**Business Problem:**

You need to predict the type of program a student is in based on other attributes --> mdata.csv

**Data:**

Program: is a categorical variable indicating what type of program a student is in: It is output feature.

1. General
2. Academic
3. Vocational

Ses: is a categorical variable indicating someone’s socioeconomic class: Categorical data so convert them into 3 features.

1. Low
2. Middle
3. High

Read, write, math and science are their scores on different tests.

Honours: Whether they have enrolled or not. It is also categorical data so convert them into 2 features.

Female features are also categorical so converted them into 2 features.

Data doesn’t have NA and Outliers.

**Building the Model:**

Considered all the features except few dummy variables due to degree of freedom.

pgm.prog <- multinom(prog ~ pgm$fem+pgm$ses\_low+pgm$ses\_mid+pgm$schtyp\_public+

pgm$honors\_enroll+pgm$read+pgm$write+pgm$math+pgm$science, data=pgm)

Taking baseline as general.

pgm$prog <- relevel(pgm$prog, ref= "general")

**Predicted Results:** From Cross Table.

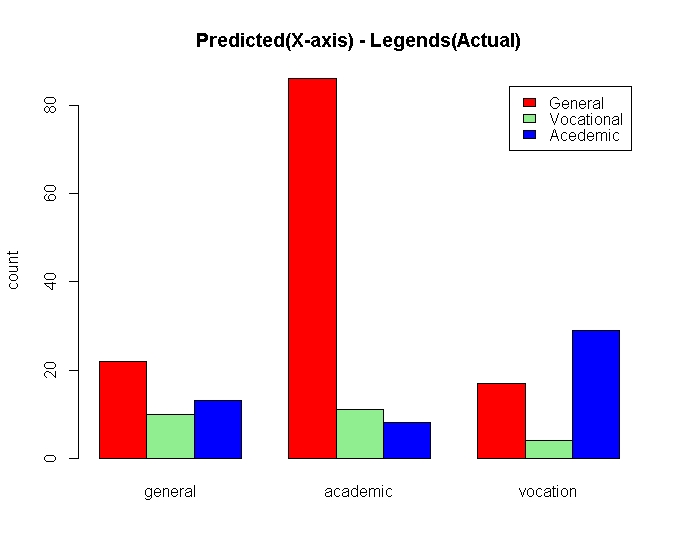
Pred\_name general academic vocation

academic 22 86 17

general 10 11 4

vocation 13 8 29

Visualization:



Probability values are more than accepted values for few features but we have to discuss with Subject matter expert ( SME) to delete them as of now, I’m keeping all of them.

Accuracy is 62.5% when we taken “general” as baseline.

**Checking with other base line: vocation**

Pred\_name vocation general academic

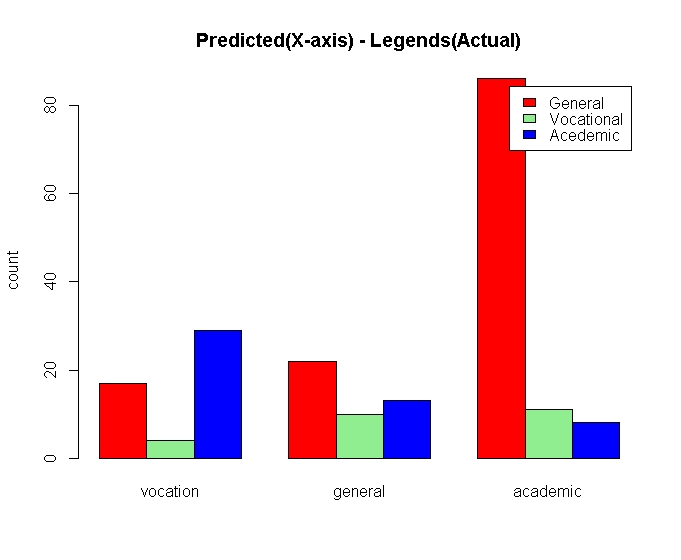
academic 17 22 86

general 4 10 11

vocation 29 13 8

Accuracy is 62.5% when we taken “vocation” as baseline. There was no difference in Predicted values and accuracy percentage.

Visualization:



**Checking with other base line: academic**

**Predicted values from cross table.**

Pred\_name academic vocation general

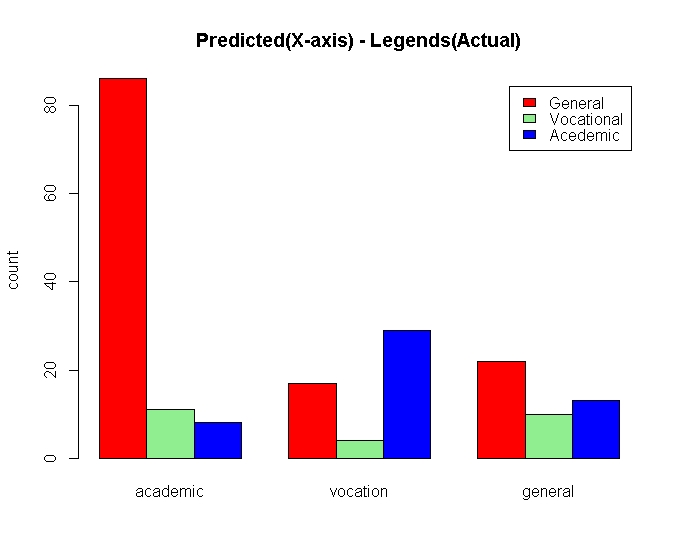
academic 86 17 22

general 11 4 10

vocation 8 29 13

Accuracy is 62.5% when we taken “vocation” as baseline. There was no difference in Predicted values and accuracy percentage.

**Visualization:**

****

Conclusion:

Need some more data to improve the performance and also need SME support to delete uncorrelated features, there are few features have the more than the accepted probability value.