Educational Data Mining on Performance of under Graduate Students of Dibrugarh University using R

Sadiq Hussain
Examination Branch
Dibrugarh University
Assam India

Jiten Hazarika
Department of Statistics
Dibrugarh University
Assam, India

Pranjal Buragohain
Department of Education
Dibrugarh University
Assam, India

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

ABSTRACT

Most of the Universities and Educational Institutions of repute adopt educational data mining for the improvement of quality in education in general and students' improvement in particular. The research was carried out with the data of the students enrolled in different affiliated colleges of Dibrugarh University. The study explores the effect of performance on the basis of gender and caste. Another analysis was carried out to examine the trends of performance with respect to time using ARIMA Model. The authors also investigated impact of some of the socio-demographic factors on the performance of the students.

Keywords

Educational Data Mining, R Programming, ARIMA, Two-way ANOVA

1. 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. 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. (The Dibrugarh University Website)

There are more than one hundred fifty numbers of Colleges/ Institutes offering TDC (Three Year Degree) Course affiliated/ permitted under the Dibrugarh 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 2011 to 2013. This paper studies performance of students' gender wise as well as caste wise. The trends of the results of the B.A. course since 1970 till 2014 are also evaluated. Out of several data mining tools and statistical models available, this paper focuses on R programming and attempted to find out the statistical models for such knowledge discovery.

2. METHODOLOGY

2.1 Data Mining

Data Mining is an effective tool to extract meaningful and interesting patterns from the current and historical data stored in data warehouses which may be analyzed to predict future trends [1]. In today's world it is next to impossible to extract hidden patterns without the use of data mining tools and programs from the large data marts and data warehouses. It is like searching for diamonds from the mountains of data. The knowledge support system is based on such knowledge discovery from databases and helpful in decision making process of the organisations including future trend predictions.

Statisticians used manual techniques over the years to predict trends and analyzed the data for the benefit of the business houses and corporate world. The business houses produce huge amount of data everyday that becomes data tombs as the data is never transformed to information. But with the help of data mining tools and emerging research trends in this field, the data miner may extract knowledge from the large data marts / data warehouses very efficiently and quickly which may be used for the betterment of the organisations and the society.

2.2 Educational Data Mining

Data mining plays an important role in extracting hidden pearls from the sea of data warehouses [2]. The same is true for Educational Data Mining. Data mining is a part of Knowledge Discovery in Databases (KDD) process. Data mining had touched many fields including bioinformatics, e-commerce, fraud detection and lately in the field of education as well. The data mining in the field of educational research is known as Educational Data Mining (EDM) [3]. EDM often tries to simulate a student models which may be used for the improvement of students by predicting the future trends [4].

2.3 R Programming

R is a programming language for the purpose of statistical computations and data analysis. The R language is widely used by the data miners and statisticians on high dimensional pattern extraction. R's popularity has increased substantially in recent years which proved by the polls and surveys. R is inspired by Scheme and an implementation of S programming language combined with lexical scoping semantics. S was designed by John Chambers while at Bell Labs. The creator of R was Ross Ihaka and Robert Gentleman at the University of

Auckland, New Zealand. R is named after the first names of the two creators.

R is freely available under the GNU General Public License and the source code is written in FORTRAN, C and R. It is a GNU project. The pre-compiled binary versions are freely available for various flavors of operating system. R is basically command line interface (CLI) and various GUI interfaces are also available nowadays.

R provides numerous statistical techniques from modeling to analysis, clustering, classification and the list goes on. The packages developed by the R community plays an important role in this regards. The C,C++, Java, .NET or Python programmers may write their own code to manipulate the R objects. Advanced users may use algorithms of their choices for any computationally intensive tasks.

Graphical packages are also available in R. R produces dynamic, interactive and publication quality graphs for the data miners and statisticians. [5]

2.4 R Studio

R Studio is a free and open source integrated development environment (IDE) for R, the statistical computing language for the data miners. There are two editions of R Studio. One is R Studio Server, which may be accessed through web browser from a remote Linux Server. Another Edition is R Studio Desktop which available for Microsoft Windows, Mac OS X, and Linux. R Studio Desktop runs locally. R Studio uses the Qt framework for the GUI and is written in C++ language. [6]

2.5 Two-way ANOVA

For comparing the means of populations that are classified in two different ways, or the mean responses in an experiment with two factors, the two-way ANOVA is used. The aov() function is used to fit two-way ANOVA models in R. For example, the command:

```
> aov(Response ~ FactorA + FactorB)
```

fits a two-way ANOVA model without interactions. In contrast, the command

```
> aov(Response ~ FactorA + FactorB + FactorA*FactorB)
```

includes an interaction term. Here both FactorA and FactorB are categorical variables, while response is quantitative.

Two-Way Analysis of Variance (ANOVA) is a technique for studying the relationship between a quantitative dependent variable and two qualitative independent variables. Usually it is interesting to find whether the level of the dependent variable differs for different values of the qualitative variables.

2.6 ARIMA

ARIMA (autoregressive integrated moving averages) models, commonly known as the Box–Jenkins approach is a forecasting model. It comprises of three stages. They are model identification, parameter estimation and diagnostic checking. These steps are repeated until an appropriate model is identified for prediction. At the very outset the time series data is analyzed to find out the autocorrelations (ACF) and partial autocorrelations (PACF). R provides acf and pacf functions for this purpose.

In the ARIMA (p,d,q) model, where the p, d, and q are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. There are many attempts over the years to automate ARIMA modelling. For a stationary series, Hannan and Rissanen [7] proposed a method to recognize the order the ARIMA model. The innovations can be obtained by fitting a long autoregressive model to the data, and then the likelihood of potential models is computed via a series of standard regressions by using their method. Gomez and Maravall [8] developed software TRAMO and SEATS for automatic identification procedure. , the algorithm attempts to find the model with the minimum BIC for a given series.

Liu [9] proposed an algorithm that is used in the SCA-Expert software for identification of seasonal ARIMA models using a filtering method. Melard and Pasteels [10] develops another approach for univariate ARIMA models that allows intervention analysis. The name of the software package where their method is implemented is called “Time Series Expert" (TSE-AX). Another excellent automatic ARIMA algorithm is Forecast Pro [11]. The algorithm was in the M3-forecasting competition [12]. Ord and Lowe [13] review some of the commercial software having facility of automatic ARIMA forecasting. Reilly [14] develops another proprietary algorithm which is implemented in Autobox.

3. EXPERIMENTS

3.1 The Dataset

The authors had included a small part of the Category and Gender based tables termed as Table II for the data analysis. The Examination Branch of Dibrugarh University provides various College Codes for different Colleges under its jurisdiction. The fields and their meaning are given below:

Table –I : The Field Name of the Data with their meaning

Field Name	Meaning
ExamName	Final Examination B.A. Part-III.
Year	Year of Examination
Gender	Sex of the Candidate
Category	Caste of the Candidate
Pass Percentage	Aggregate Percentage of the particular Candidate for all the three year examinations viz, B.A. Part-I, II and III Examinations
First Class	Pass Percentage is above 60%
Second Class	Pass Percentage is above 45%
Simple Pass	Pass Percentage is below 45% and above 30%.

Table –II: Year wise Gender wise Category wise College wise Pass Percentage of Students

Exam Name	Year	Gender	Category	CollegeCode	PassPercentage	Result
B.A. PART-III	2011	F	General	115	45.56	II
B.A. PART-III	2011	M	General	115	35.33	P

The Following is the SQL Procedure to extract the data from the Microsoft SQL Server Database.

```
create procedure TDC_Querry
@Year varchar(4),           --like 2013, 2014
@CoCode varchar(5),                 --BA or BSc or BCom
@Sub varchar(6),            --(optional) Subject code - Branch code in case of BCom
@Gender varchar(5),         --(optional) M/F
@Cate varchar(5)            --(optional) Cate Code Like ST, SC, MOBC
as
set nocount on
if exists (select * from dbo.sysobjects where id = object_id(N'[tmpData]') and OBJECTPROPERTY(id, N'IsUserTable') = 1)
drop table [tmpData]
declare @ShId int, @EmId int
select @ShId = ShId, @EmId = EmId
from Schedules, Sessions,
(select EmId = Max(EmId) from Exams where EmCoCode = @CoCode) as Em
where ShEmId = EmId and SnNo = ShSnNo and SnYear = 2013

select RgId, EfId, ExamName = EmName, ExamYear = @Year, CandName = RgName,
        Gender = RgSex, Category = CdDesc, ColgCode = RgCgCode,
        Sub = StrCode, MajorSub = StrName, TotalMarks = EfAObtMarks,
        FirstYr = 0, SecondYr = 0, ThirdYr = 0, OverallPC = EfDivPc,
        Division = EfDiv
into tmpData
from ExamStr, ExamForms, RegnForms, Exams, CodeTable
where StrCoCode = @CoCode and StrCode <> 'GEN'
        and EfShId = @ShId and EfStrId = StrId
        and EfResult = 'P' and EfExcp = "
        and RgId = EfRgId and EmId = @EmId
        and CdType = 'CS' and CdSeq <> 0 and CdCode = RgCaste
update tmpData set Sub = EsCode, MajorSub = EsName
from FormSub, ExamSub
where FsEfId = EfId and EsId = FsEsId
        and EsType = 'S' and EsName like 'MAJOR%'
update tmpData set FirstYr = ObtMarks
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 1
update tmpData set SecondYr = ObtMarks
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 2
update tmpData set ThirdYr = ObtMarks
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 3
select top 1000
        ExamName, ExamYear, CandName, Gender, Category, ColgCode,
        MajorSub, TotalMarks, FirstYr, SecondYr, ThirdYr, OverallPC, Division
from tmpData
where (Sub = @Sub or @Sub = ")
        and (Gender = @Gender or @Gender = ")
        and (Category = @Cate or @Cate = ")
order by EfId + RgId
return (0)
```

The R Codes for the first Experiment i.e. the trend analysis of B.A. students of the Colleges affiliated to Dibrugarh University are given below. The Data was collected for 45 years of pass percentage of the students starting from 1970. ARIMA Model was used for the trend analysis and the forecasting. The following steps are adopted for this.

1. The dataset in the form of .csv file is imported to the R Studio environment.
2. Store the data in a time series object in R using ts ( ) function.
3. The time series data is plotted using plot.ts ( ) function.

The commands are given below:

```
> PassPercentage1 <- read.table("C:/reseach/PassPercentage1.csv", header=TRUE,quote="\"")
> View(PassPercentage1)

> pp <- ts (PassPercentage1, start=c(1970))

> plot.ts(pp)
```

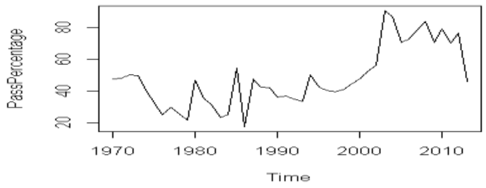


Figure –1: plot.ts(pp) output

The trend component of the time series was estimated by smoothing using a simple moving average. SMA( ) function in the TTR package of R is best suited for this job. The following commands in R to smooth the time series using a simple moving average of order 3 and to plot it.

```
> library(TTR)

> ppSMA3 <- SMA (Passparcent,n=3)

> plot.ts (ppSMA3)
```

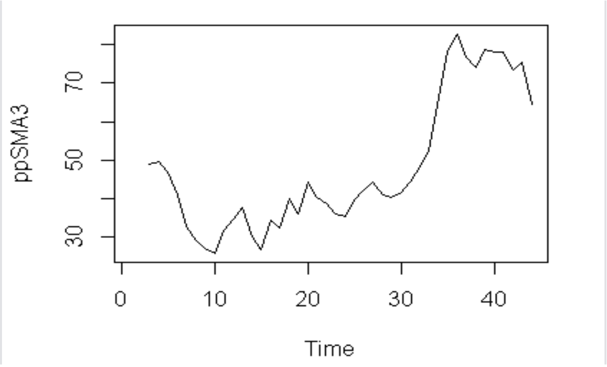


Figure –2 : plot.ts (ppSMA3) output

To get a clearer picture of trend component, a simple moving average of order 8 was applied. The pass percentage had decreased from 40 to below 30 after 1980 and gradually performing well up to 80 percentage till 2013. After introduction of Semester System, there was a sharp decline in 2014.

```
> ppSMA8 <- SMA (Passparcent,n=8)

> plot.ts (ppSMA8)
```

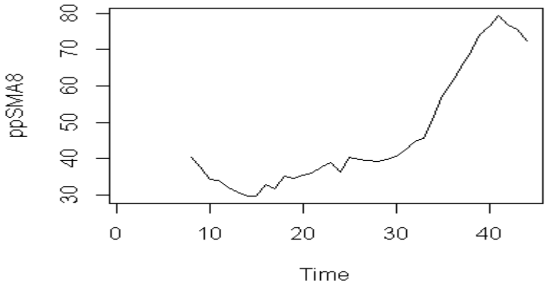


Figure –3 : plot.ts (ppSMA8) output

As time series was described using an additive model with constant level and non seasonality, simple exponential smoothing was used to make short term forecasts. The simple exponential smoothing method provides a way of estimating the level at current time point. The smoothing is controlled by the parameter alpha and the value lies between 0 and 1. The authors had fit simple exponential using HoltWinters ( ) function in R. The beta and gamma parameters were set to false. The following are the R code:

```
> ppfor <- HoltWinters (pp,beta=FALSE,gamma=FALSE)

> ppfor

Holt-Winters exponential smoothing without trend and
without seasonal component.

Call:

HoltWinters(x = pp, beta = FALSE, gamma = FALSE)

Smoothing parameters:

 alpha: 0.5123064

 beta : FALSE

 gamma: FALSE

Coefficients:

     [,1]

a 60.25329
```

The authors had plotted the original time series in black and forecasts as a red line. The time series of the forecasts were much smoother.

```
> plot(ppfor)
```

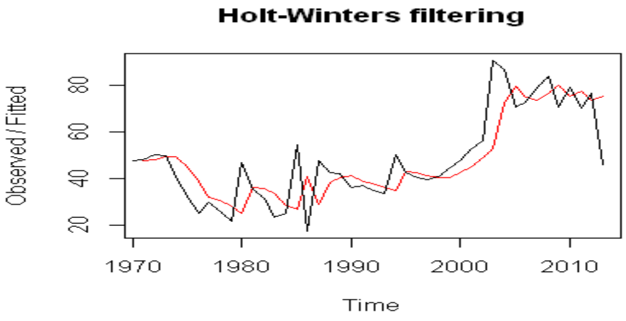


Figure –4 : plot(ppfor) output: Holt-Winters Filtering

The authors used the forecast package in R for ARIMA modelling. The auto.arima ( ) function was used to find the appropriate ARIMA model. The appropriate model found was ARIMA (0,1,1) with p=0,d=1,q=1 where d was the order of differencing required. The R code is as follows:

```
> library(forecast)

> auto.arima(Passparcent)

Series: Passparcent

ARIMA(0,1,1) with drift

Coefficients:

          ma1   drift

     -0.4853  0.3053

s.e.   0.1382  0.9755

sigma^2 estimated as 145.3:  log likelihood=-163.79
```

```
AIC=333.57   AICc=334.19   BIC=338.85
```

The authors had differenced the time series once using diff() function and plotted it.

```
> ppseriesdiff1 <- diff(Passparcent, differences=1)

> plot.ts(ppseriesdiff1)
```

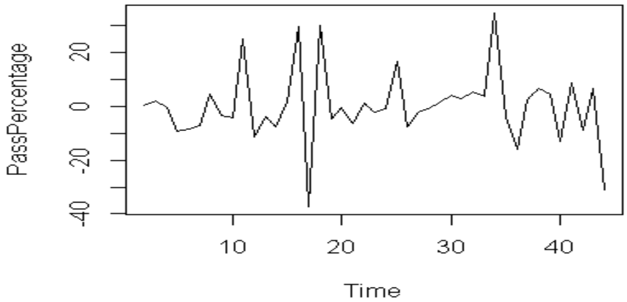


Figure –5 : plot.ts(ppseriesdiff1) output

The time series did not appear to be stationary in mean. So, the authors differenced the time series once. The resulting time series of first difference appeared to be stationary in mean.

The Authors used acf( ) and pacf ( ) functions in R to plot a correlogram and partial correlogram for lags 1-20 respectively.

```
> acf (ppseriesdiff1,lag.max=20)
```

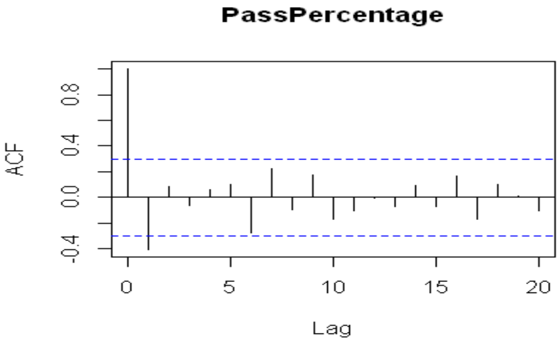


Figure –6 : acf (ppseriesdiff1,lag.max=20) output

```
> pacf (ppseriesdiff1,lag.max=20)
```

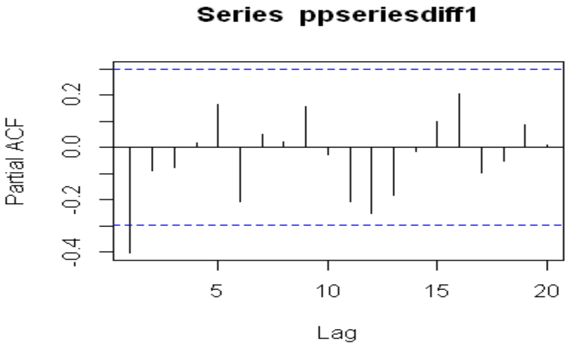


Figure –7: pacf (ppseriesdiff1,lag.max=20) output

The authors had fitted an ARIMA (0,1,1) model to the time series. An ARIMA (0,1) model is written as X_t – mu = Z_t – (theta-Z_t-1), here the theta parameter needs to be estimated. The value of theta given as ma1 in R output is -0.4768.

```
> ppseriesarima <- arima(pp, order=c(0,1,1))

> ppseriesarima

Series: pp

ARIMA(0,1,1)


Coefficients:

          ma1

      -0.4768

s.e.   0.1329

sigma^2 estimated as 142.3:  log likelihood=-167.73

AIC=339.47   AICc=339.77   BIC=342.99
```

The authors had used forecast.Arima() function to make future forecast for next 5 years. The ARIMA model had forecasted pass percentage for the next 5 years as 59.93. The authors had plotted the next 5 years forecasted value as well.

```
> ppseriesforecasts <- forecast.Arima(ppseriesarima, h=5)

> plot.forecast(ppseriesforecasts)
```

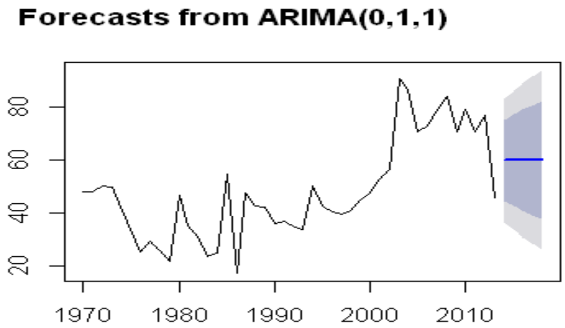


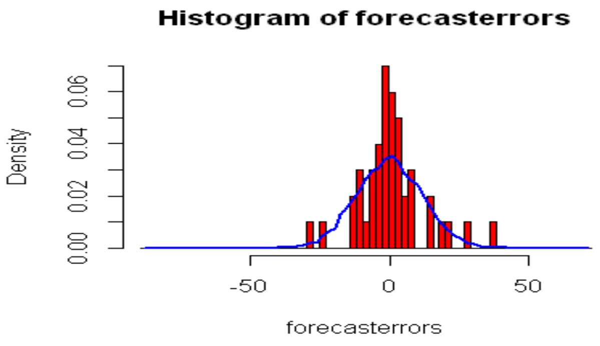
Figure –8: plot.forecast(ppseriesforecasts) output

The authors had also performed the Ljung-Box Test for lags 1-20.

```
> Box.test(ppseriesforecasts$residuals, lag=20, type="Ljung-Box")

Box-Ljung test

data:  ppseriesforecasts$residuals

X-squared = 13.3585, df = 20, p-value = 0.8615
```

The histogram of the time series showed that the forecast errors are normally distributed and mean was closed to zero.

```
 > plotForecastErrors(ppseriesforecasts$residuals)
```

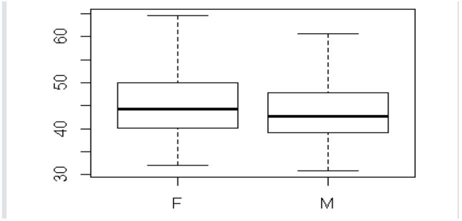


Figure–9 : plot Forecast Errors (pp series forecasts $residuals) output

3.2 The R Code for the Second Experiment

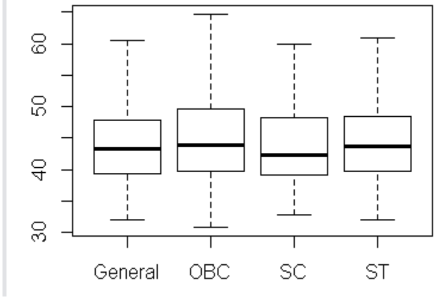
The .csv file gender_cat was imported to the R studio for finding out the performance parameter in respect to gender and category i.e. General, SC,ST, OBC. The authors were interested in whether the level of the dependant variable ‘Percentage’ differed for different values of the qualitative variables ‘Gender’ and ‘Category’.The authors tested for two way ANOVA, stored the results in a variable, and then the summary of those results were generated.

```
> Category <- as.factor (gender_cat$Category)

> Gender <- as.factor (gender_cat$Gender)

> Year <- as.factor (gender_cat$Year)

> Percentage <- as.double (gender_cat$Percentage)

> tapply (Percentage,Gender,mean)

     F       M

45.6416 44.0731

> tapply (Percentage,Category,mean)

 General     OBC       SC       ST

44.41533 45.41272 43.98241 44.55798

> attach(gender_cat)

> int <- aov(Percentage~Category*Gender)

> summary (int)

                 Df Sum Sq Mean Sq F value   Pr(>F)

Category           3    740   246.8   4.844   0.0023 **

Gender             1   1708  1708.0  33.520 7.79e-09 ***

Category:Gender    3    160    53.3   1.046   0.3711

Residuals       2992 152461    51.0

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> boxplot (Percentage~Gender)
```



Figure–10 : boxplot (Percentage~Gender) output

```
> boxplot (Percentage~Category)
```



Figure–11 : boxplot Percentage~Category)output

```
> pairwise.t.test (Percentage,Gender,p.adjust.method="none")

Pairwise comparisons using t tests with pooled SD

data:  Percentage and Gender

  F

M 2.3e-09

P value adjustment method: holm

> pairwise.t.test
(Percentage,Category,p.adjust.method="none")

Pairwise comparisons using t tests with pooled SD

data:  Percentage and Category

 General OBC    SC

OBC 0.0072  -      -

SC  0.5000  0.0151 -

ST  0.7210  0.0053 0.3430

P value adjustment method: none
```

4. EVALUATION

The study reveals that both Holt-Winter’s filtering and ARIMA (0,1,1) models can be used as competing models for studying performance of students at UG level of Dibrugarh University. However, from the diagrams it is observed that Holt-Winter’s filtering will provide better forecasts, whereas the ARIMA (0,1,1) model has forecasted past percentage for the next five years as constant values (59.93%). The increasing trend has been noticed in the performance of the study subjects with exception in few years. The results of two independent samples t-statistic used for testing whether there is significant difference in the performance of the study subjects as a whole with respect to sex, it is observed that the value of t-statistic is significant (p-value<0.01) at 1% level of significance. Using two-way ANOVA, it is observed that the values of F-statistic for gender as well as category (caste) are significant. To pinpoint the source of variations using two independent samples t-test pair-wise, these differences have been noticed between General and OBC; OBC and SC; and

OBC and ST only. In this context, an interesting result has been observed that there is no significance difference between General and SC; General and ST which is usually not expected. The mean performance of OBC students is 45.41%, whereas the same is 44.41% for General, 43.98% for SC and 44.56% for ST community. The findings of p-values based on t-test shows no significant difference in the performance of the first class holders with respect to time, whereas the same is significant among the second class holders. It reveals that as a whole there is significant difference in the performance of the students with respect to time. Here female pass percentage (45.64%) is better than male counterpart (44.07%).

5. DISCUSSION

The result of the study confirmed that there was a gradual improvement of the results of the students of Dibrugarh University in their bachelor degree examination from1980 to 2013. But there was a sharp decline of achievement of the students after the introduction of semester system. The authors hypothesized that this decline is because of an immediate change of the system. From the previous observations the authors inferred that immediate change of a system creates a short term adjustment problem and affects academic achievement. However, the analysis of the results of the future years will confirm whether this decline is because of the adjustment problem with a new system or anything else. The analysis shows that as a whole the achievement of the female students are better than the male students. The analysis of the results shows that there is no significant change in the result of the first class holders with the change of time whereas there is a change among the second class holders and simple pass students. The analysis further shows that there is no significant gender difference and caste difference among the first class holder students whereas there is a significant difference among the second class holder and simple pass students. The authors observed that there is a continuous supervision of the teachers and parents especially for the students showing good academic achievement. Teachers give importance to such students. Society also takes care of them. This pushes them to carry on the same result or to improve the result in future. So, their results are either steady or better than the earlier. Whereas the result of the second class holder and simple pass students show that they have been lacking a favour of continuous supervision. The authors proposed that if they are supervised and motivated continuously there will be a continuity of the achievement among the second class holder and simple pass students. The authors further concluded that gender and caste are not determining factor if students get proper supervision and motivation. On the other hand gender and caste can be determining factor if the students are low achiever and lack proper supervision. For future study, authors planned to study different examinations results with different parameters.

6. 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.

7. REFERENCES

[1] John Silltow, (2006): Data Mining 101: Tools and Techniques, http://www.internalauditoronline.org/

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

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

[4] 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.

[5] http://en.wikipedia.org/wiki/R_(programming_language)

[6] http://www.stat.columbia.edu/~martin/W2024/R8.pdf

[7] Hannan EJ, Rissanen J (1982). "Recursive Estimation of Mixed Autoregressive-Moving Average Order." Biometrika, 69(1), 81-94.

[8] Gomez V, Maravall A (1998). "Programs TRAMO and SEATS, Instructions for the Users." Working paper 97001, Ministerio de Economia y Hacienda, Direccion General de Analisis y Programacion Presupuestaria.

[9] Liu LM (1989). "Identification of Seasonal Arima Models Using a Filtering Method." Communications in Statistics: Theory & Methods, 18, 2279-2288.

[10] Melard G, Pasteels JM (2000). "Automatic ARIMA Modeling Including Intervention, Using Time Series Expert Software." International Journal of Forecasting, 16, 497-508.

[11] Goodrich RL (2000). "The Forecast Pro Methodology." International Journal of Forecasting, 16(4), 533-535.

[12] Makridakis S, Hibon M (2000). "The M3-Competition: Results, Conclusions and Implications." International Journal of Forecasting, 16, 451-476.

[13] Ord K, Lowe S (1996). "Automatic Forecasting." The American Statistician, 50(1), 88-94.

[14] Reilly D (2000). "The Autobox System." International Journal of Forecasting, 16(4), 531-533.