International Institute of Information Technology, Bangalore

Analytics On A Subset Of The 2011 SSLC Data Set Report

Abhijith Madhav Saumya Tayal Nikita Shah

December 3, 2014

Contents

1 Data Preparation And Characterization

1.1 Data Preparation

Prior preparation of data using a spreadsheet software

1. Have removed the '*'s in all of the marks column.
2. Absentees have been given 0 marks in the respective subjects replacing the 888 marker.
3. Correcting totalling errors for about 21 records.
4. Added 'class' fields for each of the marks by using the following discretization. All marks attributes were scaled to 100.
   - < 35 : FAIL
   - < 50 : PASS
   - < 60 : 2nd class
   - < 85 : 1st class
   - > 84 : Distinction
5.

Note : The following columns have the said number of rows with NA data. Shouldn't really matter as they are not being used in any of the experiments

- DOB = 1
- NRC_MOTHER_NAME = 42
- NRC_FATHER_NAME = 32
- L1_RESULT = 2
- L2_RESULT = 32
- L3_RESULT = 38

1.2 Data Set Characterization

1.3 Code

`https://github.com/AbhijithMadhav/SSLC-Data-Analysis`

2 Discretization + Classification

Objective

To build a classification model for predicting the class of a student based on his course marks.

Procedure

- Marks of all languages and subjects have already been discretized as in /ref
- Built a classification model using a decision tree algorithm with L1_CLASS, L2_CLASS, L3_CLASS, S1_CLASS, S2_CLASS, S3_CLASS as predictors.
  - Proportion of sizes of the training and test data sets were respectively 2/3 and 1/3.
  - Data records themselves were put into both the above data sets using random sampling.

Observations

- Tabulation of the count of the predicted versus actual values for the test data set

```
{
       pred
true      1    2 D    FAIL PASS
  1    2717  299   50    1    9
  2     613  920    0   23  350
  D     212    1  274    0    0
  FAIL    9   49    0 1894  558
  PASS   58  458    0  424 2081
}
```

- Misclassification error

```
0.2830909
```

- Decision tree generated

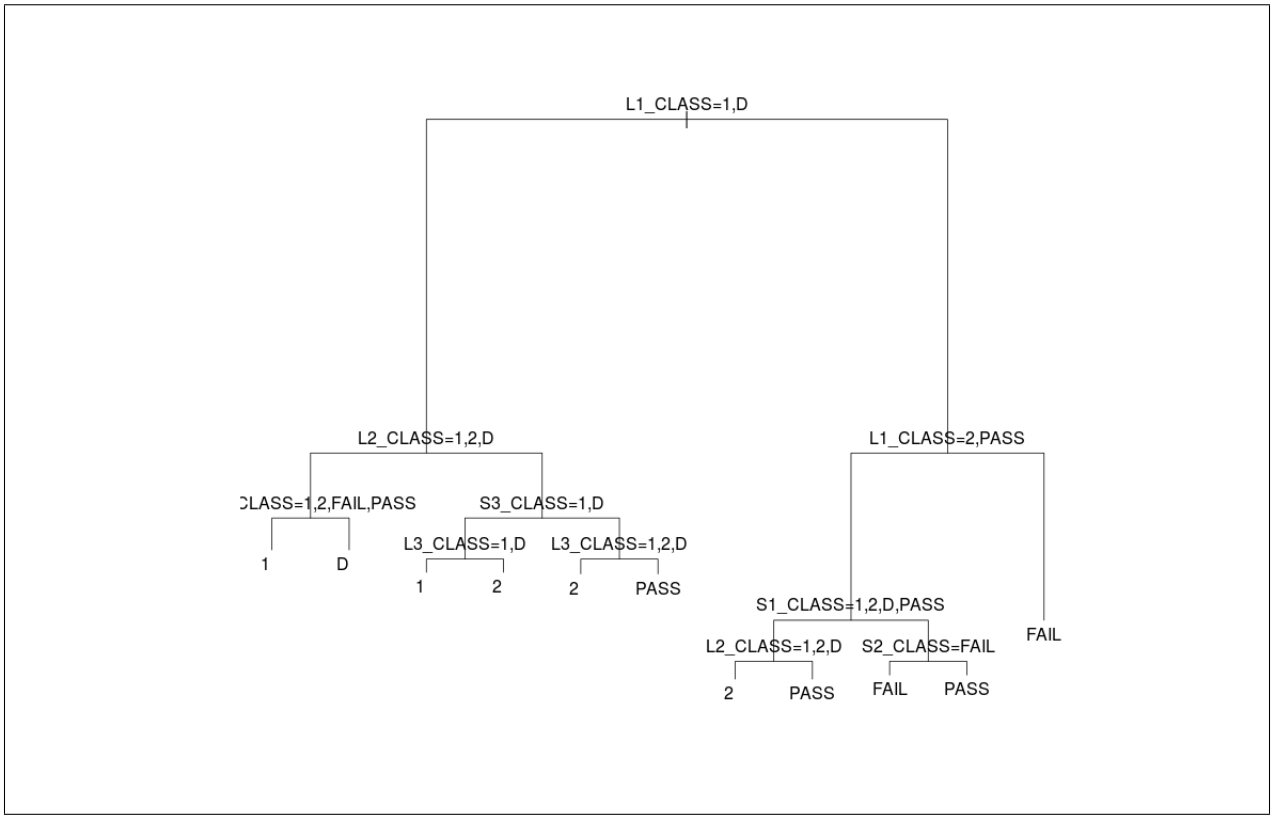


Figure 2.1: Decision Tree

Conclusions

TODO

3 Regression + Classification

Objective

Classification models using a lot of predictors give a model which might be complex. The objective thus is to try to use a simple linear regression model to determine the least number of attributes which affect the class to be predicted and then construct classification models.

Here the marks of the students are used to predict the NRC_CLASS variable using a classification model.

Procedure

- Marks of all languages and subjects have already been discretized as in /refchap:chapter1.
- Typically a linear regression model with more number of predictors gives a better fitted model. Thus for every count of the number of courses, from 1 to 5(not including all the courses), the best combination of predictors are chosen.
  - For a number n,the best combination of predictors are found building regression models for all possible combination of attributes and the standard deviation of their residuals.
- From each of these best combination of predictors a classification model is built.
  - The classification model chosen is KNN as the predictors are numeric attributes. k was chosen to be 10 after manual inspection with several values.
- The accuracy of each of these classification models is compared to find out the best classification model.

Observations

- Best predictors obtained

```
> best_models
[[1]] # For n = 1
lm(formula = TOTAL_MARKS ~ S2_MARKS)

[[2]] # For n = 2
```

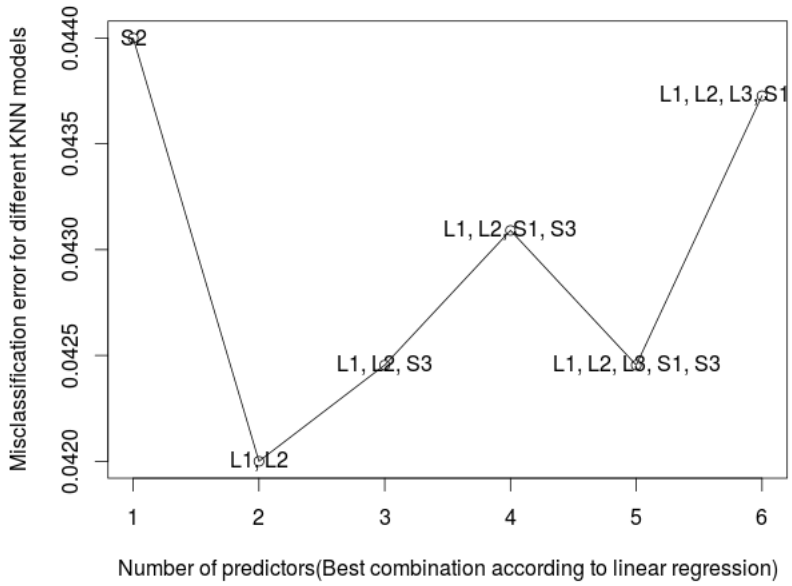
```
lm(formula = TOTAL_MARKS ~ L1_MARKS + L2_MARKS)

[[3]] # For n = 3
lm(formula = TOTAL_MARKS ~ L1_MARKS + L2_MARKS + S3_MARKS)

[[4]] # For n = 4
lm(formula = TOTAL_MARKS ~ L1_MARKS + L2_MARKS + S1_MARKS + S3_MARKS)

[[5]] # For n = 5
lm(formula = TOTAL_MARKS ~ L1_MARKS + L2_MARKS + L3_MARKS + S1_MARKS
                           + S3_MARKS)
```

- Comparision of misclassification errors for different number of predictors(Best combination)



Conclusions

- Marks of the first two languages(Typically Kannada and English) are the most accurate predictors of the NRC_CLASS of a student.

4 Clustering + Association Rules

Objective

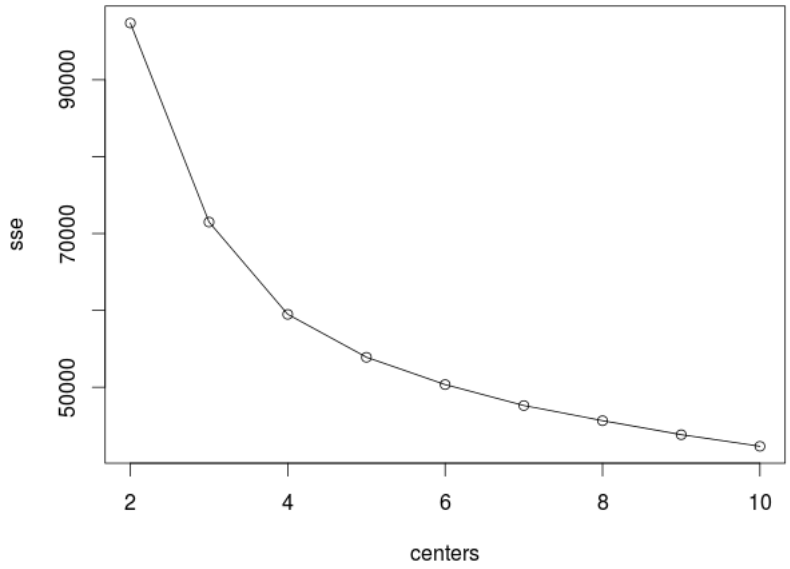
Cluster characterization is most difficult part of clustering. Descriptive statistics are the standard way doing it. The objective here is to see if association rules will (hopefully) give a more intuitive explanation.

Procedure

- Create a clustering of the students data set using the marks attributes.
- KMeans is used and k is determined by using an elbow plot. The elbow plot was reasonably sharp at 4.
- An extra column is added to the student dataset which contains the id of the cluster to which a particular record belongs to.
- Association rules are generated by forcing the cluster id to be the antecedent to get the characterization of each cluster.

Observations

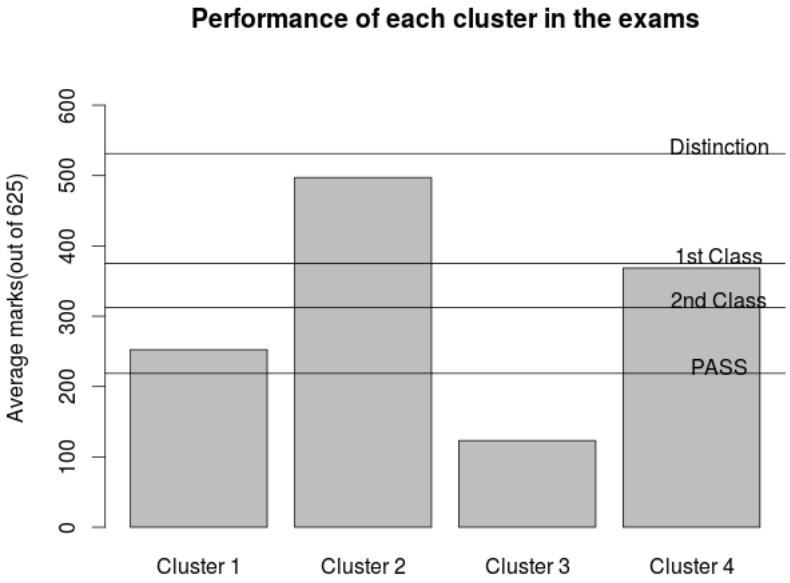
- Elbow plot obtained



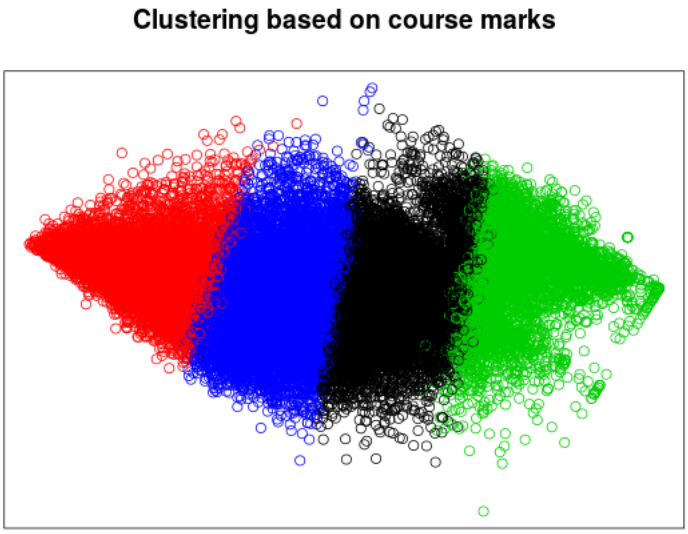
- Distribution around NRC_CLASS

	1	2	D	FAIL	PASS
1	1	57	0	2878	8780
2	4554	0	1439	1	0
3	0	0	0	4950	5
4	4386	5669	1	32	249

- Characterization using mean marks



- Visualization after reducing dimension to two using SVD



- Association rules for each cluster

```
> inspect(sort(cluster.rules))
    lhs                    rhs                                 support   confidence
lift
1   {CLUSTER_ID=1} => {NRC_PHYSICAL_CONDITION=N} 0.3540695   0.9973540 1.0003245
2   {CLUSTER_ID=1} => {CANDIDATE_TYPE=RF}        0.3011333   0.8482417 0.9576707
3   {CLUSTER_ID=1} => {NRC_MEDIUM=K}             0.2771953   0.7808126 1.1236864
4   {CLUSTER_ID=1} => {NRC_CLASS=PASS}           0.2660445   0.7494025 2.7376336

5   {CLUSTER_ID=2} => {CANDIDATE_TYPE=RF}        0.1815042   0.9993327 1.128253
6   {CLUSTER_ID=2} => {NRC_PHYSICAL_CONDITION=N} 0.1813526   0.9984985 1.001472
7   {CLUSTER_ID=2} => {NRC_CASTE_CODE=4}         0.1536877   0.8461795 1.196368
8   {CLUSTER_ID=2} => {NRC_CLASS=1}              0.1379916   0.7597598 2.804339

9   {CLUSTER_ID=3} => {NRC_CLASS=FAIL}           0.1499909   0.9989909 4.1939573
10  {CLUSTER_ID=3} => {NRC_PHYSICAL_CONDITION=N} 0.1489607   0.9921292 0.9950841
11  {CLUSTER_ID=3} => {S2_CLASS=FAIL}            0.1335374   0.8894046 3.4110554
12  {CLUSTER_ID=3} => {S1_CLASS=FAIL}            0.1203866   0.8018163 3.9028825
13  {CLUSTER_ID=3} => {NRC_MEDIUM=K}             0.1198715   0.7983855 1.1489760
14  {CLUSTER_ID=3} => {L1_CLASS=FAIL}            0.1105994   0.7366297 4.2330232

15  {CLUSTER_ID=4} => {NRC_PHYSICAL_CONDITION=N} 0.3126477   0.9981619 1.001135
16  {CLUSTER_ID=4} => {CANDIDATE_TYPE=RF}        0.3112842   0.9938086 1.122017
17  {CLUSTER_ID=4} => {NRC_CASTE_CODE=4}         0.2308345   0.7369643 1.041954
18  {CLUSTER_ID=4} => {NRC_MEDIUM=K}             0.2207139   0.7046532 1.014084
```

Conclusions

- Generating association rules on clusters is a satisfactory categorization technique.
- The lift is sometimes close to 1 which suggests independence. But this does not matter as the rules are not being interpreted as correlations between the lhs and rhs. The rhs just gives the categorization of the cluster. Lift close to one suggests the same categorization for other clusters.
- The low support is also not a matter of concern as it is reflective of the measure w.r.t the whole dataset and not to the particular cluster.
- Categorization : The categorization is as shown in the observations. The physical condition of students across clusters being normal is result of the lopsided distribution.

5 Cross Cluster Analysis

Objective

To compare the cluster characteristics of a subset of the dataset with that of the whole dataset

Procedure

- Partition the student dataset on a particular basis
  - GENDER_CODE
  - URBAN_RURAL
- Cluster the while dataset and the partitioned dataset separately and categorize the clusters
  - Clustering and categorization is done as specified in refchap:experiment3
- Compare the categorization of clusters created out of each partitioned dataset with that of the clusters constructed using the whole dataset.

Observations

- Partitioning based on GENDER_CODE

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Whole population(33002)	{L1_CLASS=FAIL} {NRC_CLASS=FAIL} {S1_CLASS=FAIL} {S2_CLASS=FAIL} {NRC_MEDIUM=K}	{NRC_CLASS=PASS} {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	{CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4} {NRC_MEDIUM=K}	{NRC_CLASS=1} {NRC_CASTE_CODE=4} {CANDIDATE_TYPE=RF}
Boys(54%)	Avg. Marks = 107 {S3_CLASS=FAIL} {S1_CLASS=FAIL} {L1_CLASS=FAIL} {NRC_CLASS=FAIL} {S2_CLASS=FAIL} {L2_CLASS=FAIL} {NRC_MEDIUM=K}	Avg. Marks = 237 {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	Avg. Marks = 353 {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4} {NRC_MEDIUM=K}	Avg. Marks = 486 {NRC_CLASS=1} {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4}
Girls(46%)	Avg. Marks = 138 {NRC_CLASS=FAIL} {S1_CLASS=FAIL} {S2_CLASS=FAIL} {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	Avg. Marks = 264 {NRC_CLASS=PASS} {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	Avg. Marks = 376 {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4} {NRC_MEDIUM=K}	Avg. Marks = 502 {S1_CLASS=1} {NRC_CLASS=1} {NRC_CASTE_CODE=4} {CANDIDATE_TYPE=RF}

- Partitioning based on URBAN_RURAL

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Whole population	Avg Marks = 123 {L1_CLASS=FAIL} {NRC_CLASS=FAIL} {S1_CLASS=FAIL} {S2_CLASS=FAIL} {NRC_MEDIUM=K}	Avg Marks = 252 {NRC_CLASS=PASS} {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	Avg Marks = 368 {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4} {NRC_MEDIUM=K}	Avg Marks = 497 {NRC_CLASS=1} {NRC_CASTE_CODE=4} {CANDIDATE_TYPE=RF}
Urban(43%)	Avg Marks = 128 {L1_CLASS=FAIL} {NRC_CLASS=FAIL} {S1_CLASS=FAIL} {S2_CLASS=FAIL}	Avg Marks = 256 {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4}	Avg Marks = 382 {NRC_MEDIUM=E} {NRC_CASTE_CODE=4} {CANDIDATE_TYPE=RF}	Avg Marks = 515 {CANDIDATE_TYPE=RF} {NRC_CASTE_CODE=4}
Rural(57%)	Avg Marks = 118 {L1_CLASS=FAIL} {NRC_CLASS=FAIL} {S1_CLASS=FAIL} {S2_CLASS=FAIL} {L2_CLASS=FAIL} {NRC_MEDIUM=K}	Avg Marks = 250 {NRC_CLASS=PASS} {NRC_MEDIUM=K} {CANDIDATE_TYPE=RF}	Avg Marks = 358 {CANDIDATE_TYPE=RF} {NRC_MEDIUM=K}	Avg Marks = 477 {NRC_CLASS=1} {S1_CLASS=1} {NRC_CASTE_CODE=4} {CANDIDATE_TYPE=RF} {NRC_MEDIUM=K}

Conclusions

- Partitioning based on URBAN_RURAL
  - Cluster 1 - Candidates who fail
    - Urban - Their medium is not necessarily Kannada
    - Rural - They also fail in L2
  - Cluster 2 - Candidates who just pass
    - Urban - Medium is not necessarily kannada and students belong to the general category.

    - Rural - Repeaters do not just pass. They either secure higher grades or fail(more likely).
  - Cluster 3 - Just missing First Class
    - Urban - Medium is english.
    - Rural – Medium is kannada. Are not necessarily general category.
  - Cluster 3 - Just missing distinction
    - Urban - There are a significant number of students who have gotten distinction.
    - Rural - Medium is kannada and have scored atleast 60 in mathematics.
- Partitioning based on GENDER_CODE
  - Cluster 1 - Candidates who fail
    - Boys - Fail in all subjects except L3, which is typically Hindi.
    - Girls - Do not fall in L1(Mostly Kannada). Most are regular freshers.
  - The other clusters have the same characteristics.