Predicting Employee Job Performance

1. Introduction

This report aims to use statistical learning techniques to analyse a comprehensive dataset and create a predictive model for the outcome of interest. The analysis first explores the respective data in order to gain a better understanding of the information contained within the dataset. It then goes on to create a statistical model suitable for answering the formulated research question. As a final step, the chosen model is assessed for its predictive accuracy, and the model is adjusted based on the results of the assessment.

The dataset in question holds demographic and job-related information about 1470 employees at a single company. Demographic data recorded includes information about gender, age, education, income, and marital status. Additionally, the dataset holds records of the length and types of employees' careers, their performance at the company, and their satisfaction with their career. Both qualitative and quantitative data were collected, but variables recorded in a qualitative format were transformed into numeric factors in order to conduct a valid analysis. A summary of the variables in the formatted dataset and their respective descriptions are included in table 1 in appendix 1.

2. Research Question

The variables recorded in the employee dataset contain significant information about employee characteristics and job features. A variable that is especially intriguing is that which is derived from employee performance evaluations. The binary outcome, identified as $Employee\_ExcExpectations$, takes the value of 1 if the employee 'exceeded expectations' in their last performance evaluation, and 0 if they either just 'met expectations' or were 'inconsistent'. Even more interesting is that of 1470 employees, only 243 were able to achieve this label. Therefore, the question naturally arises: what distinguished these 243 employees from the remaining 1227?

Economic research and intuition tells us that an individual's success in the workplace is determined by their intrinsic ability and motivation, among other things. Unfortunately, these features are not directly observable, and when an attempt to observe them is made, it is usually done with error. Due to this, indirect estimates of these features are often used in determining whether an employee will be successful in their work. An example of this is job-market signalling, where individuals use information like their education level to signal their ability to employers. Therefore, individual demographics and job features may indirectly affect the outcome of performance evaluations conducted by employers, through an effect on or correlation with employee ability and performance itself. This analysis aims to determine whether the features available in the data can be used to predict the outcome of employee performance evaluations. The formulated research question is stated formally below.

Research Question: Are employee demographics and job features helpful determinants of whether employees will 'exceed expectations' in performance evaluations?

3. Methods

The empirical analysis required to answer the research question is conducted in three stages. The first consists of an exploratory analysis of the formatted dataset, where principal component analysis (PCA) is used to visualise the data in two dimensions. The second stage involves setting up an appropriate model using regression techniques to predict the outcome of interest. The third builds on this analysis by first assessing the chosen model for its robustness as a predictive tool, and then using LASSO regression analysis to enhance the predictive accuracy of the model. All three methods and their relevance to the data and research question are described briefly below.

3.1 Exploration – Principal Component Analysis

Where data is high-dimensional, PCA is used to express the dataset in fewer dimensions by identifying correlated variables that seem to be partly recording the same information. This tool can help identify a smaller set of variables that capture as much information in the dataset as possible, i.e., identify linear combinations of the variables that have the largest variance. The method aims to maximise

$$Var(Z_1) = \frac{\sum_{i=1}^{n}(z_{i1} - \bar{z}_1)^2}{n},$$

where $z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$ is the score of the ith observation on the first principal component, and ϕ_{j1} are the loadings of the variable on the component. The tool is also helpful in visualising data in order to gain some preliminary insights and determine areas for further analysis. As the research question requires identification of features that support the prediction of employee performance, combining variables may have a counter-intuitive effect on the results. Therefore, in this analysis, PCA is used solely for visualisation of the data, and inferences are drawn from the plot of the first two principal components.

3.2 Predictive Modelling – Logistic Regression Model

Logistic regressions are popular in the modelling of a binary outcome, such as the outcome of interest in this analysis, $Employee\_ExcExpectations$. They estimate the log of the odds ratio as a linear combination of the explanatory variables, i.e.,

$$\log - \text{odds ratio (logit)} = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where the odds ratio is the probability of the outcome being realized against the probability that it isn't. All available variables are used in the regression, and the parameter on each variable is interpreted as the change in the log of the odds ratio for a unit change in the respective variable. However, the analysis conducted in this report is mainly concerned with the predictive accuracy of the model rather than the interpretation of the relationships found. The regression is therefore estimated to minimise the classification error in the data using the intuition behind the Bayes Classifier, which allocates observations to its most likely class given the inputs used, i.e., classifies an employee as 'exceeding expectations' in their performance evaluation if $P(Y = 1|X) > 0.5$. The formula used to obtain this probability is described below, where Y is the variable $Employee\_ExcExpectations$, and parameters are estimated via maximum likelihood in order to identify the model that is most likely to have produced the given data, based on the assumption that the underlying model is, in fact, logit.

$$\widehat{Pr}(Y = 1) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p}}$$

3.3 Model Assessment & Adjustment – Cross-validation & LASSO Regression

Statistical models are often subject to the bias-variance trade-off, where bias is an error introduced by estimating a simple model not representative of a complex underlying reality, and variance is introduced due to the sensitivity of the estimated model to the training set used, i.e., overfitting. Cross-validation aims to assess the variance of the model by estimating the same model on resampled training sets and examining the average misclassification rate on validation sets. A low average misclassification rate in the validation sets would suggest that the model has good predictive accuracy outside of the training sets and is likely not overfitted to the data it was estimated on. The method used in this analysis is k-fold cross-validation, which resamples the data k times, training the model on a distinct set of $((k-1)/k)^{th}$ of the n observations each time, and using the remaining observations as a validation set in which predictions are made. $k = 10$ is used in this analysis. Moreover, shrinkage methods such as the Lasso regression can lower the variance of model estimates by minimising the residual sum of squares of the model and applying a shrinkage penalty to the absolute values of the parameters. The optimal choice for this penalty is dependent on the value that leads to the lowest cross-validation error. Shrinkage improves model prediction accuracy and can also be helpful in subset selection, where parameters that shrink to 0 can be removed from the model. However, as in any case, a lower variance in the model leads to an increase in bias, a risk taken to possibly improve the model's predictive capabilities.

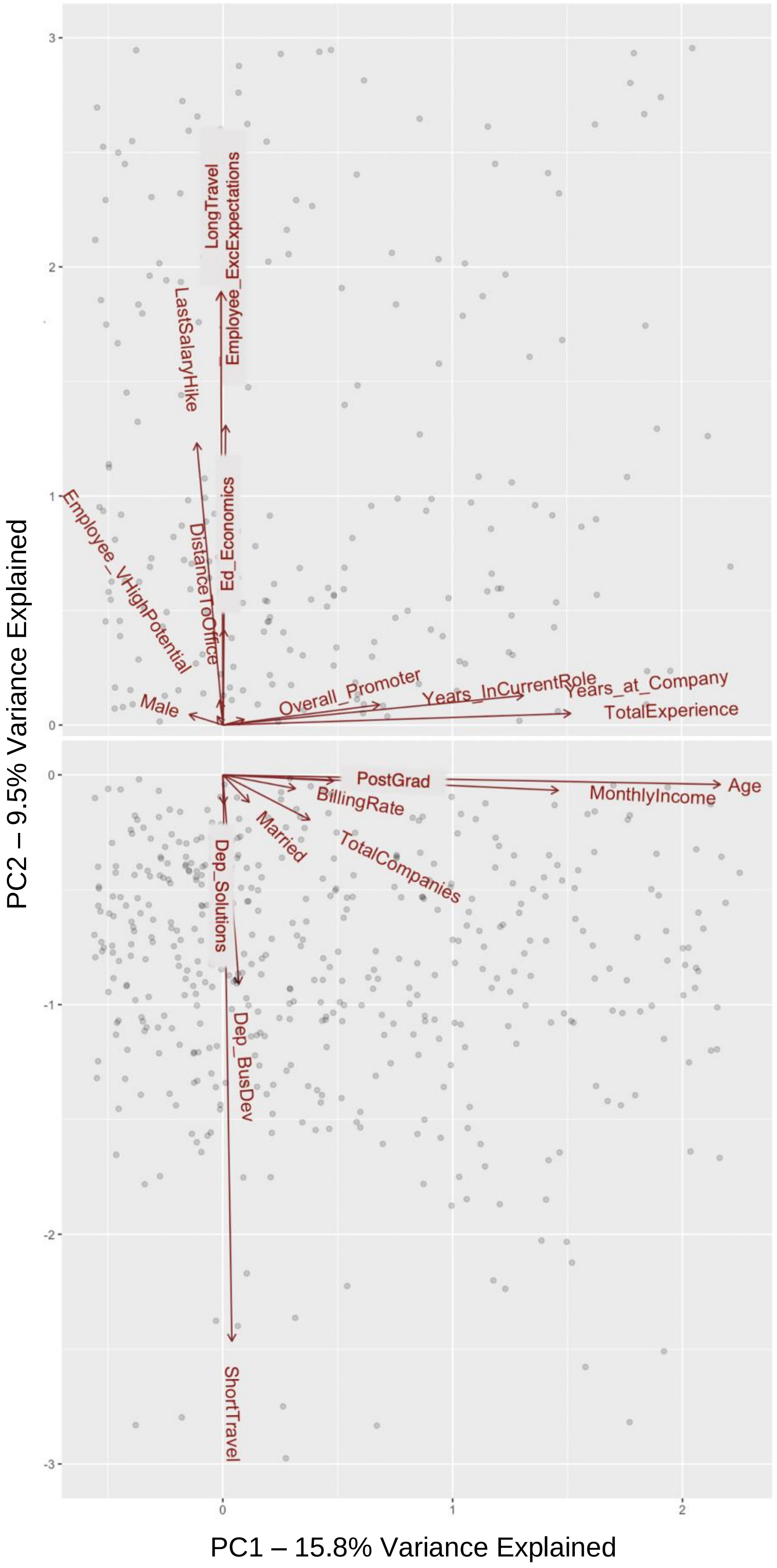
4. Results

4.1 Principal Component Analysis

Figure 1 plots the first principal component against the second, where the points represent the scores of the observations on the principal components, and the arrows represent the loadings of the respective variable on the principal components. The length and direction of the arrows in the plot can be used to identify correlated variables in the data. Variables projected in the same direction are positively correlated with each other, and variables of greater length have greater loadings on the respective principal component, e.g., variables with a greater horizontal length help explain a larger amount of variance in the data through principal component 1.

The first principal component shows correlation between the length of an employee's career, their age, and their monthly income. Intuitively, it can be seen that these variables partly measure the same thing: experience. However, none of these variables seem to be correlated with the variable of interest, performance evaluation outcomes. Looking at the plot alone, it can be inferred that employee performance may be better predicted through features like the employee's last salary hike, their degree subject, the department they work in, and the travel commitments they make for the company.

Figure 1 – Principal Component Biplot



4.2 Logistic Regression Analysis

Table 2 reports the parameter estimates and the in-sample misclassification rate of the logistic model. The results show that only the parameter on the $LastSalaryHike$ variable is significantly different from zero at the 5% level. Intuitively, it can be argued that $LastSalaryHike$ and $Employee\_ExcExpectations$ have a simultaneous effect on each other, as an employee's salary hike is likely affected by their performance just as performance is likely affected by salary changes, introducing a bias that was not accounted for in the analysis. Regardless, only a select few of the features provided in the data seem to be significant in helping determine the outcomes of employee performance evaluations. As well as that, the low in-sample misclassification rate of 94.6% may be the result of an

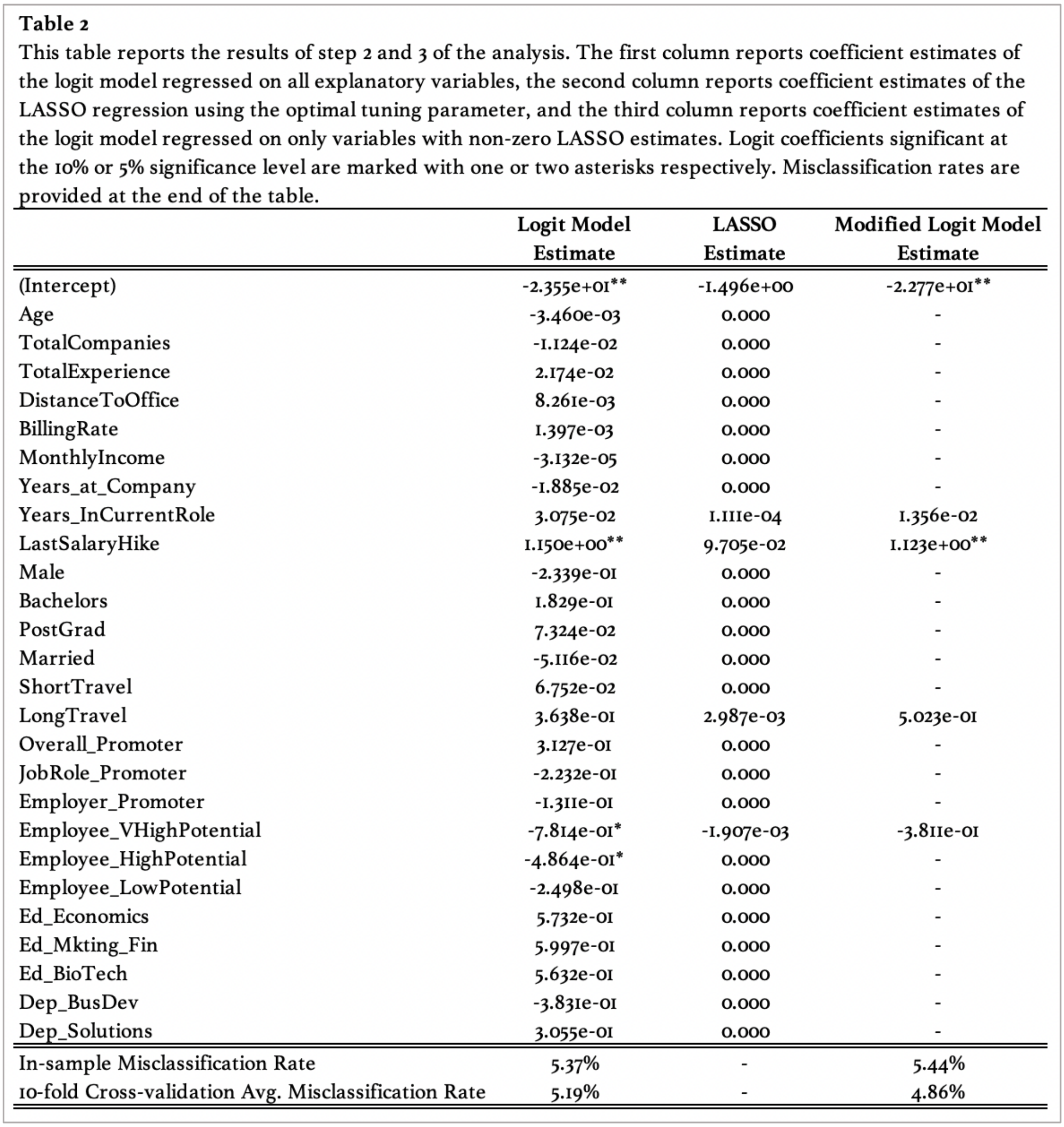


Table 2

This table reports the results of step 2 and 3 of the analysis. The first column reports coefficient estimates of the logit model regressed on all explanatory variables, the second column reports coefficient estimates of the LASSO regression using the optimal tuning parameter, and the third column reports coefficient estimates of the logit model regressed on only variables with non-zero LASSO estimates. Logit coefficients significant at the 10% or 5% significance level are marked with one or two asterisks respectively. Misclassification rates are provided at the end of the table.

	Logit Model Estimate	**LASSO Estimate**	**Modified Logit Model Estimate**
(Intercept)	-2.355e+01**	-1.496e+00	-2.277e+01**
Age	-3.460e-03	0.000	-
TotalCompanies	-1.124e-02	0.000	-
TotalExperience	2.174e-02	0.000	-
DistanceToOffice	8.261e-03	0.000	-
BillingRate	1.397e-03	0.000	-
MonthlyIncome	-3.132e-05	0.000	-
Years_at_Company	-1.885e-02	0.000	-
Years_InCurrentRole	3.075e-02	1.111e-04	1.356e-02
LastSalaryHike	1.150e+00**	9.705e-02	1.123e+00**
Male	-2.339e-01	0.000	-
Bachelors	1.829e-01	0.000	-
PostGrad	7.324e-02	0.000	-
Married	-5.116e-02	0.000	-
ShortTravel	6.752e-02	0.000	-
LongTravel	3.638e-01	2.987e-03	5.023e-01
Overall_Promoter	3.127e-01	0.000	-
JobRole_Promoter	-2.232e-01	0.000	-
Employer_Promoter	-1.311e-01	0.000	-
Employee_VHighPotential	-7.814e-01*	-1.907e-03	-3.811e-01
Employee_HighPotential	-4.864e-01*	0.000	-
Employee_LowPotential	-2.498e-01	0.000	-
Ed_Economics	5.732e-01	0.000	-
Ed_Mkting_Fin	5.997e-01	0.000	-
Ed_BioTech	5.632e-01	0.000	-
Dep_BusDev	-3.831e-01	0.000	-
Dep_Solutions	3.055e-01	0.000	-
In-sample Misclassification Rate	5.37%	-	5.44%
10-fold Cross-validation Avg. Misclassification Rate	5.19%	-	4.86%

overfitted model, as the model's out-of-sample performance has not yet been analysed. The robustness of these results is therefore examined further.

4.3 Cross-validation & LASSO

Table 2 also reports the results of the k-fold cross-validation analysis. The low average misclassification rate of 5.2% across the different resampled sets of data suggests the model is not highly sensitive to differences in training sets and is a reliable predictive tool. A plot of the misclassification rates for each k can be seen in Figure 2 in appendix 1. However, in order to answer the research question and improve this variance further, the results of the LASSO regression reported in Table 2 are analysed, and non-zero estimates are only found across four variables: $Years\_InCurrentRole$, $LastSalaryHike$, $LongTravel$, and $Employee\_VHighPotential$. The logistic regression is run again using just these variables. A classification rate of 94.6% is found, and 10-fold cross-validation provides a slightly lower average misclassification rate of 4.9%, also seen in Table 2. These results provide suggestive evidence that these four variables are most helpful in determining the outcome of performance evaluations.

5. Limitations

The results of this analysis suggest that certain information about employees can help determine their prospective performance at the company. However, the analysis faces certain limitations in making robust conclusions. The first is the limited amount of observations used. The data is collected across only one firm, and any conclusions made cannot be generalised across other firms. As well as that, additional variables that better measure employee skill and output are needed for a more intuitive analysis. This limitation also extends to the PCA, as the assumptions made in PCA are not appropriate for binary or count data, and the analysis could be improved if more continuous variables were recorded. Therefore, the analysis is limited by the no. of observations and types of variables observed.

The method of analysis itself can also be improved. As mentioned earlier, the $LastSalaryHike$ variable may be simultaneously related to $Employee\_ExcExpectations$, introducing a simultaneity bias. A two-stage approach would need to be applied for a valid analysis. As well as that, the model estimated may have high bias due to its non-complexity.

There may have been non-linear relationships in the model that were not accounted for, and variables like employee output not included in the regression may have caused an omitted variable bias, reducing the model's predictive capability outside of the sample. A variety of more flexible models using more data need to be tested in order to obtain the model that achieves the optimal bias-variance trade-off and better answers the research question.

6. Conclusion

The analysis conducted in this report provides intuitive information about the dataset and the outcome of interest, $Employee\_ExcExpectations$. The PCA provided an opportunity to draw inferences about what variables would be most important in the predictive model, and the inferences made were mostly supported by the estimation and assessment of the logit model, where only $Years\_InCurrentRole$, $LastSalaryHike$, $LongTravel$, and $Employee\_VHighPotential$ were found to have non-zero coefficients.

The report concludes that for the employees of this firm, only a few of the recorded features help determine whether they will 'exceed expectations' in performance evaluations, and employee demographics are not as important in the prediction of this outcome.

Appendix 1 – Supplementary Tables & Figures

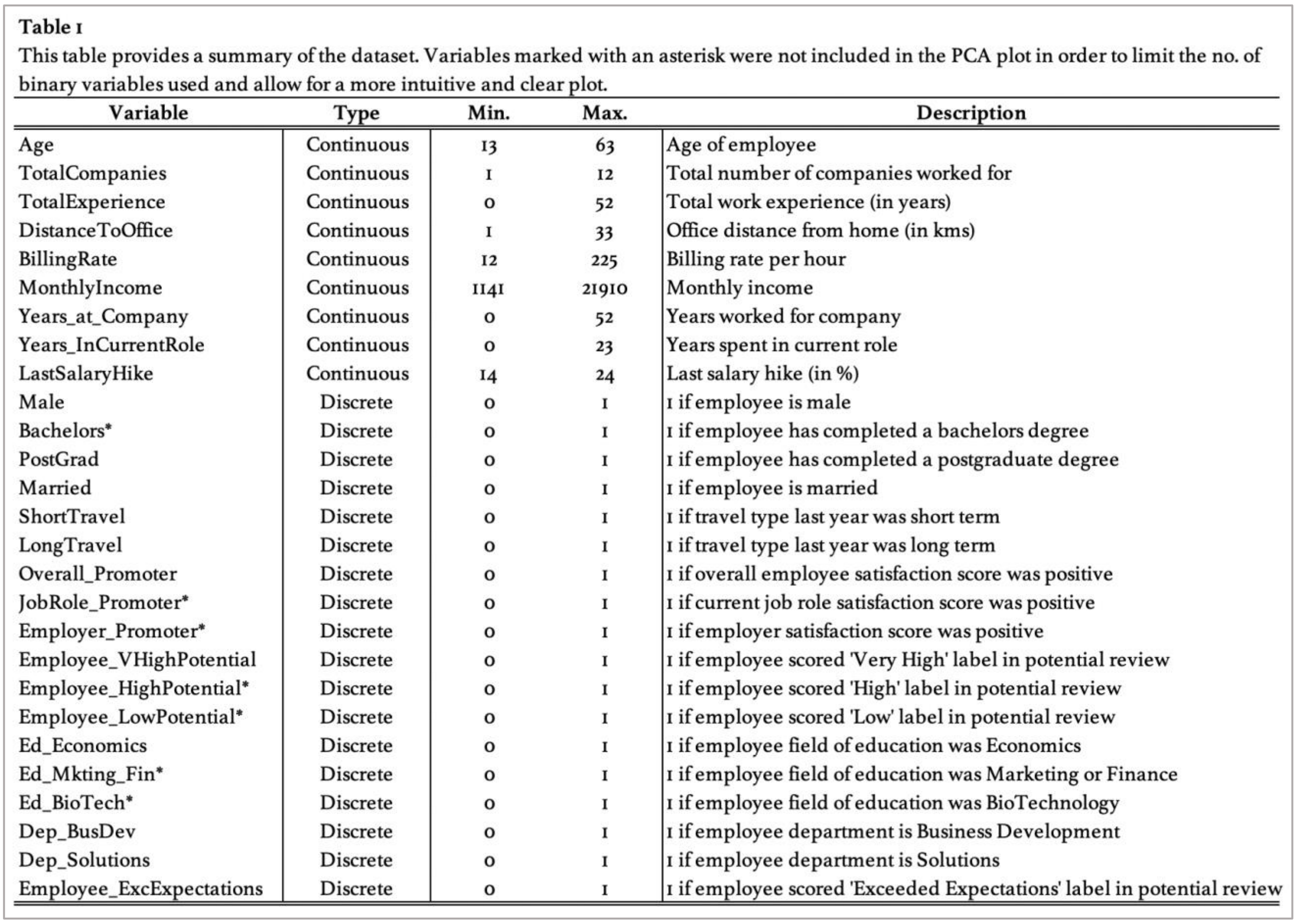
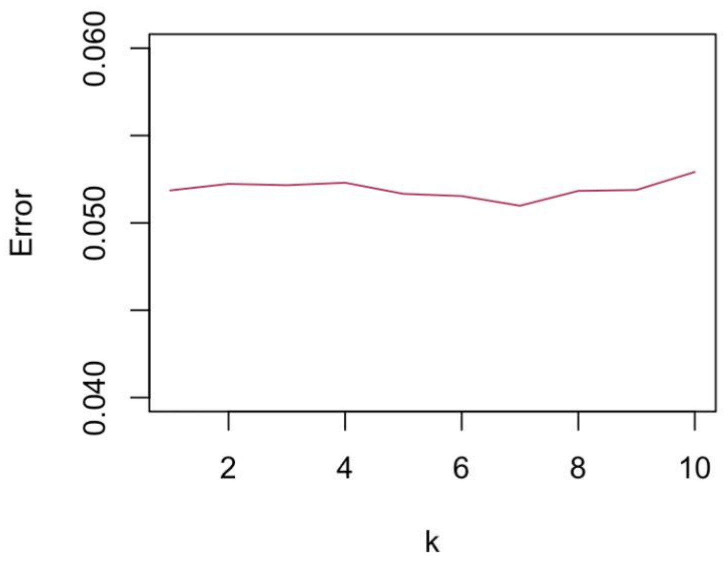


Table 1

This table provides a summary of the dataset. Variables marked with an asterisk were not included in the PCA plot in order to limit the no. of binary variables used and allow for a more intuitive and clear plot.

Variable	Type	Min.	Max.	Description
Age	Continuous	13	63	Age of employee
TotalCompanies	Continuous	1	12	Total number of companies worked for
TotalExperience	Continuous	0	52	Total work experience (in years)
DistanceToOffice	Continuous	1	33	Office distance from home (in kms)
BillingRate	Continuous	12	225	Billing rate per hour
MonthlyIncome	Continuous	1141	21910	Monthly income
Years_at_Company	Continuous	0	52	Years worked for company
Years_InCurrentRole	Continuous	0	23	Years spent in current role
LastSalaryHike	Continuous	14	24	Last salary hike (in %)
Male	Discrete	0	1	1 if employee is male
Bachelors*	Discrete	0	1	1 if employee has completed a bachelors degree
PostGrad	Discrete	0	1	1 if employee has completed a postgraduate degree
Married	Discrete	0	1	1 if employee is married
ShortTravel	Discrete	0	1	1 if travel type last year was short term
LongTravel	Discrete	0	1	1 if travel type last year was long term
Overall_Promoter	Discrete	0	1	1 if overall employee satisfaction score was positive
JobRole_Promoter*	Discrete	0	1	1 if current job role satisfaction score was positive
Employer_Promoter*	Discrete	0	1	1 if employer satisfaction score was positive
Employee_VHighPotential	Discrete	0	1	1 if employee scored 'Very High' label in potential review
Employee_HighPotential*	Discrete	0	1	1 if employee scored 'High' label in potential review
Employee_LowPotential*	Discrete	0	1	1 if employee scored 'Low' label in potential review
Ed_Economics	Discrete	0	1	1 if employee field of education was Economics
Ed_Mkting_Fin*	Discrete	0	1	1 if employee field of education was Marketing or Finance
Ed_BioTech*	Discrete	0	1	1 if employee field of education was BioTechnology
Dep_BusDev	Discrete	0	1	1 if employee department is Business Development
Dep_Solutions	Discrete	0	1	1 if employee department is Solutions
Employee_ExcExpectations	Discrete	0	1	1 if employee scored 'Exceeded Expectations' label in potential review

Figure 2 – k-fold Cross-validation Error Plot



Appendix 2 – R Code

```
########################################
####        DATA & LIBRARIES        ####
library(pacman)
pacman::p_load(MASS, tidyverse, tidyr, dplyr, stats, ggplot2, gridExtra,
        ggbiplot, boot, ROCR, leaps, glmnet)
employee_dataset <- read_csv("employee_dataset.csv")
data <- separate(employee_dataset, col=1, into=c("EmployeeID", "Gender", "Age",
"Education", "EducationType", "MaritalStatus", "TotalCompanies", "TotalExperience",
"DistanceToOffice", "Department", "Traveltype_last_year", "BillingRate",
"MonthlyIncome", "Years_at_Company", "Years_InCurrentRole", "LastSalaryHike",
"PotentialReview", "PerformanceReview", "SatisfactionScore", "JobRole_SatisfactionScore",
"Overall_SatisfactionScore"), sep=";")
########################################
#######################################
#########  DATA FORMATTING   ########
data$Male <- ifelse(data$Gender=="Male", 1, 0)
data$Bachelors <- ifelse(data$Education=="Graduation" | data$Education=="Masters /
PHD", 1, 0)
data$PostGrad <- ifelse(data$Education=="Masters / PHD", 1, 0)
data$Married <- ifelse(data$MaritalStatus=="Married", 1, 0)
data$ShortTravel <- ifelse(data$Traveltype_last_year=="No" |
data$Traveltype_last_year=="LongTermProject", 0, 1)
data$LongTravel <- ifelse(data$Traveltype_last_year=="LongTermProject", 1, 0)
data$Overall_Promoter <- ifelse(data$Overall_SatisfactionScore=="Promoter", 1, 0)
data$JobRole_Promoter <- ifelse(data$JobRole_SatisfactionScore=="Promoter", 1, 0)
data$Employer_Promoter <- ifelse(data$SatisfactionScore=="Promoter", 1, 0)
data$Employee_VHighPotential  <- ifelse(data$PotentialReview=="Very High", 1, 0)
data$Employee_HighPotential  <- ifelse(data$PotentialReview=="High", 1, 0)
data$Employee_LowPotential  <- ifelse(data$PotentialReview=="Low", 1, 0)
data$Employee_ExcExpectations <- ifelse(data$PerformanceReview=="Exceed
Expectations", 1, 0)
data$Employee_IncExpectations <- ifelse(data$PerformanceReview=="Inconsistent", 1, 0)
#not included in dataset
```

```
data$Ed_Economics <- ifelse(data$EducationType=="Economics", 1, 0)
data$Ed_Mkting_Fin <- ifelse(data$EducationType=="Marketing / Finance", 1, 0)
data$Ed_BioTech <- ifelse(data$EducationType=="Bio-technology", 1, 0)
data$Dep_BusDev <- ifelse(data$Department=="BusinessDevelopment", 1, 0)
data$Dep_Solutions <- ifelse(data$Department=="Support", 1, 0)
#dummies only created for k-1 possibilities for each variable to avoid dummy variable trap
########################################
######### NUMERIC DATA SUBSET ####
num_data <- data[c("Age", "TotalCompanies", "TotalExperience", "DistanceToOffice",
"BillingRate", "MonthlyIncome", "Years_at_Company", "Years_InCurrentRole",
"LastSalaryHike", "Male", "Bachelors", "PostGrad", "Married", "ShortTravel",
"LongTravel", "Overall_Promoter", "JobRole_Promoter", "Employer_Promoter",
"Employee_VHighPotential", "Employee_HighPotential", "Employee_LowPotential",
"Ed_Economics", "Ed_Mkting_Fin","Ed_BioTech", "Dep_BusDev", "Dep_Solutions",
"Employee_ExcExpectations")]
num_data[] <- lapply(num_data, function(x) as.numeric(as.character(x)))
sapply(num_data, class)
summary(num_data)
########################################
########################################
#######     EXPLORATION – PCA      #######
pca_data <- data[c("Age", "TotalCompanies", "TotalExperience", "DistanceToOffice",
"BillingRate", "MonthlyIncome", "Years_at_Company", "Years_InCurrentRole",
"LastSalaryHike", "Male", "PostGrad", "Married", "ShortTravel", "LongTravel",
"Overall_Promoter", "Employee_VHighPotential", "Employee_ExcExpectations",
"Ed_Economics", "Dep_BusDev", "Dep_Solutions")]
pca_data[] <- lapply(pca_data, function(x) as.numeric(as.character(x)))
sapply(pca_data, class)

pca4 <- prcomp(pca_data, rank=5, scale=TRUE)
biplot(pca4, col=c("grey", "black"), cex=c(0.2,0.9))
biplot1 <- ggbiplot(pca4, alpha=0.2, var.axes = TRUE, varname.size=4.7)
biplot1 <- biplot1 + xlim(-0.56, 2.25) + ylim(0, 3) +
  xlab("PC1 - 15.8% Variance Explained") + ylab("PC2 - 9.5% Variance Explained")
```

```
print(biplot1)
biplot2 <- ggbiplot(pca4, alpha=0.2, var.axes = TRUE, varname.size=4.7)
biplot2 <- biplot2 + xlim(-0.56, 2.25) + ylim(-3, 0) +
  xlab("PC1 - 15.8% Variance Explained") + ylab("PC2 - 9.5% Variance Explained")
print(biplot2)
grid.arrange(biplot1, biplot2, ncol=1, nrow=2, heights=c(2.5,2.5))
########################################
########################################
#######     MODELLING – LOGIT       #########
#LOGIT MODEL
logit <- glm(Employee_ExcExpectations ~ ., data=num_data, family=binomial)
summary(logit)
tab_model(logit)
#predict p(X)
logit_phat <- predict(logit, type="response")
#predict Y
logit_yhat <- ifelse(logit_phat>0.5, 1,0)
table(logit_yhat, num_data$Employee_ExcExpectations)
#misclassification rate
((table(logit_yhat, num_data$Employee_ExcExpectations)[2]+
    table(logit_yhat, num_data$Employee_ExcExpectations)[3])/
    sum(table(logit_yhat, num_data$Employee_ExcExpectations)))*100
########################################
########################################
####### TWIST – CROSS-VALIDATION  #########
#Logit - crossvalidation
cv.error.10 <- rep(0,10)
for (i in 1:10){
  logit.fit <- glm(Employee_ExcExpectations ~ ., data=num_data, family=binomial)
  cv.error.10[i] <- cv.glm(num_data, logit.fit, K=10)$delta[1]
}
cv.error.10
mean(cv.error.10) #avg. misclassification rate, quite low (5.2%)
```

```
plot(1:10, cv.error.10,
    type='l', col='maroon', ylim=c(0.04,0.06),
    xlab='k', ylab='Error')
#######################################
#######################################
#######  TWIST – LASSO   #########
#lambda search grid + set up variable matrices
grid <- 10^seq(10,-2,length=100)
x <- model.matrix(Employee_ExcExpectations ~ ., data=num_data)[,-1]
y <- num_data$Employee_ExcExpectations
#train and test subset
train <- sample(1:nrow(x), nrow(x)/2)
test <- (-train)
y.test <- y[test]
lasso.mod <- glmnet(x[train,], y[train], alpha=1, lambda=grid)
plot(lasso.mod, xvar="lambda") #coefficient values at different lambda values
cv.out <- cv.glmnet(x[train,], y[train], alpha=1)
plot(cv.out)
bestlam <- cv.out$lambda.min
lasso.pred <- predict(lasso.mod, s=bestlam, newx=x[test,])
mean((lasso.pred-y.test)^2)
#Fit model with best lambda
out <- glmnet(x, y, alpha=1, lambda=grid)
lasso.coef <- predict(out, type="coefficients", s=bestlam)[1:27,]
lasso.coef
lasso.coef[lasso.coef!=0] #nonzero variables

#LOGIT model with variable subsets
logit.lasso <- glm(Employee_ExcExpectations ~ Years_InCurrentRole + LastSalaryHike +
                  LongTravel + Employee_VHighPotential, data=num_data, family=binomial)
summary(logit.lasso)
#predict p(X)
logit.lasso_phat <- predict(logit.lasso, type="response")
#predict Y
```

```
logit.lasso_yhat <- ifelse(logit.lasso_phat>0.5, 1,0)
table(logit.lasso_yhat, num_data$Employee_ExcExpectations) #95% classification instead of 96%
#misclassification rate
((table(logit.lasso_yhat, num_data$Employee_ExcExpectations)[2]+
    table(logit.lasso_yhat, num_data$Employee_ExcExpectations)[3])/
    sum(table(logit.lasso_yhat, num_data$Employee_ExcExpectations)))*100

cv.error.10.2 <- rep(0,10)
for (i in 1:10){
  logit.lasso.fit <- glm(Employee_ExcExpectations ~ Years_InCurrentRole + LastSalaryHike +
                           LongTravel + Employee_VHighPotential, data=num_data, family=binomial)
  cv.error.10.2[i] <- cv.glm(num_data, logit.lasso.fit, K=10)$delta[1]
}
cv.error.10.2
mean(cv.error.10.2)
########################################
```