Project VI: GAM, MARS, and PPR

DS 5494 - Statistical Data Mining II

Willliam Ofosu Agyapong[*] University of Texas at El Paso (UTEP)

November 22, 2022

Contents

[*] woagyapong@miners.utep.edu

1 Introduction

In this project, we shall be considering various models including generalized additive models (GAM), multivariate adaptive regression splines (MARS), and projection pursuit regression (PPR). We aim, among other things, to **obtain a model with the best predictive power and also identify the important drivers of employee turnover or retention.**

We make use of a human resource data set concerning employee retention from one Kaggle data analytics competition. The data set contains 14,999 observations for 10 variables as shown in Table 1.

1.1 Data Description

```r
data.frame(
    rbind(
        c("satisfaction_level", "Satisfaction Level"),
        c("last_evaluation","Last evaluation"),
        c("number_project", "Number of projects"),
        c("average_montly_hours", "Average monthly hours"),
        c("time_spend_company  ", "Time spent at the company"),
        c("Work_accident       ", "Whether they have had a work accident"),
        c("left                ", "Whether the employee has left"),
        c("promotion_last_5years", "Whether had a promotion in the last 5 years"),
        c("sales               ", "Departments (column sales)"),
        c("salary              ", "Salary")
    )
) %>%
    kable(caption = "Varaibles and their meanings", booktabs = T,
          col.names = c("Variable name", "Description")) %>%
    kable_styling(latex_options = c("HOLD_position")) %>%
    kable_classic()
```

Table 1: Varaibles and their meanings

Variable name	Description
satisfaction_level	Satisfaction Level
last_evaluation	Last evaluation
number_project	Number of projects
average_montly_hours	Average monthly hours
time_spend_company	Time spent at the company
Work_accident	Whether they have had a work accident
left	Whether the employee has left
promotion_last_5years	Whether had a promotion in the last 5 years
sales	Departments (column sales)
salary	Salary

- **For the left, work accident and promotion in last 5 years binary variables, 0 and 1 should be interpreted as "No" and "Yes", respectively**.

1.2 Data Preparation

```r
# bring in the data
hr <- read.csv("HR_comma_sep.csv")
# dim(hr)
# head(hr)
# names(hr)

# 1. change the categorical variable salary to ordinal
# 2. change name for variable sales to department.
# 3. make left variable a factor variable
hr_new <- hr %>%
    mutate(salary = factor(salary, levels = c("low","medium","high"),ordered = T),
           left = as.factor(left)) %>%
    rename(department = sales)

# str(hr_new)
```

```
# get data types
output <- NULL
for(i in seq_along(hr_new)) {
  output <- rbind(output, c(names(hr_new)[i],
                  paste(class(hr_new[[i]]), collapse = " "),
                  length(unique(hr_new[[i]]))
                   )
                  )
}
output <- as.data.frame(output)

# checking for missing values
output %>% left_join(
naniar::miss_var_summary(hr_new), by=c("V1"="variable")) %>%
  kable(booktabs = T, linesep="", align = "lcc",
        col.names = c("Variable name", "Type", "levels", "Number missing", "Percent missing"),
        cap = "Data types and amount of missing values in the HR data") %>%
kable_styling(latex_options = c("HOLD_position"))
```

Table 2: Data types and amount of missing values in the HR data

Variable name	Type	levels	Number missing	Percent missing
satisfaction_level	numeric	92	0	0
last_evaluation	numeric	65	0	0
number_project	integer	6	0	0
average_montly_hours	integer	215	0	0
time_spend_company	integer	8	0	0
Work_accident	integer	2	0	0
left	factor	2	0	0
promotion_last_5years	integer	2	0	0
department	character	10	0	0
salary	ordered factor	3	0	0

We do not have any missing values in the data.

From the output above, it can be seen that among all the predictors, 2 are continuous, 5 variables are integer counts among which 3 (number_project, average_monthly_hours and time_spend_company) can be reasonably treated as continuous, while the other 2 together with department and salary will be treated as categorical variables.

2 Explaratory Data Analysis (EDA)

In this section, we explore the underlying data set with the hope of discovering any interesting patterns or insights to aid our understanding of the data.

```
# prepare data for EDA
hr_eda <- hr_new %>%
    mutate(left = as.factor(ifelse(left == 0, "No", "Yes")),
           Work_accident = as.factor(ifelse(Work_accident == 0, "No", "Yes")),
           promotion_last_5years =
               as.factor(ifelse(promotion_last_5years==0, "NO", "Yes")))
```

2.1 Part (a): How does satisfaction level relate to number of projects?

```
ggplot(hr_eda, aes(number_project, satisfaction_level, color=left)) +
    geom_point(alpha=0.7) +
    scale_color_brewer(palette = "Set1") +
    labs(x="Number project", y="Satisfaction level",
         title = "How does satisfaction level relate to number of projects?")
```

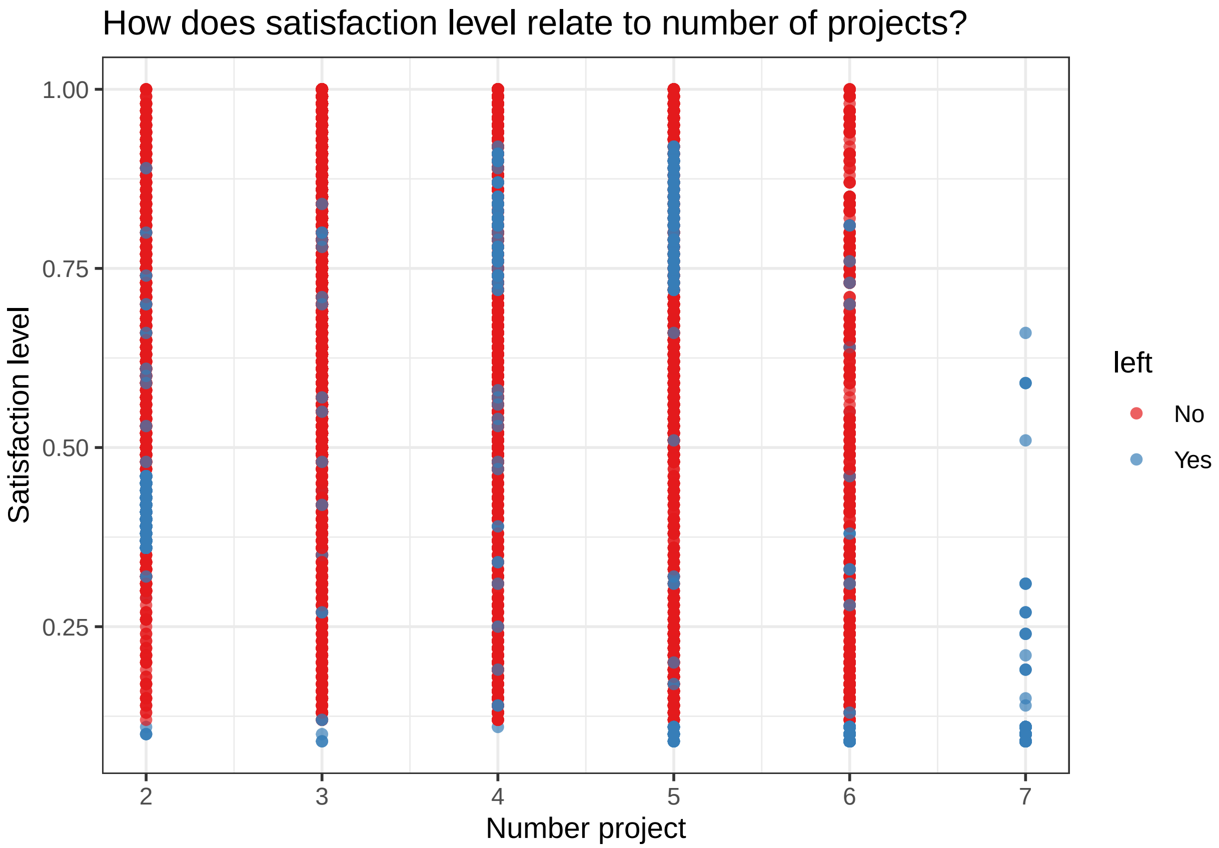


Figure 1: Relationship between satisfaction level and the number of projects.

In general, we do not observe any clear increasing or decreasing patterns between satisfaction level and number of projects as either of them increases or decreases. That is, there does not appear to be a linear relationship between satisfaction level and the number of project. However, it is seen that satisfaction level decreased with increasing number of projects after the 5^{th} project only for employees who left, indicating some kind of effect between the two variables.

As expected, employees who did not leave during the review period recorded the highest satisfaction levels. It turns out however that employees who worked on the most number of projects (7 projects) tend to have low satisfaction levels and all those workers also left the company.

2.2 Part (b): Association between variables

Since the data contain different types of variables, both continuous and categorical, the Goodman and Kruskal tau measure was used to investigate the association among the variables instead of the pearson correlation. Another benefit of this association measure is that it is sensitive to directional associations. Applying the `GKtauDataframe()` function to the HR data set yielded the association plot shown below, which brings to the light some interesting details.

```r
# install.packages("GoodmanKruskal")
library(GoodmanKruskal)
# data1<- diabetes %>% select(class)
dat <- GKtauDataframe(hr_new)
plot(dat, colorPlot=T)
```

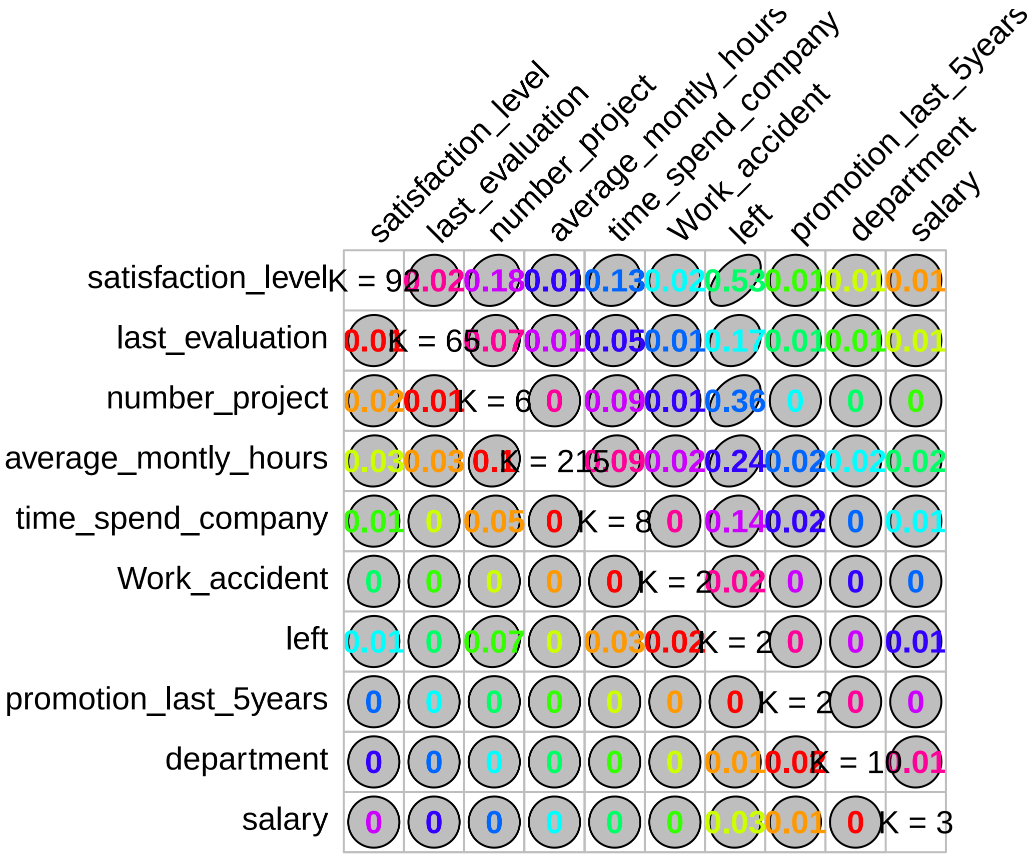


Figure 2: Association between variables using the Goodman and Kruskal tau measure.

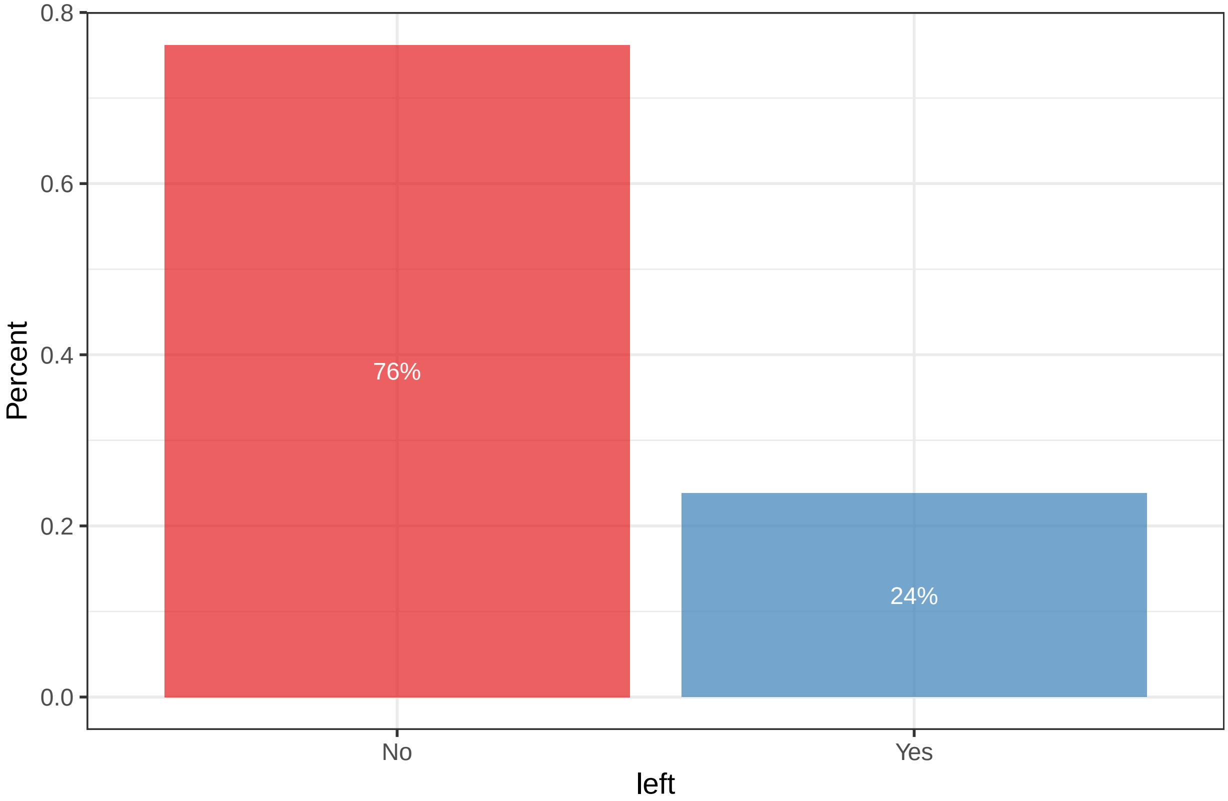
We observe from the 6^{th} column in Figure 2 that ***satisfaction level***, ***last evaluation***, ***number of projects***, ***average monthly hours***, and ***time spent at the company*** are moderately associated with employee turnover (left). The opposite is not true which confirms the asymmetric nature of the Goodman and Kruskal tau association values. This suggests that these variables have the ability to explain variation in employee turnover, a sign that such variables, especially satisfaction level and number of projects having the highest values of **0.53** and **0.36** respectively, are likely to be most influential in the modelling phase of the project.

Generally, there is no high association among the predictors, suggesting that multicollinearity will not be an issue for us. It is interesting to note that only satisfaction level and number of projects show some association with a tau value of **0.18** measuring the influence of satisfaction level on the number of projects employees work on. Note that the strength of the association in the reverse case is weaker for these two variables.

2.3 Other findings

2.3.1 Distribution of Target variable

```
library(scales)
hr_eda %>%
  group_by(left) %>%
  summarise(n=n()) %>%
  mutate(pct = n/sum(n),
         lbl = percent(pct)) %>%
ggplot(aes(left, pct,fill=left)) +
    geom_bar(stat = "identity",position = "dodge",alpha=0.7) +
    scale_fill_brewer(palette = "Set1") +
  geom_text(aes(label = lbl), size=3, color = "white",
            position = position_stack(vjust = 0.5)) +
   labs(y="Percent", title = "") +
    theme(legend.position = "none")
```



It can be observed that about 76% of employees stayed while 24% of employees left the company.

2.3.2 Distribution of continuous predictors by employee turnover

```
hr_eda %>%
    dplyr::select(-c(salary, department, Work_accident,promotion_last_5years)) %>%
    pivot_longer(-left, names_to = "variable", values_to = "value") %>%
    ggplot(aes(left, value, fill = left)) +
    geom_boxplot(alpha=0.7) +
    scale_fill_brewer(palette = "Set1") +
    facet_wrap(vars(variable),  scales = "free") +
    labs(x="", y="") +theme_bw() +
    theme(axis.text.x = element_blank(),
          axis.ticks.x = element_blank())
```

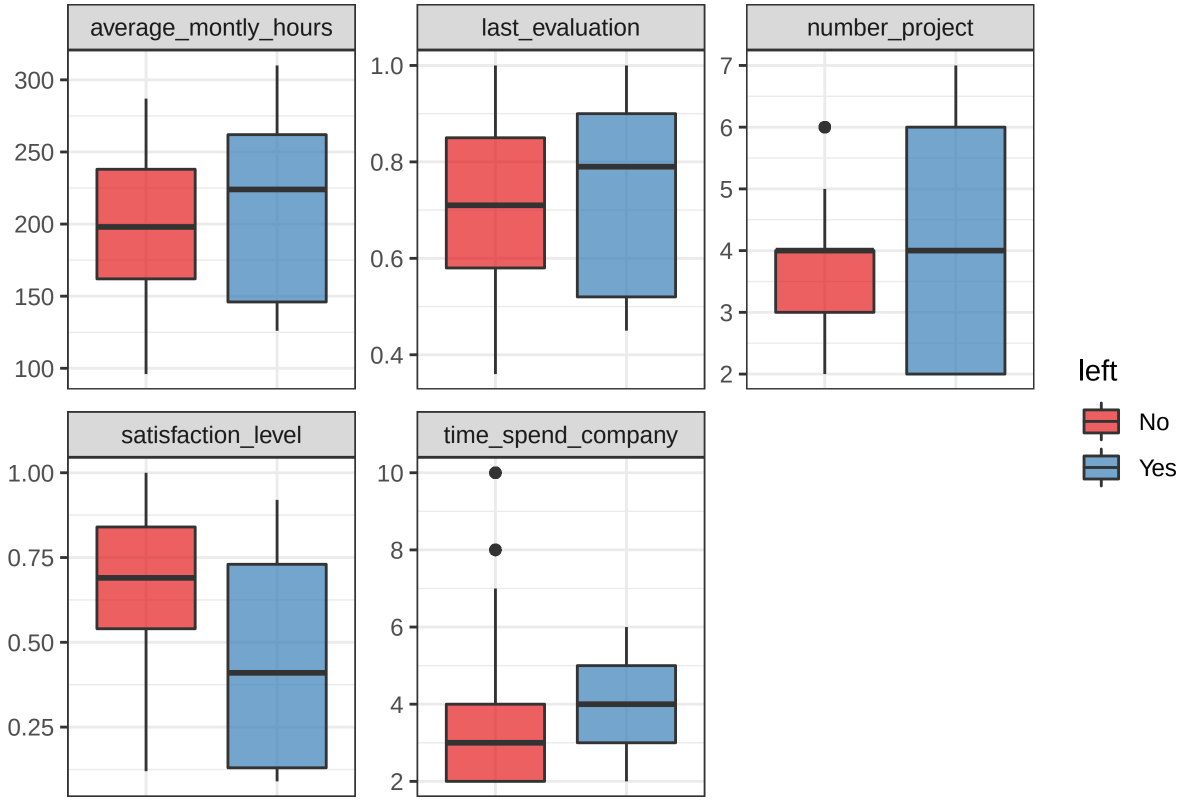


Figure 3: Distribution of continuous predictors by employee turnover (left)

From Figure 3, the following observations can be made:

- Except for time spent at the company, there is high variability in the measures for employees who left the company during the period under consideration compared to those who remained as indicated by the sizes of the box plots.
- On average, employees who left recorded higher values for average monthly working hours, last evaluation, and time spent at the company and low satisfaction levels.
- Older employees in the company have spent up to 10 years, suggesting that the company is relatively young. Employees who left the company typically spent about 4 years, while typical active employees have remained in the company for about 3 years.

2.3.3 Distribution of categorical predictors by employee turnover

```
plt1 <- ggplot(data = hr_eda, aes(x = department, fill=left)) +
  geom_bar(position = "dodge", alpha=0.7) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "",y = "Number of employees",
       title = "Department versus employee turnover") +
  theme(axis.text.x = element_text(angle = 45),
        legend.position = "right", plot.title = element_text(hjust = .5))
```

```
plt2 <- ggplot(data = hr_eda, aes(x = salary, fill=left)) +
  geom_bar(position = "dodge", alpha=0.7) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "",  y = "Number of employees",
       title = "Salary") + theme(legend.position = "none", plot.title = element_text(hjust = .5))

plt3 <- ggplot(data = hr_eda, aes(x = factor(Work_accident), fill=left)) +
  geom_bar(position = "dodge", alpha=0.7) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "", y = "",
       title = "Work accident") +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = .5))

plt4 <- ggplot(data = hr_eda, aes(x = factor(promotion_last_5years), fill=left)) +
  geom_bar(position = "dodge", alpha=0.7) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "", y = "",
       title = "Promotion in last\n 5 years") +
  theme(legend.position = "none", plot.title = element_text(hjust = .5))

plt1
```

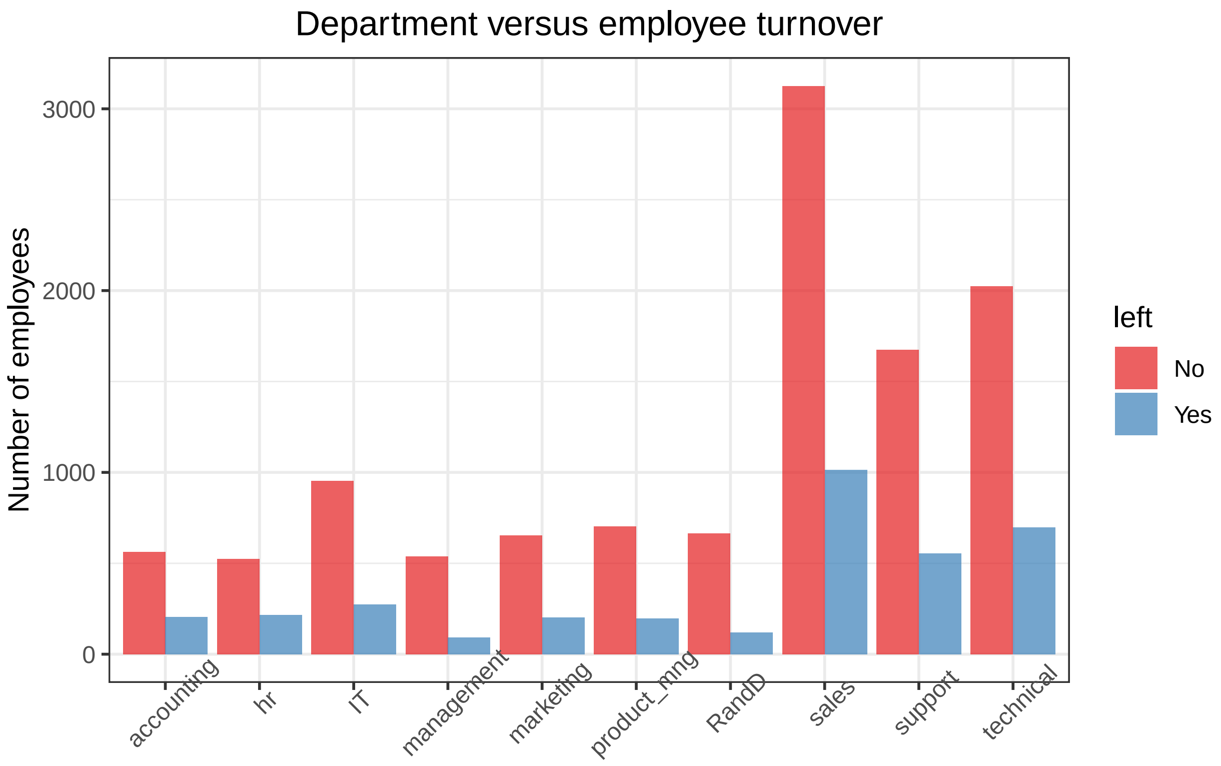


Figure 4: Distribution of department by employee turnover (left)

- Sales, technical, and support department were the top three departments to have employee turnover whiles the management department had the smallest amount of turnover. This could probably due to the fact that those three departments happen to have the largest number of employees.

```
(plt2 + plt3 + plt4)
```

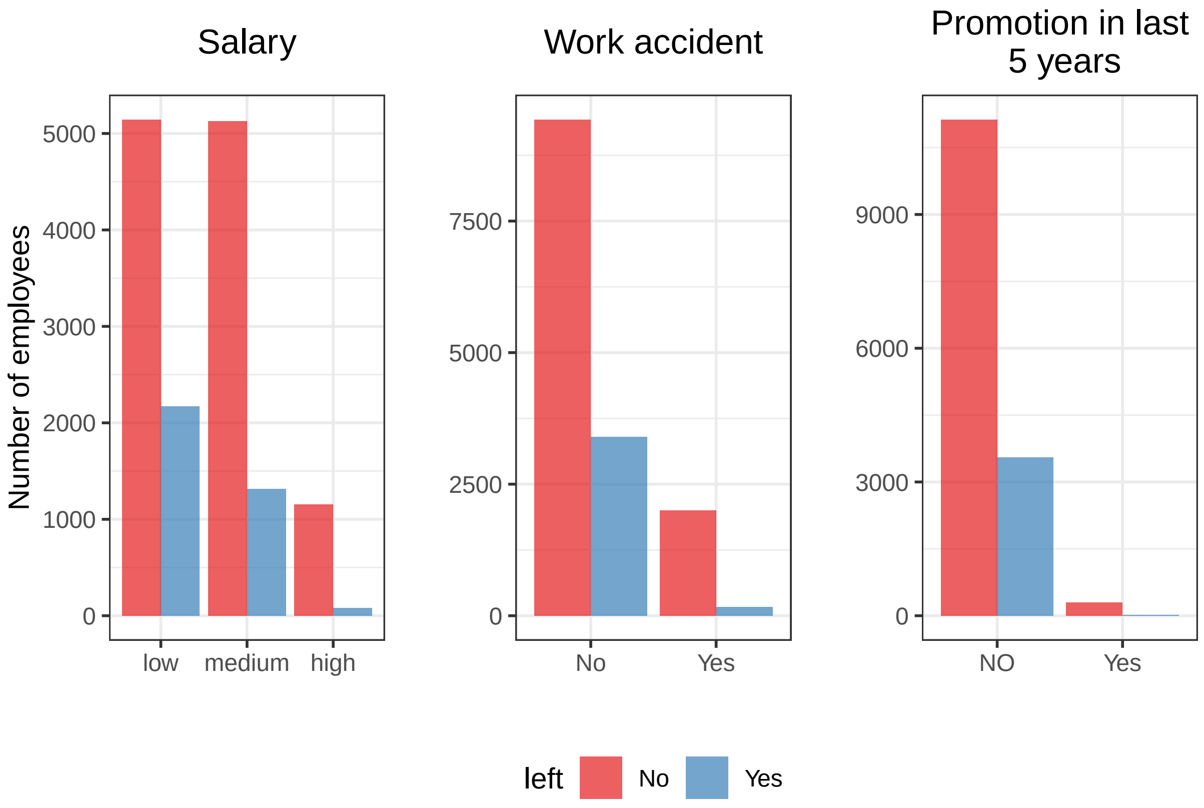


Figure 5: Distribution of other categorical predictors by employee turnover (left)

Figure 5 suggests that:

- Most employees receive low to medium salaries and the majority of employees who left the company belong to this same salary group, while few employees with high salaries left. This shows that salary levels have a high tendency to influence employee turnover.
- Majority are the employees who have never had a work related accident.
- Only a small proportion of employees got promoted in the last 5 years during the review period.

3 Data Partitioning: Randomly splitting data into training and test sets

For the purpose of model validation, we randomly partitioned the data D into the training set D_1 and the test set D_2 with a ratio of approximately 2:1 on the sample size. The output below shows that 9999 observations and 5000 observations were allocated to the resulting training and test sets, respectively. A ***9940*** seed was used to ensure reproducibility of results affected by random splitting of the data.

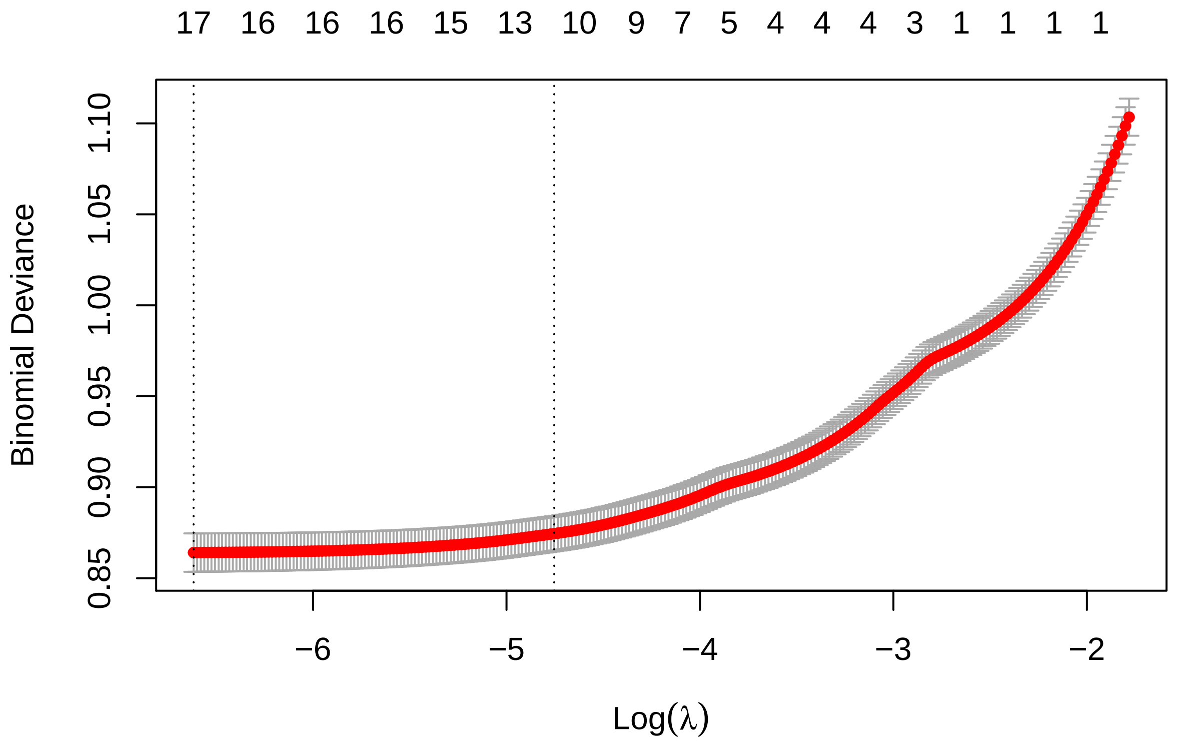
```
set.seed(9940)
ratio <- 2/3
train_ind <- sample(1:NROW(hr_new), size = NROW(hr_new)*ratio)
train_set <- hr_new[train_ind, ]
test_set <- hr_new[-train_ind, ]
dim(train_set); dim(test_set)
```

```
## [1] 9999   10
## [1] 5000   10
```

4 Logistic Regression: A baseline classifier

A 10-fold cross validation regularized logistic regression model with LASSO penalty was fitted to the training set D_1.

```
library(glmnet)
# select target and create the design matrix
y <- train_set$left
X <- model.matrix(~ . -left, data = train_set)
# fit the model
set.seed(125)
cv.lasso <- cv.glmnet(X, y, nfolds = 10, family="binomial", alpha=1, lambda.min=.0001,
                      thresh = 1e-07, nlambda=500, standardize=T, maxit=2000, type.measure = "deviance")
# best tuning parameter
best.lambda <- cv.lasso$lambda.1se

plot(cv.lasso)
```



The best tuning parameter λ was obtained as 0.0086 based on the minimum cross-validated such that error is within 1 standard error of the minimum. From the graph above we observe that 13 terms are selected with this choice of λ.

```
fit.lasso <- glmnet(x=X, y=y, family="binomial", alpha = 1, lambda=best.lambda, standardize = T, thresh =
fit.lasso$beta
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)            .
## satisfaction_level    -3.672146
## last_evaluation        0.267254
## number_project        -0.157487
## average_montly_hours   0.002461
## time_spend_company     0.187753
## Work_accident         -1.138058
## promotion_last_5years -0.706800
## departmenthr           0.010072
## departmentIT           .
## departmentmanagement  -0.205603
## departmentmarketing    .
## departmentproduct_mng  .
## departmentRandD       -0.213453
## departmentsales        .
## departmentsupport      .
## departmenttechnical    .
## salary.L              -0.788295
## salary.Q               .
```

Applying a zero (0) cutoff on the absolute values of the coefficients as a selection criterion, all 9 predictors will be retained in our final model. It is worthy to note that although the coefficient associated with some of the levels of `department` and `saalary` are zero (0), the nonzero ones justify their inclusion in the model. We therefore derive our final logistic regression model as follows:

```
logistic <- glm(left ~ ., data = train_set, family = "binomial")

# summary(logistic)

logistic %>% broom::tidy(conf.int=T,conf.level=0.95) %>%
    kable(booktabs=T, linesep="", caption = "Parameter estimates for the Logistic model")%>%
    kable_styling(latex_options =c("HOLD_position"))
```

Table 3: Parameter estimates for the Logistic model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3895	0.1880	-2.0716	0.0383	-0.7593	-0.0221
satisfaction_level	-4.1856	0.1204	-34.7641	0.0000	-4.4233	-3.9513
last_evaluation	0.8286	0.1823	4.5467	0.0000	0.4720	1.1865
number_project	-0.3075	0.0262	-11.7402	0.0000	-0.3591	-0.2564
average_montly_hours	0.0044	0.0006	7.0246	0.0000	0.0032	0.0057
time_spend_company	0.2561	0.0188	13.6486	0.0000	0.2194	0.2930
Work_accident	-1.4872	0.1098	-13.5387	0.0000	-1.7072	-1.2762
promotion_last_5years	-1.5708	0.3121	-5.0325	0.0000	-2.2273	-0.9947
departmenthr	0.2224	0.1619	1.3735	0.1696	-0.0950	0.5400
departmentIT	-0.1174	0.1479	-0.7940	0.4272	-0.4068	0.1733
departmentmanagement	-0.4618	0.1949	-2.3690	0.0178	-0.8480	-0.0832
departmentmarketing	-0.0654	0.1605	-0.4076	0.6836	-0.3803	0.2492
departmentproduct_mng	-0.1765	0.1590	-1.1103	0.2669	-0.4885	0.1351
departmentRandD	-0.5214	0.1766	-2.9528	0.0031	-0.8701	-0.1774
departmentsales	-0.0177	0.1250	-0.1414	0.8876	-0.2608	0.2293
departmentsupport	0.1199	0.1330	0.9018	0.3671	-0.1392	0.3823
departmenttechnical	0.0756	0.1307	0.5787	0.5628	-0.1789	0.3336
salary.L	-1.3699	0.1133	-12.0874	0.0000	-1.5994	-1.1543
salary.Q	-0.3670	0.0736	-4.9865	0.0000	-0.5145	-0.2256

From the above table of results, there is enough evidence at 5% significance level to conclude that all the predictors are statistically significant in the model as evidenced by the small p-values. Even though most of the levels for the department variable did not show significance, we consider the whole variable to be significant.

The sign of the coefficients show that most of the predictors including satisfaction level and number of projects have negative effect on turnover. From the odds table below, we learn that the estimated odds for satisfaction level is $e^{-4.1856} = 0.0152$, meaning for every unit increase in satisfaction level, the odds (likelihood) of an employee turnover decreases by a factor of 0.0152, holding all other factors constant. Similar interpretations can be made for the other variables.

```
exp(cbind(OR = coef(logistic), confint(logistic))) %>%
    kable(booktabs=T, linesep="", caption =
    "Odds ratio based on parameter estimates from the logistic model") %>%
    kable_styling(latex_options =c("HOLD_position"))
```

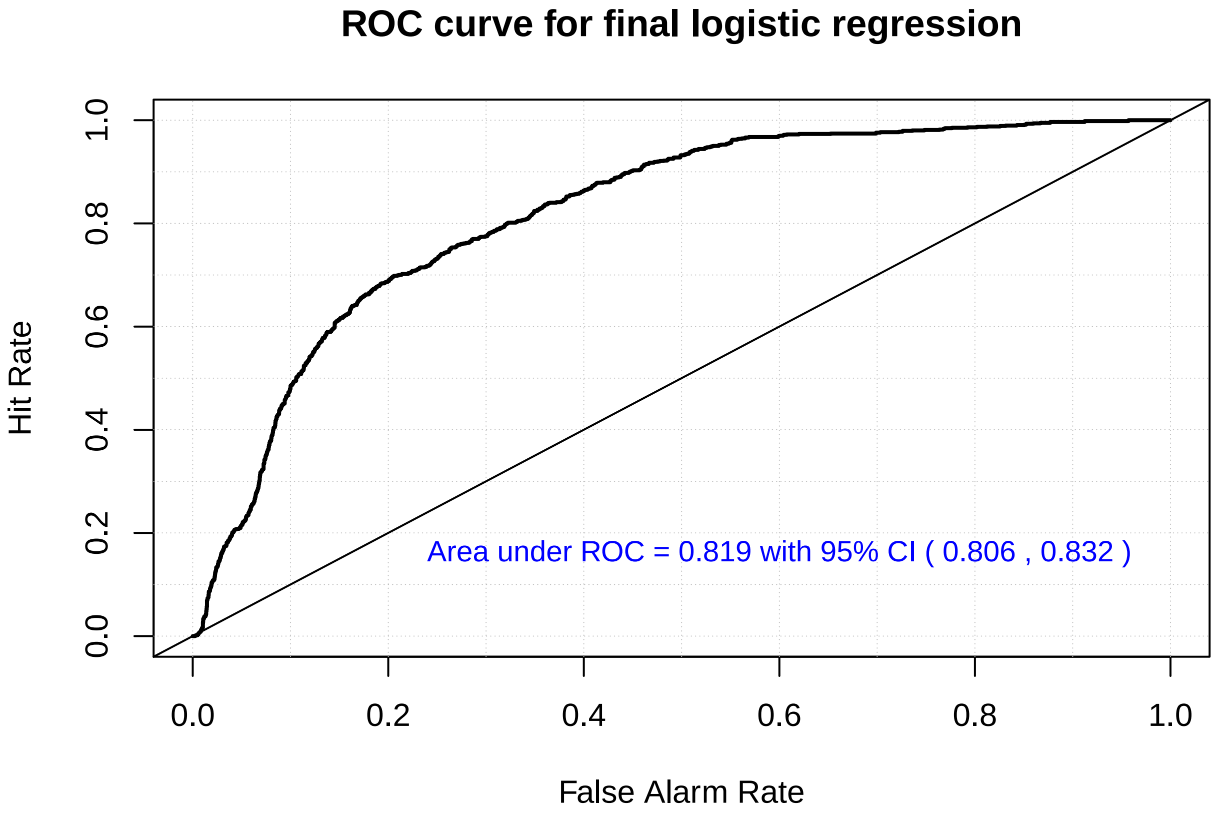
Table 4: Odds ratio based on parameter estimates from the logistic model

	OR	2.5 %	97.5 %
(Intercept)	0.6774	0.4680	0.9781
satisfaction_level	0.0152	0.0120	0.0192
last_evaluation	2.2902	1.6031	3.2755
number_project	0.7353	0.6983	0.7738
average_montly_hours	1.0044	1.0032	1.0057
time_spend_company	1.2919	1.2453	1.3404
Work_accident	0.2260	0.1814	0.2791
promotion_last_5years	0.2079	0.1078	0.3698
departmenthr	1.2491	0.9094	1.7160
departmentIT	0.8892	0.6658	1.1892
departmentmanagement	0.6302	0.4283	0.9202
departmentmarketing	0.9367	0.6837	1.2831
departmentproduct_mng	0.8382	0.6136	1.1446
departmentRandD	0.5937	0.4189	0.8375
departmentsales	0.9825	0.7705	1.2577
departmentsupport	1.1274	0.8701	1.4656
departmenttechnical	1.0786	0.8362	1.3959
salary.L	0.2541	0.2020	0.3153
salary.Q	0.6928	0.5978	0.7981

4.1 Applying the fnal model to the test data and presenting the ROC curve

```
library(verification)
library(cvAUC)

phat.logit <- predict(logistic, newdata = test_set, type="response") # predicted probabilities

# a custom function for computing and plotting ROC curve
roc_curve <- function (phat, main = "", col="blue", roc_val=F) {
   yobs <- as.integer(as.character(test_set$left))
   AUC <- ci.cvAUC(predictions=phat, labels=yobs, folds=1:NROW(test_set),
                confidence=0.95)

   if(roc_val) return(AUC$cvAUC)

   auc.ci <- round(AUC$ci, digits=3) # confidence interval for cross-validated
                                     # Area Under the ROC Curve
   mod <- verify(obs=yobs, pred=phat)
   roc.plot(mod, plot.thres = NULL, main=main)
  text(x=0.6, y=0.16, paste("Area under ROC =", round(AUC$cvAUC, digits=3),
    "with 95% CI (", auc.ci[1], ",", auc.ci[2], ")",
    sep=" "), col=col, cex=.9)
}

roc_curve(phat.logit, main="ROC curve for final logistic regression")
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```



The Area under the ROC curve is obtained as **0.819**, which is indicative of a relatively good predictive performance. With 95% confidence level, the ROC is estimated to lie between **0.806** and **0.833**.

5 Random Forest (RF): Another baseline model

```
 library(randomForest)
fit.rf <- randomForest(left ~., data=train_set,importance=T, ntree=500)
print(fit.rf)
```

```
##
## Call:
##  randomForest(formula = left ~ ., data = train_set, importance = T,      ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 1.08%
## Confusion matrix:
##      0    1 class.error
## 0 7575   17    0.002239
## 1   91 2316    0.037806
```

5.1 Partial dependence plots

```
# partial dependence plot
par(mfrow=c(2,3))
partialPlot(fit.rf, pred.data = train_set, x.var = satisfaction_level, rug = T,
            main = "", xlab = "Satisfaction level")
partialPlot(fit.rf, pred.data = train_set, x.var = number_project, rug = T,
            main = "", xlab = "Number of projects")
partialPlot(fit.rf, pred.data = train_set, x.var = last_evaluation, rug = T,
            main = "", xlab = "Last evaluation")
partialPlot(fit.rf, pred.data = train_set, x.var = time_spend_company, rug = T,
            main = "", xlab = "Time spent")
partialPlot(fit.rf, pred.data = train_set, x.var = average_montly_hours, rug = T,
            main = "")
```

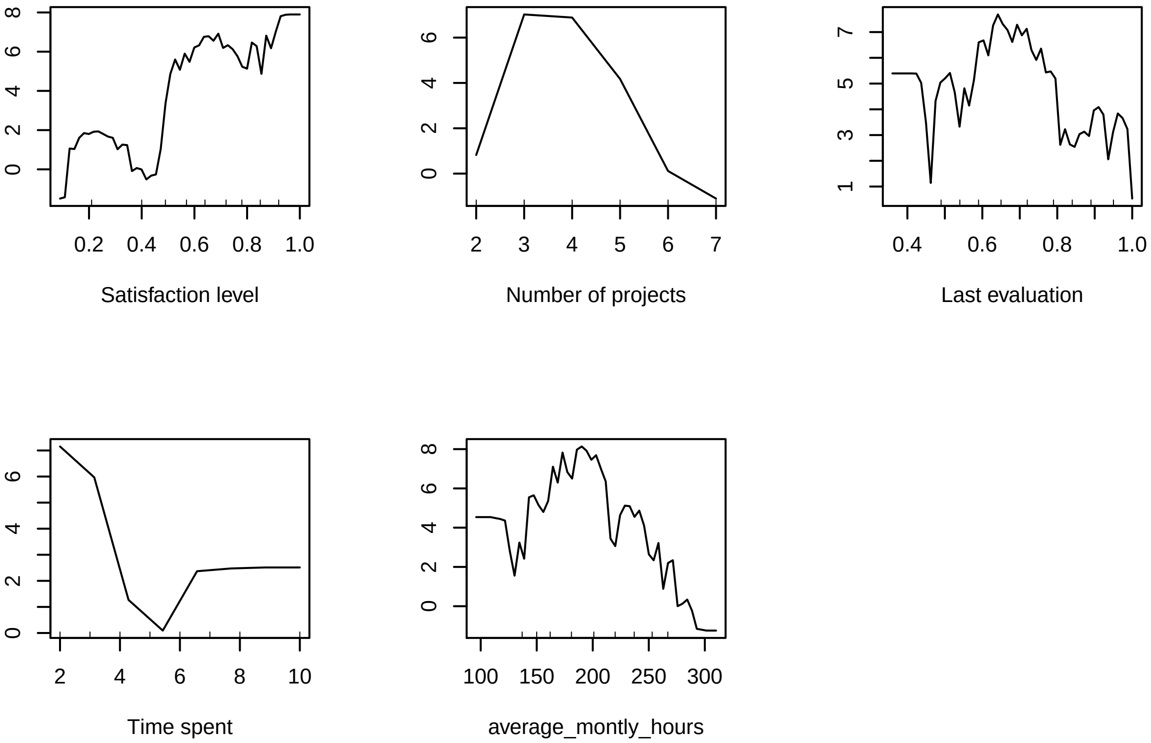
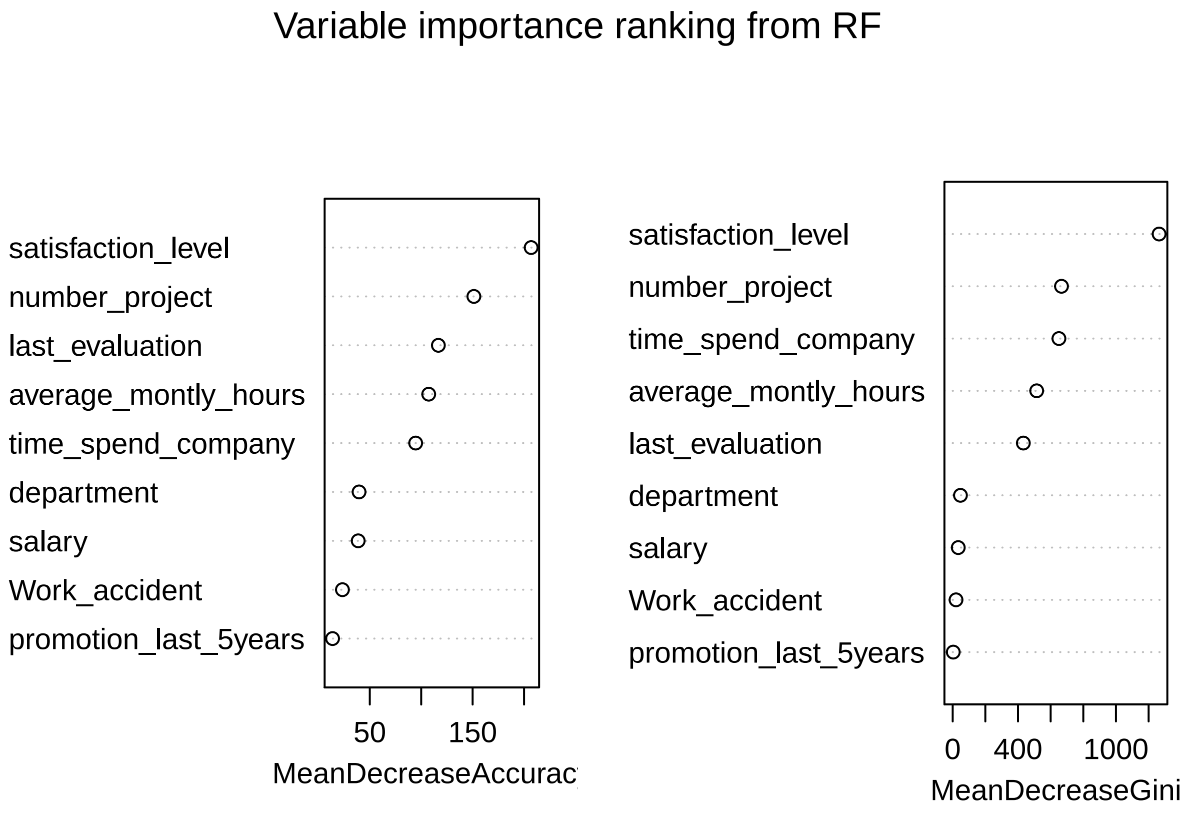


Figure 6: Partial dependence plots for continuous predictors

We observe strong non-linear patterns in all the plots in Figure 6, which signifies that a linear model such as the logistic regression is not appropriate to model the relationship between employee turnover (left) and the variables indicated in the plots.

5.2 Variable importance rankings

```
varImpPlot(fit.rf, main = "Variable importance ranking from RF", cex=0.89)
```

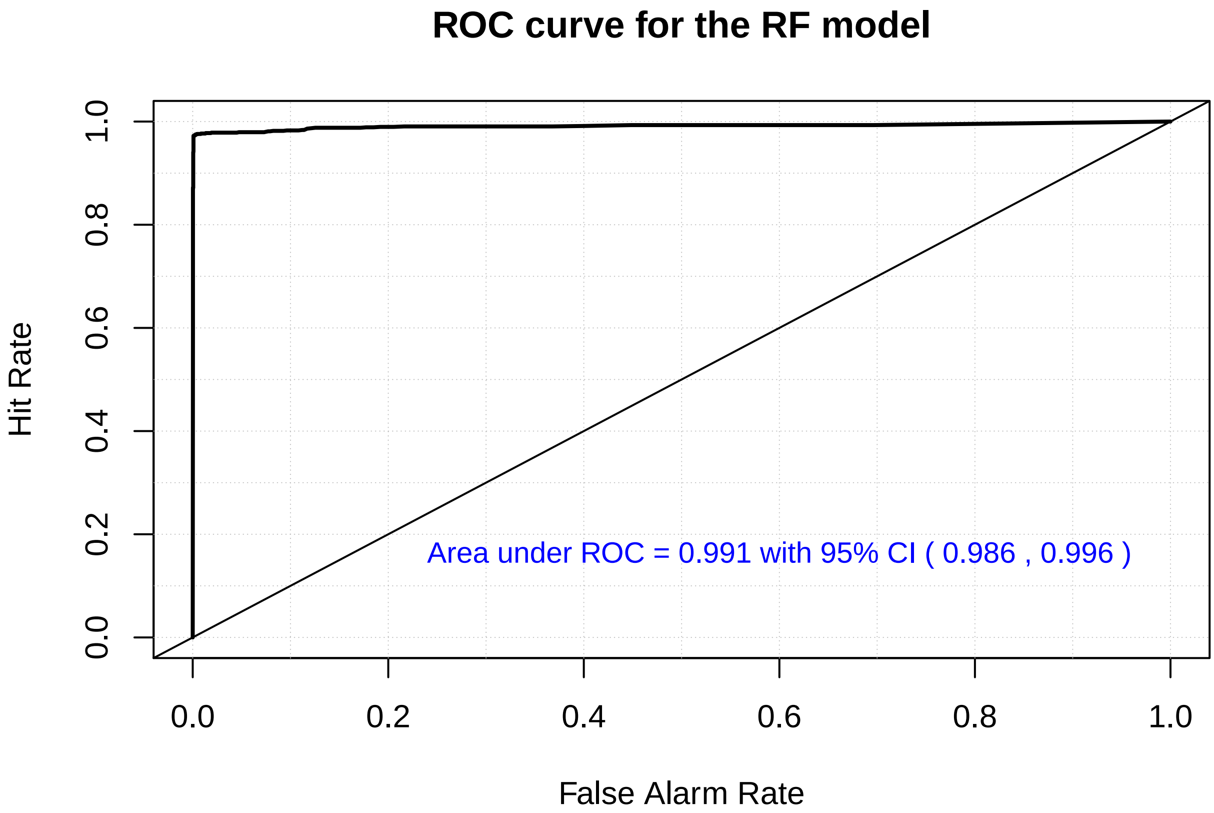


According to the mean decrease accuracy, the top four variables are `satisfaction level`, `number of projects`, `last evaluation`, and `average monthly hours`. Satisfaction level and number of projects happen to be the two most important determinants of employee turnover or retention. Based on the mean decrease GINI, satisfaction level remains the most influential variable.

5.3 ROC curve based on the predictions on the test set

```r
# get predicted probabilities
phat.rf <- predict(fit.rf, newdata=test_set, type="prob")[, 2]
# create ROC curve
roc_curve(phat.rf, main="ROC curve for the RF model")
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```



The AUC value of **0.992** shows that the random forest model has very high predictive performance for the problem at hand.

6 Generalized Additive Model (GAM)

A stepwise selection procedure with AIC was employed to select the best fitting GAM model. The `scope` argument for the `step.Gam()` function allowed us to choose the smoothing parameters adaptively in the backfitting algorithm by specifying whether a term could either appear not at all, linearly, or as a smooth function estimated non-parametrically via smoothing splines or loess smoother.

```
library(gam)
# Create a GAM object for use in the stepwise selection. Note that the smoothing
# terms will be specified later in the selection stage following.
fit.gam <- gam( left ~ satisfaction_level + number_project + time_spend_company +
department + last_evaluation + average_montly_hours + Work_accident +
    promotion_last_5years + salary , family = binomial,
    data=train_set, trace=T, control = gam.control(epsilon=1e-04, bf.epsilon = 1e-04,
                                               maxit=50, bf.maxit = 50))

#--- perform a stepwise selection
# register parallel backend for paralel execution
require(doMC)
registerDoMC(cores = (detectCores()-8))
fit.step.gam <- step.Gam(fit.gam, scope=list(
    "satisfaction_level"=~1 + satisfaction_level + lo(satisfaction_level) +
        s(satisfaction_level),
    "last_evaluation"=~1+ last_evaluation + lo(last_evaluation)+ s(last_evaluation),
    "number_project"=~1 + number_project + lo(number_project) + s(number_project),
    "average_montly_hours"=~1 + average_montly_hours + lo(average_montly_hours) +
        s(average_montly_hours),
    "time_spend_company"=~1 + time_spend_company + lo(time_spend_company) +
        s(time_spend_company)),
            scale =2, steps=1000, parallel=T, direction="both", trace = F)

summary(fit.step.gam)
```

```
##
## Call: gam(formula = left ~ salary + s(satisfaction_level) + s(last_evaluation) +
##     lo(number_project) + s(average_montly_hours) + s(time_spend_company),
##     family = binomial, data = train_set, control = gam.control(epsilon = 1e-04,
##         bf.epsilon = 1e-04, maxit = 50, bf.maxit = 50), trace = FALSE)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08887 -0.31039 -0.12434 -0.00151  3.64731
##
## (Dispersion Parameter for binomial family taken to be 1)
##
##     Null Deviance: 11037 on 9998 degrees of freedom
## Residual Deviance: 4061 on 9975 degrees of freedom
## AIC: 4109
##
## Number of Local Scoring Iterations: 1
##
## Anova for Parametric Effects
##                           Df Sum Sq Mean Sq F value  Pr(>F)
## salary                     2     75      38    37.2 < 2e-16 ***
## s(satisfaction_level)      1     26      26    26.1 3.2e-07 ***
## s(last_evaluation)         1     57      57    56.0 7.9e-14 ***
## lo(number_project)         1    104     104   102.6 < 2e-16 ***
## s(average_montly_hours)    1     75      75    74.3 < 2e-16 ***
## s(time_spend_company)      1    370     370   366.6 < 2e-16 ***
## Residuals               9975  10077       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                         Npar Df Npar Chisq P(Chi)
## (Intercept)
## salary
## s(satisfaction_level)         3        496 <2e-16 ***
## s(last_evaluation)            3        370 <2e-16 ***
## lo(number_project)            4        737 <2e-16 ***
## s(average_montly_hours)       3        346 <2e-16 ***
## s(time_spend_company)         3        305 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

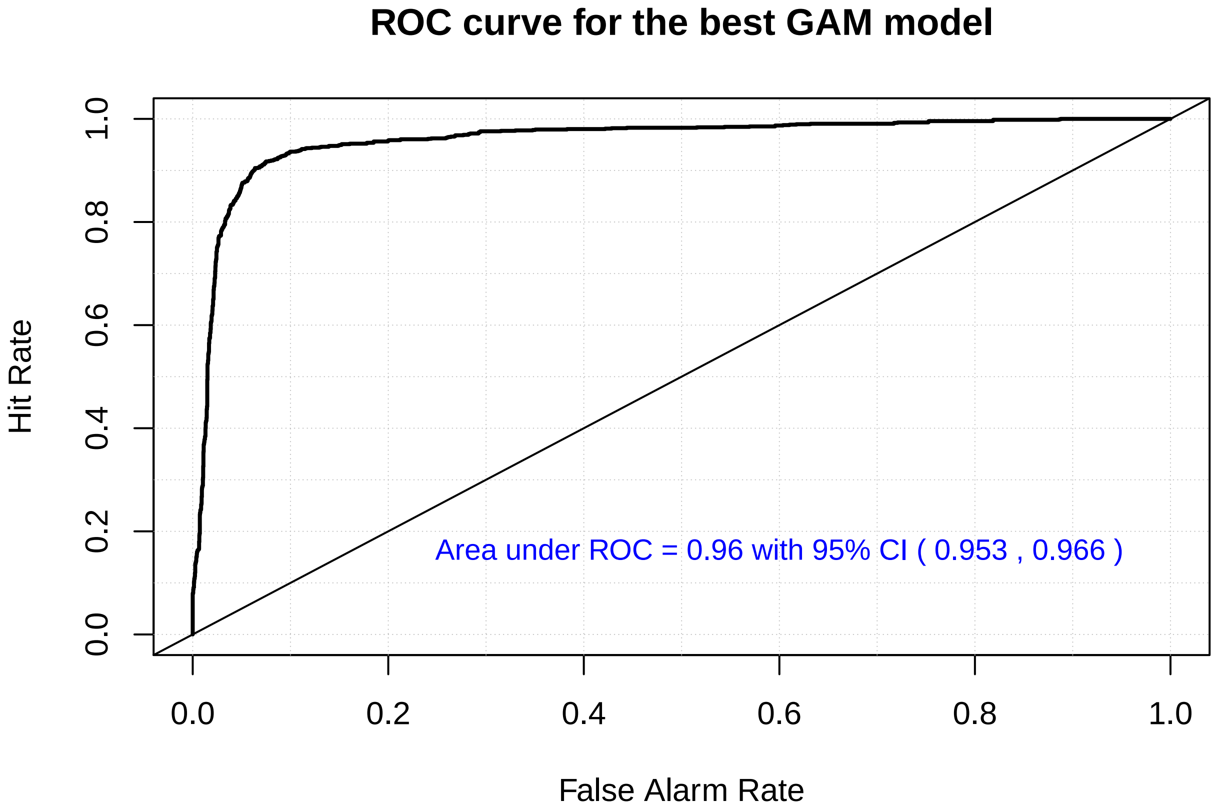
The results from the anova for nonparametric effects' table indicate that the loess smoother is appropriate for the nonparametric smoothing terms in `number of projects`. On the other hand, smoothing splines were chosen for `satisfaction level`, `last evaluation`, `average monthly hours`, and `time spent at the company` as the best smoothing functions with 3 degrees of freedom for each term.

6.1 ROC curve based on predictions on the test set

```
# get predicted probabilities
phat.gam <- predict(fit.step.gam, newdata=test_set, type="response", se.fit=F)
```

```
roc_curve(phat.gam, main="ROC curve for the best GAM model")
```

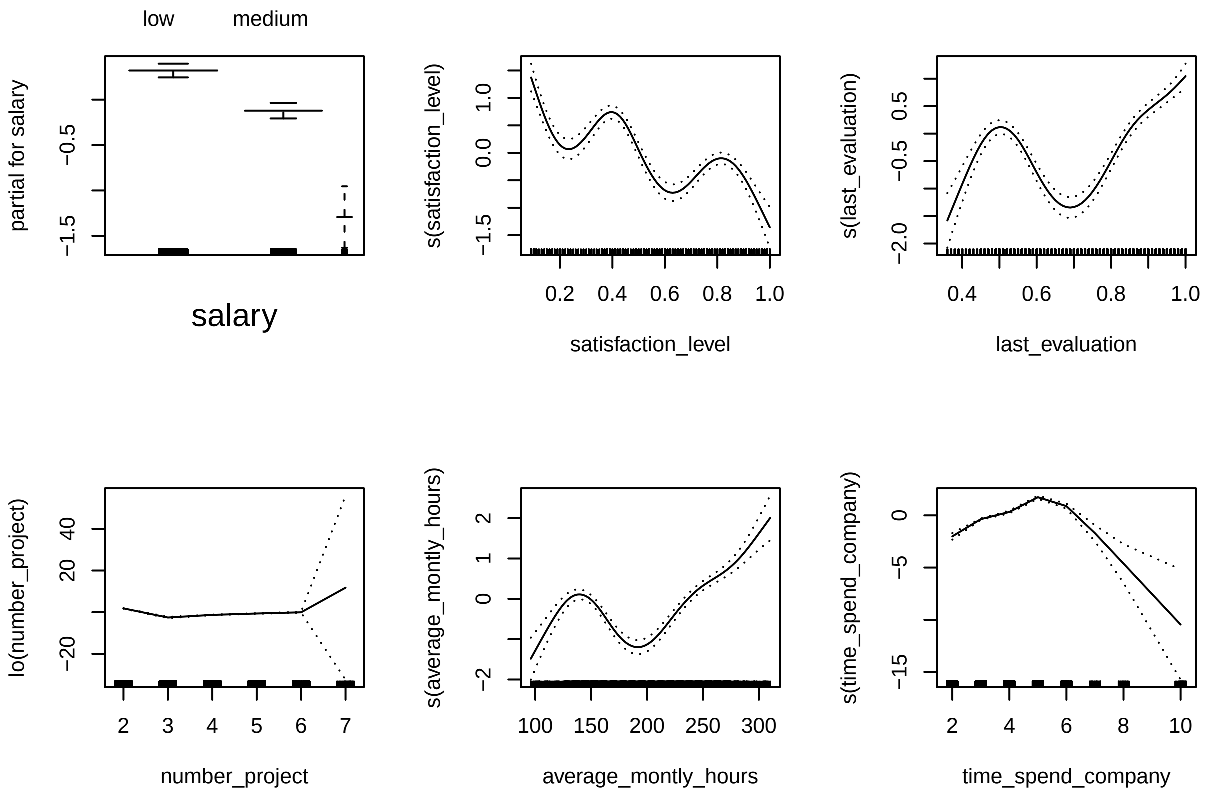
```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```



From the plot, with AUC value of **0.957**, we notice that the final GAM model obtained performed well predicting the probability of an employee leaving or remaining in the company on the test data.

6.2 Plots of the functional forms for continuous predictors

```
par(mfrow=c(2,3))
plot(fit.step.gam, se=T)
```



The functional forms for the continuous predictors show non-linear relationships, indicating the inadequacy of linear models for our classification problem.

7 Multivariate Adaptive Regression Splines (MARS)

We allowed maximum degree of interactions up to 3 by setting `degree` to 3.

```
library(earth)

fit.mars <- earth(left ~ ., data=train_set, degree=3,
                  glm=list(family=binomial(link = "logit")))
fit.mars
```

```
## GLM (family binomial, link logit):
##  nulldev   df       dev   df   devratio     AIC iters converged
##    11040 9998      2359 9971       0.79    2420    19         1
##
## Earth selected 28 of 34 terms, and 5 of 18 predictors
## Termination condition: Reached nk 37
## Importance: satisfaction_level, number_project, time_spend_company, ...
## Number of terms at each degree of interaction: 1 4 14 9
## Earth GCV 0.03687    RSS 363.6    GRSq 0.7983    RSq 0.801
```

```
summary(fit.mars) %>% .$coefficients %>% head(10)
```

```
##                                                                 1
## (Intercept)                                             -0.009439
## h(number_project-3)                                      0.045349
## h(3-number_project)                                      1.088727
## h(number_project-3)*h(time_spend_company-5)             -0.023808
## h(satisfaction_level-0.39)*h(3-number_project)          -2.359664
## h(0.39-satisfaction_level)*h(3-number_project)          -2.221185
## h(0.23-satisfaction_level)*h(number_project-3)                      0.290139
## h(satisfaction_level-0.23)*h(number_project-3)*h(time_spend_company-5) -0.154943
## h(satisfaction_level-0.23)*h(number_project-3)*h(5-time_spend_company) -0.016829
## h(satisfaction_level-0.23)*h(number_project-3)*h(time_spend_company-4)  0.142782
```

7.1 Variable importance ranking

```
library(vip)
# generating variable importance plot
vip(fit.mars, num_features = 10, aesthetics = list(fill="dodgerblue",
                                                   color="dodgerblue")) +
    ggtitle("Variable importance (GCV) ranking from MARS") +
    scale_fill_brewer(palette = "Set1") +
    theme(plot.title = element_text(hjust = .5))
```

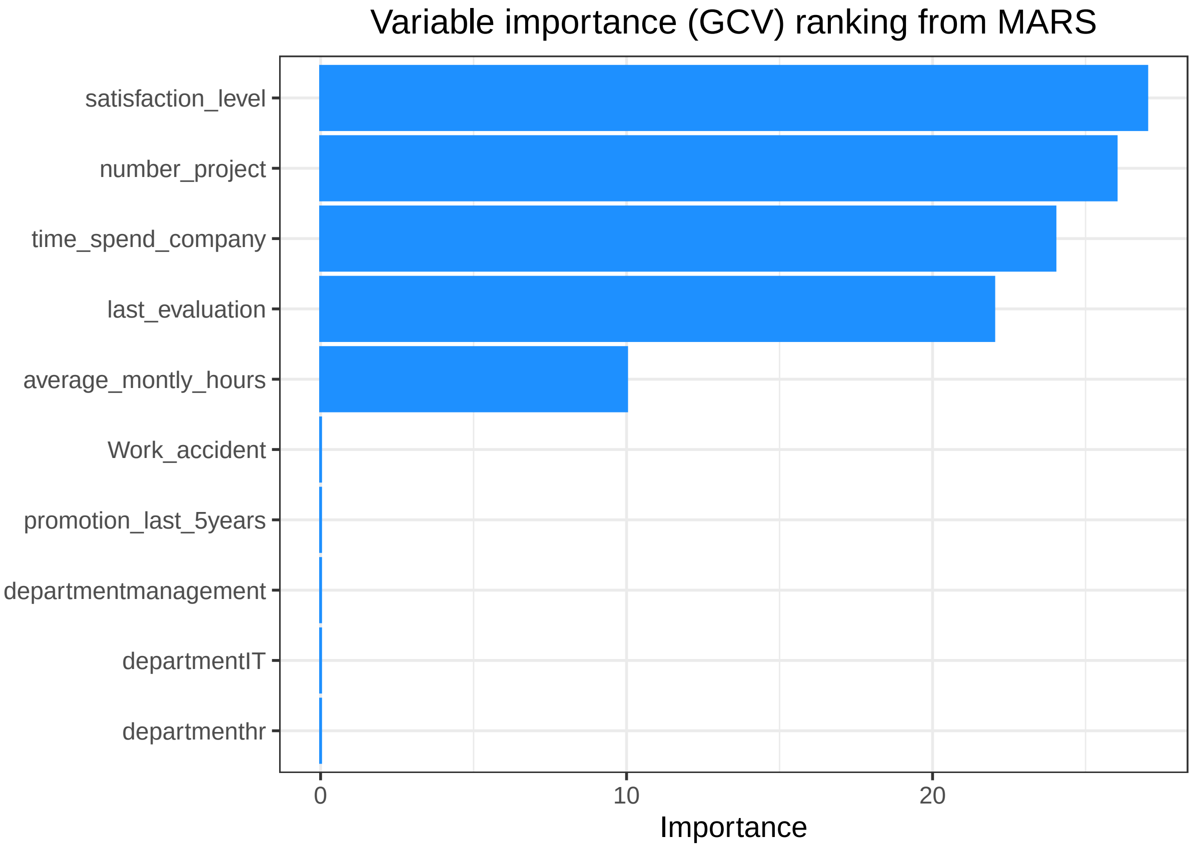


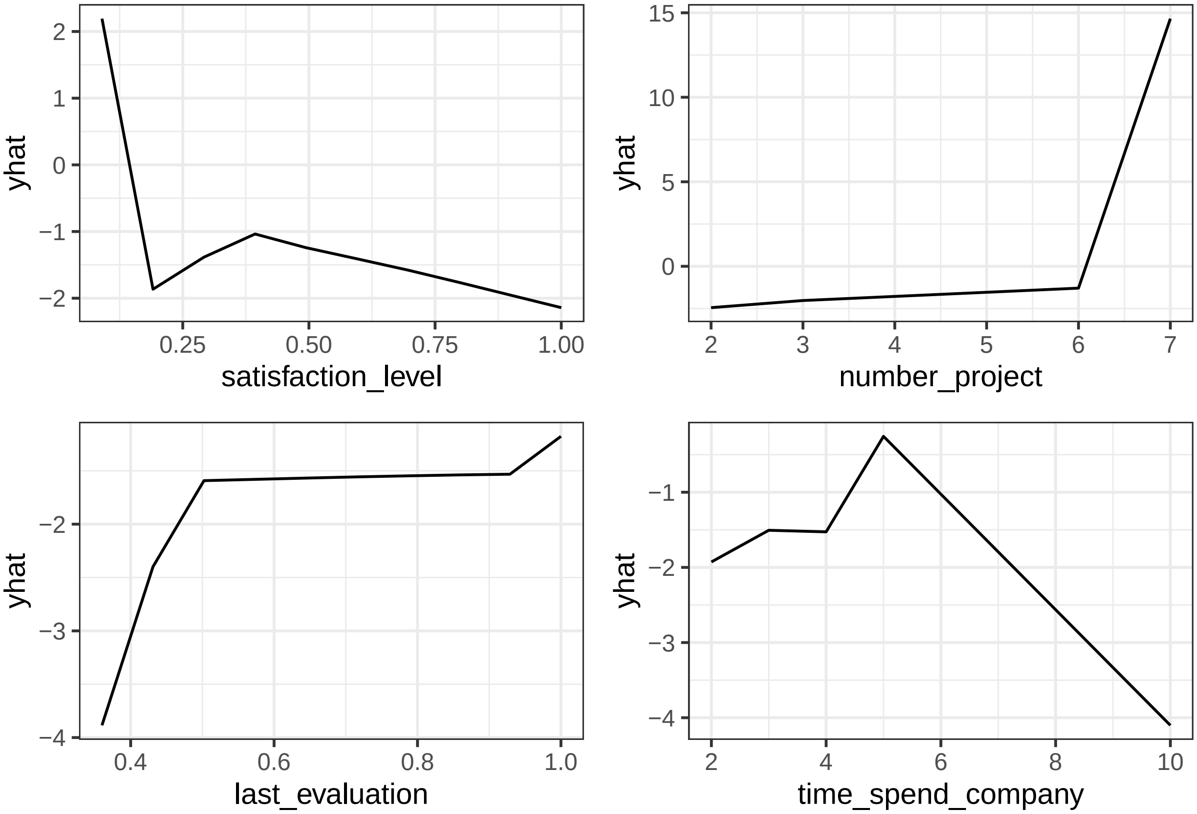
Figure 7: Variable importance based on impact to GCV as predictors are added to the model

We see here that the top four important variables include `satisfaction level` and `number of projects`, `time spent at the company`, and `last evaluation`, signifying that these variables played major role in predicting employee turnover or retention on the test data. Once again, the satisfaction level and number of projects are the top two most important variables.

7.2 Partial dependence plots

```
library(pdp)
# partial dependence plot
(partial(fit.mars, pred.var = "satisfaction_level", grid.resolution = 10)%>%autoplot() +
partial(fit.mars, pred.var = "number_project", grid.resolution = 10)%>%autoplot()) /
```

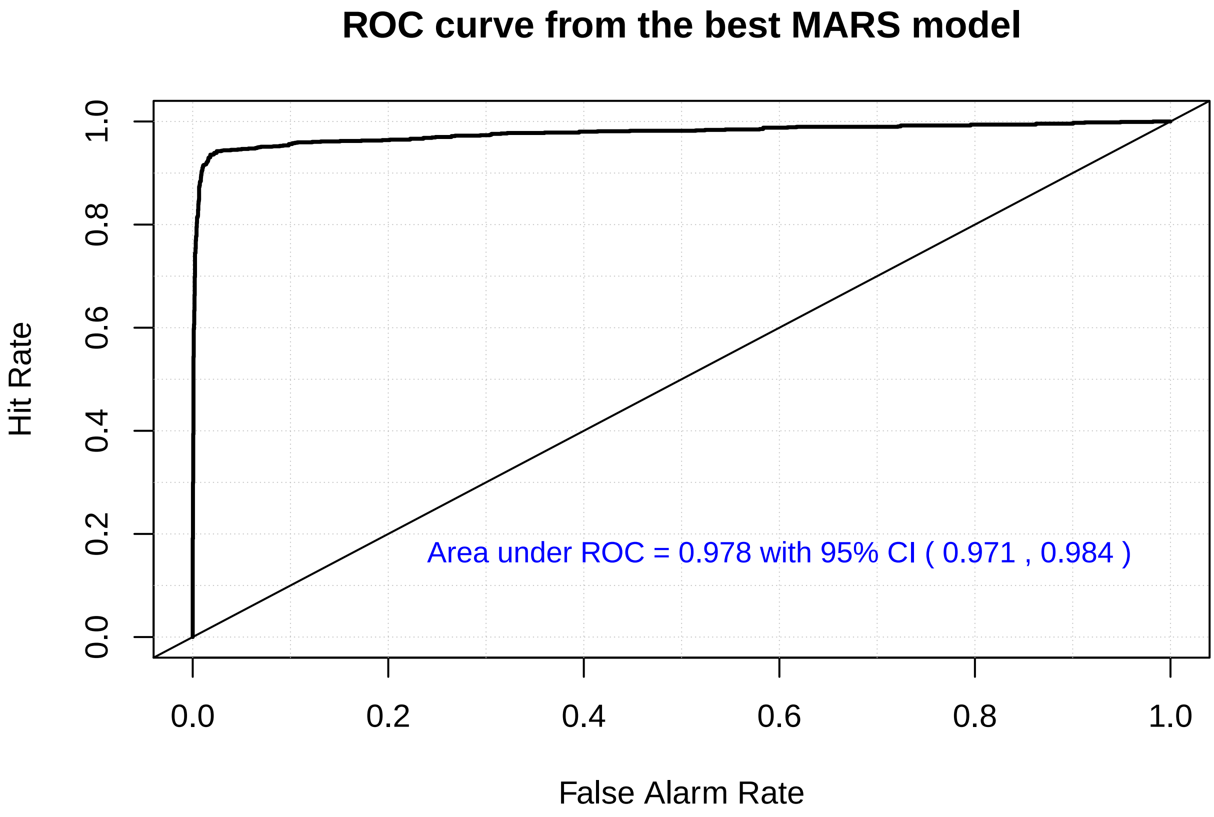
```
(partial(fit.mars, pred.var = "last_evaluation", grid.resolution = 10)%>%autoplot() +
partial(fit.mars, pred.var = "time_spend_company", grid.resolution = 10)%>%autoplot())
```



These plots help us to assess the marginal effect of each variable on predicting employee turnover. For example, higher satisfaction levels and time spent have negative effect on the the likelihood of employee turnover.

```
#  get predicted probabilities
phat.mars <- predict(fit.mars, newdata=test_set, type="response")
# create ROC curve
roc_curve(phat.mars, main="ROC curve from the best MARS model")
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```



With AUC of **0.978**, it is clear that the MARS model performed well predicting the likelihood of employee turnover on the the test data.

8 Projection Pursuit Regression

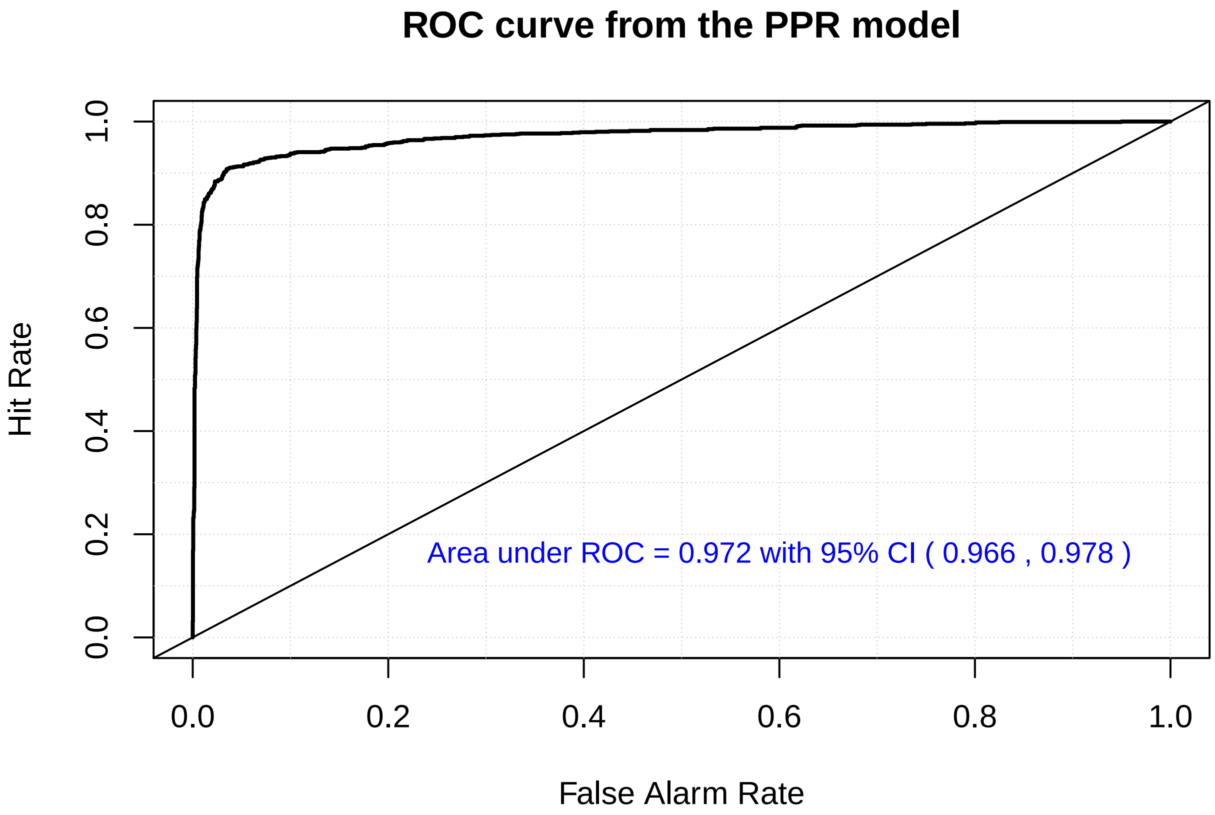
```
train_set2 <- train_set %>%
    mutate(left = as.integer(as.character(left)))

fit.ppr <- ppr(left ~ ., sm.method = "supsmu",
    data = train_set2, nterms = 2, max.terms = 12, bass=3)

# summary(fit.ppr)

fit2.ppr <- update(fit.ppr, bass=5, nterms=4)

summary(fit2.ppr)
```

```
## Call:
## ppr(formula = left ~ ., data = train_set2, sm.method = "supsmu",
##     nterms = 4, max.terms = 12, bass = 5)
##
## Goodness of fit:
##  4 terms  5 terms  6 terms  7 terms  8 terms  9 terms 10 terms 11 terms
##    451.4    446.6    453.2      0.0      0.0      0.0      0.0      0.0
## 12 terms
##      0.0
##
## Projection direction vectors ('alpha'):
##                       term 1     term 2     term 3     term 4
## satisfaction_level    -0.2331359  0.1110459  0.0457991  0.5311360
## last_evaluation         0.1443914  0.2082609  0.0303084  0.2695419
## number_project         -0.0330944  0.0874888  0.0309969  0.0552125
## average_montly_hours    0.0008021  0.0008039  0.0001825  0.0006572
## time_spend_company      0.1021438  0.0494171  0.0205818 -0.2808617
## Work_accident          -0.0951757 -0.0024925  0.0020970  0.0117135
## promotion_last_5years -0.0370700 -0.0278142  0.0009451 -0.0263496
## departmentaccounting  -0.3101242 -0.2999029 -0.3125504  0.2452198
## departmenthr           -0.2635642 -0.2925437 -0.3082662  0.1952116
## departmentIT           -0.3042922 -0.3080432 -0.3182266  0.2351986
## departmentmanagement  -0.2778582 -0.3119968 -0.3168444  0.2467050
## departmentmarketing   -0.2865438 -0.3105636 -0.3177001  0.2369756
## departmentproduct_mng -0.3513256 -0.3066011 -0.3160292  0.2441861
## departmentRandD        -0.3401794 -0.3091553 -0.3136593  0.2456992
## departmentsales        -0.2944225 -0.3041843 -0.3129795  0.2294481
## departmentsupport      -0.2751116 -0.3104359 -0.3240571  0.2503886
## departmenttechnical    -0.2803193 -0.3011082 -0.3147186  0.2346803
## salary.L               -0.0718595 -0.0074482 -0.0011939  0.0324935
## salary.Q               -0.0047179 -0.0034719 -0.0033493  0.0207855
##
## Coefficients of ridge terms ('beta'):
## term 1 term 2 term 3 term 4
## 0.1190 0.3479 0.1428 0.1677
```

8.1 ROC curve based on predictions on the test set

```
#  get predicted probabilities
phat.ppr <- predict(fit2.ppr, newdata=test_set)
phat.ppr <- scale(phat.ppr,center = min(phat.ppr),
                  scale = max(phat.ppr)-min(phat.ppr))
# create ROC curve
roc_curve(phat.ppr, main="ROC curve from the PPR model")
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```



The AUC obtained based on the PPR model is **0.972**, showing a better predictive performance.

9 Summary of results

```
roc_values <- c(roc_curve(phat.logit, roc_val = T),
                roc_curve(phat.rf, roc_val = T),
                roc_curve(phat.gam, roc_val = T),
                roc_curve(phat.mars, roc_val = T),
                roc_curve(phat.ppr, roc_val = T)
                )

data.frame("Model"= c("Logistic (LASSO)","Random Forest","GAM","MARS","PPR"), "AUC"= roc_values) %>%
 kable(booktabs=T, align = "lc")%>%
    kable_styling(latex_options =c("HOLD_position"))
```

Model	AUC
Logistic (LASSO)	0.8191
Random Forest	0.9914
GAM	0.9598
MARS	0.9776
PPR	0.9722

Among all the five supervised learning approaches considered, the random forest (RF) model yielded the best predictive performance since it provides the largest AUC value for determining the likelihood of employee turnover or retention in the company. However, among GAM, MARS, and PPR, MARS is the best performing model followed closely by PPR. The regularized logistic regression model performed poorly relative to the other methods in terms of predictive performance, however, it must be noted that its results are highly interpretable compared to the others. For instance, using the odds ratios obtained from the logistic coefficients, we can explain how exactly a particular predictor influences employee retention.

Overall, satisfaction level and number of projects emerged as the top two variables that predict an employee's turnover or retention. Therefore, it is recommended that employers take these factors seriously into account in their bid to reducing turnover rate or improving retention.