Data Analysis Report

This report is written to present the key findings of TWO projects of **ANALYTICS7**. The first project is to study the contributing factors of employee's demographic characteristics of **TassPaperMill** (TPM) company as well as other numerical attributes on how the salary increment was measured. In addition, TPM was also interested to learn the major contributors of its high employee turnover rate. The second project is to locate the future demand of energy resources of **Australian Power** (AP) which helps the company to visualize the up-coming need in 2020. This report is intended to Mr. Hugo Barra who leads data analysis team at ANALYTICS7. The technical details of this report are given in response to the meeting minutes hosted by him, with reference number: AP-211 TPM project.

The following parts of the report covers the details of each minute item (*in underlined and italicized format*) and followed by their key findings of the respective item. Since there was no specified accuracy requirement, I decided to use the rule of thumb to determine the level of alpha value which, in this case, was 5%. Next, the overall summary of the project objectives and findings will be restated to conclude this report. In addition, limitations of data analysis will also be provided in conjunction to the summary section. Appendices are also attached to provide visual representations for a better insight and serve as evidential references for all the claims made within this report.

Minutes of ANALYTICS7 Team Meeting
TOPIC: TPM and AP Research Projects – Analytics Details

ITEM 1:

1. *Providing an overall summary of the following two variables:*

1.1. *Percentage increase in salary (PercentSalaryHike)*

PercentSalaryHike is a discrete numerical variable whose values ranging from 11 to 25 with the mean of 15.21. The majority of the observations contained low values of the range and the proportion of observations decreased as the values of salary hike went higher (see **Figure 1**). Overall, the distribution of **PercentSalaryHike** was strongly positive skewed with standard deviation of 3.65. From the distribution, we can say that most of the employees were intended to low hike. As seen in **Table 1**, approximately 62.52% of all 1470 employees received the lowest range of salary hike from 11% to 15%.

1.2. *Attrition*

Attrition is a nominal categorical variable whose values are 0 and 1, representing the decision of employee whether they continued or left the job respectively. In the dataset, the proportion of employees who continued working for TPM outnumbers those who left the job. More specifically, 16% of all 1470 sample population decided to leave TPM and the rest of 84% remained with the company (see **Figure 2**).

ITEM 2:

2. *Identify potential variables that may influence **PercentSalaryHike**:*

2.1. *Identify a list of possible variables that influence percentage increase in salary. Which three independent variables have the more impactful linear relationship with **PercentSalaryHike**? What form of relationship(s) exist between the independent variable(s) and **PercentSalaryHike**? Are there any potential multi-collinearity problems? If so, which variables are they?*

To find potential influence of any independent variables over **PercentSalaryHike**, I ran the correlation estimation for all variables. Based on the result table, I found that **PerformanceRating** had the most impact on **PercentSalaryHike** with correlation value of 76.48% (see **Figure 3**). After plotting, I realized that this relationship was not linear but rather a quadratic one. The second and the third most influential variables were **JobSatisfaction** and **JobInvolvement** with positive linear relationship of 67.45% and 55.04% respectively. There were also two other variables which had moderate linear relationship with

PercentSalaryHike such **Gender** and **NumCompaniesWorked**. However; unlike **Gender**, **NumCompaniesWorked** had a negative linear form of relationship with **PercentSalaryHike**. The rest of the remaining variables tended to have weak or very weak linear relationships with **PercentSalaryHike**. Beside these linear relationships, I also found that there was a multi-collinearity relationship at the value of 75.88% between independent variables namely, **YearsInCurrentRole** and **YearsAtCompany** (see **Table 2**).

2.2. Build a regression model to estimate percentage increase in salary.

Building a regression model is a repetitive process. In my work, I tried to eliminate any insignificant variables once at a time to ensure that the final model would contain all important variables. Throughout the process, nine variables had been removed including **YearSinceLastPromotion**, **YearsInCurrentRole**, **Education**, **EnvironmentSatisfaction**, **Attrition**, **WorkLifeBalance**, J**obLevel**, **OverTime** and **MaritalStatus**. So only the remaining seven independent variables were kept in the model. This model produced a significant outcome with R Squared value up to 0.8267 which explained its high reliability in predicting the percentage of salary hike (see **Table 3**). Below is the final linear regression formula obtained from the model:

Y = - 4.124 + 0.011(**Age**) + 0.842(**Gender**) + 0.633(**JobInvolvement**) + 0.932(**JobSatisfaction**) - 0.28(**NumCompaniesWorked**) + 4.8.06(**PerformanceRating**) - 0.017(**YearsAtCompany**)

With this model, we can interpret the prediction of **PercentSalaryHike** as follow:

- ☐ Holding other variables constant, a unit increase in **Age** will additionally increase the **PercentageSalaryHike** by 0.011%
- ☐ Holding other variables constant, a unit increase in **Gender** will additionally increase the **PercentageSalaryHike** by 0.842%
- ☐ Holding other variables constant, a unit increase in **JobInvolvement** will additionally increase the **PercentageSalaryHike** by 0.633%
- ☐ Holding other variables constant, a unit increase in **JobSatisfaction** will additionally increase the **PercentageSalaryHike** by 0.932%
- ☐ Holding other variables constant, a unit increase in **NumCompaniesWorked** will descrease the **PercentageSalaryHike** by 0.28%
- ☐ Holding other variables constant, a unit increase in **PerformanceRating** will additionally increase the **PercentageSalaryHike** by 4.806%
- ☐ Holding other variables constant, a unit increase in **YearsAtCompany** will descrease the **PercentageSalaryHike** by -0.017%

2.3. Perform residual analysis. Based on your residual plots, does there appear to be any problems with the regression model?

Based on the model built in 2.2 above, I conducted the residual plots to check whether there was any problem that missed the standard of residual normality. As the results, all plots were acceptable, except that of **PerformanceRating** plot. But this was problem that I expected. At section 2.1, I learned that there was a quadratic form of relationship between **PerformanceRating** and the **PercentSalaryHike**. So, a pure linear model would result in some irregularities of the residual. To fix the issue, I ran another test model using Polynomial (Order 2) by adding extra PerformanceRating^2 variable. The outcome was more promising with the R-Squared value of 0.8388, compared to the linear one of 0.8267 (see Table 4). With this model, the residual plot (see **Figure 5**) of **PerformanceRating** looked more standardized comparing to the previous model (see **Figure 4**).

ITEM 3:

3. *Hugo has performed some preliminary analysis and discovered that the performance rating is a significant predictor of the Percentage increase in salary. Prior research shows that the strength of the relationship between performance rating and percentage increase in salary may vary according to satisfaction with the job. Generally speaking increased job satisfaction creates a more productive workforce as they are more motivated to improve their job performance. Therefore, Hugo believes that the relationship between performance rating and percentage increase in salary should be stronger for employees who are satisfied with their jobs.*

3.1. *Model the interaction between the variables to test Hugo's assumption.*

In order to see the effect of interaction of **JobSatisfaction** and **PerformanceRating**, I formulated two different Linear Regression models. One model excluded the interaction term (model-1) and the other one was included (model-2). So, there were a total of four variable in model 2: **JobSatisfaction** and **PerformanceRating** as independent variables, the product of **JobSatisfaction** and **PerformanceRating** as the interaction term, and **PercentSalaryHike** as dependent variable.

3.2. *Comment on whether there is sufficient evidence to conclude that the interaction term in the model is statistically significant.*

The test results of both models yielded the significant evidences of **JobSatisfaction** and **PerformanceRating** (with p-value less than alpha = 5%) which affected the value of **PercentSalaryHike**. More importantly, the model-2 (with interaction term) revealed a more reliable result with R-Squared value of 0.748 versus 0.744 of the model-1.

In addition to the summary outcome table, I plotted the interaction effects between **JobSatisfaction** and **PerformanceRating** to have a clear view about the nature of the interaction (see **Figure 6**). As seen in the figure, in either case of low or high performance rate, salary percentage was boosted up by the effect of job satisfaction. This meant that we always expected better percentage of salary hike in all performance rating cases whenever job satisfaction was higher. More remarkably, at the high level of job satisfaction, the increment of salary percentage jumped more sharply than that at low level of job satisfaction. Therefore, the interacting effect of job satisfaction on performance rate did exist and it contributed to various increments of salary percentage.

ITEM 4:

4. *A model to predict the likelihood of an employee leaving the TPM*

4.1. *Hugo has already performed an analysis with Attrition and Age, Environment Satisfaction, Overtime and Years in current role as the independent variables. Continue to refine his work and develop a model to ascertain the likelihood of an employee leaving the TPM.*

To predict the likelihood of an employee leaving the TPM, Logistic Regression is more reliable than Linear Regression model. Therefore, in order to visualize what had been done by Mr. Hugo, I setup the exact model he mentioned using Logistic Regression (with the four independent variables). His model showed the evidence that all variables were significant. The practical accuracy was 85.37% and R-Squared(L) was at 0.13447 (see **Table 5**). The low R-Squared(L) received from the model explained the unreliability of the model in predicting the likelihood. But I also accepted the fact that dataset has the unbalance proportion of Attrition values (16% of “success” vs. 84% of “failure” observations). So, it was not the problem of the model but the insufficient proportion of “success” data. If the improvement is required, more data needed to be collected to increase the amount of the “success” respondents.

It seemed that his model relied on very few variables when there were much more variables leaving behind. Thus, in order to refine his work, I started to build the model employing all independent variables and gradually eliminate the least significant variable once at a time. The repetition was continued until none of the remaining independent variables were insignificant. Based on this iterative process, I finally completed the model with eight remaining independent variables including: **Age**,

EnvironmentSatisfaction, **JobLevel**, **MaritalStatus**, **OverTime**, **WorkLifeBalance**, **YearsInCurrentRole**, and **YearsSinceLastPromotion**. This model produced the practical accuracy of 85.71% and R-Squred(L) of 0.1847 (see **Table 6**) which was an improvement on the previous model formulated by Mr. Hugo using only four independent variables. Based on these properties of the model, I was certain that it gave a more reliable prediction of **Attrition**. The model was formulated as follow:

Logit = - 1.576 - 0.031(**Age**) - 0.351 (**EnvironmentSatisfaction**) - 0.376(**JobLevel**)
+ 0.642(**MaritalStatus**) + 1.555(**OverTime**) - 0.231(**WorkLifeBalance**)
- 0.163(**YearsInCurrentRole**) + 0.146(**YearsSinceLastPromotion**)
odds = exp(Logit)
P(y) = Odds/(1+Odds)

Interpretation:

- ☐ Holding other variables constant, a unit increase in **Age** will decrease the odds of employee leaving the company by 3.07%
- ☐ Holding other variables constant, a unit increase in **EnvironmentSatisfaction** will decrease the odds of employee leaving the company by 28.58%
- ☐ Holding other variables constant, a unit increase in **JobLevel** will decrease the odds of employee leaving the company by 31.32%
- ☐ Holding other variables constant, a unit increase in **MaritalStatus** will increase the odds of employee leaving the company by 90.04%
- ☐ Holding other variables constant, a unit increase in **OverTime** will increase the odds of employee leaving the company by 373.59%
- ☐ Holding other variables constant, a unit increase in **WorkLifeBalance** will decrease the odds of employee leaving the company by20.61%
- ☐ Holding other variables constant, a unit increase in **YearsInCurrentRole** will decrease the odds of employee leaving the company by 15.05%
- ☐ Holding other variables constant, a unit increase in **YearsSinceLastPromotion** will increase the odds of employee leaving the company by 15.76%

4.2. Hugo is specifically interested in understanding how the following aspects drive employee attrition.

a) Medium satisfaction level with their working environment and job, and 5 years since their last promotion
b) Number of years in current roles and whether they work overtime
c) 45 years old married employee with a very-high level job classification and maintaining a good work-life balance.

In order to gain an edge in the current very competitive talent market, Hugo believes attaining a very good understanding in what drives employees to quit is well worth the time and investment. In addition, TPM should take prompt actions to mitigate increasingly high employee turnover costs which could be up to twice an employee's salary depending on their position. Accordingly, your job is to visualize the predicted likelihood of employee attrition with the specific attributes described above.

To predict the likelihood of employee leaving TMP based on the preference given by Mr. Hugo, I created another model which included only the intended independent variables, namely **EnvironmentSatisfaction**, **JobSatisfaction**, **YearsSinceLastPromotion**, **YearsInCurrentRole**, **OverTime**, **Age**, **MaritalStatus**, **JobLevel**, and **WorkLifeBalance**. However, **JobStatisfaction** was not significant enough (p-value of 0.08) to be part of the model. Therefore, I removed it from the group of independent variables. Finally, this model was exactly the same as the one I built previously at 4.1.

To interpret the results from this model, I calculated the cut off percentage which I believed to be reasonable enough to predict whether an employee would be leaving TPM. I decided to take the standard rule of thumb which required the cut off value to be the combination of the probability by chance plus an additional one-fourth of that probability. Since the proportion of **Attrition** data had no balance, I picked **Proportional Chance Criterion** (PCC) strategy to locate the best probability by chance. As the result, PCC

produced the probability of 72.95%. Based on this, the cut off probability went up to (72.95% + 18.24%) 91.19%.

I experimented the model with the given setting from Mr. Hugo as follow:

EnvironmentSatisfaction	=	2.5 (Medium)
YearsSinceLastPromotion	=	5
Age	=	45
Dummy_MaritalStatus	=	2 (Married)
JobLevel	=	4 (High)
WorkLifeBalance	=	2 (Good)

Plus all values of YearsInCurrentRole and OverTime.

The results were obtained and separated into two lists of values which I presented in **Table 7**. Further, I also visually plot the results using line charts for a better insight (see **Figure 7**). As seen in the table and figure, the employee who worked overtime were more likely to leave the company than those who did not, given that Mr. Hugo's criteria above were constant. It was also noticeable that the more recent the employee was hired, the more likely that they would leave the job. Another interesting evidence was that when the employees had stayed longer with TPM, those who worked overtime had a very low probability of leaving the job which was very similar to those who did not work overtime. But this similarity faded with the decreasing amount of years in their current roles. In conclusion, the probability of leaving TPM was too low (from 2.6% to 33.3%, in respect to the increasing order of **YearsInCurrentRole**, for those who worked overtime and 0.6% to 9.5% for those who did not work overtime) compared to the cut off probability of 91.19%. In other words, the employees (under the above scenario) would not, in any regular circumstances, be the potential person to leave the company.

ITEM 5:

5. *Develop a time-series model to forecast AP's energy consumption for the next 12 months. How are summer predictions different from those for winter?*

I'd looked at the visuality of the data and I found that the data contained the combination of a linear trend and seasonal variations. In addition, what we were trying to predict was a 12-months period which I classified it a long-term prediction based on monthly data. Therefore, I decided to build a linear trend-based forecasting model to predict the amount of energy consumption of AP. In this model, I used 7-moving average to smooth the data points. I believed this 7-moving average was suitable with our large amount of data points (since we could afford losing some data points) in order to get a better smooth trend and eliminate the effects of seasonality from our model prediction.

After deseasonalizing the data, I found that the energy consumption had a negative linear trend which the amount of energy consumption tended to reduce over time in a constant manner (see **Figure 8**). From the linear equation, it showed that every increase in one unit of time period (1 month), the energy consumption would decrease approximately 10,957 megawatts. There would also be fluctuations in regard to seasonal effects. For instance, there would be additional demand of 7% in Summer and 6% in Winter of energy consumption in average. However, these demands dropped down during Autumn and Spring by 6% and 7% less than the average respectively.

In our forecast for 2020, the need of energy demand for the coming Summer would be about 11,252,350 megawatts and another 11,152,605 megawatts for Winter (see **Figure 9**). In other words, an additional power of 665,532 megawatts for Winter and 765,277 megawatts for Summer would be required on the top of the average (of 10,487,073 megawatts).

Summary:

So far, all aspects of the meeting minutes have been covered within this report. First of all, detailed descriptions of the two important variables, **PercentSalaryHike** and **Attrition**, were given. This section is important as it gives foundation knowledge for other following sections. Next the report reveals the relationships between all independent variables and **PercentSalaryHike**, and produces linear regression model which can explain the significant factors that contributed to the percentage increase in salary. The regularity of the model was also been inspected under the visuality of the residual plots. A quadratic relationship was found between **PerformanceRating** and **PercentSalaryHike** after spot checking of the residual plots which gave a reform of the model to have a better predicting reliability. Another evidence was also found that **JobSatisfaction** had an indirect effect on **PercentSalaryHike** in addition to the main effect of **PerformanceRating**. In other word, there was an interacting effect of **JobSatisfaction** and **PerformanceRating** over **PercentSalaryHike**.

Factors which contributed to **Attrition** were also studied by using Logistic Regression model. Within this context, the model suggested by Mr. Hugo was refined based on the improvement of the statistical evidences. The final model was built to contain eight important independent variables instead of four independent variables suggested earlier. The same model was then used to predict the likelihood of the employees leaving TPM under a specified setting provided by Mr. Hugo. As yielded by the model, none of the specified employees were likely to leave the company, given that no unknown factors occurred.

Lastly, this report also details how a time-series model was built and reveals the predicting results of energy consumption of AP for 2020. As explained by the model, energy consumption decreased over time and there were variations between seasons. Summer was the highest season of energy consumption and followed by 1% of the average lower in Winter. Spring had the lowest energy consumption but not far away from Autumn figure.

Informatively, all predictions were conducted to a certain level of confidence and there is always a chance that an unaccounted factor or sampling errors may influence the occurrence of the reality. Furthermore, the dataset for studying **Attrition** was insufficient to bring the Logistic Regression model to a promising level of reliability. In my suggestion, a better balance of **Attrition** values is crucial should the practicality of model is sought.

The end

Appendices:

Table 1. *Count of each observation of PercentSalaryHike*

Row Labels	Count of PercentSalaryHike	
11	14.29%	
12	13.47%	
13	14.22%	
14	13.67%	sub total:
15	6.87%	62.52%
16	5.31%	
17	5.58%	
18	6.05%	
19	5.17%	
20	3.74%	
21	3.27%	
22	3.81%	
23	1.90%	
24	1.43%	
25	1.22%	
Grand Total	**100.00%**	

Table 2. *Correlation table of all independent variables and PercentSalaryHike*

	Age	Dummy_Attrition	Education	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel	JobSatisfaction	Dummy_MaritalStatus	NumCompaniesWorked	Dummy_OverTime	PerformanceRating	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	PercentSalaryHike
Age	1																
Dummy_Attrition	-0.15921	1															
Education	0.04752	0.01149	1														
EnvironmentSatisfaction	0.01015	-0.10337	-0.00754	1													
Gender	0.00035	0.01259	-0.0473	-0.031	1												
JobInvolvement	-0.01183	-0.04113	0.00238	-0.0095	0.33093	1											
JobLevel	0.5096	-0.1691	0.03179	0.00121	-0.01402	-0.05606	1										
JobSatisfaction	0.00888	-0.04119	-0.038	-0.04174	0.36158	0.3907	-0.02301	1									
Dummy_MaritalStatus	-0.09503	0.16207	0.00962	-0.00359	-0.02196	-0.00694	-0.07677	0.01967	1								
NumCompaniesWorked	0.07866	0.02674	0.05222	0.01345	-0.28891	-0.35116	0.03604	-0.38203	0.00085	1							
Dummy_OverTime	0.02806	0.24612	-0.06563	0.07013	-0.02601	0.00838	0.00054	-0.01187	-0.01752	-0.01547	1						
PerformanceRating	-0.00687	0.00557	-0.04805	-0.03536	0.34808	0.33595	-0.01041	0.42164	-0.00596	-0.24505	0.00625	1					
WorkLifeBalance	-0.02149	-0.06394	-0.02017	0.02763	0.02114	-0.01934	0.03782	0.02513	0.01471	0.01432	-0.02709	0.0029	1				
YearsAtCompany	0.31131	-0.13439	-0.00607	0.00146	0.00438	-0.06846	0.53474	-0.03088	-0.05999	-0.00183	-0.01169	0.0033	0.01209	1			
YearsInCurrentRole	0.2129	-0.16055	0.0484	0.01801	0.01606	-0.03037	0.38945	-0.0029	-0.06582	-0.00221	-0.02976	0.03307	0.04986	0.75875	1		
YearsSinceLastPromotion	0.21651	-0.03302	0.00826	0.01619	-0.00695	-0.0653	0.35389	-0.00834	-0.03092	0.00971	-0.01224	0.012	0.00894	0.61841	0.54806	1	
PercentSalaryHike	0.00363	-0.01348	-0.04893	-0.0317	0.50325	0.55039	-0.03473	0.67446	0.01249	-0.51057	-0.00543	0.76477	-0.00328	-0.03599	-0.00152	-0.02215	1

Table 3. *Summary Output of the final linear regression model*

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.909229							
R Square	0.826698							
Adjusted R Square	0.825868							
Standard Error	1.527259							
Observations	1470							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	7	16267.32	2323.903	996.3053	0			
Residual	1462	3410.146	2.332521					
Total	1469	19677.47						
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-4.124	0.393	-10.483	0.000	-4.896	-3.353	-4.896	-3.353
Age	0.011	0.005	2.403	0.016	0.002	0.020	0.002	0.020
Gender	0.842	0.092	9.134	0.000	0.661	1.023	0.661	1.023
JobInvolvement	0.633	0.050	12.766	0.000	0.536	0.730	0.536	0.730
JobSatisfaction	0.932	0.042	22.164	0.000	0.849	1.014	0.849	1.014
NumCompaniesWorked	-0.280	0.018	-15.185	0.000	-0.317	-0.244	-0.317	-0.244
PerformanceRating	4.806	0.121	39.749	0.000	4.568	5.043	4.568	5.043
YearsAtCompany	-0.017	0.007	-2.403	0.016	-0.030	-0.003	-0.030	-0.003

Table 4. *Summary Output of the final Polynomial Order 2 model with additional PerformaceRating^2*

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.915856					
R Square	0.838793					
Adjusted R Square	0.83791					
Standard Error	1.473505					
Observations	1470					
ANOVA						
	df	*SS*	*MS*	*F*	*gnificance F*	
Regression	8	16505.32	2063.165	950.2338	0	
Residual	1461	3172.149	2.171218			
Total	1469	19677.47				
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	19.394	2.278	8.513	0.000	14.925	23.862
Age	0.010	0.004	2.161	0.031	0.001	0.018
Gender	0.803	0.089	9.021	0.000	0.629	0.978
JobInvolvement	0.601	0.048	12.541	0.000	0.507	0.695
JobSatisfaction	0.935	0.041	23.049	0.000	0.855	1.014
NumCompaniesWorked	-0.281	0.018	-15.763	0.000	-0.316	-0.246
PerformanceRating	-9.168	1.340	-6.843	0.000	-11.796	-6.540
PerformanceRating ^2	2.052	0.196	10.470	0.000	1.668	2.437
YearsAtCompany	-0.016	0.007	-2.444	0.015	-0.029	-0.003

Table 5. *Summary output of Logistic Regression model by Mr. Hugo*

LL0	-649.291		Classification Table				
LL1	-561.981						
				Suc-Obs	Fail-Obs		
Chi-Sq	174.6209		Suc-Pred	36	14	50	
df	4		Fail-Pred	201	1219	1420	
p-value	0.000			237	1233	1470	
alpha	0.05						
sig	yes		Accuracy	0.151899	0.988646	0.853741	
R-Sq (L)	0.13447		Cutoff	0.5			
R-Sq (CS)	0.112005						
R-Sq (N)	0.190933						
	coeff b	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
Intercept	-0.696	0.397	3.067	0.080	0.499		
Age	-0.048	0.009	28.067	0.000	0.953	0.937	0.970
EnvironmentSatisfaction	-0.324	0.070	21.655	0.000	0.723	0.631	0.829
Dummy_OverTime	1.481	0.156	90.392	0.000	4.396	3.239	5.964
YearsInCurrentRole	-0.123	0.026	22.777	0.000	0.885	0.841	0.930

Table 6. *Summary output of Logistic Regression of the refined model*

LL0	-649.291		Classification Table				
LL1	-529.314						
				Suc-Obs	Fail-Obs		
Chi-Sq	239.9552		Suc-Pred	51	24	75	
df	8		Fail-Pred	186	1209	1395	
p-value	0.000			237	1233	1470	
alpha	0.05						
sig	yes		Accuracy	0.21519	0.980535	0.857143	
R-Sq (L)	0.184782		Cutoff	0.5			
R-Sq (CS)	0.150608						
R-Sq (N)	0.256739						
	coeff b	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
Intercept	-1.576	0.584	7.283	0.007	0.207		
Age	-0.031	0.010	9.324	0.002	0.969	0.950	0.989
EnvironmentSatisfaction	-0.351	0.072	23.593	0.000	0.704	0.611	0.811
JobLevel	-0.376	0.105	12.858	0.000	0.687	0.559	0.843
Dummy_MaritalStatus	0.642	0.114	31.941	0.000	1.900	1.521	2.374
Dummy_OverTime	1.555	0.162	92.721	0.000	4.736	3.451	6.499
WorkLifeBalance	-0.231	0.109	4.524	0.033	0.794	0.642	0.982
YearsInCurrentRole	-0.163	0.033	25.055	0.000	0.850	0.797	0.906
YearsSinceLastPromotion	0.146	0.033	19.491	0.000	1.158	1.085	1.235

Table 7. *Lists of experimental results of Logistic Regression model with the intended variables by Mr. Hugo.*

Without Over Time = 1			
YearsInCurrentRole	Logit	Odds	P(y)
0	-2.249	0.105	9.54%
1	-2.412	0.090	8.22%
2	-2.575	0.076	7.08%
3	-2.738	0.065	6.08%
4	-2.901	0.055	5.21%
5	-3.064	0.047	4.46%
6	-3.227	0.040	3.81%
7	-3.390	0.034	3.26%
8	-3.554	0.029	2.78%
9	-3.717	0.024	2.37%
10	-3.880	0.021	2.02%
11	-4.043	0.018	1.72%
12	-4.206	0.015	1.47%
13	-4.369	0.013	1.25%
14	-4.532	0.011	1.06%
15	-4.695	0.009	0.91%
16	-4.858	0.008	0.77%
17	-5.021	0.007	0.66%
18	-5.184	0.006	0.56%

With Over Time = 2			
YearsInCurrentRole	Logit	Odds	P(y)
0	-0.694	0.500	33.32%
1	-0.857	0.424	29.80%
2	-1.020	0.361	26.50%
3	-1.183	0.306	23.45%
4	-1.346	0.260	20.65%
5	-1.509	0.221	18.11%
6	-1.672	0.188	15.81%
7	-1.835	0.160	13.76%
8	-1.998	0.136	11.94%
9	-2.161	0.115	10.33%
10	-2.324	0.098	8.91%
11	-2.487	0.083	7.67%
12	-2.651	0.071	6.60%
13	-2.814	0.060	5.66%
14	-2.977	0.051	4.85%
15	-3.140	0.043	4.15%
16	-3.303	0.037	3.55%
17	-3.466	0.031	3.03%
18	-3.629	0.027	2.59%

Figure 1. *Histogram of PercentSalaryHike*

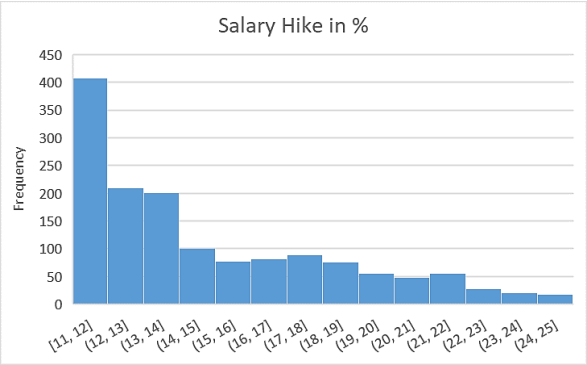

Figure 2. *The proportion of employees who left and remained working*

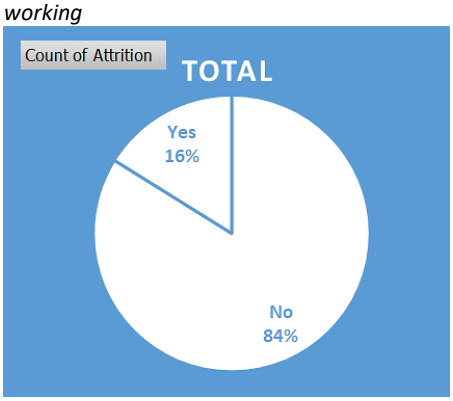

Figure 3. *Quadratic relationship between PerformanceRating and PercentSalaryHike*

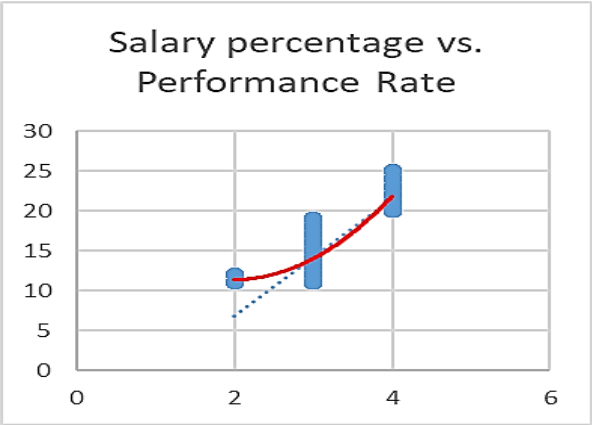

Figure 4. *Residual plot of PerformanceRating with Linear Regression model (with irregularity)*

Figure 5. *Residual plot of PerformanceRating with Polynomial Order 2 model*

Figure 6. *Interaction effect between JobSatisfaction and PerformanceRating over PercentSalaryHike*

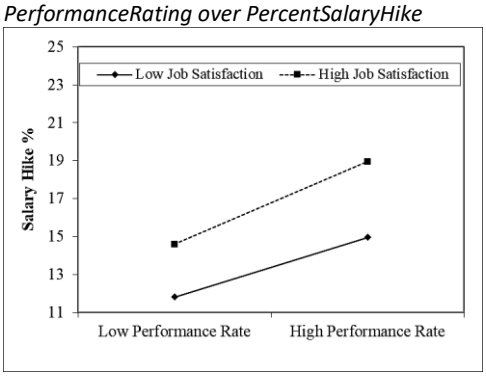

Figure 7. *Line plot presenting the results of Logistic Regression model with intended variables by Mr. Hugo*

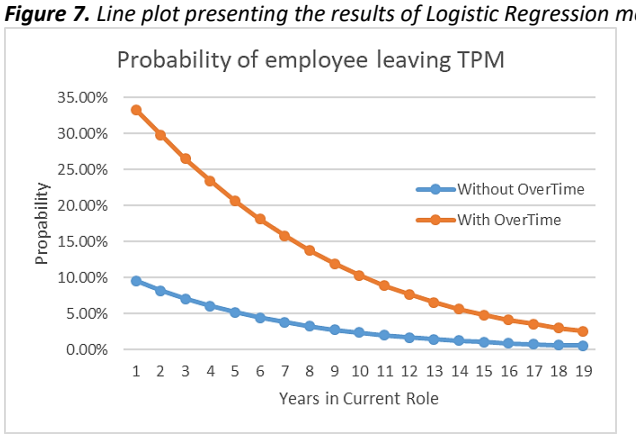

Figure 8. *Negative linear trend of energy consumption and the predictions*

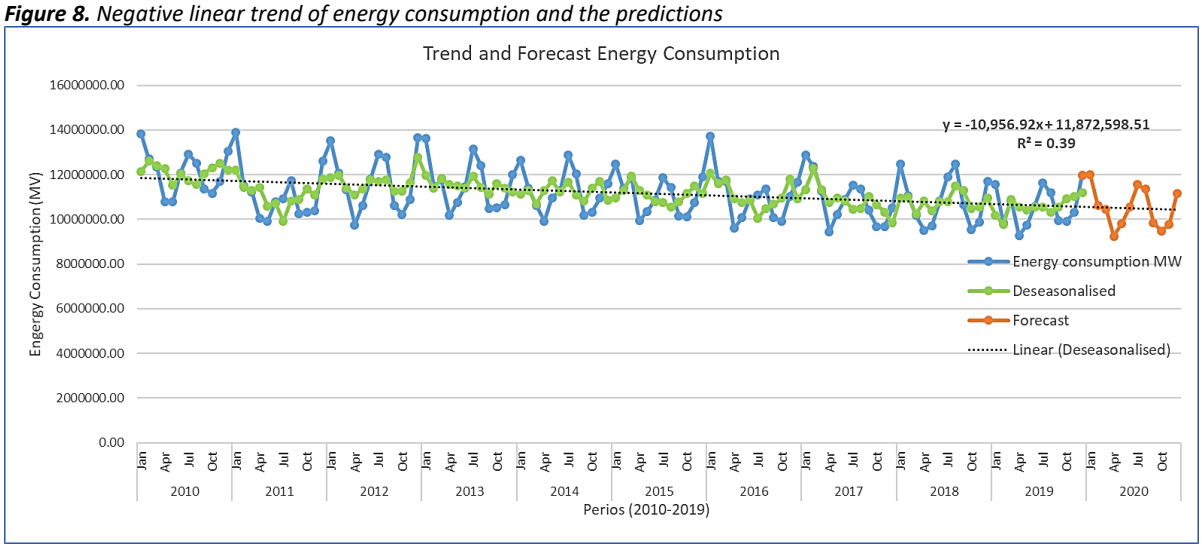

Figure 9. *Prediction of 2020 energy consumption by seasons*