A DATA MINING STUDY OF RANKING WITHIN SOCIAL NETWORKS

A THESIS IN

Computer Science

Presented to the Faculty of the University
of Missouri Kansas City in partial fulfillment of
the requirements for the degree
MASTER OF SCIENCE

by

RAMA DEVI RAGHAVAN

B. Tech., Anna University, 2008

Kansas City, Missouri

2014

A DATA MINING STUDY OF RANKING WITHIN SOCIAL NETWORKS

Rama Devi Raghavan, Candidate for the Master of Science Degree

University of Missouri – Kansas City, 2014

ABSTRACT

Social networks have become very popular in the past few years and have become a significant part of our personal and professional lives. As the number of participants in social networks has grown, they have become a virtual space for exerting influence. Studies in sociology and marketing have stressed the vital role of influence for businesses to survive; organizations and businesses are constantly seeking to establish and expand their presence by exploiting social networks. This has led to an implicit competition for higher visibility and ranking within social networks. Ideally, the ranking of participants within social networks should mirror the real world. However, this may or may not be true because of different degrees of participation, overrepresentation due to self-promotion and the possibility of unreliable or false information.

This thesis addresses the following related questions. What are appropriate measures for ranking participants in social networks? Does the ranking within networks mirror those based on traditional measures for ranking organizations? We use data mining and statistical analysis to evaluate several measures, including a new measure based on the H-index, for ranking participants within social networks against established benchmarks for university programs. We find that prominence within social networks correlates in general with prominence in the real world. We identify the best measures for predicting prominence in the real world, and perform preliminary outlier analysis. While the observations are not proven to be causal, they offer insights of potential value to social network marketing.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of School of Computing and Engineering, have examined a thesis titled “A Data Mining Study of Ranking Within Social Networks” presented by Rama Devi Raghavan, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Dinakarpandian Deendayal, Ph.D., Chair
School of Computing and Engineering

Vijay Kumar, Ph.D.
School of Computing and Engineering

Praveen Rao, Ph.D.
School of Computing and Engineering

TABLE OF CONTENTS

ABSTRACT i
APPROVAL PAGE ii
LIST OF TABLES vi
LIST OF ILLUSTRATIONS vii

CHAPTERS
1. INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement 2
1.3 Thesis Outline 3
2. RELATED WORK 4
3. RANKING OF USERS IN SOCIAL NETWORKS 6
3.1 Data Sources 7
3.2 Consistency Check within Networks 10
3.3 Consistency Check between Networks 12
4. RELIABILITY OF SOCIAL NETWORK USER RANKING 14
4.1 Accepted Ranking Methods 14
4.1.1 U.S. News ranking 14
4.1.2 Research Ranking 16
4.1.3 University Ranking by Academic Performance (URAP) 17
4.2 Additional Features Collection 18
4.3 Correlation Analysis 19
4.3.1 With U.S. News Rank 19
4.3.2 With Microsoft Academic Search Rank 20
4.3.3 With URAP Rank 21
5. PROPOSED METHODS OF SOCIAL NETWORK RANKING 22
5.1 H-index 22
5.2 Age of Account 23
5.3 Popularity of followed users 23

5.4 Average Re-tweet 24
5.5 Comparison of New Methods with TwitterCounter 24
5.6 Correlation Analysis with Accepted Ranking Methods 25
6. COMPOSITE SOCIAL NETWORK RANKING 26
6.1 Methodology 26
6.2 Correlation Analysis with Accepted Ranking Methods 27
7. DATA MINING STUDY 28
7.1 Rank Difference Calculation 28
7.2 Universities that have Positive Correlate with U.S. News Rank 29
7.3 Universities that have Negative Correlation with U.S. News Rank 29
7.4 Universities that have Positive Correlation with Research Rank 30
7.5 Universities that have Negative Correlation with Research Rank 30
7.6 University Information Sources 31
7.6.1 General Information 31
7.6.2 Student Ratings 35
7.6.3 Research Information 36
7.7 Classification Based on U.S. News Rank 37
7.7.1 Positive and Negative Correlation Graphs 37
7.7.2 Low Marketing and High Marketing Graphs 42
7.7.3 Feature Selection 48
7.7.4 Linear Regression 49
7.7.5 Decision Tree Rules 60
7.7.5.1 Decision Tree for Classifying Positive and Negative Correlation 61
7.7.5.2 Decision Tree for Classifying High and Low Marketing 64
7.8 Classification Based on Research Rank 67
7.8.1 Positive and Negative Correlation Graphs 68
7.8.2 Low Marketing and High Marketing Graphs 73
7.8.3 Linear Regression 78
7.8.4 Decision Tree Rules 85
7.8.4.1 Decision Tree for Positive and Negative Correlation classes 86
7.8.4.2 Decision Tree for Classifying High and Low Marketing 88

8. CONCLUSION AND FUTURE WORK 92
8.1 Summary 92
8.2 Future Work 93
REFERENCES 94

LIST OF TABLES

1. Comparison of Related Work 4
2. Explanation of social network data collected 9
3. Harvard University Social Network Data 19
4. Sample H-index Data 23
5. Sample Social Network Account Duration Data 23
6. Sample Twitter Average Following Popularity Data 24
7. Sample Average Re-tweet for a Tweet Data 24
8. Correlation of Social Network Rank with Accepted Ranking 27
9. Rank Difference Calculation between Social Network and U.S. News 28
10. Universities that have Positive Correlate with U.S. News Rank 29
11. Universities that Inverse Correlate with U.S. News rank 29
12. Universities that Correlate with Research Rank 30
13. Universities that Inverse Correlate with Research Rank 31
14. List of Universities General Information Data 31
15. Universities Research Information Description 36
16. Universities that Correlate with U.S. News rank 38
17. Universities that Inverse Correlate with U.S. News rank 39
18. Universities in Low Marketing w.r.t U.S. News rank 42
19. Universities in High Marketing w.r.t U.S. News rank 43
20. Neural Networks output for High/Low marketing classification (USNews) 67
21. Rank difference calculation between Research and Social Networks 68
22. Universities that Correlate with Research rank 69
23. Universities that Inverse Correlate with Research rank 70
24. SAT 25th percentile distribution for Research Correlate & Inverse Correlate 73
25. Universities in High Marketing w.r.t Research rank 74

LIST OF ILLUSTRATIONS

1. Example Facebook page 6
2. Example Twitter Page 6
3. Example YouTube Page 7
4. Facebook top users data 8
5. Twitter top users data 8
6. YouTube top users data 9
7. Correlation within Random Twitter Data 11
8. Correlation within Random Facebook Data 11
9. Correlation within random YouTube Data 11
10. Correlation between Likes, Followers, Subscribers 12
11. Correlation between Tweets, Videos, People Talking 13
12. Correlation between Re-tweets, Post likes, Views 13
13. Correlation between Post shares, Favorites, Comments 13
14. U.S. News Undergrad College Ranking Algorithm 16
15. Microsoft Academic Search snapshot 17
16. University Ranking by Academic Performance (URAP) snapshot 18
17. Harvard University Home Page 18
18. Social Network Rank Correlation with U.S. News score 20
19. Social Network Rank Correlation with Research Publications Count 20
20. Social Network Rank Correlation with Research Citation Count 21
21. Social Network Rank Correlation with URAP 21
22. H-index for Social Networks 22
23. Correlation Analysis of New Methods with TwitterCounter 25
24. Correlation Analysis of New Methods with Accepted Ranking 25
25. Hierarchical Clustering of Social Network Features Output 26
26. Unigo Student Ratings Snapshot 35
27. USNews Correlate & Inverse Correlate histogram 40
28. SAT percentile distribution for USNews Correlate & Inverse Correlate 40
29. Faculty and Financial Resources distribution for USNews Correlate & Inverse Correlate 41

30. Freshman in top 25% distribution for USNews Correlate & Inverse Correlate 41
31. Retention and Graduation rate distribution for USNews Correlate & Inverse Correlate 41
32. Low & High Marketing w.r.t USNews histogram 45
33. Acceptance rate distribution for Low & High Marketing w.r.t USNews 45
34. Academia Features distribution for Low & High Marketing w.r.t USNews 46
35. Graduation and Selectivity distribution for Low & High Marketing w.r.t USNews .. 46
36. SAT percentile distribution for Low & High Marketing w.r.t USNews 47
37. Student Faculty ratio and Under 20 classes distribution for Low & High Marketing w.r.t USNews 47
38. General information distribution for Low & High Marketing w.r.t USNews 48
39. Feature selection for Correlate & Inverse correlate (USNews) 49
40. Feature selection for Low & High marketing (USNews) 49
41. Bivariate fit of Rank difference by Freshman Retention rate (USNews) 51
42. Bivariate fit of Rank difference by 6-year Graduation rate (USNews) 52
43. Bivariate fit of Rank difference by Freshman Retention rate (USNews) 53
44. Bivariate fit of Rank difference by Predicted Graduation rate (USNews) 54
45. Bivariate fit of Rank difference by Freshman in top 10% (USNews) 55
46. Bivariate fit of Rank difference by Undergraduates (USNews) 56
47. Bivariate fit of Rank difference by Unigo Sports score (USNews) 57
48. Bivariate fit of Rank difference by Acceptance rate (USNews) 58
49. Bivariate fit of Rank difference by Retention rank (USNews) 59
50. Linear Regression Model for Rank difference (USNews) 60
51. Decision Tree to Classify Correlate and Inverse Correlate (USNews) 62
52. Decision Tree Leaf Report to Classify Correlate and Inverse Correlate (USNews) .. 63
53. ROC curve and Confusion Matrix for Correlate and Inverse Correlate classifier (USNews) 64
54. Decision Tree to Classify Low & High marketing (USNews) 65
55. Decision Tree Leaf Report to Classify Low & High marketing (USNews) 65
56. ROC curve & Confusion Matrix for Low & High marketing classifier (USNews) ... 66
57. Research Correlate & Inverse Correlate histogram 71

58. Retention rate distribution for Research Correlate & Inverse Correlate..... 71
59. Full time faculty distribution for Research Correlate & Inverse Correlate 72
60. Retention rank distribution for Research Correlate & Inverse Correlate 72
61. SAT 25th percentile distribution for Research Correlate & Inverse Correlate..... 73
62. Low & High Marketing w.r.t Research histogram 75
63. Acceptance rate distribution for Low & High Marketing w.r.t Research..... 76
64. Under 20 classes distribution for Low & High Marketing w.r.t Research 76
65. Financial Resources rank distribution for Low & High Marketing w.r.t Research.... 76
66. Unigo student ratings distribution for Low & High Marketing w.r.t Research..... 77
67. Setting and Private/Public histograms for Low & High Marketing w.r.t Research ... 77
68. Student Faculty ratio distribution for Low & High Marketing w.r.t Research..... 78
69. Tuition distribution for Low & High Marketing w.r.t Research..... 78
70. Bivariate fit of Rank difference by 6-year graduation rate (Research)..... 80
71. Bivariate fit of Rank difference by Student-faculty ratio (Research) 81
72. Bivariate fit of Rank difference by Undergraduates (Research)..... 82
73. Bivariate fit of Rank difference by Unigo sports score (Research) 83
74. Bivariate fit of Rank difference Financial Resources rank (Research)..... 84
75. Linear Regression Model for Rank difference (Research) 85
76. Decision Tree to Classify Correlate and Inverse Correlate (Research) 86
77. Decision Tree Leaf Report to Classify Correlate and Inverse Correlate (Research).. 87
78. ROC curve and Confusion Matrix for Correlate and Inverse Correlate classifier (Research) 88
79. Decision Tree to Classify Low & High marketing (Research)..... 89
80. Decision Tree Leaf Report to Classify Low & High marketing (Research)..... 89
81. Histogram for Classes with under 20 students..... 90
82. ROC curve & Confusion Matrix for Low & High marketing classifier (Research)... 90
83. Decision tree for High/Low marketing with Gain ratio split..... 91

ACKNOWLEDGEMENTS

"The only time you fail is when you fall down and stay down"
-Stephen Richards.

When I started my graduate studies, I remember reading graduate school experiences from the web and seeing a lot of "thought about dropping out" and "too overwhelming to bear". But I was high in spirit and motivation and I used to think they were joking. But when I was working on my Thesis and my motivation started ebbing, I understood what they meant, I felt like I was in an abyss of void. I considered myself a failure as I was struggling to get motivated. And I survived through that phase. There are a couple of things that helped me get through. An advisor who is genuinely interested in helping you succeed and who trusts your ability even when you don't, family and friends who believe your potential and always stand by your side. I understood it is ok to fall down as it doesn't mean I failed.

Hence I take this opportunity to thank my advisor Dr. Deendayal Dinakarpandian and mentors Dr. Vijay Kumar and Dr. Praveen Rao. I would like to express my gratitude to the great authors whose pearls of wisdom have given me clarity and confidence. I'd also like to thank the SCE department of UMKC for giving me the education and setting the right stage for me to grow into a graduate.

I will always be grateful and proud of my family as they taught me valuable lessons. My father R. V Raghavan has taught me dedication, my mother Gokila Raghavan has taught me patience, my husband Rajasekar Rajendran has taught me humility, my brother Janardhanan Raghavan has taught me maturity, and they have loved me no matter what.

CHAPTER 1

INTRODUCTION

1.1 Background

From sports to education, we are captivated by rankings. Everyone wants their favorite team to be ranked number one. Every parent wants their child to attend the best college. A new business becomes a success or failure based on how it is ranked by Google. A sports team stepping to rank one makes billions of people excited. Hence it is getting increasingly important to analyze the techniques of various ranking method to ensure fairness. Ranking has universal draw because of the abundance of interesting datasets in nearly every field imaginable. “Evolution rewards those who make quick comparisons, those who thought more slowly (or even those who were quick but incorrect) were no doubt removed from the gene pool by the swifter-thinking predator” [1]

Some famous ranking techniques include Google’s PageRank algorithm to rank webpages, Netflix and Internet Movie Database (IMDb) to rank movies, college football teams ranking (Bowl Championship Series), college basketball ranking (Rating Percentage Index), and college and university ranking by U.S.News, etc.

In our daily life we frequently have the need to choose one or more things from several alternatives. And obviously we would like to choose the best option from the alternatives. The best option could be an easy choice sometimes but more often than not it is not. As David Hume discusses in his book “A Treatise of Human Nature”, humans get a malicious joy by comparison, as “the misery of another gives us a more lively idea of our happiness and his happiness of our misery” [36]. Therefore we involuntarily rank things without much thought, which might affect a lot of people. For example, Billboard Hot 100 is

the American music industry popularity chart issued weekly by Billboard magazine which is based on sales. When a song is on top of this list, we tend to think that it is better than others down the list and we tend to buy the album and it increases sales and it becomes a loop promoting this song which may or may not be liked by everybody, and other album sales are also affected by it. There are other things that we use to rank individuals/businesses involuntarily like social networks. Social networking sites are online communication tool that have taken over the internet by being an incredibly efficient gossip engine. Nearly 30% of world population uses social networking sites [35], and by 2017, 2.33 (31.3%) billion people will use social networks. Since social networks have such a wide reach among us, we involuntarily judge about a lot of things through them because of our innate nature of comparison. Hence it is important to study ranking in social networks and identify unbiased methods to rank users in social networks, and do a comparative study with golden standards to understand the features of social networks.

1.2 Problem Statement

Social networking is ranked as the most popular content category for worldwide engagement with market penetration of 85 percent in 41 out of 43 markets. It is the most popular online activity worldwide, 1 in every 5 minutes is spent on social networking sites globally [3]. “Social media is not only growing, it’s absolutely here to stay,” according to Erik Qualman, author of the best-selling book, “Socialnomics: How social media transforms the way we live and do business” [37], which breaks down the power of the medium with the help of real-world examples.

Due to its outreach and intense activity, social networks are becoming the most popular data source available for mining. They predominantly have live user-generated

content which is very powerful when put to right use. Social network serves as the right medium for self-promotion and marketing due to the abundance of users logging into it every day. Ranking users in social networks needs to be standardized as it is ever-available and updated live and therefore very exciting and interesting. However, in spite of its usefulness, there are not many references for algorithms to rank users in social networks, which are disposed to huge bias of self-promotion and marketing. There are some tools that are popular for ranking Twitter users but they are quite simple and straight forward. However, Facebook with much more global popularity doesn't have many studies devoted to it (not much research has been done) on ranking user popularity. A comprehensive study on ranking users popularity in social network is missing. Therefore, it is an unexplored area and we have some novel ideas to share.

1.3 Thesis Outline

In Chapter 2 we discuss related work on measuring user popularity on various social networks (Twitter, Facebook). Chapter 3 introduces to current measures of ranking in social networks with sample data and we do a consistency check for these measures. In Chapter 4 we compare the collected measures of social networks with the same users' corresponding ranking in accepted ranking methods like U.S. News' academic ranking and Microsoft Academic Search's research ranking. We introduce some novel measures of ranking in Chapter 5 and do a quick study on how it compares with other popular ranking tools and accepted methods. We propose a new measure of ranking a user in social network which is explained in Chapter 6 and we compare it with other methods. We perform a data mining study with the collected data and it is explained in-depth in Chapter 7 and the conclusion and summary is in Chapter 8.

CHAPTER 2

RELATED WORK

In this chapter, we discuss about ranking in general and about work done on measuring popularity and ranking in social networks.

Ranking is defined as the relationship between a set of items such that for any two items, the first is either "ranked higher than", "ranked lower than", or "ranked equal to" the second. In mathematics it is known as preorder of objects. [38] There are many simple strategies for assigning rankings:

Say, A ranks ahead of B and C (which compare equal) which are both ranked ahead of D.

Standard competition ranking: A gets 1, B gets 2, C gets 2 and D gets 4

Dense ranking: A gets 1, B gets 2, C gets 2 and D gets 3

Ordinal ranking: A gets 1, D gets 4, either B gets 2 and C gets 3 or B gets 3 and C gets 2

Fractional ranking: A gets 1, B and C each get average (2+3/2 = 2.5), and D gets 4

Table 1: Comparison of Related Work

CITATIONS	STUDY
Eysenbach [2]	Data mining on tweets (updates on Twitter) to predict future citation count and correlation with impact factor. Impact of journal is analyzed from tweets and not from single user. Social impact measures based on tweets to compliment traditional citation metrics.
Chat et al [3]	Measuring user's influence in Twitter. 3 measures of influence: indegree, retweeets and mentions were compared with dynamics of user across topics and time. Temporal analysis was performed and no definite measure of influence was concluded.
Sarma et al	Mechanisms for ranking status messages (tweets) using user reviews (pairwise

CITATIONS	STUDY
[4]	comparisons instead of rating)
Leavvitt [5]	Measuring different features of Twitter (followers, following, follower following ratio) and reasoning each feature, inspecting 12 users for a 10 day period and how time difference changes each feature.
Weng et al [6]	Studying general characteristics of follower following network (reciprocity) and topics of tweets. Influence calculated by recursively querying over the network of followers
Bakshy et al [7]	Tracking word of mouth information – spread through Twitter follower network
Uysal et al [8]	Constructing a model to predict whether a user will retweet a particular tweet from the list of new tweets by following retweeting nature of user. Ranking incoming tweets
Bandari et al [9]	Model for predicting popularity of news items on Twitter when news items are mentioned as URLS. Compares news item popularity with retweet popularity.

Most work was based on Twitter data and not much work has been done on user popularity measure on Facebook except network analysis. Combining data from different social networks for a user is novel in our thesis and work hasn't been done on this topic.

CHAPTER 3

RANKING OF USERS IN SOCIAL NETWORKS

Since their introduction, social network sites (SNSs) such as MySpace, Facebook, Cyworld, and Bebo have attracted millions of users, many of whom have integrated these sites into their daily practices. Some of the most popular social networks today are Facebook, Twitter, LinkedIn, Google Plus [39]. For this thesis we will be using Facebook, Twitter and YouTube as the primary source of social network data.

Figure 1: Example Facebook page

Figure 2: Example Twitter Page

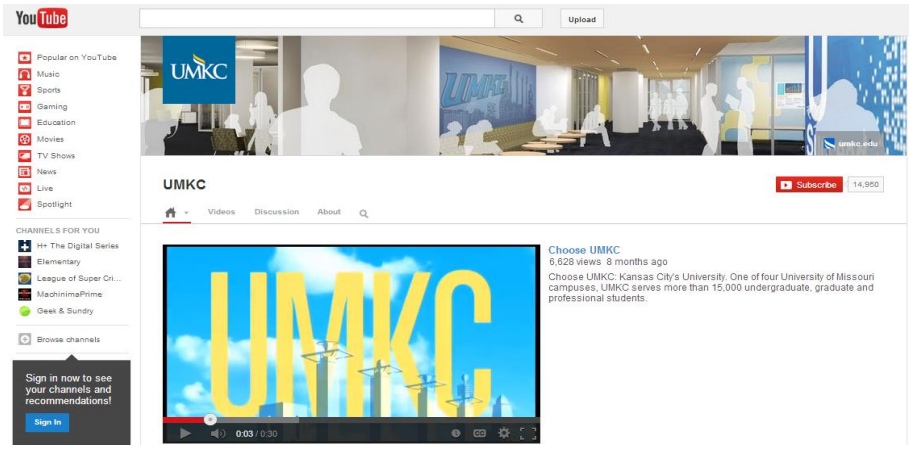


Figure 3: Example YouTube Page

3.1 Data Sources

To analyze the characteristics of different features of social networks we need same users that are present in different social networks to be consistent. Hence we selected users that could be found in Twitter, Facebook and YouTube and whose accounts are verified. We need the accounts to be verified as otherwise there are multiple accounts and we cannot find which account is the legitimate account of the user. We selected 17 users randomly based on their verified accounts in Twitter, Facebook and YouTube (social network sites verify famous users' accounts and mark them as verified). Facebook, Twitter and YouTube have public APIs that was used for data collection. Facebook account data collected were number of likes for pages, number of people talking about the page, number of likes for the retrievable number of posts and number of shares for the posts. Twitter account data collected were number of followers, number of tweets, number of re-tweets and number of favorites for their tweets. YouTube account data collected were number of subscribers for the page, number of videos published, number of views for the videos and number of comments for the videos. Below is a snapshot of the first few rows of data.

FACEBOOK					
Facebook id	Likes	People talking	Posts	Post likes	Post shares
katyperry	60641609	368322	365	11916181	638216
JustinBieber	58326482	525600	463	30605134	1608944
ladygaga	61020906	860418	359	28506606	755246
barackobama	37614739	2501491	258	39013453	2725257
TaylorSwift	51183970	1441163	366	30583178	691398
britneyspears	33830667	948588	430	31365666	1137634
rihanna	81110960	481341	387	40280566	1376029
justintimberlake	29113927	536088	275	12218416	306092
jenniferlopez	28345844	159169	375	14644714	585893
ellentv	11604305	478544	800	42476865	2496833
Cristiano	67122125	3525026	443	81876665	4506652
shakira	75356175	1655446	390	61278672	2316846
oprahwinfrey	9112935	675208	393	4905769	221221
pink	26932628	497121	228	17873823	843147
OnlyDemetriaLovat(	14646	1827	429	47268	1316
adele	50024860	232312	26	2350531	195156
KimKardashian	15929620	674905	800	17469358	246739
harrystyles	9583736	93554	1	93744	17419

Figure 4: Facebook top users data

TWITTER				
Twitter id	Followers	Tweets	Retweets	Favorites
katyperry	48389969	5294	4661169	8073195
justinbieber	47656569	25264	77812675	121747553
ladygaga	40866249	4250	9938153	16056334
BarackObama	40481991	10546	2757550	6967682
taylorswift13	37506785	2072	7322643	10038653
britneyspears	34645889	3374	1945806	3931253
rihanna	33164455	8730	3646048	7906700
jtimberlake	29121366	2063	1066780	1755673
JLo	25919059	2648	554732	1015881
TheEllenShow	24050579	8145	3909561	4911549
Cristiano	23238438	1667	1553159	3005534
shakira	23090158	2166	694251	1137138
Oprah	22459506	8314	331135	512700
Pink	21323926	5128	1058644	2390681
ddlovato	20244973	10355	19728207	34654104
OfficialAdele	18879087	195	263278	409290
KimKardashian	18830146	16426	1473556	1600707
Harry_Styles	18138399	4592	72683116	69318038

Figure 5: Twitter top users data

YOUTUBE				
Youtube id	View	Comment	Subscriber	Video
KatyPerryVEVO	2166719070	46338	7989018	71
JustinBieberVEVO	4272151505	471769	7273636	94
LadyGagaVEVO	2568913348	194145	4129870	91
TaylorSwiftVEVO	1845932982	39460	6685661	61
BritneySpearsVEVO	1668022264	47202	1899808	74
RihannaVEVO	4461570876	78272	11538140	83
justintimberlakeVEVO	677226664	3132	2636140	33
JenniferLopezVEVO	1797350926	18258	1970045	33
TheEllenShow	2320531257	137083	7645268	4647
CristianoRONALDO	12153554	13091	187185	30
shakiraVEVO	2237924294	29521	1830823	83
PinkVEVO	1213153221	8155	4022300	77
DemiLovatoVEVO	565702519	10533	3798915	56
AdeleVEVO	1170692973	16212	4070423	25

Figure 6: YouTube top users data

Explanation for the features collected is given below:

Table 2: Explanation of social network data collected

Social Network	Features	Explanation
Facebook	Page likes	Users can like a public page to receive their updates and this is total number of users who have liked a particular Facebook page
	People talking	This is the total count of users in a seven day span who like a page, post on the wall, like/comment/share a post, mention the page, tag on a photo, check-in, etc.
	Post likes	Total number of likes for all the posts (that are retrievable through the API)
	Post shares	Total number of shares (will appear on the timeline of the user who shares) for all the posts (that are retrievable through the API)
Twitter	Followers	Users can follow a public twitter account to receive their tweets and this is the total number of followers who have followers a particular Twitter

Social Network	Features	Explanation
		account
	Tweets	Total number of updates of the user
	Re-tweets	A re-tweet is someone else's Tweet that a user chooses to share with all of your followers. This feature is total number of re-tweets for all tweets (retrievable from Twitter API)
	Favorites	Favorites, represented by a small star icon next to a Tweet, are most commonly used when users like a Tweet. Favoriting a Tweet can let the original poster know that you liked their Tweet, or you can save the Tweet for later. This feature it total number of favorites for all the tweets (retrievable from Twitter API)
YouTube	Subscribers	Total number of subscribers for the YouTube channel
	Videos	Total number of videos published by the user in the channel
	Views	Total number of views for all the videos in the channel
	Comments	Total number of comments received for all the videos published in the channel (we do not distinguish positive and negative comments in this analysis)

3.2 Consistency Check within Networks

The first step in analysis the collected data is to check for consistency of the features extracted. For each social network we have four features of the users and we did a Spearman correlation [44] analysis without multiple testing error correction, between the features to understand them.

For Twitter we did correlation between ranks of number of followers, number of tweets, number of re-tweets and number of favorites for 117 users. And we could see that number of followers, re-tweets and favorites correlated with each other well but number of tweets (updates of user) did not.

Variable	by Variable	Spearman ρ	Prob>|ρ|
TWITTER-Tweets	TWITTER-Followers	0.0089	0.9246
TWITTER-Retweets	TWITTER-Followers	0.6371	<.0001*
TWITTER-Retweets	TWITTER-Tweets	0.2435	0.0087*
TWITTER-Favorites	TWITTER-Followers	0.6408	<.0001*
TWITTER-Favorites	TWITTER-Tweets	0.2157	0.0206*
TWITTER-Favorites	TWITTER-Retweets	0.9514	<.0001*

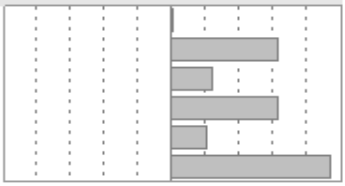


Figure 7: Correlation within Random Twitter Data

For Facebook we did correlation between number of page likes, number of people talking, number of post likes and number of post shares. And we could see that all the features correlate well with each other significantly.

Variable	by Variable	Spearman ρ	Prob>|ρ|
People talking	Likes	0.7527	<.0001*
Post likes	Likes	0.7796	<.0001*
Post likes	People talking	0.8949	<.0001*
Post shares	Likes	0.7387	<.0001*
Post shares	People talking	0.8222	<.0001*
Post shares	Post likes	0.9135	<.0001*

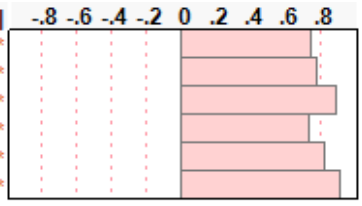


Figure 8: Correlation within Random Facebook Data

For YouTube we did correlation between number of subscribers for the channel, number of videos, number of views and number of comments. We were able to see that number of subscribers, comments and views correlate well with each other but number of videos (updates of the user) doesn't correlate as well as others.

Variable	by Variable	Spearman ρ	Prob>|ρ|
YOUTUBE-Comment	YOUTUBE-View	0.7669	<.0001*
YOUTUBE-Subscriber	YOUTUBE-View	0.9101	<.0001*
YOUTUBE-Subscriber	YOUTUBE-Comment	0.7312	<.0001*
YOUTUBE-Video	YOUTUBE-View	0.3576	0.0032*
YOUTUBE-Video	YOUTUBE-Comment	0.3249	0.0078*
YOUTUBE-Video	YOUTUBE-Subscriber	0.2863	0.0198*

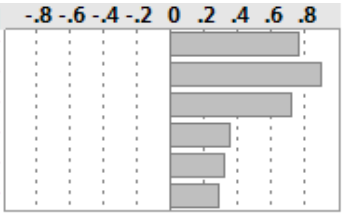


Figure 9: Correlation within random YouTube Data

3.3 Consistency Check between Networks

Different social network features are used for ranking and there are different ways or ranking users. We have analyzed how each feature correlates with each other in each social network, and what is more interesting is to see how features correlate between social networks. And this will show us how consistent different social networks are in describing the prominence of users and also if users are equally present in different social networks. We grouped similar features together to do this analysis.

Number of followers in Twitter, number of page likes in Facebook and number of subscribers in YouTube are similar to each other because all of them say how many users have found this user useful enough to get their updates and to keep track of their social network activity. And by the correlation analysis below, we can see that these features correlate very well with each other.

Variable	by Variable	Spearman ρ	Prob>|ρ|	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
FACEBOOK-Likes	TWITTER-Followers	0.4739	<.0001*	
YOUTUBE-Subscriber	TWITTER-Followers	0.6749	<.0001*	
YOUTUBE-Subscriber	FACEBOOK-Likes	0.5799	<.0001*	

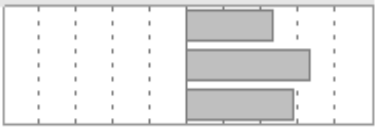


Figure 10: Correlation between Likes, Followers, Subscribers

Number of tweets in Twitter and number of videos in YouTube are similar to each other because they give us the activity of the user. Number of people talking in Facebook is indirectly giving us the activity of the user. Because if the user is more active more people visit the user's page which increases the "Number of people talking" count. Therefore these measures can be grouped together. And by the correlation analysis results below, we can see that these features do not correlate very well with each other, and using the analysis we did within social networks we can say that user updates do not correlate with other people driven features.

Variable	by Variable	Spearman ρ	Prob>|ρ|
People talking	Tweets	-0.1194	0.2099
Video	Tweets	0.0509	0.6851
Video	People talking	0.2174	0.0844

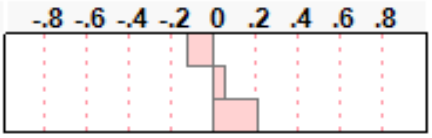


Figure 11: Correlation between Tweets, Videos, People Talking

Number of re-tweets in Twitter, number of post likes in Facebook and number of views in YouTube are very similar to each other. They represent the quality of the updates made by the user which makes followers to visit, re-tweet and like the update. And by the correlation analysis below, we can say that these feature correlate fairly with each other.

Variable	by Variable	Spearman ρ	Prob>|ρ|
FACEBOOK-Post likes	TWITTER-Retweets	0.2405	0.0096*
YOUTUBE-View	TWITTER-Retweets	0.3250	0.0078*
YOUTUBE-View	FACEBOOK-Post likes	0.5376	<.0001*

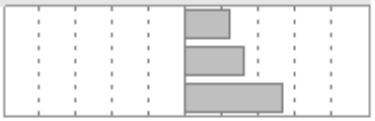


Figure 12: Correlation between Re-tweets, Post likes, Views

Number of favorites in Twitter, number of shares in Facebook and number of comments in YouTube are similar. People share or make the updates as their favorites when they find the update amusing. However, comments can be both positive and negative and we haven't differentiated it, but vaguely assume that more the number of comments, better the quality of the published video. And by the correlation analysis below we can say that number of comments correlates well with favorites and shares but they do not correlate with each other as well.

Variable	by Variable	Spearman ρ	Prob>|ρ|
FACEBOOK-Post shares	TWITTER-Favorites	0.1022	0.2769
YOUTUBE-Comment	TWITTER-Favorites	0.5268	<.0001*
YOUTUBE-Comment	FACEBOOK-Post shares	0.4873	<.0001*

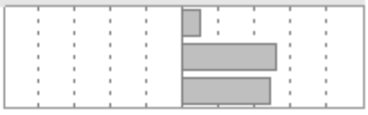


Figure 13: Correlation between Post shares, Favorites, Comments

CHAPTER 4

RELIABILITY OF SOCIAL NETWORK USER RANKING

As discussed in previous chapter, we found, we have many properties of users in different social networks which users commonly use to rank the accounts, but the question is if we can trust the social network ranking that is prevalent. And in order to do this we can compare it with accepted ranking methods and depending on how it compares we can judge on its trustworthiness.

4.1 Accepted Ranking Methods

4.1.1 U.S. News ranking

USNews ranking is called as the "Granddaddy" of the college rankings. It was established in 1983 and it is the most influential of all the college rankings. The popularity of U.S. News & World Report's college rankings is reflected in its release [45].

10 million page views in the first three days

80% of visitor access the ranking section of the website

A one-rank improvement leads to a 0.9% increase in number of participants

The printed issue of the college rankings sells 50% more than its normal issues.

Undergrad universities ranking criteria: it is the weighted average of 8 quantitative and qualitative criteria:

- Faculty resources (20%)
- Retention (20%)
- Peer assessment (15%)
- Student selectivity (15%)
- Financial resources (10%)

- High school guidance counselor assessment (7.5%)
- Graduation rate (7.5%)
- Alumni giving (5%)

Ranking Category	Category Weight		Subfactor	Subfactor Weight	
	National Universities and National Liberal Arts Colleges	Regional Universities and Regional Colleges		National Universities and National Liberal Arts Colleges	Regional Universities and Regional Colleges
Undergraduate academic reputation	22.5%	25%	Peer assessment survey	66.7%	100%
			High school counselors' ratings	33.3%	0%
Student selectivity for fall 2011 entering class	15%	15%	Acceptance rate	10%	10%
			High school class standing in top 10%	40%	0%
			High school class standing in top 25%	0%	40%
			Critical Reading and Math portions of the SAT and the composite ACT scores	50%	50%
Faculty resources for 2011-2012 academic year	20%	20%	Faculty compensation	35%	35%
			Percent faculty with top terminal degree in their field	15%	15%
			Percent faculty that is full time	5%	5%
			Student/faculty ratio	5%	5%
			Class size, 1-19 students	30%	30%
			Class size, 50+ students	10%	10%
Graduation and retention rates	20%	25%	Average graduation rate	80%	80%
			Average freshman retention rate	20%	20%
Financial resources	10%	10%	Financial resources per student	100%	100%
Alumni giving	5%	5%	Average alumni giving rate	100%	100%
Graduation rate performance	7.5%	0%	Graduation rate performance	100%	0%
Total	100%	100%	—	100%	100%

Figure 14: U.S. News Undergrad College Ranking Algorithm

4.1.2 Research Ranking

Microsoft academic search is a free academic search engine developed by Microsoft Research. It covers more than 48 million publications and over 20 million authors.

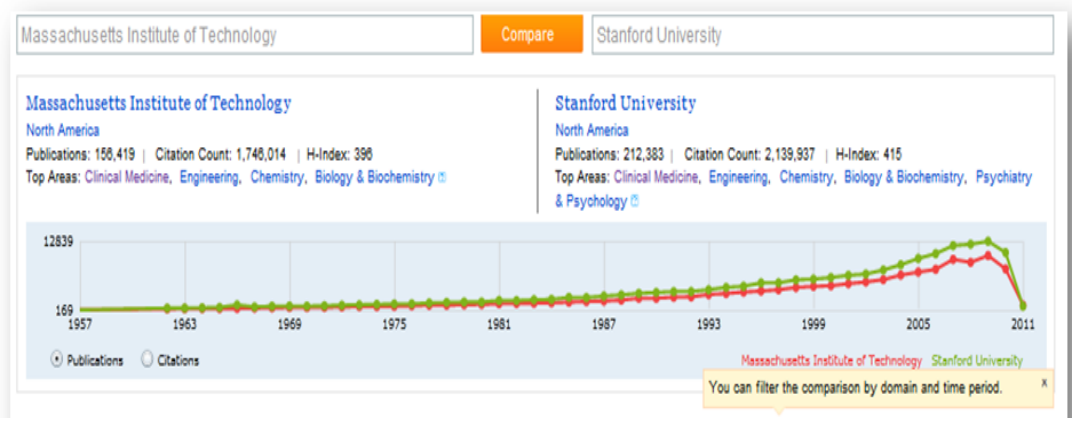


Figure 15: Microsoft Academic Search snapshot

We get two features for each university from Microsoft academic search and they are: Number of publications: Total number of publications of all the authors in the university. This number will give us a collective value of research happening in the university. And it doesn't show us the significance of the publications, in other words, quality of research is now a characteristic of this feature.

Citation count: Total number of citation of all the papers published by all the authors of the university. This number is a collective sum of all the research publications' citations. While it is highly dependent on many attributes, we could say that it will give us a picture of the quality of research of the university. However it is dependent on the popularity of the university, the subject of the paper, the popularity of the author, etc.

4.1.3 University Ranking by Academic Performance (URAP)

First published in 2010, the University Ranking by Academic Performance (URAP) was developed in the Informatics Institute of Middle East Technical University in Turkey and ranked 2,000 universities according to an aggregation of six academic research performance indicators: current productivity (number of published articles), long-term productivity (total documents from Institute for Scientific Information),

research impact (citations from Institute for Scientific Information), impact (cumulative journal impact), quality (Journal Citation Impact Total), and international collaboration.

Country Ranking	University Name	World Ranking	Category	Article	Citation	Total Document	JIT	JCIT	Collaboration	Total
1	Harvard University	1	A++	126.00	126.00	60.00	108.00	90.00	90.00	600.00
2	Johns Hopkins University	2	A++	119.54	124.15	57.57	107.89	86.34	74.95	570.45
3	Stanford University	4	A++	114.94	125.87	54.20	103.30	89.91	71.90	560.12
4	University of California Berkeley	5	A++	115.15	121.48	51.85	101.02	87.50	77.69	554.69
5	University of Washington Seattle	7	A++	117.15	122.52	54.96	101.60	84.58	70.55	551.36
6	University of Michigan Ann Arbor	8	A++	121.65	119.27	57.72	97.25	79.33	71.66	546.89
7	University of California Los Angeles	9	A++	114.16	120.33	55.40	96.93	80.81	72.41	540.05

Figure 16: University Ranking by Academic Performance (URAP) snapshot

4.2 Additional Features Collection

We needed a new dataset as the accepted ranking methods we chose were specific for universities. Therefore, we collected 150 universities' Twitter, Facebook and YouTube accounts that were mentioned in their website homepage.

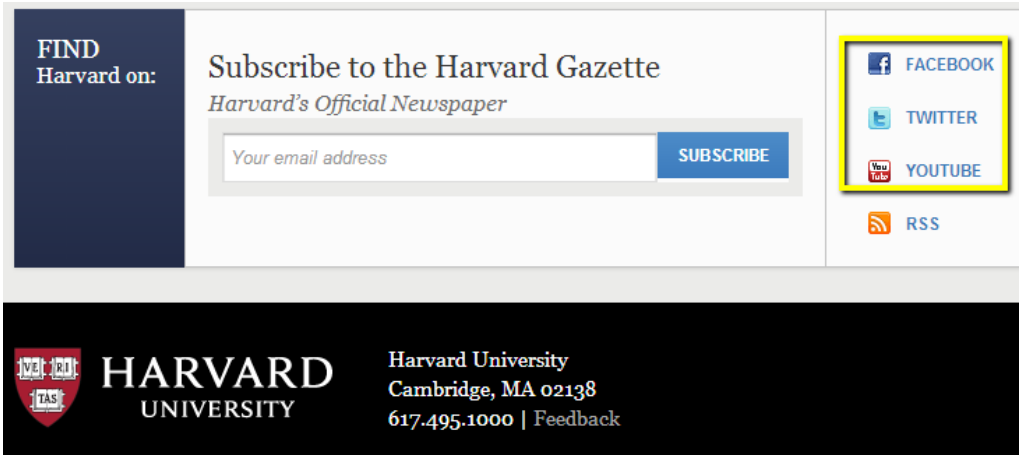


Figure 17: Harvard University Home Page

The new dataset has the following columns for each university:

- USNews: Rank
- Microsoft academic search: Number of publications, Citation count
- URAP: Score

- Twitter: Followers, Tweets, Re-tweets, Favorites
- Facebook: Page likes, people talking, post likes, post shares
- YouTube: Subscribers, videos, comments, views

Table 3: Harvard University Social Network Data

UNIVERSITY	Harvard
PUBLICATIONS	598771
CITATION COUNT	9,339,998
USNEWS RANK	1
SUM	600
FOLLOWERS	295703
TWEETS	16457
RE-TWEET COUNT	65496
FAVORITES	30839
LIKES	2380527
PEOPLE TALKING	51388
POST LIKES	2119500
POST SHARES	253991
VIEW	19122330
COMMENT	448
SUBSCRIBER	124177
VIDEO	1442

4.3 Correlation Analysis

We did an analysis between all the accepted ranking measures and all the social network features and the using the results we can decide on reliability of social networks.

4.3.1 With U.S. News Rank

We have 150 universities' social network and U.S. News data. Twelve measures of social networks were compared with the same user's USNews score. Twitter followers and YouTube subscribers were correlated with significant p-value and other measures had weak correlation.

USNEWS-USNEWS SCORE	TWITTER-FOLLOWERS	0.3480	<.0001*
USNEWS-USNEWS SCORE	TWITTER-TWEETS	0.1043	0.2072
USNEWS-USNEWS SCORE	TWITTER-RE-TWEET COUNT	0.0979	0.2364
USNEWS-USNEWS SCORE	TWITTER-Favorites	0.2776	0.0006*
USNEWS-USNEWS SCORE	FACEBOOK-LIKES	0.2680	0.0010*
USNEWS-USNEWS SCORE	FACEBOOK-PEOPLE TALKING	0.2179	0.0078*
USNEWS-USNEWS SCORE	FACEBOOK-Postlikes	0.1888	0.0220*
USNEWS-USNEWS SCORE	FACEBOOK-Postshares	0.2150	0.0089*
USNEWS-USNEWS SCORE	YOUTUBE-VIEW	0.2984	0.0002*
USNEWS-USNEWS SCORE	YOUTUBE-COMMENT	0.1721	0.0365*
USNEWS-USNEWS SCORE	YOUTUBE-SUBSCRIBER	0.3916	<.0001*
USNEWS-USNEWS SCORE	YOUTUBE-VIDEO	0.1834	0.0257*

Figure 18: Social Network Rank Correlation with U.S. News score

4.3.2 With Microsoft Academic Search Rank

The same twelve measures of social network were compared with the same user's Microsoft academic search number of publications and citation count score for 150 universities. All the social network features were correlated with significant p-value.

Microsoft Academic-Publications	TWITTER-FOLLOWERS	0.4780	<.0001*
Microsoft Academic-Publications	TWITTER-TWEETS	0.1931	0.0209*
Microsoft Academic-Publications	TWITTER-RE-TWEET COUNT	0.2124	0.0109*
Microsoft Academic-Publications	TWITTER-Favorites	0.4315	<.0001*
Microsoft Academic-Publications	FACEBOOK-LIKES	0.4905	<.0001*
Microsoft Academic-Publications	FACEBOOK-PEOPLE TALKING	0.4384	<.0001*
Microsoft Academic-Publications	FACEBOOK-Postlikes	0.4375	<.0001*
Microsoft Academic-Publications	FACEBOOK-Postshares	0.4430	<.0001*
Microsoft Academic-Publications	YOUTUBE-VIEW	0.3948	<.0001*
Microsoft Academic-Publications	YOUTUBE-COMMENT	0.2989	0.0003*
Microsoft Academic-Publications	YOUTUBE-SUBSCRIBER	0.5006	<.0001*
Microsoft Academic-Publications	YOUTUBE-VIDEO	0.2823	0.0006*

Figure 19: Social Network Rank Correlation with Research Publications Count

Microsoft Academic-Citation count	TWITTER-FOLLOWERS	0.4650	<.0001*
Microsoft Academic-Citation count	TWITTER-TWEETS	0.1837	0.0281*
Microsoft Academic-Citation count	TWITTER-RE-TWEET COUNT	0.1702	0.0421*
Microsoft Academic-Citation count	TWITTER-Favorites	0.4060	<.0001*
Microsoft Academic-Citation count	FACEBOOK-LIKES	0.4538	<.0001*
Microsoft Academic-Citation count	FACEBOOK-PEOPLE TALKING	0.4059	<.0001*
Microsoft Academic-Citation count	FACEBOOK-Postlikes	0.3930	<.0001*
Microsoft Academic-Citation count	FACEBOOK-Postshares	0.3932	<.0001*
Microsoft Academic-Citation count	YOUTUBE-VIEW	0.3821	<.0001*
Microsoft Academic-Citation count	YOUTUBE-COMMENT	0.2674	0.0012*
Microsoft Academic-Citation count	YOUTUBE-SUBSCRIBER	0.4892	<.0001*
Microsoft Academic-Citation count	YOUTUBE-VIDEO	0.2716	0.0010*

Figure 20: Social Network Rank Correlation with Research Citation Count

4.3.3 With URAP Rank

12 measures of social network were compared with the same URAP research score for 150 universities and all the social network features were correlated with significant p-value.

RESEARCH-SUM	TWITTER-FOLLOWERS	0.5904	<.0001*
RESEARCH-SUM	TWITTER-TWEETS	0.3289	<.0001*
RESEARCH-SUM	TWITTER-RE-TWEET COUNT	0.3033	0.0002*
RESEARCH-SUM	TWITTER-Favorites	0.5405	<.0001*
RESEARCH-SUM	FACEBOOK-LIKES	0.5895	<.0001*
RESEARCH-SUM	FACEBOOK-PEOPLE TALKING	0.5635	<.0001*
RESEARCH-SUM	FACEBOOK-Postlikes	0.5385	<.0001*
RESEARCH-SUM	FACEBOOK-Postshares	0.5533	<.0001*
RESEARCH-SUM	YOUTUBE-VIEW	0.4167	<.0001*
RESEARCH-SUM	YOUTUBE-COMMENT	0.2001	0.0147*
RESEARCH-SUM	YOUTUBE-SUBSCRIBER	0.5298	<.0001*
RESEARCH-SUM	YOUTUBE-VIDEO	0.2490	0.0023*

Figure 21: Social Network Rank Correlation with URAP

The results are quite interesting and it gives us the idea that social network ranking can be trusted.

CHAPTER 5

PROPOSED METHODS OF SOCIAL NETWORK RANKING

Apart from the standard four features of Twitter we collected and derived four more methods of ranking and they are explained as follows:

5.1 H-index

The idea was derived from the Hirsch index. The definition of h-index is as follows: A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np − h) papers have no more than h citations each.

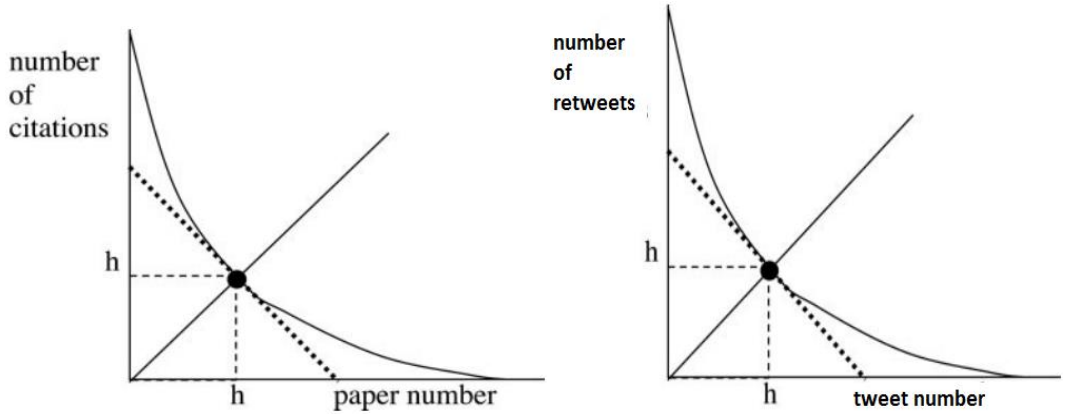


Figure 22: H-index for Social Networks

We calculated the h-index for each university based on the university's tweets and number of re-tweets just like papers and number of citations in the Hirsch index. Below in the steps we used to calculate the h-index for each university:

Create a list of all tweets and number of re-tweets (for each tweet)

Re-order the list by descending order of the number of re-tweets

Add an index starting from 1 for each tweet

H-index is the index value at which the re-tweet count of a tweet is equal to or larger than the line number of a given tweet

Table 4: Sample H-index Data

UNIVERSITY	TWITTER-H-INDEX
Columbia	7
University of Chicago	7
University of Pennsylvania	10
Duke	6
Dartmouth	6
Northwestern University	13
John Hopkins University	5
Washington University in St. Louis	8
Cornell University	17

5.2 Age of Account

Age of the social network is how old the Twitter account is. We got the account creation date and calculated Twitter account age in the "year.month" format. Age of account cannot be used to rank the user but as an additional social network information. Therefore we obtained the age of the Twitter account for each university in terms of years and months.

Table 5: Sample Social Network Account Duration Data

UNIVERSITY	TWITTER ACC DURATION
Columbia	2.2
University of Chicago	3
University of Pennsylvania	4.1
Duke	4
Dartmouth	4.2
Northwestern University	4
John Hopkins University	5
Washington University in St. Louis	4.5
Cornell University	4.5

5.3 Popularity of followed users

Each Twitter user has a total following number. This number is the total number of Twitter users that the current user is following. This number could mean multiple things, if the number was high we could think that this user is active in Twitter and likes to read others' tweets, or that the user reciprocates by following those that follow him, or that the user clicked on follow button on many users but never keeps track on the tweets, etc. Hence, this is also not a measure to rank a user so we went one level deeper. We calculated the total

number of followers that all the users (that current user is following) have and averaged it to total following count to get a following popularity index.

Table 6: Sample Twitter Average Following Popularity Data

UNIVERSITY	TWITTER-AVERAGE FOLLOWING POPULARITY
Columbia	142338.72
University of Chicago	42768.25
University of Pennsylvania	158837.96
Duke	90036.87
Dartmouth	11958.81
Northwestern University	87991.68
John Hopkins University	43967.37
Washington University in St. Louis	1556.20
Cornell University	168451.37

5.4 Average Re-tweet

Total re-tweet count is comparable to number of citations in research world, and even though number of citations is important, average citation for each publication gives more information on the quality of the author. Likewise in Twitter, average re-tweet count for tweets give a better picture of the user's tweeting quality. Because when a user is a frequent tweeter, the total re-tweets count might be because of the sheer number of tweets so average re-tweet count is another measure we added.

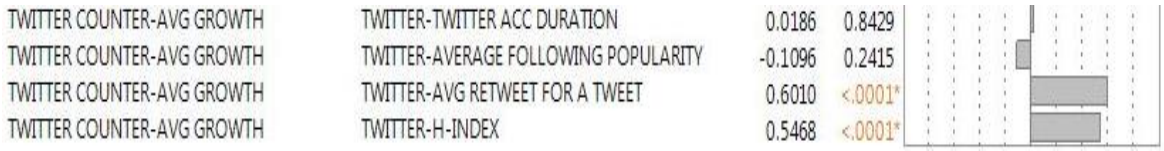
Table 7: Sample Average Re-tweet for a Tweet Data

UNIVERSITY	AVG RETWEET FOR A TWEET
Columbia	1.93
University of Chicago	2.29
University of Pennsylvania	2.64
Duke	2.17
Dartmouth	1.80
Northwestern University	5.37
John Hopkins University	2.00
Washington University in St. Louis	1.36
Cornell University	4.68

5.5 Comparison of New Methods with TwitterCounter

Twitter Counter is an analytics service for Twitter. It is the number 1 stats site for Twitter users. It provides statistics of Twitter usage and tracks over 94 million users and

counting. We compared the Twitter ranking methods with Twitter Counter's rank. H-index and average re-tweet correlated well with Twitter Counter score but Twitter account duration and average following popularity score had no or weak correlation.

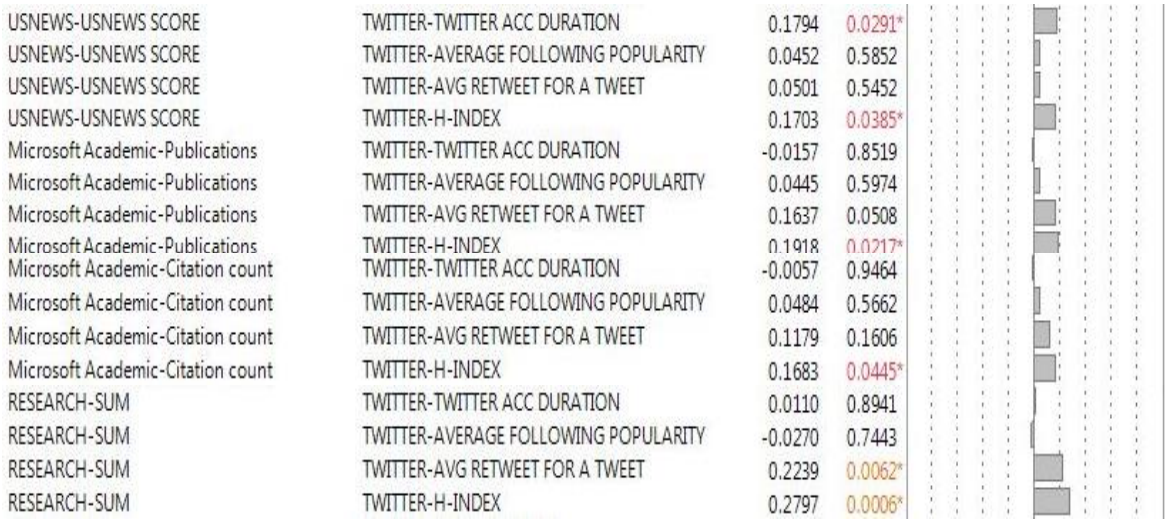


TWITTER COUNTER-AVG GROWTH	TWITTER-TWITTER ACC DURATION	0.0186	0.8429
TWITTER COUNTER-AVG GROWTH	TWITTER-AVERAGE FOLLOWING POPULARITY	-0.1096	0.2415
TWITTER COUNTER-AVG GROWTH	TWITTER-AVG RETWEET FOR A TWEET	0.6010	<.0001*
TWITTER COUNTER-AVG GROWTH	TWITTER-H-INDEX	0.5468	<.0001*

Figure 23: Correlation Analysis of New Methods with TwitterCounter

5.6 Correlation Analysis with Accepted Ranking Methods

We did correlation analysis of all the new ranking methods with accepted ranking methods. H-index measure has correlation with all accepted ranking methods, average re-tweet has correlation with some measures but the other two measures do not seem to correlate well. With USNews rank Twitter account duration seem to correlate weakly.



USNEWS-USNEWS SCORE	TWITTER-TWITTER ACC DURATION	0.1794	0.0291*
USNEWS-USNEWS SCORE	TWITTER-AVERAGE FOLLOWING POPULARITY	0.0452	0.5852
USNEWS-USNEWS SCORE	TWITTER-AVG RETWEET FOR A TWEET	0.0501	0.5452
USNEWS-USNEWS SCORE	TWITTER-H-INDEX	0.1703	0.0385*
Microsoft Academic-Publications	TWITTER-TWITTER ACC DURATION	-0.0157	0.8519
Microsoft Academic-Publications	TWITTER-AVERAGE FOLLOWING POPULARITY	0.0445	0.5974
Microsoft Academic-Publications	TWITTER-AVG RETWEET FOR A TWEET	0.1637	0.0508
Microsoft Academic-Publications	TWITTER-H-INDEX	0.1918	0.0217*
Microsoft Academic-Citation count	TWITTER-TWITTER ACC DURATION	-0.0057	0.9464
Microsoft Academic-Citation count	TWITTER-AVERAGE FOLLOWING POPULARITY	0.0484	0.5662
Microsoft Academic-Citation count	TWITTER-AVG RETWEET FOR A TWEET	0.1179	0.1606
Microsoft Academic-Citation count	TWITTER-H-INDEX	0.1683	0.0445*
RESEARCH-SUM	TWITTER-TWITTER ACC DURATION	0.0110	0.8941
RESEARCH-SUM	TWITTER-AVERAGE FOLLOWING POPULARITY	-0.0270	0.7443
RESEARCH-SUM	TWITTER-AVG RETWEET FOR A TWEET	0.2239	0.0062*
RESEARCH-SUM	TWITTER-H-INDEX	0.2797	0.0006*

Figure 24: Correlation Analysis of New Methods with Accepted Ranking

CHAPTER 6

COMPOSITE SOCIAL NETWORK RANKING

6.1 Methodology

In this chapter we have described our methodology of ranking social network user. We have 8 measures of ranking for Twitter, 4 measures for Facebook and 4 measures for YouTube. The steps are as follows:

- Finding clusters
- Selecting one attribute for each cluster based on variance

Hierarchical clustering method was used to find clusters and 4 clusters were detected based on linkage criteria on Ward's criterion [43].

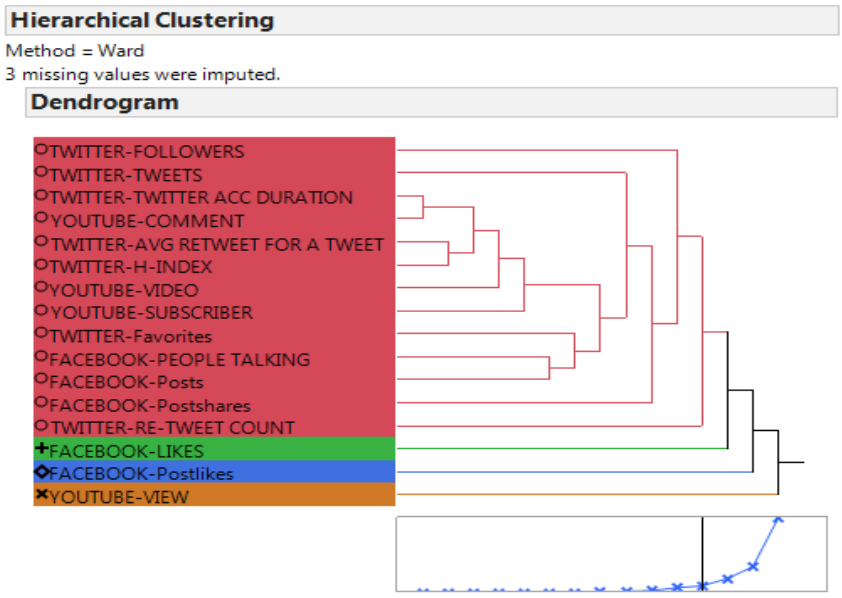


Figure 25: Hierarchical Clustering of Social Network Features Output

Therefore 4 clusters detected were:

1. Twitter: followers, tweets, comments, acc duration, average re-tweet, favorites, h-index, YouTube: subscriber, video. Facebook: people talking, post shares
2. Facebook: likes

3. Facebook: posts likes
4. YouTube: views

In the first cluster, Twitter followers was the attribute with maximum variance and therefore the 4 attributes that rank social networks are:

1. Twitter Followers
2. Facebook page likes
3. Facebook post likes
4. YouTube views

6.2 Correlation Analysis with Accepted Ranking Methods

All the four social network ranking measures correlate well with USNews rank, Microsoft academic search rank (publications, citations) and URAP score with very significant p-value.

Table 8: Correlation of Social Network Rank with Accepted Ranking

Accepted rank	Social Network rank	Spearman coefficient	p value
USNews score	Twitter Followers	0.348	<0.0001
USNews score	Facebook page likes	0.268	0.001
USNews score	Facebook post likes	0.1888	0.022
USNews score	YouTube views	0.2984	0.0002
Citation count	Twitter Followers	0.465	<0.0001
Citation count	Facebook page likes	0.4538	<0.0001
Citation count	Facebook post likes	0.393	<0.0001
Citation count	YouTube views	0.3821	<0.0001
Publications	Twitter Followers	0.478	<0.0001
Publications	Facebook page likes	0.4905	<0.0001
Publications	Facebook post likes	0.4375	<0.0001
Publications	YouTube views	0.3948	<0.0001
URAP score	Twitter Followers	0.5904	<0.0001
URAP score	Facebook page likes	0.5895	<0.0001
URAP score	Facebook post likes	0.5385	<0.0001
URAP score	YouTube views	0.4167	<0.0001

CHAPTER 7

DATA MINING STUDY

7.1 Rank Difference Calculation

In this chapter, we are going to describe the data mining process and interesting patterns observed. By the end of last chapter, we described the 4 measures that we will be using to rank social networks which are followers, page likes, post likes and views. We have to define a method to find rank difference for classification purpose. And that is described as follows:

- Calculate rank difference between one accepted rank and all four social network ranks.
- Calculate standard deviation of the ranks
- Calculate total rank difference
- Filter out users with large deviation in ranks
- Using the rest of the users with minimum deviation, use the rank difference to find classes

Table 9: Rank Difference Calculation between Social Network and U.S. News

UNIVERSITY	Standard deviat	Difference total	UN-Followers	UN-Likes	UN-Postlikes	UN-Views
Colorado School of Mines	-33.97	-100	-24	-25	-22	-29
Lehigh University	-33.66	-197	-46	-56	-52	-43
Rensselaer Polytechnic Inst	-23.55	-245	-69	-60	-69	-47
Wake Forest University	-18.93	-257	-62	-57	-54	-84
Brandeis University	-14.58	-234	-36	-66	-73	-59
SUNY College of Environmer	-13.86	-112	-35	-33	-27	-17
Washington University in S	-12.17	-215	-58	-61	-68	-28
University of California-Sar	-9.87	-231	-66	-23	-74	-68
Stevens Institute of Technol	-8.59	-90	-33	-29	-18	-10

7.2 Universities that have Positive Correlate with U.S. News Rank

Universities that fall under this category will be those with minimum rank difference total (sum of rank differences between USNews and 4 social network ranks) and having less standard deviation among the four rank differences. University name and total rank difference is given below:

Table 10: Universities that have Positive Correlate with U.S. News Rank

UNIVERSITY	Rank difference
University of San Diego	-14
University of California-Davis	-18
Tulane University	-16
Northeastern University	1
Drexel university	4
University of Notre Dame	8
University of North Carolina-Chapel Hill	11
Rutgers, The State University of New Jersey - New Brunswick	25
University of California - Santa Cruz	19

7.3 Universities that have Negative Correlation with U.S. News Rank

Universities that fall under this category will be those with maximum rank difference total (sum of rank differences between USNews and 4 social network ranks) and having less standard deviation among the four rank differences. University name and total rank difference is given below:

Table 11: Universities that Inverse Correlate with U.S. News rank

UNIVERSITY	Difference total
Colorado School of Mines	-100
Lehigh University	-197
Rensselaer Polytechnic Institute	-245
Wake Forest University	-257
Brandeis University	-234
SUNY College of Environmental Science and Forestry	-112
Washington University in St. Louis	-215
St. Johns Fisher College	168
University of Kansas - Lawrence	376
Oregon State University	478
University of Cincinnati	287
University of South California	385
Arizona State University	501

7.4 Universities that have Positive Correlation with Research Rank

Universities that fall under this category will be those with minimum rank difference total (sum of rank differences between Microsoft citations count rank and 4 social network ranks) and having less standard deviation among the four rank differences. University name and total rank difference is given below:

Table 12: Universities that Correlate with Research Rank

UNIVERSITY	Rank difference
Northwestern University	-33.00
University of Southern California	-20.00
University of Iowa	-2.13
Clarkson University	-1.88
University of San Francisco	-1.17
Binghamton University - SUNY	0.00
University of St. Thomas	0.00
Temple University	0.04
University of Colorado-Boulder	0.17
University of the Pacific	0.71
Fordham University	0.79
Miami University - Oxford	0.94
Washington State University, Pullman	1.18
Colorado School of Mines	1.36
University of Virginia	1.45
Stevens Institute of Technology	1.72

7.5 Universities that have Negative Correlation with Research Rank

Universities that fall under this category will be those with maximum rank difference total (sum of rank differences between Microsoft citations count rank and 4 social network ranks) and having less standard deviation among the four rank differences. University name and total rank difference is given below:

Table 13: Universities that Inverse Correlate with Research Rank

UNIVERSITY	Rank difference
University of Vermont	-267
St. Louis University	-211
Wake Forest University	-301
Brandeis University	-274
Washington University in St. Louis	-263
University of Illinois, Chicago	-211
University of Notre Dame	260
University of Illinois-Urbana Champaign	295
Purdue University - West Lafayette	202
Virginia Tech	197
University of Tennesse	374

7.6 University Information Sources

7.6.1 General Information

This chapter details the list of information obtained about the subjects of interest to perform data mining to try to understand and reason the rank difference. Below is a table which contains different details of universities collected from their websites and from U.S News website.

Table 14: List of Universities General Information Data

ATTRIBUTE	DESCRIPTION
Selectivity	The college's relative performance in this measure of how much competition applicants face for admission. (Most selective/More selective/Selective/Less selective)
High school counselor score	The school's average score on a survey asking high school counselors to rate its undergraduate academic quality on a scale of 1 (marginal) to 5 (distinguished) or "don't know"
Average	The percentage of freshmen who returned to the college the following fall,

ATTRIBUTE	DESCRIPTION
freshman retention rate	averaged over the first-year classes entering between fall 2008 through fall 2011.
Classes with fewer than 20 students	The percentage of undergraduate classes, excluding class subsections, with fewer than 20 students enrolled during fall 2012.
Classes with 50 or more students	The percentage of undergraduate classes, excluding class subsections, with 50 students or more enrolled during fall 2012.
Student-faculty ratio	The number of FTE undergraduate students per FTE faculty
Fall 2012 acceptance rate	The ratio of the number of students admitted to the number of applicants for fall 2012 admission. The acceptance rate is equal to the total number of students admitted divided by the total number of applicants.
6-year graduation rate	The percentage of an entering class that graduated within six years for the most recent cohort of students.
Predicted graduation rate	The percentage of students who should graduate from the college, based on characteristics of the entering class and the characteristics of the institution.
Overperformance(+)/Underperformance	The difference between the actual six-year graduation rate for students entering in the fall of 2005 and the U.S. News predicted graduation rate.

ATTRIBUTE	DESCRIPTION
(-)	
Graduation and retention rank	The college's relative performance in these measures of the percentages of students who graduate in six years and freshmen who return the following fall.
Peer assessment score	The school's average score on a survey asking top college administrators to rate its undergrad academic quality on a scale of 1 (marginal) to 5 (distinguished) or "don't know"
Faculty resources rank	The school's relative performance in this measure that includes class sizes, student-faculty ratio, and faculty salaries.
Percent of faculty who are full-time	Percentage of faculty who are working full-time in the University
Student selectivity rank	The college's relative performance in this measure of how much competition applicants face for admission.
SAT/ACT 25th-75th percentile	25 percent of the college's students scored at or below the lower end of this range, and 25 percent scored at or above the upper end
Freshmen in top 10 percent of	The proportion of students enrolled for the academic year beginning in fall 2012 who graduated in the top 10 percent of their high school class

ATTRIBUTE	DESCRIPTION
high school class	
Freshmen in top 25 percent of high school class	The proportion of students enrolled for the academic year beginning in fall 2012 who graduated in the top 25 percent of their high school class
Financial resources rank	The college's relative performance in this measure of the average spent per student on instruction, research, student services, and educational expenditures
Alumni giving rank	The college's relative performance in this proxy measure for student satisfaction.
Average alumni giving rate	The average percentage of undergraduate alumni of record who donated money to the college or university. Alumni of record are former full- or part-time students who received an undergraduate degree and for whom the college or university has a current address.
Tuition and fees	Total tuition/fees
Students enrolled	Total number of undergraduates enrolled in Fall 2012
Private/Public	If the school is private or public

ATTRIBUTE	DESCRIPTION
Setting	College Town/Urban/Rural/Suburban

7.6.2 Student Ratings

To make a university active in social networks, type of students must have a large influence and therefore we collected data from Unigo which is a free online college resource guide and student platform claiming to cover more than 1,600 colleges and universities in the United States [26]. The Unigo website is used by college students to share photos, videos, documents, and reviews of their school [27]. High school students and parents use the site as a research tool to explore college options. Unigo's main purpose is to create a student-generated online college guide that does not have the limitations that its print counterparts do [28][29]. This allows college students to update information about their school on a continuous basis and cover topics not found in traditional guidebooks.

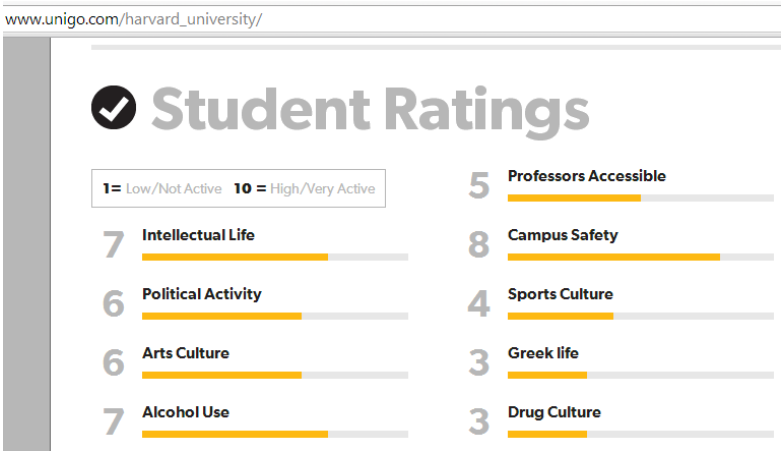


Figure 26: Unigo Student Ratings Snapshot

The student ratings information collected from Unigo are given below:

- Intellectual Life
- Political activity

- Arts culture
- Alcohol use
- Professors accessibility
- Campus safety
- Sports culture
- Greek life
- Drug culture

For each of the above factor, the university is given a score between 1 and 10, 1 meaning low/not active and 10 meaning high/very active.

7.6.3 Research Information

Apart from student ratings and other general information about the university, we wanted to collect the research statistics also because it would provide more insights on the quality of education. Hence we included the following details for each university. All the data was obtained from www.urapcenter.org

Table 15: Universities Research Information Description

ATTRIBUTE	DESCRIPTION
Journal impact	Measure of scientific impact which is derived by aggregating the impact factors of journals in which a university published articles between 2008 and 2012.
Journal citation impact	Measure of received citation quality which is based on the impact factors of journals where the citing articles are published.
International collaboration	Measure of global acceptance of a university. International collaboration data is based on the total number of publications made in collaboration with foreign universities.

7.7 Classification Based on U.S. News Rank

In this chapter, we are going to be finding patterns in the data collected for classifying rank differences between social network and USNews. For recap, we define the following again:

- We have selected 4 features to represent social network ranking. They are Twitter followers, Facebook page likes, Facebook post likes and YouTube video views.
- To find rank difference, we subtract all the 4 social network feature ranks from USNews rank for each university.
- We calculate standard deviation between all the 4 rank differences which tell us if a university is consistently different from USNews rank or not.
- We define rules for each class based on our goal defined below in each topic after carefully studying the dataset properties.

7.7.1 Positive and Negative Correlation Graphs

In this section, we would like to find patterns in the dataset for classifying universities that correlate well in their social network presence and their USNews ranking with the universities that have inverse correlate in their social network presence and USNews ranking (we do not consider direction in this section). The rules for classification are defined below.

Rules for correlate class:

- The ratio (Sum of rank differences/Standard deviation) must be between -4 to +4
- Rank difference must be between -50 and +50

After filtering, we had a total of 19 universities that fell under the correlate class.

Table 16: Universities that Correlate with U.S. News rank

UNIVERSITY	Sum of rank / Std deviation	Std Deviation	Sum rank difference
Boston College	-1.46	29.95	-44
University of Virginia	-3.51	9.67	-34
University of California-Davis	-0.85	21.11	-18
Tulane University	-0.84	18.93	-16
University of San Diego	-1.54	9.03	-14
Northeastern University	0.04	24.93	1
Drexel university	0.31	12.72	4
University of Notre Dame	0.71	11.19	8
University of North Carolina Chapel Hill	0.77	14.26	11
Cornell University	1.96	7.63	15
University of California Santa Cruz	1.50	12.63	19
University of Vermont	2.60	8.05	21
University of Southern California	3.04	7.87	24
Rutgers, SUNY New Brunswick	1.16	21.51	25
St. Louis University	2.72	9.17	25
Illinois Institute of Technology	3.75	8.52	32
University of the Pacific	3.97	9.81	39
University of Delaware	3.48	13.78	48
University of Miami	1.17	42.57	50

Rules for inverse correlate class:

- The ratio (Sum of rank differences/Standard deviation) must be less than -12 and greater than 30
- Rank difference must be less than -100 or greater than +130

After filtering, we had a total of 17 universities that fell in inverse correlate class.

Table 17: Universities that Inverse Correlate with U.S. News rank

UNIVERSITY	Sum of rank/Std deviation	Std Deviation	Sum rank difference
Wake Forest University	-18.93	13.57	-257
Rensselaer Polytechnic Institute	-23.54	10.40	-245
Brandeis University	-14.57	16.05	-234
Washington University in St. Louis	-12.16	17.67	-215
Lehigh University	-33.66	5.85	-197
SUNY College of Environmental Science and Forestry	-13.85	8.08	-112
Colorado School of Mines	-33.96	2.94	-100
South Carolina State University	75.05	1.73	130
St. Johns Fisher College	35.81	4.69	168
Virginia Tech	32.66	5.90	193
Purdue University – West Lafayette	34.26	6.24	214
University of Cincinnati	41.31	6.94	287
University of Tennessee	31.89	10.47	334
University of Kansas – Lawrence	37.47	10.03	376

UNIVERSITY	Sum of rank/Std deviation	Std Deviation	Sum rank difference
University of South California	42.28	9.10	385
Oregon State University	40.16	11.90	478
Arizona State University	59.21	8.46	501

Using simple visualization techniques, the data was analyzed first and below are some interesting histograms that are worth explaining.

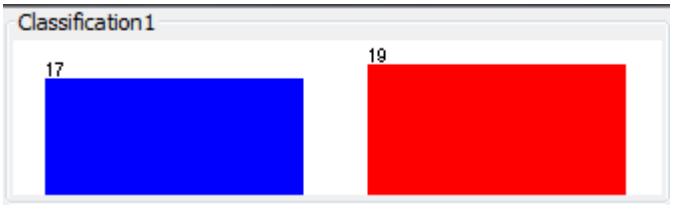


Figure 27: USNews Correlate & Inverse Correlate histogram

The histogram colored in red represents Correlate class with 19 data points and the one colored in blue represents Inverse correlate class with 17 data points.

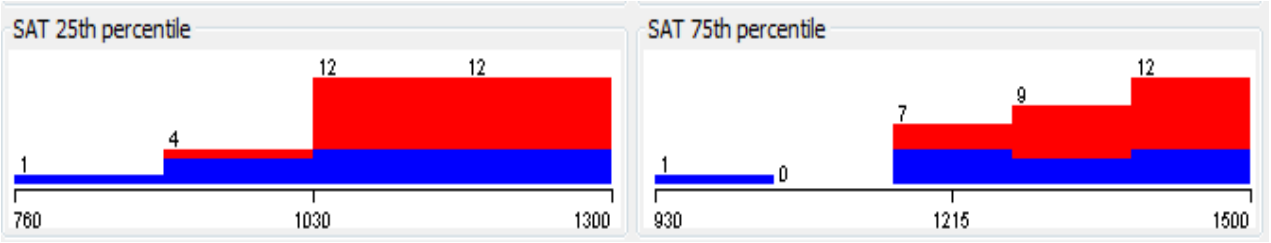


Figure 28: SAT percentile distribution for USNews Correlate & Inverse Correlate

The correlate class seems to be inclined towards higher SAT 25th and 75th percentile score whereas the inverse correlate class seems to be equally distributed through the middle and high scores with some data falling on the lower end of the score also.

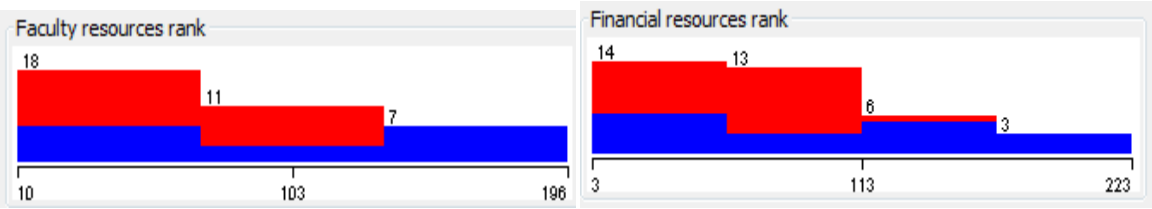


Figure 29: Faculty and Financial Resources distribution for USNews Correlate & Inverse Correlate

Correlate class is denser on the lower end of the faculty and financial resources rank which suggests that universities that fall in this category have better faculty and financial resources. The inverse correlated universities are spread across the limits of faculty and financial resource ranks.

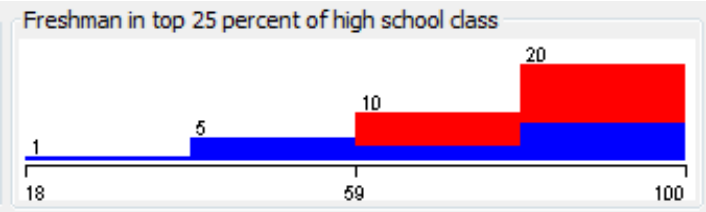


Figure 30: Freshman in top 25% distribution for USNews Correlate & Inverse Correlate

This graph suggests that universities that fall under correlate class have more number of their freshman in top 25 percent of high school class. So universities that have consistent ranking in USNews and social networks tend to have more students from top 25 percent of high school class. And the inverse correlated universities are widespread from 50 to 100%.

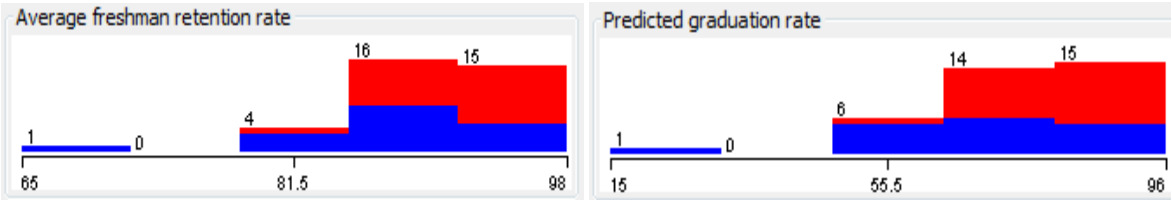


Figure 31: Retention and Graduation rate distribution for USNews Correlate & Inverse Correlate

This graph also suggests that universities that are consistent in their academic ranking and social network ranking have better freshman retention rate and graduation rate on the whole

than the ones that are not consistent. The lower retention and graduation rates are found in universities that are inversely correlated.

7.7.2 Low Marketing and High Marketing Graphs

In this section, we would like to find patterns in the dataset for classifying universities that are ranked high in USNews and low in social networks (Low marketing) and that are ranked high in social networks and low in USNews (High marketing) assuming that the presence of social network activity means marketing. The rules for classification are defined below.

Rules for low marketing class:

- The ratio (Sum of rank differences/Standard deviation) must be large with less standard deviation.
- Rank difference must be less than or equal to -90

After filtering, we had a total of 17 universities that fell under the low marketing class.

Table 18: Universities in Low Marketing w.r.t U.S. News rank

UNIVERSITY	Sum of rank/Std deviation	Std Deviation	Sum rank difference
Wake Forest University	-18.93	13.57	-257
Rensselaer Polytechnic Institute	-23.54	10.40	-245
Brandeis University	-14.57	16.05	-234
University of California-San Diego	-8.54	27.14	-232
University of California-Santa Barbara	-9.86	23.41	-231
Washington University in St. Louis	-12.16	17.67	-215
Lehigh University	-33.66	5.85	-197

UNIVERSITY	Sum of rank/Std deviation	Std Deviation	Sum rank difference
Yeshiva University	-6.07	30.77	-187
Tufts University	-6.36	28.42	-181
Rice University	-5.11	30.70	-157
University of Chicago	-6.39	21.40	-137
SUNY College of Environmental Science and Forestry	-13.85	8.08	-112
Worcester Polytechnic Institute	-4.72	23.70	-112
University of Rochester	-3.18	32.99	-105
Colorado School of Mines	-33.96	2.94	-100
Duke	-3.61	25.44	-92
Stevens Institute of Technology	-8.59	10.47	-90

Rules for high marketing class:

- The ratio (Sum of rank differences/Standard deviation) must be greater than +15
- Rank difference must be greater than or equal to +200

After filtering, we had a total of 20 universities that fell in high marketing class.

Table 19: Universities in High Marketing w.r.t U.S. News rank

UNIVERSITY	Sum of rank/Std deviation	Std Deviation	Sum rank difference
University of Dayton	15.66	13.59	213

Purdue University – West Lafayette	34.26	6.24	214
University of Albany-SUNY	15.24	14.16	216
Michigan State University	20.50	11.26	231
Hofstra University	15.53	18.01	280
University of Cincinnati	41.31	6.94	287
North Carolina State University – Raleigh	15.46	19.91	308
Washington State University, Pullman	16.15	19.5	315
University of Rhode Island	24.61	13.40	330
University of Tennessee	31.89	10.47	334
Ohio University	16.97	19.79	336
University of Utah	17.24	19.88	343
University of Kansas – Lawrence	37.47	10.03	376
University of South California	42.28	9.10	385
University of Kentucky	17.48	22.47	393
Kansas State University	28.91	14.45	418
Colorado State University	21.46	19.70	423
Oklahoma State University	25.72	17.41	448
Oregon State University	40.16	11.90	478
Arizona State University	59.21	8.46	501

Using simple visualization techniques, the data was analyzed first and below are some very striking differences that can be identified easily using these histograms.

Figure 32: Low & High Marketing w.r.t USNews histogram

The histogram colored in blue represents Low marketing class with 17 data points and the one colored in red represents High marketing correlate class with 20 data points.

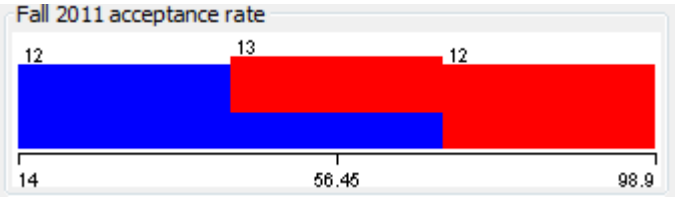


Figure 33: Acceptance rate distribution for Low & High Marketing w.r.t USNews

The acceptance rate is strikingly lower for the Low marketing class (universities that are ranked higher in USNews but have low social network presence) and High marketing class have higher acceptance rate.

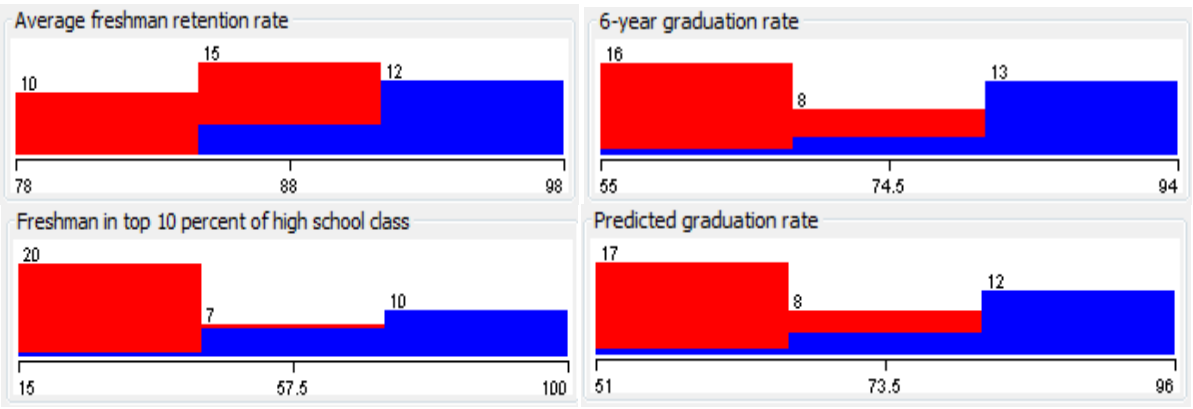


Figure 34: Academia Features distribution for Low & High Marketing w.r.t USNews

All the above attributes represent the quality of education academically. And the low marketing class has higher retention rate, graduation rate and has their freshman in top 10 percent of high school class. The difference is visibly large and the high marketing class is distinctly lower in values for these attributes that measure the academic performance.

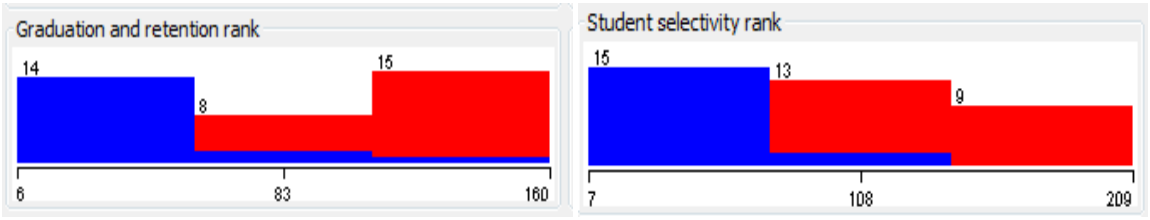


Figure 35: Graduation and Selectivity distribution for Low & High Marketing w.r.t USNews

High marketing class has better ranks for graduation and retention rank and student selectivity rank which mean that they have strict selection criteria to achieve better graduation rank. And the low marketing class is on the other side of the spectrum.

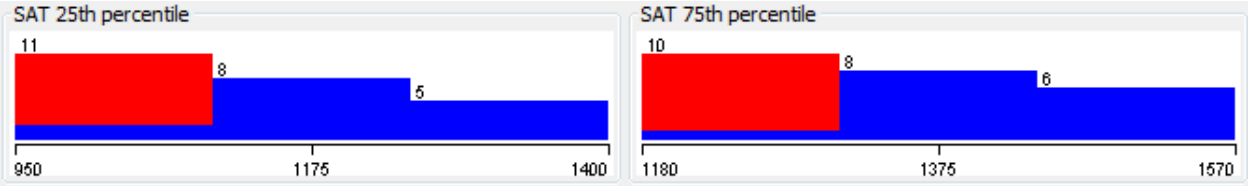


Figure 36: SAT percentile distribution for Low & High Marketing w.r.t USNews

Just as we would have guessed, the SAT percentile is much higher for low marketing class and lower for high marketing class.

Figure 37: Student Faculty ratio and Under 20 classes distribution for Low & High Marketing w.r.t USNews

Student faculty ratio is the number of number of FTE students per FTE faculty and for low marketing universities they have low student faculty ratio then high marketing universities. Percentage of classes with under 20 students are higher (near 75%) for low marketing universities than high marketing ones.

Other interesting information is explained in the histograms below:

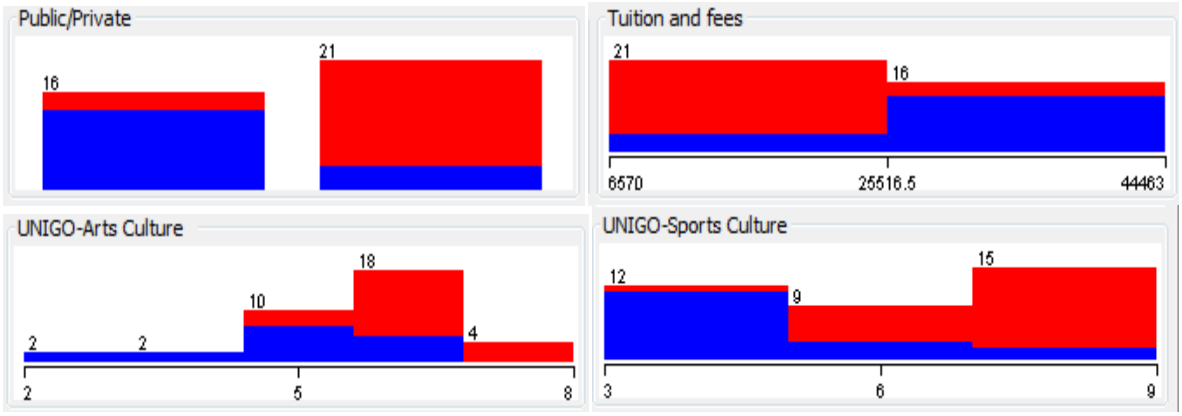


Figure 38: General information distribution for Low & High Marketing w.r.t USNews

Number of private universities in low marketing class is higher and number of public universities in high marketing class is higher. Low marketing universities fall in the higher end of the tuition spectrum than the low marketing ones. We can see a pattern in UNIGO student ranking for these 2 classes. Both for arts and sports low marketing universities are ranked lower than the high marketing universities.

7.7.3 Feature Selection

Before applying any data mining algorithm, it is important to do dimensionality reduction if we have a large set of features like in our case. Therefore this chapter aims to explain the methods we used for feature selection. We have used information gain as the measure to select features. Information gain of an attribute is the change in entropy from the dataset before applying the attribute as a condition for split and the dataset after split. We want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned. Information gain tells us how important a given attribute of the feature vectors is.

We have two classification problems, one is correlate and inverse correlate classes and the other one is low and high marketing classes. When we deployed information gain

feature selection method for the first set of data the following attributes were selected in the order of information gain of each attribute.

```
Ranked attributes:
 0.2491   14 Freshman in top 25 percent of high school class
 0.2491   32 Faculty resources rank
 0.2244    6 6-year graduation rate
 0.2077   34 Financial resources rank
 0.1441   15 Setting
 0.0547   16 Public/Private
 0.0393   18 Selectivity
 0        13 Freshman in top 10 percent of high school class
```

Figure 39: Feature selection for Correlate & Inverse correlate (USNews)

Next, we selected features for the low/high marketing dataset and the following are the attributes that have significant information gain for classification.

```
Ranked attributes:
 0.8385   13 Freshman in top 10 percent of high school class
 0.8385   11 Predicted graduation rate
 0.746    33 Student selectivity rank
 0.7339   30 Fall 2011 acceptance rate
 0.7339   32 Faculty resources rank
 0.6656   14 Freshman in top 25 percent of high school class
 0.648    17 Undergraduates
 0.648    31 Graduation and retention rank
 0.648     6 6-year graduation rate
 0.5215   19 Tuition and fees
 0.5075    5 Average freshman retention rate
 0.5075    8 Student-faculty ratio
 0.5016    7 Classes with under 20 students
 0.4674   18 Selectivity
 0.3928   34 Financial resources rank
 0.3589   10 Undergraduate academic reputation index
 0.3589    4 High school counselor score
 0.3494   35 Alumni giving rank
 0.3416   24 UNIGO-Sports Culture
 0.3239   41 Average alumni giving rate
 0.2955   16 Public/Private
 0.2955    9 Peer assessment score
 0.2638   38 SAT 75th percentile
 0.2162   37 SAT 25th percentile
 0.214    25 UNIGO-Arts Culture
 0.1005   15 Setting
```

Figure 40: Feature selection for Low & High marketing (USNews)

7.7.4 Linear Regression

In this chapter we are going to perform regression analysis on our dataset to find interesting equations. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [42].

On the feature selected dataset we perform linear regression with a statistical analysis tool called JMP and below are the results. The predictor variable in this case is the rank difference and the dependent variables are:

- Average freshman retention rate
- 6 year graduation rate
- Student faculty ratio
- Predicted graduation rate
- Freshman in top 10 percent of high school class
- Total undergraduates
- UNIGO sports score
- Fall 2011 acceptance rate
- Graduation and retention rate

It is important we understand the meaning of the explanatory variable (rank difference here). Rank difference varies from -300 to +600. This is the sum of rank differences between USNews rank and the four social network ranks. Therefore a negative rank difference means it is ranked higher in USNews and lower in social network and a positive rank difference means it is ranked lower in USNews and higher in social network. Therefore universities that end up in the lower end of the distribution in the limits (-300,600) are those that have lower marketing and those universities that end up in the higher end of the

distribution in the limits (-300,600) are those that have higher marketing. Below are the graphs and linear regression equations for the above mentioned dependent variables.

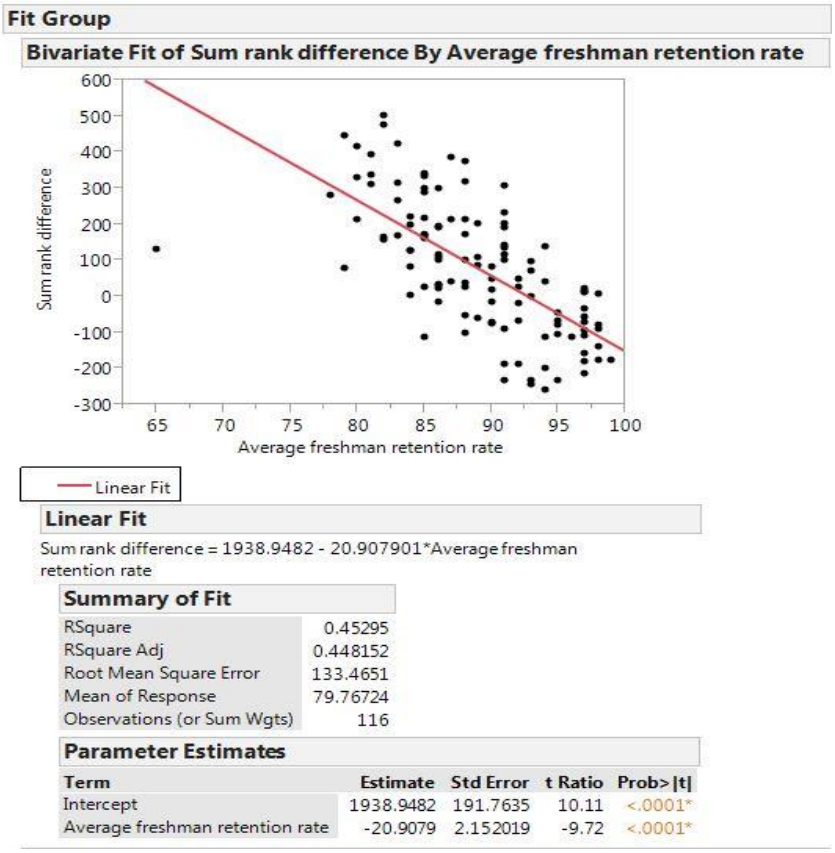
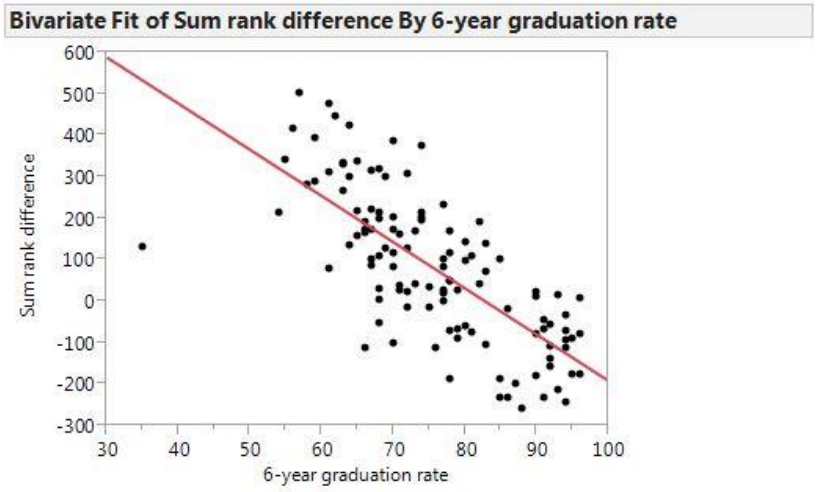


Figure 41: Bivariate fit of Rank difference by Freshman Retention rate (USNews)

From the above graph, we can interpret that average freshman retention rate inversely correlate with the rank difference (i.e.) data points with negative rank differences (low marketing) have higher average freshman retention rate and data points with positive rank differences (high marketing) have lower average freshman retention rate. The coefficient of the linear fit for average freshman retention rate is approximately -20 for predicting the rank difference.



Linear Fit

Sum rank difference = 923.08207 - 11.148093*6-year graduation rate

Summary of Fit

RSquare	0.530811
RSquare Adj	0.526696
Root Mean Square Error	123.6027
Mean of Response	79.76724
Observations (or Sum Wgts)	116

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	923.08207	75.13916	12.28	<.0001*
6-year graduation rate	-11.14809	0.981639	-11.36	<.0001*

Figure 42: Bivariate fit of Rank difference by 6-year Graduation rate (USNews)

From the above graph, we can interpret that 6 year graduation rate also inversely correlates with the rank difference (i.e.) data points with negative rank differences (high marketing) have higher 6 year graduation rate and data points with positive rank differences (low marketing) have lower 6 year graduation rate. The coefficient of the linear fit for 6 year graduation rate is approximately -11 for predicting the rank difference.

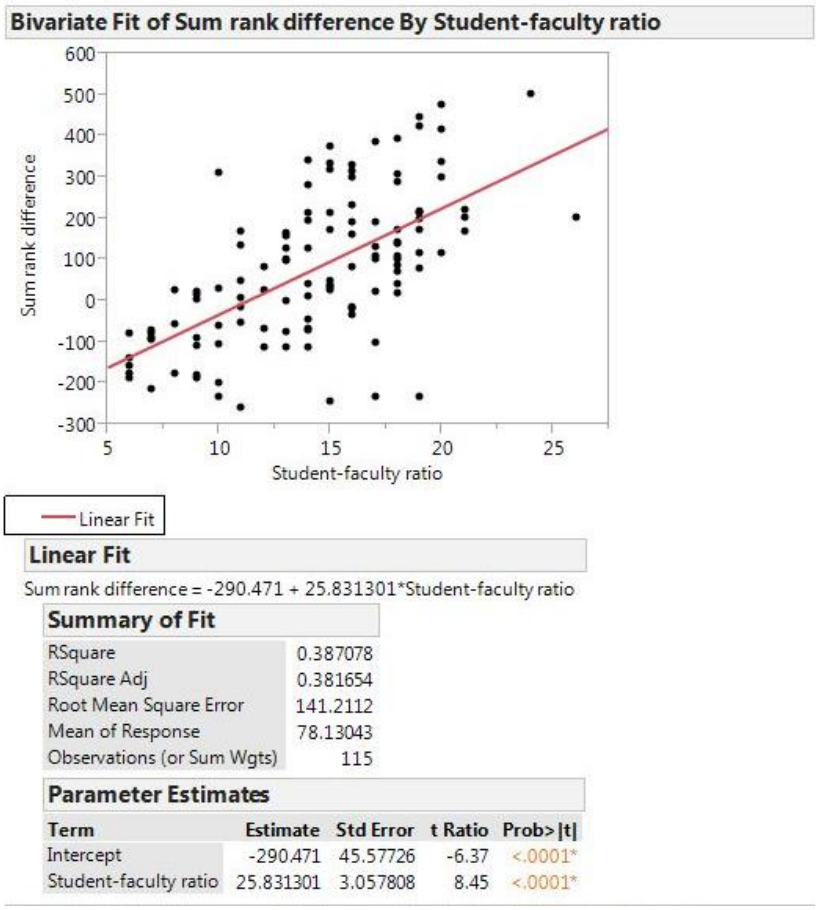


Figure 43: Bivariate fit of Rank difference by Freshman Retention rate (USNews)

From the above graph, we can interpret that student faculty ratio directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher student faculty ratio (more students per faculty) and data points with negative rank differences (low marketing) have lower student faculty ratio (less students per faculty). The coefficient of the linear fit for student faculty ratio is approximately +26 for predicting the rank difference.

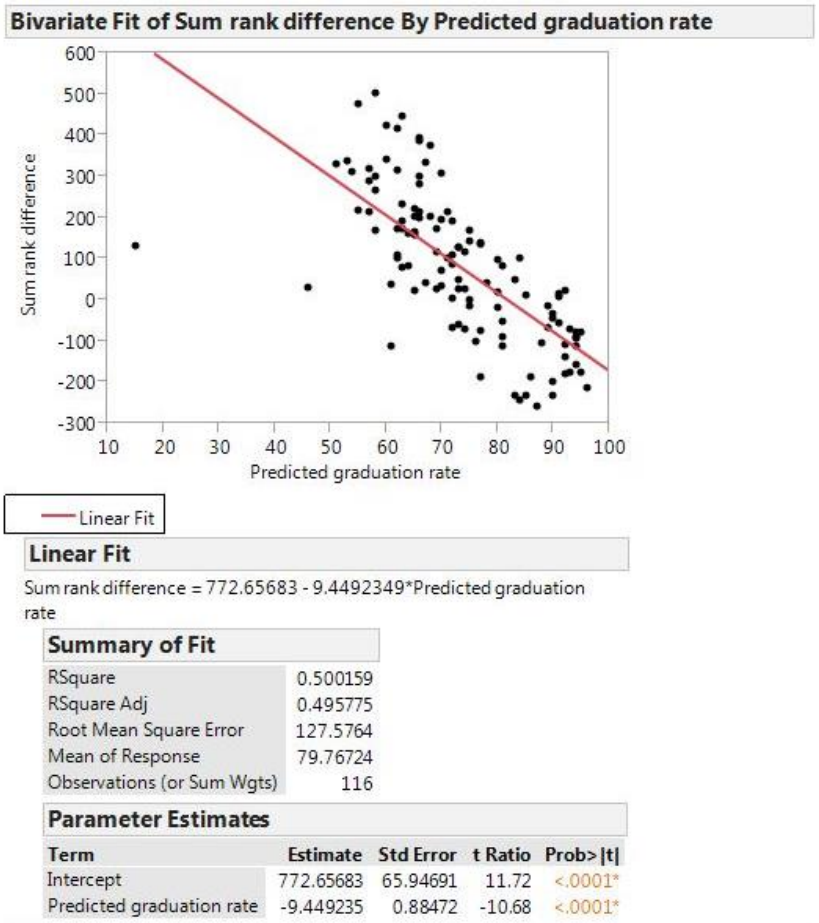


Figure 44: Bivariate fit of Rank difference by Predicted Graduation rate (USNews)

From the above graph, we can interpret that predicted graduation rate inversely correlate with the rank difference (i.e.) data points with negative rank differences (low marketing) have higher predicted graduation rate and data points with positive rank differences (high marketing) have lower predicted graduation rate. The coefficient of the linear fit for predicted graduation rate is approximately -9 for predicting the rank difference.

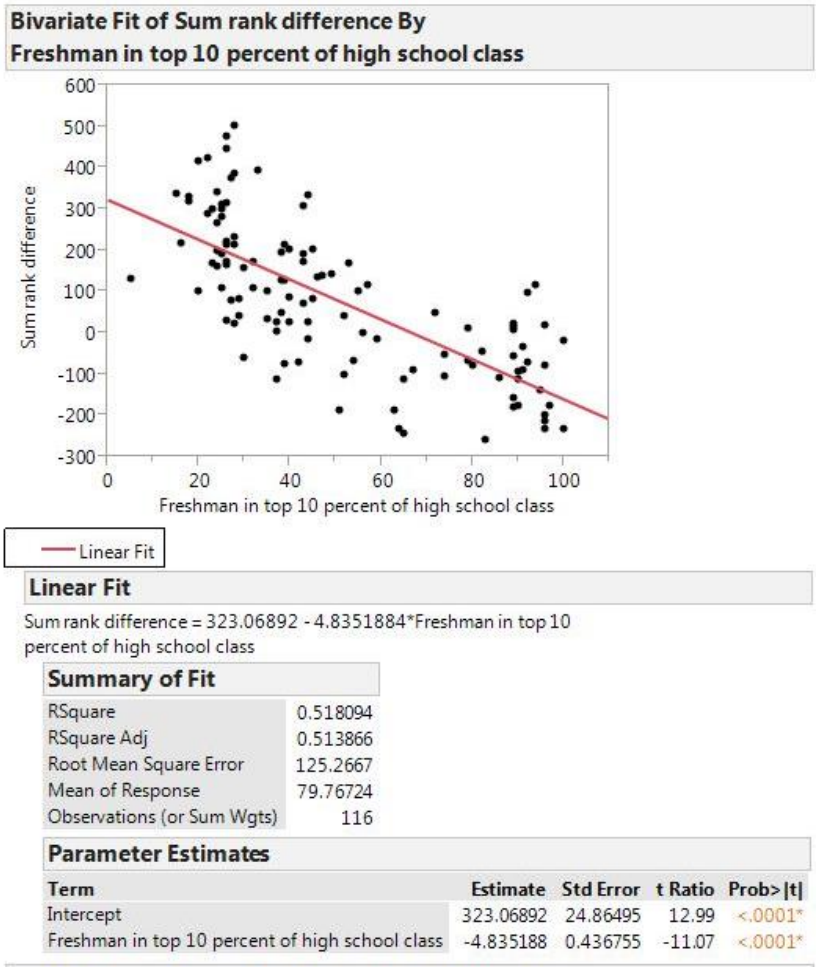


Figure 45: Bivariate fit of Rank difference by Freshman in top 10% (USNews)

From the above graph, we can interpret that percentage of freshman in top 10 percent of high school class inversely correlate with the rank difference (i.e.) data points with negative rank differences (low marketing) have higher percentage of freshman in top 10 percent of high school class and data points with positive rank differences (high marketing) have lower percentage of freshman in top 10 percent of high school class. The coefficient of the linear fit for percentage of freshman in top 10 percent of high school class is approximately -5 for predicting the rank difference.

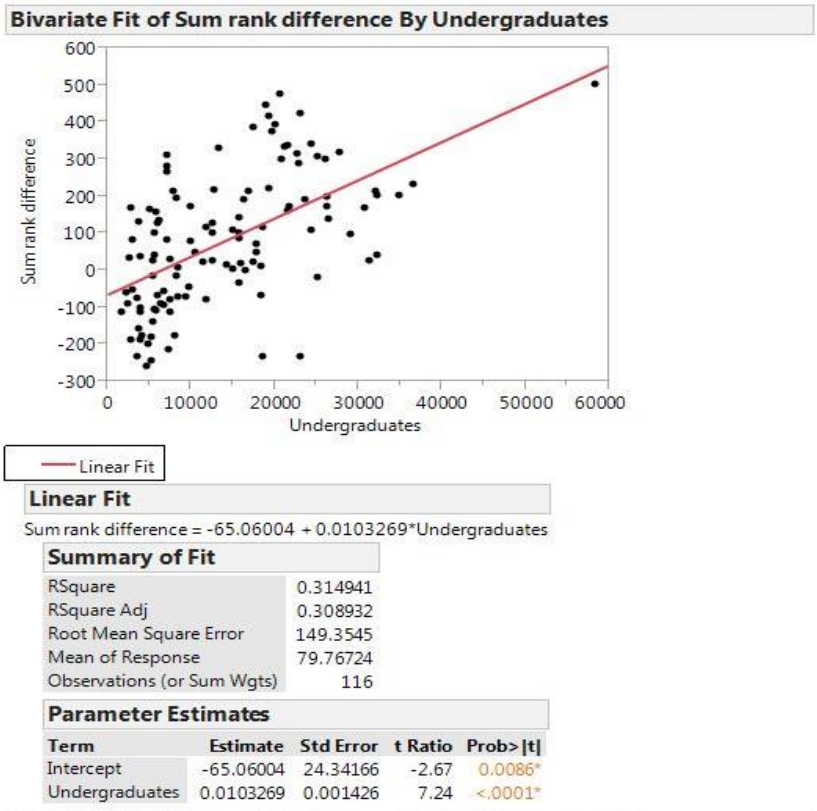


Figure 46: Bivariate fit of Rank difference by Undergraduates (USNews)

From the above graph, we can interpret that total undergraduates directly correlate with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher total undergraduates (more population) and data points with negative rank differences (low marketing) have lower total undergraduates (less population). This could mean both, low marketing universities are more selective in selecting their students and also that more number of students select schools with high marketing. The coefficient of the linear fit for total undergraduates is approximately +0.01 for predicting the rank difference.

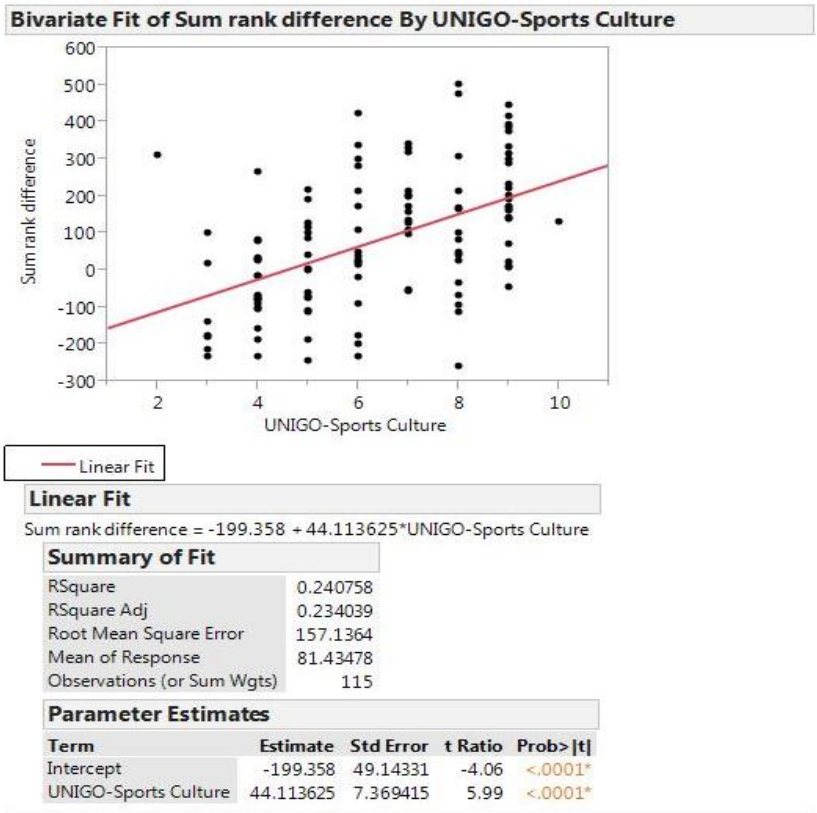


Figure 47: Bivariate fit of Rank difference by Unigo Sports score (USNews)

From the above graph, we can interpret that UNIGO sports culture students score directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher UNIGO sports culture students score and data points with negative rank differences (low marketing) have lower UNIGO sports culture students score. This suggests that universities that are active in sports tend to be more popular in social networks. The coefficient of the linear fit for UNIGO sports culture students score is approximately +44 for predicting the rank difference.

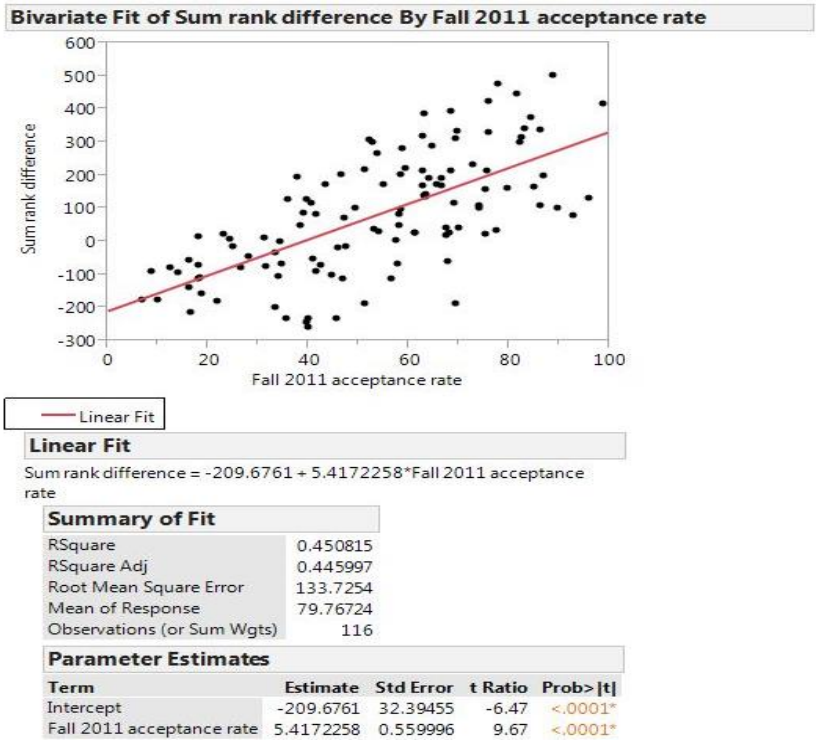


Figure 48: Bivariate fit of Rank difference by Acceptance rate (USNews)

From the above graph, we can interpret that Fall 2011 acceptance rate directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher Fall 2011 acceptance rate and data points with negative rank differences (low marketing) have lower Fall 2011 acceptance rate. This suggests that universities that are more popular in social networks have lesser selectivity criteria compared to those that are not so popular. The coefficient of the linear fit for Fall 2011 acceptance rate is approximately +5 for predicting the rank difference.

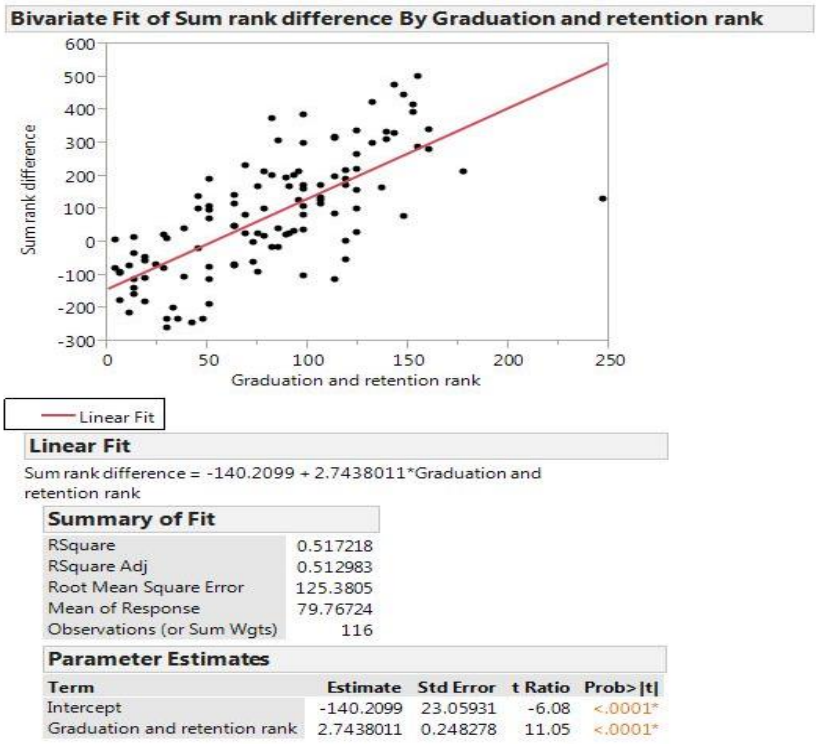


Figure 49: Bivariate fit of Rank difference by Retention rank (USNews)

From the above graph, we can interpret that graduation and retention rank directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher graduation and retention rank (ranked lower) and data points with negative rank differences (low marketing) have lower graduation and retention rate (ranked higher). The coefficient of the linear fit for graduation and retention rate is approximately +3 for predicting the rank difference.

```
Linear Regression Model

Sum rank difference =

     -5.3566 * Average freshman retention rate +
    -13.1167 * 6-year graduation rate +
     -5.729  * Student-faculty ratio +
     18.5167 * Predicted graduation rate +
     -0.8634 * Freshman in top 10 percent of high school class +
     31.943  * Setting=Rural,College Town +
      0.0054 * Undergraduates +
     21.2033 * UNIGO-Sports Culture +
     20.0971 * UNIGO-Arts Culture +
      1.3305 * Graduation and retention rank +
      1.0313 * Faculty resources rank +
      1.3131 * Student selectivity rank +
      0.5391 * Financial resources rank +
      0.2929 * Alumni giving rank +
      9.3758 * Overperformance/Underperformance +
      0.4748 * SAT 25th percentile +
     17.4827 * ACT 25th percentile +
  -1389.3347
```

Figure 50: Linear Regression Model for Rank difference (USNews)

The above equation is the multi-variable regression equation for the rank difference predictor variable. Just as a reminder, positive rank difference means it is ranked lower in USNews and ranked higher in social networks and negative rank difference means it is ranked higher in USNews and ranked lower in social networks. Therefore, according to the regression equation a negative rank difference is influenced by higher average freshman retention rate, 6 year graduation rate, etc. And positive rank difference is influenced by higher score in UNIGO sports and arts culture, College town or rural university setting, etc. These are some interesting observations we can make from linear regression model results.

7.7.5 Decision Tree Rules

Decision tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In this chapter we have used

decision tree model to create rules to predict 2 classifications. One is to create rules to predict correlate and inverse correlate classes and the other one is to create rules to predict high and low marketing classes.

The algorithm finds the most significant split in each recursive step and it is determined by the largest likelihood-ratio Chi-square statistic [41]. The split is chosen to maximize the difference in the responses between the two branches of the split (maximum information gain)

7.7.5.1 Decision Tree for Classifying Positive and Negative Correlation

After filtering the selected features out of the dataset we apply decision tree modeling. In this chapter, we are going to be classifying correlate and inverse correlate class using the selected features. Below are the rules and other results obtained by the J48 decision tree algorithm using 10-fold cross-validation [40]

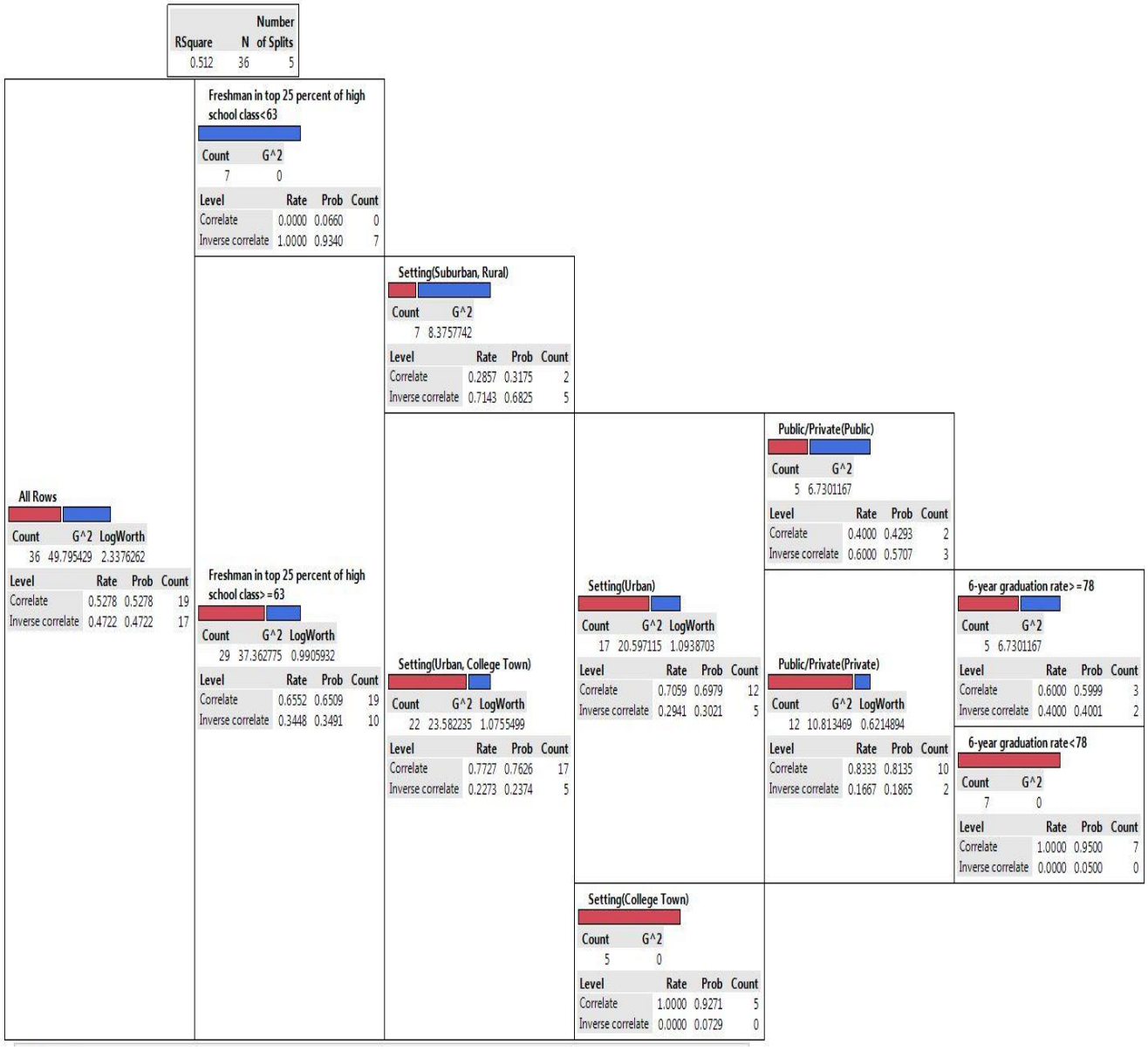


Figure 51: Decision Tree to Classify Correlate and Inverse Correlate (USNews)

Leaf Report

Response Prob

Leaf Label	Correlate	Inverse correlate
Freshman in top 25 percent of high school class<63	0.0660	0.9340
Freshman in top 25 percent of high school class>=63&Setting(Suburban, Rural)	0.3175	0.6825
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Public)	0.4293	0.5707
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Private)&6-year graduation rate>=78	0.5999	0.4001
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Private)&6-year graduation rate<78	0.9500	0.0500
Freshman in top 25 percent of high school class>=63^&Setting(College Town)	0.9271	0.0729

Response Counts

Leaf Label	Correlate	Inverse correlate
Freshman in top 25 percent of high school class<63	0	7
Freshman in top 25 percent of high school class>=63&Setting(Suburban, Rural)	2	5
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Public)	2	3
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Private)&6-year graduation rate>=78	3	2
Freshman in top 25 percent of high school class>=63^&Setting(Urban)&Public/Private(Private)&6-year graduation rate<78	7	0
Freshman in top 25 percent of high school class>=63^&Setting(College Town)	5	0

Figure 52: Decision Tree Leaf Report to Classify Correlate and Inverse Correlate (USNews)

The inverse correlate class's first decision tree rule is defined by less than 63% of freshman in top 25% of high school class. And for ones that are higher they have a suburban setting and are public universities.

The correlate class fall under the rule greater than 63% of freshman in top 25% of high school class. And their setting is college town. If the university is set in urban, it is a private university and have greater than 78% 6-year graduation rate.

The above rules can be generalized that universities that have correlation between their social network ranking and academic ranking have better students (top 25% in high school) and better 6 year graduation rate and are mostly set in college towns and are private universities. Universities that have inverse correlation between their social network ranking and academic ranking have lesser students in top 25% of high school, lesser than 78% 6 year graduation rate and are set in urban/suburban and are public universities.

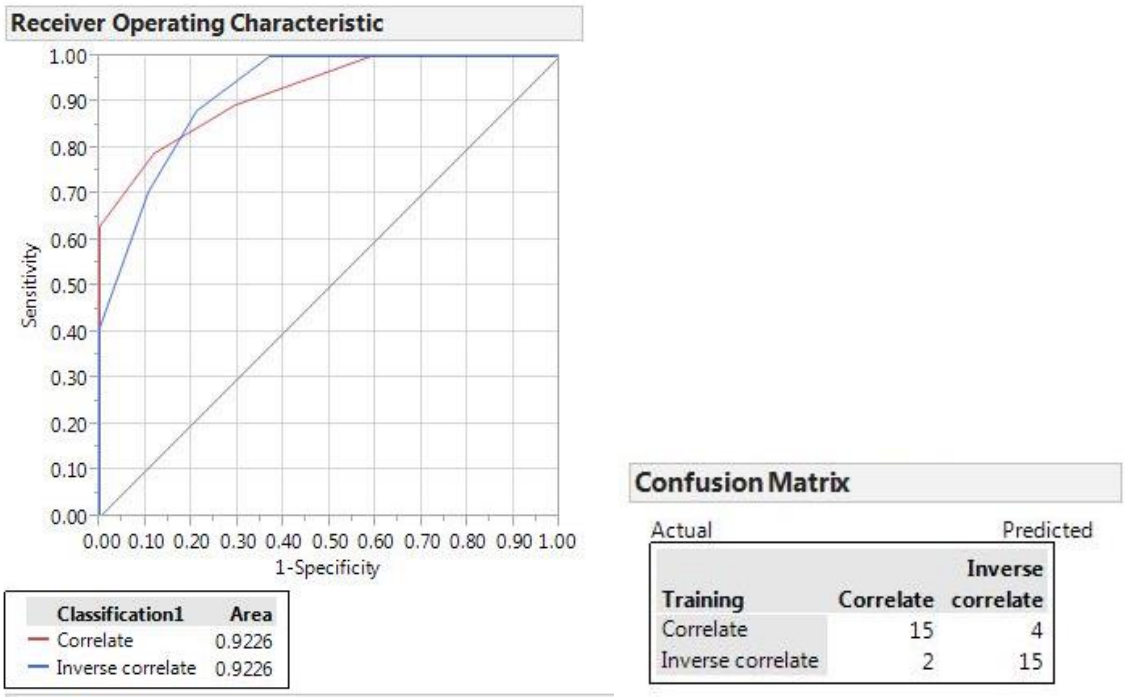


Figure 53: ROC curve and Confusion Matrix for Correlate and Inverse Correlate classifier (USNews)

The ROC curve of the decision tree says that it is a good model and the confusion matrix gives the true positive/negative and false positive/negative number of prediction from 10 fold cross-validation.

7.7.5.2 Decision Tree for Classifying High and Low Marketing

After filtering the selected features out of the dataset we apply decision tree modeling. In this chapter, we are going to be classifying high and low marketing class using the selected features. Below are the rules and other results obtained by the J48 decision tree algorithm.

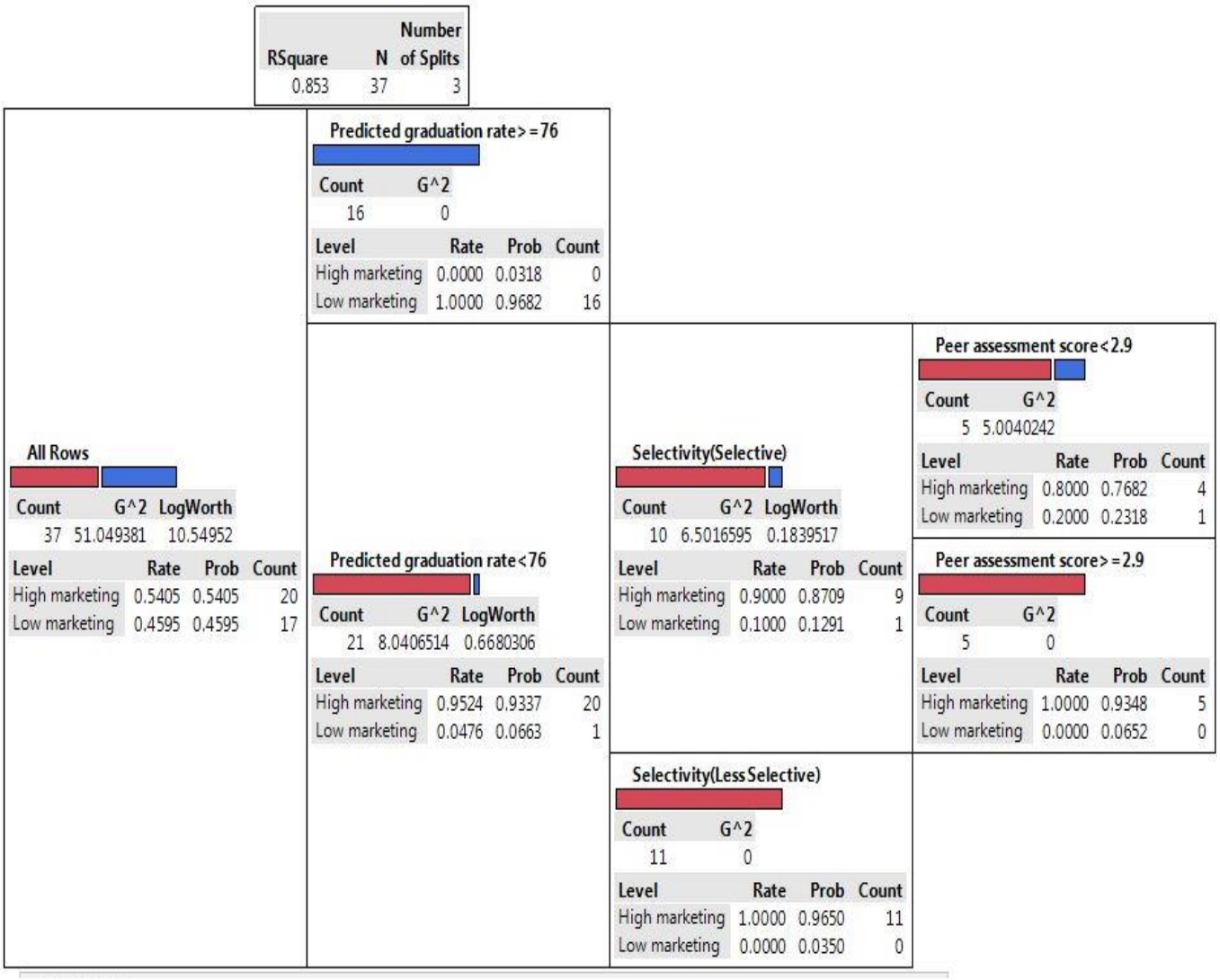


Figure 54: Decision Tree to Classify Low & High marketing (USNews)

Leaf Report

Response Prob

Leaf Label	High marketing	.2 .4 .6 .8	Low marketing	.2 .4 .6 .8
Predicted graduation rate>=76	0.0318		0.9682	
Predicted graduation rate<76&Selectivity(Selective)&Peer assessment score<2.9	0.7682		0.2318	
Predicted graduation rate<76&Selectivity(Selective)&Peer assessment score>=2.9	0.9348		0.0652	
Predicted graduation rate<76&Selectivity(Less Selective)	0.9650		0.0350	

Response Counts

Leaf Label	High marketing	Low marketing
Predicted graduation rate>=76	0	16
Predicted graduation rate<76&Selectivity(Selective)&Peer assessment score<2.9	4	1
Predicted graduation rate<76&Selectivity(Selective)&Peer assessment score>=2.9	5	0
Predicted graduation rate<76&Selectivity(Less Selective)	11	0

Figure 55: Decision Tree Leaf Report to Classify Low & High marketing (USNews)

With 0.96 probability the classification is Low marketing when predicted graduation is greater than or equal to 76%, (i.e.) when the university's predicted graduation rate is better than average the university doesn't exhibit social network marketing. And their selectivity criterion is most selective or more selective than the universities that have high social network marketing.

High marketing universities on the other hand have less than average predicted graduation rate and their selectivity criteria for applicants is less stringent than its counterpart. But their peer assessment score (varies from 2 to 5) is generally more than average (>3).

The above rules can be generalized that universities that are ranked higher in academia but lower in social networks have higher predicted graduation rate and are more selective in their application selection process than the universities that are ranked higher in social networks and lower in academia. But in general they have good peer assessment scores.

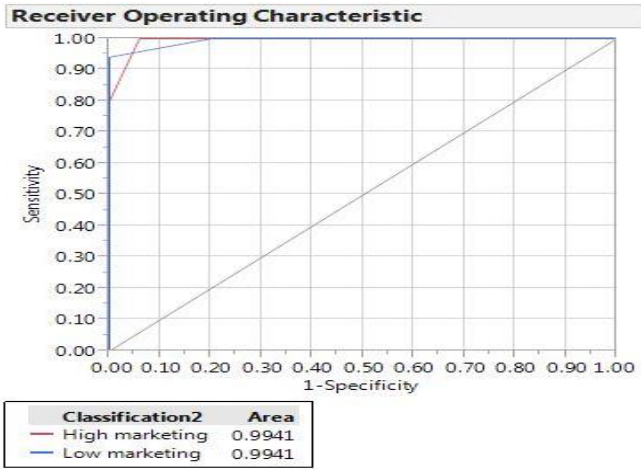


Figure 56: ROC curve & Confusion Matrix for Low & High marketing classifier (USNews)

We also did modelling using neural networks (single layer perceptron) which gives us weights for each attribute for the prediction. And after 25 recursions, with rank difference as the predictor variable the following are the weights given for each attribute.

Table 20: Neural Networks output for High/Low marketing classification (USNews)

Inputs	**Weights**
Threshold	-0.33
High school counselor score	-0.055
Average freshman retention rate	-0.22
6-year graduation rate	-0.14
Undergraduate academic reputation index	-0.06
Predicted graduation rate	-0.19
Percent of faculty who are full-time	-0.14
Freshman in top 25 percent of high school class	-0.01
Setting=Urban	-0.02
Setting=College Town	0.22
Setting=Rural	0.41
Undergraduates	0.24
Selectivity=Selective	-0.089
Selectivity=Less Selective	0.18
Tuition and fees	-0.04
Classes with 50 or more students	0.14
Fall 2011 acceptance rate	0.04
Financial resources rank	-0.008
Alumni giving rank	-0.007

The results are similar to all the previous analyses that we have performed. Negative rank differences (low marketing) class is driven by better graduation, freshman retention rate, with urban setting and with more selective application process. Positive rank differences (high marketing) class is driven by more undergraduates, less selective application process, more acceptance rate and college town setting.

7.8 Classification Based on Research Rank

We have defined research rank as number of citations per publication for the university. We have total number of publications and total citation count for each university from Microsoft Academic Search and we have taken a ratio of citation count to total publications and ranked the attribute for the 150 universities to get Research rank for each university. In this chapter we are going to be studying the patterns of rank differences of universities between their social network phenomenon and research. We are going to be finding patterns in the data collected for classifying rank differences between social network and research.

- We have selected 4 features to represent social network ranking. They are Twitter followers, Facebook page likes, Facebook post likes and YouTube video views.
- To find rank difference, we subtract all the 4 social network feature ranks from research rank for each university.
- We calculate standard deviation between all the 4 rank differences which tell us if a university is consistently different from Research rank or not.
- We define rules for each class based on our goal defined below in each topic after carefully studying the dataset properties.

Table 21: Rank difference calculation between Research and Social Networks

UNIVERSITY	Rank difference/Std Dev	Difference	R-Followers	R-Likes	R-Postlikes	R-Views
University of Vermont	-33.14	-267	-56	-66	-75	-70
St. Louis University	-22.99	-211	-63	-58	-45	-45
Wake Forest University	-22.17	-301	-73	-68	-65	-95
Duquesne University	-21.80	-132	-36	-37	-24	-35
South Carolina State University	-17.32	-30	-8	-9	-5	-8
Brandeis University	-17.07	-274	-46	-76	-83	-69
St. Johns Fisher College	-17.06	-80	-25	-22	-14	-19
University of Cincinnati	-16.84	-117	-29	-38	-29	-21
Washington University in St. Lou	-14.88	-263	-70	-73	-80	-40

7.8.1 Positive and Negative Correlation Graphs

In this section, we would like to find patterns in the dataset for classifying universities that correlate well in their social network presence and their Research ranking with the universities that have inverse correlate in their social network presence and Research ranking (we do not consider direction in this section). The rules for classification are defined below.

Rules for correlate class:

The ratio (Sum of rank differences/Standard deviation) must be between -4 to +4

Rank difference must be between -30 and +30

After filtering, we had a total of 18 universities that fell under the correlate class.

Table 22: Universities that Correlate with Research rank

UNIVERSITY	Rank difference/Std Deviation	Difference total
University of Iowa	-2.13	-23
University of Southern California	-2.54	-20
University of San Francisco	-1.17	-16
Clarkson University	-1.87	-11
Binghamton University – SUNY	0	0
University of St. Thomas	0	0
Temple University	0.04	1
Colorado School of Mines	1.35	4
University of Colorado-Boulder	0.16	5
University of the Pacific	0.71	7
Fordham University	0.78	10

University of Virginia	1.44	14
Stevens Institute of Technology	1.71	18
Miami University – Oxford	0.93	23
Washington State University, Pullman	1.17	23
University of Minnesota – Twin Cities	1.85	25
Missouri University of Science & Technology	3.16	26
Cornell University	3.53	27

Rules for inverse correlate class:

- The ratio (Sum of rank differences/Standard deviation) must be less than -10 or greater than 10
- Rank difference must be less than -200 or greater than +200

After filtering, we had a total of 18 universities that fell in inverse correlate class.

Table 23: Universities that Inverse Correlate with Research rank

UNIVERSITY	Rank difference/Std Deviation	Difference total
Yeshiva University	-11.92	-367
University of California-San Diego	-12.82	-348
Wake Forest University	-22.17	-301
Brandeis University	-17.06	-274
University of Vermont	-33.13	-267
Washington University in St. Louis	-14.88	-263
University of Pittsburgh	-11.19	-224

St. Louis University	-22.98	-211
University of Illinois, Chicago	-13.40	-211
Purdue University – West Lafayette	32.34	202
Clemson University	20.59	208
Ohio University	10.90	216
University of Kansas – Lawrence	22.72	228
University of Notre Dame	23.22	260
Oklahoma State University	15.15	264
University of Illinois-Urbana Champaign	29.07	295
University of North Carolina-Chapel Hill	23.19	331
University of Tennessee	35.71	374

Using simple visualization techniques, the data was analyzed first and below are some interesting histograms that are worth explaining.

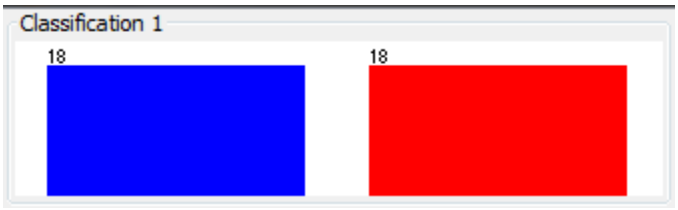


Figure 57: Research Correlate & Inverse Correlate histogram

Inverse correlate class is represented in blue color in the classification histogram above and correlate class is represented in red color. Both classes have equal number of data points (18)

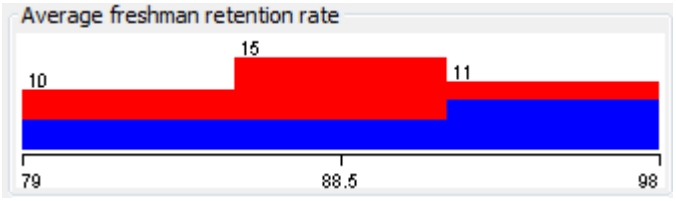


Figure 58: Retention rate distribution for Research Correlate & Inverse Correlate

Unlike USNews-Social network we did not see any striking difference between these two classes however we observed some slightly interesting patterns which are described here. Average freshman retention rate graph is shown above and we can note that inverse correlate class is slightly more prevalent on the upper end of the spectrum between around 90 to 98% and correlate class is slightly more pronounced on the middle scale of 80-100% limit. Universities that have research ranking and social network ranking correlate with each other have marginally lesser retention rate than the universities which do not have their rankings correlate.

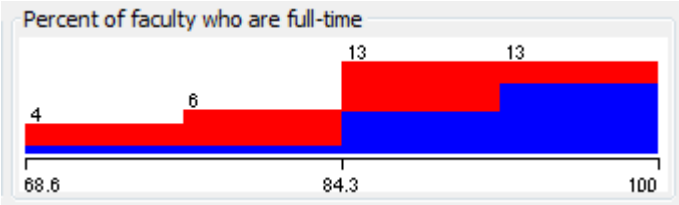


Figure 59: Full time faculty distribution for Research Correlate & Inverse Correlate

From the percent of full-time faculty graph above, we can notice that universities that have inverse correlation have almost 90 to 100% of their faculty that are full-time whereas universities that correlate are widespread across the limit (70 to 100%). This suggests that universities that have their research and social network ranking correlate have marginally lesser FTE faculty than the universities that do not have their ranking correlate.

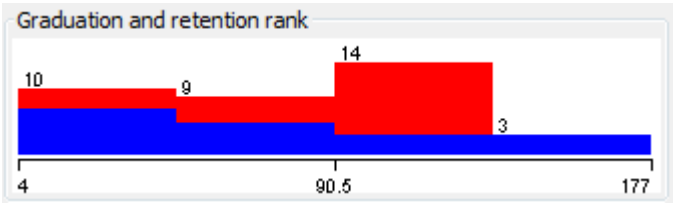


Figure 60: Retention rank distribution for Research Correlate & Inverse Correlate

Inverse correlate class has better graduation and retention rank than correlate class. This pattern is interesting because one would think that universities that are equally ranked in

research and social network field will have better graduation and retention rank. Universities that do not have their rankings correlate have better graduation and retention rank.

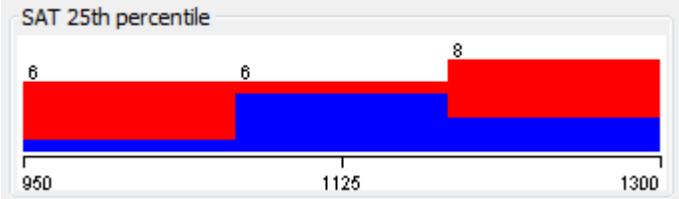


Figure 61: SAT 25th percentile distribution for Research Correlate & Inverse Correlate

Correlate universities have lower SAT 25th percentile than universities that are inversely correlated. All the histograms above suggest that universities that are inversely correlated have better academic standards than universities that correlate and this is an interesting pattern.

7.8.2 Low Marketing and High Marketing Graphs

In this section, we would like to find patterns in the dataset for classifying universities that are ranked high in Research and low in social networks (Low marketing) and that are ranked high in social networks and low in Research (High marketing) assuming that the presence of social network activity means marketing. The rules for classification are defined below.

- The ratio (Sum of rank differences/Standard deviation) must be less than -11
- Rank difference must be less than -100

After filtering, we had a total of 16 universities that fell under the low marketing class.

Table 24: SAT 25th percentile distribution for Research Correlate & Inverse Correlate

UNIVERSITY	Rank difference/Std Deviation	Difference total
Yeshiva University	-11.92	-367
University of California-San Diego	-12.82	-348

UNIVERSITY	Rank difference/Std Deviation	Difference total
Wake Forest University	-22.17	-301
Brandeis University	-17.06	-274
University of Vermont	-33.13	-267
Washington University in St. Louis	-14.88	-263
University of Pittsburgh	-11.19	-224
St. Louis University	-22.98	-211
University of Illinois, Chicago	-13.40	-211
University of California – Santa Cruz	-11.79	-149
Stony Brook University – SUNY	-10.05	-146
Clark University	-10.82	-139
Duquesne University	-21.79	-132
Drexel university	-10.05	-128
University of Cincinnati	-16.84	-117
University of Massachusetts – Amherst	-12.88	-114

Rules for high marketing class:

The ratio (Sum of rank differences/Standard deviation) must be greater than or equal to 10.

Rank difference must be greater than or equal to 100

After filtering, we had a total of 15 universities that fell in high marketing class.

Table 25: Universities in High Marketing w.r.t Research rank

UNIVERSITY	Rank difference/Std Deviation	Difference total
Michigan State University	13.04	147

UNIVERSITY	Rank difference/Std Deviation	Difference total
Arizona State University	20.44	173
College of William and Mary	18.33	183
University of Dayton	13.60	185
Kansas State University	12.86	186
Virginia Tech	33.33	197
Purdue University – West Lafayette	32.34	202
Clemson University	20.59	208
Ohio University	10.90	216
University of Kansas – Lawrence	22.72	228
University of Notre Dame	23.22	260
Oklahoma State University	15.15	264
University of Illinois-Urbana Champaign	29.07	295
University of North Carolina-Chapel Hill	23.19	331
University of Tennessee	35.71	374

Using simple visualization techniques, the data was analyzed first and below are some interesting histograms that are worth explaining.

Figure 62: Low & High Marketing w.r.t Research histogram

The above histogram shows us the total class population, blue represents low marketing class with 16 data points and red represents high marketing class with 15 data points.

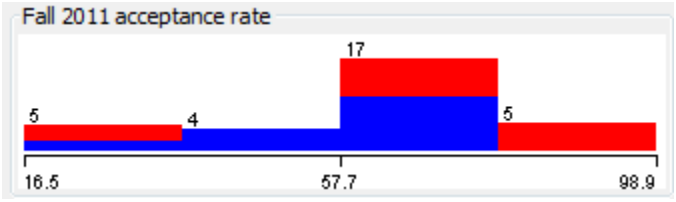


Figure 63: Acceptance rate distribution for Low & High Marketing w.r.t Research

The difference is not striking here but we can observe some pattern here. Acceptance rate in high marketing universities are marginally higher than low marketing class. Low marketing universities do not have any data point in the highest end of the acceptance rate category and are spread across the lower end.

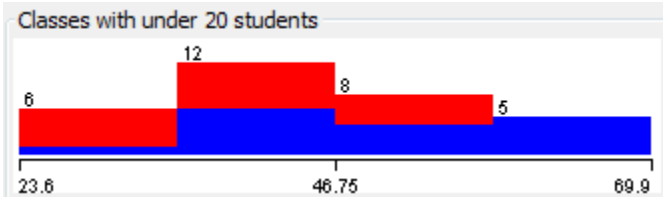


Figure 64: Under 20 classes distribution for Low & High Marketing w.r.t Research

Percentage of classes with fewer than 20 students tells how many classes are there in the university that has fewer than 20 students and like we would expect low marketing class falls on the upper end of this spectrum and low marketing class on the lower and middle limits. Because the acceptance rate is lower we could guess that number of students could be lower for low marketing universities.

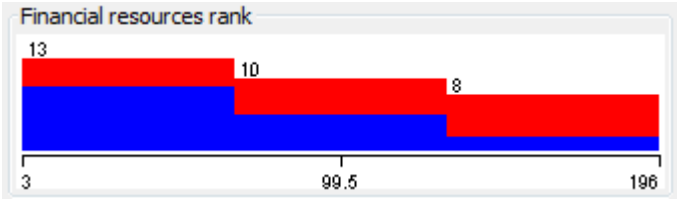


Figure 65: Financial Resources rank distribution for Low & High Marketing w.r.t Research

Universities that fall in low marketing class have higher rank in financial resources mostly than the universities that fall in high marketing class. If we try to make sense of this observation, we could say that since the low marketing class universities have higher financial resources rank they do not have to invest in social network marketing as much as the high marketing class universities which do not have financial resources like the former class.

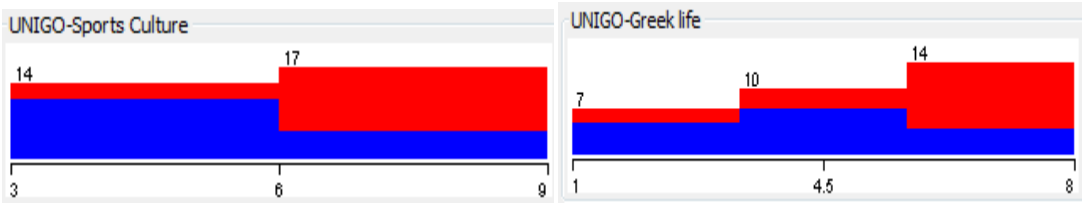


Figure 66: Unigo student ratings distribution for Low & High Marketing w.r.t Research

UNIGO student ratings also give us some interesting patterns in these histograms. Students have ranked sports and Greek life culture of universities on a scale of 1 to 10, 10 being most active. And high marketing universities seem to get a better score on both the ratings than low marketing ones.

Figure 67: Setting and Private/Public histograms for Low & High Marketing w.r.t Research

The number of low marketing universities that fall under private class is much higher than high marketing universities and the number of high marketing universities that fall under public class is much higher than low marketing universities. This suggests that private universities market lesser compared to public universities. And most college town and rural universities seem to be high marketing universities.

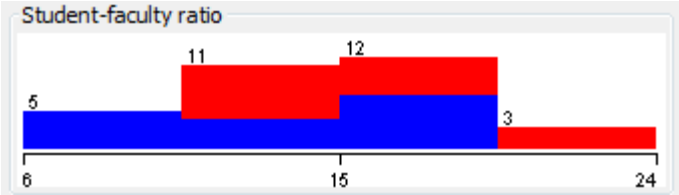


Figure 68: Student Faculty ratio distribution for Low & High Marketing w.r.t Research

Student faculty ratio is the number of students per FTE faculty and low marketing universities have lower student faculty ratio than high marketing universities. This could be because of the population of undergraduate students, as they are higher in high marketing universities.

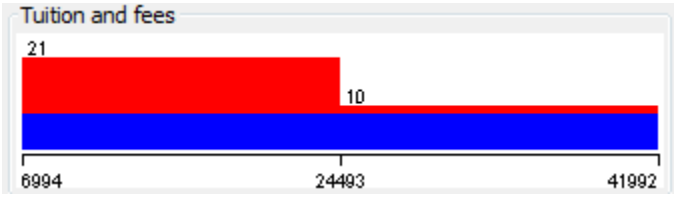


Figure 69: Tuition distribution for Low & High Marketing w.r.t Research

This is another interesting histogram. Low marketing class is equally spread across the spectrum of tuition and fees whereas high marketing universities fall mostly under the lower end of the spectrum which suggests that universities that are popular in social networks have comparatively lower tuition.

7.8.3 Linear Regression

In this chapter we are going to perform regression analysis on our dataset to find interesting equations. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. On the feature selected dataset we perform linear regression with a statistical analysis tool called JMP and below are the results. The predictor variable in this case is the rank difference and the dependent variables are:

- 6 year graduation rate

- Student faculty ratio
- Total undergraduates
- UNIGO sports score
- Financial resources rank

It is important we understand the meaning of the explanatory variable (rank difference here). Rank difference varies from -400 to +400. This is the sum of rank differences between Research rank and the four social network ranks. Therefore a negative rank difference means it is ranked higher in Research ranking and lower in social network and a positive rank difference means it is ranked lower in Research ranking and higher in social network. Therefore universities that end up in the lower end of the distribution in the limits (-400,+400) are those that have lower marketing and those universities that end up in the higher end of the distribution in the limits (-400,+400) are those that have higher marketing. Below are the graphs and linear regression equations for the above mentioned dependent variables.

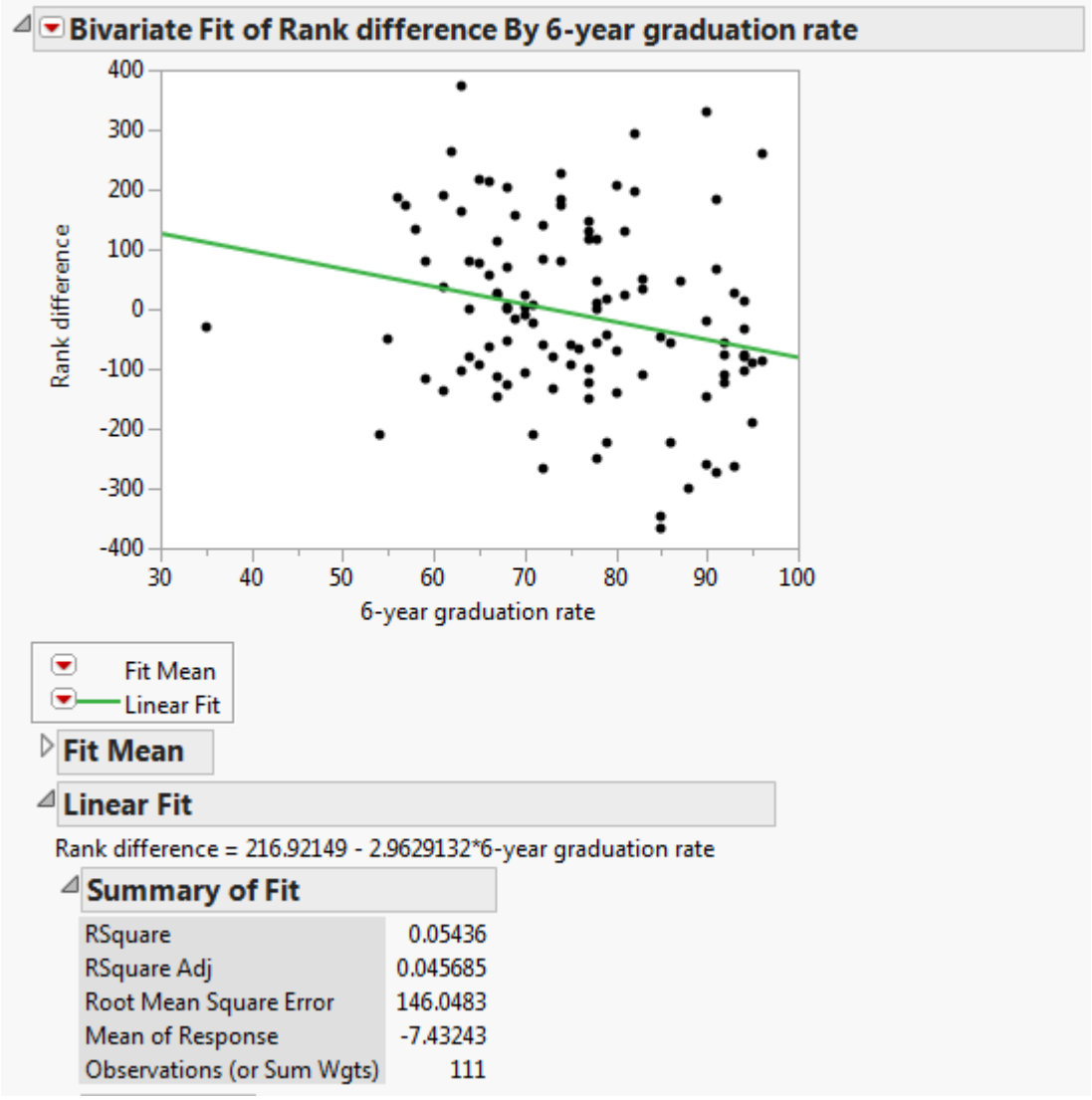


Figure 70: Bivariate fit of Rank difference by 6-year graduation rate (Research)

From the above graph, we can interpret that 6 year graduation rate inversely correlates with the rank difference (i.e.) data points with negative rank differences (low marketing) have higher 6 year graduation rate and data points with positive rank differences (high marketing) have lower 6 year graduation rate. The coefficient of the linear fit for 6 year graduation rate is approximately -3 for predicting the rank difference.

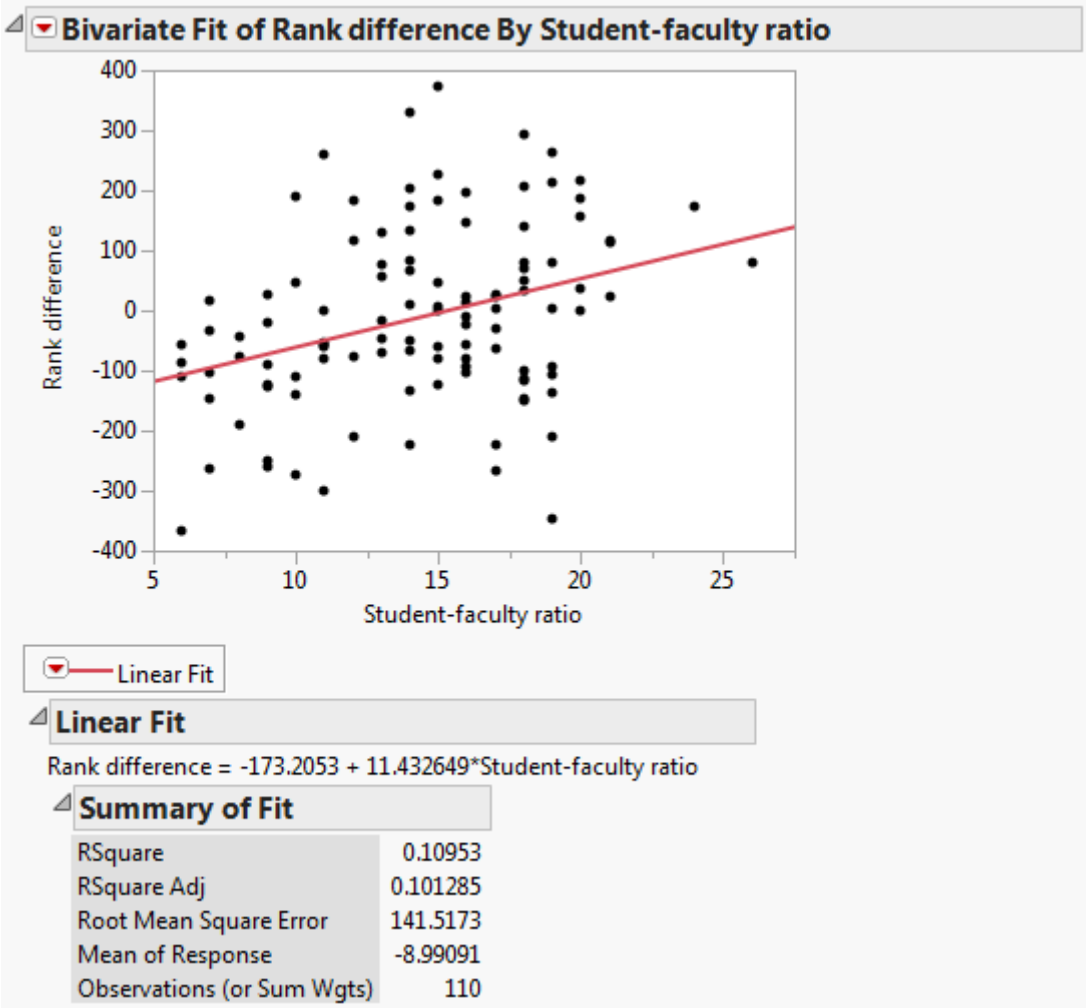


Figure 71: Bivariate fit of Rank difference by Student-faculty ratio (Research)

From the above graph, we can interpret that student faculty ratio directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher student faculty ratio (more students per faculty) and data points with negative rank differences (low marketing) have lower student faculty ratio (less students per faculty). The coefficient of the linear fit for student faculty ratio is approximately +11 for predicting the rank difference.

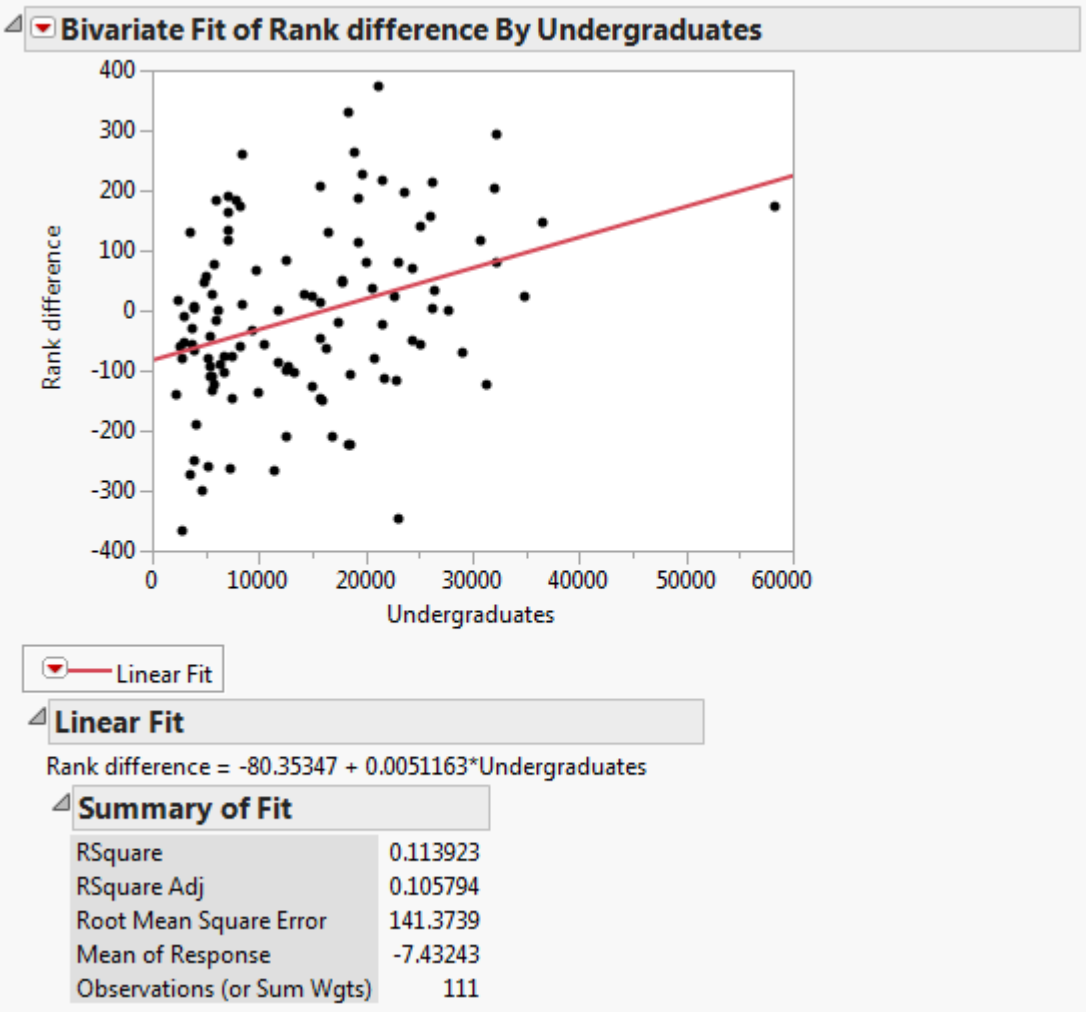


Figure 72: Bivariate fit of Rank difference by Undergraduates (Research)

From the above graph, we can interpret that total undergraduates directly correlate with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher total undergraduates (more population) and data points with negative rank differences (low marketing) have lower total undergraduates (less population). This could mean both, low marketing universities are more selective in selecting their students and also that more number of students select schools with high marketing. The coefficient of the linear fit for total undergraduates is approximately +0.005 for predicting the rank difference.

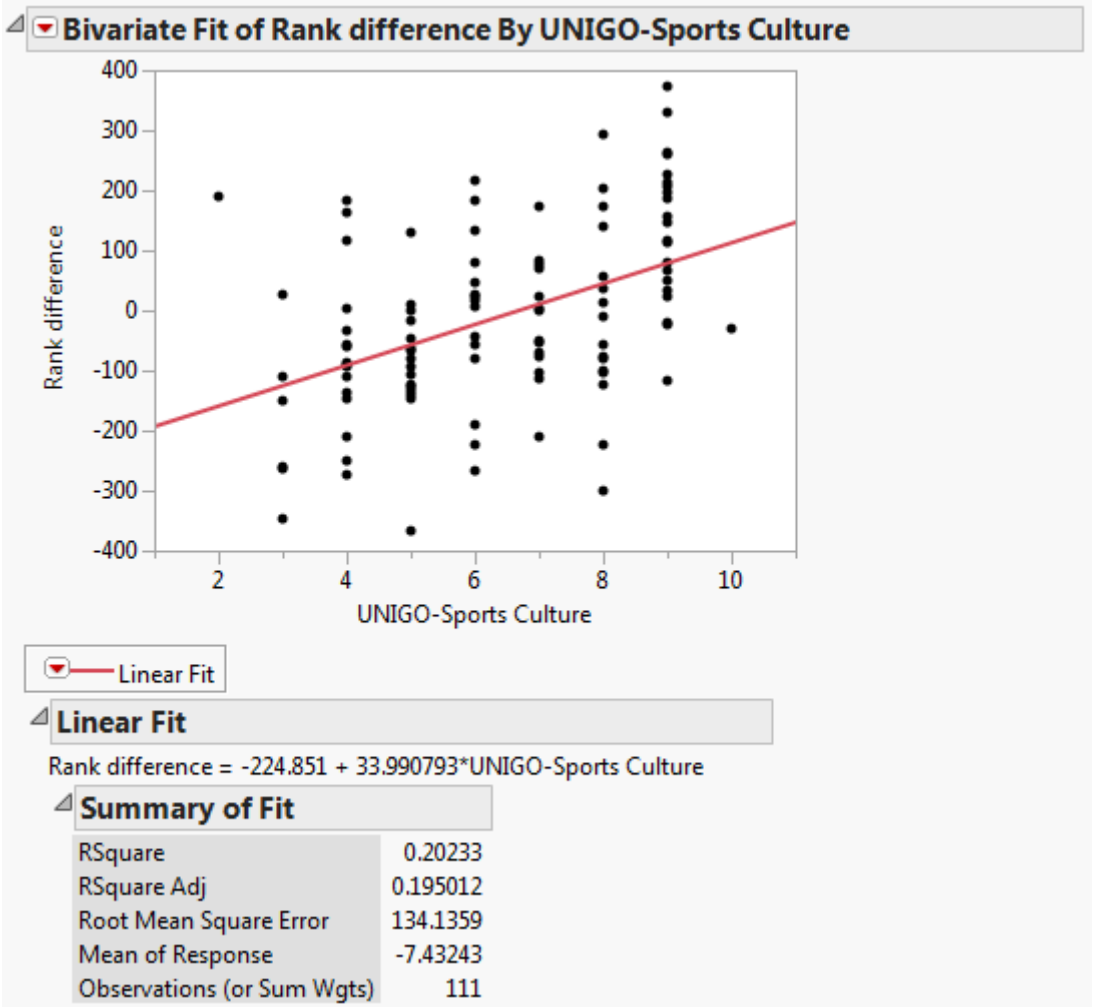


Figure 73: Bivariate fit of Rank difference by Unigo sports score (Research)

From the above graph, we can interpret that UNIGO sports culture students score directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher UNIGO sports culture students score and data points with negative rank differences (low marketing) have lower UNIGO sports culture students score. This suggests that universities that are active in sports tend to be more popular in social networks. The coefficient of the linear fit for UNIGO sports culture students score is approximately +34 for predicting the rank difference.

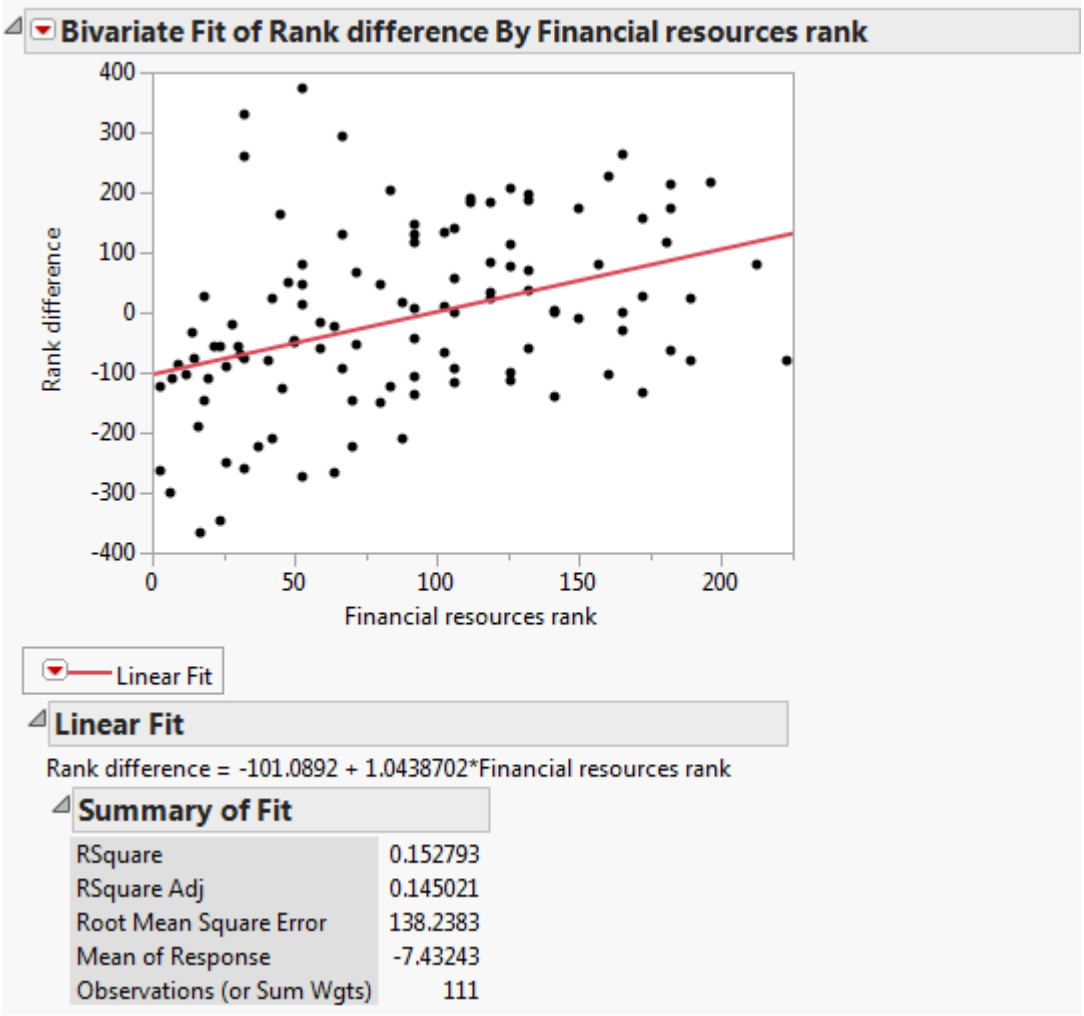


Figure 74: Bivariate fit of Rank difference Financial Resources rank (Research)

From the above graph, we can interpret that financial resources rank directly correlates with the rank difference (i.e.) data points with positive rank differences (high marketing) have higher financial resources rank and data points with negative rank differences (low marketing) have lower financial resources rank. This suggests that universities that are ranked higher in financial resources tend to have lower social network presence than those with lower financial resources rank. The coefficient of the linear fit for financial resources rank is approximately +1 for predicting the rank difference.

```
Linear Regression Model

Difference total =

     -2.9266 * 6-year graduation rate +
     -3.4243 * Freshman in top 10 percent of high school class +
      1.7717 * Freshman in top 25 percent of high school class +
     73.093  * Setting=College Town +
     61.0964 * Selectivity=More Selective,Less Selective +
     26.2227 * UNIGO-Sports Culture +
     -1.84   * Student selectivity rank +
      0.9147 * Financial resources rank +
      1.1303 * Alumni giving rank +
      8.024  * Average alumni giving rate +
   -120.1748
```

Figure 75: Linear Regression Model for Rank difference (Research)

The above equation is the multi-variable regression equation for the rank difference predictor variable. Just as a reminder, positive rank difference means it is ranked lower in Research and ranked higher in social networks and negative rank difference means it is ranked higher in Research and ranked lower in social networks. Therefore, according to the regression equation a negative rank difference is influenced by higher average 6 year graduation rate, stringent selectivity, etc. And positive rank difference is influenced by higher score in UNIGO sports culture, college town university setting, easier selectivity, etc. These are some interesting observations we can make from linear regression model results.

7.8.4 Decision Tree Rules

In this chapter we have used decision tree model to create rules to predict 2 classifications. One is to create rules to predict correlate and inverse correlate classes and the other one is to create rules to predict high and low marketing classes. The algorithm finds the most significant split in each recursive step and it is determined by the largest likelihood-ratio Chi-square statistic. The split is chosen to maximize the difference in the responses between the two branches of the split (maximum information gain)

7.8.4.1 Decision Tree for Positive and Negative Correlation classes

After filtering the selected features out of the dataset we apply decision tree modeling. In this chapter, we are going to be classifying correlate and inverse correlate class using the selected features. Below are the rules and other results obtained by the J48 decision tree algorithm.

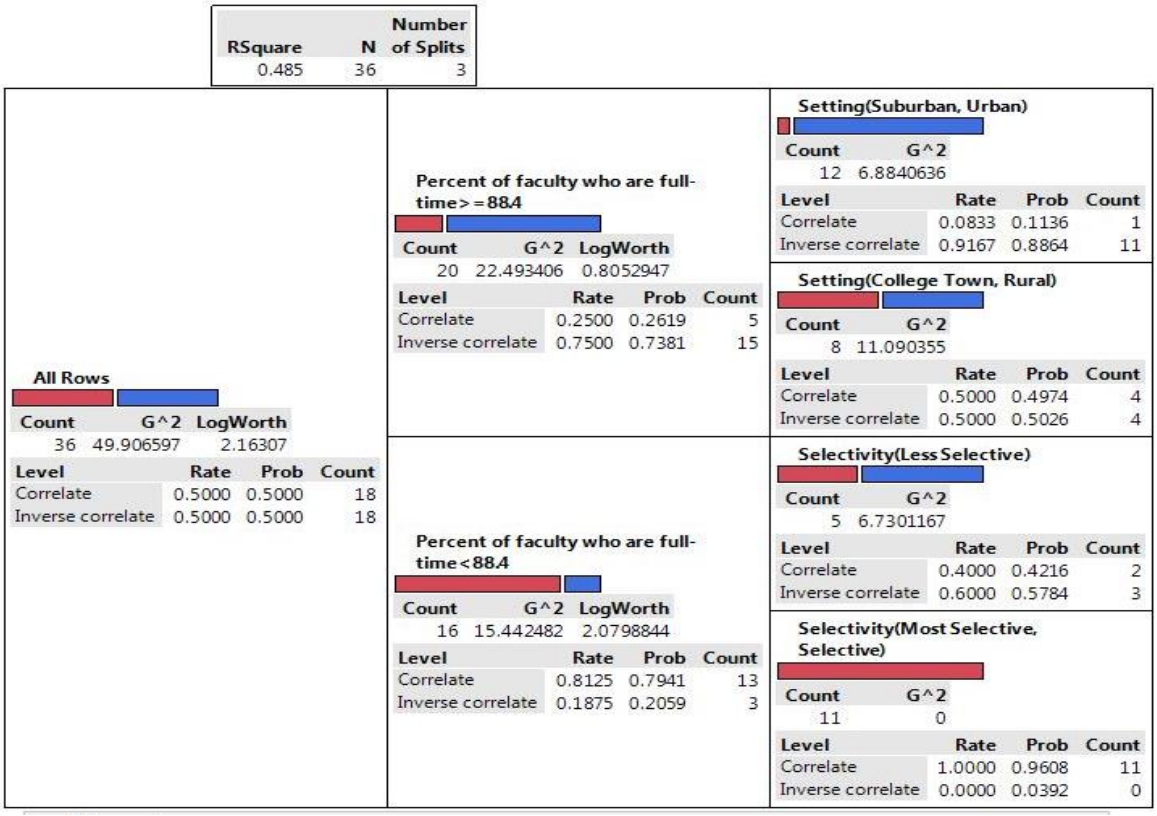


Figure 76: Decision Tree to Classify Correlate and Inverse Correlate (Research)

This model is a good predictor for inverse correlate class as it has 100% accuracy after doing a 10 folds cross validation. Let us review the rules generated by the model.

Leaf Report

Response Prob

Leaf Label	Correlate	Inverse correlate
Percent of faculty who are full-time>=88.4&Setting(Suburban, Urban)	0.1136	0.8864
Percent of faculty who are full-time>=88.4&Setting(College Town, Rural)	0.4974	0.5026
Percent of faculty who are full-time<88.4&Selectivity(Less Selective)	0.4216	0.5784
Percent of faculty who are full-time<88.4&Selectivity(Most Selective, Selective)	0.9608	0.0392

Response Counts

Leaf Label	Correlate	Inverse correlate
Percent of faculty who are full-time>=88.4&Setting(Suburban, Urban)	1	11
Percent of faculty who are full-time>=88.4&Setting(College Town, Rural)	4	4
Percent of faculty who are full-time<88.4&Selectivity(Less Selective)	2	3
Percent of faculty who are full-time<88.4&Selectivity(Most Selective, Selective)	11	0

Figure 77: Decision Tree Leaf Report to Classify Correlate and Inverse Correlate (Research)

From the above decision trees we can interpret the following. Universities that have more than 88.4% (very high percentage) of their faculty as full-time and that have suburban or urban setting tend to have inverse correlation (ranked opposite in research and social networks). This is opposite to the finding in USNews as universities that have high FTE faculty had correlation between USNews and social network ranking. However when it comes to selectivity it is similar to USNews as universities that correlate have higher selectivity criteria than universities that have inverse correlation.

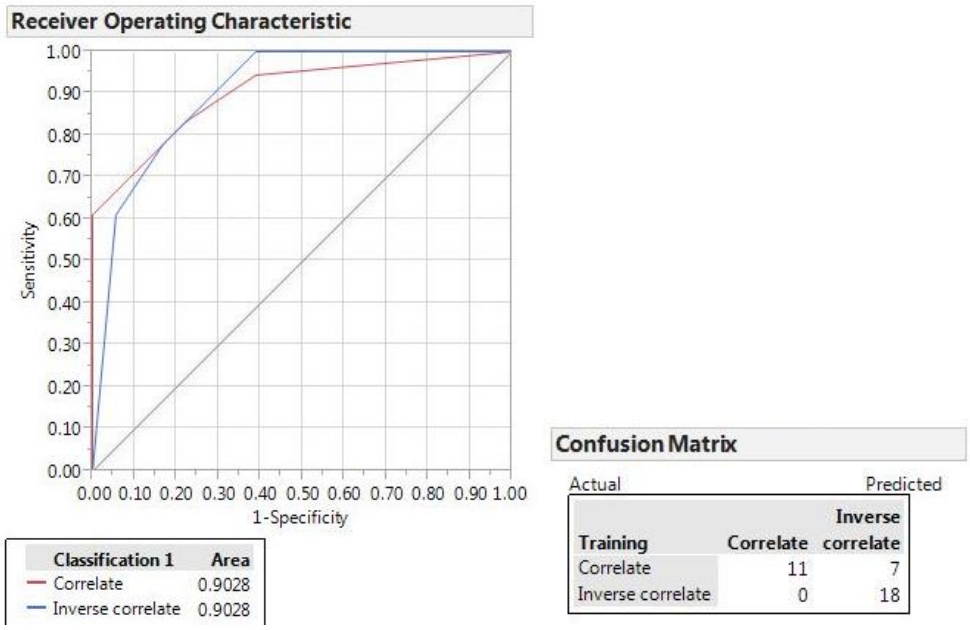


Figure 78: ROC curve and Confusion Matrix for Correlate and Inverse Correlate classifier (Research)

We can see from the ROC curve that the model is a fairly acceptable model and the confusion matrix shows us that after 10 fold cross validation the inverse correlate class has 100% accuracy.

7.8.4.2 Decision Tree for Classifying High and Low Marketing

After filtering the selected features out of the dataset we apply decision tree modeling. In this chapter, we are going to be classifying high and low marketing classes using the selected features. Below are the rules and other results obtained by the J48 decision tree algorithm.

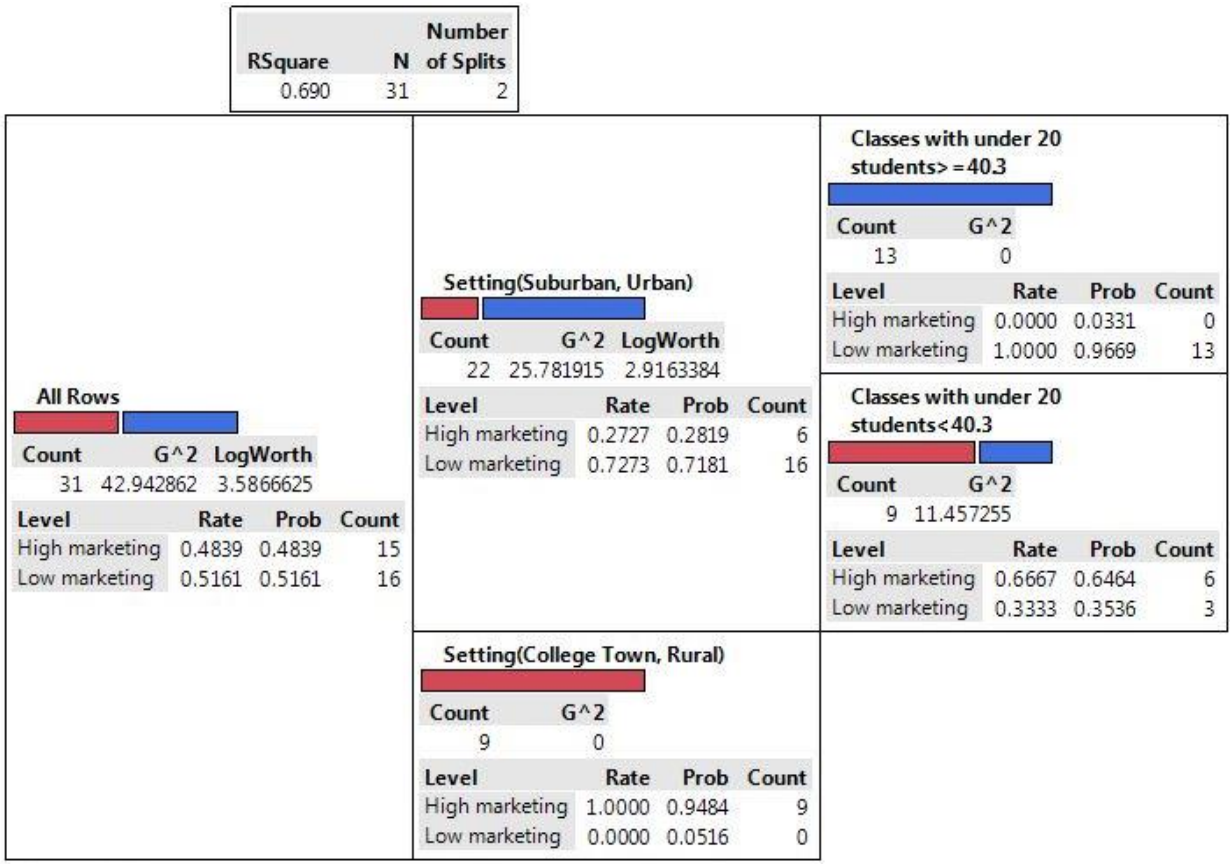


Figure 79: Decision Tree to Classify Low & High marketing (Research)

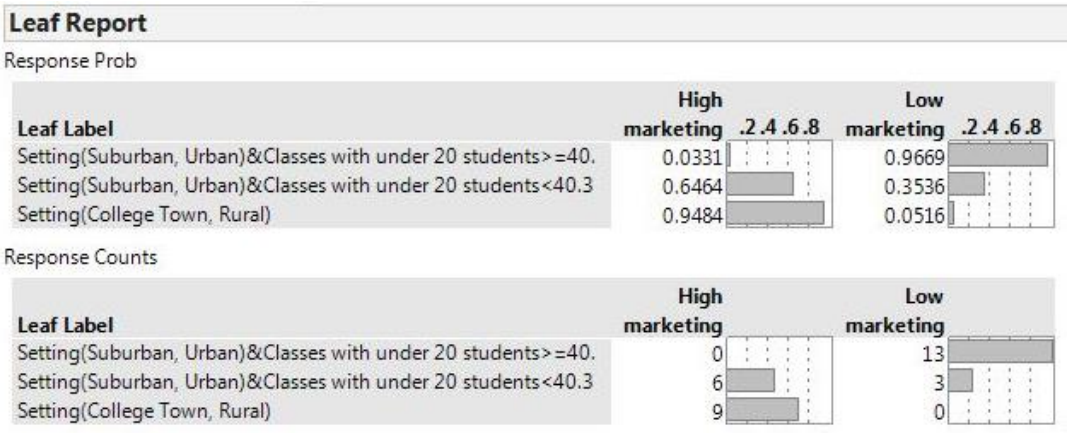


Figure 80: Decision Tree Leaf Report to Classify Low & High marketing (Research)

From the decision tree model above we have several interesting results. When the university has a college of rural setting it will mostly fall under high marketing category with 0.05 probability and when it is has a suburban or rural setting depending on its class size it is

classified as high or low marketing university. When most of the classes have size under 20 then it falls under low marketing category. Let us look at the distribution of the attribute classes with under 20 students to get a better picture here.

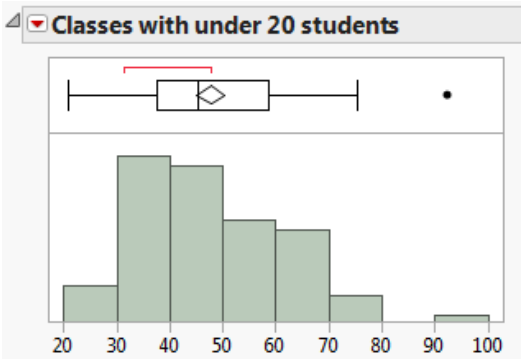


Figure 81: Histogram for Classes with under 20 students

Classes with under 20 students has mean 47 and the decision tree rule says if the university has urban setting and classes with under 20 students >= 40 (which is more than 50% of the distribution) it will fall under low marketing class.

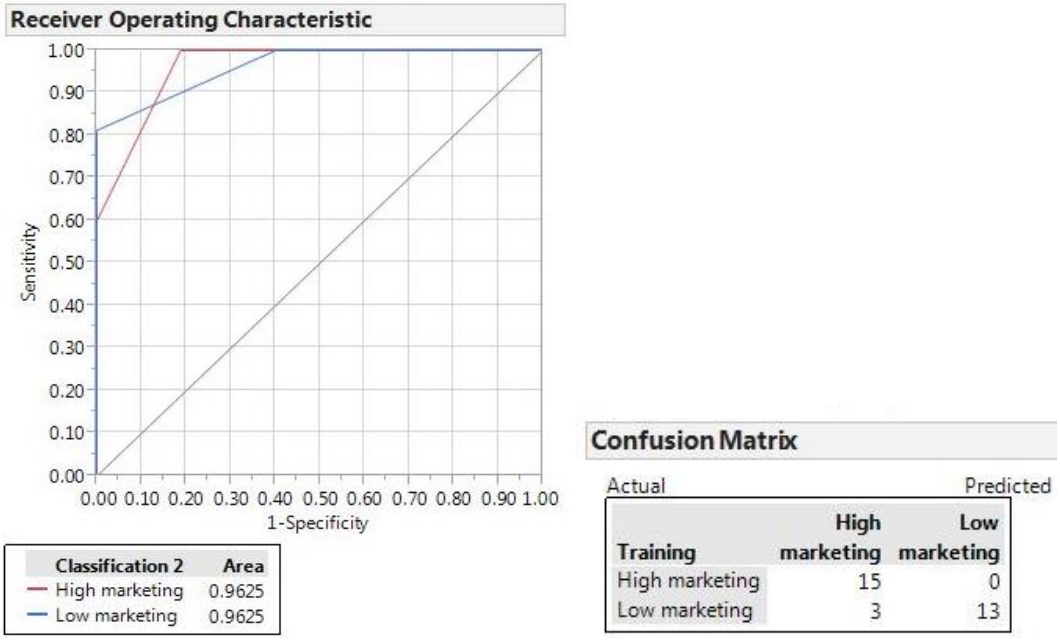
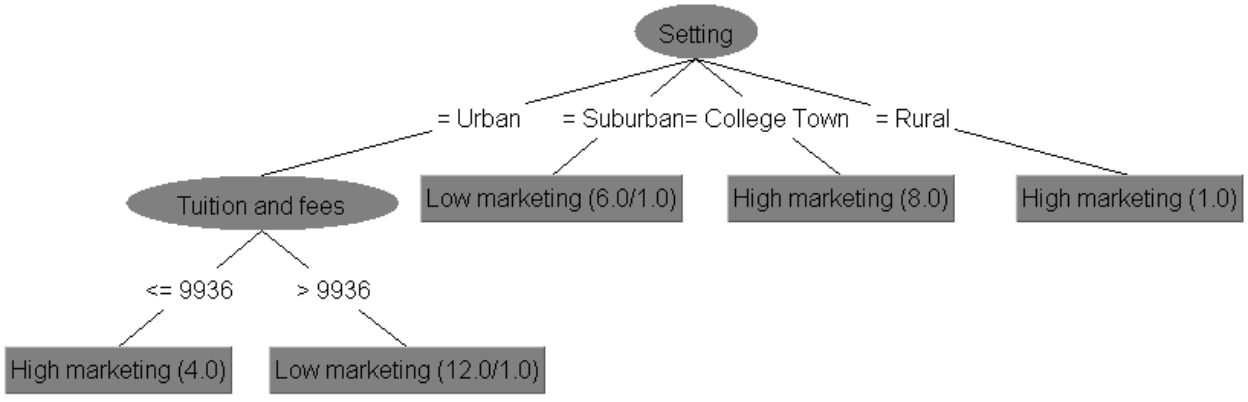


Figure 82: ROC curve & Confusion Matrix for Low & High marketing classifier (Research)

From the ROC curve and confusion matrix we can say that the above decision tree model has good predictive capability and it has given some interesting rules that we have discussed.

We also did a decision tree modeling with gain ratio as the deciding factor to find the optimum split and we have discussed the results below.




```
=== Confusion Matrix ===

  a  b   <-- classified as
 10  6 |  a = Low marketing
  6  9 |  b = High marketing
```

Figure 83: Decision tree for High/Low marketing with Gain ratio split

As the model before, if setting is college town or rural it is predicted as a university that has high marketing. When the setting is urban, if tuition is higher it is predicted as low marketing class. It is interesting to find tuition in the rules here as it suggests that university that has higher tuition tends to market lesser compared to university that has lower tuition which is similar to the results we had discussed in classification based on USNews rank difference from social networks.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Summary

In this thesis, we main addressed 2 questions. First one is, what are appropriate measures for ranking participants in social networks? We found that substantial research hasn't been done in this area, and factors that users usually use to rank users in social networks and how they correlate with each other. We found that most factors correlate with each other consistently except for user's participation which seems to contain different information. Second question is, does the ranking within networks mirror those based on traditional measures for ranking organizations? We selected universities' data and compared different ranking methods and compared it to USNews rank and research rank. We found some inconsistencies and deviations but the most popular social network features seem to correlate well. We have studied common ranking methodologies and we have tried to find a fair and optimum way to rank popular social networks and in this process we have tried to understand social network ranking by comparing it with accepted methods of ranking. Then we have performed some analysis using data mining techniques and have discussed the results.

Some interesting results from data mining that need causal analysis are: universities that have high sports ranking by students, low tuition, higher acceptance rate, lower SAT/ACT score cut-off and have a college-town setting have high social network rank. Universities that have better financial resources, that are private, that have lower acceptance rate and that have more classes with under 20 students have low social network rank and higher USNews and Research ranks.

All our results are based on statistical modeling of observational data and do not imply causality. We have assumed USNews college ranking as an accepted measure of ranking however there are some controversies in USNews ranking methodology which we have not considered.

8.2 Future Work

The work that is done here is a beginning that could pave way for better understanding of ranking criteria in different social networks. As a future work the following could be done to get more insights and to validate the results that we have in this project.

We could get temporal data by monitoring social networks over a period of time and get interesting insights of temporal data by comparing it with standard ranking methods and seeing how it changes over

We could use LinkedIn (professional social network) to get data about where graduates of universities are working currently and perform more analysis based on that.

Get dynamic data from social network and finding temporal opinion differences and to see how it is reflected in other standard ranking methods.

REFERENCES

[1] Langville, Amy N., and Carl Dean Meyer. Who's# 1?: the science of rating and ranking. Princeton University Press, 2012.

[2] Eysenbach, Gunther. "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact."Journal of medical Internet research 13.4 (2011).

[3] Cha, Meeyoung, et al. "Measuring User Influence in Twitter: The Million Follower Fallacy." ICWSM 10 (2010): 10-17.

[4] Das Sarma, Anish, et al. "Ranking mechanisms in twitter-like forums."Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.

[5] Leavitt, Alex, et al. "The influentials: New approaches for analyzing influence on twitter." Web Ecology Project 4.2 (2009): 1-18.

[6] Weng, Jianshu, et al. "Twitterrank: finding topic-sensitive influential twitterers."Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.

[7] Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[8] Uysal, Ibrahim, and W. Bruce Croft. "User oriented tweet ranking: a filtering approach to microblogs." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[9] Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity." ICWSM. 2012.

[10] Anger, Isabel, and Christian Kittl. "Measuring influence on Twitter." Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies. ACM, 2011.

[11] Subrahmanyam, Kaveri, et al. "Online and offline social networks: Use of social networking sites by emerging adults." Journal of Applied Developmental Psychology 29.6 (2008): 420-433.

[12] Geng, Xiubo, et al. "Feature selection for ranking." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[13] Social network is here to stay: Nabil Hassan El-Ghoroury (March 2010). Available from: http://www.apa.org/gradpsych/2010/03/matters.aspx

[14] Social media is here to stay: 10 things you should know. Heather Timmerman (February 2009). Available from: http://www.previewnetworks.com/blog/social-media-stay-10/

[15] How social can we get? What evolutionary psychology says about social networking? Available from: http://www.nbcnews.com/id/20642550/ns/technology_and_science-innovation/t/how-social-can-we-get/#.UWRliJOG18E

[16] Why are social networks so popular? Available from: http://blog.sfgate.com/techchron/2011/11/15/why-are-social-networks-so-popular/

[17] U.S. News college ranking. Available from: http://www.usnews.com/rankings

[18] Rojstaczer, Stuart. "College rankings are mostly about money." San Francisco Chronicle 3 (2001).

[19] Better games, more viewers raise ratings, NCAA to air basketball games. Available from: http://web1.ncaa.org/web_files/NCAANewsArchive/1981/19810215.pdf

[20] Mollett, Amy, Danielle Moran, and Patrick Dunleavy. "Using Twitter in university research, teaching and impact activities: A guide for academics and researchers." London School of Economics and Political Science: LSE Public Policy Group (2011).

[21] Ellison, Nicole B. "Social network sites: Definition, history, and scholarship."Journal of Computer-Mediated Communication 13.1 (2007): 210-230.

[22] Available from: http://en.wikipedia.org/wiki/U.S._News_%26_World_Report

[23] Monks, James, and Ronald G. Ehrenberg. "The impact of US News & World Report college rankings on admissions outcomes and pricing policies at selective private institutions." (1999).

[24] Roy, Senjuti Basu, et al. "The microsoft academic search dataset and kdd cup 2013." Proceedings of the 2013 KDD Cup 2013 Workshop. ACM, 2013.

[25] Hirsch, Jorge E. "An index to quantify an individual's scientific research output."Proceedings of the National academy of Sciences of the United States of America 102.46 (2005): 16569-16572. [26] Unigo.com Available from: http://www.unigo.com/Colleges/Default.aspx

[27] Glazowski, Paul. "Unigo Puts Users in Charge of College Reviews (The Startup Review)" - Mashable Sept. 19, 2008

[28] Moran, Caitlin. "Web Site Features College Reviews by Students, for Students" - The Chronicle of Higher Education Sept. 23, 2008

[29] Hockenberry, John and Adaora Udoji. "Unigo.com reviews colleges drawing from those who know them best: students" -The Takeaway Sept. 18, 2008

[30] Available from: http://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

[31] Available from: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[32] Available from: http://www.jmp.com/software/pdf/102565_pointsofinterest.pdf

[33] Available from: http://www.autonlab.org/tutorials/dtree18.pdf

[34] Available from: http://scikit-learn.org/stable/modules/tree.html

[35] Available from: http://www.computerworld.com/s/article/9244251/ One_out_of_seven_people_use_so cial_networks_study_shows

[36] Hume, David. A treatise of human nature. Courier Dover Publications, 2012.

[37] Qualman, Erik. Socialnomics: How social media transforms the way we live and do business. John Wiley & Sons, 2012.

[38] Dubois, Didier, and Henri Prade. "Ranking fuzzy numbers in the setting of possibility theory." Information sciences 30.3 (1983): 183-224.

[39] Available from: http://news.discovery.com/tech/apps/top-ten-social-networking-sites.htm

[40] Quinlan, John Ross. C4. 5: programs for machine learning. Vol. 1. Morgan kaufmann, 1993.

[41] Pearson, Karl. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50.302 (1900): 157-175.

[42] Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates

[43] Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function."Journal of the American statistical association 58.301 (1963): 236-244.

[44] Zar, Jerrold H. "Spearman rank correlation." Encyclopedia of Biostatistics(1998).

[45] Available from: http://www.nytimes.com/2007/12/05/education/05education.html?_r=0

VITA

Rama Devi Raghavan received B.Tech. in Information Technology from Anna University, India. Soon after her bachelors, she worked for Infosys Technologies Limited, Chennai for 2 years. She continued her education and worked on her Masters degree in Computer Science from Spring 2011. Her majors in her Master's program were Bioinformatics and Machine learning and she is currently working for Kansas University Medical Center as Bioinformatics Specialist / IT programmer since February 2013.