



OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

LINEAR REGRESSION FOR TRENDS IN TECHNICAL SKILLS

Chirag Ashar¹, Omkar Dhuri², Hrishikesh Bihani³, Uday Rote⁴

IT Department, K J Somaiya Institute of Engi. & IT, Sion, Mumbai, India.^{1 2 3}

Head of Department, IT Department, K J Somaiya Institute of Engi. & IT, Sion, Mumbai, India.⁴

chirag.ashar@somaiya.edu¹, omkar.dhuri@somaiya.edu², h.bihani@somaiya.edu³, udayrote@somaiya.edu⁴

Abstract: This paper recognizes the theory and practice of regression techniques for prediction of technical domain trends by using a converted data set in ordinal data format. The data formats in technical proficiency and skill level provide a process for calculation of technical domain trends. The converted data set contains only a standardized ordinal data type which offers a process to measure rankings of technical skills. The primary design is based on regression analysis from WEKA machine learning software. The technical domain trends from Alumni Portal, KJSIEIT is used as our research setting. The data sources are alumni technical proficiency which included recently used skill and its level. The variables included in the data set were formed based on technical domain trends from the alumni profiles. Classifiers in WEKA were used as algorithms to produce the outcomes. This learning showed that the results of regression techniques can be used for the prediction of technical skills which would be trending in the future by using a dataset in standardized ordinal data format

Keywords: regression techniques; machine learning; fundamental analysis; classifiers; linear regression, technical domain trends

I INTRODUCTION

Before we forge into the nitty-gritty of the primary subject of the project-Alumni Portal, let us first get ourselves conditioned to the etymology of the word, Portal. In the context of Computers and Networks, it is a Website that helps you find other sites! Hence, it forms a sort of a gateway to all the other places a user wish to explore!

The need to classify, segment and manage large conglomerate consisting of diverse entities has always been an important and at the same time labyrinthine! Public institutions form a very good example of that including the one we currently are a part of! So the moment we attach Alumni in front of the portal, the range and reach spreads across the 16 years that institution has been running and gets under it a sizeable amount of students, both the pass outs and the current batches. Moreover, Alumni Portal becomes the nodal junction for interactions and discussions amongst them. Alumni Portal is a website that doubles up as a dynamic intermediary database of the Alumni and the correlated data that comes with them. The portal acts as a one-stop destination for information on the students of the college. The data is added up every year and hence updating and removal of redundant data becomes a priority!

The sheer size of entities and the not to forget the upcoming years of data is reason enough for the creation of a portal for the Alumni. However, if we are going to look closely; it is obvious of the volatility of the existing data. Alumni from the earlier batches are spread across the length and breadth of the world! The companies or organizations they work in, keep on changing as we are talking about careers in Information Technology sector! Moreover, some may altogether alter their career paths ending up in non-technical fields. Therefore, the data once collected isn't the final one and needs alteration at regular interval of time. Therefore, the need arises of Alumni Portal where the data entry, retrieval and deletion is efficient and regular

II LITERATURE SURVEY

Technical analysis method identifies chart patterns based on alumni's recently used skill set. This approach does not gain insight into the business side of how much a particular skill is valued in the industry; it assumes the available information does not offer a competitive advantage in a financial way. This technique predicts trends in advance through chart patterns [1].

The research on technology trends prediction techniques has eventually moved into the technological

realm. Machine learning approach is one of the common practices. The approach of machine learning is by observing a potentially linear or non-linear relationship exists with the availability of enough indicators [2]. Machine learning is a branch of artificial intelligence. This approach find patterns in training datasets and form their own rules which are then used for making forecasts in testing datasets [3].

In statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X*. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.[4] (This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.)[5]

Common regression analysis involves inputs of numerical data which may consist of infinite or a wide range of values. In this research, we start by gathering string data in real-valued format using the fundamental analysis approach. After that we apply a new transformation process to convert the string into numerical values and then those values into ordinal values. The ordinal values contain only a range of categorical enumerated values. The relationships between the dependent and the independent ordinal variables are correlated based on the enumerated values.

III COMPARISON BETWEEN EXISTING SYSTEM & PROPOSED SYSTEM

TABLE 1: EXISTING SYSTEM VS. PROPOSED SYSTEM

Existing System	Proposed System
The data in the existing system is outdated and redundant.	In proposed system, timely updating of data along with regular verification is proposed.
Data collection in the earlier system was a tiresome task that required more physical work and fewer entries by the Alumni	In the Alumni Portal, the facility to 'Sign Up' brings down the efforts and is an assured way of getting maximum Alumni registrations
The communication channel existed but was rarely used owing to unawareness of the upcoming and ongoing events	Robust communication channel is one of the primary aim of the Alumni Portal aided sufficiently by the message services and discussion forums
The existing system gave little exposure on the journey of the Alumni post-graduation.	Alumni Portal will regularly feed in the required updated data about the whereabouts of the Alumni

Link to connect the Alumni and students was present but was seldom capitalized to its potential.	Constant interaction will be fruitful in organizing meets at city level and college level thereby increasing the network between the Alumni and students.
The present data in the existing system is in the raw form, meaning pattern finding and conclusion making is almost impossible using that.	The Alumni Portal is proposed to use regression techniques to find trends in the recent domain studies.
The existing system lacked the function of creating useful information from the present data	Using linear regression and taking help of data mining will help us procure useful data in line with our needs

IV METHODOLOGY

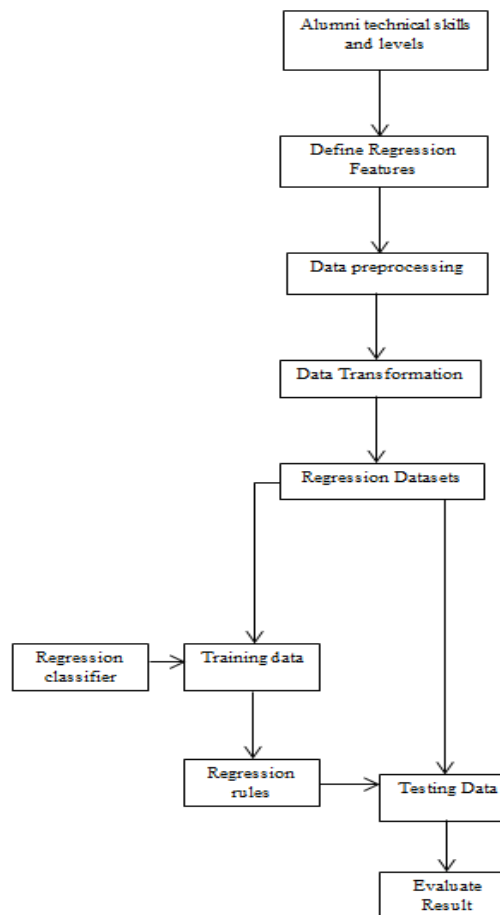


Figure 1 – Methodology

Statistics on corporations and dataset features are generated through fundamental analysis. Data was screened and pre-processed to remove out-of-bound values. This process can prevent problems of producing misleading results [8].

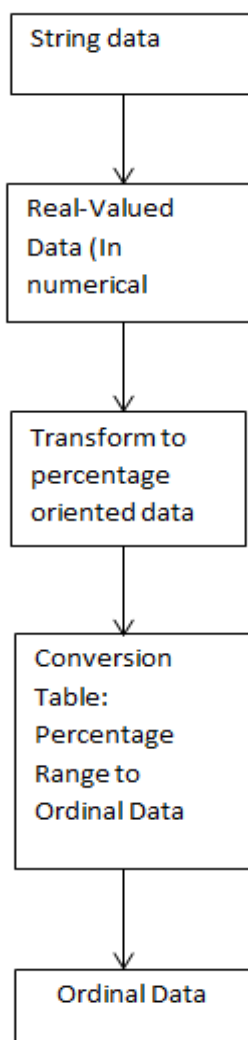


Figure 2 – Data Transformation Method

The objective of the transformation process is to make the data more structured. Pre-processed data contains general string data which include the technical skill and percentage formats. In the data transformation process, the pre-processed data is standardized into numerical value data. A percentage and to ordinal conversion table contains the mapping values for percentage values associated with their ordinal enumerated values. Each enumerated value is assigned to the dataset based on the conversion table. This approach maps each skill to a enumerated value for. This approach also clarifies the categories within a variable where its numerical values swing widely from one range to another range.

During the data training stage, one after another each regression classifier was used as predictive analytic on the dataset. A percentage split specifies a regression classifier to split the dataset into training data and testing data proportionally. Training data provides learning process for each classifier to formulate its own regression rules. The

regression rule was used on the testing data for predictions of future technology trends. The test result was then evaluated [9].

Example:


technical_proficiency		
	id	int
	skill	varchar(60)
	level	varchar(12)

Figure 3 : Technical Proficiency Schema.

Consider table technical proficiency contains these values.

Table 2: Populated technical proficiency table.

id	skill	level
1	java	beginner
2	python	advanced
3	php	intermediate

One method of converting numbers stored as strings into numerical variables is to use a string function called real that translates numeric values stored as strings into numeric values Stata can recognize as such. The first line of syntax reads in the dataset shown above. The second generates a new variable read_n that is equal to the value of the number stored in the string variable read. The real(s) is the function that translates the values held as strings, where s is the variable containing strings [10].

V CONCLUSION

Hence after going through the research in the aforementioned techniques, the outcomes of the present data that is being fed up in the portal will be improved when the mechanisms of linear regression can be applied and the finer details from a haystack of data present obtained which is suited to our needs and requirements.

In the above studies we have used WEKA regression techniques to categorize data and bring in front of us the recent trends domain wise. Hence we get tailored data that can be utilized effectively in channeling the resources for the interested subjects and also towards improving the exposure to the concerned entities. Overall the portal will be used in addition of the above uses to gather data of the past students at one place enabling effective and efficient management and retrieval of useful information as required. Also since the data is going to expand further, various techniques can be compared and used selectively for further research

REFERENCES

[1] Kirkpatrick and Dahlquist. Technical Analysis: The Complete Resource for Financial Market Technicians.

- Financial Times Press, 2006, page 3. ISBN 0-13-153113-1.
- [2]MacKay, D.J.C. (2003). Information Theory, Inference, and Learning Algorithms, Cambridge University Press. ISBN 0-521-64298-1.
- [3]Alpaydin, Ethem (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press, ISBN 0-262-01211-1.
- [4]David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. “A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient”.
- [5] Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.
- [6] Alpaydin, Ethem (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press, ISBN 0-262-01211-1.
- [7] Freedman, David (2005) Statistical Models: Theory and Practice, Cambridge University Press.
- [8] Kotsiantis, S.; Kanellopoulos, D. ;Pintelas, P. (2006) "Data Preprocessing for Supervised Learning", International Journal of Computer Science.
- [9] Theodoridis, Sergios; Koutroumbas, Konstantinos (2009) "Pattern Recognition", 4th Edition, Academic Press, ISBN 978-1-59749-272-0.
- [10] Introduction to SAS. UCLA: Statistical Consulting Group. from <https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/>
- [11] Han Lock Siew, Md Jan Nordin “Regression techniques for the prediction of stock price trend”, ISBN 978-1-4673-1582-1