Analyze the various factors that influences an employee to Exit the company

- Shreyas Rewagad, Utkarsh Kumar Singh & Saheli Saha

Table of Contents

Introduction

Workforce is the primary asset for any company. It is a vital pillar in determining the company's state of business. Employee sentiment is one of the key metric that influences the status of the company and determines whether the company is destined for profits/loss in the up-coming fiscal year. Our main concern are the factors influencing employee to exit the company. It is quintessential for the company to determine these factors and improve on them, thus, retaining the employee. Making a list of trends [1] that are common across most businesses today, we see that:

- One-third of the new hires quit their jobs after about six months.
- More than half of all organizations globally have difficulty retaining some of their most marketable employee groups.
- Referred employees have a 45% retention rate after two years.
- Nearly four out of five (78%) of business leaders rank employee retention as important or urgent.
- Some 35% of employees will start looking for a job if they don't receive a pay raise in the next 12 months.

According to 2016 Employee Engagement/Retention Statistics [2][3], we can get the following statistics:

- 21% of millennials say they've changed jobs within the past year, more than 3x the number of non-millennials
- 93% of millennials left their company the last time they changed roles.
- 74% of all workers are satisfied with their jobs; 66% of those are still open to new employment
- 47% of workers report that they have had to replace more than 20% of their workforce during the past 12 months
- 24% of workers say their employers are putting in less effort to retain them; while 15% counter this statement.

Given the current market condition and the approaching recession [4], it is vital that we try to curb the rising issue of employee attrition. As it is a well-known fact that Human Resource information of any company is confidential, and this is something that can never be made available to the public; very few repositories hold such data. Our source for this data is the Kaggle website.

Data at Glance

The dataset was procured from the Kaggle, a recent Google acquisition, the website possessed some of the most interesting and challenging datasets for the purpose of analytical study. The Hum Resource Analytics dataset is rich w.r.t. the number of records, it consists records of 14999 individuals/employees. And a total of 10 features/attributes per employee. Let us take a dive deeper and understand the various factors/attributes mentioned in the dataset.

Data Description

We have seven-numeric type data and three-character type attributes. Enlisting them below along with their respective range of values and the description:

Attribute	Range of Values	Description
Satisfaction Level	0 - 1	0 indicative of 0%(Least Satisfied) and 1 of 100% (Most Satisfied) respectively.
Last Evaluation	0 – 1	0 indicative of 0%(Lowest Score) and 1 of 100% (Highest Score) respectively.
Average Monthly Hours	96 – 310	Time an employee spends on an average per month in the company.
Time Spend in Company	2 – 10	Number of years spent in the company by an employee
Number of Projects	2 – 7	Number of projects undertaken by the employee during his tenure
Promotion Last 5 Years	0/1	0 indicates that the employee hasn't had a promotion in the last 5 years and 1 indicates otherwise.
Work Accidents	0/1	0 indicates that the employee hasn't had a Work Accident and 1 indicates otherwise.
Salary	High/Medium/Low	Indicates the Salary category of the employee
Department	Accounting/ HR/ IT/ Management/ Marketing/ Product Management/ R & D/ Sales/ Support/ Technical	Indicates the Department of the employee
EXIT	0/1	0 - Indicates that the employee won't exit; 1 - indicates that the employee wishes to exit the company.

Problem Statement

Our aim is to determine the factors that lead/influence an employee to exit the company. The feature (variable) - "Exit" in the dataset is our response variable.

Proposed Solution

As our response variable is binary-nominal variable, we intend to arrive at a probability of the employee exiting the company w.r.t. the 9 attributed mentioned above. We have several was to model the data, enlisting them below:

- OLS regression [5]: The linear regression can be used with the dependent variable – EXIT to model the employees that exit the company. This approach is referred as the linear probability model. This is used to describe the conditional probabilities. Thus, we should logically get output within the range of 0 to 1, whereas we can end up getting output well beyond 1 and less than 0. Here, our assumption of OLS of homoscedasticity and normality of errors is violated. Thus, resulting in invalid standard errors and hypothesis tests. Hence, we cannot move forward with this approach.
- Probit/Logit regression: This method produces comparable results to logistic regression. The modeling technique differs. Hence, the interpretation of results is different for both the types of regression. It is a matter of personal preference as to which model is to be used.

Furthermore, Logistic regression does not make many of the key assumptions regarding linearity, normality, homoscedasticity as in case of GLM/OLS. Logistic regression applies a non-linear log transformation to the predicted odds ratio. Hence, it can handle any kind of relationship. Also, the error terms (the residuals) do not need to be normally distributed. Thus, we decide to proceed with Logistic regression to model our data.

The Model

Initially, we start by creating dummy variables for the categorical variables in our dataset. We create (n-1) dummies to escape the dummy variable trap. Subsequently, considering all the variables to be significant, we run a logistic regression. Our hypothesis in this case would be:

H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 \ldots\ldots = 0$ (jointly)

H_1: $\beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \ldots\ldots \neq 0$

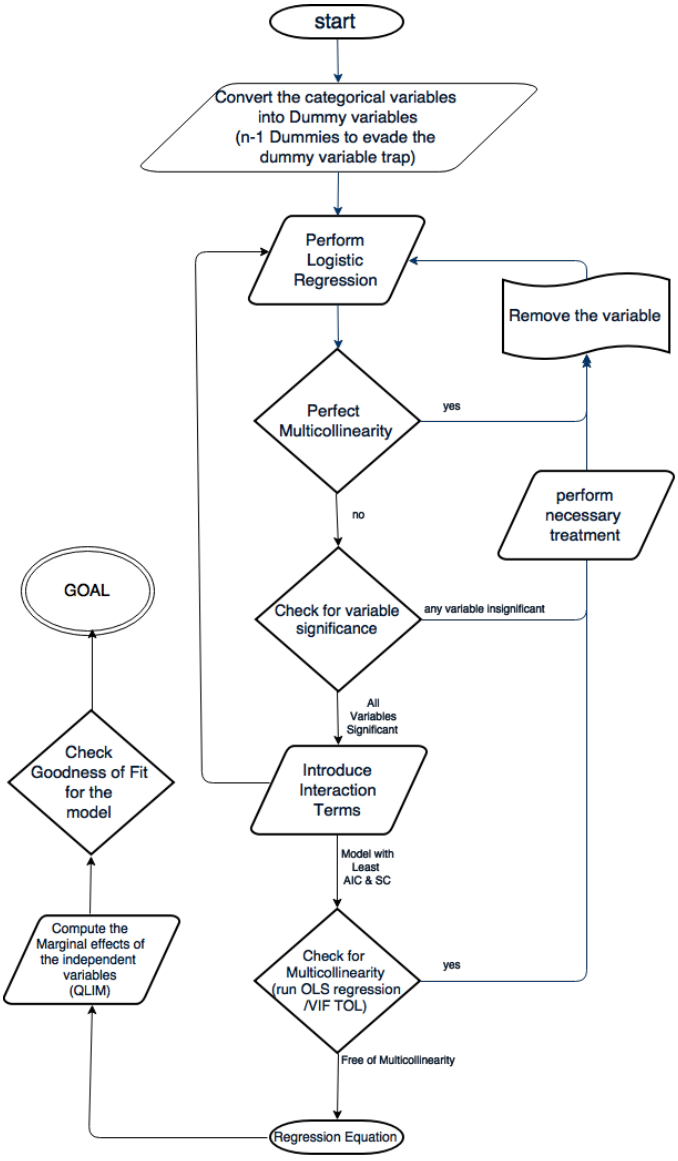
i.e. whether the coefficients of all the variables included in the model are jointly significant or not. As per our analysis we rejected the null hypothesis, indicating that all the variables included in the model are jointly significant.

Moving forward, SAS automatically removes the perfectly collinear variables. We re-run the model with the reduced variables and review the individual coefficients and chi-square statistic to find the significant variables. Our hypothesis here are:

H_0: $\beta_i = 0$;

H_1: $\beta_i \neq 0$.

Subsequently, performing proc logit with the reduced variables. Checking the AIC & SC statistics at every iteration and correct the model for specification errors. We also do a proc reg and generate the VIF and TOL statistic values to check for severe multicollinearity. To further enhance the predictive power of the model we try to introduce interaction terms. We have categorized the continuous variables (satisfaction level, last evaluation, time spend in company and average monthly hours) in 3 divisions: High, Medium and Low, creating 2 dummies for each to evade the dummy variable trap. We have run multiple iterations of the model incorporating the possible interaction terms and accessing the model via AIC & SC statistics in the Intercept and Covariates table; ensuring that our model does not possess any specification errors.



Furthermore, we need to check for potential vulnerabilities w.r.t. heteroscedasticity [6] and autocorrelation. The logistic regression models for the probability of the employee leaving the company (EXIT =1). A probability signifies uncertainty, which is attributed to all variables we have not included in our model.

A logistic regression could have issues related to unequal variance i.e. different variance than the one assumed by the model. This generally occurs when we deal with counts of the number of successes/failure (depends on the outcome that we model for). Which most likely end up with Over-dispersion, which arises when data exhibit variances larger than those assumed under the binomial distribution.

This occurs due to:

- Correlated data
- Heterogeneity among observations
- Large number of 0's: more failures as compared to the successes i.e. very fewer employees exiting the company.

A model that suffers from over-dispersion; exhibits under-estimated Standard errors which leads to under-estimated P-values. This increases the Type-I errors i.e. high false positive numbers. These cascading effects get serious when we intent to use the model for medical test prediction E.g. Predict whether the patient is diagnosed with cancer or not.

But in our case, we need not worry for this issue, though we model to predict the probability of the employee exiting the company, we are not interested in the count of the number of successes.

Moving on to problem of auto-correlation, we do not have time series data not we possess a panel data. Our dataset consists of 14999 individual employee records. Even if two or more employees exhibiting strongly correlated attributes we need not treat it as an anomaly. As, this can very well occur in regular scenario. Thus, we can rest assured that our model is not plagued by the above discussed generic problems.

Once we have finalized our model, we use Proc QLIM to compute the marginal effect [7][8][9] for each of the variables considered in the model. Marginal effect of an attribute X_i on our dependent variable Exit is calculated from the resulting estimated parameter coefficient (β_x); which is then transformed into predicted odds ratio. Thus, providing us with a statistic equivalent to the standard estimated in Linear regression, to judge the impact of an attribute on the probability of the employee exiting the company. We also check the classification table to look at the accuracy and errors (Type I & II) committed by our model.

Regression Analysis Results

As per the procedure explained above, we end up with the below model statistics. From the below tables we observe that, our model does a fairly good job of modelling the data. Looking by the AIC & SC statistics, we have managed to significantly reduce these stats after adding the interaction terms. Thus, signifying no specification errors. We do have the high VIF statistics for the variables – high salary and high salary*HR department. But, we need not be concern as both the variables are significant at alpha level 99%. Also, as stated above, our model cannot suffer from Heteroscedasticity and Autocorrelation.

Model Statistics	Full Model	W/O Dept-Sales	With Interaction terms
AIC	12921.308	12919.362	11552.679
SC	13020.312	13010.751	11697.378
Accuracy	80.10%	80.10%	82.40%
True Positives	1627	1622	2049
True Negatives	10390	10392	10314
False Positives	1038	1036	1114
False Negatives	1944	1949	1522

Variable	Full Model	W/O Dept-Sales	With Interaction terms	VIF Statistic
Intercept	0.4373(0.1207)***	0.4415(0.1193)***	-5.3282(0.2653)***	0
satisfaction_level	-4.1288(0.0978)***	-4.1287(0.0978)***	-4.9141(0.133)***	1.73543
last_evaluation	0.7265(0.1488)***	0.7258(0.1488)***	6.5504(0.2823)***	3.09907
number_project	-0.3139(0.0213)***	-0.314(0.0213)***	-0.2447(0.0229)***	1.43334
average_montly_hours	0.00447(0.000515)***	0.00447(0.000515)***	0.00624(0.000561)***	1.29581
time_spend_company	0.2623(0.0154)***	0.2625(0.0154)***	0.3496(0.0185)***	1.21895
Work_accident	-1.5351(0.0895)***	-1.5353(0.0895)***	-1.5779(0.0943)***	1.00845
promotion_last_5year	-1.4934(0.256)***	-1.4942(0.256)***	-1.2914(0.2596)***	1.01844
dept_sales	0.0117(0.0505)NO			
dept_accounting				
dept_hr	0.2826(0.0965)***	0.2788(0.0951)***		
dept_tech				

dept_support				
dept_mgnt				
dept_IT	-0.1304(0.0835)NO	-0.1342(0.0819)NO		
dept_prodMgnt				
dept_mkt				
salary_high	-2.0039(0.1277)***	-2.005(0.1276)***	-1.2679(0.1684)***	27.67348
salary_mid	-0.5394(0.0456)***	-0.5395(0.0456)***	-0.1432(0.0773)*	2.00478
salary_high*dept_hr			1.6372(0.5354)***	25.55539
salary_mid*dept_hr			0.7778(0.1431)***	1.12467
dept_hr*sat_low			-0.7892(0.2421)***	1.13684
salary_mid*sat_low			-1.1425(0.1293)***	1.856
salary_high*sat_low			-1.1897(0.3296)***	1.18559
salary_mid*sat_mid			-0.4537(0.0912)***	1.87777
salary_hig*eval_high			-1.0542(0.2748)***	2.51441
time_low*eval_mid			-0.8675(0.3351)***	1.83743
eval_mid*time_mid			2.8599(0.1045)***	2.9482

Final Regression Equation

log[p/(1-p)] = -5.3282 - 4.9141* satisfactionLevel + 6.5504* lastEvaluation -0.2447* numberProject + 0.00624* averageMonthlyHours + 0.3496 * timeSpendCompany -1.5779 * workAccident -1.2914 * promotionLast5years -1.2679 * salaryHigh -0.1432 * salaryMid + 1.6372 * highHr + 0.7778 * medHr -0.7892*(satLow*deptHr) -1.1425*(salaryMid*satLow) -1.1897*(salaryHigh*satLow) -0.4537*(salaryMid*satMid) -1.0542 *(salaryHig*evalHigh) -0.8675 *(timeLow*evalMid) + 2.8599 *(evalMid*timeMid)

Interpretation

We can interpret the parameter estimates as follows: for a one unit change in the predictor variable, the difference in log-odds for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

- Intercept - If every variable in the model has value of 0 then Difference in log-odd ratio will be -5.3282
- Satisfaction Level - This is the estimated logistic regression coefficient comparing satisfaction level, given the other variables are held constant in the model. 1% increase in the satisfaction level will decrease difference in log-odd ratio by 4.9141 units given that all other variables in the model are held constant.
- Last Evaluation - This is the estimate logistic regression coefficient for a one unit change in last_evaluation, given the other variables in the model are held constant. If last evaluation increases by 1 % then difference in log odd ratio will increase by 6.5504 given other variables are held constant.
- Number of Projects - This is the estimate logistic regression coefficient for a one unit change in number_project, given the other variables in the model are held constant. With increase of 1 project in number of project variable difference in log odd ratio is supposed to decrease by 0.2447 units keeping all other variables constant.
- Average Monthly Hours - This is the estimate logistic regression coefficient for a one hour change in average_monthly_hours, given the other variables in the model are held constant. 1 hour increase in average monthly hours leads to 0.00624 units increase in the difference in log odd ratio while all other variables are kept constant.
- Time Spend in Company - This is the estimate logistic regression coefficient for a one yeat change in time_spend_company, given the other variables in the model are held constant. With every 1 year increase in time spent in company variable difference in log odd ratio will increase by 0.3496 units when all other variables are kept constant.
- Work Accident - This is the estimate logistic regression coefficient for a one unit change in work_accident i.e. the employee suffers/doesn't suffer a work accident, given the other variables in the model are held constant. If employee suffers a work accident difference in log odd ratio is supposed to decrease by 1.5779 units given that other variables are held constant.
- Promotion Last 5 years - This is the estimate logistic regression coefficient for a one unit change in promotion_last_5years i.e. the employee either receives a promotion in the last 5 years or not, given the other variables in the model are held constant. If an employee get a promotion then difference in log odd ratio will decrease by 1.2914 units keeping all other variables constant.

- High Salary - This is the estimate logistic regression coefficient for a one unit change in salary_high i.e. the salary range of the employee changes to high from any other range, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 1.2679 for all the employee with high salary given other variables are held constant.
- Medium Salary - This is the estimate logistic regression coefficient for a one unit change in salary_mid i.e. the salary range of the employee changes to mid from any other range, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 0.1432 units for all the employee with medium salary given other variables are held constant.
- HR department with high salary - This is the estimate logistic regression coefficient for a one unit change in HR department with high salary, given the other variables in the model are held constant. Difference in log odd ratio expected to increase by 1.6372 units for all the employee from HR department with high salary given other variables are held constant.
- HR department with medium salary - This is the estimate logistic regression coefficient for a one unit change in HR department with medium salary, given the other variables in the model are held constant. Difference in log odd ratio expected to increase by 0.7778 units for all the employee from HR department with medium salary given other variables are held constant.
- sat_low*dept_hr - This is the estimate logistic regression coefficient for a one unit change in sat_low*dept_hr, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 0.7892 units for all the employee from HR department with low satisfaction level given other variables are held constant.
- salary_mid*sat_low - This is the estimate logistic regression coefficient for a one unit change in salary_mid*sat_low, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 1.1425 units for all the employee with medium salary and low satisfaction level given other variables are held constant.
- salary_high*sat_low - This is the estimate logistic regression coefficient for a one unit change in salary_high*sat_low, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 1.1897 units for all the employee with high salary and low satisfaction level given other variables are held constant.
- salary_mid*sat_mid - This is the estimate logistic regression coefficient for a one unit change in salary_mid*sat_mid, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 0.4537 units for all the employee with medium salary and medium satisfaction level given other variables are held constant.
- salary_hig*eval_high - This is the estimate logistic regression coefficient for a one unit change in salary_hig*eval_high, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 1.0542 units for all the employee with high salary and high evaluation given other variables are held constant.
- time_low*eval_mid - This is the estimate logistic regression coefficient for a one unit change in time_low*eval_mid, given the other variables in the model are held constant. Difference in log odd ratio expected to decrease by 0.8675 units for all the employee who have spent less time in the company and having medium evaluation given other variables are held constant.
- eval_mid*time_mid - This is the estimate logistic regression coefficient for a one unit change in eval_mid*time_mid, given the other variables in the model are held constant. Difference in log odd ratio expected to increase by 2.8599 units for all the employee who have spent approximately 3.5 years in the company and having medium evaluation given other variables are held constant.
- Interpreting point estimates:
  - **satisfaction_level -** For 1 % increase in the satisfaction level, odd ratio for exit of an employee from the company increases by 0.007 units given other variables in the model are kept constant.
  - **last_evaluation -** with 1 % increase in last evaluation for an employee, odd ratio of his exit from company increases by 699.541 units given other variables in the model are kept constant.
  - **number_project -** For each additional project, odds ratio of employee exit from the company increases by 0.783 units given other variables in the model are kept constant.
  - **average_montly_hours -** With every 1 hour increase in the average monthly hours, odds ratio will increase by 1.006 units given other variables in the model are kept constant.
  - **time_spend_company -** With every 1 year increase in the time spend in the company, odds ratio of his exit from the comapny will increase by 1.418 units given other variables in the model are kept constant.
  - **Work_accident -** If work accident happens with an employee, odds ratio of his exit supposed to increase by 0.206 units given other variables in the model are kept constant.
  - **promotion_last_5year -** If an employee gets promotion odds ratio of his exit from the company will increase by 0.275 units.

Association of Predicted Probabilities and Observed Responses

- Percent Concordant - A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value (EXIT = 0) has a lower predicted mean score than the observation with the higher ordered response value (EXIT = 1). Higher the value of it better the model is, for out model the value is 86.5.

- Percent Discordant - If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant. Lower the value of it better the model is, for out model the value is 13.4.
- Percent Tied - If a pair of observations with different responses is neither concordant nor discordant, it is a tie.
- Pairs - This is the total number of distinct pairs in which one case has an observed outcome different from the other member of the pair. In the Response Profile table in the Model Information section above, we see that there are 3571 observations with EXIT=1 and 11428 observations with EXIT=0. Thus, the total number of pairs with different outcomes is 3571 *11428 = 40809388.
- Somers' D - Somer's D is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as (nc-nd)/t where nc is the number of pairs that are concordant, nd the number of pairs that are discordant, and t is the number of total number of pairs with different responses. In our example, it equals the difference between the percent concordant and the percent discordant divided by 100: (86.5-13.4)/100 = 0.731.
- Gamma - The Goodman-Kruskal Gamma method does not penalize for ties on either variable. Its values range from -1.0 (no association) to 1.0 (perfect association). Because it does not penalize for ties, its value will generally be greater than the values for Somer's D.
- Tau-a - Kendall's Tau-a is a modification of Somer's D that takes into the account the difference between the number of possible paired observations and the number of paired observations with a different response. It is defined to be the ratio of the difference between the number of concordant pairs and the number of discordant pairs to the number of possible pairs (2(nc-nd)/(N(N-1)). Usually Tau-a is much smaller than Somer's D since there would be many paired observations with the same response.
- c - c is equivalent to the well-known measure ROC. c ranges from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response.

Marginal Effect of Attributes

To standardize our estimates and determine the most important factors affecting the employee's decision to exit the company; we compute the marginal effects of each variable w.r.t. the probability of success (EXIT = 1). In the below table, we have discussed the various factors and their effect on employee's decision.

Variable Name	Mean	Interpretation
satisfaction_level	-0.607734	one unit increase in satisfaction level the employee is 60.77 % less likely to leave the company.
last_evaluation	0.8115537	one unit increase in last evaluation score the employee is 81.15 % more likely to leave the company.
number_project	-0.0302697	one unit increase in number_projects the employee is 3.02% less likely to leave the company.
average_montly_hours	0.000770648	one hour increase in average_monthly_hours the employee is 0.08% more likely to leave the company.
time_spend_company	0.043189	one year increase in time_spend_company the employee is 4.32% more likely to leave the company.
Work_accident	-0.1950544	employee who has incurred work accident is 19.5% less likely to leave the company.
promotion_last_5year	-0.1613746	employee who has got promotion in last five years is 16.13% less likely to leave the company.
salary_high	-0.1552084	employee whose salary is high is 15.52% less likely to leave the company than low salary employee.
salary_mid	-0.0176905	employee whose salary is in medium bracket is 1.77% less likely to leave the company than low salary employee.
salary_high*dept_hr	0.165754	employee from HR department & whose salary is high is 16.57% more likely to leave the company.
salary_mid*dept_hr	0.096276	employee from HR department & whose salary is in medium bracket is 9.62% more likely to leave the company.

dept_hr*sat_low	-0.0967689	employee from HR department & whose satisfaction level is low is 9.67% less likely to leave the company.
salary_mid*sat_low	-0.1412442	employee whose salary is medium & satisfaction level low is 14.12% less likely to leave the company.
salary_high*sat_low	-0.1494416	employee whose salary is high & satisfaction level low is 14.94% less likely to leave the company.
salary_mid*sat_mid	-0.0560504	employee whose salary is medium & satisfaction level medium is 5.6% less likely to leave the company.
salary_hig*eval_high	-0.1267102	employee whose salary is high & last evaluation score high is 12.67% less likely to leave the company.
time_low*eval_mid	-0.1015953	employee whose time spend in company is low & last evaluation score medium is 10.16% less likely to leave the company.
eval_mid*time_mid	0.3540741	employee whose time spend in company is medium & last evaluation score medium is 35.40% more likely to leave the company.

It appears to be a great model with high accuracy of 82.4%. But, we still need to be concerned as we get inflated False Positives i.e. Type-I error. We would need to employ different modelling technique to efficiently model our data as this is outside the scope of logistic regression.

Conclusion

From the above regression analysis, we managed to shortlist the factors that influences the employee's decision to exit the company.
Factors significantly motivating the Employee not to Exit i.e. continue with the company:

- Satisfaction Level: An increase in the satisfaction level motivates the employee continue with the company.
- Work Accident: If the employee has a work accident, the employee is less likely to exit the company.
- Promotion Last 5 Years: If the employee has had a promotion in the last five years, he/she is less likely to quit.
- Salary High: higher compensation makes the employee stay longer in the company.
- Salary high & Satisfaction Low: Employees with high salary would continue to work; despite of low satisfaction level.
- Salary mid & Satisfaction Low Employees with mediocre salary would continue to work; despite of low satisfaction level.
- Salary high & High - Last Evaluation: Employees with high salary and high last evaluation score are likely to stay.
- Low Time spend in company & Mid Last Evaluation: Employees who have recently joined and have received a mediocre last evaluation are more likely to stay with the company.

Factors significantly motivating the employee to Exit the company:

- High Salary & HR Department: If the employee of HR department earns a high pay, he is likely quit.
- Mid Last Evaluation and Mid Time spend in company: An employee who has spent nearly 3.5 years in the company, receives a mediocre last evaluation score. It is likely that employee to exit the company.
- Last Evaluation: Increase in Last Evaluation rating motivates an employee to quit the company.

Future Work

We find that some variables do not make intuitive sense, such as, Work Accident. Our initial assumption w/o exploring the data was that work accident would influence the employee to exit the company. But, we discover a counter intuitive effect. We would want to procure more data w.r.t. work accident to gain a better understanding.
As mentioned earlier, Logistic regression doesn't seem to model the data well. As per our study, the False negative count is inflated as we try to add significant variables to the model, despite of the AIC and SC statistics yield a considerably lower value. We should consider different approach to modelling our problem.

Technical Appendix

Looking at the initial 10 rows in the dataset:

We proceed as per the flow-chart described above, make dummy variables and run a logistic model.

satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	EXIT	promotion_last_5years	department	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.80	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low
0.41	0.50	2	153	3	0	1	0	sales	low
0.10	0.77	6	247	4	0	1	0	sales	low
0.92	0.85	5	259	5	0	1	0	sales	low
0.89	1.00	5	224	5	0	1	0	sales	low
0.42	0.53	2	142	3	0	1	0	sales	low

The perfectly collinear variables are removed by SAS.

Full Model

The LOGISTIC Procedure

Model Information	
Data Set	PROJECT.DATADUMMY
Response Variable	EXIT
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	14999
Number of Observations Used	14999

Response Profile		
Ordered Value	EXIT	Total Frequency
1	1	3571
2	0	11428

Probability modeled is EXIT='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	16466.691	12921.308
SC	16474.306	13020.312
-2 Log L	16464.691	12895.308

R-Square	0.2118	Max-rescaled R-Square	0.3178

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3569.3829	12	<.0001
Score	3195.6437	12	<.0001
Wald	2460.8682	12	<.0001

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

dept_accounting =	0
dept_tech =	0
dept_support =	0
dept_mgnt =	0
dept_prodMgnt =	0
dept_mkt =	0

We can evaluate our full model as per the below Goodness of fit for the full model can be determined by the below classification table:

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.450	1627	10390	1038	1944	80.1	45.6	90.9	38.9	15.8

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.4373	0.1207	13.1283	0.0003	
satisfaction_level	1	-4.1288	0.0978	1782.8543	<.0001	-0.5660
last_evaluation	1	0.7265	0.1488	23.8273	<.0001	0.0686
number_project	1	-0.3139	0.0213	217.8336	<.0001	-0.2133
average_montly_hours	1	0.00447	0.000515	75.2529	<.0001	0.1230
time_spend_company	1	0.2623	0.0154	289.9485	<.0001	0.2112
Work_accident	1	-1.5351	0.0895	294.4482	<.0001	-0.2977
promotion_last_5year	1	-1.4934	0.2560	34.0252	<.0001	-0.1188
dept_sales	1	0.0117	0.0505	0.0541	0.8161	0.00289
dept_accounting	0	0	.	.	.	.
dept_hr	1	0.2826	0.0965	8.5773	0.0034	0.0337
dept_tech	0	0	.	.	.	.
dept_support	0	0	.	.	.	.
dept_mgnt	0	0	.	.	.	.
dept_IT	1	-0.1304	0.0835	2.4366	0.1185	-0.0197
dept_prodMgnt	0	0	.	.	.	.
dept_mkt	0	0	.	.	.	.
salary_high	1	-2.0039	0.1277	246.2753	<.0001	-0.3039
salary_mid	1	-0.5394	0.0456	140.0223	<.0001	-0.1472

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
satisfaction_level	0.016	0.013	0.020
last_evaluation	2.068	1.545	2.768
number_project	0.731	0.701	0.762
average_montly_hours	1.004	1.003	1.005
time_spend_company	1.300	1.261	1.340
Work_accident	0.215	0.181	0.257
promotion_last_5year	0.225	0.136	0.371
dept_sales	1.012	0.917	1.117
dept_hr	1.327	1.098	1.603
dept_IT	0.878	0.745	1.034
salary_high	0.135	0.105	0.173
salary_mid	0.583	0.533	0.638

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.9	Somers' D	0.640
Percent Discordant	17.9	Gamma	0.641
Percent Tied	0.2	Tau-a	0.232
Pairs	40809388	c	0.820

We need to remove the perfectly collinear variables and the variable – dept_sales as it is not significant.

Model without the strongly colinear variables - Dept<sales>

The LOGISTIC Procedure

Model Information	
Data Set	PROJECT.DATADUMMY
Response Variable	EXIT
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	14999
Number of Observations Used	14999

Response Profile		
Ordered Value	EXIT	Total Frequency
1	1	3571
2	0	11428

Probability modeled is EXIT='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	16466.691	12919.362
SC	16474.306	13010.751
-2 Log L	16464.691	12895.362

R-Square	0.2118	Max-rescaled R-Square	0.3178

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3569.3288	11	<.0001
Score	3195.3026	11	<.0001
Wald	2460.7638	11	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	0.4415	0.1193	13.6847	0.0002	
satisfaction_level	1	-4.1287	0.0978	1782.7859	<.0001	-0.5659
last_evaluation	1	0.7258	0.1488	23.7910	<.0001	0.0685
number_project	1	-0.3140	0.0213	217.9534	<.0001	-0.2134
average_montly_hours	1	0.00447	0.000515	75.2476	<.0001	0.1230
time_spend_company	1	0.2625	0.0154	290.5420	<.0001	0.2113
Work_accident	1	-1.5353	0.0895	294.5284	<.0001	-0.2977
promotion_last_5year	1	-1.4942	0.2560	34.0589	<.0001	-0.1189
dept_hr	1	0.2788	0.0951	8.5959	0.0034	0.0333
dept_IT	1	-0.1342	0.0819	2.6866	0.1012	-0.0203
salary_high	1	-2.0050	0.1276	246.9315	<.0001	-0.3041
salary_mid	1	-0.5395	0.0456	140.1481	<.0001	-0.1473

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
satisfaction_level	0.016	0.013	0.020
last_evaluation	2.066	1.544	2.766
number_project	0.731	0.701	0.762
average_montly_hours	1.004	1.003	1.005
time_spend_company	1.300	1.261	1.340
Work_accident	0.215	0.181	0.257
promotion_last_5year	0.224	0.136	0.371
dept_hr	1.322	1.097	1.592
dept_IT	0.874	0.745	1.027
salary_high	0.135	0.105	0.173
salary_mid	0.583	0.533	0.637

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.9	Somers' D	0.640
Percent Discordant	17.9	Gamma	0.641
Percent Tied	0.2	Tau-a	0.232
Pairs	40809388	c	0.820

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.450	1622	10392	1036	1949	80.1	45.4	90.9	39.0	15.8

Looking at the AIC and SC statistics we consider the latest model as the AIC & SC values are lower. Also, all vaiables except dept_IT are significant. We may remove dept_IT, but, as the chi-square statistic is slightly over alpha of 90%. To improve the predictive power of our model, we decide to incorporate interaction terms. There are 3 continuous variables: satisfaction level, last evaluation and average monthly hours. We have split them in 3 categories high, medium and low. Subsequently, making dummies. There are myriad possible interactions, hence, via exploratory data analysis we have determined few significant interactions and included it in out model. As discussed above, the issues such as heteroscedasticity and autocorrelation will not occur in our model. But, we may encounter the issue of multicollinearity. We test this using the linear regression's VIF and TOL statistics. A VIF value above 10 is considered as severe multicollinearity.

Logistic with Interaction variables

The LOGISTIC Procedure

Model Information	
Data Set	PROJECT.DATA_INTR
Response Variable	EXIT
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	14999
Number of Observations Used	14999

Response Profile

Ordered Value	EXIT	Total Frequency
1	1	3571
2	0	11428

Probability modeled is EXIT='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	16466.691	11552.679
SC	16474.306	11697.378
-2 Log L	16464.691	11514.679

R-Square	0.2811	Max-rescaled R-Square	0.4218

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4950.0117	18	<.0001
Score	4144.3431	18	<.0001
Wald	2721.4398	18	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-5.3282	0.2653	403.3326	<.0001	
satisfaction_level	1	-4.9141	0.1330	1365.0326	<.0001	-0.6736
last_evaluation	1	6.5504	0.2823	538.3373	<.0001	0.6182
number_project	1	-0.2447	0.0229	114.1295	<.0001	-0.1663
average_montly_hours	1	0.00624	0.000561	123.4768	<.0001	0.1718
time_spend_company	1	0.3496	0.0185	355.2075	<.0001	0.2814
Work_accident	1	-1.5779	0.0943	279.8774	<.0001	-0.3060
promotion_last_5year	1	-1.2914	0.2596	24.7432	<.0001	-0.1027
salary_high	1	-1.2679	0.1684	56.6876	<.0001	-0.1923
salary_mid	1	-0.1432	0.0773	3.4339	0.0639	-0.0391
salary_high*dept_hr	1	1.6372	0.5354	9.3516	0.0022	0.0494
salary_mid*dept_hr	1	0.7778	0.1431	29.5280	<.0001	0.0655
dept_hr*sat_low	1	-0.7892	0.2421	10.6295	0.0011	-0.0352
salary_mid*sat_low	1	-1.1425	0.1293	78.0533	<.0001	-0.1441
salary_high*sat_low	1	-1.1897	0.3296	13.0318	0.0003	-0.0617
salary_mid*sat_mid	1	-0.4537	0.0912	24.7621	<.0001	-0.0916
salary_hig*eval_high	1	-1.0542	0.2748	14.7212	0.0001	-0.1249
time_low*eval_mid	1	-0.8675	0.3351	6.7029	0.0096	-0.1322
eval_mid*time_mid	1	2.8599	0.1045	749.1499	<.0001	0.7176

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
satisfaction_level	0.007	0.006	0.010
last_evaluation	699.541	402.255	>999.999
number_project	0.783	0.749	0.819
average_montly_hours	1.006	1.005	1.007
time_spend_company	1.418	1.368	1.471
Work_accident	0.206	0.172	0.248
promotion_last_5year	0.275	0.165	0.457

Association of Predicted Probabilities and Observed Responses

Percent Concordant	86.5	Somers' D	0.731
Percent Discordant	13.4	Gamma	0.732
Percent Tied	0.1	Tau-a	0.265
Pairs	40809388	c	0.866

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.450	2049	10314	1114	1522	82.4	57.4	90.3	35.2	12.9

From the model statistics, we can confirm that this is the best model so far. The AIC and SC statistic values are the least as compares to any of the previous models. Even after addition of interaction terms our AIC and SC statistics are lower. It is to be notes that AIC and SC criterion penalizes for any additional independent terms. Thus, signifying that our model is free of specification errors. The model accuracy is 82.4% and True positive, True negative, false positive & false negative counts are 2049, 10314, 1114 and 1522 respectively. Indicating that the model does a fairly good job of predicting the employee's probability to exit the company.

Further, we check the issue of multicollinearity via Proc Reg.

From the below parameter estimate table, we can see that salary_high*dept_hr and salary_high show a VIF value greater than 25. This gives us an evidence of high multicollinearity. But, we need not do any transformation to the data as both these variables are highly significant.

Checking for multicollinearity

The REG Procedure
Model: MODEL1
Dependent Variable: EXIT

Number of Observations Read	14999
Number of Observations Used	14999

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	751.78070	41.76559	317.75	<.0001
Error	14980	1969.02655	0.13144		
Corrected Total	14998	2720.80725			

Root MSE	0.36255	R-Square	0.2763
Dependent Mean	0.23808	Adj R-Sq	0.2754
Coeff Var	152.27981		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-0.19153	0.06313	-3.03	0.0024	.	0
satisfaction_level	1	-0.72576	0.01569	-46.27	<.0001	0.57623	1.73543
last_evaluation	1	0.70338	0.03045	23.10	<.0001	0.32268	3.09907
number_project	1	-0.02265	0.00288	-7.88	<.0001	0.69767	1.43334
average_montly_hours	1	0.00073948	0.00006748	10.96	<.0001	0.77172	1.29581
time_spend_company	1	0.04101	0.00224	18.32	<.0001	0.82037	1.21895
Work_accident	1	-0.14177	0.00845	-16.77	<.0001	0.99162	1.00845
promotion_last_5years	1	-0.10800	0.02071	-5.22	<.0001	0.98189	1.01844
salary_high	1	-0.01996	0.05661	-0.35	0.7244	0.03614	27.67348
salary_mid	1	-0.01728	0.00847	-2.04	0.0413	0.49881	2.00478
sal_dept_hr	1	0.12741	0.05533	2.30	0.0213	0.03913	25.55539
sal_mid_hr	1	0.11413	0.02054	5.56	<.0001	0.88915	1.12467
sat_low_hr	1	-0.09113	0.03898	-2.34	0.0194	0.87963	1.13684
sat_low_sal_mid	1	-0.17396	0.01763	-9.87	<.0001	0.53879	1.85600
sat_low_sal_high	1	-0.31638	0.03426	-9.24	<.0001	0.84346	1.18559
sat_mid_sal_mid	1	-0.10315	0.01108	-9.31	<.0001	0.53255	1.87777
eval_high_sal_high	1	-0.00816	0.02184	-0.37	0.7086	0.39771	2.51441
time_low_eval_mid	1	0.09891	0.01452	6.81	<.0001	0.54424	1.83743
time_mid_eval_mid	1	0.33533	0.01117	30.02	<.0001	0.33919	2.94820

Logit Marginal Effects

The QLIM Procedure

Discrete Response Profile of EXIT

Index	Value	Total Frequency
1	0	11428
2	1	3571

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	EXIT
Number of Observations	14999
Log Likelihood	-5758
Maximum Absolute Gradient	3.35178
Number of Iterations	172
Optimization Method	Quasi-Newton
AIC	11553
Schwarz Criterion	11698

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	4949.7	2 * (LogL - LogL0)
Upper Bound of R (U)	16465	- 2 * LogL0
Aldrich-Nelson	0.2481	R / (R+N)
Cragg-Uhler 1	0.2811	1 - exp(-R/N)
Cragg-Uhler 2	0.4218	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.3246	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.3222	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.3006	R / U
Veall-Zimmermann	0.4742	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.8009	

N = # of observations, K = # of regressors

Algorithm converged.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-5.339949	0.265452	-20.12	<.0001
satisfaction_level	1	-4.918774	0.133057	-36.97	<.0001
last_evaluation	1	6.568415	0.282507	23.25	<.0001
number_project	1	-0.244991	0.022908	-10.69	<.0001
average_montly_hours	1	0.006237	0.000561	11.11	<.0001
time_spend_company	1	0.349556	0.018548	18.85	<.0001
Work_accident	1	-1.578698	0.094351	-16.73	<.0001
promotion_last_5years	1	-1.306107	0.260636	-5.01	<.0001
salary_high	1	-1.256199	0.168031	-7.48	<.0001
salary_mid	1	-0.143180	0.077293	-1.85	0.0640
salary_high*dept_hr	1	1.341551	0.562727	2.38	0.0171
salary_mid*dept_hr	1	0.779223	0.143189	5.44	<.0001
dept_hr*sat_low	1	-0.783212	0.242595	-3.23	0.0012
salary_mid*sat_low	1	-1.143178	0.129356	-8.84	<.0001
salary_high*sat_low	1	-1.209525	0.329182	-3.67	0.0002
salary_mid*sat_mid	1	-0.453651	0.091184	-4.98	<.0001
salary_high*eval_high	1	-1.025545	0.273049	-3.76	0.0002
time_low*eval_mid	1	-0.822275	0.329444	-2.50	0.0126
eval_mid*time_mid	1	2.865745	0.104567	27.41	<.0001

Now, we have reached the final stage of our analysis and arrived to our final model. We need to determine the impact of the various variables on the employee's exit probability. We use Proc QLIM to yield a marginal effect value for each variable w.r.t. individual employee. We need to computer the mean of these values to find the impact that one attribute/variable is creating on across all employees. We use Proc Mean to procure this statistic.

Logit Marginal Effects

The MEANS Procedure

Variable	Label	Mean	Std Dev
Meff_P2_satisfaction_level	Marginal effect of satisfaction_level on the probability of EXIT=2	-0.6077340	0.4314674
Meff_P2_last_evaluation	Marginal effect of last_evaluation on the probability of EXIT=2	0.8115537	0.5761715
Meff_P2_number_project	Marginal effect of number_project on the probability of EXIT=2	-0.0302697	0.0214903
Meff_P2_average_montly_hours	Marginal effect of average_montly_hours on the probability of EXIT=2	0.000770648	0.000547130
Meff_P2_time_spend_company	Marginal effect of time_spend_company on the probability of EXIT=2	0.0431890	0.0306625
Meff_P2_Work_accident	Marginal effect of Work_accident on the probability of EXIT=2	-0.1950544	0.1384810
Meff_P2_promotion_last_5years	Marginal effect of promotion_last_5years on the probability of EXIT=2	-0.1613746	0.1145697
Meff_P2_salary_high	Marginal effect of salary_high on the probability of EXIT=2	-0.1552084	0.1101919
Meff_P2_salary_mid	Marginal effect of salary_mid on the probability of EXIT=2	-0.0176905	0.0125595
Meff_P2_salary_high_dept_hr	Marginal effect of salary_high_dept_hr on the probability of EXIT=2	0.1657540	0.1176789
Meff_P2_salary_mid_dept_hr	Marginal effect of salary_mid_dept_hr on the probability of EXIT=2	0.0962760	0.0683522
Meff_P2_dept_hr_sat_low	Marginal effect of dept_hr_sat_low on the probability of EXIT=2	-0.0967689	0.0687021
Meff_P2_salary_mid_sat_low	Marginal effect of salary_mid_sat_low on the probability of EXIT=2	-0.1412442	0.1002778
Meff_P2_salary_high_sat_low	Marginal effect of salary_high_sat_low on the probability of EXIT=2	-0.1494416	0.1060977
Meff_P2_salary_mid_sat_mid	Marginal effect of salary_mid_sat_mid on the probability of EXIT=2	-0.0560504	0.0397936
Meff_P2_salary_high_eval_high	Marginal effect of salary_high_eval_high on the probability of EXIT=2	-0.1267102	0.0899593
Meff_P2_time_low_eval_mid	Marginal effect of time_low_eval_mid on the probability of EXIT=2	-0.1015953	0.0721287
Meff_P2_eval_mid_time_mid	Marginal effect of eval_mid_time_mid on the probability of EXIT=2	0.3540741	0.2513788

Logit Predicted Probabilities

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
EXIT		14999	0.2380825	0.4259241	0	1.0000000
pred	Estimated Probability	14999	0.2380828	0.2404759	0.000111658	0.9761757

The above table on Marginal effects, the mean value for the variables signify the effect a unit increase in the variable creates on the employee's decision to quit the company.

The predicted probability table on the left helps us determine the goodness of fit for out model.

References

[1] Employee Retention Statistics: https://www.eremedia.com/tlnt/9-employee-retention-statistics-that-will-make-you-sit-up-and-pay-attention/

[2] 2016 Employee Engagement & Loyalty statistics: http://blog.accessperks.com/2016-employee-engagement-loyalty-statistics#1

[3] Employee Renton Statistics: http://blog.bonus.ly/10-surprising-employee-retention-statistics-you-need-to-know

[4] Bloomberg Recession Forecast: https://www.bloomberg.com/view/articles/2016-10-14/the-next-recession-is-coming-big-deal

[5] UCLA Logit Regression: http://stats.idre.ucla.edu/sas/dae/logit-regression/

[6] Heteroscedasticity discussion by Maarten Buis: http://www.stata.com/statalist/archive/2010-11/msg00996.html

[7] Ani Katchova: https://www.youtube.com/watch?v=iy8nG8ylzCY

[8] SAS Support – Marginal Effect Estimator for Logit Model: http://support.sas.com/kb/22/604.html

[9] Discussion by Gina Nicolosi: https://groups.google.com/forum/#!topic/comp.soft-sys.sas/nNoLSGms9bQ

[10] Over-dispersion discussion and presentation by Jessica Harwood, UCLA: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&uact=8&ved=0ahUKEwjC4aHenr3TAhUo54MKHTSIA3EQFggg9MAQ&url=http://chipts.ucla.edu/downloads/738&usg=AFQjCNFKoNtFK_ws6t36-kODfDagN34FWA&sig2=vcdZf_bKu1bUTfKt4C9p0Q

[11] SAS Support – Overdispersion: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect068.htm

[12] Theory and Adjustment: https://onlinecourses.science.psu.edu/stat504/node/162

[13] Discussion on Heteroscadasticity by Dale McLerran: https://groups.google.com/forum/#!topic/comp.soft-sys.sas/cN49H1NtJaA