Available online at: https://ijact.in

Date of Submission	21/11/2018
Date of Acceptance	27/12/2018
Date of Publication	07/01/2019
Page numbers	2969-2975 (7 Pages)

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

FUZZY-FILTER FEATURE SELECTION FOR ENVISIONING THE EARNINGS OF HIGHER EDUCATION GRADUATES

Tahseen A. Wotaifi, Eman S. Al-Shamery
Department of Software, University of Babylon, Iraq
tahseen.ubabylon@gmail.com, emanalshamery@itnet.uobabylon.edu.iq

Abstract: With the great change in the labor market and contemporary education, a high consideration is paid for the issue of graduates' employability and their earning. However, educational institutes and universities have different polices for preparing graduates to compete strongly in the labor market based on their academic skills and experience. This paper aims at (1) predicting the earnings of alumni for six years after graduation, and (2) identifying the most important factors that directly affect the earnings. Hence, the accuracy of the prediction can be improved based on the higher education system in the United States. Unlike previous research, this study contributes to the use of fuzzy logic on three filter methods: Relief Attribute Evaluation, Correlation Attribute Evaluation, and classifier Attribute Evaluation. Because these three methods provide different weight to the same attribute, the fuzzy logic is used to obtain a single weight. The proposed system is applied on higher education of United States; specifically the college scorecard data set that contains nearly (8000) educational institutions and (1825) feature for each. The proposed Fuzzy-Filter method has selected 45 features only. Accordingly, two models are used to predict graduates' earnings, which are Random Forest and Linear Regression with Mean Absolute Error (MAE) (0.055) and (0.065) respectively. The research findings are better in terms of Mean Absolute Error (MAE) and reducing the number of features in comparison to previous studies.

Keywords: Earning prediction of Alumni; Fuzzy-Filter Feature Selection; Linear Regression; Mining Higher Education; Random Forest.

I. INTRODUCTION

Higher education is a good investment for graduates that should be motivated by schools, families, policymakers, and communities so that researchers show that the wages of university graduates have doubled and there is a large gap in earning between the median high-school educated worker and the median college-educated worker as well as explained that higher education "makes life better" during family stability, a host of social benefits in community relations, and health. This means that the common denominator of all educational systems is to provide job opportunities. It should be noted that the prestige of the university or educational institution is one of the most important factors affecting the income of graduates and the increase in employment opportunities [4].

Many students believe that the reputation of the education institution helps them more after graduation in obtaining employment opportunities. In fact, this is not the only factor, where there are many factors associated with an educational institution that affect the earning of graduates. The United States Department of Education has shown its assistance to students and their families in this field by launching college scorecard dataset in September 2015. This dataset contains hundreds or thousands of factors for educational institutions in the United States as well as factors specific to students who have entered educational institutions in the United States such as demographics, race, tuition, financial aid, fees, family income, and much more. This dataset designed for putting the power in the hands of the students from those choosing colleges to those improving college quality to see how well different schools are serving their students[1][2].

Because the main benefit of data mining is the discovery of hidden patterns from large databases, there are growing research interests in the use of data mining techniques in education[5][14].This emerging field (Educational Data Mining) is interested in developing techniques to discover valuable information from the educational database in order to analyze the student's direction and behaviors towards education, in another word, what is meant in educational databases are student databases, so educational data mining produces many techniques that can help the educational system and discovery of any information related to students such as improve the learning experience of students, earning of alumni, student performance, the failure of students and thus sending early alerts to faculty of the poor performance of their students and more. [3][6].

According to higher education system in the United States, this study has focused on the important factors which affect the earning of graduates according to higher education system in the United States through the applying many machine learning techniques on the educational database (which is called collage scorecard dataset) then analysis and extraction of the most important features and thus enable prospective students to benefit and select the educational institution before entering it.

II. RELATED WORK

As dataset is a newly-released dataset, so there are not many studies on it [1]. Through the study of previous works, it has been found that many studies have been interested in the effect of the selection of the educational institution on the future of students in terms their income during the study and after graduation.

Miranda Strand, Tommy Truong, (2016), have produced more than a model to predict graduates' earnings. The dataset has used in this study is the college scorecard dataset where the focus was on data released from the US Treasury Department only after that has been determined the factors affecting the earning of graduates. The studies have been valuable but in fact that many other factors are found directly affect the earning of graduates.

Monica Agrawal, Priya Ganesan, Keith Wyngarden, (2017),have attempted to use a variety of machine learning models such as linear regression and support vector machine to make predictions regarding post-college earnings of alumni. The dataset used in this study are also college scorecard dataset. The analysis of the dataset in this project has been good but the number of factors has been identified as impressive on the earnings of graduates has been (170) attributes and this is considered very large in addition to the errors of the models have been used in this study have been unconvincing.

John M. Nunley, Adam Pugh, Nicholas Romero, and Richard Alan Seals, Jr, (2015), have used data mining techniques to create models that predict the impact of academic specialization on employment. The weakness in this study is the few data used, where this data has been compiled by submitting random messages on the Internet to advertise vacant posts and analyzing the responses.

Denise Jackson and Grant Michelson, (2015), have presented a model predicting the identification of the most important factors affecting the employment of graduates. It was an important study where they found that demographic characteristics had a greater impact. Determining Ph.D. students only by analyzing underestimated the importance of this project.

Patrick Premand, Stefanie Brodmann, Rita Almeida, RebekkaGrun, and Mahdi Barouni, (2015), have relied on the randomized assignment of university students and have tried to extract an additional factor, namely, entrepreneurship education for students. They have found that this feature contributed to the increase in employment opportunities

Ewan Wright, Qiang Hao, Khaled Rasheed, and Yan Liu, (2017), have provided more than one method from feature selection methods to determine the factors affecting the earnings of graduates. The steps followed in this study have been good but when they have been neglected four important categories of college scorecard dataset the importance has been significantly reduced.

III. THEORETICAL BACKGROUND

A. Higher Education System

Higher Education System in the United States is one of the important educational systems in the world and occupies ranks high in the rankings. The system of higher education in America is characterized by many advantages, including the time required to obtain a certificate is longer than the rest of the systems, financial assistance has received by students, and others. The most important elements that make the higher education system in the United States more powerful is academic freedom, there are a vast number of scientists and researchers in the United States, equality of opportunity between individuals, and much more[7]. It is worth mentioning that the database of higher education or the database of students has increased rapidly as a result of technological progress. In order to extract important information from this dataset, a new field has emerged for this purpose, called Educational Data Mining[4][5].

B. Feature Selection

It is one of the most important methods that can be used in data preprocessing to effectively reduce data or is the process of selecting the most important set of attributes from all original attributes because the features are usually categorized as: strongly relevant, weakly relevant, but not redundant, redundant, and irrelevant. The features that strongly relevant to the target are selected through three broad-use strategies: filter, embedded and wrapper. There are many types of filter methods such as correlation attribute evaluation, relief attribute evaluation, and others [8][9].

C. Data Mining Techniques

1. Random forest is a regression and classification technique based on the pooling of a large number of decision trees. These large numbers of trees are created from the training set and validated to predict future observations, and can have a categorical or a continuous value output. There are many strong points of the Random forest such as can be used for both classification and regression task, handle missing value and maintains accuracy for missing data, and handle large data set with higher dimensionality[11][12].
2. Linear Regression is a statistical procedure used to evaluate the relationship between two variables. The first variable is called independent variable (feature) and the second variable is called dependent variable(target or class)[16]. Linear regression takes the following formula:

   $Y=a+bX$...

Where Y is the dependent variable, X is the independent variable, 'a' is called the Y-intercept, or simply the intercept, and 'b' is the slope of the line [10].

IV. Research Methodology

A. Dataset

Dataset is collected by the US Department of Education in 2015 and has been named College Scorecard[1]. These datasets contain much valuable information that helps students and families determine which university suits their abilities. More than 8000 educational institutions and exactly 1825 variable including the degrees and majors offered, demographics' about the students at each college, the cost, the financial aid, loan and pall grant of students and more divided into nine categories: costs, financial aid, admission, school, academics, repayment, student, completion and earning. Many of the details of this dataset are given in the table (1) listed below.

Table1: Number of factors in each category

category	*No. of features*	*description*
cost	*52*	*important information about tuition fees and other costs*
financial aid	*40*	*information about educational institutions that provide financial aid to students such as loans, pall grant, and others*
admission	*25*	*information about admissions rates and SAT/ACT scores of students*
school	*170*	*basic information about the educational institution*
academics	*228*	*information about types of academic offerings*
repayment	*131*	*information on the percentage of loan repayment by students*
student	*96*	*demographic information about the student body*
completion	*1013*	*information includes the percentage of completion or drops out of students and others*
earning	*70*	*Information includes student earnings, family income, and others*

In Table 1 above there are 70 features in the category of earning. These features are the mean and median of earnings of graduates in educational institutions for 6, 7, 8, 9 and 10 years after graduation. In other words, this study faced the problem of determining the target. Therefore, these features have been studied in detail and the median earning for six years after graduation has been identified as the target class.

B. Preprocessing

The biggest challenge in this project is the curse of dimensionality, so preprocessing has been done. Preprocessing play the key role in data reduction and have used two steps. The first step has been to remove features that were not useful in prediction such as ID number as well as removing any feature containing a single value for all instances. The second step is to delete any attribute containing missing values of more than fifty percent. Then the missing values have been processed in each feature, applying two methods to calculate the missing values which are mode method and the mean method. Each method has been applied depending on the type of feature. Algorithm (1) shown below summarizes the above.

Algorithm (1): Data Preprocessing

Input: College Scorecard Dataset
Output: valuable features

Let: (Dij) is the array of features values
where: (i ⟶ n) [n] is the number of features
(j ⟶ m)[m] is the number of institutions
(V) represent feature value
(F) represent feature
(MV) represent missing values
(VFij) is the array of valuable features
(NFij)is the array of neglected features
(μ) represent mean
(M) represent mode

For each V∈ Dij.
Begin
If (all V in F are equal or MV in F > 0.5)
NFijF ← // neglected features
Else
VFijF ← // valuable features
End for
For each MV ∈ VFij
Compute (μ), identify (M)

10. MV = μ or M//According to the attribute type
11. End for
12. Return VFij

V. DEVELOPED FILTERING METHOD FOR FEATURE SELECTION

The median earning 6 years after graduation was the target in this study. In order to reduce the dimensions of the college scorecard dataset, three filter methods of feature selection techniques have been applied in this study which are Relief Attribute Evaluation, Correlation Attribute Evaluation, and classifier Attribute Evaluation, each technique gives different weight to the same feature. Minimize the dimensionality of data and select related features are done in two stages:

The first stage depends on the technique of feature selection itself. Using the relief attribute evaluation method that gives weight to the feature between (1) and (-1).Any feature with a negative weight has been deleted either in the correlation attribute evaluation technique that gives weight to the feature between (0-1).The attributes with a weight equal to zero have been deleted. A classifier Attribute Evaluation method that evaluates the worth of an attribute by using a user-specified classifier. The Zero R has been used as a classifier and then has removed irrelevant attributes.

The second and most fundamental stage is to use the fuzzy logic and give one weight to each feature. This is done by the different weights whose attribute have been evaluated by the three algorithms above. It should be noted that the membership function has been used to find a fuzzy value for each crisp value explained in Figure 1 (crisp is the weight of feature).

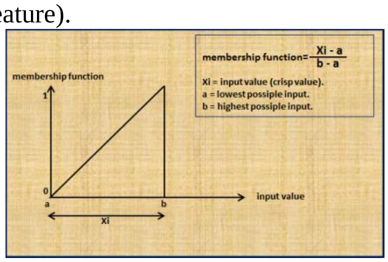

Figure1: fuzzy logic to identify the significant features

The following formula has also usedto calculate the new weight (de-fuzzification), which represents the final weight of the feature.

$$\text{new weight} = \frac{\sum \mu Xi * Xi}{\sum \mu Xi}$$

After each feature had only one weight by applying the fuzzy logic. Our main goal is to obtain the highest accuracy possible and with a few low features. For these, the attributes which have a weight less than (0.5) have been omitted. Algorithm 2 illustrates the algorithmic descriptive.

Algorithm (2): feature selection

Input: VFij// the output of the data preprocessing algorithm
Output: significant features
Let: (VFij) is the array of features values
where: (i ⟶ n) [n] is the number of features
(j ⟶ m)[m] is the number of institutions
(V) represent feature value
(F) represent feature
(W) represent the weight
(MV) represent missing values
(SF1ij) significant features //Relief
(N-SF1ij)no significant features //Relief
(SF2ij) significant features //correlation
(N-SF2ij)no significant features //correlation
(SF3ij) significant features //classifier
(MS) membership function
(DFU) defuzzification
(SIGFij) is the array of output

Relief Attribute Evaluation

1.For each F ∈ VFij
2. Compute W of feature
3. if (W < 0)
4. N-SF1ij ⟵ W //non-significant features.
5. else
6. SF1ij ⟵ W // significant features.
7.End for

Correlation Attribute Evaluation

8.For each F ∈ VFij
9. Compute W of feature
10. if (W1 = 0)
11. N-SF2ij ⟵ W //non-significant features.
12. else
13. SF2ij ⟵ W // significant features.
14.End for

Classifier Attribute Evaluation

15. For each F ∈ VFij
16. if (F not prune from classifier)
17. Compute W of feature
18. SF3ij ⟵ W// significant features.
19. End for

Fuzzy logic

20. for each (F∈SF1ij or SF2ij or SF3ij)
21. MS=(Xi-b)/(a-b) //a= the least possible value, b=the greatest possible value, x=weight
22. DFU=(∑MSi ∗ Wi)/(∑MSi)
23. if (DFU >= 0.5)
24. SIGFij ⟵ DFU
25. End for
26. Return SIGFij

In order to validate the accuracy of the selection of these features and its effect on student earnings, two different approaches have been used. This study sought to represent the data in the form of a tree through the use of random forest technique and also used a linear regression technique to know the relationship between features and target.

VI. EXPERIMENTAL RESULT AND DISCUSSION

The main purpose of our project is to reduce the factors to the lowest possible while maintaining high accuracy, so the process of neglecting the weak relevant features and the selection of strong relevant features have been applied in stages. The first reduction of features with data pre-processing where the number of features remaining after this process is (912) features. The second reduction of the factors has been through the use of three filter techniques for feature selection. The last reduction of the attributes has been done through the use of fuzzy logic and giving one weight for each feature. Then any feature with a weight of less than 0.5 has been discarded. After this process, (46) features only remained by reviewing the features chosen by each method separately, it has been found that these (46) features have been nominated by the three methods above.

The table (2) below shows two different methods has been used to predict the graduates' earnings which are Random Forest technique and Linear Regression technique. The basic error metrics are both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). There has been a difference in error ratios in both methods as shown.

Table2: Error for earning across the two models

Models	MAE	RMSE
Random forest	0.055	0.078
Linear Regression	0.065	0.087

These results have been compared with previous research, for example, the second research in the related works above mentioned ("Prediction of Post-Collegiate Earnings and Dept. (2017)" from Stanford University). In this research, five models have been used to predict the earnings of graduates: linear regression, weighted linear regression, SVM, KNN, and neural network. Mean Percent Error for each model is (13.42), (9.60), (14.11), (23.80), and (11.93) respectively with 170 features.
As mentioned above, this study identified only 45 features that are strongly relevant to the target. The table (3) illustrates these factors.

Table3: The important factors

No	category	features	weights
1	Student	DEP_INC_AVG	0.6254
2	School	AVGFACSAL	0.6154
3	Completion	NOLOAN_COMP_ORIG_YR3_RT	0.6065
4	Completion	NOLOAN_COMP_ORIG_YR4_RT	0.6030
5	Admission	ACTMT25	0.6002
6	F. aid	PELL_EVER	0.5989
7	Repayment	MALE_RPY_7YR_RT	0.5976
8	Admission	ACTMTMID	0.5973
9	F. aid	PCTPELL	0.5950
10	Student	FAMINC	0.5939
11	Admission	SATMT25	0.5739
12	Admission	ACTMT75	0.5727
13	Admission	SATMTMID	0.5672
14	F. aid	MD_FAMINC	0.5544
15	Student	DEP_INC_PCT_LO	0.5542
16	Student	INC_PCT_LO	0.5517
17	Repayment	CDR3	0.5489
18	Admission	SAT_AVG_ALL	0.5475
19	Completion	COMP_ORIG_YR6_RT	0.5425
20	Admission	SATMT75	0.5404
21	Completion	NOT1STGEN_COMP_ORIG_YR6_RT	0.5396
22	Completion	MALE_COMP_ORIG_YR6_RT	0.5386
23	Student	PAR_ED_PCT_HS	0.5356
24	Admission	ACTCM25	0.5350
25	Completion	DEP_COMP_ORIG_YR4_RT	0.5317
26	Completion	NOT1STGEN_COMP_ORIG_YR4_RT	0.5310
27	Admission	SAT_AVG	0.5295
28	Completion	COMP_ORIG_YR4_RT	0.5272
29	Completion	MALE_ENRL_ORIG_YR2_RT	0.5266
30	Admission	ACTEN25	0.5248
31	Completion	MALE_COMP_ORIG_YR4_RT	0.5245
32	Completion	DEP_COMP_ORIG_YR6_RT	0.5238
33	Completion	FIRSTGEN_COMP_ORIG_YR4_RT	0.5208
34	Admission	ACTCMMID	0.5193
35	Student	INC_PCT_H2	0.5155
36	Completion	FIRSTGEN_COMP_ORIG_YR6_RT	0.5152
37	Completion	NOT1STGEN_COMP_ORIG_YR8_RT	0.5129
38	Completion	LO_INC_COMP_ORIG_YR6_RT	0.5123
39	Completion	MALE_COMP_ORIG_YR8_RT	0.5110
40	Completion	NOT1STGEN_ENRL_ORIG_YR2_RT	0.5101
41	Completion	FEMALE_COMP_ORIG_YR6_RT	0.5077
42	Completion	FEMALE_COMP_ORIG_YR4_RT	0.5054
43	Completion	LO_INC_COMP_ORIG_YR4_RT	0.5041
44	Completion	FIRSTGEN_ENRL_ORIG_YR2_RT	0.5021
45	Completion	COMP_ORIG_YR8_RT	0.5019

By looking at Table (2), it has been found that some categories have been repeated more than once (student, completion, admission, repayment, financial aid, school) and others have not been influential in earnings of graduates (academic, cost). The table (4) shows the categories that have been repeated and the number of factors in each category.

Table 4: the number of factors in each category

Category	Number of factors
Student	6
Completion	22
Admission	11
Repayment	2
financial aid	3
school	1

More details, the dataset recently updated by the US Department of Education (called college scorecard dataset) has been used in this project to identify the most important features (or the attributes that have more influence) on the median earning 6 years after graduation. Before any of the above prediction methods are used, three filter methods have been applied to the selection of the attribute:1) Relief Attribute Evaluation. 2) Correlation Attribute Evaluationand3) classifier Attribute Evaluation.
The results of each algorithm have then been reviewed. It has been found that "Relief Attribute Evaluation Technique" is interested in factors belonging to 1) completion category, 2) admission category, 3) students category 4) financial aid category, 5) school category and

6) repayment category (ranking is better to worse). "Correlation Attribute Evaluation technique" is perfectly consistent in class arrangement with "Relief Attribute Evaluation Technique". "Classifier Attribute Evaluation" is interested in factors belonging to 1) admission category, 2) completion category, 3) financial aid category, 4) students category, 5) repayment category, and 6) school category.

This study has used a different technique to determine the important factors affecting the target by applying the fuzzy logic to the weights resulting from the feature selection techniques above.

The fuzzy logic technique has been applied because it takes into consideration any feature that has been nominated by any of the three methods mentioned previously. Emphasis has been placed on factors with a weight greater than or equal to (0.5) resulting from the fuzzy logic technique. The number of these factors is only (45). Among the top ten features selected have, for example, the rate of salaries of faculty and the percentage of students who did not receive PELL grant, the average of family income, and SAT & ACT scores.

If the effect of each feature is discussed separately on the earnings of graduates, the salaries of the faculty may reflect the quality of the institution itself. In other words, it is possible that prestigious educational institutions attract a more efficient faculty. Consequently, this affects directly the earning of alumni. It is clear that there is a high correlation between the two features of the percentage of students who did not receive PELL grant and the average of family income because the students who did not need a PELL grant may have family with a high income and the students who do not suffer from the burden of money may have more than an opportunity such as choosing the appropriate college and others. Previous research has shown the strength of the relationship between graduates' earnings and SAT score[13]. Educational institutions determine their own SAT & ACT score[15], so students with higher SAT scores tend to attend more selective colleges and are therefore more likely to receive higher income after graduation. Through the use of two different approaches to predict the earning of graduates (linear regression and random forests) and according to the results, it has been found that the method of random forests slightly better than linear regression method. The reason for this may be because the dataset is too large and random forest method performs a type of feature selection as well as this method also performs an important function of pruning the nodes.

VII. CONCLUSION

Some educational institutions as a result of their importance in the US may be considered the most important criterion for enhancing the income of the graduate.This study aims at identifying the most important factors affecting the earnings of alumni based on a college scorecard dataset recently released by the US Department of Education. This dataset contains a huge number of factors, making it difficult to understand. The curse of dimensionality is the most important challenge we faced in the analysis of this dataset. However, the key contribution of this research is using the fuzzy logic technique on three filtering methods for feature selection which are Relief Attribute Evaluation, Correlation Attribute Evaluation, and classifier Attribute Evaluation. These three methods provide different weight to the same attribute, so the fuzzy logic has used to obtain a single weight. Through the final weights resulting from the fuzzy logic technique, the features with higher weight have been adopted. As such, this study has identified the most important factors affecting the graduates' income and it has been found that saving rate of selection more than 91% for features.. Through the method used in removing the non-important factors, it is found that all factors belonging to the categories of academics and cost are not important. To prove that the features identified are relevant with the target, two techniques have been used to predict post-graduate employments (earnings of alumni) which are "Random forests technique" and "Linear regression technique". The performance of the two models applied was different, although both agreed that some factors such as family income, SAT and ACT scores are important. However, the method of random forest has yielded slightly better results than linear regression in terms of Mean Absolute Error (MAE) because it performed a type of attribute selection. It is hoped that our work complements with the rest of the research in this field by offering more detailed insights into post-graduation incomes.

VIII. REFERENCES

[1] Monica Agrawal, PriyaGanesan, Keith Wyngarden.Prediction of Post-Collegiate Earnings and Debt, (2017).

[2] Miranda Strand, Tommy Truong.Predicting Student Earnings After College, (2016).

[3] Bhise R.B., Thorat S.S., Supekar A.K. Importance of Data Mining in Higher Education System, (2013).

[4] Ewan Wright, QiangHao, Khaled Rasheed, & Yan Liu. Feature Selection of Post-Graduation Income of College Students in the United States, (2018).

[5] Brijesh Kumar Baradwaj, Saurabh Pal. Mining Educational Data to Analyze Students‟ Performance, (2011).

[6] M.Kavitha and , and , P.T.Srinivasan, G. Renuga and L.V.Jayakumar. "Evaluation of Antimitotic Activity of Mukiamaderaspatana L. Leaf Extract in Allium cepa Root Model." International Journal of Pharmacy Research & Technology 4 (2014), 01-04.

[7] Peter D. Eckel, Jacqueline E. King .an overview of higher education in united states, (2005).

[8] A. Jović, K. Brkić,and N. Bogunović. A review of feature selection methods with applications, (2014).

[9] Mark A. Hall. Feature Selection for Discrete and Numeric Class Machine Learning, (1998).

[10] HarveyMotulsky, Arthur Christopoulos. Fitting Models to Biological Data Using Linear and Nonlinear Regression, (2002).

[11] Anne-Laure Boulesteix, SilkeJanitza, JochenKruppa, and Inke R. König. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics, (2012).

[12] Bin Luo, Qi Zhang, Somya D. Mohanty. Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment, (2018).

[13] Michael T. French, Jenny F. Homer, IoanaPopovici, and Philip K. Robins. What You Do in High School Matters: High School GPA, Educational Attainment, and Labor Market Earnings as a Young Adult, (2015).

[14] Eman Al-Shamery and Ameer Al-Haq Al-Shamery. A New Deep Neural Network Regression Predictor Based Stock Market, (2018)

[15] J., Yakubu H., and AboiyarT.. "A chaos based image encryption algorithm using Shimizu-Morioka system." International Journal of Communication and Computer Technologies 6 (2018), 7-11.

[16] Eman Al-Shamery and Ameer Al-Haq Al-Shamery. Enhancing Prediction of NASDAQ Stock Market Based on Technical Indicators, (2018).