

***BC2407 Analytics II: Advanced Predictive Techniques***

***Semester 2, AY 2017/18***

**Semester Team Project**

|  |  |  |
| --- | --- | --- |
| Prepared by | **Chong Shu Yi**  **Joey Tay Yi Qin**  **Yvonne Yeo**  **Toh Xing Yu**  **Tan Zhi Jie** | U1610102J  U1610214C  U1610963K  U1510365G  U1610104C |
| Seminar Group | **01** | |
| Team | **02** | |
| Tutorial Instructor | **Professor Neumann Chew** | |
| Date of Submission | **14 April 2018** | |

# 

[EXECUTIVE SUMMARY 2](#_Toc511570204)

[1 BUSINESS PROBLEM 3](#_Toc511570205)

[1.1 The Indian Labour Market 3](#_Toc511570206)

[1.2 Business Problem Statement 4](#_Toc511570207)

[1.3 Desired Business Outcomes 4](#_Toc511570208)

[INTRODUCTION TO THE ANALYTICS SOLUTION 5](#_Toc511570209)

[2 DATA PREPARATION 5](#_Toc511570210)

[2.1 Reclassification of Job Designation 5](#_Toc511570211)

[2.2 Choosing Variables 6](#_Toc511570212)

[2.3 Missing Values 7](#_Toc511570213)

[2.4 Final Cleaning 7](#_Toc511570214)

[3 VARIABLE SELECTION 8](#_Toc511570215)

[3.1 Multivariate Adaptive Regression Splines (Mars) 8](#_Toc511570216)

[3.2 CART 9](#_Toc511570217)

[3.3 Results 11](#_Toc511570218)

[4 NEURAL NETWORK 13](#_Toc511570219)

[4.1 Data Import & Partition Nodes 13](#_Toc511570220)

[4.2 Neural Network Node 14](#_Toc511570221)

[5 IMPLEMENTATION AND IMPLICATIONS 17](#_Toc511570222)

[5.1 Short Term Strategy - A Hybrid Portal 17](#_Toc511570223)

[5.2 Long Term Strategy - A Career Portal 17](#_Toc511570224)

[6 LIMITATIONS & FUTURE IMPROVEMENTS 18](#_Toc511570225)

[6.1 Classification of Jobs 18](#_Toc511570226)

[6.2 Predicting Different Areas 18](#_Toc511570227)

[6.3 Premium Service 18](#_Toc511570228)

[APPENDIX 19](#_Toc511570229)

[A. Data Dictionary 19](#_Toc511570230)

[B. Misclassification Rate of Variable Selection Models on Testset 21](#_Toc511570231)

[REFERENCES 22](#_Toc511570232)

# 

# 

# EXECUTIVE SUMMARY

***Opportunity***

The Indian IT labour market relies heavily on job portals to allocate human resources efficiently and effectively, a task made more difficult due to the sheer size of the Indian economy. However, as human resource requirements shift towards heavier emphasis on soft skills and interpersonal aptitude as evaluation criteria of candidate potential, traditional job portals are slow to adapt existing functionality to facilitate this mode of screening, i.e traditional resumes are still the primary form of data collected. Consequences of the the shift in hiring requirements include worsening attrition rates and wasted costs in inefficient hiring.

This creates an opportunity for our team to infuse HR analytics into a traditional job portal. Our team proposes analysing the suitability of candidates for different job positions based on market-wide available information that has thus far been underutilised. This information will span the domains of background, cognitive and personality traits, to both meet the present expectations of hirers and investigate the depth of correlation between personality traits and job suitability. The scope of our proposal is limited to modelling only for the IT and engineering sector in India.

***Approach & Outcome***

Our analytics solution aims to develop a model capable of accurately predicting a candidate’s suitability for multiple job positions. The desired output are the predicted probabilities of suitability. A dataset of four thousand Indian graduates employed in the IT and engineering sector is used for our study.

The core design is built around a Neural Network in the SAS Enterprise Miner environment. To overcome a neural network’s lack of variable selection functionality, we first performed variable selection using Multivariate Adaptive Regression Splines (MARS) models and Decision Trees. The models are compared on Misclassification Rate and Improvement to the Neural Network (when the variables in the respective models are used). The Random Forest model produced the best results.

With a best subset of input variables, the neural network is developed and optimised in SAS EM. Our final model gives us a very favourable Misclassification Rate of only 20.9%.

The implementation plan is to provide candidates information on how suitable they are for different roles, and hiring companies - our customers - the ability to view the applicants for their job listings in a list ranked by the candidates’ suitability for the role.

# 1 BUSINESS PROBLEM

## 1.1 The Indian Labour Market

In the Indian market, internet job boards are responsible for producing the highest amounts of hires for Indian companies. The usage of such job boards that incorporate resume databases are relatively more prevalent there as compared to India’s global counterparts. It is estimated that, in India, 47% of the hires come from internet resume databases whilst that of the global population stands at 26% (Gager et. al., 2015). Therefore, the online job portal market is a sizeable one and can be said to be one of the main drivers for recruitment in India.

However, the current business proposition of the traditional job portal is highly irrelevant in the face of the changing landscape of human resource needs in India’s IT industry. A traditional job portal is one that is characterized by being a one-stop search site for job listings and usually incorporates the *Apply* *Now* function which allows for applicants to upload their resumes for the perusal of recruiters (Matas, 1993). Therefore, traditional job portals usually have a resume database of their users. Recruiters from the various companies then scour through these resumes and then pick candidates from there.

The Indian IT industry is moving up the value chain: instead of engaging in jobs such as manufacturing, a lot of major Indian technological companies are looking to become innovation and research hubs for blockchain technology, artificial intelligence etc (IBEF, 2018). This therefore leads to an emergence of demand for soft-skills amongst the staff of the IT industry. Entry level staff are expected to display an extremely high standard of teamwork as a project’s success is highly contingent on it (Agrawal & Thite, 2003). After a few years, these entry level staff are then expected to move into managerial positions that deviate greatly from the technical nature of their entry level job scope.

This shift has led to significant attrition rates amongst the clients we have in the IT industry. The shift from a highly technical job scope to one that is non-technical in nature is deemed to be a downgrade from the employee’s perspective despite the more attractive pay packages and benefits. This is attributed to the widely held belief that having a technical-based IT job in India is prestigious. Therefore, when the technical employees are promoted into managerial positions, they would leave the company in search for a new job that aligns with their interests. Many Indian technological companies incur significant costs in grooming technical talent that eventually leave the company. As such, many are on the lookout for potential ‘hybrid managers’ which are characterised by traits such as extraversion (Lebowitz, 2016).

Therefore, the supply of resumes coming from such sites are irrelevant as they usually include a great deal of technical skills and hardly touch on the soft skills applicants possess. This is evidenced by the staff’s reluctance to give up their technical careers for managerial positions, implying that technical skills are their main selling point. Furthermore, companies nowadays hire based on potential rather than demonstrable experience (Aggarwal, 2016). These key indicators cannot be derived from resumes and the current method of recommending candidates to our clients is not aligned with the future directives of the IT industry’s human resource needs which focuses more on the intangible.

Consequently, attrition rates are extremely high (25%) in the Indian IT industry. A commonly cited explanation for this phenomenon is the poor fit between the job and the staff. A proposed solution would be to hire the right people (Purohit, 2016).

## 1.2 Business Problem Statement

Therefore, we aim to leverage upon this opportunity by introducing a more holistic applicant screening process so as to allow our clients to hire the right people in accordance to their current preference for soft skills. More specifically, we would recommend suitable candidates to the companies. Suitability is determined through considering the applicant’s AMCAT scores and resume information. AMCAT scores are the chosen metric at this current stage as it is India’s most common employability test. Furthermore, a lot of job seekers get themselves accredited with it, making it easy to obtain the data. The incorporation of this value proposition will ensure that our job portal will not lose out on growth opportunities, thereby allowing us to have higher business revenue.

We plan to leverage upon the capabilities of analytics in the form of predictive hiring which is the utilisation of data to project candidate’s future successes. These predictions are based on the traits of existing employees (Jarret, 2017). HR analytics is best known for reducing biases in the recruitment process (Fineman, Tsuchida & Collins, 2017).

This value proposition therefore gives the discipline of analytics a competitive advantage in assessing candidate personalities which can be highly subjective from the view of the hiring manager.

## 1.3 Desired Business Outcomes

Given the above business proposition, the desired outcome from the implementation of this solution into our traditional job portals would be as follows:

***Higher Customer Satisfaction***

The immediate benefit our solution offers would be enhanced efficiency in the recruitment process for clients. The metric governing this outcome would be cost of hire. Cost of hire is calculated by considering internal and external costs. Internal costs include secondary management cost of time for recruiting (SHRM, 2018). The use of our targeted analytics solution can cut down recruitment time as it allows recruiters to concentrate their resources on candidates who are more likely to take up the jobs. This likelihood is being underscored by past data.

The long-term benefit of our solution will be evidenced by lower attrition rates amongst our clients. This is consequent of the idea of finding the ‘right’ person. This will therefore reduce attrition costs which are substantial, given how rampant attrition is in the Indian market.

***Growth of our Portal***

By embracing the disruptions in the job portal market, we aim to differentiate ourselves from our competitors. The key feature of our solution would be the holistic assessment of a job candidate. Our solution takes into consideration academics, skills and personalities. We aim to making this feature our unique selling point that companies are willing to pay for. Given that the subscription model is our revenue model, this added feature would be a premium service which clients would be willing to pay for, thereby, opening up more revenue streams.

# 

# INTRODUCTION TO THE ANALYTICS SOLUTION

To achieve our desired business outcomes, we define the analytics problem as the need for a model capable of predicting a candidate’s suitability for a position (i.e Probability of being suitable), given a candidate’s background, cognitive and personality information.

Our analytics solution is designed around the use of a neural network to predict for a multinomial response variable - *Designation*. With the expansion of scope beyond just academic and cognitive parameters, variable selection (a function inherently absent from neural networks) is first conducted using other predictive models to discern significant variables. The neural network is then optimised and benchmarked using its Misclassification Rate.

To develop our solution, we will use a dataset of 4000 Indian graduates largely employed in the engineering and IT sectors.

# 2 DATA PREPARATION

For our study, we have chosen to use a dataset which contains various information about a set of engineering candidates from India and their employment outcomes. For every candidate, the data contains both the individual’s profile information along with their employment outcome information.

Candidate Profile Information includes:

* Aspiring Minds Computer Adaptive Tests (**AMCAT**)[[1]](#footnote-1): Cognitive (*English Language, Logical Ability, Quantitative Ability*), Skill and Personality (*Big Five Traits*) assessment scores.
* Personal Information: Gender, Date-of-Birth
* Pre-University Information: Grade 10/12[[2]](#footnote-2) Results, School Board, Year of Graduation
* University Information: GPA, University ID, University Tier, Degree, Specialisation, University Location, Year of Graduation

Employment Outcome Information includes:

* Annual Salary, Date of Joining/Leaving the Company, Job Designation and Job Location.

As our study focuses on predicting a suitable job designation for a candidate, we will be cleaning the dataset to keep only the variables that we deem significant and useful, and at the same time, remove any unnecessary data. The following describes our data preparation process.

## 2.1 Reclassification of Job Designation

Several issues were identified with the variable *Designation* i.e the job designation that the different candidates held:

* Job designations that were referring to the same role but spelled differently (e.g. ‘ase’ and ‘assistant software engineer’).
* Job designations that were misspelled (e.g. ‘business development manager’ and ‘business development managerde’).
* Out of the 4,000 records there were 420 unique designations, but certain designations only had 1 record.
* Job designations that had similar job scope but given different titles (e.g. ‘customer service manager’ and ‘customer service representative’).

To prevent our results from being skewed in predicting only certain designations and improving accuracy, we decided to reclassify all 420 designations in our dataset according to the Singapore Standard Occupational Classification Version 2018 (SSOC).

Subsequently, we grouped designations based only on the first four digits of the classification codes that we have assigned to each record. For example, Embedded Engineer (SSOC Code 21526) and Implementation Engineer (SSOC Code 21527) will be classified in the same designation - Electronics Engineer (SSOC Code 2152).

After reclassification, our dataset has 62 unique designations.

## 2.2 Choosing Variables

In the next part of our data cleaning, we kept variables we deemed necessary in predicting a candidate’s suitability to a job designation. We kept variables per traditional requirements found on job portals as well as the AMCAT’s cognitive, skill and personality results.

A typical job posting will include information such as job designation, job description, skill, knowledge, education and experience (Roberts, 2017). However, for our study, we have removed the education variables *Degree* and *Specialization* as our dataset comprises mostly engineering students and the records for the two variables are near homogeneous, which resulted in a extremely skewed results when first included.

We also dropped most of AMCAT Skills tests (eg. *Computer Programming*) as the tests are optional and the number of candidates who took each specific test made up less than 5% of the dataset. Domain was kept as a sufficiently adequate substitute as it considered only the aggregated percentile of each candidate for the test they did take. Lastly, we took out irrelevant variables in (e.g. *Date of Leaving Job, College State*).

As such, here are the variables that we have retained for **SECTION 3 VARIABLE SELECTION**:

|  |  |
| --- | --- |
| **Target Variable:** Designation | |
| **Independent Variables** | |
| Background Information | Gender  CollegeTier  CollegeGPA  CollegeCityTier |
| AMCAT Cognitive Tests | English  Logical  Quant |
| AMCAT Skills Tests | Domain |
| AMCAT Personality Tests | Conscientiousness  Agreeableness  Extraversion  Nueroticism  Openness\_To\_Experience |

## 

## 2.3 Missing Values

For the variable *Domain*, there were candidates with missing values as the variable is computed based on an optional test and some candidates may not have taken any AMCAT Skills test. To account for the missing values, we used the ***mice*** packageto impute the missing data.

## 2.4 Final Cleaning

For our final stages of data cleaning, we removed designations that made up less than 2.5% of the dataset (i.e. 100 records). The removal of the extreme observations left us with seven designations in the dataset.

Refer to **Appendix A** for the Data Dictionary of our final cleaned dataset.

# 3 VARIABLE SELECTION

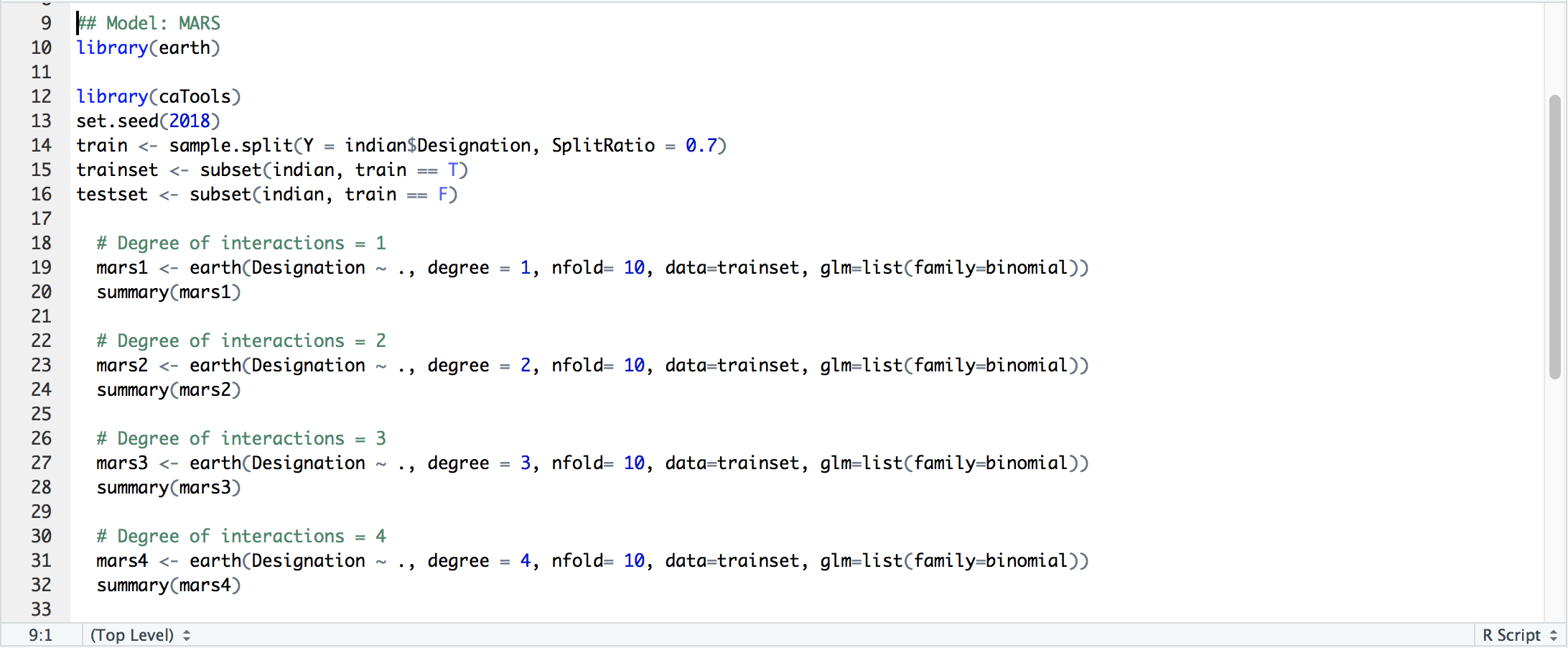
Neural networks regularly outperform more familiar classification models like logistic regression models, multi-adaptive regression splines (MARS) and classification and regression trees (CART). However, neural networks lack an inherent variable selection feature, which raises the issue of variable irrelevance and a diminished explanation of response variable behaviour through our neural network (May, Dandy & Maier, 2011).

In our proposed solution, we will separately build, optimise and compare the following models to choose the best subset of input variables for our neural network: MARS, and CART pruned via optimal complexity parameter or through applying Random Forest techniques.

The best subset will be chosen on the criteria of (i) Selection Model Accuracy, and (ii) Effect on Neural Network Accuracy. Each model is initialised with the same set of variables.

## 3.1 Multivariate Adaptive Regression Splines (Mars)

***Build & Optimise***



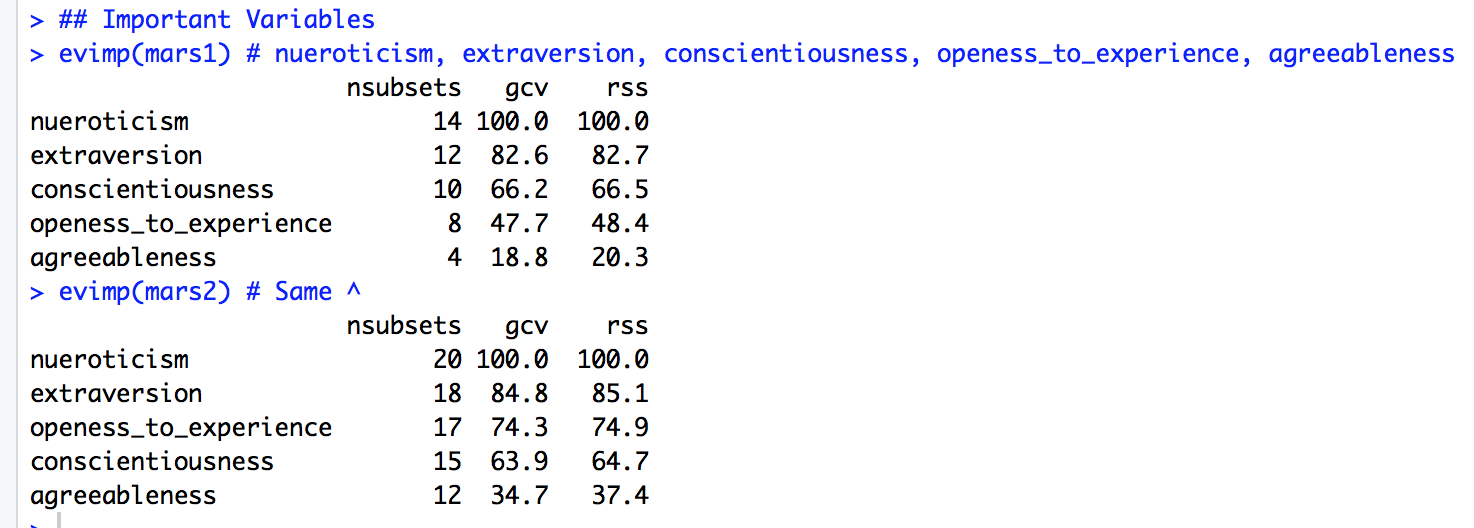
*Figure 3-1. Building the four MARS models.*

With MARS, the ***earth*** package was used. Since our response variable is a factor, we invoke the argument ‘*glm=list(family=binomial)’* and through *earth()* our factor with seven levels is converted into seven indicator columns of 1s and 0s (Milborrow, 2018).

Earth builds seven simultaneous models each having the same set of basis functions but different coefficients. The models are pruned by minimising the sum of GCVs across all seven models.

To test for significant interaction terms, models were built for degrees up till four (i.e mars1 to mars4). N-fold cross validation is also inbuilt into *earth()* function, which helps reduce trainset bias.

***Significant Variables***



*Figure 3-2. evimp() function to view important variables.*

The *evimp()* function is used to view important variables in each of the MARS models ranked by their nsubset scores, and normalised incremental effect on GCV and RSS values (averaged across the multiple response models and subsets).

## 3.2 CART

***Build & Optimise***

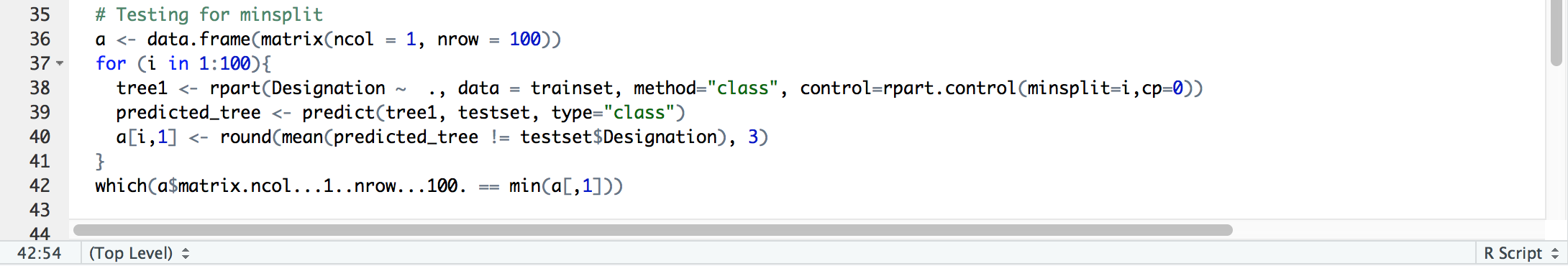


*Figure 3-3. Building the three different classification trees.*

With CART, the ***rpart*** and ***randomForest*** packages were used. First, the unpruned *tree1* is grown by selecting the splitting variables and split points that minimise the weighted average Gini impurity of the resultant child nodes.

Growing stops when all cases in the node are pure, have the same set of input variable values or number of cases in the node fall below the *minsplit* value. Pruning *tree1* using minimal cost complexity pruning produces *tree2*.

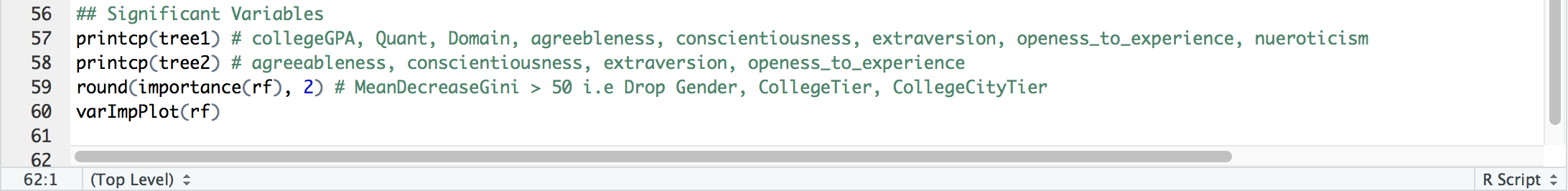
Considering the inherent bias of a single iteration of decision tree to the training set used, we chose to further explore the use of a Random Forest.



*Figure 3-4. For-loop to identify minsplit value that minimises misclassification error.*

To optimise the *rpart()* trees, we ran a for-loop testing *minsplit* values between 1 and 100. The optimal *minsplit* value minimises the misclassification rates of *tree1*. The derived value is 73.

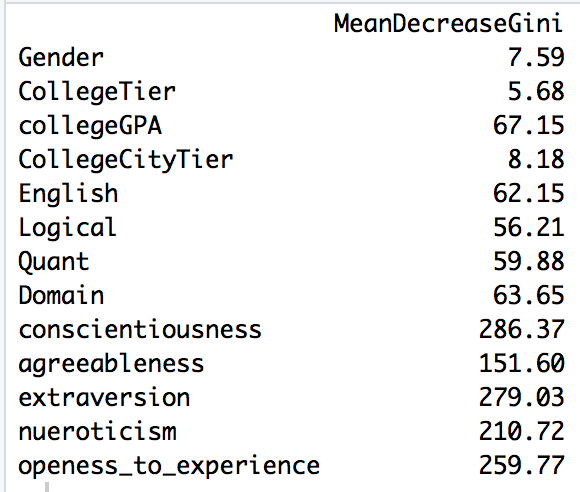
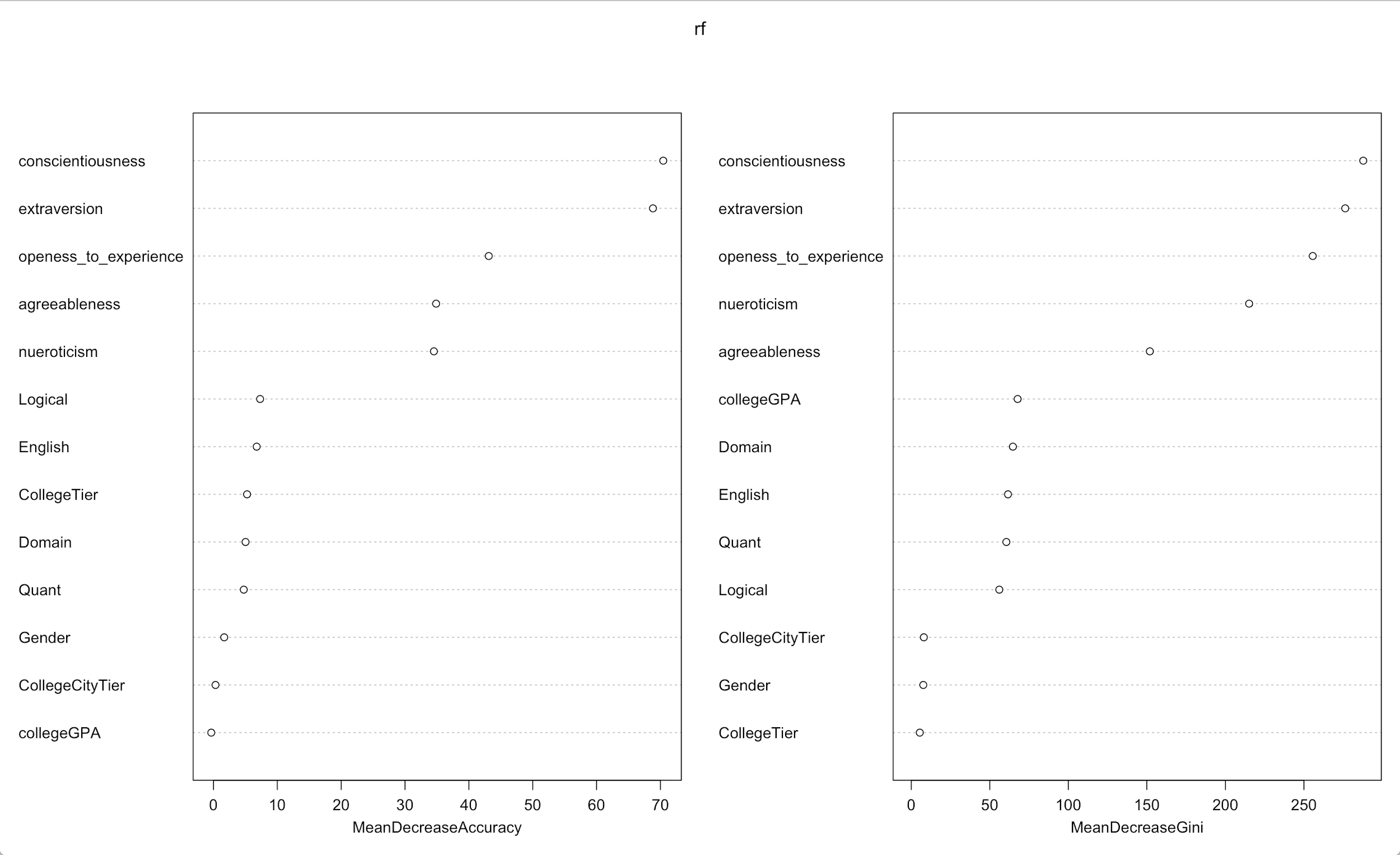
***Significant Variables***



*Figure 3-5. printcp() and importance() functions to view important variables.*

The *printcp()* function is used to view variables actually used in the rpart trees.

For the Random Forest model, variable importance is represented by incremental decrease on the mean Gini Coefficient of all models generated. We view this via the *importance()* function and *varImpPlot()*, as shown below. Setting a threshold of 50, we select only variables exhibiting an incremental decrease in the mean Gini Index above this.



*Figure 3-5. Results of importance() and varImpPlot() functions on Random Forest model.*

## 3.3 Results

The respective models’ accuracies are tested on a testset. Further, the incremental improvement of the SAS Neural Network when using the models’ subset of variables are also recorded. This step is similar to building the Neural Network in **SECTION** **4 NEURAL NETWORK**, except now the network is rebuilt with different subset of input variables.The improvement is calculated as the difference between the original model (i.e all thirteen input variables included) and the subsetted model (i.e variables are chosen from the variable selection models).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | | **Misclassification**  **Rate**  **(%)** | **Absolute Improvement in Neural Network (%)** | **Subset of Variables** |
| **MARS** | *Degree = 1* | 18.1 | 1.8 | nueroticism  extraversion  conscientiousness  openess\_to\_experience  agreeableness |
| *Degree = 2* | 19.4 | Same ^ | Same ^ |
| *Degree = 3* | 19.4 | Same ^ | Same ^ |
| *Degree = 4* | 19.4 | Same ^ | Same ^ |
| **Decision Tree** | *Not pruned* | 21.8 | 5.7 | collegeGPA  Quant  Domain  agreeableness  conscientiousness  extraversion  openess\_to\_experience  nueroticism |
| *Pruned* | 20.3 | 1.6 | extraversion  conscientiousness  openess\_to\_experienc  agreeableness |
| *Random Forest* | 17.8 | 5.6 | collegeGPA  English  Logical  Quant  Domain  agreeableness  conscientiousness  extraversion  openess\_to\_experience  nueroticism |

*Figure 3-6. Table of misclassification rates and attributable improvement in SAS Neural Network.*

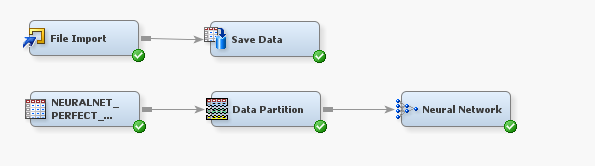
Based on misclassification rates, the top three variable selection models are Random Forest, MARS (*1st Degree*) and MARS (*2nd to 4th Degree*) in descending order of merit. Based on degree of improvement observed in the Neural Network when model’s subset of variables are used, the top three are CART (*Not Pruned*), Random Forest and MARS (*All Degrees*) in descending order of merit.

The best subset of variables is chosen from the Random Forest model. Its accuracy ranks as the top by a significant extent from the rest and while CART (*Not Pruned*) outperforms it in incremental improvements to the neural network, the difference of 0.1% is negligible. Furthermore, Random Forest accounts for individually-generated decision trees’ sample set bias, resulting in better predictive ability with unexpected new test data.

# 4 NEURAL NETWORK

The core of our solution - a Neural Network - is developed within the SAS Enterprise Miner (EM) environment. The input variables are derived from the best performing model in **SECTION 3 VARIABLE SELECTION** - Random Forest. The following steps were also performed when evaluating the variable selection models before, bar optimisation.

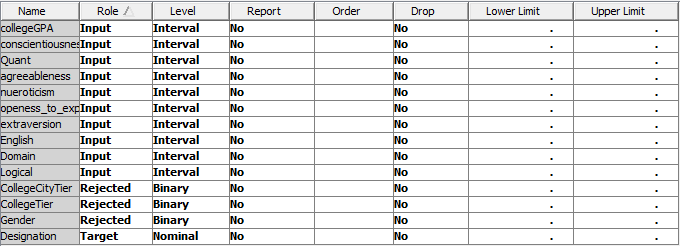
## 4.1 Data Import & Partition Nodes



*Figure 4-1. Overview of nodes used in SAS EM.*

***Data Import***

The ‘**indian\_7jobs\_perfect.csv**’ is imported into SAS EM through the *Data Import* node. Under the variables property panel, the variables are assigned the roles, as shown below. In particular, *Designation* is assigned the role of *Target*. The other variables are assigned *Input* if it was included in the Random Forest best subset and assigned *Rejected* if otherwise. SAS EM automatically excludes *Rejected* variables from any further analysis.

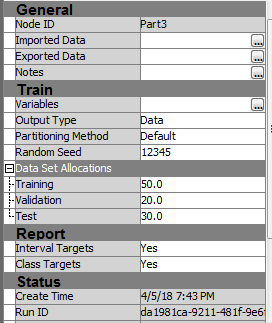


*Figure 4-2. Variable import.*

The file is saved within the existing SAS User library as a SAS data set using the *Save Data* node. This allows the data set to be analyzed within SAS EM.

***Data Partition***

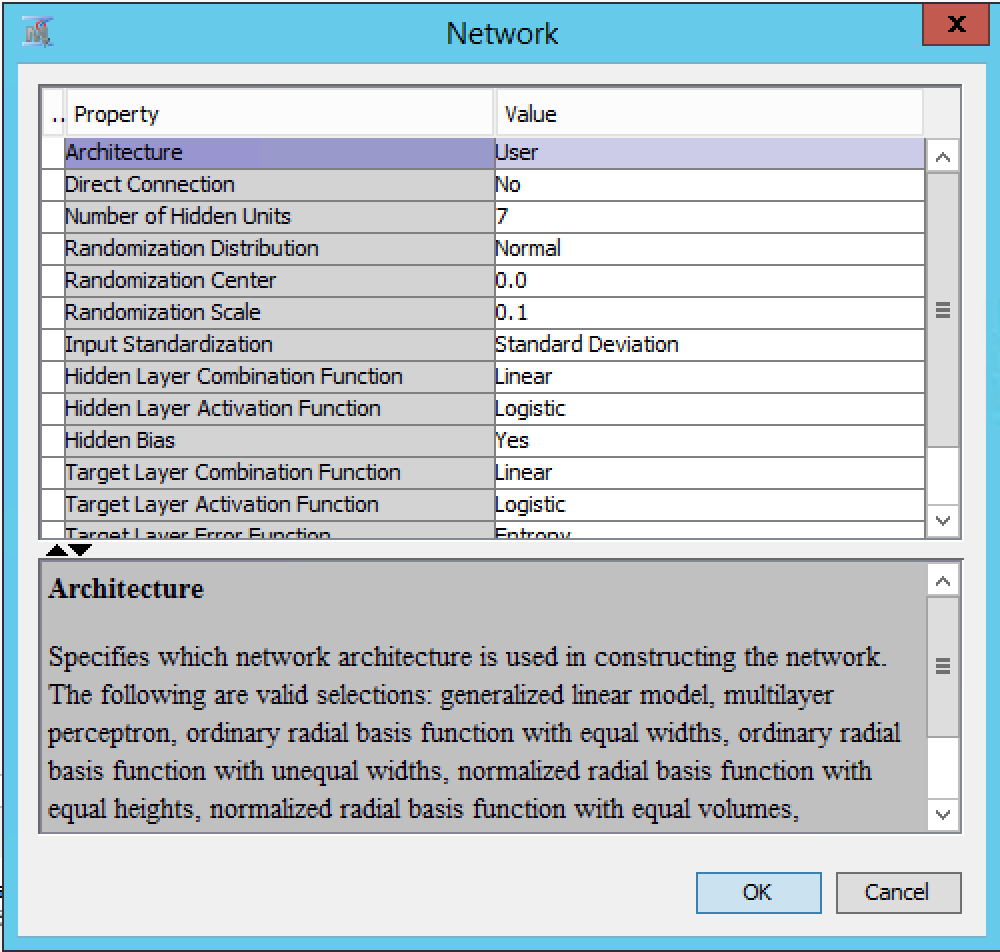
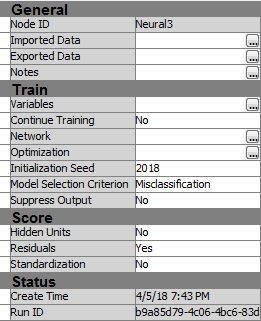
Before further analysis, there is a need to create a training and test set. This can be performed using the *Data Partition* node. We apply the train-test ratio of 70:30 as per our variable selection models. Based on the default settings of the *Data Partition* node, the partitioning method used will be *stratified* partitioning if the *Target* variable is a categorical variable.



*Figure 4-3. Data partition settings.*

## 4.2 Neural Network Node

The *Neural Network* node in SAS EM is able to develop a neural network model for our dataset. We selected *Misclassification* as our model selection criterion because our *Target* variable - *Designation* - is a categorical variable. SAS naturally restricts the number of hidden layers to one.



*Figure 4-4. Neural network node’s general and network-related settings.*

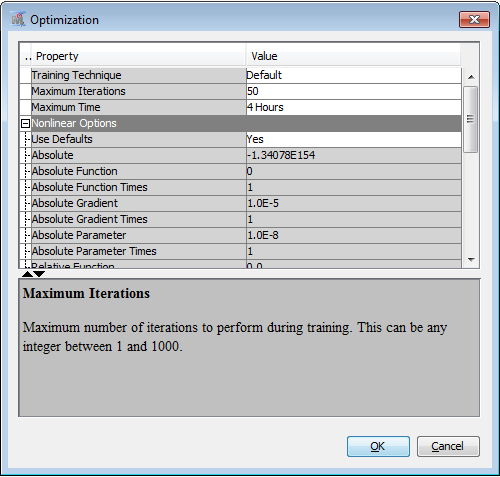
Parameters are set within the network options panel. We set the *Number of Hidden Units* based on the ⅔ Rule as research shows it to allow for stable learning within an Adaptive Resonance Theory (ART) architecture (Carpenter & Grossberg, 1991), of which many neural network models are based upon. In this case, our baseline neural network will have seven hidden units, given that there are ten variables in total.

We choose both the *Hidden Layer / Target Layer Combination Functions* to be *Linear*. With regards to the *Target Layer Error Function*, we choose *Entropy*, which is the most commonly used error function when the *Target* variable is categorical (Sadowski, N.d). We proceed with optimising other parameters.

***Activation Function***

For the *Hidden Layer Activation Function*, we compared *Logistic* against *Hyperbolic Tangent*. Ceteris paribus, using *Logistic* produces a lower misclassification rate of 23.8% as compared to 29.6% when the function is *Hyperbolic Tangent*. Hence, our choice will be *Logistic.*

***Iterations***



*Figure 4-4. Neural network node’s optimisation settings.*

In the optimization panel, there is an option for *Maximum Iterations*. This option allows us to specify the number of times the dataset is passed through the application to develop the neural network model.

Holding all else equal, we selected 1000 as our maximum number of iterations instead the default of 50 because it gave us a better misclassification rate of 20.9% instead of 23.8%, with minimal detriment to training time.

***Training Technique***

The default training technique that is used in the development is the *Quasi-Newton* method. The *Quasi-Newton* method performs significantly better in optimization problems than other gradient-descent techniques (Likas & Stafylopatis, 2000), such as resilient backpropagation *(RProp)* and classical backpropagation *(BackProp)*. This improvement is also seen in the model generated in SAS EM.

The Quasi-Newton method resulted in a model of lower misclassification rate at 23.8% as compared to *RProp*’s 27.4% and *BackProp*’s 33.8%, with all other parameters being held at the default settings.

Based on the optimisation parameters selected in the section above, we developed our final neural network model. Our model has a misclassification rate of 20.6%.

# 5 IMPLEMENTATION AND IMPLICATIONS

## 5.1 Short Term Strategy - A Hybrid Portal

A hybrid portal is likened to a recruitment consultant where a guaranteed response would be given to the company looking to recruit (Fahey, 2016).

At the point of application, the applicant inputs information relating to the variables we have identified as significant in our model. Our system returns the candidate’s suitability ratings for all job categories, highlighting both (i) job categories that the candidate has indicated interest in and, (ii) job categories that the candidate is highly suited for. This implores the candidate to consider positions they may not have otherwise explored.

From the recruiter’s perspective, applicants for their job listings are returned in a list ranked by their suitability for the position.

Overall, our approach value-adds to the company’s HR department by reducing biases on the part of both the applicant (i.e applying for jobs they deem themselves suitable for) and the recruiter (i.e when evaluating applications) in the hiring process. The desired outcome is a more efficient labour market where the most suitable candidates apply and companies hire the most suitable.

Recruiters can concentrate their efforts on a targeted pool of applicants, allowing for time savings in the screening process which takes an estimated 23 hours per candidate presently (Fahey, 2016). Additionally, attrition rates are likely to fall due to more effective allocation of human resources according to suitability.

## 5.2 Long Term Strategy - A Career Portal

A career portal provides a full suite of career resources such as content, courses, event invitations and even psychometric testing (Ideal, 2016). Through constructing a career portal, additional data can be harnessed from the various interactions the applicant has with the portal, thereby allowing for further profiling to be done. For example, a highly skilled IT candidate may actively peruse the learning resources for accounting, this is possibly indicative of a hybrid manager as it shows an active interest in the non-technical business aspects.

This adds value to our analytics solution as additional information harnessed under non-test conditions may be more authentic than those garnered from formal testing where there is a possibility of applicants choosing answers perceived preferable.

From the job portal company’s point of view, the data we harness is proprietary and insights we harnessed are exclusive for our usage. These insights will improve the accuracy of our models and even open new areas for prediction, generating additional revenue streams.

# 6 LIMITATIONS & FUTURE IMPROVEMENTS

## 6.1 Classification of Jobs

***Overcoming Sparse Data***

Currently, in our data set, the classification of designations are done manually according to our interpretations of the job titles as there are (i) many different positions and (ii) many positions which seem to have similar jobscope. Previously, it was difficult to get an accurate result with sparse data. However, this also means our results are limited by our scope and consistency of classification.

When more data is available in the future, it will be possible to drill down to more detailed classifications to have meaningful interpretations from the varying levels. Also, machine learning may be used to interpret the job titles and classify them without any bias.

***Segment Jobs by Seniority***

In our current analysis, we have consolidated our designations to exclude seniority. (eg. senior software developer vs junior software developer) This segregation of seniority will improve our predictive ability and further pinpoint the optimal fit for both the applicants and the employers. However, as we are limited in the amount of data we have for each designation according to seniority, our prototype serves as a proof of concept before we explore seniority roles in the future once more data is made available.

## 6.2 Predicting Different Areas

***Salary***

Currently our prediction model focuses on establishing a fit for the applicant’s designation according to the variables provided. This has yielded a reasonably good result and provides a valuable tool to both the job applicant and the employers, giving the job portal a reliable profit making model.

To further expand on this, we can possibly try to predict the possible salary range upon entry to a specific role (applicant use) or provide a range estimate for the salary of an applicant for the employers according to their perceived fit for the job (employer use).

## 6.3 Premium Service

***Targeted Head Hunting***

While our current solution provides an employer with the list of applicants ranked by suitability, this limits the talent pool available to just candidates who have applied. We aim to extend to companies the option to view viable candidates from our site’s talent pool, who have not applied for their position but may be more suitable than the current applicants. This headhunting option will be offered as a premium service to complement the basic functions available to a subscribed company, creating a new revenue stream.

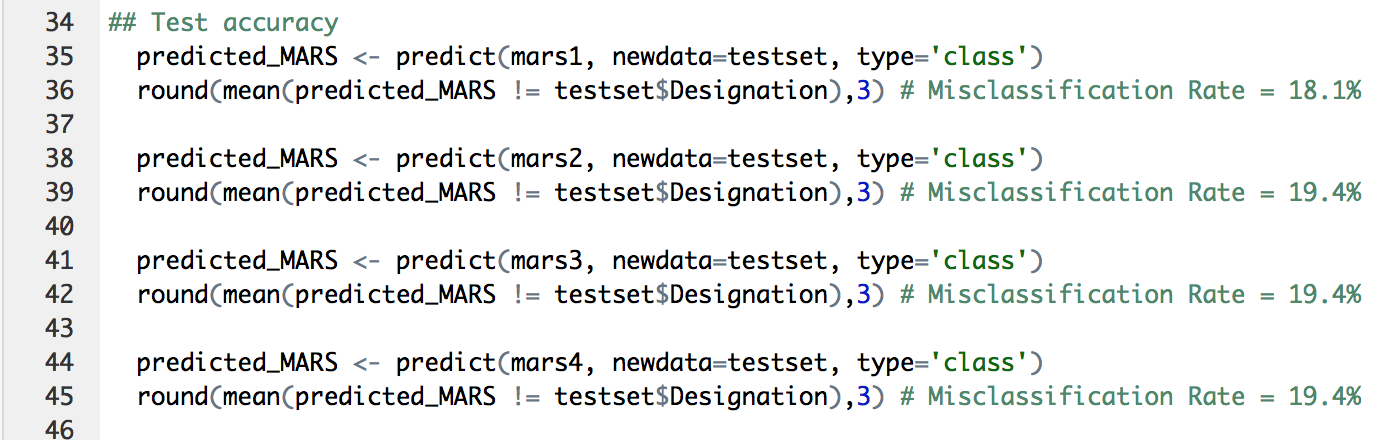
# APPENDIX

## A. Data Dictionary

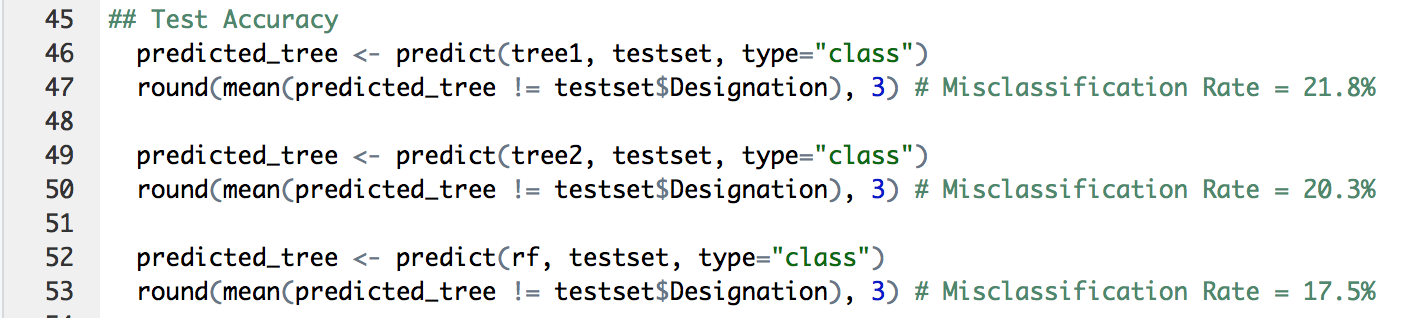
|  |  |
| --- | --- |
| **Variable** | **Description** |
| Designation | Position candidate holds at current job  Levels: Applications and System Programmers, General Engineering, Quality Checkers and Testers, Sales, Business and Marketing Development, Software Developer, Systems Analysts, Web Developer |
| Gender | Levels: Male (m), Female (f) |
| CollegeTier | Colleges in India have been given a rating of 1 or 2. The rating is computed based on the average AMCAT scores obtained by the students attending the college. Colleges with an average score above a threshold (not known) are given a rating of 1 while those that fall below the threshold are given a rating of 2.  Levels: 1, 2 |
| collegeGPA | Final GPA of the candidate at Graduation Date. The most common India Grading System is as follows:   |  |  | | --- | --- | | GPA Description | GPA | | First Class | 60.00 - 100.00 | | Second Class | 50.00 - 59.00 | | Third Class | 40.00 - 49.00 | | Below Minimum Pass | 0.00 - 39.00 |   Values: [0 , 100] |
| CollegeCityTier | Cities in India have been given a rating of 0 or 1. The rating is based on the population of the cities.  Levels: 0, 1 |
| English | Scores that candidate attained in the AMCAT English Comprehension section  Values: [0 , 900] |
| Logical | Scores that candidate attained in the AMCAT Logical Reasoning section  Values: [100 , 900] |
| Quant | Scores that candidate attained in the AMCAT Quantitative Ability section  Values: [100 , 900] |
| Domain | Percentile of candidates in their respective domain-specific tests (e.g. Computer Programming, Electrical Engineering)  Values: [0 , 1] |
| conscientiousness | Z-scores of candidate’s trait scores for the AMCAT Personality Test (Aspiring Minds Personality Inventory). The personality assessment is based on the Big Five Model personality test.  **Conscientiousness**: Reflects how careful and orderly a candidate is. High scorers tend to be well-organized, self-disciplined and careful. Low scorers tend to be disorganised and negligent.  **Agreeableness**: Reflects how much a candidate likes and tries to please others. High scorers tend to be good natured, forgiving and courteous. Low scorers tend to be critical, rude and harsh.  **Extraversion**: Reflects how outgoing a candidate is and likelihood of a candidate deriving satisfaction from interacting with others. High scorers tend to be sociable, fun-loving and talkative. Low scorers tend to be introverted, reserved and quiet.  **Neuroticism**: Reflects the tendency of a candidate to experience negative emotions. High scorers tend to be nervous, insecure and worry too much. Low scorers tend to be calm, relaxed and hardy.  **Openness to Experience**: Reflects how receptive a candidate is to new experiences. High scorers tend to be creative and curious. Low scorers tend to be conventional, down-to-earth and inflexible.  Values: [-2 , 2] |
| agreeableness |
| extraversion |
| nueroticism |
| openess\_to\_experience |

\**Variable Names are based on the naming used in original dataset*

## B. Misclassification Rate of Variable Selection Models on Testset

****

*Figure B-1. Testing MARS models.*

****

*Figure B-2. Testing Classification Trees.*

# REFERENCES

Aggarwal, H. (2016). *Why Job Portals Are Outliving Their Usefulness*. *HuffPost India*. Retrieved 14 April 2018, from <https://www.huffingtonpost.in/himanshu-aggarwal/why-job-portals-are-outli_b_9466466.html>

Agrawal, N., & Thite, M. (2003). Human resource issues, challenges and strategies in the Indian software industry. *International Journal Of Human Resources Development And Management*, *3*(3), 249. <http://dx.doi.org/10.1504/ijhrdm.2003.003664>

Carpenter, G., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge, Mass: MIT Press.

*Cost-per-Hire Standard*. (2012). *Shrm.org*. Retrieved 14 April 2018, from <https://www.shrm.org/ResourcesAndTools/business-solutions/Documents/shrm_ansi_cph_standard.pdf>

Fahey, M. (2016). *Job board evolution: Seven possible future models*. *Madgex.com*. Retrieved 14 April 2018, from <https://www.madgex.com/labs/knowledge-hub/blog/job-board-evolution-seven-possible-future-models>

Fineman, D., Tsuchida, A., & Collins, L. (2017). *People analytics: Recalculating the route*. *Deloitte Insights*. Retrieved 14 April 2018, from <https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2017/people-analytics-in-hr.html>

Gager, S., Bowley, R., Nayak, M., & Chhabria, P. (2015). *Business.linkedin.com*. Retrieved 14 April 2018, from <https://business.linkedin.com/content/dam/business/talent-solutions/global/en_US/c/pdfs/india-recruiting-trends-final1.pdf>

*IT & ITeS Industry in India: Market Size, Opportunities, Growth...IBEF*. (2018). *Ibef.org*. Retrieved 14 April 2018, from <https://www.ibef.org/industry/information-technology-india.aspx>

Jarrett, J. (2017). *How to Use Data to Make Better Hires*. *TLNT*. Retrieved 14 April 2018, from <https://www.tlnt.com/how-to-use-data-to-make-better-hires/>

Lebowitz, S. (2016). *The ‘Big 5’ personality traits could predict who will and won’t become a leader*. *Business Insider Singapore*. Retrieved 14 April 2018, from <http://www.businessinsider.sg/big-five-personality-traits-predict-leadership-2016-12/?r=US&IR=T>

Likas, A., & Stafylopatis, A. (2000). Training the random neural network using quasi-Newton methods. *European Journal Of Operational Research*, *126*(2), 331-339. <http://dx.doi.org/10.1016/s0377-2217(99)00482-8>

Milborrow, S. (2018). *Notes on the earth package*. *Milbo.org*. Retrieved 14 April 2018, from <http://www.milbo.org/doc/earth-notes.pdf>

Purohit, M. (2016). A Study on - Employee Turnover in IT Sector with Special Emphasis on Wipro and Infosys. *IOSR Journal Of Business And Management (IOSR-JBM)*, *Volume 18*(Issue 4 .Ver. I), PP 47-51.

*Resume Screening: A How-To Guide For Recruiters | Ideal*. (2016). *Ideal*. Retrieved 14 April 2018, from <https://ideal.com/resume-screening/>

Sadowski, P. (2018). *Notes on Backpropagation*. *Ics.uci.edu*. Retrieved 14 April 2018, from <https://www.ics.uci.edu/~pjsadows/notes.pdf>

Carpenter, G., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge, Mass: MIT Press.

*India Grading System*. *Classbase.com*. Retrieved 14 April 2018, from <https://www.classbase.com/countries/India/Grading-System>

Friedman, H., & Schustack, M. (2016). *Personality: Classic Theories and Modern Research* (6th ed., p. p. 191). Pearson.

May, R., Dandy, G., & Maier, H. (2011). *Review of Input Variable Selection Methods for Artificial Neural Networks*. *InTech*. Retrieved 14 April 2018, from <http://cdn.intechweb.org/pdfs/14882.pdf>

Roberts, M. (2017). *Definition and Specs of a Job Posting*. *The Balance*. Retrieved 15 April 2018, from <https://www.thebalance.com/what-information-is-included-on-a-job-posting-1669554>

1. India’s first employability test which aims to help companies find the ‘right’ employee. [↑](#footnote-ref-1)
2. India’s Grade 10 and Grade 12 is Singapore’s equivalent of Secondary 2 and Secondary 4 [↑](#footnote-ref-2)