset.



The regression model was inappropriate. Because, a very important input "Gender" has a lot of missing values. Regression model by default filter out the missing values. So, in this case the model does not explain any-thing and misclassification rate or the average squared error are more than 85%.

Similarly, Neural Network model shows similar result, as our data set is not appropriate for Neural Network model.

The best model that is coming out is Decision Tree model with very less misclassification rate and Average Squared error rate. The sensitivity came out as 94%, which is no higher side. The reason behind this is the data set, which is available for only those students for which invitation has been sent. Hence, False Negative values are very less.

From the model analysis, it is observed that 'OSU Legacy' is the most important variable followed by variable name 'First Generation'. After First Generation, the next important variable is Hispanic, percentage clicked, academic talent, percentage delivered, gender, active region and Race in same order.

SUGGESTIONS

The analysis that has been done on the data-set is limited and need further consideration. The text email that are send to the prospective candidates is a general email. Hence, a personalized email to the candidates will impact positively and improve the chances of converting the prospect to the applicant. Because of the scarcity of text availability, text analytics was not performed in the analysis. Hence, a more depth insight is expected if text analytics is performed on the text that has been communicated with the prospective candidates. Over all the scope of this paper is very limited and there is good opportunity to draw more insights from the further analysis on larger data set if available.

CONCLUSION

From the analysis, it can be concluded that the list of email sender need to be considered on many parameters. Different factors like, location, gender, race, click trough rate, delivery rate, number of day gaps in enquiries all plays important role in deciding the application completion of the applicants. Hence, these factors should be considered on top priority while sending the application or advertisement email. Further, the application rate also depends on the reputation of the university and the quality of education. These factors should also be considered while sending the promotional email and should be taken into account while doing the analysis.

REFERENCES

1) Department of Marketing and Students Communication, Oklahoma State University, Student pool data [page 1 -5].

2) *Annual Student Data Report*, ASCO, Academic year 2017-18

3) Ma, Jennifer and Baum, Sandy. April 2016.*Trend in Community Colleges, Research Brief*

4) Patel, Pooja and Clinedinst , Melissa.2018. *State of College Admission*