ANALYSIS ON EMPLOYEE SALARY & ATTRITION

BY NIKHITA KANDIKONDA and VENKATA TAMVADA

I. Introduction

Data Scientists at IBM corporation designed a data set to represent human resource data from a hypothetical company. The data contains various attributes of the employees that will help in studying employee behavior. Understanding the employee needs and adjusting the company terms and policies accordingly has been a very important concern for the human resource department. The problem of salary projections is very old in various sectors and is important for the company and employees to understand the market conditions. Also, Employee attrition has been a major concern for all the companies and various contour measures have been proposed to help reduce attrition.

The general objective of this study is to develop models to predict employee salaries and employee attrition and various factors that affect them. To properly analyze the data and to make a satisfactory prediction, it is essential to understand the data. For this purpose, we have performed various exploratory analysis and predictive analysis on the data.

II. Data Pre-processing and Exploration

On broad overview of the dataset, we understand that there are 2940 observations with 35 variables. There are no NA values in the dataset. Further, we find that columns like 'over 18', 'employee count', 'standard hours', 'Employee Number', 'Environment Satisfaction' are not informative and remove it. Columns like hourly rate, daily rate and monthly rate have redundant information and have been removed. Few column names have been changed to more meaningfully short names and categorical columns have been factored.

Before modelling, we need to find out the variables that could be important in predicting the outcome. Therefore, we do some univariate and bivariate data analysis to discover insights and try to correlate the data. Figure 1 shows the correlation plot of the data. From this figure, we can understand that the most outstanding result is between 'Job Level' and 'Monthly income', whose correlation is 0.95. The more performance rating, the more Performance salary hike, whose correlation is 0.773. The more total working hours, the more Job Level, whose correlation is 0.782. The more total working hours, the more monthly Income, whose correlation is 0.772. The more 'yearswithcurrmanager', the more 'yearsatcompany', whose correlation is 0.769. The more 'yearsatcompany', the more 'yearsInCurrentRole', whose correlation is 0.758. The more 'yearswithcurrmanager', the more 'yearsincurrentrole', whose correlation is 0.71. To avoid multi-collinearity problems, one of the highly correlated variables are excluded (Performance Salary Hike, 'yearswithcurrmanager', 'yearsatcompany'). Job level was used in predicting Monthly Income, but was exclude in predicting Employee Attrition.

III. Models

1. Ordinary Least Squares

The ordinary least square (OLS) method is a basic approach to estimate β. Its expression is given by

$$\hat{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\,.$$

The OLS estimator is widely used and can serve as the initial estimator in many other methods.

2. Ridge Regression

Ridge regression uses an ℓ2-norm penalty to improve OLS when the covariates are correlated. Like OLS, the ridge estimator has an explicit form

$$\widehat{\beta}_{\lambda} = \arg\min_{b} \{||y - Xb||_2^2 + \lambda||b||_2^2\}$$

where λ > 0 is the tuning parameter and $I_p$denotes the p × p identical matrix. Here we select λ by minimizing the generalized cross-validation criterion.

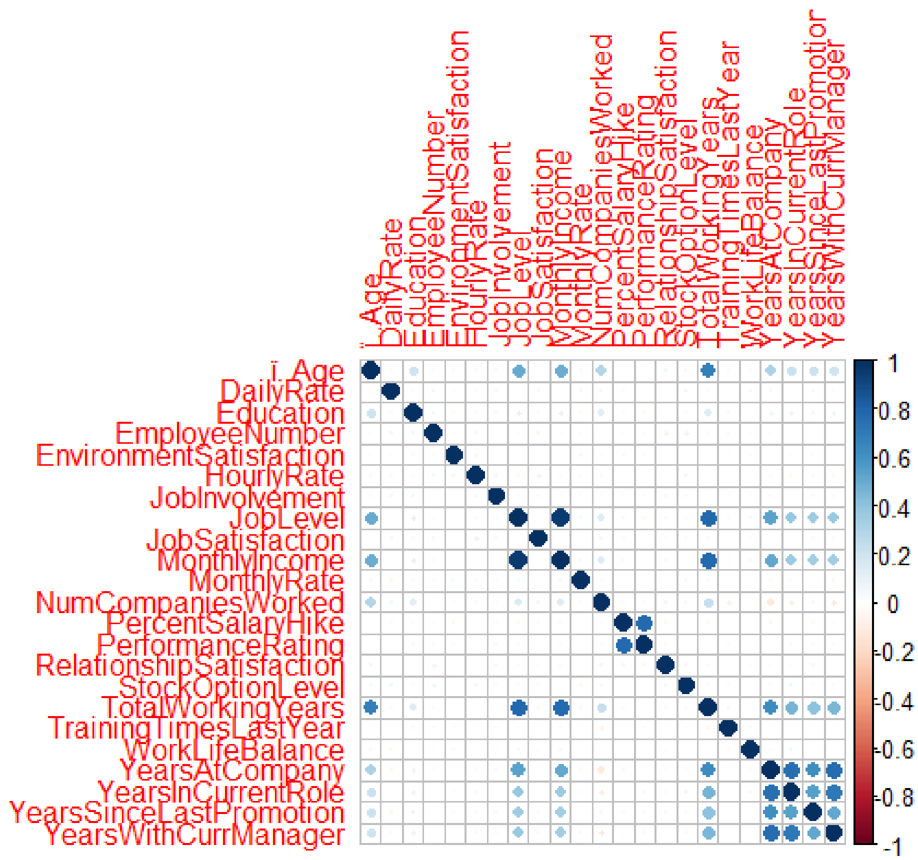

Figure 1

3.Lasso Regression

Lasso regression uses an ℓ1-norm penalty to improve OLS when the covariates are correlated. Like OLS, the lasso estimator has an explicit form

$$\widehat{\beta}_{\lambda} = \arg\min_{b} \{||y - Xb||_2^2 + \lambda||b||_1\}$$

where λ > 0 is the tuning parameter and I_p denotes the p × p identical matrix. Here we select λ by minimizing the generalized cross-validation criterion.

4.Logistic Regression

Logistic regression is useful when you are predicting a binary outcome from a set of continuous predictor variables. It is frequently preferred over discriminant function analysis because of its less restrictive assumptions. Suppose we have observations X and the responses Y. The objective function for the Gaussian family is

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda\left[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1\right].$$

where λ≥0λ≥0 is a complexity parameter and 0≤α≤10≤α≤1 is a compromise between ridge (α=0α=0) and lasso (α=1α=1).

To evaluate and compare learning algorithms, we divided the current dataset into two segments: one used to learn and validate the model (70%) and the other one used to test the final model (30%). Also, we split the 70% of training data to two segments: one used to train the model and the other one used to validate the model. We applied 10-fold cross-validation on the train and validation data set. We partitioned the data into ten equally sized folds. We held one-fold out for validation and used the other nine-folds for training. (Refaeilzadeh P et al., 2009) We repeated the process ten times and held a different validation fold every time. We used the predicted R-squared (formula given below) as the evaluation metric to determine how good the model predicts responses for new observations. We evaluated the performance of various models by comparing their predicted R-squared values and Mean Square Error (MSE). A larger value of predicted R-squared and small vale of (MSE) shows a model of greater predictive ability. In the end, we applied the model with the highest R-squared value to the test data set (30% of the original data set) and calculated its predicted R-squared.

IV. Results

1. Monthly Income

To predict the Monthly Income, we first used the linear regression model with all the variables after data procession. Figure 2 shows the summary of the model. A common way to summarize how well a linear regression model performed is via the coefficient of determination. This can be calculated as the square of the correlation between the observed values and the predicted values.

If the predictions are close to the actual values, we would expect R^2 to be close to 1. On the other hand, if the predictions are unrelated to the actual values, then R^2 =0. In all cases, R2 lies between 0 and 1. The model has a R^2 of 95.14% which indicates that 95% of the variance in data is captured. The p- Value of the F-statistic is significantly small indicating that at least one of the variable is related to the response variable and also, we can see that there are more than 5% variables where the p-value of the t-statistic is very less (below 0.05 considering 95 confidence interval). Figure 3 shows the plot of residuals of the fitted model. As the figure shows, the residuals are unbiased and homoscedastic and this add to explain the R^2 of the model. But, as we see from figure 2, most of the independent variables are statistically insignificant and it is difficult to interpret the model.

To further improve the model, we used Lasso and Ridge models. These two models are closely related and used to prevent overfitting and regularize the coefficients. Like OLS, Lasso and Ridge models try to minimize the residual square errors. The best λ for ridge and lasso are determined by choosing the value that minimizes k-fold cross validation error and plots of errors for both models are shown in figure 4. The plot on left shows MSE vs log(lambda) for Ridge and on the right, shows the plot for lasso. The LASSO regression also tends to “shrink” the regression coefficients to zero as λ increases. The reader can tell this by looking at the numbers in the upper part of the plot which again mean the number of non-zero coefficients in the regression model. The best lambda is shown by the vertical lines. We can see that MSE increases in lasso as we shrink the model. From this plot we select the best λ and fit lasso and ridge models. The best lambda for Ridge and Lasso are 395.76 and 41.46 respectively.

```
> summary(lmmod)

Call:
lm(formula = train_ibm$MonthlyIncome ~ ., data = train_ibm)

Residuals:
    Min      1Q  Median      3Q     Max 
-3021.4  -631.7   -55.6   611.4  4382.9 

Coefficients:
                                 Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                     4339.6019   676.1106   6.418  0.00000000021748391 ***
Age                                0.3218     5.3232   0.060              0.95181    
AttritionYes                    -137.5530   107.1373  -1.284              0.19949    
BusinessTravelTravel_Frequently  133.0134   131.8025   1.009              0.31314    
BusinessTravelTravel_Rarely      144.4330   112.8443   1.280              0.20088    
DepartmentResearch & Development -14.5289   573.2573  -0.025              0.97979    
DepartmentSales                  -95.9552   583.4321  -0.164              0.86940    
DistanceFromHome2                 73.6981   124.1515   0.594              0.55291    
DistanceFromHome3                 30.9439   165.9480   0.186              0.85212    
DistanceFromHome4                117.8417   186.4803   0.632              0.52759    
DistanceFromHome5                100.9236   176.9180   0.570              0.56851    
DistanceFromHome6                -16.1181   191.0598  -0.084              0.93279    
DistanceFromHome7               -106.5954   165.0923  -0.646              0.51865    
DistanceFromHome8               -111.1051   174.1320  -0.638              0.52359    
DistanceFromHome9                 -7.3741   160.0754  -0.046              0.96327    
DistanceFromHome10              -135.2657   161.2662  -0.839              0.40181    
DistanceFromHome11              -221.2972   233.8211  -0.946              0.34417    
DistanceFromHome12               104.1498   330.7198   0.315              0.75289    
DistanceFromHome13                63.5096   309.6921   0.205              0.83756    
DistanceFromHome14               -29.2019   297.8648  -0.098              0.92192    
DistanceFromHome15              -215.4928   268.8687  -0.801              0.42306    
DistanceFromHome16              -280.8944   237.2915  -1.184              0.23681    
DistanceFromHome17               131.1300   273.3700   0.480              0.63157    
DistanceFromHome18               134.3641   244.1969   0.550              0.58229    
DistanceFromHome19               -27.7790   290.8940  -0.095              0.92394    
DistanceFromHome20               183.0527   288.3101   0.635              0.52564    
DistanceFromHome21               150.4320   286.0426   0.526              0.59908    
DistanceFromHome22                54.8776   301.6590   0.182              0.85569    
DistanceFromHome23              -376.5016   272.6948  -1.381              0.16771    
DistanceFromHome24              -556.9232   276.4925  -2.014              0.04427 *  
DistanceFromHome25              -278.1200   278.6181  -0.998              0.31843    
DistanceFromHome26                35.4055   274.6811   0.129              0.89747    
DistanceFromHome27              -193.7089   347.0344  -0.558              0.57685    
DistanceFromHome28               259.6058   292.6818   0.887              0.37531    
DistanceFromHome29               125.4283   283.9840   0.442              0.65883    
Education2                      -182.2802   126.3934  -1.442              0.14959    
Education3                      -150.2777   114.3509  -1.314              0.18910    
Education4                       -43.1702   121.5967  -0.355              0.72265    
Education5                      -317.9171   202.0928  -1.573              0.11602    
EducationFieldLife Sciences     -233.9807   330.1606  -0.709              0.47869    
EducationFieldMarketing         -148.2632   352.0212  -0.421              0.67372    
EducationFieldMedical           -301.9346   331.4898  -0.911              0.36261    
EducationFieldOther             -307.3744   352.0692  -0.873              0.38286    
EducationFieldTechnical Degree  -203.3442   346.2026  -0.587              0.55710    
EnvironmentSatisfaction2         -65.6403   108.4939  -0.605              0.54532    
EnvironmentSatisfaction3         -71.8462    97.5668  -0.736              0.46168    
EnvironmentSatisfaction4         -69.2136    97.7154  -0.708              0.47892    
GenderMale                       101.8628    68.9147   1.478              0.13971    
JobInvolvement2                 -313.7162   155.7029  -2.015              0.04420 *  
JobInvolvement3                 -423.2876   147.2982  -2.874              0.00415 ** 
JobInvolvement4                 -436.1312   177.6509  -2.455              0.01427 *  
JobLevel2                       1575.0433   130.5262  12.067 < 0.0000000000000002 ***
JobLevel3                       4869.9589   182.7200  26.653 < 0.0000000000000002 ***
JobLevel4                       8681.0899   281.2499  30.866 < 0.0000000000000002 ***
JobLevel5                      11031.3108   326.5297  33.783 < 0.0000000000000002 ***
JobRoleHuman Resources          -940.7771   585.9387  -1.606              0.10870    
JobRoleLaboratory Technician   -1358.9718   172.6502  -7.871  0.00000000000000959 ***
JobRoleManager                  3226.7228   262.0141  12.315 < 0.0000000000000002 ***
JobRoleManufacturing Director   -295.2277   166.5992  -1.772              0.07670 .  
JobRoleResearch Director        3329.4411   220.5158  15.098 < 0.0000000000000002 ***
JobRoleResearch Scientist      -1387.0262   173.1237  -8.012  0.00000000000000332 ***
JobRoleSales Executive          -100.7486   309.5553  -0.325              0.74490    
JobRoleSales Representative    -1566.5333   350.6740  -4.467  0.00000888307246232 ***
JobSatisfaction2                -120.3472   110.6136  -1.088              0.27687    
JobSatisfaction3                 -97.6079    96.9743  -1.007              0.31442    
JobSatisfaction4                 -27.3707    95.3670  -0.287              0.77417    
MaritalStatusMarried              54.9320    89.5087   0.614              0.53956    
MaritalStatusSingle              147.9925   148.1885   0.999              0.31821    
NumCompaniesWorked                30.6445    14.7646   2.076              0.03821 *  
OverTimeYes                       89.4986    78.0159   1.147              0.25160    
PerformanceRating4                59.7524    93.6950   0.638              0.52380    
RelationshipSatisfaction2        109.6075   108.6770   1.009              0.31344    
RelationshipSatisfaction3         43.7628    97.3330   0.450              0.65309    
RelationshipSatisfaction4         25.4816    98.4034   0.259              0.79573    
StockOptionLevel1                180.6176   119.0669   1.517              0.12962    
StockOptionLevel2                124.1396   149.8759   0.828              0.40772    
StockOptionLevel3                -17.2890   177.8928  -0.097              0.92260    
TotalWorkingYears                 26.2534     9.1237   2.877              0.00410 ** 
TrainingTimesLastYear             -8.8806    25.7807  -0.344              0.73057    
WorkLifeBalance2                  30.4382   159.8286   0.190              0.84900    
WorkLifeBalance3                  80.9688   150.6987   0.537              0.59119    
WorkLifeBalance4                  29.0352   176.6234   0.164              0.86946    
YearsInCurrentRole                16.7718    12.2145   1.373              0.17005    
YearsSinceLastPromotion            0.5169    12.8836   0.040              0.96801    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1035 on 945 degrees of freedom
Multiple R-squared:  0.9554,	Adjusted R-squared:  0.9514 
F-statistic: 243.6 on 83 and 945 DF,  p-value: < 0.00000000000000022
```

Figure 2

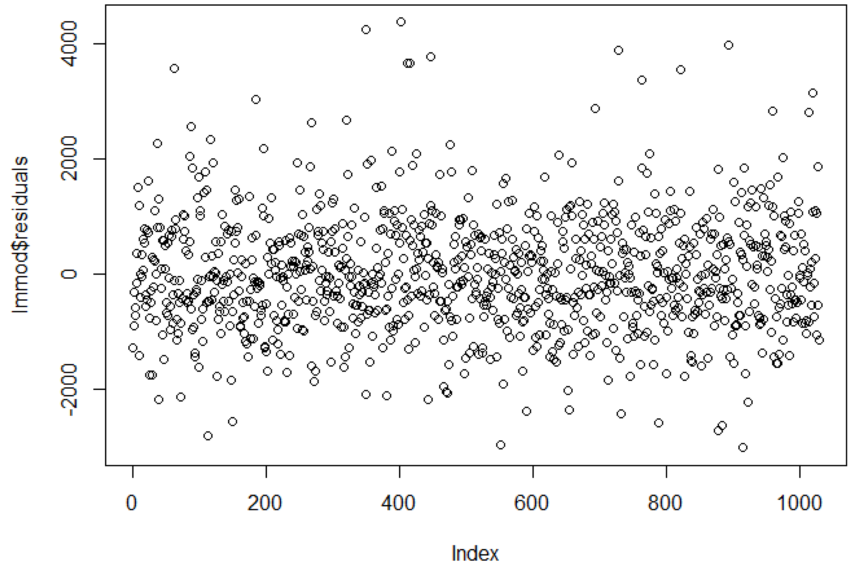

Figure 3

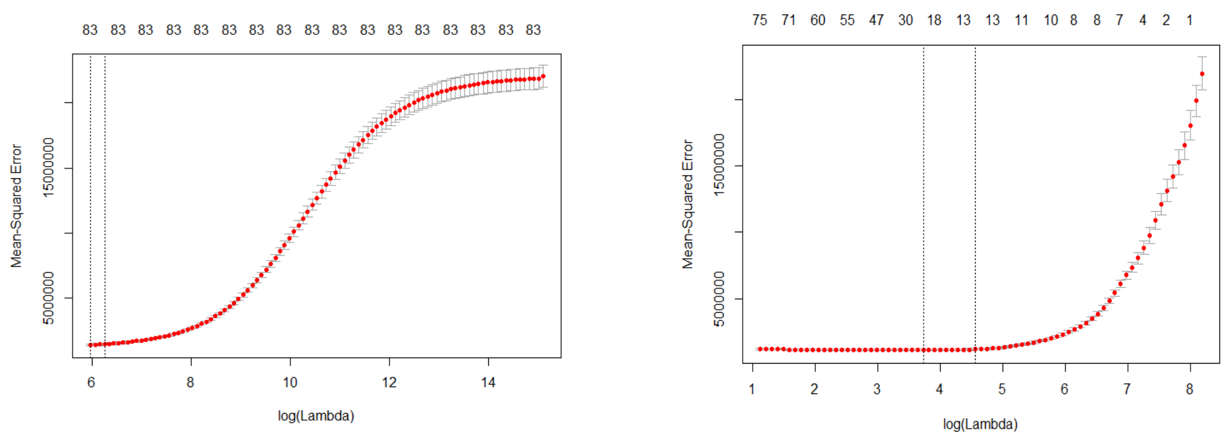

Figure 4

Lasso model reduced the coefficients and the non-zero lasso coefficients shown in table 1.

JobLevel2	995.49734
JobLevel3	3978.82272
JobLevel4	7313.67666
JobLevel5	9592.50558
JobRoleManager	3464.55494
JobRoleResearch Director	3510.11191
JobRoleSales Executive	70.01073
TotalWorkingYears	67.05742
YearsInCurrentRole	4.50465

Table 1

Let us now compare the Residual Sum of Square value for all the models. We will first calculate the RSS on the train set and then move to the test set. Table 2 documents the RSS for all the models.

RSS	OLS	Ridge – Best	Ridge – 1se	Lasso - Best	Lasso – 1se
Train	1070871392	1257521581	1323429244	1120166309	1228423098
Test	501988064	593711640	624537631	493186614	523654496

Table 2

Here the smallest RSS value on the train set is predictably achieved by the OLS regression since unlike Ridge and LASSO the OLS does not impose penalties on the coefficients. RSS for Ridge and LASSO is again predictably greater when λ is selected using the "one-standard error" rule. It is interesting that the errors on the test set are ordered differently - the minimum is achieved by LASSO at λ = lasso.best, the second is OLS. In this case, LASSO perform better than OLS on the test set. The variables in table 1 impact the income level as expected. JobLevel5 which is 'Very High' has the greatest contribution i.e. 9592.5 compared to job level 1 'Very Low'. Also, the other variables which are Manager Role, Research Director also have great impact.

2. Attrition Prediction

Generalized Linear model (GLM) was used to predict the possibility of Attrition of employees. We believe that large gap between management and production employees is one of the strongest reason for attrition. The GLM model of binomial family and link function logit is fitted to the data. The deviance of the model is 434 and null deviance is 886. Since the deviance is much less than the null deviance, our model explains a large proportion of the outcome. Also, since the deviance is less than the degrees of freedom 942, the model fits the data well and is not overfitted. Next, we found the variables that are statistically significant (probability that the estimates are due to chance is less than 5%, AND that have a significant effect: Odds Ratio greater than 1.00. From this we deduced that employees who are working overtime (presumably a lot), have had many previous employments, have to travel often, and/or are male are most likely to be associated with our outcome variable - attrition. Figure 5 plots the confusion matrix of the predicted Attrition on test data. The model performed pretty well with an accuracy of 84.13%.

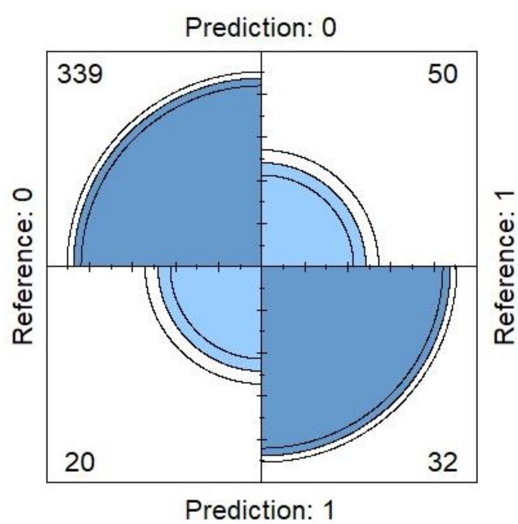

Figure 5

V. Conclusion

Our results provide predictions for Employee Monthly Income and Employee Attrition. From the results discussed in section IV, Lasso model works well with least RSS. 'JobLevel' and position of the employee have strong contribution to Employee Monthly Income. On the other hand, Employee Attrition depends mainly on factors like working overtime, many previous employees and have to travel often are more likely to leave the company. Hence, the HR department should work towards a better work life balance and hire people with fewer previous employers. Employees should also plan to live close to the companies and this would help them continue longer in their job.

VI. References

How to perform a Logistic Regression in R. (2015, September 13). Retrieved December 19, 2017, from https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

P. (2017, March 31). Retrieved December 19, 2017, from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Ridge Regression and the Lasso. (2017, May 23). Retrieved December 19, 2017, from https://www.r-bloggers.com/ridge-regression-and-the-lasso/

Logit Regression | R Data Analysis Examples. (n.d.). Retrieved December 19, 2017, from https://stats.idre.ucla.edu/r/dae/logit-regression/

Confusion matrix and ROC curves. (n.d.). Retrieved December 19, 2017, from https://stats.stackexchange.com/questions/156036/confusion-matrix-and-roc-curves

https://www.kaggle.com/dave1216/employee-attrition-as-an-epidemiologic-problem

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: with applications in R. New York: Springer.