Satisfaction analysis and prediction of Taught Masters Students’ in Ireland

Pratik (18250375)

August 4, 2019

# Libraries

Creating a function to get all the libraries into R Studio.

run\_lib <- function(){  
 library(randomForest)  
 library(e1071)  
 library(rpart)  
 library(leaps)  
 library(pls)  
 library(VGAM)  
 library(MLmetrics)  
 library(glue)  
 library(glmnet)  
 library(ggplot2)   
 library(ISLR)  
 library(snakecase)  
 library(GPArotation)  
 library(psych)  
 library(MASS)  
 library(mvtnorm)  
 library(dplyr)  
 library(skimr)   
 library(caret)   
 library(tidyverse)   
 library(haven)   
 library(nnet)   
 library(Hmisc)   
 library(reshape2)   
 library(klaR)   
 library(ranger)  
}

# Reading Data

Reading Raw Data in Sav Format.

eurostudent <- read\_sav("E\_V\_Data.sav")

# Filtering Dataset

Selecting relevant columns.

euro\_filtered <- subset(eurostudent[,c(3,4,6,7,14,15,16,17,18,19,20,21,22,23,24,  
 25,57,65,66,67,77,87,97,110,112,135,136,137,  
 218,219,220,223,243,244,245,246,247,248,249,  
 252,253,258,259,260,261,262,263,272,273)])  
  
# Step 1. Taught Masters Course and complete data for relevant columns.  
  
taught\_masters <- subset(euro\_filtered,(Q1.1 == "6"))   
taught\_masters <- taught\_masters[,c(-1,-6:-10,-43:-47)]  
  
# Skimming - To check missing values  
  
Skimmed <- suppressWarnings(skim\_to\_wide(taught\_masters))  
  
# Step 2- Average monthly expenses missing  
  
taught\_masters <- subset(taught\_masters,Q3.8.1c1 >= 0 & Q3.8.1c2 >= 0)  
  
# Step 3. Students who haven't filled satisfaction rating missing  
  
taught\_masters <- subset(taught\_masters,Q6.2.5 > 0 & Q6.2.1 > 0 & Q6.2.2 > 0)  
  
# Step 4 - Q5.1 - What age are you (years)? missing  
  
taught\_masters <- subset(taught\_masters,Q5.1 > 0) # 607 Observations

# Creating Backup

Data is stored and read back in CSV format.

write.csv(taught\_masters, "taught\_masters.csv", row.names=T)  
taught\_masters <- read.csv("taught\_masters.csv", header=TRUE)  
taught\_masters$X = NULL;  
dim(taught\_masters)

## [1] 607 38

# Final Dataset has 607 rows  
  
taught\_masters\_n <- taught\_masters

# Variable Conversion

Grouping Response Levels and Categorical variables for analysis and interpretation.

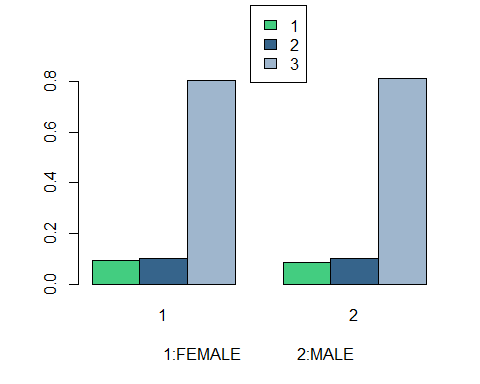
# Re-grouped response to a scale of 1-3 as some of the groups had limited responses (unbalanced data).   
  
taught\_masters\_n$Q6.2.1 <- ifelse(taught\_masters\_n$Q6.2.1 < 3,1,  
 ifelse((taught\_masters\_n$Q6.2.1 == 3), 2,  
 ifelse(taught\_masters\_n$Q6.2.1 >= 4, 3,4)))  
  
taught\_masters\_n$Q6.2.2 <- ifelse(taught\_masters\_n$Q6.2.2 < 3,1,  
 ifelse((taught\_masters\_n$Q6.2.2 == 3 ), 2,  
 ifelse(taught\_masters\_n$Q6.2.2 >= 4, 3,4)))  
taught\_masters\_n$Q6.2.5 <- ifelse(taught\_masters\_n$Q6.2.5 < 3,1,  
 ifelse((taught\_masters\_n$Q6.2.5 == 3 ), 2,  
 ifelse(taught\_masters\_n$Q6.2.5 >= 4, 3,4)))  
  
# Creating factors and groups for categorical predictors  
  
taught\_masters\_n$Q5.4 <- ifelse(taught\_masters\_n$Q5.4 < 3,1,  
 ifelse(taught\_masters\_n$Q5.4 >= 3 , 2,3))  
  
taught\_masters\_n$Q7.5 <- ifelse(taught\_masters\_n$Q7.5 < 5,1,  
 ifelse(taught\_masters\_n$Q7.5 >= 5 & taught\_masters\_n$Q7.5 < 8,2,  
 ifelse(taught\_masters\_n$Q7.5 >= 8, 3,4)))  
  
taught\_masters\_n$Q5.4 <- factor(taught\_masters\_n$Q5.4 )  
taught\_masters\_n$Q7.5 <- factor(taught\_masters\_n$Q7.5 )

# Plots

Creating Plots for visualization.(1: - Unsatisfied; 2: - Neutral; 3: - Satisfied)

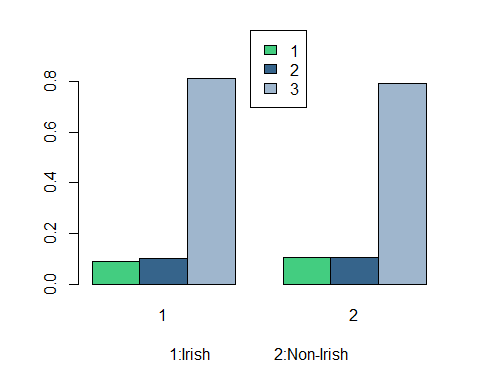
## Gender

##   
## 1 2  
## 1 0.09383378 0.08547009  
## 2 0.10187668 0.10256410  
## 3 0.80428954 0.81196581



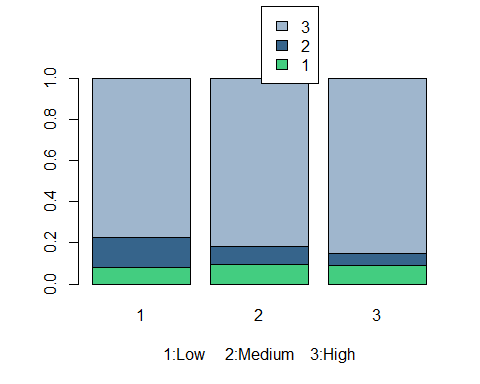
## Nationality

##   
## 1 2  
## 1 0.08757637 0.10344828  
## 2 0.10183299 0.10344828  
## 3 0.81059063 0.79310345



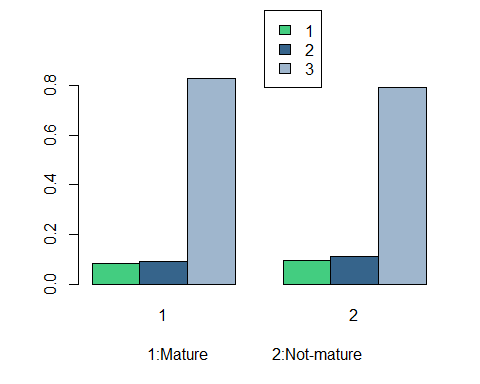
## Social Standing

##   
## 1 2 3  
## 1 0.08219178 0.09367681 0.08823529  
## 2 0.14383562 0.09133489 0.05882353  
## 3 0.77397260 0.81498829 0.85294118



## Mature Student

##   
## 1 2  
## 1 0.08267717 0.09631728  
## 2 0.09055118 0.11048159  
## 3 0.82677165 0.79320113



# Splitting Data

Data is partitioned in 2:1 ratio with training and testing set.

trainIndex1 <- createDataPartition(taught\_masters\_n$Q6.2.5, p = .67,  
 list = FALSE,times = 1)  
Train1\_n <- taught\_masters\_n[trainIndex1,]  
Test1\_n <- taught\_masters\_n[-trainIndex1,]  
  
trainIndex2 <- createDataPartition(taught\_masters\_n$Q6.2.1, p = .67,  
 list = FALSE,times = 1)  
Train2\_n <- taught\_masters\_n[trainIndex2,]  
Test2\_n <- taught\_masters\_n[-trainIndex2,]  
  
trainIndex3 <- createDataPartition(taught\_masters\_n$Q6.2.2, p = .67,  
 list = FALSE,times = 1)  
Train3\_n <- taught\_masters\_n[trainIndex3,]  
Test3\_n <- taught\_masters\_n[-trainIndex3,]  
  
rm(trainIndex1,trainIndex2,trainIndex3)

# Data Modelling

## 1. Satisfaction with College (Reponse)

### Linear Models

In this thesis, I first check the associations with the utilization of classical regression analysis. This methodology gives us an initial understanding of dependent and independent factors. But in our case the data is not normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

##   
## Call:  
## lm(formula = Q6.2.5 ~ Q1.14.1 + Q1.14.2 + Q1.14.5 + Q1.14.6,   
## data = Train1\_n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1227 -0.2156 0.1090 0.2783 1.2662   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.74598 0.14115 5.285 2.07e-07 \*\*\*  
## Q1.14.1 0.15430 0.04337 3.558 0.000418 \*\*\*  
## Q1.14.2 0.09291 0.02957 3.143 0.001799 \*\*   
## Q1.14.5 0.16939 0.03373 5.021 7.73e-07 \*\*\*  
## Q1.14.6 0.07731 0.02739 2.822 0.005005 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5075 on 402 degrees of freedom  
## Multiple R-squared: 0.3554, Adjusted R-squared: 0.349   
## F-statistic: 55.41 on 4 and 402 DF, p-value: < 2.2e-16

The response is “Students perceived satisfaction with college” and significant predictors are derived , where P-Value<0.05, i.e. in our case Satisfaction with Organisation of timetable and studies, Teaching Quality, Teaching staff’s attitude as well as Study amenities (e.g. computers, library, classrooms). For all the significant predictors , coefficients are positive, i.e. Satisfaction with college increases with an increase in satisfaction with the predictors.

### SVM

##   
## Call:  
## svm(formula = factor(Q6.2.5) ~ Q1.14.1 + Q1.14.2 + Q1.14.4 +   
## Q1.14.5 + Q1.14.6, data = Train1\_n)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 1   
## gamma: 0.2   
##   
## Number of Support Vectors: 175  
##   
## ( 95 44 36 )  
##   
##   
## Number of Classes: 3   
##   
## Levels:   
## 1 2 3

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 9 0 8  
## 2 2 0 16  
## 3 1 1 163  
##   
## Overall Statistics  
##   
## Accuracy : 0.86   
## 95% CI : (0.8041, 0.9049)  
## No Information Rate : 0.935   
## P-Value [Acc > NIR] : 0.9999551   
##   
## Kappa : 0.3724   
##   
## Mcnemar's Test P-Value : 0.0001227   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7500 0.0000 0.8717  
## Specificity 0.9574 0.9095 0.8462  
## Pos Pred Value 0.5294 0.0000 0.9879  
## Neg Pred Value 0.9836 0.9945 0.3143  
## Prevalence 0.0600 0.0050 0.9350  
## Detection Rate 0.0450 0.0000 0.8150  
## Detection Prevalence 0.0850 0.0900 0.8250  
## Balanced Accuracy 0.8537 0.4548 0.8589

## [1] 0.6206897

Multi-category SVM algorithm from the e1071 package (7) is used for analysis. SVM fits the test data correctly and misclassified only 22.5% observations. Again, in this case, Group 2 (“Neutral” in terms of satisfaction with the college) is misclassified 100%. Overall accuracy is 77.5% and it does a great job in predicting “Unsatisfied” and “Satisfied” students.

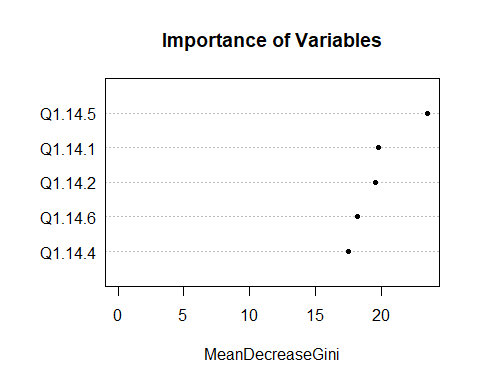
### Random Forest

“randomForest” function inside “randomForest” package is used for analysis.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 10 0 7  
## 2 4 3 11  
## 3 2 7 156  
##   
## Overall Statistics  
##   
## Accuracy : 0.845   
## 95% CI : (0.7873, 0.8922)  
## No Information Rate : 0.87   
## P-Value [Acc > NIR] : 0.87483   
##   
## Kappa : 0.4279   
##   
## Mcnemar's Test P-Value : 0.05343   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.6250 0.3000 0.8966  
## Specificity 0.9620 0.9211 0.6538  
## Pos Pred Value 0.5882 0.1667 0.9455  
## Neg Pred Value 0.9672 0.9615 0.4857  
## Prevalence 0.0800 0.0500 0.8700  
## Detection Rate 0.0500 0.0150 0.7800  
## Detection Prevalence 0.0850 0.0900 0.8250  
## Balanced Accuracy 0.7935 0.6105 0.7752

## [1] 0.6060606

Random Forest fits the test data correctly and misclassified only 19% observations. Group 2 (“Neutral” in terms of satisfaction with the college) is misclassified 100%. Overall accuracy is 81% and it does a great job in predicting “Unsatisfied” and “Satisfied” students.



“Importance ranking” for predictors is also shown in the above plot.

### Linear Discriminant Analysis

LDA assumes that the predictors are normally distributed. As stated earlier that we have data that is not normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.

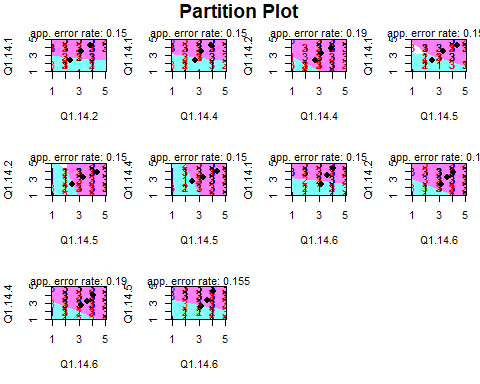
## Call:  
## lda(factor(Q6.2.5) ~ Q1.14.1 + Q1.14.2 + Q1.14.4 + Q1.14.5 +   
## Q1.14.6, data = Train1\_n)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.09336609 0.10810811 0.79852580   
##   
## Group means:  
## Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6  
## 1 2.894737 2.315789 2.526316 2.684211 2.973684  
## 2 3.590909 3.181818 3.363636 3.500000 3.590909  
## 3 4.224615 3.895385 3.883077 4.323077 4.101538  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Q1.14.1 0.5111665 -0.03561930  
## Q1.14.2 0.2806533 -0.17440987  
## Q1.14.4 0.1613915 -0.95091470  
## Q1.14.5 0.4925170 0.96725703  
## Q1.14.6 0.2300676 -0.02992203  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9958 0.0042

## [1] 0.5789474

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3. LD1 is the first linear discriminant which is a linear combination of below parameters: -

(Q1.14.1*0.3611706) + (Q1.14.2*0.3136274) + (Q1.14.4*0.1703964) + (Q1.14.5*0.5724015) + (Q1.14.6\*0.2461356)

Teaching staff’s behavior towards students, Teaching Quality and Organisation of timetable and studies have the largest coefficients which signify them as the most important parameters in the estimation of response. All the coefficients are positive, this resonates the result that increases in satisfaction of any of the included predictors in the model lead to an overall increase in student satisfaction with college. LD1 explains 97.54% of the overall variance.



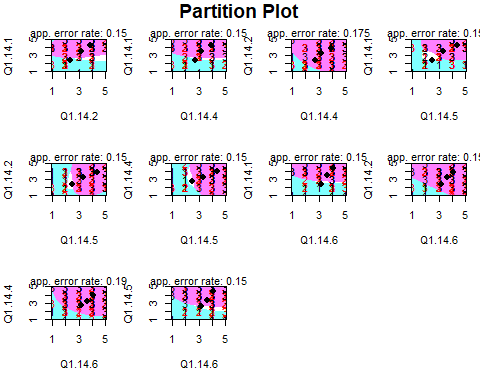
Partition plots are a series of plots for all combinations of categories. The plot shows the three groups with overlapping boundary. It shows that Group 1 and 3 are fairly distinguishable, but Group 2 has major overlapping with the boundary of both the remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the college.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 11 0 6  
## 2 5 0 13  
## 3 5 0 160  
##   
## Overall Statistics  
##   
## Accuracy : 0.855   
## 95% CI : (0.7984, 0.9007)  
## No Information Rate : 0.895   
## P-Value [Acc > NIR] : 0.9705518   
##   
## Kappa : 0.4262   
##   
## Mcnemar's Test P-Value : 0.0004213   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.5238 NA 0.8939  
## Specificity 0.9665 0.91 0.7619  
## Pos Pred Value 0.6471 NA 0.9697  
## Neg Pred Value 0.9454 NA 0.4571  
## Prevalence 0.1050 0.00 0.8950  
## Detection Rate 0.0550 0.00 0.8000  
## Detection Prevalence 0.0850 0.09 0.8250  
## Balanced Accuracy 0.7451 NA 0.8279

This model fit the test data correctly and misclassified only 17% observations. But it doesn’t predict group 2 correctly as there is a lot overlap between the adjacent classes hence it is very difficult to predict if a student is “Neutral” in terms of satisfaction with the college. Otherwise, it does a great job of predicting “Unsatisfied” or “Satisfied” students accurately. The model performs comprehensively well with the test data (not used for modeling) having an overall classification accuracy of 83%.

### Quadratic Discriminant Analysis

## Length Class Mode   
## prior 3 -none- numeric   
## counts 3 -none- numeric   
## means 15 -none- numeric   
## scaling 75 -none- numeric   
## ldet 3 -none- numeric   
## lev 3 -none- character  
## N 1 -none- numeric   
## call 3 -none- call   
## terms 3 terms call   
## xlevels 0 -none- list



Group 1 and 3 are fairly distinguishable, but Group 2 has an overlapping boundary with remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the college.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 11 2 4  
## 2 4 2 12  
## 3 6 5 154  
##   
## Overall Statistics  
##   
## Accuracy : 0.835   
## 95% CI : (0.7762, 0.8836)  
## No Information Rate : 0.85   
## P-Value [Acc > NIR] : 0.7596   
##   
## Kappa : 0.4226   
##   
## Mcnemar's Test P-Value : 0.2670   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.5238 0.2222 0.9059  
## Specificity 0.9665 0.9162 0.6333  
## Pos Pred Value 0.6471 0.1111 0.9333  
## Neg Pred Value 0.9454 0.9615 0.5429  
## Prevalence 0.1050 0.0450 0.8500  
## Detection Rate 0.0550 0.0100 0.7700  
## Detection Prevalence 0.0850 0.0900 0.8250  
## Balanced Accuracy 0.7451 0.5692 0.7696

## [1] 0.5789474

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3. QDA model predicts the test data correctly but misclassified slightly more that LDA with 22.7% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it does a fair job in predicting “Unsatisfied” or “Satisfied” students accurately with an accuracy of 77.3%, but not as good as LDA.

### Multinomial Logistic Regression

Now once we have established the final set of predictors using the above chapter, we can analyze their relationship with the response.

## # weights: 21 (12 variable)  
## initial value 447.135201   
## iter 10 value 239.940923  
## iter 20 value 187.613637  
## final value 187.535729   
## converged

## Call:  
## multinom(formula = factor(Q6.2.5) ~ Q1.14.1 + Q1.14.2 + Q1.14.4 +   
## Q1.14.5 + Q1.14.6, data = Train1\_n)  
##   
## Coefficients:  
## (Intercept) Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6  
## 2 -4.273673 0.2913448 0.3699819 0.4611358 0.1397360 0.2011129  
## 3 -8.981374 0.9107614 0.6053995 0.5238860 0.6809016 0.4740576  
##   
## Std. Errors:  
## (Intercept) Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6  
## 2 1.172216 0.3208549 0.2527120 0.2433659 0.2666127 0.2219381  
## 3 1.323130 0.3130482 0.2275298 0.2230125 0.2479882 0.2073337  
##   
## Residual Deviance: 375.0715   
## AIC: 399.0715

## (Intercept) Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6  
## 2 0.01393052 1.338226 1.447708 1.585874 1.149970 1.222763  
## 3 0.00012573 2.486215 1.831984 1.688577 1.975658 1.606499

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 10 1 6  
## 2 4 1 13  
## 3 5 0 160  
##   
## Overall Statistics  
##   
## Accuracy : 0.855   
## 95% CI : (0.7984, 0.9007)  
## No Information Rate : 0.895   
## P-Value [Acc > NIR] : 0.970552   
##   
## Kappa : 0.4261   
##   
## Mcnemar's Test P-Value : 0.001912   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.5263 0.50000 0.8939  
## Specificity 0.9613 0.91414 0.7619  
## Pos Pred Value 0.5882 0.05556 0.9697  
## Neg Pred Value 0.9508 0.99451 0.4571  
## Prevalence 0.0950 0.01000 0.8950  
## Detection Rate 0.0500 0.00500 0.8000  
## Detection Prevalence 0.0850 0.09000 0.8250  
## Balanced Accuracy 0.7438 0.70707 0.8279

## [1] 0.5555556

## (Intercept) Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6  
## 2 2.665543e-04 0.363864072 0.143181456 0.05811559 0.600197787 0.3648478  
## 3 1.137160e-11 0.003622009 0.007796723 0.01881721 0.006038154 0.0222281

Neutral vs Unsatisfied Students -

With P-values (< 0.05) only Q1.14.1 is significant.

For a one-level increase in satisfaction for Quality of teaching, we expect to see an 87% (OR = 1.87) rise in the odds of being Neutral vs Unsatisfied in terms of satisfaction from college.

Satisfied vs Unsatisfied Students -

With P-values (< 0.05) all predictors Q1.14.1, Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 are significant. For example, we expect to see 113% (OR = 2.13) increase in the odds of being Satisfied vs Unsatisfied for one level increase in satisfaction of Quality of teaching. Similarly, odds of being Satisfied vs Unsatisfied for one level increase in satisfaction of Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 increases by 93%, 54%, 97% and 91% (OR = 1.937122 / 1.549803 / 1.971980 / 1.915410).

### Ordinal regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.

## Call:  
## polr(formula = factor(Q6.2.5) ~ Q1.14.1 + Q1.14.2 + Q1.14.4 +   
## Q1.14.5 + Q1.14.6, data = Train1\_n, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## Q1.14.1 0.7098 0.2193 3.237  
## Q1.14.2 0.3830 0.1578 2.427  
## Q1.14.4 0.3001 0.1545 1.943  
## Q1.14.5 0.5752 0.1748 3.291  
## Q1.14.6 0.3991 0.1444 2.763  
##   
## Intercepts:  
## Value Std. Error t value  
## 1|2 5.9973 0.8698 6.8948   
## 2|3 7.3818 0.9136 8.0795   
##   
## Residual Deviance: 374.6879   
## AIC: 388.6879

## Q1.14.1 Q1.14.2 Q1.14.4 Q1.14.5 Q1.14.6   
## 2.033562 1.466722 1.350053 1.777398 1.490442

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 10 0 7  
## 2 5 0 13  
## 3 3 0 162  
##   
## Overall Statistics  
##   
## Accuracy : 0.86   
## 95% CI : (0.8041, 0.9049)  
## No Information Rate : 0.91   
## P-Value [Acc > NIR] : 0.9926812   
##   
## Kappa : 0.4205   
##   
## Mcnemar's Test P-Value : 0.0002054   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.5556 NA 0.8901  
## Specificity 0.9615 0.91 0.8333  
## Pos Pred Value 0.5882 NA 0.9818  
## Neg Pred Value 0.9563 NA 0.4286  
## Prevalence 0.0900 0.00 0.9100  
## Detection Rate 0.0500 0.00 0.8100  
## Detection Prevalence 0.0850 0.09 0.8250  
## Balanced Accuracy 0.7585 NA 0.8617

## [1] 0.5714286

## Value Std. Error t value  
## Q1.14.1 0.7097889 0.2192531 3.237304  
## Q1.14.2 0.3830301 0.1578271 2.426897  
## Q1.14.4 0.3001438 0.1544870 1.942842  
## Q1.14.5 0.5751504 0.1747745 3.290814  
## Q1.14.6 0.3990728 0.1444122 2.763429  
## 1|2 5.9973238 0.8698368 6.894769  
## 2|3 7.3817612 0.9136443 8.079469

## Value Std. Error t value p value  
## Q1.14.1 0.7097889 0.2192531 3.237304 1.206650e-03  
## Q1.14.2 0.3830301 0.1578271 2.426897 1.522856e-02  
## Q1.14.4 0.3001438 0.1544870 1.942842 5.203523e-02  
## Q1.14.5 0.5751504 0.1747745 3.290814 9.989807e-04  
## Q1.14.6 0.3990728 0.1444122 2.763429 5.719750e-03  
## 1|2 5.9973238 0.8698368 6.894769 5.395242e-12  
## 2|3 7.3817612 0.9136443 8.079469 6.504914e-16

With P-values (< 0.05) all predictors Q1.14.1, Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 are significant.

We expect to see 61% (OR = 1.61) increase in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in satisfaction of Quality of teaching (Q1.14.1). Due to the property of proportional odds assumption, the same increase, 1.61 times, is found between Satisfied vs Neutral or Unsatisfied.

Similarly, odds of being Unsatisfied vs Neutral or Satisfied for one level increase in satisfaction of Q1.14.2, Q1.14.4, Q1.14.5, Q1.14.6 increases by 57%, 38%, 89.9% and 58.7% (OR =1.57 / 1.37 / 1.899 / 1.587). Similar conclusion for Satisfied vs Neutral or Unsatisfied category due to proportional odds assumption.

## 2. Satisfaction with Accommodation (Response)

### Linear Models

In this thesis, I first check the associations with the utilization of classical regression analysis. This methodology gives us an initial understanding of dependent and independent factors. But in our case the data is not normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

##   
## Call:  
## lm(formula = Q6.2.1 ~ Q3.8.1c1 + Q3.12 + Q5.2 + Q6.1.1, data = Train2\_n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.04147 -0.01185 0.20514 0.36246 0.77194   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.836e+00 1.700e-01 16.685 < 2e-16 \*\*\*  
## Q3.8.1c1 1.149e-04 4.335e-05 2.650 0.008356 \*\*   
## Q3.12 -9.501e-02 2.571e-02 -3.695 0.000251 \*\*\*  
## Q5.2 -1.699e-01 6.456e-02 -2.631 0.008844 \*\*   
## Q6.1.1 8.034e-02 2.513e-02 3.197 0.001499 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6296 on 402 degrees of freedom  
## Multiple R-squared: 0.09565, Adjusted R-squared: 0.08665   
## F-statistic: 10.63 on 4 and 402 DF, p-value: 3.384e-08

The response is students perceived satisfaction with accommodation which is derived from predictors (P-Value<0.05).

Gender [1- Female and 2- Male] - Coefficient is negative which signifies that satisfaction level is less for a male when compared to female students.

Perceived level of financial difficulties - Negative coefficient signifies that satisfaction level decreases when students are facing more financial difficulty.

Feeling of being in good spirits and cheerful over last two weeks - Positive coefficient signifies that satisfaction level increases when students have felt more cheerful on a scale of 1-6 (At no Time - All of the Time).

Average monthly expenditure (Own Pocket) - Positive coefficient signifies that satisfaction level increases when student’s expenditure increases. This is a bit out of the box and we need to re-validate this outcome when we derive final conclusions after comparing other models.

### SVM

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 0 0 26  
## 2 0 0 15  
## 3 0 0 159  
##   
## Overall Statistics  
##   
## Accuracy : 0.795   
## 95% CI : (0.7323, 0.8487)  
## No Information Rate : 1   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity NA NA 0.795  
## Specificity 0.87 0.925 NA  
## Pos Pred Value NA NA NA  
## Neg Pred Value NA NA NA  
## Prevalence 0.00 0.000 1.000  
## Detection Rate 0.00 0.000 0.795  
## Detection Prevalence 0.13 0.075 0.795  
## Balanced Accuracy NA NA NA

## [1] NaN

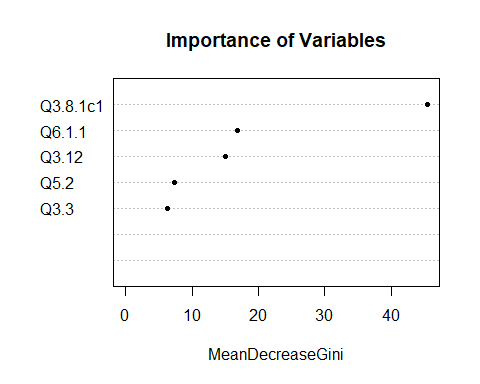
Multi-category SVM algorithm from the e1071 package is used for analysis. SVM fits the test data correctly and misclassified only 20% observations. In this case, Group 1 and 2 are misclassified 100% of the times. Hence it is not a good model for satisfaction prediction with accommodation for desired reasons.

### Random Forest

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 1 1 24  
## 2 0 0 15  
## 3 2 1 156  
##   
## Overall Statistics  
##   
## Accuracy : 0.785   
## 95% CI : (0.7215, 0.8398)  
## No Information Rate : 0.975   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0323   
##   
## Mcnemar's Test P-Value : 5.587e-07   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.33333 0.0000 0.80000  
## Specificity 0.87310 0.9242 0.40000  
## Pos Pred Value 0.03846 0.0000 0.98113  
## Neg Pred Value 0.98851 0.9892 0.04878  
## Prevalence 0.01500 0.0100 0.97500  
## Detection Rate 0.00500 0.0000 0.78000  
## Detection Prevalence 0.13000 0.0750 0.79500  
## Balanced Accuracy 0.60321 0.4621 0.60000

## [1] 0.06896552

Random Forest fits the test data correctly and misclassified only 21% observations. In this case, Group 2 is misclassified 100% of the times. Overall accuracy is 79% and it does a great job in predicting “Unsatisfied” and “Satisfied” student categories for satisfaction with the accommodation.



“Importance ranking” for predictors is also shown in the above plot.

### Linear Discriminant Analysis

LDA assumes that the predictors are normally distributed. As stated earlier that we have data that is not normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.

## Call:  
## lda(factor(Q6.2.1) ~ Q3.3 + Q3.12 + Q3.8.1c1 + Q6.1.1 + Q5.2,   
## data = Train2\_n)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.11056511 0.08353808 0.80589681   
##   
## Group means:  
## Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 1 1.200000 3.888889 555.8222 3.400000 1.533333  
## 2 1.500000 3.617647 588.8706 3.382353 1.352941  
## 3 1.432927 3.134146 858.4009 4.006098 1.371951  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Q3.3 0.6195383091 -1.6449756664  
## Q3.12 -0.4400635520 -0.0832476584  
## Q3.8.1c1 0.0005683843 0.0002859345  
## Q6.1.1 0.3892727280 0.3536452387  
## Q5.2 -0.8784316218 0.8608864115  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.8511 0.1489

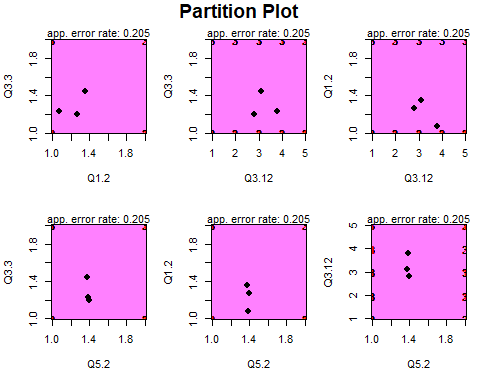
##   
## 1 2 3   
## 45 34 328

There are 407 total observations, 45 in class 1, 34 in class 2 and 328 in class 3. LD1 is the first linear discriminant which is a linear combination of below parameters: -

(Q3.32\* 0.4182167604 + Q3.12*-0.4737921525 + Q3.8.1c1* 0.0005513997 + Q6.1.1\* 0.3628719222 + Q5.22\*-1.0241931415)

Q5.22, Q3.32, and Q3.12 have the largest coefficients, hence these are the most significant predictors in influencing the response.

Q3.12 and Q5.22 have negative coefficients and the rest of the parameters have positive coefficients, this resonates the result from linear regression for predicting student satisfaction with the accommodation. LD1 explains 94.72% of the overall variance in the data.



In this case, all the 3 Groups have overlapping boundaries and hence the model predicts all the students to be in group 3 “Satisfied” with the accommodation. It is eventually difficult to separate the classes and predict unless we have balanced data.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 1 0 25  
## 2 0 0 15  
## 3 1 0 158  
##   
## Overall Statistics  
##   
## Accuracy : 0.795   
## 95% CI : (0.7323, 0.8487)  
## No Information Rate : 0.99   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0314   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.50000 NA 0.79798  
## Specificity 0.87374 0.925 0.50000  
## Pos Pred Value 0.03846 NA 0.99371  
## Neg Pred Value 0.99425 NA 0.02439  
## Prevalence 0.01000 0.000 0.99000  
## Detection Rate 0.00500 0.000 0.79000  
## Detection Prevalence 0.13000 0.075 0.79500  
## Balanced Accuracy 0.68687 NA 0.64899

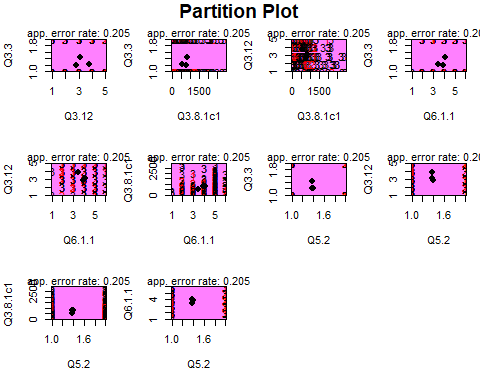
## [1] 0.07142857

This model fit the test data correctly and misclassified only 20% observations. But it doesn’t predict group 2 correctly 100% of times as there is a complete overlap between the adjacent classes, hence it is very difficult to predict if a student is “Neutral” in terms of satisfaction with the accommodation. The model accuracy is 80% on the test data.

### Quadratic Discriminant Analysis

## Length Class Mode   
## prior 3 -none- numeric   
## counts 3 -none- numeric   
## means 15 -none- numeric   
## scaling 75 -none- numeric   
## ldet 3 -none- numeric   
## lev 3 -none- character  
## N 1 -none- numeric   
## call 3 -none- call   
## terms 3 terms call   
## xlevels 0 -none- list

## Call:  
## qda(factor(Q6.2.1) ~ Q3.3 + Q3.12 + Q3.8.1c1 + Q6.1.1 + Q5.2,   
## data = Train2\_n)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.11056511 0.08353808 0.80589681   
##   
## Group means:  
## Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 1 1.200000 3.888889 555.8222 3.400000 1.533333  
## 2 1.500000 3.617647 588.8706 3.382353 1.352941  
## 3 1.432927 3.134146 858.4009 4.006098 1.371951



Again, the model predicts most of the students to be in group 3 “Satisfied” with the accommodation with indistinguishable boundaries with its neighbors.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 2 0 24  
## 2 0 0 15  
## 3 2 2 155  
##   
## Overall Statistics  
##   
## Accuracy : 0.785   
## 95% CI : (0.7215, 0.8398)  
## No Information Rate : 0.97   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0466   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.50000 0.0000 0.79897  
## Specificity 0.87755 0.9242 0.33333  
## Pos Pred Value 0.07692 0.0000 0.97484  
## Neg Pred Value 0.98851 0.9892 0.04878  
## Prevalence 0.02000 0.0100 0.97000  
## Detection Rate 0.01000 0.0000 0.77500  
## Detection Prevalence 0.13000 0.0750 0.79500  
## Balanced Accuracy 0.68878 0.4621 0.56615

## [1] 0.1333333

There are 407 observations, 43 in class 1, 33 in class 2 and 331 in class 3. QDA model assists in predicting the test data correctly but misclassified slightly more that LDA with 23% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it predicts student satisfaction accommodation with an accuracy of 77%.

### Multinomial Logistic Regression

Now once we have established the final set of predictors using the above chapter, we can analyze their relationship with the response.

## # weights: 21 (12 variable)  
## initial value 447.135201   
## iter 10 value 249.593014  
## iter 20 value 226.760727  
## final value 226.760237   
## converged

## Call:  
## multinom(formula = factor(Q6.2.1) ~ Q3.3 + Q3.12 + Q3.8.1c1 +   
## Q6.1.1 + Q5.2, data = Train2\_n)  
##   
## Coefficients:  
## (Intercept) Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 2 -0.2250486 1.457694 -0.2058391 0.0001881674 -0.0350865 -0.8485874  
## 3 1.9371312 1.107488 -0.5108312 0.0006969825 0.3191668 -0.8651455  
##   
## Std. Errors:  
## (Intercept) Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 2 0.3988785 0.4542553 0.1802206 0.0004100711 0.1645992 0.4353774  
## 3 0.6197189 0.3754109 0.1425845 0.0003111711 0.1247624 0.3177048  
##   
## Residual Deviance: 453.5205   
## AIC: 477.5205

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 1 0 25  
## 2 0 0 15  
## 3 1 0 158  
##   
## Overall Statistics  
##   
## Accuracy : 0.795   
## 95% CI : (0.7323, 0.8487)  
## No Information Rate : 0.99   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0314   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.50000 NA 0.79798  
## Specificity 0.87374 0.925 0.50000  
## Pos Pred Value 0.03846 NA 0.99371  
## Neg Pred Value 0.99425 NA 0.02439  
## Prevalence 0.01000 0.000 0.99000  
## Detection Rate 0.00500 0.000 0.79000  
## Detection Prevalence 0.13000 0.075 0.79500  
## Balanced Accuracy 0.68687 NA 0.64899

## [1] 0.07142857

## (Intercept) Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 2 0.572615721 0.001332089 0.2533913189 0.64633088 0.83119970 0.051285341  
## 3 0.001773087 0.003177032 0.0003401195 0.02509946 0.01052164 0.006467026

## (Intercept) Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2  
## 2 0.7984774 4.296041 0.8139640 1.000188 0.9655219 0.4280191  
## 3 6.9388160 3.026746 0.5999967 1.000697 1.3759808 0.4209903

Neutral vs Unsatisfied Students

With P-values (< 0.05) only Q3.12 is significant. For a one-level increase in perceived financial difficulties on a scale 1-5 (No Difficulty to Serious Difficulties), we expect to see a 53% (OR = 0.53) decrease in the odds of being Neutral vs Unsatisfied in terms of satisfaction with the accommodation. There exists an inverse relationship between satisfaction with accommodation and level of financial difficulty, which seems logical.

Satisfied vs Unsatisfied Students With P-values (< 0.05) predictors Q5.22, Q6.1.1, Q3.12, and Q3.8.1c1 are significant. For understanding the relationships, we need to analyze the odds ratio for all the predictors as below: -

Q3.12- We expect to see 53% (OR = 0.53) decrease in the odds of being Satisfied vs Unsatisfied for one level increase in perceived financial difficulties on a scale of 1-5.

Q5.22 - For Male vs Female, we expect to see 37% (OR = 0.53) decrease in the odds of being Satisfied vs Unsatisfied with accommodation. Female students are more satisfied with their accommodation comparatively.

Q6.1.1 - We expect to see 43% (OR = 1.43) increase in the odds of being Satisfied vs Unsatisfied for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All of the Time). Hence, cheerful students are almost twice as more satisfied with their accommodation compared to other students.

Q3.8.1c1 - With the increase in average monthly expenditure (Own Pocket in Euros) on a nominal scale, we expect to see 0.1% (OR = 1.001) increase in the odds of being Satisfied vs Unsatisfied with accommodation. This may be due to the fact that students with higher expenditure may be paying higher rents as well for better houses, in turn leading to more satisfaction with the accommodation.

### Ordinal regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.

## Call:  
## polr(formula = factor(Q6.2.1) ~ Q3.3 + Q3.12 + Q3.8.1c1 + Q6.1.1 +   
## Q5.2, data = Train2\_n, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## Q3.3 0.5135772 0.2774063 1.851  
## Q3.12 -0.4395595 0.1241446 -3.541  
## Q3.8.1c1 0.0006393 0.0002048 3.122  
## Q6.1.1 0.3331046 0.1033509 3.223  
## Q5.2 -0.5898395 0.2672243 -2.207  
##   
## Intercepts:  
## Value Std. Error t value  
## 1|2 -2.1201 0.8394 -2.5258  
## 2|3 -1.3896 0.8322 -1.6698  
##   
## Residual Deviance: 461.7868   
## AIC: 475.7868

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 1 0 25  
## 2 0 0 15  
## 3 1 0 158  
##   
## Overall Statistics  
##   
## Accuracy : 0.795   
## 95% CI : (0.7323, 0.8487)  
## No Information Rate : 0.99   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0314   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.50000 NA 0.79798  
## Specificity 0.87374 0.925 0.50000  
## Pos Pred Value 0.03846 NA 0.99371  
## Neg Pred Value 0.99425 NA 0.02439  
## Prevalence 0.01000 0.000 0.99000  
## Detection Rate 0.00500 0.000 0.79000  
## Detection Prevalence 0.13000 0.075 0.79500  
## Balanced Accuracy 0.68687 NA 0.64899

## [1] 0.07142857

## Value Std. Error t value  
## Q3.3 0.513577169 0.277406341 1.851353  
## Q3.12 -0.439559540 0.124144641 -3.540705  
## Q3.8.1c1 0.000639332 0.000204792 3.121860  
## Q6.1.1 0.333104645 0.103350872 3.223046  
## Q5.2 -0.589839482 0.267224278 -2.207283  
## 1|2 -2.120134651 0.839404124 -2.525762  
## 2|3 -1.389638469 0.832233857 -1.669769

## Value Std. Error t value p value  
## Q3.3 0.513577169 0.277406341 1.851353 0.0641187309  
## Q3.12 -0.439559540 0.124144641 -3.540705 0.0003990596  
## Q3.8.1c1 0.000639332 0.000204792 3.121860 0.0017971266  
## Q6.1.1 0.333104645 0.103350872 3.223046 0.0012683506  
## Q5.2 -0.589839482 0.267224278 -2.207283 0.0272943265  
## 1|2 -2.120134651 0.839404124 -2.525762 0.0115447814  
## 2|3 -1.389638469 0.832233857 -1.669769 0.0949650346

## Q3.3 Q3.12 Q3.8.1c1 Q6.1.1 Q5.2   
## 1.6712589 0.6443202 1.0006395 1.3952933 0.5544163

With P-values (< 0.05) predictors Q5.22, Q6.1.1, Q3.12 and Q3.8.1c1 are significant. [Figure 33]. We expect to see 70% (OR = 0.70) decrease in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in perceived financial difficulties (Q3.12) on a scale of 1-5. Due to the property of proportional odds assumption, the same decrease, 0.7 times, is deduced between Satisfied vs Neutral or Unsatisfied. Similarly, odds of being Unsatisfied vs Neutral or Satisfied for other predictors can be analyzed as below: -

Q5.22 - For Male vs Female, we expect to see 46% (OR = 0.46) decrease. Again, this suggests that female counterparts are comparatively more satisfied with their accommodation.

Q6.1.1 - We expect to see 37% (OR = 1.37) increase for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All of the Time). Hence, students who haven’t felt in good spirits (over the last two weeks) are less satisfied with their accommodation.

Q3.8.1c1 - With an increase in average monthly expenditure (Own Pocket in Euros) on a nominal scale, we expect to see a 0.1% (OR = 1.001) increase.

Because of the proportional odds assumption, the similar results apply for predicting Satisfied vs Neutral or Unsatisfied students.

## 3. Satisfaction with Financial/Material wellbeing (Response)

### Linear Models

In this thesis, I first check the associations with the utilization of classical regression analysis. This methodology gives us an initial understanding of dependent and independent factors. But in our case the data is not normally distributed and has a non-linear shape, thus findings of the analysis are indecisive and not used in results formulation and model comparison at a later stage.

##   
## Call:  
## lm(formula = Q6.2.2 ~ Q3.12 + Q6.1.1, data = Train3\_n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6223 -0.4362 -0.1021 0.3777 1.8170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.06601 0.14346 21.372 < 2e-16 \*\*\*  
## Q3.12 -0.47977 0.02574 -18.638 < 2e-16 \*\*\*  
## Q6.1.1 0.10316 0.02588 3.986 7.99e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6395 on 404 degrees of freedom  
## Multiple R-squared: 0.4944, Adjusted R-squared: 0.4919   
## F-statistic: 197.6 on 2 and 404 DF, p-value: < 2.2e-16

Students perceived satisfaction with financial and material wellbeing is derived from predictors (P-Value<0.05).

Perceived level of financial difficulties - Negative coefficient sign suggests that satisfaction level decreases for students facing higher levels of financial difficulty.

Feeling of being in good spirits and cheerful (over last two weeks) - Positive coefficient suggests that satisfaction level rises for students have felt more cheerful over last two weeks on a scale of 1-6 (At no Time - All of the Time).

### SVM

##   
## Call:  
## svm(formula = factor(Q6.2.2) ~ Q3.12 + Q6.1.1, data = Train3\_n)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 1   
## gamma: 0.5   
##   
## Number of Support Vectors: 238  
##   
## ( 88 79 71 )  
##   
##   
## Number of Classes: 3   
##   
## Levels:   
## 1 2 3

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 77 5 16  
## 2 11 1 15  
## 3 11 5 59  
##   
## Overall Statistics  
##   
## Accuracy : 0.685   
## 95% CI : (0.6157, 0.7487)  
## No Information Rate : 0.495   
## P-Value [Acc > NIR] : 4.163e-08   
##   
## Kappa : 0.4581   
##   
## Mcnemar's Test P-Value : 0.04251   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7778 0.09091 0.6556  
## Specificity 0.7921 0.86243 0.8545  
## Pos Pred Value 0.7857 0.03704 0.7867  
## Neg Pred Value 0.7843 0.94220 0.7520  
## Prevalence 0.4950 0.05500 0.4500  
## Detection Rate 0.3850 0.00500 0.2950  
## Detection Prevalence 0.4900 0.13500 0.3750  
## Balanced Accuracy 0.7849 0.47667 0.7551

## [1] 0.7817259

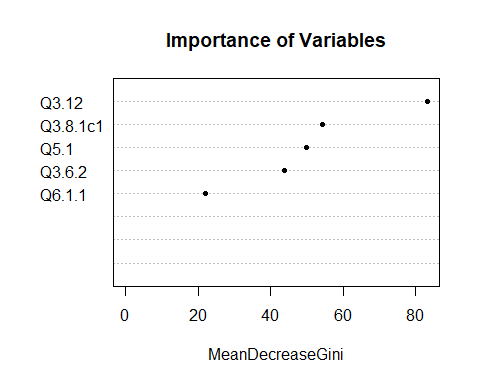
Multi-category SVM algorithm from the e1071 package (7) is used for analysis. SVM fits the test data correctly and misclassified 26% observations. Again, in this case, Group 2 (“Neutral” in terms of satisfaction with wellbeing) is misclassified 100% of the times. Overall accuracy is 74% and it does a reasonable job in predicting “Unsatisfied” and “Satisfied” students.

### Random Forest

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 83 7 8  
## 2 11 2 14  
## 3 13 10 52  
##   
## Overall Statistics  
##   
## Accuracy : 0.685   
## 95% CI : (0.6157, 0.7487)  
## No Information Rate : 0.535   
## P-Value [Acc > NIR] : 1.112e-05   
##   
## Kappa : 0.4627   
##   
## Mcnemar's Test P-Value : 0.4325   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7757 0.10526 0.7027  
## Specificity 0.8387 0.86188 0.8175  
## Pos Pred Value 0.8469 0.07407 0.6933  
## Neg Pred Value 0.7647 0.90173 0.8240  
## Prevalence 0.5350 0.09500 0.3700  
## Detection Rate 0.4150 0.01000 0.2600  
## Detection Prevalence 0.4900 0.13500 0.3750  
## Balanced Accuracy 0.8072 0.48357 0.7601

## [1] 0.8097561

Random Forest fits the test data correctly and misclassified more than a quarter (27%) of observations. This time it does not completely misclassify all observations in Group 2 (“Neutral” satisfaction. Overall accuracy is 73% and it does a fair job in predicting all classes of students for satisfaction with wellbeing.



“Importance ranking” for predictors is also shown in the above plot.

### Linear Discriminant Analysis

LDA assumes that the predictors are normally distributed. As stated earlier that we have data that is not normally distributed, but we are still interested in the prediction accuracy, hence we implement LDA.

## Call:  
## lda(factor(Q6.2.2) ~ Q3.12 + Q6.1.1, data = Train3\_n)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.4299754 0.1941032 0.3759214   
##   
## Group means:  
## Q3.12 Q6.1.1  
## 1 4.091429 3.571429  
## 2 3.101266 4.113924  
## 3 2.176471 4.241830  
##   
## Coefficients of linear discriminants:  
## LD1 LD2  
## Q3.12 -1.0663369 -0.2839222  
## Q6.1.1 0.2328299 -0.7980674  
##   
## Proportion of trace:  
## LD1 LD2   
## 0.9961 0.0039

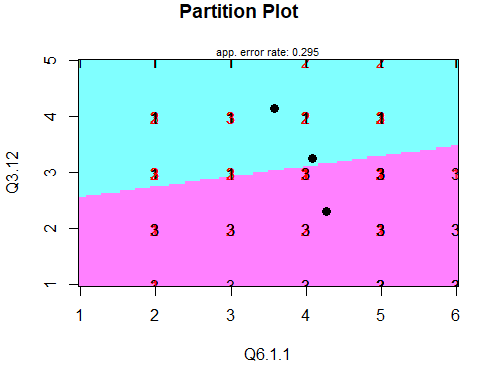
##   
## 1 2 3   
## 175 79 153

There are 407 total observations, 181 in class 1, 73 in class 2 and 153 in class 3.

LD1 is the first linear discriminant which is a linear combination of below parameters: -

(Q3.12*-1.0395162 + Q6.1.1* 0.2145408)

Q3.12 has negative coefficient and Q6.1.1 has a positive coefficient, this is similar to the results obtained from linear regression. LD1 explains almost all (99.58%) of the variability in the data.



Group 1 and 3 are fairly distinguishable, but Group 2 has complete overlapping with neighboring boundaries and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the wellbeing.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 82 0 16  
## 2 12 0 15  
## 3 16 0 59  
##   
## Overall Statistics  
##   
## Accuracy : 0.705   
## 95% CI : (0.6366, 0.7672)  
## No Information Rate : 0.55   
## P-Value [Acc > NIR] : 5.073e-06   
##   
## Kappa : 0.4749   
##   
## Mcnemar's Test P-Value : 5.887e-06   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7455 NA 0.6556  
## Specificity 0.8222 0.865 0.8545  
## Pos Pred Value 0.8367 NA 0.7867  
## Neg Pred Value 0.7255 NA 0.7520  
## Prevalence 0.5500 0.000 0.4500  
## Detection Rate 0.4100 0.000 0.2950  
## Detection Prevalence 0.4900 0.135 0.3750  
## Balanced Accuracy 0.7838 NA 0.7551

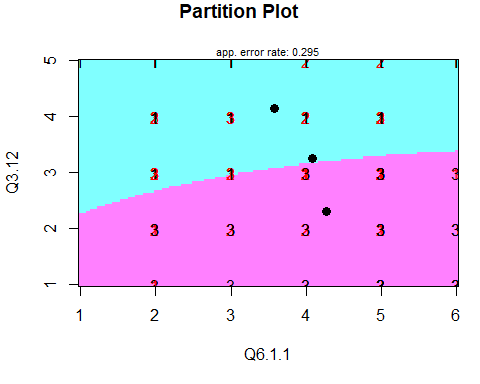
## [1] 0.7884615

This model fit the test data correctly and misclassified 26% observations. But it doesn’t predict group 2 correctly 100% of times, as there is a complete overlap between the adjacent classes. The model has reasonably well accuracy on the test data of 74% for predicting of student satisfaction with the wellbeing.

### Quadratic Discriminant Analysis

## Call:  
## qda(factor(Q6.2.2) ~ Q3.12 + Q6.1.1, data = Train3\_n)  
##   
## Prior probabilities of groups:  
## 1 2 3   
## 0.4299754 0.1941032 0.3759214   
##   
## Group means:  
## Q3.12 Q6.1.1  
## 1 4.091429 3.571429  
## 2 3.101266 4.113924  
## 3 2.176471 4.241830

##   
## 1 2 3   
## 175 79 153

 Group 1 “Unsatisfied” or Group 3 “Satisfied”, but Group 2 “Neutral” has an overlapping boundary with remaining groups and hence it is difficult to correctly predict if a student is “Neutral” in terms of satisfaction with the wellbeing.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 82 0 16  
## 2 12 0 15  
## 3 16 0 59  
##   
## Overall Statistics  
##   
## Accuracy : 0.705   
## 95% CI : (0.6366, 0.7672)  
## No Information Rate : 0.55   
## P-Value [Acc > NIR] : 5.073e-06   
##   
## Kappa : 0.4749   
##   
## Mcnemar's Test P-Value : 5.887e-06   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7455 NA 0.6556  
## Specificity 0.8222 0.865 0.8545  
## Pos Pred Value 0.8367 NA 0.7867  
## Neg Pred Value 0.7255 NA 0.7520  
## Prevalence 0.5500 0.000 0.4500  
## Detection Rate 0.4100 0.000 0.2950  
## Detection Prevalence 0.4900 0.135 0.3750  
## Balanced Accuracy 0.7838 NA 0.7551

## [1] 0.7884615

There are 451 observations, 43 in class 1, 42 in class 2 and 366 in class 3.QDA performs same as LDA, in this case, misclassified 26% observations. Again Group 2 (“Neutral” in terms of satisfaction with the college) is incorrectly predicted 100% times. Overall it predicts student satisfaction with wellbeing with an accuracy of 74%.

### Multinomial Logistic Regression

Now once we have established the final set of predictors using the above chapter, we can analyze their relationship with the response.

## # weights: 12 (6 variable)  
## initial value 447.135201   
## iter 10 value 294.826696  
## final value 294.756613   
## converged

## Call:  
## multinom(formula = factor(Q6.2.2) ~ Q3.12 + Q6.1.1, data = Train3\_n)  
##   
## Coefficients:  
## (Intercept) Q3.12 Q6.1.1  
## 2 2.241428 -1.258175 0.3988026  
## 3 5.011590 -2.193698 0.4851124  
##   
## Std. Errors:  
## (Intercept) Q3.12 Q6.1.1  
## 2 0.7868669 0.1827338 0.1263246  
## 3 0.8104319 0.2073274 0.1324503  
##   
## Residual Deviance: 589.5132   
## AIC: 601.5132

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 82 0 16  
## 2 12 0 15  
## 3 16 0 59  
##   
## Overall Statistics  
##   
## Accuracy : 0.705   
## 95% CI : (0.6366, 0.7672)  
## No Information Rate : 0.55   
## P-Value [Acc > NIR] : 5.073e-06   
##   
## Kappa : 0.4749   
##   
## Mcnemar's Test P-Value : 5.887e-06   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7455 NA 0.6556  
## Specificity 0.8222 0.865 0.8545  
## Pos Pred Value 0.8367 NA 0.7867  
## Neg Pred Value 0.7255 NA 0.7520  
## Prevalence 0.5500 0.000 0.4500  
## Detection Rate 0.4100 0.000 0.2950  
## Detection Prevalence 0.4900 0.135 0.3750  
## Balanced Accuracy 0.7838 NA 0.7551

## [1] 0.7884615

## (Intercept) Q3.12 Q6.1.1  
## 2 4.39192e-03 5.767121e-12 0.0015941961  
## 3 6.25562e-10 3.656699e-26 0.0002496685

## (Intercept) Q3.12 Q6.1.1  
## 2 9.406758 0.2841722 1.490039  
## 3 150.143312 0.1115037 1.624358

Neutral vs Unsatisfied Students - With P-values (< 0.05) significant parameters are.

Q3.12 - For a one-level increase in perceived financial difficulties on a scale 1-5 (No - Serious Difficulties), we expect to see a 32% (OR = 0.32) decrease in the odds in terms of satisfaction with wellbeing.

Q6.1.1 - We expect to see 48% (OR = 1.48) increase in the odds for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All of the Time).

Satisfied vs Unsatisfied Students - With P-values (< 0.05) significant parameters are.

Q3.12- We expect to see 14% (OR = 0.14) decrease in the odds for one level increase in perceived financial difficulties on a scale of 1-5. There exists an inverse relationship with financial difficulty and satisfaction with wellbeing.

Q6.1.1 - We expect to see 53% (OR = 1.53) increase in the odds for one level increase in perceived feeling more cheerful on a scale of 1-6 (At no Time - All of the Time). Hence, satisfaction with accommodation increases if the student has felt more cheerful over the last couple of weeks.

### Ordinal regression

After having considered the nominal approach, we will now discuss the ordinal approach that helps to answer another interesting question that how the satisfaction responses vary when ranked from 1-3.

## Call:  
## polr(formula = factor(Q6.2.2) ~ Q3.12 + Q6.1.1, data = Train3\_n,   
## Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## Q3.12 -1.5885 0.13153 -12.078  
## Q6.1.1 0.3457 0.09233 3.744  
##   
## Intercepts:  
## Value Std. Error t value   
## 1|2 -4.2372 0.5682 -7.4572  
## 2|3 -2.8267 0.5417 -5.2184  
##   
## Residual Deviance: 588.4668   
## AIC: 596.4668

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 82 0 16  
## 2 12 0 15  
## 3 16 0 59  
##   
## Overall Statistics  
##   
## Accuracy : 0.705   
## 95% CI : (0.6366, 0.7672)  
## No Information Rate : 0.55   
## P-Value [Acc > NIR] : 5.073e-06   
##   
## Kappa : 0.4749   
##   
## Mcnemar's Test P-Value : 5.887e-06   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.7455 NA 0.6556  
## Specificity 0.8222 0.865 0.8545  
## Pos Pred Value 0.8367 NA 0.7867  
## Neg Pred Value 0.7255 NA 0.7520  
## Prevalence 0.5500 0.000 0.4500  
## Detection Rate 0.4100 0.000 0.2950  
## Detection Prevalence 0.4900 0.135 0.3750  
## Balanced Accuracy 0.7838 NA 0.7551

## [1] 0.7884615

## Value Std. Error t value  
## Q3.12 -1.5885431 0.13152676 -12.077718  
## Q6.1.1 0.3457018 0.09232917 3.744232  
## 1|2 -4.2371540 0.56819353 -7.457237  
## 2|3 -2.8267077 0.54168255 -5.218384

## Value Std. Error t value p value  
## Q3.12 -1.5885431 0.13152676 -12.077718 1.385099e-33  
## Q6.1.1 0.3457018 0.09232917 3.744232 1.809467e-04  
## 1|2 -4.2371540 0.56819353 -7.457237 8.835548e-14  
## 2|3 -2.8267077 0.54168255 -5.218384 1.804906e-07

## Q3.12 Q6.1.1   
## 0.2042229 1.4129812

With P-values (< 0.05) both the predictors are considered significant.

Q3.12 - We expect to see 23% (OR = 0.23) decrease in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in perceived financial difficulties on a scale of 1-5. There exists an inverse relationship with financial difficulty and satisfaction with wellbeing.

Q6.1.1 - We expect to see 41% (OR = 1.41) increase in the odds of moving from Unsatisfied to Neutral or Satisfied for one level increase in the perceived feeling of being cheerful on a scale of 1-6 (At no Time - All of the Time). Hence, satisfaction with accommodation increases if a student has felt more cheerful over the last couple of weeks.

Because of the proportional odds assumption, similar results apply for predicting Satisfied vs Neutral or Unsatisfied students.