Assignment Submission Coversheet

Faculty of Business and Law

Student ID: 220129254

Student Name: Alina Kriutchenko

Campus: ☒ Burwood ☐ Waterfront ☐ Waurn Ponds ☐ Warrnambool ☐ Cloud

Assignment Title: Assignment two

Due Date: Thursday 17th September 2020, 8:00 pm

Course Code/Name: Master of Data Science

Unit Code/Name: MIS770

Unit Chair/Campus Coordinator: Dr Lasitha Dharmasena

Introduction

This report includes the overall summary of the following variables: percentage increase in salary and attrition. Identify three most influential variables for the percent salary hike; its form is relationships and potential multi-collinearity problem. Also built the regression model to estimate the salary increase and perform residual analysis.

It has a model to test whether the employee satisfaction with his job makes the relationship between the performance rating and salary increase stronger.

It states whether this model is statistically significant and define the likelihood of an employee leaving the TPM. First, in the situation with the medium satisfaction level with their working environment and job and 5 years since the last promotion. The second situation is the number of years in current roles and whether they work overtime. The third situation is 45 years old, married employee with a very-high level job classification and maintaining a good work-life balance. In addition, there is a visualization of these likelihoods.

It has a time-series model to forecast AP's energy consumption for the next 12 months and investigate how the summer predictions different from the winter predictions.

The aim of this analysis is to identify the interaction between percentage salary increase and other various variables. Based on these findings the company will improve the employee’s condition and optimize the work of the human resources department in the future. Also with the second dataset it investigates the future AP's energy consumption.

The report contains the detailed results of the analysis and methods used for it. In addition, the limitations of this analysis discussed. The report structure is introduction, the main body, conclusion and appendixes.

Question 1.2

The overall summary of the salary increase shows that the hike between 11-14 percentage (image 1) is more common. The percentage increase in salary has the mean of 15.21%, the mode of 11%, the minimum of the 11%, the maximum of the 25% and the standard deviation of the 3.6. The number of employees leaving the company is 16% of the total sample (image 2).

Question 2.1

This question covers the identification of the independent variables performed through visual scatter plots. It showed that there are interactions between the percent salary hike and five variables. These variables influence the percentage increase in salary with the positive relationships: performance rating, job satisfaction, gender and job involvement (image 3). The negative relationship is with the variable named number of companies, which means that the percentage salary hike decrease with the every next company (image 4). The method used to answer this question is multiple regression.

Three independent variables that have the more impactful linear relationship with Percent Salary Hike are performance rating, job satisfaction and job involvement (image 5). For every performance unit goes 15 percent salary increase. The form of relationship of independent variables is strong positive (0,77) for the performance rating, moderate positive (0.76) for the job satisfaction and moderate positive (0,55) for the job involvement retrospectively. The maximum correlation is 0.76 between years at company and years in the current role. There are no other variables with the correlation over 0.8. Therefore, this model does not have any potential multi-collinearity problem.

Question 2.2

The next step is to build a regression model to estimate the percentage increase in salary. The dependent variable is percent salary hike (% salary) or increase in salary. The independent variables are the three variables mentioned above.

The Null hypothesis states that there is no linear relationship between variables, while the alternative hypothesis states that at least one independent variable affects dependent variable. Since the p-value is less than 0.5 for intercept, we reject the null hypothesis. Therefore, it can be stated that at least one independent variable explains significant variation in the dependent variable and has a predictive power.

The p-value for the performance rating, job satisfaction and job involvement is equal to zero and less than alpha. It means that all independent variables are statistically significant in this model. Therefore, we can conclude that performance rating, job satisfaction and job involvement produce a significant amount of variation on dependent variable and have a predictive power.

The R Square is 0.78 or 78% of variation in the percent salary hike. The adjusted R square is also 78 percentage. The multiple R is 0.88%. The variation pick up by the model is 15320.39 and the variation do not picked up by the model is 4357.08. The total variation is 19677.47. Therefore, about 22 percent of variation does not catch by the model. The model has 1470 observations. The 78% of the number of the independent variables on sample size. The

standard error is 1.72, which is less than two. Therefore, it means that the fit (goodness) is good for this model.

Coefficient for the percentage salary intake on average -5.72 points. The average increase is 5.17 for each additional unit of performance rating, 1.11 for each additional unit of the job satisfaction and 0.91 for each additional unit of the job involvement. The intercept's standard error is 0.38. The standard error is 0.13 for the performance rating, 0.04 for the job satisfaction and 0.05 for the job involvement.

Question 2.3

This question covers the performance of residual analysis. There is no obvious pattern in the residual plots, which means that it fits relatively well (image 6). The normal probability plot is close to the straight line, which is good for the model. It is not far away from perfectly normal (image 7). Therefore, residuals plotting area approximately normal. The residual output shows that there are some potential outliers in the residuals, which are more than two standard errors away. Therefore, these numbers, which are bigger than 3.44 highlighted.

Question 3

The next question considers whether the relationship between performance rating and percentage increase in salary should be stronger for employees who are satisfied with their jobs. The method used to answer this question is multiple regression. In the first model the dependant variable is percentage salary hike again and the independent variables are job satisfaction square, job satisfaction and performance rating.

The R Square is 0.75 or 75% of variation in the percent salary hike. Compared to the regression model developed in Q2, R^2 decreased. This indicates a slight decrease of improvement in the explanatory (predictive) power of the regression model. The adjusted R square is 75 %. The multiple R is 0.86%. The variation picked up by the model is 14676.88 and the variation do not pick up by the model is 5000.59. The total variation is 19677.47. Therefore, about 25.4 percent of variation does not caught by the model. This is slightly higher than in the previous model. The number of observations is 1470 as well. The 75% is the number of the independent variables in sample size. The standard error is 1.82, which is less than two. Therefore, it means that the fit (goodness) is good for this model.

All other predictors held constant, for an employee with zero job satisfaction and performance rating the average salary increase expected to be -6.65 points. All other predictors held constant, the average increase is -0.18 for each additional unit of job satisfaction square, 2.27 for each additional unit of the job satisfaction and 5.60 for each additional unit of the performance rating. The intercept's standard error is 0.42. The standard error is 0.02 for the job satisfaction square, 0.13 for the job satisfaction and 0.14 for the performance rating.

The p-value is less than 0.05 and we Reject Ho. Overall, the model has some predictive power. All individual variables are individual variables have significance since the p-values is less than 0.055.

The regression equation for this model:

$$y = -6.65 - 0.18x1^2 + 2.27x2 + 5.60x3$$

For finding an answer, we can choose the medium performance rating level and test it with every job satisfaction level. The salary hike in this case has an obvious relationship with the

job satisfaction. Hugo stated in his assumption than the percent salary increase is stronger for the employee who are satisfied with their jobs. For the same level of performance rating, the salary hike is higher with every higher unit of the job satisfaction (from 6.54 to 10.75%) (image 8, image 9). Therefore, the relationship is stronger with the highest level of job satisfaction. The normal probability plot is curvier, which shows less linear relationship.

The second model (List Q3 (b)) is a comparison of four different models. These four models created based on the job satisfaction levels low, medium, high, very high. Based on the Multiple R we can judge how strong linear relationships for every model.

The model with the low jobs satisfaction (image 10) has a multiple R equal to 0.22, the model with the medium jobs satisfaction (image 11) has a multiple R equal to 0.52, the model with the high jobs satisfaction (image 12) has a multiple R equal to 0.72 and the model with the very high jobs satisfaction (image 13) has a multiple R equal to 0.86.

With the low job satisfaction, the R square of 5% means that 5% of the variation of percent salary hike around the mean is explained by the Performance rating. For the medium job satisfaction, the R square of 27% means that 27% of the variation of percent salary hike around the mean is explained by the Performance rating. For the high job satisfaction, the R square of 53% means that 53% of the variation of percent salary hike around the mean is explained by the Performance rating. For the very high job satisfaction, the R square of 73% means that 73% of the variation of percent salary hike around the mean is explained by the Performance rating.

Therefore, the Hugo assumption that the stronger relationship between the performance rating and percent salary high is stronger with the higher job satisfaction level is true.

Question 4.2 (a)

The question 4.1 covers the medium satisfaction level with their working environment and job, and 5 years since their last promotion results in 0.18 probability of an employee leaving the company. The hit ratio is 18%. The method used to answer this question is logistic regression.

The logistic logarithmic equation. These coefficients are logarithmic terms; it is a straight-line equation. The direction: the age, the environment satisfaction and year in current role are decreasing, while the overtime is increasing. Therefore, the younger employees are slightly more likely to leave the job. The lower environment satisfaction has a high effect and year in the current role has a noticeable effect. The increasing overtime also make people to leave their jobs. The exponential regression coefficiency exp(b) demonstrate the positive outcome for the overtime, where it is greater than one.

The higher wald, the lover p-value. The wald is pretty significant. However, possibly is less significant statistically because of the high square error. Possibly, there are some outliers in the data and we need to investigate the data and remove some values.

The percentage of change in Odds: One unit of decrease in attrition decrease the likelihood (odds) of average age by 4.8%. One unit of decrease in attrition decrease the likelihood (odds) of environment satisfaction by 27.7%. One unit of decrease in attrition decrease the likelihood (odds) of average years in the current job by 17,55%. One unit of decrease in attrition increase the likelihood (odds) of overtime by 339.55%.

The LL0 equal -649.291 is far away from the zero. The new LL1 with more information being explained has more variation in deviance being picked up as expected. It is slightly closer to zero (equals to -639.235).

The likelihood ration (improvement in deviance) is 20.11. The p-value is less than alpha and this is a statistically significant result. The new model with four variables predicts the outcome significantly better than the baseline model. The R square measures indicate that less than 1% of variation between two groups can be explained by the logistic regression model.

The 101 employees who actually left the company as it was predicted, but it was predicted for 136. Of the 904 employees who do not leave the company as it was predicted.

Overall, the percentage of successfully predicted people who left the company is 42.6% and the percentage of successfully predicted people who does not leave the company is 73.3%. The total classification accuracy of the model's predictions is 68.4%.

The area under the curve (AUC) (image 13) is not so close to the 1 and shows show that the data fit is pretty average.

The conditions are medium job satisfaction, medium working environment and five years since last promotion. The logit is equal to -1.53, the odds equal to 0.22.

Therefore, the probability of the employee leaving the job while having the medium satisfaction level with their working environment and job, and 5 years since their last promotion is 18%.

Question 4.2 (b)

The question 4.2 investigates whether with the increase in the number of years in current roles and while working overtime, the probability of an employee leaving the company decreases. The hit ratio is 18%.

The LL0 equal -649.291 is pretty far away from the zero. The new LL1 with more information being explained and more variation in deviance being picked up as expected and it is slightly mode closer to zero (equals to -588.059). The 120 employees who actually left the company as it was predicted, but it was predicted for 117. Out of the 990 employees who do not leave the company it was predicted.

Overall, the percentage of successfully predicted people who left the company is 50.6% and the percentage of successfully predicted people who does not leave the company is 80.3%. The total classification accuracy of the model's predictions is 75.5%. The area under the curve (AUC) (image 14) is not so close to the 1, but the data fit is better than in the previous model (image).

The conditions are working overtime and the number of years in current role. The logit is equal to -0.43, the odds equal to 0.65. Therefore, the probability of the employee leaving the job while working overtime and 1 year since their last promotion is 39%. It slowly goes down, stops at 7-8 years in the current role and then continue to go down 5% probability when "years in current role" reach 18 years (image 15). The relationships form is negative.

Question 4.2 (c)

The question 4.1 covers whether the 45 years old, married employee with a very-high level job classification and maintaining a good work-life balance has a probability of leaving the company. The hit ratio is 21%.

The LL0 equal -649.291 is pretty far away from the zero. The new LL1 with more information being explained and more variation in deviance being picked up as expected and it is slightly mode closer to zero (equals to -601.087). The 120 employees who actually left the company as it was predicted, but it was predicted for 117. Out of the 950 employees who do not leave the company it was predicted.

Overall, the percentage of successfully predicted people who left the company is 50.6% and the percentage of successfully predicted people who does not left the company is 77.1%. The total classification accuracy of the model's predictions is 72,8%. The area under the curve (AUC) (image 16 is not so close to the 1 and the data fit is average (image). To answer these questions I used the logistic regression method.

The probability of 45 years old, married employee with a very-high level job classification and maintaining a good work-life balance leaving the company is 7%.

Question 5

In the question Q5, I did a time-series model to forecast AP's energy consumption for the next 12 months. The method used to answer this question is time series and forecasting. These predictions are: 10086427.74 in January 2020, 8880201.33 in February 2020, 8707354.652 in March 2020, 7576595.941 in April 2020, 7977033.59 in May 2020, 8587131.262 in June 2020, 9290200.654 in July 2020, 9205144.57 in August 2020, 8024482.387 in September 2020, 7814182.199 in October 2020, 8157704.677 in November 2020 and 9240040.476 in December 2020 (image 17).

Through the years the average energy consumption during the summer was 8.7% higher than the monthly average. In particular, in December the energy consumption was 7.8% higher, in January 16% higher, in February 2.23% higher (image 18). Overall, the consumption was lower during the spring (93,4%) and autumn (93,12).

In 2020, the energy consumption during the summer months expected to be slightly higher than in the winter months. In particular, the consumption in December, January and February is 28% of the total yearly consumption. While the consumption in the June, July and August is 26% of the total consumption. Moreover, energy consumption during the autumn or spring is only 23% of the total consumption.

The forecast for the next 12 months has a downward trend as in the previous years. In 2020, the energy consumption during the summer months expected to be slightly higher than in the winter months (image 19). The MAPE is 23,03% and historical MAPE is 17%.

To answer his question I used forecasting quantitative time series method with simple moving average and smoothing.

Conclusion

There are five possible variables that influence the percentage increase in salary and three the most impactful of them are performance rating, job satisfaction and job involvement with from medium to strong positive form of relationship. There is no a potential multi-collinearity problem. The percentage increase in salary is 5.17 for each additional unit of performance rating, 1.11 for each additional unit of the job satisfaction and 0.91 for each additional unit of the job involvement. The residual analysis is satisfactory.

The Hugo assumption that the stronger relationship between the performance rating and percent salary high is stronger with the higher job satisfaction level is true and the model is statistically significant.

The probability of the employee leaving the job while having the medium satisfaction level with their working environment and job, and 5 years since their last promotion is 18%.

The probability of the employee leaving the job while working overtime and 1 year since their last promotion is 39%. It slowly goes down, stops at 7-8 years in the current role and then continue to go down 5% probability when “years in current role” reach 18 years. The relationships form is negative.

The probability of 45 years old, married employee with a very-high level job classification and maintaining a good work-life balance leaving the company is 7%.

The forecast for the next 12 months has a downward trend as in the previous years. In 2020, the energy consumption during the summer months expected to be slightly higher than in the winter months (image 18).

One of limitation of this analysis is a small dataset of energy consumption, which covers information only of the last decade. The second limitation is lack of the data about employee’s living area due to privacy. It could improve the attrition prediction.

Appendices

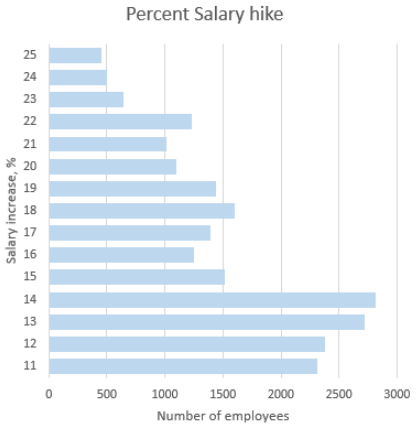

Image 1. Percent Salary Hike

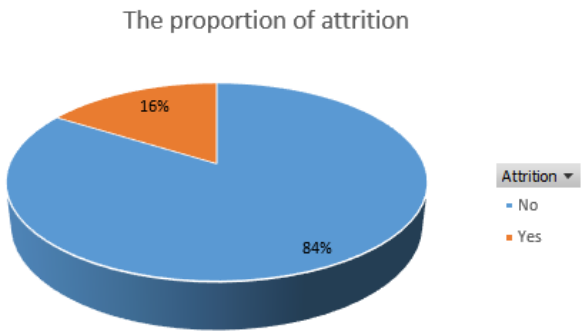

Image 2. The proportion of attrition

Image 3. Relationships of Performance salary Hike

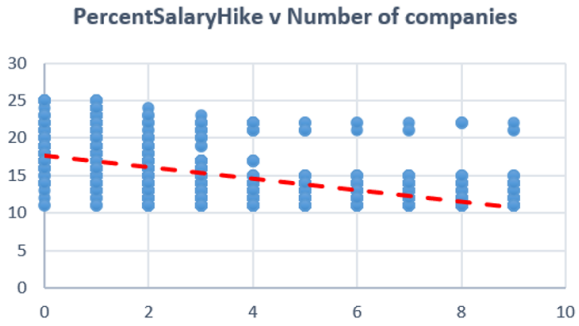

Image 4. Relationship of Performance salary Hike

Image 5. Relationships of Performance salary Hike

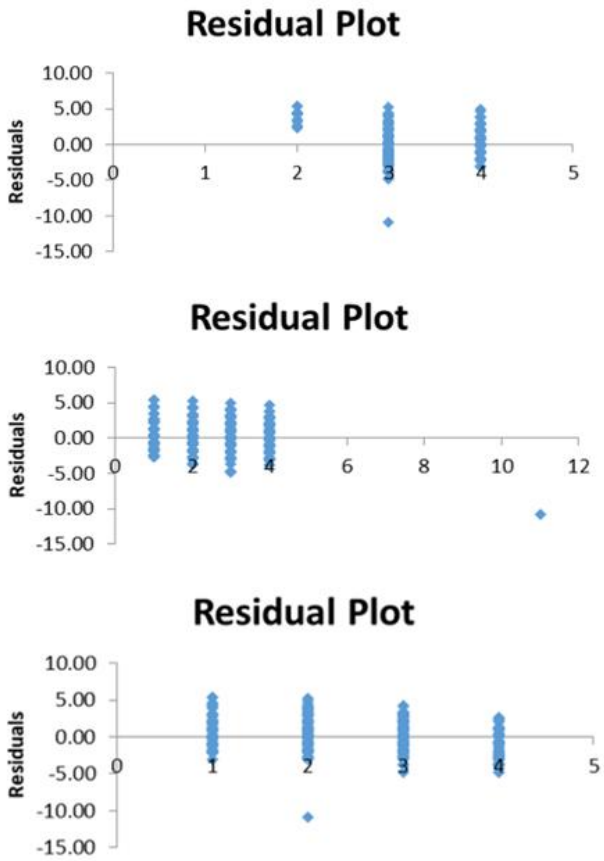

Image 6. Residual plots

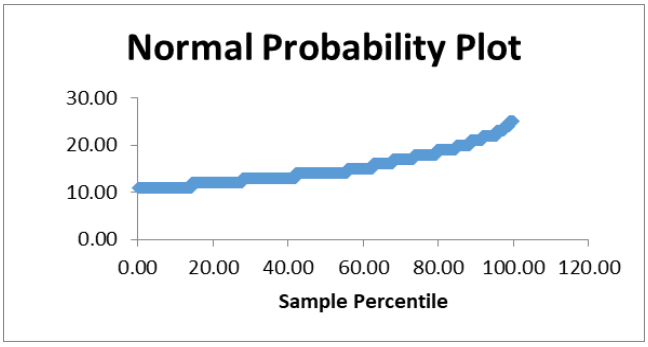

Image 7. Normal Probability plot

SalaryHike	JobSatisfaction	PerformanceRating
6,64	1	2
8,37	2	2
9,74	3	2
10,75	4	2

Image 8. Variation of salary hike by Job satisfaction level

Image 9. Variation of salary hike by Job satisfaction level

Image 10. Salary Hike v Performance rating with Low job satisfaction

Image 11. Salary Hike v Performance rating with Medium job satisfaction

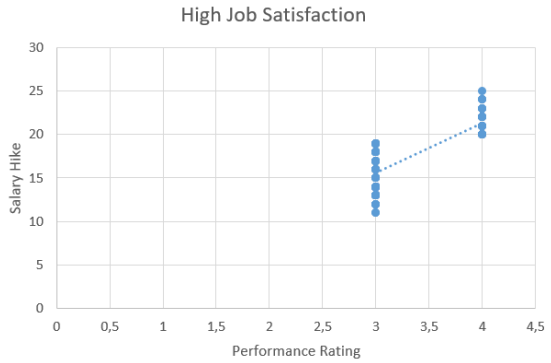

Image 12. Salary Hike v Performance rating with High job satisfaction

Image 13 Salary Hike v Performance rating with Very High job satisfaction

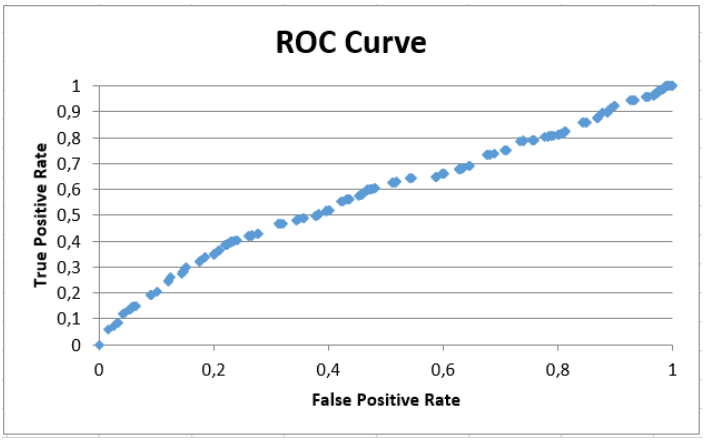

Image 13. ROC curve (3.1 a)

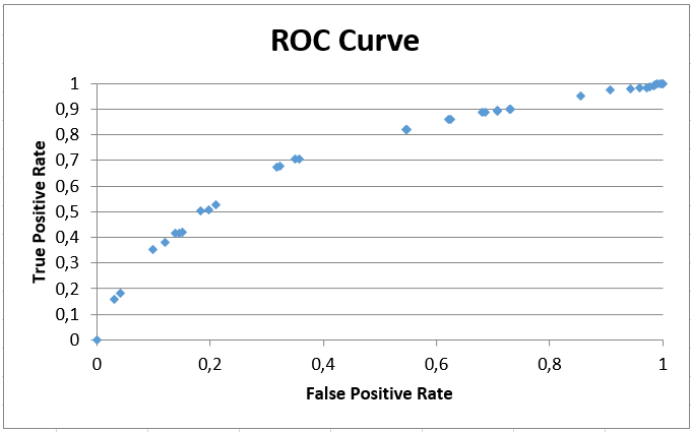

Image 14. ROC curve (3.1 b)

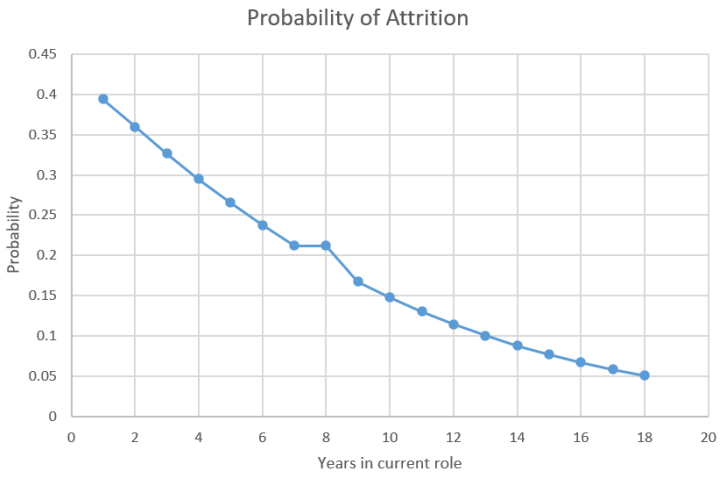

Image 15. Probability of attrition

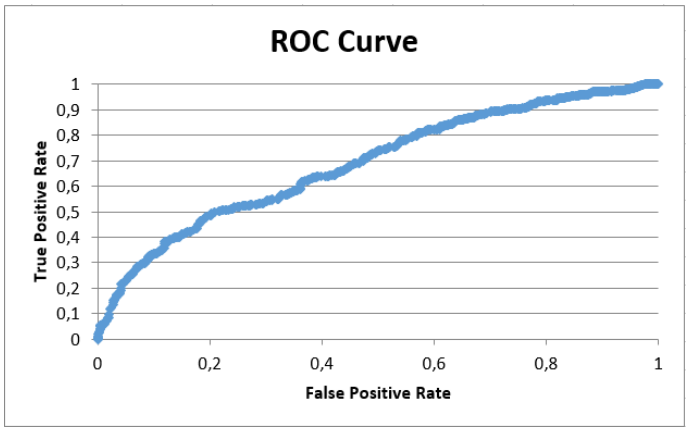

Image 16. ROC curve (3.1 c)

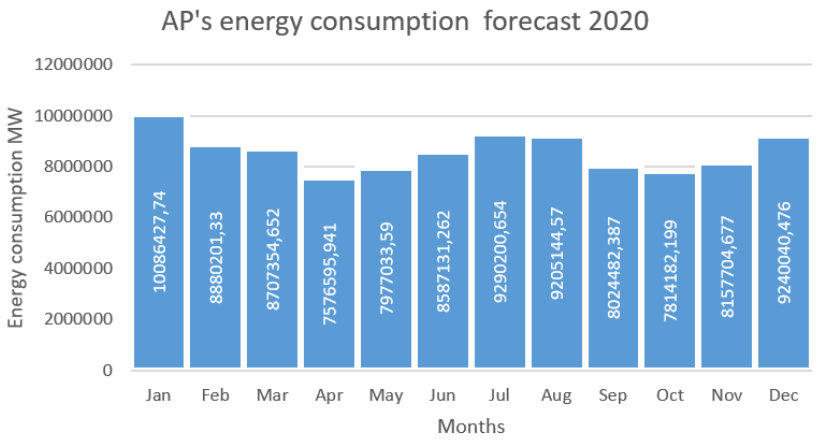

Image 17. Energy consumption forecast 2020, MV

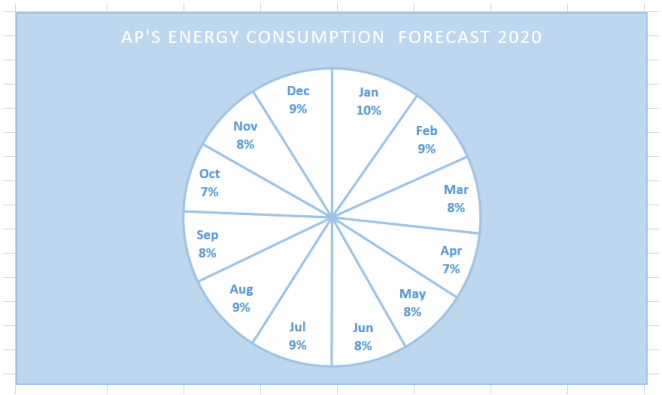

Image 18. Energy consumption forecast 2020, %

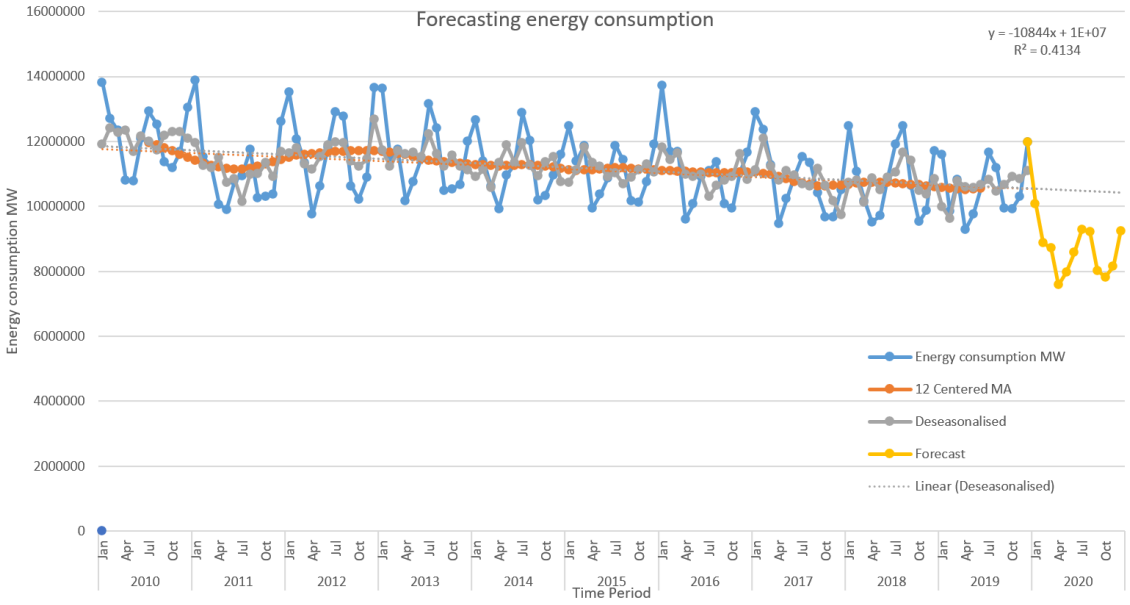

Image 19. Forecasting energy consumption