DS707 Data Analytics

Project Report

SSLC Data Analysis

Masters of Technology
in
Information Technology

Submitted by

Roll No	Names of Students
MT2013025	Apoorwa Mishra
MT2013026	Arjun S Bharadwaj
MT2013140	Shah Ankitkumar

Under the guidance of
Prof. Chandrashekar R

International Institute of Information Technology
Bangalore, Karnataka, India – 560 100

Fall Semester 2014

Contents

Descriptive Report

This section describes the dataset that was used for performing various experiments using R programming language. Some of the notable features of the dataset are as follows:

- The data set has 33,002 observations of 38 variables.
- The data set has 32,003 valid observations that is suitable for analytics.
- The maximum marks scored by the student is 620 and the minimum marks is 6.
- The mean of total marks is 321.6.
- Every subject has at least one person with maximum score, but there is no one with maximum score in all the subjects.
- Just 4.23% of students have scored distinction, 21.23% have failed. First and Second class have about 28% of students each.
- About 40% of students belong to government schools, 30% belong to aided schools and remaining to unaided schools.
- About 57.41% belong to rural area and rest to the urban area.
- 70% of candidates belong to general category. 17% belong to SC, rest belong to ST and Category 1.
- About 53% of candidates are boys, rest are girls.
- 69% of students belong to Kannada medium, 25% belong to English. Rest belong to other medium schools.
- 99.78% of candidates are normal and others have disability.
- 90.26% of candidates are regular freshers.
- In all the subjects, the pass percentages of individual subjects are in the range of 84% - 92%. But still, more than 22% have failed in overall result! This indicates the candidates who have failed have to focus more on few subjects.

Experiment 1

Discretizaion + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Discretize subject marks into discrete attributes S (use NRC_CLASS as domain)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 2

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
    - Determine the least number of attributes S that give the best possible regression equation (least error)
    - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
    - Data is loaded and cleansed by removing invalid and missing rows.
    - Regression analysis is performed on the data by using the marks data.
    - Synergy/Interaction effect of all the marks are obsereved and the combination of marks having least p-value is chosen for classification.
    - Marks are rounded off for improving the speed of classification.
    - Classification is performed on the data based on the class variable combination having least p-value.
    - The results of confusion matrices are compared.
- Results Obtained:
    - All Subjects are used for classification:
        - Accuracy: 90.2%
        - 95% CI: (0.8962, 0.9076)

- Best case:
  L1_MARKS, L2_MARKS, S2_MARKS, S3_MARKS (p-value = 0.0732) are used for classification:
  * Accuracy: 84.03%
  * 95% CI: (0.8332, 0.8472)
- Worst case:
  L3_MARKS, S1_MARKS (p-value = 0.94523) are used for classification:
  * Accuracy: 69.46%
  * 95% CI: (0.6857, 0.7033)

- Conclusion:
  - Taking all the subjects marks for classification gives the highest accuracy.
  - Taking the combination of the subjects having low p-value offers the next highest accuracy for classification.
  - Conversely, the combination of subjects having highest p-value gives the least accuracy.

Experiment 3

Clustering + Association Rules

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.

- Objective:
  - Create clusters based on M
  - Characterize each cluster individually by creating association rules (Use discretized subject marks as ITEMS)
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 4

Confidence Interval

Calculate the confidence intervals.

- Objective:
  - On pass percentage across different districts (DIST_CODE)
  - Repeat the above for school type (SCHOOL_CODE)
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 5

Urban / Rural Characterization

What are the characteristics of students from urban and rural areas, respectively? For antecedent, try with demographic info (SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE) Also try with subject performance in the antecedent

- Objective:
  Identify association rules with URBAN_RURAL in the consequent of the rule
- Procedure followed:
  - Data is loaded and cleansed by removing invalid and missing rows.
  - The values of all the columns are factored so that it's suitable for association rules analysis.
  - Apriori algorithm is run on the data by forcing URBAN_RURAL=Rural rule in consequent.
  - Apriori algorithm is run on the data by forcing URBAN_RURAL=Urban rule in consequent.
  - The rules generated with high confidence and lift are compared for both the cases.
- Results Obtained:
  For URBAN_RURAL = Urban in the consequent, the following were the results:

Antecedant	Support	Confidence	Lift
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English	0.1305375	0.8029823	1.883977
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English, NRC_PHYSICAL_CONDITION = Normal	0.1297800	0.8025108	1.882871
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English, NRC_CASTE_CODE = General	0.1130235	0.8002575	1.877584

For URBAN_RURAL = Rural in consequent, the following were the top three results:

Antecedant	Support	Confidence	Lift
SCHOOL_TYPE = Government, NRC_GENDER_CODE=Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE=Regular Fresher, L1_RESULT = Pass, L2_RESULT = Pass, S2_RESULT = Pass, S3_RESULT = Pass	0.1018423	0.8611325	1.500797
SCHOOL_TYPE = Government, NRC_GENDER_CODE = Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE = Regular Fresher, L1_RESULT = Pass, L2_RESULT = Pass, L3_RESULT = Pass, S1_RESULT = Pass, S3_RESULT = Pass	0.1015393	0.8609969	1.500561
SCHOOL_TYPE = Government, NRC_GENDER_CODE = Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE = Regular Fresher, L1_RESULT = Pass, L2_RESULT = Pass, L3_RESULT = Pass, S1_RESULT = Pass, S2_RESULT = Pass	0.1012969	0.8609323	1.500448

- Conclusion:
  - Students in Urban area mainly belong to Unaided English medium schools.
  - Students in Rural area are mainly boys who belong to Government Kannada medium schools.

Experiment 6

Performance characteristics

Use D (Distinction) and FAIL in the consequent of association rule. For antecedent, try with demographic info (SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE). Also try with subject performance in the antecedent.

- Objective:
  - What properties characterize high and poor performers?
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 7

Decision tree vis-a-vis A-rules

Can we use decision trees to validate association rules or vice-versa?

- Objective:
  - Note the strongest rules that have been found
  - Then induce a decision tree based on those attributes
  - Validate the decision tree using standard metrics
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 8

Cross - cluster analysis

Create a clustering C1 of the overall population. Then create a clustering C2 of partitioned population separately (e.g., gender-based)

- Objective:
  - Compare cluster C1 with C2.
  - Are the characteristics same? Show it by statistical analysis.
- Procedure followed:
  - The data file is loaded, the invalid data is removed and the L1_MARKS is normalised to 100.
  - The data is split into three parts: Overall population data, Boys data, Girls data.
  - The value of k = 5 is selected and k-means is run of all the three datasets.
- Results Obtained:
  - The range of mean of marks for all the subjects across the three datasets are as follows:

Dataset \Cluster	1	2	3	4	5
Overall population data	54.6 - 75.79	43.38 - 62.23	73.36 - 87.19	19.39 - 27.20	35.05 - 43.52
Boys data	72.66 - 85.43	16.86 - 24.90	32.93 - 41.34	42.26 - 58.07	54.01 - 72.40
Girls data	21.11 - 29.24	74.6 - 88.71	36.53 - 47.63	44.61 - 67.27	55.74 - 79.22

- When the cluster are analysed with NRC_CLASS, the following matrix is obtained:
  - Overall Population data:

Class \Cluster	Distinction	Fail	First	Pass	Second
1	0	0	6065	9	677
2	0	129	95	2992	4967
3	1356	0	2927	0	0
4	0	4164	0	0	0
5	0	2502	0	6120	0

  - Boys data:

Class \Cluster	Distinction	Fail	First	Pass	Second
1	597	0	1485	0	0
2	0	2139	0	0	0
3	0	2048	0	2692	0
4	0	96	0	2399	1961
5	0	2	2646	8	842

  - Girls data:

Class \Cluster	Distinction	Fail	First	Pass	Second
1	0	1821	0	0	0
2	759	0	1327	0	0
3	0	651	0	3296	0
4	0	38	470	725	2761
5	0	0	3159	1	80

- Conclusion:
  The characteristics are slightly different, however the pattern across clusters are similar.
  - The average score range is lesser in boys data compared to overall population data and girls data indicationg girls performing better in every cluster.
  - The width of the range of average scores is more is boys data and overall population data than girls data indicating the standard deviation is low for boys in every cluster.
  - In the boys data, the fail and pass class are almost equally distributed in the clusters.
  - In the girls data, the pass class has more distribution than fail class in the clusters.

Experiment 9

Score Prediction - Additional Experiment

Prediction of the score using regression equation.

- Objective:
  - Predict the total marks of the candidate from the regression equation.
- Procedure followed:
  - The data file is loaded and the invalid data is removed.
  - The regression formulation is done by specifying the class variables and the predictors.
  - The data is passed to the lm function and the equation is obtained based on the training data.
  - The equation is now used to predict TOTAL_MARKS of the test data.
- Results Obtained:

  The topper data is given to the model for prediction. The following result is obtained:

Actual Score	Predicted Score	Actual Score - Predicted Score
610	610.2218	-0.2217667
610	610.2105	-0.2105289
612	612.2313	-0.2313323
611	611.2340	-0.2339893
613	613.2204	-0.2203704
619	619.2363	-0.2362730
611	611.2160	-0.2160255
612	612.2188	-0.2187575
620	464.1491	155.8508606
615	447.1857	167.8142784

- Conclusion:

  – The generated regression equation is an accurate equation and it can be seen with the predicted data.

  – The predicted data in the last two observations indicate a large difference, this shows that the total marks was tampered artificially.

  – The co-efficients of all the intercepts are almost equal to 1. This makes it highly accurate.