**UNSW Business School**

**ZZBU6511 Predictive Analytics**

**Assessment 3**

Predictive Modelling

Author: Geoffrey, Pang

Student Number: z5383289

Due Date: 20 June 2022 @5pm (Sydney time)

Word Count: 2107 words

Submission file should be in Word format and uploaded to Moodle before the due date.

**TABLE OF Contents**

1 Executive summary 4

2 Introduction 5

3 The top three predictive models 6

3.1 First Logistic Regression Predictive (Model3b) 7

3.2 Second Logistic Regression Predictive (Model4b) 11

3.3 Decision Tree Predictive Model (Model6) 14

4 conclusions & Recommendations 17

5 REferences 18

6 Appendixes 19

6.1 List of R Codes First Logistic Prediction Model 19

6.1.1 Codes to generate Correlation Matrix 19

6.1.2 Codes to do Logistic Regression Prediction using all variables 20

6.1.3 Codes to do Logistic Regression using up to 3\* Statistically Significant variables. 21

6.1.4 Codes to do Logistic Regression using up to 3\* Statistically Significant variables minus NumCompaniesWorked. 22

6.1.5 Codes to do AUC calculation and plotting for Logistic Regression for Model 1-3. 23

6.2 List of R Codes for Second Logistic Prediction Model 24

6.2.1 Codes to generate AUC plot for Model 4 and 5. 24

6.2.2 Codes to do Logistic Regression using up to 2\* statistically significant variables 25

6.2.3 Codes to do Logistic Regression using up to 1\* statistically significant variables 26

6.3 List of R Codes for Decision Tree Prediction Model 27

6.3.1 Codes to plot AUC and best threshold probability for Decision Tree with all variables 27

6.3.2 Codes to do Decision Tree Prediction using all variables minus Over18 and Attrition 28

# Executive summary

Three prediction models are proposed to predict the likelihood of whether staff will leave the company or not.

Based on the prediction model results, overtime is one of key factor in causing staff to leave. Company needs to figure out ways to reduce overtime so that this factor can be reduced.

Furthermore, company need to look into why the younger group of staffs (<22 years old, less than 4 years in company or less than 1 year with current manager) are leaving. Find out what the company can do to reduce attrition among this group.

Finally, employees in the job role of Sales Representative, Laboratory Technician and Sales Executive need special attention as more staff are leaving from these roles. Find out what are the triggers. Is it overtime, business travel, job satisfaction, work environment, job involvement, salary, or work life balance? Then find ways to rectify the situation.

By addressing these three areas, Globex Pharma should be able to reduce the attrition to an acceptable level of 10% or less.

# Introduction

As the newly hired consulting firm for a pharmaceutical company, Globex Pharma, we have been asked to build the best 3 models for predicting whether or not an employee will leave the company, and to glean insights from which predictors are used in the model about what they might do to reduce attrition.

After exploring various predictive models, the following three models are selected for being the most accurate ones:

1. Using logistic regression model up to three asterisk (\*\*\*) statistically significant variables minus NumCompaniesWorked.
2. Using logistic regression model up to two asterisk (\*\*) statistically significant variables.
3. Using decision tree predictive model using all available variables minus Over18 and Attrition.

The methods, assumptions and results are presented below in this report. Finally, conclusions and recommendations are made based on the predictions results on ways to reduce further attrition in the company.

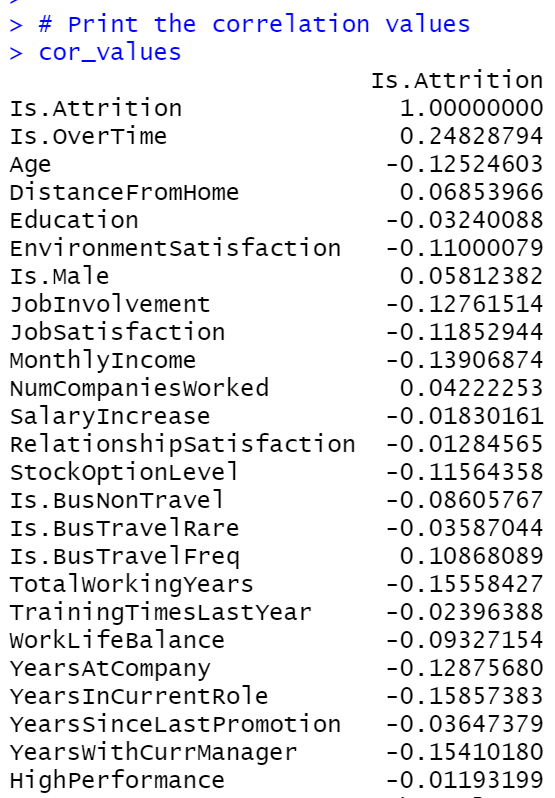
# The top three predictive models

We are predicting the value of a categorical variable, i.e., whether an employee will be leaving the company or not, hence linear regression which predict the value of a continuous variable is not suitable and only logistic regression and decision tree are appropriate for this task (UNSW 2022). DataRobot can be used too but it is not used for this prediction.

The first approach is to inspect the dataset, employee2.csv. It is noted that Over18 column can be ignored because all the value are Yes and hence doesn’t add any value in trying to differentiate whether an over 18 will leave the company or not. So, this column is ignored in all the prediction models.

We ran cross correlation of Attrition to all the possible variables and the values are tabulated below (Figure 1). None are above 0.5 meaning poor correlation between Attrition to any of these variables. Almost all are negative values meaning they are inversely correlated. OverTime has the highest positive value meaning the higher the OverTime (value are 1 for Yes and 0 for No) value, the higher likelihood of employees leaving the company.

We then split the dataset, 80% as training dataset while the rest as testing dataset. We make model prediction using the training dataset and later test the prediction using the testing dataset. We judge how good the logistic regression predictive models based on AUC (Area Under the ROC curve) (Starmer 2019) and the confusion matrix (Starmer 2018) computation for accuracy, sensitivity and specificity. The closer these numbers are to the value of 1, the better the prediction model fit the data. Finally, we also ran test on what is the most optimum threshold probability to use in make prediction for staff leaving.



**Figure 1: Cross Correlations between all numeric variables (some of the non-numeric variables are converted). Refer to Appendix 6.1.1 on how these calculations were done using R.**

## First Logistic Regression Predictive (Model3b)

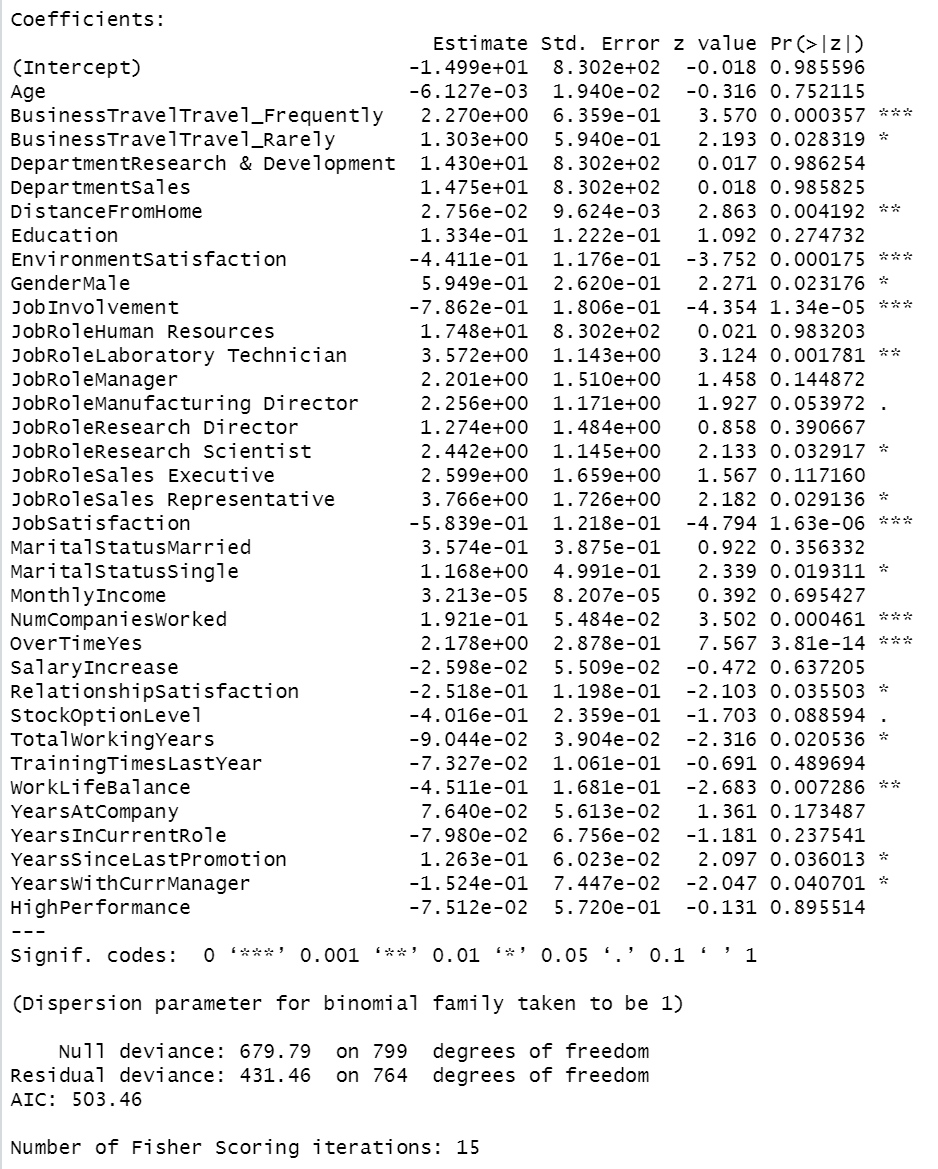
Prior to selection of the variables for the first logistic regression model prediction, all available variables are used and tested which ones are statistically significant in affecting the decision to quit (Figure 2). We can choose all the variables but we will run into the danger of overfitting the model. Our goal is to build a model that is robust enough to predict any new data. So, using a small but optimal number of variables would be ideal as long as we achieve the desired goal of prediction. We don’t want to choose too few variables also because we might end up under fitting the model too.

Logistic regression utilizing all variables (Model1a) (Figure 3a) with threshold probability set to 0.5 showed good accuracy (0.895) and specificity (0.953) on the testing dataset but was only able to predict about half the number of people leaving (sensitivity of 0.552). To improve sensitivity, threshold probability is tested from 0.1 to 0.9 range and the corresponding accuracy, sensitivity and specificity numbers are recorded (Fig 4). The test suggests that threshold probability of 0.15 is a good compromise in term of yielding the best result for sensitivity with some sacrifice to the accuracy and specificity numbers.

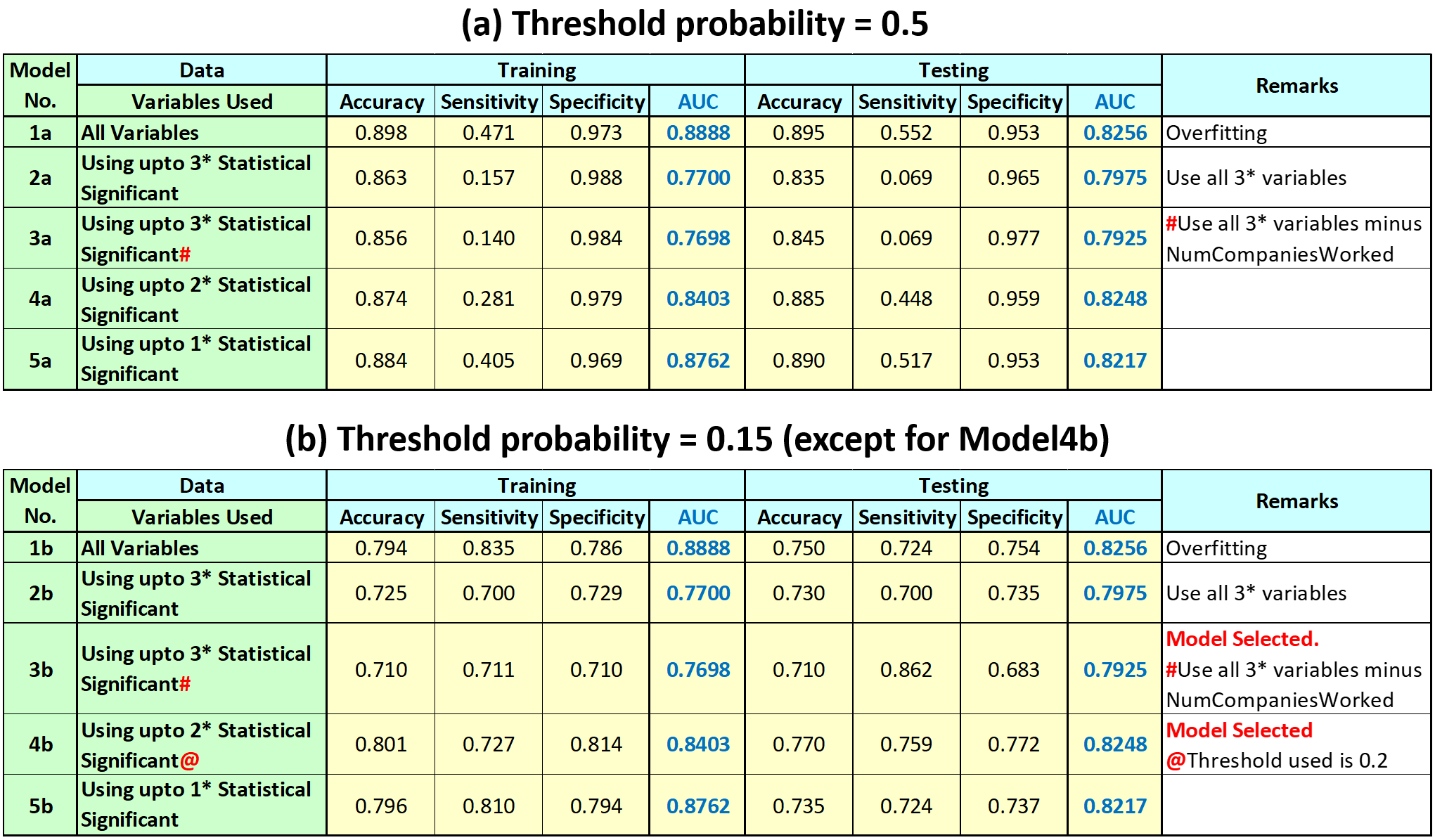
Choosing the 6 variables with coefficients that are statistically significant (Galak 2020) at 99.9% CI since p-value is <0.001 (3 asterisk only) (Figure 2) yield an acceptable prediction (Model2a). In fact, the logistic regression model (Figure 5) suggest that we can drop NumCompaniesWorked variable since it is statistically insignificant (Figure 5). The first logistic regression model is using just 5 variables (Model3b) (Figure 6) namely, BusinessTravel, OverTime, JobSatisfaction, EnvironmentSatisfaction and JobInvolvement. The logistic regression model suggest that these are the most important variables in determining whether employees leave or stay in the company.

The threshold probability is also tested and found that 0.15 is the most optimum to use (Figure 7). This first model (Model3b) is able to predict the testing dataset with accuracy of 0.71, sensitivity of 0.86 and specificity of 0.68 (Figure 3). AUC numbers calculated showed that logistic model using the 5 variables (Model3a & 3b) show good result for the training (0.77) and testing dataset (0.79) (Figure 3 & 8). Hence, this first model (Model3b) is able to predict with 86% probability whether an employee will leave the company if we have those 5 variables.

We can conclude from this model (Figure 6) that employees that are likely to leave the company are those that travel frequently, work overtime, not satisfied with their job involvement, not satisfied with their job and not satisfied with their work environment in decreasing order of importance.



**Figure 2: Table showing Coefficients and Statistically Significant for all the variables used in the Logistic Regression. Refer to Appendix 6.1.2 for R codes.**

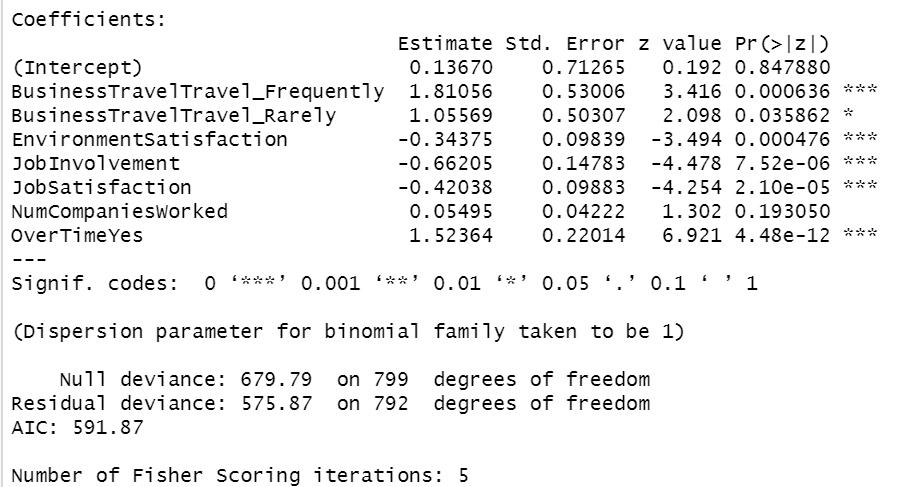


**Figure 3(a) Threshold Probability 0.5 used and 3(b) 0.15 (except for Model4b (0.2))**

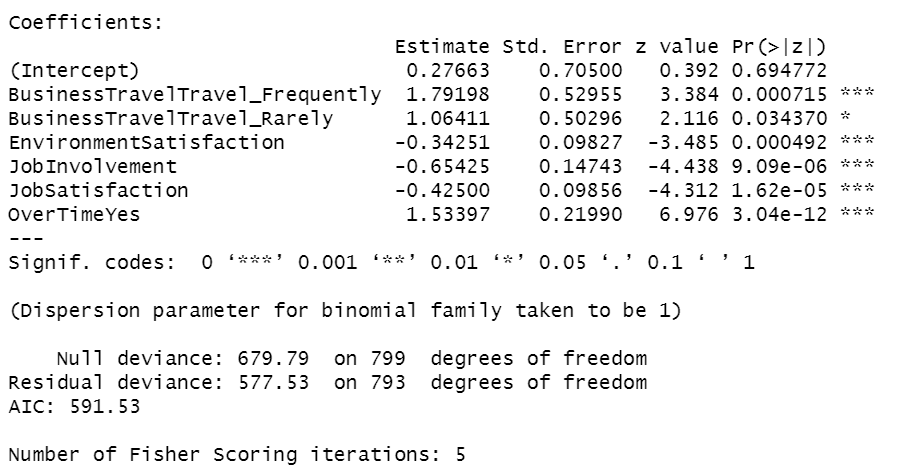
**Tabulation of different models using various variables for Logistic Regression. Model3b is selected as the first Logistic Regression model while Model4b is selected at the second Logistic Regression model. AUC column is Area under ROC curve is used to judge which is the best prediction model. The higher the better. Refer to Appendix 6.1.2 to 6.1.5 and 6.2.2 to 6.2.3 for R codes.**



**Figure 4: Tabulation of logistic regression prediction using all variables for (a)Training and (b)Testing dataset. Threshold probability of 0.15 give the most optimum prediction result. Refer to Appendix 6.1.2 for R codes.**



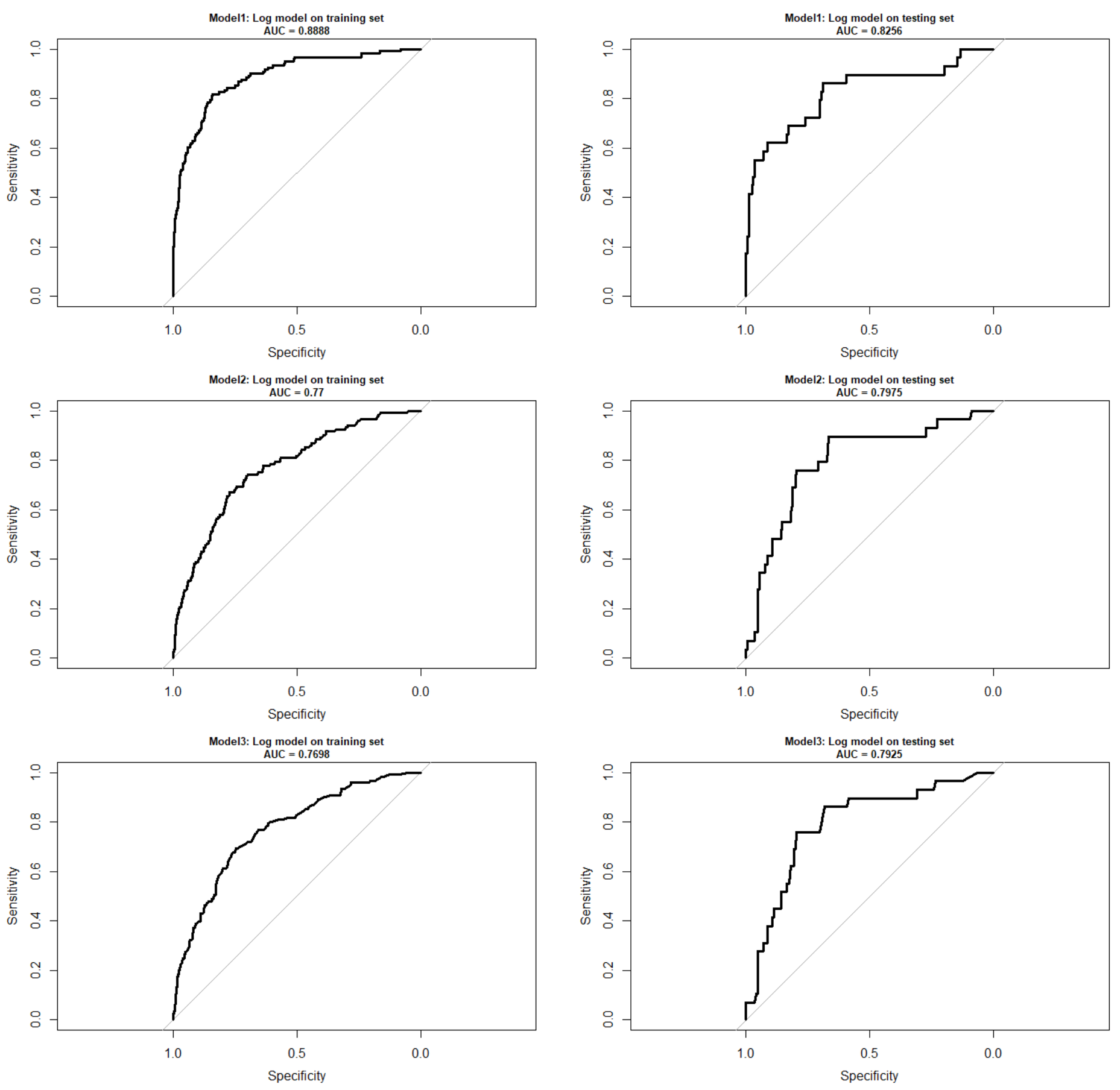
**Figure 5: Table showing the Coefficients and Statistically Significant variables when only 3\* statistically significant variables used in the Logistic Regression. Result suggest that we can drop the NumCompaniesWorked variable. Refer to Appendix 6.1.3 for R codes.**



**Figure 6: Table showing the Coefficients and Statistically Significant variables when only 3\* statistically significant variables minus NumCompaniesWorked are used in the Logistic Regression. Refer to Appendix 6.1.4 for R codes.**



**Figure 7: Tabulation of logistic regression prediction using 5 variables for (a)Training and (b)Testing dataset. Threshold probability of 0.15 give the most optimum prediction result. Refer to Appendix 6.1.4 for R codes.**



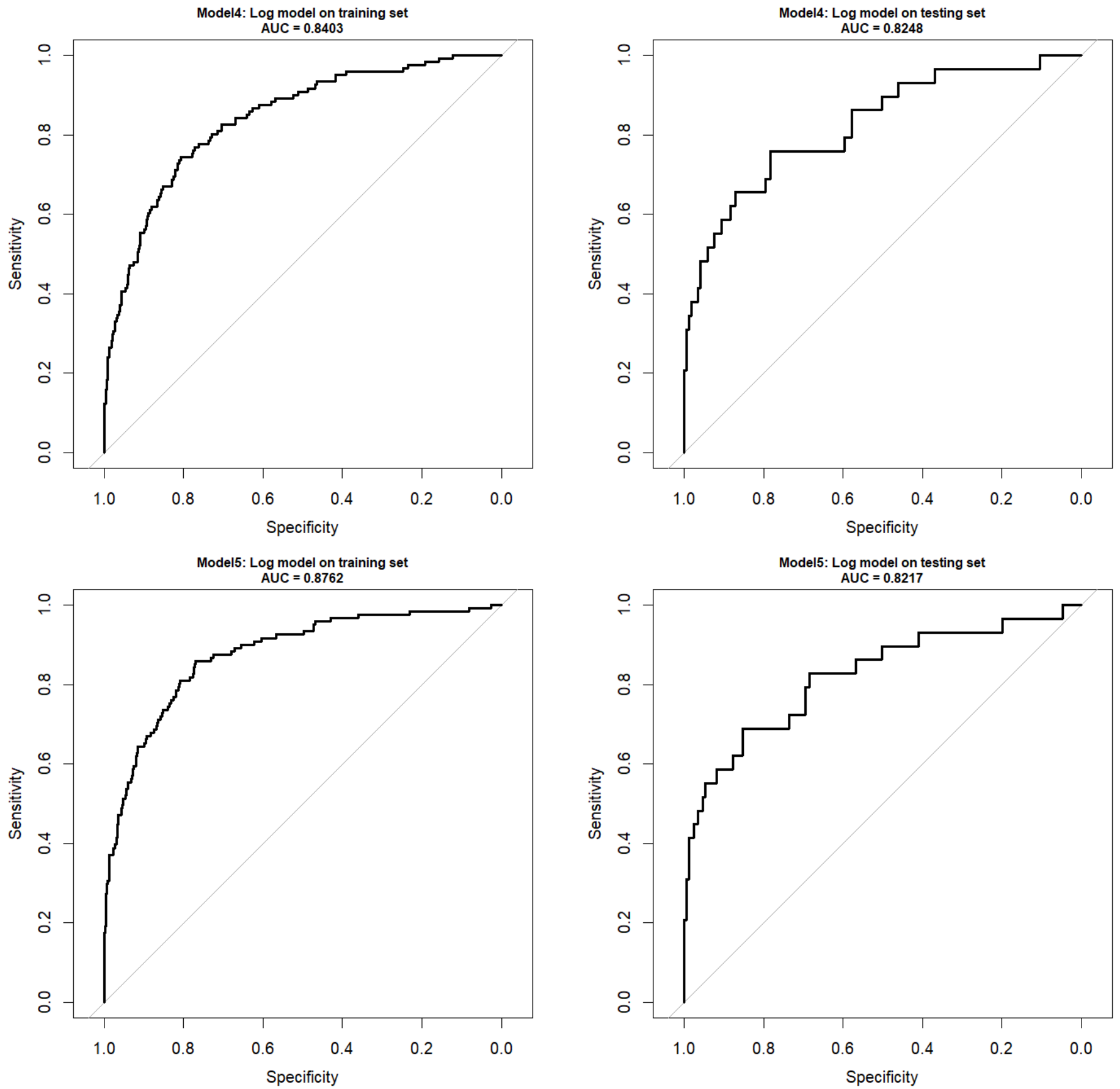
**Figure 8: AUC plots for Model 1 to 3 for Training and Testing dataset. Refer to Appendix 6.1.5 on how plots were done using R.**

## Second Logistic Regression Predictive (Model4b)

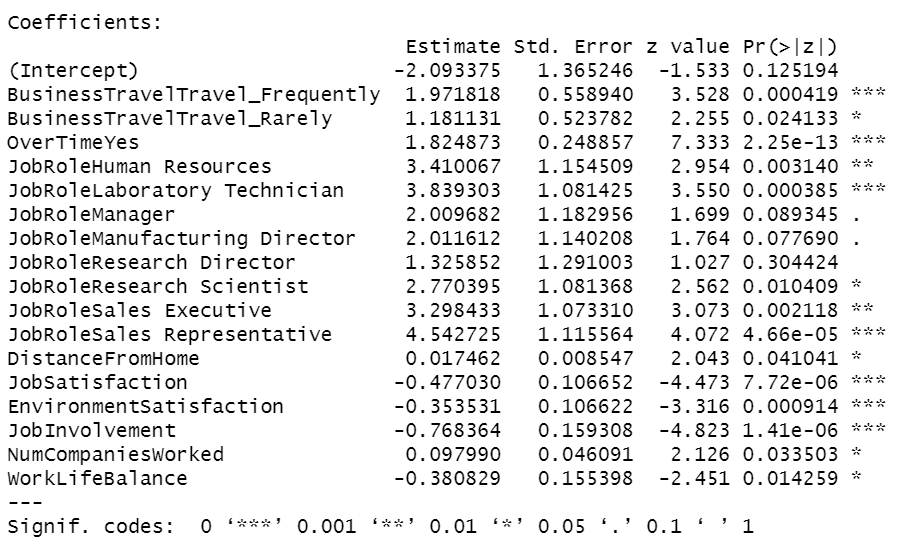
In the second logistic regression, to further improve the overall prediction, more variables are used. The AUC numbers (Figure 9) suggest that we should get better prediction by using more variables with value of 0.8248 (testing dataset) for using up to 2 asterisk (\*\*) statistically significant variables at 95%% CI of the variables p-value <0.05 (Figure 2). Adding up to 1\* statistically significant variables actually decrease the AUC number slightly to 0.8217. Due to this, we recommend that the second logistic regression model is using up to 2\* statistically significant variables.

Running the logistic regression up to the 2\* statistically significant, suggest that we don’t need to drop any variables (Figure 10). We tested the threshold probability and 0.2 is the most optimum to use to get the best overall prediction result (Figure 11). Using the testing dataset (Figure 3b) (Model4b), there are improvement in the prediction as seen in the accuracy (0.77 vs 0.71), specificity (0.76 vs 0.68) and AUC (0.8248 vs 0.7925) compared to Model3b but the sensitivity number reduce slightly (0.77 vs 0.86). The variables that are excluded from the prediction model are Department, MaritalStatus, RelationshipStatisfaction, MonthlyIncome, Education, Age, SalaryIncrease, StockOptionLevel, TrainingTimesLastYear and YearsAtCompany. Our logistic regression model suggest that these variables are not important in determining whether employees are leaving or staying.

In addition to the previous variables in Model3b, we have included JobRole, DistanceFromHome, NumCompaniesWorked and WorkLifeBalance in this Model4b. This model (Model4b) further suggests that those who worked as Sales Representative, Laboratory Technician, Human Resources and Sales Executive, doesn’t have good work life balance, have worked for multiple companies and stayed quite a distance from office in decreasing order of importance (Figure 10) is further likely to increase the likelihood of leaving the company.



**Figure 9: AUC plots for Model 4 to 5 for Training and Testing dataset. Refer to Appendix 6.2.1 on how this plot is done using R.**



**Figure 10: Table showing the Coefficients and Statistically Significant when up to 2\* statistically significant variables used in the Logistic Regression. Refer to Appendix 6.2.2 for R codes.**



**Figure 11: Threshold probability testing. 0.2 is the most optimum. Refer to Appendix 6.2.2 for R codes.**

## Decision Tree Predictive Model (Model6)

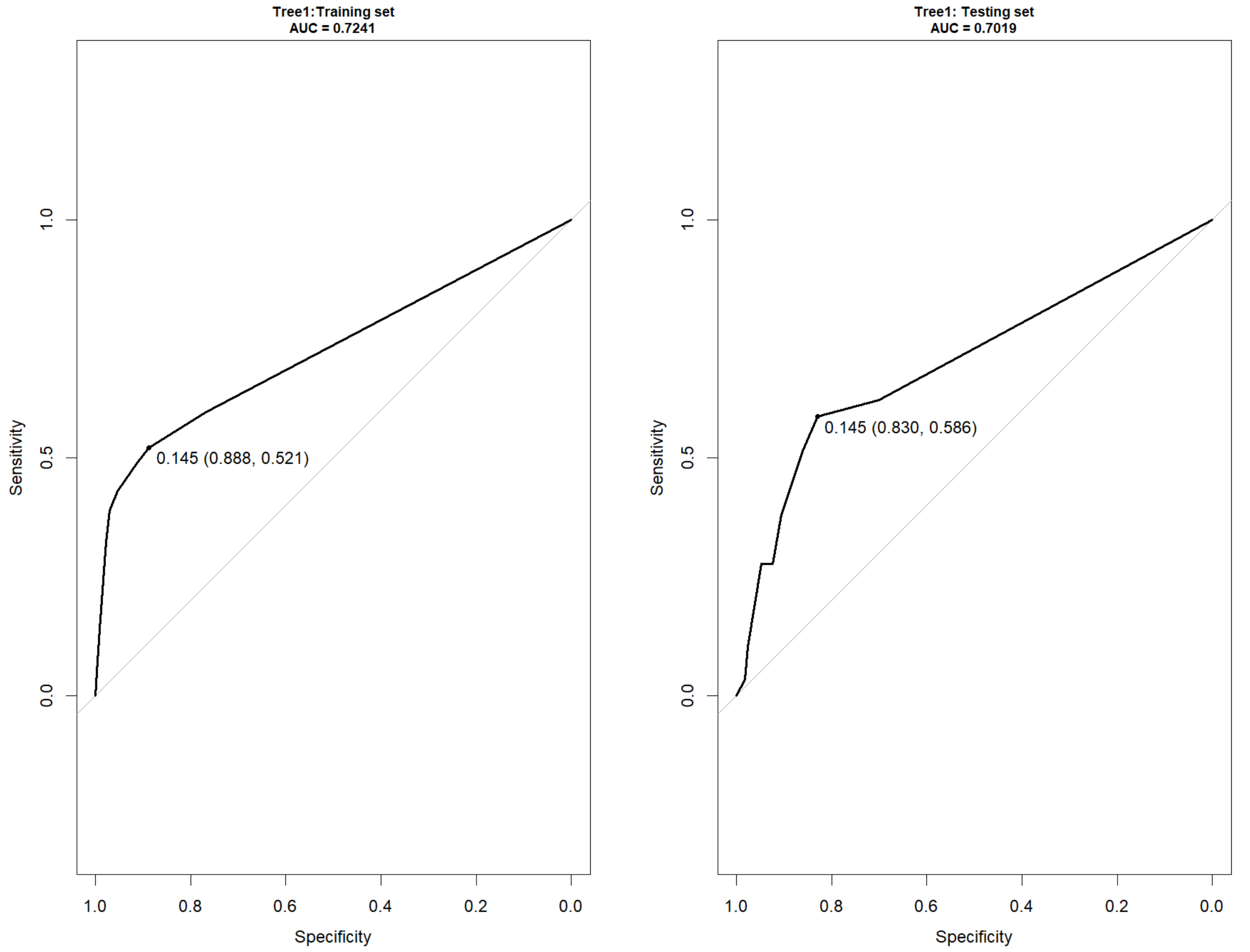
For the decision tree prediction model, we run the model with all available variables minus the Over18 and Attrition. The AUC (Figure 12) from the testing dataset is 0.70 which is lower than the previous best logistic prediction at 0.82 (Model4b) and the best threshold probability is calculated as 0.145 which give specificity of 0.830 and sensitivity of 0.586 for the testing dataset (Figure 13). Overall, the prediction is not as good as the previous two logistic prediction models (Model3b & 4b).

Since our interest is only finding out people that are leaving, we focus on our node investigation on the predicted class of Yes (i.e., leaving) rather than No (i.e., not leaving) especially on the leaf nodes. We focus on the most important node to the least judging by the probability of occurrence and percentage data set falls under that leaf node (UNSW 2022).

Looking at the Decision Tree (Figure 14), we can conclude the following:

1. The most significant leaf node is the one on the right most with 4% of the dataset falls under this node with 0.71 probability that an employee will leave since the predicted class of the node is Yes (i.e., leaving). This node was the result of a person not working overtime, not working as either Healthcare Representative, Manager, Manufacturing Director, Research Director and Research Scientist, and staying more than 18km from office.
2. The second most significant leaf node is the 2nd one from the right. There are 2% dataset falls under this node with 0.73 probability that an employee will leave. This node is the result of not working overtime, not working as either Healthcare Representative, Manager, Manufacturing Director, Research Director and Research Scientist, stay less than 18km from office, not having stock option and have not work more than 4 years in the company.
3. The third most important leaf node with predicted class of Yes is the 2nd from the left. This was the result of a person is working overtime and age is over or equal to 22. There is 0.62 probability that an employee will leave and there are 2% of data set falls under this node.
4. The last leaf node with predicted class of Yes is the 5th node from the left. This was the result of a person not working overtime, works as either Healthcare Representative, Manager, Manufacturing Director, Research Director and Research Scientist, with monthly income less than $3514 and have not been more than a year with current manager. There is 0.75 probability that an employee will leave and there are 1% of the data set falls under this node.

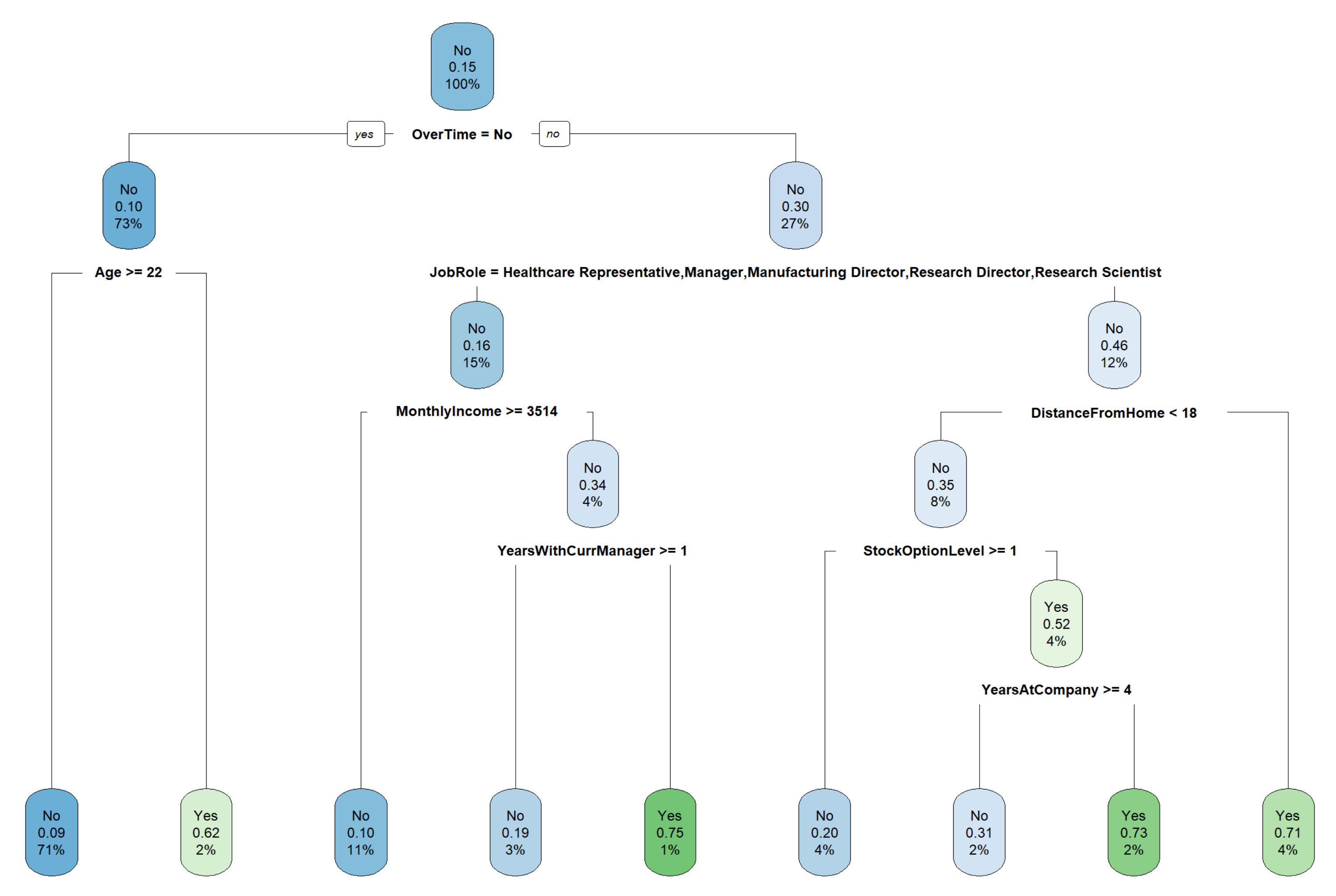
Judging from this prediction, company need to pay attention to these group of people: those who are either stay further than 18km from office, less than 4 years at the company, no stock options, younger than 22, worked less than 1 year with current manager, drawing monthly income of less than $3514, worked as Healthcare Representative, Manager, Manufacturing Director, Research Director and Research Scientist and working overtime. The model predicts that employees falling under these categories are likely to leave the company.



**Figure 12: AUC plots for training and testing dataset using al variables. Refer to Appendix 6.3.1 on how these plots were done using R.**



**Figure 13: Decision Tree key measures. Refer to Appendix 6.3.2 on how these plots were done using R.**



**Figure 14: Decision Tree using all variables. Refer to Appendix 6.3.2 on how these plots were done using R.**

# conclusions & Recommendations

We proposed two predictive (Model 3b & 4b) models using logistic regression and one using decision tree (Model6) in determining whether staff will leave the company or not. Based on the AUC, accuracy, sensitivity and specificity, the best prediction model to determine whether employee will leave the company is the second logistic prediction model (Model4b). This model is able to predict with 76% probability that an employee will leave or not based on the available data on whether employee is working overtime, job satisfaction, job involvement, job role, business travel, environment satisfaction, number of companies worked, distance from home and work life balance in decreasing order of importance. Earlier, Model3a had identified that the five most important variables are BusinessTravel, OverTime, JobSatisfaction, EnvironmentSatisfaction and JobInvolvement will determine whether staff will leave or not.

Furthermore, the decision tree prediction model suggests that those working less than 4 years at the company, not having stock options, younger than 22, worked less than 1 year with current manager and drawing less than $3514 monthly income are important consideration too.

Due to these predictions, we recommend the following actions to be taken by the company to reduce the likelihood of employee leaving:

1. Monitor those who worked overtime. Why are they working overtime? Is it necessary? Find ways to reduce overtime. Can the workload be distributed? Do we need to hire more people?
2. Find out why the younger staffs are leaving. Are the younger staffs especially those under 4 years with company are given sufficiently suitable, challenging and offered good career progression? Can they be incentivised with stock options? Can their salary be increase?
3. Find out why those who worked as Sales Representative, Laboratory Technician and Sales Executive are not happy? Is it related to overtime, business travel, job satisfaction, work environment, job involvement, salary, or work life balance? Find out what company might be able to offer to help alleviate any of the problem.

These observations and recommendations seem to agree with Oliver, 1998 who had mentioned before that high staff turnover is usually a function of negative job attitudes and/or low job satisfaction. She further mentioned that statistic show that one-fifth of employees quit within the first 6 months of employment due to bad job fit among other reasons.

With more information, company then will be able to take necessary actions to rectify the situation so that the reason for leaving can be minimized. We will never be able to eliminate staff attrition but less than 10% would be a healthy target to achieve (Insight Global 2021).

# REferences

Galak, J, 2020, Statistical Significance and p-Values Explained Intuitively, Data Demystified.

<https://www.youtube.com/watch?v=DAkJhY2zQ3c> [Assessed 19 June 2022]

Oliver, J 1998, ‘High staff turnover - find out why your staff are leaving’, Management today, p. 84.

Insight Global, 2021, Employee Attrition Rate: How to Calcualte & Improve.

<https://insightglobal.com/blog/employee-attrition-rate-how-to-calculate-improve/#:~:text=While%20it%E2%80%99s%20difficult%20to%20define%20a%20%E2%80%9Cgood%E2%80%9D%20attrition,and%20industry%20to%20industry%2C%20depending%20on%20the%20circumstances>. [Assessed 19 June 2022]

Starmer, J, 2018, Machines Learning Fundamentals: The Confusion Matrix, Clearly Explained, StatQuest.

<https://youtu.be/Kdsp6soqA7o> [Assessed 7 June 2022]

Starmer, J, 2019, ROC and AUC, Clearly Explained, StatQuest.

<https://www.youtube.com/watch?v=4jRBRDbJemM>

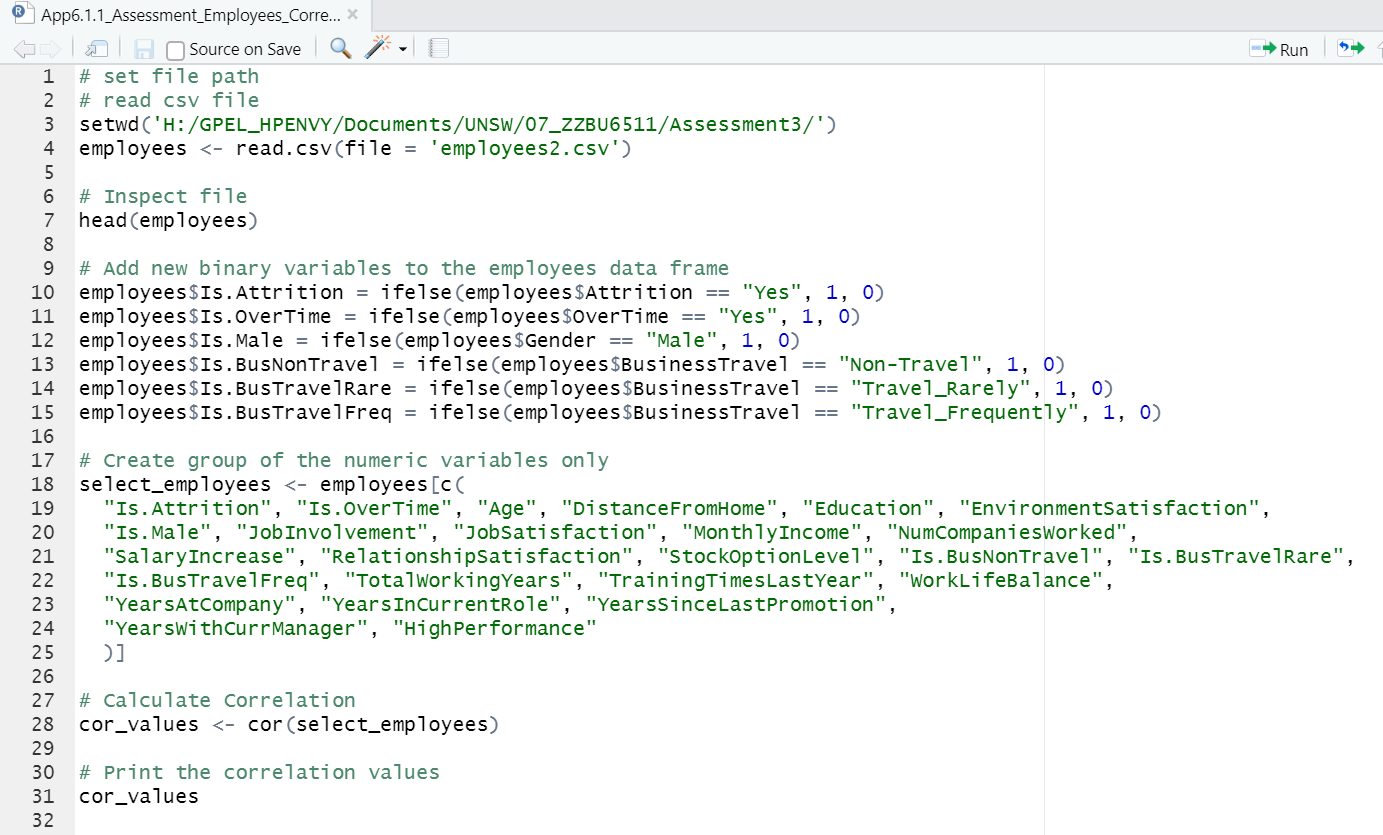
UNSW, 2022, ZZBU6511-Predictive Analytics (H322 Online) Week 4: Prediction – Logistic regression course note.

UNSW, 2022, ZZBU6511-Predictive Analytics (H322 Online) Week 5: Prediction with decision trees course note.

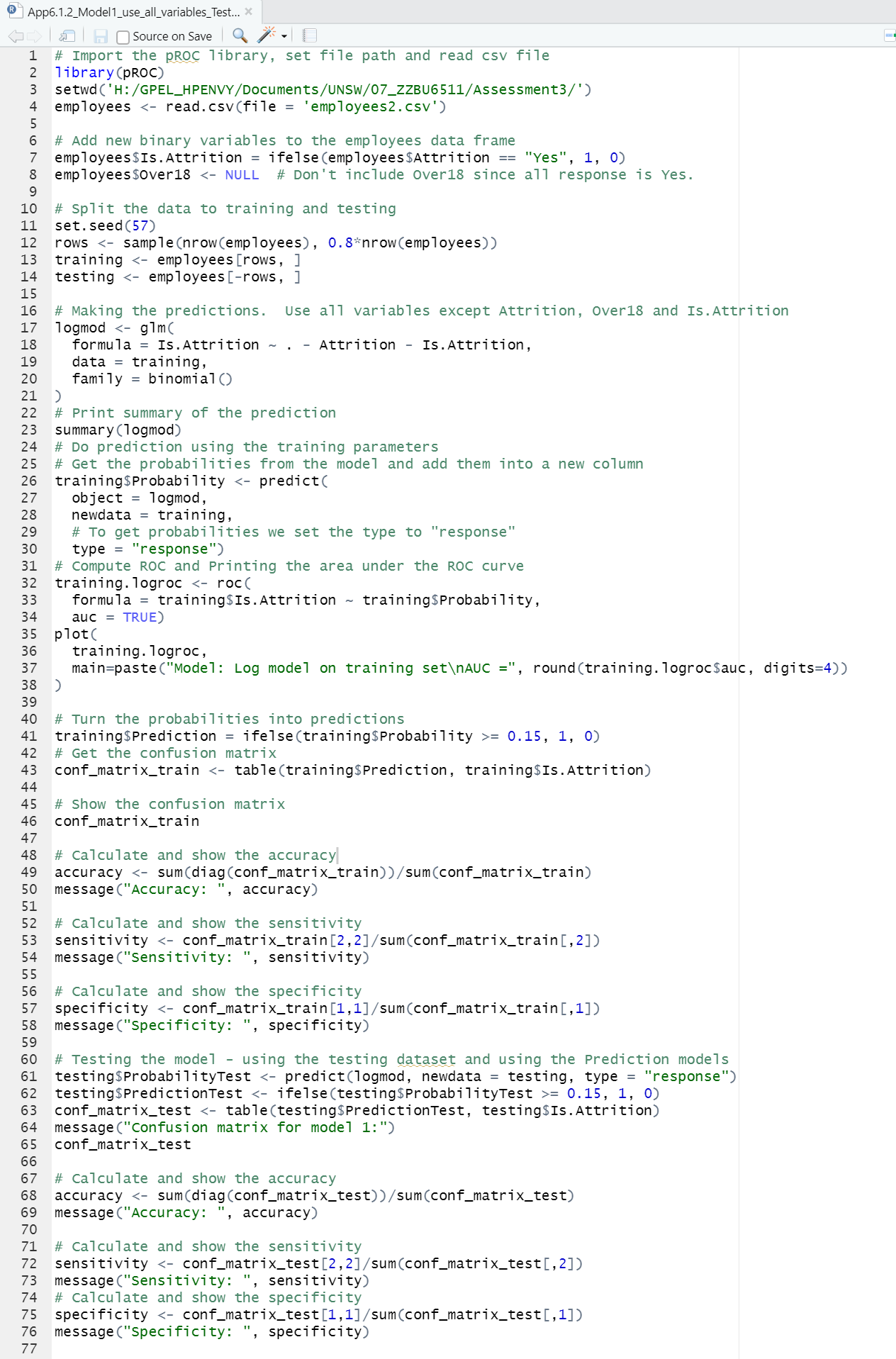
# Appendixes

## List of R Codes First Logistic Prediction Model

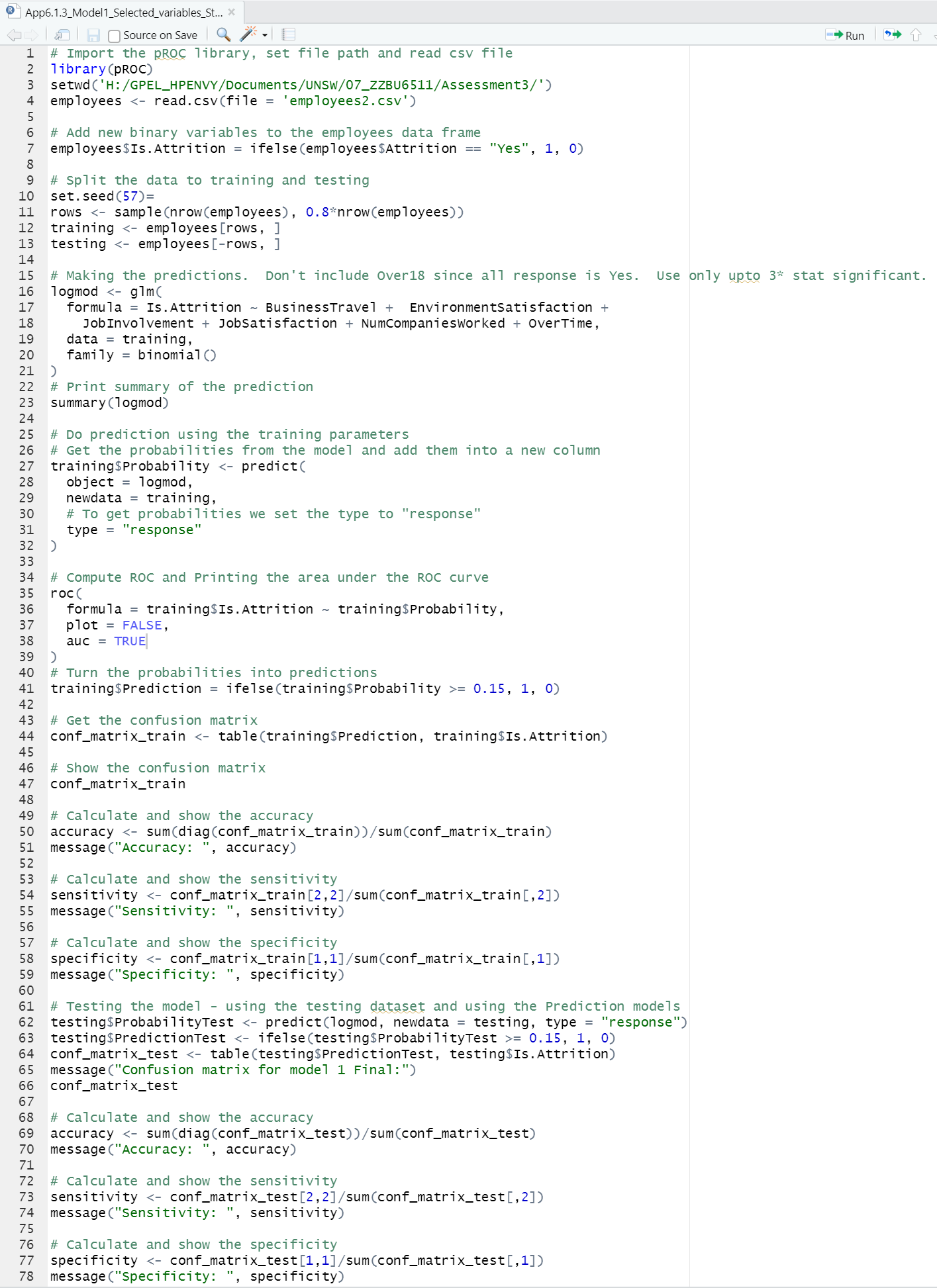
### Codes to generate Correlation Matrix



### Codes to do Logistic Regression Prediction using all variables



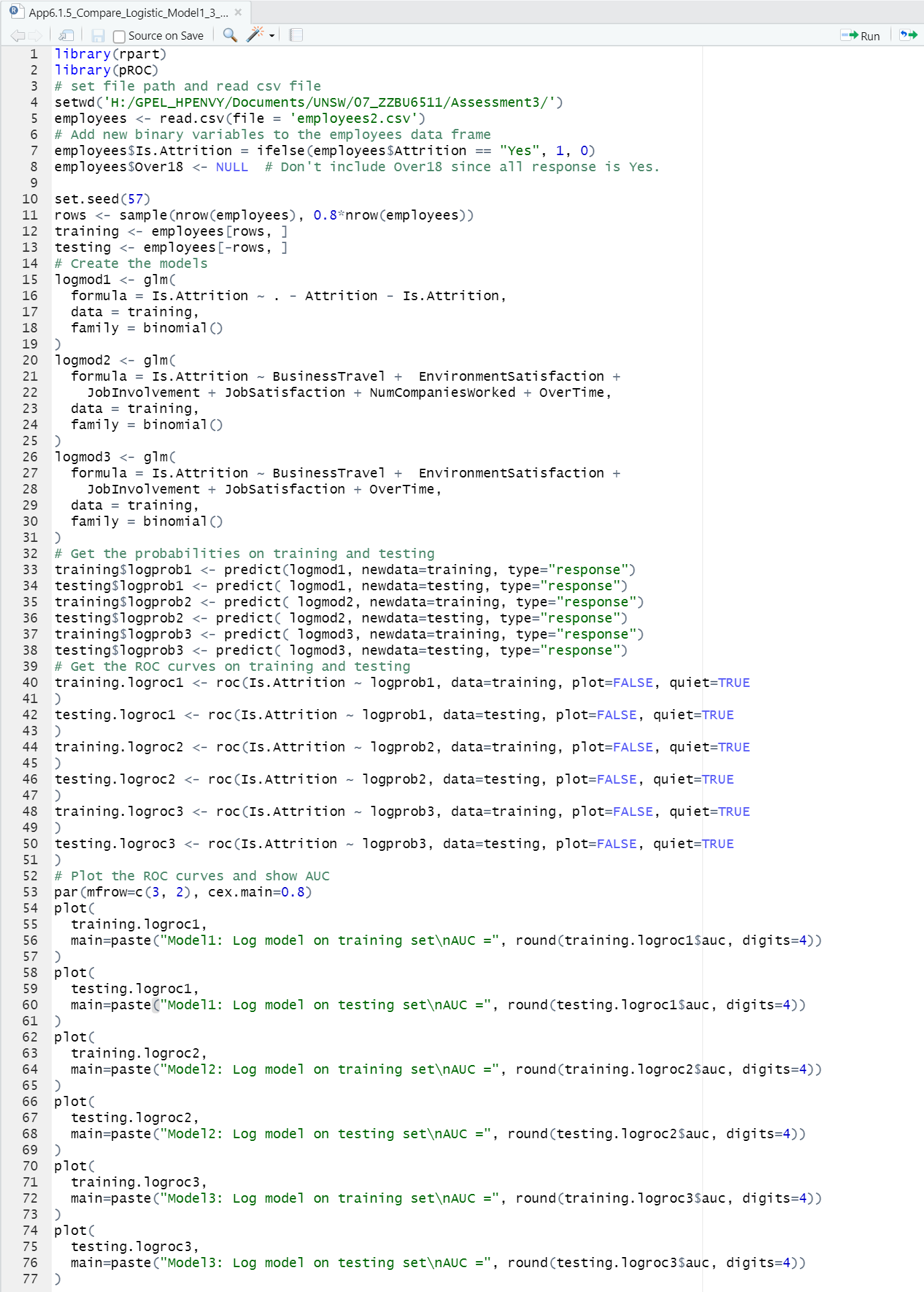
### Codes to do Logistic Regression using up to 3\* Statistically Significant variables.



### Codes to do Logistic Regression using up to 3\* Statistically Significant variables minus NumCompaniesWorked.

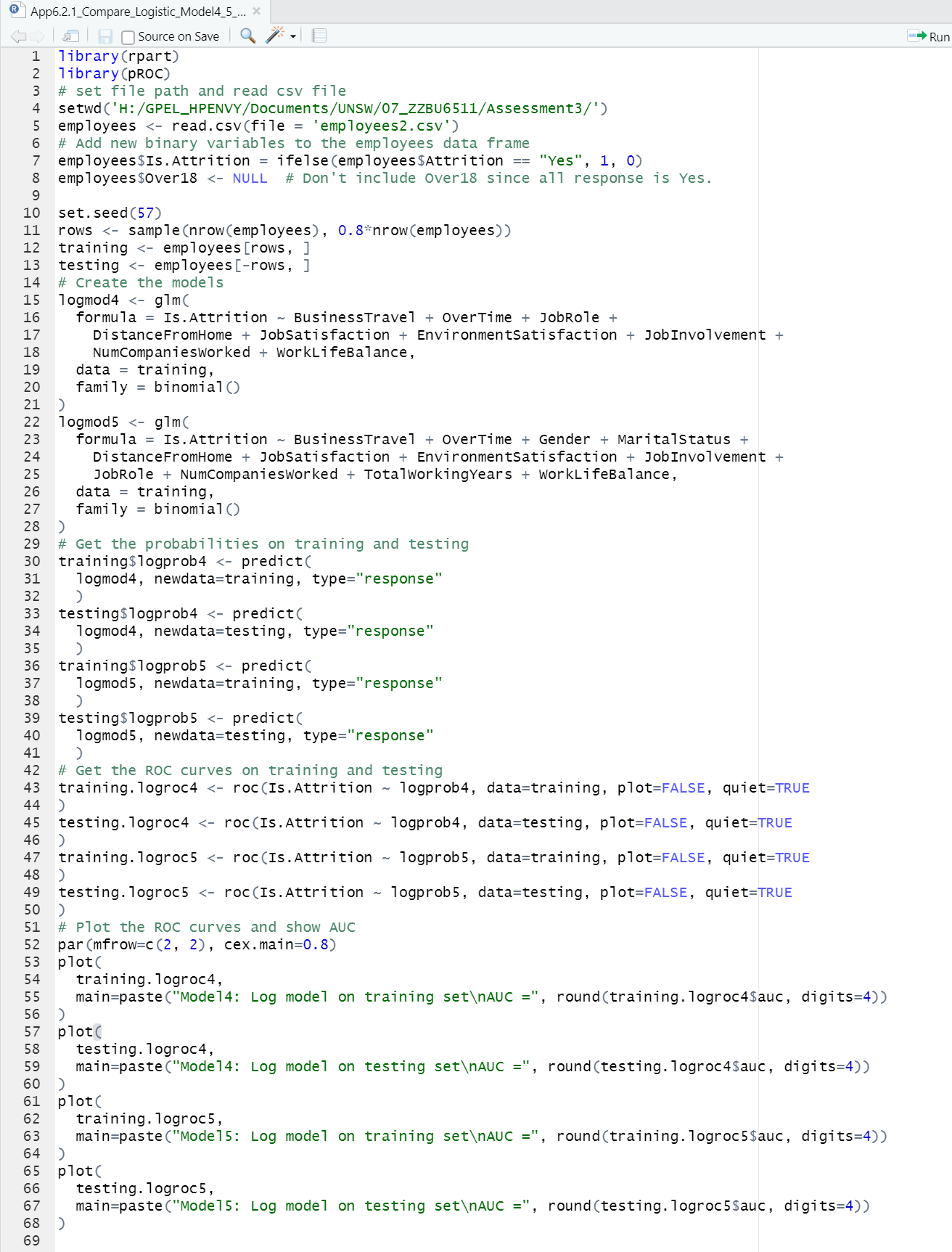


### Codes to do AUC calculation and plotting for Logistic Regression for Model 1-3.

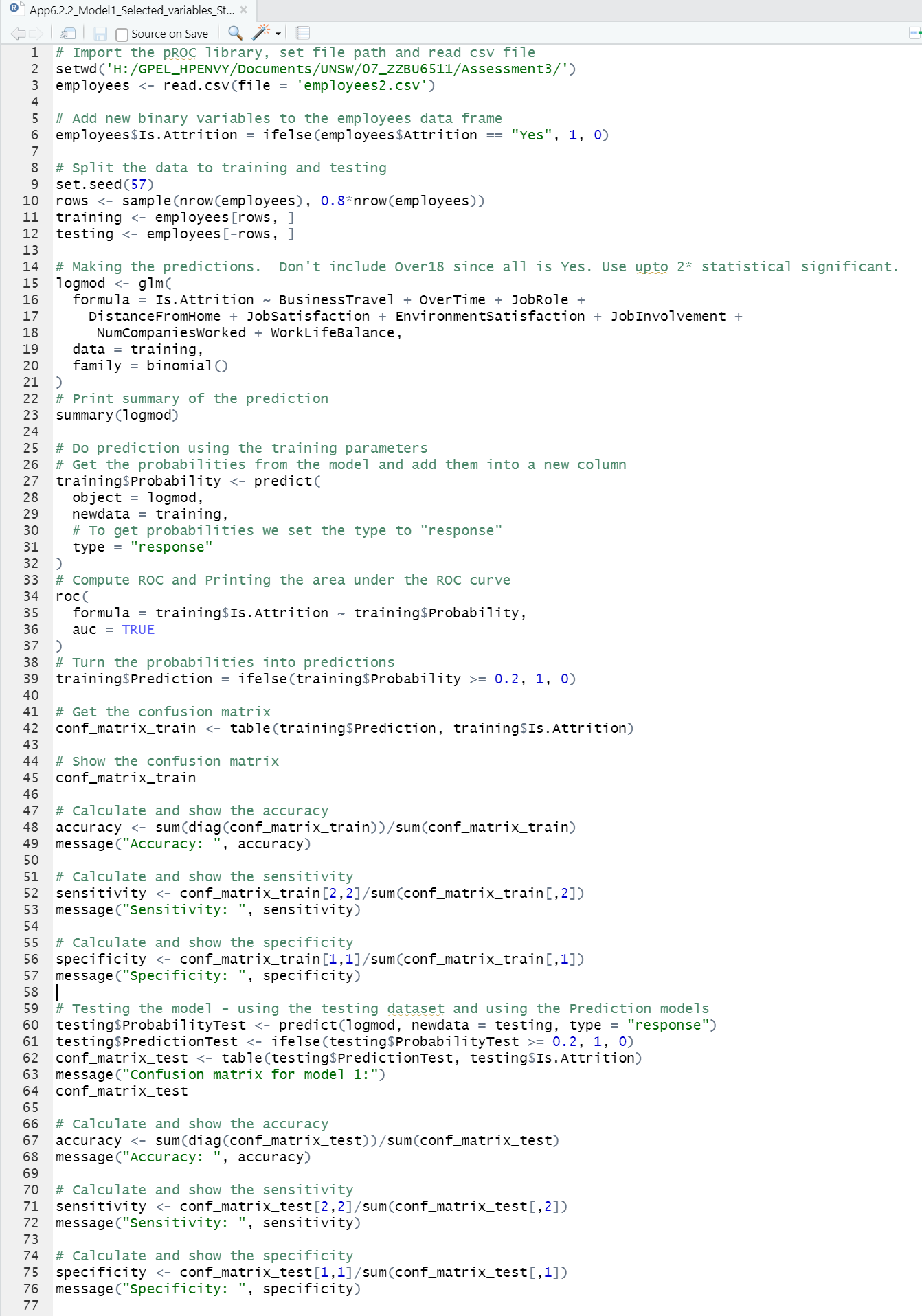


## List of R Codes for Second Logistic Prediction Model

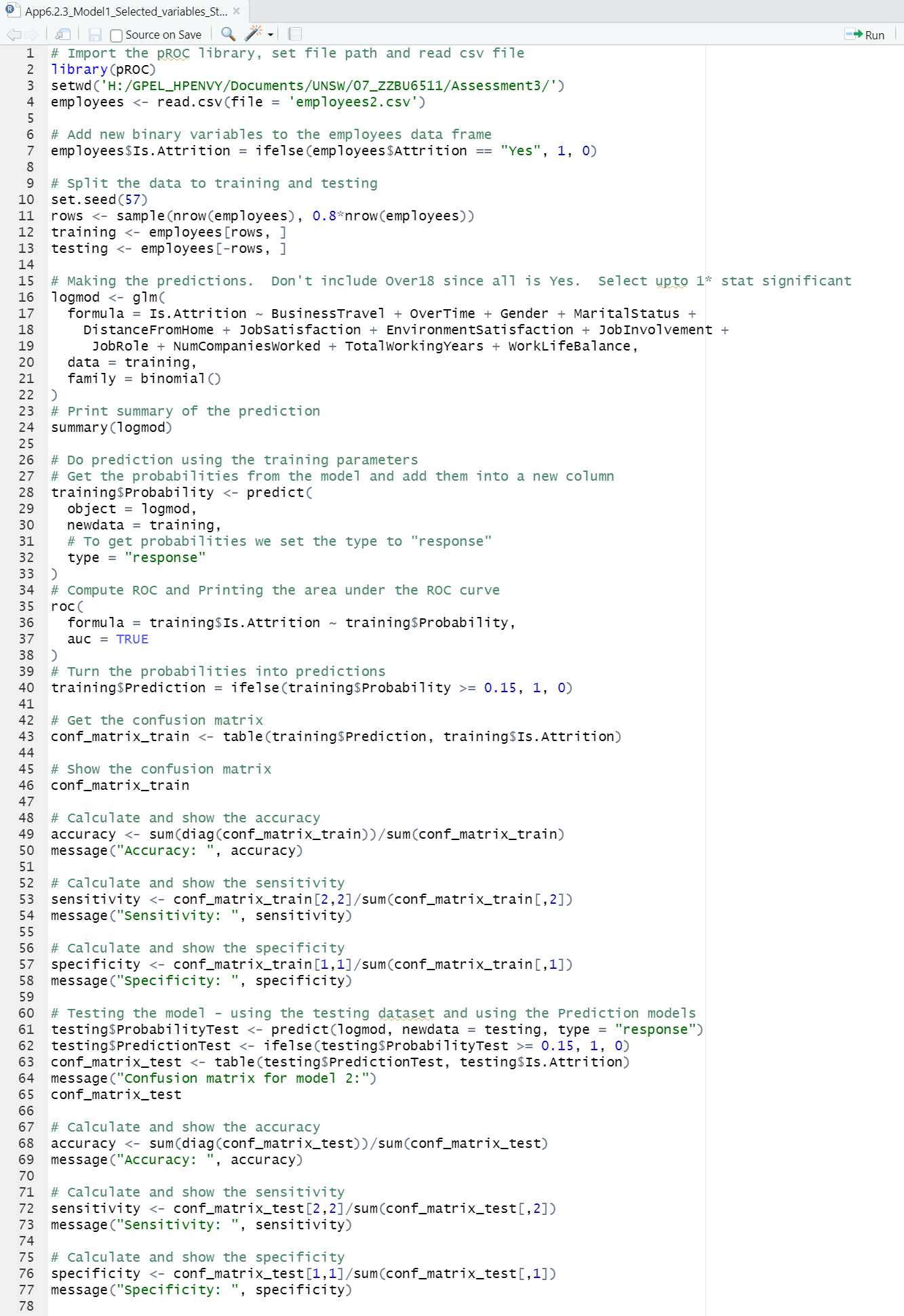
### Codes to generate AUC plot for Model 4 and 5.



### Codes to do Logistic Regression using up to 2\* statistically significant variables

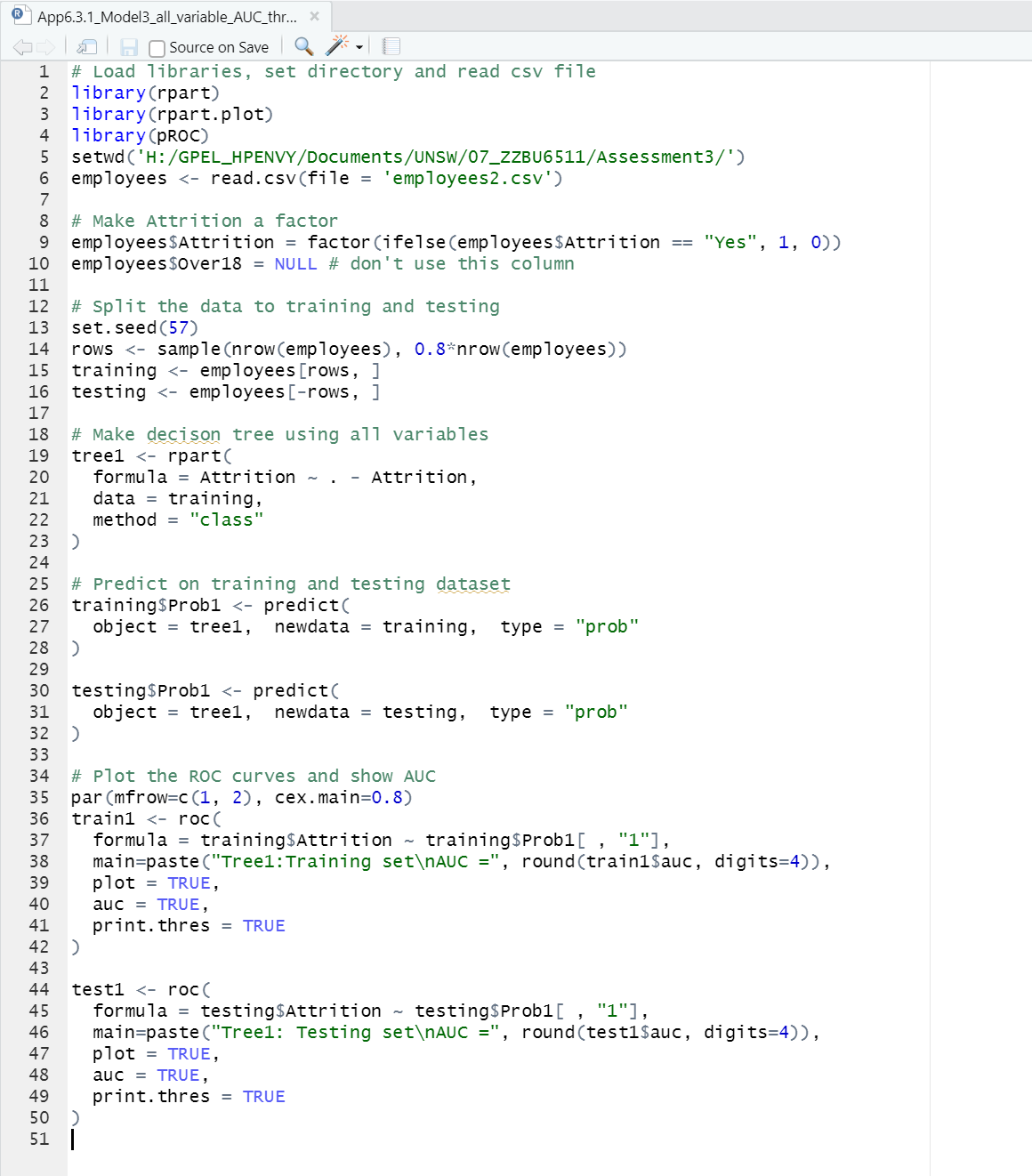


### Codes to do Logistic Regression using up to 1\* statistically significant variables



## List of R Codes for Decision Tree Prediction Model

### Codes to plot AUC and best threshold probability for Decision Tree with all variables



### Codes to do Decision Tree Prediction using all variables minus Over18 and Attrition

