# ***Project 3 – Prediction of Future Employee Resignations***

*Samuel Coyne, Yao Yao, and Joseph Denney*

## Problem Statement

Why do employees quit? Perhaps the adage suggesting those who voluntarily leave a company are not leaving a specific job, but rather a boss holds merit. According to the researchers from the Harvard Business School, “employee turnover is associated with decreased performance, as measured by profit margin and customer service.” (1) IBM states, “HR analytics enable organizations to use their wealth of employee data to make better decisions about their workforces and improve operational performance.”(2) There is a clear role in predictive analytics to employ statistical processes in an attempt to identify employees who might be inclined to leave. Management would be able to intervene prior to the departure of high-risk employees and reduce the cost associated with turnover and increase employee satisfaction. Identification of factors leading to the departure of employees can also engender organizational change to address issues highly correlated with departure and improve employee satisfaction. It is a maxim of management that your most valuable customers are your employees. In fact, Sir Richard Branson in Inc. magazine stated, “If the person who works at your company is 100% proud of the job they're doing, if you give them the tools to do a good job, they're proud of the brand, if they were looked after, if they're treated well, then they're gonna be smiling, they're gonna be happy and therefore the customer will have a nice experience.”(3) Our research question of interest will be to predict if an employee will leave the company. We will also identify the variable(s) with the greatest predictive ability related to leaving with an eye toward recommending changes to reduce the likelihood of employee resignations.

## Data Set Description

In this paper we will be using a dataset of simulated data from Kaggle.com entitled “Human Resource Analytics, Why are our best and most experienced employees leaving prematurely?”(4)

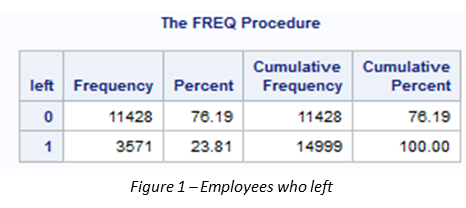
The raw data set consists of the following variables:

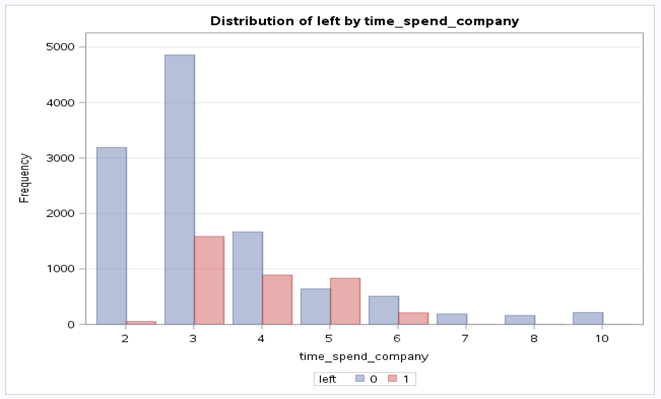
1. satisfaction level - score of an employee’s satisfaction with his or her job, range is between 0 and 1 – 1 represents the highest satisfaction
2. last evaluation – the performance score of an employee given in his or her last performance review
3. number of projects – the number of projects each employee contributed to
4. average monthly hours – the average number of hours spent working in a month
5. time spend company – the number of years an employee has been with the company
6. work accident – This is a binary explanatory variable in which 1 = yes, a work accident occurred or 0 = no, a work accident did not occur
7. left – This is the binary dependent (response variable) in which 1 = yes, the employee left the company or 0 = no, the employee did not leave the company.
8. promotion last five years – This is a binary explanatory variable in which 1 = yes, a promotion has been given in the last five years or 0 = no, a promotion has not been given in the last five years
9. job function – department to which each employee reports. Please note, in the raw data, this variable is mistakenly entered as ‘sales.’ Please reference the SAS appendix import statement for the renaming of ‘sales’ to ‘job function’
10. salary – this is a categorical explanatory variable corresponding to an employee’s salary level (low🡪medium🡪high)

The collected data can be thought of as falling into either the continuous category (i.e. satisfaction level) or discrete/categorical (i.e. work accident). To begin, we will take a high-level view of our variables to understand some general patterns. A more in-depth look at the variables will follow to gain an understanding of the variables by their decision to stay or leave the company.

## Exploratory Data Analysis

Our initial step is to be sure there are no missing observations that will need to be accounted for. None were found; the data is tidy, and we have no concerns to address before we begin. Next, we start to explore the data by determining the actual number of employees who have left. Figure 1, below, shows approximately 24% of employees left during the sample period.



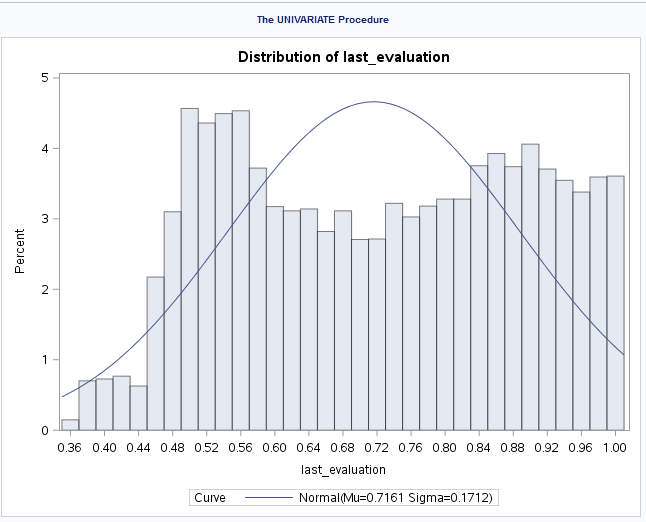
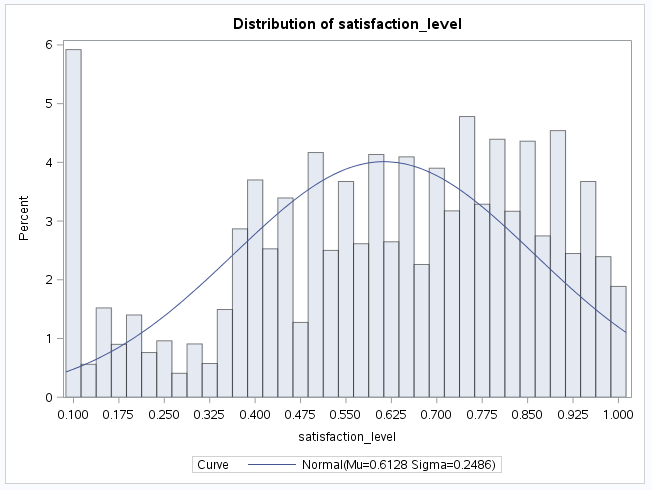
Other items of note discovered are the number of projects worked on is normally distributed, with most of the values in the range of 3 to 5 projects. The number of years spent with the company is right skewed, showing a short tenured work force with the mode being three years. As seen in Figure 2, something seems to be happening at the 3 to 5-year mark, as these are the periods of greatest percentage departures. Once an employee gets to 6 or more years, the departure rates fall to almost zero. Since long term employees tend to say with the company, answering our original question of why employees leave becomes more interesting.

*Figure 2 – Distribution of length of employment*

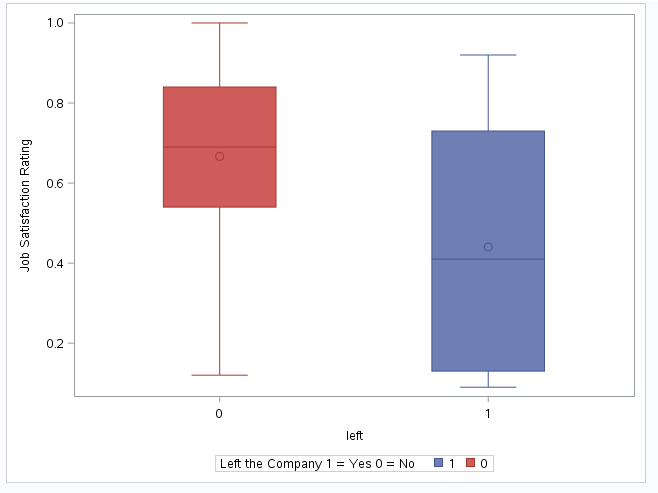
The number of accidents records is small, with about 14% of the employees having a reported accident. The number of promotions is small, with only 2% of employees receiving a promotion. Job functions are primarily concentrated in sales (27.6%), support (14.9%), and technical (18.1%), for a total of about 60% of the jobs concentrated in 3 of the 10 categories. The salary variable has 3 levels, with most employees falling into the low and medium salary categories.

The continuous variables (job satisfaction, evaluation, and average monthly hours) show some definite patterns. Job satisfaction is left skewed with a high number of employees reporting low satisfaction, then an increase in satisfaction levels afterwards. This is a potential area we are going to want to examine further, given the high number of low satisfaction scores seen. See Figure 3.

The evaluation variable is also skewed, although not as much as satisfaction. Evaluation does show some bimodal distribution, as seen below in Figure 4. These departures from normality cause no issues with performing logistic regression, but do give us more insight into the data.

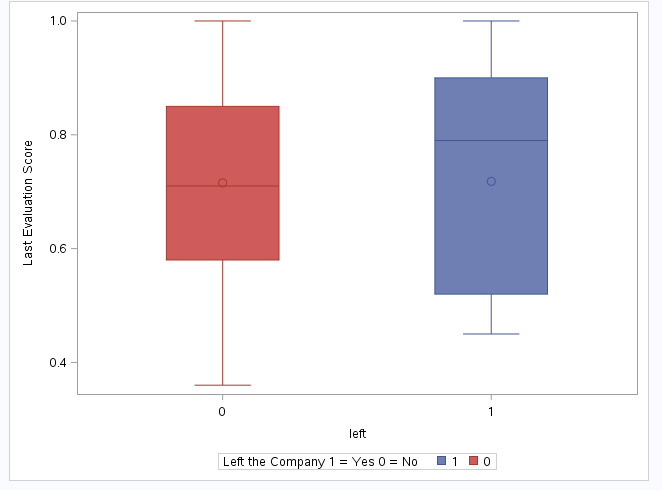


*Figure 3 – Satisfaction level scores Figure 4 – Bimodal distribution of last\_evaluation*

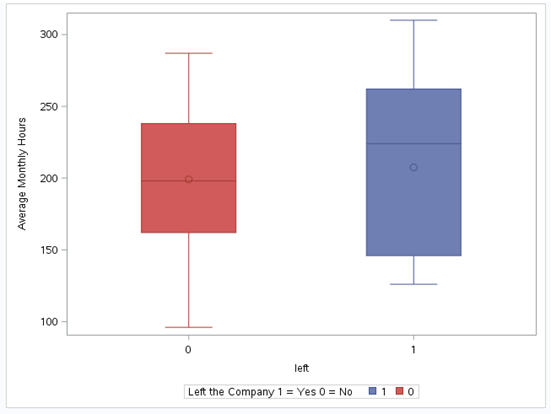
The next step in our exploratory analysis is to look for patterns and relationships between our predictor and response variable. There seems to be a relationship between the number of projects and leaving the company. Those employees who had 3, 4, or 5 projects left the company at much lower rates than those with less than 2 or more than 5 projects.

*Figure 5 – Box plot of job satisfaction vs left or stayed*

For the final part of the exploratory data analysis, we will look at the continuous explanatory variables by leave/stay status. First, Figure 5 gives us some interesting takeaways: of those who did leave, there were no perfect scores of satisfaction. Second, a comparison of both the means and medians indicates that those who did stay with the company had their job satisfaction score is higher.

The next variable, last evaluation score, does reveal an unexpected result: of those who left the company, both the mean and median evaluation scores were higher than those who stayed. This finding begs the question “why are the best employees leaving?”

*Figure 6 -Box plot of evaluation vs left or stayed*

The final variable in our exploratory data analysis is a measure of time spent working. The plot suggests that of the employees leaving, longer hours could be playing a role. We see a higher mean and median of hours worked for those who did leave.

*Figure 7 – Box Plot of Hours vs left or stayed*

To conclude the exploratory data analysis, the data paints a picture of a company with some unexpected correlations in the data. We see experienced employees with long tenure are not common. More concerning, the employees with a higher evaluation score are leaving. If those who scored average or below average on their last evaluation stay with the company, the enterprise cannot expect to excel at a faster rate than their competition.

## Assumptions

The analysis method we will use for this study is logistic regression. The assumptions of logistic regression are not as rigid as ordinary least squares (OLS) regression. The assumptions to be met are as follows: the response variable must be binary, independence, and linearity between continuous variables and the logit of the dependent variable. A further check we will address here is multicollinearity.

First, this study does have a binary outcome as the dependent variable: employees either left the company or they did not (1=left the company 0 = didn’t leave the company). Second, nothing in the study indicates the observations are not independent. Third, the linearity assumption can be assumed to be met at this point. We will revisit this assumption once we discuss the goodness of fit test. To address this assumption, we turn our attention to the Hosmer and Lemmeshow Goodness of Fit test with the following hypotheses:

: the model indicates a good fit

: the model does not indicate a good fit.

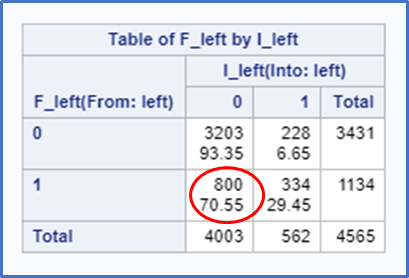
The p-value found in our selected model, as well as all models evaluated, is <.0001. This finding suggests us to reject the null hypothesis. On the surface, this would appear to indicate the model does not meet the assumptions of logistic regression. However, this goodness of fit test has one critical shortcoming: sensitivity to large data sets. Our master data set is made up of nearly 15,000 observations. Because our data set is so large, this test is oversensitive. The size of the data set means we will have to examine other indicators of model fit. A strong test performance indicates our final model is robust and modeled correctly with logistic regression.

## Logistic Regression Analysis

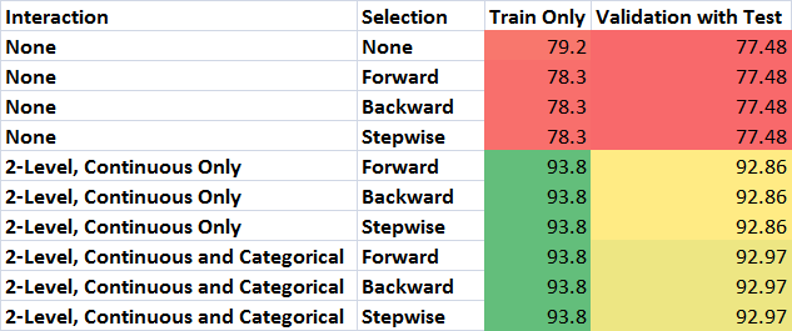
With the assumptions met and armed with the insight gained from the exploratory data analysis, we proceeded to the model development phase. While this data set is not a comprehensive picture of all data points in the human resources spectrum, the data set offers a wealth of variables that are associated with an employee’s decision to stay or leave a company. It is because of this wealth we hypothesized the model with the most predictive power would result from applying various variable selection techniques.

***Model Selection***

First, from our master data, we chose to employ a 70% training/30% testing data partition. The assignment of observations to train/test was decided by a random number generator (see SAS code lines 99-105). The model building process began with the development of a preliminary model, followed by interaction terms between continuous only as well as continuous and categorical. An overview of model results split by training and test data partitions can be seen below in Figure 8.

For the preliminary model, we first ran the logistic regression on the whole data set with no interaction and no selection and the prediction accuracy was 79.2% for the training set and 77.48% with test. The data set was already a good indicator above a 50-50 chance for accurately predicting whether an employee leaves a company. However promising the aggregate results are, the false negatives leave much to be desired. We see that of the 1,134 individuals who left the company in the test data set, 800 were misclassified.

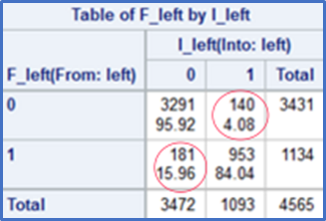
*Figure 8 – Preliminary Model Confusion Matrix*

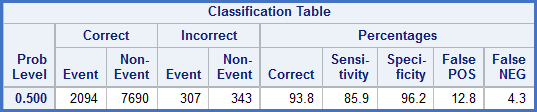
We were quite concerned about the high number of misclassifications in the preliminary model. A combination of this concern and to increase the predictive power of the model, we ran a number of auto selections with and without various interactions to try to improve the model, including forward, backward, and stepwise with no interaction terms. This resulted in no significant increase or change in prediction accuracy from the preliminary model. For 2-level interaction auto-selection, where only the continuous variables are interacting, the prediction accuracy jumped to 93.8% for the training set only and 92.86% for the test data. For 2-level interaction auto-selection, where the continuous variables are interacting with categories that are 0 or 1, the prediction accuracy stayed the same at 93.8% for the training set only and 92.97% for the test data. Because all the selection method came up with the same model per interaction method, we chose to summarize with the model having the highest predictive score, which is 2-level interaction with continuous and categorical variables using the stepwise selection method. See Figure 9 for details. Figure B in the appendix shows the various statistical values associated with each variable in the model.

*Figure 9 – Tabular view of the various considered models*

Our selected model has the following predictive characteristics:

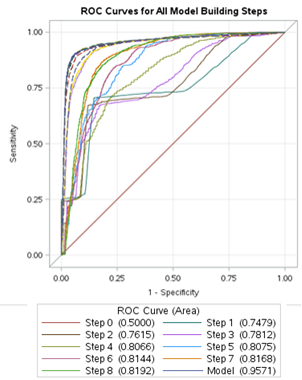
* the prediction accuracy of the training set (n=10,434) is 93.8%.
* The test data set (n=4,565) has a prediction accuracy of 92.97%, as found by the total number of correct classifications (3291+953) divided by the total in the test data set (4,565).
* There were 140 false positives found in the test performance and 181 false negatives.

Figures 10 and 11 below show the classification table and confusion matrix for our final model.

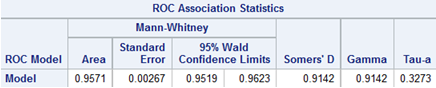
 *Figure 10 -Classification Table Final Model*

*Figure 11 – Confusion Matrix Final Model*

To conclude the final model selection, we visualized the predictive power of the model by generating ROC curves. The ROC curves for each step of the stepwise selection are based on the final model. As interaction terms for both categorical and continuous variables are implemented, the full 21-step model increased to an area under the curve of 0.9571 with a standard error of 0.00267 for the training set only. This visualization reinforced what we saw with the other measures of predictive power.

* Figure 12 – ROC Curve Final Model*

*Figure 13 – ROC Statistics*



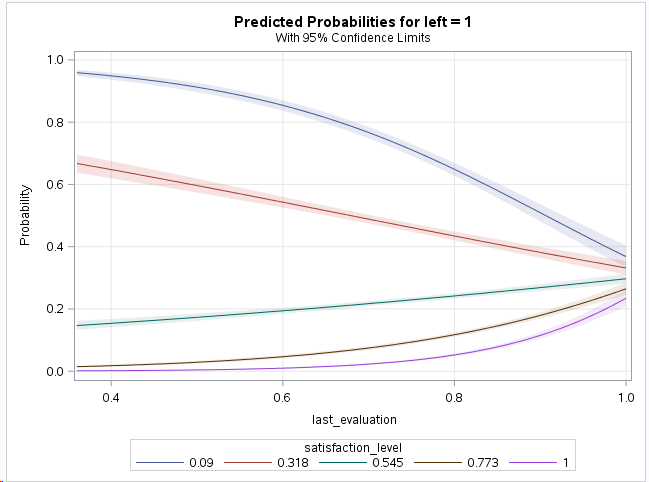
***Limitations of the Model –***

As with any model, simplifying a complicated reality leads to some shortcomings that need further explanation. This model does have several variables that are highly correlated. “Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors.” (5) The source of the multicollinearity can be traced back to the inclusion of the interaction terms with their components. As the assumptions check revealed, the explanatory variables are not highly correlated on their own.

## Conclusion and Recommendations

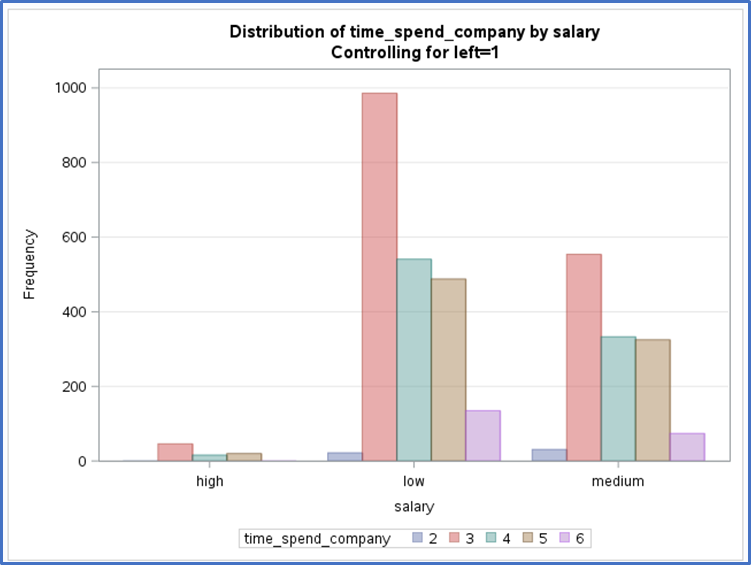
We began by asking why employees leave, and if we can predict which ones are likely to leave. Our final logistic regression model shows we can predict with a good deal of accuracy which employees are likely to leave. While the statistical discussion of logistic regression for prediction can be rather imposing, the bottom line is using this data set; we have a good idea of the factors which strongly influence the model telling us which employees are at risk for departure. For a full tabular view of the various components of the regression equation, please see Figure A in the appendix.

**Recommendations –**

The satisfaction level of an employee and the last evaluation score are the factors most impactful to the model. High satisfaction levels lead to a log-odds decrease of 26.5% chance of an employee leaving, assuming all other variables remain constant. In practice, we recommend examining the constituent components of satisfaction level to determine other factors besides the ones captured in this data set that can improve satisfaction. This might lead to an increase in satisfaction scores and a corresponding decrease in employee turnover. The last evaluation score also is highly impactful, although the story here is less clear. As shown in Figure 14, below, there is an unexpected result. The probability of leaving is low with a high satisfaction (purple line), as we would expect. With low satisfaction (blue line), the likelihood of leaving decreases with a good last evaluation. The data shows employees with high satisfaction having a higher likelihood of leaving when the last evaluation score is 90 or above. Additional research should be undertaken to understand why some of our best employees (high satisfaction levels and high evaluation scores) are leaving.

*Figure 14 – Probability of leaving interaction of evaluation and satisfaction scores*

Most of the turnover is in low and medium salary employees with relatively short times at the company. The variable most impactful by itself in predicting the likelihood of an employee leaving was low pay, which makes intuitive sense. Low pay by itself raised the log odds by 51%, assuming all other variables were held constant. It is important to keep this in context. Since most employees have short tenure with the company, and most employees are in the low salary group, it makes sense that low salary employees would be more likely to leave. There may be business reasons why a higher turnover rate among these employees is acceptable. See Figure 15 for a visual grouping of salary and time with the company for those employees who left.



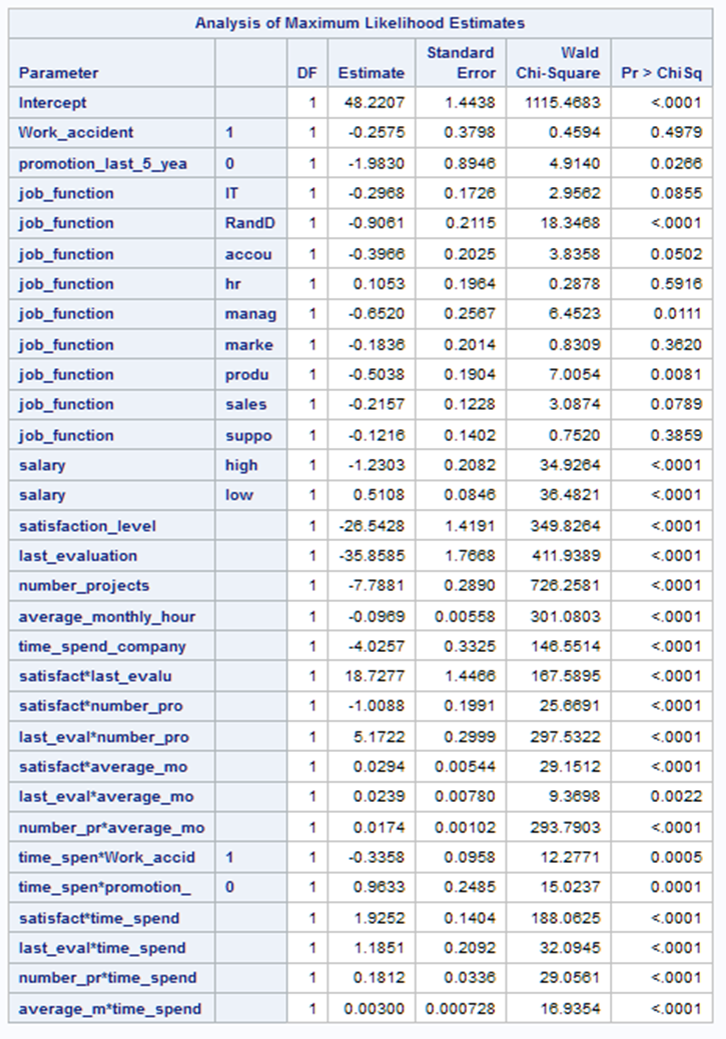
*Figure 15 – Employees who left by salary and time at company*

Using this data set, we have a high likelihood of correctly identifying which employees are likely to leave the company. Our model gives us the ability to identify which employees are likely to leave with an accuracy of 93%. Just as importantly, our rates of error are low, with the model not flagging employees as ‘stay’ who left is only 4% of the time. The model also flagged employees as ‘leave’ but who stayed 3% of the time. The high predictive value of the model, when combined with the low error rates means the company can operationalize this process with a high level of confidence in the model. From a business perspective, management now has another tool to use to help make decisions about staffing. Retaining those long-term, highly evaluated employees at risk of leaving is the most unexpected finding of our research. Solving this issue should help improve both the operations of the company and makes its employees more loyal and productive.

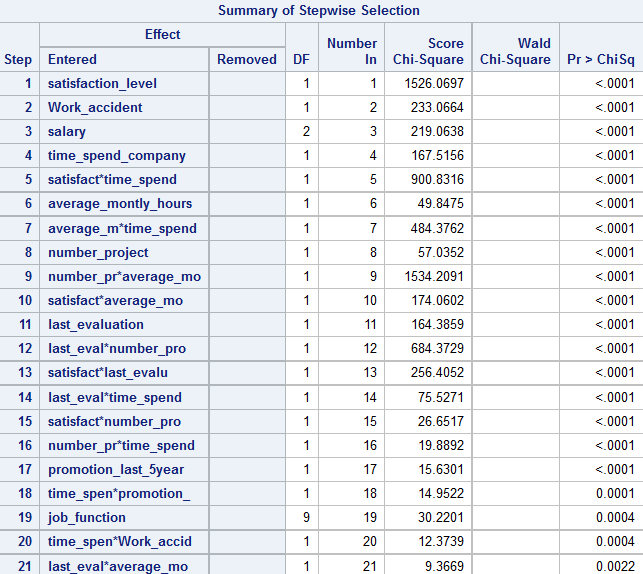
## Appendix A– References and Figures

References:

1. on, Zeynep, and Robert S. Huckman. ["Managing the Impact of Employee Turnover on Performance: The Role of Process Conformance."](http://www.people.hbs.edu/rhuckman/ton_huckman.pdf) *Organization Science* 19, no. 1 (January–February 2008): 56–68
2. IBM.com, downloaded from <https://www-01.ibm.com/software/analytics/solutions/operational-analytics/hr-analytics/> on August 3, 2017
3. Eric Shruenberg, “Richard Branson: Why Customers Come Second at Virgin”, downloaded from [https://www.inc.com/eric-schurenberg/sir-richard-branson-put-your-staff-first-customers-second-and-shareholders-third.html August 3](https://www.inc.com/eric-schurenberg/sir-richard-branson-put-your-staff-first-customers-second-and-shareholders-third.html%20August%203), 2107
4. Downloaded from Kaggle.com <https://www.kaggle.com/ludobenistant/hr-analytics>
5. Habshah Midi , S.K. Sarkar & Sohel Rana (2010) Collinearity diagnostics of binary logistic regression model, Journal of Interdisciplinary Mathematics, 13:3, 253-267, DOI: 10.1080/09720502.2010.10700699



*Figure A- Full model of stepwise selection, 2-level interaction for continuous and categorical*



*Figure B- summary of stepwise selection, 2-level interaction for continuous and categorical*

## Appendix B– Code

\*read in raw data;

data raw;

length job\_function $12;

infile '/home/slcoyne0/Data Sources/HR\_comma\_sep.csv' DLM=',' FIRSTOBS=2;

input satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company work\_accident $ left promotion\_last\_5\_years job\_function $ salary $;

run;

\*make a frequency table of dependent variable to understand balance;

proc freq data=raw;

tables left;

run;

\*baseline EDA for discrete variables;

ods graphics on;

proc freq data=raw;

tables number\_projects time\_spend\_company work\_accident left promotion\_last\_5\_years job\_function salary /plots=Freqplot(scale=freq);

run;

\*baseline EDA for continuous variables pt.1;

proc univariate data=raw;

var satisfaction\_level;

hist/normal;

run;

\*baseline EDA for continuous variables pt.2;

proc univariate data=raw;

var last\_evaluation;

hist/normal;

run;

\*baseline EDA for continuous variables pt.3;

proc univariate data=raw;

var average\_monthly\_hours;

hist/normal;

run;

\*EDA by response level for discrete variables part 1;

proc freq data=raw;

tables left\*number\_projects / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for discrete variables part 2;

proc freq data=raw;

tables left\*time\_spend\_company / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for discrete variables part 3;

proc freq data=raw;

tables left\*work\_accident / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for discrete variables part 4;

proc freq data=raw;

tables left\*promotion\_last\_5\_years / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for discrete variables part 5;

proc freq data=raw;

tables left\*job\_function / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for discrete variables part 6;

proc freq data=raw;

tables left\*salary / plots=FreqPlot(twoway=cluster scale=Freq);

run;

\*EDA by response level for continuous variables part 1;

proc sgplot data=raw;

vbox satisfaction\_level / category=left group=left;

yaxis label = "Job Satisfaction Rating";

keylegend / Title = "Left the Company 1 = Yes 0 = No";

run;

\*EDA by response level for continuous variables part 2;

proc sgplot data=raw;

vbox last\_evaluation / category=left group=left;

yaxis label = "Last Evaluation Score";

keylegend / Title = "Left the Company 1 = Yes 0 = No";

run;

\*EDA by response level for continuous variables part 3;

proc sgplot data=raw;

vbox average\_monthly\_hours / category=left group=left;

yaxis label = "Average Monthly Hours";

keylegend / Title = "Left the Company 1 = Yes 0 = No";

run;

\*correlation matrix to understand multicollinearity;

proc corr data=raw;

var satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company;

run;

\*test for VIF;

proc reg data = raw;

model left = satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company/vif tol collin;

run;quit;

\*create train/test partition using 70\_30 split;

DATA train test;

SET raw;

Random1 = RANUNI(14380132);

IF Random1 < 0.7 THEN output train;

ELSE output test;

Run;

\*==preliminary\_model==;

\*make preliminary model and store specify outmodel to train\_results1;

ods graphics on;

proc logistic data=train outmodel = train\_results1 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company / LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results1 and output predictions to testpred1;

ods graphics on;

proc logistic inmodel=train\_results1;

score data=test out=testpred1 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred1;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==no interaction==;

\*==forward\_selection==;

\*make a model and store specify outmodel to train\_results2\_1;

ods graphics on;

proc logistic data=train outmodel = train\_results2\_1 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company / selection = forward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results2\_1 and output predictions to testpred2\_1;

ods graphics on;

proc logistic inmodel=train\_results2\_1;

score data=test out=testpred2\_1 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred2\_1;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==backward\_selection==;

\*make a model and store specify outmodel to train\_results2\_2;

ods graphics on;

proc logistic data=train outmodel = train\_results2\_2 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company / selection = backward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results2\_2 and output predictions to testpred2\_2;

ods graphics on;

proc logistic inmodel=train\_results2\_2;

score data=test out=testpred2\_2 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred2\_2;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==stepwise\_selection==;

\*make a model and store specify outmodel to train\_results2\_3;

ods graphics on;

proc logistic data=train outmodel = train\_results2\_3 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company / selection = stepwise LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results2\_3 and output predictions to testpred2\_3;

ods graphics on;

proc logistic inmodel=train\_results2\_3;

score data=test out=testpred2\_3 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred2\_3;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==with interaction==;

\*==forward\_selection==;

\*make a model and store specify outmodel to train\_results3\_1;

ods graphics on;

proc logistic data=train outmodel = train\_results3\_1 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = forward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results3\_1 and output predictions to testpred3\_1;

ods graphics on;

proc logistic inmodel=train\_results3\_1;

score data=test out=testpred3\_1 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred3\_1;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==backward\_selection==;

\*make a model and store specify outmodel to train\_results3\_2;

ods graphics on;

proc logistic data=train outmodel = train\_results3\_2 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = backward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results3\_2 and output predictions to testpred3\_2;

ods graphics on;

proc logistic inmodel=train\_results3\_2;

score data=test out=testpred3\_2 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred3\_2;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==stepwise\_selection==;

\*make a model and store specify outmodel to train\_results3\_3;

ods graphics on;

proc logistic data=train outmodel = train\_results3\_3 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = stepwise LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results3\_3 and output predictions to testpred3\_3;

ods graphics on;

proc logistic inmodel=train\_results3\_3;

score data=test out=testpred3\_3 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred3\_3;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==interaction w categorical=;

\*==forward\_selection==;

\*make a model and store specify outmodel to train\_results4\_1;

ods graphics on;

proc logistic data=train outmodel = train\_results4\_1 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company work\_accident | promotion\_last\_5\_years | satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = forward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results4\_1 and output predictions to testpred4\_1;

ods graphics on;

proc logistic inmodel=train\_results4\_1;

score data=test out=testpred4\_1 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred4\_1;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==backward\_selection==;

\*make a model and store specify outmodel to train\_results4\_2;

ods graphics on;

proc logistic data=train outmodel = train\_results4\_2 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company work\_accident | promotion\_last\_5\_years | satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = backward LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results4\_2 and output predictions to testpred4\_2;

ods graphics on;

proc logistic inmodel=train\_results4\_2;

score data=test out=testpred4\_2 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred4\_2;

table F\_left\*I\_left / nocol nocum nopercent;

run;

\*==stepwise\_selection==;

\*make a model and store specify outmodel to train\_results4\_3;

ods graphics on;

proc logistic data=train outmodel = train\_results4\_3 descending plots(only)=(roc(id=obs) effect);

class work\_accident (ref = "0") promotion\_last\_5\_years (ref = "1") job\_function salary / param = ref;

model left(event='1')= work\_accident promotion\_last\_5\_years job\_function salary satisfaction\_level last\_evaluation number\_projects average\_monthly\_hours time\_spend\_company work\_accident | promotion\_last\_5\_years | satisfaction\_level | last\_evaluation | number\_projects | average\_monthly\_hours | time\_spend\_company @2/selection = stepwise LACKFIT outroc=troc ctable pprob = (0.5);

effectplot fit / obs(jitter(y=0.02)) link;

roc; roccontrast;

run;

\*proc logistic with inmodel statement using train\_results4\_3 and output predictions to testpred4\_3;

ods graphics on;

proc logistic inmodel=train\_results4\_3;

score data=test out=testpred4\_3 outroc=vroc;

run;

\*make a confusion matrix;

proc freq data=testpred4\_3;

table F\_left\*I\_left / nocol nocum nopercent;

run;