A Model to Predict Pay Scale Fixation in Job Market Based On Educational Excellence

Abishek kamalanathan, Archana Tamizharasan

Abstract: ***The students of current generation worry a lot about their future salary when they get employment. The salary of an employee has always remained the major concern. In this paper we study and understand the major factors which are influencing the pay scales of the employees. This Research paper is based on predicting the pay scales of employees working in the southern part of India. The pay scale of an employee with below 4 lakh per annum is treated as medium pay scale and pay scale with above 4 lakh per annum is treated as good pay scale. In this paper the machine learning model is applied to predict the pay scale of an employee with the features which best suits the model. The year of passing, English marks, Quant marks, logical marks scores in AMCAT's exam, programming skills and college CGPA were the major factors influencing the pay scales. Three machine learning algorithms were applied to the dataset Naïve bayes, Decision tree and Bagging .Naive bayes and bagging model gave the best accuracy of 85.2%***

Index Terms: Feature Selection, job market, pay scale forecast, predicting salaries

I. INTRODUCTION

The broad mind set of almost every student is that they believe good CGPA will help them to get good salaries. This mind set of the students prompted me to take to this topic and understand the factors which really have an impact in the salary apart the CGPA. In this paper we found that solely CGPA cannot influence the pay scales of an employee though CGPA had an impact in the model but factor like experience, marks obtained in AMCAT's examination and programming skills influenced more than the CGPA. The salary predicting model will help the upcoming students to analyse about the salary trends in job market and will actuate the students towards the right factors which will influence them to get good salaries [1]. This dataset was obtained from AMEO. This dataset consist of 37 independent features out of which after data pre-processing and feature selection only 6 significant attributes were selected passout_year, eng_mark, quant_mark, prog_skills, cgpa_clg and logic_mark. The entire coding was done in R. The dataset was split into train and test model in the ratio 80 and 20.The predictor labels are medium pay and good pay, the model performed well for the medium class labels. Three models were used in this paper Naïve Bayes, Decision Tree, Bagging.

.

II. BACKGROUND

Previous Researchers used statistical techniques methods to predict the salaries Young-joo Lee et al. applied regression method to predict the salaries considering details of student's academic performances and comparing the satisfaction with salary [2].P. khongchai et al.[4] build a salary predicting system using k-NN, SVM used to predict the pay scales. This paper consider factors like gender,gpa,branch of study.[3] used multiple regression approach to predict the salaries of two groups of one who are Graduates and the other groups are drop out and compared the results of expected salaries by the adults and the actuals salaries of the adults.[9] This paper used statistical approach and considers factors like GMAT scores, age, Gender to Predicted and analyse how the salaries differ from the students graduating from the MBA programs. Most of the research works in this field were related to salaries hence they followed regression method to predict the continuous variable. This paper uses classification approach by binning the continuous values of salaries into categorical values. Hence making classification of pay scales and other control variables which are contributing to build a good model were made into categorical from continuous variable. Classification method can provide more deeper understanding ,insights and helps us to analyse with different machine learning model than the regression method .Machine learning model like Naïve Bayes, Bagging accept categorical values than continuous values. Bagging is a type ensemble which aggregates all the results and votes for the best predicted outcomes for each test data which gave the best accuracy. There were 37 control variables in the dataset which were fed into random forest model for feature selection to subset the importance factors which contribute to the model. Surprisingly the results we got that the academic performances in the schools did not have an impact in the model. This is was not done by other researchers as far our knowledge.

III. DATASET DESCRIPTION

This dataset was obtained from Aspiring Minds'

Revised Manuscript Received on July 05, 2019.

Abishek kamalanathan, Computer Science, Vit University,Vellore , India

Archana Tamizharasan,, Computer Science, Asst. Professor, Vit University, Vellore ,India

Attribute name	Description
Salary	The salary of employee INR per annum
DOJ	The date of joining of employee in the company
DOL	The Date of leaving of employee from the company
Designation	The designation of the employee in the company
Job city	The city where the employee works
Gender	The Gender of the employee
DOB	The date of birth of an employee
10 percentage	The mark obtained by the employee in his 10^{th} standard
10 board	The board in which the employee studied in 10 standard
12 graduation	The year in which the employee graduated
12 percentage	The marks obtained by the employee in 12^{th} standard
12 board	The board in which the employee studied in 12 standard
CollegeID	The college ID of the employee
CollegeTier	The college Tier of the employee With value 1 and 2.Tier 1 college has better standards as compared to Tier 2 colleges
Degree	The degree he has completed
Specialization	The department which she studied
CollegeCGPA	The cgpa the employee obtained
CollegeCityTier	The Tier of the college in that city
CollegeState	The state in which the employee studied his college
Graduation Year	The graduation of an employee
English	The marks scored by the employee in the AMCAT exam
Logical	
Quant	
Domain	
ComputerProgramming	The marks scored by the employees in their respective domain stream in AMCATS exam
ElectronicsAndSemicon	
ComputerScience	
MechanicalEngg	
ElectricalEngg	
TelecomEngg	
CivilEngg	
Conscientiousness	The score obtained by the employee in the AMCATS'S personality test.
Agreeableness	
Extraversion	
Neuroticism	
openess_to_experience	

Table i.) Table Description

Employability Outcomes (AMEO).This dataset has the details of the employees from their 10^{th} standard marks to college CGPA. The dataset before pre-processing consisted of 37 features.

The description of each feature is given in the table i.

IV. DATA PREPROCESSING

The attribute Jobcity had many ambiguous spelling for the cities like Chennai, Madras and Banglore, Bengaluru which was corrected and had one correct spelling for all the City.

The College Cgpa attribute had CGPAs calculated for 100 and some were calculated for 10.Since most of the colleges had CGPAs out of 100 all CGPAs were scaled to 100.The continuous factor like Graduation year, English, Logical, Quant, College CGPA, Computer Programming were binned into categorical values. A new attribute IsCapital attribute was introduced this attribute has values 1 if it is capital city (Chennai, Bengaluru, Hyderebad, Thiruvananthapuram) and 0 if is not a capital city. In the box plot in Fig:1 a. shows that salaries of employee working in capital City is greater than the non-capital City. The boxplot Fig: 1b. shows the comparison of salaries between male and female, there was no difference in mean salaries between male and female.Pay scale attribute was introduced and the salaries below 4 lakhs were labelled as medium pay and salaries above 4 lakhs were labelled as Good pay The continuous attributes English,

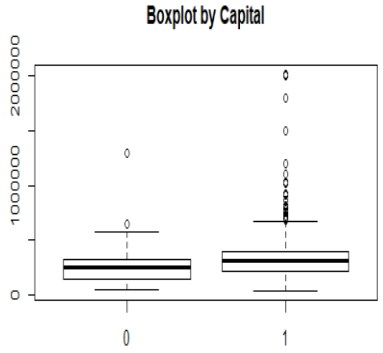

Fig 1 a.)

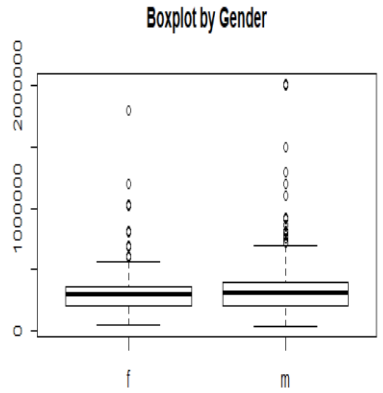

Fig 1 b.)

Logical, Quant, Domain were converted into new attributes English mark, Logical mark, Quant mark, Domain mark with categorical values having labels >=500 or <500.The Graduation Year attribute was binned into passout_year with values 2007-2011 and 2012-2016. The attribute Computer Programming has continuous variable of marks scored in AMCAT's programming examination which is binned in a new attribute Prog_skills with values >=425 and <425. The attribute College_CGPA had continuous variables which was categorised as >=70 and <70 in a new attribute cgpa_clg.

V.FEATURE SELECTION

The feature Selection was done by using Random Forest model. The package RandomForest in R was used where the function importance () gives the measure of Mean Decrease Accuracy and Mean Decrease Gini of each attributes. The attribute with larger value of Mean Decrease Accuracy (MDA) is more important feature for the model. The value of MDA tells the decrease Accuracy (MDA) is more important feature for the model. The value of MDA tells the decrease in the accuracy if the attribute is removed from the feature list. The Gini index is the measure of the purity or the homogeneity of the attribute contributing towards the split of the node. A low Gini has higher Decrease Gini thus that attribute contributes to the split of the node. The attributes IsCapital, Gender, 12th board, domain range do not have significance in the model. The attributes which has values less than zero in mean decrease accuracy is not considered in building the machine learning model. The attribute passout_year represents the experience of an employee was the most significant attribute which becomes the root node in the decision tree. The pre-processed data had 1219 records of medium class labels and only 219 records of good class labels, the number of training examples of good class labels were less compared to the medium class labels which make the model difficult to predict for good pay labels. The Fig 2 and 3 below shows the variable importance and measure of (MDA) and MeanDecreaseGini.

VI. PROPOSED MODEL

The dataset obtained after pre-processed is used for building the machine learning model. The target attribute has two values medium pay and good pay which is the main objective of this paper to classify the employees pay range.

Three machine learning algorithm was used in making the model. The control variables are selected from the Feature Selection, so the features which contribute significantly in building the model are passout_year, eng_mark, quant_mark, prog_skills,cgpa_clg,logic_mark. The Fig4 depicts the model flow.

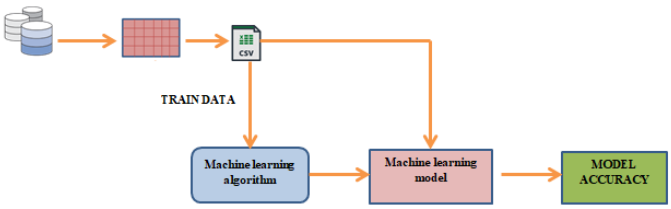

Fig: 4 proposed model

A. Naïve Bayes

Naïve Bayes algorithm is one of the best algorithms for classi fication problems with categorical values. It is based on the Bayes probability theorem works with an assumption that the probability of a feature does not affect the other feature that is the probability of the feature given the predictor is indepe ndent of other features.

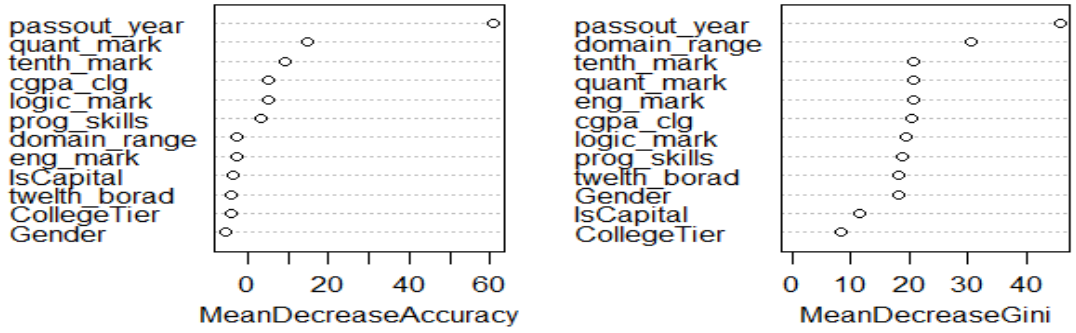

Fig 2.) The variable importance of each attribute for feature selection

	medium	good	MeanDecreaseAccuracy	MeanDecreaseGini
passout_year	46.5151999	47.7638334	62.80802932	42.288258
eng_mark	-0.1642035	-4.5376596	-2.63491836	18.860618
quant_mark	12.5730755	7.8701374	15.48306062	19.621353
prog_skills	5.6539247	-3.1711680	3.30275697	18.394902
cgpa_clg	5.2653090	3.2599144	6.41710356	20.058143
logic_mark	5.9899770	2.1165524	6.87343111	20.692045
IsCapital	-0.8818059	-3.6931333	-2.89262413	11.292799
Gender	-1.5682036	-2.6557114	-2.81753679	18.787091
tenth_mark	13.3361059	-5.0783190	9.42866451	20.556442
twelth_borad	-0.1905726	0.3229840	-0.08286838	18.160597
CollegeTier	-4.8861300	10.8389931	1.99732354	8.504583
domain_range	-1.2577186	-0.4984501	-1.39824924	29.454667

Fig 3.) The value of MeanDecreaseAccuracy and MeanDecreaseGini

The prior probability of the medium pay class label is 0.78 a nd prior probability of good pay class label is 0.21.The condi tional probability of each attribute is given in Table ii.). The model had an accuracy of 85.2%.

Feature Attribute		Class Attribute	
		Medium Pay	Good Pay
passout_year	2007-2011	0.1523438	0.4676259
	2012-2016	0.8476562	0.5323741
eng_mark	<500	0.5585938	0.4388489
	>=500	0.4414062	0.5611511
quant_mark	<500	0.5517578	0.3093525
	>=500	0.4482422	0.6906475
prog_skills	<425	0.5566406	0.3920863
	>=425	0.4433594	0.6079137
cgpa_clg	<70	0.3789062	0.2913669
	>=70	0.6210938	0.7086331
logic_mark	<500	0.5439453	0.4388489
	>=500	0.4560547	0.5611511

Table ii.) Conditional probability of each feature

B. Decision Tree

Decision Tree is one of the best algorithm for classification. Decision trees can handle both continuous variables and categorical variables in-case of continuous variables it becomes regression tree and in the other case it becomes classification tree. Classification works well with datasets with limited number of records. Classification tree works on the entropy approach. The attribute passout_year has the highest information gain so it becomes the root node. The marks obtained in the quant has the second greatest information gain value, it formed the next split after the root node. The cgpa_clg and programming_skills formed the subsequent splits respectively. The information gain is the difference between the entropy of the class attribute and entropy of the feature.

The entropy of attribute is calculated as

$$-\log_2\left(p\left(\frac{good\ pay}{total\ observation}\right)\right)-\log_2\left(p\left(\frac{medium\ pay}{total\ observation}\right)\right)$$

From the decision tree it can be inferred that The probability of a employee being in good pay is high if passes out in year

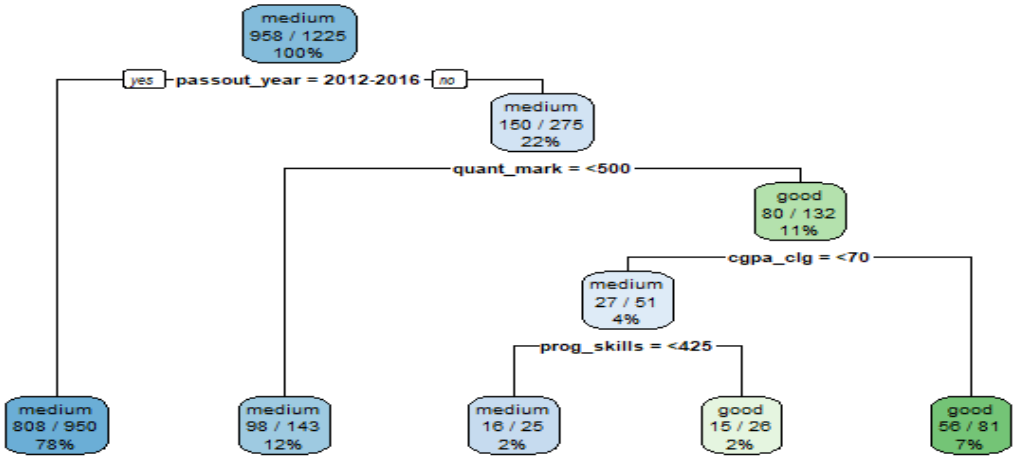

Fig:5.The decision tree showing the node splits the right side split denotes YES and left side split denotes NO

2007-2011 and his quant mark is greater than 500 and cgpa is greater than 70.

If an employee has graduated from 2011-2016 then it comes under the category of medium pay, also if the employee does not have good marks in quant and programming skills is more likely to be in medium pay scale. Even if the cgpa of an employee is less than 70 and his programming skills is good he is likely to be having a good pay scale. The accuracy of the model was 84.7%.The visualization of decision Tree is shown in Fig 5.

C. Bagging

Bagging is a type of ensemble learner. Ensemble classifier combines various algorithms and provides the best solution. It is a combination of bootstrap and aggregation were bootstrap creates a random sample data from the train data and aggregate all the outcomes from the model. Each model votes for the test instances, the vote here is predicting the class labels. The maximum votes get elected and becomes the final outcome of the model. The accuracy of model was 85.9%..

VII. CONCLUSION

The model performed well for the Medium pay Labels as it had more training labels. The model showed no difference in pay scales with respect to Gender, also the marks obtained in schools were insignificant in building the model. There was a significant decrease in the number of employee who scored good marks in class 10 when compared to class 12. The mean salary of employee who holds B.Tech degree was higher than the M.tech. The mark obtained in the respective domain exam was insignificant in predicting the mode. The employees working in the capitals cities had large number of outliers than the employees working in non-capital areas. The personality test(extraversion,Agreeableness,neuroticism,openess_to_experience and agreeableness) marks had no impact in the model. The college tier had no significance in deciding the pay scales. The experience of the employee was the important factor in deciding the pay scales. The marks obtained in logical, quant, computer programming and college CGPA had an impact in predicting the pay scales. It can be concluded that the employee with more experience likely to get a good pay scale and other major factors influencing the pay scales are the employees programming skills and his college CGPA. Since bagging has better accuracy, the best model is chosen as bagging. In the test data there were 255 instances of medium class labels out which the bagging model predicted 246 instances correctly and Decision tree predicted 244 instances correctly and the test data had only 52 good class labels out which the bagging model could able to predict the 22 instances correctly and Naïve Bayes predicted only 18 and decision Tree predicted 19. In both class labels bagging model had an edge over other two models. The comparison is shown in Table iii

		Naïve Bayes		Decision tree		Bagging	
		Actual Label		Actual Label		Actual Label	
		Medium Pay	Good Pay	Medium Pay	Good Pay	Medium Pay	Good Pay
Model prediction	Medium Pay	246	34	244	33	246	30
	Good Pay	9	18	11	19	9	22
Accuracy		85.2%		84.7%		85.9%	

Table iii.) comparison of algorithms used.

REFERENCES

1. P. Khongchai, P. Songmuang, "Random Forest for Salary Prediction System to Improve Students? Motivation", *016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 637-642, 2016.
2. Y. Lee, M. Sabharwal, "Education—Job Match Salary and Job Satisfaction Across the Public Non-Profit and For-Profit Sectors: Survey of recent college graduates", Public Management Review, vol. 18, no. 1, pp. 40-64, 2014.
3. J. Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", Education Economics, vol. 23, no. 2, pp. 162-179, 2013.
4. P. khongchai, P. Songmuang, "Improving Students'Motivation to Study using Salary Prediction System", 13th International Joint Conference on Computer Science and Software Engineering (Preprint), 2016.
5. T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging Boosting and Randomization", Machine Learning, vol. 40, no. 2, pp. 139-157, 2001.
6. Betts, J. "What Do Students Know About Wages? Evidence from a Survey of Undergraduates." Journal of Human Resources 31 (1): 27–56 doi: 10.2307/146042, 1996
7. Jerrim, J "The Wage Expectations of UK Students: Are They Realistic?." FiscalStudies 32 (4): 483–509 doi:

10.1111/j.1475-5890.2011.00148,2011
8. Karla R. H. and W. A. Hamlen, “Faculty salary as a predictor of student outgoing salaries from MBA programs", Journal of Education for Business, 91:1, p 38-44, 2015
9. B. Xu, J. Li, Q. Wang, X. Chen, "A Tree Selectllion Model for Improved Random Forest", *Bulletin of Advanced technology research*, vol. 6, no. 2, 2012
10. O. Villacampa, "Feature Selection and Classification Methods for Decision Making: A Comparative Analysis", *College of Engineering and Computing Nova Southeastern University*, 2015.
11. Guyon I, Elisseeff A: An introduction to variable and feature selection. *JMachLearnRes*,3:1157–82.10.1162/153244303322753616, 2003
12. Archer, K.J., Kimes, R.V.: Empirical characterization of random forest variable importance measures.Comput.Stat. Data Anal. 52, 2249–2260, 2008
13. Biau, G., Devroye, L., Lugosi, G.: Consistency of random forests and other averaging classifiers. J. Mach. Learn. Res., 2015–2033,2008
14. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Mach. Learn. 40, 139–157,2000.

AUTHORS PROFILE

Abishek kamalanathan, He is pursuing his 4th year computer science in Vellore Institute of Technology. He is an Machine learning and data science enthusiast. He has done his academic internship on Data Science in National University of Singapore. His aspiration is do Masters in data science and purse research works data science and machine learning.

Archana Tamizharasan is an assistant professor in School of computer science and engineering at Vellore Institute of Technology. She is pursuing her Ph.D. She has publications in interdisciplinary research on Market analysis in e-business and psychological pattern mining in virtual learning environment. Her current research focus is exploratory data analysis and applications of deep and machine learning.