ST4248
Statistical Learning II

Final Report

Group 8
Lee Xin Qi
Leona Ang Qiao En
Leong Jia Wei, Marcus
Ng Mei Ting

Title of Project: Predicting IBM Employee Attrition

Summary

Using the IBM HR Analytics Employee Attrition & Performance dataset fictionally created by IBM data scientists, this project aims to be able to predict employee attrition using significant variables and achieving an Accuracy, Sensitivity, Specificity and AUC of more than 0.800 each. The dataset was first pre-processed, and statistical models such as Logistic Regression, Decision Trees and Support Vector Machines were subsequently used to predict attrition. Logistic Regression with Lasso was selected as the final prediction model based on its overall performance. With these results, companies would thus be able to better cope with employee attrition and its consequences. *(99 words)*

Introduction

After COVID-19 had hit Singapore, The Straits Times released an article reporting that a local survey showed employee attrition as the biggest talent-related challenge facing Singapore's Science and Technology industry (Yee, 2020). Attrition is a process in which a number of people retire or resign and it often leads to a company losing the knowledge and skill capitals one brings. With high attrition, other implications such as the cost required for onboarding new employees may cost a company billions. Socially, a team's morale may also be affected when valuable members leave the company. Thus, it is of great interest for companies to understand the significant variables behind employee attrition and predict when one is likely to leave. This would then better help companies implement measures to remedy the situation, be it minimising employee attrition or planning for recruitment in advance. For this project, the dataset used is the IBM HR dataset adapted from Kaggle based in the U.S context. Though the study recognized that it may be unrepresentative of all companies, it still serves as a great synthetic data for predictive modelling and validation.

Goals & Hypotheses

This project aims to predict how likely an employee may leave the company, with an accuracy, sensitivity, specificity and area under curve (AUC) of more than 0.80 each. Additionally, it seeks to identify the top-emerging significant variables influencing one's decision to leave the company. Lastly, the hypothesis this project is interested to verify is: JobSatisfaction, a potential indicator of attrition based on existing subject knowledge, is a significant variable in the prediction of attrition.

Data Cleaning & Data Exploration

Based on preliminary data exploration, the dataset contains 1470 observations and 35 variables, as shown in Figure 1. It was also noted that 7 of the highlighted numerical variables were categorical ordinal variables encoded (refer to Appendix for more details).

Numerical Variables (26)		Categorical Variables (9)
Age	Number of Companies Worked	Attrition
Daily Rate	Percent Salary Hike	Business Travel
Distance From Home	Performance Rating	Department
Education	Relationship Satisfaction	Education Field
Employee Count	Standard Hours	Gender
Employee Number	Stock Option Level	Job Role
Environment Satisfaction	Total Working Years	Marital Status
Hourly Rate	Training Times Last Year	Over 18
Job Involvement	Work Life Balance	Over Time
Job Level	Years at Company	
Job Satisfaction	Years in Current Role	
Monthly Income	Years Since Last Promotion	
Monthly Rate	Years With Current Manager	

Figure 1: A table showing the Categorical and Numerical variables

The dataset was then observed to have no missing values, and variables such as 'EmployeeCount', 'Over18' and 'StandardHours' were removed as all the observations have the same constant value. Additionally, the variable 'Employee Number' was also removed since it is solely an identifier. Next, for further analysis and statistical modelling, the response variable, 'Attrition', was numerically encoded with 'Yes' and 'No' represented by 1 and 0 respectively. The correlation matrix of the numerical variables (excluding the 7 encoded variables) and attrition was then retrieved, with MonthlyIncome & Job Level emerging with the highest correlation of 0.950 as highlighted in Figure 2. With that, it can be foreseen that variable selection may potentially lead to one of them being removed from the final model.

	Age	Attrition	DailyRate	DistanceFr	HourlyRat	JobLevel	MonthlyIn	MonthlyRa	NumComp	PercentSal	StockOptic	TotalWork	TrainingTi	YearsAtCo	YearsInCu	YearsSinc	YearsWith
Age	1	-0.15921	0.010661	-0.00169	0.024287	0.509604	0.497855	0.028051	0.299635	0.003634	0.03751	0.680381	-0.01962	0.311309	0.212901	0.216513	0.202089
Attrition	-0.15921	1	-0.05665	0.077924	-0.00685	-0.1691	-0.15984	0.01517	0.043494	-0.01348	-0.13714	-0.17106	-0.05948	-0.13439	-0.16055	-0.03302	-0.1562
DailyRate	0.010661	-0.05665	1	-0.00499	0.023381	0.002966	0.007707	-0.03218	0.038153	0.022704	0.042143	0.014515	0.002453	-0.03405	0.009932	-0.03323	-0.02636
DistanceFromHome	-0.00169	0.077924	-0.00499	1	0.031131	0.005303	-0.01701	0.027473	-0.02925	0.040235	0.044872	0.004628	-0.03694	0.009508	0.018845	0.010029	0.014406
HourlyRate	0.024287	-0.00685	0.023381	0.031131	1	-0.02785	-0.01579	-0.0153	0.022157	-0.00906	0.050263	-0.00233	-0.00855	-0.01958	-0.02411	-0.02672	-0.02012
JobLevel	0.509604	-0.1691	0.002966	0.005303	-0.02785	1	0.9503	0.039563	0.142501	-0.03473	0.013984	0.782208	-0.01819	0.534739	0.389447	0.353885	0.375281
MonthlyIncome	0.497855	-0.15984	0.007707	-0.01701	-0.01579	0.9503	1	0.034814	0.149515	-0.02727	0.005408	0.772893	-0.02174	0.514285	0.363818	0.344978	0.344079
MonthlyRate	0.028051	0.01517	-0.03218	0.027473	-0.0153	0.039563	0.034814	1	0.017521	-0.00643	-0.03432	0.026442	0.001467	-0.02366	-0.01281	0.001567	-0.03675
NumCompaniesWorked	0.299635	0.043494	0.038153	-0.02925	0.022157	0.142501	0.149515	0.017521	1	-0.01024	0.030075	0.237639	-0.06605	-0.11842	-0.09075	-0.03681	-0.11032
PercentSalaryHike	0.003634	-0.01348	0.022704	0.040235	-0.00906	-0.03473	-0.02727	-0.00643	-0.01024	1	0.007528	-0.02061	-0.00522	-0.03599	-0.00152	-0.02215	-0.01199
StockOptionLevel	0.03751	-0.13714	0.042143	0.044872	0.050263	0.013984	0.005408	-0.03432	0.030075	0.007528	1	0.010136	0.011274	0.015058	0.050818	0.014352	0.024698
TotalWorkingYears	0.680381	-0.17106	0.014515	0.004628	-0.00233	0.782208	0.772893	0.026442	0.237639	-0.02061	0.010136	1	-0.03566	0.628133	0.460365	0.404858	0.459188
TrainingTimesLastYear	-0.01962	-0.05948	0.002453	-0.03694	-0.00855	-0.01819	-0.02174	0.001467	-0.06605	-0.00522	0.011274	-0.03566	1	0.003569	-0.00574	-0.00207	-0.0041
YearsAtCompany	0.311309	-0.13439	-0.03405	0.009508	-0.01958	0.534739	0.514285	-0.02366	-0.11842	-0.03599	0.015058	0.628133	0.003569	1	0.758754	0.618409	0.769212
YearsInCurrentRole	0.212901	-0.16055	0.009932	0.018845	-0.02411	0.389447	0.363818	-0.01281	-0.09075	-0.00152	0.050818	0.460365	-0.00574	0.758754	1	0.548056	0.714365
YearsSinceLastPromotion	0.216513	-0.03302	-0.03323	0.010029	-0.02672	0.353885	0.344978	0.001567	-0.03681	-0.02215	0.014352	0.404858	-0.00207	0.618409	0.548056	1	0.510224
YearsWithCurrManager	0.202089	-0.1562	-0.02636	0.014406	-0.02012	0.375281	0.344079	-0.03675	-0.11032	-0.01199	0.024698	0.459188	-0.0041	0.769212	0.714365	0.510224	1

Figure 2: A correlation matrix of numerical variables and 'Attrition'

From the summary statistics, it was found that the attrition rate of the company was around 16%. Exploring the levels of JobSatisfaction, which is the focus of this report's hypothesis, it appears that higher JobSatisfaction comes with lower attrition. However, more analysis and the use of statistical modelling needs to be done to confirm this. Other summary statistics, shown in Figure 3, suggested that a higher monthly income can potentially serve as a good incentive to keep employee attrition low.

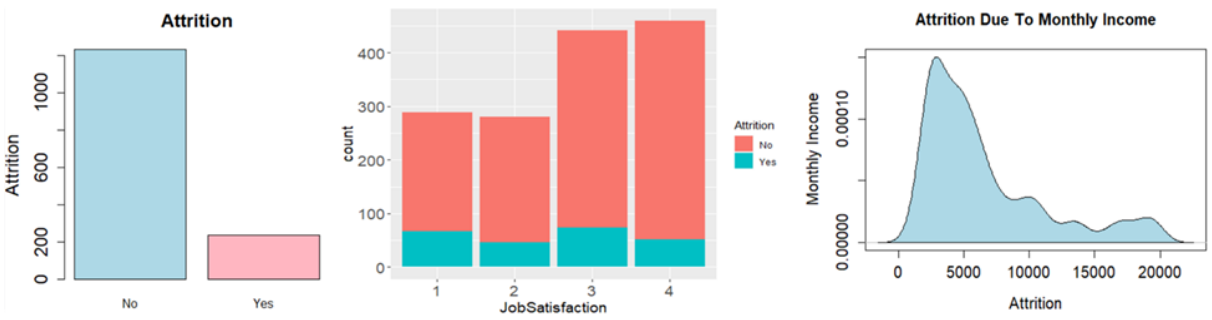


Figure 3: Descriptive plots on some of the variables in the dataset

Synthetic Minority Oversampling Technique (SMOTE)

After data cleaning and exploration, the cleaned dataset was then split equally for training and testing, with each having a size of 735. However, it was observed that the training dataset was unbalanced, with around 606 non-attrition and 129 attrition observations. As this might adversely affect the evaluation metrics, the observations of attrition from the training dataset were up-sampled using the Synthetic Minority Oversampling Technique (SMOTE) algorithm. This technique creates synthetic samples from the minority class, attrition, by randomly choosing one of the k-nearest-neighbours to create a similar but randomly tweaked new observation. As such, the final training dataset has a total of 903 observations, with 516 non-attrition and 387 attrition observations.

Logistic Regression (LR)

As this is a classification prediction problem, a logistic regression model was first fitted onto the training dataset after SMOTE using all the variables and the response variable Attrition. The logistic model obtained from the training dataset was then used to predict attrition in the test dataset, and the results are as shown below. (refer to Appendix for LR on training data without SMOTE)

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	496	29
Attrition (1)	131	79

Accuracy	0.782
Sensitivity	0.731
Specificity	0.791
Area under Curve	0.826

Figure 4: Confusion matrix and performance metrics of the full logistic model

Next, it was observed that not all the variables in the logistic model were significant (p-values < 0.01). As such, variable selection needs to be done so as to reduce overfitting and potentially improve the

evaluation metrics. The Lasso (Least Absolute Shrinkage and Selection Operator) was performed to mitigate this and the results are as shown below.

Lasso (Least Absolute Shrinkage and Selection Operator)

As Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model, variables that contribute to overfitting can be eliminated using Lasso. In addition, it offers a neat way to model the dependent variable while automatically selecting significant variables by shrinking the coefficients of unimportant predictors to zero. (Kailash, 2017) As such, after standardizing the features and tuning for the best lambda parameter for Lasso using cross-validation and predicting attrition of the test dataset, the results of the logistic regression model are as shown below.

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	518	32
Attrition (1)	109	76

Accuracy	0.808
Sensitivity	0.704
Specificity	0.826
Area under Curve	0.819

Figure 5: Confusion matrix and performance metrics of the logistic model after performing lasso

In comparison to the first logistic regression model, the accuracy and specificity obtained from the model after Lasso performed better even though the sensitivity and AUC values were slightly compromised. As a result, the logistic regression with Lasso model is preferred.

Additionally, the logistic regression with Lasso model eliminated insignificant variables such as DepartmentSales, Education, EducationFieldLife Sciences, GenderMale, HourlyRate, JobRoleManager, JobRoleResearch Director, JobRoleSales Executive, MaritalStatusMarried, MonthlyIncome, MonthlyRate and TotalWorkingYears by shrinking their coefficients to zero. It could also be seen that variables such as OverTime and JobInvolvement are deemed significant by the model since their coefficients are relatively large as compared to the other variables.

Decision Trees

Single Decision Tree

Next, this report moves on to more flexible methods such as the Decision Tree. Using the *tree()* function without any parameter adjustments from the *tree* library in R, the Single Decision Tree was

first trained to serve as a benchmark against the Decision Tree Ensemble methods to be used, i.e. Bagging, Random Forests and Bagging. The confusion matrix and values of model evaluation metrics obtained from the Single Decision Tree are as shown below.

Predicted \ Actual	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	410	32
Attrition (1)	217	76

Accuracy	0.661
Sensitivity	0.704
Specificity	0.654
Area under Curve	0.693

Figure 6: Confusion matrix and performance metrics of the Single Decision Tree

Additionally, it was noted that the pruned version of this Single Decision Tree performed slightly worse than the unpruned Single Decision Tree for all the model evaluation metrics considered above.

Random Forests and Bagging

To mitigate the high variance that Decision Trees suffer from, Decision Tree Ensembles such as Bagging and Random Forests were introduced. To compare the evaluation metrics (accuracy in this case) of each model trained, the *randomForest()* function from the *randomForest* library in R was used, where the number of trees (*ntree*) varied from 100 to 4000, in intervals of 100. Additionally, three different values for the maximum number of predictors chosen as split candidates (*mtry)* were considered: the full $p=30$ predictors (for Bagging), and $floor(\sqrt{p})=5$ and $p/2=15$ (for Random Forests). It was found that the accuracy of each of these models stabilised at around *ntree=4000.*

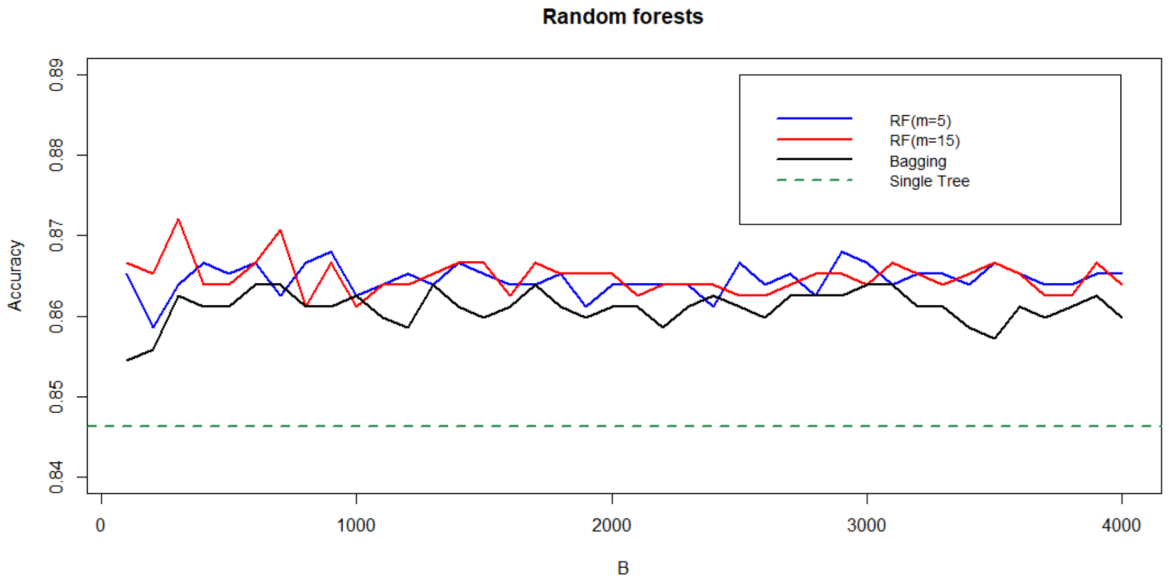


Figure 7: Plot of Accuracy of each Bagging and Random Forests model

In general, there were improvements in the accuracy, specificity and AUC values of the Bagging and Random Forest models, but a decrease in sensitivity when compared to the Single Decision Tree. The results of each model when *ntree=4000* are displayed below. Overall, the Random Forests model with *mtry=15* performed the best, with all metrics each having a higher value than 0.6.

Predicted \ Actual	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	549	48
Attrition (1)	78	60

Accuracy	0.829
Sensitivity	0.556
Specificity	0.876
Area under Curve	0.818

Figure 8: Confusion matrix and performance metrics of Random Forests (mtry=5)

Predicted \ Actual	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	543	42
Attrition (1)	84	66

Accuracy	0.829
Sensitivity	0.611
Specificity	0.866
Area under Curve	0.810

Figure 9: Confusion matrix and performance metrics of Random Forests (mtry=15)

Predicted \ Actual	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	537	44
Attrition (1)	90	64

Accuracy	0.818
Sensitivity	0.593
Specificity	0.856
Area under Curve	0.801

Figure 10: Confusion matrix and performance metrics of Bagging (mtry=30)

The importance of each predictor was then obtained using the *importance()* and *varImpPlot()* functions from the *randomForest* package, where the ordering of importance of each predictor is determined by the total decrease in the Gini index due to splits over the predictor, averaged over the total number of trees. From the variable importance plots and top 10 most important variables of each model obtained, it can be seen that the features JobRole, Age, Overtime, and DailyRate are consistently ranked within the top five most important features in each of the aforementioned models, while MonthlyIncome, Distance, MonthlyRate, and HourlyRate are consistently ranked within the top ten. (refer to Appendix for the various plots obtained)

Boosting

Next, Boosting was also used to attempt to mitigate the problem of high variance of the Single Decision Tree model. To train the Gradient Boosted Trees, the *gbm()* function was used. Similar to Bagging and Random Forests, the number of trees (*n.trees*) were set from 100 to 4000, in intervals of 100. Additionally, different combinations of the number of splits in each tree (*interaction.depth, d*)

and learning rate (*shrinkage, s*) were also used, where *d* was set to either 1 or 2, and *s* was set to either 0.1 or 0.01. Although the accuracy was not as high with a larger number of trees compared to a smaller number of trees, other metrics such as sensitivity increased as the number of trees increased. There were also no signs of significant overfitting as the number of trees increased.

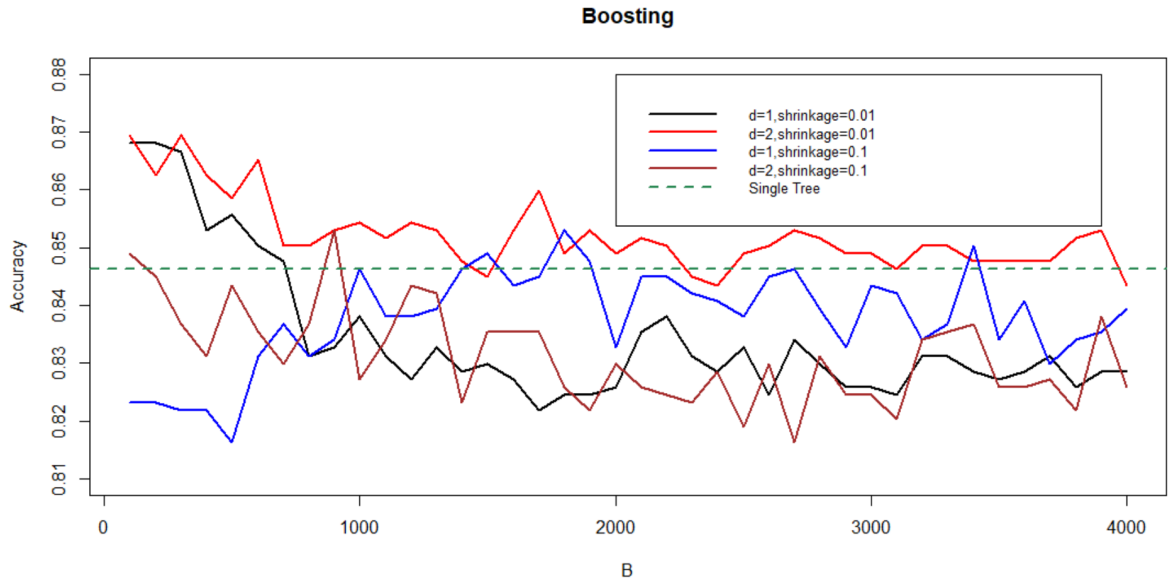


Figure 11: Plot of Test Errors for each Gradient Boosted Trees model

It was found that compared to the Single Decision Tree, every Gradient Boosted Trees model performed better in every model evaluation metric. The results of each model when *n.tree=4000* are displayed below. Of all these models, the model with *s=0.1* and *d=1* performed the best overall, with all metrics each having a higher value than 0.7.

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	536	36
Attrition (1)	91	72

Accuracy	0.827
Sensitivity	0.667
Specificity	0.855
Area under Curve	0.851

Figure 12: Confusion matrix and performance metrics of Gradient Boosted Trees model (s=0.01, d=1)

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	552	37
Attrition (1)	75	71

Accuracy	0.848
Sensitivity	0.657
Specificity	0.880
Area under Curve	0.859

Figure 13: Confusion matrix and performance metrics of Gradient Boosted Trees model (s=0.01, d=2)

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	536	32
Attrition (1)	91	76

Accuracy	0.833
Sensitivity	0.704
Specificity	0.855
Area under Curve	0.826

Figure 14: Confusion matrix and performance metrics of Gradient Boosted Trees model (s=0.1, d=1)

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	541	41
Attrition (1)	86	67

Accuracy	0.827
Sensitivity	0.620
Specificity	0.863
Area under Curve	0.826

Figure 15: Confusion matrix and performance metrics of Gradient Boosted Trees model (s=0.1, d=2)

From the variable importance plots and top 10 most important variables of each model obtained, it can be seen that the features JobRole and YearsSinceLastPromotion are consistently ranked within the top five most important features in each of the aforementioned models, while *Overtime* is consistently ranked within the top ten. (refer to Appendix for the various plots)

To conclude this section, out of all the Decision Tree Ensembles, it appears that the Gradient Boosted Trees model with *s=0.1, d=1* performed the best, with all metrics except sensitivity each having a higher value than 0.8.

Support Vector Machines

Next, since the attrition response variable is binary with 2 classes, Support Vector Machine (SVM) with different kernels, namely linear, radial and polynomial were also experimented. The model obtained from the training dataset was then used to predict attrition in the test dataset.

For SVM with linear kernel, a log grid of 0.001 to 100 was used to tune for the cost parameter. The optimal cost and error obtained were 0.04641 and 0.207 respectively, which gave a model with 508 support vectors. The accuracy, sensitivity, specificity and AUC values obtained from the test dataset using this model are as shown below.

Actual / Predicted	Non-Attrition (0)	Attrition (1)
Non-Attrition (0)	522	35
Attrition (1)	105	73

Accuracy	0.810
Sensitivity	0.676
Specificity	0.833
Area under Curve	0.817

Figure 16: Confusion matrix and performance metrics of the linear SVM model

For SVM with radial kernel, gamma values of 0.5, 1, 2, 3 and 4, and a log grid of 0.001to 100 were similarly used to tune for the respective parameters. The optimal cost and error obtained were 2.154435 and 0.1796 respectively, corresponding to a radial kernel with gamma = 0.5. On the test

dataset, the model gave an accuracy of 0.842, sensitivity of 0.0741 and specificity of 0.974. . However, it is also crucial to note that this model used all the training observations as support vectors.

For SVM with polynomial kernel, degrees 2, 3 and 4, and a log grid of 0.001to 100 were used to tune for the respective parameters.. The optimal cost and error obtained were 27.82559 and 0.0842 respectively, which corresponds to a polynomial kernel with degree = 4. In addition, this model uses 573 support vectors and has an accuracy of 0.741, sensitivity of 0.5 and specificity of 0.783 on the test dataset.

As it was noted that the sensitivity of SVM with radial and polynomial kernels were significantly less than 0.8 relative to the SVM with linear kernel, these two models will not be taken into consideration for model selection.

Model Selection

Although none of the above statistical models achieved 0.800 and above for all evaluation metrics, based on the accuracy, sensitivity, specificity and AUC values, the Logistic regression with Lasso and Gradient Boosted decision tree (with s=0.1, d=1) seemed to predict attrition with the highest metric values. Generally, Logistic regression works well for many business applications which have a simple decision boundary. Moreover, because of its simplicity it is less prone to overfitting than flexible methods such as decision trees. Furthermore, as this report has shown, variables that contribute to overfitting can be eliminated using lasso regularisation, without compromising out-of-sample accuracy. Given these advantages and its inherent simplicity, although it was observed that the Gradient Boosted decision tree had slightly better metric values than the Logistic regression with Lasso, the Logistic regression with Lasso model was chosen to be the final model, with accuracy of 0.808 sensitivity 0.704, specificity 0.826 and AUC 0.819.

Checking Logistic Model Assumptions

The logistic regression method assumes that there is a linear relationship between the logit of the outcome and predictor variables, that there are no influential values and no high multicollinearity among the predictors. As such, the following parts of this section shall show that these assumptions

hold true for the dataset. Firstly, the linear relationship between the logit of the outcome and each quantitative predictor variable can be checked by visually inspecting the scatter plot between each predictor and the logit values. The scatter plots obtained showed that variables such as Age and YearsAtCompany are quite linearly associated with the outcome in logit scale, whereas variables such as MonthlyRate are not as linearly related and might need some transformations. However, since the Lasso model takes into account variables that are insignificant and contribute to overfitting, variables such as Monthly rate were removed in the final model. Next, influential values can be examined by obtaining the Cook's distance values. However, though about 64 influential points in the dataset were detected using Cook's distance, they were not removed due to a lack of subject knowledge and to avoid creating a perfect separation for the logistic regression model. Lastly, multicollinearity was assessed by examining the variance inflation factors (VIF). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In the dataset, all predictor variables have a value of VIF well below 10 except the variables Department, JobLevel, JobRole and MonthlyIncome. However, as seen from the section Logistic Regression on Lasso, these variables were removed from the model as they were deemed insignificant and contribute to overfitting. As such, this section thus showed that the logistic regression model assumptions indeed hold true.

Conclusion & Recommendations

Based on the variables chosen from the logistic regression model with lasso, variables such as OverTime and JobInvolvement are strong indicators of Attrition. As such, a company could look into providing more manpower so as to lower the need to overtime and potential employee burnouts. Profiling can also be done to draw out the strengths and weaknesses of employees. Suitable work tasks can then be subsequently allocated to encourage more job involvement, engagement and ownership. Lastly, contrary to existing subject knowledge, JobSatisfaction is not a significant indicator in the prediction of employee attrition. However, as job satisfaction is a subjective variable and there is a chance that employees might not be truthful in data collection surveys in the real world, it is powerful to note that our final prediction model obtained for this project does not rely on this variable.

Citations

Yee, Y. (2020, February 04). Survey shows talent retention is a big challenge for Singapore's science and technology industry. Retrieved April 02, 2021, from https://www.straitstimes.com/tech/survey-shows-talent-retention-is-a-big-challenge-for-singapores-science-and-technology-industry

Kailash, A. (2017, October 06). A gentle introduction to logistic regression and lasso regularisation using r. Retrieved April 02, 2021, from https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

Appendix

Categorical Ordinal Variables

7 of the highlighted numerical variables were categorical ordinal variables encoded.

Numerical Value / Variable	1	2	3	4	5
Education	Below College	College	Bachelor	Master	Doctor
Environment Satisfaction	Low	Medium	High	Very High	
Job Involvement	Low	Medium	High	Very High	
Job Satisfaction	Low	Medium	High	Very High	
Performance Rating	Low	Good	Excellent	Outstanding	
Relationship Satisfaction	Low	Medium	High	Very High	
Work Life Balance	Bad	Good	Better	Best	

Figure 17: Encoding for the categorical ordinal variables

Outlier Detection

Using Cook's distance, observations that had a cook's distance greater than 5 times the mean were classified as influential points. From the detection, 64 influential points were identified. However, none of the influential points have been removed due to a lack of subject knowledge on the dataset. Also, these points could potentially serve insights in the statistical models we attempt to fit. Hence, SMOTE was performed instead to remedy the situation.

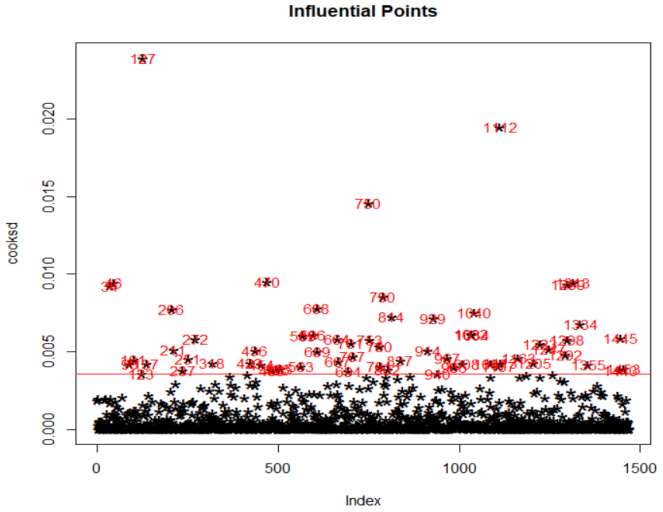


Figure 18: Plot of the influential points

Baseline Model Logistic Regression

As presented in the proposal, a logistic regression model was fitted onto the training dataset without SMOTE, using all the variables, and the results are as shown below.

```
Call:
glm(formula = Attrition ~ ., family = binomial, data = hr.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8416  -0.4627  -0.2467  -0.0847   3.2091

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -9.790e+00  8.345e+02  -0.012 0.990640
ï..Age                         -6.121e-02  2.122e-02  -2.885 0.003919 **
BusinessTravelTravel_Frequently  1.777e+00  5.894e-01   3.015 0.002572 **
BusinessTravelTravel_Rarely      1.100e+00  5.355e-01   2.055 0.039908 *
DailyRate                      -3.891e-04  3.162e-04  -1.231 0.218490
DepartmentResearch & Development 1.329e+01  8.345e+02   0.016 0.987290
DepartmentSales                  1.440e+01  8.345e+02   0.017 0.986238
DistanceFromHome                 4.934e-02  1.631e-02   3.026 0.002477 **
Education                      -1.865e-02  1.283e-01  -0.145 0.884425
EducationFieldLife Sciences    -4.002e-01  1.108e+00  -0.361 0.717955
EducationFieldMarketing        -2.058e-01  1.178e+00  -0.175 0.861365
EducationFieldMedical          -7.346e-01  1.115e+00  -0.659 0.509908
EducationFieldOther            -5.023e-01  1.218e+00  -0.412 0.679981
EducationFieldTechnical Degree   3.082e-01  1.155e+00   0.267 0.789486
EnvironmentSatisfaction        -4.223e-01  1.185e-01  -3.563 0.000366 ***
GenderMale                       9.974e-02  2.578e-01   0.387 0.698792
HourlyRate                       2.815e-03  6.490e-03   0.434 0.664441
JobInvolvement                 -5.196e-01  1.705e-01  -3.047 0.002309 **
JobLevel                       -2.576e-01  4.616e-01  -0.558 0.576754
JobRoleHuman Resources           1.477e+01  8.345e+02   0.018 0.985876
JobRoleLaboratory Technician     1.353e+00  6.433e-01   2.103 0.035456 *
JobRoleManager                 -3.024e-01  1.137e+00  -0.266 0.790180


JobRoleManager                 -3.024e-01  1.137e+00  -0.266 0.790180
JobRoleManufacturing Director  -4.992e-01  7.601e-01  -0.657 0.511363
JobRoleResearch Director       -1.463e+00  1.207e+00  -1.212 0.225372
JobRoleResearch Scientist       6.686e-02  6.706e-01   0.100 0.920574
JobRoleSales Executive         -3.123e-01  1.471e+00  -0.212 0.831905
JobRoleSales Representative     9.546e-01  1.546e+00   0.617 0.536961
JobSatisfaction                -4.479e-01  1.183e-01  -3.785 0.000154 ***
MaritalStatusMarried            5.063e-01  4.031e-01   1.256 0.209081
MaritalStatusSingle             1.344e+00  5.217e-01   2.576 0.010001 *
MonthlyIncome                   7.942e-05  1.184e-04   0.671 0.502224
MonthlyRate                    -4.054e-06  1.778e-05  -0.228 0.819656
NumCompaniesWorked              2.343e-01  5.889e-02   3.979 6.93e-05 ***
OverTimeYes                     2.063e+00  2.831e-01   7.286 3.19e-13 ***
PercentSalaryHike              -4.277e-02  5.700e-02  -0.750 0.453022
PerformanceRating               3.158e-02  5.832e-01   0.054 0.956820
RelationshipSatisfaction       -1.537e-01  1.195e-01  -1.287 0.198228
StockOptionLevel               -2.414e-01  2.401e-01  -1.005 0.314677
TotalWorkingYears              -4.238e-02  4.258e-02  -0.995 0.319507
TrainingTimesLastYear          -1.949e-01  1.046e-01  -1.862 0.062575 .
WorkLifeBalance                -4.287e-01  1.808e-01  -2.372 0.017712 *
YearsAtCompany                  1.001e-01  5.032e-02   1.989 0.046680 *
YearsInCurrentRole             -7.765e-02  6.035e-02  -1.287 0.198218
YearsSinceLastPromotion         1.934e-01  5.708e-02   3.388 0.000703 ***
YearsWithCurrManager           -1.730e-01  5.849e-02  -2.959 0.003088 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 682.84  on 734  degrees of freedom
Residual deviance: 434.24  on 690  degrees of freedom
AIC: 524.24

Number of Fisher Scoring iterations: 15
```

Figure 19: Results of the baseline logistic regression model

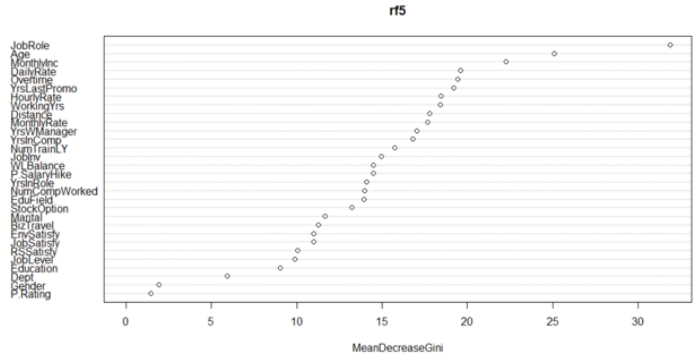
The baseline logistic model obtained was then used to predict attrition in the test dataset, and the results are as shown below.

glm.pred \ hr.test	0	1
0	586	51
1	41	57

Figure 20: Confusion Matrix of baseline logistic regression model

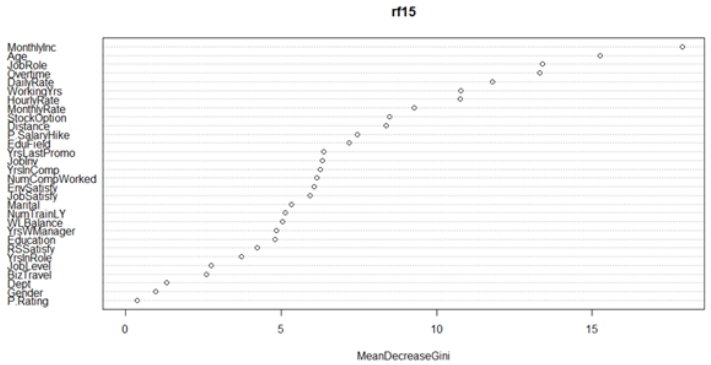
The accuracy obtained from the logistic model is approximately 0.875. Additionally, the sensitivity and specificity values are 0.528 and 0.935 respectively. Although these values are relatively high, problems relating to imbalanced data and insignificant variables needed to be addressed. Methods such as Synthetic Minority Oversampling Technique (SMOTE) and variable selection were hence performed in the main report.

Bagging and Random Forests: Variable Importance Plots and Top 10 Predictors of each model



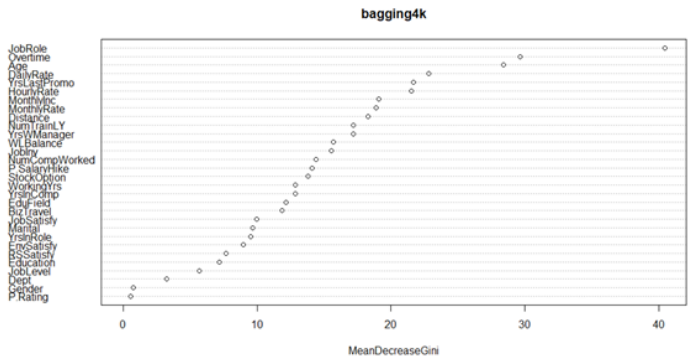
variable	MeanDecreaseGini
JobRole	31.86557
Age	25.09285
MonthlyInc	22.25002
DailyRate	19.59817
Overtime	19.41618
YrsLastPromo	19.18407
HourlyRate	18.45846
WorkingYrs	18.41160
Distance	17.80309
MonthlyRate	17.67387

Figure 21: Variable Importance Plot and Top 10 Important Variables of Random Forests (mtry=5)



variable	MeanDecreaseGini
MonthlyInc	17.908178
Age	15.297974
JobRole	13.428608
Overtime	13.237873
DailyRate	11.702859
WorkingYrs	11.188044
HourlyRate	10.637315
MonthlyRate	9.388303
StockOption	8.774419
Distance	8.326123

Figure 22: Variable Importance Plot and Top 10 Important Variables of Random Forests (mtry=15)



variable	MeanDecreaseGini
JobRole	40.56496
Overtime	30.29969
Age	28.21295
DailyRate	23.44678
YrsLastPromo	21.86563
HourlyRate	21.29123
MonthlyInc	19.04776
MonthlyRate	18.53200
Distance	18.19103
YrsWManager	17.46910

Figure 23: Variable Importance Plot and Top 10 Important Variables of Bagging (mtry=30)

Boosting: Variable Importance Plots and Top 10 Predictors of each model

	var	rel.inf
YrsLastPromo	YrsLastPromo	9.627993
JobRole	JobRole	9.191518
Overtime	Overtime	8.140658
NumTrainLY	NumTrainLY	5.776048
Age	Age	5.320927
WLBalance	WLBalance	5.197257
JobInv	JobInv	4.965278
BizTravel	BizTravel	4.914384
NumCompWorked	NumCompWorked	3.885328
Distance	Distance	3.873367

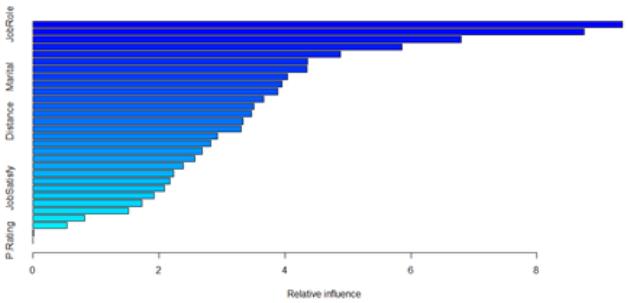


Figure 24: Variable Importance Plot and Top 10 Important Variables of Gradient Boosted Trees (s=0.01, d=1)

	var	rel.inf
JobRole	JobRole	9.360053
YrsLastPromo	YrsLastPromo	8.750302
Overtime	Overtime	6.800626
NumTrainLY	NumTrainLY	5.855046
Age	Age	4.879088
YrsWManager	YrsWManager	4.362052
WLBalance	WLBalance	4.353821
Marital	Marital	4.038260
JobInv	JobInv	3.957332
BizTravel	BizTravel	3.891321

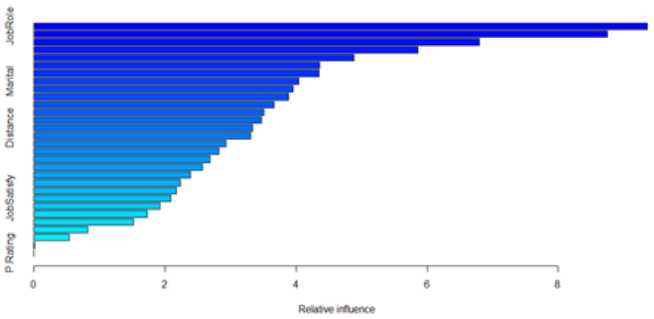


Figure 25: Variable Importance Plot and Top 10 Important Variables of Gradient Boosted Trees (s=0.01, d=2)

	var	rel.inf
JobRole	JobRole	8.406782
YrsLastPromo	YrsLastPromo	7.419572
DailyRate	DailyRate	6.737098
MonthlyRate	MonthlyRate	6.724115
Age	Age	5.806411
NumTrainLY	NumTrainLY	4.897965
MonthlyInc	MonthlyInc	4.461768
HourlyRate	HourlyRate	4.336258
P.SalaryHike	P.SalaryHike	4.043507
Overtime	Overtime	3.979428

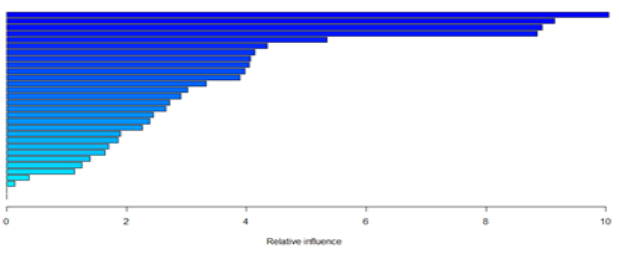


Figure 26: Variable Importance Plot and Top 10 Important Variables of Gradient Boosted Trees (s=0.1, d=1)

	var	rel.inf
JobRole	JobRole	8.636147
YrsLastPromo	YrsLastPromo	8.538167
NumTrainLY	NumTrainLY	6.115464
Overtime	Overtime	5.574974
Age	Age	5.514650
MonthlyRate	MonthlyRate	4.458481
WLBalance	WLBalance	4.066216
BizTravel	BizTravel	3.940564
YrsWManager	YrsWManager	3.803897
DailyRate	DailyRate	3.786392

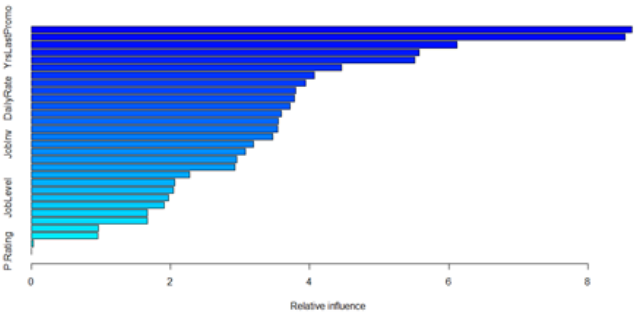


Figure 27: Variable Importance Plot and Top 10 Important Variables of Gradient Boosted Trees (s=0.1, d=2)

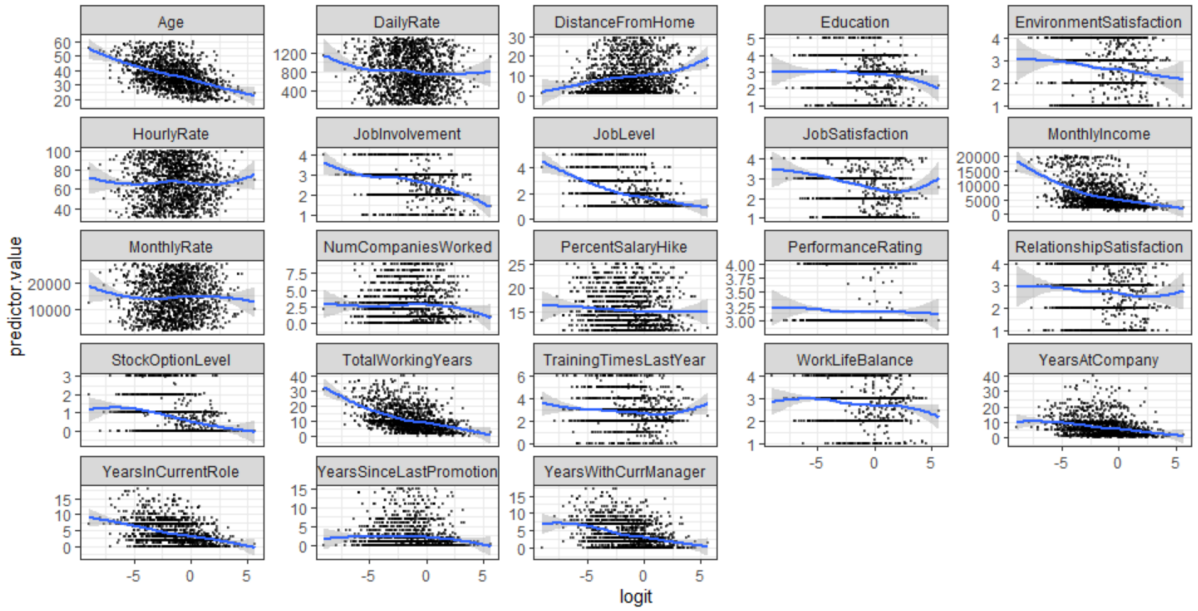


Figure 28: Scatter plots of predictor values VS outcome in logit scale