Assignment 2 – Predictive Analytics

Contents

[Executive Summary: 1](#_Toc493707237)

[Data Exploration + Prep: 2](#_Toc493707238)

[Discover Relationships / Cluster Analysis: 3](#_Toc493707239)

[Create Models in SAS: 4](#_Toc493707240)

[Evaluate Models: 5](#_Toc493707241)

[Solution: 8](#_Toc493707242)

# Executive Summary: [One Page] - 10 marks

The main objective of this report is to analyse employee statistics of the health service company, “CHS Care” in order to recognize which employees are most valuable and then be able to predict which employees are most likely to leave.

Employee value is defined by Minchington as: “… the skills, capabilities and experiences an employee brings to the organization.” (Minchington, 2010) In the case study that we have analysed, we have found that the variables which best prescribe Employee value are “JobInvolvement” and “PerformanceRating”.

Following our analysis we have found that CHS should focus on its youngest employees as they are the most valuable, but also the most likely to leave the company. Part of the reason for this is the increased number of workers who are working overtime; we believe that these workers feel undervalued and that is the primary reason they are leaving.

To combat leaving workers, we propose that CHS do the following: Offer quick promotions to Junior members of staff, offer pay rises alongside promotions and hire more workers to reduce the number of employees working overtime.

# Data Exploration + Prep: [One Page] - 30 marks

|  |  |
| --- | --- |
| Graph 1 – “Distributions of all Variables” | Graph 2 – “Distribution of Attrition” |

The distributions of the data that will be used in this report are explored in *Graph 1 – “Distributions of all Variables”*. The majority of records had nearly no missing values, although “RefsGivenLast12Months” had very few records; therefore we dropped the variable from any further analysis. Similarly, the variables “Over18”, “StandardHours”, “EmployeeCount” and “EmployeeNumber” were all pretty arbitrary so we also dropped them from any further analysis.

The best variables to measure Employee worth are likely “JobInvolvement” and “PerformanceRating”. Performance rating only had values of 3 and 4, therefore we decided that it was effectively a binary variable than indicated if an employee was valuable or not.

The primary variable that is being analysed in this project is “Attrition.” This is a binary variable with the values “Yes / No” that indicate whether an employee has left employment during the observation period, often referred to as “Churning”. By analysing Attrition we aim to discover which variables help drive an employee’s decision to leave and thus be able to offer recommendations to reduce those characteristics and therefore keep the employee working.

*Graph 2 – “Distribution of Attrition”* shows the distribution of Attrition. Unsurprisingly most employees did not leave CHS in the observation period, however there were 237 records with “Yes” values; this means that 16% of the workforce left during the observation period, a significant amount.

We experimented with transforming some of the data, however this had not much of an impact on the final result so we decided to exclude any transformations from the final model. We tried performing log transformations on the variables “Distance” and “MonthlyIncome” as both of the variables had a smaller number of high scoring results, however ultimately this didn’t help the predictive power much.

# Discover Relationships / Cluster Analysis: [One Page] - 10 marks

|  |  |
| --- | --- |
| Graph 3 – “Cluster Segments” | Graph 4 – “Cluster Distance Graph” |
| Graph 5 – “Attrition – Variable Worth Graph” | |

In order to understand how different groups of employees were organised, we undertook some cluster analysis. The data was first split into 4 segments as shown in *Graph 3 – “Cluster Segments”*. The clustering was slightly flawed in that groups 1/3 + 2/4 were very similar, as *Graph 4 – “Cluster Distance Graph”* illustrates; arguably they could have been merged into 2 groups. The clusters were primarily grouped by employee age (in order running youngest to eldest): 2, 4, 1, 3. This clustering by age suggests that when targeting groups of employees it is best to target entire age groups.

Using Segment Profiling we were able to generate *Graph 5 – “Attrition – Variable Worth Graph”*. This graph suggests that the main variables which drive employee churn are: “TotalWorkingYears”, “YearsAtCompany”, “JobLevel”, “YearsInCurrentRole”. Using the Cluster analysis, we were able to establish a few insights.

Firstly, those who had worked for longer, and or were older, tended to churn slightly less. This suggests that older workers tend to stay in the same job for longer as they are likely paid well and value the stability of staying in the same job, to pay mortgages etc. On the other side, new workers are at high risk of churning and should be focused on to retain employee numbers.

Secondly, those who earned higher monthly wages were less likely to churn; this suggests that the health service industry has high wage competitiveness

Thirdly, the variables “YearsWithCurrManager”,“YearsInCurrentRole” and “JobLevel” all had decent contributions. This suggests that employees greatly valued a sense of progression in their careers, and if they felt they were stagnating, they would likely churn and look to move elsewhere.

Finally we found that workers who were younger were more likely to work Overtime and had higher PerformanceRating scores.

# Create Models in SAS: [Two Pages] - 10 marks

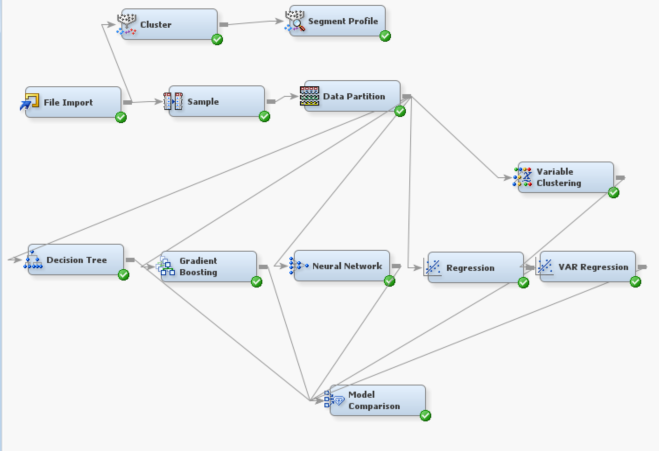


Figure 1 – “SAS Enterprise Miner Workflow”

To predict when an employee would churn, we created several models in SAS and compared their performance to find the model with the most predictive power, resulting in the final workflow shown in *Figure 1 – “SAS Enterprise Miner Workflow”*. We chose to use “Attrition” as the target variable and Misclassification Rate as my target assessment for each model as this is the best way of measuring a model’s predictive power for a binary variable. We first used the data partition module to split the data set into Training and Validation sets of 70% and 30% respectively; this was done so that the models could be tested on data that it had not been trained on, making it a fair test. We then began adding models to use these data sets and all report into a model comparison node at the end to find the best model.

The first model we added was a Decision Tree, set to “Assessment” construction method to prune the tree to create the simplest model with a high assessment value. Next we created a Gradient Boosting model with default settings, a Neural Network and a Regression. Finally we added another Regression model, however we placed a Variable Clustering node before the model in order to improve the predictive power of the variables

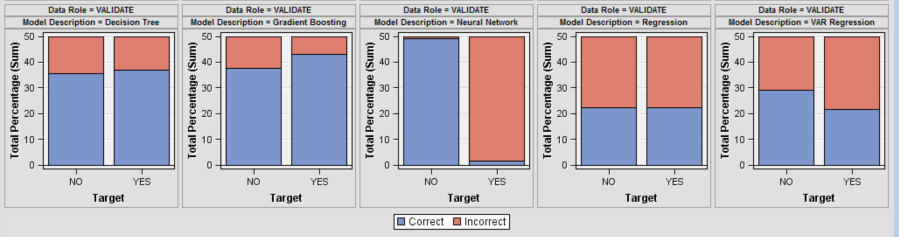
In order to prevent over-training on records of employees who don’t churn, we used Under-Sampling to have an equal number of Churning and Non-Churning records in the training data set. To do this, we used the Sampling node set to “Equal” strata (aka a balanced number of records for both “Yes” and “No”) derived from 100% of the data. Whilst this did reduce the number of records in our analysis, it gives much more balanced dataset that has equal weighting of employees who do and don’t churn, therefore improving the training performance of the models and reducing bias in “Attrition”.

In order to improve the regression performance, we first set the regressions to run Logistic regressions and then used the variable clustering node on the second regression to group together similar variables; ultimately however this had limited impact as the regressions still performed poorly.

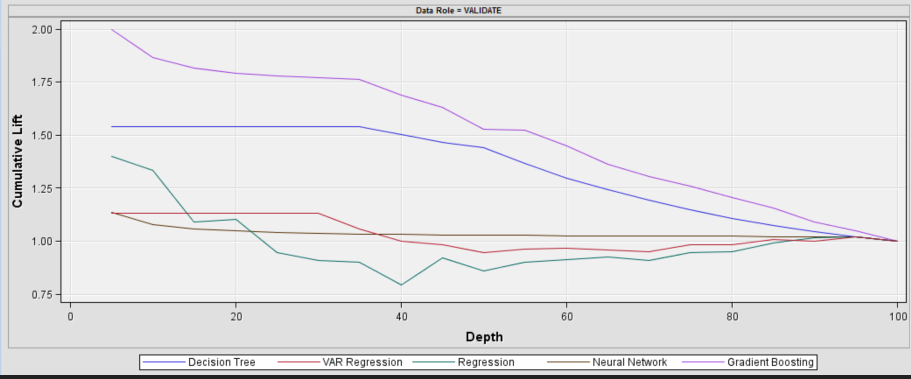
# Evaluate Models: [Three Pages] - 15 marks

The following table displays the results of the model comparison node for each iteration of the model building phase. The results on the right hand side of the table are for the validation data set and best performing model as described in the “Best Model” column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model Iteration* | *ROC Chart* | *Best Model* | *Misclassification Error* | *Average Square Error* |
| 1  (Initial) | C:\Users\Rob\Documents\Work\Deakin\Deakin - T2\Predictive Analytics\Assignment 2\Rob - Outputs\Initial Model ROC Chart _ Attrition.png | Gradient Boosting | *0.14* | *0* |
| *False Negatives* |  |
| *63* |  |
| *2*  *(Under-Sampling)* |  | *Best Model* | *Misclassification Error* | *Average Square Error* |
| *Gradient Boosting* | *0.19* | *0.16* |
| *False Negatives* |  |
| *10* |  |



Graph 6 – “Classification Chart (Iteration 2)”



Graph 7 – “Cumulative Lift Chart”

In the initial iteration of Models (1), we achieved ok performance for most models. Iteration 2 utilised Under-Sampling and whilst this did have the negative impact of slightly increasing the overall Misclassification rate and Average Square Error; it gave better False Negative results and the ROC chart line improved. Comparing the ROC charts for the Train and Validate data sets we can see that the Train data set is closer to the top left corner than the Validate graph; this suggests that the models are over training. Using the ROC chart to compare performance between models, we can see that Gradient Boosting has the highest Sensitivity and lowest Specificity scores; followed by Decision tree, Regression and lastly Neural Network.

Interpreting *Graph 7 – “Cumulative Lift Chart”,* we can see that the Gradient Boosting is over-performing and is much more sensitive than the other models.

With regards to the model’s misclassifications, having higher False Negative results would be far more costly to CHS than the higher False Positive results. This is because it is better to be over cautious and put more employees through retention schemes than it is to unexpectedly have a valuable member of staff leave the company. Therefore we are aiming to primarily reduce False Negative results. As shown in *Graph 6 – “Classification Chart (Iteration 2)”*, Gradient Boosting was correct most of the time. However if the objective was primarily to get the lowest number of Incorrect “No” churn predictions, then perhaps Neural Networks could be considered; the model managed to highlight nearly all of those who would churn, most likely through.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Variable | Importance | | OverTime | 1 | | Age | 0.913 | | JobRole | 0.898 | | EnvironmentSatisfaction | 0.846 | | JobLevel | 0.689 | | StockOptionLevel | 0.684 | | YearsWithCurrManager | 0.683 | | JobInvolvement | 0.608 | | YearsAtCompany | 0.532 |   Table – “Gradient Boosting – Variable Importance” | |  |  |  |  | | --- | --- | --- | --- | |  |  | Overtime | | |  |  | Yes | No | | Attrition | Yes | 127 | 289 | | No | 110 | 944 |   Table – “Matrix of Attrition / Overtime” |

If we choose Gradient Boosting as the most powerful model, analysis of the Variable Importance may help us understand which variables have an influence on Employee Churn. As shown in Table 1 – “Gradient Boosting – Variable Importance”, the model ranks “Overtime”, “Age” and “Job Role” as the most important factors. These make logical sense:

* Those who tend to work Overtime are likely to be more stressed and look to change their job. As shown in *Table 2 – “Matrix of Attrition / Overtime”* when an employee is not working any overtime, they are fairly unlikely to Churn; however of those who do work overtime, it is an even split between those who churn and those don’t. This may suggest that some employees are being overworked and so “Burning out” and leaving the job.
* The high importance of the “Age” variable means model agrees with the initial Cluster Analysis.
* Job role also likely plays a large part of determining an employee’s chances of leaving as if employees are having a tough time, it is likely the sentiment is shared by the rest of their department too. If employees were to leave, this would have an effect on the rest of the employees who are in a similar job role.

# Solution: [One Page] - 10 marks

Using the Gradient Boosting model, we should be able to predict if an employee is likely to quit in the future. By keeping data on current employees up to date, we could run the data through the model on a monthly basis and check to see if the total number of “Yes” values in “Attrition” changes. If so, the model would be predicting that an employee will churn, giving CHS a small timeframe to make some changes and / or offer some incentives to continue working.

In order to minimize the loss of workers, CHS needs to focus on the following points:

* Aim efforts at retaining employees at the youngest members of the staff, as these tend to be the workers that are most at risk, whilst also being the hardest workers and most valuable to the company.
* Look into offering quick promotions to new employees so that they feel that they are progressing at the company and have some motivation to stay.
* Alongside promotions, offer pay rises to those are at risk of Churning, as “MonthlyIncome” has a strong effect on “Attrition” and may encourage workers to stay.
* As a lot of workers who work Overtime are churning, look into hiring more employees to support these workers / spread the workload and thus reducing Overtime workers and therefore reduce churn.

# References:

 Minchington, B (2010) Employer Brand Leadership – A Global Perspective, Collective Learning Australia.