DePaul University – Fall, 2020
DSC 424 Class Project - Final Report

IBM EMPLOYEE ATTRITION PREDICTION

Group Name: IBM
Group Members:
Serena Yang, Yueting Zhao

Table of Contents

Non-technical Parts

Introduction

Attrition is defined as Both the voluntary and involuntary reduction of a company's workforce. In a company, attrition is a human resource problem. Those employees who voluntarily resign may have different reasons; some are due to the company's overall working atmosphere; some are personal reasons. While some attrition is expected in everyday business operations, a high level of reduction can lead to problems and a lack of human resources. Some of the ways human resources professionals do their part to keep top-performing employees happy and attrition rates low is to design and implement company compensation programs, motivation systems, and company culture. Besides retaining top-performing employees, business owners keep their attrition rates as low as possible to keep from spending money on advertising for, hiring, training, and completing paperwork for new employees. For this human resource problem, to find the common factors that affect those workers who leave is the essential step for the HR department to anticipate employees with potential reasons. The goals of this project are first uncovered the factors that lead to IBM employee attrition; secondly, predict if an employee is likely to quit.

In this project, we mainly used software R to achieve these goals. During the process, first, we created the basic regression model and used Forward selection and Backward elimination to check the variables selected. Secondly, we checked the overfitting problem and used Lasso to choose the best lambda, consequently handling regression models' overfitting problem. Thirdly, we have checked the multicollinearity problem based on the model by analyzing high correlations among independent variables. Applied Principal Component Analysis and Linear Discriminant analysis to achieve dimensionality reduction. We used Ordinal analysis to analyze those ordinal variables. Finally, make a logistic regression model and use the receiver operating characteristic to predict on the test dataset.

Data Preparation

The dataset that we used was published by the Human Resource department of IBM. The download link is from Kaggle website. Before we did the data cleaning, the dataset contained 37 variables, including numeric and categorical, and 23,436 rows in total.

After we decided to use the dataset that we are planning to use for this project, we determined to delete those NAs rows and some variables to make dataset cleaning including:

- Daily Rate - Since we can calculate this by using hourly rate and standard hours
- Employee count - Data in this column are all equal to 1, not helpful for analysis
- Application ID - this variable is not helpful for our goal

- Over 18 - Everyone in this dataset is over 18
- Standard Hours - Every object's standard hours are 80 hours
- Monthly Rate - We already have variable "Monthly Income"

By continually finished data cleaning, we converted some variables from categorical to numeric type. After this step, the dataset has 23 variables, including 11 categorical and 12 numeric variables.

Type	Variable	Description
numeric	Age	age
categorical	Attrition	Voluntary Resignation/Current Employee
categorical	BusinessTravel	Frequency of work travel
categorical	Department	employee's department
numeric	DistanceFromHome	distance from home to company
categorical	Education	years of education
categorical	EducationField	education field
numeric	EmployeeNumber	the total number of employees in the company's branch where the employee is located
numeric	EnvironmentStatisfaction	employee's satisfaction with the environment
categorical	Gender	employee's gender
numeric	HourlyRate	employee's hourly rate
numeric	JobInvolvement	the contribution of the employee to the work
numeric	JobLevel	employee's job level
categorical	JobRole	employee's job role
numeric	JobSatisfaction	employee satisfaction with the job
categorical	MaritalStatus	employee's marital status
numeric	MonthlyIncome	employee's monthly income
numeric	NumCompaniesWorked	how many companies have the employee worked in his career?
categorical	OverTime	does the employee work overtime?
numeric	PercentSalaryHike	average annual salary growth rate for the employee
numeric	PerformanceRating	the measurement in which the analyst oberves the worker's perfornance
numeric	RelationshipSatisfaction	employee's satisfaction with the relationship in the IBM branch
categorical	StockOptionLevel	level of IBM stock the employee holds in the company
numeric	TotalWorkingYears	how many years does the employee worked in his career?
numeric	TrainingTimesLast Year	how long does the empoyee trained last year?
numeric	WorkLifeBalance	employee balance between life and the work
numeric	YearsAtCompany	the years of the employee work in IBM
numeric	YearsInCurrentRole	the years of the employee work in the current role
numeric	YearsSinceLastPromotion	the years of the employee work since last promotion
numeric	YearsWithCurrManager	the years of the employee work with current manager
categorical	Employee.Source	what resources does the employee use to get hired

These two data types describe each employee's background and characteristics and label whether they are still in the company or leave. Moreover, we deleted those NAs rows, which is about 230 objects. Before we analyze this dataset, we separated the dataset into two parts. 80% of data counts for the training dataset, and the rest are the test dataset.

Technical Parts

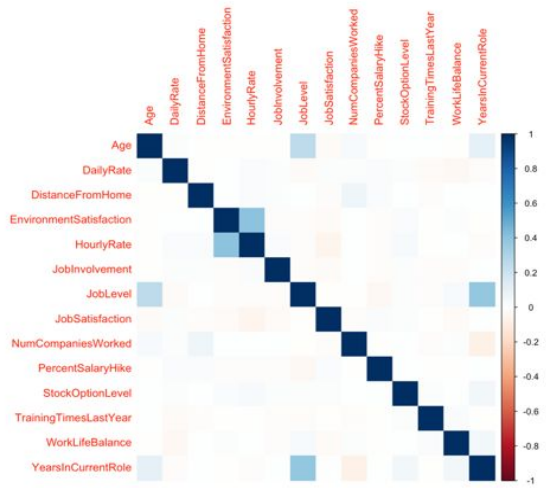
2.1 Principal Component Analysis

We choose to use Principal Component Analysis because our dataset is pretty large and has many variables and multicollinearity/overfitting. Dimension reduction is useful for all sorts of analyses in science and data science. Rather than care what each component means, we hope to throw away many components and not lose much. Principal Component Analysis (PCA) is a statistical technique used to reduce the data's dimensionality by selecting the essential features that capture maximum information about the dataset. Overfitting mainly occurs when there are too many variables in the dataset. PCA also helps in overcoming the overfitting issue by reducing the number of features.

```
> coef(lasso, s = lasso$lambda.1se)
26 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                4.504442e-01
Age                        4.946087e-03
DailyRate                  3.142392e-05
DistanceFromHome          -2.174808e-03
Education                  .
EmployeeNumber             .
Application ID             .
EnvironmentSatisfaction    9.786299e-03
HourlyRate                -3.015882e-05
JobInvolvement             2.434104e-02
JobLevel                   5.385355e-03
JobSatisfaction            1.110766e-02
MonthlyIncome              .
MonthlyRate                .
NumCompaniesWorked        -2.591582e-03
PercentSalaryHike          7.224965e-04
PerformanceRating          .
RelationshipSatisfaction   .
StockOptionLevel           2.868161e-02
TotalWorkingYears          .
TrainingTimesLastYear      7.555285e-03
WorkLifeBalance            1.933916e-03
YearsAtCompany             .
YearsInCurrentRole         4.180213e-03
YearsSinceLastPromotion    .
YearsWithCurrManager       .
```

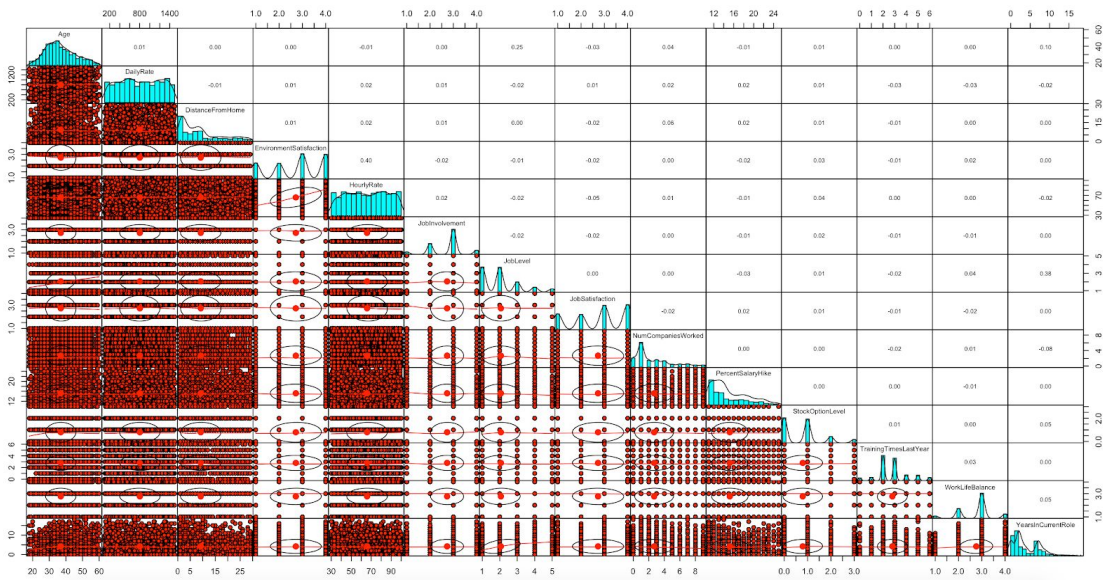
2.1.1 Selection Elimination

Lasso can perform feature selection. Lasso adds a penalty term for having large coefficients, which have a tuning parameter that decides how important the penalty is concerning the squared error term. In this case, we choose lasso$lambda.1se = 0.007877239. Look at the left graph; the variable Education, EmployeeNumber, Application ID, MonthlyIncome, MonthlyRate, PerformanceRating, RelationshipSatisfaction, TotalWorkingYears, YearsAtCompany are not selected. As a result, the total number of numerical variables is reduced from 26 to 14. Also provided Stepwise Selection(both backward and forward selection) based on the logistic regression model, since the response variable is binary. It received the same result as Lasso.



2.1.2 Correlation Matrix and Corrplot

Look at the correlation matrix and corrplot on the left, we can see there are some fairly significant correlations. In particular, the variables EnvironmentSatisfaction and HourlyRate, JobLevel, YearsInCurrentRole and Age. From the Scatter Plots and histograms(below), we can see the original variables are mostly not normalized distributed, histograms are some right-skewed, some are left-skewed. High correlations among independent variables lead to Multicollinearity problems.



2.1.3 Scale of the data

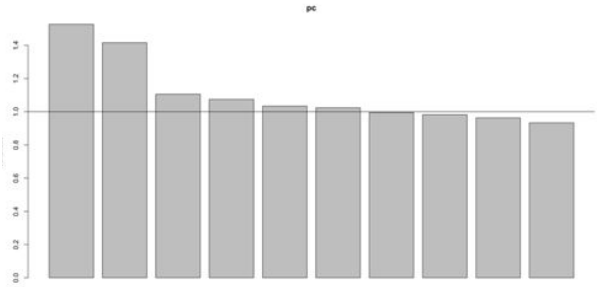
In our dataset, the individual parameters vary widely; scaling advantages account for very large ranges that would dominate as a single component. It also equalizes variance, so disparate variable units/scales do not skew the result. So, we will apply the PCA with the scaled dataset.

2.1.4 Summary of prcomp

From the summary of prcomp, 14 variables are coming in, 14 principal components are coming out, and the total is 100 percent. We get that the eigenvectors are sorted by decreasing the proportion of variance. The first component has about 11% of the variance; the second component has 10% of the variance; the third component has about 8% of the variance, etc. By the Seventh component, we have got about 58% of the data.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.2340 1.1892 1.04943 1.03572 1.01570 1.01131 0.99696 0.99109
Proportion of Variance 0.1088 0.1010 0.07866 0.07662 0.07369 0.07305 0.07099 0.07016
Cumulative Proportion  0.1088 0.2098 0.28846 0.36508 0.43877 0.51182 0.58282 0.65298
                          PC9   PC10    PC11    PC12    PC13    PC14
Standard deviation     0.98317 0.9707 0.96398 0.9196 0.77270 0.76001
Proportion of Variance 0.06905 0.0673 0.06638 0.0604 0.04265 0.04126
Cumulative Proportion  0.72202 0.7893 0.85570 0.9161 0.95874 1.00000
```

2.1.5 Scree Plot



Visualizing the principal components can tell how much variance is concentrated in these first few PCs. It helps to analyze how many principal components we are going to use. They are presented in decreasing order and show the relative sizes of the components. Knee in the scree plot where the variance levels off, draw the line at var = 1, components with variances below the line are sometimes discarded. We use the criterion of

drawing the line at the variance equals one. In our dataset, and based on the var = 1 criterion, the first 7 PCs are recommended for the analysis. The first 7 PSs take 58% of the total variance.

2.1.6 Rotation

Look at the data's Rotation table (In this case, we only consider the first 7 PCs). We get that the eigenvectors are sorted by decreasing the proportion of variance. The rotation gives us the contributions of each of the variables to the components. All the values in each component are normalized eigenvectors. So each principal component is a mixture like a regression model.

	PC1	PC2	PC3	PC4	PC5
Age	-0.421990594	0.085093141	-0.27511976	-0.007367868	-0.154476445
DailyRate	0.053191159	0.031681630	-0.24765115	-0.494904277	-0.075324700
DistanceFromHome	0.013360446	0.058912560	-0.49713696	0.147612088	0.382045376
EnvironmentSatisfaction	0.121863226	0.676626232	0.09463573	-0.044701409	0.013938318
HourlyRate	0.140110803	0.681743291	0.03245954	-0.050689562	-0.006525504
JobInvolvement	0.020865928	0.025304992	-0.16541810	-0.059512876	-0.302716445
JobLevel	-0.649541897	0.109058802	-0.04612967	-0.029678714	-0.024447961
JobSatisfaction	0.006129449	-0.135644565	0.20344476	-0.420942966	0.302173171
NumCompaniesWorked	0.066083502	0.024635588	-0.63810413	0.230597210	0.088918323
PercentSalaryHike	0.040867082	-0.062795815	-0.08096381	-0.272681000	0.591282153
StockOptionLevel	-0.055698976	0.126532108	0.01325782	-0.177684827	0.367655787
TrainingTimesLastYear	0.013761235	-0.009171707	0.28819793	0.358755932	0.207495339
WorkLifeBalance	-0.095661235	0.062128662	0.12920199	0.502662881	0.306496091
YearsInCurrentRole	-0.586154293	0.101239645	0.14588857	-0.081026025	0.105062435

For example, PC1 is a mixture of the 14 original variables. The formula for PC1 is PC1 = -.42*Age + 0.05*DailyRate + 0.01*DistanceFromHome + … -0.49*YearsInCurrentRole. We can see that Age, JobLevel, and YearsInCurrentRole are the top three variables contributing to PC1. Furthermore, they are all negatively correlated with PC1. All other variables are mixed evenly small. In PC2, the EnvironmentSatisfaction and HourlyRate are the top two contributions. And positively related to PC2. In PC3, the DistanceFromHome and NumCompaniesWorked are the two most contributions, and they are negatively correlated with PC3, etc.

2.1.7 Factor Analysis with factor rotation

```
Loadings:
                     RC1    RC2    RC3    RC4    RC5    RC6    RC7
Age                   0.571
DailyRate                                0.688
DistanceFromHome                  0.409         0.570
EnvironmentSatisfaction        0.830
HourlyRate                     0.828
JobInvolvement                                          0.885
JobLevel              0.815
JobSatisfaction
NumCompaniesWorked                0.802
PercentSalaryHike                                 0.666
StockOptionLevel                                                 0.781
TrainingTimesLastYear                                            0.547
WorkLifeBalance                             -0.605
YearsInCurrentRole    0.715

                RC1   RC2   RC3   RC4   RC5   RC6   RC7
SS loadings   1.518 1.408 1.084 1.068 1.050 1.021 1.019
Proportion Var 0.108 0.101 0.077 0.076 0.075 0.073 0.073
Cumulative Var 0.108 0.209 0.286 0.363 0.438 0.511 0.583
```

PFA is based on the correlation matrix, not the covariance matrix. The result on the left is the correlation of each of the variables with PCs. In our dataset, 58% of all of the explainable variants that we have with seven components. Chi-squared test for adequacy. Use the factor rotation technique to improve the PFA interpretability. Set cutoff = .4 to clean up loadings print hides values with small loadings and interpretation easier. We have much simpler and clear loadings, with all less correlated variables hidden, now.

```
Call:
multinom(formula = Attrition ~ PC1 + PC2 + PC3 + PC4 + PC5 +
    PC6 + PC7, data = trg)

Coefficients:
             Values  Std. Err.
(Intercept)  1.80276057 0.02233498
PC1         -0.34249592 0.01882579
PC2          0.11557200 0.01739453
PC3          0.11136127 0.01947761
PC4         -0.21440847 0.01964165
PC5          0.03663934 0.02075981
PC6         -0.18180723 0.02068250
PC7         -0.27489668 0.02141510

Residual Deviance: 15368.58
AIC: 15384.58
```

2.1.8 Fit Model with PCs

Our target is to predict the company's Attrition of employees based on the HR dataset in our project. We have already generated the seven components/factors; next, we need to fit a linear model with those PCs with our response variable to analyze which variables strongly affect the response

variable. By using the linear model, we can easily come to some observations; for example, PC1 has a negative correlation with Attrition, and in the PC1, there is also a negative correlation between PC1 and Age, JobLevel, YearsInCurrentRole. So we can summarize it into Age, JobLevel, and YearsInCurrentRole have a strong positive correlation with Attrition. In other words, Employees who are at a younger age, lower job level, and work fewer years in the current role show higher potential to leave. Similarly, we can also get results with other PCs.

2.1.9 Misclassification

```
> p_train <- predict(mymodel, trg)
> tab <- table(p_train, trg$Attrition)
> 1-sum(diag(tab))/sum(tab)
[1] 0.1561778
> p_test <- predict(mymodel, tst)
> tabb <- table(p_test, tst$Attrition)
> 1-sum(diag(tabb))/sum(tabb)
[1] 0.1628565
```

In the last step of the Principal Component Analysis, we need to consider how accurate the model is in predicting new data. In other words, when using the model in predicting the attrition condition of a new set of employees, how strongly they pretend to leave the company. We get a 15.6% error rate in the training dataset by calculating the misclassification error and a 16.2% error rate in the testing dataset. In other words, for any new dataset, our PCA model will provide about 85% accuracy of prediction in Attrition.

2.2 Linear Discrimination Analysis

2.2.1 Non-Scaling LDA

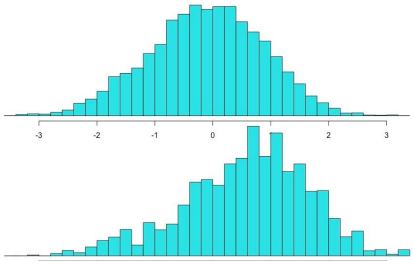
After we applied PCA on our dataset, we used another dimension reduction technology LDA. Look at the output on the left side below; it is clear that the current employee is about 84%, and the rest are voluntary resignations. The "Group Means" is the means in each group of the independent variables. These are what LDA is trying to separate those means. Then we got the actual computation of the coefficients of the linear discriminant. Moreover, the formula would be -0.066 times age plus 0.035 times distance from home and so forth. We can see there are strong relationships between the dependent variable and PercentSalaryHike, JobInvolvement, HourlyRate, and StockOptionLevel. Since we only have one component, we did not get a proportion of trace.

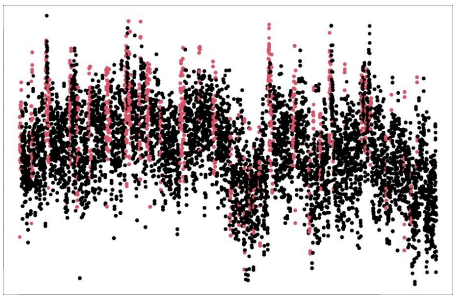
```
Prior probabilities of groups:
        0         1
0.1582759 0.8417241

Group means:
       Age DistanceFromHome EnvironmentSatisfaction logHourlyRate JobInvolvement JobLevel JobSatisfaction NumCompaniesWorked
0 33.71977        10.510076               2.594499      4.144607       2.619281 1.853486        2.595316           2.928649
1 37.52207         8.942186               2.742319      4.132428       2.750973 2.104721        2.753021           2.644152
  logPercentSalaryHike StockOptionLevel TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
0             2.679266        0.6097495              2.655501        2.718137       6.140795           3.509532               2.032952
1             2.698178        0.8285539              2.824406        2.768896       7.178462           4.363939               2.208265
  YearsWithCurrManager
0             3.579248
1             4.233152
Coefficients of linear discriminants:
                                LD1
Age                      0.06617132
DistanceFromHome        -0.03545971
EnvironmentSatisfaction  0.24782997
logHourlyRate           -0.44001807
JobInvolvement           0.40997637
JobLevel                 0.16657077
JobSatisfaction          0.20859075
NumCompaniesWorked      -0.06412468
logPercentSalaryHike     0.60414436
StockOptionLevel         0.42682205
TrainingTimesLastYear    0.15477313
WorkLifeBalance          0.13227236
YearsAtCompany          -0.02907330
YearsInCurrentRole       0.09075408
YearsSinceLastPromotion -0.05608601
YearsWithCurrManager     0.03807279
```

From the table below, loadings, which the component contributes, we can see the linear discriminant has the mostly Hourly rate on the negative side. On the positive side, we got 0.4 times job involvement, 0.4 times stock options level, and 0.6 times percentage salary hike.

```
        logHourlyRate      NumCompaniesWorked YearsSinceLastPromotion       DistanceFromHome          YearsAtCompany
           -0.44001807            -0.06412468             -0.05608601           -0.03545971           -0.02907330
YearsWithCurrManager                     Age      YearsInCurrentRole        WorkLifeBalance   TrainingTimesLastYear
            0.03807279             0.06617132              0.09075408             0.13227236             0.15477313
              JobLevel        JobSatisfaction EnvironmentSatisfaction          JobInvolvement        StockOptionLevel
            0.16657077             0.20859075              0.24782997             0.40997637             0.42682205
logPercentSalaryHike
            0.60414436
```

After we applied it to the training dataset and looked at the histogram below, there is a massive overlap with a quilt bit more confusion between these two groups. Furthermore, plot the transformed data, we can see the classification below. There is a very slight separation, but most categories overlap, as we got from the histogram. To test overfitting, we reserved a test set and training set.





2.2.2 Scaling LDA

```
Prior probabilities of groups:
       0       1
0.158028 0.841972

Group means:
       Age DistanceFromHome EnvironmentSatisfaction logHourlyRate JobInvolvement JobLevel JobSatisfaction NumCompaniesWorked
0 33.61200        10.569042               2.577566      4.146561       2.620184 1.845892        2.606887           2.960450
1 37.55071         8.928521               2.745633      4.130892       2.752672 2.102067        2.757471           2.650413
  logPercentSalaryHike StockOptionLevel TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
0             2.676790        0.6266621              2.668599        2.718377       6.112854           3.504944                2.019093
1             2.697336        0.8215268              2.821975        2.768990       7.163371           4.367185                2.208933
  YearsWithCurrManager
0             3.566996
1             4.233186
```

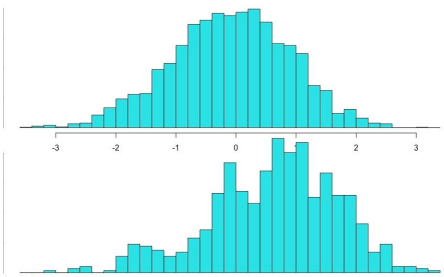
Coefficients of linear discriminants:

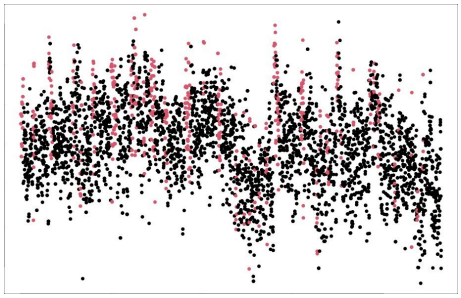
	LD1
Age	0.06788411
DistanceFromHome	-0.03639222
EnvironmentSatisfaction	0.27262953
logHourlyRate	-0.50917822
JobInvolvement	0.41255268
JobLevel	0.17103396
JobSatisfaction	0.19588034
NumCompaniesWorked	-0.07245684
logPercentSalaryHike	0.64332032
StockOptionLevel	0.37661561
TrainingTimesLastYear	0.14059526
WorkLifeBalance	0.12420802
YearsAtCompany	-0.03431667
YearsInCurrentRole	0.09058620
YearsSinceLastPromotion	-0.05255021
YearsWithCurrManager	0.04251593

```
logHourlyRate NumCompaniesWorked YearsSinceLastPromotion DistanceFromHome YearsAtCompany
   -0.43356011        -0.06450862             -0.06184281      -0.03897913   -0.03466500
YearsWithCurrManager         Age      YearsInCurrentRole    WorkLifeBalance TrainingTimesLastYear
          0.04942304    0.06232772              0.09432567         0.13073370          0.16455091
       JobSatisfaction    JobLevel EnvironmentSatisfaction       JobInvolvement     StockOptionLevel
          0.19241426    0.21041778              0.24723163         0.38014283          0.41779109
logPercentSalaryHike
          0.51230354
```

After separate data into the training and test dataset, We got the sample of 84 percent and the rest for the test set. After we built the model on the training set, look at the output. We got a similar output from the parameters and loadings compared to the non-scaling outputs. Things are looking pretty stable at this stage.

The histogram and classification plot are also very similar. We did a little better than the previous one, but we cannot separate the data with more details. This attrition dataset is not very separable. There are many variables in similar distances with similar sizes and with similar magnitudes, and the size of objects is enormous. In N-dimensional space, these classes separate themselves not very well.





```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8420   0.3403   0.4900   0.6279   1.4200

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.998966   0.403232  -4.957 7.15e-07 ***
Age                       0.048365   0.002541  19.036  < 2e-16 ***
DistanceFromHome         -0.022966   0.002471  -9.294  < 2e-16 ***
EnvironmentSatisfaction   0.181166   0.020582   8.802  < 2e-16 ***
logHourlyRate            -0.337726   0.068932  -4.899 9.61e-07 ***
JobInvolvement            0.274427   0.028796   9.530  < 2e-16 ***
JobLevel                  0.123037   0.024347   5.053 4.34e-07 ***
JobSatisfaction           0.136294   0.018776   7.259 3.90e-13 ***
NumCompaniesWorked       -0.043395   0.008028  -5.406 6.46e-08 ***
logPercentSalaryHike      0.375689   0.092566   4.059 4.94e-05 ***
StockOptionLevel          0.325617   0.026674  12.207  < 2e-16 ***
TrainingTimesLastYear     0.113591   0.016459   6.901 5.15e-12 ***
WorkLifeBalance           0.067245   0.028948   2.323 0.020181 *
YearsAtCompany           -0.027127   0.007045  -3.851 0.000118 ***
YearsInCurrentRole        0.072599   0.009741   7.453 9.14e-14 ***
YearsSinceLastPromotion  -0.038306   0.008727  -4.389 1.14e-05 ***
YearsWithCurrManager      0.023597   0.010048   2.348 0.018856 *
---
```

```
(Intercept)            -0.005172425 -0.04180397 -0.07518149 -0.101270057 -0.123644021 -0.144031240 -0.1626083 -0.179535862 -0.194960548
Age                     .           .           .           .           .           .           .          .           .
DistanceFromHome        .           .           .           .           .           .           .          .           .
EnvironmentSatisfaction .           .           .           .           .           .           .          .           .
logHourlyRate           .           .           .           .           .           .           .          .           .
JobInvolvement          .           .           .           .           .           .           .          .           .
JobLevel                .           .           .           .           .           .           .          .           .
JobSatisfaction         .           .           .           .           .           .           .          .           .
NumCompaniesWorked      .           .           .           .           .           .           .          .           .
logPercentSalaryHike    .           0.01359305  0.02597860  0.037168522 0.047333825 0.056596479 0.0650367  0.072727538 0.079735557
StockOptionLevel        .           .           .           .           .           .           .          .           .
TrainingTimesLastYear   .           .           .          -0.001456041 -0.003253428 -0.004891219 -0.0063836 -0.007743489 -0.008982656
WorkLifeBalance         .           .           .           .           .           .           .          .           .
YearsAtCompany          .           .           .           .           .           .           .          .           .
YearsInCurrentRole      .           .           .           .           .           .           .          .           .
YearsSinceLastPromotion .           .           .           .           .           .           .          .           .
YearsWithCurrManager    .           .           .           .           .           .           .          .           .

(Intercept)            -0.212285047 -0.233662101 -0.265232518 -0.294000262 -0.320214156 -0.344100904 -0.36586707 -0.38570086
Age                     .           .           .           .           .           .           .          .
DistanceFromHome        .           .           .           .           .           .           .          .
EnvironmentSatisfaction .           .           .           .           .           .           .          .
logHourlyRate           .           0.000723603 0.004307163 0.007572550 0.010548018 0.013259308 0.01572987 0.01798109
JobInvolvement          0.001119932 0.003021683 0.004718519 0.006264706 0.007673621 0.008957453 0.01012731 0.01119330
JobLevel                .           .           .           .           .           .           .          .
JobSatisfaction         .           .           .           .           .           .           .          .
NumCompaniesWorked      .           .           .           .           .           .           .          .
logPercentSalaryHike    0.086190752 0.092136598 0.097589261 0.102557961 0.107085647 0.111211454 0.11497104 0.11839691
StockOptionLevel        .           .           .           .           .           .           .          .
TrainingTimesLastYear  -0.010102657 -0.011115829 -0.012038544 -0.012879369 -0.013645572 -0.014343774 -0.01498001 -0.01555977
WorkLifeBalance         .           .           .           .           .           .           .          .
YearsAtCompany          .           .           .           .           .           .           .          .
YearsInCurrentRole      .           .           .           .           .           .           .          .
YearsSinceLastPromotion .           .           .           .           .           .           .          .
YearsWithCurrManager    .           .           .           .           .           .           .          .
```

2.3 Logistic Regression

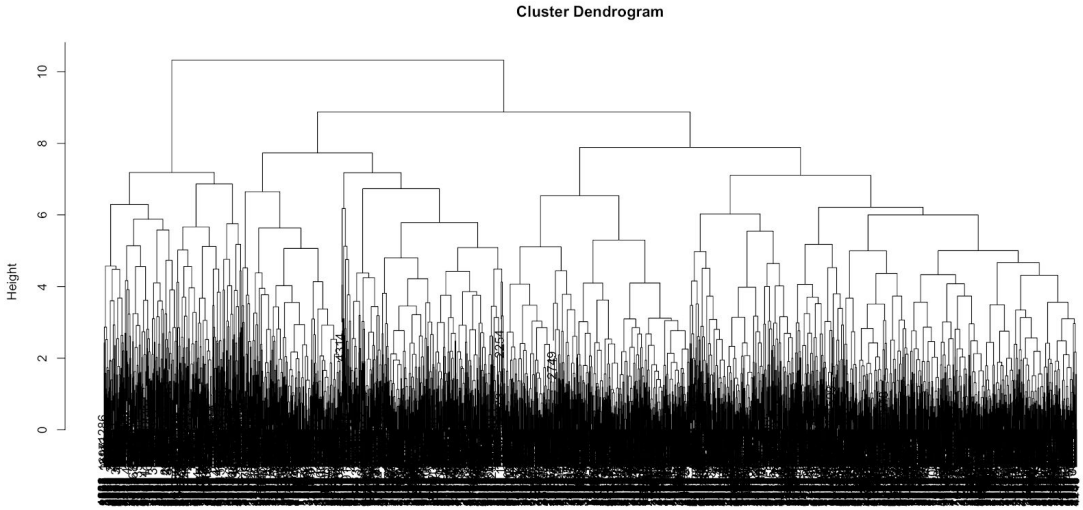
After we applied PCA and LDA, we also did logistic regression; first, we used a general linear model function with a binomial family. From the result on the left side, we got almost every variable that is significant. The accuracy is coming out to 84.6 percent. However, too many influencing factors can consume a considerable amount of human and financial resources. Since we needed a more precise result, we decided to use logistic regression with LASSO to study the data further.

Since LASSO can help us to further select variables, we tried logistic regression with LASSO. From the table above, we can see at lambda.max, there are four independent variables that are significant, including HourlyRate, JobInvolvement, PercentSalaryHike, and TrainingTimesLastYear. As the lambda goes down, four variables drop to one, PercentSalaryHike, as we can find the same result from the result table on the left. By checking the accuracy of predicting test dataset with the lambda.min, we got about 24%.

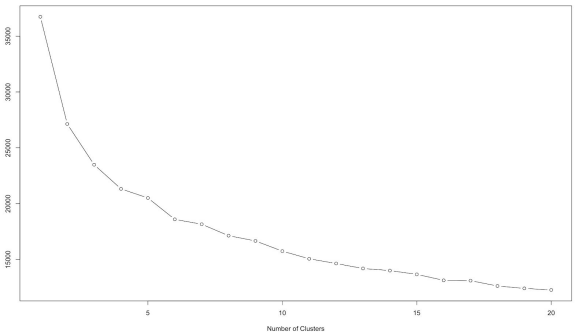
2.4 Cluster Analysis

We also decided to perform the Cluster Analysis on a subset of our dataset. Cluster Analysis is an essential and useful technique that places objects into groups(clusters), suggested by the data, not defined as a priority. Objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. It helps us to identify homogeneous groups of cases if the grouping is not previously known. This feature will provide us a new visualized way to understand our data and be helpful for analysis. We chose a sub-dataset that included the response variable: Attrition and independent variables: Age, Education, EmployeeNumber, HourlyRate, JobLevel, MonthlyIncome, TotalWorkingYears. This dataset includes the essential features(income, education, age, etc.) of an employee. We think it is useful for prediction and analysis. We also decided to normalize our dataset since it contains a large variety of values. Nevertheless, we will also consider un-normalized data analysis, too.

2.4.1 Cluster Dendrogram



First of all, we calculated the distances between each observation and generated the Cluster Dendrogram plot. Because this is hierarchical clustering, the algorithm finds which one is the closest in terms of distance. So the lowest level of the plot is many, many small groups of two observations. (our dataset is too big so that we can see the detail of the lower-level groups), at level 10 (the highest), the total observations are grouped by one cluster.



*2.4.2 Scree Plo*t

The Scree Plot gives us an overview of all possible clusters and within the group sum of squares. We want to reduce the within-cluster variability, so there are significant drops when we go from one cluster to two clusters. Somewhere starting from 4 or 5 clusters, the curve is going to be stable. In our case, we decided to choose 4 clusters for future analysis, based on the Scree Plot.

2.4.3 Cluster Means

2.4.3.1 Cluster Means with Scaled data

From the result of Cluster Means we can get the average values for the four clusters for each variable with Normalized Values. So this will help us in characterizing these four clusters. If we can't see too much variation among these four averages for a variable, that variable is not really playing a very significant role in deciding cluster membership. For example, look at Total Working Years in Group2, it shows a high value 1.72 and some negative values in other groups. Obviously, TotalWorkingYears have a significant impact on

	Group.1	Age	Education	EmployeeNumber	HourlyRate	JobLevel	MonthlyIncome
1	1	34.64045	2.946228	11156.33	67.03130	2.359551	7411.644
2	2	43.19018	3.004601	10626.98	64.54448	4.082822	15605.252
3	3	39.70733	3.176696	12684.56	68.64168	1.520788	4184.242
4	4	29.42330	2.146482	10756.58	58.05421	1.302191	3536.055

	TotalWorkingYears	YearsInCurrentRole
1	12.925361	6.670947
2	24.877301	6.397239
3	8.178337	2.772976
4	5.897347	2.318339

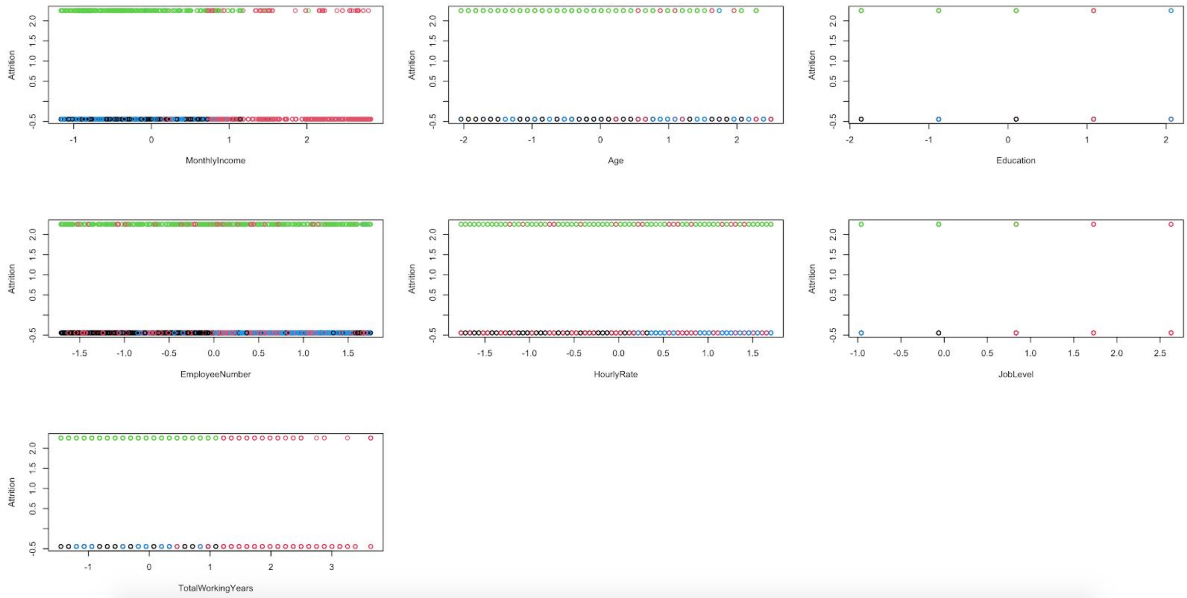
clustered membership. So, in other words, employees in groups 3 have longer total working years. Whereas, employees belonging to group 4 have shorter total working years.

2.4.3.2 Cluster Means with un-Scaled data

Group.1	Age	Education	EmployeeNumber	HourlyRate	JobLevel	MonthlyIncome	TotalWorkingYears	YearsInCurrentRole
1	-0.2426444	0.04992674	-0.06742062	0.06976085	0.2589799	0.1790472	0.1934314	0.6689866
2	0.6812384	0.10712488	-0.14540734	-0.05356926	1.8043641	1.8990136	1.7150973	0.5930672
3	0.3048818	0.27575492	0.15772767	0.14962542	-0.4932004	-0.4984348	-0.4109379	-0.4122095
4	-0.8064088	-0.73371874	-0.12631431	-0.37544420	-0.6892318	-0.6344993	-0.7013430	-0.5383139

We can also use the original value, which makes the interpretation more user-friendly. For example, employees of older age belong to group 2, and younger employees belong to group4. Also, employees who belong to group 2 at an older age with the highest income per month.

2.3.4 Result of Cluster Analysis



For this project, our goal is predicting the employees' features related to the dependent variable, Attrition. From the scatter plots above, the Cluster algorithm separated the observation into four clusters, showing with different colors. We can see that employees who have higher income monthly pretend to stay. Companies with larger numbers of employees(big companies) have a lower attrition rate. Employees with lower Education, hourly rate, lower job level, and a small number of total working years are a trend to contribute to the company's attrition.

Conclusion

In this project, we used four different techniques to analyze and predict our data. They are Principal Component Analysis, Linear Discriminant Analysis, Logistic Regression, and Cluster Analysis.

From PCA, we found that employees at a younger age(25-34), who work in a lower job lever(level1) and have less working experience, seem likely to leave. Employees who feel less environment satisfaction and receiving lower hourly rates show a strong trend to leave. From LDA, we found people who voluntarily resign, are not satisfied with their salary hike percent, hourly rate, and stock level. Employees' job involvement also has a substantial impact on the response variable, Attrition. From Logistic Regression with LASSO, when lambda has reached the max, there are four significant independent variables, including HourlyRate, JobInvolvement, PercentSalaryHike, and TrainingTimesLastYear. As the lambda goes down, four variables eventually drop to one, PercentSalaryHike, which means salary hike percent greatly influences employees' voluntary resignation. The results collected from Cluster Analysis is that Companies with larger numbers of employees(big branch companies) have a lower attrition rate. Employees with lower Education, lower hourly rate, lower job level, and a small number of entire working years contribute to its attrition the most. On the other hand, employees who enjoy higher monthly income show a strong trend of stay.

In sum, there are three aspects affect the dependent variable, Attrition, the most:

- Years - age, total working years
- Income - monthly, hourly, stock, salary hike, job level
- Satisfaction - environment satisfaction, job involvement

Appendix

2.1 Yueting Zhao

I play the role of Team Member in this project, majorly focused on Principal Component Analysis and Cluster Analysis, and discussed the data cleaning/preparation, conclusion with the other team member. Besides the content which is presented in this report, I also tried several other techniques, like applying PCA on different sub-dataset especially with highly correlated variables, ordinal analysis, logistic regression, stepwise selection based on lasso, ElasticNet and logistic regression, visualization of single variables.

Based on the successful analysis on the HR dataset, both PCA and Cluster analysis draw very clear and abundant results. Especially the Cluster Analysis, the strong power of its visualization of the distance of each pair of variables by color groups, I can easily tell how the differences of the employees features in different groups, in other words, each group collects the similar feature of the observations, this is very helpful for data analysis. Because of the goal of this project, I only focused on how the employee features affect the attrition, for example, employees who have higher monthly income show a strong trend of stay. Companies with larger numbers of employees(big branch companies) have a lower attrition rate. Employees with lower Education, lower hourly rate, lower job level, and a small number of entire working years contribute to its attrition the most. The interpretation of PCA is a little complicated, I used the factor rotation to cut off some of the less important variables in each PC, so the loading contributions are much clearer, and a lot easier for understanding the hidden information. I received results like employees at a younger age, who work in a lower job lever and have less working experience, seem likely to leave. Employees who feel less environment satisfaction and receiving lower hourly rates show a strong trend to leave.

I learn a lot from this project. Practicing is the best way to learn. In order to do a good job on this project, I reviewed the whole course all over again, tried each technique which could help me analyze and predict this dataset. This process gives a solid foundation of when facing a new dataset, I have a general idea that which technique could help. And what kind of results I can gain from those various techniques. Then from the practice of analyzing the dataset which is a large one with over 30 columns, I met several significant problems. Categorical/numerical, binary/ordinal, those all bring me some issue when applying the analysis. After data cleaning, which dataset I should choose, cause me a lot of time. Based on what I have learnt from this course, I decided to use PCA to capture the most information of all the numerical variables. Because Principal Component Analysis brings the advantage to reduce the data's dimensionality by selecting the essential features that capture maximum information about the dataset. It also helps in overcoming the overfitting issue by reducing the number of features. Next problem is that my response variable is categorical and with two levels only. It cannot be applied directly into the PCA dataset, since PCA only works for numerical variables. After searching, I found a way to fit the categorical response variable and all selected PCs into a linear model. So that I have a way to understand the underlying correlation among each variable and the target variable.

Another benefit of fitting the linear model with PCs is I can calculate for the misclassification error, which gives me a direct understanding of how accurate this model is for new data analysis. By the calculation on both training and testing dataset, I received about 83% prediction accuracy, which is pretty good.

During learning data science, I have been concerned with a couple of the issues like how can I predict the values of response variables based on a set of independent variables. And how can fit a model to analyze the relationships between them, in order to understand what is happening in the dataset. In this class, the Principal component analysis and factor analysis provide me the new angle of predicting the data, which is the biggest gain from this class. Also the factor rotation even gives the way of cutoff the less important variables, and making the interpretation only focus on the fewer variables and make things many times easier.From Cluster analysis, it opens a new window and lets me have a way to discover new categories of data that I haven't explicitly measured. So that I can name them and distinguish them and understand more detail in the dataset. The idea of cluster analysis is to identify how similar and different the data are to each other. and even better, the cluster analysis is very strong at visualization, by calculating the “distance” of each observation and plotting them in a two-demesion surface, which is a very user-friendly view to let us see what is the difference between the observations by groups. This class is covering a lot of data analysis knowledge, which is worth me to review time by time. and each time I review those models, they provide me something “new”, like a deeper understanding of a small aspect or gaining some new idea of data analysis.

2.2 Serena Yang

During this project analysis progress, I mainly did data cleaning and completed basic model building, used Stepwise selection on the basic model, checked model overfitting problem, utilized Ordinal Correlation, applied Linear Discrimination Analysis, Logistic Regression, and Receiver Operating Characteristic. In a good data analysis process, visualization is an important and efficient way to interpret data. I performed visualizations in Linear Discrimination Analysis, Ordinal Correlation, and Receiver Operating Characteristic in the algorithms that I was responsible for.

From the primary model building stage, I realized that I could not just consider the general regression model in the primary model building, as in this project, we also need to consider the logistic regression model. When we encounter a large dataset, I should check the correlation in several different ways based on the dataset situation first to select the best contribution variables to save time and cost. After I built the basic model, I utilized Stepwise selection to check the variables selected. After I improved the final model, I used RMSE to check whether the model has the overfitting problem.

While my teammate analyzed the data using PCA, I used LDA to interpret the data from a different perspective. In this process, I reviewed the lectures repeatedly to understand further and learn every step code of LDA. Even though the final result, whether scaling or not, is not remarkably consistent with our expectations, we could get several general ideas from the product that people who voluntarily resign are not satisfied with their salary hike percent, hourly rate, and stock level. Employees' job involvement also has a substantial impact on the response variable. When I found that more techniques are needed to analyze the data further, I decided to apply Ordinal Correlation to check the relationship between each ordinal and dependent variable and then using Logistic Regression since our respondent variable is binary.

After I converted each ordinal variable into factor level and generated a correlation plot in the Ordinal Correlation step, there are only two variables, stocks option level and marital status have a strong relationship. However, there is no evidence showing that marital status has a strong relationship with our respondents variable after deeper analysis. Therefore, I decided to keep the variable stocks option level to get into the next technique step.

In the Logistic Regression stage, I realized that the general regression model uses a longer time to execute the code than the Logistic Regression model. I first generate all selected variables into the general logistic regression model with a specific basis for selecting variables. Very surprisingly, I got a perfect result with all the variables significant, and by using the model to predict on the test dataset, I got about 80% accuracy. After I discussed with my teammate, we found that it takes a lot of time and energy to further verify each variable. For this reason, I built a logistic regression model with LASSO since LASSO could help select variables more precisely. During this process, I encountered a series of problems; the code that I found from lectures and online did not fit our model since the dataset is too large and the model has several independent variable types. By analyzing the code of each step and searching every error in R on

the Internet, I finally got the final result. When lambda has reached the max, there are four significant independent variables, including HourlyRate, JobInvolvement, PercentSalaryHike, and TrainingTimesLastYear. As the lambda goes down, four variables eventually drop to one, PercentSalaryHike, which means salary hike percent greatly influences employees' voluntary resignation. In Logistic Regression step, I learnt a lot, first I learnt several way to predict the value of the dependent variable based on independent variables.

From this course, I learned more about multicollinearity and overfitting problems and how to alleviate them. By applying what I have learned to this project, to further deal with the data that is different situations from class and homework. Compared with the previous lesson, I have learned these two concepts on a deeper level. Moreover, I learned a few more analysis arithmetic to analyze different categories and how to visualize the data to interpret the data from different angles.