ETM 538/638

DSS: DATA WAREHOUSING

Winter 2017

Instructor: Daniel Sagalowicz, Mike Freiling

Application of Data Mining to Virtual Viking Student Engagement Data

By

Devender Kaur

Jacob Yang

Rashi Tiwari

Sudipta Tripathy

Table of Content

Introduction

Virtual Vikings is a weekly e- newsletter sent to Portland state university students by the internal office of university Communications every Sunday. The newsletter is published in the fall, winter and spring of each academic school year. It contains information about the recent past week and upcoming week in form of snippets and links .The main aim of the Virtual Viking is to keep students engaged with the campus news, events, achievements etc.

This project was done on the data provided by the BI team at PSU which manages the enterprise data warehouse and reporting suite (IBM Cognos) along with the university's budgeting and planning software (IBM Cognos TM1). The enterprise data warehouse consists of data from the university's ERP system Banner along with data brought in from additional campus systems such as the campus CRM Talisma.

One of the Key business Goals of the Analytics Department at Portland State University is to engage students in attempts to help them achieve better outcomes. This is done through reporting all sorts of data being collected by the University.

Project Objective

This project is to analyze whether there is a correlation between students opening and clicking through URLs in the Virtual Viking email campaign and their performance that term in attempt to determine if engagement with the Virtual Viking email campaign is indicative of their engagement with the university and their studies.

Virtual Viking email is sent every Sunday from domain **virtualv@pdx.edu** and contains information about the recent past week and upcoming week in form of snippets and links. The main aim of the Virtual Viking is to keep students engaged with the campus news, events, achievements etc.

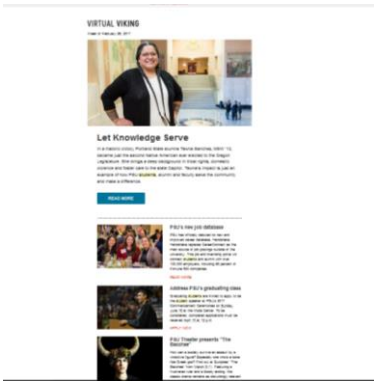
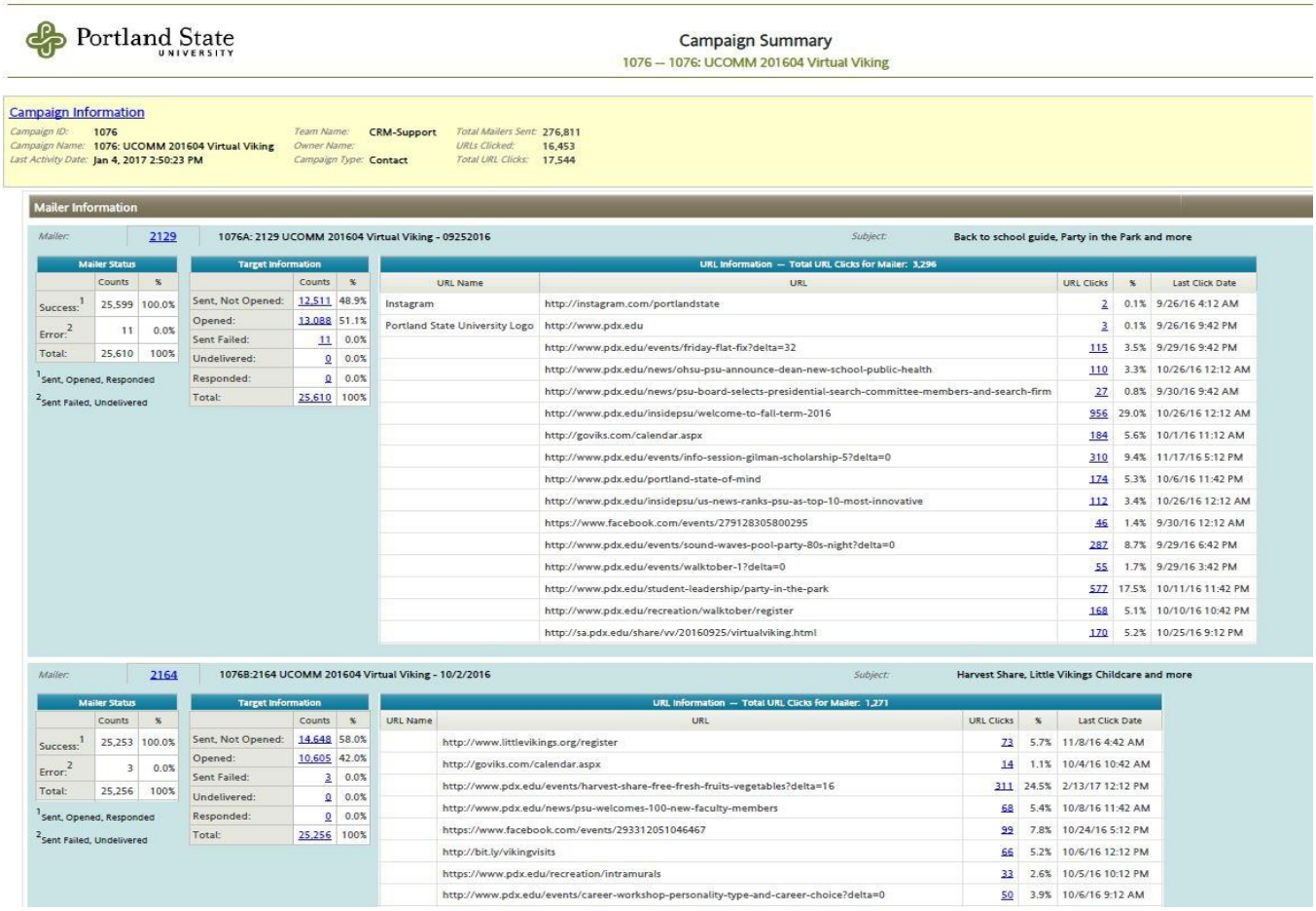


Figure 1: Sample virtual Viking Newsletter Email

Current Approach

Business doesn't have any analytics tool or has any analysis mechanism in place. IBM Cognos is used to report the number of emails sent, the number of emails read and unread.

Currently there are no systems to analyze the correlation between a student engagement and their academic proficiency.



Portland State UNIVERSITY

Campaign Summary

1076 -- 1076: UCOMM 201604 Virtual Viking

Campaign Information

Campaign ID:	1076	Team Name:	CRM-Support	Total Mailers Sent:	276,811
Campaign Name:	1076: UCOMM 201604 Virtual Viking	Owner Name:		URLs Clicked:	16,453
Last Activity Date:	Jan 4, 2017 2:50:23 PM	Campaign Type:	Contact	Total URL Clicks:	17,544

Mailer Information

Mailer: 2129 — 1076A: 2129 UCOMM 201604 Virtual Viking - 09252016 — Subject: Back to school guide, Party in the Park and more

Mailer Status

	Counts	%
Success:[1]	25,599	100.0%
Error:[2]	11	0.0%
Total:	25,610	100%

[1] Sent, Opened, Responded

[2] Sent Failed, Undelivered

Target Information

	Counts	%
Sent, Not Opened:	12,511	48.9%
Opened:	13,088	51.1%
Sent Failed:	11	0.0%
Undelivered:	0	0.0%
Responded:	0	0.0%
Total:	25,610	100%

URL Information -- Total URL Clicks for Mailer: 3,296

URL Name	URL	URL Clicks	%	Last Click Date
Instagram	http://instagram.com/portlandstate	2	0.1%	9/26/16 4:12 AM
Portland State University Logo	http://www.pdx.edu	3	0.1%	9/26/16 9:42 PM
	http://www.pdx.edu/events/friday-flat-fix?delta=32	115	3.5%	9/29/16 9:42 PM
	http://www.pdx.edu/news/ohsu-psu-announce-dean-new-school-public-health	110	3.3%	10/26/16 12:12 AM
	http://www.pdx.edu/news/psu-board-selects-presidential-search-committee-members-and-search-firm	27	0.8%	9/30/16 9:42 AM
	http://www.pdx.edu/insidepsu/welcome-to-fall-term-2016	956	29.0%	10/26/16 12:12 AM
	http://goviks.com/calendar.aspx	184	5.6%	10/1/16 11:12 AM
	http://www.pdx.edu/events/info-session-gilman-scholarship-5?delta=0	310	9.4%	11/17/16 5:12 PM
	http://www.pdx.edu/portland-state-of-mind	174	5.3%	10/6/16 11:42 PM
	http://www.pdx.edu/insidepsu/us-news-ranks-psu-as-top-10-most-innovative	112	3.4%	10/26/16 12:12 AM
	https://www.facebook.com/events/279128305800295	46	1.4%	9/30/16 12:12 AM
	http://www.pdx.edu/events/sound-waves-pool-party-80s-night?delta=0	287	8.7%	9/29/16 6:42 PM
	http://www.pdx.edu/events/walktober-1?delta=0	55	1.7%	9/29/16 3:42 PM
	http://www.pdx.edu/student-leadership/party-in-the-park	577	17.5%	10/11/16 11:42 PM
	http://www.pdx.edu/recreation/walktober/register	168	5.1%	10/10/16 10:42 PM
	http://sa.pdx.edu/share/vv/20160925/virtualviking.html	170	5.2%	10/25/16 9:12 PM

Mailer: 2164 — 1076B:2164 UCOMM 201604 Virtual Viking - 10/2/2016 — Subject: Harvest Share, Little Vikings Childcare and more

Mailer Status

	Counts	%
Success:[1]	25,253	100.0%
Error:[2]	3	0.0%
Total:	25,256	100%

[1] Sent, Opened, Responded

[2] Sent Failed, Undelivered

Target Information

	Counts	%
Sent, Not Opened:	14,648	58.0%
Opened:	10,605	42.0%
Sent Failed:	3	0.0%
Undelivered:	0	0.0%
Responded:	0	0.0%
Total:	25,256	100%

URL Information -- Total URL Clicks for Mailer: 1,271

URL Name	URL	URL Clicks	%	Last Click Date
	http://www.littlevikings.org/register	73	5.7%	11/8/16 4:42 AM
	http://goviks.com/calendar.aspx	14	1.1%	10/4/16 10:42 AM
	http://www.pdx.edu/events/harvest-share-free-fresh-fruits-vegetables?delta=16	311	24.5%	2/13/17 12:12 PM
	http://www.pdx.edu/news/psu-welcomes-100-new-faculty-members	68	5.4%	10/8/16 11:42 AM
	https://www.facebook.com/events/293312051046467	99	7.8%	10/24/16 5:12 PM
	http://bit.ly/vikingvisits	66	5.2%	10/6/16 12:12 PM
	https://www.pdx.edu/recreation/intramurals	33	2.6%	10/5/16 10:12 PM
	http://www.pdx.edu/events/career-workshop-personality-type-and-career-choice?delta=0	50	3.9%	10/6/16 9:12 AM

Figure 2: Sample Cognos Report

Data set and Variable Selection

For analysis, out of different type of students which is a whole data set, a subset of Undergraduate students are being considered and of different terms, fall of 2016 has been considered.

Virtual_viking_random.csv was the file with undergrad students who have received mailers through the Virtual Viking email campaign for fall 2016 term. Students are identified through an identifier which has been randomized for privacy.

Campaign_students_random.csv had more information on the student's. This is to give perspective on the student's performance for fall 2016 term and whether they went on to register for winter 2017.

For the scope of our analysis we combined the two csv files to get a picture of how the two files are related and introduced a few derived columns like Count of emails read, count of emails open, count of emails not opened, count of emails failed and GPA Level. The total data set that was used was 18089.
Apart from that we made buckets of varying sizes for different attributes which have been described below in the sections that they have been used.

The key terms and their definitions have been explained in appendix A

The columns and their descriptions are explained in Table I and Table II

Table I : Virtual-Viking-Random- Column Attribute

COLUMN NAME	Data Type	Miss/Null	Description
CAMPAIGN_TALISMA_ID	Integer	No	Type of email campaign ID
CAMPAIGN_NAME	Text	No	Type of email campaign in which there can be many mailers sent.
RANDOM_UID	Integer	No	This is the "Foreign Key" field linked with campaign_student table. This id is randomized identifier for the student.
MAILER_TALISMA_ID	Integer	No	Email ID on with unique content and subject that was sent to students.
MAILER_NAME	Text	No	Email Name related to the MAILER_TALISMA_ID
MAILER_SUBJECT	Text	No	Email Subject related to the MAILER_TALISMA_ID
MAILER_STATUS_DESC	Text	No	Sent(Not open) or Opened
MAILER_STATUS_DATE	Date	No	Email status date and time, sent or opened
MAIL_TALISMA_CREATED_DATE	Date	No	Email created date and time
URL_TALISMA_ID	Integer	Yes	URL Link ID clicked by student
URL	Text	Yes	URL Address of the URL_TALISMA_ID
URL_NAME	Text	Yes	URL description of the URL_TALISMA_ID
URL_NUMBER_OF_CLICKS	Integer	Yes	Number of clicks for the URL

Table II : Campaign-Student-Random - Column Attributes

COLUMN NAME	Data Type	Miss/Null	Description
RANDOM_UID	Integer	No	This is the "Primary Key" field for campaign_student table. This id is randomized identifier for the student. This field links to the virtual_viking_email_data table, which contains detail records for each email.
TERM_GPA	Number	No	This is calculated by taking the number of grade points.
TERM_CREDITS	Integer	No	Credit hours taking classes for the term based on the number of "contact hours" per week in class
GRADUATED_YN	Yes/No	No	Graduation after this term.
REGISTERED_NEXT_TERM_YN	Yes/No	No	Student enrollment in the next term.
STUDENT_LEVEL	Text	No	Undergraduate or graduate
STUDENT_CLASSIFICATION	Integer	No	Grades of students by number
STUDENT_CLASSIFICATION_DESC	Text	No	Grades of students by text
ACADEMIC_STANDING	Integer	No	Student's academic status based on cumulative and current grade by number
ACADEMIC_STANDING_DESC	Text	No	Student's academic status based on cumulative and current grade by text
ADMIT_ACADEMIC_PERIOD	Integer	No	Admitted term for the study

Initial data analysis

We wanted to find the high predictors that indicate student engagement levels and so an initial manual analysis of the data was done before we used any of the data mining algorithms .These Indicators are further used while doing the analysis.

The tables were connected so that the student attributes and the newsletter attributes could be analyzed together .The ERD for the tables relationship is represented below

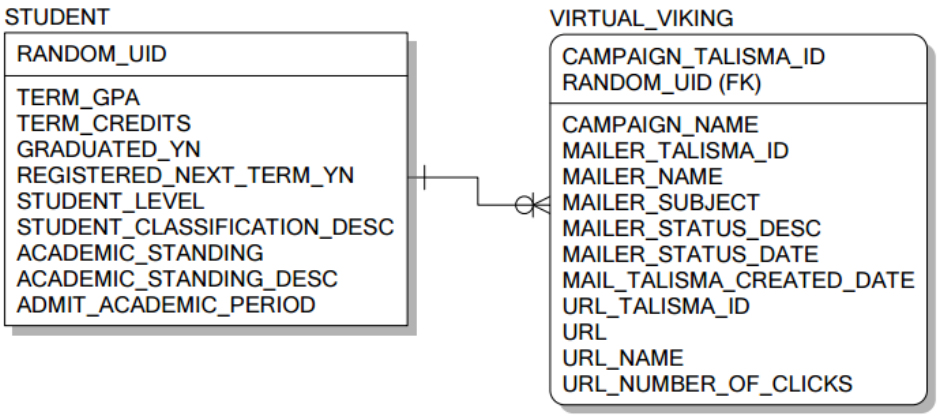


Figure 3: ERD Diagram Representing the Relations Between the tables

We then set out to finding patterns manually using pivot tables and data .To identify the distribution of students in different years was very important since that would be a big indicator of who could be the largest audience of the emails , based purely on numbers .The distribution of Students is shown in figure 4. Number of students in their senior year, is the highest followed by junior and sophomore.

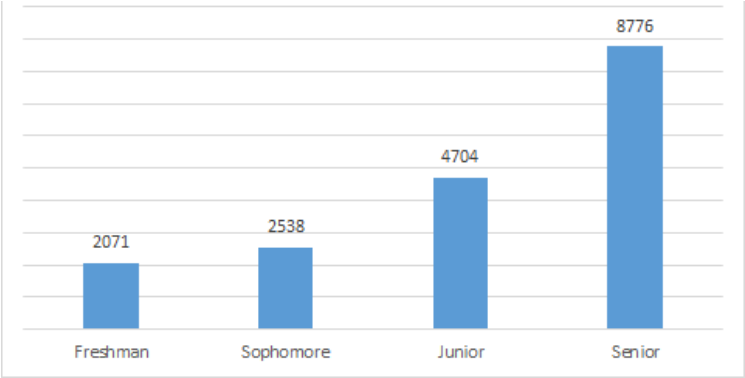


Figure 4: Student Count Distribution by year

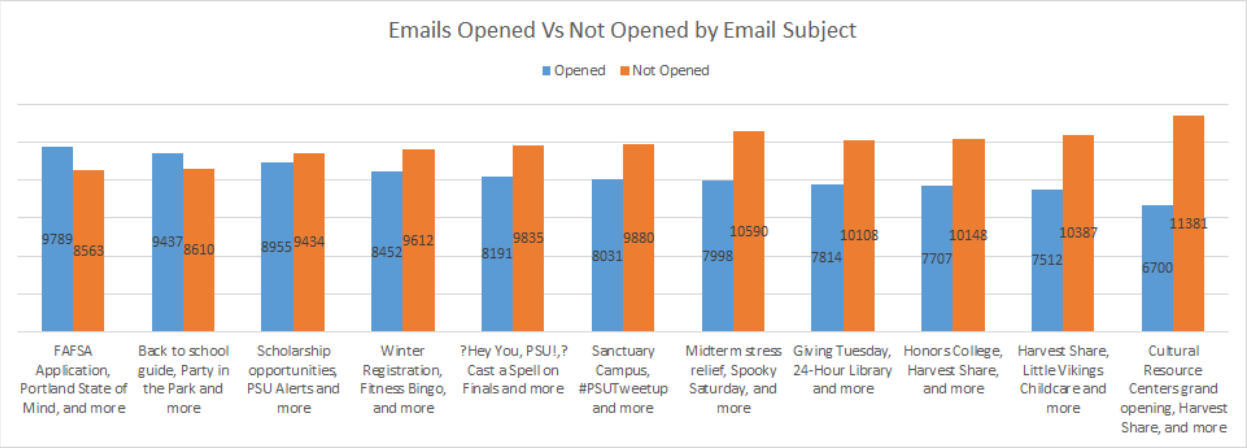


Figure 5: Student population distribution

The data was then analyzed based on what emails were the most popular amongst students. This was based on the fact that from the emails that were opened which were the ones that were accessed the most. Party in the park, FAfsa applications and scholarship applications were the most popular amongst students. Harvest share emails were one of the most infrequently read emails amongst others.

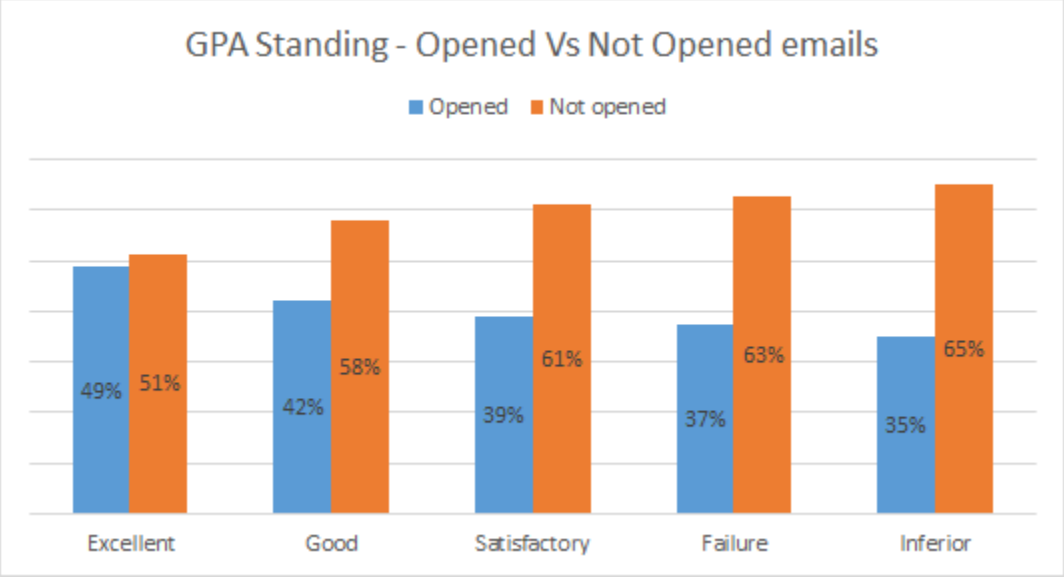


Figure 6: Opened vs Unopened Mails and Student GPA Level Correlation

Figure 6 here represents the distribution of students based on their GPA levels and how often they read the virtual Viking emails. It is evident that the students in the excellent and good category read much more emails compared to those in the other three categories. The Difference between opened and unopened Emails was the most apparent in the Inferior and failed Grade levels.

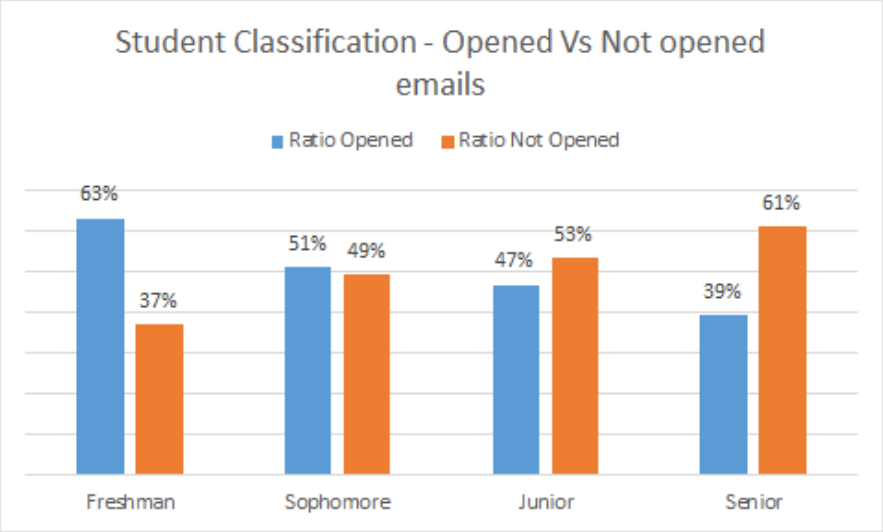


Figure 7: Opened vs Unopened Mails and Student Classification Correlation

Similarly the data was analyzed for which class of students read the most emails and the results are as represented in Figure 7.freshman and sophomore students were the biggest readers of the Virtual Vikings newsletters whereas senior students were reading these emails less often.

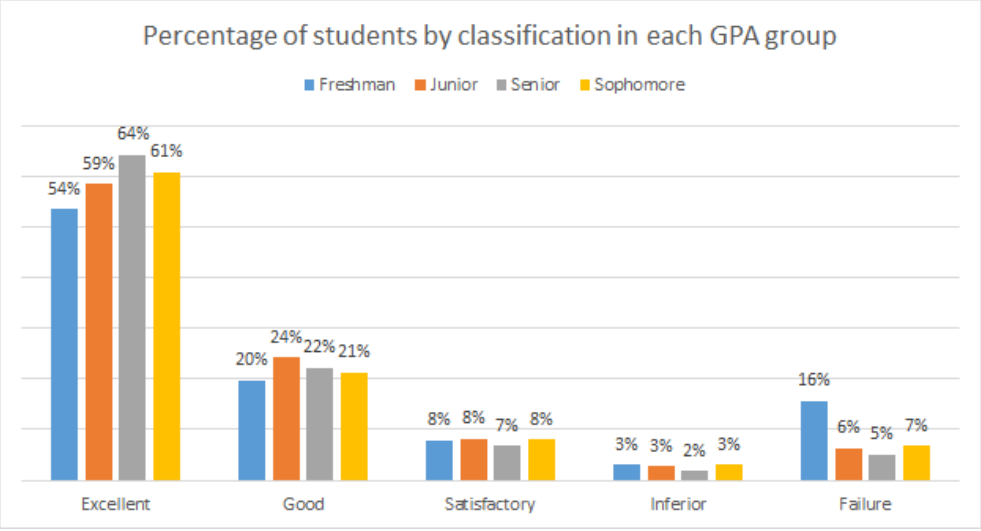


Figure 8: Student Academic Standing

Figure 8 above shows the distribution of students according to their academic standing and their frequency of reading the Virtual Viking email. It is very evident that the high performing students i.e. those in the Excellent and Good categories, wanted to be actively involved with ongoing campus activities and hence had the most engagement out of the group.

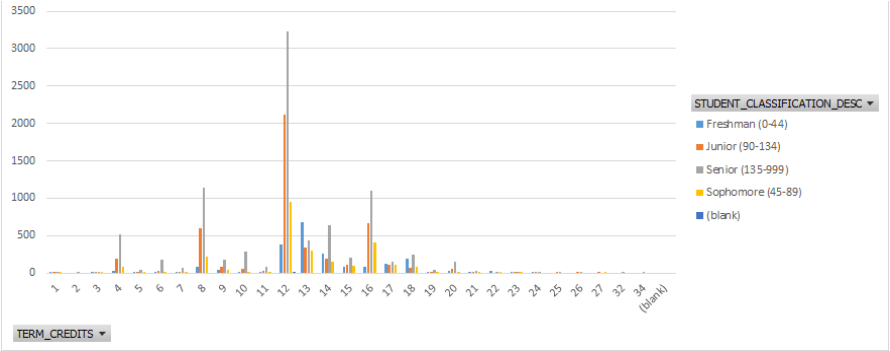
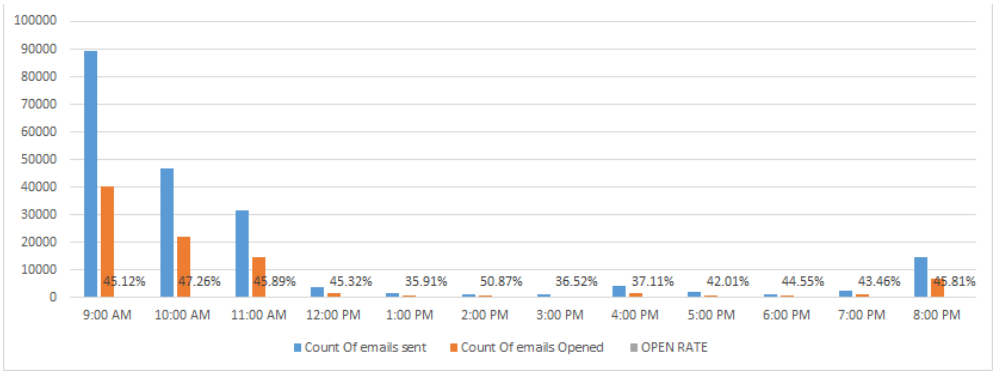


Figure 9: Term credits classified by student Yea distribution

Distribution Of student and the credits taken that quarter are shown in figure 9. On an average students with 12-17 credits and in the juniors and seniors years read the most emails.

We then analyzed the data based on what time the emails were sent and when they were actually read .Since the Mailing Job runs in batches therefore not emails were sent at the same time. The Columns had timestamps to help analyze these attributes



.

Figure 10: Rate of Emails read based on the sent Time

Figure 10 here gives the distribution of emails sent by time and how many of them are opened depending on the time of the day they were sent .Most of the emails being Opened are the ones that were sent early in the day and then the pattern falls to a low Read count as the day progresses, however it picks up for the emails being sent at 8Pm in the evening.

This would be helpful in finding a suitable time to send most of the emails based on how many of them are being read early in the day .

We also did an analysis on how soon do students tend to open the Virtual Viking emails and we found out that most often people read more emails within 12 hours to 1 day of it being sent.

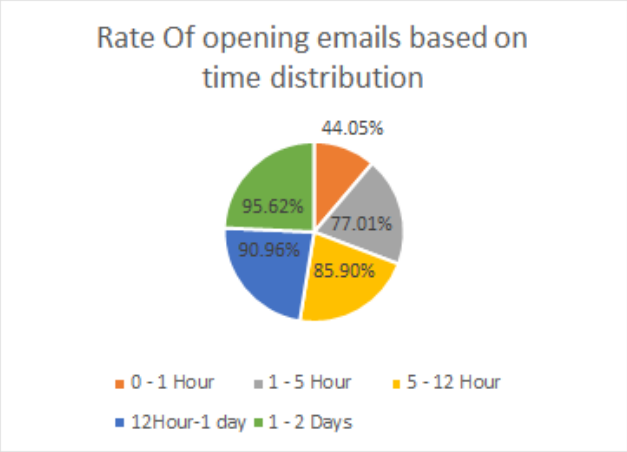


Figure 11: Rate of emails being opened by Time distribution

Figure 11 shows the amount of emails being opened within a few hours of being sent. The Highest Emails are opened within 1-2 days of being sent, which shows that people do not treat the virtual Viking emails as top priority, however 44% of the emails that do get opened are the ones being accessed in 0-1 Hour. Almost 91 % of read emails are opened within 12 hour-24 hours

Data Mining Methodology

We came up with a number of algorithm that were well suited for our analysis, however we used three that gave us better results. The reason for selecting these algorithms was to

1. Find a pattern in the existing data
2. Form a predictive model for new data for subsequent terms.

The following Grade level buckets have been used all throughout for the Algorithms, the rest of the buckets used in different sections are shown in the individual sections.

Table III: Grade Level Buckets

GPA Desc	Rule
Excellent	GPA between >3 & <= 4
Good	GPA between >2 & <=3
Inferior	GPA between >0 & <=1
Failure	GPA = 0

Table IV: Email Read Count Buckets

Email Classification	Rule
A	Emails read is 0
B	Emails read >0<=3
C	Emails read >3<=6
D	Emails read >6<=11
E	Emails read >11

Algorithms used

1. The 1R algorithm

The 1R algorithm was used since we wanted to come up with one single rule that suggests a pattern on the basis of a single predictive attribute.

This algorithm was used to understand relationship between Student performance for the term (GPA_Levels) and the Virtual Viking emails read.

Count of STUDENT_CLASSIFICATION	Column Labels										
Row Labels	**A**	**B**	**C**	**D**	**E**	**Grand Total**	**Max(A,B,C,D,E)**	**MAX/TOTAL**	**TOTAL-MAX**	**Error Rate**	
Sat	266	435	268	371	21	1361	435	0.31961793	926	0.680382072	B
Excellent	1468	2913	2196	3967	420	10964	3967	0.3618205	6997	0.638179497	D
Fail	421	344	157	319	10	1251	421	0.33653078	830	0.663469225	A
Good	750	1208	772	1268	82	4080	1268	0.31078431	2812	0.689215686	D
Inf	108	139	76	106	3	432	139	0.32175926	293	0.678240741	B
Grand Total	**3013**	**5039**	**3469**	**6031**	**536**	**18088**	6031	0.33342548	12057	0.666574525	D

Students with excellent and good grades read emails between 6 and 11 with the error of .63 and .68. This shows high engagement through emails with student with GPA >3

Students with satisfactory and inferior grades appear to read 3 emails with error of .68 and .67

Students who fail appears not to read the emails.

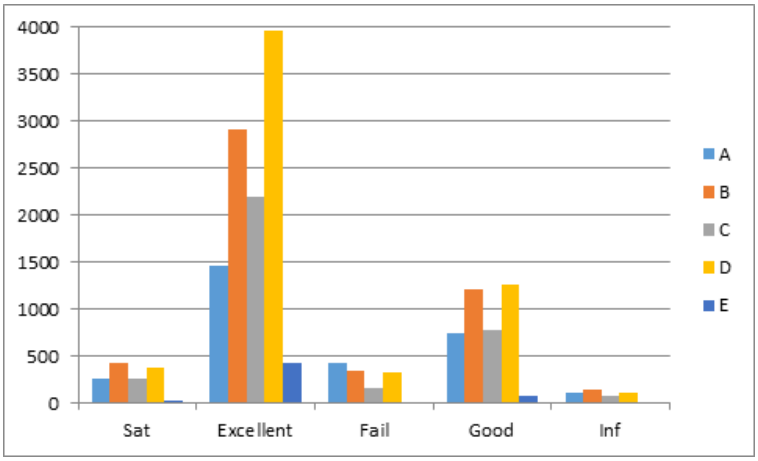


Figure 12 : Student performance and count of emails Read

2. Covering Algorithm

Covering is a Rule Based algorithm that works by concentrating on a particular class at a time and by maximizing the probability of the desired classification.

Applying the Covering algorithm on the Virtual Viking data, we did analysis on the level of email activity as described in table IV .For the purpose of this report, the analysis done below are For level E & D which are the highest level of student engagement (opening and checking a large number of newsletter sent throughout the term). The best rules were selected for every Iteration.

The term credits were segregated into buckets as

Table V: Credit Level Buckets

Term credit level	Rule
Low	Term Credit 5 or below
Average	Term Credit between 5 & <=12
High	Term Credit between 12 & 20
Very High	Term Credit between 20 &34

Level E

Antecedent
Student Classification
Term Levels
GPA levels
Consequent = Level E

Support: Number of cases that match all antecedents

Confidence: Total Number of cases in the antecedents divided by the Total number of cases for that instance
The Algorithm has been done in a number of Iterations as mentioned below
1st level Iteration

The attributes and their values are shown in the table below

Attribute	Value	Support	Confidence (E)
Student_Classification	freshman	2071	5.3%
Student_Classification	Junior	4704	3.4%
Student_Classification	Senior	8776	1.9%
Student_Classification	Sophomore	2537	3.9%
Term credit levels	Average	9911	2.6%
Term credit levels	High	7100	3.7%
Term credit levels	Very High	966	1.3%
Term credit levels	Low	112	5.4%
GPA Level	Sat	1361	1.5%
GPA Level	Excellent	10965	3.8%
GPA Level	Fail	1251	0.8%
GPA Level	Good	4080	2.0%
GPA Level	Inf	432	0.7%

Figure 13

The highest Confidence is for 5.4% for term credit level = Low, however the support for the same is very low so we are going to take the second best value that is student-classification = Freshman and consider only those cases. Here we found the best rule to be GPA_level =Excellent with a high confidence rating of 76.4%

Attribute	Value	Total Count	Confidence (E)
Term credit levels	Average	110	21.8%
Term credit levels	Low	110	1.8%
Term credit levels	Very High	110	1.8%
Term credit levels	High	110	74.5%
GPA Level	Sat	110	5.5%
GPA Level	Excellent	110	76.4%
GPA Level	Fail	110	1.8%
GPA Level	Good	110	16.4%

Figure 14

Ongoing down one level further and analyzing only freshman with GPA_Level = excellent.

Attribute	Value	Total Count	Support (E)	Confidence (E)
Term credit levels	Average	84	15	17.9%
Term credit levels	High	84	66	78.6%
Term credit levels	low	84	1	1.2%
Term credit levels	Very High	84	2	2.4%

Figure 15

Rule1

If student classification = freshman and Term credit level = High and GPA_Level = Excellent then Email Read = E

II Level Iteration

Removing the 84 records from the first rule above and analyzing the results.

Attribute	Value	Support	Confidence (E)
Student_Classification	freshman	2005	2.2%
Student_Classification	Junior	4704	3.4%
Student_Classification	Senior	8776	1.9%
Student_Classification	**Sophomore**	**2537**	**3.9%**
Term credit levels	Average	9911	2.6%
Term credit levels	High	7034	2.8%
Term credit levels	Very High	966	1.3%
Term credit levels	Low	112	5.4%
GPA Level	Sat	1361	1.5%
GPA Level	Excellent	10899	3.2%
GPA Level	Fail	1251	0.8%
GPA Level	Good	4080	2.0%
GPA Level	Inf	432	0.7%

Figure 16

Using the rule If student classification = sophomore and going down a level further.

Attribute	Value	Total Count
Term credit levels	Average	99
Term credit levels	Low	99
Term credit levels	Very High	99
Term credit levels	High	99
GPA Level	Sat	99
GPA Level	**Excellent**	**99**
GPA Level	Fail	99
GPA Level	Good	99
GPA Level	Inf	99

Figure 17

The rule now states If student classification = sophomore and GPA_Level = Excellent.
Ongoing down one level further for the above rule

Attribute	Value	Total Count
Term credit levels	Average	81
Term credit levels	**High**	**81**
Term credit levels	low	81
Term credit levels	Very High	81

Figure 18

Rule 2: If student classification = sophomore and Term credit level = High and GPA_Level = Excellent then Email Read = E

Similarly the rest of the rules for Consequent = E

Rule 3: If student classification = Junior and Term credit level = Average and GPA_Level = Excellent then Email Read = E

If student classification = freshman and Term credit level = High and GPA_Level = Excellent then Email Read = E
If student classification = sophomore and Term credit level = High and GPA_Level = Excellent then Email Read = E
If student classification = Junior and Term credit level = Average and GPA_Level = Excellent then Email Read = E
If student classification = Senior and Term credit level = High and GPA_Level = Excellent then Email Read = E
If student classification = Junior and Term credit level = High and GPA_Level = Excellent then Email Read = E
If Student classification = sophomore and Term credit level = Average and GPA_Level = Excellent then Email Read = E

This relates to the fact that students with high level of interest in College activity and those that are reading the virtual Viking emails regularly are the most academically well performing students. The dataset above gave all results for GPA_Level = Excellent and a good credit level for the term, both indicators of academic proficiency.

Level D

Antecedent

Student Classification
Term Levels
GPA levels
Consequent = Level D

Attribute	Value	Support	Confidence D
Student_Classification	**freshman**	**2071**	**52.1%**
Student_Classification	Junior	4704	34.1%
Student_Classification	Senior	8776	26.9%
Student_Classification	Sophomore	2537	38.9%
Term credit levels	Average	9911	29.5%
Term credit levels	**High**	**7100**	**40.5%**
Term credit levels	Very High	966	18.0%
Term credit levels	**Low**	**112**	**49.1%**
GPA Level	Sat	1361	27.3%
GPA Level	Excellent	10965	36.2%
GPA Level	Fail	1251	25.5%
GPA Level	Good	4080	31.1%
GPA Level	Inf	432	24.5%

Figure 19

The highest Confidence is for Student Classification= freshman, and on using only those records we get

Freshman			
Attribute	Value	Total Count	Confidence (E)
Term credit levels	Average	1078	19.4%
Term credit levels	Low	1078	76.3%
Term credit levels	Very High	1078	1.4%
Term credit levels	High	1078	2.9%
GPA Level	Sat	1078	5.9%
GPA Level	**Excellent**	**1078**	56.7%
GPA Level	Fail	1078	15.9%
GPA Level	Good	1078	19.2%
GPA Level	inf	1078	2.3%

Figure 20

The Student Classification= Freshman & GPA_Level = Excellent have the following records

Attribute	Value	Total Count	Suppor	Confidence (E)
Term credit levels	Average	611	127	20.8%
Term credit levels	High	611	477	78.1%
Term credit levels	low	611	1	0.2%
Term credit levels	Very High	611	2	0.3%

Figure 21

The Highest confidence is for term credit level = High and hence the First rule for D is

Rule 1

If student classification = freshman and GPA_Level = Excellent and Term credit level = High THEN Email Read = D
Removing the 477 records and starting over.

II Iteration

Removing the above records and starting over, the Highest Confidence is for Term Credit Levels = Very High

Attribute	Value	Support	Support D	Confidence D
Student_Classification	freshman	1595	602	37.7%
Student_Classification	Junior	4704	1606	34.1%
Student_Classification	Senior	8776	2361	26.9%
Student_Classification	**Sophomore**	**2537**	**986**	**38.9%**
Term credit levels	Average	9911	2927	29.5%
Term credit levels	High	6624	2399	36.2%
Term credit levels	Low	966	174	18.0%
Term credit levels	**very High**	**112**	**55**	**49.1%**
GPA Level	Sat	1361	371	27.3%
GPA Level	Excellent	10489	3491	33.3%
GPA Level	Fail	1251	319	25.5%
GPA Level	Good	4080	1268	31.1%
GPA Level	Inf	432	106	24.5%

Figure 22

Using only the records for Term Credit level = Very high and going down a level further, we get freshman with the highest support and confidence.

Attribute	Value	Support	Support D	Confidence D
Student_Classification	**freshman**	**55**	**31**	**56.4%**
Student_Classification	Junior	55	8	14.5%
Student_Classification	Senior	55	13	23.6%
Student_Classification	Sophomore	55	3	5.5%
GPA Level	Sat	55	2	3.6%
GPA Level	Excellent	55	19	34.5%
GPA Level	Fail	55	29	52.7%
GPA Level	Good	55	4	7.3%
GPA Level	Inf	55	1	1.8%

Figure 23

Rule 2: If Term credit level =Very High & If student classification = freshman and GPA_Level = fail THEN Email Read = D

III Iteration

Attribute	Value	Support	Support D	Confidence D
Student_Classification	freshman	1566	573	36.6%
Student_Classification	Junior	4704	1606	34.1%
Student_Classification	Senior	8776	2361	26.9%
Student_Classification	**Sophomore**	**2537**	**986**	**38.9%**
Term credit levels	Average	9911	2927	29.5%
Term credit levels	High	6624	2399	36.2%
Term credit levels	Low	966	174	18.0%
Term credit levels	very High	83	26	31.3%
GPA Level	Sat	1361	371	27.3%
GPA Level	Excellent	10489	3491	33.3%
GPA Level	Fail	1222	290	23.7%
GPA Level	Good	4080	1268	31.1%
GPA Level	Inf	432	106	24.5%

Figure 24

Removing the 55 records from Rule 2, we get the Highest Confidence for Student classification = sophomore, using records only for student classification = sophomore, we get the following records.

Sophomore				
Attribute	Value	Total Count	Support (E)	Confidence (E)
Term credit levels	Average	986	444	45.0%
Term credit levels	Low	986	508	51.5%
Term credit levels	Very High	986	31	3.1%
Term credit levels	High	986	3	0.3%
GPA Level	Sat	986	64	6.5%
GPA Level	**Excellent**	**986**	**632**	**64.1%**
GPA Level	Fail	986	40	4.1%
GPA Level	Good	986	232	23.5%
GPA Level	inf	986	18	1.8%

Figure 25

Drilling down on student classification = sophomore, the highest confidence is for GPA_Level = Excellent and the final rule is

Rule3

If student classification = Sophomore & GPA_Level = Excellent & Term credit level =Low and THEN Email Read = D

Rules for Level D

If Term credit level =Very High & If student classification = freshman and GPA_Level = fail THEN Email Read = D
If student classification = freshman and Term credit level = High and GPA_Level = Excellent THEN Email Read = D
If student classification = Sophomore & GPA_Level = Excellent & Term credit level =Low and THEN Email Read = D

The results for Activity level = D also indicate that a high level of engagement is a result of high academic standards. One exceptions to that was rule 2, however the rest of the rules are consistent with the pattern that activity level increases with academic grades level.

Using the above two Activity Levels, the covering algorithm suggests that

1. GPA = excellent is the highest Indicator of high Engagement
2. Freshman and Sophomore students have the highest Involvement with the Viking newsletter
3. Students with a term credit level = high/average have a high engagement.

3. Bayesian Algorithm

Using the Bayesian algorithm, the relationship between the following factors was studied with respect to the activity level-
1> GPA of the Term
2> whether the student registered for next term?
3> Student category (which year of undergrad - freshman/sophomore/junior/senior).

The students' activity around the Virtual Viking emails was classified as follows

Table Voicemails read Activity Level

Activity Level	Rule
High Act.	If ratio of the count of number of times emails were opened / count of not opened emails >= 0.7 and <=1
Med Act.	If ratio of the count of number of times emails were opened / count of not opened emails >= 0.2 and < 0.7
Low Act.	If ratio of the count of number of times emails were opened / count of not opened emails < 0.2

Now, in order to use the Bayesian algorithm we need to find the product of the probabilities of all the factors that we decided to use in our model. For this, we created individual tables of the probabilities for each factor. This was done so we could easily automate the calculation by building a predictive model for our dataset using the excel function VLOOKUP on the probability (Bayesian) tables of the factors being considered

Probability of GPA categories factor per activity level

GPA Category | Activity Level	Count of GPA Category | Activity level	Activity level Count	Probability
Excellent | High Act.	3704	5565	0.665589
Excellent | Low Act.	3501	6504	0.538284
Excellent | Med Act.	3760	6020	0.624585
Failure | High Act.	341	5565	0.061276
Failure | Low Act.	601	6504	0.092405
Failure | Med Act.	309	6020	0.051329
Good | High Act.	1110	5565	0.199461
Good | Low Act.	1609	6504	0.247386
Good | Med Act.	1361	6020	0.22608
Inferior | High Act.	86	5565	0.015454
Inferior | Low Act.	208	6504	0.03198
Inferior | Med Act.	138	6020	0.022924
Satisfactory | High Act.	324	5565	0.058221
Satisfactory | Low Act.	585	6504	0.089945
Satisfactory | Med Act.	452	6020	0.075083

Figure 26

Probability of registered next term factor per activity level.

Reg Next Term | Activity Level	Count of Reg Next Term | Activity level	Activity level Count	Prob
N | High Act.	530	5565	0.095238
N | Low Act.	1242	6504	0.190959
N | Med Act.	714	6020	0.118605
Y | High Act.	5035	5565	0.904762
Y | Low Act.	5262	6504	0.809041
Y | Med Act.	5306	6020	0.881395

Figure 27

Probability factor based on student Classification

Student Category | Activity Level	Count of Student Category | Activity level	Activity level Count	Prob
Freshman | High Act.	1071	5565	0.192453
Freshman | Low Act.	382	6504	0.058733
Freshman | Med Act.	618	6020	0.102658
Junior | High Act.	1501	5565	0.269721
Junior | Low Act.	1640	6504	0.252153
Junior | Med Act.	1563	6020	0.259635
Senior | High Act.	2060	5565	0.370171
Senior | Low Act.	3717	6504	0.571494
Senior | Med Act.	2999	6020	0.498173
Sophomore | High Act.	933	5565	0.167655
Sophomore | Low Act.	765	6504	0.11762
Sophomore | Med Act.	840	6020	0.139535

Figure 28

Probability for the activity levels is mentioned in fig 29

Activity Level	Count of Students in the Activity level	Prob
High Act.	5565	0.307646
Low Act.	6504	0.359556
Med Act.	6020	0.332799

Figure 29

Below is the excel implementation that helps automate the calculation of the combined probabilities of the factors we have selected to establish a relationship between.

Observation:	GPA	Registered Next Term	Student Classification	Prob Act. Level	Product (x 10^{-4})	Likelihood (%)
	Excellent	Y	Freshman			
High Act.	0.6655885	0.904761905	0.19245283	0.30764553	356.5455559	55.3%
Med Act.	0.624584718	0.881395349	0.102657807	0.359555531	203.1982979	31.5%
Low Act.	0.538284133	0.80904059	0.058733087	0.332798939	85.12294729	13.2%
TOTAL					644.8668011	100.0%

GPA Desc	GPA	Reistered Next Term	Student Classification
GPA between >3 & <= 4	Excellent	Y	Freshman
GPA between >2 & <=3	Good	N	Sophomore
GPA between >1 & <=2	Satisfactory		Junior
GPA between >0 & <=1	Inferior		Senior
GPA between = 0	Failure		

Figure 30

Using the above predictive model we could see the following:

1. **Freshmen/Seniors** -
   What we observed is that freshmen irrespective of their performance actively use the Virtual Vikings emails. Seniors on the other hand do not usually get actively involved with the Virtual Vikings email program. It was also observed that there is a high likelihood that seniors not signing up for the next term show low email activity level.
2. **Sophomores & Juniors** -
   It was observed that lower email activity meant lower GPAs and less signing up for next term.

Evaluation Criteria

For analysis different bucketing was used. This bucketing ensured that we were able to correlate the data without altering the content.

Also it ensured that we were looking at different criteria's based on different types of academic activities such as the GPA bucket was built according to the grading mechanism for undergraduates as per Portland State University and used in manual analysis and all the algorithms.

GPA Desc	Rule
Excellent	GPA between >3 & <= 4
Good	GPA between >2 & <=3
Inferior	GPA between >0 & <=1
Failure	GPA = 0

For Bayesian, the emails classification was according to the ratio. The students' activity around the Virtual Viking emails was classified.

Activity Level	Rule
High Act.	If ratio of the count of number of times emails were opened / count of not opened emails >= 0.7 and <=1
Med Act.	If ratio of the count of number of times emails were opened / count of not opened emails >= 0.2 and < 0.7
Low Act.	If ratio of the count of number of times emails were opened / count of not opened emails < 0.2

In 1R and covering algorithm the bucketing of email was considered on the read emails.

Email Classification	Rule
A	Emails read is 0
B	Emails read >0<=3
C	Emails read >3<=6
D	Emails read >6<=11
E	Emails read >11

The manual analysis through SQL, graphs and charts helped with the initial assessment and 1R, Bayesian and covering algorithms helped to conclude and predict.

Conclusion

Based on our research below are the findings

- Senior population is highest compared to freshman, sophomore and juniors. And the percentage of unopened emails is higher in Senior Students. Freshman Students have higher open ratio compared to the others.
- Email regarding culture center grand opening, harvest share was the least read and the email with FASA, Portland State of Mind and party in the Park were the most read.
- Irrespective of the class standing the students who did not open their emails is high compared to students who opened their emails. However this difference is low for excellent and good GPA students but very high between Satisfactory, Failure and Inferior GPA students.
- The Highest Emails are opened within 1-2 days of being sent and almost 91 % of read emails are opened within 12 hour-1 day. Only 44% of the emails get opened in 0-1 Hour.
- Using 1R algorithm it was found that students with excellent and good grades read emails between 6 and 11 with the error of .63% and .68%. This shows high engagement through emails with student with GPA >3 Students with satisfactory and inferior grades appear to read 3 emails with error of .68% and .67%.
- The covering algorithm suggests that the Students with GPA Level Excellent are the highest Indicator of high Engagement .Freshman and sophomore students have the highest Involvement with the Viking newsletter and Students with a term credit level high/Average have a high engagement.
- Using Bayesian it was found that
    1) Freshmen irrespective of their performance actively use the Virtual Vikings emails.
    2) Seniors on the other hand do not usually get actively involved with the Virtual Vikings email program. It was also observed that there is a high likelihood that seniors not signing up for the next term show low email activity level.
    3) For Sophomores & Juniors it was predicted that lower email activity is related to lower GPAs and less signing up for next term.

Recommendations

Based on these finding it can be concluded that students who don't read these emails are more when compared to students who do, and also the students who are more engaged in academics through Good and Excellent GPA's are more likely to read the emails. We recommend that -

Target High Population Students

- Department can make efforts to reach out to the senior students as they have high population and less readers.
- Involve senior students by including articles of their interest like on campus interviews etc.

Delivery Time

- Perhaps choose a different time for the email to be sent as the maximum number of emails are opened after 12 hours of being sent.
- It is most likely that emails are read in the morning rather than during the latter part of the day according to the data. Hence Saturday night or Sunday evening could be better time to send email.

Segment Audience

- Have a forum for students to decide what they would like to read about in the emails.
- Encourage student involvement through frequent feedbacks.
- Check what content students like reading based on different patterns too.

Make Emails Mobile Friendly

- Since most of the emails are read on mobile phones, and students are mostly interested in on campus activities, it makes much more sense to have the main events of the week as top link along with Dates and times of events.
- People are more unlikely to open the different browser links individually on a mobile device and hence it would be better to have the most information on the first link itself.

Further Research

The data that was provided was for the fall 2016 and Undergraduate Students. There is a need for using this as a training data and establish test data for other terms or similar fall term for other academic years as well as other programs and develop rules that can be further used to make changes to the Campaign in order to make it more effective.

Appendix A

Definition of terms

Academic Period: Term (ie Fall 2016 term, winter 2017 term etc). Codes for this have 01 for winter, 02 for spring, 03 for summer and 04 for fall term. For example 201604 is the code fall 2016.

Campaign: Type of email campaign in which there can be many mailers sent.

Mailer: Email on with unique content and subject that was sent to students

Random UID: Randomized identifier for the student

CAMPAIGN_TALISMA_ID - only one

MAILER_TALISMA_ID related to MAILER_SUBJECT (one to one)

1- Good Graduate Standing
AD- Accad Dismissal
AR -Acad Disq-Reinstated on Prob
AP-Academic Probation
AW -Academic Warning

Reference

[1]Witten, Frank, and Hall. Data Mining (3rd edition). Morgan Kaufman, Burlington, MA; 2011.

[2]W. Inmon, *Building the data warehouse*, 1st ed. Indianapolis: Wiley, 2011.