**MAS 6V10.005**

**Special Topics in Management Information Systems**

**(Fall 2015)**

**Customer Analytics Using R**

**Pratik Gandhi**

**Acknowledgement**

I would like to thank Dr. Atanu Lahiri for his guidance and support throughout the course of this project. His timely suggestions and resources proved to be valuable in the achieving the goals of the project. I would also like to express my gratitude to Mrs. Jennifer Abramson for giving me this opportunity to do the project for ATD, which provided me with a great learning experience.

**Introduction**

The Association for Talent Development (ATD) is a professional membership organization supporting those who develop the knowledge and skills of employees in organizations around the world. They offer a wide variety of products and services and have customers engaging with them throughout the world.

ATD has gathered lot of data about their customers, pertaining to different aspects of their behavior, such as Visits and Page Views on their site, blogs and forum, Webcast Registration and Attendance, Amount and Quantity of purchases, attendance in conferences, interaction with customer care, and many more. They also have customers’ demographic data along with membership start and end date, consecutive years of membership and their current and past statuses.

The objective of the project is to help ATD with analyzing their customer base and derive insights, which they can use to proactively act on and improve customer experience and at the same time increase their business. The marketing team will likely be the biggest beneficiary of the analysis, as it will help them to better target their ad campaigns towards customers according to their tastes and interest.

The project would involve below two form of analysis

Membership Retention – Using the data points available to try and predict the likelihood of someone renewing their membership. The output of the predictive analysis would likely be the probability of a customer being retained, using different engagement variables as input and ‘Retention’ variable as output, based on various machine-learning algorithms. ATD’s job will be to act throughout the year to push customers to actions that will push the number up.

Customer Segmentation - Leveraging customer information surrounding various engagement metrics and other customer traits to identify key differentiators that divide their customers into different groups.

The customer data were scrubbed of any contact info for the individuals to protect privacy. The project uses R programming language to create and implement various statistical methods needed for performing the analysis.

**Predictive Analysis**

Following variables were selected to train the different classifiers used in building predictive models.

* **MemberType** – Individual, Group, Partner
* **MembershipTypeCode** – PROF and PLUS codes across all the three membership types were combined into two type codes, and rest all were marked as ‘Others’
* **ConsecutiveYearsAsMember** - No. of consecutive years customer has been a member
* **AllocationUsage** – Ratio of benefits used to benefits allocated
* **SetCOPPreference** – 0, 1; Preference set or not
* **Purchase Qty** – Quantity purchased during the membership year. *Missing values were filled by multiple imputation using Predictive Mean Matching method.*
* **Purchase Amt -**Amount for stuff purchased during the respective membership year. *Missing values were filled by multiple imputation using Predictive Mean Matching method.*
* **Webcast Registration** - No. of webcasts registered for during the membership year
* **Web Visits** - Web visits in days during the membership year
* **Web Page Views** - No. of Web Page Views during the membership year
* **Blog Visits** - Blog visits in days during the membership year
* **Blog Page Views** - No. of Blog Page Views during the membership year
* **Forum Visits** - Forum Visits in days during the membership year
* **Forum Page Views** - No. of Forum Page Views during the membership year
* **Parature Tickets** - No. of Parature Tickets raised by the customer during the membership year
* **Parature Chats** - No. of Parature Chats by the customer during the membership year
* **Registered for Edu** - No. of registrations for Education programs during the membership year
* **Attended ICE** – 0, 1; attended ICE Conference or not
* **Attended TK** – 0, 1; attended TK Conference or not
* **Frugal** – 0, 1; purchased membership during conference or not. Missing values were replaced by 0
* **Chapter Board** – 0, 1; part of chapter board or not. Missing values were replaced by 0
* **ProfileCompletion** - % of membership profile completed
* **NumberofLogins** – No. of logins to the website during the membership period

Downloads was not considered since it had around 90% of values missing in the dataset.

Attached file displays bar and box plots for various variables against ‘Retained’ and ‘Lost’ status.



**Variables Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **MemberType** | **MembershipTypeCode** | **ConsecutiveYearsAsMember** | **AllocationUsage** |
| Individual : 126503 | Prof : 23876 | Min. : -1.000 | Min. : 0.00000 |
| Group : 24138 | Plus : 36927 | 1st Qu. : 0.000 | 1st Qu. : 0.00000 |
| Partner : 21235 | Others : 111073 | Median : 1.000 | Median : 0.00000 |
|  |  | Mean : 2.597 | Mean : 0.03294 |
|  |  | 3rd Qu. : 4.000 | 3rd Qu. : 0.00000 |
|  |  | Max. :28.000 | Max. : 11.00000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **SetCOPPreference** | **Purchase.Qty** | **Webcast.Registration** | **Web.Visits** |
| No : 105111 | Min. : 1.000 | Min. : 0.000 | Min. : 0.00 |
| Yes : 66765 | 1st Qu. : 2.000 | 1st Qu. : 0.000 | 1st Qu. : 0.00 |
|  | Median : 3.000 | Median : 0.000 | Median : 0.00 |
|  | Mean : 6.612 | Mean : 1.034 | Mean : 17.27 |
|  | 3rd Qu. : 7.000 | 3rd Qu. : 0.000 | 3rd Qu. : 16.00 |
|  | Max. : 1663.000 | Max. : 183.000 | Max. : 1889.00 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Web.Page.Views** | **Blog.Visits** | **Blog.Page.Views** | **Forum.Visits** |
| Min. : 0.00 | Min. : 0.0000 | Min. : 0.000 | Min. : 0.00 |
| 1st Qu. : 0.00 | 1st Qu. : 0.0000 | 1st Qu. : 0.000 | 1st Qu. : 0.00 |
| Median : 0.00 | Median : 0.0000 | Median : 0.000 | Median : 0.00 |
| Mean : 35.48 | Mean : 0.8619 | Mean : 1.075 | Mean : 18.14 |
| 3rd Qu. : 28.00 | 3rd Qu. : 0.0000 | 3rd Qu. : 0.000 | 3rd Qu. : 16.00 |
| Max. :3 942.00 | Max. : 388.0000 | Max. : 868.000 | Max. : 1949.00 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Forum.Page.Views** | **ParatureTickets** | **ParatureChats** | **Registered.for.Edu** |
| Min. : 0.00 | Min. : 0.0000 | Min. : 0.00000 | Min. : 0.0000 |
| 1st Qu. : 0.00 | 1st Qu. : 0.0000 | 1st Qu. : 0.00000 | 1st Qu. : 0.0000 |
| Median : 0.00 | Median : 0.0000 | Median : 0.00000 | Median : 0.0000 |
| Mean : 36.56 | Mean : 0.1231 | Mean : 0.01228 | Mean : 0.1452 |
| 3rd Qu. : 29.00 | 3rd Qu. : 0.0000 | 3rd Qu. : 0.00000 | 3rd Qu. : 0.0000 |
| Max. : 4179.00 | Max. : 147.0000 | Max. : 13.00000 | Max. : 93.0000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Attended. ICE** | **Attended.TK** | **Frugal** | **Chapter.Board** |
| No : 157033 | No : 169286 | No : 164064 | No : 168634 |
| Yes : 14843 | Yes : 2590 | Yes : 7812 | Yes : 3242 |

|  |  |  |  |
| --- | --- | --- | --- |
| **ProfileCompletion** | **NumberofLogins** | **Purchase.Amt** | **Target** |
| Min. : 5.00 | Min. : 0.00 | Min. : 0.00 | Lost : 44348 |
| 1st Qu. : 55.00 | 1st Qu. : 0.00 | 1st Qu. : 147.00 | Retained : 77504 |
| Median : 60.00 | Median : 2.00 | Median : 209.00 | Other : 50024 |
| Mean : 56.93 | Mean : 21.35 | Mean : 637.30 |  |
| 3rd Qu. : 60.00 | 3rd Qu. : 8.00 | 3rd Qu. : 428.00 |  |
| Max. : 100.00 | Max. : 162896.00 | Max. : 39255.90 |  |

**Method & Results**

The data set was prepared for mining in the following manner:

* All records with status ‘Active’, ‘Awaiting Renewal’ and ‘Future Dated’ were moved in to a new scoring data set. The model with best performance will be used on this data set to make predictions.
* Remaining data were randomly split in a ratio of 70/30 into training and testing sets respectively. Training data was used to train different classifiers whose performances were then evaluated using the testing set.

Below are the results from various classification models when evaluated using test partition. Definitions and formulae for measures can be found at the end of document.

***Logistic Regression***

**Confusion Matrix**

Actual

Prediction 0 1

0 9446 4942

1 5532 20697

**Accuracy** – **74.21%**

**Sensitivity/Recall** – 0.8072

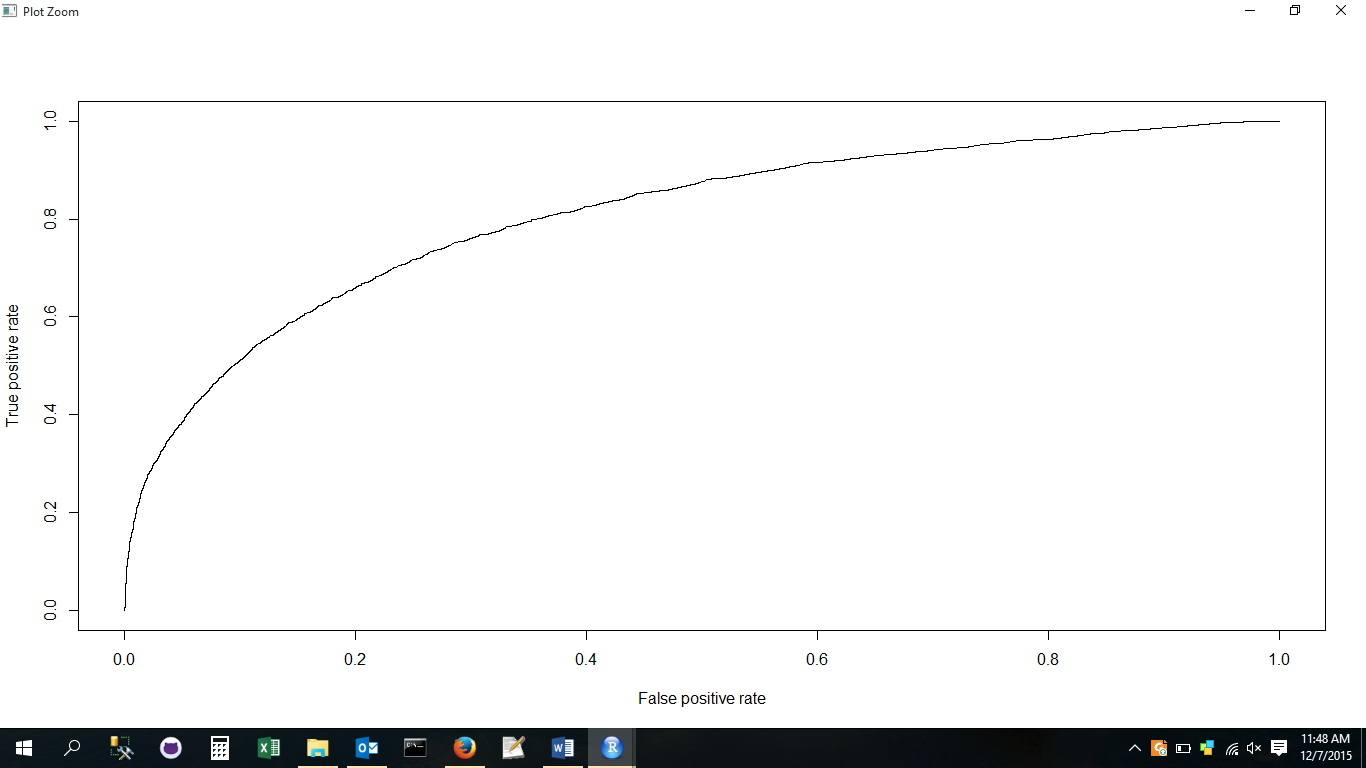
**Specificity** – 0.6307

**Precision** – 0.7891

**Cost based classification** – 4942 \* 2 + 5532 \* 1 = **15416**

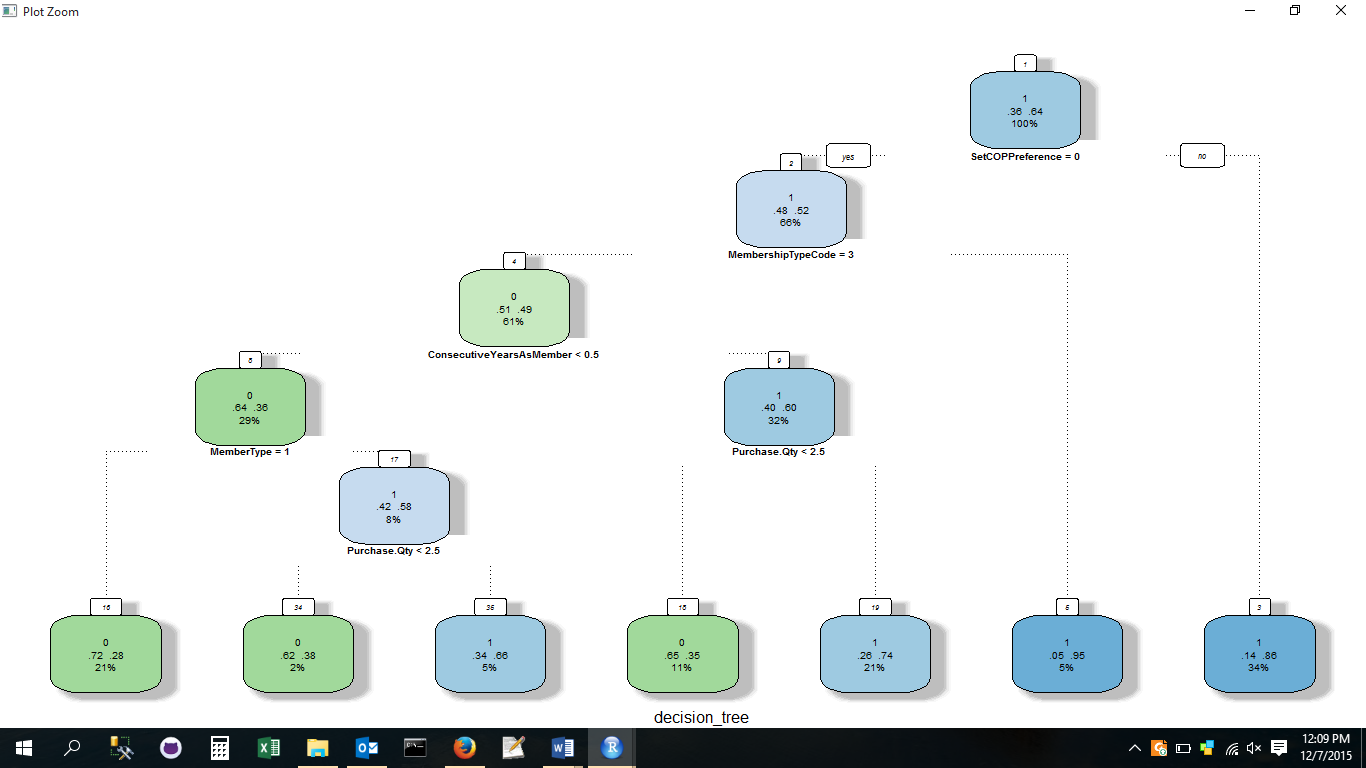
**AUC** – **0.8088**

**ROC curve**



***Decision Tree***

*C4.5 algorithm*



**Confusion Matrix**

Actual

Prediction 0 1

0 9932 4289

1 5046 21350

**Accuracy** – **77.02%**

**Sensitivity/Recall** – 0.8327

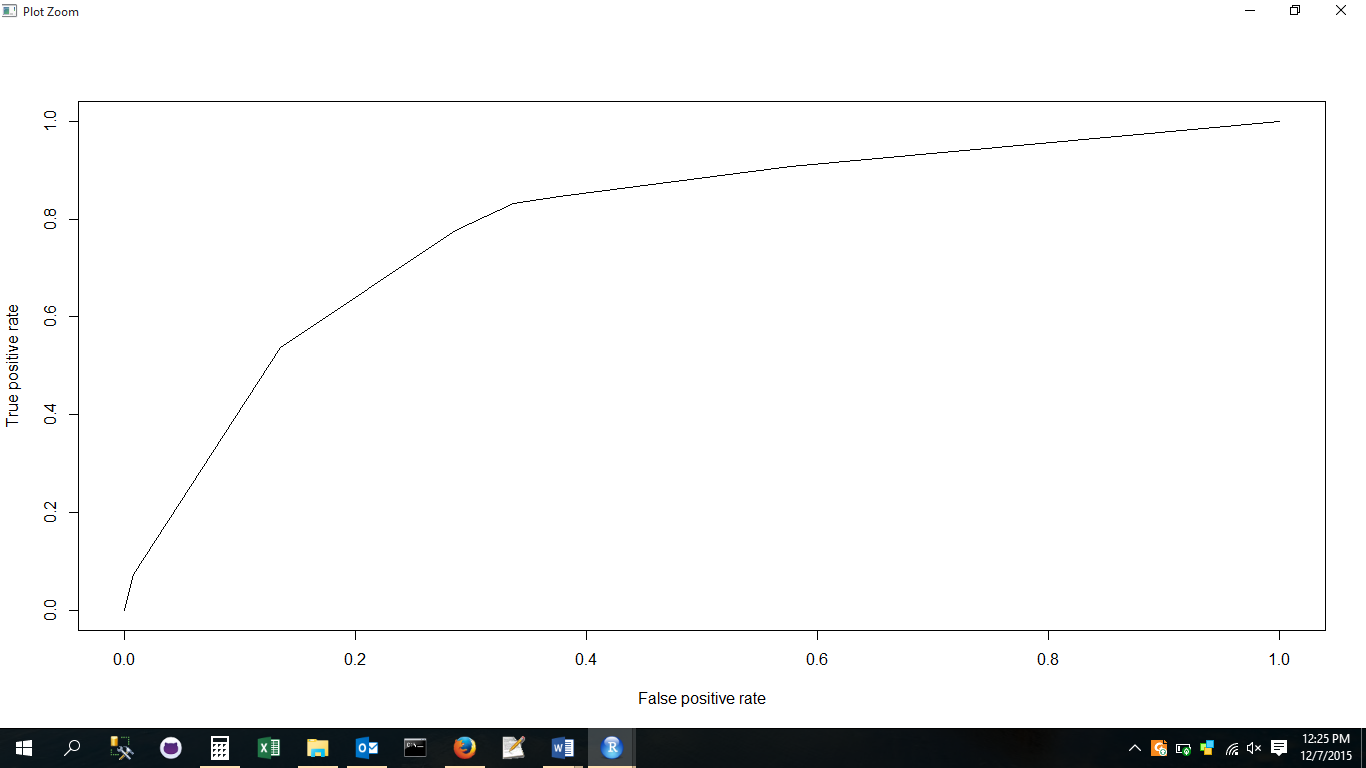
**Specificity** – 0.6631

**Precision** – 0.8088

**Cost based classification** – 4289 \* 2 + 5046 \* 1 = **13624**

**AUC** – **0.7922**

**ROC curve**



*C5.0 algorithm* ***(Updated after including Purchase.Amt)***

**Confusion Matrix**

Actual

Prediction 0 1

0 11564 2570

1 3414 23069

**Accuracy** – **85.27%**

**Sensitivity/Recall** – 0.8998

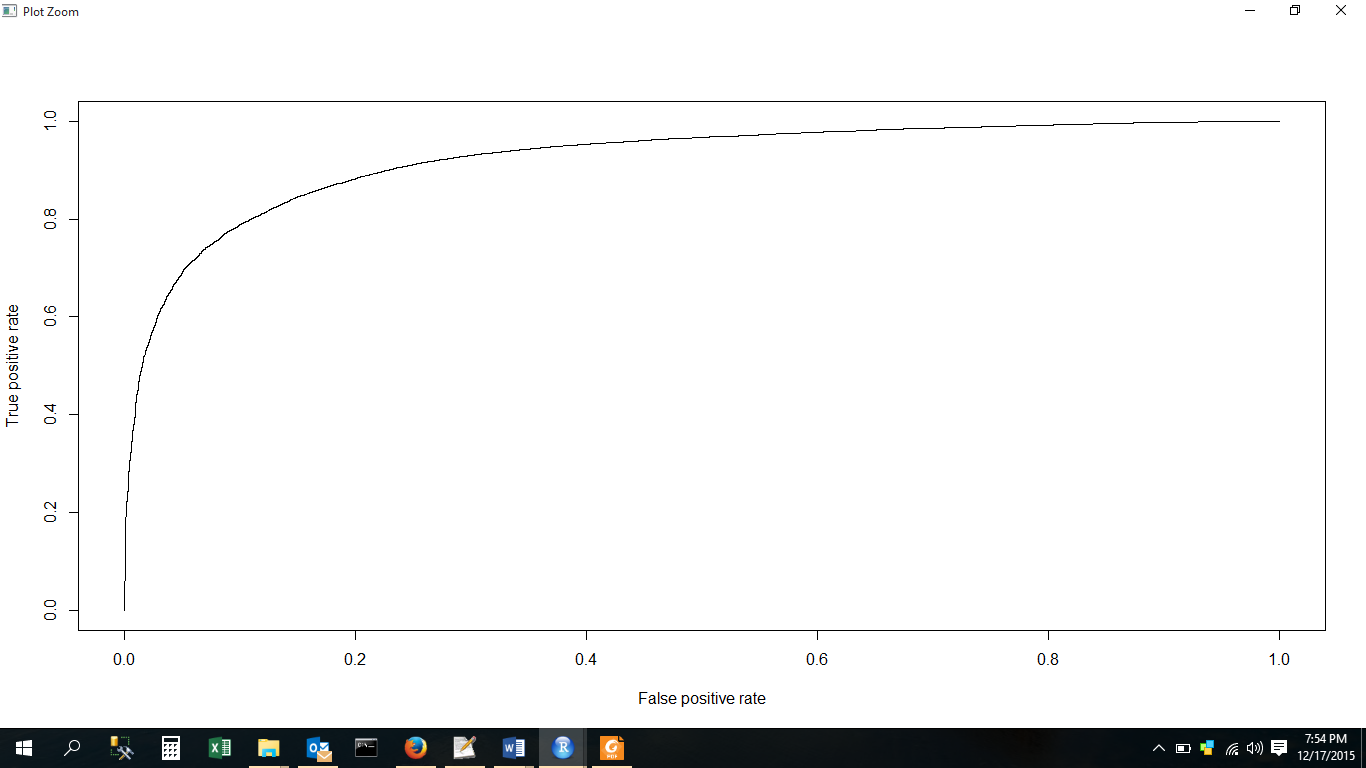
**Specificity** – 0.7721

**Precision** – 0.8711

**Cost based classification** – 2570 \* 2 + 3414 \* 1 = **8554**

**AUC** – **0.8613**

**ROC curve**



***Random Forests***

**Confusion Matrix**

Actual

Prediction 0 1

0 10434 4175

1 4544 21464

**Accuracy** – **78.53%**

**Sensitivity/Recall** – 0.8372

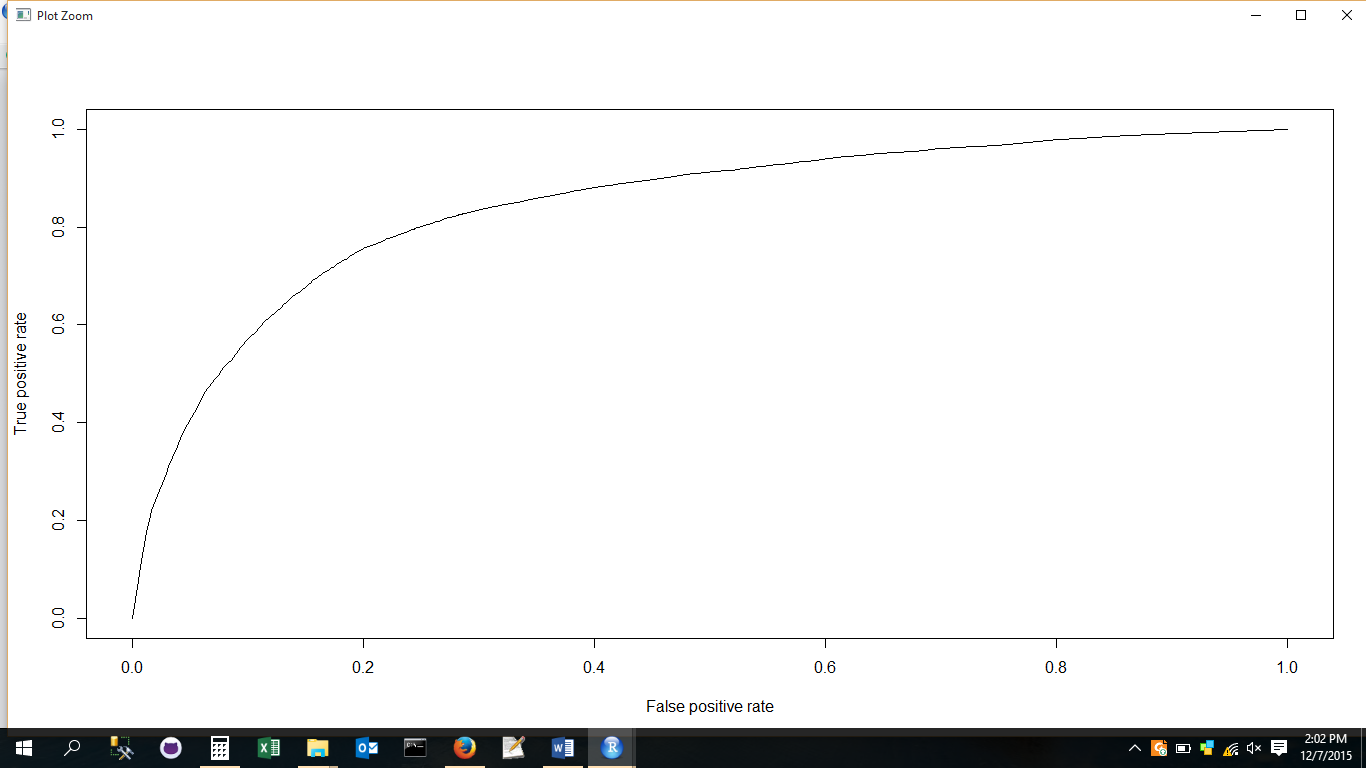
**Specificity** – 0.6966

**Precision** – 0.8253

**Cost based classification** – 4175 \* 2 + 4544 \* 1 = **12894**

**AUC** – **0.8429**

**ROC curve**



***Gradient Boosting***

**Confusion Matrix**

Actual

Prediction 0 1

0 10209 4198

1 4769 21441

**Accuracy** – **77.92%**

**Sensitivity/Recall** – 0.8363

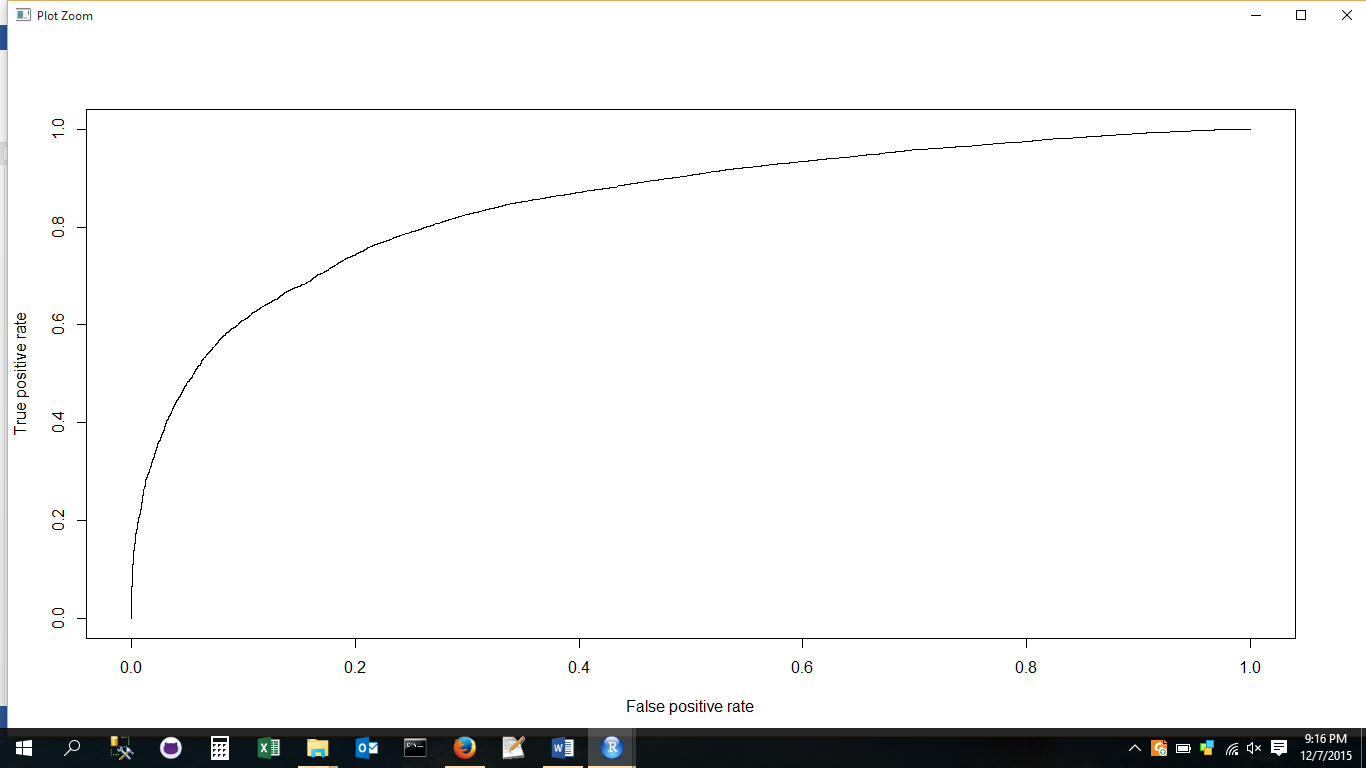
**Specificity** – 0.6816

**Precision** – 0.8180

**Cost based classification** – 4198 \* 2 + 4769 \* 1 = **13165**

**AUC** – **0.8469**

**ROC curve**



***Naïve Bayes***

**Confusion Matrix**

Actual

Prediction 0 1

0 14370 20878

1 608 4761

**Accuracy** – **47.1%**

**Sensitivity/Recall** – 0.1857

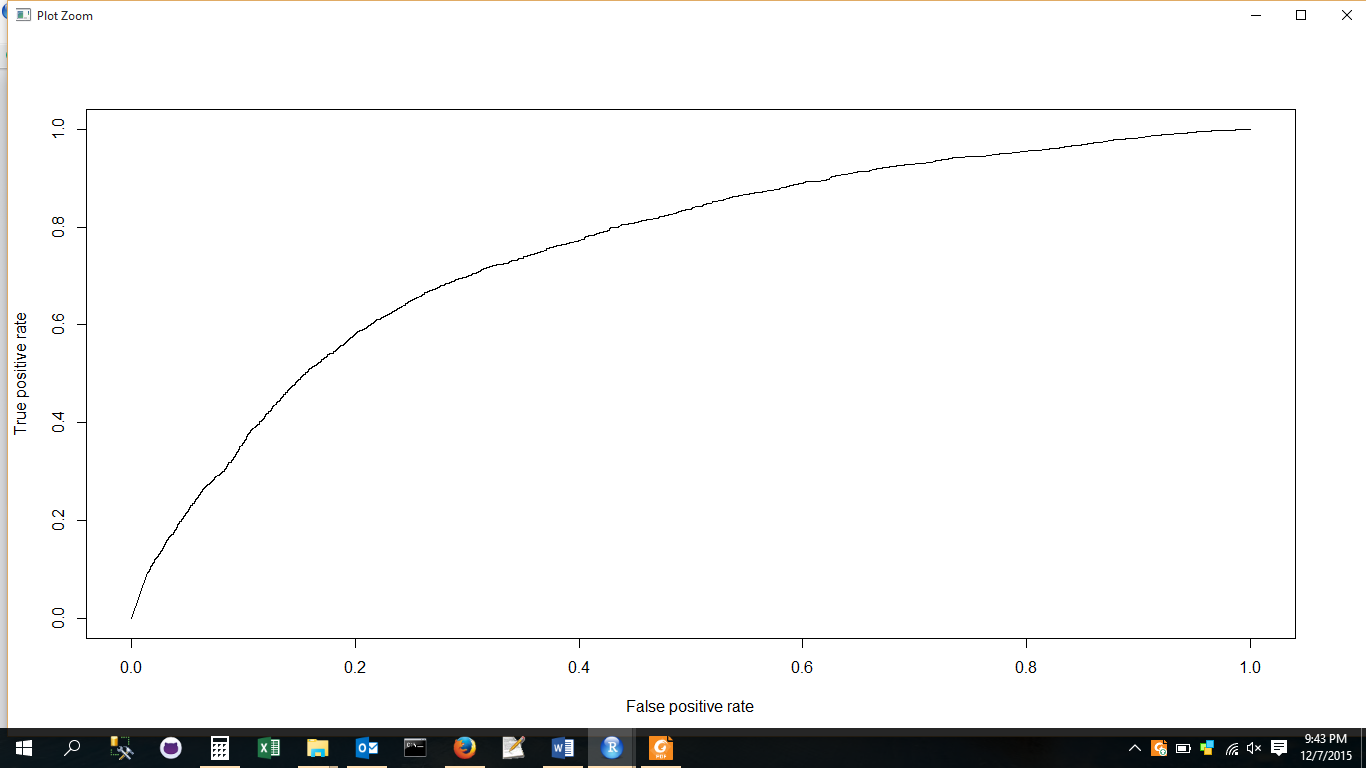
**Specificity** – 0.9594

**Precision** – 0.8868

**Cost based classification** – 20878 \* 2 + 608 \* 1 = **42634**

**AUC** – **0.75612**

**ROC curve**



***Support Vector Machines (SVM)***

**Confusion Matrix**

Actual

Prediction 0 1

0 8451 3972

1 6527 21667

**Accuracy** – **74.15%**

**Sensitivity/Recall** – 0.8451

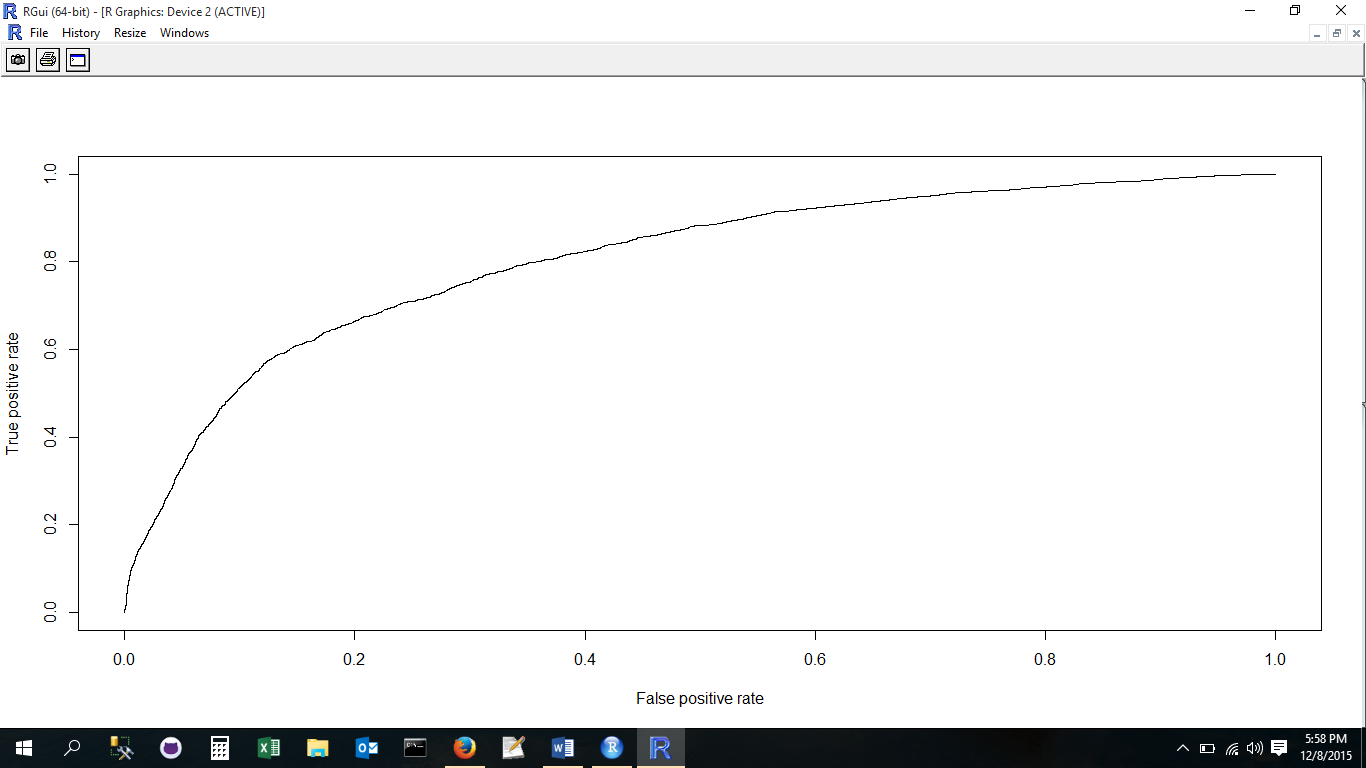
**Specificity** – 0.5642

**Precision** – 0.7685

**Cost based classification** – 3972 \* 2 + 6527 \* 1 = **14471**

**AUC** – **0.8055**

**ROC curve**



***K-Nearest Neighbor***

**Confusion Matrix**

Actual

Prediction 0 1

0 9763 4136

1 5215 21503

**Accuracy** – **76.98%**

**Sensitivity/Recall** – 0.8387

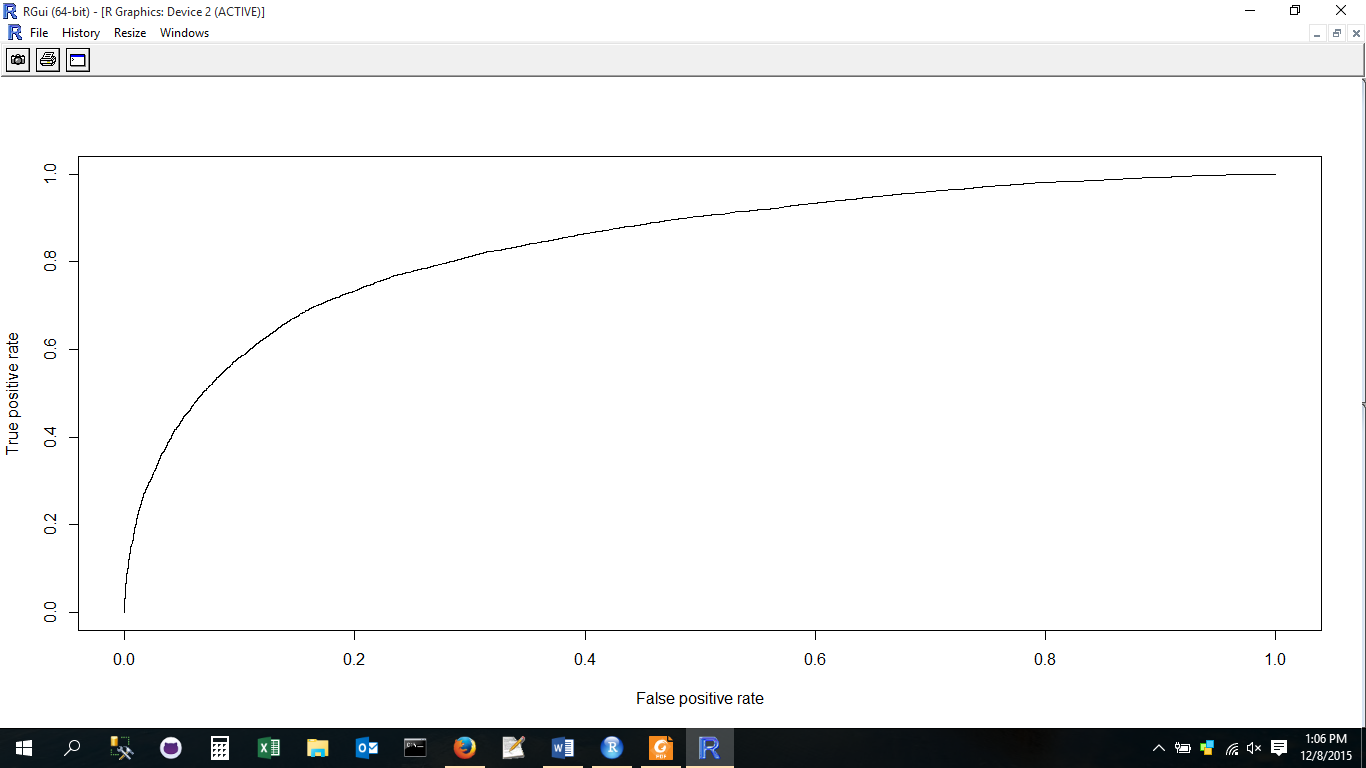
**Specificity** – 0.6518

**Precision** – 0.8048

**Cost based classification** – 4136 \* 2 + 5215 \* 1 = **13487**

**AUC** – **0.8400**

**ROC curve**



Decision tree with C5.0 implementation turned out to be the best amongst the models tested, and was used on scoring data set, while Naïve Bayes turned out to be the worst performer.

**Evaluation Measures**

***Confusion Matrix*** - A contingency table of actual and predicted values showing no. of cases that are correctly and incorrectly classified.

***True Positive (TP) -*** Items correctly predicted as belonging to positive class, here it means retained.

***True Negative (TN)*** - Items correctly predicted as belonging to negative class, here it means lost.

***False Positive (FP)*** - Items from negative class that are incorrectly classified as positive, i.e. lost customers are being predicted as retained.

***False Negative (FN)*** - Items from positive class that are incorrectly classified as negative, i.e. retained customers are being predicted as lost.

***Accuracy*** - TP + TN/TP + FN + TN + FP (*higher the better)*

***Sensitivity/Recall/True Positive Rate*** - TP/TP + FN

***Specificity*** - TN/FP + TN

***Precision*** - TP/TP + FP

***False Positive Rate*** - FP/FP + TN

***Cost based classification*** - Assigning a cost to FP and FN. Here, a cost of 2 is assigned to cases where positive values are incorrectly classified as negative, and a cost of 1 to cases where negative values are misclassified as positive; *lower the better*.

***Receiver Operating Characteristic (ROC) Curve*** - ROC curve is a graph of true positive rate (benefits) plotted against false positive rate (costs) at different threshold values for probabilities/scores obtained from the model. It shows the tradeoff between sensitivity and specificity. *The closer the curve is to the upper left corner, and then to the top border, the better is the performance of the classifier*.

***Area Under Curve (AUC)*** - It is the area under the ROC curve. It represents the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation. *Values closer to 1 are considered to be better*.

**Clustering**

Cluster analysis is an unsupervised technique that consists of finding group of objects such that the objects in the group are similar to one another and different from the objects in other groups. Similarity in this case is expressed in terms of a distance function. For the project, engagement variables were considered to try to find clusters that would segment the customers according to their engagement levels with the company.

**K-means clustering**

Variables were first standardized and then used with k-means algorithm for clustering. Number of clusters were set to 3. On comparing the clustering result with the class label (Target) to check whether similar objects are grouped together, we get the following result.

Cluster

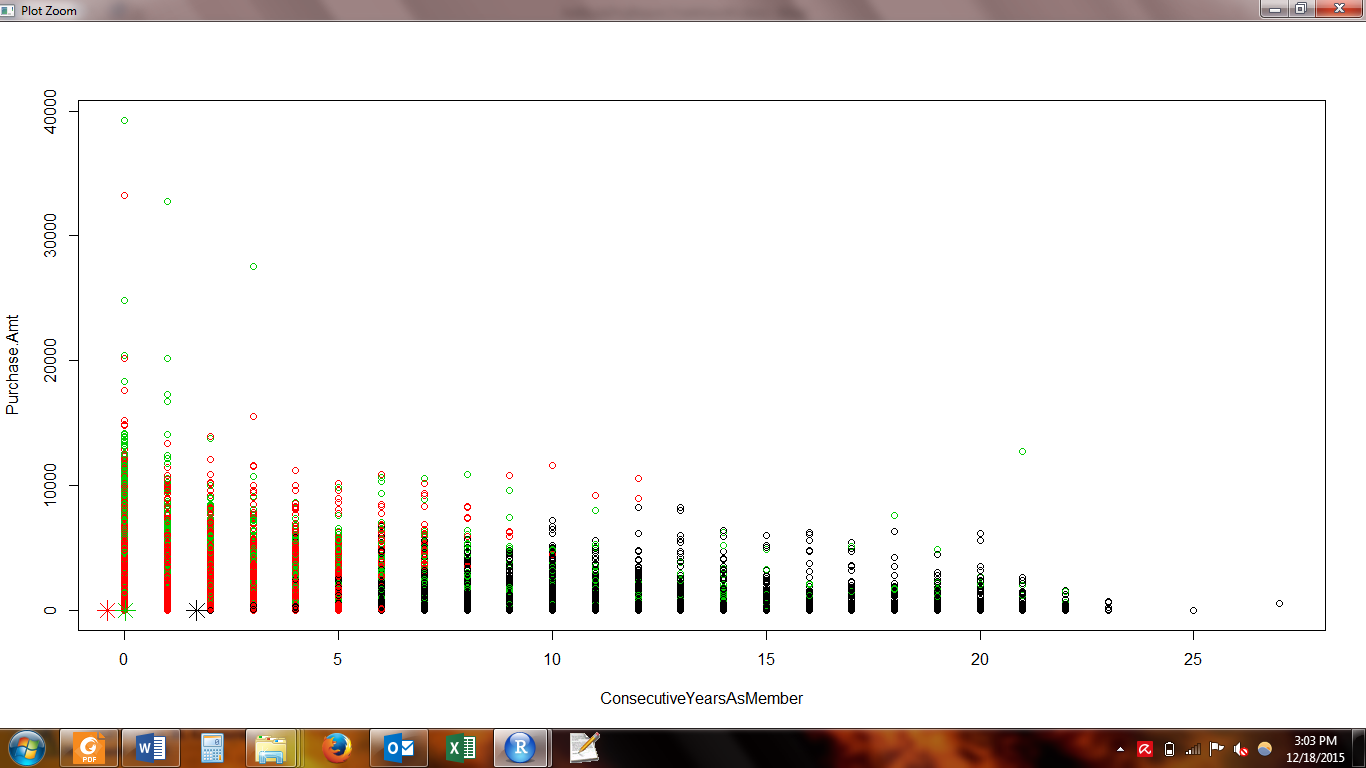
1 2 3

0 4205 39823 320

1 18208 54674 4622

The above table shows that clusters are not clearly able to separate ‘Lost’ (0) customers form the ‘Retained’ (1) ones. This might be because none of the variables has a significant direct bearing on the outcome variable.

Below plot shows the clusters and the centers for two of the dimensions – ConsecutiveYearsAsMember and Purchase.Amt



**References**

R and Data Mining: Examples and Case Studies, Yanchang Zhao

Bayesian Network Classifiers - Nir Friedman, Dan Geiger, Moises Goldszmidt

A scaling law for the validation-set training-set size ratio, Isab elle Guyon

<http://www.r-bloggers.com/identifying-potential-customers-with-classification-techniques-in-r-language/>

<http://trevorstephens.com/post/72916401642/titanic-getting-started-with-r>

<https://github.com/wehrley/wehrley.github.io/blob/master/SOUPTONUTS.md>

<http://www.ats.ucla.edu/stat/r/faq/R_pmm_mi.htm>