SHOULD I STAY OR SHOULD I GO: UNDERSTANDING EMPLOYEE ATTRITION IN THE WORKPLACE

THEWINNINGTEAM()

EMILY GOULD
DANYLO KOVALENKO
JAMIE DANCE
JENNA WILKES
PRAKASH JHA
WILLIAM HEMSLEY

MSC DATA SCIENCE: SEMESTER 1 2020
COMM053 PRACTICAL BUSINESS ANALYTICS

TABLE OF CONTENTS

INTRODUCTION

The following report will approach the business problems associated with employee attrition within organisations, using practical business techniques and a data-driven approach. The business problem and objectives will be defined in detail, the data and data mining techniques used to meet these business objectives discussed, and a guided step-by-step outline of the process shared. The project will follow the CRISP-DM iterative methodology, with each choice and assumption explained. By the end of the report, the business problem will be evaluated, with results and conclusions explored.

1 CLARITY OF BUSINESS UNDERSTANDING

1.1 BUSINESS PROBLEM OUTLINE

The recruitment process can be both costly and time-consuming for an organisation, with research by Robert Half UK estimating that HR Directors in the UK spend close to 28 days on average to recruit within their organisation [3]. There are also the costs of advertising the role, internal and external recruiter expenses, training the employees when they arrive and valuable HR staff time.

On average, around one in ten people change jobs each year [4] with more than half of UK workers planning on changing careers in the next 5 years [5]. According to research by Investec, millennials are likely to have 12 different jobs during their working lives [6] and in the US, a LinkedIn study states, young workers are likely to change jobs four times in their first 10 years after graduating [5]. The prospect of having a job or career for life is increasingly slim, and the likelihood of employees leaving the organisation or changing jobs is on the rise.

To reduce the rate of recruitment within an organisation, reducing employee attrition is key. Here, *employee attrition* is defined to be the number of employees that left the organisation in the previous year. If an employee has recently left the company, then attrition has occurred.

Employee attrition can be costly for an organisation, with research by Oxford Economics estimating the loss of an employee carrying an average financial impact of over £30,000 [7]. Also, in the nine months before an employee leaves, their overall engagement at work plummets [8] severely impacting productivity.

Furthermore, employee attrition may mean a loss of talent in the organisation. The progress and growth of an organisation rely on the skills and strength of the workforce; therefore, talent retention and hiring the right employees are drivers of company performance. 83% of HR leaders agree that talent is the top priority at their company, highlighting talent retention is ever-more pertinent [9].

1.2 BUSINESS OBJECTIVES

The first core business objective is to gain a clear understanding of the key factors that affect the problem of employee attrition in an organisation.

When the importance of these factors in attrition is identified, the next key objective is to be able to predict employee attrition based on these factors.

With this knowledge, the final objective is to identify steps that can be implemented by a business to reduce future employee attrition. The end-goal of this is to reduce the costs associated with attrition and to retain talented employees, benefitting the company in both the short and long-term.

1.3 DATA MINING AND MACHINE LEARNING GOALS

Through the use of data mining processes in the R programming language, we aim to be able to provide a clear answer to the business objectives.

A key criterion for success is producing a working machine learning model that is accurate and can generalise well for future datasets. A model can be said to *generalise* the data effectively if it works well with unseen data - not overfitting or underfitting the data used. The model should be able to discover underlying properties, relationships and patterns of the data; so it can predict attrition for future data as well. Without a generalised model, the model will not accurately predict attrition for other datasets.

To achieve this objective, metrics associated with precision will be explored and evaluated; and the performance of the model will be analysed using a test data set.

2 DATA UNDERSTANDING

2.1 INITIAL DATA COLLECTION

The dataset chosen to explore the business objectives was sourced online from Kaggle (https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study). The dataset is titled ‘HR Analytics Case Study’ and was last updated 2 years ago. This file was chosen as it had an appropriate size and each file had an appropriate number of fields. The dataset directly related to the business objectives as it included an attrition field for every record.

2.2 DESCRIBE DATA

The HR Analytics Case Study contained the following files:

- *employee_survey_data.csv* - survey data conducted on all employees within the organisation asking employees to rate the working environment, job satisfaction and work-life balance.
- *general_data.csv* - information on each of the employees (demographics, company role, job role, monthly salary, etc.)
- *manager_survey.csv* - a survey from managers, scoring employee job involvement and performance.

- *in_time.csv* - log-in times for each employee (from 01/01/2015 - 31/12/2015)
- *out_time.csv* - log-out times for each employee (from 01/01/2015 - 31/12/2015)

The data in each CSV file is either numeric or categorical, thus the data is structured. The field '*Employee ID*' is common between all of the CSV files and uniquely describes each of the employees within the organisation. Therefore, this field is called a *primary key* and can be used to link each table to form a single data frame. To verify '*Employee ID*' is the primary key, the number of records in each dataset was compared to the number of unique IDs and it was checked whether these unique IDs matched in each dataset.

2.3 INITIAL DATA EXPLORATION

Initial data exploration began by examining the data dictionary provided with the dataset. '*Relationship Satisfaction*' was listed in the data dictionary but was not used in the given dataset, so this variable was deleted from the data dictionary. The CSV files were then loaded into R as a data frame, allowing the identification of the fields and format of the rows.

After inspecting *'in_time.csv'* and *'out_time.csv'*, the dates were not formatted to be easily read in R, so they were manipulated to be legible date and time formats. A numeric value representing the hours spent at work that day was calculated, then the mean hours worked per day across 260 working days provided in the files found. This represented how long an employee tended to work each day and would enable us to test the hypothesis that working time is a significant factor in attrition. Missing values were ignored using *na.rm* condition during the calculation of the row means and the final data was stored in a new data frame using *'Employee ID'*.

There exists a field in '*general_data.csv'* called '*Standard Hours*' containing only the value *'8'* for all records, suggesting that employees are expected to work 8 hours as standard per day. Despite the standard 8 hours per day, the median of the data collected for worked hours falls below this value, suggesting the majority of staff do not work overtime.

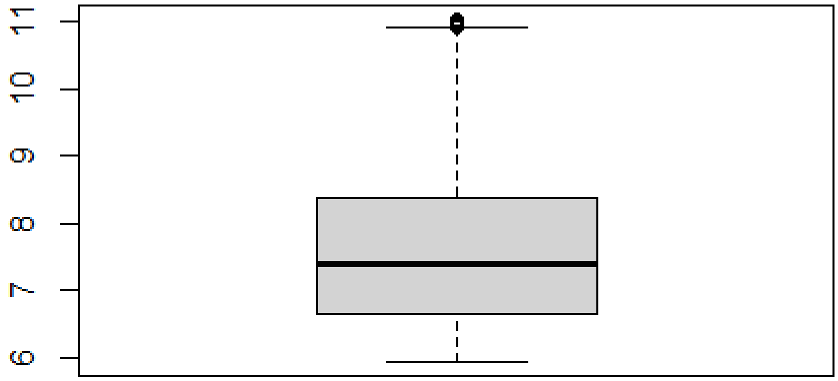


Figure 1: Boxplot representing the dispersion of average hours worked by employees

Figure 1 shows the data has a slight negative skew. The lowest average hours worked was approximately 6 hours with the largest average hours worked 11 hours. Hence, the range should be broad enough to find valuable relationships between attrition and number of hours worked.

To better relate to the business objectives and to further explore the hypothesis that working hours impact attrition, the mean working hours were encoded into three categories:

- '*Early Logout'* (average working hours of less than 7 hours)
- *'Standard'* (average working hours between 7 and 9 hours)
- '*Overtime'* (average working hours greater than 9 hours)

In summary, the data from *'in_time.csv'* and *'out_time.csv'* had been reduced to a single variable containing 3 possible values, {*Early Logout*, *Standard*, *Overtime*}. Feature construction had therefore been used to reduce the initial dimensionality, deriving new fields that were more informative to the business objectives.

Finally, to compare and use the data from all four CSV files given, they were merged into a single data frame using '*Employee ID*'. This was exported as a new CSV file into the working directory and is the file used in our models.

2.4 EXPLORING THE MERGED DATA

The dimensions of the new data frame were established: 31 columns and 4410 rows. Summary statistics and structure were noted for each of the variables in the data frame.

2.4.1 Exploring Field Attributes:

At this stage of the data exploration, there were 22 numeric columns. The following fields were found to be numeric:

- EmployeeID
- EnvironmentSatisfaction
- JobSatisfaction
- WorkLifeBalance
- JobInvolvement
- PerformanceRating
- Age
- DistanceFromHome
- Education
- EmployeeCount
- JobLevel
- MonthlyIncome
- NumCompaniesWorked
- PercentSalaryHike
- StandardHours
- StockOptionLevel
- TotalWorkingYears
- TrainingTimesLastYear
- YearsAtCompany
- YearsSince LastPromotion
- YearsWithCurrManager
- meanWorkingHours

To explore the numeric fields, the quartiles for each were calculated, demonstrating the quartiles in which most of the data lie. Null or missing values were ignored for this procedure. There were also 9 categorical columns:

- Attrition
- BusinessTravel
- Department
- EducationField
- Gender
- JobRole
- MaritalStatus
- Over18
- Undertime/Overtim

2.4.2 Initial Data Quality Check

The number of missing values in the data was found to determine the initial quality of the data frame. 111 values were missing, therefore the overall percentage missing in the data frame is 0.084%. These missing values were addressed in the data cleaning stage of the project.

Attrition is the dependent variable and was a field in the data containing values *'Yes'* and *'No'*. Employee attrition within the dataset was 16.1%, suggesting an unbalanced dataset - a potential source of bias when modelling.

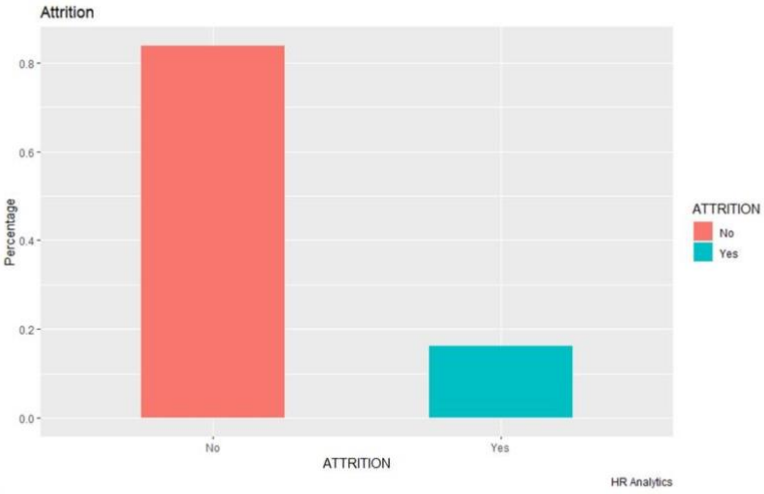


Figure 2: Balance of Attrition in the Dataset

2.5 DATA CLEANING

After initial data exploration, data cleaning was required before further graphical illustrations or pre-processing for modelling. Data cleaning improves readability and reduces the number of redundant rows and columns.

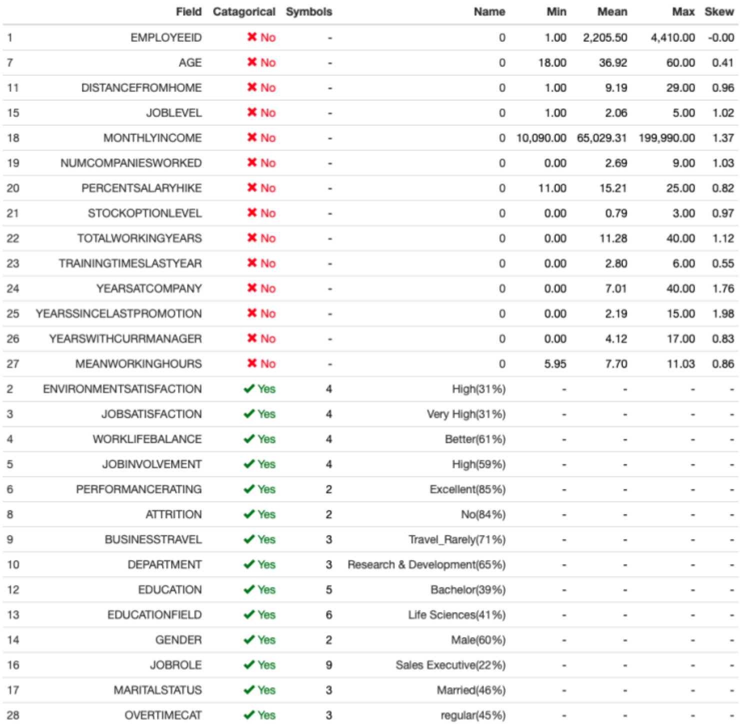
First, to improve the readability of the field names and to check consistency, the field names were capitalised and any punctuation removed.

Secondly, the rows were checked for duplicates so that any redundant information could be removed. However, no duplicate rows were found therefore no further action was necessary.

Next, the number of unique values in each field was calculated. After exploring the uniqueness of the variables, it was clear that '*Over18*' should be omitted as it only contained the value *'Y'* (i.e. all employees are adults). This field was therefore removed from the data as it offered no relevant information. Similarly, '*EmployeeCount*' and 'StandardHours' were redundant fields as they only included the values '*1*' and '*8'* respectively. Thus, '*EmployeeCount*' and '*StandardHours*' were also removed from the data frame.

The values returned from managers to describe employee performance, *'PerformanceRating',* were either '3' (high) or '4' (very high) out of a possible range of 1-4. As it is difficult to distinguish between these scores and no employee was underperforming, the field was removed from the data frame.

There were some instances of the formats of the categorical variables not matching the data dictionary, so these were changed to match. For example, '*Environmental Satisfaction*' was rated on a scale of *{'low', 'medium', 'high', 'very high'}* but was changed to the scale *{1,2,3,4}*. This affected 5 fields: '*Education*', 'Environment Satisfaction', *'Job Involvement*', '*Job Satisfaction*' and '*Relationship Satisfaction*'. At this stage, the labels of each of the variables were altered to match the data dictionary, resulting in 14 categorical variables, shown in figure 3.



	Field	Catagorical	Symbols	Name	Min	Mean	Max	Skew
1	EMPLOYEEID	✗ No	-	0	1.00	2,205.50	4,410.00	-0.00
7	AGE	✗ No	-	0	18.00	36.92	60.00	0.41
11	DISTANCEFROMHOME	✗ No	-	0	1.00	9.19	29.00	0.96
15	JOBLEVEL	✗ No	-	0	1.00	2.06	5.00	1.02
18	MONTHLYINCOME	✗ No	-	0	10,090.00	65,029.31	199,990.00	1.37
19	NUMCOMPANIESWORKED	✗ No	-	0	0.00	2.69	9.00	1.03
20	PERCENTSALARYHIKE	✗ No	-	0	11.00	15.21	25.00	0.82
21	STOCKOPTIONLEVEL	✗ No	-	0	0.00	0.79	3.00	0.97
22	TOTALWORKINGYEARS	✗ No	-	0	0.00	11.28	40.00	1.12
23	TRAININGTIMESLASTYEAR	✗ No	-	0	0.00	2.80	6.00	0.55
24	YEARSATCOMPANY	✗ No	-	0	0.00	7.01	40.00	1.76
25	YEARSSINCELASTPROMOTION	✗ No	-	0	0.00	2.19	15.00	1.98
26	YEARSWITHCURRMANAGER	✗ No	-	0	0.00	4.12	17.00	0.83
27	MEANWORKINGHOURS	✗ No	-	0	5.95	7.70	11.03	0.86
2	ENVIRONMENTSATISFACTION	✓ Yes	4	High(31%)	-	-	-	-
3	JOBSATISFACTION	✓ Yes	4	Very High(31%)	-	-	-	-
4	WORKLIFEBALANCE	✓ Yes	4	Better(61%)	-	-	-	-
5	JOBINVOLVEMENT	✓ Yes	4	High(59%)	-	-	-	-
6	PERFORMANCERATING	✓ Yes	2	Excellent(85%)	-	-	-	-
8	ATTRITION	✓ Yes	2	No(84%)	-	-	-	-
9	BUSINESSTRAVEL	✓ Yes	3	Travel_Rarely(71%)	-	-	-	-
10	DEPARTMENT	✓ Yes	3	Research & Development(65%)	-	-	-	-
12	EDUCATION	✓ Yes	5	Bachelor(39%)	-	-	-	-
13	EDUCATIONFIELD	✓ Yes	6	Life Sciences(41%)	-	-	-	-
14	GENDER	✓ Yes	2	Male(60%)	-	-	-	-
16	JOBROLE	✓ Yes	9	Sales Executive(22%)	-	-	-	-
17	MARITALSTATUS	✓ Yes	3	Married(46%)	-	-	-	-
28	OVERTIMECAT	✓ Yes	3	regular(45%)	-	-	-	-

Figure 3: Field information generated by 'formattable' package

The missing values and issues in the quality of the data were then addressed. To help determine how to deal with these values, the percentage of missing values in each field was calculated. The following figure describes where these missing values were located:

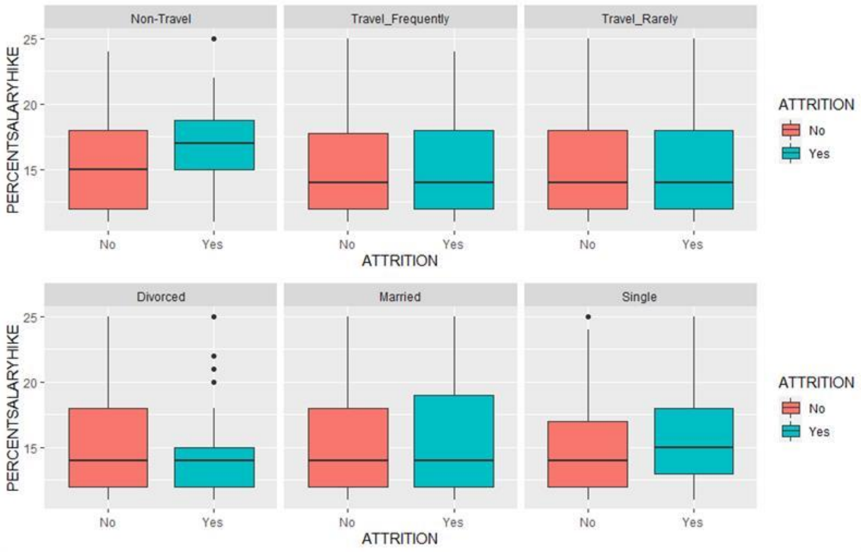
Variable Name	Number of Missing Values	Percentage Missing (%)
Environment Satisfaction	25	0.567%
Job Satisfaction	20	0.453%
Work-Life Balance	38	0.862%
Total Working Years	9	0.204%
Number of Companies Worked At	19	0.431%

Figure 4: Missing Values Table

Since the percentage of missing values in each field was less than 1%, any rows containing null or missing values were removed. Removing such a low percentage of data would have no significant consequences for the results. If the percentage of missing values had been above 1%, they would have been replaced with the median (for numeric data) or the mode (for categorical data). A final check that the missing values had been removed was performed, confirming all of the missing values had been dealt with.

2.6 INITIAL DATA VISUALISATION

After the initial data cleaning, the data was visualised. The following plots were produced:



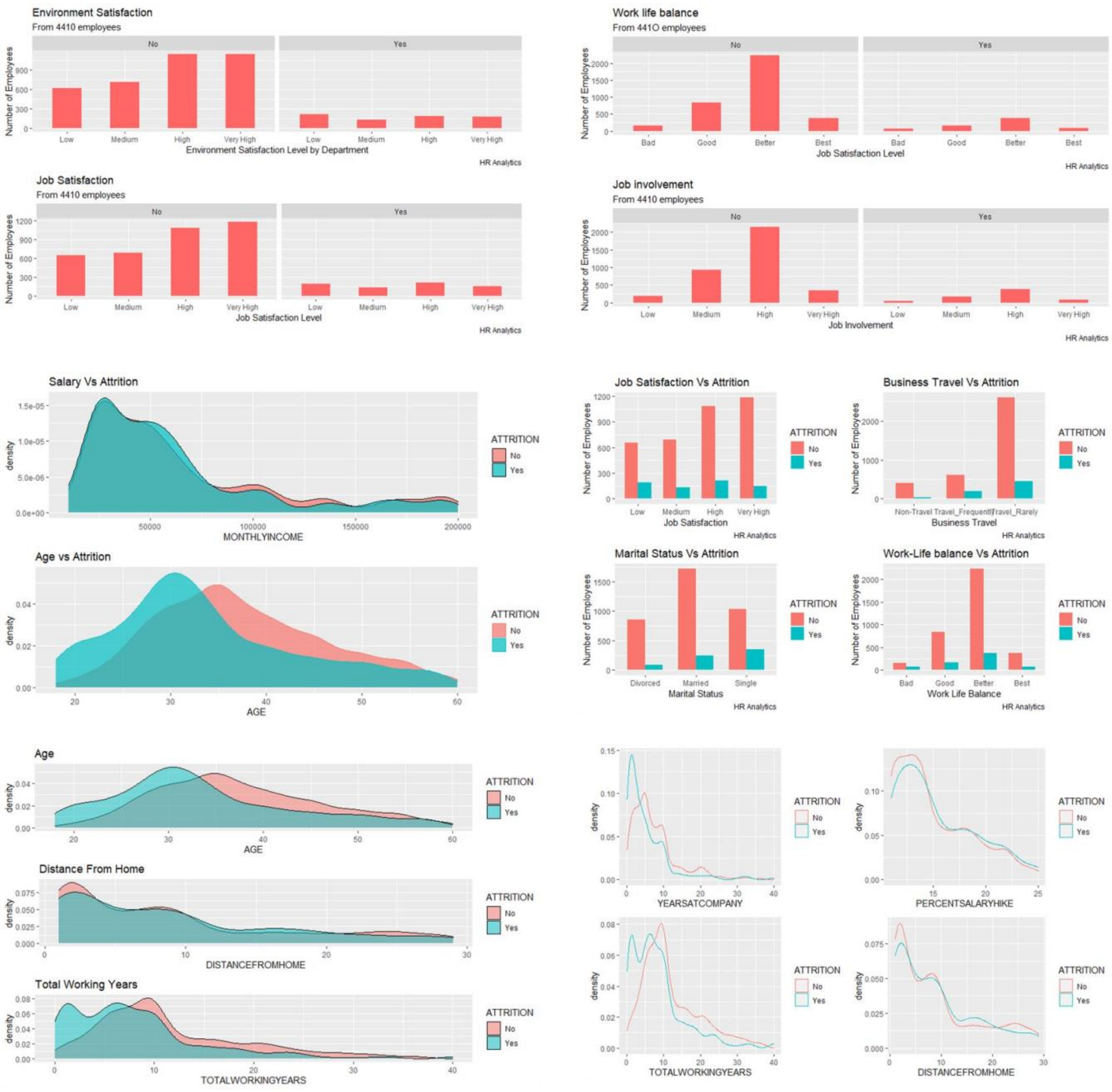


Figure 5: Array of Examples of Data Visualisations

3 DATA PRE-PROCESSING

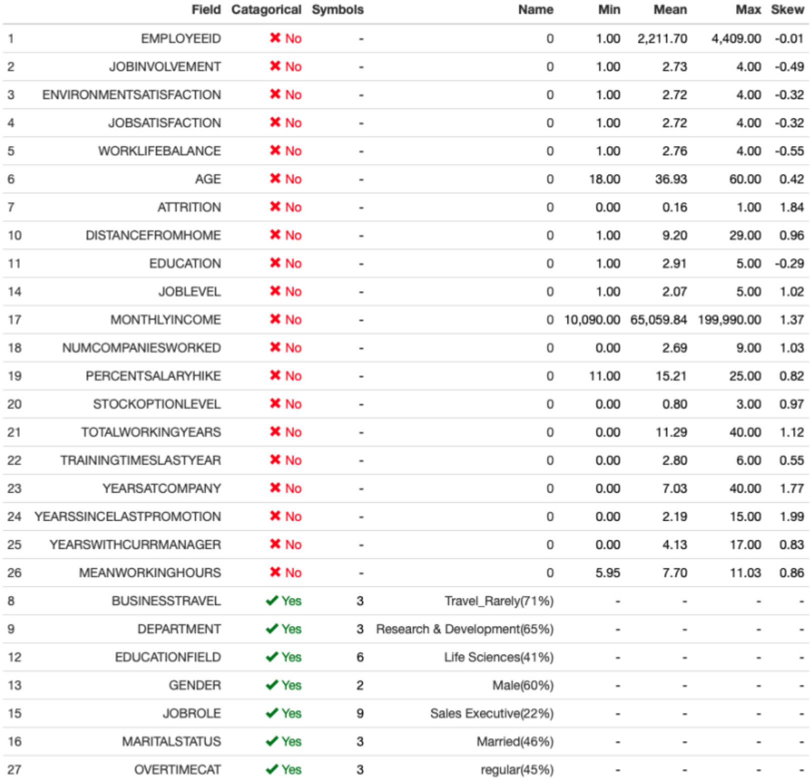
The next stage was to pre-process the data to ensure it would be suitable for modelling.

3.1 FIELD LABELLING

A high number of fields can cause issues with dimensionality, making modelling difficult. *One-hot encoding* is a method of splitting categorical data into several binary fields. Thus, the fields previously changed to be categorical to match the data dictionary were changed back to their original format. For example, '*Job Satisfaction*' was changed back from {*Low, Medium, High, Very High*} to {*1,2,3,4*}. A high number of categorical fields would have grown rapidly due to the one-hot encoding stage of the pre-processing.

As *'Attrition'* has two values: *'Yes'* and *'No'*, the field was converted to be numeric and relabeled to {*0,1*}. The minority class ('*Yes*') was assigned to *'0' and 'No'* to *'1'*, making the field binary. A binary output variable is required for several machine learning models.

As a result of these changes, there were then just 7 categorical fields. The fields were then assigned into two separate data frames based on their data type: numeric or categorical, as these types are pre-processed differently.



	Field	Catagorical	Symbols	Name	Min	Mean	Max	Skew
1	EMPLOYEEID	✗ No	-	0	1.00	2,211.70	4,409.00	-0.01
2	JOBINVOLVEMENT	✗ No	-	0	1.00	2.73	4.00	-0.49
3	ENVIRONMENTSATISFACTION	✗ No	-	0	1.00	2.72	4.00	-0.32
4	JOBSATISFACTION	✗ No	-	0	1.00	2.72	4.00	-0.32
5	WORKLIFEBALANCE	✗ No	-	0	1.00	2.76	4.00	-0.55
6	AGE	✗ No	-	0	18.00	36.93	60.00	0.42
7	ATTRITION	✗ No	-	0	0.00	0.16	1.00	1.84
10	DISTANCEFROMHOME	✗ No	-	0	1.00	9.20	29.00	0.96
11	EDUCATION	✗ No	-	0	1.00	2.91	5.00	-0.29
14	JOBLEVEL	✗ No	-	0	1.00	2.07	5.00	1.02
17	MONTHLYINCOME	✗ No	-	0	10,090.00	65,059.84	199,990.00	1.37
18	NUMCOMPANIESWORKED	✗ No	-	0	0.00	2.69	9.00	1.03
19	PERCENTSALARYHIKE	✗ No	-	0	11.00	15.21	25.00	0.82
20	STOCKOPTIONLEVEL	✗ No	-	0	0.00	0.80	3.00	0.97
21	TOTALWORKINGYEARS	✗ No	-	0	0.00	11.29	40.00	1.12
22	TRAININGTIMESLASTYEAR	✗ No	-	0	0.00	2.80	6.00	0.55
23	YEARSATCOMPANY	✗ No	-	0	0.00	7.03	40.00	1.77
24	YEARSSINCELASTPROMOTION	✗ No	-	0	0.00	2.19	15.00	1.99
25	YEARSWITHCURRMANAGER	✗ No	-	0	0.00	4.13	17.00	0.83
26	MEANWORKINGHOURS	✗ No	-	0	5.95	7.70	11.03	0.86
8	BUSINESSTRAVEL	✓ Yes	3	Travel_Rarely(71%)	-	-	-	-
9	DEPARTMENT	✓ Yes	3	Research & Development(65%)	-	-	-	-
12	EDUCATIONFIELD	✓ Yes	6	Life Sciences(41%)	-	-	-	-
13	GENDER	✓ Yes	2	Male(60%)	-	-	-	-
15	JOBROLE	✓ Yes	9	Sales Executive(22%)	-	-	-	-
16	MARITALSTATUS	✓ Yes	3	Married(46%)	-	-	-	-
27	OVERTIMECAT	✓ Yes	3	regular(45%)	-	-	-	-

Figure 6: Summary Table of the Fields

3.2 ORDINAL AND DISCRETE ORDINAL SUBTYPES

Numeric fields can be further classified into two subtypes: *Ordinal (Continuous Numeric)* and *Discrete Ordinal*. Ordinal denotes a closed interval with ordered values, whereas discrete ordinal denotes a set of unordered elements, which are termed to be categorical. An example of ordinal is the 1-4 ranking values in 'Job Satisfaction'.

One method to determine discrete ordinal fields is called *histogram equalisation* which aggregates the field into several bins. The histograms are then evaluated to determine if the field is continuous ordinal or discrete ordinal.

The numeric fields were sorted into ordinal and discrete ordinal using the histogram equalisation method. To implement this method, a function with parameter '*number of bins'* was used. The '*number of bins*' parameter defines the number of empty bins needed to determine discrete values. Using a trial and error iterative method, 7 bins was deemed appropriate as 6 bins created too many discrete fields increasing dimensionality. The histograms were plotted, providing a visual representation of the numeric fields (example shown in figure 7). The numeric fields had then been split into ordinal and discrete ordinal subtypes.

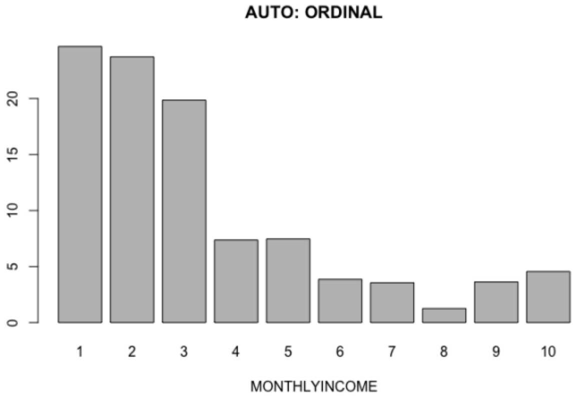


Figure 7: Ordinal and Discrete Histogram Equalisation Plot

3.3 TESTING FOR OUTLIERS IN THE ORDINAL FIELDS

Next, a subframe containing only the ordinal data fields was created to further process this data type. Data imputation was used to replace outlier values with the mean value. The outlier detection process was completed using hypothesis testing.

Hypothesis testing formally examines two opposing hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1), using statistical analysis. H_0stated that a data point was not an outlier and H_1stated that it was an outlier. Chi-squared tests at the 5% significance level were used, demonstrating an outlier was present with 95% confidence. Some outliers were found, as shown highlighted in red in figure X.

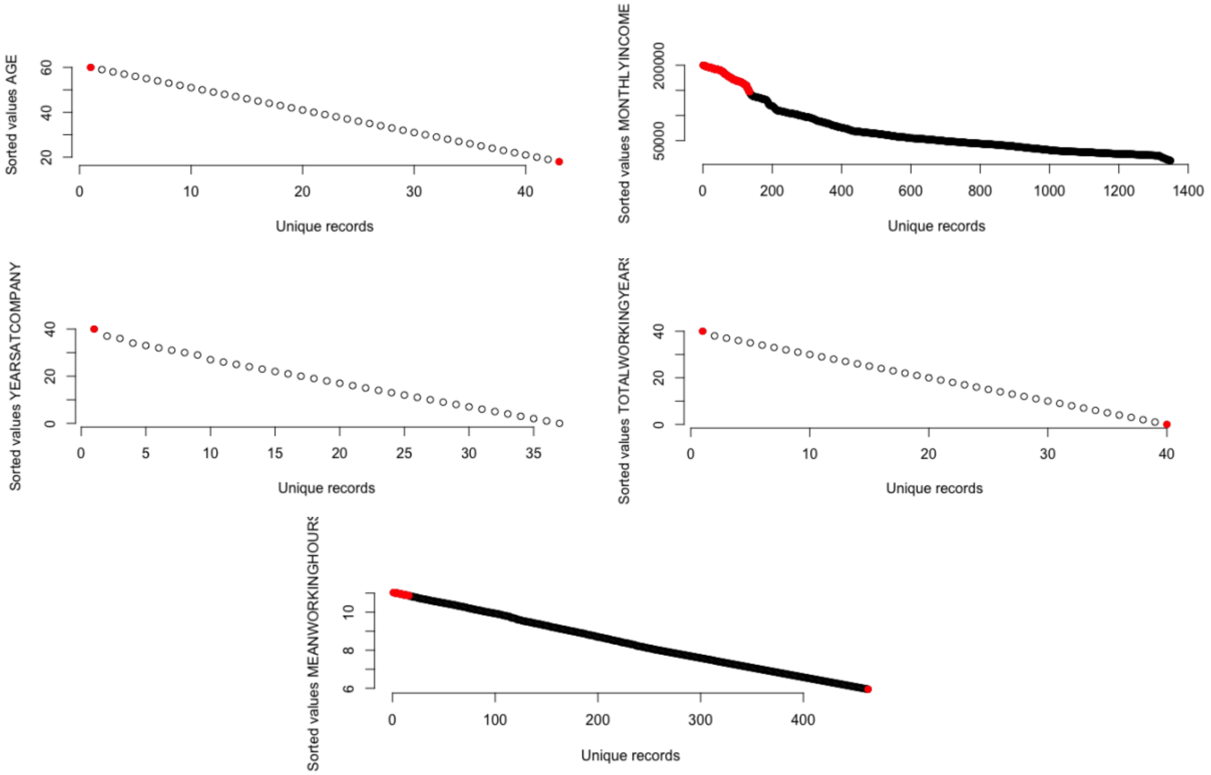


Figure 8: Examples of Outlier Detection Plots

3.4 NORMALISATION

Many machine learning models require the data to lie within the interval [0,1]. To ensure all of the data fit within this interval, the data needed to be normalised. *Normalisation* is a method that scales ordinal fields by using a standard score (called a *z-score*). The z-score identifies the number of standard deviations a value is over the mean. Normalisation of the data was achieved, likely improving the numerical stability and reducing the training times of future models.

3.5 ONE-HOT ENCODING

Next, the categorical fields were addressed. If there were two or more unique literals within the categorical fields, one-hot encoding was used, whereas if there were less than 2, a binary output was used. One-hot-encoding worked effectively as there were only a small number of categorical fields that needed to be transformed. These pre-processing steps converted 9 categorical fields to 29 binary fields. Finally, the newly transformed categorical and numeric fields were combined into a single data frame and the number of fields noted as 39.

3.6 CHECKING FOR FIELD REDUNDANCY

Fields with high multicollinearity are likely to be redundant. *Multicollinearity* is the property of fields being correlated to one another which can skew or change results to become inaccurate.

Hence, when independent variables are highly correlated with one another, these fields can be removed.

A function using linear correlation statistics with a cut-off parameter was used to remove redundant fields. A cut-off of 0.6 was chosen through trial and error as a higher cut-off did not seem to further reduce dimensionality.

The rows in the data frame were then randomised and a CSV file containing the pre-processed data was created.

4 FINDING THE MOST SIGNIFICANT INPUT VARIABLES FOR ATTRITION

The pre-processed dataset had 4300 cases and 39 variables. It was critical to reduce the dimensionality for modelling purposes so that the models can generalise more effectively. Therefore, key variables to find attrition were identified and a further subset of data created to be used for modelling.

4.1 MULTIPLE LINEAR REGRESSION

As the initial step to reduce the dimensionality, a multiple linear regression model was used. This was not planned, however, our methodology was changed due to the high dimensionality. The multiple linear regression model learns the linear relationships between multiple input variables and the output variable. This aimed to find explanatory variables that had little correlation with attrition by evaluating the model metrics.

As multiple linear regression is the simplest machine learning algorithm, it can be poor with real-world data relationships as these relationships are often non-linear. This model was only to be used as an initial pointer for our investigation.

The R^2 metric is the *residual sum of squares*, which represents the difference between the data and the model estimate. This metric lies within the interval [0,1] where ‘1’ indicates a perfect fit and ‘0’ a very poor fit. The R^2 value was 0.1357, hence the model was an extremely poor fit to the data. This is likely due to the data having a binary output variable and it being discrete.

Due to the poor fit, the significance of each variable to the model was not informative and the model was abandoned in favour of a logistic regression model. The logistic model would be more appropriate to the modelling of the binary classification problem.

4.2 MULTIPLE LOGISTIC REGRESSION

Logistic regression is a classification learning algorithm that can be used to predict a binary classification based on an input variable. The extension of this model to multiple input variables is formally called multiple logistic regression, however, will be referred to as logistic regression throughout this report. Logistic regression models perform well with non-linear relationships. As the aim is to solve a real-world binary classification problem, the logistic regression model suited the problem well.

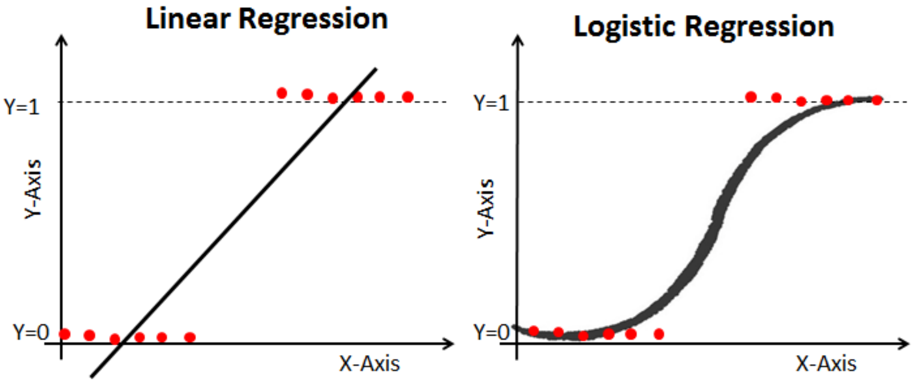


Figure 9: Linear Regression and Linear Regression Example (https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python)

The logistic regression model was trained on the entire pre-processed data to identify the key contributors to employee attrition. Initial statistics of the logistic regression model were then calculated to interpret the significance of variables in the model.

Several variables were seen to be insignificant to attrition, so were the most obvious to remove, however, a *stepwise selection* was performed before removal. Stepwise selection is a combination of forward and backward selection techniques, used to check the importance of all variables in the model.

It was then possible to remove insignificant variables consecutively. Variables needed to be removed one-by-one as the removal of a variable in the model can change the significance of the remaining variables. This was continued until only the variables significant at the 99% confidence level were included in the model. The final independent variables were then explored to see if there was any multicollinearity.

VIF (*Variance Inflation Factor*) scores explain the multicollinearity of a variable in a model. The higher the VIF score, the higher the collinearity of an independent variable to another in the model. Including a variable with a high VIF score would be an issue in modelling because it can mean that the independent variables influence each other, affecting the significance of other independent variables to the model. Variables with a VIF score above 5 would be excluded, as this is very high and would suggest multicollinearity is an issue.

Two variables had high VIF, suggesting they were related. The field with the highest VIF was removed, then VIF scores of the resulting variables were found to be below 2. This process caused several variables to become insignificant and the consecutive removal method was used until all variables were significant to the model.

The 11 variables deemed to be highly significant in determining attrition at the 99% significance level were:

- Environment Satisfaction
- Job Satisfaction
- Work-life Balance
- Training Time Last Year
- Years Since Last Promotion
- Years with Current Manager
- Frequent Business Travel
- No Business Travel
- Marital Status - Single
- Overtime
- Under Time

Another logistic regression model was then created to test if the overall model performance had decreased. Since overall performance was not significantly different, the reduced number of fields could be used going forward.

5 MODELLING AND MODEL EVALUATION

5.1 MODEL PERFORMANCE METRICS

Before creating the machine learning models, it is important to discuss the metrics used in the model analysis:

- ***True Positive (TP):***

  The number of correctly identified values of the positive class (non-attrition).

- ***False Positive (FP):***

  The number of falsely identified values of the positive class (non-attrition).

- ***True Negative (TN):***

  The number of correctly identified values of the negative class (attrition).

- ***False Negative (FN):***

  The number of falsely identified values of the negative class (attrition).

- ***False Positive Rate (FPR):***

  The number of false positives divided by the total number of negatives. Also called fall-out.

- ***True Positive Rate (TPR):***

  The number of true positives divided by the total number of positives. Also known as sensitivity.

- ***Matthews Correlation Coefficient (MCC):***

  MCC is calculated by:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(Tp + FP) \times (FN + TN) \times (TP + FN) + (FP + TN)}}$$

MCC has a value between -1 and 1. MCC = -1 shows no agreement between the prediction and the actual value, MCC = 0 is as good as a coin toss and MCC = 1 is a perfect classifier. The higher the MCC value, the better the model is at classifying attrition. It is a reasonable single metric even for unbalanced data, so can be used to compare models.

- ***Accuracy:***

  The number of correctly classified divided by the total number of classifications made. A weak approach for unbalanced data hence is not used in the model evaluation.

- ***Precision:***

  The number of true positives divided by the total sum of true positives and false positives.

5.2 PARTITIONING METHOD

For modelling, the data needed to be split into training and testing sets. The subsequent models would be learned using the train sets and evaluated using the test sets. There are various methods to split the data, which are discussed below.

The *Holdout method* randomises the order of the rows in the data and splits 70% of the data for training and 30% for testing. As it only performs one split, resulting in models may not generalise well, therefore this method was not adopted.

The *K-Fold Cross-Validation method* splits the data into k different, randomly selected groups of equal size. This uses a group to test and all other groups to train the model. This is repeated, with a different group becoming the test data in each iteration. The k test datasets never overlap records. K results are generated and the mean is calculated. Since our data is unbalanced (only 17% attrition), this method was not appropriate to adopt. Attrition may not have been well represented in each group, resulting in a poor model fit. The *Stratified K-fold Cross-Validation method* addresses this issue because the output classes are separated and recreated (so the ratio is maintained each fold). Hence, the Stratified K-fold Cross-Validation method was the optimal way to partition the data. This method reduces overfitting and can be performed for all models. 5 folds were used for all models - appropriate for the dimensions of the data frame. Had there been more records, more folds would have been necessary.

5.3 MULTIPLE LOGISTIC REGRESSION MODEL

A logistic regression model was then used to predict employee attrition, directly addressing the business objective. This model assumed that the relationship between the dependent and independent variables may not be linear.

The model correctly predicted 514 employees left and 101 stayed, however, incorrectly predicted 37 employees left that stayed and 206 stayed that left. This is summarised in a confusion matrix. FPR and TPR were 27% and 71% respectively.

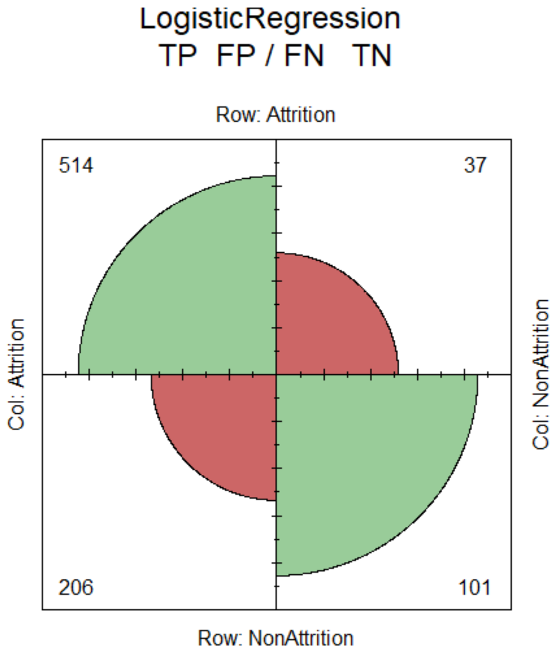


Figure 10: Confusion Matrix for Multiple Logistic Regression Model

A *Receiver Operator Characteristic* (ROC) chart, calculated by plotting the *sensitivity* (TPR) against the *specificity* (1-FPR), can be used to assess classification model performance. The higher the *Area Under the ROC Curve* (AUC), the better the model. AUC ranges from 0.0 to 1.0, with values over 0.5 denoting a good classification model.

The ROC chart for the multiple logistic regression model was plotted for each fold, with a mean AUC of 0.77, so the model can be described as a good classifier.

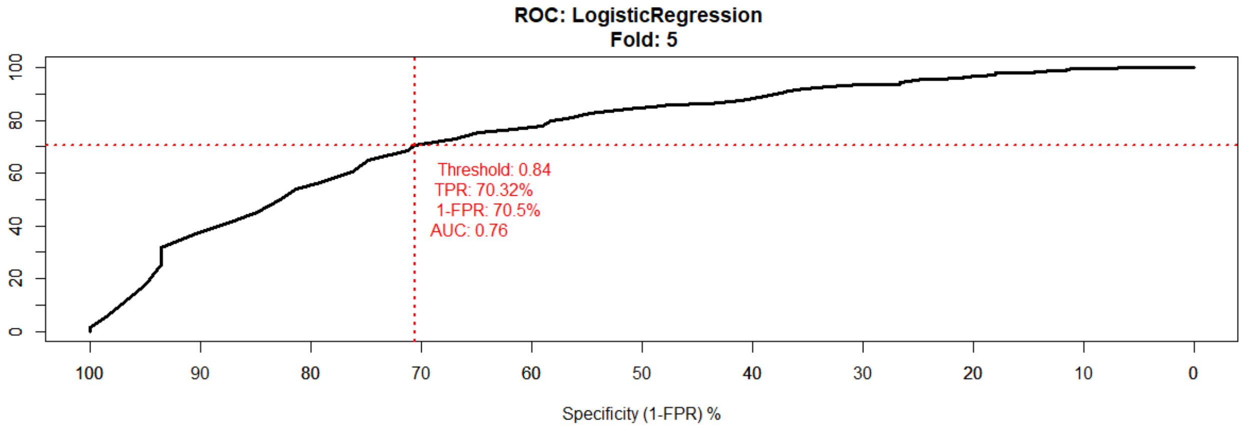


Figure 11: ROC chart for Multiple Logistic Regression Model for Fold 5

MCC was 0.34, suggesting a good fit. To examine the significance of each variable in prediction, the importance of each was calculated and plotted (figure X). According to the multiple logistic regression model, marital status (single), years with current manager and working overtime were the top 3 factors in predicting attrition.

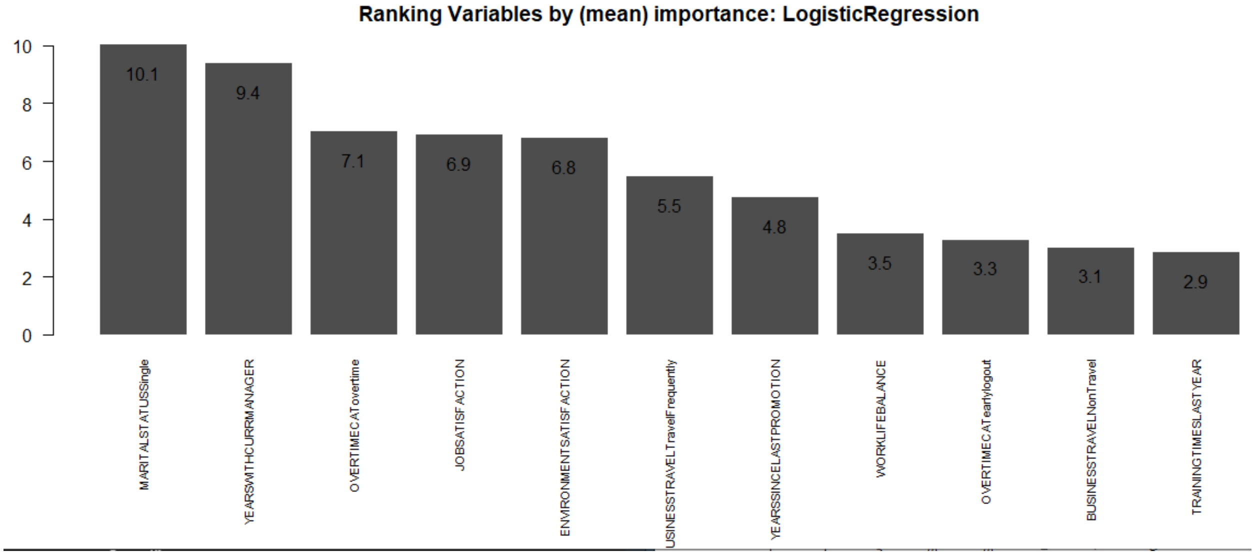


Figure 12: Ranked Variables by Importance in the Multiple Logistic Regression Model

5.4 RANDOM FOREST MODEL

A *random forest* is an ensemble method consisting of multiple boosted decision trees. A simple decision tree contains nodes representing decisions, with each decision splitting into branches representing possible outcomes and splits occurring until a further split is not statistically significant to the overall model.

The *boosting decision tree method* trains multiple simple decision trees sequentially, with these trees then voting for a class and the majority winning. Records that are misclassified are used to train the next tree until a maximum number of trees is reached. The *random forest method* selects a random subset of input variables with trees voting as in the boosting method, however, the individual trees are then joined to create a forest.

Random forests work well with large datasets and are less likely to overfit the model than a decision tree. Random forests can still overfit if there is too much noise (unmeaningful data), so it was assumed that the removal of outliers and insignificant variables in pre-processing would avoid overfitting.

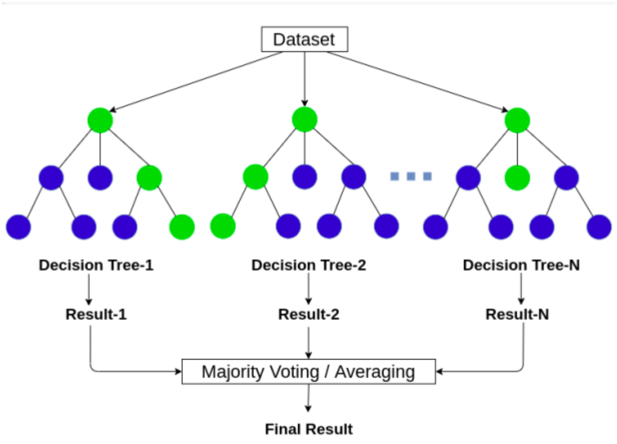


Figure 13: Visual Representation of a Random Forest https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/)

Initially, 1000 trees were chosen for the random forest; however, after running the model with 100 trees, the model was considerably faster with little change to the output. A larger forest size increases the likelihood of overfitting, so a random forest size of 100 trees was selected.

The model correctly predicted 683 employees left and 130 stayed, however, incorrectly predicted 8 employees left that stayed and 37 stayed that left. FPR and TPR were 5.8% and 98% respectively.

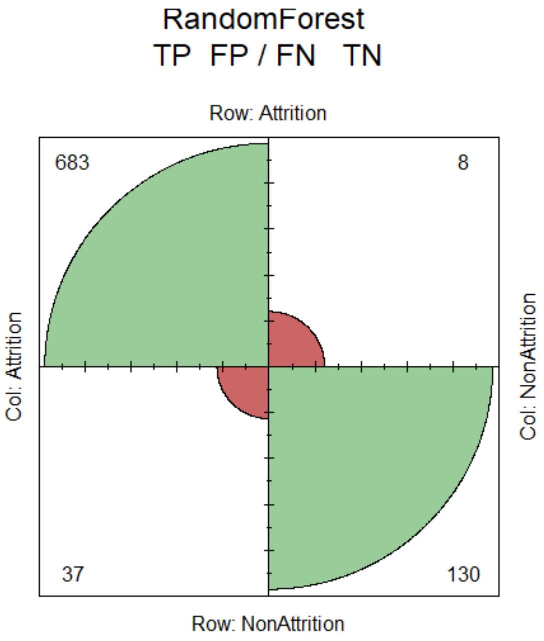


Figure 14: Confusion Matrix for Random Forest Model

For each fold, the ROC chart was plotted and the AUC calculated. The fifth fold had AUC of 0.99, suggesting a near-perfect fit but that the model could be overfitting the data.

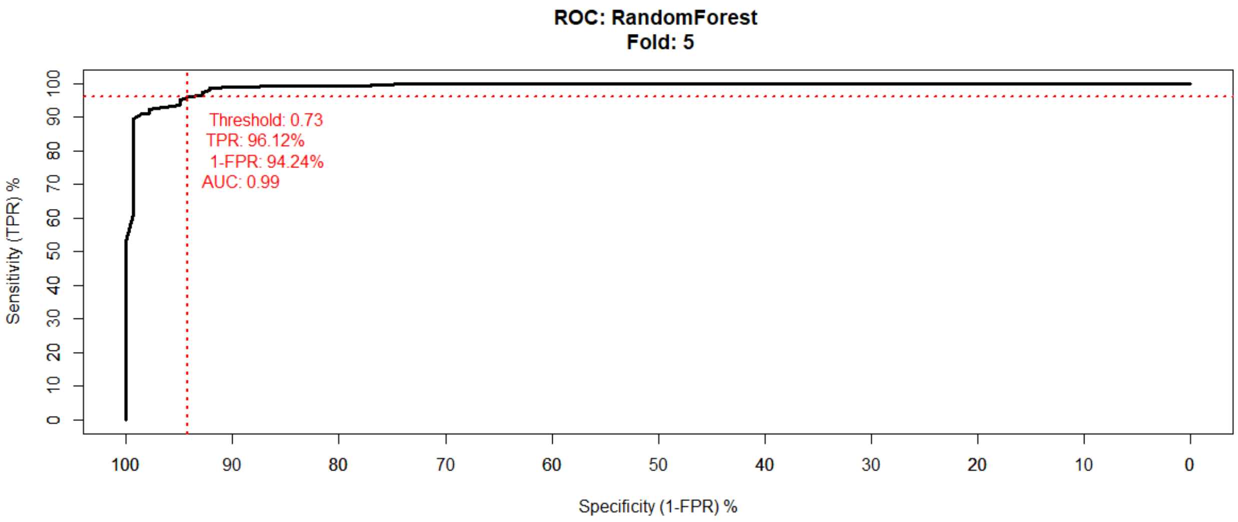


Figure 15: ROC Chart for Random Forest Model

The variables were then ranked by importance to the model. According to the random forest model, the years spent with the same manager, environment satisfaction, and the number of years since the employees last promotion were the three variables with the top importance in determining attrition.

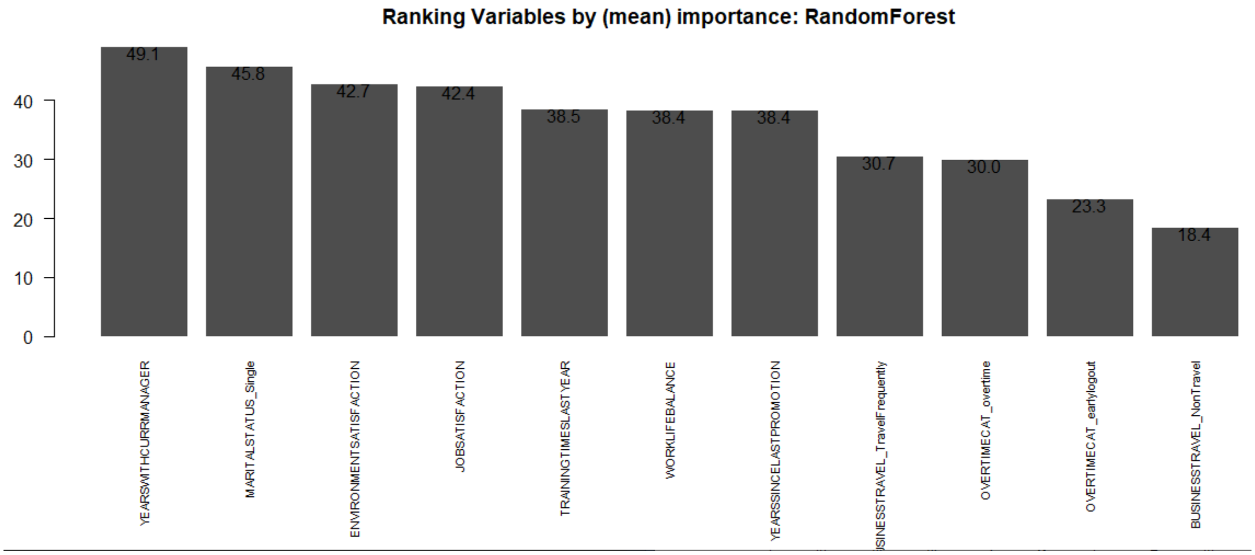


Figure 16: Ranked Variables by Importance in the Random Forest Model

Finally, the MCC was calculated to be 0.82, suggesting that the model is an excellent classifier of the current data and may be an excellent predictor of employee attrition.

5.5 NEURAL NETWORK MODEL

The workings of a neural network are largely based on the biological nervous system, learning to recognise patterns within a dataset, similarly to how the human brain learns from experiences.

They use feedback loops to find patterns in a dataset and predict future outcomes, abstracting information well from data that is noisy or with outliers.

The simplest form of a neural network is called a simple perceptron model, which takes weighted inputs to create a 'neuron' and classifies the output based on a threshold. Whereas, a more complex model, a shallow neural network (SNN), can be created by adding a further layer of neurons, called a hidden layer. A hidden layer makes a neural network hard to explain since weights and biases are being added in the hidden layer to the variables. Then, the activation function determines whether the weighted sum is above a threshold value.

Due to the binary classification supervised learning problem, the neural network would have a single output value for each employee. For the model to work, the data needs to be scaled and normalised, which was addressed in pre-processing. Neural networks can handle categorical datasets and can be highly efficient once they are trained.

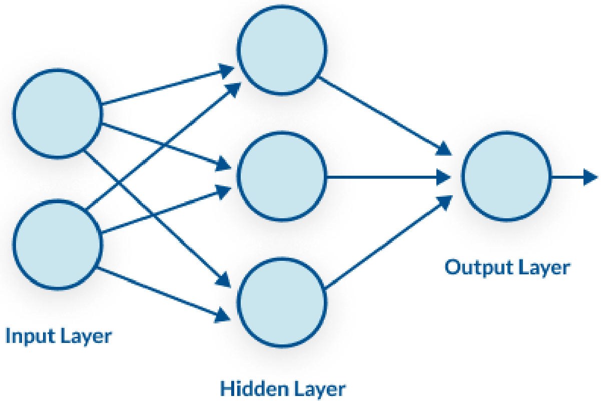


Figure 17: Example of a Shallow Neural Network (https://missinglink.ai/guides/neural-network-concepts/perceptrons-and-multi-layer-perceptrons-the-artificial-neuron-at-the-core-of-deep-learning)

After a period of trial and error, iterating the parameter settings, the neural network that performed best had a single hidden layer containing 9 neurons. This iteration was in-line with the CRISP-DM methodology and improved the model's performance metrics. The logistic (sigmoid) activation function was chosen as it works best for binary classification problems.

The model correctly predicted 627 employees left and 82 stayed but incorrectly predicted 94 employees left that stayed and 56 stayed that left. FPR and TPR were 40% and 87% respectively.

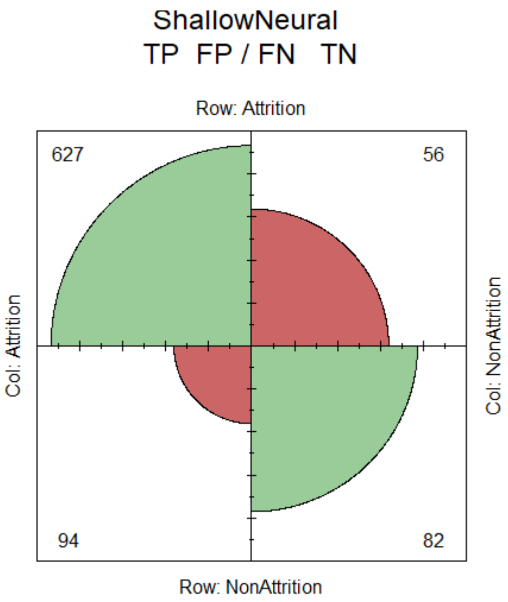


Figure 18: Confusion Matrix for Deep Neural Network

An AUC of 0.76 and an MCC of 0.49 was observed, demonstrating that the model was a good classifier, however, not nearly as good as the random forest model trained previously.

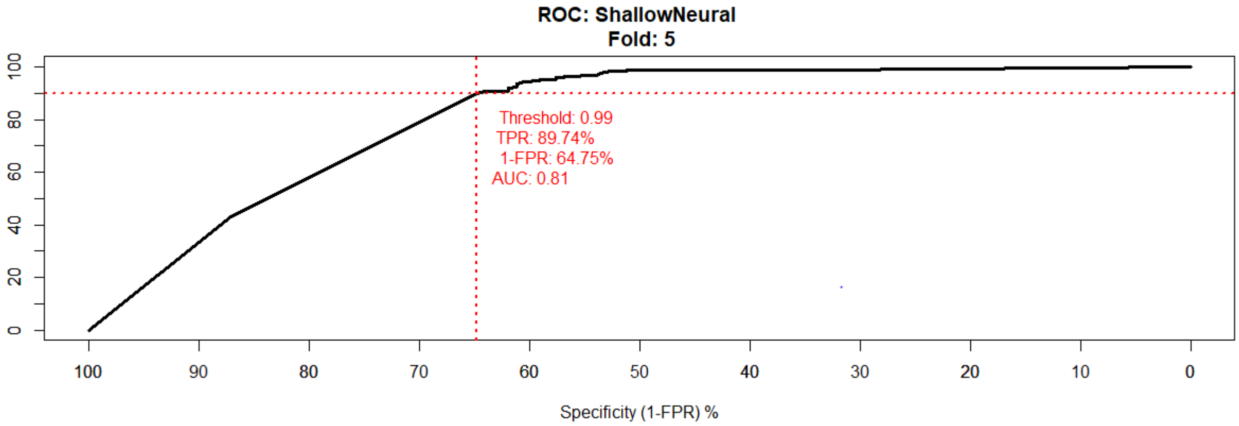


Figure 19: ROC Chart for Shallow Neural Network

Due to the black-box nature of the ‘neuralnet’ package used to create this model in R, the importance of each variable to the SNN would be difficult to calculate - a significant drawback to the business objective of identifying the key factors for attrition. Hence, the SNN could be used as a means to predict if an employee would stay/leave the company, however not to identify individual variable significance.

5.6 EXTENDING THE SHALLOW NEURAL NETWORK

A deeper neural network extends the SNN by adding further hidden layers. Due to the extensive running time of the ‘neuralnet’ package, a different package (‘h2o’) was used to create this model. The aim of this was to improve upon the SNN with one hidden layer.

Using an iterative trial and error method to find the optimal combination of neurons and hidden layers, a model with 2 hidden layers (4 neurons in the first layer and 6 neurons in the second layer) performed the best. Neuron numbers in models with 2 and 3 hidden layers were tested, however adding a fourth hidden layer was not explored as the number of combinations of neurons to test increases exponentially with each layer added. Three hidden layers were found to not be as successful as a model with two hidden layers.

Typically, deep neural networks have 3 or more hidden layers, with earlier layers building *feature abstraction* (relationships between features) and later layers combining these to make a classification decision. Using an iterative method to find the optimal combination of neurons and hidden layers, a model with 2 hidden layers (4 neurons in the first layer and 6 neurons in the second layer) performed the best. Neuron numbers in models with 2 and 3 hidden layers were tested and 2 hidden layers proved to be more successful. Therefore, a deeper shallow neural network proved better than a formally defined deep neural network.

The model correctly predicted 503 employees left and 96 stayed, however, incorrectly predicted 42 employees left that stayed and 217 stayed that left. FPR and TPR were 31% and 70% respectively.

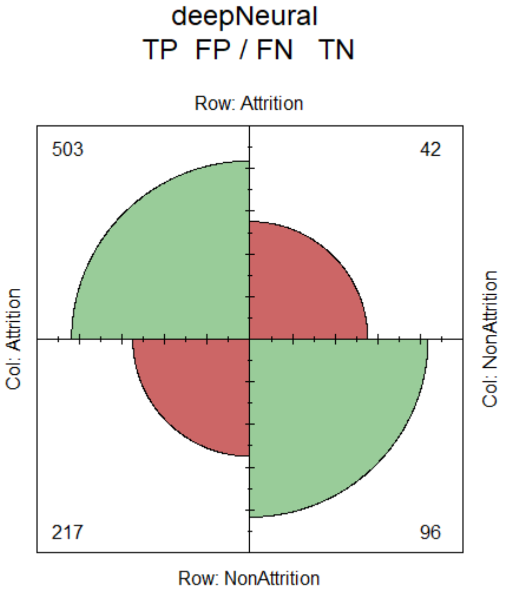


Figure 20: Confusion Matrix for Deep Neural Network

The ROC chart was plotted for each fold, revealing an average AUC of 0.74. The MCC value was 0.31, meaning the deeper neural network did not perform better than the shallow neural network across any metrics.

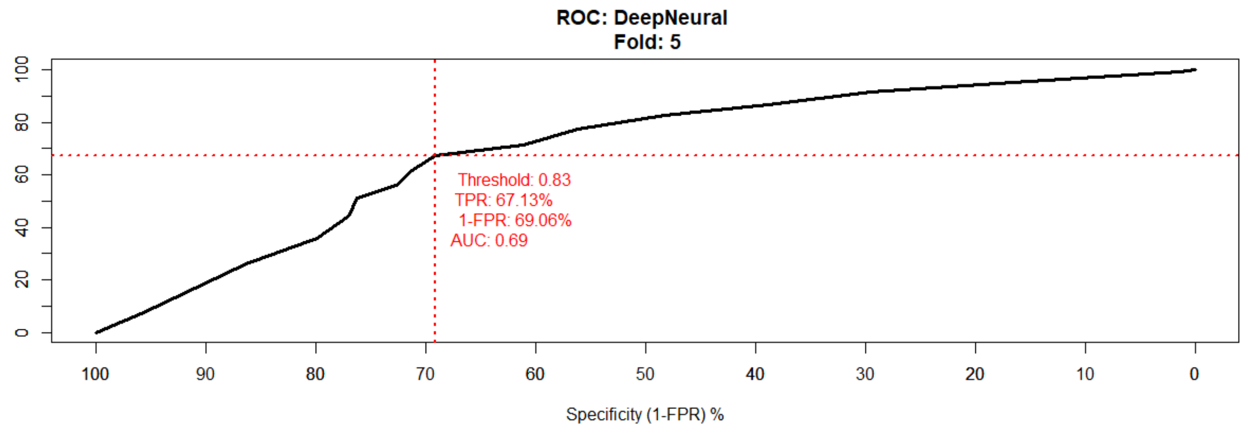


Figure 21: ROC Chart for Deep Neural Network

Using a property of the h2o package, each variable's importance to attrition was then ranked. This highlighted that years with the current manager and marital status being single were by far the most important factors like the logistic regression and random forest model suggested.

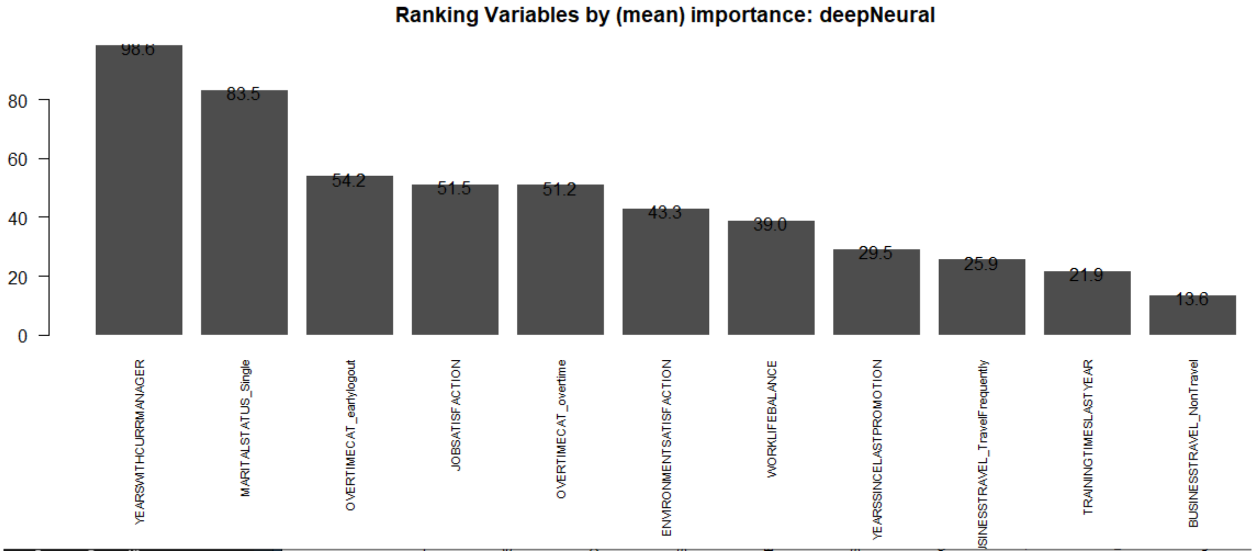
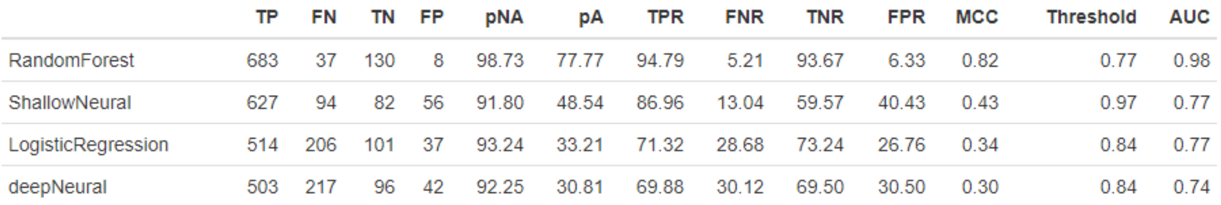


Figure 22: Ranked Importance for Deeper Neural Network

5.7 MODELLING SUMMARY



	TP	FN	TN	FP	pNA	pA	TPR	FNR	TNR	FPR	MCC	Threshold	AUC
RandomForest	683	37	130	8	98.73	77.77	94.79	5.21	93.67	6.33	0.82	0.77	0.98
ShallowNeural	627	94	82	56	91.80	48.54	86.96	13.04	59.57	40.43	0.43	0.97	0.77
LogisticRegression	514	206	101	37	93.24	33.21	71.32	28.68	73.24	26.76	0.34	0.84	0.77
deepNeural	503	217	96	42	92.25	30.81	69.88	30.12	69.50	30.50	0.30	0.84	0.74

Figure 23: Table of Calculated Metrics for All Models

Despite logistic regression correctly predicting more attrition cases than the shallow neural network, the shallow neural network had significantly less false negatives. However, the random forest model had the highest MCC value, lowest FPR and highest TPR of all of the models explored, so the model could be described as the best predictor of attrition. This model could also directly link to the business objectives: identifying the key variables for attrition and predicting employee attrition successfully.

6 EVALUATION

After selecting the best classification model for the data, the business objectives and success criteria were re-evaluated to determine the success of the business analytics project.

6.1 KEY FACTORS FOR ATTRITION

The first business objective was to understand the key factors for attrition. Using significance testing and VIF to build a multiple logistic regression model and reduce the dimensionality to just 11 input variables (ranked by importance by each of the models). The most effective model, the random forest, saw that the most important variables in order were:

1. Years with Current Manager
2. Environment Satisfaction
3. Years Since Last Promotion
4. Job Satisfaction
5. Marital Status - Single
6. Work-Life Balance
7. Training Time in the Last Year
8. Frequent Business Travel
9. Working (on average) Overtime
10. Working (on average) Under the Standard Hours
11. No Business Travel

This suggests that the hypothesis that employees working under or overtime was not a key factor in attrition, whereas, years with current manager, environment satisfaction and the years since last promotion were. Business travel was also not a major factor for attrition.

6.2 PREDICTING ATTRITION

The second business objective was to build a model to be able to successfully predict employee attrition. The random forest model was found to be an effective classification model and is therefore an effective predictor of attrition. This model had FPR and FNR rates of just 6% and 5% respectively and was considered an excellent classifier by MCC and AUC. Hence, a successful machine learning algorithm for attrition prediction has been achieved.

6.3 REDUCING ATTRITION

The final business objective was to be able to use a machine learning model to define areas that a business can improve/change to reduce employee attrition rate.

The random forest model was the best predictor of attrition; the top 5 factors for attrition were: years with current manager, years since last promotion, environment satisfaction, job satisfaction and marital status - single.

Firstly, job satisfaction is key. It is reasonable to assume that if an employee enjoys their work, they are less likely to leave. To improve employee job satisfaction, job roles should be carefully defined when recruiting so that only those interested in the work apply for the role. Another way to improve job satisfaction could be to reward and celebrate employees' working achievements.

Another significant factor is environment satisfaction, which concerns satisfaction with the workplace. Again, it is reasonable to presume that a higher environmental satisfaction means an employee is less likely to leave. Workplaces differ greatly by industry, however, a good way to understand what can be done to create a pleasant working environment is an environment satisfaction survey featuring a suggestions section can be given to employees. Through such a survey, more granular detail into improvements or issues with the workplace can be explored. For example, employees may wish for more plants in the office or for an improved kitchen area.

Unfortunately, despite the importance of the fields being known, it is not known if marital status of single has an adverse or a negative correlation to attrition. For example, it is not known if those who are single or those who are married are more likely to leave.

Similarly, the years spent with the current manager and number of years since last promotion are known to be important, however the relationship between these fields and attrition is unknown. This was a major difficulty found after evaluating the models created to predict employee attrition and meant that the final business objective could not be fully answered.

6.4 KEY AREAS FOR IMPROVEMENT

If the investigation into attrition was to be re-run, it would be highly beneficial to further break down some fields into more meaningful detail. For example, several fields could have been broken down into a more useful 'low', 'medium' or 'high', such as the years with current manager or environment satisfaction fields. Alternatively, the number of empty bins needed to determine ordinal or discrete ordinal numeric fields in the pre-processing could have been reduced. If this idea were to be implemented, the final business objective of creating valuable advice for a business regarding reducing future employee attrition could have been answered in more granular detail.

One of the key areas of improvement for the project relates to the last business objective which could not be fully addressed by the project. If the project were to be reran, more detail about the relationships between the significant variables and attrition should be known. Without this, mere assumptions are needed to accurately provide plausible recommendations to a business.

7 ACKNOWLEDGEMENTS

We would like to thank Professor Nick Ryman-Tubb for his continuous support throughout this project. Professor Ryman-Tubb allowed us to use some of his prewritten functions in R to aid the delivery of our results. These functions have been prefixed with a ‘N’ in the R code.

8 APPENDIX

8.1 PROJECT PLAN

PROBLEM DEFINITION

The Human Resources (HR) team within an organisation is the team responsible for recruitment, payroll, employee policies and benefits [1].

As the department handling recruitment, the HR team must first identify the need for a new employee and develop a description of the position. Then, the recruitment process has to be planned and the role advertised. As applications are sourced, they have to be assessed and numerous potential candidates have to be interviewed. Once an applicant is selected for the role, the work continues as an offer of employment must be drafted, all before an employee can begin working [2].

The recruitment process can be both costly and time consuming for an organisation. The research by Robert Half UK calculates that HR Directors in the UK spend close to 28 days on average to recruit within their organisation [3]. The cost of recruitment involves not only the time taken by the HR department, but also the costs of advertising the role, internal and external recruiter expenses, and training the employees when they arrive.

However, how often do employees really leave an organisation? How often does this recruitment process occur? One key avenue to understanding the rate of recruitment is to examine how often people change jobs.

On average, close to one-in-ten people change jobs each year [4] with another survey adding that more than half of UK workers are planning on changing careers in the next 5 years [5]. According to research by Investec, millennials are likely to have 12 different jobs during their working lives [6] and in the US, a LinkedIn study states, young workers are likely to change jobs four times in their first 10 years after graduating [5]. Thus, the prospect of having a job or career for life is increasingly slim, and the likelihood of employees leaving the organisation or changing jobs is on the rise.

A key way of reducing the rate of recruitment within an organisation is to reduce the rate of employee attrition. For this project, ‘employee attrition’ is defined as the number of employees that have left in the previous year. For example, if an employee has recently left the company, then attrition has occurred.

Employee attrition can be costly for an organisation, with research by Oxford Economics on the financial impact of losing staff estimates that the loss of an employee carries an average financial impact of just over £30,000 [7]. Also, in the nine months before an employee leaves their job, their overall engagement at work plummets [8] severely impacting productivity.

Furthermore, employee attrition may mean a loss of talent in the organisation. The progress and growth of an organisation depend heavily on the skills and strength of the workforce; therefore, talent retention and hiring the right employees are factors directly linked to company performance. For example, 83% of HR leaders agree that talent is the top priority at their company, showing that retaining talent is ever-more pertinent [9].

The reasons for an employee leaving vary, with a study suggesting that a better work-life balance, career advancement, higher remuneration, better location and better corporate culture are among the main reasons [10].

PROJECT OBJECTIVE

Our core business objective is to better understand the key factors that affect employee attrition in an organisation, using a data-driven approach. Once the significance of these factors is understood, the aim is to be able to both predict and reduce future employee attrition rates, with the key business objective of reducing the costs associated with attrition.

BRIEF INTRODUCTION TO THE DATASET

To conduct our analysis, the project will make use of a dataset sourced from Kaggle. The dataset that we have chosen is the ‘HR Analytics Case Study’ dataset (https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study).

The dataset includes 5 csv Excel spreadsheet files, as well as the data dictionary (metafile) and totals approximately 45Mb in size. These files include:

- **employee survey data** (*employee_survey_data.csv)* – responses from employees on environment satisfaction, job satisfaction and work/life balance
- **general data** (*general_data.csv)* – detailing employee demographics
- **in-time data** *(in_time.csv)* – employee login time on specific dates
- **out-time data** *(out_time.csv)* – employee logout time on specific dates
- **manager survey data** *(manager_survey_data.csv)* – responses from managers on job involvement and performance rating

The data will be analysed and interpreted to answer the business objective using the following iterative and data-driven business analytics methodology.

BUSINESS ANALYTICS TASKS IDENTIFIED

Throughout the project, we will follow the CRISP-DM Model approach [11]. The initial steps in the CRISP-DM Model regarding business and data understanding have been discussed in the Problem Definition. In this section, we will discuss Data Preparation and Modelling methodology.

Firstly, the data will need to be prepared. To do this, we are going to:

- Merge all non-time datasets via the unique identifier "Employee ID"
- Merge the "in_time" and "out_time" datasets, whilst keeping in mind that column names are not unique and will need to be edited to display "in" and "out" labels
- Clean the data using sample analysis to determine if there are any cases that should not be included in the investigation, such as missing cases or incorrect cases as a result of sampling error
- Conduct descriptive analysis to determine the usability of each variable. This may include looking at the distribution of data to check for anomalies or determine normality in the data (maybe look into transformations of the data if necessary) and looking at basic descriptive statistics such as Mean/Median, Standard Deviation, Skewness, etc...
- Using the k-fold method, we are able to have a testing set and a training set to evaluate how well our data does at predicting attrition. We also need to make sure if we are looking at each individual attribute that we take out any that are not statistically significant

PROPOSED MACHINE LEARNING MODELS

To identify the model which best predicts the key variables that contribute to an employee leaving the company, we will use a variety of machine learning methods. The aim is to find the model that most accurately predicts employee attrition.

Using the Confusion Matrix, we will test the accuracy of each model by finding the TPR (True Positive Rate) and TNR (True Negative Rate) as metrics to evaluate each method. These metrics will inform us as to how well our model performs at correctly predicting if an employee will leave the organisation, given the attributes in the dataset.

The proposed machine learning models that the project will utilise are:

- **Classification (logistic regression)**
  - The use of logistic regression to predict the probability of an employee leaving, based on the variables provided in data.
- **Random forest**
  - The use of the random forest method to predict which variables given in the dataset are the largest contributors to attrition, performed by analysing the statistical significance of each on attrition.
- **Neural network**
  - The use of a neural network to find any significant patterns in employee attrition rates to be able to predict if a given employee will leave the company, given a set of variables.
  - Covariates should have the same scale to create a more stable neural network, and outliers should be removed.

EXPECTATION

A number of different sources highlight that remuneration is often not the main factor of attrition. M. Russel suggests in his article that although 89% of bosses believe that employees quit because they are unsatisfied with their compensation, only 12% of employees leave because of this reason [12].

From our dataset, we expect to find a number of correlations related to attrition. Our expectation is that the most significant factors are low job involvement and low work environment satisfaction, which can result in low performance rating and ultimately attrition. We also appreciate that combinations of numerous other factors will help us predict whether an employee will leave the company. Our model will help reduce attrition by focusing on improving the most significant factors.

We also aim to explore how the number of hours an employee is present at work impacts the likelihood of attrition. The data required will be sourced from the in-time and out-time datasets. We can compare this to the standard hours the employees are expected to work. We can also explore how the dynamics of their presence at work change towards their decision to leave. We expect that people who complete less hours than expected of them are likely to have less job involvement and satisfaction, thus be more likely to leave the company. Furthermore, we expect that there is a correlation between the change in the number of hours worked and attrition.

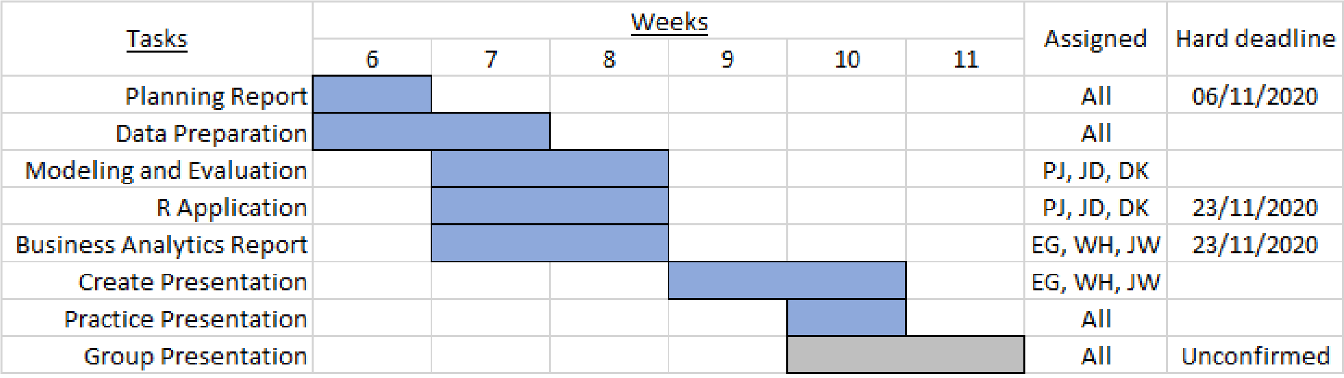
PROJECT TIMESCALE

Throughout the project, there will be bi-weekly meetings with the whole group. These bi-weekly meetings will discuss any updates, challenges and potential issues at each stage of the project - ensuring that the project remains on track and meets the listed deadlines, as well as ensuring that the whole group is updated on the status of each deliverable.

To ensure that the project runs smoothly and falls within the allocated time intervals, we shall be following a project plan, outlined below.

The week timetable outlined aligns with the module ‘weeks’ with week 1 commencing on Monday 28th September. For ease of readability, initials are used to denote each group member.

PROJECT GANTT CHART



Tasks	Weeks						Assigned	Hard deadline
	6	7	8	9	10	11		
Planning Report	🟦						All	06/11/2020
Data Preparation	🟦	🟦					All	
Modeling and Evaluation		🟦	🟦				PJ, JD, DK	
R Application		🟦	🟦				PJ, JD, DK	23/11/2020
Business Analytics Report		🟦	🟦				EG, WH, JW	23/11/2020
Create Presentation				🟦	🟦		EG, WH, JW	
Practice Presentation					🟦		All	
Group Presentation					▨	▨	All	Unconfirmed

8.2 DATA DICTIONARY

The data dictionary that was used was taken directly from the datasets found on Kaggle.

Variable	Meaning	Level
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees travelled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kms	
Education	Education Level	1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'

Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
JobInvolvement	Job Involvement Level	
JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	
JobSatisfaction	Job Satisfaction Level	1 'Low' 2 'Medium' 3 'High' 4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	
Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock Option Level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad' 2 'Good' 3 'Better' 4 'Best'

YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrentManager	Number of years under current manager	

8.3 REFERENCES

[1] Unum website, What does Human Resources do?, 21 April 2020
https://www.unum.co.uk/resources/what-does-human-resources-do

[2] Iowa State University website, Recruitment and Selection Process, Accessed 5 November 2020
https://www.hr.iastate.edu/employing-units/recruitment-selection

[3] Robert Half, HR Directors Spend 28 Days on Recruitment Process, 2 November 2016
https://www.roberthalf.co.uk/press/hr-directors-spend-28-days-recruitment-process

[4] Office for National Statistics (ONS), Analysis of Job Changers and Stayers, 29 April 2019
https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/economicreview/april2019/analysisofjobchangersandstayers#job-changers-and-stayers

[5] The Financial Times, Plan for five careers in a lifetime, Helen Barrett, 5 September 2017
https://www.ft.com/content/0151d2fe-868a-11e7-8bb1-5ba57d47eff7

[6] Talint International, Millennials likely to have 12 jobs in their working lives, 20 November 2017
https://www.recruitment-international.co.uk/blog/2017/11/millennials-likely-to-have-12-jobs-in-their-working-lives-research-finds)

[7] Oxford Economics, The Cost of Brain Drain (Understanding the financial impact of staff turnover) Report, February 2014

[8] The Financial Times / Peakon, Communication Breakdown: Why Your Employees Really Quit, Accessed 4 November 2020
https://www.ft.com/brandsuite/peakon/communication-breakdown-why-your-employees-really-quit.html

[9] LinkedIn Global Recruiting Trends Report 2017
https://www.slideshare.net/pedrooolito/linkedin-global-recruiting-trends-report-2017

[10] The Financial Times, In A Blink: the reasons employees leave, Carola Hoyos, 3 September 2015
https://www.ft.com/content/721356ae-5106-11e5-8642-453585f2cfcd

[11] The CRISP-DM Process Model, Chapman, P, Kerber R, et al., 1999.

[12] Russell, M., 2018. Why Employees Quit: 20 Stats Employers Need To Know. [online] Medium.
https://medium.com/@checkli/why-employees-quit-20-stats-employers-need-to-know-b921c253f767
[Accessed 5 November 2020].

Title Page Image: https://greatpeopleinside.com/employees-leaving-their-jobs/