Capstone Project - Choose Your Own

Stevonne Nugent

1/7/2021

Executive Summary

The aim of this paper is to predict the likelihood of an employee meeting greater than 80% of their key performance indicators given an employee's age; gender; average training scores received; length of service; department; region; education level; channel of recruitment; receiving a promotion; receiving awards; rating from the previous year; and the number of trainings received.

A human resource dataset consisting of 54,808 observations was cleaned and separated into train and test datasets which accounted for 80% and 20% of the transformed dataset, respectively, to mitigate the possibility of overfitting.

Training and tuning were done for three machine learning models – Logistic Regression, Linear Discriminant Analysis and Random Forest – on the training dataset using 5-fold cross validation. The model that has the highest area under the receiver operating characteristic curve (ROC) across the 5 resamples was chosen to predict the likelihood of an employee meeting more than 80% of their key performance indicators in the test dataset.

Data analysis was undertaken to examine and explore the data to looks for trends and potential relationships/interactions with the dependent variable.

The best performing model was the random forest model with an ROC of 0.7612. This model was selected and then applied to the test dataset and produced an ROC of 0.689, and an accuracy of 0.7470. Variable importance was calculated for all three models, which revealed for the random forest model that the variable previous year rating was the most important in determining the likelihood an employee would meet more than 80% of their key performance indicators (KPIs).

Section 1: Introduction

Companies and institutions worldwide are all interested in having an efficient and hardworking staff to produce the best results to improve or maintain an overall good performance. It is against this background that companies and institutions worldwide employ an evaluation system for their staff typically using a list of key performance indicators. Given the desire to be top performing companies and institutions it would be beneficial for management teams to know the likelihood their staff will meet more than 80% of their KPIs, which is considered a gold standard for employee performance.

In this report, the goal is to build a model that will predict the likelihood of an employee meeting more than 80% of their KPIs (variable – kpi). The HR dataset "train" from (Shivan Kumar 2020)[1] was used, which contained information on 54,808 employees and if they met more than 80% of their KPIs.

The remainder of the paper is outlined as follows:

- Section 2 – Data Description, Exploration & Wrangling
- Section 3 – Methodology

[1] Accessed 2020-12-14: https://www.kaggle.com/shivan118/hranalysis?select=train.csv

- Section 4 – Results
- Section 5 – Conclusion.

Section 2: Data Description, Exploration & Wrangling

This section gives a concise data description of the HR dataset that will be used for the analysis, as well as data wrangling undertaken to clean the data for use.

Data Description

The HR dataset used contained 54,808 rows and 14 columns with missing values. For simplicity of the data analysis all observations with missing values were removed, resulting in a dataset containing 48,660 rows and 14 columns. The 14 columns were the following variables:

- employee id – unique employee identification number, there are 48,660 employees
- department – department names, there are 9 unique departments
- region – the regions employees are in, there are 34 unique regions
- education – the education level of employees, there are 3 unique levels
- gender – the gender of an employee, male or female
- recruitment channel – the channel through which an employee was recruited, there are 3 unique channels
- number of trainings – represents the number of training sessions an employee attended
- age – age of the employee
- previous year rating – the rating an employee received the previous, there are 5 unique ratings ranging from 1 to 5
- length of service – the number of years an employee has been employed
- kpi – represents whether or not an employee met $> 80\%$ of their KPIs for the year, 1 means yes and 0 means no
- awards won – represents whether or not an employee received an award, 1 means yes and 0 means no
- average training score – the average training score received by an employee
- is promoted – represents whether or not an employee received a promotion, 1 means yes and 0 means no.

Data Wrangling

Initial data exploration highlighted that some variable names were not in a tidy format, the format of the dependent variable needed to be converted, the dataset contained NAs, as well as categorical variables were listed as numeric. The following changes were made:

1. delete all observations that had NAs for simplicity in the analysis
2. convert the following variables to categorical/factor, variables: department, region, education, gender, recruitment channel, kpi, awards won and is promoted
3. convert the kpi from ones and zeros to yes and no
4. clean the variable names, that is, all variable names were written in lower cases and joined with an underscore
5. employee_id was drop from the analysis as it represents the unique identification of a employee that was chosen randomly.

Table 1a
First 6 Rows and 7 Columns of the Processed HR Dataset

department	region	education	gender	recruitment_channel	no_of_trainings	age
Sales & Marketing	region_7	Master's & above	f	sourcing	1	35
Operations	region_22	Bachelor's	m	other	1	30
Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34
Sales & Marketing	region_23	Bachelor's	m	other	2	39
Technology	region_26	Bachelor's	m	other	1	45
Analytics	region_2	Bachelor's	m	sourcing	2	31

Table 1b
First 6 Rows and Last 6 Columns of the Processed HR Dataset

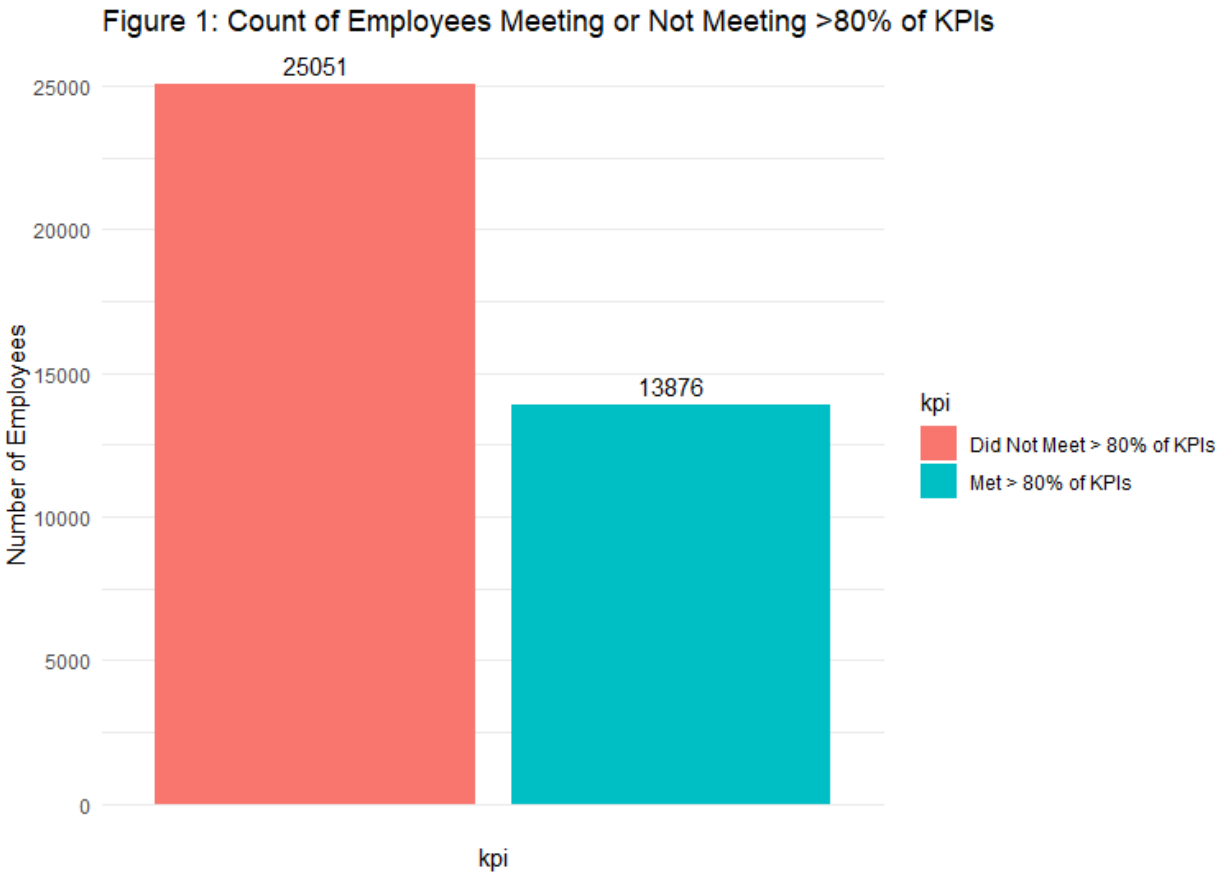
previous_year_rating	length_of_service	kpi	awards_won	avg_training_score	is_promoted
5	8	yes	0	49	0
5	4	no	0	60	0
3	7	no	0	50	0
1	10	no	0	50	0
3	2	no	0	73	0
3	7	no	0	85	0

The aim of this paper is to build a model with the ability of predicting the likelihood of an employee meeting more than 80% of their KPIs, i.e. kpi is the outcome variable and the other variables will be predictor variables. To mitigate against the possibility of overfitting the models the transformed HR dataset, which contains 48,660 employees was randomly split on the dependent variable, kpi, to preserve the overall class distribution of the data into hr_train and hr_test, which represented 80% and 20% of the transformed dataset, respectively. The former was used to train and tune the model and the latter was used to evaluate the final model. Given the size of the data, the 80/20 split for training and testing, respectively took into consideration the computational cost of training and evaluating the model, as well as ensuring that variance is not too high in parameter estimates and the model performance statistics.

Data Exploration

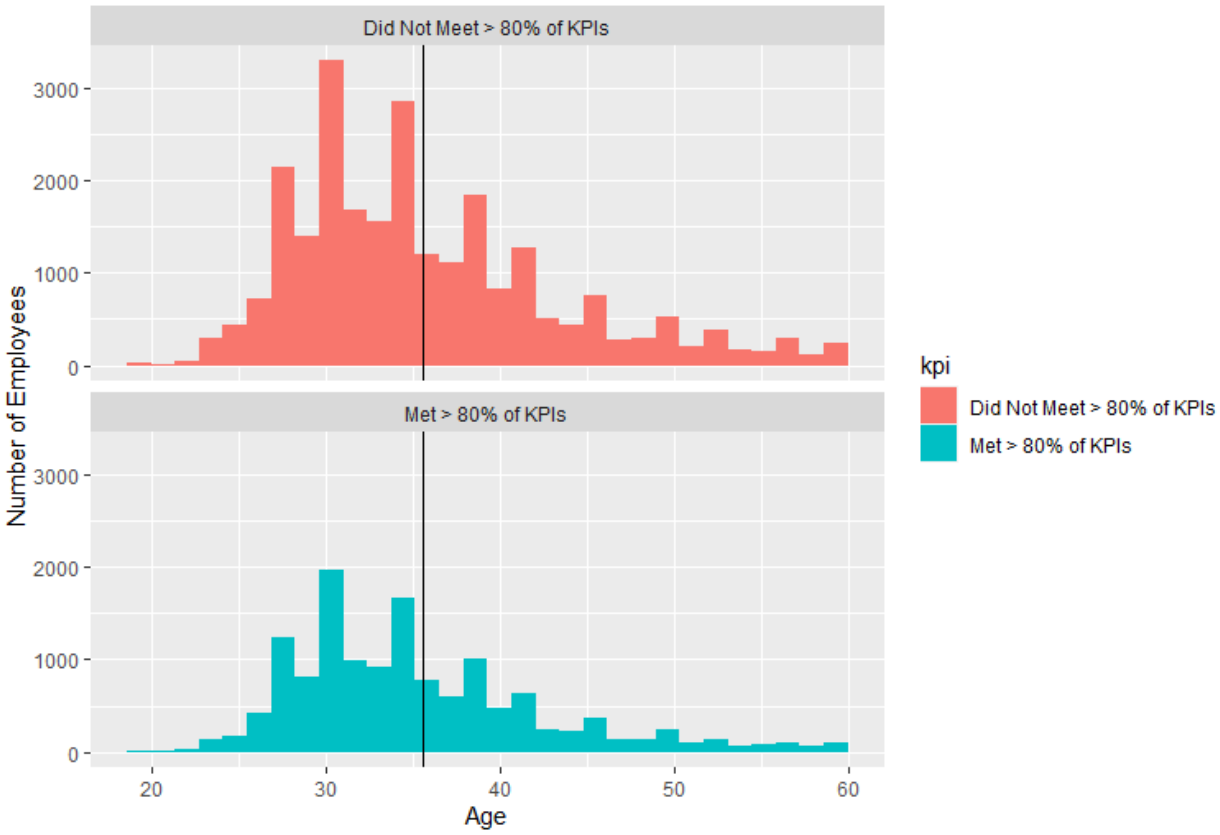
The variables in the hr_train dataset were explored individually with the dependent variable (kpi) to examine possible relationships.

- of the 38,927 employees 13,876 met more than 80% of their KPIs (kpi), which represented a share of 35.6% (Figure 1)



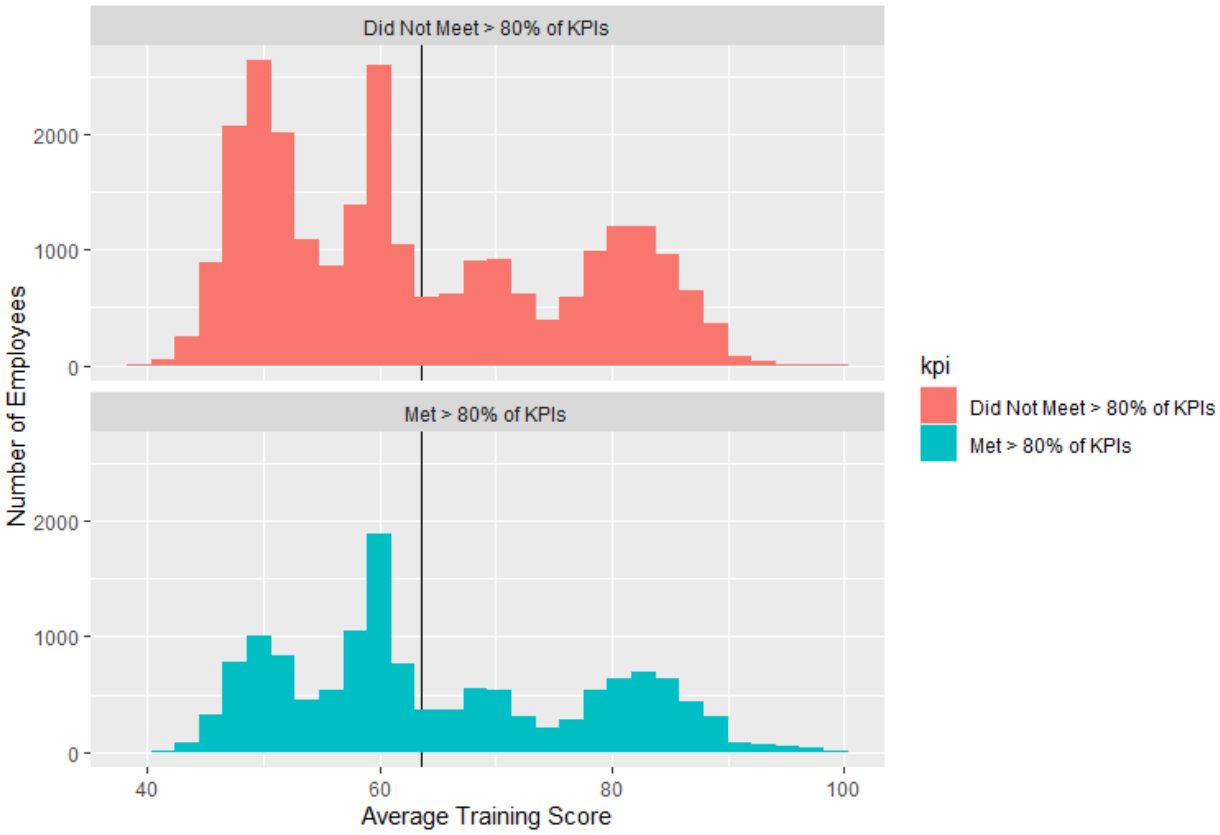
- age: the largest share of employees is between the ages of 30 and 35
- as shown in Figure 2 the age distribution does not vary by much whether employees met more than 80% of their KPIs or did not meet more than 80%

Figure 2: Employee Age Distribution by kpi

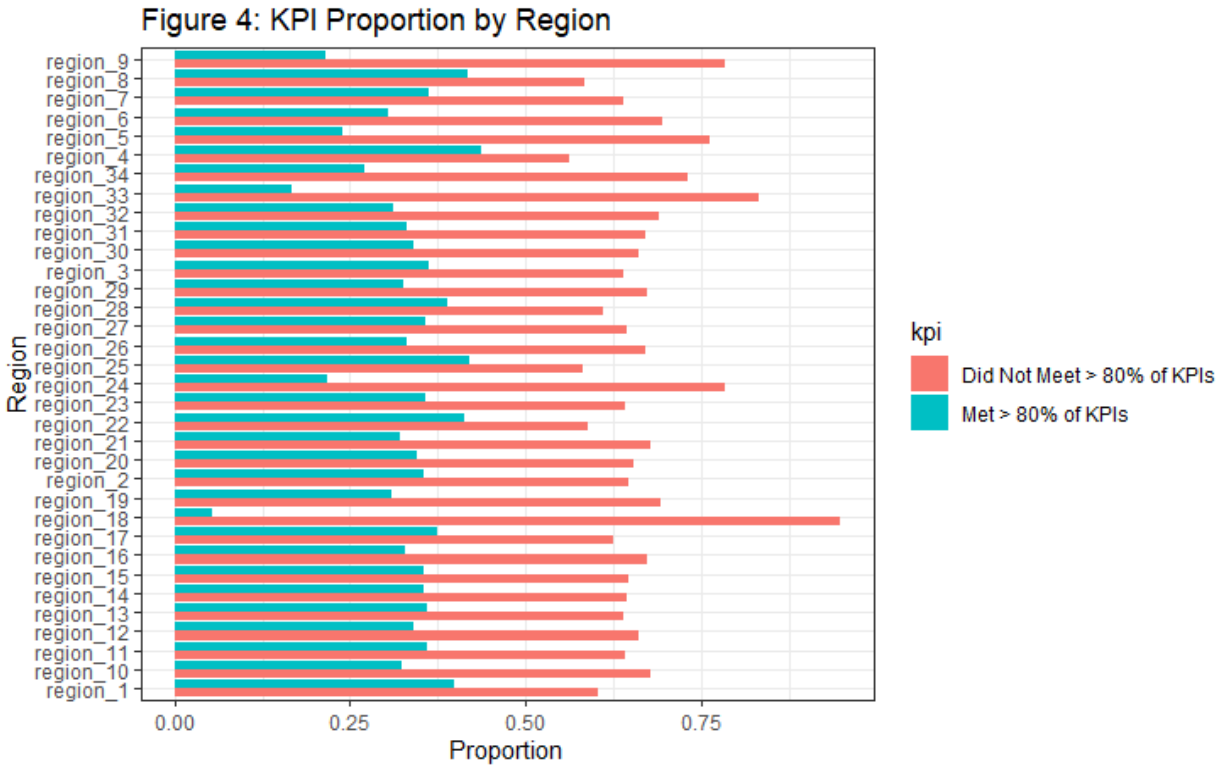


- average training score: the distribution of the average training score for persons who met greater than 80% of their KPIs seems to be different than those who did not meet greater than 80% of their KPIs (Figure 3)

Figure 3: Employee Average Training Score Distribution by kpi



- the proportion of meeting more than 80% of their KPIs by region shows that there is some variation, for example region_22 had an approximately 41.0% of its employees meeting more than 80% of their KPIs, while region_34 had approximately 27.0% of their employees meeting more than 80% of their KPIs (Figure 4)

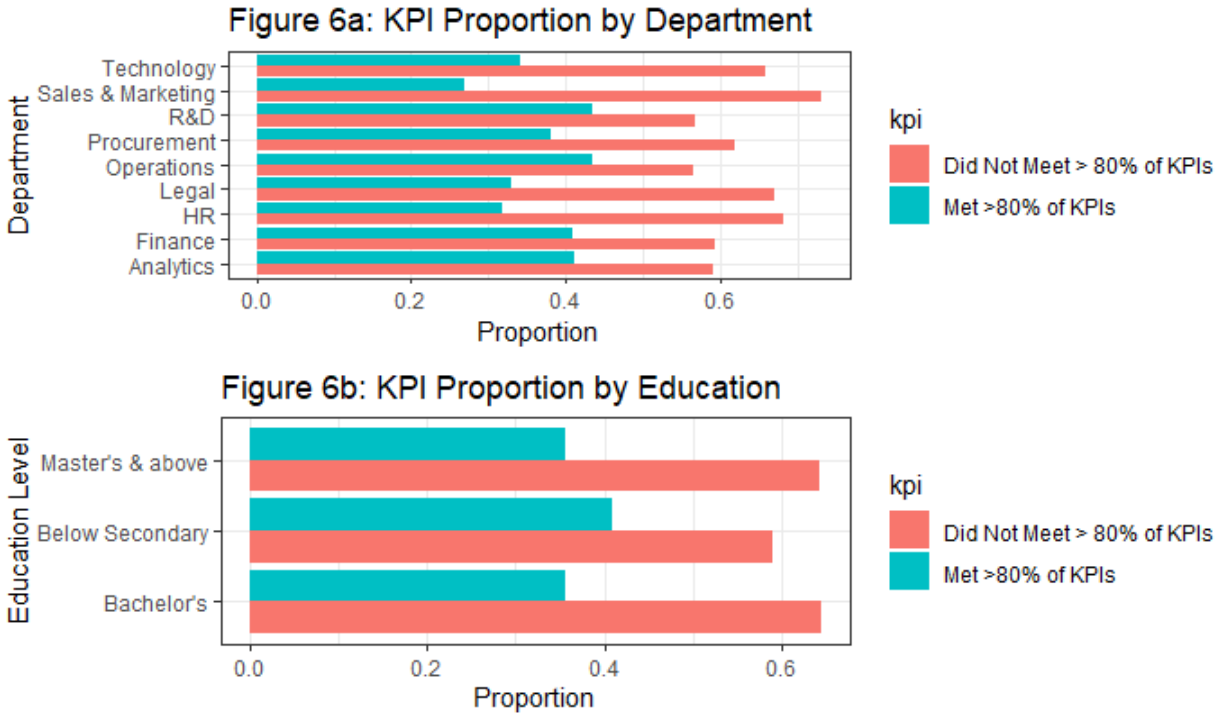


- the median age of employees was similar for both groups, those that met more than 80% of their KPIs and those that did not (Figure 5a)
- The median length of services for employees was similar for those that met more than 80% of their KPIs and those that did no (Figure 5b)
- The median average training score for employees that met more than 80% of their KPI was slightly above persons that did not meet above 80% of their KPI (Figure 5c).

Figure 5: Boxplots Showing Age, Length of Service & Average Training Score by KPIs

- based on the proportion, department for an employee appears to matter in an employee meeting more than 80% of their key performance indicators (Figure 6a)
- persons with Below Secondary Education have a marginally higher proportion of persons that met greater than 80% of their KPIs, relative to persons with Bachelors and Masters & above (Figure 6b)

Figure 6: Bar Charts Showing KPI Proportion for Department and Education Level

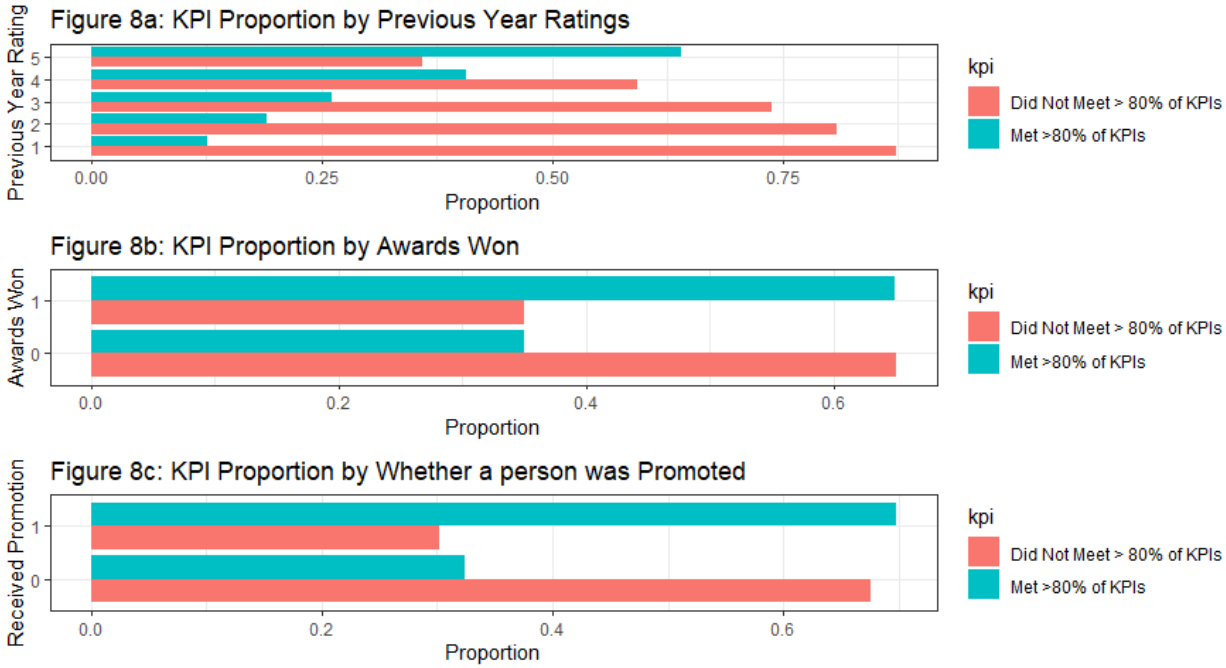


- the proportion of females that met more than 80% of their KPIs was greater than the proportion of males (Figure 7a)
- recruitment channel showed the largest differential for when an employee was referred (Figure 7b)
- the number of trainings received seems to influence whether or not an employee meets more than 80% of their KPIs. For example, no person that received 9 or more trainings met greater than 80% of their KPIs (Figure 7c).

Figure 7: Bar Charts Showing KPI Proportion for Gender, Recruitment Channel and Number of Training Received

- the rating received in the previous year seems to have a large impact on an employee meeting more that 80% of their KPIs, as shown by the wide differential between meeting more than 80% of KPIs and not meeting more that 80% of KPIs (Figure 8a)
- receiving an award seems to influence whether someone met more than 80% of their KPIs, while persons who did not receive an award are more likely to not meet more than 80% of their KPIs (Figure 8b)
- similar to awards won, persons who received a promotion are more likely to meet more than 80% of their KPIs, while a person who was not promoted are more likely not to meet over 80% of their KPIs (Figure 8c).

Figure 8: Bar Charts Showing KPI Proportion for Previous Year Ratiing, Receiving an Award and Receiving a Promotion



Section 3: Methodology

The data exploration showed that the target variable was a categorical variable with two levels as its output "1 – yes" or "0 – no" and as such, we have a classification problem, therefore the modelling approaches used needed to consider this. The modelling approaches used were:

- General Linear Model – Logistic Regression
- Linear Discriminant Analysis
- Random Forest Model.

All the models were built and tuned using the training dataset (hr_train) with a 5-fold cross validation (CV) to improve the robustness of the results. A 5-fold CV was used as indicated by the work of (Krstajic et al. 2014), where in their work of comparing 5-fold CV and 10-fold CV showed there was not a presence of a significant difference in model results between a 5-fold CV and a 10-fold CV.

The model that has the highest area under the receiver operating characteristic (ROC) curve was chosen to predict the likelihood of an employee meeting more than 80% of their KPIs in the test dataset.

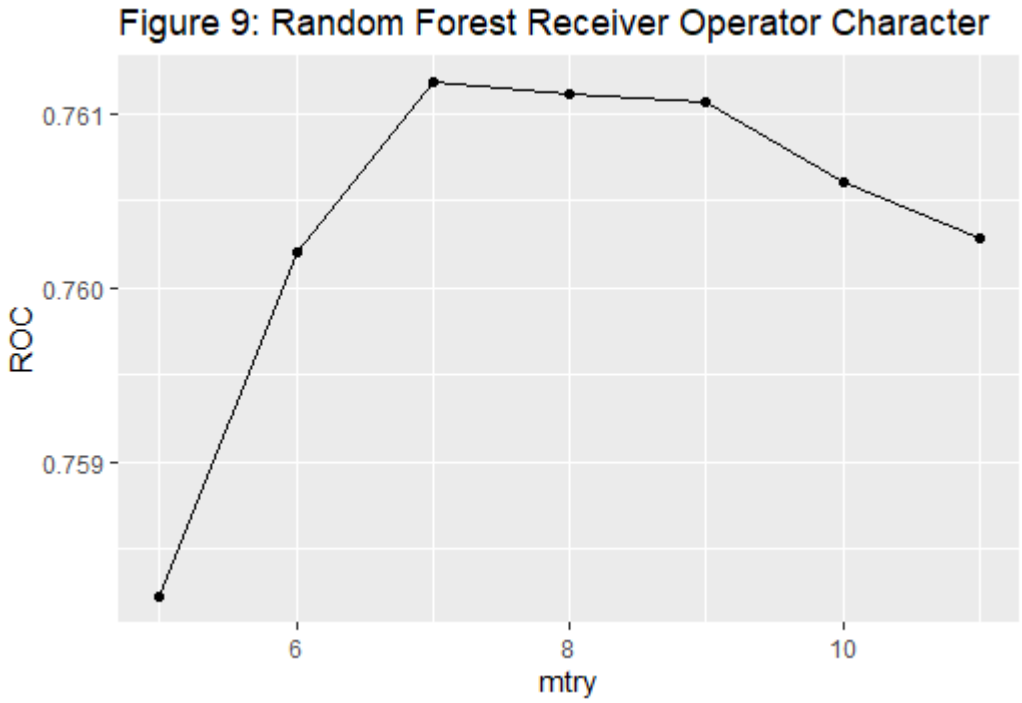
In training and tuning the Random Forest model the parameters the optimal tuning parameter that produced the largest ROC was chosen.

Variable importance was calculated for each model and each model was compared to examine which model produced the best area under the curve for the receiver operating characteristic (ROC). The receiver operation curve is a probability curve that plots the true positive rate against the false positive rate at different threshold settings. It tells how much a model is capable of distinguishing between classes. Higher the ROC, better the model is at predicting 0s as 0s and 1s as 1s.

Section 4: Results

This section presents the modeling results and discusses model performance. The random forest was the only model used that had a tuning parameter.

The tuning parameter that resulted in the highest area under the receiver operating curve (ROC) was chosen as the final model for the Random Forest algorithm (Figure 9). The optimal ROC of 0.7612 was chosen when mtry (number of randomly selected predictors) was 7.



A plot of variable importance showed that for both the General Linear Model and the Linear Discriminant Analysis the top 2 most important variables were the average training score and length of service, while the 2 least important variables were age and the recruitment channel of the employee (Figure 10a and 10b). However, for the Random Forest model the 2 most important variables were the previous year rating and and receiving a promotion, while the 2 least important variables were number of trainings received and age (Figure 10c).

Figure 10: Variable Importance

Figure 10a: GLM

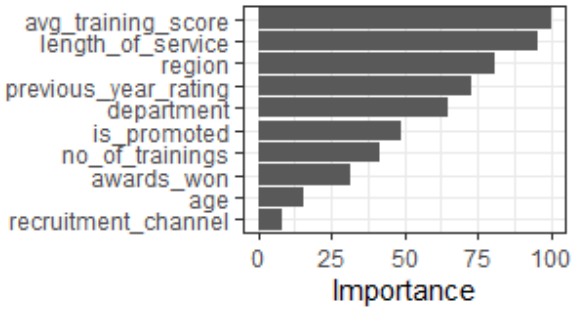


Figure 10b: LDA

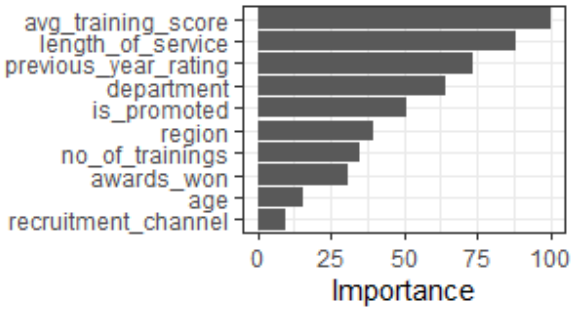
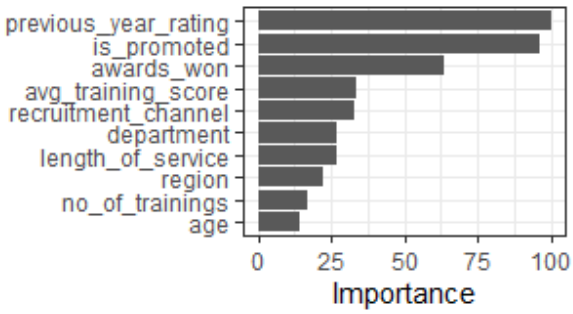


Figure 10c: Random Forest



Comparing the metrics of each model revealed that the random forest model produced the largest out-turn for 2 out of 3 of the metrics measured, the ROC and the sensitivity (Figure 11). An ROC of 0.7581 was yielded for the General Linear Model; 0.7583 for the Linear Discriminant Analysis model; and 0.7612 for the Random Forest model. While a sensitivity of 0.8654 was yielded for the General Linear Model; 0.8632 for the Linear Discriminant Analysis model; and 0.8775 for the Random Forest model. The Random Forest Model was chosen as the best model based on the out-turn of its ROC.

Figure 11: Comparing Metrics for Each Model

Figure 11a: Comparing the ROC for Each Model

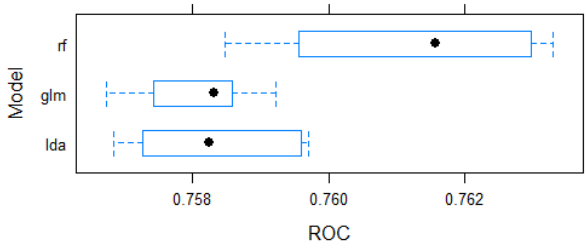


Figure 11b: Comparing the Sensitivity for Each Model

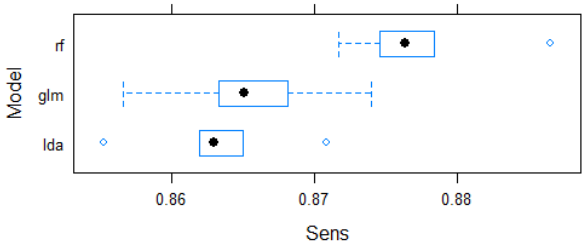
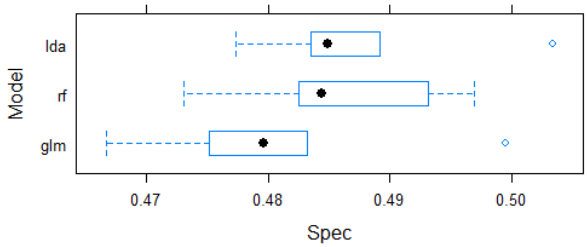


Figure 11c: Comparing the Specificity for Each Model



Applying the best model, the random forest model, to the test dataset (hr_test) returned an ROC of 0.689 with an accuracy of 0.7470 (Figure 12 and Table 2). The accuracy was higher than the no information rate of 0.6435, as shown below in Table 2, indicating that the model added value.

Figure 12: Plot Calculating ROC

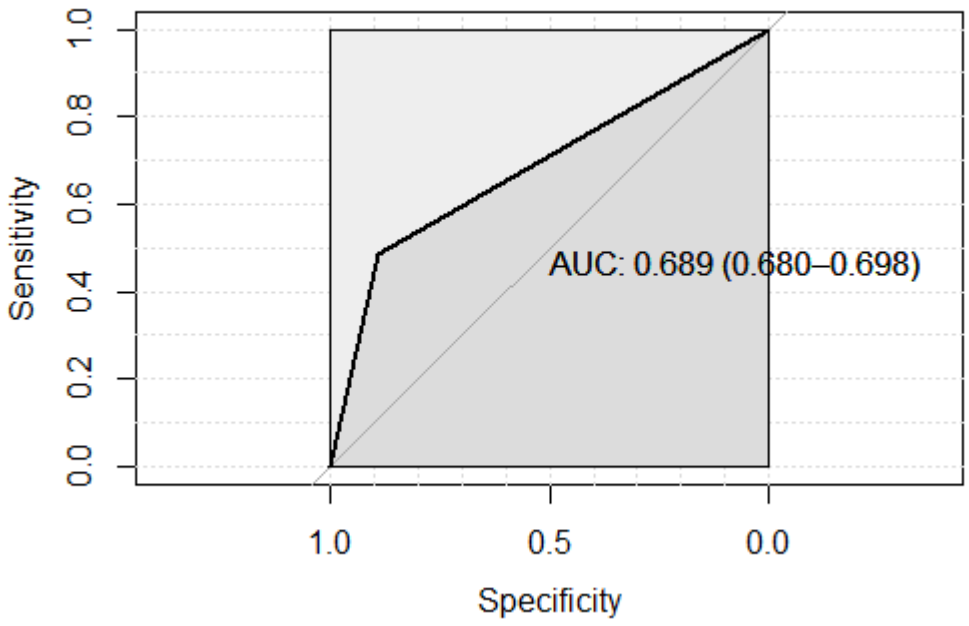


Table 2: Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  5580 1779
##        yes  683 1691
##
##                Accuracy : 0.747
##                  95% CI : (0.7383, 0.7557)
##     No Information Rate : 0.6435
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4069
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4873
##             Specificity : 0.8909
##          Pos Pred Value : 0.7123
##          Neg Pred Value : 0.7583
##              Prevalence : 0.3565
##          Detection Rate : 0.1737
##    Detection Prevalence : 0.2439
##       Balanced Accuracy : 0.6891
```

```
##
##        'Positive' Class : yes
##
```

Section 5: Conclusion

The project's objective was to predict the likelihood of an employee meeting more than 80% of their KPIs from a dataset of 48,660 observations. To achieve this the impact of age; gender; average training scores received; length of service; department; region; education level; channel of recruitment; receiving a promotion; receiving awards; rating from the previous year; and the number of trainings received was examined.

The data was cleaned, then split into a training and test set. A total of three machine learning algorithms were trained and tuned using 5-fold cross validation. The 5-fold cross validation yielded an ROC of 0.7581 for the General Linear Model; 0.7583 for the Linear Discriminant Analysis model; and 0.7612 for the Random Forest model (largest ROC, best model). The Random Forest model was then applied to the test/hold out dataset which produced an ROC of 0.689 and an accuracy of 0.7470.

Limitations

- 6,148 employees where excluded from the analysis because of missing variables, this has the potential of biasing the results.
- as indicated by (Muschelli 2019) using the area under the curve as a performance measuring metric for categorical variables can be misleading, and as such this might have posed a potential bias for the results of this paper

Future Work

To improve the rigour of the analysis, one could have estimated the missing values using various techniques, for example, median imputation and k-nearest neighbour (KNN) imputation.

References

Krstajic, Damjan, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. 2014. "Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models." *Journal of Cheminformatics* 6 (1). https://doi.org/10.1186/1758-2946-6-10.

Muschelli, John. 2019. "ROC and AUC with a Binary Predictor: A Potentially Misleading Metric." *Journal of Classification* 37 (3): 696–708. https://doi.org/10.1007/s00357-019-09345-1.