

**BC2406 Analytics I: Visual & Predictive Techniques Report**

**Prepared for: Professor Pinar Darendeli**

**Seminar Class 7, Team 6**

**Class 7, Team 6**

|  |  |
| --- | --- |
| **Lum Yao Zong, Edward** | **U1810185G** |
| **Harvinder Singh Chahal** | **U1910668J** |
| **Chan Shi Hui Amanda** | **U1910759K** |
| **Lee Yi Shan** | **U1810232H** |
| **Fung Kai Xiang Daniel** | **U1921788J** |

[**1. Executive Summary**](#_axd22m8ju2v5) **3**

[**2. Introduction**](#_qvibof2o23rf) **4**

[**2.1 Business Problem Statement**](#_571p31sb602l) **4**

[**2.2 Why is the Business Problem Significant?**](#_3bcap177739f) **4**

[**2.3 Job Embeddedness Theory**](#_b6jzkefs1fb5) **4**

[**2.4 Benefits of Job Embeddedness Theory Approach**](#_qktw6x1n9ice) **5**

[**3. Data Preparation**](#_846m0mi1eyuc) **6**

[**3.1 Data Cleaning**](#_4icxe83qgx3b) **6**

[**3.1.1 Duplicated Rows**](#_fjy0gwe39ck0) **6**

[**3.1.2 Test Values**](#_huo59dkvdonl) **6**

[**3.1.3 Wrong Values**](#_5ddboljfqz7x) **6**

[**3.1.4 Dropping Unnecessary Columns**](#_quwvx911qy1g) **6**

[**3.1.5 Illogical Values**](#_hqfr9dw26iwe) **7**

[**3.1.6 NA and Missing Values**](#_1llw68tipnrv) **7**

[**3.2 Data Correlation**](#_s1394dslw9k9) **7**

[**3.3 Train-Test Split**](#_1s3bdjb5vfnp) **7**

[**4. Fit**](#_j5spr5g6asm5) **8**

[**4.1 Data Exploration**](#_qrzp4iejt55d) **8**

[**4.1.1 Insight 1: Analysis of YearsAtCompany**](#_hiuox081l6np) **8**

[**4.1.2 Insight 2: Analysis of JobLevelNew**](#_2p0vhe4tow54) **8**

[**4.1.3 Insight 3: Analysis of MonthlyIncomeNew**](#_nl15z59g3z2k) **8**

[**4.1.4 Insight 4: Analysis of JobRolesNew**](#_n54sktprzl45) **9**

[**4.2 Data Modelling (Linear Regression)**](#_z1ifmcqvyjxg) **9**

[**4.2.1 Linear Model 1**](#_hdhyquam6yp) **9**

[**4.2.2 Linear Model 2**](#_cku3lgptfl7z) **9**

[**4.2.3 Linear Model 3**](#_38nwcwgoq4xd) **10**

[**4.2.4 Linear Model 4**](#_mvj4b1o98cl) **10**

[**4.2.5 Effectiveness of model (RMSE)**](#_wouiousk5uzj) **10**

[**5. Link**](#_96s7nvhqc7ra) **10**

[**5.1 Data Exploration**](#_eog9h6cgj1ej) **10**

[**5.1.1 Insight 1: Analysis of JobInvolvementNew**](#_cgfg8p3tsoaz) **10**

[**5.1.2 Insight 2: Analysis of Age**](#_srqykdosmign) **10**

[**5.1.3 Insight 3: Analysis of MonthlyIncomeNew**](#_gmuajviylqn7) **11**

[**5.1.4 Insight 4: Analysis of PercentSalaryHikeNew**](#_g2sw9oy6cy51) **11**

[**5.1.5 Insight 5: Analysis of YearsInCurrentRoleNew**](#_y5mpegmsros8) **12**

[**5.2 Data Modelling (CART)**](#_6upnzgt9eb7d) **12**

[**5.2.1 Phase 1**](#_r6iv63bqjpoy) **12**

[**5.2.2 Phase 2**](#_2xwrcnxrvs8o) **13**

[**5.2.3 Data Prediction**](#_9caaj5jy7cdq) **13**

[**6. Sacrifice**](#_o2mp17bpz0xt) **14**

[**6.1 Data Exploration**](#_begln6gib184) **14**

[**6.1.1 Insight 1: Analysis of JobSatisfactionNew**](#_hgaf5r3d90eb) **14**

[**6.1.2 Insight 2: Analysis of HourlyRateNew**](#_ungv40j401k3) **14**

[**6.1.3 Insight 3: Analysis of MonthlyIncomeNew**](#_d9wcwfgzmjnv) **14**

[**6.1.4 Insight 4: Analysis of TotalWorkingYearsNew**](#_2urtdvhxjw9u) **15**

[**6.1.5 Insight 5: Analysis of YearsInCurrentRoleNew**](#_b3s4lyik07zl) **15**

[**6.2 Data Modelling (CART)**](#_wmt483nnb7cq) **16**

[**6.2.1 Phase 1**](#_p7uzbbkbhztr) **16**

[**6.2.2 Phase 2**](#_78xr034yn630) **16**

[**6.2.3 Data Prediction**](#_ml8yhpb1sb1x) **17**

[**7 Attrition**](#_3ahhaeup9yjm) **17**

[**7.1 Data Exploration**](#_zi8sskl9npgz) **17**

[**7.1.1 Insight 1: Analysis of AttritionNew**](#_2ej7zicvfv3u) **17**

[**7.1.2 Insight 2: Analysis of JobSatisfactionNew**](#_aulao0x2gdv3) **17**

[**7.1.3 Insight 3: Analysis of YearsAtCompany**](#_9g6mm2tbpuee) **17**

[**7.1.4 Insight 4: Mean and Median of YearsAtCompany**](#_8k89z6jdscak) **18**

[**7.2 Data Modelling (Logistic Regression)**](#_fpylzeu6jqe6) **18**

[**7.2.1 Logistic Model One (Imbalanced data)**](#_h9wrwtjbqrum) **18**

[**7.2.2 Logistic Model One (Balanced data)**](#_vn4plp682djp) **19**

[**8. Business Recommendations**](#_ar3hh1wqr50) **19**

[**8.1 Solution Overview for Data Gathering**](#_vajbtc5p48gc) **19**

[**8.2 Solution Overview for Mitigating Flight Risk**](#_ea5gdb32bcdk) **20**

[**8.2.1 Explanation on How to Use Our Model**](#_g0hmhjeu064u) **20**

[**8.2.2 Recommendations**](#_nafs7nf7f2op) **21**

[**8.3 Limitations**](#_9x65iugt1k3o) **21**

[**9. Conclusion**](#_v82gs2tpex8y) **22**

[**10. References**](#_mggwdtr70mfq) **23**

[**11. Appendix**](#_4ymfg2ygwtc9) **25**

#### 1. Executive Summary

White Rock is a company in the asset management industry and is currently facing an issue of high employee turnover.

This report seeks to identify solutions and business recommendations to help White Rock reduce its employees’ flight risk. Our flexible and customisable solution is based on Job Embeddedness Theory which has been proven to be predictively valid from past research (Lee et al., 2014). It consists of three components - fit, link and sacrifice. These components refer to the congruence of common values, the integratedness between employee and the company and loss suffered by the resigning employee respectively.

Analytics will be incorporated within the job embeddedness theory. The first step is to conduct data cleaning which ensures the completeness and accuracy of the dataset. Next, fit is measured using years at company through the linear regression while link and sacrifice are measured by job involvement level and job satisfaction level respectively using Classification and Regression Trees (CART) model. Lastly, the three variables (YearsAtCompany, JobInvolvement and JobSatisfaction) will be combined to predict employee’s attrition on whether he/she is more inclined to stay on or resign through logistic regression.

To ease White Rock’s transition in using our solution, we have designed a questionnaire to gather data such as employees’ JobSatisfaction level, JobInvolvement level and work life balance that is usually not stored in their employee database. With the data, White Rock will be able to follow up with appropriate remedies based on individual employee’s fundamental reasons for leaving and hence reduce employee turnover.

White Rock will expect a reduction in turnover costs, an improvement in its uptime with respect to operational activities and a boost in its employees’ morale with the above interventions. However, inherent limitations that are not within our control could potentially reduce the effectiveness of the solution. One such example is the company’s inability to capture data relating to external circumstances causing the employee to resign.

All in all, embedding analytics into job embeddedness theory aims to reduce White Rock’s employee turnover and mitigate employees’ high flight risk. It is desired that through this solution, White Rock’s financial, economic and human resources are maximised to their potential to reinforce White Rock’s strong positioning in the asset management industry.

#### 2. Introduction

##### 2.1 Business Problem Statement

Employee turnover rates at investment and asset management companies have risen rapidly over the years. This holds true even during the current covid-19 pandemic where employee turnover rates have not declined (Globe Newswire, 2020). Some probable causes include better job opportunities, ill-fit with the company culture and unsatisfactory pay.

In order to stay competitive in the asset management industry, White Rock cannot afford to lose its best employees. Currently, the company faces an issue of employee retention as it is unable to identify employees that have higher flight risk. Therefore, the company requires an analytics-driven solution that is able to identify current, as well as predict future high flight risk individuals based on employee factors.

As analytics consultants to White Rock, we aim to study the dataset of employee factors to identify the underlying causes of flight risk. This allows us to assist White Rock’s management in categorizing new and current employees into levels of flight risk and highlight potential reasons behind employees’ high flight risk.

##### 2.2 Why is the Business Problem Significant?

Employee turnover is a significant issue to White Rock as it operates in the asset management sector, where a great personalized clientele experience is paramount to its business success. Employees play a pivotal role in providing this experience as they establish and maintain close relationships with the clients. As such, the ability to retain its competent employees contributes directly to White Rock’s profitability in the long run.

Moreover, replacing employees incurs significant costs, such as separation costs, replacement costs, and re-training costs. Losing an employee can result in turnover costs ranging between 1.5 to 2 times the employee’s salary (Heinz, 2020). By retaining employees, White Rock will be able to avoid these significant costs.

Lastly, high employee turnover negatively influences the morale of the remaining employees. When a department head leaves, other employees in that department may also follow suit, causing waves of employee turnover. This not only weakens White Rock’s corporate culture, but also erodes the remaining employees' trust in the organisation.

##### 2.3 Job Embeddedness Theory

In response to the high flight risk issue faced by White Rock, we have adopted an approach based on the Job Embeddedness Theory (Young et al., 2013). This theory suggests that employee turnover is affected by three main components, namely the employees’ fit with the organisation, their social links, and the sacrifices they need to make to change jobs.

First, fit is the congruence of common values between the employee and company. We choose YearsAtCompany as the y-variable as employees tend to stay longer when they match company culture, indicating a commonality of values and beliefs between them and White Rock (Skapinker, 2020). We use the linear regression model as YearsAtCompany is a continuous variable, with higher YearsAtCompany signifying a greater fit in the company.

Next, link is the integratedness between an employee and the organisation. We choose JobInvolvement as the y-variable as higher JobInvolvement levels signify greater responsibilities and opportunities to build relationships in the workplace. Thus, it is a proxy for how linked an employee is. We use the CART model as JobInvolvement is a categorical variable, with higher JobInvolvement signifying greater links with the company.

Last, sacrifice is what the employee stands to lose. We choose JobSatisfaction as the y-variable as it represents an overall value that employees derive from their job. This includes contentment with their colleagues, sense of achievement and remuneration. We use the CART model again as JobSatisfaction is a categorical variable, with higher JobSatisfaction implying greater sacrifice ought to be made by the employee.

After completing the data analysis process using the job embeddedness theory, we combine all the 3 y-variables (YearsAtCompany, JobInvolvement and JobSatisfaction) to predict employee’s Attrition. As Attrition is a binary categorical variable, we use the logistic regression model to predict it. This helps White Rock to determine if an employee is likely to stay or voluntarily resign, and then let them decide whether or not to take action to retain these employees.

##### 2.4 Benefits of Job Embeddedness Theory Approach

Using Job Embeddedness Theory makes it easier to pinpoint areas that require more attention. Managers at White Rock will be able to see which variables are negatively affecting an employee’s job embeddedness and save resources by targeting those variables. By doing so, less wastage of resources is incurred from targeting at the insignificant variables. This approach is also specific, since only those significant adverse indicators are being focused on. This helps to improve job embeddedness directly and efficiently.

With the theory, we can also obtain a predictive summary of the employees’ job embeddedness such as their JobSatisfaction and JobInvolvement levels which are more objective. This can avoid conflicts of interest where some employees may feel deterred to voice out dissatisfaction to White Rock during their annual performance review.

In addition, with the use of our suggested questionnaire, which will be further explained in section 8, White Rock does not need to manually conduct interviews to find out what their employees truly feel about their job, collecting more truthful employee sentiments. This reduces opportunity costs in terms of the time spent manually conducting interviews where the time could have been better spent on coming up with solutions to retain the employees or driving the company’s operations.

These will help White Rock save on turnover costs by improving employee’s job embeddedness and hence, reducing employee attrition. At the same time, happier employees are likely to be more productive (Bellet et al., 2019). Overall, our model will make it easier for White Rock to improve employee’s experience.

#### 3. Data Preparation

##### 3.1 Data Cleaning

Data cleaning is first performed with the information currently available, to ensure a more accurate analysis.

###### 3.1.1 Duplicated Rows

We check for the presence of entirely duplicated rows by comparing the total number of rows in the dataset with the total number of unique rows in the dataset. Since the comparison returns false, duplicated rows are detected. We choose to drop all duplicated rows instead of keeping one unique row of each duplicated set as we are unsure if these rows are ‘test’ rows (see Section 3.1.2).

###### 3.1.2 Test Values

We check the values in each column and note irregularities such as ‘test’, ‘TEST’ and ‘test 456’ which are likely inserted into the dataset during system testing. These values are dropped since they are not genuine records of employees.

###### 3.1.3 Wrong Values

We also note that ‘?????’ is present in several rows of Application.ID. During the usual hiring process, an applicant would have to submit their application (identified by a key, Application.ID) before they are considered for employment in the company. Therefore, the creation of EmployeeNumber and its corresponding records are dependent on the existence of Application.ID. As such, we drop the rows corresponding to invalid Application.ID values.

###### 3.1.4 Dropping Unnecessary Columns

We drop EmployeeNumber, EducationField, EmployeeCount, Application.ID, StandardHours, Employee.Source and Over18. EmployeeCount, StandardHours and Over18 have the same values for all rows which do not value-add to our analysis. EmployeeNumber, EducationField, Application.ID and Employee.Source are also irrelevant to our analysis.

We also notice some irregularities in these columns such as multiple values of “Marketing” being erroneously entered into the EmployeeCount column. However, we do not drop these rows because unlike Application.ID, EmployeeCount is not a unique identifier and its irregularity does not render the entire row invalid.

###### 3.1.5 Illogical Values

We check for illogical values and drop them since they are outliers based on common sense. (Table 1)

###### 3.1.6 NA and Missing Values

When the dataset is imported into R, missing values are displayed as NA and blank fields. NA values also arise from when we set appropriate data types (Table 2) but these are dropped as it would mean the data in the wrong data type is invalid in the first place.

We investigate the rows pertaining to NA and missing values and found that these errors are random instead of systematic. Furthermore, NA values occupy a small percentage of the dataset (~200/20000 = 1%). As such, these values are replaced with the population mode and mean for categorical and continuous variables respectively (Table 3). The cleaned data are represented in another column, named as <VariableName>New (for example, JobLevel is represented as JobLevelNew) to preserve the original dataset information (Table 4).

We reperform data cleaning for illogical values in cleaned columns and drop those rows that are found to contain them. The final cleaned data is written to “IBM\_cleaned.csv”.

##### 3.2 Data Correlation

As part of data exploration, we generate the correlation matrix (Fig 5) to better understand which variables are more related to the fit, link and sacrifice framework. Insights from the correlation matrix are presented in Table 6. From the insights, we construct data tables for each component of the job embeddedness theory.

##### 3.3 Train-Test Split

For each of the three data tables, train-test split is done based on a generally accepted convention of a 0.7 split-ratio. This way, we have 30% (number of observations) as unseen data to test the accuracy of our models. We use set.seed function to make sure the same train and test sets are generated each time the code is run. A comparison of the summary statistics between train set and test set is conducted afterwards to ensure that the distribution of Y is similar in both sets so as to avoid unreliable tests and high RMSE.

#### 4. Fit

##### 4.1 Data Exploration

###### 4.1.1 Insight 1: Analysis of YearsAtCompany

We begin by exploring the number of years that each employee has worked at the company, which is the y variable for fit analysis. Through a bar graph as shown in Fig 7, we observe how long employees have worked at the company and how many of them fall into each category. Through the summary function, we discover that both the median and mode is 5 years while average is 6.91 years. This suggests that a majority of employees decide to leave the company by the 10 year mark and that the current workforce at the company is relatively new to their roles. Since the goal of our analysis is to retain current employees, it is crucial to improve these numbers. Hence, we will further explore the various factors which could potentially influence the YearsAtCompany results.

###### 4.1.2 Insight 2: Analysis of JobLevelNew

We obtain an understanding of the current situation in the company by looking at JobLevelNew independently before analyzing how it affects YearsAtCompany. By plotting a bar graph as shown in Fig 8, we observe that more than two-thirds of the company’s workforce are employees with lower job levels, with 37.26% of employees only holding a position with JobLevelNew of 1 and 36.67% being able to achieve a JobLevelNew of 2.

However, when we further explore JobLevelNew in relation to YearsAtCompany, as shown by the boxplot in Fig 9, we observe that the average YearsAtCompany increases as JobLevelNew increases, with 3.99 years at level 1 to 14.55 years at level 5.

This supports what we found in 4.1.1 and also suggests YearsAtCompany could be higher if a majority of the employees are not stuck at lower job levels.

###### 4.1.3 Insight 3: Analysis of MonthlyIncomeNew

Once again we explore another variable by observing the bar graph as shown in Fig 10, through which we can notice another trend. A majority of the employees make $7,500 a month or less, with the mean being $6,417 and the median being $4,898.

However, when we plot a smooth line as observed in Fig 11, we discover that the number of years an employee chooses to work at a company increases as their MonthlyIncomeNew increases. Those making $7500 monthly still spend less than 10 years at the company, supporting what we found in 4.1.1.

###### 4.1.4 Insight 4: Analysis of JobRolesNew

In similar fashion, we take a look at another variable, JobRolesNew. The bar graph in Fig 12, shows the number of roles filled for every position with 56% of employees holding non-leadership roles and 44% holding leadership roles (generally we can classify Executives, Managers and Directors into leadership roles and the rest as non-leadership roles). As such, we can observe that there are more non-leadership roles being filled as compared to leadership roles.

We then go on to view this in relation to YearsAtCompany, through the boxplot in Fig 13. The averages in this figure depict that employees with leadership roles opt to spend more years at the company than those without leadership roles.

Since the majority of employees do not hold leadership roles, this supports the low value of YearsAtCompany in section 4.1.1.

##### 4.2 Data Modelling (Linear Regression)

###### 4.2.1 Linear Model 1

Following data exploration, we apply Linear Regression model on our Fit trainset as the y-variable, YearsAtCompany, is continuous in nature. The model shows that all our independent variables are statistically significant, with the model returning an adjusted R-squared value of 78.01 which means that 78.01% of the data can be explained by the model.

However, we notice that the coefficients for DailyRateNew, JobLevelNew, JobRoleNew, MonthlyIncomeNew and MonthlyRateNew are all negative and this does not make sense as YearsAtCompany should be increasing with increases in these variables. Hence, we conduct a multi-collinearity test to determine which variables are causing this problem. JobLevelNew, JobRoleNew and MonthlyIncomeNew all return VIF values greater than 10, with JobLevelNew being the highest with a value of 32.20.

###### 4.2.2 Linear Model 2

As such, we remove JobLevelNew as one of the variables and run the model as well as the multi-collinearity test on the variables again. This time, the model returns an adjusted R-squared value of 77.8% with all variables returning VIF values less than 10. We proceed with conducting a diagnostic check on the linear model and realize that the graph for Normal Q-Q does not provide a result which follows a normal distribution with mean 0. In an attempt to rectify the problem, we first identify it by running a residual check, through which we determined that the distribution skewed to the right.

###### 4.2.3 Linear Model 3

This leads us to run the model once again but with the dependent variable as log(YearsAtCompany+1). This time, the model returns an adjusted R-squared value of 77.9%. However, the variables DailyRateNew & MonthlyRateNew are suggested to be statistically insignificant.

###### 4.2.4 Linear Model 4

We run our model one last time, without the statistically insignificant variables mentioned above. The model returns an adjusted R-squared value of 77.9%, while suggesting that all variables are statistically significant. The multi-collinearity test also returns VIF values less than 10 for all variables. Diagnostic checks also suggest that all 4 graphs produced acceptable results. Residual checks also reflect that the new residual values resemble more closely to a standard normal distribution as compared to that of linear model 2.

###### 4.2.5 Effectiveness of model (RMSE)

Finally, we check how good our model is at predicting YearsAtCompany and get a Root Mean Squared Error of 0.355 for the Fit trainset and 0.352 for the Fit testset (Fig 18), suggesting that our model is indeed effective in predicting accurate results for Fit.

#### 5. Link

##### 5.1 Data Exploration

###### 5.1.1 Insight 1: Analysis of JobInvolvementNew

We begin by exploring the JobInvolvementNew of the employees in the company, which is the variable measured in our link analysis. By plotting a bar graph as shown in Fig 19, we observe that most employees have high JobInvolvementNew levels. This is followed by the category of employees displaying medium JobInvolvementNew levels. This suggests that the majority of employees in the company are relatively participative in their daily operations. We further explore various factors that may potentially cause these JobInvolvementNew results.

###### 5.1.2 Insight 2: Analysis of Age

Next, we explore the distribution of employees’ Age across the 4 categories of JobInvolvementNew, to understand the relationship between Age and JobInvolvementNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 20, we observe that employees who display the lowest JobInvolvementNew levels tend to be the youngest in the company, centered around Age 30. In comparison, employees who display the highest JobInvolvementNew levels tend to be the oldest in the company, centered around age 40.

Furthermore, senior employees, those Age 50 and above, mostly display high to very high JobInvolvementNew levels. This suggests a possible positive correlation between Age and JobInvolvementNew, and that older employees tend to be more engaged and involved in their daily operations (Kim & Kang, 2017).

###### 5.1.3 Insight 3: Analysis of MonthlyIncomeNew

Next, we explore the distribution of employees’ MonthlyIncomeNew across the 4 categories of JobInvolvementNew, to understand the relationship between MonthlyIncomeNew and JobInvolvementNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 21, we observe that employees who display the highest JobInvolvementNew levels are mostly paid the lowest in terms of MonthlyIncomeNew, centered around $3000. In comparison, employees who display the lowest JobInvolvementNew levels are paid significantly higher, centered around $6000 and some at $10,000. This is counter-intuitive, as we expect employees who are more involved with their responsibilities to attain better recognition in the form of remuneration.

As seen from the boxplot shown in Fig 22, we may also observe a similar trend, where the median MonthlyIncomeNew of employees who are least involved are the highest, centered around $6000, while the median MonthlyIncomeNew of employees who are most involved are the lowest, centered around $3000. This suggests that employees are not adequately recognised for their level of contribution and involvement in the company, and there is inequity with respect to the company’s remuneration policy.

###### 5.1.4 Insight 4: Analysis of PercentSalaryHikeNew

Next, we explore the distribution of employees’ PercentSalaryHikeNew across the 4 categories of JobInvolvementNew, to understand the relationship between PercentSalaryHikeNew and JobInvolvementNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 23, we observe that employees who display the lowest JobInvolvementNew levels have attained the highest PercentSalaryHikeNew, centered around 12% and 22%. In comparison, employees who display the highest JobInvolvementNew levels attained significantly lower PercentSalaryHikeNew, centered at only 6% and 17%. This follows a similar counter-intuitive trend to the analysis of MonthlyIncomeNew, as we expect employees who are more involved with their responsibilities to attain better recognition in the form of better PercentSalaryHikeNew.

As seen from the boxplot shown in Fig 24, we may observe that the median PercentSalaryHikeNew of employees who are least involved are approximately the same as the median PercentSalaryHikeNew of employees who are most involved, at 15%. The median PercentSalaryHikeNew of employees who have displayed neutral and high JobInvolvementNew levels are however, lower than that of employees who are least involved, at 10%. This likewise suggests that employees who are least involved in their work are attaining higher pay raises as opposed to employees who are more engaged in their work, which implies that the company is not allocating its financial resources equitably.

###### 5.1.5 Insight 5: Analysis of YearsInCurrentRoleNew

Next, we explore the distribution of employees’ YearsInCurrentRoleNew across the 4 categories of JobInvolvementNew, to understand the relationship between YearsInCurrentRoleNew and JobInvolvementNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 25, we observe that employees who display neutral to very high levels of JobInvolvementNew have mostly worked between 1-3 years in their current roles, while some have worked between 6-7 years in their current roles. The right skewness of the density plot suggests that the employees of the company are mostly new in their current roles, where most have only worked for 1-2 years in their current role. The data will be more informative, with respect to employees’ JobInvolvementNew changes, if the company has more employees that have remained in their current roles for longer periods of time.

##### 5.2 Data Modelling (CART)

Following data exploration, we adopt the CART model for our link analysis. Since the y-variable, JobInvolvementNew, is categorical in nature, the model used is specifically a classification tree. This is completed in two phases.

###### 5.2.1 Phase 1

In Phase 1, we grow the tree to the maximum. At each node, the CART algorithm will consider and test all x variables, all possible values in that x variable to determine the best splitting variable and binary split, which ideally results in the purest possible child nodes, on average. The best splits are then used to generate 2 child nodes, and the x variables with the highest associations with the y variable will be repeatedly splitted first to generate the splitting rules. The process continues for each child node, until a predetermined stopping criteria is met. These criteria include all terminal nodes being completely pure and the number of observations in each terminal node reaches the pre-specified minimum. This methodology is technically called recursive partitioning (Strobl, C., Malley, J., & Tutz, G., 2009, December). The fully grown tree will likely overfit the train set used and the resulting model may be overly complex and difficult to interpret, leading to poor test set performance (Boehmke & Greenwell, 2020). This is solved via Phase 2, where the tree is pruned to its optimal level.

We first import the data.table, rpart and rpart.plot libraries. We then construct the maximal tree (Fig 26) with our y-variable, JobInvolvementNew, against all other x-variables using our link\_trainset data. The minimum number of observations in each node that must exist before a split can be attempted (minsplit) is set to 12, and the complexity parameter (cp) is set to 0 to ensure we do not stop the tree growing process prematurely. This will be called link.cart1.

We observe the maximal tree in link\_trainset to be shown in Fig 27.

###### 5.2.2 Phase 2

In Phase 2, we prune to simplify the overfitted tree by applying the optimal complexity parameter (cp). As all trees whose 10-fold Cross Validation (CV) error is below the CV error cap are statistically significant in terms of error, the optimal tree will be the simplest tree with the highest CV error right within the CV error cap (Kumar., n.d.). The CV error cap is determined by adding a 1 standard error (xstd) to the minimum CV error. The pruned optimal tree will be the most stable tree that still performs well.

We plot the CV errors of each prune sequence against their respective prune triggers (cp) based on geometric mean as a chart to determine the optimal cp as shown in Fig 28 .

As it is unclear to identify the optimal cp just based solely on observation, we automate the search for the optimal tree (Fig 29). First, we compute the CV error cap by identifying the minimum CV error and adding 1 standard error (xstd) to it. This will be called link\_CVerror.cap. Next, we check all CV errors against the CV error cap, starting from the first tree onward, using a while loop. The loop will stop when the CV error is smaller than the CV error cap, to identify the optimal cp region where the CV error is just below the CV error cap in the maximal tree. Finally, we compute the geometric mean of the 2 identified cp values in the optimal region, which will be the optimal cp value used in pruning. This will be called link\_cp.opt. In this case, the 47th pruning sequence gives the optimal tree.

We proceed to prune the maximal tree using the link\_cp.opt, and this will be called link.cart2 (Fig 30). By observing the cp chart of the pruned tree (Fig 31) , we may observe that the size of the tree is slightly smaller, and there is only 1 point that has the CV error being just below the CV error cap, demarcated by the horizontal dotted line. Based on the printcp table (Fig 32), we observe that the pruned tree is left with 633 splits, the root node error is 0.40764, the train set misclassification error is 0.02988 (=0.073300 \* 0.40764), and the test set misclassification error is 0.05340 (=0.13100 \* 0.40764)

###### 5.2.3 Data Prediction

Lastly, we test the link.cart2 model using our link\_testset data (Fig 33). This is done by predicting the outcome for the JobInvolvementNew using the model, and the following tabulates the results of the prediction against the actual JobInvolvementNew levels in the test set data (Fig 34). The overall accuracy of the model is 0.9461328, and MonthlyRateNew is observed to be the most important x variable, followed by MonthlyIncomeNew (Fig 35).

#### 6. Sacrifice

##### 6.1 Data Exploration

###### 6.1.1 Insight 1: Analysis of JobSatisfactionNew

We begin by exploring the JobSatisfactionNew levels of the employees in the company, which is the variable measured in our sacrifice analysis. By plotting a bar graph as shown in Fig 36, we observe that employees that have high and very high JobSatisfactionNew levels make up most of the group, and that there are slightly more employees with very high JobSatisfactionNew level compared to high JobSatisfactionNew level. This is followed by the category of employees displaying neutral JobSatisfactionNew levels. This suggests that the majority of employees in the company are relatively satisfied with their jobs and the company, and that employees who eventually decide to leave White Rock may potentially sacrifice much of this satisfaction. We will further explore various factors that may potentially cause these JobSatisfactionNew results.

###### 6.1.2 Insight 2: Analysis of HourlyRateNew

Next, we explore the distribution of employees’ HourlyRateNew across the 4 categories of JobSatisfactionNew, to understand the relationship between HourlyRateNew and JobSatisfactionNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 37, we observe that most employees who display the lowest JobSatisfactionNew levels are paid the highest hourly rates in the company, with most of them being paid around $83 per hour. In comparison, most employees who display the highest JobSatisfactionNew levels are paid the lowest in terms of hourly, with most of them being paid around $45 per hour. Interestingly, we may observe a peak towards the upper bracket of hourly rates for employees who are least satisfied, indicating that most employees who are least satisfied are in fact paid higher than their colleagues who are more satisfied. This suggests a possible negative correlation between HourlyRateNew and JobSatisfactionNew, and that employees who are receiving the highest remuneration, most likely the senior management, are the least satisfied with their jobs.

###### 6.1.3 Insight 3: Analysis of MonthlyIncomeNew

Next, we explore the distribution of employees’ MonthlyIncomeNew across the 4 categories of JobSatisfactionNew, to understand the relationship between MonthlyIncomeNew and JobSatisfactionNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 38, we observe that employees who display the lowest JobSatisfactionNew levels tend to be paid the most in the company, with a relatively higher proportion of employees being paid above $10,000 per month. In comparison, employees who display the highest JobSatisfactionNew levels tend to be paid the least, with a relatively lower proportion of employees being paid above $10,000 a month. This is consistent with our previous analysis of HourlyRateNew, where employees who are receiving higher remuneration, likely the senior management, are least satisfied with their jobs in the company.

However, as seen from the boxplot shown in Fig 39, we may observe that the median MonthlyIncomeNew of employees as well as the distribution of outliers across all 4 categories of JobSatisfactionNew are similar, which suggests varying satisfaction levels across employees, regardless of MonthlyIncomeNew levels.

###### 6.1.4 Insight 4: Analysis of TotalWorkingYearsNew

Next, we explore the distribution of employees’ TotalWorkingYearsNew across the 4 categories of JobSatisfactionNew, to understand the relationship between TotalWorkingYearsNew and JobSatisfactionNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 40, we observe a similar distribution of employees in terms of their TotalWorkingYearsNew, across each category of JobSatisfactionNew, with a significantly high proportion of employees having worked for 10 years in total.

The right skewness of the density plot suggests that the employees of the company are mostly of a younger demographic, having only worked around 10 years in total thus far. The data will be more informative, in respect of employees’ JobSatisfactionNew changes, if the company has more employees that have worked for longer periods of time. This is exemplified by the differences in distribution when we observe employees that have worked above 15 years in total.

As seen from the boxplot shown in Fig 41, we may observe a similar trend, where distribution of TotalWorkingYearsNew across all categories are similar, with interquartile ranges of approximately 6-15 total working years.

###### 6.1.5 Insight 5: Analysis of YearsInCurrentRoleNew

Next, we explore the distribution of employees’ YearsInCurrentRoleNew across the 4 categories of JobSatisfactionNew, to understand the relationship between YearsInCurrentRoleNew and JobSatisfactionNew, and perhaps identify any patterns of correlation. By plotting a density distribution graph as shown in Fig 42, we observe a similar distribution across all categories of JobSatisfactionNew, where there are largely 2 peaks in terms of the number of years in their current role, specifically 2 and 6 years respectively. Similar to the above analyses, the data will be more informative, in respect of employees’ JobSatisfactionNew changes, if the company has more employees that have remained in their current roles for longer periods of time.

##### 6.2 Data Modelling (CART)

Following data exploration, we also adopt the CART model on our sacrifice analysis as well. As the y-variable here is JobSatisfactionNew, which is categorical in nature, the model used is specifically a classification tree. This is likewise completed in two phases.

###### 6.2.1 Phase 1

Similar to link analysis, we construct the maximal tree (Fig 43) with our y-variable, JobSatisfactionNew, against all other x-variables using our sac\_trainset data. The minimum number of observations in each node that must exist before a split can be attempted (minsplit) is set to 12, and the complexity parameter (cp) is set to 0 to ensure we do not stop the tree growing process prematurely. This will be called sac.cart1.

We observe the maximal tree in sac\_trainset as shown in Fig 44.

###### 6.2.2 Phase 2

Similar to link analysis, we prune to simplify the overfitted tree. We plot the CV errors of each prune sequence against their respective prune triggers (cp) based on geometric mean as a chart to determine the optimal cp in the following (Fig 45).

As it is unclear to identify the optimal cp just based solely on observation, we automate the search for the optimal tree (Fig 46). First, we compute the CV error cap by identifying the minimum CV error and adding 1 standard error (xstd) to it. This will be called sac\_CVerror.cap. Next, we check all CV errors against the CV error cap, starting from the first tree onward, using a while loop. The loop will stop when the CV error is smaller than the CV error cap, to identify the optimal cp region where the CV error is just below the CV error cap in the maximal tree. Finally, we compute the geometric mean of the 2 identified cp values in the optimal region, which will be the optimal cp value used in pruning. This will be called sac\_cp.opt. In this case, the 63rd pruning sequence gives the optimal tree.

We proceed to prune the maximal tree using the sac\_cp.opt, and this will be called sac.cart2 (Fig 47). By observing the cp chart of the pruned tree (Fig 48) , we may observe that the size of the tree is slightly smaller, and there is only 1 point that has the CV error being just below the CV error cap, demarcated by the horizontal dotted line. Based on the printcp table (Fig 49), we observe that the pruned tree is left with 824 splits, the root node error is 0.6886, the train set misclassification error is 0.05359 (=0.077820 \* 0.6886), and the test set misclassification error is 0.08619 (=0.12516 \* 0.6886).

###### 6.2.3 Data Prediction

Last, we test the sac.cart2 model using our sac\_testset data (Fig 50). This is done by predicting the outcome for the JobSatisfactionNew using the model, and the following tabulates the results of the prediction against the actual JobSatisfactionNew levels in the test set data (Fig 51). The overall accuracy of the model is 0.9137261, and MonthlyIncomeNew is observed to be the most important x-variable, followed by MonthlyRateNew (Fig 52).

#### 7 Attrition

##### 7.1 Data Exploration

###### 7.1.1 Insight 1: Analysis of AttritionNew

To visualize the distribution of AttritionNew, we plotted a bar graph (Fig 53). We found that the majority class, “Current employee”, makes up 84.3% of the observations while the minority class, “Voluntary Resignation”, makes up only 15.7% of the observations. This is close to the typical 16.7% turnover rate in the finance industry (salary.com, 2018). This imbalance in the dataset may pose a challenge in fitting a model to predict AttritionNew because our model may predict only the majority class, resulting in an accuracy of over 80% even though it has no predictive ability.

###### 7.1.2 Insight 2: Analysis of JobSatisfactionNew

Next, we explore how employees’ JobSatisfactionNew can affect their decision to leave. We used a bar graph, faceted by the levels in AttritionNew to obtain (Fig 54). From the graphs, we see that those who choose to stay on with the company (“Current employee”) tend to have higher JobSatisfactionNew. We calculated that out of those who stayed, 62% had a JobSatisfactionNew rating of 3 or 4. In contrast, out of those who left (“Voluntary Resignation”), 57% had a JobSatisfactionNew rating of 3 or 4. Those who left tended to have lower JobSatisfactionNew, suggesting a possible link between lower JobSatisfactionNew and Voluntary Resignation.

###### 7.1.3 Insight 3: Analysis of YearsAtCompany

Similar to Insight 2, we explore how YearsAtCompany can affect AttritionNew. We plotted a density distribution graph of YearsAtCompany (Fig 55) and found that there is a noticeable peak in the zero to five years range for employees who chose to leave the company. This shows that out of those who leave, a larger proportion stayed for a relatively short time. This could be showing job-hopping behaviour which is becoming common (Half, 2018). Hence, there could be a relation where a longer time spent working at a company leads to lower likelihood of voluntary resignation.

###### 7.1.4 Insight 4: Mean and Median of YearsAtCompany

We also wanted to explore any correlations between the combinations of JobInvolvementNew and JobSatisfactionNew, and the number of YearsAtCompany. We did this by plotting two heatmaps, where one’s fill color was being linked to the mean of YearsAtCompany for each group (Fig 56), and the other being linked to the median of YearsAtCompany of each group.

Looking at the mean YearsAtCompany, we see that the group with highest mean years is for JobSatisfactionNew = 2, JobInvolvementNew = 3, and JobSatisfactionNew = 4, JobInvolvementNew = 1. There are two opposite groups: one group has high satisfaction but low involvement, and the other is the opposite. This shows that employees who stay longer at the company stay for different reasons (Williams & Scott, 2012) - it could be that high involvement employees enjoy greater connection with the workplace while high satisfaction employees value the achievement that their job provides.

From the median YearsAtCompany heatmap, we can see that the highest median years is for JobSatisfactionNew = 1, JobInvolvementNew = 1, and JobSatisfactionNew = 4, JobInvolvementNew = 1. This, again, seems to be two opposite groups because it suggests that those who stay at the company longer are either highly satisfied with their job, or are highly dissatisfied with their job. The group with the low satisfaction and involvement levels could be anomalous. This small group of employees (count=229) may be staying at the company because of reasons other than involvement or satisfaction such as skill growth (Accenture, 2011). Nevertheless, we included these employees because we want to learn why this group tends to stay longer.

There is a disparity between the two plots because there could be outliers within each cell. For example for JobSatisfactionNew = 2 and JobInvolvementNew = 3, there are likely to be outliers who stayed for a long time at the company, bringing the mean up. Conversely, in JobSatisfactionNew = 4 and JobInvolvementNew = 2, there are outliers who stayed for only a short time, bringing the mean down.

##### 7.2 Data Modelling (Logistic Regression)

To predict AttritionNew, we opted for the logistic regression model because it can be used to predict binary categorical variables like AttritionNew. A value of “Current employee” was coded as level 0 and “Voluntary Resignation” was coded as level 1. This is because we are more interested in predicting employees who are likely to resign so that we could intervene and reduce turnover.

###### 7.2.1 Logistic Model One (Imbalanced data)

We fit a logistic model of AttritionNew against our three predictors from the fit, link, and sacrifice components, namely YearsAtCompany, JobInvolvementNew and JobSatisfactionNew respectively (Fig 57). YearsAtCompany is a continuous variable while the other two are categorical. Thus, there are dummy variables in our logistic model for the categorical variables. The logistic model predicts by combining the independent variables linearly, much like in linear regression, then placing this into a logistic function. The result is an S-shaped curve that takes any value from 0 to 1. A threshold is then set to make a binary classification for AttritionNew.

After running the regression, we computed the odds ratio (Fig 58) and each odds ratio 95% confidence interval (Fig 59). From these, we can be confident that all the variables used are statistically significant as all the odds ratio confidence intervals are less than 1. An odds ratio smaller than 1 signifies that any of these variables increasing by one unit would decrease the predicted AttritionNew. For instance, being at JobInvolvementNew level 2 would decrease the predicted probability of AttritionNew = 1 by 0.67 times.

We also calculated the p-value of each coefficient and found that all of them were less than 0.05, and hence significant (Fig 60). For att.log1, we obtained prediction results shown in Fig 61.

Our model did not predict any “Voluntary Resignation”. This resulted in a relatively high accuracy of 0.84, which is incidentally the proportion of majority class (“Current employee”). As we are interested in predicting for the resignation, we set that as our positive case. Despite the good accuracy, our true positive rate is 0. This is likely due to the imbalanced dataset.

###### 7.2.2 Logistic Model One (Balanced data)

We resolved this using the SMOTE function found in the DMwR package. SMOTE works by over-sampling the minority class by generating new data based on their nearest neighbours and under-sampling the majority class (Chawla et. al. 2002). We over-sampled the minority class by 200%, and sampled 130% of the newly generated minority class from the majority class (Fig 62). Again, we verify the odds ratio, odds ratio confidence interval and p-value (Fig 63, 64, 65). They do not differ much from the first model, showing that our variables are still significant. The p-values are also all very small, meaning each variable is statistically significant for our prediction.

Finally, looking at the results for the new model, att.log2 (Fig 66). Although the overall accuracy is only 0.51, our true positive rate is now 0.18. As opposed to focusing on all employees, White Rock would use the predicted information to focus more on those predicted to resign, saving its resources. The 2955 false positive cases can be seen as a well-being check on the employees. We would be able to reach out to 60% of employees who intended to resign.

#### 8. Business Recommendations

##### 8.1 Solution Overview for Data Gathering

To execute the above analysis, White Rock has to obtain data to train the models. One suggested data gathering method is to ask the employees to fill up a questionnaire (Fig 67). Each employee is required to fill up the questionnaire, starting with typing their employee.ID as a unique identifier. Demographic information such as age, address (in which postal code can be used to derive distance from home) are not asked since these are data that the company usually would have. Likewise, work profiles such as years at company, business travel and monthly salary are not asked to keep the questionnaire succinct. Instead, the questionnaire focuses on gathering data which the company usually does not have such as job satisfaction. The results from the questionnaire should be similar to the IBM dataset used above, such as the levels of categorical variables, to facilitate ease of executing our analysis proposed.

##### 8.2 Solution Overview for Mitigating Flight Risk

###### 8.2.1 Explanation on How to Use Our Model

With the linear regression model predicting the employee’s degree of fitness with White Rock, CART models predicting the employee’s links and sacrificial levels with White Rock, the management of White Rock can apply the following 5-step approach to predict the employee’s overall flight risk, by means of attrition, and understand its potential cause.

First, management may input the employee’s JobRole, MonthlyIncome, TotalWorkingYears, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrManager into the linear model, fit.lm4. The model will predict the expected total number of years the employee will stay in White Rock, measured by YearsAtCompany, which serves as a proxy for measuring the degree of commonality in terms of values and beliefs the employee has with White Rock. The longer the expected total number of years, the more common values the employee and White Rock share, the greater the level of fit.

Second, management may input the employee’s Age, DailyRate, HourlyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, TotalWorkingYears, YearsInCurrentRole and YearsWithCurrManager into the CART model, link.cart2. The model will predict how involved the employee is in his job scope, measured by JobInvolvement, which serves as an indicator for the extent of integratedness the employee is with White Rock. The higher the job involvement level, the more connected the employee is with White Rock in terms of his value to White Rock, the greater the level of link.

Third, management may input the employee’s Age, DailyRate, HourlyRate, MonthlyIncome, MonthlyRate, TotalWorkingYears and YearsInCurrentRole into the CART model, sac.cart2. The model will predict the extent of sacrifice the employee will stand to lose should he leave his current job in White Rock, measured by JobSatisfaction. The higher the job satisfaction level, the greater the likelihood of the employee being preferentially remunerated and treated, the higher the opportunity cost should he choose to forsake his current job, the greater the level of sacrifice.

Fourth, with the predicted variables YearsAtCompany, JobInvolvement and JobSatisfaction used to measure fit, link and sacrifice respectively, management may input them into the logistic model, att.log2. The model will predict the attrition rate of the employee, measured by Attrition, which serves as a direct proxy to flight risk. The higher the attrition rate, the greater the employee’s flight risk.

Last, with the predicted flight risk, management may decide whether to replace the employee, or pour in resources to retain the employee. Based on the predicted results of linear and CART models (fit.lm4, link.cart2, sac.cart2), management may observe the particular component that is most significant in causing the employee to increase his flight risk. With the component identified, management may implement remedies tailored to the variables that were used to predict the component. This will be further explained in the next section.

With the remedies in place, management may reuse the models to re-establish the employee’s fit, link, sacrifice and attrition levels to ascertain if the remedies are effective in reducing the employee’s flight risk.

###### 8.2.2 Recommendations

Management of White Rock may pursue a tailored remedy which directly targets the factor that is causing a deficiency in the particular component of the job embeddedness theory. For example, suppose the employee has a predicted high level of flight risk, and management further observes that his fit level is significantly lower than link and sacrifice levels, thus increasing his attrition risk. Instead of implementing a universal approach in its employee retention efforts like raising a fixed amount of his income, management may adopt a more targeted approach based on the significant factors that affect the employee’s YearsAtCompany, or his level of fit. For instance, management may improve the employee’s YearsInCurrentRole and YearsSinceLastPromotion by promoting him to a more senior position. This effectively resets both of these variables to 0 and changes their JobRole level. Management can then re-establish the employee’s fit level by inputting the revised variables into the linear model, fit.lm4. This targeted approach thus ascertains and reassures management that its retention efforts are effective in reducing employee’s flight risk.

##### 8.3 Limitations

While we are confident in our models’ ability to predict flight risk, there may be some limitations in implementing our model. We recognize that White Rock only has one data scientist and expecting him to analyze such a huge data set and run four models on this data set can be highly taxing. This is mainly because only the data scientist has the knowledge to understand how to import the data, run the models and interpret the statistics for the data. If the data scientist has other projects to manage, running these models will be highly time consuming and may result in him being overworked. As such, we recommend that the data scientist trains Human Resources (HR) employees to perform the basic processes such as importing data and running the data through the models so that, if necessary, he will only be required to interpret certain statistics.

We are also aware that the model may be ineffective in cases where anomalies exist. For example, an employee with outstanding results for fit, link and sacrifice may still decide to leave the company even though the models suggest otherwise. This is mainly because the flight risk predictions derived from the models only take into account variables based on the Job Embeddedness Theory and fail to consider many other factors such as family situation, natural disasters, crime rates, etc. As such, the prediction may not be a complete representation of the employee’s actual flight risk as external circumstances cannot be captured in the company’s system and hence cannot be accounted for.

Finally, it is important to note that our models have been trained on a hypothetical dataset sourced from IBM (Kaggle, 2017), a technology and consulting multinational form headquartered in America. This means that the insights may be influenced by American workplace culture and social norms in terms of variables such as MonthlyIncome, JobRole, JobInvolvement and so on. White Rock should certainly re-train these models on their own dataset which will be more representative of their workforce.

#### 9. Conclusion

To conclude, our analytics models provide a data-driven approach to cost-effective HR management and reduced Attrition rates. We accomplish this by first obtaining a complete, cleaned dataset. Then, we performed preliminary data exploration to decide our predictor variables and predicted variables for each component of the Job Embeddedness Theory. For Fit, we predicted years stayed in the company; For Link, job involvement level; For Sacrifice, job satisfaction level. These models allow us to pre-emptively address any deficiencies with a given employee’s overall embeddedness in a targeted manner, even allowing us to pinpoint the exact variable to address. Our final model which predicts voluntary resignation serves as a tool for management to find employees at higher risk of leaving the company, and implement measures to retain them, if desired. Because of the use of data analytics, our solution overcomes human’s natural biases and provides accurate and timely information for better decision-making.

All in all, the data analysis would help White Rock reduce turnover and hence decrease turnover costs, reduce manpower and time spent on employee retention, and improve the overall well-being of their employees. This could help improve employee productivity, and improve the overall quality of White Rock’s services.

#### 10. References

Accenture. (2011, March 4). *International Women’s Day Accenture research shows that women and men around the globe share similar career perspectives*. Accenture. https://newsroom.accenture.com/subjects/accenture-corporate/despite-low-job-satisfaction-employees-unlikely-to-seek-new-jobs-accenture-research-reports-prefer-to-focus-on-creating-opportunities-with-current-employers.htm

Bellet, C., Neve, J. D., & Ward, G. (2019, October 17). Does Employee Happiness have an Impact on Productivity? *Said Business School WP 2019-13*, 45. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3470734#

Boehmke, B., & Greenwell, B. (2020). Decision Trees. In *Hands-On Machine Learning With R* (p. 7). Github. https://bradleyboehmke.github.io/HOML/DT.html

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321-357.

Globe Newswire. (2020, August 12). *Even in a Pandemic, Companies Still Struggle with Employee Turnover*. GlobeNewswire News Room. Retrieved October 29, 2020, from https://www.globenewswire.com/news-release/2020/08/12/2077241/0/en/Even-in-a-Pandemic-Companies-Still-Struggle-with-Employee-Turnover.html

Half, R. (2018, April 5). *Does Job Hopping Help Or Hurt Your Career?* roberthalf. Retrieved October 29, 2020, from http://rh-us.mediaroom.com/2018-04-05-Does-Job-Hopping-Help-Or-Hurt-Your-Career?utm\_campaign=Press\_Release

Heinz, K. (2020, March 22). *Dangers Of Turnover: Battling Hidden Costs*. Builtin. Retrieved October 29, 2020, from https://builtin.com/recruiting/cost-of-turnover

Kaggle. (2017). *IBM HR w/more Rows*. kaggle. Retrieved November 29, 2020, from https://www.kaggle.com/dgokeeffe/ibm-hr-wmore-rows

Kim & Kang. (2017, October). *Older and More Engaged: The Mediating Role of Age-Linked Resources on Work Engagement*. Retrieved October 29, 2020, from https://www.researchgate.net/publication/295455164\_Older\_and\_More\_Engaged\_The\_Mediating\_Role\_of\_Age-Linked\_Resources\_on\_Work\_Engagement

Kumar, V. (2020). Classification: Basic Concepts, Decision Trees, and Model Evaluation. In *Introduction to Data Mining* (p. 61). University of Minnesota. Retrieved October 29, 2020, from https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf

Lee, T. W., Burch, T. C., & Mitchell, T. R. (2014, March). The Story of Why We Stay: A Review of Job Embeddedness. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*, 199-216. https://www.annualreviews.org/doi/full/10.1146/annurev-orgpsych-031413-091244

Skapinker, M. (2020, April 30). *Why people stay with the same company for decades*. Retrieved October 29, 2020, from https://www.ft.com/content/3ff770d6-6815-11e9-9adc-98bf1d35a056

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, *14*(4), 323-348. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927982/

Williams, D. K., & Scott, M. M. (2012, October 18). *10 Reasons to Stay in a Job for 10 Years*. HBR. https://hbr.org/1973/07/why-employees-stay

Young, J. A., Stone, J., Aliaga, O., & Shuck, B. (2013, August). *Job Embeddedness Theory: Can It Help Explain Employee Retention Among Extension Agents?* Journal of Extension. Retrieved October 29, 2020, from https://joe.org/joe/2013august/a7.php#:~:text=Overview%20of%20Job%20Embeddedness%20Theory,the%20job%2C%20organization%2C%20and%20community

#### 11. Appendix

Table 1: Illogical values

|  |  |
| --- | --- |
| **Illogical Values** | **Treatment** |
| YearsAtCompany > TotalWorkingYears | Dropped 6 rows |
| YearsInCurrentRole > TotalWorkingYears | 0 rows found |
| YearsInCurrentRole > YearsAtCompany | 0 rows found |
| YearsWithCurrManager > YearsAtCompany | 0 rows found |
| TotalWorkingYears > Age-18  (Since Over18 columns are all Yes values, we assume that employees’ earliest working age is 18) | Dropped 2795 rows |
| HourlyRate > DailyRate | 0 rows found |
| DailyRate > MonthlyRate | 0 rows found |

Table 2: Setting appropriate data types

|  |  |
| --- | --- |
| **Data Type** | **Variables** |
| Integer | DistanceFromHome, HourlyRate, JobSatisfaction, MonthlyIncome, PercentSalaryHike |

Table 3: List of categorical and continuous variables

|  |  |
| --- | --- |
| **Types of Variable** | **Variables** |
| Categorical | Attrition, BusinessTravel, Department, Education, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, OverTime, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, WorkLifeBalance |
| Continuous | DistanceFromHome, HourlyRate, JobSatisfaction, MonthlyIncome, PercentSalaryHike |

Table 4: Data cleaning for each column

|  |  |
| --- | --- |
| **Variables** | **Replaced With (Mode/Mean)** |
| Attrition | * Blank fields are replaced with mode * Rows where Attrition = “Termination” are dropped as our problem statement relates to flight risk (i.e. current employees and voluntary resignation) * “Termination” level is dropped subsequently |
| BusinessTravel | * Blank fields are replaced with mode * Levels with 0 rows are dropped |
| DailyRate | * NA values are replaced with mean |
| Department | * Blank fields are replaced with mode * Unreasonable levels of “1296” and an empty level which have 0 rows are dropped |
| DistanceFromHome | * NA values are replaced with mean |
| Education | * NA values are replaced with mode * Level of “6” has 0 rows but seems reasonable as previous levels range from 1 to 5, so not dropped |
| EnvironmentSatisfaction | * NA values are replaced with mode * Levels of “127249” and “129588” are unreasonable and dropped |
| Gender | * Blank values are replaced with mode * Levels of “1” and “2” are inconsistent and dropped |
| HourlyRate | * NA values are replaced with mean |
| JobInvolvement | * NA values are replaced with mode * Levels of “47” and “54” are unreasonable and dropped |
| JobLevel | * NA values are replaced with mode |
| JobRole | * Blank fields are replaced with mode * Levels of “4” and “5” are unreasonable and dropped |
| JobSatisfaction | * NA values are replaced with mode |
| MaritalStatus | * Blank fields are replaced with mode * Level of 4 is unreasonable and dropped |
| MonthlyIncome | * NA values are replaced with mean |
| MonthlyRate | * NA values are replaced with mean |
| NumCompaniesWorked | * NA values are replaced with mean |
| OverTime | * Blank fields are replaced with mode * Level of “Y” is unreasonable and dropped |
| PercentSalaryHike | * NA values are replaced with mean |
| PerformanceRating | * NA values are replaced with mode * Level of “11” and “13” are unreasonable and dropped |
| RelationshipSatisfaction | * NA values are replaced with mode |
| StockOptionLevel | * NA values are replaced with mode * Level of “80” is unreasonable and dropped |
| TrainingTimesLastYear | * NA values are replaced with mean |
| WorkLifeBalance | * NA values are replaced with mode |
| YearsInCurrentRole | * NA values are replaced with mean |
| YearsSinceLastPromotion | * NA values are replaced with mean |
| YearsWithCurrManager | * NA values are replaced with mean |

Fig 5: Correlation Matrix

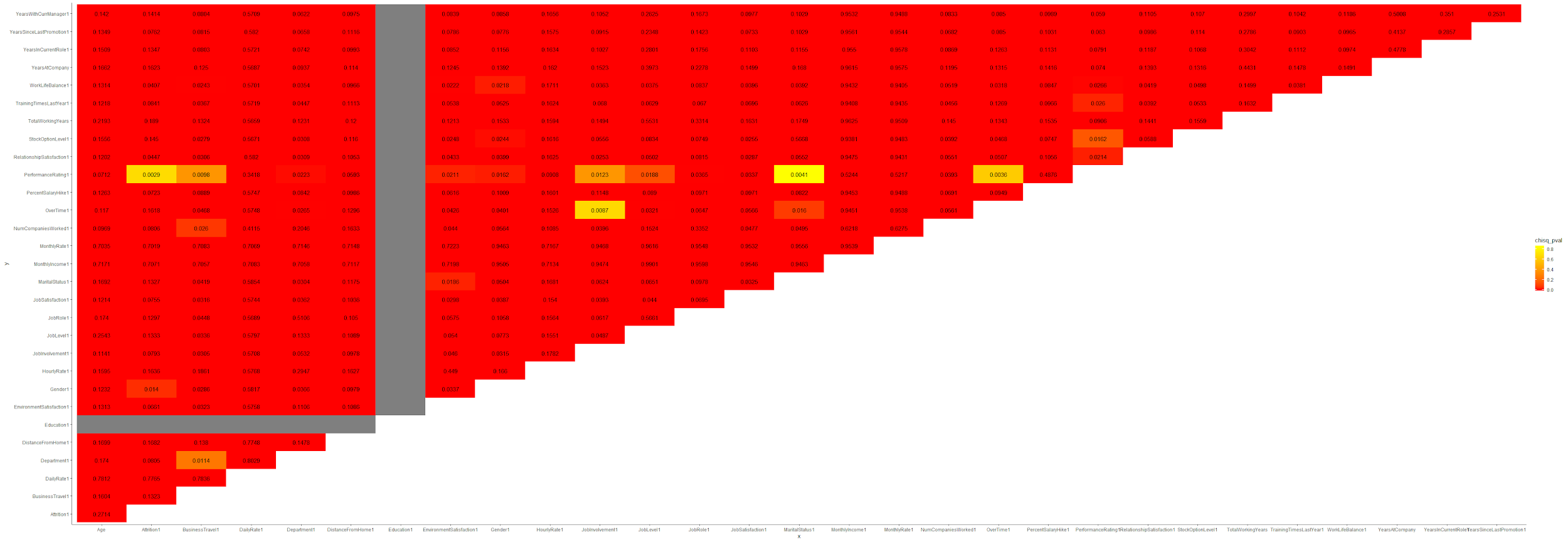
****

Table 6: Insights from correlation matrix

|  |  |  |
| --- | --- | --- |
| **Component** | **Y-Variable** | **Closely Related Variables (exclude 3 Y-Variables)** |
| Fit | YearsAtCompany | DailyRate, JobLevel, JobRole, MonthlyIncome, MonthlyRate, TotalWorkingYears, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager |
| Link | JobInvolvement | Age, HourlyRate, DailyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, TotalWorkingYears, YearsInCurrRole, YearsWithCurrManager |
| Sacrifice | JobSatisfaction | Age, DailyRate, HourlyRate, MonthlyIncome, MonthlyRate, TotalWorkingYears, YearsInCurrentRole |

Fig 7: Analysis of YearsAtCompany

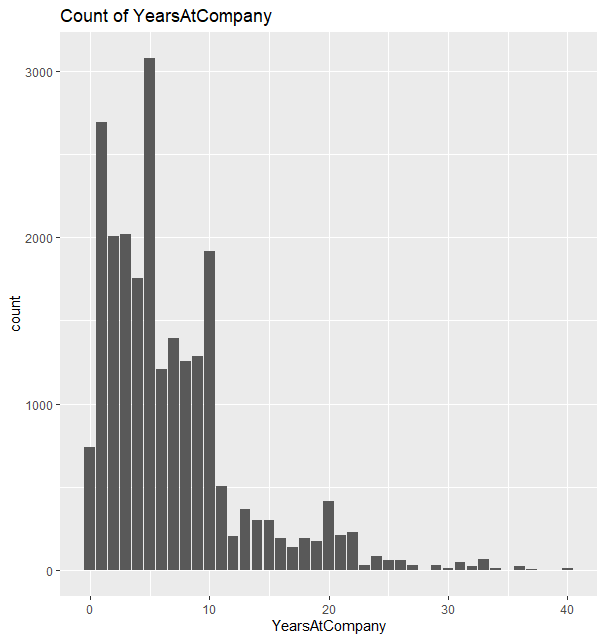




Fig 8: Analysis of JobLevelNew

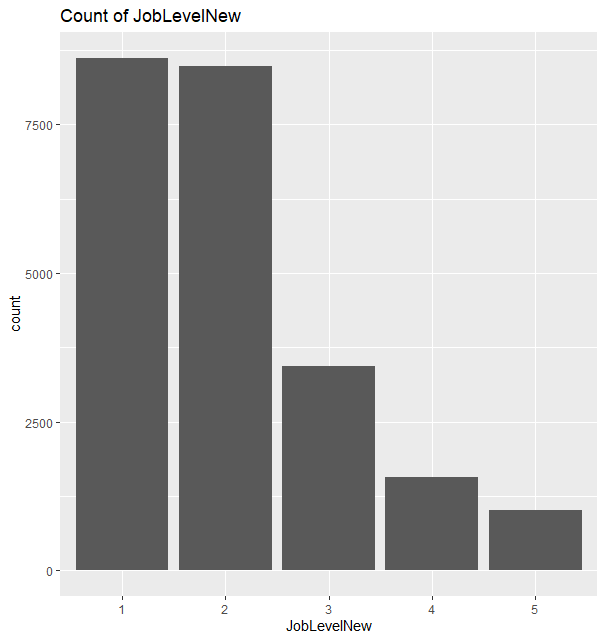




Fig 9: Analysis of JobLevelNew

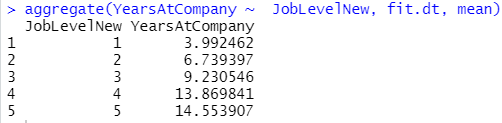
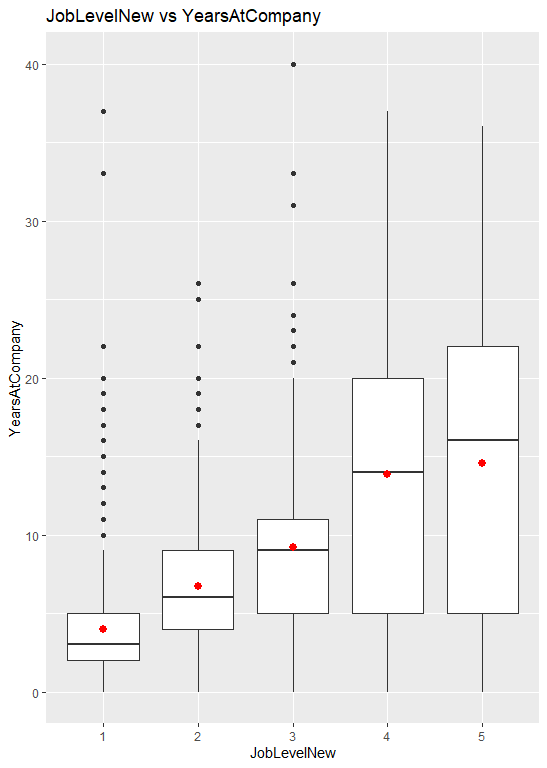


Fig 10: Analysis of MonthlyIncomeNew

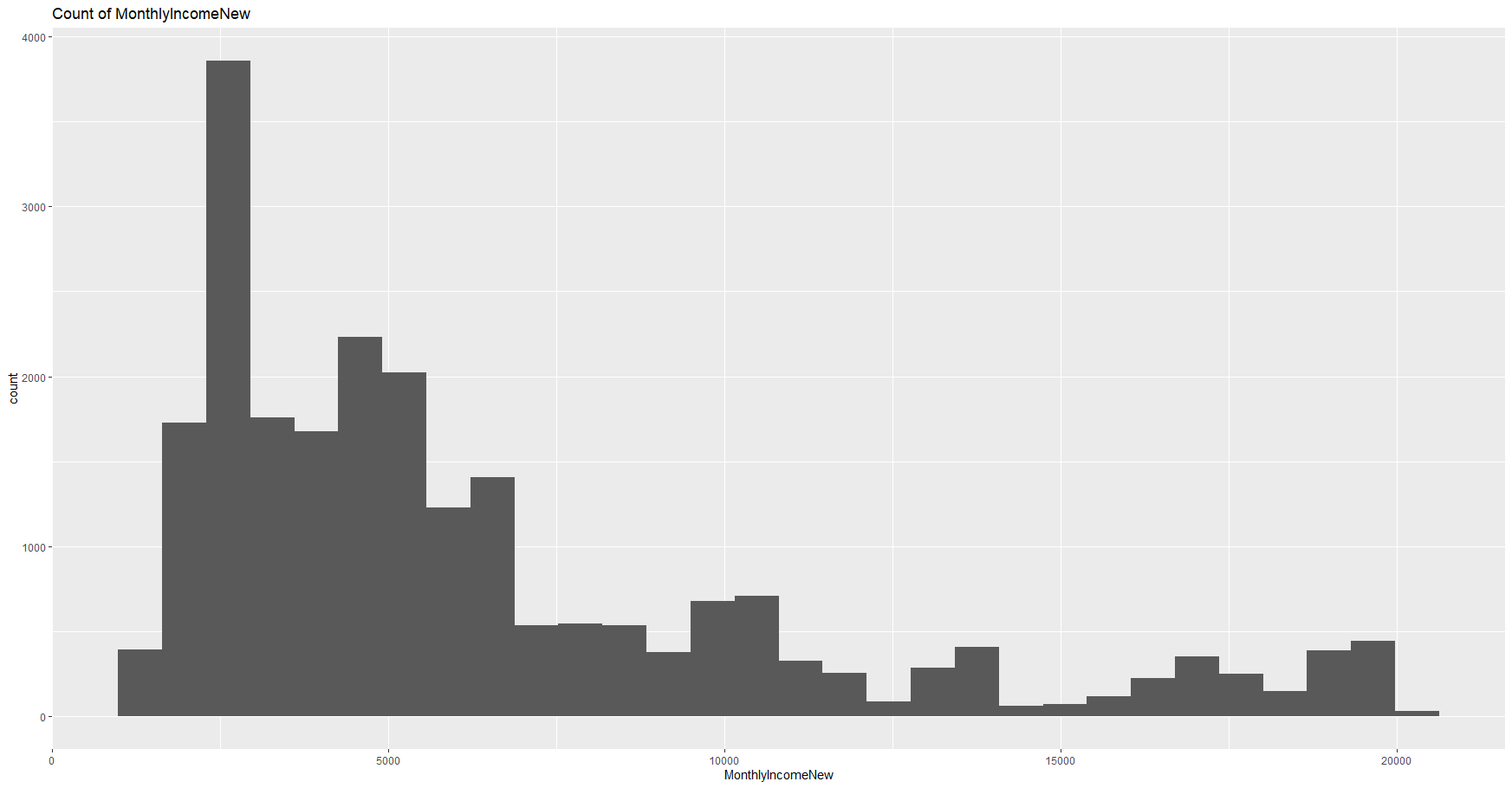


Fig 11: Analysis of MonthlyIncomeNew

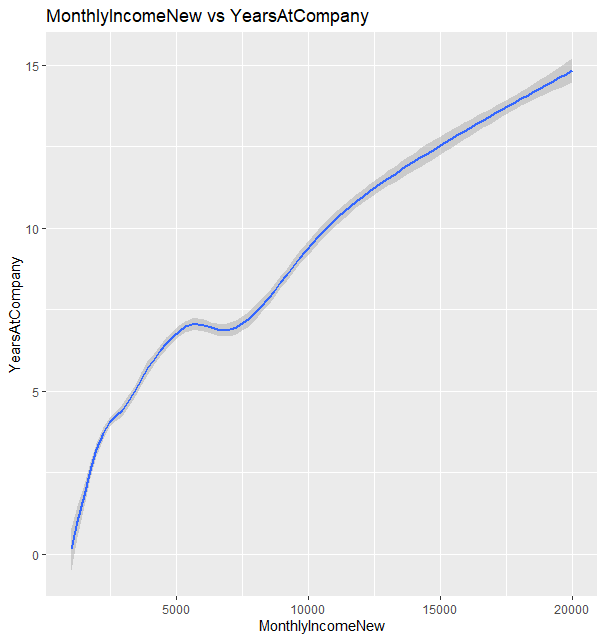


Fig 12: Analysis of JobRolesNew

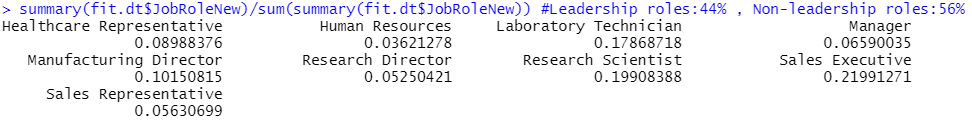
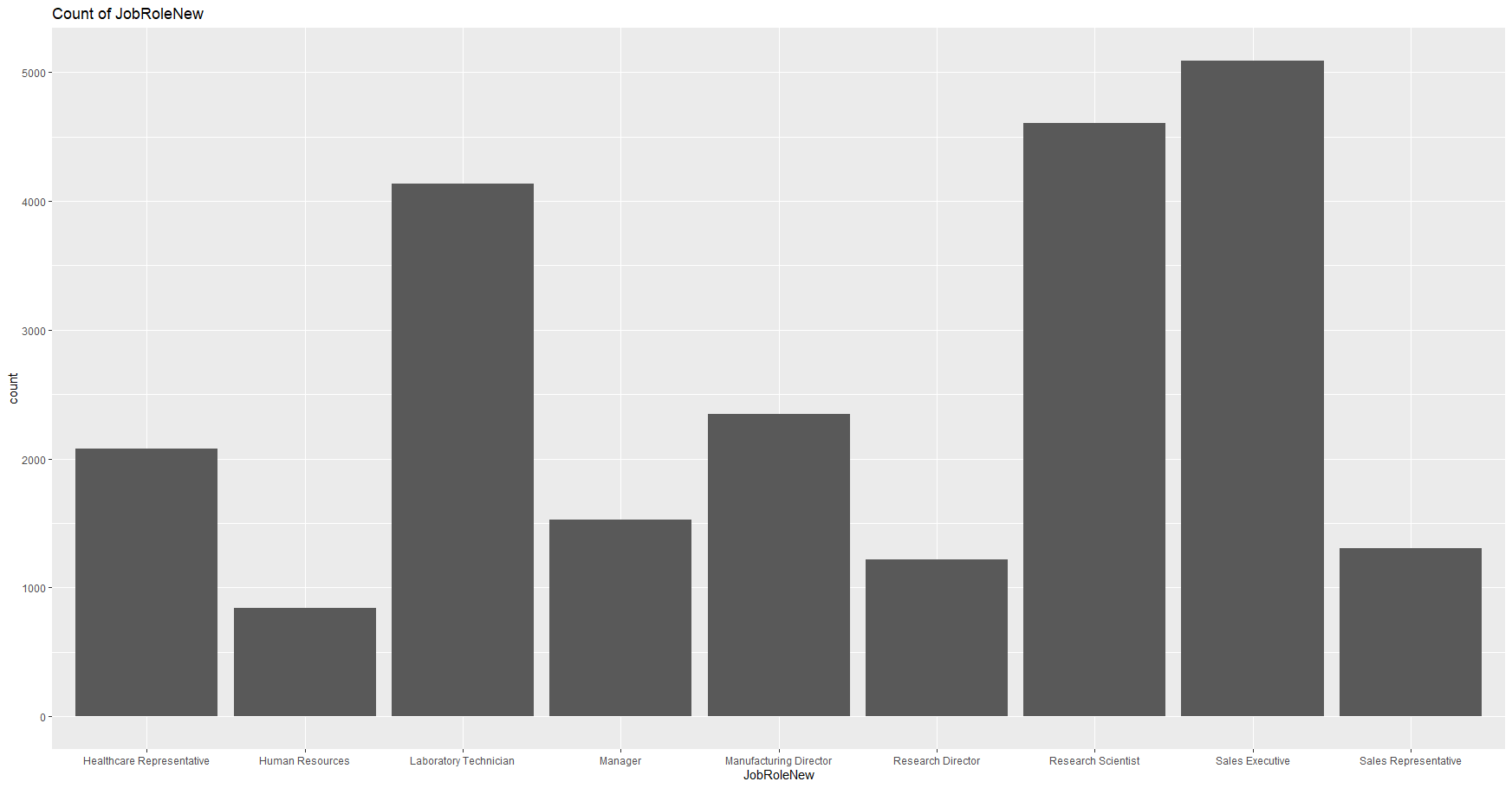


Fig 13: Analysis of JobRolesNew

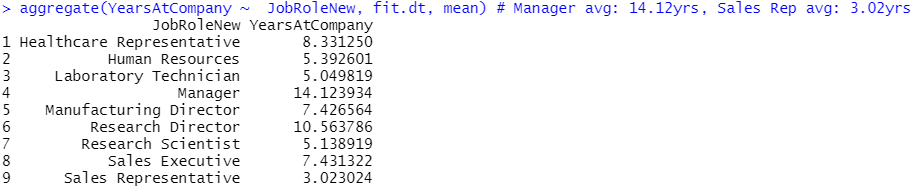
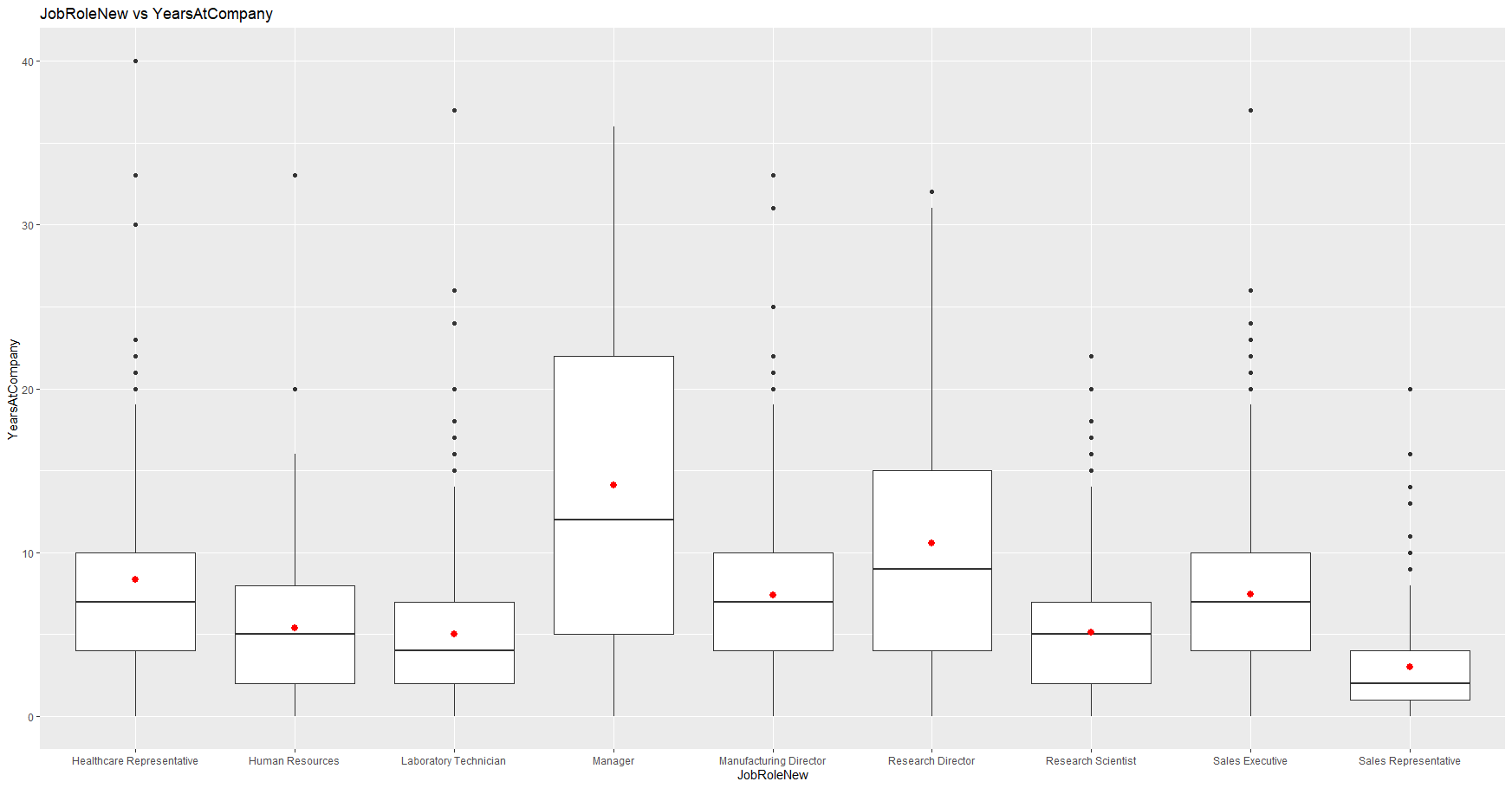


Fig 14: Analysis of fit.lm1

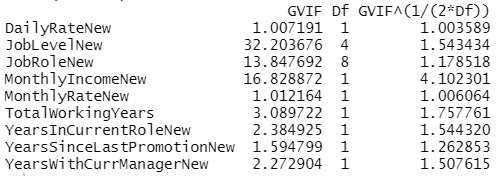
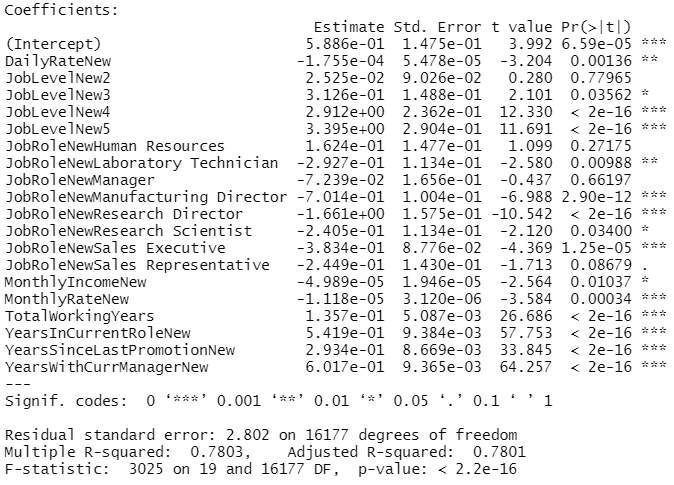
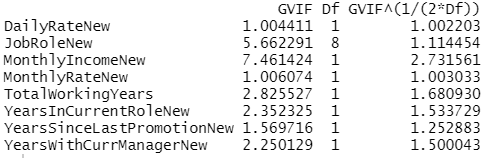
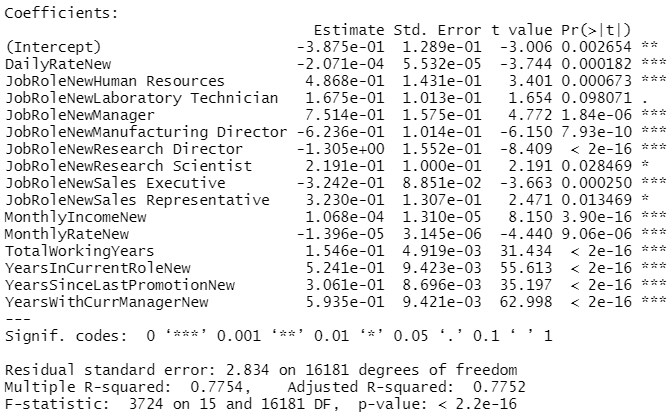
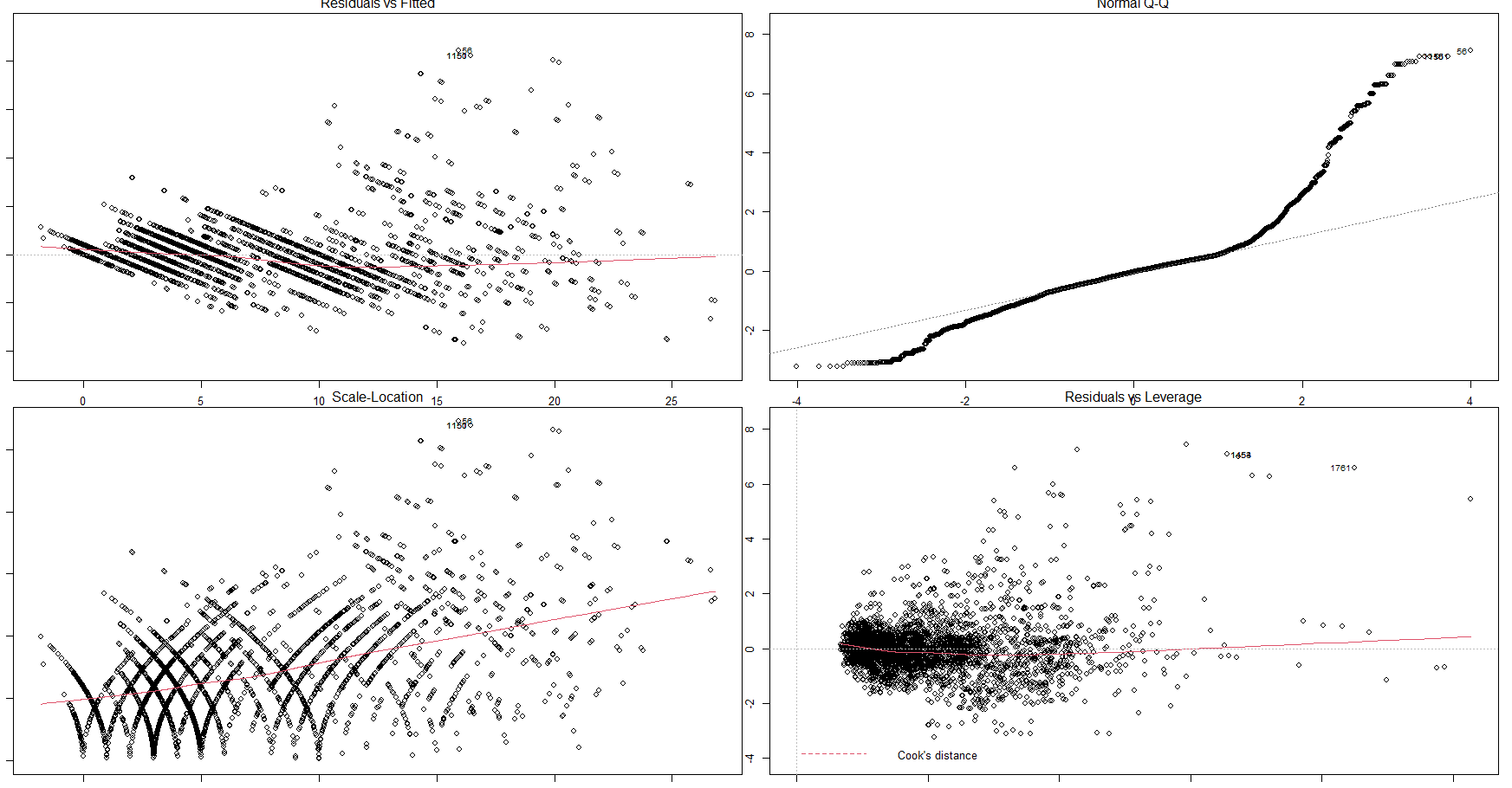


Fig 15: Analysis of fit.lm2





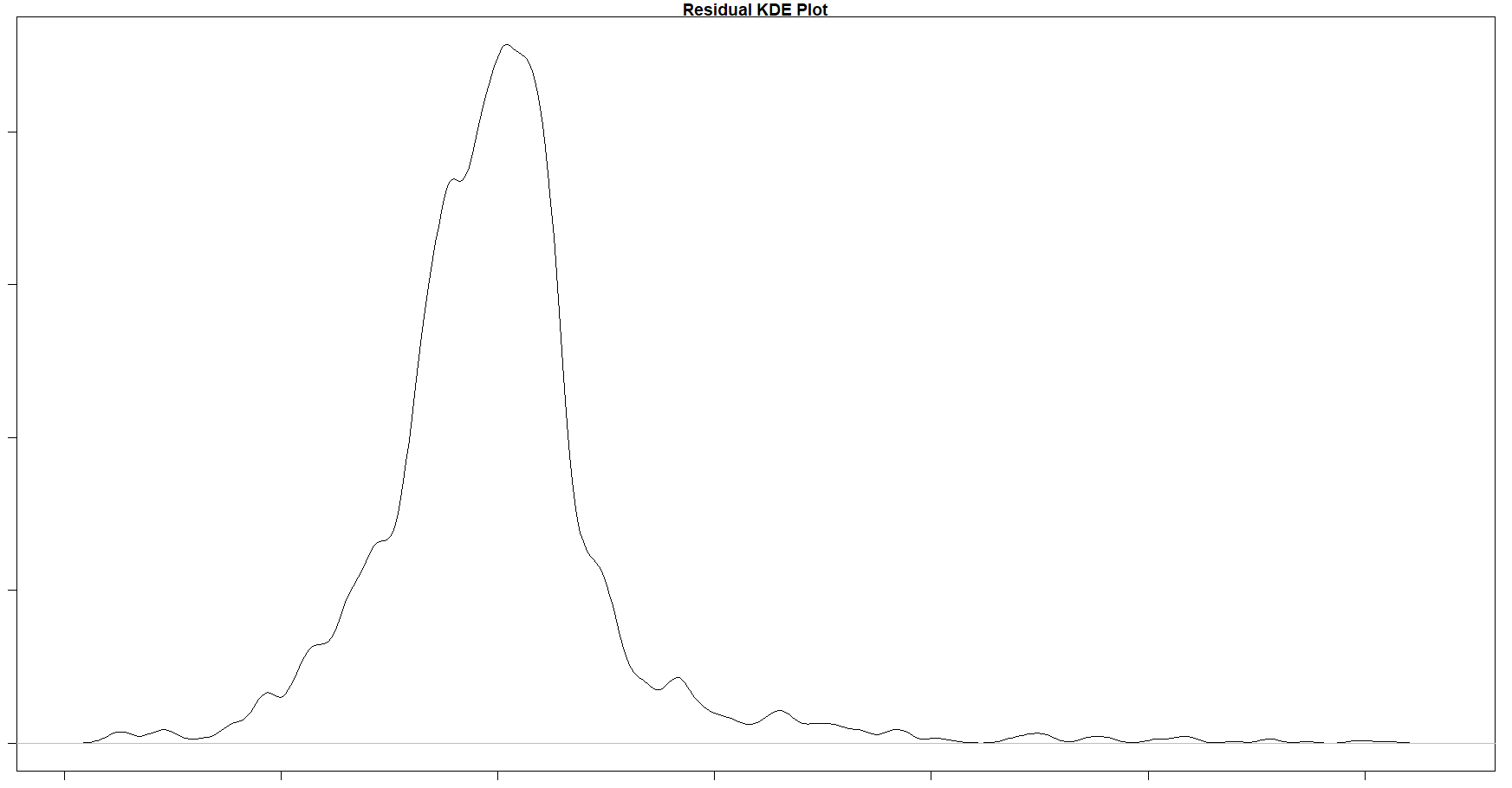


Fig 16: Analysis of fit.lm3

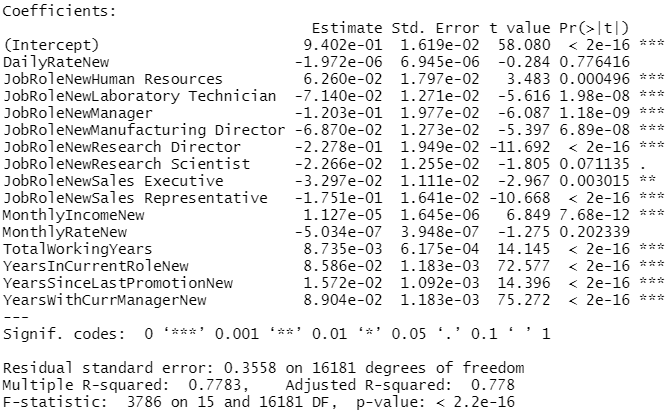
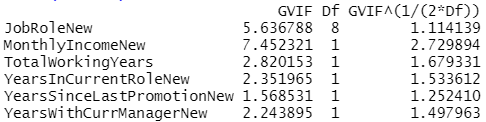
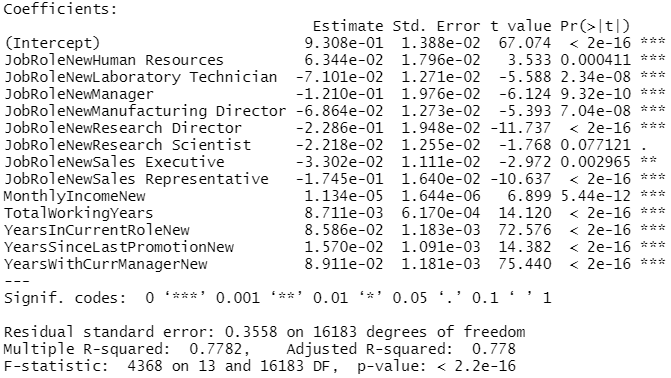
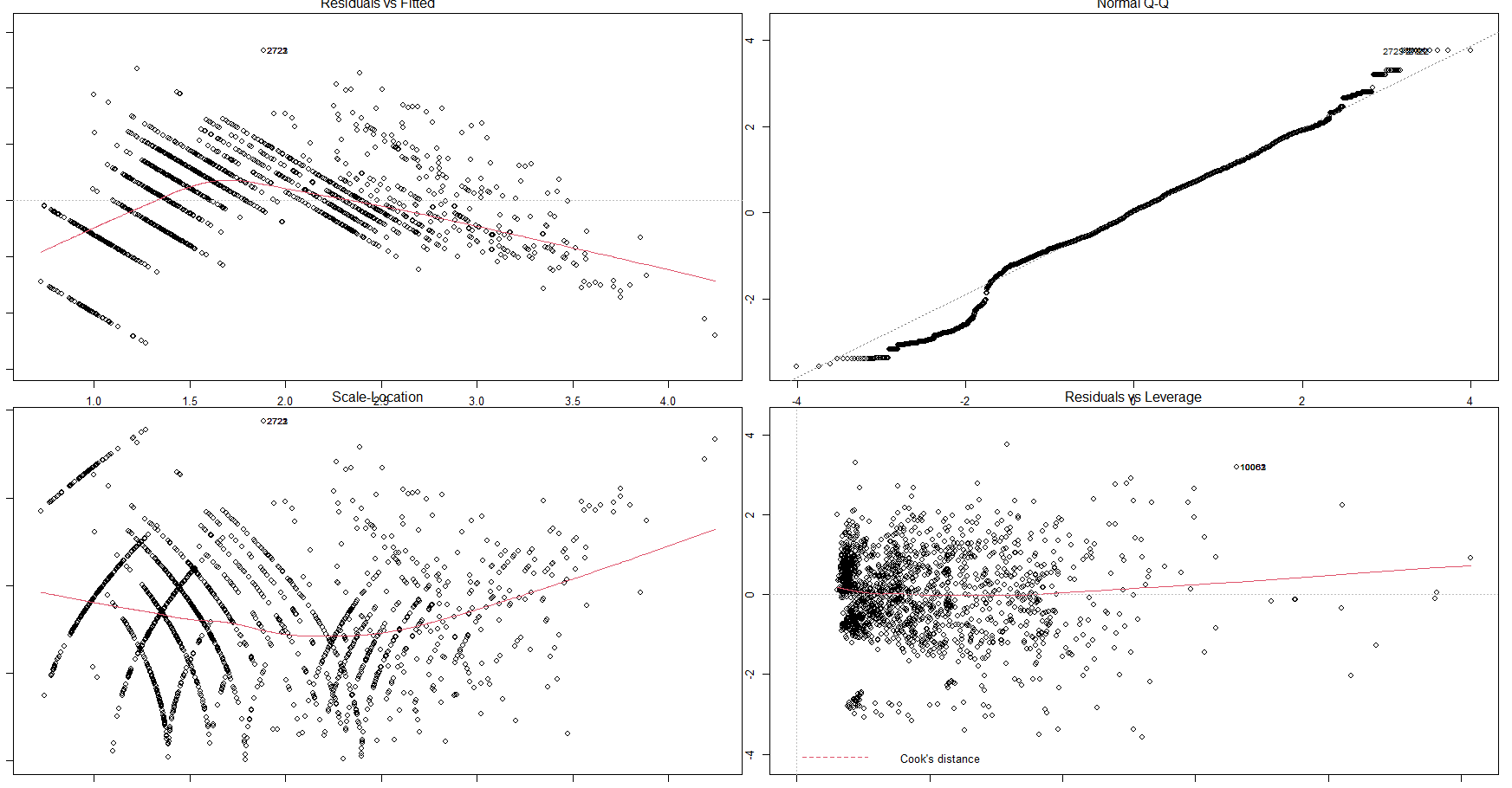


Fig 17: Analysis of fit.lm4





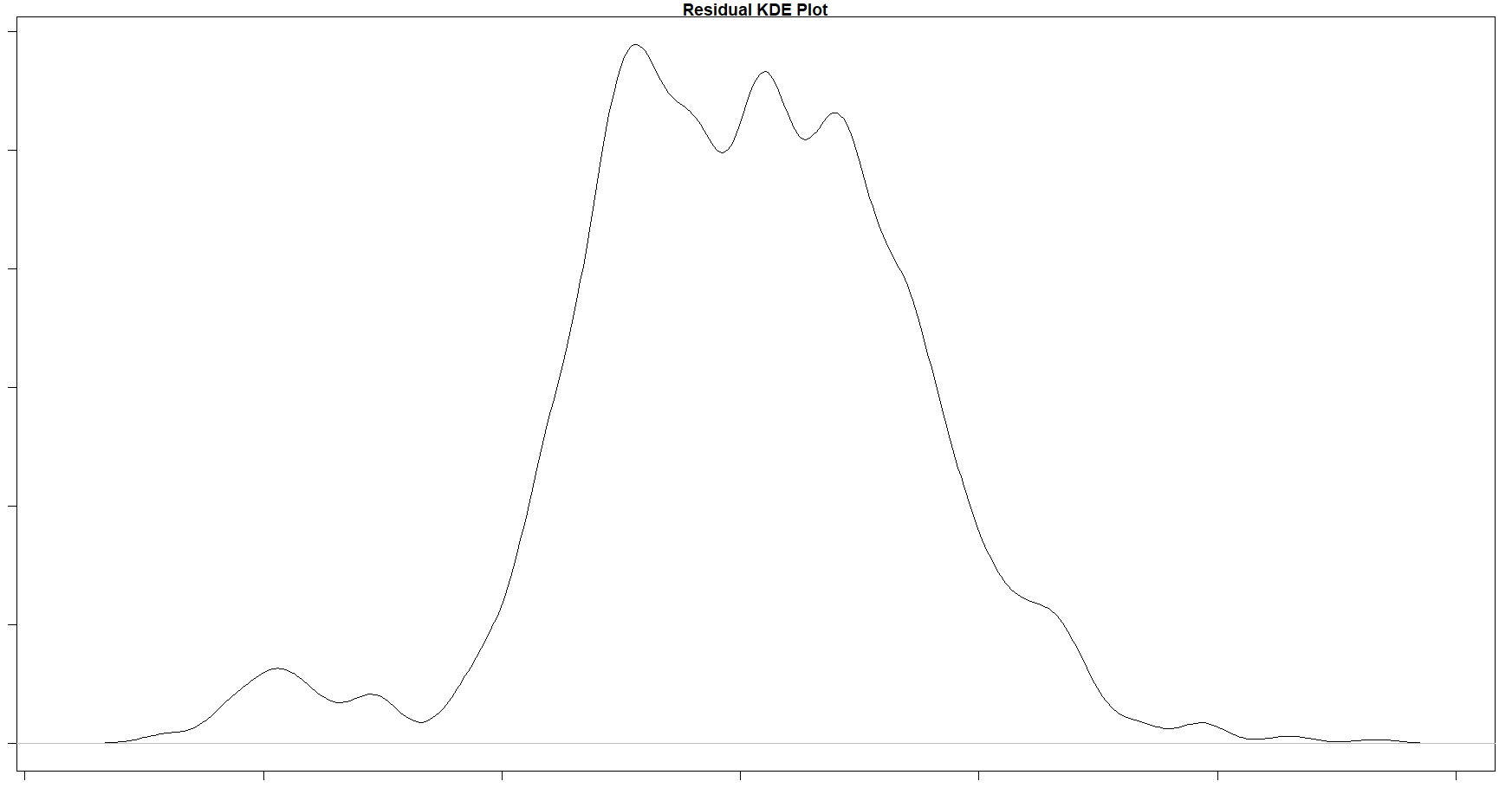


Fig 18: Fit trainset & testset RMSE

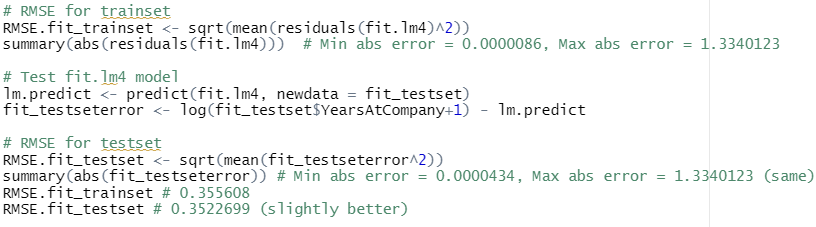


Fig 19: Analysis of JobInvolvementNew

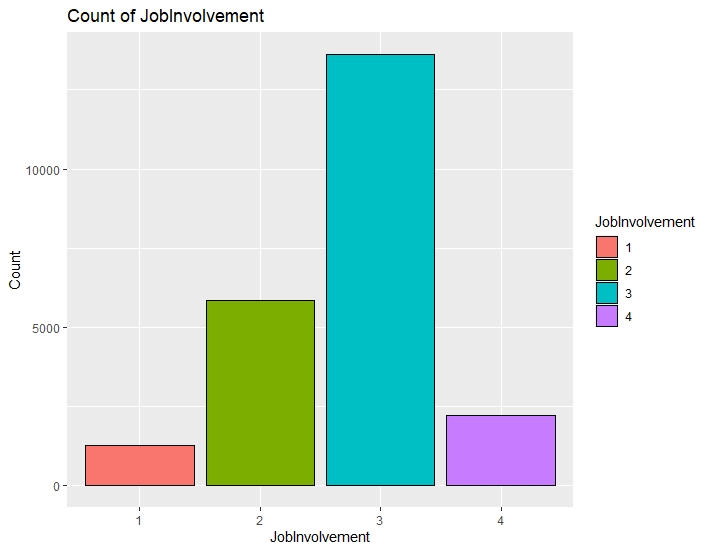
****

Fig 20: Analysis of Age

****

Fig 21: Analysis of MonthlyIncomeNew

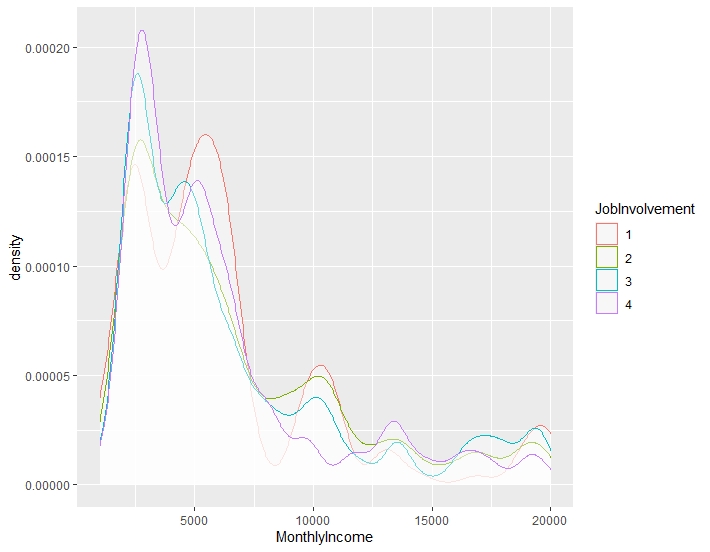
****

Fig 22: Analysis of MonthlyIncomeNew

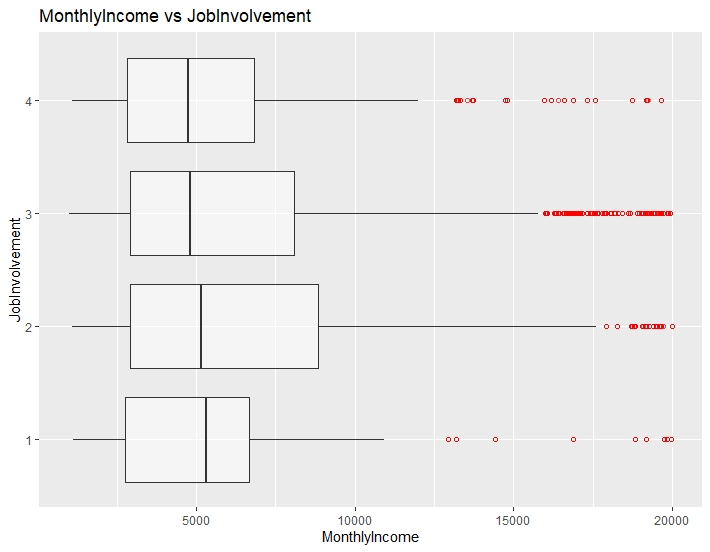
****

Fig 23: Analysis of PercentSalaryHikeNew

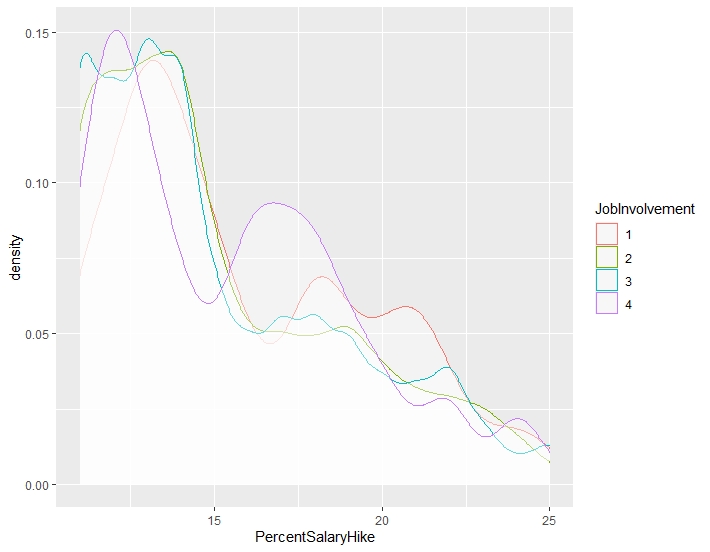
****

Fig 24: Analysis of PercentSalaryHikeNew

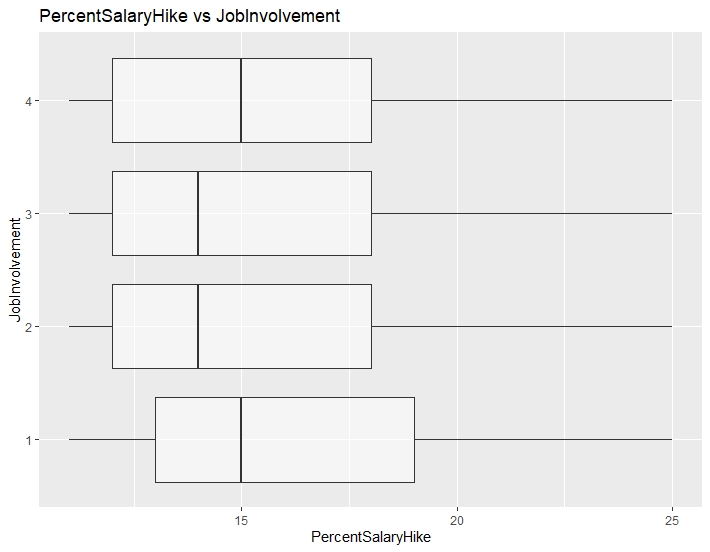
****

Fig 25: Analysis of YearsInCurrentRoleNew

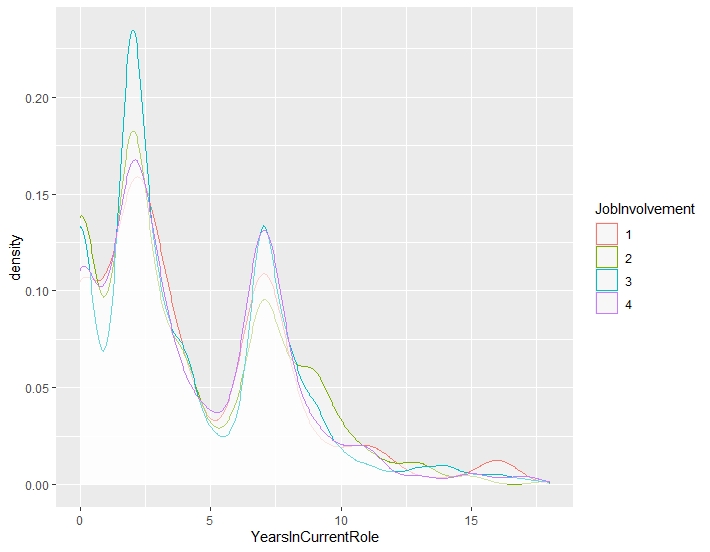
****

Fig 26: Growing the maximal tree



Fig 27: Maximal tree in link\_trainset



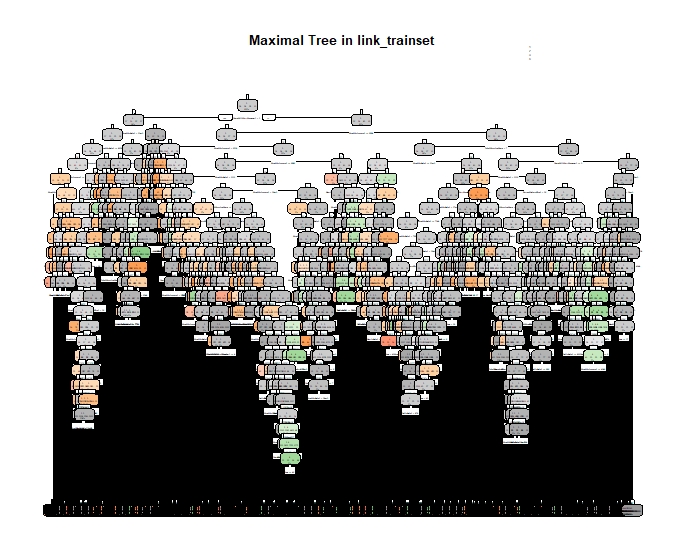


Fig 28: Subtrees in link\_trainset



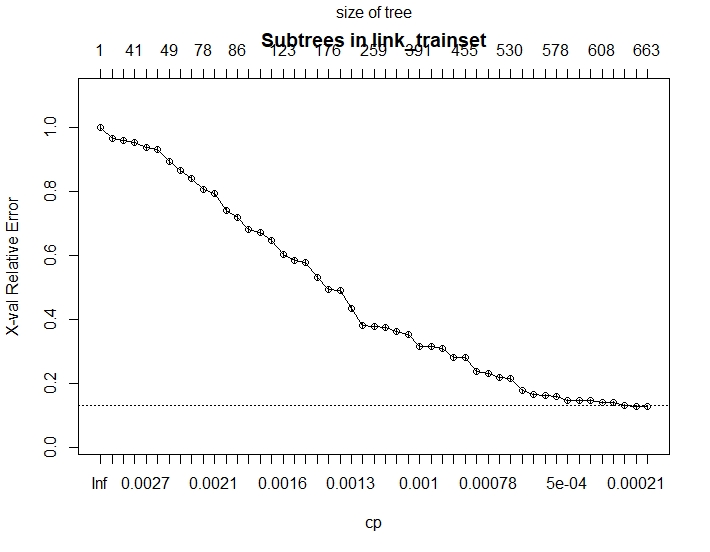


Fig 29: Automate the search for optimal tree

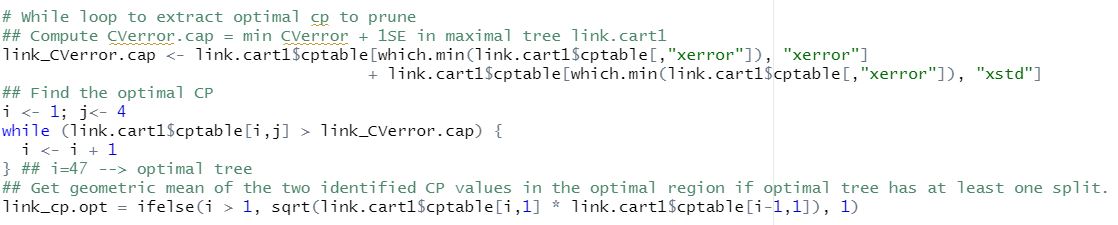


Fig 30: Pruning to its optimal tree

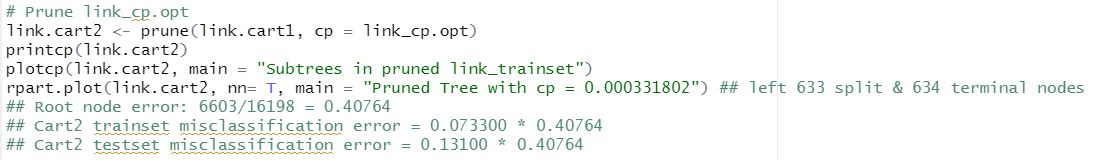


Fig 31: Subtrees in pruned link\_trainset

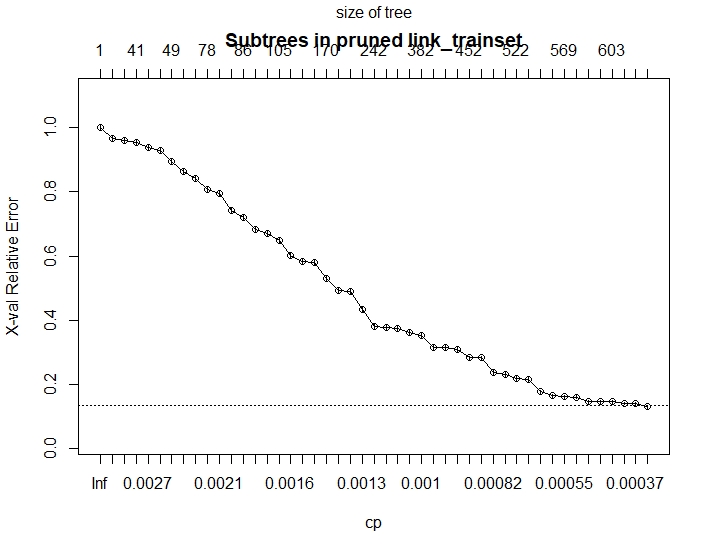


Fig 32: Printcp table for link.cart2

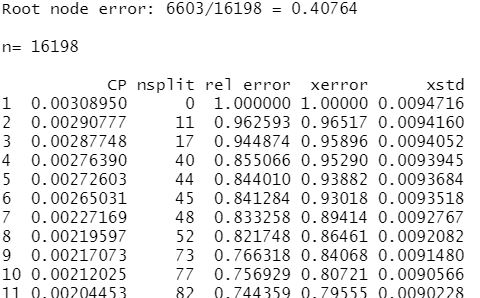


Fig 33: Testing link.cart2

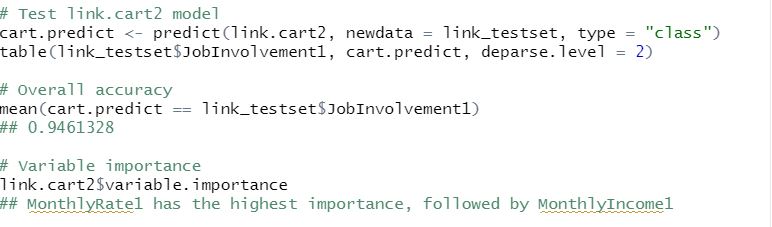


Fig 34: Prediction results for link.cart2

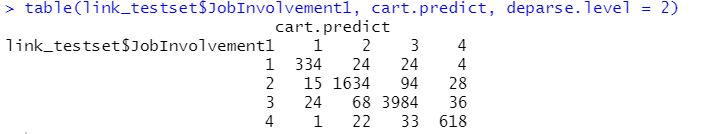


Fig 35: Variable importance for link.cart2

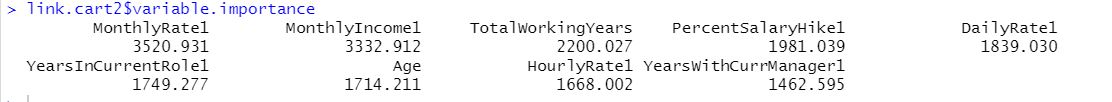


Fig 36: Analysis of JobSatisfactionNew

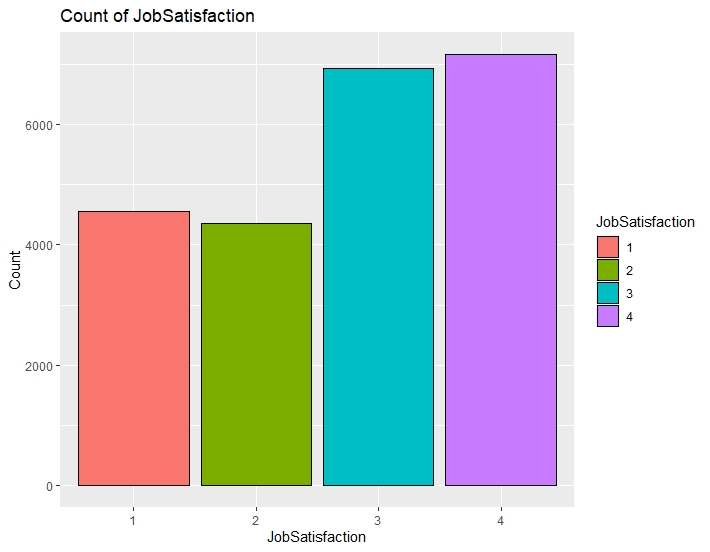
****

Fig 37: Analysis of HourlyRateNew

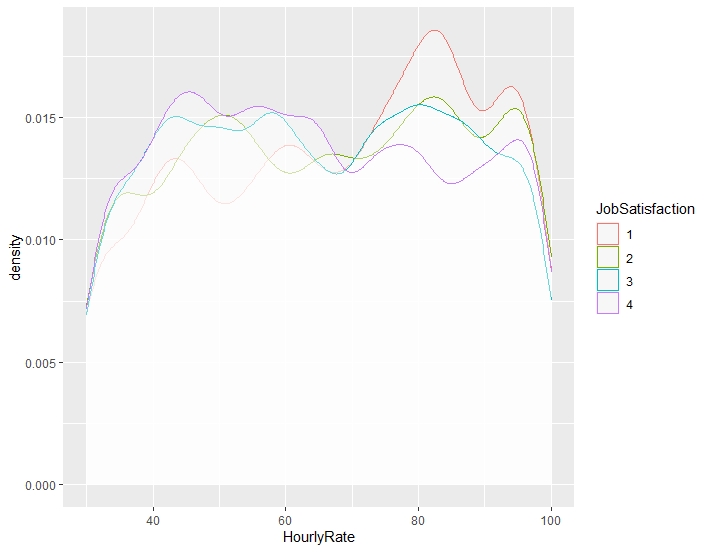
****

Fig 38: Analysis of MonthlyIncomeNew

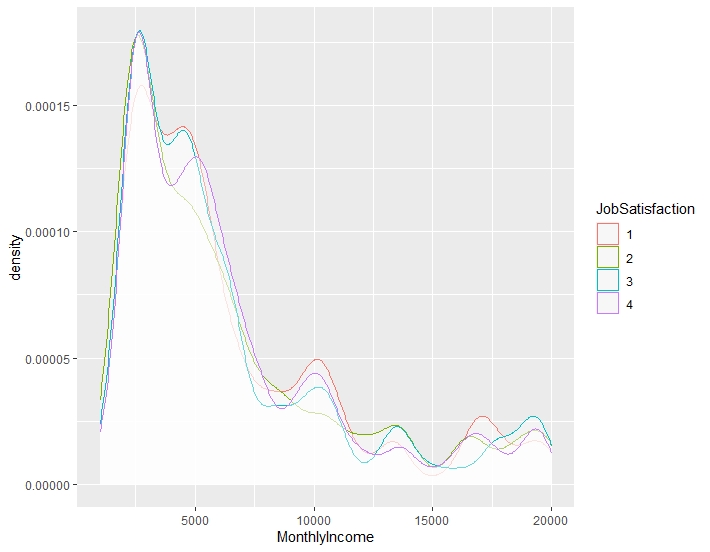
****

Fig 39: Analysis of MonthlyIncomeNew

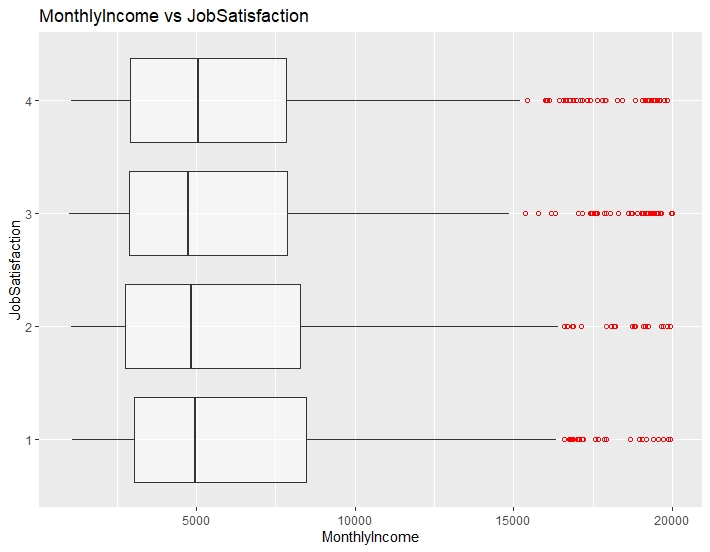
****

Fig 40: Analysis of TotalWorkingYearsNew

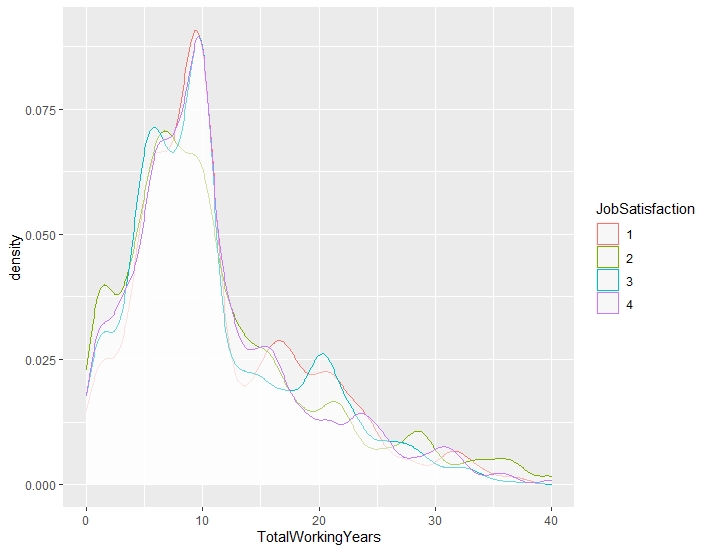
****

Fig 41: Analysis of TotalWorkingYearsNew

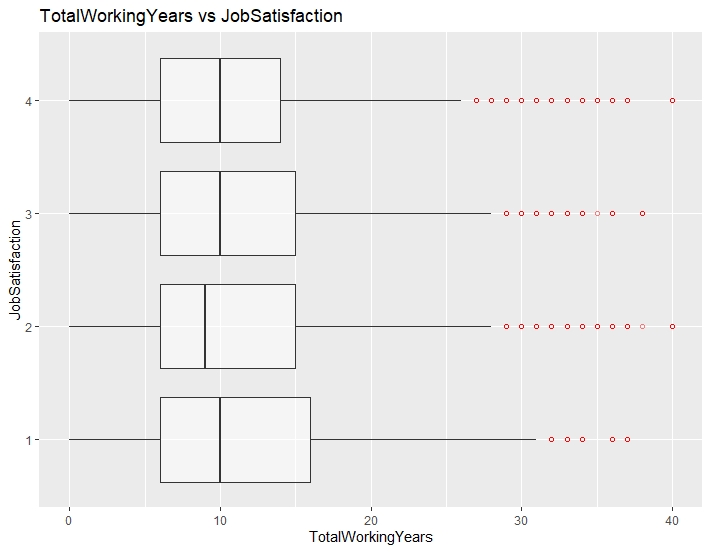
****

Fig 42: Analysis of YearsInCurrentRoleNew

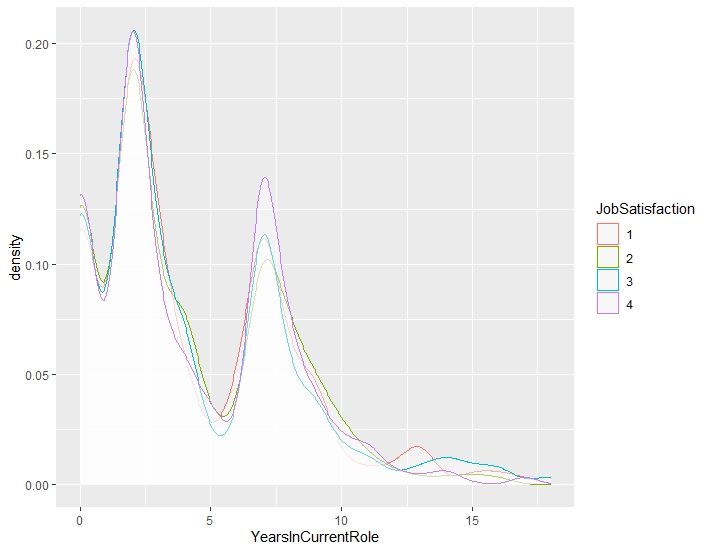
****

Fig 43: Growing the maximal tree

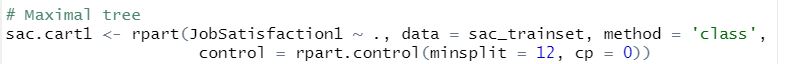


Fig 44: Maximal tree in sac\_trainset



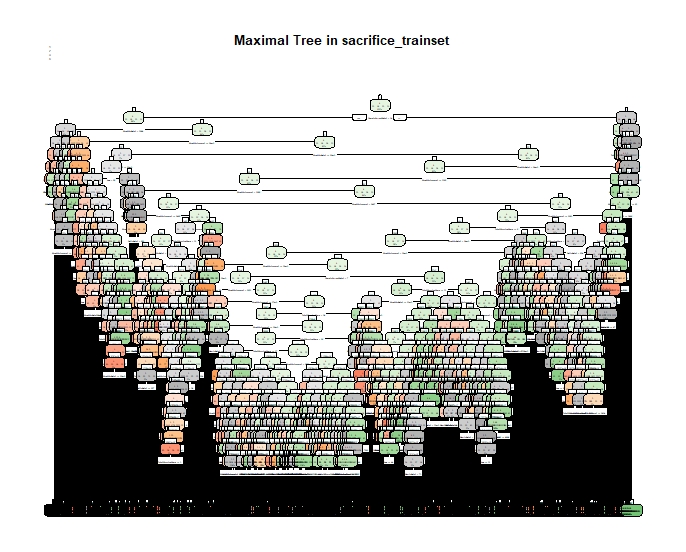


Fig 45: Subtrees in sac\_trainset



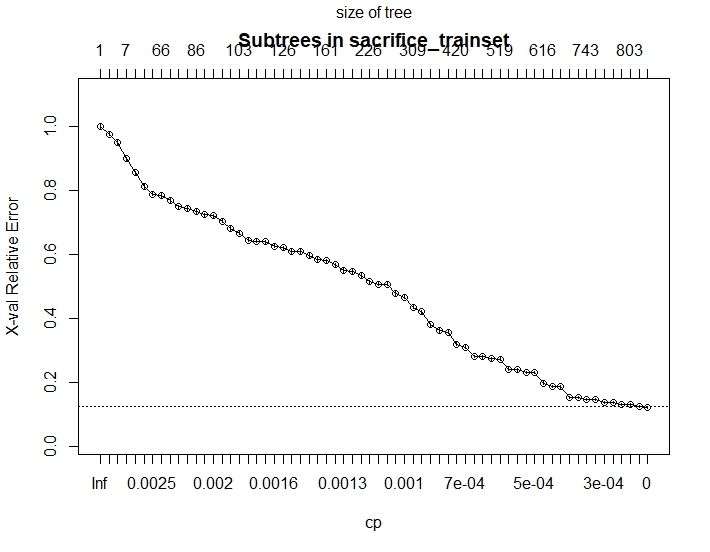


Fig 46: Automate the search for optimal tree

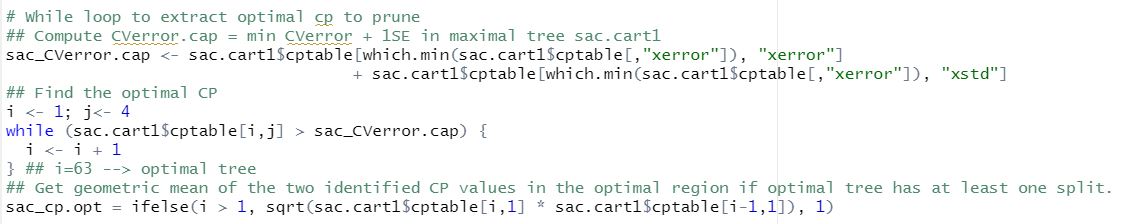


Fig 47: Pruning to its optimal tree

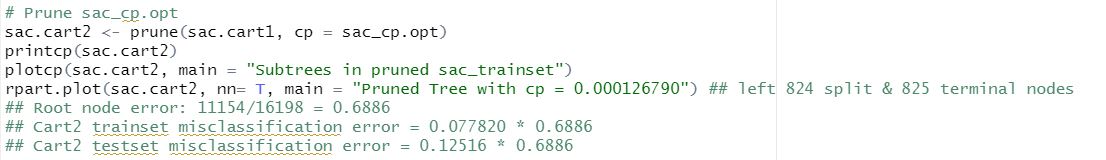


Fig 48: Subtrees in pruned sac\_trainset

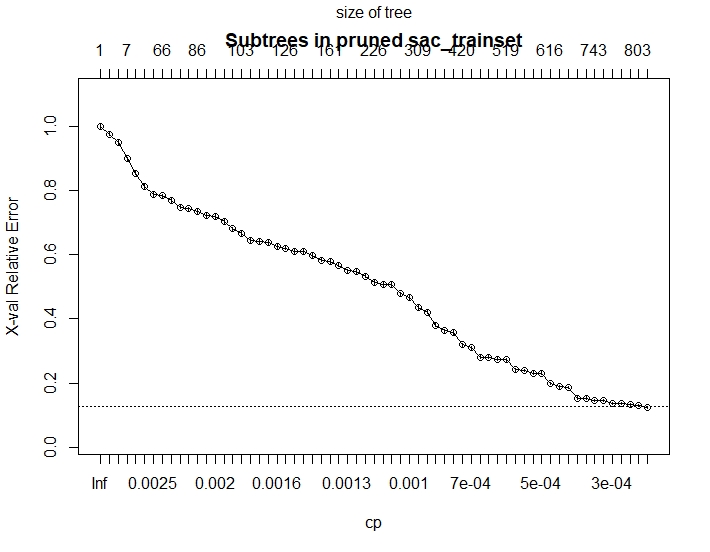


Fig 49: Printcp table for sac.cart2

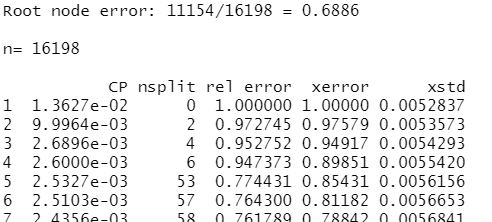


Fig 50: Testing sac.cart2

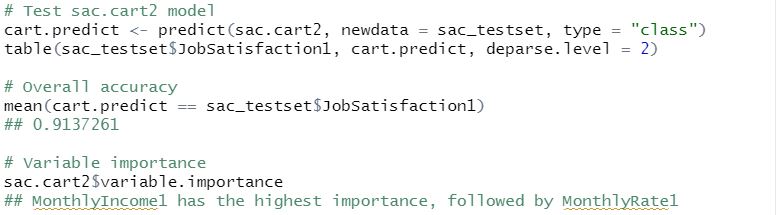


Fig 51: Prediction results for sac.cart2

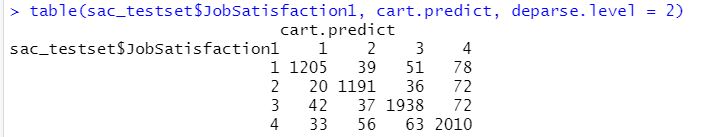


Fig 52: Variable importance for sac.cart2

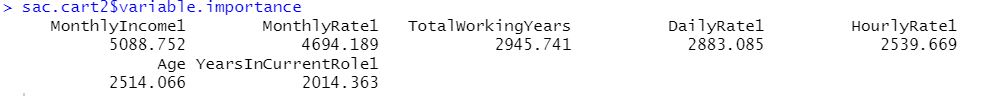


Fig 53: Analysis of AttritionNew

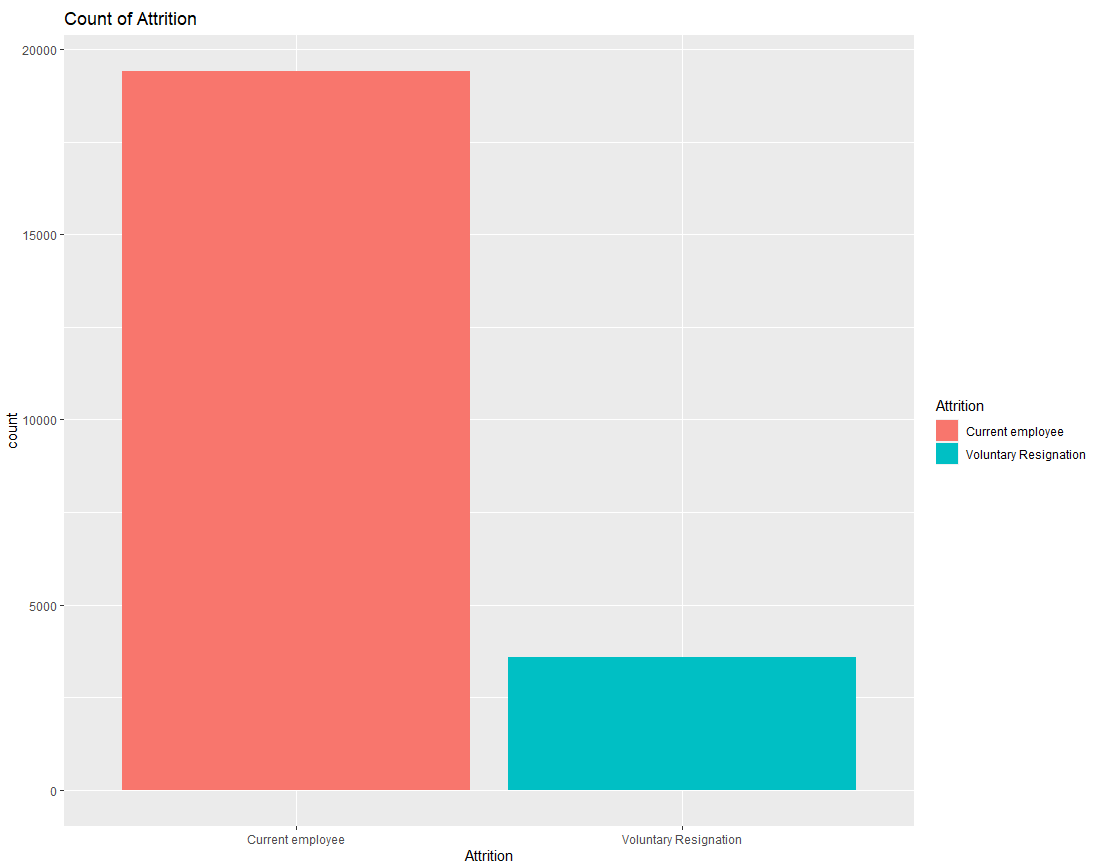
****

Fig 54: Analysis of JobSatisfactionNew

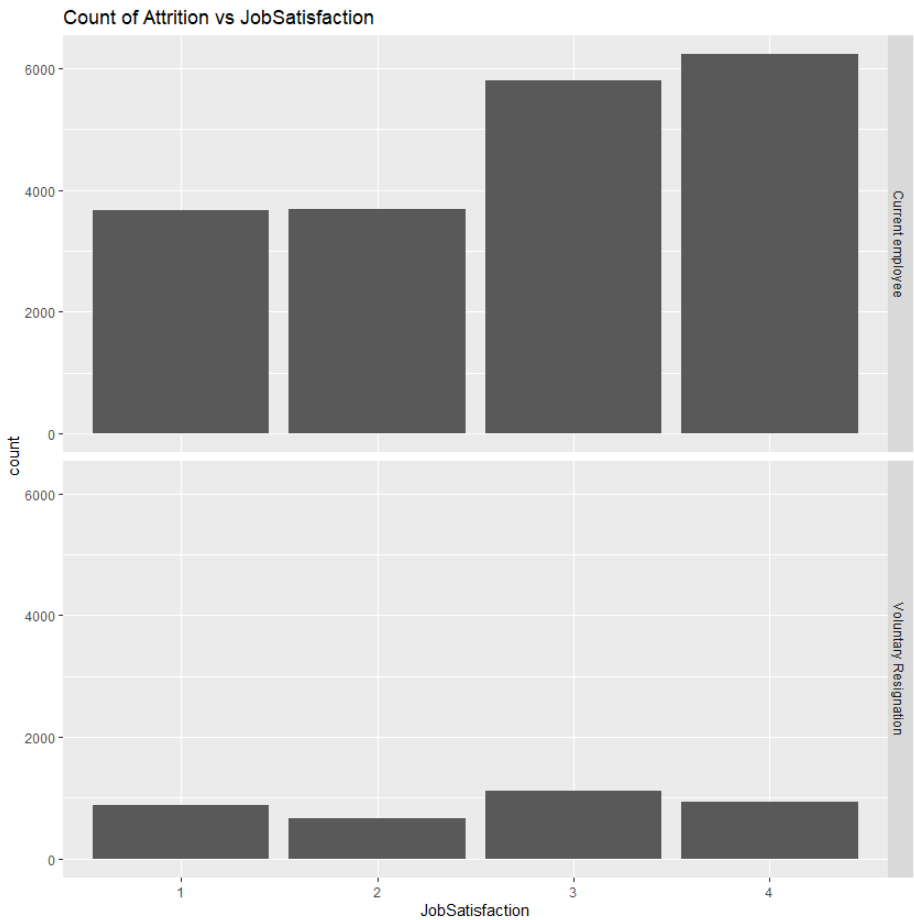
****

Fig 55: Analysis of YearsAtCompany

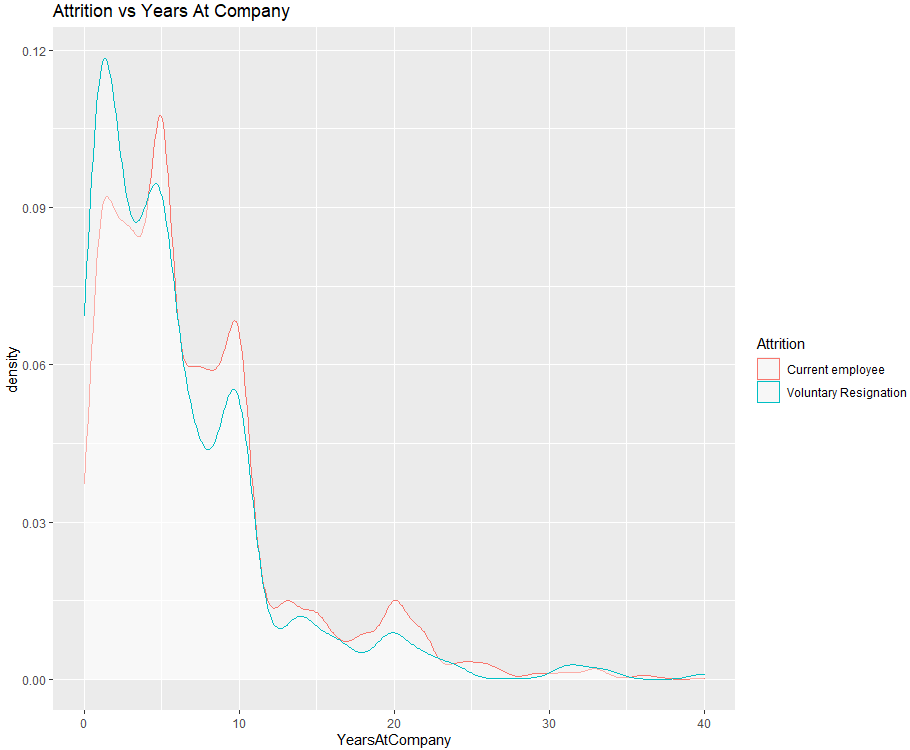
****

Fig 56: Mean and median of YearsAtCompany

****

Fig 57: Logistic regression model for AttritionNew, att.log1



Fig 58: Odds ratio for att.log1

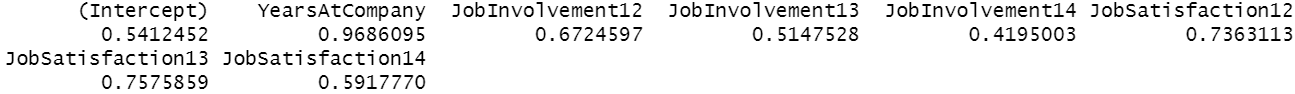


Fig 59: Odds ratio for att.log1, 95% confidence interval

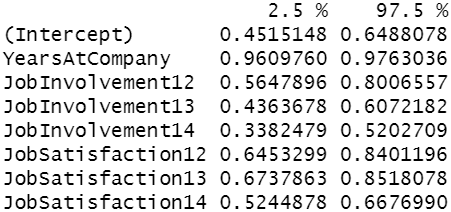


Fig 60: P-value for variables in att.log1

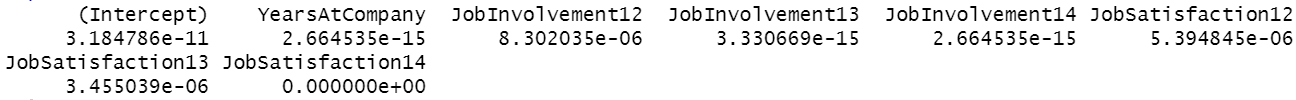


Fig 61: Train set prediction results using att.log1

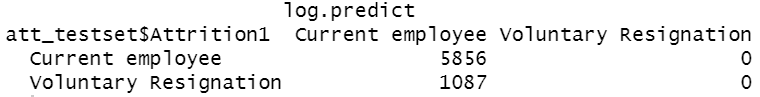


Fig 62: SMOTE, Logistic regression model for AttritionNew, att.log2



Fig 63: Odds ratio for att.log2

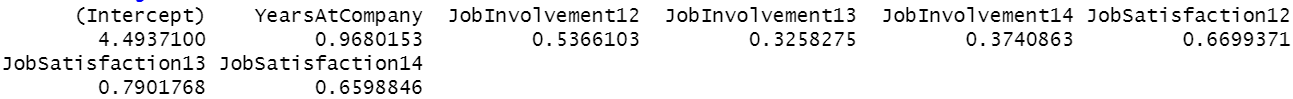


Fig 64: Odds ratio for att.log2, 95% confidence interval

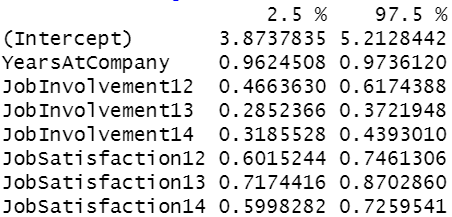


Fig 65: P-value for variables in att.log2

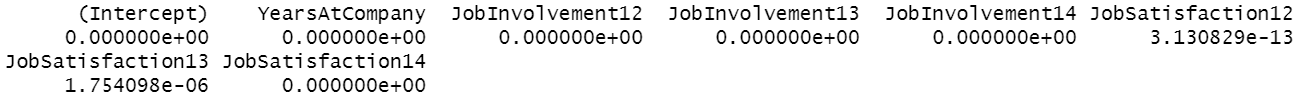


Fig 66: Train set prediction results using att.log2

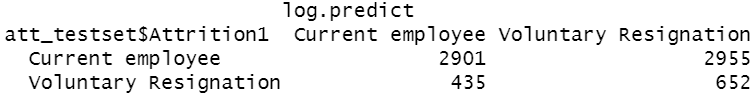
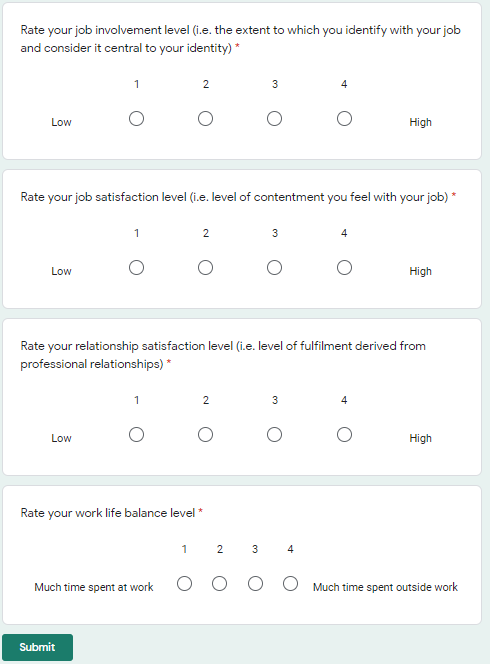
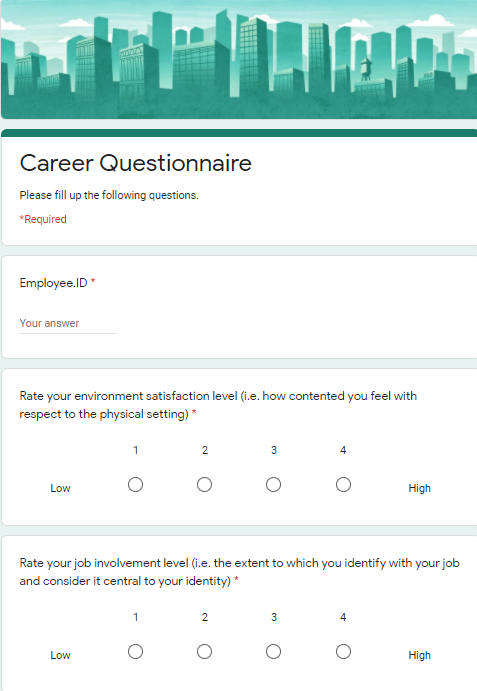


Fig 67: Career questionnaire

****