ISSN No: - 2456 – 2165

A Comparative Study on Prediction Models for Educational Institution Rankings

Akarsh Prabhu K.
SJCE Mysuru
aks7695@gmail.com

Anukarsh G. Prasad
SJCE, Mysuru
anukarsh.prasad@gmail.com

N. R. Prashanth
Compiler Tree Technologies
naduthota@gmail.com

Dr.S.K.Padma
SJCE, Mysuru
skp@sjce.ac.in

***Abstract* - Recent Trends in decision making have made use of various prediction models and algorithms to ensure accurate decisions. The prediction models are based on historic data aggregations. Statistical analysers train the models and portray results as predictions using need specific algorithms. The standout models use adaptive techniques to blend irregular variations in the data into the model; usage of weights during model design institutes the same. This paper conducts a thorough comparative study on different prediction models. It presents the results of time-series prediction models and conclusively shows implementation of the linear regression modelling for Common Entrance Test results of Karnataka Education Authority. The model is built to forecast the future occurrences and rank colleges based on the predictions and compares the results of the three fore-mentioned modelling techniques for the application. It incorporates outlier detection techniques for data pre-processing. Further the data is optimised by assigning specific weights to particular data series based on trends. The paper introduces a breakthrough in the genre of decision support for career planning.**

Keywords—Linear Regression Model, Time-series, Rank Predictor, Holt-Winters, ARIMA, R Programming.

I. INTRODUCTION

Everyday humans make a lot of predictions, maybe knowingly or unknowingly, like “Will it rain today?” “Probably not” or decisions on the route from office to home. But its not the ideal way of making all decisions, the decisions which are made daily are based on personal intuitions and some recent occurrences. As the importance of the decisions increase or the value associated with the decision increases, more analysis will go into making that particular decision. So the machines are used to obtain predictions for the future trends which are analysed to make an informed decision, such decisions would be ideal as the repercussions of the decisions are at least roughly known.

The success of a career is based on crucial decisions made during initial stages of life. The advent of modern era with immense competition among students to select the best institution and among the institutions to be the best institution presents a need for an automated algorithm which predicts the trends of growth of an institution. The students are posed with hundreds of options in terms the choice of college or a course. But the decision has be to made with utmost information available as the choice shapes the future of the student. Such a decision would need a lot of background research, the predictions on the same for upcoming years. This process cannot be made manually as its prone to errors which is unacceptable and also very cumbersome.

The paper is aimed at automated learning of trends which might occur in historical data using three algorithms and making a comparative study on the results of the same and obtaining a best fit model from the accuracy of predictions, two sets of predictions are considered for evaluation namely the top ten engineering college in the State of Karnataka and another being the cut-off ranks of particular courses in all colleges. The model is further improved by reinforcements in terms of distributed weight-age on the trends present in the data set. The proposed system considers trends and thus any data entry which are outliers or noise points in the dataset can make the prediction model fail tragically, so a careful set of pre-processing and formatting should be implemented before the learning phase to achieve desirable results.

The theory of time series was first thought off early with the development of stochastic processes[1]. The moving averages concept was used to remove periodic fluctuations in the time series. Then the classic book Time Series Analysis by Box [2] gave a thorough description of the modelling procedure. So by 1970 the Box-Jenkins Model became widespread for a lot of applications as it covers most forecasting and adjustments with respect to seasons. The first generalization was the ARMA model by Whittle (1951) [3] which could work only with stationary models. Later dynamic models were built which had wider applications. In the case of regression models, these started in with least square methods introduced by Gauss(1809) and Legendre(1805). Regression as a word was first used by Galton(1908) when he spoke about heredity and mediocrity in biology [4].

II. CHARACTERISTICS OF A GOOD PREDICTOR

Any good predictor is supposed to use the best fit model for training and make accurate predictions based on user requirements and present the results in the simplest way possible to enable easy analysis.

ISSN No: - 2456 – 2165

An ideal case situation for the proposed topic would be an algorithm which would make the most accurate predictions of trends as well as the cut-off ranks and top colleges list. The model evaluation should lead to an error within the acceptable range. The algorithm should adhere to time and space complexity constraints and work with provided basic system resources. It should be platform independent and shall be scalable based on needs. The inclusions of any new features should not hinder the accuracy and should maintain uniformity in analysis at all times. The model selection or accuracy that is expected off the predictor is never a constant and is still a subject of research. The accuracy is modelled based on the sensitivity of the data. For example a prediction about the income of a small scale industry which is yet to break even would be very sensitive as the company is not capable of handling any loss but the same prediction for a multinational company with abundant funds would not be sensitive, as the company will be able to bear a loss in the range of millions. Thus an ideal predictor should be able to sense the data sensitivity and should maintain an accuracy based on it. Although the development of an ideal predictor might not be practically possible due to lack of ideal conditions. The paper however tries to bring out most of the desirable features in the proposed model.

III. EXISTING SYSTEM

An institutional ranking website would take into consideration the previous year results and makes a certain number of assumptions on such data. The predictor should be basically able to predict and provide a user-friendly statistical picture. The present system for students who need to find each college's status is to visit kea.kar.nic.in website and go through unstructured documents having all the cut-off ranks of college and category wise in a single PDF in a tabular form. This kind of formatting makes it difficult for the user to deal with such enormous data. It also is a challenge to integrate multiple year data and draw inferences on which institute is consistently successful.

A better approach in dealing with this data has been compassed by a student made website called gopuc.com. Here all the data is well organised and presented with good filters. In the section where our paper ventures, the website doesn't manage to predict the cut-off ranks as such as it simply gives conclusions based on the immediate previous year's data. If unfortunately the previous year was the outlier year, then all the conclusion drawn will be very inaccurate. The website also provides a list of top colleges based on the previous year cut-off ranking solely. This will give a biased comparison as a college with certain adverse conditions like a student's untimely withdrawal of admission into the college might cause an outlier entry in the ranking cut-off , judging the rank of the college based on such conditions would not be ideal. Thus the decision based on such adversities would result to erroneous results.

After observing such existing systems we can conclude that, for successful predictions to be made, a thorough analysis is to be made of the data. One cannot simply say that every year is going to be same as the previous years.

IV. PROPOSED SYSTEMS

Due to the limitations of the existing system and the inaccuracies in predictions, a new model is proposed for the educational institutional ranking. The proposed system can be broken down into modules like in which the first module would be the pre-processing which is to order, normalise and format the data. Due to highly unstructured nature of the data that is available this step is a necessity. The next module is the model selection i.e., the picking of a training and prediction models for the dataset depending upon the characteristics. The first set of models considered were the time-series models as the data aggregation was yearly and this confers to form a time-series data collection. The ARIMA and the Holt-winters were the models which were considered where the dynamic nature of the ARIMA model favoured our calculation as the data was very much irregular. The next model considered was the linear regression model, which uses arithmetic functions to analyse and fit the data into models. The linear regression model also failed under some conditions due to surges in the data. The next component was designed to minimise errors by weighting the various sections based on the trend analysis and human intuition. The results section thus give a comparative viewpoint to the various algorithms that are considered and suggest the best model in case of such data acquisitions.

A. Pre-processing

Data pre-processing plays an important role in obtaining accurate prediction in data analytics. First and foremost getting the dataset in the right format is necessary. The data is tuned suitably for the purpose. Further it is sometimes possible that there will be certain years when that one person will get admitted to a college even though he does not meet the usual cut off scores trend of that course during previous years. This might be due to certain policy changes in the educational institution or due to untimely withdrawal by a candidate from the course. Hence the cut-off rankings for such courses show a sudden surge and disturb the data. The data of previous KCET exams held by K.E.A was obtained from their official website. The data under consideration is the final round cut-off ranks of the present top 10 colleges in Karnataka for the basic branches Electronics and Communication Engineering, Computer Science and Engineering, Electrical and Electronics Engineering, Mechanical Engineering and Civil Engineering for the General Merit Category from 2009 to 2014. As the available data was not formatted, a manual formatting process was carried out to structure the data for suitable needs. The procedure followed by us for handling anomalies in data is to detect it using box plots as shown in Fig. 1.

ISSN No: - 2456 – 2165

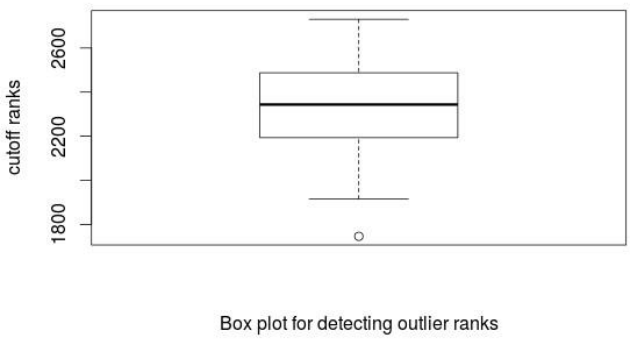

Fig. 1: Box Plot representing SJCE cut-off ranks.

The method used for detecting anomalies is to find the quartiles of the vector containing the cut-off ranks. Any point outside of the whisker's limit, which denote the quartiles, are called outliers. These are either eliminated from the dataset or noted and later ignored.

B. Model Section

The detailed discussion of the shortcomings and the assumptions made by the existing system poses a need for a better model and better analysis. This paper provides a comparative study of two time series model namely Holtwinters and ARIMA and then moves onto a linear regression model. Time series analysis is the study or probing for the internal structure of data which is assimilated over a defined period of time. A time series data may consist of various trends , seasonal patterns or correlations which are useful to train the models and use them for predictions. Our implementation of this variant is using the constructs mentioned above, the CET dataset was trained by a split of 5:1 using the models and then the prediction was made based on the findings of those models. This approach gave us a detailed understanding of the trends which have occurred in the cut-off variations over the time period of 2009 to 2010. The shortcomings of this model was found to be that the absence of trends in the data provided would lead to inaccurate predictions and the CET dataset has shown absence of continuous trends in the time period considered and were composed of many irregularities. Consideration of the irregularities of the time-series ultimately gives the better accuracy. The results presents a clear picture of the accuracy of these models. Linear regression is a mathematical approach used to correlate two or more variables irrespective of trends or internal structures that are present. The detailed description of the models and their implementations are provided in the sub-sections.

1) Holt-Winters: The pre-processing step formats the data in the required format. In time-series analysis requires the data to be smoothed before a model fitting is implemented. Smoothing is the process of replacing each data entry with the average of all the nearest data points [5]. Smoothing is used make sure the irregular component of the fitted time series data is nullified and a smooth plot is obtained for analysis. The Holt's model uses exponential smoothing which is the usage of more than two pass filters with exponential window functions to smooth the data so that the analysis of the trends are easier[6]. The Holt-Winters smoothing makes the model learn the slope at the specific data point, to accomplish that task two variables called alpha and beta are used. The alpha variable is used to recognise and present the level of the data and the beta variable is used to analyse and present the trend of the data. The implementation function uses a Holt-Winters function with the gamma variable set to false as only alpha and beta are the only variables used. The function is used to fit the dataset w.r.t to the time period specified.

The fitted model is then used to forecast or predict the future events by using the forecast function with the HoltWintersfilter. The forecast is produced with minimum and maximum errors which are expected to occur.

The Fig. 2. shows the CET dataset fitted with the HoltWinters model and the forecast is made using the same model. The forecast in the figure is shown with the possible minimum and maximum errors. It can be clearly seen that trend which is observed in the BMSCE college cut-off ranking is the major factor to identify and predict the model statistics. The predictions however cannot be controlled by any weights or specific altercations, the model automates both the training and the prediction models thus leaving no room for programmer interaction or modification.

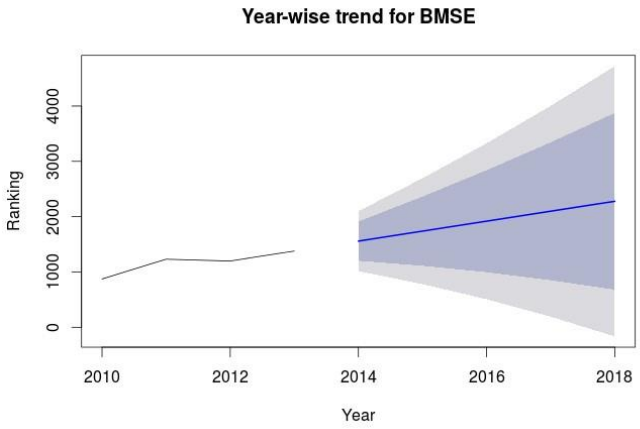

Fig. 2: Holt-Winters model fitting and prediction

The model specifies the prediction to be based on the trend and the absence of that can lead to smoothing to generate trends and thus it will lead to higher errors in the predictions made by the model.

2) ARIMA: The Holt-Winters model uses exponential smoothing methods and do not make any assumptions about correlation present in the data elements. But in some cases where

ISSN No: - 2456 – 2165

data is not seasonal or does not contain trends evenly it is better to consider them as it would generate a better predictive model. The Auto-regressive Integrated Moving Average(ARIMA) [7] is used so do that as it incorporates the irregular component of time-series and it finds correlations between the neighbouring data elements there as a way of learning. The model trains the data and forecasts the futuristic results based on the same.

The ARIMA model[8] varies based on three variables namely p,d and q. The variable p represents the order of the auto regressive model i.e., a model where the a random model which best describes certain time-varying distributions in the practical world. The auto-regressive model[9] considers that each output variable will be linearly dependent on other previous variables in the fitted model. This is very similar to the concept of linear regression which is explained in the other section. The variable d is the order of differencing of the timeseries required to obtain a stationery time-series, a stationery time-series would be a distribution with a mean of roughly zero. The variable q would be the order of the moving average model. A moving-average model is a linear regression of the current value or the predicted value in comparison with the current and the previous present noise points or as we call the irregular component. The value of q would be determined by

$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$ where μ is the mean of the time-series ,θ is the variables of the time-series which represent the series and the ε is the irregularities or the noise points.[10]

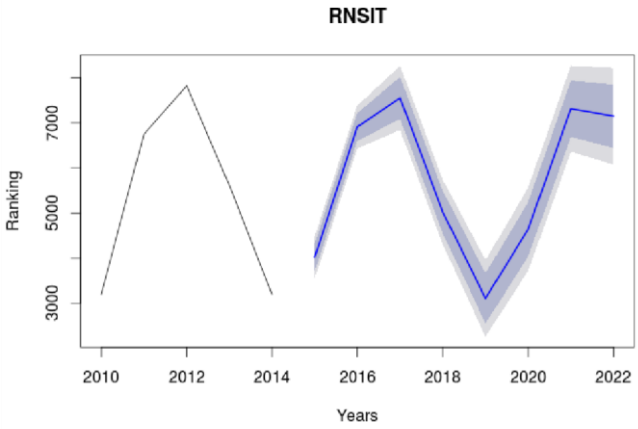

Fig. 3: ARIMA model fitting and prediction

The plot in Fig. 3. shows the ARIMA model fit onto the CET dataset for the RNSIT college. The ARIMA model for this was decided by the various computations that are to be performed, first to find the value of d , the timeseries is differenced a series whose mean is roughly zero is obtained, this can be seen in the Fig. 4. which shows the time-series of the CET dataset plotted for the PESIT college after differencing it once. The plot clearly shows symmetry about the x=0 line thus the value of d = 1. The value of p is determined by the observing at what lag value auto-correlation exceeds the set significance of the plotted correlogram. In Fig. 5. we can observe that the lag at 0.0 exceeds the significance bounds thus the value of p is zero. The value of q is determined by the observing at what lag value auto-correlation exceeds the set significance of the plotted partial correlogram. In Fig. 6. it is observed that the lag does not exceed the significance bounds at any point , thus the value of q is also zero. So the best fit model for the considered time series is ARIMA(0,1,0). The accuracy of the model is based on the model of the ARIMA which was used and the errors were significantly reduced as compared to Holt-Winters as the component which was eliminating the irregular component was truncated and the irregular component was used and fitted into the training model and significant weight age was placed on the same. The results of the model are presented in the results section.

3) Linear Regression Model: Linear regression[11] is a statistical approach to define a relationship between two or more variables. Here we have a dependent variable say y, and an independent or explanatory variable say x. Such a model where there is only one dependent variable is called simple linear

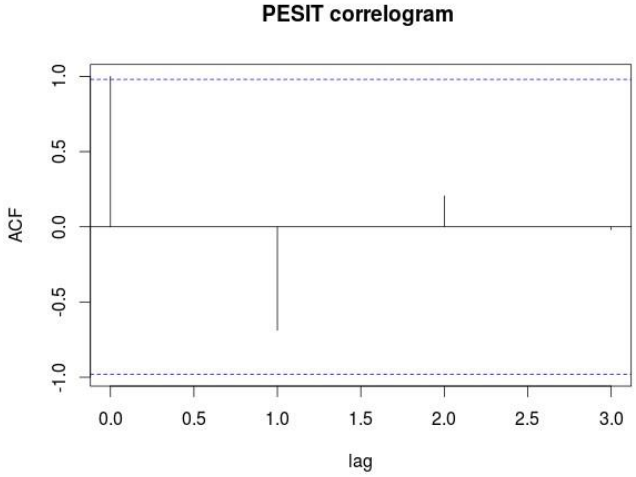

Fig. 5: Correlogram to find value of ARIMA variable.

regression. Basically a set of points are considered and using their x and y co-ordinates the model tries to fit a straight line through it. The general form of a simple linear regression equation is,

$h_i = \beta_0 + \beta_1 x + \varepsilon,$

ISSN No: - 2456 – 2165

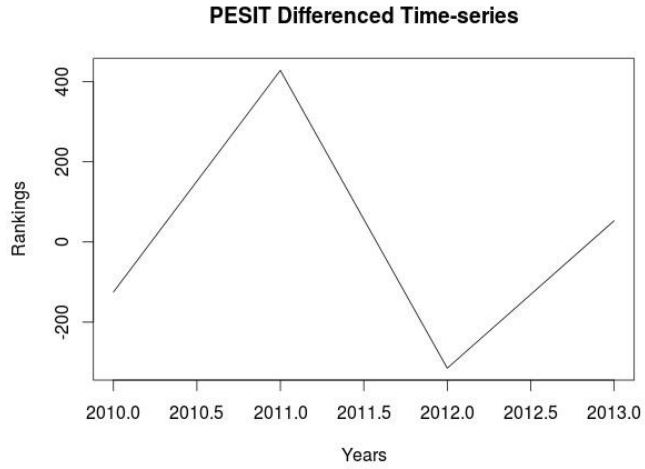

Fig. 4: Differencing of the Time Series of PESIT data.

here y is the dependent variable, β_0 is the y intercept $\beta_1$is the slope of the line. The least squares method is used to fit the regression model in the implementation concerning to this paper. For the purpose of predictions, we can use linear regression to fit a predictive model to a data-set with values of both variables. Eventually when the value of the dependent variable is not given, it can be estimated with the given independent variable value and the fitted model.

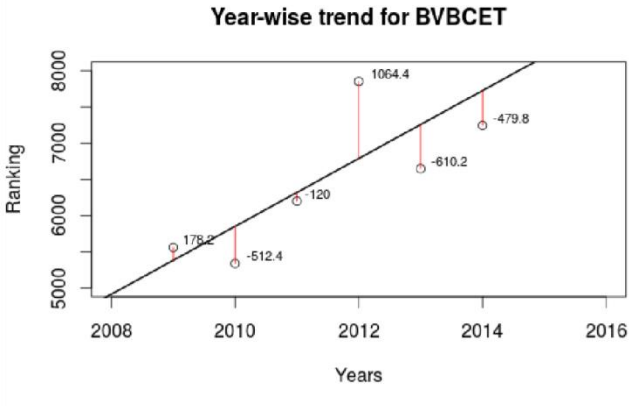

Fig. 7: Plot representing BVBCET EEE branch using ordinary LRM.

In our problem we take the year-wise cut-off ranks of each college as the variable vector to be supplied to the regression model. We fit the points into a straight line with least possible error. Further we obtain an equation satisfying the best fit, using which we get future values. Finally we plot residuals from the points to the fitted line which represent the difference between the predictions and the real values, lesser the better. One example from our dataset is shown in Fig. 4.

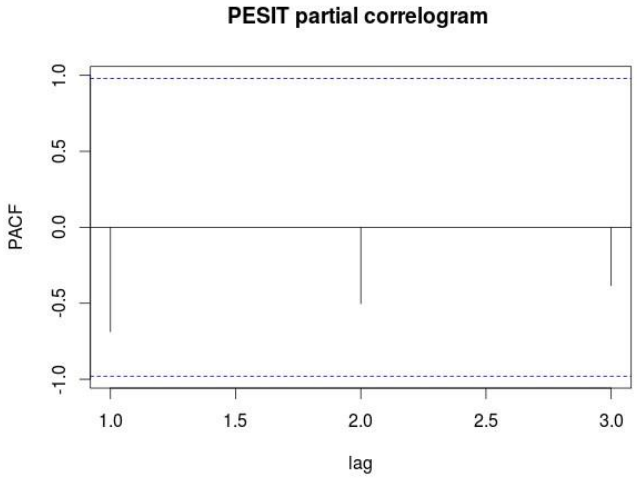

Fig. 6: Partial Correlogram to find value of ARIMA variable.

In Fig. 7 data used is from year 2009 to 2013 and predicting the expected cut-off of 2014 and matching it with the actual cut-off of 2014. The college shown here is BVBCET, Hubli. The plot shows the difference is 479 ranks when the range of ranks we are looking at is 5k to 8k . The lines dropping from each point of our fitted lines are the residuals plotted. This is an unweighted regression model where every year's data is given equal weightage irrespective of its nature. As we use the lm function in R for the purpose, we get very flexible options to implement the model. Further we set weights to our linear model to give more weightage to values which are more relevant and ignore the unusual surges which aggravate the model. With the right balance between human and computer interaction in building the relevant model each time, we improve the previous results.

V. RESULTS

The comparative study was conducted on the models mentioned above and the results are presented as shown below. The implementation of the concepts explained was done using R programming language using a number of useful R libraries for the purpose. For the Holt Winters model we use forecast[12], [13] and ttr package [16] ARIMA model is implemented using the tseries [14] and xts [15] packages. For regression models we use built in packages like stat and graphics which ship with RStudio[17], the IDE which was used.

The implementation takes into consideration the first and foremost problem i.e., to obtain the predictions on the cut-off rankings for the upcoming year, which would set targets for students and help them to plan their study patterns. The cut-off ranking from 2009 to 1013 are fitted onto the line or made into a time series and then the prediction is made based on the model.

ISSN No: - 2456 – 2165

Suppose an example of a branch of Computer Science &Engg. of all colleges is considered.

Model	RMSE	Normalised RMSE
Holt-Winters	1997	0.019
Linear Regression	2244	0.022
ARIMA	586	0.005

Cut-off ranks prediction RMSE

where Normalised RMSE = [RMSE/(max(DV)] DV is the decision variable i.e., rankings given to students in the entrance test. This can range from 0 to 100,000.

The results given above are RMSE of the CS&E branch from all colleges. The data from years 2009-2013 are considered. As one can see the time series models produce a satisfactory result. The ARIMA gives the best fit as the data here are year wise trends and have an irregularity factor.

College	Actual Value	Predicted Value
RVCE	270	581
PESIT	464	746
BMSCE	760	1023
MSRIT	1093	1643
SJCE	1476	2041
BIT	2496	2864
UVCE	2833	4442
NIE	2844	3293
BVBCET	3811	3915
RNSIT	3192	3910

College wise Cut-Offs prediction for CS&E branch using ARIMA

After a thorough comparison of results given by the 3 models we prefer the ARIMA mode to predict such a result. Therefore table here shows our predictions based on the ARIMA model. Here we have the list off all colleges which we have chosen along with their actual cut-offs of CS&E branch for the year 2014. The next column is the predictions we make for the corresponding college.

Model	RMSE	Normalised RMSE
Holt-Winters	1.26	0.12
Linear Regression	1.18	0.11
ARIMA	1.18	0.11

College rank prediction RMSE

where Normalised RMSE = [RMSE/(max(DV)]
DV is the decision variable i.e., rankings given to colleges based on the cut-offs for all branches. This can range from 0 to 10.

Once we obtain the cut-off rankings, prediction for overall rankings of the college are made. This is done based on the predicted cut-offs of the college. The error is based on the positions given to the colleges. These positions given to each institute is completely based on the student cut-offs and should not be misunderstood for other factors.

A. Reinforcement

The above presented results for a regression model shows an error which is acceptable for most colleges in the above average region and all colleges below it as the cut-off ranks for them will be more than 5000. An error in the prediction of a thousand ranks will not matter as the maximum rank given by KCET is as high as 100,000. However when it comes to top colleges the accepted percentage of error is relatively low as the range of cut-off ranks in this region is in the range of 0-5000. Here such an error is very significant.

For this reason we remodelled our idea to obtain better results using the simple concept of regression. As mentioned earlier there will be few years which shoot up from the normal trend for various reasons. Such data points are either ignored or given less

ISSN No: - 2456 – 2165

weightage during the linear model construction. This paper also presents an opinion that data of the near past is more significant than relatively older data. Accordingly we increase weights of data points proportional to the year they represent.

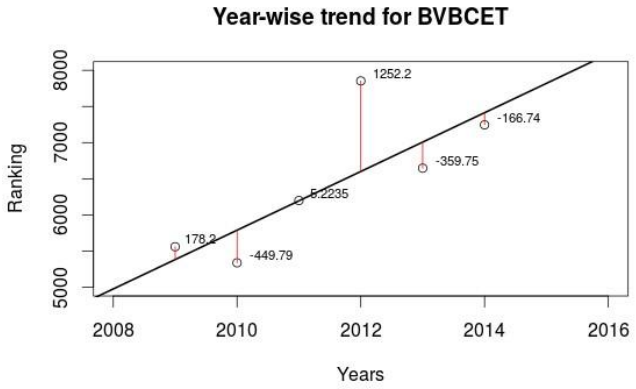

Fig. 8: Plot representing BVBCET EEE branch using weighted LRM.

Fig. 8 represents the same course in the same institute as tested previously is refitted into our reinforced model with weights. Weights for years from 2009-2013 were assigned as 2 ,2,2,1, 2.

Even though the results aren't as good as ARIMA, a simple model of regression could be improved in this manner by using a little bit of help from human intuition.

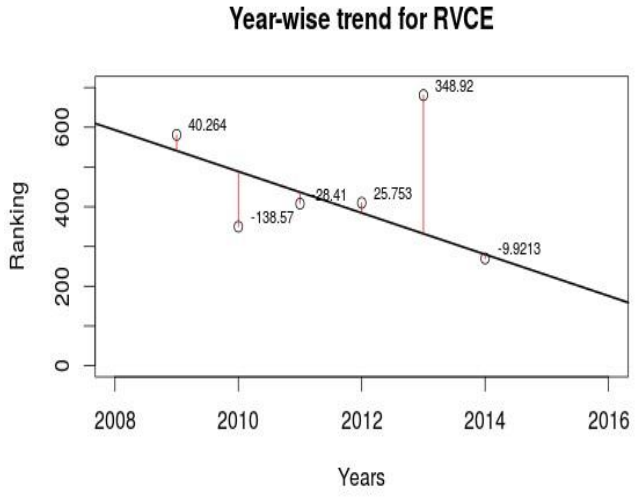

Fig. 9: Plot representing RVCE CS&E branch using weighted LRM.

Fig. 9 represents the topmost college in the state from the past few years. We consider the Computer Science branch. By applying weights we obtain an offset from the real value of just 10 ranks. We use weights for the years as 3 ,1,3,4, 0.

VI. CONCLUSION

We successfully completed a comparative study for obtaining a solution to modelling a predictor for evaulating educational institutes. The inferences we can draw is,

1) The Holt Winters model which is a time series approach to the problem is successful when we do not want to consider the irregular component of our data. We only consider the trend component for fitting the model. In the problem we rarely have a particular trend which repeats itself periodically. This makes the the Holt Winters prediction unsuitable for the given data.
2) The ARIMA model considers both the irregular component as well as the trend. This is a better fit than the Holt-Winters as the data we are using it on does not have a clear trend but has a significant irregularity component which is taken care by ARIMA. We can conclude it is the best fit for this problem.
3) The Linear regression model gives us a lot of flexibility when it comes to assigning priority or weights to particular data series over others, which is suggested in this paper. It also is a relatively simple model to implement which in turn makes it the favourable in terms of cost and computing speed along with a satisfactorily good accuracy.

REFERENCES

[1] Karlin, Samuel & Taylor, Howard M. (1998). An Introduction to Stochastic Modeling, Academic Press. ISBN 0-12-684887-4.
[2] Box, George EP Time series analysis: forecasting and control. John Wiley & Sons, 2015.
[3] Whittle, P. (1951). Hypothesis testing in times series analysis. Uppsala: Almqvist & Wiksells Boktryckeri AB.
[4] Bulmer, Michael (2003). Francis Galton: Pioneer of Heredity and Biometry. Johns Hopkins University Press. ISBN 0-8018-7403-3.
[5] Simonoff, Jeffrey S. (1998) Smoothing Methods in Statistics, 2nd edition. Springer ISBN 978-0387947167
[6] "Forecasting Trends and Seasonal by Exponentially Weighted Averages". Office of Naval Research Memorandum 52. reprinted in Holt, Charles C.
[7] https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
[8] "Notation for ARIMA Models". Time Series Forecasting System. SAS Institute.
[9] Pandit, Sudhakar M.; Wu, Shien-Ming (1983). Time Series and System Analysis with Applications. John Wiley & Sons.
[10] Enders, Walter (2004). "Stationary Time-Series Models". Applied Econometric Time Series (Second ed.). New York: Wiley. pp. 48107. ISBN 0-471-45173-8.

ISSN No: - 2456 – 2165

[11] Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 12, ISBN 9789812834119.
[12] Hyndman RJ and Khandakar Y (2008). Automatic time series forecasting: the forecast package for R. Journal of Statistical Software, 26(3), pp. 122. http://www.jstatsoft.org/article/view/v027i03.
[13] Adrian Trapletti and Kurt Hornik (2016). tseries: Time Series Analysis and Computational Finance. R package version 0.10-35.
[14] Jeffrey A. Ryan, Joshua M. Ulrich(2014). xts: eXtensible Time Series. R package version 0.9-7
[15] Joshua Ulrich(2016). TTR: Technical Trading Rules, Functions and data to construct technical trading rules with R. R package version 0.23-1
[16] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.