|  |
| --- |
| **TEAM BANANA** |
| FINAL REPORT |
| **Analysis of Influential Factors on Employment** |

|  |
| --- |
| **Haeryoung Kim, Ohsang Yoo, Taewoo Kim, Soyeon Kim, Hyeongeun Lee** |

**ABSTRACT**

We wonder what things have the powerful effect on Graduates’ job employment among various variables. Our team has narrowed down to five variables, gender, age, GPA, English score and internship experience. For analysis, using models are Logistic Regression, Quadratic Discriminant Analysis, and Classification Tree. Since the most predictive of model is the Classification Tree, it is selected as the final model to predict the employment success. As we expected, GPA scores are significant to get a job and English capacity is also important. In conclusion, we suggest that graduates who are looking for a job should focus on not only getting high GPA but also acquiring or improving their general capability.

**Ⅰ. Introduction**

In these days, getting a job is the first priority for many university students. Students work hard on their own way for better job opportunity. Since there are so many influential variables for job opportunities, we wonder what things have the powerful effect on Graduates’ job employment. Among various variables, our team have narrowed down to five variables, gender, age, GPA, English score and internship experience that we judge as influential variables. We expect to find the relationship between those variables and the employment through some analysis. Our goal is to find the relationship between employment, response variable, and main factors, predictors, what we are interested in. The dataset we are going to analyze is from ‘2013 Graduates Occupational Mobility Survey in Korea' researched by Korea Employment Information Service. To find out the relationship, we choose five factors with 5273 observations among hundreds of variables in the raw data; GENDER, AGE, GPA, ENG\_SCORE and INTERN\_DUMMY.

**Ⅱ. About Dataset**

Factor 'GENDER' is one of what we want to seek. In Korea, sexual discrimination in employment field is severe. We assume that so called 'Glass Ceiling' can affect when getting a job in Korea. Therefore, we pick ‘1’ as male, and ‘0’ as female in gender variables. ‘GPA’ is students' grades so it is essential source for entering company. 'ENG\_SCORE' variable is Certified English Test including TOEIC, TOEFL PBT·CBT·IBT, and TEPS which is converted by standard criteria. ‘AGE’ variable is Korean birth standard. This factor influences employment field. Because we consider that internship experience can affects employment. Factor 'INTERN\_DUMMY' is also included in our dataset. The value '1' means the graduate student who has internship experience, and the other is opposite. From now with these five variables, we will analyze the effects to job employment.

**Ⅲ. Methods**

The dataset contains attributes of Korean undergraduate students, and whether they are employed or not. Students were asked questions regarding their GPA, experience of internship, average score of the English exam (i.e. TOEIC), gender, and age. Some variables are continuous, and some are categorical. We researched the dataset, students’ attributes regarding job employment with logistic regression model, Quadratic Discriminant Analysis and classification tree model. These models can provide predictions regarding employment, and we have compared powers of these models. With this process, we have concluded most appropriate model to the dataset, and get conclusion of the research.

Firstly, ‘Logistic Regression Model’ can handle categorical (binary) responses when original regression model cannot research them effectively. We found correlations between job employment and all independent variables including categorical variables. The logistic regression model provided coefficients of each predictor that explains a possibility of the event, 'getting a job' occurs. And then, our team took fitting process to find our analysis was appropriate. However, we cannot sure that this model is appropriate for the research. So we took other methods to find the right model. To get the most significant results, we optimized Logistic Regression Model, Quadratic Discriminant Analysis, and Classification Tree Model. To compare the power of each model and choose the most powerful model, we will use ‘misclassification ratio’ as a criterion. This misclassification ratio can be derived by comparing the predicted values of job employment and the real values. By random sampling the whole 5273 in the rate of 6:4, training dataset and test dataset have 3163 and 2110 observations for each.

**ⅰ) Logistic regression model**

The logistic regression analysis provides coefficients of each prediction that explain a possibility of the event, “get a job” occurs. Firstly, the following <table1> shows the results of logistic regression analysis. From the table, we can conclude that there exist not influential variables since the p-values of ‘GENDER’ and ‘GPA’ are not significant. Therefore, to analyze further, we used ‘Stepwise Selection Method’ for selecting significant variables.

|  |  |  |
| --- | --- | --- |
| **Predictors** | **Estimate Value** | **P-value** |
| **Intercept** | 2.0148396 | 0.000392 |
| **GENDER** | 0.1994438 | **0.020409** |
| **AGE** | -0.0604207 | 0.000435 |
| **GPA** | 0.0232155 | **0.805629** |
| **ENG\_SCORE** | -0.0003073 | 0.005259 |
| **INTERN\_DUMMY** | 0.2725206 | 0.001982 |

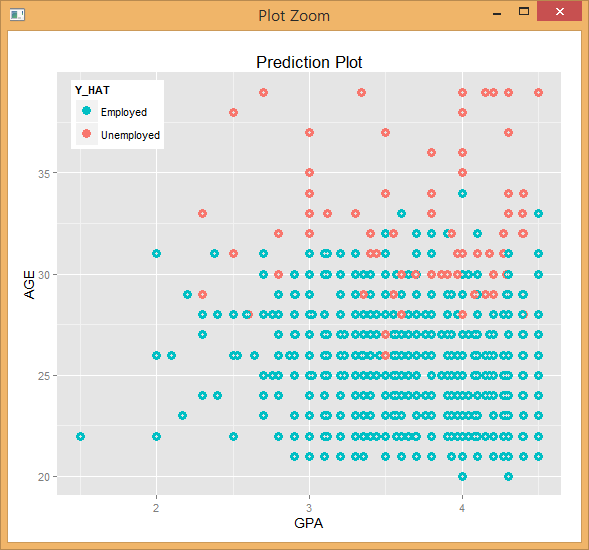
**<table.1>** Results of logistic regression analysis

The following <table2> shows the result of Stepwise Selection Method. In the coefficient values in the table below, plus signed values mean that relatively higher possibility of getting a job, and vice versa. If you want to calculate more accurate values, you have to adopt exponential function. For example, in case of GENDER variable, men are likely to get a job 1.22 times () more than women are. Finally, misclassification rate of the model after stepwise selection is 0.4118483.

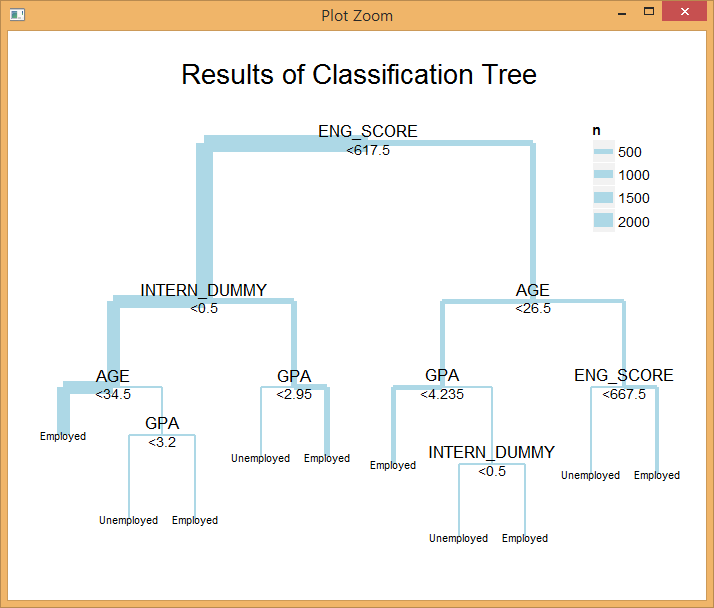
|  |  |  |
| --- | --- | --- |
| **Predictors** | **Estimate Value** | **P-value** |
| **Intercept** | 2.1075523 | 0.000000731 |
| **GENDER** | 0.1961855 | 0.020967 |
| **AGE** | -0.0606780 | 0.000401 |
| **ENG\_SCORE** | -0.0003066 | 0.005353 |
| **INTERN\_DUMMY** | 0.2732888 | 0.001912 |

**<table.2>** Results of Stepwise Selection Method

**ⅱ) Quadratic Discriminant Analysis (QDA)**

Quadratic Discriminant Analysis deducting the function classifies groups of employment or unemployment which the cases are corresponded to. In this process, misclassification rate derived from relationship between employment results and other variables are crucial to conclude appropriate research model. From the result of Quadratic Discriminant Analysis, students who are expected to get a job are 1817 of entire 5273 observations, and the rest of students are expected to be unemployed. Figure at the right show us the visualization of the prediction. As you see on the figure, GPA(x-axis) score is not affective when you get a job. However, age does matter if you are under 30. The misclassification rate of Quadratic Discriminant Analysis Model is 0.3507109.

**ⅲ) Classification Tree Analysis**

 Classification Tree Analysis is used when the predicted outcome is the class which the data belongs to. In tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. As the results of conducting algorithm based on ‘Gini’s coefficients’ in QDA, independent variables used to construct classification model are ‘ENG\_SCORE’, ‘AGE’, ‘INTERN\_DUMMY’, and ‘GPA’. The number of terminal node is 5. The result of the visualization is as right above. Interpretation the results of Classification tree analysis are very intuitive. With comparing the value corresponding to each independent variable, we proceed step by step. If the figure gets true value, go left and if it is false, go right. Misclassification rate from Classification Tree is 0.3464455.

**Ⅳ. Discussion**

Until now, we have fitted employment of graduates applying various analysis models. Looking at misclassification ratio of all three models, they get 41.1%, 35.1%, and 34.6% each. Therefore the most suitable model to predict employment is Classification Tree that has lowest misclassification rate. Judging from classification tree, English score is essential factor distinguishing whether graduates get employed. Other variables, Internship experience, age, and GPA also can be important.

From each node, the following judgement is suggested. For most graduates having low English score and having no internship experience, getting more than some degree of GPA is crucial. Especially at the 7th node, even though GPA is exceptionally high, graduates who have no internship experience are classified as the unemployed. This result means practical efforts like internship experience is more required than only academic studies.

**Ⅴ. Conclusions and Future Directions**

This paper researched job employment with different attributes of undergraduates. We used several statistical models to find appropriate prediction model, and to discover which attribute is most affective when we get a job. Our team made a conclusion that the Classification Tree model is appropriate for the research and GPA scores are meaningful when we get a job. Of course, English capacity and internship experience are important. So we suggest, graduates, who are looking for a job, should focus on not only getting high GPA but also acquiring or improving their general capability.

**Ⅵ. References**

- “Conversion Table of Seoul University TEPS Governing Committee” from TOEIC, TOEFL PBT·CBT·IBT, TEPS. (http://www.teps.or.kr/Teps/Public/conversion\_table.aspx)

- Korea Employment Information Service http://survey.keis.or.kr

- Job Korea. http://www.freedomsquare.co.kr/1561

- National Statistical Office ‘ Korea Standard Certification Occupation(KSCO, following 6th reform)’