Educational Data Mining Model Using Rattle

Sadiq Hussain
System Administrator
Dibrugarh University
Dibrugarh Assam

G.C. Hazarika
Department of Mathematics
Dibrugarh University
Dibrugarh Assam

***Abstract*—Data Mining is the extraction of knowledge from the large databases. Data Mining had affected all the fields from combating terror attacks to the human genome databases. For different data analysis, R programming has a key role to play. Rattle, an effective GUI for R Programming is used extensively for generating reports based on several current trends models like random forest, support vector machine etc. It is otherwise hard to compare which model to choose for the data that needs to be mined. This paper proposes a method using Rattle for selection of Educational Data Mining Model.**

Keywords—Educational Data Mining; R Programming; Rattle; ROC Curve; Support Vector Machine; Random Forest

I. Introduction

Dibrugarh University, the easternmost University of India was set up in 1965 under the provisions of the Dibrugarh University Act, 1965 enacted by the Assam Legislative Assembly. It is a teaching-cum-affiliating University with limited residential facilities. The University is situated at Rajabheta at a distance of about five kilometers to the south of the premier town of Dibrugarh in the eastern part of Assam as well as India. Dibrugarh, a commercially and industrially advanced town in the entire northeastern region also enjoys a unique place in the fields of Art, Literature and Culture. The district of Dibrugarh is well known for its vast treasure of minerals (including oil and natural gas and coal), flora and fauna and largest concentration of tea plantations. The diverse tribes with their distinct dialects, customs, traditions and culture form a polychromatic ethnic mosaic, which becomes a paradise for the study of Anthropology and Sociology, besides art and culture. The Dibrugarh University Campus is well linked by roads, rails, air and waterways. The National Highway No.37 passes through the University Campus. The territorial jurisdiction of Dibrugarh University covers seven districts of Upper Assam, viz, Dibrugarh, Tinsukia, Sivasagar, Jorhat, Golaghat, Dhemaji and Lakhimpur. [1]

There are more than hundred numbers of Colleges/ Institutes offering TDC (Three Year Degree) Course affiliated/ permitted under the University. Since the number of students in the Arts Stream is larger in comparison to the other stream (B.Sc., B.Com., B.Tech. etc) we considered the data for the B.A. (Bachelor of Arts) course for our present study of educational data mining. The required digitized data are collected from Dibrugarh University Examination Branch for the affiliated colleges of the University B.A. programme from 2010 to 2013. This paper evaluates performance gender wise as well as caste wise of the students. The Colleges are categorized as Urban as well as Rural depending on their locations. In case of caste wise observations, the binomial operators are Urban and Rural.

There are several data mining tools and statistical models available. This paper focuses one which data mining tools shall be the best suited and what would be the statistical models for such knowledge discovery.

II. Literature Review

A. Data Mining

Data Mining detects the relevant patterns from databases / data warehouses using different programs and algorithms to look into current and historical data which can be analyzed to predict future trends [2]. It is very difficult for any organization to extract hidden patterns from the huge data marts and data ware houses without the help of data mining tools and programs. It is like searching for the pearls in the sea of data. This knowledge set is extremely useful in developing a knowledge support system and making important decisions regarding the future trends predictions.

Statisticians have used different manual techniques for the benefit of the business, predicting trends and results based on data over the years. The business houses had developed huge databases or data warehouses to become “data tombs”. The data was never transformed into information. But with the help of data mining tools and algorithms now professionals from different areas may extract knowledge quickly and at ease.

B. Educational Data Mining

Data mining, often called knowledge discovery in database (KDD), is known for its powerful role in uncovering hidden information from large volumes of data [3]. Its advantages have landed its application in numerous fields including e-commerce, bioinformatics and lately, within the educational research which commonly known as Educational Data Mining (EDM) [4]. EDM is defined by The Educational Data Mining community website, www.educationaldatamining.org as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting, and using those methods to better understand students, and the settings which they learn in. EDM often stresses with the improvement of student models which denote the student’s current knowledge, motivation and attitudes [5].

C. Rattle: A Data Mining GUI for R

The data miner draws heavily on methodologies, techniques and algorithms from statistics, machine learning, and computer science [6]. R programming language is a powerful tool for data mining. Rattle (the R Analytical Tool To Learn Easily) provides GUI for the R programming environment. We have to use the library (rattle) and rattle () brings up the GUI for the programmers. Highly skilled Statisticians may efficiently use the R Programming Language. So, it is out of reach for many people without in depth knowledge of Statistics. But Rattle provides sophisticated GUI for data analysis and provides the necessary graphs with a click. Rattle provides another magnitude to the R programming and a platform for the novice data miners to work efficiently. Rattle's user interface provides an entry into the power of R as a data mining tool. [6]

D. ROC Curves Analysis

To determine a cutoff value, Receiver operating characteristic (ROC) curves is used in many areas. We may use the ROC curve for the selection of best suited models. In our educational data mining experiment, we use the ROC curve to determine the selection of model.

III. Experiments and Evaluation

A. The Data Set

We have included a small part of the Category and Gender based tables termed as Table 1 and Table 2 for which the suitable models needs to be selected. The Examination Branch of Dibrugarh University provides various College Codes for different Colleges under its jurisdiction. The field 'Appeared' means the number of candidates appeared for that examination and 'Passed' means the number of candidates passed for that particular examination. The field 'PassPercentage' is the Passed Percentage of the Candidates for a particular category. We define various terms in their codes as below:

a) Category

Category	Code
General	1
MOBC	2
OBC	3
SC	4
ST	5

b) Performance

Pass Percentage	Performance
>= 90%	1
>=75%	2
>=60%	3
>=45%	4
< 45%	5

c) Location

Location	Code
urban areas colleges	0
rural areas colleges	1

d) Gender

Gender	Code
Male Candidates	0
Female Candidates	1

The meaning of the data fields as depicted in the sample Table 2 are same Table 1 except one field i.e. 'Gender'. Now the stage is set and ready to perform.

B. Experiments performed by Rattle

The main objective in this paper is to select the best suited models for performing the statistical analysis of the datasets. We used one Xeon based Database Server for the experiments. The rattle package was used for the same. The data is imported to R which was stored in .csv format. The target data was categorical data and the partition chosen was 70/30/0. If one explores the data, one may visualise the data by using box plot, histogram, cumulative and benford curves. The histogram, the cumulative and benford curves are presented in the figures I,II,III and IV. Now, one may use the Model tab and select all the models for the comparison. The models are of type tree, random forest, boost, support vector machine, regression models and neural network. The data is evaluated through all the models. Our goal is to find the best suited models for the data through ROC curve.

C. Evaluation of the Experiments

In the figure V, we have placed one of the ROC curves for the category data. The followings are the actual findings using the Rattle based on the category wise data.

```
Area under the ROC curve for the rpart model on
categoryba.csv [validate] is 0.8814

Rattle timestamp: 2014-05-06 06:48:54 sadiq

============================================

Area under the ROC curve for the ada model on categoryba.csv
[validate] is 0.9425

Rattle timestamp: 2014-05-06 06:48:55 sadiq

============================================

Area under the ROC curve for the rf model on categoryba.csv
[validate] is 0.9221

Rattle timestamp: 2014-05-06 06:48:55 sadiq

============================================

Area under the ROC curve for the ksvm model on
categoryba.csv [validate] is 0.9301

Rattle timestamp: 2014-05-06 06:48:55 sadiq

============================================

Area under the ROC curve for the glm model on
categoryba.csv [validate] is 0.8980

Rattle timestamp: 2014-05-06 06:48:55 sadiq

============================================

Area under the ROC curve for the nnet model on
categoryba.csv [validate] is 0.7393

Rattle timestamp: 2014-05-06 06:48:55 sadiq
```

From the above ROC curve analysis, it is quite clear that whose area under ROC curve for a particular model is 1 or close to 1, that model is best suited for that data. The Statisticians can further analyze the data based on that model.

The models best suited for our category-wise data are ada model (with area value is 0.9425), rf model (0.9221), ksvm model (0.9301).

If we generate the ROC curve for the gender specific data, then we find the following:

```
Area under the ROC curve for the rpart model on
Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

=============================================

Area under the ROC curve for the ada model on
Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

=============================================

Area under the ROC curve for the rf model on
Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:53 sadiq

=============================================

Area under the ROC curve for the ksvm model on
Gender_BA.CSV [validate] is 0.9982

Rattle timestamp: 2014-05-07 19:55:54 sadiq

=============================================

Area under the ROC curve for the glm model on
Gender_BA.CSV [validate] is 1.0000

Rattle timestamp: 2014-05-07 19:55:54 sadiq

=============================================

Area under the ROC curve for the nnet model on
Gender_BA.CSV [validate] is 0.9999

Rattle timestamp: 2014-05-07 19:55:54 sadiq
```

We may conclude from the above that almost all the models are would deliver better results, but rpart, ada, rf and glm models are best suited.

IV. Conclusions and Future Work

The Rattle package provides a GUI platform toward using R as a programming language. Rattle is open source data mining tools packed under the regime of R. In this paper, two data sets were mined. If one compares the two data sets results, then it may be concluded that ada, rf models are best suited for the data that were mined. We hance found that the female candidates of the University did better than the boys' candidates and the rural candidates did better performance than the urban candidates'(Refer to the figures below). Moreover, as this paper dealt with only one examination i.e. Bachelor of Arts, there are lots of another Examinations to deal with as well as one may extract valuable patterns and information from them. The future plan is to compare entry and exit data of TDC students of different colleges affiliated to Dibrugarh University.

V. Acknowledgments

The authors express their gratefulness to Prof. Alak Kr. Buragohain, Vice-Chancellor, Dibrugarh University for his inspiring words and allowing them to use the Examination data of the University. They generously thank Mr. N.A. Naik, Senior Programmer, Mumbai based firm for helping us to extract the .csv files from the SQL Server database.The authors would like to offer gratitude to Prof. Jiten Hazarika Department of Statistics, Dibrugarh University for his valuable ideas.

References

[1] The Dibrugarh University website: www.dibru.ac.in

[2] John Silltow, August 2006 : Data Mining 101: Tools and Techniques, http://www.internalauditoronline.org/

[3] Witten, I.H. and Frank, E. 1999. Data Mining:Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman, San Francisco, CA.

[4] Baker, R.S.J.d.: Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (eds.) To appear in International Encyclopedia of Education, 3rd edn. Elsevier, Oxford (2010)

[5] Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.

[6] Graham Williams, Rattle: A Data Mining GUI for R, The R Journal Vol. 1/2, December 2009 ISSN 2073-4859.

TABLE I. SAMPLE DATA FOR YEAR-WISE COLLEGE-WISE CATEGORY-WISE LOCATION-WISE DATA OF THE B.A. CANDIDATES

Year	CollegeCode	Category	Appeared	Passed	PassPercentage	Performance	Location
2010	103	1	2	2	100	1	1
2010	103	2	3	3	100	1	1
2010	103	3	25	25	100	1	1
2010	103	4	4	4	100	1	1
2010	103	5	11	8	72.73	3	1

TABLE II. SAMPLE DATA FOR YEAR-WISE COLLEGE-WISE GENDER-WISE DATA OF THE B.A. CANDIDATES

Year	CollegeCode	Gender	Appeared	Passed	PassPercentage	Performance
2010	101	0	46	42	91.3	1
2010	101	1	57	51	89.47	1
2010	102	0	57	47	82.46	1
2010	102	1	66	58	87.88	1

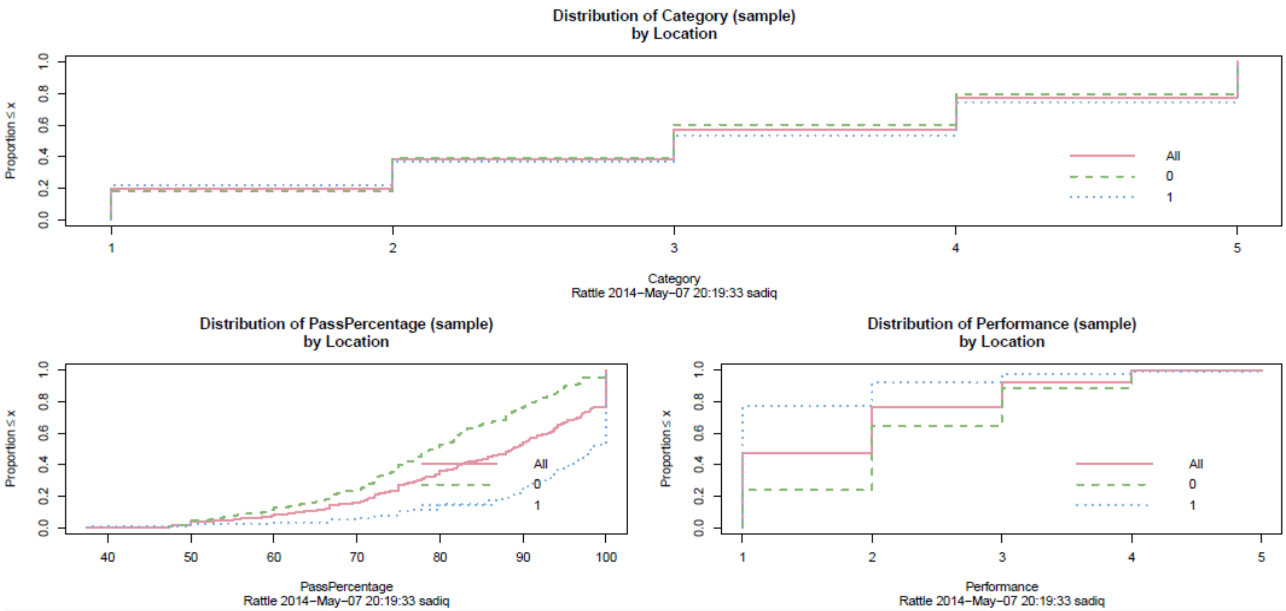

Fig. 1. Cumulative Diagram showing category-wise, Pass Percentage-wise, Performance-wise distribution on the basis of Location

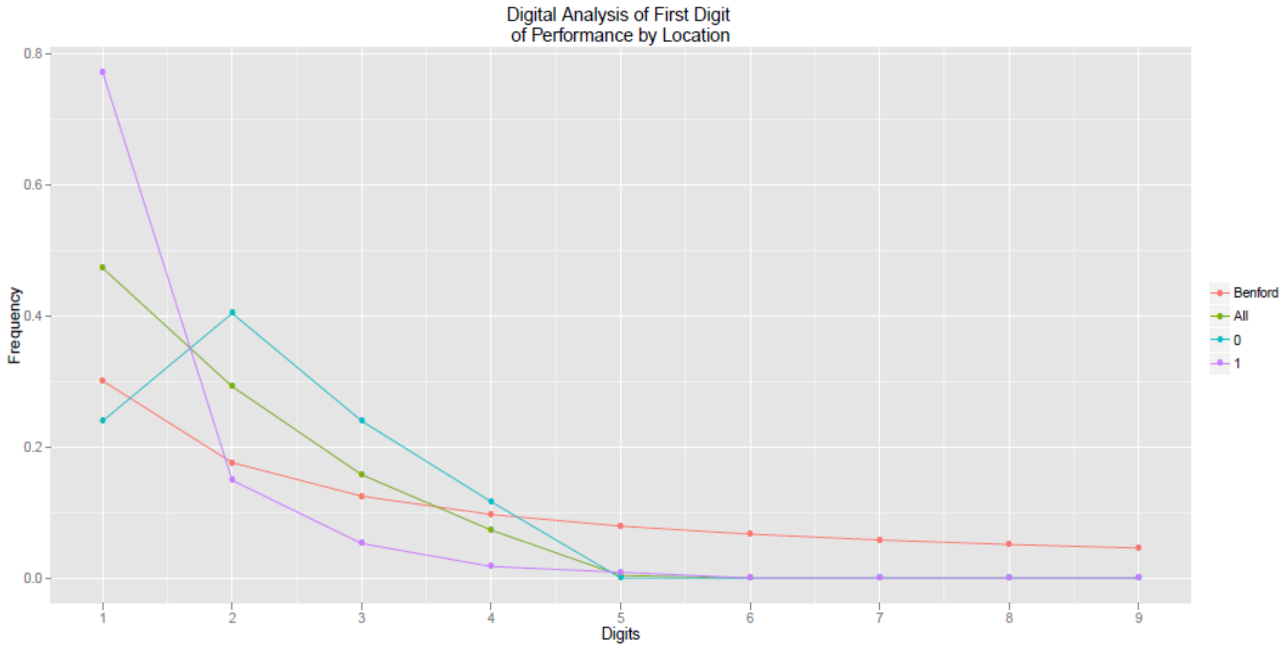

Fig. 2. Benford Diagram showing the performance by Location of the Candidates.

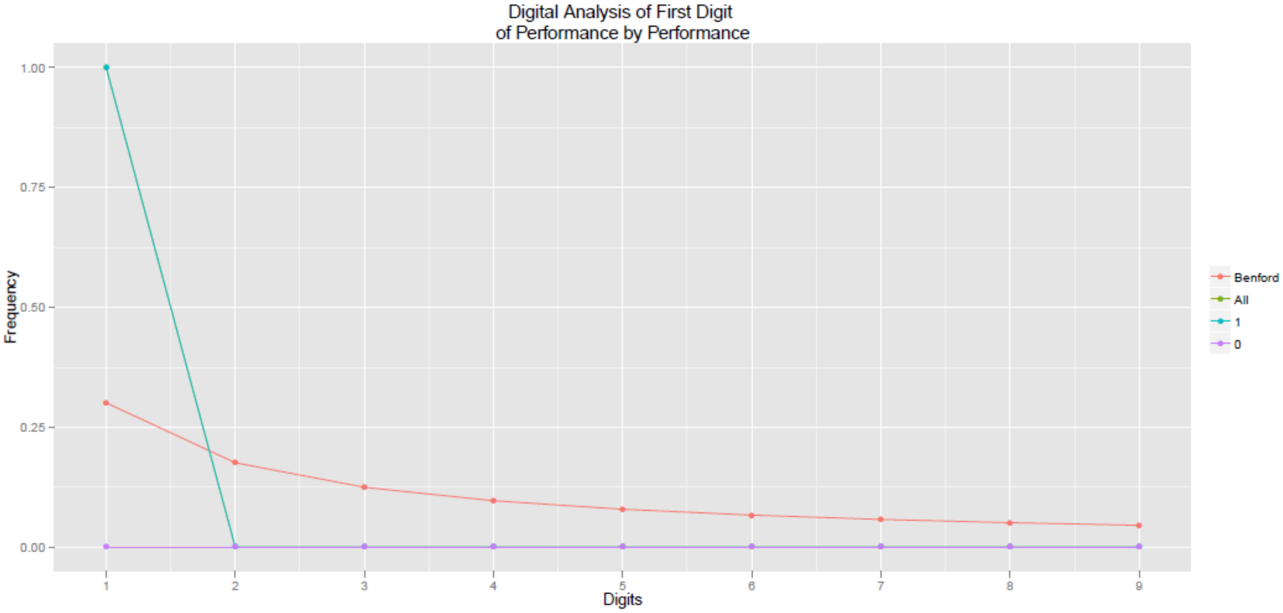

Fig. 3. Benford Diagram showing the performance by Gender of the Candidates.

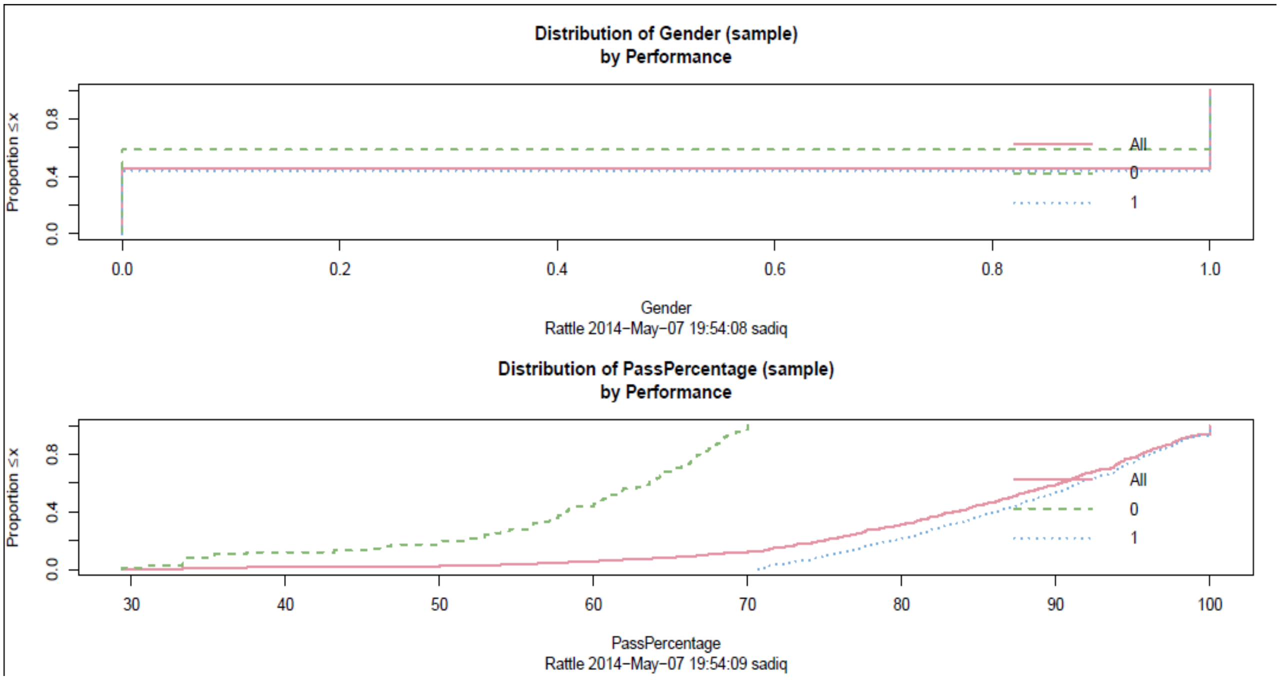

Fig. 4. Cumulative Diagram showing the performance by Pass Percentage and Gender wise.

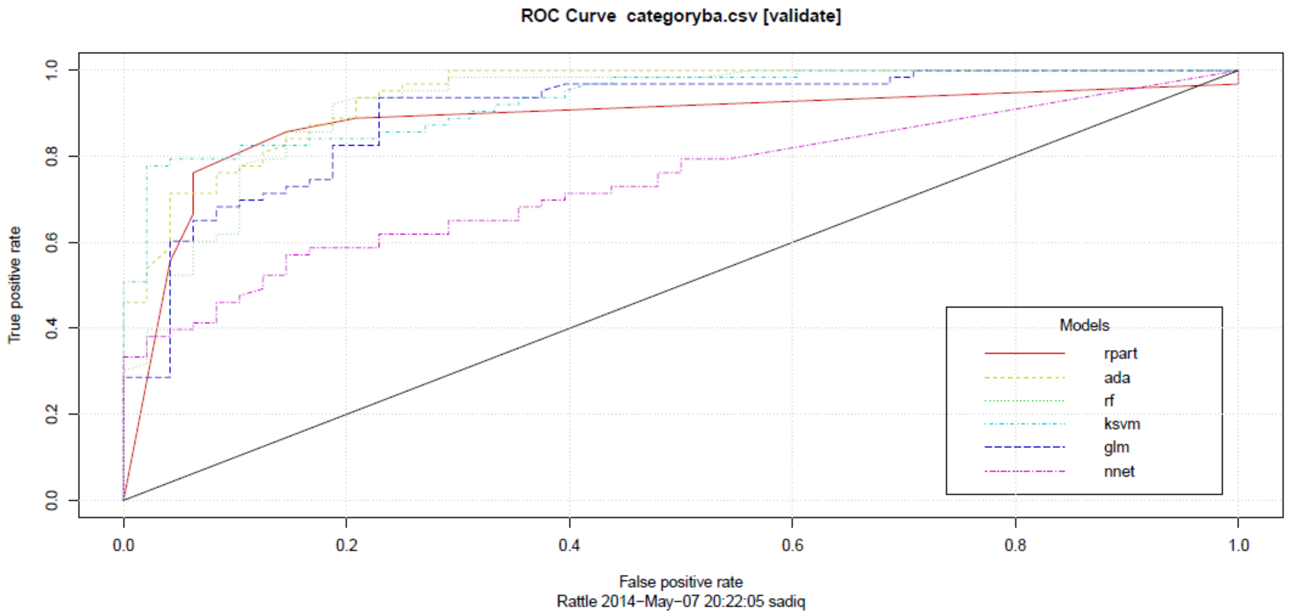

Fig. 5. ROC Curve for the first experiment i.e. performance by category.