**Selecting Candidates to Provide Optimum Business Success**

Omar Elfeky, Isabella T Oakes, and Ashutosh Singh

ADS-502 Applied Data Mining

Shiley-Marcos School of Engineering

University of San Diego

**Abstract**

One of the most important concerns for a successful business is hiring employees. The money delegated toward employees should lead to success and profit for the company. If training courses are provided to potential employees, it should ultimately provide profit to the company. Each candidate comes with a range of skills, experience, and history that can provide clues as to whether they will be a potential addition to the company. By properly assessing these factors, candidates can be chosen who are most likely to benefit the company.

**Contents**

Introduction…………………………………………………………………………………………………………………………………………..4

Data Exploration…………………………………………………………………………………………………………………………………….4

Data Validation………………………………………………………………………………………………………………………………………8

Data Mining Methods…………………………………………………………………………………………………………………………….8

Performance Metrics……………………………………………………………………………………………………………………………10

Data Mining Results……………………………………………………………………………………………………………………………..12

Reference…………………………………………………………………………………………………………………………………………….13

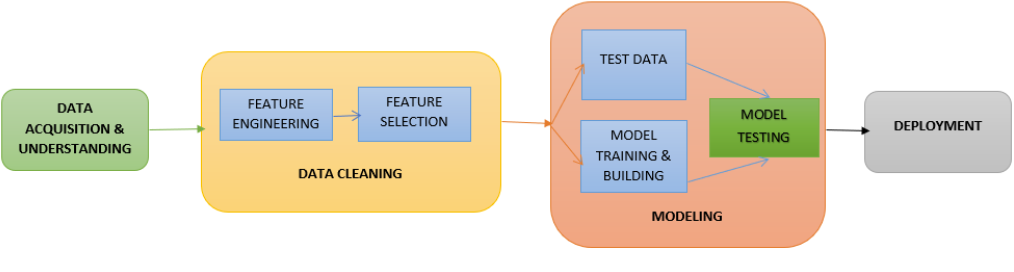
Appendix……………………………………………………………………………………………………………………………………………..14

**Selecting Candidates to Provide Optimum Business Success**

When selecting job candidates for training, it is important to choose candidates who will likely join the company, otherwise the cost of training them is lost. The data examined looked at candidates who signed up for courses conducted by a big data company. The goal is to determine which candidates that sign up for the courses will go on to work for the company, and which are likely to stay at their current jobs. By determining which candidates are most likely to work for the company, the time and money for the courses can be allocated only to candidates most likely to go on to work for the company. The process to create models to uncover this information is shown in Figure 1. Using data about the candidates, models will be created and tested to determine the optimum method to filter potential candidates and select the best ones.

**Figure 1**

*Flowchart of Information Discovery*



**Data Exploration**

To begin understanding the data, the features of the test and training data sets are generated. The training data consists of 19,158 entries of 12 variables and the result, while the test data set consists of 2,129 entries of the 12 variables. After looking at the general shape of the data, the individual variables can be assessed, which is shown in Table 1.

**Table 1**

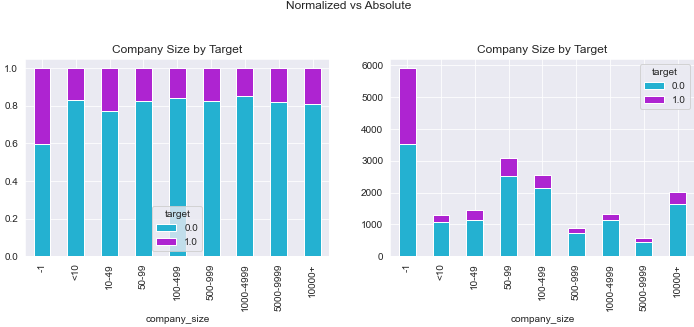
*Variable Descriptions*

|  |  |
| --- | --- |
| Variable | Description |
| City | 102 numbered city options, ie ‘city\_1’ |
| City\_development\_index | Values range from 0.448 to 0.949 |
| Gender | Female: 1,238; Male: 13,221; NA: 191 |
| Relevant\_experience | Has relevant experience: 13,792; No relevant experience: 5,366 |
| Enrolled\_university | Full time: 3,757; Part time: 1,198; No enrollment: 13,817 |
| Education\_university | Primary: 308; High school: 2,017; Graduate: 11,598; Masters: 4,361; Phd: 414 |
| Major\_discipline | Arts: 253; Business degree: 327; Humanities: 669; No major: 223; Other: 381; STEM: 14,492 |
| Experience | Begins with <1 year and goes from 1-20 in one-year increments, then 20< |
| Company\_size | Categorized by <10, 10-49, 50-99, 100-500, 500-999, 1000-4999, 5000-9999, 10000+ |
| Company\_type | Early Stage Startup: 603; Funded startup: 1001; NGO: 521; Other: 121; Public Sector: 955; Pvt Ltd: 9,817 |
| Last\_new\_job | By years: 1, 2, 3, 4, >4, never |
| Training\_hours | Values between 1 and 336 |
| Target | 0 (not looking for job change) or 1 (looking for a job change) |

After examining the variables, we also look at how many NA values exist within the data so that we can determine the best ways to handle NA values within each variable before data mining. Company size and type have the most NA values (5,938 and 6,140 respectively). Gender has 4,508 and major discipline has 2,813. Enrolled in university, education level, and last new job have similar amounts of NA values (386, 460, and 423). Experience has 65 NA values, which is the smallest amount other than the variables without NA values (city, city development index, relevant experience, and training hours). To continue processing the data, the variables are turned into factors, setting NA values to -1. To examine how the variables relate to one another correlation and a scatter matrix are used. The variables with the highest correlation values to the target are company\_size (-0.183), city\_development\_index (-0.342), enrolled\_university (0.141), and experience (-0.177). Figures 1-4 show normalized and absolute distributions of these variables with the target overlayed.

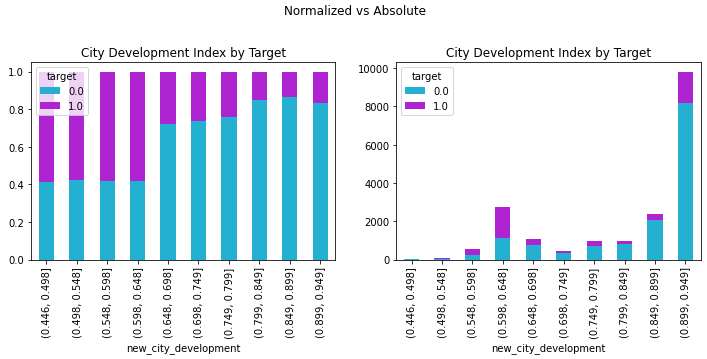
**Figure 1**

*Company Size Normalized and Absolute Distributions*



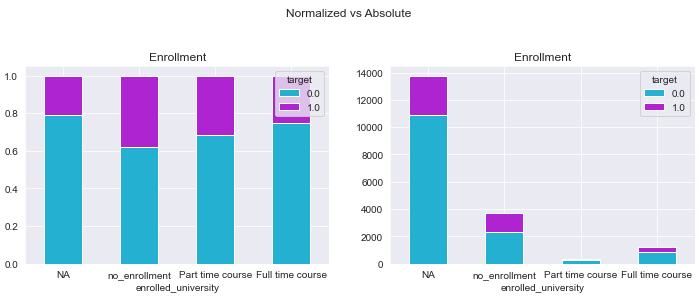
**Figure 2**

*City Development Index Normalized and Absolute Distributions*



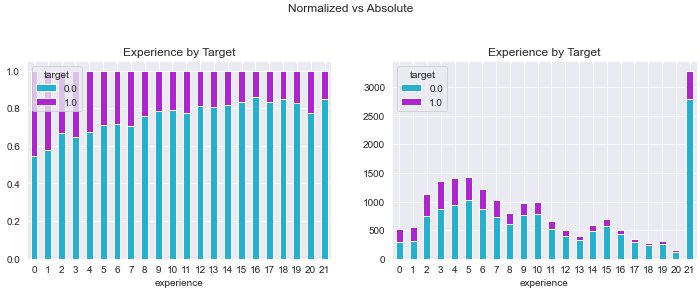
**Figure 3**

*Enrollment Normallized and Absolute Distributions*



**Figure 4**

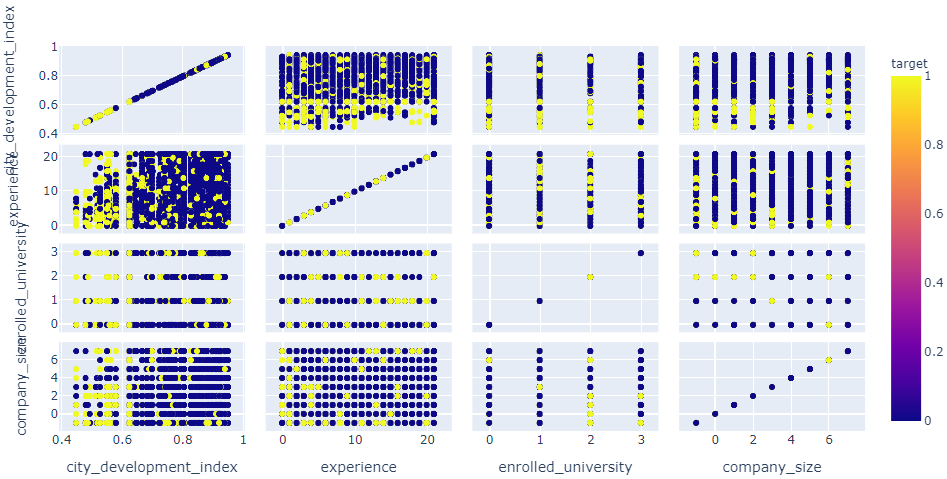
*Experience Normalized and Absolute Distributions*



A scatter matrix is generated to visualize the relationship between the independent variables, with blue dots representing 0, or will not work for the company, and yellow dots representing 1, or will work for the company, as shown in Figure 5.

**Figure 5**

*Scatter Matrix for City Development Index, Experience, University Enrollment, and Company Size*



These visualizations additional information, such as city development index being a strong factor in the target variable, including when compared to other variables.

**Data Validation**

Validating the data includes running code to describe null values to show that null values have been removed. After partitioning, the data partitions are validated, and the training data consisted of 50% target = 1 after oversampling. This code is shown in Appendix A.

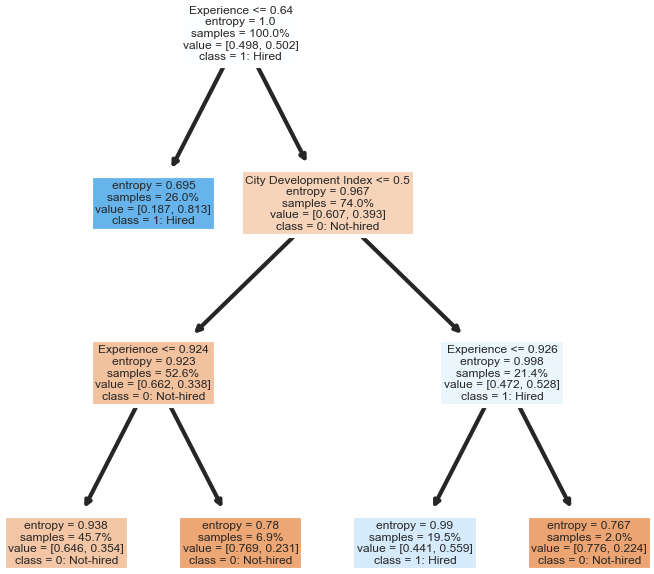
**Data Mining Methods**

To begin data mining, the training data is split into 80% training data and 20% test data, with an oversample for the target = 1. To oversample the data, it is randomly duplicated, as other methods produced results that were not significantly different. Variables were chosen based on a high correlation to the target and low correlation with each other. The variables that have the most statistical significance to create models with are ‘enrolled\_in\_university’, ‘city\_development\_index’, and ‘experience’. The first model that is created is a logistical regression model. The model generated with logistical regression is:

For the logistical regression occasion, is the target (with values less than 0.5 belonging to target group 0 and values equal to or greater than 0.5 belonging to the target group 1), is the variable ‘enrolled\_university’, is the variable ‘city\_development\_index’, and is the variable ‘experience’. The next models created are a Random Forest and a Naïve Bayes model. The fourth and final model created is a CART5.0 decision tree, shown in Figure 6.

**Figure 6**

*Cart5.0 Decision Tree of Target Using University Enrollment, City Development Index, and Experience*

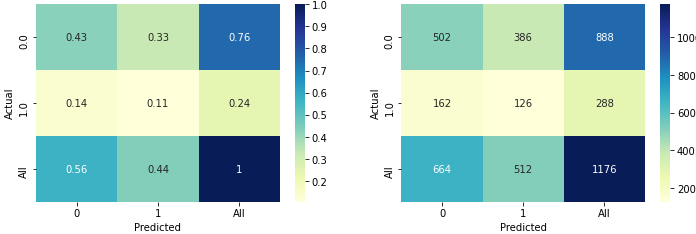


**Performance Metrics**

To determine how effective each model is at categorizing candidates as 0 (will work for another company) and 1 (will be hired), confusion matrices and ROC curve graphs are generated for each model and the accuracy, error rate, sensitivity, specificity, and precision are calculated so that model performance can be compared. The results are shown in Figures 7-10 and Table 2.

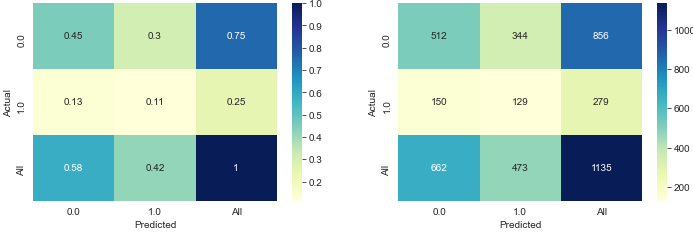
**Figure 7**

*Confusion Matrices for Logistical Regression Model (Normalized and Absolute)*



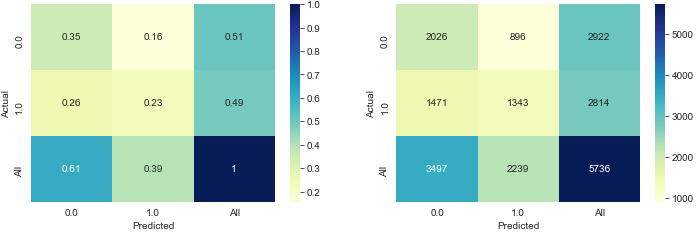
**Figure 8**

*Confusion Matrices for Random Forest Model (Normalized and Absolute)*



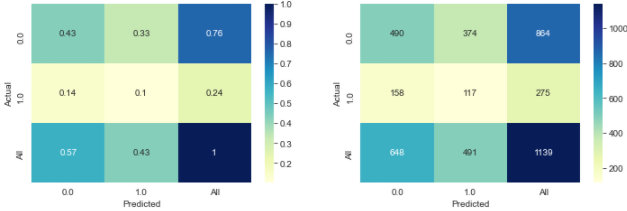
**Figure 9**

*Confusion Matrices for Naïve Bayes Model (Normalized and Absolute)*



**Figure 10**

*Confusion Matrices for Cart5.0 Model (Normalized and Absolute)*



**Table 2**

*Metrics for Data Models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Logistical Regression | Random Forest | Naïve Bayes | Cart5.0 Decision Tree |
| Accuracy | 0.534 | 0.565 | 0.587 | 0.533 |
| Error Rate | 0.466 | 0.435 | 0.413 | 0.467 |
| Sensitivity | 0.438 | 0.462 | 0.477 | 0.425 |
| Specificity | 0.565 | 0.598 | 0.693 | 0.567 |
| Precision | 0.246 | 0.273 | 0.600 | 0.238 |
| Area under ROC Curve | 0.67 | 0.73 | 0.58 | 0.58 |

There is not a huge range between the four models. The largest difference is in the precision value for Naïve Bayes, which is significantly higher than the other models. The accuracy of the models ranges from 53.4% correct predictions to 58.7% correct predictions. Logistical regression performed lowest among the models, followed by the Cart5.0 Decision Tree. Random Forest performed slightly better, with Naïve Bayes performing highest by a decent margin in all categories. The area under the ROC curve was higher for the logistical regression and random forest models, while Naïve Bayes and Cart5.0 Decision Tree models had the same value.

**Data Mining Results**

While the models performed in a similar range to each other, there are many differences that would influence which model would be chosen. If it is financially expensive to provide the training to candidates who will not be hired, a model with the lowest false positive rate, meaning highest sensitivity, would be ideal so that training was not provided when the candidate would not choose to work for the company. The Naïve Bayes model had the highest sensitivity at 0.477, so would be the ideal choice. The precision is also much higher than it is for the other models. Overall, Naïve Bayes performed best among all models except the ROC curve. Random Forest had the best ROC curve results and was second to Naïve Bayes in all other metrics. When it comes to choosing a model, the Naïve Bayes would be the best choice to ensure that the number of false positives was limited and the cost of conducting training was not wasted.

**Reference**

HR Analytics: Job Change of Data Scientists. (2021, January). https://www.kaggle.com/arashnic/hr-

analytics-job-change-of-data-scientists

**Appendix A**

