Case Study 3 – Machine Learning

DS501: Intro to Data Science

Martin Blatz

Summer 2020

https://martinblatz.shinyapps.io/casestudy3/

Introduction

The data used for this project is IBM HR Analytics Employee Attrition and Performance data. It can be found at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset. The data is fictional data created by data scientists at IBM for research and learning purposes. I chose the data because it is rich in numerical data and the concept of employee and customer retention is interesting to me. I expect that current world affairs will significantly alter trends in employee retention. As a resource manager, employee retention and attrition directly affect my everyday work.

Principal Component Analysis

Given the assignment to create a shiny application which gives the end user the ability to change parameters in a machine learning algorithm to explore how those parameters affect the results, I decided to approach this data set with a principal components analysis. The data has a relatively large number of columns – 26 numerically valued columns in all. Principal component analysis is a multivariate exploratory data analysis technique which seeks to explain the variance of a data set with fewer "components" by grouping correlated variables together. These extracted principal components help to reduce the data set to a more manageable size and can be used as inputs to further data analysis techniques.

Principal component analysis works by treating each variable as its own dimension. Within each dimension, the scale is often variable. To correct for disproportionate weighting of variables, the standard deviation is calculated for each dimension and the inverse applied to the entire column of data. This results in a dataset with scaled variables of equal variance. Mean centering the resulting data involves subtracting the average value of the column from each entry. Mean centering results in an average value of 0, centering the entire data set around the origin.

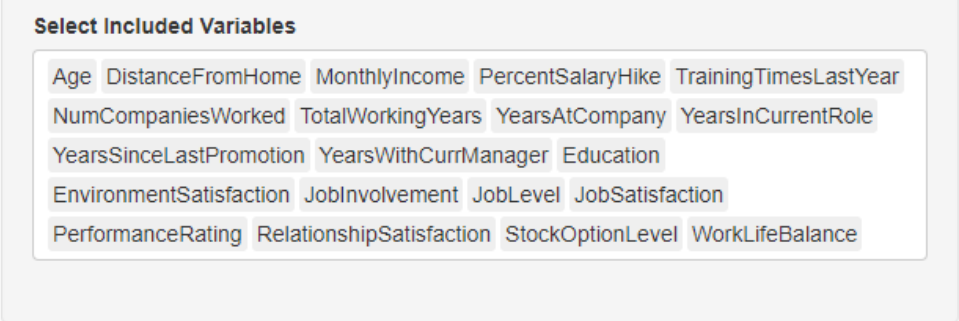
To calculate the first principal component, a best fit line is calculated using least squares. The first principal component line intersects with the average point in the data and thus explains as much variance in the data as is possible with a single straight line. The second principal component is determined in the same way, but each additional principal component must be orthogonal to those before it. Each observation is projected onto the preceding principal component lines to determine a

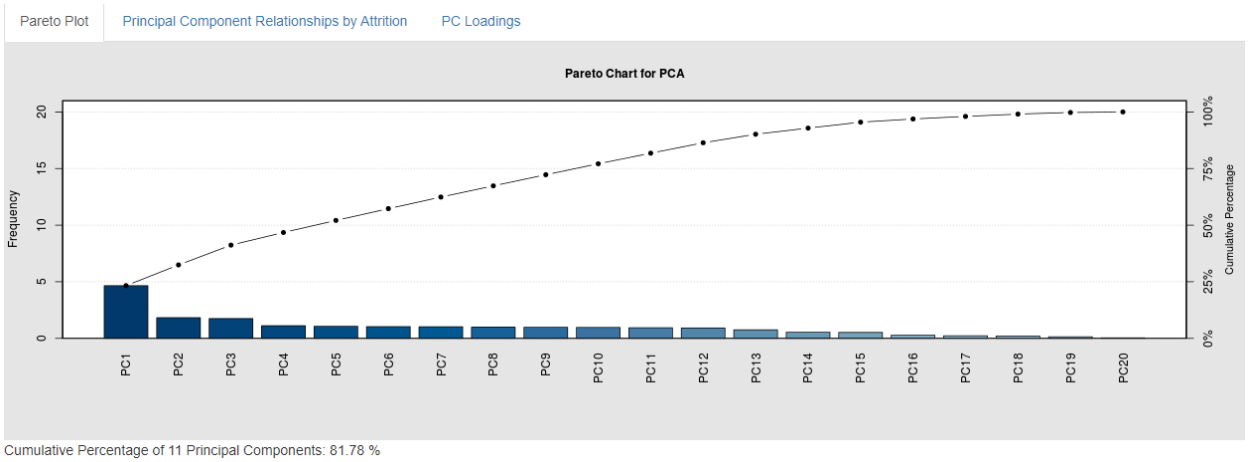
"score" representing its value along each axis created by the principal components. Each successive principal component is calculated using least squares to determine the direction of greatest variance.

Modeling a data set can be achieved and visualized by drawing a plane using the axes represented by the principal components. Observations are plotted in the plane based on the observation score on each principal component. This score plot allows a data scientist to visualize the structure of the data two dimensions at a time. By selecting different colors to represent a third value, a data scientist can discern patterns in the data – observations close to each other on the plot represent similar properties based on the principal component loading.

Application

The first step in the data analysis process is to clean the data. While the IBM data set is relatively clean, there are several data fields which are not numerical, and some which don't vary. Character fields (except attrition), standard hours, employee count, and nominal data fields such as employee number were removed before beginning the process. There are monthly income, daily rate, monthly rate, and hourly rate all appear to represent earnings but don't have an explanation in the appendix data provided. Because of this, only the monthly income was retained. After cleanup, there are 20 variables remaining to begin the principal components analysis.

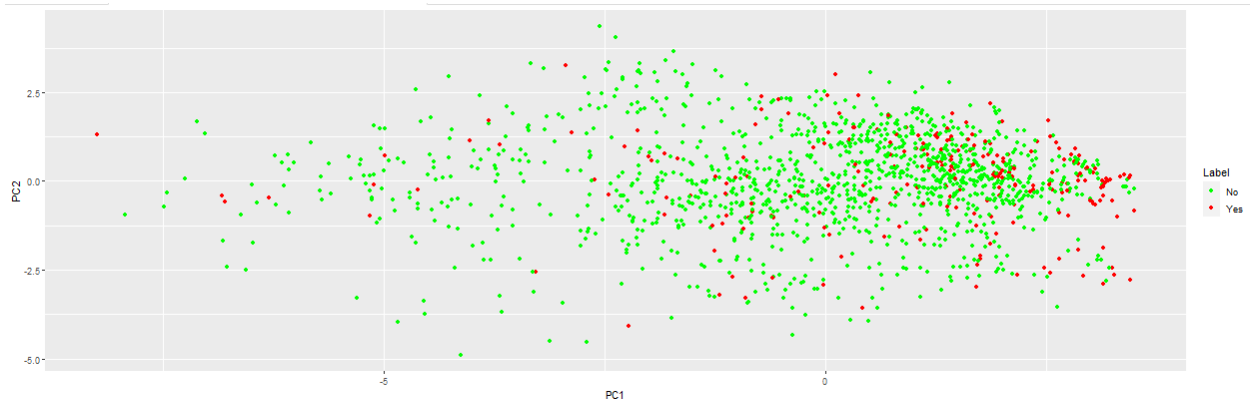




Cumulative Percentage of 11 Principal Components: 81.78 %

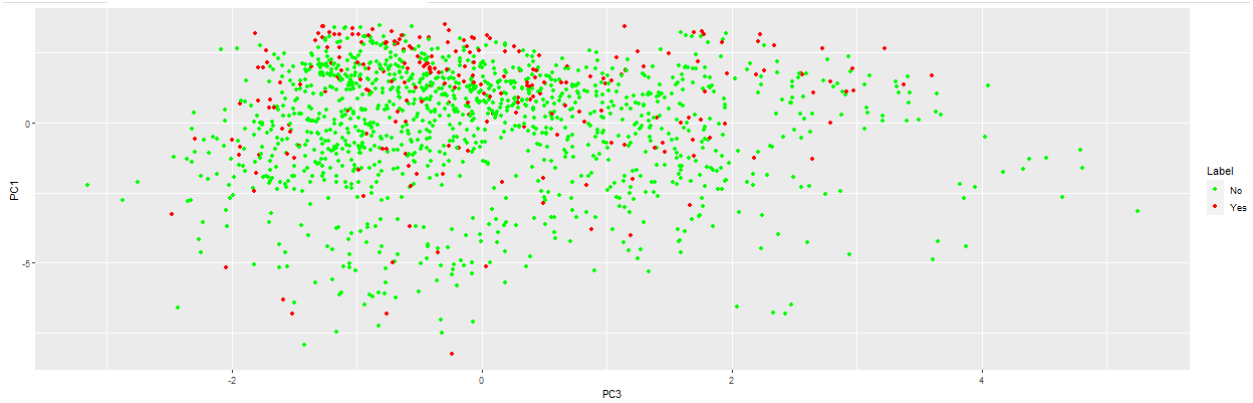
There are several approaches to interpreting PCA results. One approach is to consider as many PCs are are necessary to reach a certain percentage. At least 70-80% is necessary for exploratory data analysis, while closer to 90% is typically required for further modeling. 11 principal components are necessary to account for greater than 80% of the variance in the data set. 13 PCs are necessary to account for more than 90%. The scree plot approach stops considering PCs once there is a break in the chart where the trend flattens out. This chart is relatively flat after the 3rd PC. The two approaches yield very different results. PC1 is the only PC which accounts for greater than 10% of the variance in the data set.

Looking at the score charts confirms the findings of the first chart. A chart with PC1 and PC2 reveals that observations representing positive attrition tend towards the right as PC1 increases, suggesting a positive correlation. There also appears to be more positive attrition as PC2 increases, but the correlation is much less pronounced.



A score chart for PC1 and PC3 also shows a slightly negative correlation between PC3 and attrition, which more red observations on the left of the plot.

Beyond PC3, the distribution appears much more consistent with patterns indiscernable.



The rotation values from the pcromp function provide the variable loadings for each principal component. Each column contains an eigenvector corresponding to the principal component, with the values representing the loading or correlation between the principal component and the variable. A positive value represents a positive correlation while a negative value is a negative correlation. Higher absolute values of loading signifies a stronger correlation between the component and the underlying

variable. Variables which are loaded on the same principal component are correlated and can be reasonably represented by a single value – the principal component.

rownames(table)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
Age	-0.28	0.27	0.27	-0.06	0.00	-0.02	0.03	-0.02	-0.06	-0.04	0.10	0.03	-0.07	-0.67	0.42	0.05	0.08	0.26	-0.24	0.01
DistanceFromHome	-0.00	-0.06	0.03	-0.43	0.14	0.13	-0.16	-0.49	0.54	0.13	0.01	0.42	-0.15	0.01	-0.01	-0.01	-0.01	0.00	-0.01	0.02
MonthlyIncome	-0.37	0.19	0.16	0.18	0.10	0.03	0.01	-0.07	0.06	-0.05	-0.19	0.07	0.24	0.30	-0.11	0.03	0.06	0.17	-0.12	0.70
PercentSalaryHike	0.02	-0.46	0.53	0.02	-0.05	-0.01	0.01	-0.01	-0.04	-0.02	0.01	0.02	0.03	0.01	-0.01	-0.00	0.68	-0.20	0.03	-0.01
TrainingTimesLastYear	0.01	-0.04	-0.07	0.30	0.02	-0.41	0.55	-0.22	0.04	-0.34	0.35	0.35	-0.15	0.04	-0.04	0.00	-0.01	-0.00	0.01	0.00
NumCompaniesWorked	-0.05	0.36	0.33	-0.12	-0.13	-0.07	-0.08	0.10	-0.00	-0.02	0.16	-0.20	-0.70	0.25	-0.27	-0.02	0.01	-0.02	-0.11	-0.00
TotalWorkingYears	-0.40	0.16	0.16	0.04	0.04	0.02	0.00	-0.00	0.02	-0.04	-0.04	0.03	-0.01	-0.16	0.02	-0.07	-0.18	-0.50	0.68	-0.00
YearsAtCompany	-0.39	-0.21	-0.19	-0.02	-0.02	0.03	0.01	0.02	-0.01	0.01	0.04	-0.02	-0.02	-0.08	-0.09	-0.09	-0.12	-0.59	-0.62	-0.01
YearsInCurrentRole	-0.34	-0.27	-0.22	-0.09	-0.05	-0.05	-0.03	0.04	-0.05	0.04	0.05	-0.05	-0.16	-0.10	-0.24	0.77	0.04	0.18	0.15	0.01
YearsSinceLastPromotion	-0.30	-0.19	-0.17	-0.04	-0.08	0.02	-0.03	0.04	-0.00	0.01	0.14	-0.09	-0.20	0.47	0.72	-0.07	0.04	0.09	0.09	-0.01
YearsWithCurrManager	-0.33	-0.27	-0.23	-0.11	-0.07	0.00	0.02	0.07	-0.04	0.01	0.05	-0.04	-0.11	-0.19	-0.35	-0.61	0.11	0.39	0.16	0.03
Education	-0.08	0.14	0.11	-0.42	-0.08	-0.18	0.16	0.23	-0.18	0.46	0.46	0.20	0.37	0.16	-0.07	0.01	-0.02	-0.02	0.00	0.00
EnvironmentSatisfaction	-0.00	0.03	-0.05	0.14	-0.21	-0.16	-0.74	-0.04	-0.24	-0.29	0.26	0.37	0.09	0.02	-0.03	-0.02	-0.00	-0.01	0.00	0.01
JobInvolvement	0.00	0.05	-0.01	-0.44	-0.34	-0.01	0.22	-0.10	-0.45	-0.26	-0.51	0.28	-0.06	0.07	0.04	0.00	-0.02	-0.03	-0.01	0.00
JobLevel	-0.38	0.18	0.15	0.17	0.11	0.02	0.00	-0.09	0.06	-0.04	-0.18	0.08	0.23	0.26	-0.13	0.02	0.05	0.20	-0.11	-0.71
JobSatisfaction	0.01	-0.03	-0.00	0.10	0.51	0.20	-0.02	-0.41	-0.63	0.27	0.10	0.03	-0.18	0.03	-0.02	-0.03	-0.01	0.01	0.01	0.00
PerformanceRating	0.00	-0.47	0.52	0.03	-0.07	-0.00	-0.01	0.01	-0.03	-0.03	0.01	-0.01	0.02	0.03	-0.00	-0.01	-0.68	0.18	-0.05	0.00
RelationshipSatisfaction	-0.02	0.09	-0.00	0.06	-0.56	0.33	0.09	-0.55	-0.02	-0.03	0.29	-0.35	0.21	-0.00	-0.07	0.02	0.01	-0.00	0.02	-0.00
StockOptionLevel	-0.02	-0.01	0.01	-0.37	0.29	-0.56	-0.13	-0.28	-0.02	-0.28	-0.00	-0.50	0.19	0.03	0.01	-0.03	-0.01	-0.02	0.01	0.00
WorkLifeBalance	-0.01	-0.01	-0.02	0.28	-0.31	-0.53	-0.09	-0.25	0.02	0.59	-0.32	0.00	-0.12	-0.07	0.05	-0.05	0.00	-0.01	-0.00	0.01

PC1 has significant loading (>|0.4|) on total working years and moderate loading (>|0.3|) on monthly income, years at company, years in current role, years with current manager, and job level. It appears that this PC is most closely related to professional experience.

PC2 and PC3 both have significant loading on the most recent salary hike and performance rating, with moderate loading on the number of companies worked. These PCs appear to be related to job performance.

PC4 has significant loading on the distance from home, education, and job involvement. There is also moderate loading on stock option level.

PC5 has significant loading on job satisfaction and relationship satisfaction, and moderate loading on work life balance and job involvement.

PC6 has significant loading on stock option level and work life balance and moderate loading on training times last year.

As one progresses further to the right, the PCs represent less and less of the variance within the data set and correlate less with differences in the underlying data.

Conclusion

The shiny application I built to complete this analysis is available at https://martinblatz.shinyapps.io/casestudy3/. It provides the ability to change and chose the variables you wish to include in your analysis. Changes to the variables used results in a recalculation of all PCs, charts, and graphs. The first tab allows the user to select the number of PCs to calculate the cumulative variance represented. The second tab allows the user to select two PCs and view the score chart, color coded with red for attrition observations and green for non attrition observations. The third tab is a chart with rotation or loading values for the calculated principal components.

Principal components analysis is a useful technique for evaluating the structure and underlying relationships between variables within a data set. By performing the analysis, a data scientist can better understand which variables are most closely related and if they may be able to be represented by a single component. This is especially useful for large data sets where a reduction in the number of dimensions is desired before continuing on to further analysis.

The IBM data is best represented by the first 3 factors in my analysis, but those three components only account for 41.8% of the variance within the data set. To achieve at least 70% of the variance, 9 principal components must be used. This still represents a significant reduction in the number of variables from the original 20 used.

Appendix

Additional Information provided regarding the data set values:

Education
1 'Below College'
2 'College'
3 'Bachelor'
4 'Master'
5 'Doctor'

EnvironmentSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

JobInvolvement
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

JobSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

PerformanceRating
1 'Low'
2 'Good'
3 'Excellent'
4 'Outstanding'

RelationshipSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

WorkLifeBalance
1 'Bad'
2 'Good'
3 'Better'
4 'Best'