# Analyzing HR Analytics Data - Job Change of Data Scientists

1. **Introduction**

The Kaggle data set titled: *HR Analytics - Job Change of Data Scientists* was analyzed using various statistical methods and machine learning techniques. Essentially, a company (whose name isn’t disclosed) wants to hire data scientists from a population of candidates that have taken training courses. These training courses were conducted by the company itself, and the company wants to know which candidates are interested in working for the company and which are seeking new employment after the training process is concluded. The aim of doing this is to help reduce training costs and time, specifically as it relates to the quality of training, the planning of courses, and the categorization of candidates. Evaluation of this data set can also help provide the Human Resource department with essential information regarding the factors that lead candidates to leave current job positions. This analysis was performed to reach conclusions with the following research questions in mind:

**Research Questions**

1. What is the probability that a candidate for a data science job position will work for the company or look for a new job after having taken company training courses?
2. How can the affected factors on candidate decision making be interpreted?
3. What kind of specific practical improvements can be made to the training program based on this analysis? (i.e., should the company cut program costs, look to recruit based on certain characteristics, etc.)

**Data Description**

The entire data set is composed of two data files: (1) a train data set with 19,158 observations and 14 variables (2) a test data set with 2,129 observations and 13 variables. Since the outcome variable is only in the train data, this analysis utilizes the train data set as the “full” data set and the test data is disregarded. Thus, the full data for this analysis has a sample size of 19,158 observations and 14 variables. The predictors for this data deal with important information regarding candidate credentials, demographics, and work and education experience. Ten predictors are categorical, two are continuous numeric, there is an “ID” variable, and the outcome variable is binary and indicates if a candidate is looking for a job change or not. The following [link](https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists) can be used to access the data set, and the variables as well as their descriptions can also be seen below.

**Variables**

1. Enrollee ID: Unique candidate ID number
2. City: City code identification (123 unique cities)
3. City Development Index: Scaled development index which is a measure of the development level of the city (values range from 0.448 to 0.949)
4. Gender: Candidate gender (Male, Female, or Other)
5. Relevant Experience: Indicates if a candidate has relevant work experience or not (0 = Candidate has no relevant work experience, 1 = Candidate has relevant work experience)
6. University Enrollment: Type of University course enrollment (Full time, Part time, or No Enrollment)
7. Education Level: Candidate education level (Graduate, High School, Masters, PhD, or Primary School)
8. Major Discipline: Candidate education major (STEM, Business, Humanities, Arts, Other, or No Major)
9. Experience: Candidate total work experience in years (categorized in groups between less than 1 and greater than 20)
10. Company Size: Number of employees in current employer's company (categorized in groups between less than 10 and ≥ 10,000 employees)
11. Company Type: Current employer's company type (Early-Stage Startup, Funded Startup, NGO, Public Sector, Private Ltd, or Other)
12. Last New Job: Difference between previous job and current job in years (categorized in groups between 1 and greater than 4 years, or never)
13. Training Hours: Number of completed training hours (values range from 1 to 336 hours)
14. Target: Indicates if candidate is looking for a job change or not after havinf taken training courses (0 = Not looking for job change, 1 = Looking for job change)
15. **Statistical Methods**

**Data Preprocessing**

The variables were closely inspected and further evaluated, including the outcome variable, and it was concluded that three variables dealing with candidate ID, candidate city code ID, and candidate University major did not need to be considered for this data. Each candidate’s unique ID is a randomized number where no conclusions with respect to the outcome variable can be made, and candidate city code ID also represents unique numbers that tell a similar story[[1]](#footnote-1). Candidate University major was also removed since it’s a near-zero variance variable where most candidates are labeled as *STEM* majors, and for interpretability purposes, city development index was rescaled to represent a value out of 100 as well. Next, it was discovered that there were variables with missing values. Missing values for candidate gender were incorporated in the “Other” category and candidates with missing values for University enrollment were considered to not be enrolled. Mode imputation was then utilized on 5 other variables to deal with missing values. [Table 1](#_Table_1_1) below demonstrates variables with corresponding missing value amounts.

# Table 1

Table 1 – Variables with missing values

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | Gender | University Enrollment | Education Level | Experience | Company Size | Company Type | Last New Job |
| **NA** | 4,508 | 386 | 460 | 65 | 5,938 | 6,140 | 423 |

There were a few variables that had many different categories, or levels, and needed to be condensed as well. For example, candidate major and company type had 6 different groups, candidate major and education level had 5, and candidate experience had over 20 different groups. A total of five variables were releveled, and the original and new levels for these variables can be seen in [Table 2](#_Table_2) below. Under the “New Levels” column, whatever’s underlined represents the new level name for a given variable, and whatever’s in parenthesis is the original level name.

# Table 2

Table 2 – Demonstrates the original and releveled levels for 5 variables

|  |  |  |
| --- | --- | --- |
| **Variable** | **Original Levels** | **New Levels** |
| Education Level | Graduate, High School, Masters, PhD, Primary School | University (Graduate, Masters, PhD), Other (High School, Primary School) |
| Experience | <1, 1-20 (1, 2, etc.), >20 | Some Experience (<1,1-4), Experience (5-14), A lot of Experience (15-20, >20) |
| Company Size | <10, 10-49, 50-99, 100-500, 500-999, 1000-4999, 5000-9999, 10000+ | Small (<10, 10-49, 50-99, 100-500), Medium (500-999, 1000-4999), Large (5000-9999, 10000+) |
| Company Type | Early-Stage Startup, Funded Startup, NGO, Public Sector, Pvt Ltd, Other | Public (Public Sector), Startup (Early-Stage Startup, Funded Startup), Other (NGO, Pvt Ltd, Other) |
| Last New Job | Never, 1-4, >4 | Never (Never), Recent (1, 2), Kind of Recent (3, 4), Not Recent (>4) |

Moreover, there were four distinct predictors with significant class imbalance issues for this data, and so the sampling method of SMOTE was used to address this issue. The predictors with class imbalance issues can be seen in [Figure 1](#_Figure_1) below, which demonstrates bar plots with counts on the y-axis and the corresponding variable levels, or groups, on the x-axis. These bar plots also represent the data before it was manipulated (i.e., before any data reduction, releveling, imputation, or other data manipulation techniques were utilized), and the actual process of SMOTE was used after splitting the data into train, validation, and test sets (this will be discussed in greater detail in the analysis portion of this section).

# Figure 1

Figure 1 – Bar plots of variables with class imbalance issues

Graphical user interface, application

Description automatically generated

As we can see from [Figure 1](#_Figure_1) above, the bar plot of gender shows that there are significantly more candidates who are male than candidates of any other gender, the bar plot of relevant experience demonstrates how there are more candidates with relevant work experience than there aren’t, the bar plot of education level shows that most candidates have an undergraduate degree as opposed to any other degree type or education level, and lastly, the bar plot for University enrollment indicates that most candidates aren’t enrolled in full or part-time academic courses.

**Analysis**

The data set had a sample size of 19,158 observations and 11 variables (10 predictors and the outcome variable) before being split into 60% train, 25% validation, and 15% test sets. After splitting the data, there were 11,555 observations in the train set, 4,800 in the validation, and 2,803 in the test, and after incorporating SMOTE, there were 20,440 observations in the train set, 8,176 in the validation, and 4,823 in the test. It’s important to note that the SMOTE procedure increases data sample sizes, which is what occurred here. [Figure 2](#_Figure_2) below shows how SMOTE was able to remedy class imbalance issues since the level counts are closer together in value for the outcome variable when SMOTE is used compared to when it isn’t. Also, only model comparison results where SMOTE is utilized, and not when it isn’t, will be displayed for this analysis to prevent cluster and conserve space. This was a subjective decision.

# Figure 2

Figure 2 – Bar plots of outcome variable counts before and after SMOTE

Graphical user interface

Description automatically generated

Repeated 10-fold cross validation with 5 repeats was used for each model fitting process and each model was fit on the train data (20,440 observations). A validation set was utilized for this analysis so that optimal cut points can be identified and employed for predictive use on the test data, and models were evaluated and compared for the test data only. Model comparison was based on numerous evaluation criteria, such as accuracy, kappa statistic, AUC, brier score, sensitivity, and specificity measures, and a total of five models were considered for analysis: (1) Logistic Regression (2) Stochastic Gradient Boosting (XGBoost) (3) Penalized Logistic Regression (4) Random Forest (5) Support Vector Machine.

1. **Results**

[Table 3](#_Table_3) below is a model comparison table for logistic regression, stochastic gradient boosting, penalized logistic regression, random forest, and support vector machine models. Models were compared and analyzed based on the previously mentioned evaluation criteria (i.e., accuracy, kappa statistic, AUC, brier score, sensitivity, and specificity values). The optimal cut points for the models can be seen in the table as well.

# Table 3

Table 3 – Model comparison table for competing models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Kappa** | **AUC** | **Brier** | **Sens.** | **Spec.** | **Cutpoint** |
| LR | 71.55% | 0.4304 | 0.7802 | 0.1892 | 0.7470 | 0.6919 | 0.380 |
| Boosting | 81.9% | 0.6236 | 0.8936 | 0.1254 | 0.7160 | 0.8962 | 0.5523 |
| PLR | 71.59% | 0.4383 | 0.7804 | 0.1893 | 0.7983 | 0.6542 | 0.3493 |
| RF | 73.56% | 0.4692 | 0.8027 | 0.1812 | 0.7586 | 0.7184 | 0.4430 |
| SVM | 74.64% | 0.4788 | 0.7927 | 0.1802 | 0.6773 | 0.7983 | 0.4237 |

The stochastic gradient boosting model (highlighted above) performed the best amongst all five competing models based on the model evaluation criteria (except for sensitivity). Compared to the other models, the boosting model had the highest accuracy, best kappa statistic value, biggest AUC, lowest brier score, and highest specificity (i.e., it did the best between the models in terms of predicting candidates that weren’t looking for a job change). Also, although the boosting model had rather low sensitivity compared to other models, its sensitivity and all other model sensitivities improved significantly after implementing SMOTE, and evaluation criteria measures generally improved for all models with SMOTE as well. [Table 4](#_Table_4) below shows the optimal tuning parameters for the boosting model.

# Table 4

Table 4 – Boosting optimal tuning parameters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| nrounds | max-depth | eta | gamma | colsample-bytree | min-child-weight | subsample |
| 150 | 3 | 0.4 | 0 | 0.8 | 1 | 1 |

The calibration, ROC, and variable importance plots for the boosting model are demonstrated below in [Figure 3](#_Figure_3), [Figure 4](#_Figure_4), and [Figure 5](#_Figure_5), respectively. The calibration plot shows that the probabilities are very well-calibrated since the actual observed event rates match quite well with the binned midpoints, and the ROC plot demonstrates a very large AUC as well. The variable importance plot indicates that city development index, training hours, and no University enrollment are the most important features when predicting if a candidate is looking for a job change or not.

# Figure 3

Figure 3 – Calibration plot for the stochastic gradient boosting model

Chart, line chart

Description automatically generated

# Figure 4

Figure 4 – ROC plot for the stochastic gradient boosting model, with demonstrated AUC value

Chart, line chart

Description automatically generated

**AUC = 0.8936**

# Figure 5

Figure 5 – Variable importance plot for the stochastic gradient boosting model

Chart

Description automatically generated

[Table 5](#_Table_5) below shows the three most “important” features selected from boosting as well as their corresponding importance values. Only the three most “important” are reported since the “elbow” on the variable importance plot also demonstrates decreasing importance differences between features with less overall importance, (i.e., features after the “no University enrollment” feature in [Figure 5](#_Figure_5)). The table below shows that city development index plays a significant role when predicting the odds that a candidate’s looking for a job change.

# Table 5

Table 5 – Most important features selected from boosting

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| City Development Index | 0.5085 |
| Training Hours | 0.1366 |
| Not Enrolled in University | 0.0779 |

For the other four models, penalized logistic regression performed slightly better than logistic regression when comparing model accuracy, kappa statistic, AUC, and sensitivity values, however logistic regression performed slightly better with respect to brier score and specificity. Penalized logistic regression also had the highest sensitivity compared to all other competing models as it was able to better distinguish between candidates looking for a job change. RF and SVM performed quite similarly with respect to various criteria measures (i.e., accuracy, kappa statistic, AUC, and brier score measures) and both models generally performed better than logistic regression and penalized logistic regression. SVM performed better than RF with respect to accuracy, kappa statistic, brier score, and specificity measures, but RF had a larger AUC and better sensitivity.

Moreover, coefficient estimates for penalized logistic regression were further analyzed to gain better insight for interpreting fixed parameter effects. It’s pivotal to comprehend the extent of which predictors are affecting the outcome variable to draw more intuitive and practical conclusions. Also, the optimal tuning parameters for the PLR model were and , so it essentially utilized ridge regression since alpha is 0 and the small lambda value indicates that the coefficient estimates weren’t shrunk by much (i.e., variance didn’t increase, and bias wasn’t introduced). [Table 6](#_Table_6) below shows the significant coefficient estimates for penalized logistic regression.

# Table 6

Table 6 – Penalized logistic regression significant coefficient estimates

|  |  |
| --- | --- |
| **Name** | **Coefficient Estimate** |
| City Development Index | -0.6202 |
| Gender (Male) | -0.3494 |
| Gender (Other) | -0.1102 |
| No relevant experience | 0.2501 |
| No University enrollment | -0.3338 |
| Part-time University enrollment | -0.0936 |
| Obtained University degree | -0.1550 |
| 5-14 years of total work experience | -0.1263 |
| >15 years of total work experience | -0.1375 |
| Medium-sized company | 0.0896 |
| Large company | 0.1743 |
| Startup company | -0.2331 |
| ‘Other’ company | -0.4329 |
| Recent last new job | -0.0585 |
| Kind of recent last new job | 0.0458 |
| Training hours | -0.0868 |

From [Table 6](#_Table_6) above, the coefficient estimate for city development index shows that a unit increase in the candidate city development index will result in the biggest average decrease of the odds that a candidate is looking for a job change (holding all other predictors fixed). More specifically, when holding all other predictors fixed, a one unit increase in the city development index multiplies the odds that a candidate’s looking for a job change by , which means that it decreases the odds by an average of about 100% - 53.78% = 46.22%. Other factors seemed to play pivotal roles for decreasing the average odds as well. For example, holding all other predictors fixed, working for an “Other” company (i.e., an "NGO," "Pvt Ltd,” or "Other” company, and not a “Public” or “Startup” company) multiplies the odds of looking for a new job by , which means that it decreases the odds by 35.14% on average. Factors that seemed to have the largest effect on increasing the average odds of a candidate looking for a job change were (1) not having any relevant work experience[[2]](#footnote-2) (2) working for a large company[[3]](#footnote-3).

Lastly, SMOTE seemed to play a critical role in solving class imbalance issues since the data wasn’t highly dimensional, and another model comparison table of the results without the usage of SMOTE should have been included for this analysis to help demonstrate its impact. The process of SMOTE involves increasing class overlaps, and therefore introducing noise if the data were to be highly dimensional, which didn’t occur for this analysis.

1. **Conclusions**

There are many practical conclusions to be drawn with respect to the company’s goals and research questions in mind. Boosting performed the best between the other competing models for this analysis, and it found city development index, training hours, and not being enrolled in a University to be the most significant, or “important,” factors for predicting candidate odds of looking for a job change. City development index, working for a large or “other” company, and not being enrolled in a University were found to be significant factors when analyzing and interpreting fixed parameter effects for penalized logistic regression as well. Candidates that are from cities with higher city development indexes, aren’t enrolled in a University, or completed more training hours are more likely to, on average, not be looking for a job change. Meanwhile, the data also shows that candidates that work for large companies or have no relevant work experience in data science are more likely to look for a job change, on average.

Moreover, it’s not essential for the company to cut costs but it should potentially look to reallocate recruiting funds and improve its candidate selection strategies. An example of a practical route that the company could take is to look to recruit more candidates that have lower average odds of looking for a job change (i.e., candidates from cities with higher city development indexes, aren’t enrolled in a University, and/or work for smaller to midsize companies). The company should avoid recruiting candidates that are from cities with low city development indexes, work for large startup or public companies, and/or don’t have relevant work experience. In terms of potential candidate driving factors, one can speculate that candidates from cities with higher development indexes may have better exposure to data science work and employment, so they may be more likely to stay and work for the company. Also, candidates enrolled in universities may be too busy with school and other obligations to work for the company, while candidates with no relevant work experience would most likely look for other jobs that better align with their career interests.

There are a few important limitations to my analysis. First, it would have been essential to include an analysis both with and without the usage of SMOTE for comparison purposes, instead of just including results with SMOTE. Incorporating a second model comparison table showing results of another set of comparison values would help illustrate a more inciteful assessment of SMOTE’s impact. Not including this in my analysis was a completely subjective decision that can technically impact results. The data for this analysis without SMOTE would have comprised of 11,555 train observations, 4,800 validation observations, and 2,803 test observations (i.e., how the data was represented right after the train-validation-test split). Also, another limitation was the inability to utilize the test data given by the Kaggle competition to submit predictions and receive a Kaggle score. The competition was closed, and scores couldn’t be submitted. Variable interpretability was also ambiguous in some instances, but this analysis does an adequate job interpreting variable meaning and parameter effects. For example, replacing city code identification with an actual city, such as replacing it with “Dallas” for example, would have helped make for more intuitive results in terms of how location can affect candidate likelihood of looking for a job change.

1. There are 123 unique candidate cities that are represented numerically (i.e., city\_1, city\_2, etc.) and do not identify with a potential city of interest (i.e., Dallas, New York, Chicago, etc.). [↑](#footnote-ref-1)
2. Holding all other predictors fixed, having no relevant experience multiplies the odds of looking for a job change by , which means that it increases the odds by about 28.41% on average [↑](#footnote-ref-2)
3. Holding all other predictors fixed, working for a “large” company multiplies the odds of looking for a job change by , which means that it increases the odds by 19.04% on average [↑](#footnote-ref-3)