

# **Sakura Science Plan Report**

## **TOEIC Score Estimation Based On Reading Analysis**

**Sartaki Sinha Roy**

**Indian Statistical Institute, Kolkata**

**[sinharoy.sartaki24@gmail.com](mailto:sinharoy.sartaki24@gmail.com)**

# ***Acknowledgement:***

I would like to thank Prof. Utpal Garain for nominating me for the Sakura Science Plan and giving me an opportunity to visit Osaka Prefecture University.

I would like to thank Prof. Koichi Kise for the invitation to Osaka Prefecture University and having me in his Lab, Prof. Olivier Augereau for the guidance and mentoring. I would also like to express my gratitude to all the members of Intelligent Media Processing group.

Overall, it has been a great learning experience.

**Sartaki Sinha Roy**

**Indian Statistical Institute, Kolkata**

# ***Table of Contents:***

<b>1. Introduction.....</b>	<b>4</b>
<b>2. Problem Statement.....</b>	<b>4</b>
<b>3. Brief about the experiment .....</b>	<b>4</b>
<b>4. Brief about the data obtained .....</b>	<b>5</b>
<b>5. Approach / Analysis.....</b>	<b>6</b>
<b>6. Comments on the analysis .....</b>	<b>10</b>
<b>7. Analysis of subject wise error .....</b>	<b>10</b>
<b>8. Conclusion .....</b>	<b>11</b>
<b>9. References.....</b>	<b>12</b>

## **1. *Introduction:***

TOEIC is a language comprehension assessing test. It consists of two parts: listening and reading comprehension. The listening comprehension is of 45 mins and is evaluated from 5 to 495 points. The reading comprehension is of 75 mins and is evaluated from 5 to 495 points. The complete test is evaluated from 10 to 990 points (higher the points more is the assessed comprehension of a student).

Tests like TOEIC / TOEFL are elaborate, time consuming and costly. Being able to predict the TOEIC score from the eye movement patterns would provide a much easier and time / cost effective way to estimate a subject's understanding of the English language.

Eye movement during skilled reading is not a continuous process, but it alternates between *fixations* and short ballistic movements called *saccades*, which move the eye gaze from one word to another. Various studies by Rayner et al. have shown that eye movement patterns and reading comprehension are correlated. Hence by analysing the structure of fixations and saccades one can understand various cognitive processes including comprehension.

## **2. *Problem Statement:***

Estimation / Prediction of TOEIC score by analysing the eye movement patterns of subjects.

## **3. *Brief about the experiment:***

- The experiment consisted of 16 Japanese students (mixture of graduate and undergraduate). Each were given to read 15 comprehension passages. Each comprehension passage had four MCQ type of questions (each question had four options). The comprehension passages were sampled from standard TOEIC test
- The gaze pattern of each of the subjects reading the passages were recorded by SMI RED250 (sampling rate of 250 Hz, gaze position accuracy of  $0.4^\circ$ , spatial resolution of  $0.03^\circ$ ). The experiment was conducted in conjunction with the paper '*Towards an Automated Estimation of English Skill via TOEIC Score Based on Reading Analysis*' by Olivier et al.

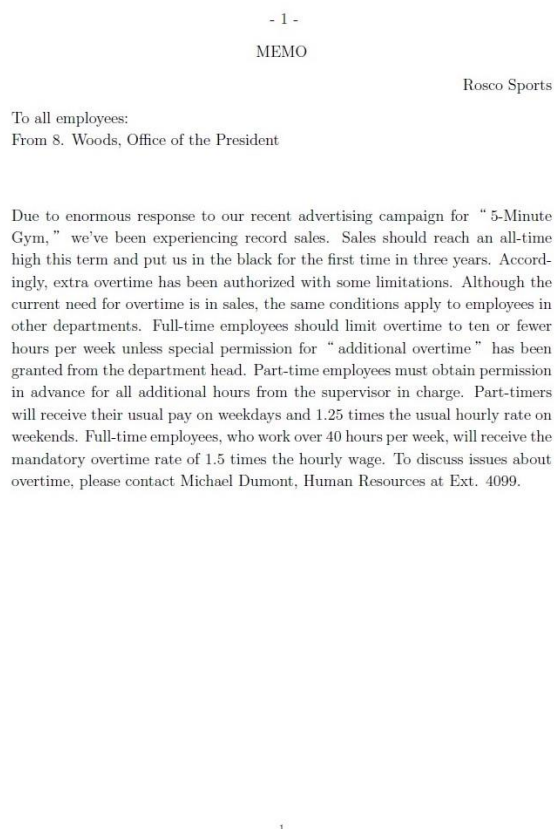


Figure 1: Example of a comprehension type question used in the experiment

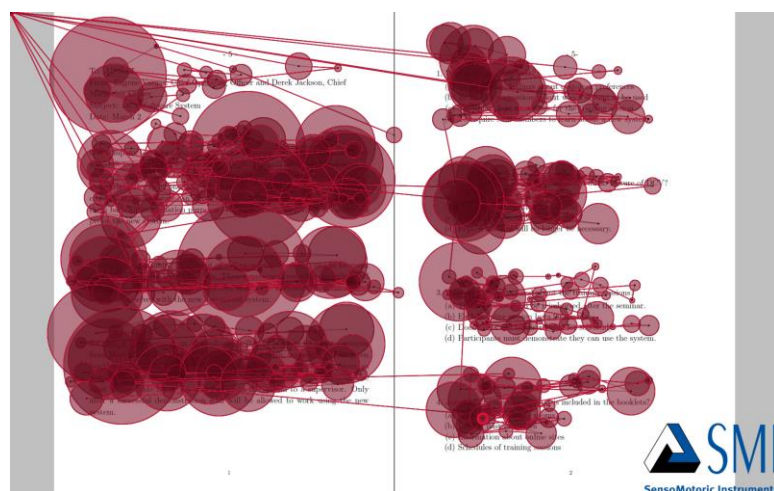


Figure 2: Eye gaze patterns on a comprehension question

## 4. Brief about the data obtained:

- The software provided by SMI was used to process the raw eye movements to extract information about fixations, saccades and blinks

- 33 statistical features are computed from the processed eye movement data such as 'Fixation count', 'Total fixation duration' etc. for each subject on each passage
- One feature is also present that indicates the number of correct answers by the subjects on each comprehension passage (the complete list of features can be found in the paper '*Towards an Automated Estimation of English Skill via TOEIC Score Based on Reading Analysis*' by Olivier et al.)
- All the 34 features are quantitative and in ratio scales (some are continuous like 'Fixation amplitude' etc. , some are discrete such as 'Saccade Count' etc.)
- The response variable is the TOEIC score. Hence the problem at hand is a value estimation / prediction problem

## 5. Approach / Analysis:

- Started with simple univariate analysis such as scatter plots between the features and response, correlations between the features and response and correlations between the features

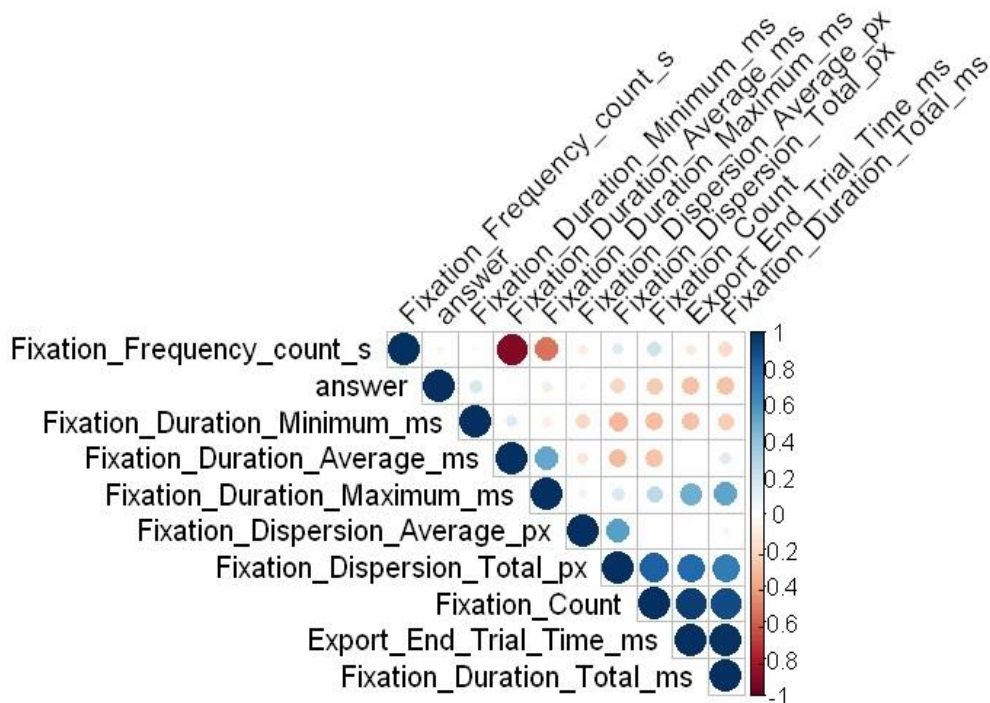


Figure 3: Correlation matrix between some of the features

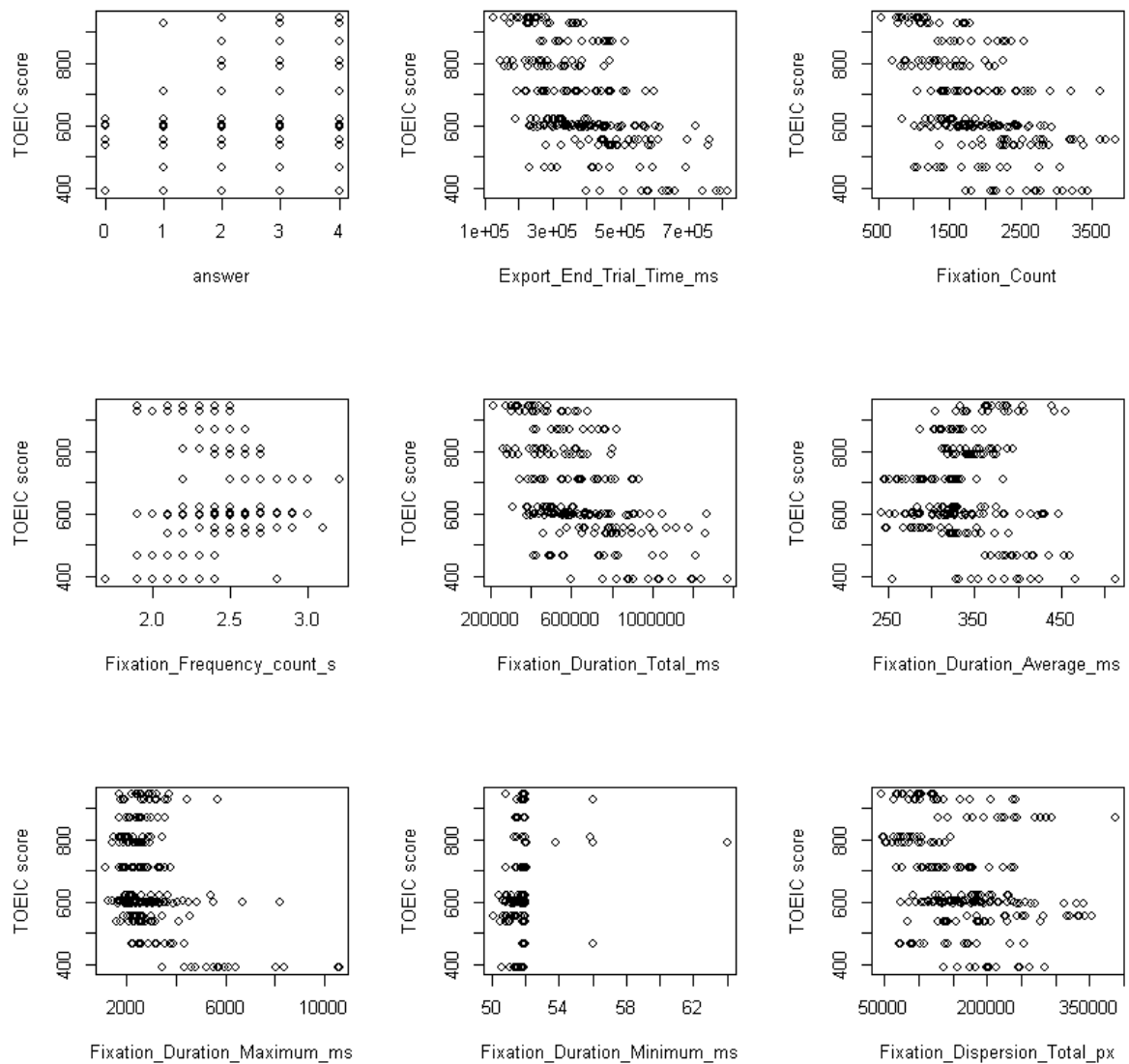


Figure 4: Scatter plots of some features and the response

Pearsonian correlation values between some features and response are as follows:

Export_End_Trial_Time_ms	Fixation_Duration_Total_ms
-0.561864559	-0.547996677
Saccade_Duration_Total_ms	Saccade_Count
-0.521099657	-0.520745747
Fixation_Count	Saccade_Amplitude_Total_deg
-0.502907650	-0.457584209
Saccade_Velocity_Total_deg_s	Scanpath_Length_px
-0.424885386	-0.423828414
answer	
0.390313619	

- Mean absolute error of the TOEIC score is considered in order to make comparison between various models

- Started by using Multivariate Regression to predict the TOEIC score. Initially built three models. One with only 'answer' and 'Export\_End\_Trial\_Time\_ms' as the predictor, another considered the first nine features having highest correlation with the response as the predictors, which also included both 'answer' and 'Export\_End\_Trial\_Time\_ms'. The nine variables having highest correlation with the response are listed above. The third is the baseline model which predicts the TOEIC score as the mean of TOEIC score of the train set.

Method used for fitting the above models are as follows :

- A subject is selected as the test set, the remaining are used to fit the model
- The mean absolute error of the TOEIC score is computed for the test set for various document combinations
- The process is looped through, for all the subjects in order to obtain an estimate of the average mean absolute error across readers for various document combination

The results of the models are as follows:

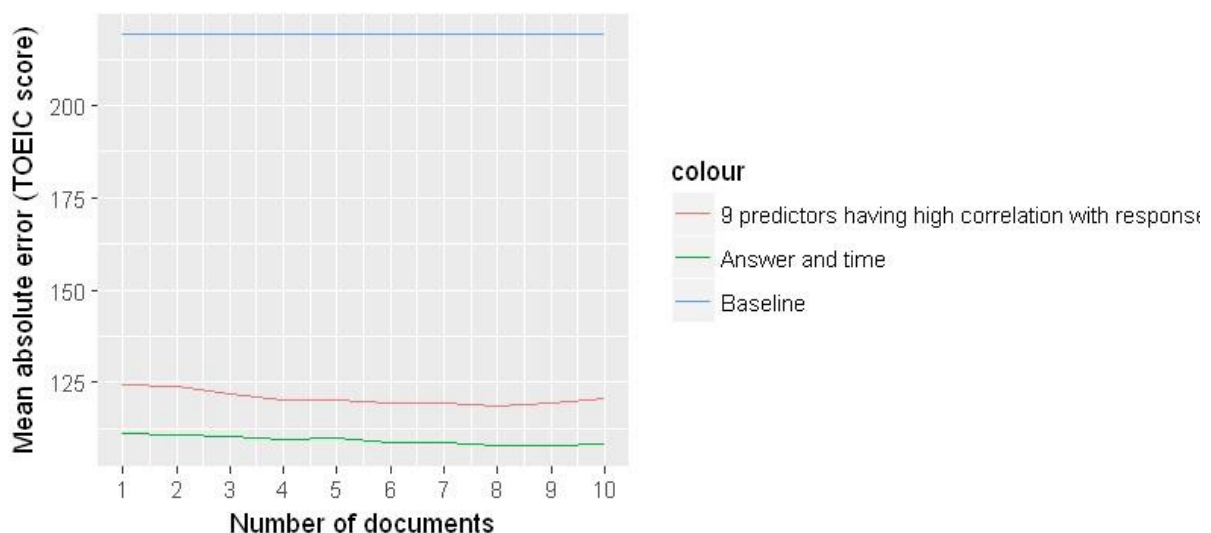


Figure 5: Results

- Next forward stepwise and backward stepwise feature selection was tried considering all the 34 features. Also, Lasso and Ridge regression was performed with the 9 features having the highest correlation with the response.

Method used for fitting the above models are as follows:

- A subject is selected as test, the remaining forms the train set
  - In the train set, a subject is selected as the Cross Validation (CV) set, the remaining are used to fit the model.
  - The CV set is used to tune the hyper-parameters. The process is looped through all subjects in the train set in order to fix on hyper-parameters that minimize the average CV mean absolute error of TOEIC score.
  - The model is fit on the entire train set with the selected



hyper-parameters

- Mean absolute error of the TOEIC score is computed for the test set for various document combinations
- The process is looped through, for all the subjects, to obtain an estimate of the average mean absolute error across subjects for various combination of documents

The results are as follows:

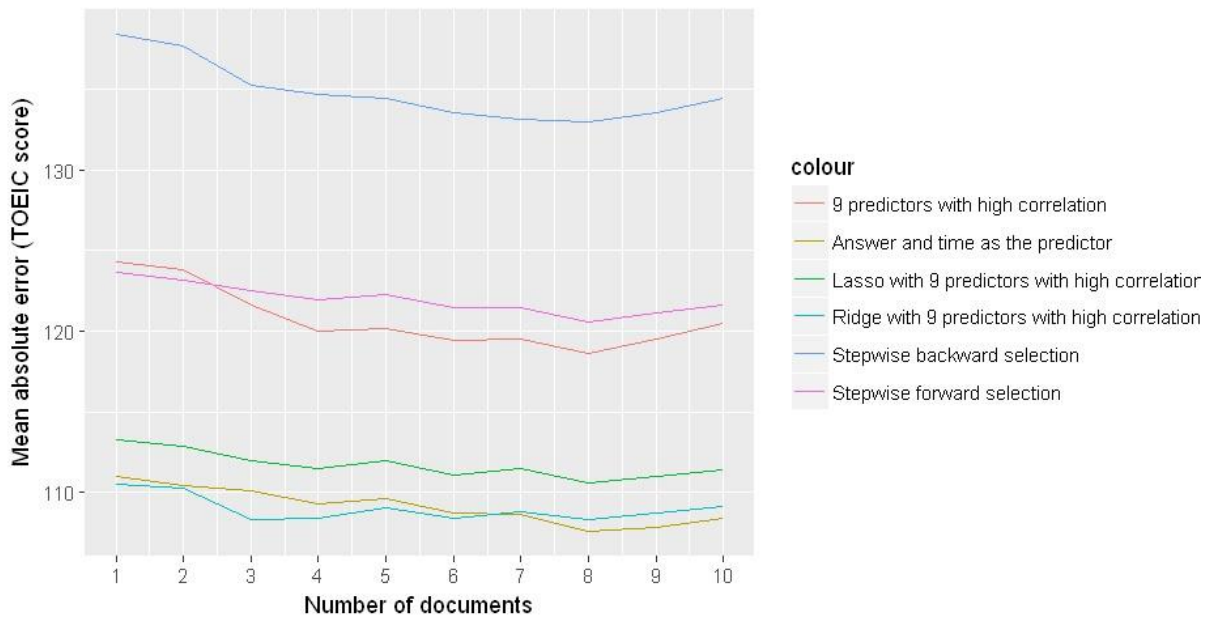


Figure 6: Results

- While trying tree based method of random forest with the nine features having highest correlation with the response, it is seen that some of the features related to saccades are resulting in the highest increase in the node purity i.e. highest reduction in the residual sum of squares. The result from the random forest is not that good. Hence following features are used to fit a Multivariate Regression, Lasso, Ridge and SVR model: 'Export\_End\_Trial\_Time\_ms', 'answer', 'Saccade\_Duration\_Total\_ms', 'Saccade\_Amplitude\_Total\_deg' and 'Saccade\_Count'. The method of fitting the mentioned models are similar to that of the earlier models. All hyper-parameters are tuned using the CV approach.

The results are as follows:

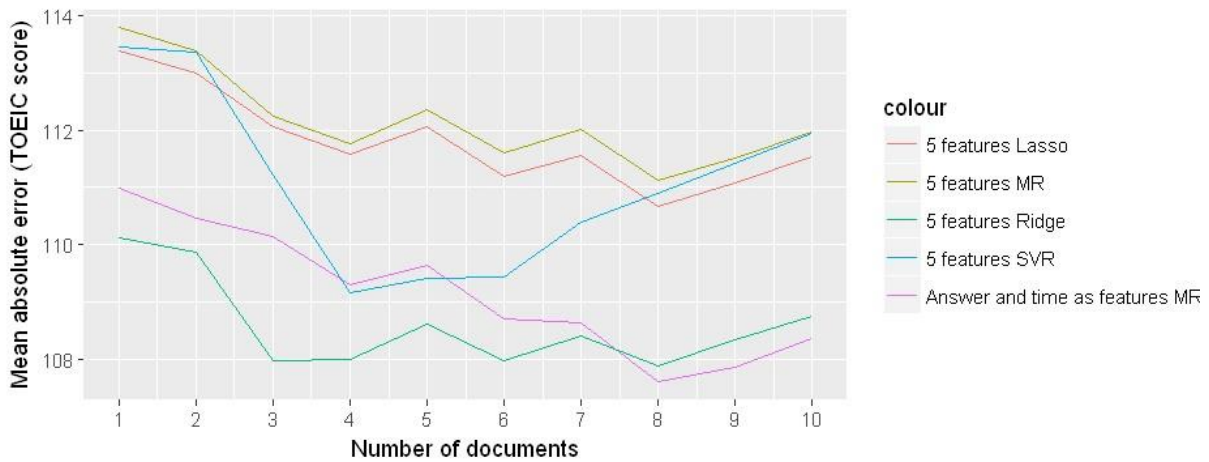


Figure 7: Results

## 6. Comments on the analysis:

- From the results it can be seen that no conclusive evidence was obtained about the eye gaze helping in the estimation of TOEIC score, from the current feature set in the Multivariate Regression setting and its variants
- The models fail to capture the within subject variability. Assumption is made that each data point is independent which might not be correct. This is because a subject might have some common reading behaviour across documents
- Very high level features are being extracted from the raw eye gaze patterns which might be leading to loss of essential information

## 7. Analysis of subject wise error:

In order to check performance of the models for various subjects, the average subject wise mean absolute error (averaged across document combinations) of TOEIC score was plotted against the corresponding TOEIC score of the subject. For all the models, following profile of the error curve was obtained with the y-axis error changing according to the type of the fitted model:

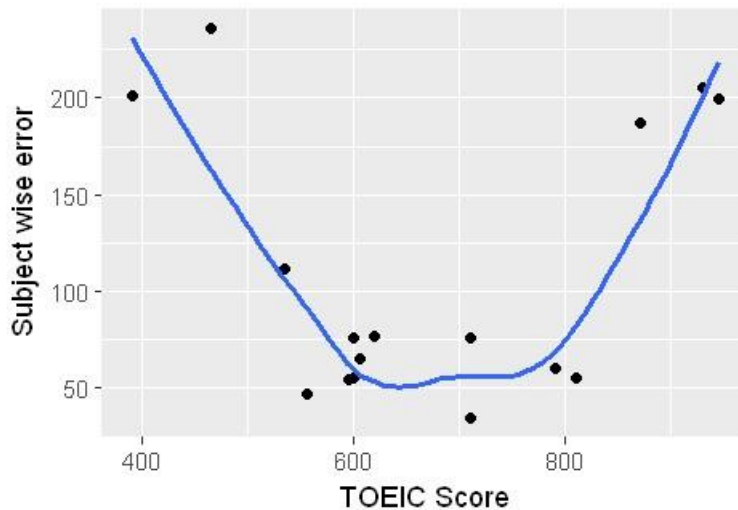


Figure 8: Subject wise error vs TOEIC score

From the above curve it is evident that models perform relatively better for subjects with TOEIC scores ~500 – 800. The models fail to perform well for subjects who have scored very high or very low in TOEIC.

## 8. Conclusion:

- In general there is a need for more involved analysis of the raw eye gaze pattern to extract informative features
- Create models / systems for automatic extraction of features from the raw eye gaze data
- Inclusion of the reading content might also help in improving the models and make them content / document independent

## **9. *References:***

- Towards an Automated Estimation of English Skill via TOEIC Score Based on Reading Analysis, 2016. Olivier Augereau, Hiroki Fujiyoshi and Koichi Kise
- An Introduction to Statistical Learning. Gareth James, Daniella Witten, Trevor Hastie and Robert Tibshirani