Initial report on analysis of the Anganwadi dataset 2010

Avishek Sen Gupta
ThoughtWorks

October 28, 2011

1 Abstract

This report summarises the results of exploration of the Anganwadi dataset provided by the Akshara Foundation. The analysis aims to characterise the structure of the data, and reveal trends (which would otherwise be obscured by the format of the source data) which may inform strategy through subsequent prediction and/or classification procedures.

2 Methodology

2.1 CRISP-DM

CRISP-DM is a process model distilled from the most common approaches used in data mining procedures. It stands for Cross Industry Standard Process for Data Mining. Not so much a prescription as a collection of 'good practices' follwed by data mining professionals, **CRISP-DM** has the following characteristics.

- Domain-neutral
- Tool-neutral
- Provides a structural approach to the data mining process

CRISP-DM segregates data mining endeavours into the following phases.

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

2.2 Relevance of CRISP-DM to this report

As far as this report is concerned, the relevant or most significant phases we focus on are:

- Data Understanding
- Data Preparation
- Modeling

Work on the Evaluation step is still preliminary, and will probably be the subject of another report. In a full-fledged project, the rest of the activities upstream and downstream to the above list will assume more importance, and require corresponding investment.

3 Data Preparation

3.1 Nature of the source data

The dataset comes from the education domain. The source data is a file, with each line corresponding to a single student evaluation record. Roughly, there are 29000 records, prior to any data sanitisation. Each line is pipe(|)-delimited into multiple fields. The fields salient to this analysis are listed below:

- Location of the student's school
- Language of the student
- Student's score before intervention
- Student's score after intervention

The score is not a single number, it is a set of 56 responses marked as 0/1. Generally, a 1 may be treated as a favourable answer, therefore, adding them up to get a single aggregate score has natural ordering: a sense of who did better. We reproduce two such records below, with the original formatting.

30915|YALLAMMANADODDI|KANNADA|1231468|Girl||1NoSection|Aug 2010 Assessment|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1
Post Test|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|

29607|BADAMAKAAN I|KANNADA|1445172|Girl|URDU|1NoSection|Aug 2010 Assessment|1|0|1|1|0|1|0|1|1|1|1|1|1|1|1|0|1|1|1|1|1|0|1|1|1|1|1|0|1|1|1|0|1|1|1|1|1|1|0|1|1|0|1|1|1|1|1|1|1|1|1|1|1|1|1|0|Anganwadi
Post Test|1|0|1|1|1|0|1|1|0|1|1|0|1|1|1|1|0|1|1|1|0|0|1|1|0|1|0|0|1|1|1|0|1|1|0|0|1|1|1|0|0|1|1|1|1|1|1|1|1|1|1|1|1|1|1|1|
...

Looking at the second row, we see that the location of the Anganwadi is BADAMAKAAN I, the student is female and speaks Urdu. The first contiguous set of 0s and 1s is the pre-intervention score, and the next set is the post-intervention one.

3.2 Data representation

3.2.1 Data store

Before any sort of sanitisation or analysis may be performed, it is important to ensure that the source data is stored in a format/datastore which makes querying and modifying the data relatively painless. This decision is largely driven by technological considerations, like:

- Scale of data (centralised/distributed store?)
- Sophistication of queries (OLAP/OLTP?)
- Structure of data, or lack thereof (SQL/NoSQL?)

We were dealing with only about 29000 records, and most of the analysis would probably be performed outside the database. Thus, we opted to use MySQL as our datastore.

3.2.2 Schema

The decisions when creating the database schema affect the ease of querying for relevant information. Apart from the attributes of interest, we wanted to store the individual binary responses as well. One way is to create one column for each response, giving us a total of 112 columns for storing these responses (56 for pre-intervention, 56 for post-intervention). The other way, and that is the one that we chose was to store this information as a 64-bit integer (bigint for MySQL). When required, we could unpack the individual response bits from this number.
We elected to not create any more schema elements like reference data for area or language at this point, because we were not sure (yet) whether there was any data corruption which could lead to duplicate reference data.

A `desc responses;` command on the table reveals the schema we ended up with.

```
+------------------+------------+------+-----+---------+----------------+
| Field            | Type       | Null | Key | Default | Extra          |
+------------------+------------+------+-----+---------+----------------+
| student_id       | int(11)    | YES  |     | NULL    |                |
| area             | char(50)   | YES  |     | NULL    |                |
| pre_performance  | bigint(20) | YES  |     | NULL    |                |
| post_performance | bigint(20) | YES  |     | NULL    |                |
| language         | char(50)   | YES  |     | NULL    |                |
| gender           | char(20)   | YES  |     | NULL    |                |
| pre_total        | int(11)    | YES  |     | NULL    |                |
| post_total       | int(11)    | YES  |     | NULL    |                |
| id               | int(11)    | NO   | PRI | NULL    | auto_increment |
+------------------+------------+------+-----+---------+----------------+
```

3.3 Data migration: Identifying invalid data

It is natural to expect missing or corrupted data. The most crucial attribute are the score data, as any misinterpretation of that data may adversely bias the quality of our analysis. Thus, specific checks were put in place to ensure that none of the binary responses was null or some string other than 0 or 1.

Using this check, we found 1067 responses which violated it. All of them had either empty pre- or post-intervention scores. We did not migrate these response records, though it may be possible to do Monte Carlo simulations to predict the missing data.

As a result, out of a total of 28535 records in the original source, 27468 were migrated to the database.
We also found a large fraction of records which did not have a LANGUAGE attribute, i.e., that field was empty. Nevertheless, they were included in the migration.

4 Bias

Analysis is most suscpetible to bias in the data collection stage. Sampling is one such activity. If, for a statistical study, participating individuals are not equally likely to have been selected, it may be difficult to distinguish between the actual phenomenon and this biased sampling. This sort of bias is called sampling bias.

4.1 Sampling bias

To find evidence for bias, we looked at a few parameters. Here is the breakdown of the population by language, with the biggest language bucket highlighted.

Unspecified=869
URDU=3564
KANNADA=18685
TELUGU=1688
TAMIL=2051
MARATHI=91
OTHER=239
HINDI=243
KONKANI=18
GUJARATHI=12
NOT KNOWN=3
ORIYA=2
MULTI LNG=1
BENGALI=1
NEPALI=1

There is an overwhelming proportion of students who speak Kannada as their mother tongue (leading by an order of magnitude), a fact that is very likely to bias any sort of analysis where language is involved. We must remain cognizant of such biases, and interpret the results accordingly.

Here is the breakdown of the population by gender.

Girl=14822
Boy=12646

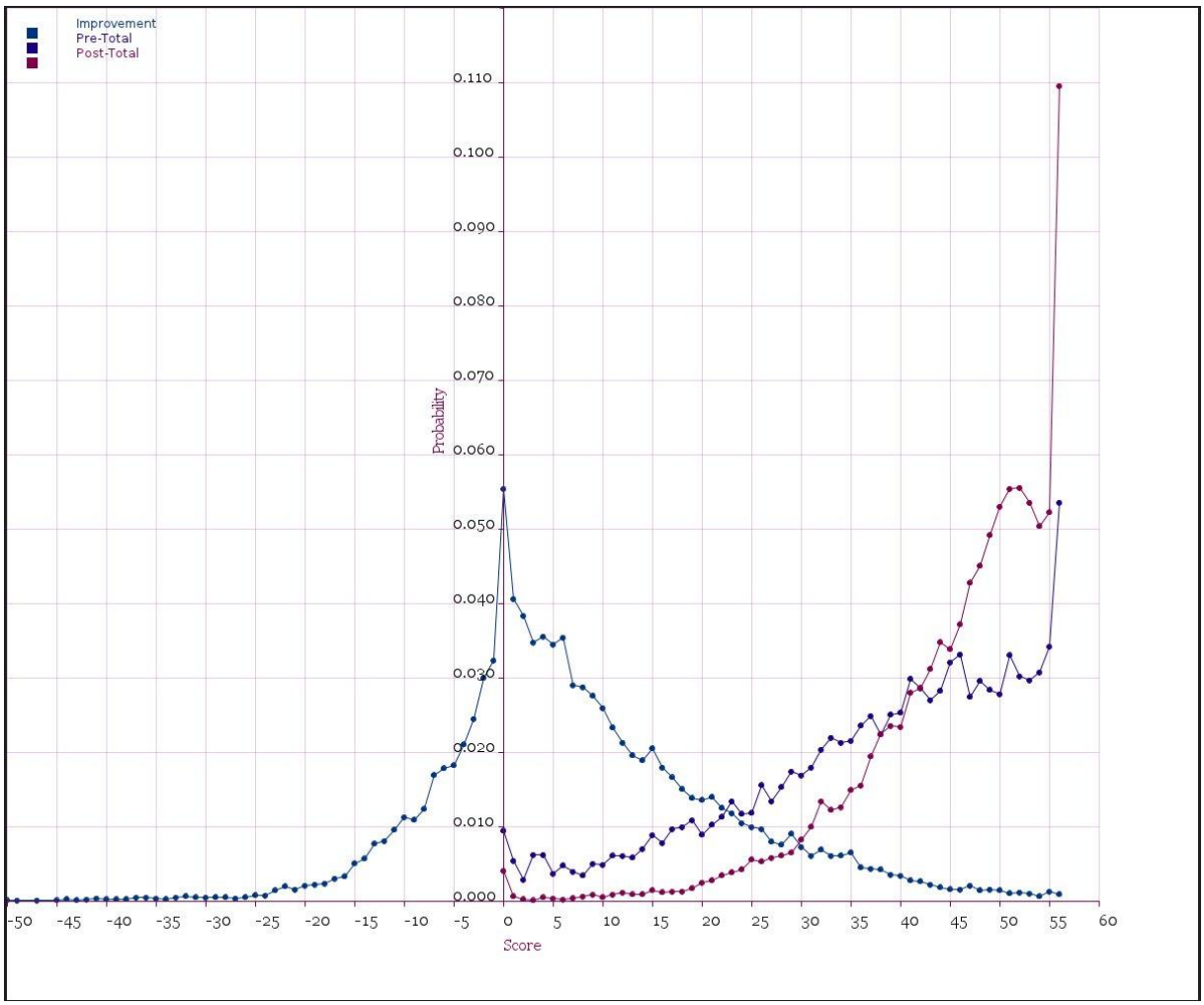
There is not a huge disparity between the two sexes, which indicates that any analysis/prediction based on gender may be less biased.

5 Shape of the Data

Before embarking on any deep analysis of data, it behooves us to look at the shape of the raw data. There are a few reasons why we want to do this.

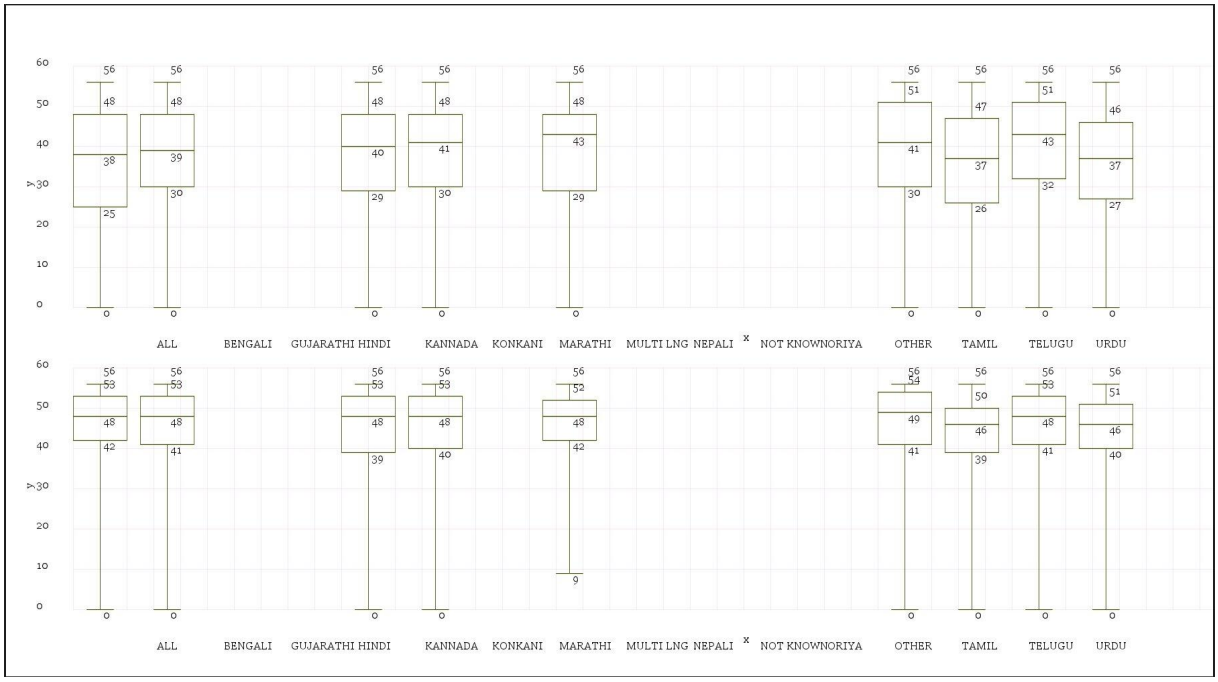
- Evident trends/outliers:
- Evidence of conformance to well-known distributions

5.1 Univariate distributions



5.2 Bivariate distribution

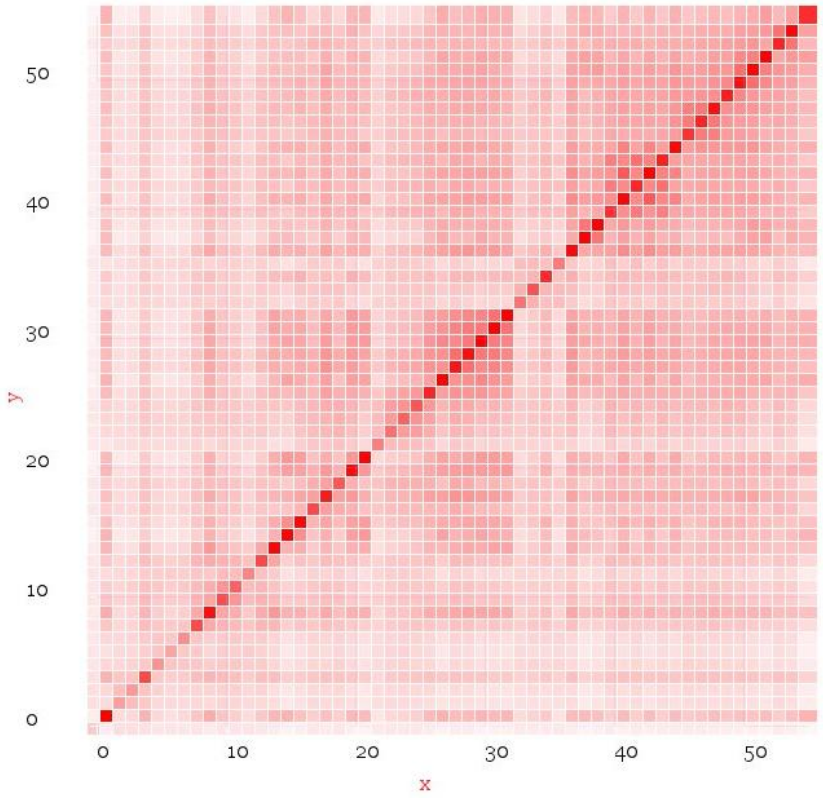
5.3 Summary plots

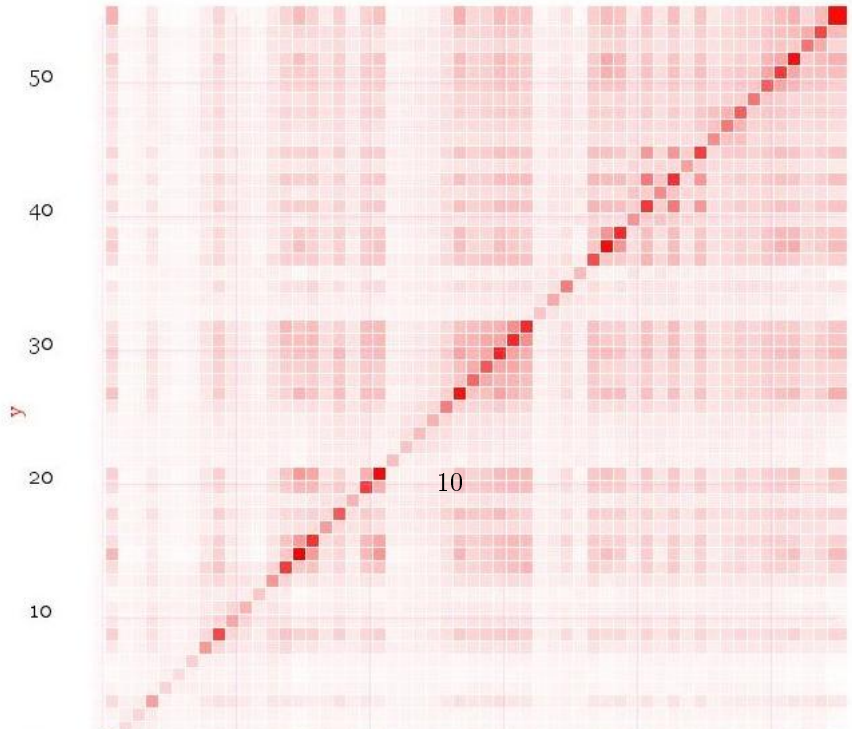


6 Data exploration

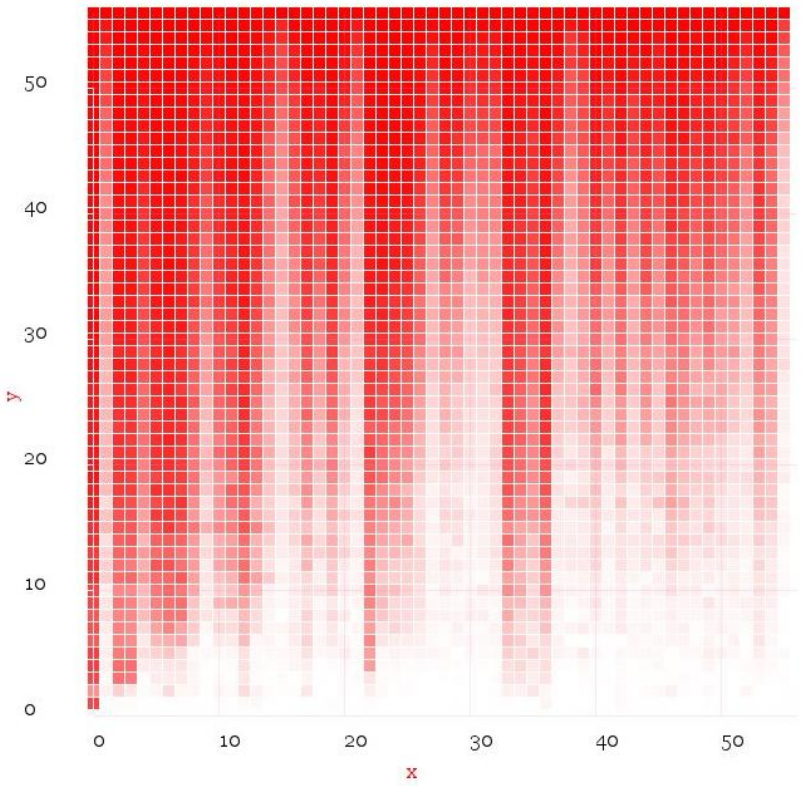
6.1 Parallel Coordinates

6.2 Covariance plot





6.3 Answer distribution

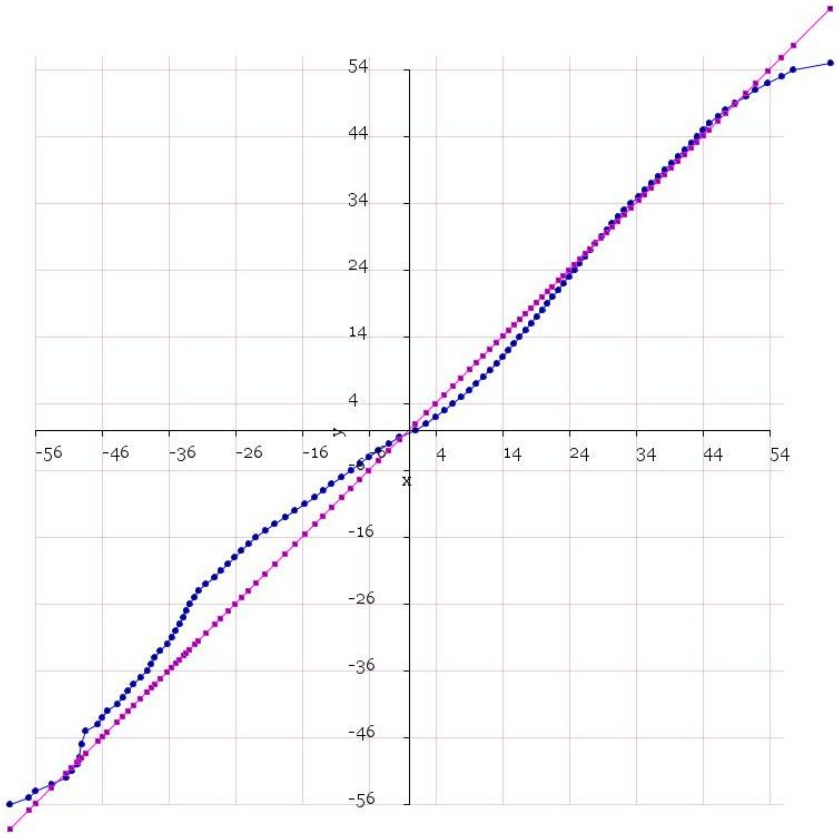


7 Tests for Univariate Normality

7.1 Evidence

7.1.1 Jarque-Bera test

7.1.2 Normal probability plot



7.2 Summary

8 Tests for variable independence

8.1 Chi-square test

Null hypothesis: Area and Improvement are NOT related.
For **area vs. improvement**
Chi-Square statistic = 56499.4692602837
X2 = 9652.9739
Degrees of freedom = 9426
Null hypothesis rejected
c
Null hypothesis: Area and Pre-Score are NOT related.
For **area vs. pre-score**
Chi-Square statistic = 58665.7089390644
X2 = 8062.2959
Degrees of freedom = 7855
Null hypothesis rejected

Null hypothesis: Area and Post-Score are NOT related.
For **area vs. post-score**
Chi-Square statistic = 38567.0016158761
X2 = 8062.2959
Degrees of freedom = 7855
Null hypothesis rejected

Null hypothesis: Language and Post-Score are NOT related.
For **language vs. post-score**
Chi-Square statistic = 280.234448946825
X2 = 96.2166
Degrees of freedom = 75
Null hypothesis rejected

Null hypothesis: Language and Improvement are NOT related.
For **language vs. improvement**
Chi-Square statistic = 232.464548410971
X2 = 113.1452
Degrees of freedom = 90
Null hypothesis rejected

Null hypothesis: Language and Pre-Score are NOT related.
For **language vs. pre-score**
Chi-Square statistic = 277.85501653079
X2 = 96.2166
Degrees of freedom = 75
Null hypothesis rejected

9 Prediction

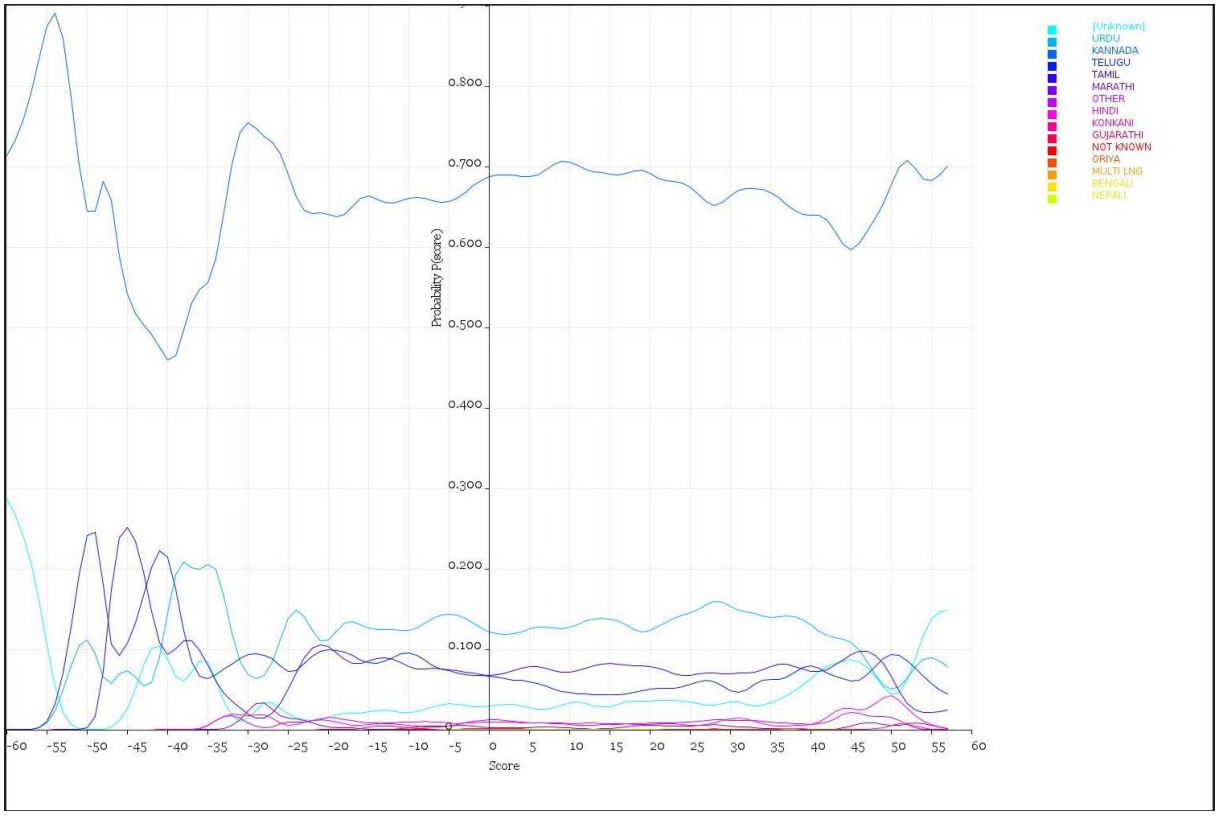
9.1 Decision Trees

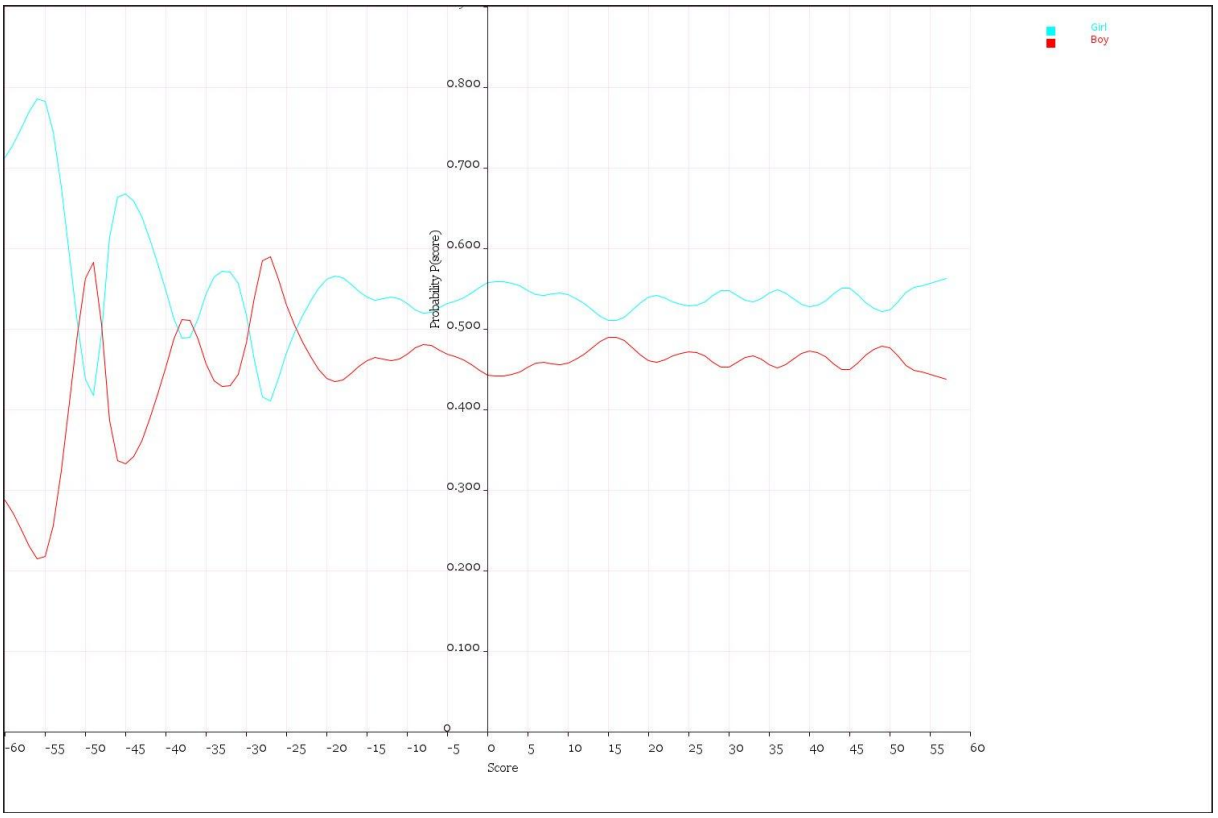
9.2 Bayes classifier

9.3 Density estimators

9.3.1 Naive Bayes density

9.3.2 Kernel density estimation




Girl
Boy
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0
Probability P(score)
-60 -55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 0 5 10 15 20 25 30 35 40 45 50 55 60
Score

10 Dimension reduction/Factor analysis

10.1 Principal Component Analysis

11 Technical notes