Analysis of Factors Affecting Employee Retention

Pauline Korukundo and Martin Vivas

2024-12-24

The dplyr and ggplot2 packages were used for the tidying, data wrangling and visualization during the Exploratory Data Analysis.

# 1. Introduction

The MFG10 Year Termination dataset contains 18 variables and 49,653 observations on employee data including their demographics, how long they have worked, those that are still employed and for terminated employees, the reason for termination is indicated.

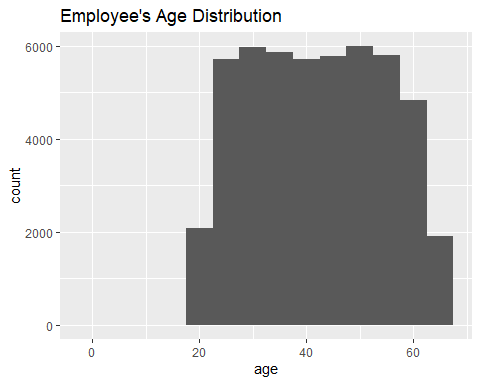
# 2. Tidying and Data Transformations

The data set does not contain any missing values. Unnecessary values such as gender\_full, birthdate\_key, terminationdate\_key, and recorddate\_key were removed. Other variables will be introduced to derive better insights on the data.

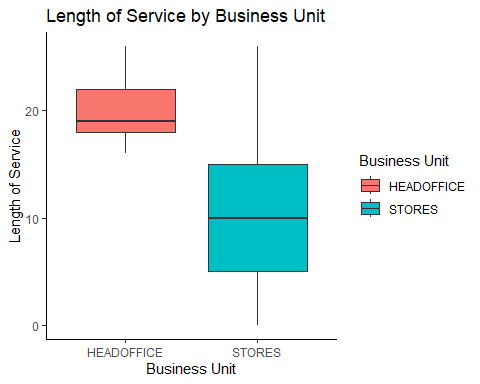
Summary statistics for age, length of service, departments, and job titles show the youngest employees are 19, the oldest are 65, and the median age is 42. The minimum length of service is 0 years, and the maximum is 26 years.

# 3. Descriptive Analytics

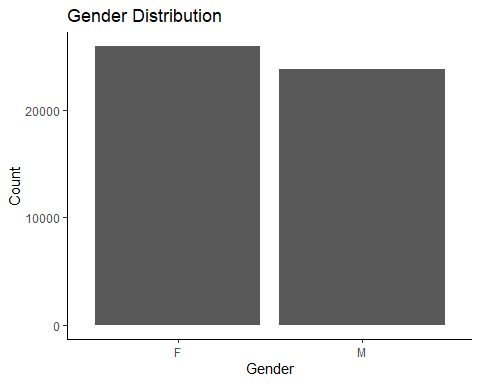
This section explores the employee data to get insights on the distribution and relationships between different variables.



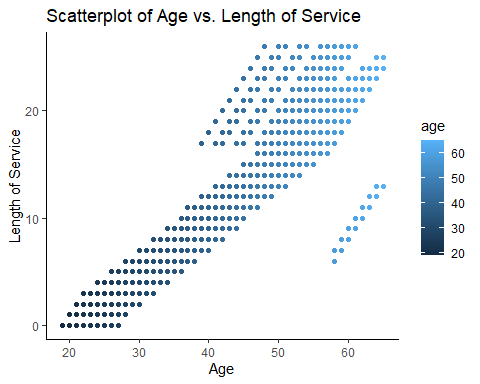
The employees are generally between the ages of 17 and 68 with most employees in their mid twenties and late fifties.



Employees working at the Head Office business unit have a longer term of service with the company, averaging 19 years, while those at the Stores business unit work for an average of 10 years and at most 15 years.



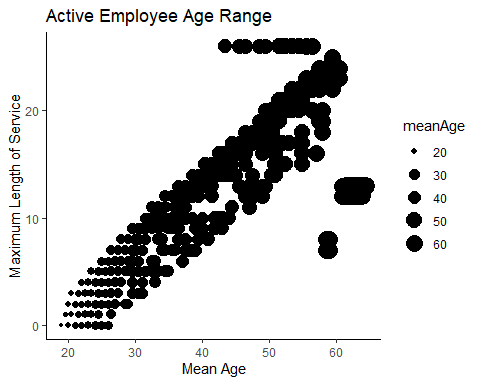
The company has slightly more females than males.



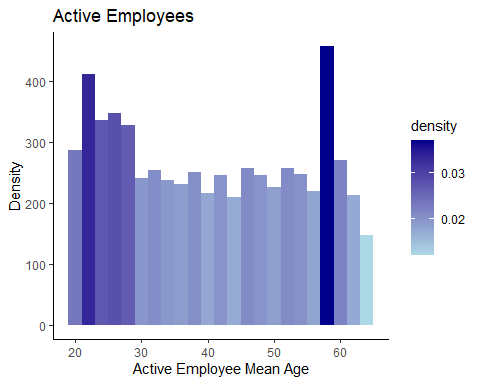
Employees in their early fifties and sixties are generally employed for a longer time than those in their early twenties and forties.

## Active vs Terminated Employee Categories

### Active Employees



The active employee age range has a bimodal distribution. Most active employees in their twenties have been working for less than 10 years, while those above 40 have been working for 15 to 26 years.

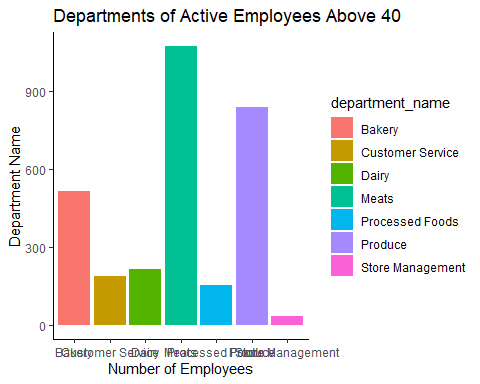


A table of Active Employees by Department

| department\_name | n |
| --- | --- |
| Bakery | 892 |
| Customer Service | 1178 |
| Dairy | 1012 |
| Meats | 1195 |
| Processed Foods | 733 |
| Produce | 1013 |

#### Active Employees above 40

## # A tibble: 3,102 × 7  
## EmployeeID job\_title gender\_short department\_name city\_name meanAge  
## <int> <chr> <chr> <chr> <chr> <dbl>  
## 1 3008 Dairy Person F Dairy Vancouver 64.5  
## 2 2645 Meat Cutter F Meats Victoria 64   
## 3 2650 Dairy Person F Dairy Vancouver 64   
## 4 2651 Meat Cutter F Meats Prince George 64   
## 5 2652 Dairy Person F Dairy Vancouver 64   
## 6 2667 Meat Cutter F Meats Burnaby 64   
## 7 2668 Meat Cutter F Meats Cranbrook 64   
## 8 2670 Dairy Person F Dairy Langley 64   
## 9 2674 Meat Cutter F Meats Kelowna 64   
## 10 2681 Dairy Person F Dairy New Westminister 64   
## # ℹ 3,092 more rows  
## # ℹ 1 more variable: maxServiceLength <int>

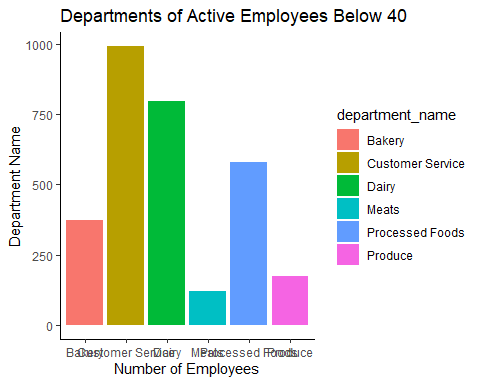


The oldest employees are female and working in the Dairy and Meats departments. They have worked for 13 years. Generally, this group of employees works in Produce, Meats, and Bakery, with very few in store Management.

#### Active Employees Below 40

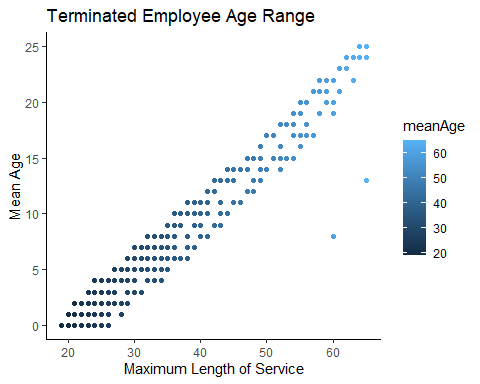
Active Employee below 40 by meanAge

| EmployeeID | job\_title | gender\_short | department\_name | city\_name | meanAge | maxServiceLength |
| --- | --- | --- | --- | --- | --- | --- |
| 8214 | Produce Clerk | F | Produce | Haney | 19.0 | 0 |
| 8296 | Cashier | F | Customer Service | Kelowna | 19.0 | 0 |
| 8321 | Cashier | F | Customer Service | Grand Forks | 19.0 | 0 |
| 8105 | Cashier | F | Customer Service | New Westminister | 19.5 | 1 |
| 8231 | Cashier | M | Customer Service | Princeton | 19.5 | 1 |
| 8258 | Dairy Person | M | Dairy | Valemount | 19.5 | 1 |
| 8279 | Cashier | F | Customer Service | White Rock | 19.5 | 1 |
| 7844 | Shelf Stocker | M | Processed Foods | Vancouver | 20.0 | 0 |



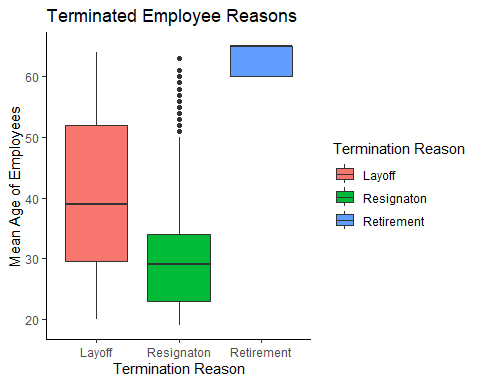
The youngest workforce has a mixture of males and females in the customer service and produce departments. The customer service department has the highest number of employees below 40 years.

### Terminated Employees



The terminated employees with a mean age of 50-60 have worked for more than 20 years. On the other hand, employees in their 20s and 30s that were terminated worked for less than 10 years.

#### Reasons for Termination



Terminated employees in their 40s were laid off, while those that resigned were mostly in their 30s with some outliers in their 60s. Others went into retirement at a mean age of 65.

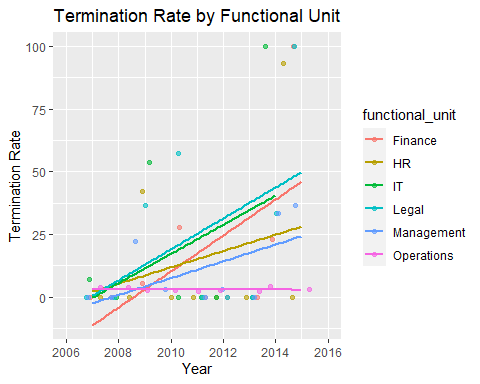
## Number of Unique Employees: 6284

Over the 10 years, the company has had a total of 6,284 different employees. To further understand these employees, we perform analytics in the next sections that will help HR professionals to know the employees at risk and those likely to stay with the company.

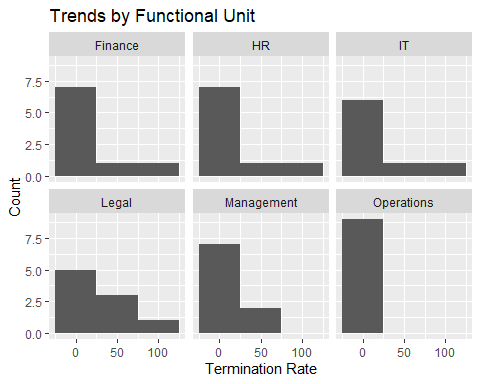
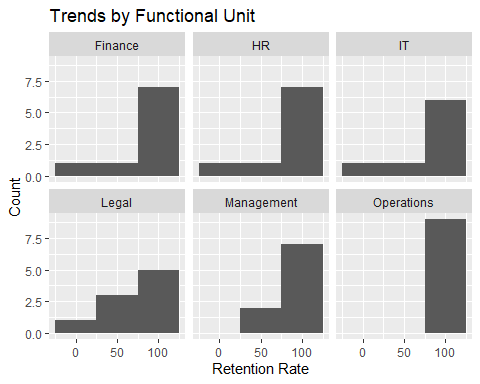
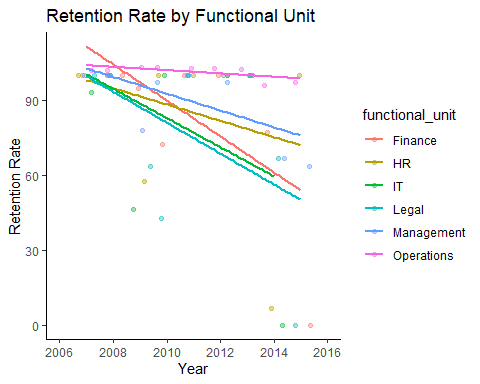
### Retention and Termination Trends by Functional Units

#### Visualizing Trends by Functional Unit

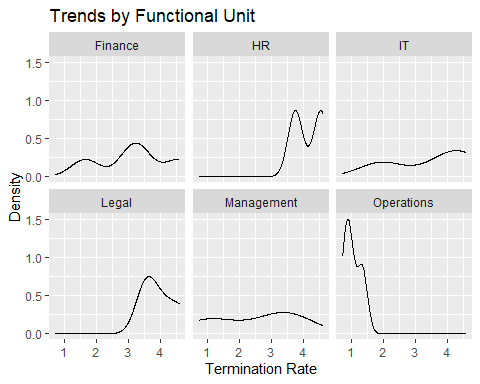
## `geom\_smooth()` using formula = 'y ~ x'



## `geom\_smooth()` using formula = 'y ~ x'

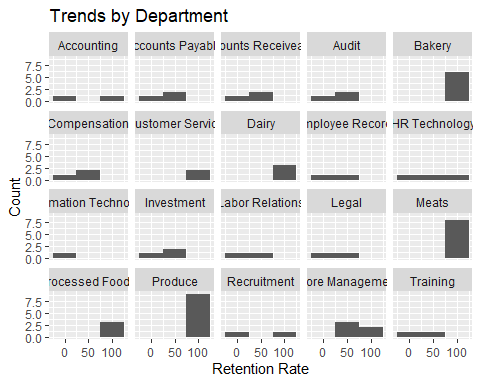


## Warning: Removed 26 rows containing non-finite values (`stat\_density()`).



#### Visualizing Trends by Department

## # A tibble: 67 × 9  
## # Groups: department\_name [20]  
## STATUS\_YEAR department\_name mean\_age mean\_length\_of\_service ACTIVE TERMINATED  
## <int> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 2007 HR Technology 60.1 19.4 8 1  
## 2 2007 Meats 41.8 10.3 1065 66  
## 3 2007 Produce 41.8 10.3 919 59  
## 4 2008 Bakery 41.8 10.3 834 16  
## 5 2008 Meats 41.8 10.3 1022 71  
## 6 2008 Produce 41.8 10.3 870 50  
## 7 2009 Audit 59.3 19.2 3 1  
## 8 2009 Bakery 41.8 10.3 829 5  
## 9 2009 Compensation 60.1 20.0 3 1  
## 10 2009 Employee Recor… 59.7 20.0 4 2  
## # ℹ 57 more rows  
## # ℹ 3 more variables: previous\_active <dbl>, termination\_rate <dbl>,  
## # retention\_rate <dbl>

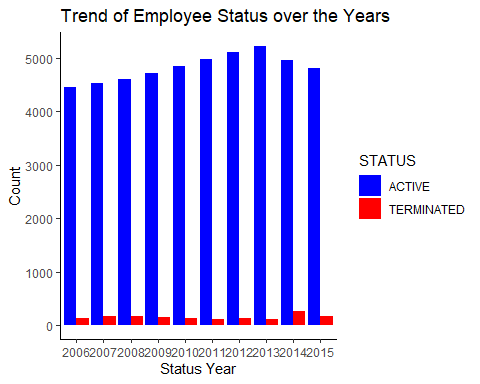


#### Summary of Findings:

* Operations Functional Unit has a stable retention rate over the years, while the Finance department’s retention rate is gradually dropping.
* Produce, Meats, and Dairy Departments have the highest retention rates.
* Termination Rates: The Operations unit started with a high termination rate, which dropped to zero towards 2015. Management has not lost many employees.

# 4. Predictive Analytics

## RQ1: Are We Losing More Employees?

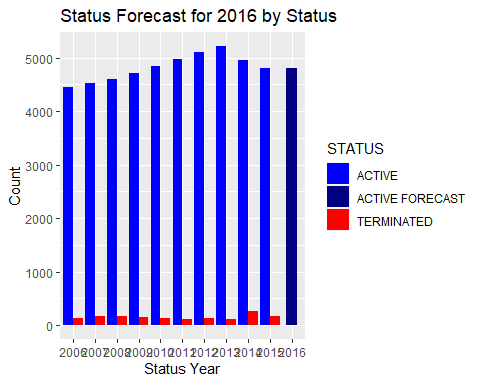


The number of active employees increases gradually with a slight drop in 2013. The company records the highest terminations during this year.

### Forecasting Active employees by 2016

## `summarise()` has grouped output by 'STATUS'. You can override using the  
## `.groups` argument.

## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 2016 4799.016 4613.55 4984.483 4515.37 5082.663



We created a model to verify what we suspected. In 2016 had about the same number of active employees as before.

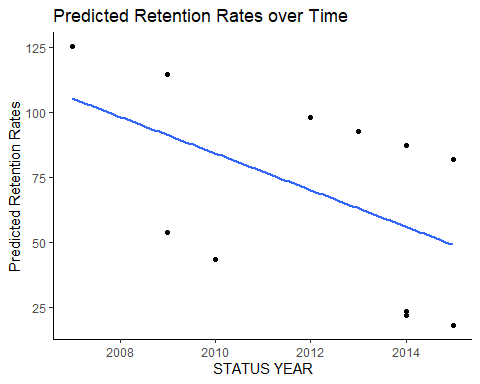
## RQ2: Which Departments are Likely to Lose the Most Employees?

### Model Selection for Retention Rate Prediction and Classification

##   
## Confidence set for the best model  
##   
## Method: raw sum of model probabilities  
##   
## 95% confidence set:  
## K AICc Delta\_AICc AICcWt  
## all 5 607.57 0 1  
##   
## Model probabilities sum to 1

AICc (Akaike Information Criterion with a correction for small sample sizes) evaluation gives a low AICc for the model with age, length of service and status as predictors of retention rate. Therefore, this is the best model. It has the lowest AICc of 607.57. The model with the functional unit was eliminated.

### Predicting Retention Rate Using Logistic Regression



The Logistic Regression model shows a reduction in the Retention Rates over the 10 year period.

### Predicting the Retention Rate for the Subsequent Years The Logistic Regression model was used along with extrapolation to predict the retention rates for the next year

## Predicted Retention Rate for 2016 - 2018 = 49.43964

The model predicts 49.43 as the retention\_rate for the subsequent year 2016 which is really low across the entire company. It must be noted however that the departments with Retention rate of 100% were eliminated from the dataset.

## RQ3: Which Employees are Likely to Leave the Company?

### Model Selection for Classification

## [1] "factor"

## K3 K5 K10 K30 K50   
## 0.7827160 0.7814815 0.7882716 0.8024691 0.8067901

This shows that all K50 has the highest accuracy of 80.6%. We use cross validation to explore more options and select the best k-value.

### Using Cross Validation to Select the Best K-value

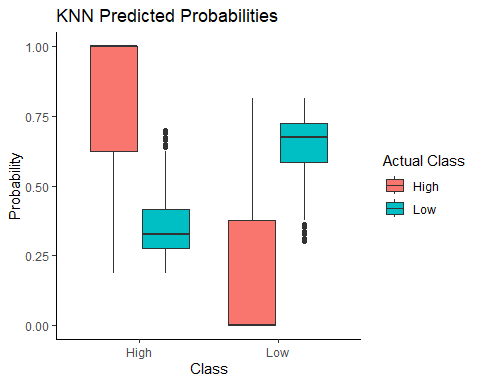
## High Low   
## 2520 1260

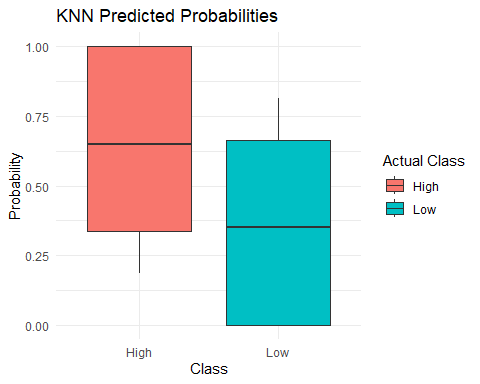
## Best k value = 80

The model predicts both classes accurately with an accuracy of 100%.

The best k-value is selected as 80 based on the cross validation. Therefore, we use this k-value to classify the employees into groups with the highest and lowest retention rates

### Applying K-NN to classify Retention Rate Classes

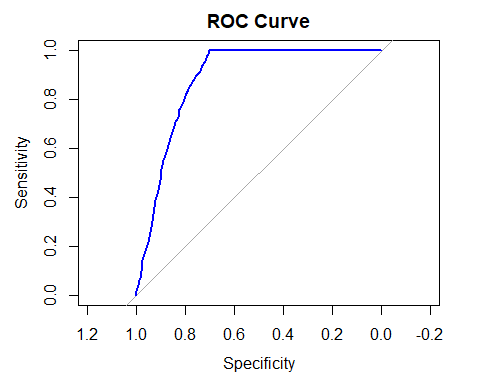




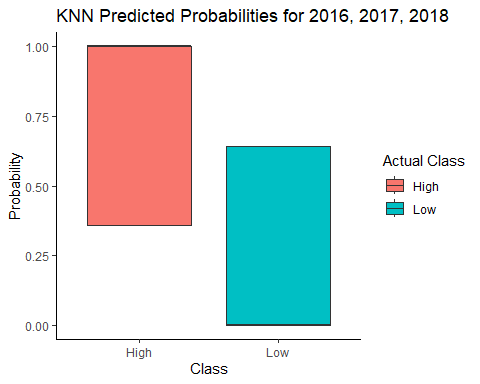
## Setting levels: control = High, case = Low

## Setting direction: controls > cases

## The AUC-ROC Area under the curve is 0.8809979

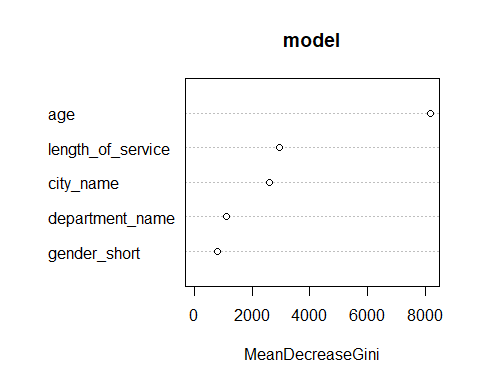


AUC\_ROC Area under the curve is 0.88 which is close to 1 therefore indicating the KNN model’s ability to correctly distinguish betweeen and predict both classes. Therefore,we can conclude that the predictions, very high accuracies and sensitivity is not a random behavior of the models.

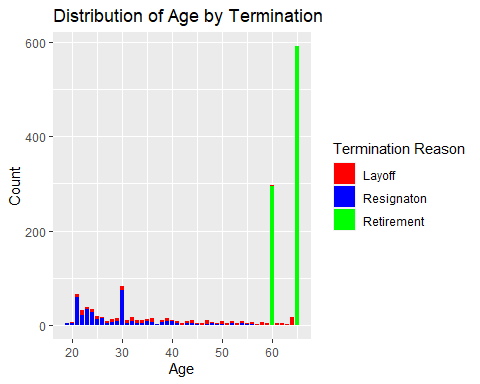


The model accurately classifies the employees with highest and lowest retention rates classes for the subsequent years. The prediction shows that all departments across the company will have a high retention rate. However, without test data, it is hard to verify these results.

## RQ4: What factors are influencing the most for Retention Rates?

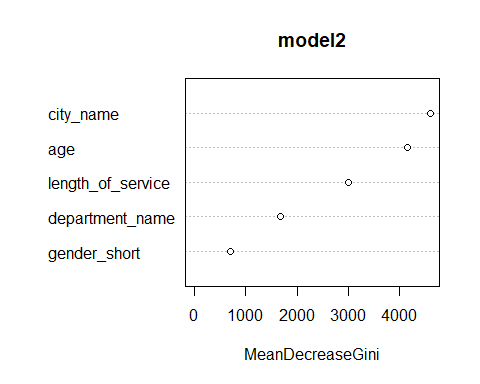


## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction ACTIVE TERMINATED  
## ACTIVE 13961 119  
## TERMINATED 489 326  
##   
## Accuracy : 0.9592   
## 95% CI : (0.9559, 0.9623)  
## No Information Rate : 0.9701   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.4981   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.9662   
## Specificity : 0.7326   
## Pos Pred Value : 0.9915   
## Neg Pred Value : 0.4000   
## Prevalence : 0.9701   
## Detection Rate : 0.9373   
## Detection Prevalence : 0.9453   
## Balanced Accuracy : 0.8494   
##   
## 'Positive' Class : ACTIVE   
##



The main factor is age, by far. Analyzing we realize that it is because most people retire at 65, if they didn’t retire earlier, usually when they were 60 years old.

To further analyze and predict we analyzed employees younger than 60 years old.



## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction ACTIVE TERMINATED  
## ACTIVE 8525 71  
## TERMINATED 394 43  
##   
## Accuracy : 0.9485   
## 95% CI : (0.9438, 0.953)  
## No Information Rate : 0.9874   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.1388   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.9558   
## Specificity : 0.3772   
## Pos Pred Value : 0.9917   
## Neg Pred Value : 0.0984   
## Prevalence : 0.9874   
## Detection Rate : 0.9438   
## Detection Prevalence : 0.9516   
## Balanced Accuracy : 0.6665   
##   
## 'Positive' Class : ACTIVE   
##

We found that the main factors are City and Age. Age seems to be still an important factor even after removing employees older than 60 years old.

