International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Data Mining For Prediction Human Performance and Skill Analysis

JAYAKRISHNAN J

Student, Dept. of Dual Degree Computer Applications, Sree Narayana Guru Institute of Science and Technology, Kerala, India

---

Abstract - *It is exceptionally basic for the organizations to guarantee the enrollment of right ability to keep up a serious edge over the others in the market. Anyway IT organizations frequently face an issue while enlisting new individuals for their progressing ventures because of absence of an appropriate structure that characterizes a standards for the choice procedure. In this paper we mean to build up a system that would permit any extend supervisor to take the correct choice for choosing new ability by connecting execution boundaries with the other space explicit properties of the competitors. Likewise, another significant inspiration driving this venture is to check the legitimacy of the choice methodology frequently followed by different large organizations in both open and private areas which center just around scholarly scores, GPA/evaluations of understudies from schools and other scholastic foundations. The extent of this work reaches out past the IT space and a comparable system can be received to build up an enrollment structure in different fields too. Information mining procedures give valuable data from the authentic undertakings relying upon which the employing supervisor can settle on choices for enrolling great workforce. This examination intends to connect this rest by building up an information mining structure dependent on an outfit learning procedure to pull together on the rules for faculty determination. The outcomes from this examination plainly exhibited that there is a need to pull together on the determination rules for quality destinations. It is effectively applied in the region of extortion location, promoting, showcasing, credit appraisal and expectation. But, it is in nascent stage in the field of education. Considerable amount of work is done in this direction, but still there are many untouched areas.*

1. INTRODUCTION

In the time of globalization work markets are a significant mainstay of success on the size of the whole populace. It very well may be contended that HR have at no other time had such a significant effect on the world. Ongoing history shows that a solitary ability can change economies and social orders over the globe. The accomplishments of Bill Gates, Angela Merkel, Henry Ford, or Oprah Winfrey are just a couple of tremendous instances of how people can influence the lives of millions. Specialized insight, social mindfulness, or enthusiasm for science—any sort of an expertise can be a significant resource when perceived, upheld, and set out to really utilize. Any place people are concerned, employments and professions are a noteworthy piece of the condition. This is consistently a matter of how one is getting by, which is legitimately reliant on what skills one has. Be that as it may, regardless of whether utilized, outsourcing, or an entrepreneur, there is no uncertainty that proficient decisions are vigorously obliged by a significant player—work showcase requests. As is obvious in the writing, minimalizing difference between aptitudes needs and flexibly is the essential worry of strategy creators around the globe. On the off chance that, Universities could recognize the elements for low execution prior and can anticipate understudies' conduct, this information can help them in taking professional dynamic activities, in order to improve the presentation of such understudies. It will be a win situation for all the stakeholders of universities/institutions i.e. management, teachers, students and parents. Students will be able to identify their weaknesses beforehand and can improve themselves. Educators will have the option to design their talks according to the need of understudies and can give better direction to such understudies. Investigation and expectation with the assistance of information mining methods have demonstrated imperative outcomes in the zone of misrepresentation discovery, foreseeing client conduct, budgetary market, credit evaluation, chapter 11 forecast, land appraisal and interruption identification. In this paper we address the issue of building up a perfect determination structure for enrolling the correct ability which carries us to the essential inquiry of what rules to follow for the choice technique.

2. EXISTING SYSTEM

With expanding multifaceted nature of programming in the business and their regularly developing requests in multidisciplinary ventures, there has been a constant advancement in research works that focus on the zones of powerful task the executives and Data mining has as of late substantiated itself as one of the most settled strategies here. Information digging systems are produced for a few applications including different parts of programming advancement and we intend to utilize this intensity of calculations to build up a determination structure. There have been a plenty of studies which fuse the devices of AI for building up a structure for Prediction of Human-Capability. Educational institutions generate and collect huge amount of data. This may include students' academic records, their personal profile, observations of their behaviour, their web log activities and also faculty profile. The extraordinary commitment of this paper is that separated from above expressed boundaries, it additionally investigates the connection between and passionate aptitudes like

affirmation, compassion, dynamic, administration, drive, stress the board to foresee employability utilizing information mining strategies. The passionate abilities like affirmation, administration, the executives, and compassion, dynamic have been incorporated utilizing standard ESAP. ESAP means "Passionate Skills Assessment Process" which is a thorough strategy to assess an understudy's Emotional Quotient (EQ). Further subtleties of this strategy are given under Experimental Setting area.

3. PROPOSED SYSTEM

MCA means "Ace of Computer Applications". It is a propelled six semester degree course in applied Computer Science offered by a few colleges across India. The 'employability' of a student has been defined as whether the student was capable of getting an on campus placement offered in V semester. The paper is organized as follows: The Introduction section gives a brief idea of the subject followed by the related work in Literature Review section. Next, Experimental Settings section discusses the data collection, data pre-processing, and data set development. This is followed by the Result and Discussion which compares different classifier's ability in predicting employability. Model developed section discusses prediction model obtained, followed by Conclusions and Future work.

In this connection, the main objectives of the present study were extracted to support the decision makers in different locations to discover potential talents of Gathering a dataset of predictive variables, Identification of different factors, which affects employees' behaviour and performance. Utilizing proposed DM order methods for building a prescient model and distinguishing connections between most significant variables influencing over entire productivity of the model. Analysis and prediction with the help of data mining techniques have shown noteworthy results in the area of fraud detection, predicting customer behaviour, financial market, loan assessment, bankruptcy prediction, real-estate assessment and intrusion detection. It can be very effective in Education System as well. It is a very powerful tool to reveal hidden patterns and precious knowledge, which otherwise may not be identified and difficult to find and comprehend with the help of statistical methods. The issue of building up a perfect determination system for enlisting the correct ability which carries us to the fundamental inquiry of what standards to follow for the choice technique. Our point is to comprehend the connection between the different undertaking individual characteristics of the up-and-comers and their expert execution boundary as appraised by their directors/administrators in the business. Our point is to distinguish the factors which have the most extreme prescient force in evaluating the exhibition abilities of newcomers. Despite the fact that there have been numerous past examinations in this area, there have been sure issues that despite everything should be tended to.

4. METHODOLOGY

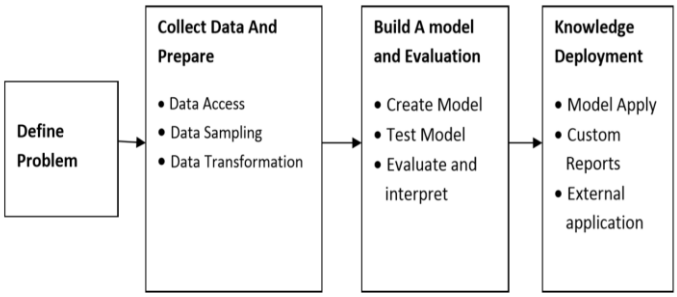


Figure -1: Architecture for Building recruitment Process

Hypothesis: Project personnel with similar skills set and capabilities will perform similarly. Project-personal information about the employees is collected from projects which use similar technology and programming language and work on similar platforms. Thus personnel under comparison here have similar capabilities and skill-sets.

Data Collection: Data was collected by using various techniques such as Form-Filling, Brainstorming, obtaining performance information about employees from the Project-leads and managers.

Data Preparation: A basic preliminary research concluded with an almost same set of attributes that must be considered under this problem-statement to obtain a correlation with performance parameters as in.

4.1 Underlying Learning Model – Random Forest

In a classification problem, we have a training sample of n observations on a class variable Y that takes values 1, 2, ... k, and p predictor variables, X1,..., Xp. Our goal is to find a model for predicting the values of Y from new X values. In theory, the solution is simply a partition of the X space into k disjoint sets, A1, A2,..., Ak, such that the predicted value of Y is j if X belongs to Aj , for j = 1, 2,..., k. If the X variables take ordered values, two classical solutions are linear discriminant analysis1 and nearest neighbor classification. These methods yield sets Aj with piecewise linear and nonlinear, respectively, boundaries that are not easy to interpret if p is large. Classification tree methods yield rectangular sets Aj by recursively partitioning the data set one X variable at a time. Pseudocode for tree construction by exhaustive search is as follows:

- Start at the root node. For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split {X∗ ∈ S∗} that gives the minimum overall X and S.
- If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

Since, this is a Regression problem, we use CART for producing individual regression trees and then build upon

this model by using the Random Forest technique to achieve better results. CART uses a generalization of the binomial variance called the Gini index. A regression tree is similar to a classification tree, except that the Y variable takes ordered values and a regression model is fitted to each node to give the predicted values of Y. This yields piecewise constant models. Although they are simple to interpret, the prediction accuracy of these models often lags behind that of models with more smoothness. These trees may not provide very high accuracies, since they have very high variance values. Randomization based ensemble methods, prove to be a good solution to this flaw. Random-Forest consists of a collection or ensemble of simple tree predictors, each of which outputs a response when presented with a set of predictor values just as the input vector X. For classification-based problems, this response can be of the forms - class membership or associations, a set of independent predictor values with one of the categories present in the dependent variable. Each tree is created from its own separate bootstrapped sample training set. The Bootstrap Sampling Method samples the given training tuples uniformly with replacement i.e. each time a tuple is selected, it is equally likely to be selected again and rendered to the training set. As the number of simple learning models within an ensemble technique increases, the overall variance of the output-value from the actual-value theoretically decreases by 1/ (number of individual models). The Mean-Decrease-in-Accuracy of a variable is evaluated during the calculation-phase of out-of-bag error. As the fall in accuracy of the random-forest increases due to the addition of a single-variable, the more important the particular variable under test is considered and hence variables with a large value for Mean-Decrease-in-Accuracy or Gini are considered as more important for data classification. The Mean-Decrease-in-Gini coefficient is a measure of how each particular variable supplements to the homogeneity of the nodes and terminal-leaves in the resulting Random-forest. The Mean-Decrease-in-Accuracy of a variable is evaluated during the calculation-phase of out-of-bag error. As the fall in accuracy of the random-forest increases due to the addition of a single-variable, the more important the particular variable under test is considered and hence variables with a large value for Mean-Decrease-in-Accuracy or Gini are considered as more important for data classification. The Mean-Decrease-in-Gini coefficient is a measure of how each particular variable supplements to the homogeneity of the nodes and terminal-leaves in the resulting Random-forest. Every time one particular variable splits a given node, the Gini coefficient for the children are calculated and compared to that of the original parent node. If the same variable causes multiple splits more than once, then the final difference in the Gini value of the topmost parent node and the bottom-most children nodes is taken as the Mean-decrease-in-Gini value. The pseudo code for generation of a random forest is as follows:

- Draw number_of_trees bootstrap samples from the original data.
- Grow a tree for each bootstrap data set. At each node of the tree, randomly select m try variables for splitting.
- Grow the tree so that each terminal node has no fewer than node size cases.
- Aggregate information from the n tree trees for new data prediction such as majority voting for classification.
- Compute an out-of-bag (OOB) error rate by using the data not in the bootstrap sample.

Using these values, a final graph for Variable Importance is plotted, where this graph represents each variable on the vertical y-axis, and their importance-values on the horizontal x-axis. They are ordered in the manner of top-to bottom as maximum-to-minimum-importance. To measure the accuracy of the classifier, we made use of Sensitivity and Specificity parameters. The following are the meaning of the variables used in the subsequent equations.

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

1. True-Positive Rate or Sensitivity is the fraction of training samples predicted correctly by model. TPR = $TP\ TP+FN$
2. False-Positive Rate or Specificity i.e. the fraction of training samples predicted incorrectly by model. FPR = $FP\ TN+FP$

Where FPR represents the false positive Rate and lower this value, the better the model is.

5. IMPLEMENTATION

For creating bootstrap samples, we used the technique of 632 Bootstrapping, which means that in any bootstrap sample generated, approximately 63.2% of the Dataset will be unique, and the rest would be placed with replacement and duplication. Studies have shown that this bootstrapping technique produces near-optimal results. The plot for Variable Importance has been obtained using R tool and the results obtained are discussed below. The model generated cannot be visualized graphically due to the large number of trees generated, each created from a separate bootstrap sample and each producing its own results. As stated earlier, the final output from the Random Forest model is the average of the results obtained from each of the predictor trees in the forest.

Based on the study conducted, data obtained was consolidated and summarized in a tabular form. The algorithm used here is Random Forest, implemented using WEKA and R under "Test options". 10 - Fold Cross validation

was applied to supplement the out-of-bag calibration mechanism of Random-Forests. The number of individual predictor-trees was set to 500 in R as no significant reduction in variance was observed beyond this value. Trees were allowed to grow completely without any pre or post pruning. The package used in R for implementing Random Forest is the "random Forest" package which is compatible with versions 4.6 and above. This package is available for download at the official R support website.

5.1 Survey of papers published in Educational Data Mining

Recent paper published in 2014 in Elsevier titled "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works" presented the survey of published papers from 2010-2013 and divided Educational Data Mining approaches in kinds of educational systems, disciplines, tasks, methods, and algorithms. Author identified that each Educational Data Mining approaches can be organized according to six functionalities student modelling, student behaviour modelling, assessment; student performance modelling; student support and feedback versus curriculum-domain knowledge sequencing, mostly focusing on academic performance.

Romero and Ventura, in 2010 published a paper in IEEE, which listed most common tasks in the educational environment resolved through data mining and some of the most promising future lines. Educational Data Mining community remained focused in North America, Western Europe, and Australia/New Zealand. They mentioned that there is a considerable scope for an increase in educational data mining's scientific Influence. They also suggested developing more unified and collaborative studies. In another paper by them in year 2007 titled "Educational data mining: A survey from 1995 to 2005" surveyed the application of data mining to traditional educational systems. They concluded that much more specialized work is needed in order for educational data mining to become a mature area.

5.2 Predicting academic performance with Pre/Post Enrollment Factors

Most of the published research papers belong to this category. Latest work published in International Journal of Computer Science and Mobile Computing, 2014 describes the process of finding the set of weak students based on graduation and post-graduation marks. Another paper published in European Journal of Scientific Research in 2010 also analysed students' learning behaviour to predict weak students. P. Ramasubramanian, K. Iyakutti and P. Thangavelu, in year 2009 also predicted weak students using rough set theory. A comprehensive evaluation method for undergraduates; that can objectively distinguish the grades of students was developed by Xiewu, Huacheng Zhang, Huimin Zhang in year 2010 [40]. Another study by Dai Shangping, Zhang Ping, in year 2008 predicted final grades of students based on features extracted from log data in web-based system and published their work in IEEE.

5.3 Comparison of Data Mining Techniques in predicting academic performance of students

In 2009 Fadzilah Siraj and Mansour Ali Abdoulha compared three techniques for understanding undergraduate's student enrolment data and published their work in IEEE. Nbtree was identified as the best classifiers to predict student sequences for course registration planning in paper published by Pathom Pumpuang, Anongnart Srivihok and Prasong Praneetpolgrang in year 2008 in IEEE. Decision tree proved to be consistently 3-12% more accurate than the Bayesian network in predicting academic performance of undergraduate and postgraduate students in a paper titled "A Comparative Analysis of Techniques for Predicting Academic Performance" in 2007, IEEE.

5.4 Other Areas of performance Analysis

International Conference on Intelligent Computational Systems published a paper titled "A Classification Model For Edu-Mining" for faculty evaluation based on different parameters [20]. Another paper published in 2009 in IEEE proposed to build evaluation index system and teaching index method based on data mining. Hua-Long Zhao, in his paper titled "Application of OLAP to the Analysis of the Curriculum Chosen by Students" published in IEEE in 2008, analysed curriculum's establishment from many angles [60]. Prediction of learning disabilities of school-age children was done by Julie M. David and Kannan Balakrishnan in 2010 [43]. In 2007, Vasile Paul Brefelean analysed students' choice in continuing their education with post University studies (Master degree, PhD) using data mining techniques.

6. RESULT AND ANAYSIS

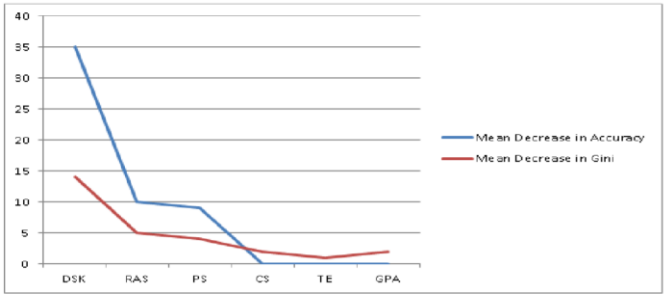
Table 1 shows the results obtained by Random forest. Table 2 shows the Variable Importance values for each of the attributes in terms of Mean-Decrease in Gini and Mean-Decrease in Accuracy.

Table -1: Accuracy Measures

CLASS	TP RATE	FP RATE	AREA Under ROC Curve
Good	0.930	0.027	0.977
Average	0.824	0.04	0.984
Poor	0.920	0.076	0.331

Firstly we look at the average area under the ROC curve and as we can see, this area is about 0.984 (average). The area here is close to 1 and for ROC curves; an area 1 refers to the highest possible accuracy of 100%. Hence, we can see that the Random-Forest Model used is highly accurate and strong conclusions can be drawn from these results. We also found through a comparative study that the model outperformed

knowledge-based decision trees and Linear Regression techniques when applied to the same data.



As we see in these plots, the variables like Domain-Specific knowledge, Reasoning and Analytical skills and Programming Skills are the most important attributes to be considered during recruitment of new personnel, as they have the maximum contribution towards the homogeneity of nodes and classification of data. Another important result obtained in this scenario is the Mean Decrease in Accuracy value for GPA which comes out to be 0. Such low scores for these values stand as a scientific base to challenge the usual recruitment procedure where maximum importance is given to GPA of candidates. What we see here is that, based on the grades and academic scores of students from universities, one cannot predict their performance in the industry. Hence GPA alone is not a very clear reflection of the candidate's capabilities as far as the software industry is concerned. It is needed that the companies separately test the other relevant attributes of the students in order to make a better decision. The software development process involves various intricate steps and complex stages where many other factors and abilities of an individual come into play. Referring to the GPA alone will not yield optimal results and this is the reason, why there has been a changing trend in the recruitment procedure today. Recruiting teams are looking for candidates with a complete package in terms of overall personality, analytical thinking abilities and good inter-personal skills apart from good grades.

```
DKA == GOOD
!       RS == ( GOOD) || (AVERAGE) : Accept
!       RS == POOR
!       !       PS == GOOD : Accept
!       !       PS == (AVERAGE) || (POOR) : Reject
DKA == AVERAGE
!       RS == GOOD : Accept
!       RS == AVERAGE
!       !       PS == GOOD
!       !       !       CS == GOOD : Accept
!       !       !       CS == (AVERAGE) || (POOR) : Reject
!       !       PS == (AVERAGE) || (POOR) : Reject
!       RS == POOR : Reject
DKA == POOR : Reject
```

As we see, during the selection process companies must initially clarify the values for each of the variable parameters that are acceptable to them. For example some companies may accept students with average programming skills but others may only want those who have great programming skills whatever may be the scores for other attributes. Once that's done, they must start classifying students based on the features in the order of their variable importance values. Another important aspect here is to remember that in numerous scenarios, there may be a very large feature set. In such a case, it is not possible for the companies to consider all of them.

7. CONCLUSIONS

Since the alignment of the model is finished utilizing the out-of-sack tests, the model doesn't experience the ill effects of over-fitting issues which encourages a prevalent exhibition in the vast majority of the situations when contrasted with Decision-trees. We plainly observe from the outcomes that the GPA/evaluations of newcomers unmistakably aren't among the most significant choice properties on which a recruiting choice can be based.

We try to build upon them and use this research to build a better and more robust model that can be not be applied to different scenarios but also work well on different data sets having varied properties with minimal or no changes. There are a wide range of algorithms in each of these categories, many of which are implemented on WEKA and R. These tools are platforms with GUI and command-line implementations respectively, with a number of machine-learning algorithms for data mining tasks, with a variety of options for regression, classification, data pre-processing, association rules, clustering and visualization.

8. FUTURE SCOPE

Future headings in our work incorporate further developed information pre-handling and cleaning methods that may additionally improve the characterization results and upgrade data extraction from sets of expectations. Since this was the primary examination of the IrishJobs.ie opportunities that utilized information mining in a R situation and printed factors, I am likewise keen on looking at the viability of this technique after some time, particularly in the context of identifying shifts in skills demands. One of the most recent and biggest challenge that higher education faces today is making students skillfully employable. Many universities/institutes are not in position to guide their students because of lack of information and assistance from their teaching-learning systems. To all the more likely regulate and serve understudy populace, the colleges/establishments need better evaluation, examination, and expectation devices.

Considerable amount of work is done in analyzing and predicting academic performance, but all of these works are segregated. There is a clear need for unified approach. Other than academic attributes, there are large numbers of factors that play significant role in prediction, which includes no cognitive factors (set of behaviors, skills, attitudes). Suitable data mining techniques are required to measure, monitor and infer these factors for prediction. Thus enriching the

input vector with qualitative values may increase the accuracy rate of prediction as well.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques, 2/e", Morgan Kaufmann Publishers, An imprint of Elsevier, 2010.

[2] C.F. Chien and L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems and Applications, vol. 34, 2008, pp. 280-290.

[3] HamidahJantan , "Human Talent Prediction in HRM using C4.5 Classification Algorithm", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, pp. 2526-2534.

[4] Suma.V, Pushpavathi T.P, and Ramaswamy. V, "An Approach to Predict Software Project Success by Data Mining Clustering", International Conference on Data Mining and Computer Engineering (ICDMCE'2012), pp. 185-190.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] P. Singh, "Comparing the effectiveness of machine learning algorithms for defect prediction", International Journal of Information Technology and Knowledge Management, 2009, pp. 481-483.

[6] J. R. Quinlan, "Introduction of decision tree", Journal of Machine learning, 1986, pp. 81-106. Pool, Lorraine Dacre, Pamela Quilter, and Peter J. Sewell. "Exploring the factor structure of the Career EDGE employability development profile." Education+ Training 56.4 (2014): 303-313.

[7] Saranya, S., R. Ayyappan, and N. Kumar. "Student Progress Analysis and Educational Institutional Growth Prognosis Using Data Mining." International Journal Of Engineering Sciences & Research Technology, 2014.

[8] Pandey, Umesh Kumar, and Brijesh Kumar Bhardwaj. "Data Mining as a Torch Bearer in Education Sector." Technical Journal of LBSIMDS (2012).

[9] Srimani, P. K., and Malini M. Patil. "A Classification Model for Edu-Mining." PSRC-ICICS Conference Proceedings. 2012.

[10] Sukanya, M., S. Biruntha, Dr S. Karthik, and T. Kalaikumaran. "Data mining: Performance improvement in education sector using classification and clustering algorithm." In International conference on computing and control engineering,(ICCCE 2012), vol. 12. 2012.

[11] Yadav, Surjeet Kumar, and Saurabh Pal. "Data Mining Application in Enrollment Management: A Case Study." International Journal of Computer Applications (IJCA) 41.5 (2012): 1-6.

[12] Minaei-Bidgoli, Behrouz, and William F. Punch. "Using genetic algorithms for data mining optimization in an educational web-based system." Genetic and Evolutionary Computation—GECCO 2003. Springer Berlin Heidelberg, 2003.

[13] Handel, M. Trends in Job Skill Demands in OECD Countries. OECD Social, Employment and Migration Working Papers, No. 143, 2012. Available online: http://dx.doi.org/10.1787/ 5k8zk8pcq6td-en (accessed on 18 October 2015).

[14] Manacorda, M.; Manning, A. Just Can't Get Enough: More on Skill-Biassed Change and Labour Market Performance; London School of Economics and Political Science: London, UK, 1999.

[15] UNESCO. International Standard Classification of Education ISCED 2011; UNESCO Institute for Statistics: Montreal, QC, Canada, 2012.

[16] Litecky, C.; Aken, A.; Ahmad, A.; Nelson, H.J. Mining for Computing Jobs. IEEE Softw. 2010, 27, 78–85.