MGT 6203 Group Project Progress Report (Team 50)

Analysis of Factors that Affect Employee Attrition Rates

TEAM MEMBERS & GITHUB

1. **Van Dang**; vdang35; Middle School Mathematics Teacher; Bachelor of Science in Nuclear Engineering
2. **Xiao Hui, Heng**; xheng3; Digital Product Manager; Bachelor of Business Management
3. **JunKai Benson, Ng** (jng87); Manager, Ministry of Health Singapore; B.A. in International Business, Finance & Economics, University of Manchester
4. **Sushil More** (smore37); Product Manager; Bachelor of Science, Information Systems
5. Github: readme (contains links to codes and folder where final report is saved)

INTRODUCTION & PROJECT OBJECTIVE

Background Information on chosen project topic:

Employee attrition creates a host of issues, ranging from hidden costs such as employee burnout and lost industry knowledge, to more quantifiable ones like lost productivity and recruitment costs. Concerning hard costs alone, some studies have shown that replacing a single employee can incur a hard cost of three to four times that of the position's salary (Navarra, 2023) while other estimates place this cost even higher (Santovec, 2010).

It's in most companies' interest to increase employee retention and therefore reduce both the hidden and hard costs of rehiring. In order to do this, employers must be able to identify and understand the factors that might affect employee attrition including, and not limited to: commute times, frequency of business trips, salary, percentage of salary increases, work-life balance, and job satisfaction.

Once the factors affecting employee attrition are identified, employers can better understand the extent of impact of these factors so that companies may enact employee retention strategies or change their recruitment criteria to filter for employees that are more likely to remain with the company for extended periods of time.

Problem Statement:

This analysis seeks to identify the key factors and the extent of these factors that lead to employee attrition so that companies can retain employees for longer to reduce personnel costs.

Research Questions:

1. Understand the key underlying factors resulting in employee attrition
2. Classify and group factors that are related to employee retention
3. Optimize relevant predictors to minimize employee attrition

Business Justification:
As mentioned previously, hiring replacement employees to compensate for attrition can be quite costly. These costs (hidden and hard costs) can be broken down into additional categories:

1. **Employee Burnout**
    - Company operations must continue with reduced staff and the consequential understaffing adds additional stress to the remaining employees who have increased workload and responsibility.

   - Employee burnout and attrition usually has an adverse impact on company morale, productivity, and motivation that is difficult to quantify (Zavgorodnii et al., 2020).
2. **Lost industry knowledge**
   - Company business practices and undocumented best business practices can be lost with the leaving employee.
3. **High cost of retraining new hires**
   - Loss in productivity during the gap of a leaving employee and retraining of new hire
   - New hires take time to fully integrate into their new role, prompting further loss of productivity
   - The costs for HR and team managers to find a suitable replacement

New hires always have a risk of a mishire which incur fees much higher than lost productivity (Erling, 2011).

DATASET

This is a fictional dataset (click here to assess the dataset) created by IBM data scientists, and contains information about employee attrition. The data includes 35 columns, including Age, DistanceFromHome, Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance, etc. For more details on the dataset, refer to Appendix A.

Dependent Variables: Attrition (Yes/ No)

Independent Variables: As the dataset contains many different variables, the team carried out exploratory data analysis ("EDA") to pick out the most relevant variables, and categorised them into groups to address the identified research questions. This will be further elaborated upon in Exploratory Data Analysis.

METHODOLOGY

Our planned approach can be broken down into four main parts, and more details are found in the subsequent sections:

1. **Data Cleaning & Transformation**
2. **Exploratory Data Analysis**
3. **Model Training & Optimization**
4. **Results Validation**

DATA CLEANING & TRANSFORMATION

The dataset is relatively clean, hence there are no missing or duplicate values to address. However, 3 key actions were taken:

No.	Step	Impact
1	Remove columns with only 1 unique value	Removal of: "EmployeeCount", "Over18", "StandardHours"
2	Convert categorical response variable into binary form to enable modelling	In the "Attribution" column, "Yes" values is transformed into "1", and "No" values into "0"
3	Convert categorical independent variables into: • Binary form • Dummy variables • Grouping into relevant range	Conversion of variables such as "BusinessTravel", Gender", "JobRole" into binary or dummary variables Grouping of "Age" into relevant age groups

EXPLORATORY DATA ANALYSIS

The key objectives of this step include:

1. Identification of high level patterns, trends and correlations
2. Selection of variables for modelling

To achieve the above, the team explored visualizations that include correlation matrix and charts. The following discussion shows the key observations and steps taken in selecting variables for modelling. Other variable reduction not discussed below are achieved during model validation and testing.

Correlation Matrix

As a first step, in the correlation matrix in Figure 1, several variables with high correlation are observed, as noted by the pairs in dark blue. If highly correlated variables are included into the models, it could introduce **multicollinearity,** which can destabilize the estimation of coefficients in regression models, and make it difficult to determine the effect of individual predictors on the response variable.

Some of the pairings showing high correlation include the bottom right hand corner, which are related to the number of years worked, including factors such as “TotalWorkingYears”, “YearsAtCompany”, “YearsInCurrentRole”, “YearsSinceLastPromotion”, “YearsWithCurrentManager”, “JobLevel”, and “MonthlyIncome”.

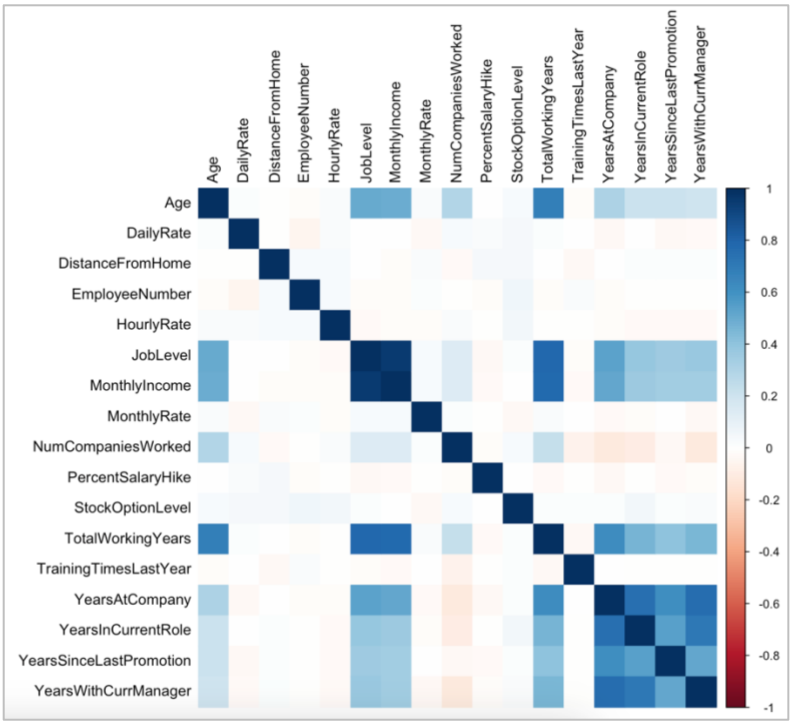


Figure 1: Correlation Matrix

Based on personal experience, the high correlation in these pairings makes sense, given that higher working experience results in higher ranks and income, as well as the number of years in the role and with the current manager. When combined with other methods, “TotalWorkingYears” was selected out of this grouping for modeling.

Charts

The team also plotted charts for different variables against the response variable. The following charts show a sample of the attributes that show clear indication of trends related to attrition, and which are ultimately used in the modelling.

Age Groups: Grouping ages into different age groups, we observed a trend of younger individuals 18-25 followed by those aged 26-35 having a higher attrition rate compared to other age groups (Figure 2).

Work Life Balance: A notable trend emerged, showing higher attrition among individuals with lower scores for work-life balance (Figure 3).

No. of Companies Worked: A higher attrition rate among individuals at their 2nd job was observed, followed by a plateau before seeing higher attrition among “serial job movers” (Figure 4).

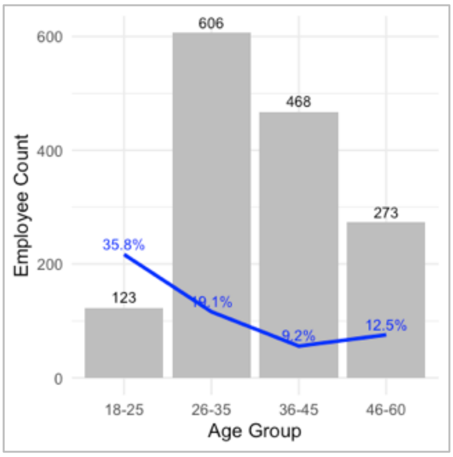


Figure 2: Age Group

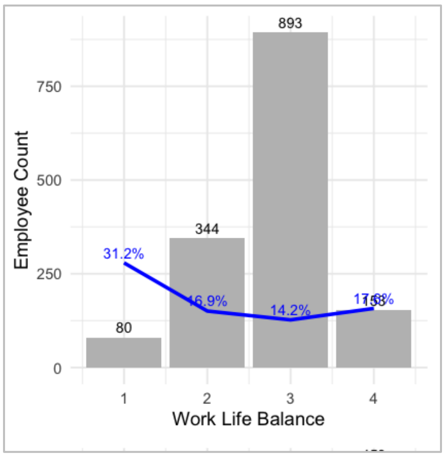


Figure 3: Work Life Balance

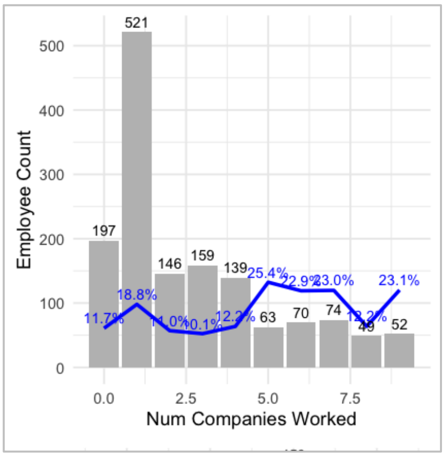


Figure 4: No. of Companies Worked

MODEL TRAINING & OPTIMIZATION

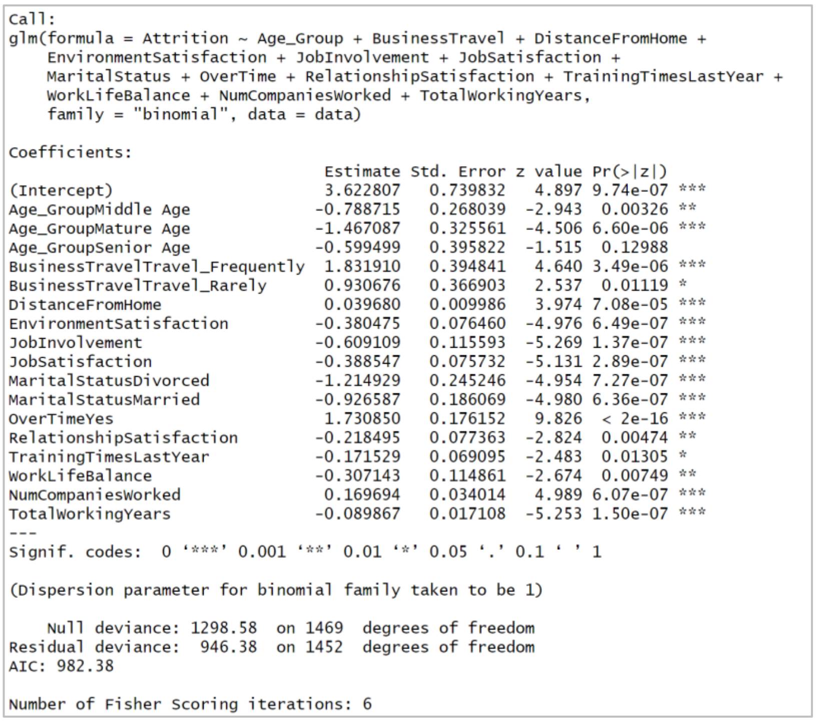
Building on the initial EDA, we conducted a **generalized linear regression (GLM)** as well as **random forest model.** Thereafter, we also used the model to run some **predictions** on employee attrition, and ran **cross-validation** to identify any issues such as overfitting or selection bias.

Initial Logistic Regression Model

The model included all relevant variables from the EDA. We then employed a systematic variable selection process, retaining only those variables with statistically significant effects (p-value < 0.05) on the likelihood of attrition. The following section presents the regression results of our initial model.

Figure 5 shows the results of the regression model, with the following key findings:

Age: Compared to employees younger than 26 years old (Young Age), those in the Middle Age (26-35 years old, -0.79) and Mature Age (36-45 years old, -1.47) groups are less likely to leave the company. There is no statistically significant difference in attrition rates between the Young Age group and the Senior Age group (46 years old and above).



```
Call:
glm(formula = Attrition ~ Age_Group + BusinessTravel + DistanceFromHome +
    EnvironmentSatisfaction + JobInvolvement + JobSatisfaction +
    MaritalStatus + OverTime + RelationshipSatisfaction + TrainingTimesLastYear +
    WorkLifeBalance + NumCompaniesWorked + TotalWorkingYears,
    family = "binomial", data = data)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     3.622807   0.739832   4.897 9.74e-07 ***
Age_GroupMiddle Age            -0.788715   0.268039  -2.943  0.00326 **
Age_GroupMature Age            -1.467087   0.325561  -4.506 6.60e-06 ***
Age_GroupSenior Age            -0.599499   0.395822  -1.515  0.12988
BusinessTravelTravel_Frequently 1.831910   0.394841   4.640 3.49e-06 ***
BusinessTravelTravel_Rarely     0.930676   0.366903   2.537  0.01119 *
DistanceFromHome                0.039680   0.009986   3.974 7.08e-05 ***
EnvironmentSatisfaction        -0.380475   0.076460  -4.976 6.49e-07 ***
JobInvolvement                 -0.609109   0.115593  -5.269 1.37e-07 ***
JobSatisfaction                -0.388547   0.075732  -5.131 2.89e-07 ***
MaritalStatusDivorced          -1.214929   0.245246  -4.954 7.27e-07 ***
MaritalStatusMarried           -0.926587   0.186069  -4.980 6.36e-07 ***
OverTimeYes                     1.730850   0.176152   9.826  < 2e-16 ***
RelationshipSatisfaction       -0.218495   0.077363  -2.824  0.00474 **
TrainingTimesLastYear          -0.171529   0.069095  -2.483  0.01305 *
WorkLifeBalance                -0.307143   0.114861  -2.674  0.00749 **
NumCompaniesWorked              0.169694   0.034014   4.989 6.07e-07 ***
TotalWorkingYears              -0.089867   0.017108  -5.253 1.50e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.58  on 1469  degrees of freedom
Residual deviance:  946.38  on 1452  degrees of freedom
AIC: 982.38

Number of Fisher Scoring iterations: 6
```

Figure 5: Linear Regression Model Results

Business Travel: Employees who travel frequently (Travel_Frequently, 1.83) are significantly more likely to attrite compared to those who don't travel for business. Employees who travel rarely (Travel_Rarely, 0.93) also show a higher propensity to leave, but to a lesser extent.

Work Environment: Higher satisfaction with the work environment (-0.38) is associated with a lower likelihood of attrition.

Job Satisfaction: Similarly, employees with higher job satisfaction (-0.39) are less likely to leave.

Job Involvement: Additionally, higher job involvement (-0.61) is associated with a lower risk of employee attrition.

Marital Status: Both divorced (-1.21) and married employees (-0.93) are less likely to attrite compared to single employees (reference group).

Work-Life Balance: Having a poor work-life balance (-0.31) increases the risk of employee attrition.

Distance from Home: Increased distance from home (0.04) is associated with a higher chance of leaving.

Overtime: Working overtime (Yes, 1.73) significantly increases the likelihood of attrition.

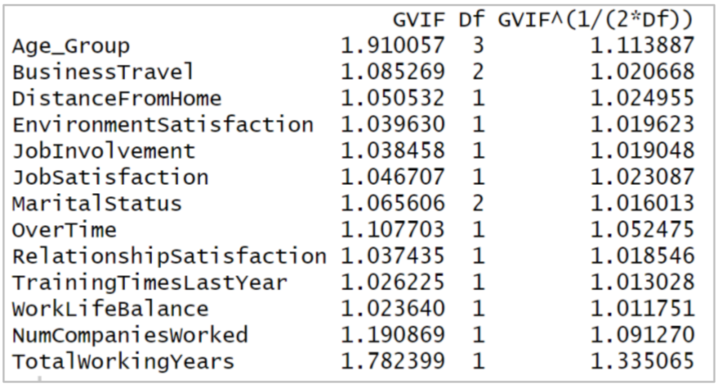
Total Working Years: Employees with more experience (Total Working Years, -0.09) tend to be less likely to attrite.

Training: Attending fewer training sessions last year (TrainingTimesLastYear, -0.17) is associated with a higher chance of leaving.

Number of Companies Worked: Employees who have worked for fewer companies (NumCompaniesWorked, 0.17) tend to be less likely to attrite.

Multicollinearity Assessment

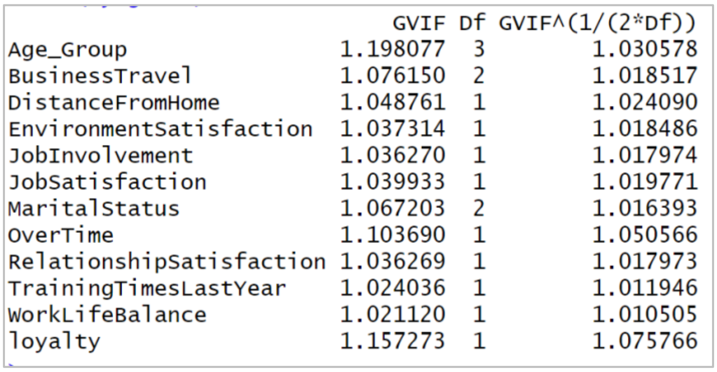
The analysis of variance inflation factors (VIFs) indicates a low risk of multicollinearity for most variables (Figure 6). All VIF scores fall between 1 and 1.1, except for Age_Group (1.91), NumCompaniesWorked (1.19), and TotalWorkingYears (1.78). These slightly higher VIFs suggest a weak to moderate correlation between these three variables, which is understandable. The number of companies worked (NumCompaniesWorked) and total working years likely have a positive relationship with age.



```
                             GVIF Df GVIF^(1/(2*Df))
Age_Group                1.910057  3        1.113887
BusinessTravel           1.085269  2        1.020668
DistanceFromHome         1.050532  1        1.024955
EnvironmentSatisfaction  1.039630  1        1.019623
JobInvolvement           1.038458  1        1.019048
JobSatisfaction          1.046707  1        1.023087
MaritalStatus            1.065606  2        1.016013
OverTime                 1.107703  1        1.052475
RelationshipSatisfaction 1.037435  1        1.018546
TrainingTimesLastYear    1.026225  1        1.013028
WorkLifeBalance          1.023640  1        1.011751
NumCompaniesWorked       1.190869  1        1.091270
TotalWorkingYears        1.782399  1        1.335065
```

Figure 6: VIFs for Initial Model

We explored mitigating this potential multicollinearity by creating a new independent variable, "loyalty" (calculated as NumCompaniesWorked divided by TotalWorkingYears). While the VIF scores (Figure 7) improved with the "Loyalty" variable, the model's Akaike Information Criterion (AIC) worsened slightly (increasing from 982.38 to 995.84). Given this trade-off, and the overall low multicollinearity risk, we will focus on exploring other avenues for model improvement moving forward.



```
                              GVIF Df GVIF^(1/(2*Df))
Age_Group                 1.198077  3        1.030578
BusinessTravel            1.076150  2        1.018517
DistanceFromHome          1.048761  1        1.024090
EnvironmentSatisfaction   1.037314  1        1.018486
JobInvolvement            1.036270  1        1.017974
JobSatisfaction           1.039933  1        1.019771
MaritalStatus             1.067203  2        1.016393
OverTime                  1.103690  1        1.050566
RelationshipSatisfaction  1.036269  1        1.017973
TrainingTimesLastYear     1.024036  1        1.011946
WorkLifeBalance           1.021120  1        1.010505
loyalty                   1.157273  1        1.075766
```

Figure 7: VIFs for Model with "Loyalty Variable"

Random Forest Model

The team also did a **random forest model**, which has been chosen for its ability to handle large amounts of data and ability to achieve a high accuracy.

A random forest model consists of a collection of decision trees, where each tree is trained on a given set of data. Each tree makes predictions on the data and the random forest aggregates the predictions from all individual trees to create a final prediction. This aggregation process produces more robust and accurate predictions compared to any single decision tree. While this method can capture complex relationships and potentially mitigate overfitting, it also is a black box process where the underlying process is hard to describe. This black box can be mitigated through validating the results with the other model, logistic regression.

Random Forest with Selected Predictors

When running the random forest function with twelve predictors, the model can rate the importance of each variable in two terms: %IncMSE and IncNodePurity. Refer to Figure 8 for the results.

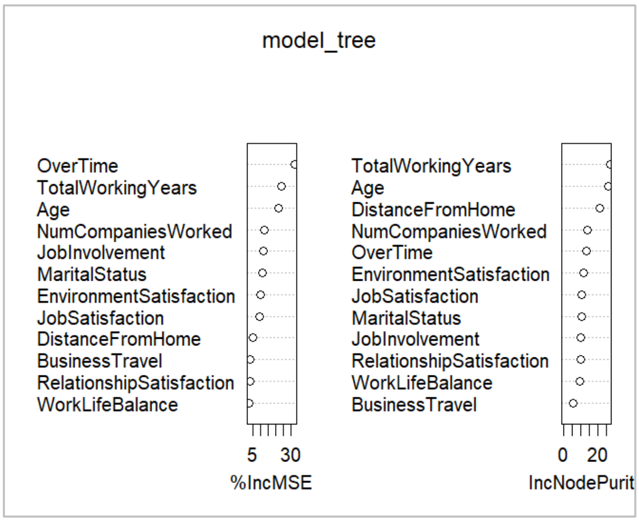


Figure 8: Results for Random Forest

%IncMSE is the increase in mean squared error and is determined by randomly shuffling the values of a particular variable in the model. If the shuffling produces a less accurate response, then that variable is more likely important for the prediction.

IncNodePurity is the increase in node purity and it measures the quality of the model in splitting the data into groups when making decisions. If splitting the data results in more homogenous groups, then it's more likely that predictor is more likely important.

This model repeatedly found TotalWorkingYears and Age to be important predictors. It also produced a psuedo R2 value of 0.207 which is relatively low, and there influence on the model. This indicates that there should be further exploration of random forest models with different predictors.

In conjunction with the logistic modeling, we will explore two more options to reduce and condense the number of predictors: combining the satisfaction ratings of employees and by creating a loyalty score based on an employee's length of tenure at a company.

Random Forest with Composite Rating Score

A composite score was made by combining WorkLifeBalance, EnvironmentSatisfaction, RelationshipSatisfaction, and JobInvolvement. Rather than keep these predictors separate and further complicating the model, they can be aggregated to create one score that can summarize the employee's rated quality of life.

Creating a composite score has reduced the number of variables and has generally increased their overall importance as shown in the figure. However, the composite score itself is not very significant which contributes to a lowed pseudo R2 value is 0.184. This method has actually decreased the predictive power of the model (refer to Figure 9).

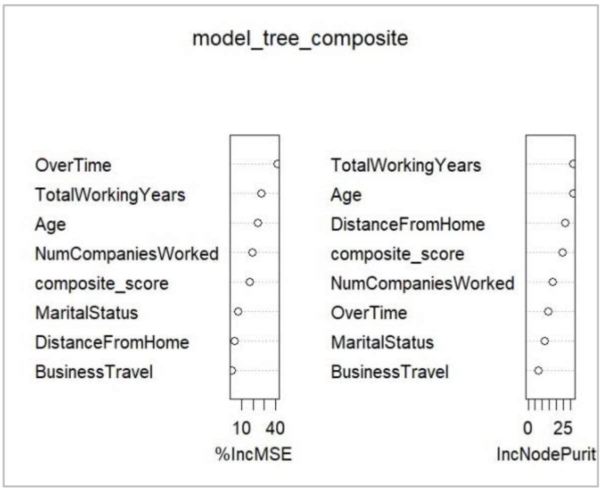


Figure 9: Results for Composite Score

Random Forest with a Loyalty Score

A loyalty score was created by taking the number of companies worked divided by the number of years worked (Figure 10). It is hypothesized that an employee who is more 'loyal' is less likely to leave the company.

A model with this loyalty score has marginally increased the pseudo R2 to 0.209. This is a slight increase of predictive power and the loyalty score does show a large degree of importance as shown on the table. Since this model has an increase in R2, it'll be worth expanding and investigating this model further.

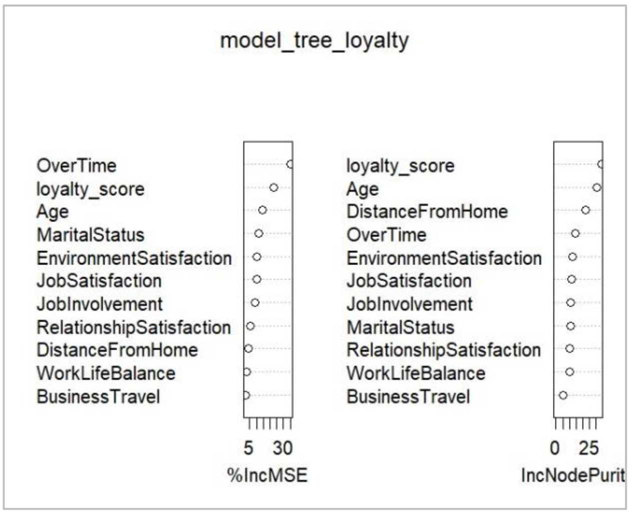


Figure 10: Results with Loyalty Score

The decision tree is included below as this may help with further improvement. For example, age is particularly important, and the level of grouping is important for employees younger than 27 years old or older than 34 years old.

Going forward, these preliminary models have indicated that we can experiment with more reduction of variables. There's also the possibility of creating a different model with both the composite score and a loyalty score. And with the preliminary model of the loyalty score, this will give us more insight on how to group age, or environmental satisfaction for logistical modeling.

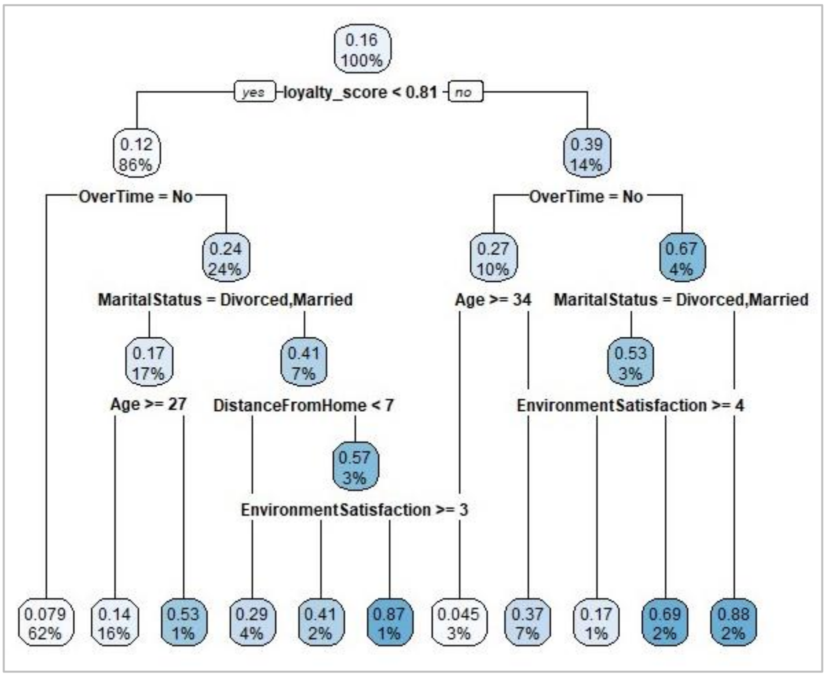
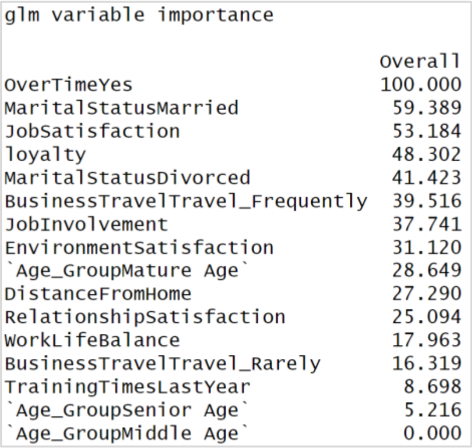


Figure 11: Random Forest Model

PREDICTIONS & CROSS VALIDATION

The previous logistic model had successfully reduced the complexity of the model and the selected predictors were validated with similarities in the random forest model. The next goal is to make the model more interpretable by continuing to reduce predictors while maintaining accuracy and to avoid overfitting. Cross-validation was chosen to achieve this objective.

A new model was created with cross-validation using the significant loyalty score and with this we were able to obtain the variable importance scores as shown in Figure 12. Since age and business travel time are factors of a larger group, these variables need to be retained for coherency with the larger model. Training time, work life balance, and relationship satisfactions are less important and therefore can be removed for an even simpler model.

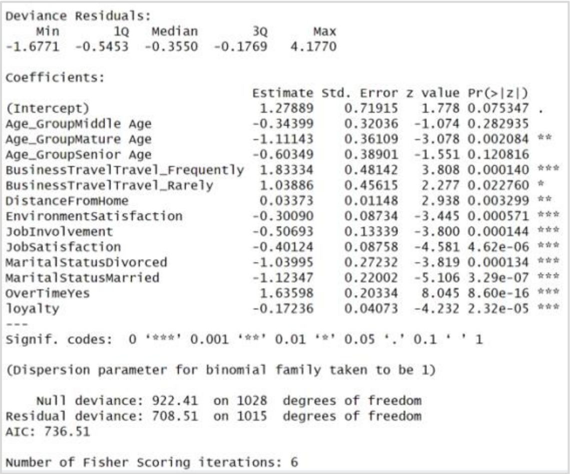


```
glm variable importance

                               Overall
OverTimeYes                    100.000
MaritalStatusMarried            59.389
JobSatisfaction                 53.184
loyalty                         48.302
MaritalStatusDivorced           41.423
BusinessTravelTravel_Frequently 39.516
JobInvolvement                  37.741
EnvironmentSatisfaction         31.120
`Age_GroupMature Age`           28.649
DistanceFromHome                27.290
RelationshipSatisfaction        25.094
WorkLifeBalance                 17.963
BusinessTravelTravel_Rarely     16.319
TrainingTimesLastYear            8.698
`Age_GroupSenior Age`            5.216
`Age_GroupMiddle Age`            0.000
```

Figure 12: Variable Importance

Analysis was then performed on this final, reduced model to measure its prediction accuracy as shown in Figure 13. The probability of attritioning was calculated for different employees, and then cutoff values were applied in steps of 0.05 to find the most accurate classification of these attrition rates.

The most optimal cutoff is 0.45 and the relevant information is displayed in the confusion matrix. This model has an extremely high prediction accuracy at 88.89%. Further analysis of this confusion matrix showed that the model is generally able to make predictions with high accuracy especially if the employee is going to remain at the company. This is reflected in the high specificity score of 0.97872 showing that the model accurately classifies these remaining employees as staying.

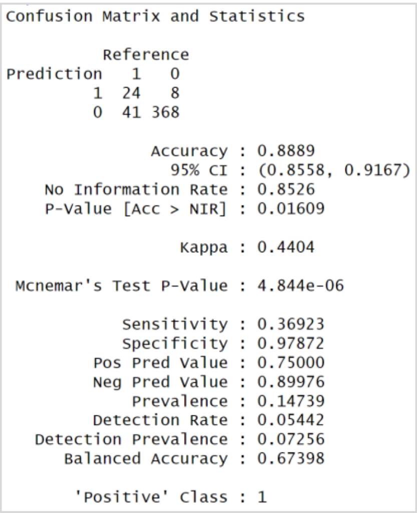


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6771  -0.5453  -0.3550  -0.1769   4.1770

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.27889    0.71915   1.778 0.075347 .
Age_GroupMiddle Age             -0.34399    0.32036  -1.074 0.282935
Age_GroupMature Age             -1.11143    0.36109  -3.078 0.002084 **
Age_GroupSenior Age             -0.60349    0.38901  -1.551 0.120816
BusinessTravelTravel_Frequently  1.83334    0.48142   3.808 0.000140 ***
BusinessTravelTravel_Rarely      1.03886    0.45615   2.277 0.022760 *
DistanceFromHome                 0.03373    0.01148   2.938 0.003299 **
EnvironmentSatisfaction         -0.30090    0.08734  -3.445 0.000571 ***
JobInvolvement                  -0.50693    0.13339  -3.800 0.000144 ***
JobSatisfaction                 -0.40124    0.08758  -4.581 4.62e-06 ***
MaritalStatusDivorced           -1.03995    0.27232  -3.819 0.000134 ***
MaritalStatusMarried            -1.12347    0.22002  -5.106 3.29e-07 ***
OverTimeYes                      1.63598    0.20334   8.045 8.60e-16 ***
loyalty                         -0.17236    0.04073  -4.232 2.32e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 922.41  on 1028  degrees of freedom
Residual deviance: 708.51  on 1015  degrees of freedom
AIC: 736.51

Number of Fisher Scoring iterations: 6
```

Figure 13: Final reduced model

However, while the model makes accurate predictions if an employee is expected to attrition, it is also more likely to misclassify the employee. The lower sensitivity rate at 0.36923 (as seen in Figure 14) shows that the model incorrectly classifies many attritioning employees as retained.

For certain businesses, it may be more important for the model to accurately predict employee attritioning employees. The cutoff value can be lowered to prioritize a higher sensitivity at the cost of being a more accurate model and lowering specificity, the accurate prediction if an employee will be retained. This model would have to be adjusted based on the priorities of that business as this study is focused on finding the key factors attributing to attrition and creating overall accurate predictions.



```
Confusion Matrix and Statistics

          Reference
Prediction   1   0
         1  24   8
         0  41 368

               Accuracy : 0.8889
                 95% CI : (0.8558, 0.9167)
    No Information Rate : 0.8526
    P-Value [Acc > NIR] : 0.01609

                  Kappa : 0.4404

 Mcnemar's Test P-Value : 4.844e-06

            Sensitivity : 0.36923
            Specificity : 0.97872
         Pos Pred Value : 0.75000
         Neg Pred Value : 0.89976
             Prevalence : 0.14739
         Detection Rate : 0.05442
   Detection Prevalence : 0.07256
      Balanced Accuracy : 0.67398

       'Positive' Class : 1
```

Figure 14: Confusion Matrix

MODEL COMPARISON

As previously stated, the decision to use two models was to provide further insight for variable selection due to the large amount of predictors in the dataset. However, both the logistic regression model and the random forest model are workable models that can be used for inference and prediction purposes.

The random forest model using loyalty score found five predictors to be most influential: loyalty score, overtime, marital status, age, distance from home, and environment satisfaction. The final logistic regression model with cross-validation used 9 predictors: age, business travel, distance from home, environment satisfaction, job involvement, job satisfaction, marital status, overtime, and loyalty score.

The most statistically influential variables in the random forest model are subsets of the logistic regression model, giving us further confidence of the importance of the selected variables. However, the random forest model has a low pseudo R-squared. Even though the final logistic regression model (in Figure 13) with cross-validation is more complex, it predicts employee attrition much more accurately so it is the preferred model.

ANSWERING RESEARCH QUESTIONS

The results from the analysis are summarized in respect to the research questions below:

1. **Understand the key underlying factors resulting in employee attrition**
   Age, business travel, distance from home, environment satisfaction, job involvement, job satisfaction, marital status, overtime, and loyalty were considered statistically significant variables to determine employee attrition.

2. **Classify and group factors that are related to employee retention**
   A loyalty score created from total working years / number of companies worked created a statistically significant grouped factor. However, creating a composite rating consisting of work life balance, environmental and relationship satisfaction was not significant for the prediction of employee attrition.

3. **Optimize relevant predictors to minimize employee attrition**
   The two largest factors that increase employee attrition likelihood are overtime and frequent business travel as shown in their large positive magnitude. Minimizing these two factors in an employee's life would have the largest effect on employee attrition.

   Alternatively, hiring employees who are married or are in the mature age category would also decrease the employee's probability of attritioning as shown with the large negative magnitude.

CHALLENGES AND INTERESTING INSIGHTS

One of the challenges the team faced was the availability of many variables in the data set that at first glance, without carrying out EDA, seemed to be related to each other based on real-life working experience. An example includes the variables DailyRate, HourlyRate, MonthlyIncome, MonthlyRate. To get around this, the team conducted extensive EDA and discussions to select the relevant variables to include into the model.

Another challenge faced was that the team was unable to find similar datasets from different industries to test out the how the attrition rates would differ across industries. This was something that was interesting to the team from the start, but lack of quality data prevented the team from exploring this aspect. One possible way that the team can explore in the future is to conduct our own surveys and gather our own data with friends and families from different industries. Even though the dataset might end up being small, it would still be an interesting experiment that the team could run to satisfy our curiosities.

CONCLUSION

In summary, the team sought to identify the key factors and their impact on employee attrition. With a Kaggle dataset, the team conducted EDA to better understand the variables and select relevant variables for analysis. The team then ran linear regression and random forest models to answer the research questions set out, and also ran cross-validation to identify any issues such as overfitting or selection bias.

The team managed to answer the research questions we set out initially, and there was also an additional layer of realism as the team was able to relate to the outcomes personally in different ways due to our own working experiences.

If given the opportunity to, the team would like to extend this analysis across different industries to better understand how employee attrition would differ.

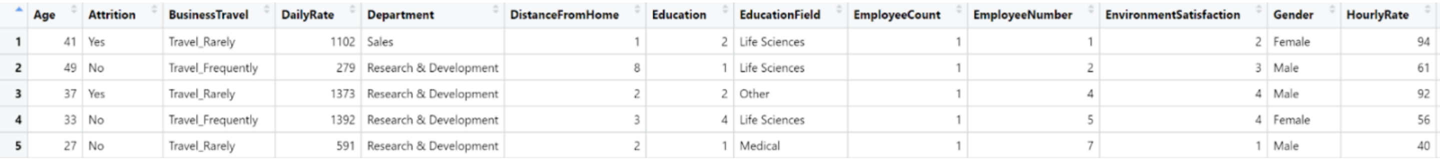
APPENDIX A: DETAILS ON DATASET

Data Sources (links, attachments, etc.):

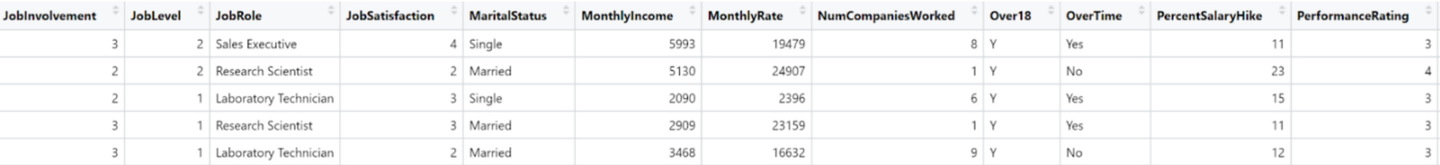
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

Data Description (describe each of your data sources, include screenshots of a few rows of data):

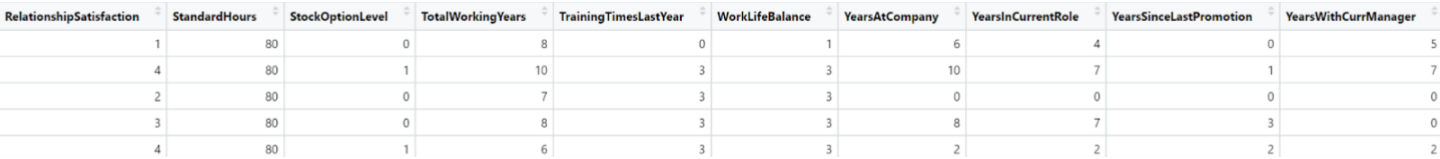
This is a fictional dataset created by IBM data scientists, and contains information about employee attrition. The data includes 35 columns, and the following figures show the column names and sample data for each column:



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male	61
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male	92
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female	56
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male	40



JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	NumCompaniesWorked	Over18	OverTime	PercentSalaryHike	PerformanceRating
3	2	Sales Executive	4	Single	5993	19479	8	Y	Yes	11	3
2	2	Research Scientist	2	Married	5130	24907	1	Y	No	23	4
2	1	Laboratory Technician	3	Single	2090	2396	6	Y	Yes	15	3
3	1	Research Scientist	3	Married	2909	23159	1	Y	Yes	11	3
3	1	Laboratory Technician	2	Married	3468	16632	9	Y	No	12	3



RelationshipSatisfaction	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
1	80	0	8	0	1	6	4	0	5
4	80	1	10	3	3	10	7	1	7
2	80	0	7	3	3	0	0	0	0
3	80	0	8	3	3	8	7	3	0
4	80	1	6	3	3	2	2	2	2

APPENDIX B: BIBLIOGRAPHY

Erling, D. (2011). Appendix III. The Cost of a Mishire: The Story of the Bad Controller. *In Match: A Systematic, Sane Process for Hiring the Right Person Every Time*. essay, Wiley. Retrieved February 15, 2024, from https://learning.oreilly.com/library/view/match-a-systematic/9780470878989/apc.html#you_know_jack.

Navarra, K. (2023, December 21). *The real costs of recruitment*. SHRM. https://www.shrm.org/topics-tools/news/talent-acquisition/real-costs-recruitment

Santovec, M. L. (2010). Build relationships to save the cost of employee attrition. *Women in Higher Education, 19(6)*, 20–21. https://doi.org/10.1002/whe.10066

Zavgorodnii, I., Lalymenko, O., Perova, I., Zhernova, P., & Kiriak, A. (2020). Identification of predictors of burnout among employees of socially significant professions. *Communications in Computer and Information Science,* 445–456. https://doi.org/10.1007/978-3-030-61656-4_30