

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

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

**Semester Team Project Proposal**

|  |  |  |
| --- | --- | --- |
| 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** | |

[**Project Timeline**](#_er3rr13litss) **3**

[**1.1 Business Problem Statement(s)**](#_lpxc6cgef491) **4**

[**2.1 Analytics Approach**](#_lxpnion4ecd0) **5**

[**3.1 Data Source (Acquired on 01/03/18)**](#_r5nxkh2fgike) **6**

[**4.1 Major Project Risks Anticipated and Mitigation Plan**](#_18j3i5ac0hnd) **7**

[**5.1 Major Issues Encountered and How You Resolve them**](#_o56p7vvjl9hd) **8**

[**6.1 Literature Review**](#_r40ovqh571k9) **9**

[**7.1 References**](#_xd6pjstvogr3) **11**

# Project Timeline

|  |  |
| --- | --- |
| **Goal** | **Deadline** |
| Brainstorming business problem and ideas for the project | 23 January |
| Finding viable data-sets - in contact with the provider of data sets, still pending approval | 5 February |
| Finalise data set for project  Finalise business problem | 7 March |
| Data Cleaning | 19 March |
| Try using Association Analysis with dataset | 26 March |
| Try using Neural Network, MARS, Logistic, CART with dataset | 29 March |
| Run all models and select variables to use in Neural Network | 7 April |
| Start writing Report | 11 April |
| Prepare Presentation Slides & Script | 12 April |
| Presentation Rehearsal | 14 April |

# 1.1 Business Problem Statement(s)

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 usually incorporate resume databases are relatively more prevalent as compared to that of India’s global counterparts. 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.

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.

# **2.1 Analytics Approach**

To solve the business problem statement, we aim to create 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 will be designed around the use of a neural network to predict for a multinomial response variable - *Designation*. To overcome the neural network’s lack of variable selection functionality, we will be performing variable selection using other models such as logistic regression and decision trees. Subsequently, we will pick the variables from the model that produces the lowest misclassification rate.

# 3.1 Data Source (Acquired on 01/03/18)

For our study, we obtained a dataset which as used for a data challenge online. The dataset 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:

* Scores on Aspiring Minds Computer Adaptive Test (AMCAT) - India’s first employability test which aims to help companies find the ‘right’ employee. The test includes cognitive (English Language, Logical Ability, Quantitative Ability), skill and personality (Big Five Traits) assessments.
* Personal Information: Gender, Date-of-Birth
* Pre-University Information: Grade 10/12[[1]](#footnote-0) 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, Job Location

For the project, we will need to do some data cleaning to filter out the records and variables that will aid us in creating a model which will help us predict the suitability of a candidate to our client’s job posting.

# **4.1 Major Project Risks Anticipated and Mitigation Plan**

Firstly, we were uncertain whether our idea would work, and whether the dataset we had would aid us in generating a model that would help predict suitability of a job position based on traditional metrics and added factors such as AMCAT scores and soft skills. Hence, we had back-up plans such as predicting salary given that our first idea didn’t work out.

Additionally, our original plan was to use RStudio to generate Neural Network models as we felt that we had more control in adjusting the parameters on R. However, given that it may not work, we set our alternative to creating a Neural Network model on SAS Enterprise Miner.

# 

# **5.1 Major Issues Encountered and How You Resolve them**

Firstly, upon brainstorming our idea, we had issues finding a dataset which incorporated both the element of soft skills and employment information. However, we were fortunate to receive a reply from an Aspiring Minds employee who was able to give us a dataset containing the information we needed fo our study.

Secondly, our original idea was based on assumptions that there was a relationship between soft skills and employment outcomes. To ascertain that there was a relationship we did a thorough research on organisational behaviour and found out more about how soft skills affected an employee’s performance at work. Additionally, we looked for more statistics to back us on the assumption that recruiters looked for soft skills during hiring.

Thirdly, the dataset that we obtained was not perfect and as a result, when we first ran our analytics approach, we ended up having a very skewed results, which is only being able to predict the suitability of an applicant for one designation. To solve this issue, we did further data cleaning and classified similar job roles to one designation group. This helped us obtain a result with lower misclassification rate, and a model which was able to predict the suitability of a candidate for more designations.

Additionally, our initial analytics approach was to only use Neural Network in helping us predict the suitability of a job position. However, we realised that Neural Network does not have the function of choosing more significant variables, but only reduced the contributing weights to the final output variable for the less significant variables. Hence, to improve the accuracy of our model, we decided to use other models to choose the variables that we will input into the Neural Network model ultimately.

At the start, we had wanted to use the **nnet** and **neuralnetwork** package on Rstudio to generate Neural Network models. However, we were limited by factors such as the need to create dummy variables for categorical variables on and that our target variable had 7 levels. Although we did try creating dummy variables to account for the categorical variables and multi-levelled target variable, the results we obtained had high misclassification rates (>45%). Hence, we decided to use the Neural Network on SAS Enterprise Miner in the end. The model generated on SAS gave us a relatively low misclassification rate (<25%).

Finally, when we wanted to compute the misclassification rate for the models we generated using Neural Network, we couldn’t directly use the results on SAS as our target variable had 7 levels. Hence, we exported our results onto Microsoft Excel and did our own computation to compare the before and after.

# **6.1 Literature Review**

|  |  |
| --- | --- |
| **Source Title (URL)** | **Review** |
| Human resource issues, challenges and strategies in the Indian software industry  <https://www.researchgate.net/publication/29457704_Human_resource_issues_challenges_and_strategies_in_the_Indian_software_industry?enrichId=rgreq-e916cebced576120b6be8d93b09ff3ff-XXX&enrichSource=Y292ZXJQYWdlOzI5NDU3NzA0O0FTOjE2MjEzODk2MjI3NjM1MkAxNDE1NjY4MzQ5NzI4&el=1_x_2&_esc=publicationCoverPdf> | This paper has presented several insights on the Indian IT industry.  Firstly, there is a demand for hybrid managers as the IT sector is moving up the value chain. Hybrid managers are IT professionals that possess the ability to lead and have strong command of technical skills coupled with business acumen.  Secondly, the Indian IT industry also requires people that possess a great amount of teamwork. Recruiters seek to identify people that has the potential to fill such roles.  Thirdly, there is an issue in the current landscape as a lot of the IT professionals in India believe that their main value proposition comes from their level of technical skill and therefore would be very unwilling to give up their technical job scopes for higher paying managerial and strategic positions due to its non-technical nature. |
| India Recruiting Trends: 3 Must-know talent acquisition trends for 2015  4th Annual Report  <https://business.linkedin.com/content/dam/business/talent-solutions/global/en_US/c/pdfs/india-recruiting-trends-final1.pdf> | This report highlights the recruiting trends and habits of Indian recruiters. Some of the notable insights presented were that Indian recruiters extensively utilised internet job boards. The utilisation rates are significantly higher than that of their global counterparts.  Given that our traditional job portal comes in the form of an internet job board, the fact that it is widely used suggests that a lot of inefficiencies that such job boards present has been incorporated into the recruitment process. Thereby validating the case for a revamp so that it suits the taste and preferences of the recruiters of today. |
| Why Job Portals Are Outliving Their Usefulness  <https://www.huffingtonpost.in/himanshu-aggarwal/why-job-portals-are-outli_b_9466466.html> | This article showcases the irrelevance of resumes in the recruiting process of today. One of the most notable insights presented was that resumes showcased the past experiences and technical skills of the applicants but recruiters are actually on the lookout for potential amongst the applicants. Potential includes the applicant's potential to lead and become a hybrid manager as well as the potential to be a great teammate. |
| Job board evolution: Seven possible future models  <https://www.madgex.com/labs/knowledge-hub/blog/job-board-evolution-seven-possible-future-models> | This web page showcases the new breeds of job portals. This site supplied us with the inspiration for the long term and short term strategy for the revamp of our job site. We aim to transform the job site into a hybrid portal first which allows us to play the role of a recruitment consultant to the company. Once we have gained a certain level of traction, we would then develop the site into a career portal so that there are several user touchpoints that will allow the site to gather more data about the applicant, making our analytics solution a more robust and holistic one. |
| Review of Input Variable Selection Methods for Artificial Neural Networks (<http://cdn.intechweb.org/pdfs/14882.pdf>) | This paper discusses the fundamental importance of variable selection in determining the optimal form of a statistical model. However, the complex nature of Artificial Neural Networks (ANN) pose a challenge in applying traditional techniques. The consequences of poorly chosen variables include irrelevance of model, increased computational effort, training difficulty and dimensionality, and worsening comprehensibility.  The paper reviews various methods proposed: Model-based Wrapper or Embedded algorithms and Model-free Filter algorithms. We apply these principles to our report, where we use the Multivariate Adaptive Splines (MARS) and Decision Tree models to identify significant variables. |

# 

# 7.1 References

SAS Institute Inc. 2018. Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach, Fourth Edition. Cary, NC: SAS Institute Inc

Raw data from online personality tests. (n.d.). Retrieved April 10, 2018, from <https://openpsychometrics.org/_rawdata/>

Aspiring Minds Employment Outcomes 2015. (n.d.). Retrieved March 29, 2018, from https://ikdd.acm.org/Site/CoDS2016/datachallenge.html

1. India’s Grade 10 and Grade 12 is Singapore’s equivalent of Secondary 2 and Secondary 4 [↑](#footnote-ref-0)