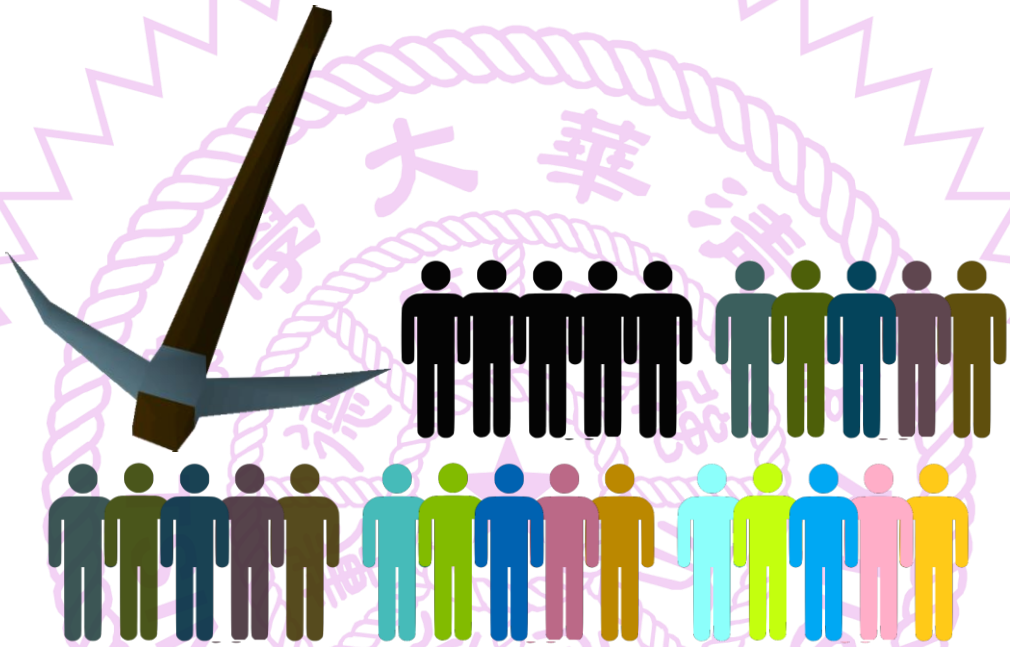
# Enhancing Efficiency of Employment by Predicting Compensation Value of Candidates



**Group5: John Liao, Jimmy Wang, Cesar Hsu, Jay Lim**

**Advisory Professor: Galit Shmueli**

# Executive Summary

## Business Problem

Companies in many industries nowadays are turning their focus to data mining and predictive-modelling technology, and the data-analytic is becoming the priceless resources for them. As a result, companies are now actively hunting for data scientists to innovate their own business. However, at the same time, the job, the data scientist, is relatively new compared to other existing occupations. Since the data science is not a standalone industry and it is normally added into other industries, companies in each industry need to consider whether the data science will bring them the advantaged or disadvantaged situation.

With the assumption that these companies already decided to implement the data science into their existing business, they will now need to hire the right person, the data scientist, who can help to achieve their goals effectively. Considering that the job, data scientist, is new position in recent years, we cannot judge data scientists' competence based on criteria of the existing job titles. Setting the appropriate criteria for hiring the right person in the field of data science is hard, and the cost of missing the right person or hiring the wrong person can be enormous.

## Data

To begin with the Kaggle website, it mainly acts as a medium between numerous competitions and participants which are all about data analysis and prediction. To find the solution to predict whether the candidate has potential to be hired, we grab roughly 16 thousand rows of survey data from the users of this website. The processes will be later described in detail. In order to exclude the bias coming from difference in country such as irregular purchasing power of units of currency, we only select the data from U.S. Accordingly, and actual data size that can be used shrinks to around 1,300 rows. The data includes data-scientist-related information along with personal information such as participants' skills, age, education status, compensation, previous job information, and current job information.

## Solution

The compensation from the survey data is the best indicator that can be used to compare among data scientists. Our model will be trained by survey data by using several data-analysis methods. Consequently, predicted value of the compensation will be based on over 70 columns which consist mainly of dimensions described above. Nonetheless, it doesn't seem to be the most ideal method to directly use the compensation as the output variable since it does not provide any comparison with existing industries. To solve the business problem mentioned above, the output variable must be binary with two outcomes which are accepted or unaccepted. Under this circumstance, compensation will be only used as the reference to be compared with the original existing industries' average compensation.

## Recommendation

- Evaluation on accuracy, cost of missing potential person and cost of finding wrong person
- Working culture in different industries and countries is not considered
- Renewing our model is necessary after a period of time
- Some components of compensation can't be evaluated in this model (such as basic salary, working KPI, bonus, and welfare)

## 1. Problem Description

a、 Background

Suppose we are a data-mining firm, Business Analysis Human Mining (BAHM), selling the models and algorithms. The main customers are human resource department of firms which are seeking data scientists to be a role to enhance their earnings power.

b、 Business goal

Improve customers' efficiency of screening out the right candidates in the first round of online assessment test, and then they can decide whether to invite those candidates to the 2nd round of interviews. The benefits to our customers is to reduce the cost of recruiting. First kind of cost is missing potential candidates, which belongs to opportunity cost, while second kind is finding wrong person which could cause sunk cost.

c、 Data mining goal

It's a supervised learning problem and we try to predict value of compensation which is our output variable to measure a new candidate is accepted or unaccepted.

d、 Assumption: High compensation means better ability in the data science field

## 2. Data Description

This dataset is from Kaggle 2017, including 16,716 observations and 228 columns. Kaggle received a lot of responses from different countries, which cause different levels of compensation based on economic level. Although it could be transformed by Purchasing Power Parity(PPP) conversion index, there is still an error. For example, someone who works in India receives compensation in USD rather than Rupee and there are no clues to confirm whether he/she works in an American company or not. Due to this reason, we focus on data scientists working in U.S.A, the sample size is 1,215 (after cleaning the dataset), and the final number of columns(predictor variables) used to build model is 79. These variables can be sorted into categories below: (1) Personal data, (2) Professional background about data mining, (3) The level of education status, (4) Working performances in the current jobs, (5) What data-mining tools they are currently using, (6) Details about the current firms where they are working.

The calculation process of average compensation of existing industries is shown below. Average compensation is weighted based on the importance of data science for each industry and it will be used as cut-off value for our customers to accept or not to accept the candidates. The average compensation in each industry is from U.S. Bureau of Labor Statistics. The result number of cutoff value is 63,869.5375. If Kaggle participant's compensation is higher than the cutoff value, he/she will be accepted in our model, and vice versa.

| Industry | count | weight | average salary(US dollar) | weighted salary |
|---|---|---|---|---|
| Financial | 329 | 0.1165 | 69,680 | 8117.8187 |
| Retail | 92 | 0.0326 | 32,120 | 1046.4023 |
| Other | 264 | 0.0935 | 57,466 | 5372.2214 |
| Technology | 586 | 0.2075 | 79,700 | 16538.3144 |
| Government | 144 | 0.0510 | 55,630 | 2836.6572 |
| Academic | 446 | 0.1579 | 54,140 | 8550.4391 |
| Internet-based | 191 | 0.0676 | 80,952 | 5475.1530 |
| Insurance | 81 | 0.0287 | 69,680 | 1998.6119 |
| CRM/Marketing | 109 | 0.0386 | 65,187 | 2516.0705 |
| Telecommunications | 96 | 0.0340 | 68,810 | 2339.1501 |
| Manufacturing | 114 | 0.0404 | 50,720 | 2047.4788 |
| Hospitality/Entertainment/Sports | 53 | 0.0188 | 36,260 | 680.5170 |
| Mix of fields | 252 | 0.0892 | 57,466 | 5128.0296 |
| Pharmaceutical | 41 | 0.0145 | 43,630 | 633.4384 |
| Military/Security | 26 | 0.0092 | 64,000 | 589.2351 |
| sum | 2824 | 1.0000 | | 63869.5375 |

Appendix 1. Cutoff Value

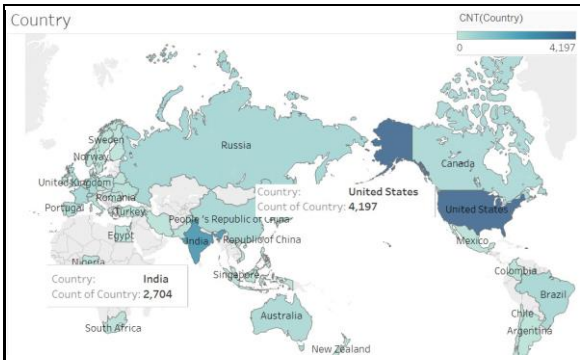
## 3. Data Preparation

Although our raw data does not contain many dimensions, our final-version data has over 200 variables. We don't have the text mining ability, so when we face the column shown in Appendix 2 and we do not want to lose any information, we cut every skill into different columns with dummies.

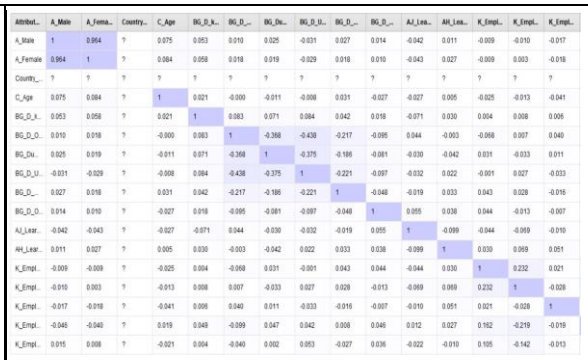
| MLTechniquesSelect |
|---|
| Evolutionary Approaches,Neural Networks - GANs,Neural Networks - RNNs |
| Bayesian Techniques,Decision Trees - Gradient Boosted Machines,Decision Trees - Random Forests,Logistic Regression,Neural Networks - CNNs,Support Vector Machines (SVMs) |
| Decision Trees - Random Forests,Ensemble Methods,Neural Networks - CNNs,Support Vector Machines (SVMs) |
| Bayesian Techniques,Decision Trees - Gradient Boosted Machines,Decision Trees - Random Forests,Ensemble Methods,Evolutionary Approaches,Gradient Boosting,Hidden Markov M |
| Bayesian Techniques,Decision Trees - Gradient Boosted Machines,Decision Trees - Random Forests,Ensemble Methods,Evolutionary Approaches,Logistic Regression,Neural Networks |
| Decision Trees - Gradient Boosted Machines,Decision Trees - Random Forests,Ensemble Methods,Gradient Boosting,Logistic Regression,Neural Networks - CNNs,Support Vector Ma |
| Bayesian Techniques,Decision Trees - Random Forests,Logistic Regression,Support Vector Machines (SVMs) |

Appendix 2. Questionnaire

We also use Tableau to find out the country which has the most responses in Appendix 3. The following is the step of dimension reduction, we have done this by using the correlation matrix in Appendix 4. We discover that most data scientists do not learn the skills from the Internet and school. As the magazine, Business Next[1], implies, the data scientist is a new job, so the school may not be prepared to cultivate this kind of people by opening relative courses. The correlation also tells us that most of them learn from work, and we get the conclusion that we need to take the weight in self-learning ability very seriously.




Appendix 3. Distribution of Data Scientist



Appendix 4. Correction Matrix

We also consult with some friends who are studying in Institute of Human Resource Management in NSYSU and NCU[2] about how we select potential people if we are a data-mining firm. Their reply is focusing on 3 dimensions, the ability, skills and study background. We don't need to consider the variable, job satisfaction, or what factors will motivate them in the workplace, which makes the whole process less complicated.

Besides the correlation matrix, we also use Gradient boosted tree shown in Appendix 5 and multiple model by regression in Appendix 6 to find out the collinearity variables and unimportant variables.

**Variable Importance**

| Variable | Relative Importance | Scaled Importance | Percentage |
|---|---|---|---|
| C_Age | 16.0213 | 1.00 | 0.22 |
| GX_RemoteWork.1 | 10.0022 | 0.62 | 0.13 |
| HA_SalaryChange | 9.9338 | 0.62 | 0.13 |
| BE_Tenure.1 | 6.6581 | 0.42 | 0.09 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| CA_WorkDatasetSize_MB | 0 | 0 | 0 |
| BL_LearningCategoryKaggle | 0 | 0 | 0 |
| GO_Visu_Work(100%) | 0 | 0 | 0 |

Appendix 5. Gradient boosted tree

*MultiModelByRegression*

MultiModelByRegression(prediction model for label GY_Compensation_USD_Dummy)

Warning:

Removed collinear columns [BE_Tenure.1.1, EC_WorkMethodsSelect_1.true, BX_WorkHarwareSelect_4.true, C_Age.range2 [44 - ∞], A_Male.true

Appendix 6. MultiModelByRegression

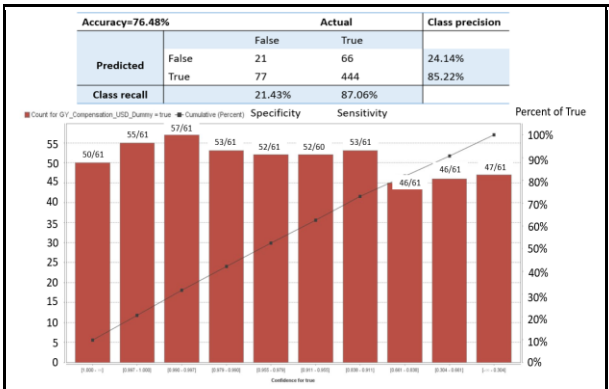
## 4. Data Mining Solution: Method and Performance Evaluation

We start to build the model of logistic regression because our outcome is whether a data scientist is accepted or not. In the beginning, as shown in Appendix 7, we get 76.48% of accuracy rate in the model. Based on the feedback from the midterm presentation, it is considered too high in the reality. Therefore, we revise our model to make sure we have the concrete proof to support out model. After comparing data scientists' compensations according to the cutoff value, the accepted data scientist is ten times as many as unaccepted ones, and therefore we decide to oversample the data to make two classes have same ratio, which is 0.5:0.5. We provide the lift chart and performance matrix before (Appendix 7) and after (Appendix 8) oversampling. In the Appendix 8, the accuracy rate is 56.91% with 56.27% sensitivity and 60.2% specificity. After using oversampling, we try ensemble by stacking

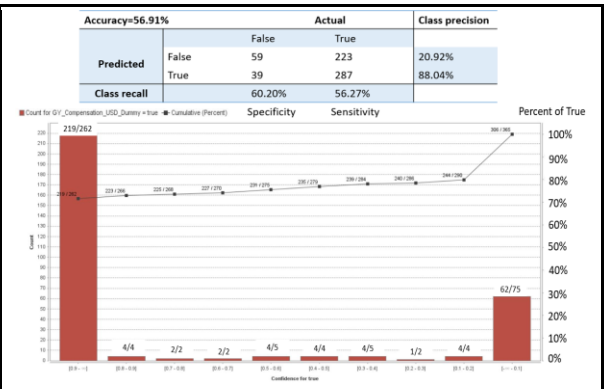
[1] 曾靈. (2017, June 07). 性感背後，資料科學家為什麼迷人？｜數位時代. Retrieved January 08, 2018, from https://www.bnext.com.tw/article/44816/why-data-scientist-is-an-attractive-job

[2] We had interviewed student Jiang from National Central and University student Chen from National Sun Yat-Sen University

the result of Naive Bayes, decision tree and Gradient boosted tree and logistic regression. The result is shown in Appendix 9, the accuracy rate is 74.79% with high sensitivity, 85.95%, and low specificity, 16.95%. In conclusion, the accuracy rate improves by nearly 18%, which increases from 56.91% to 74.79%. Simultaneously, the sensitivity rate rises substantially, while the specificity rate decreases considerably.




Appendix 7. Logistic Regression without Oversampling




Appendix 8. Logistic Regression with Oversampling

| Accuracy=74.79% | | Actual | | Class precision |
|---|---|---|---|---|
| | | False | True | |
| Predicted | False | 10 | 43 | 18.87% |
| | True | 49 | 263 | 84.29% |
| Class recall | | 16.95% | 85.95% | |
| | | Specificity | Sensitivity | |

Appendix 9. Performance Evaluation of Ensemble

## 5. Conclusions

Now, we have two different models which are ensemble model and logistic regression model, and our customers can choose either one model according to their emphasis on specificity (the rate of correctly ruling out wrong candidates) or sensitivity (the rate of correctly finding the right candidates) To be more precise, when our customers want to concentrate more on the specificity, it means that they put the emphasis on preventing hiring wrong candidates. Nonetheless, the sensitivity will drop from 85% to 56%, which deteriorates the accuracy rate of hiring the right candidates. No matter which model is deployed, there will be the trade-off in either direction.

Moreover, since the final two models only include the training data from US, once the model is used outside U.S, there will be some bias due to cultural differences which affect the predicted value of compensation. We could expand the sample of training data to other countries to make these models fit into, but it will take long time since the data scientists just appear in recent years and we lack sufficient number of participants in the Kaggle survey. In the future, we have to make the reliable survey which is suitable for each counties.

Lastly, the model has to be periodically updated due to a few reasons. Firstly, the survey needs to be kept online for additional time to boost the number of participants who fill in this Kaggle survey. This could increase the reliability of our model with enough observations. Otherwise, the size of original training data lacks reliability because of small number of observations, Furthermore, because both the cut-off value and compensation are influenced by constant effects of inflation and deflation, we have to use inflation rate to normalize compensation value and to keep updating information about compensation value in each industry that are used to calculate cut-off value.