**Data 311: Machine Learning**

Kyle Lee || 27118158

       Amat Zahra || 40429145

       Pierre Frigon || 29319150  
       Aedan Wen || 97764138

**April 4th 2019**

**UNIVERSITY OF BRITISH COLUMBIA**

**OKANAGAN**

Introduction

The dataset being analyzed is from the Job Search Intervention Study (JOBS II) that was done in 1997 to investigate the efficacy of a job training intervention program for unemployed workers across various occupation. The program is intended to increase reemployment and also enhance the mental health of those who are unemployed. The study consisted of treatment group, which received job-skills workshops and a control group that were only given a pamphlet on job-search tips. The data collected are from a pre-screening questionnaire and a follow-up interview after the program. Through Regression, the post-depression is predicted to identify if the treatment is helpful in lowering it. In addition, Classification is used to predict subject’s comply measure in attempt to determine which feature(s) are more important in predicting subjects’ willingness to comply to treatment. Finally, clustering will be done to seek any interesting patterns and clusters to explore further on the dataset.

 The following is a description of the dataset:

|  |  |
| --- | --- |
| **econ\_hard** | Level of economic hardship pre-treatment with values from 1 to 5 |
| **depress1** | Measure of depressive symptoms pre-treatment. |
| **sex** | Indicator variable for sex. 1 = female |
| **age** | Age in years. |
| **occp** | Factor with seven categories for various occupations. |
| **marital** | Factor with five categories for marital status. |
| **nonwhite** | Indicator variable for race. 1 = nonwhite. |
| **educ** | Factor with five categories for educational attainment. |
| **income** | Factor with five categories for level of income. |
| **job\_seek** | A continuous scale measuring the level of job-search self-efficacy with values from 1 to 5. The mediator variable. |
| **depress2** | Measure of depressive symptoms post-treatment. |
| **work1** | Indicator variable for employment. 1 = employed. |
| **job\_dich** | The job\_seek measure recoded into two categories of high and low. 1 = high job search self-efficacy. |
| **job\_disc** | The job\_seek measure recoded into four categories from lowest to highest. |
| **treat** | Indicator variable for whether participant was randomly selected for the JOBS II training program. 1 = assignment to participation. |
| **comply** | Indicator variable for whether participant actually participated in the JOBS II program. 1 = participation. |
| **control** | Indicator variable for whether participant was randomly selected to not participate in the JOBS II training program. 1 = non-participation. |
| **depress1** | Measure of depressive symptoms pre-treatment. |

Data Clean Up

The data collected in the study is very complete, there is no missing values. However, the comply feature is always false if the subject is part of the control group. As such, the comply feature only applies to those who were selected for selected for the treatment program (*treat* = 1) and should be used with this in mind. The *control* feature was also removed due to being redundant to the treatment feature. In addition, the *work1* feature was modified from lengthy string values to binary values for easier interpretation and conciseness.

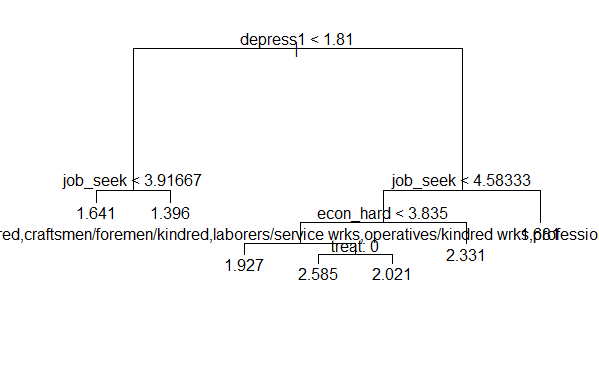
Regression

Analysis 1

Analysis 1 focuses on modeling *depress2* based on predictors: *treat*, *econ\_hard*, *depress1*, *sex*, *age*, *occp*, *marital, nonwhite, educ, income, job\_seek, depress2, work1, comply, job\_disc*. In this section, a Regression Tree, Regression Random Forest will be fitted to the training set and the results will be analyzed in terms of performance and prediction and inference trade off.

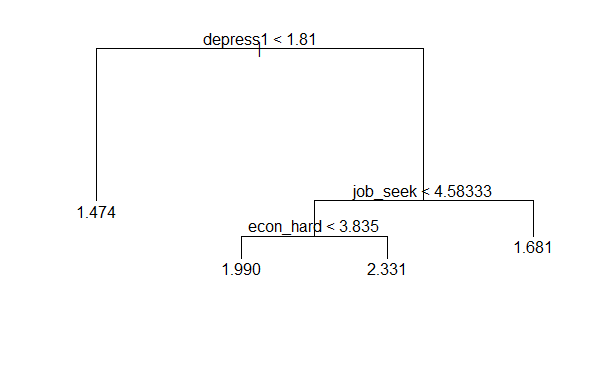
*Regressions Tree*

Regression Tree is first chosen for it easy interpretability, which provides good inference. The following is fully grown Tree fitted to the training set.



**Figure 1.12:** *Unpruned Tree*

The Test MSE of this model is 0.3543633 and this value will be used compare prediction performance with other models.. In order to avoid overfitting, cross-validation with K= 20, 50, 100, 300, LOOCV are performed. The CV results are as follows: 3,4,4,4,5 respectively, suggesting the best number of Tree nodes to keep that would minimize the model’s long run MSE. Therefore, picking the highest occurrence of suggestion: 4 nodes, the Tree is pruned accordingly and the following is the figure.



***Figure 1.12*:** *Pruned Tree*

The test MSE of the Pruned Tree is 0.3545102 which is slightly higher than the Unpruned Tree

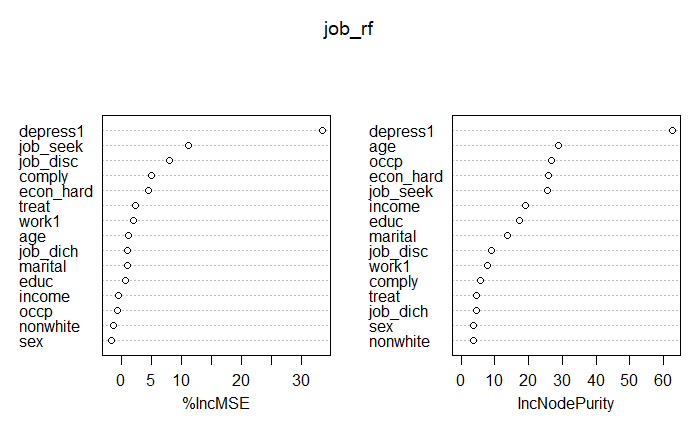
Test MSE: 0.3543633. However the difference is quite insignificant because in general, the Test MSE varies slightly when calculated from different testing sets. The Pruned Tree is easier to interpret and inference. Obtain some general inference from the diagram, the post-program depressive symptoms(*depress2*) is clearly lower if a subject had low depressive symptoms beforehand(*depress1*) or has a high job search self-efficacy(*job\_seek*). On the other hand, *depress2* is high when *econ\_hard* is high. According to this model, important factors of *depress2* are *depress1*, *job\_seek* and *econ\_hard*. However, Random Forest will be performed to further confirm this result.

*Regression Random Forest*

Random Forest is performed in attempt for better prediction performance than Tree. The mtry parameter is tuned from mtry=2 to mtry=16. This changes the size of a randomly selected set of predictors that the Trees uses to perform each split. This helps de-correlate each Tree as to increase stability of the model. It turns out, the model with a default value of mtry=5 yields the lowest OOB MSE and test MSE which are 0.34335 and 0.32069 respectively. These values are lower than the test MSE of the Tree, indicating that Random Forest does indeed predict better than the Tree. However, this is done so at the cost of interpretability and inference since Random Forest is built through averaging many Trees.

Although, one inference available is the variable importance. The following is a plot of each predictor and their importance in *depress2* prediction. The predictor with a higher %IncMSE the higher its importance with respect to the response variable. In this case, *depress1*, *job\_seek* and *job\_disc* are the most important predictors. However, since *job\_disc* is a measure based on *job\_seek* as described in the data description, these two predictors would be highly correlated. Therefore, the third most important predictor is the *econ\_hard*.

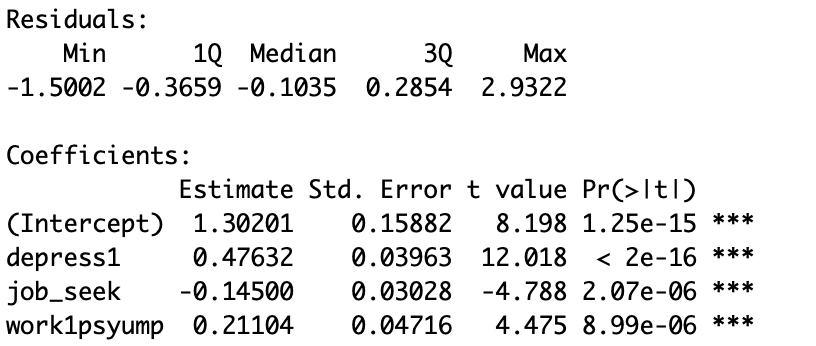
This result is consistent with the inferences made from the Tree, making it apparent that the treatment is not as indicative of depress2 levels as much as *depress1, job\_seek, econ\_hard* and *job\_seek* are.All of which indicate that subjects can influence their depressive state by their willingness to treatment or making effort in seeking employment.

**

**Figure 1.21:** *Variable Importance Plot for Regression Random Forest*

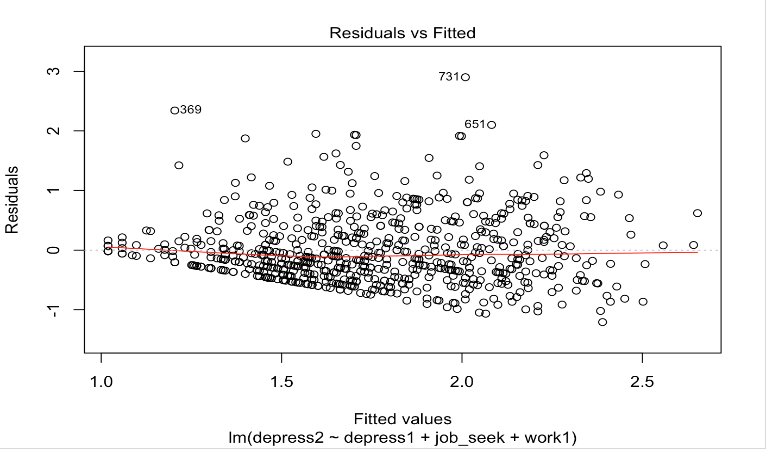
Analysis 2

Referencing the result obtained from the previous analysis, analysis 2 focuses on the three predictors with the largest effect on *depress2. This is done* to obtain better inference results. The Linear Regression model makes it clear that *job\_seek* has a negative relationship with *depress2*, coming in at about -0.2. Due to a higher *job\_seek* score being better (the score is out of 5), this indicates that the higher the self-efficacy of the subjects, the less depressed they are.



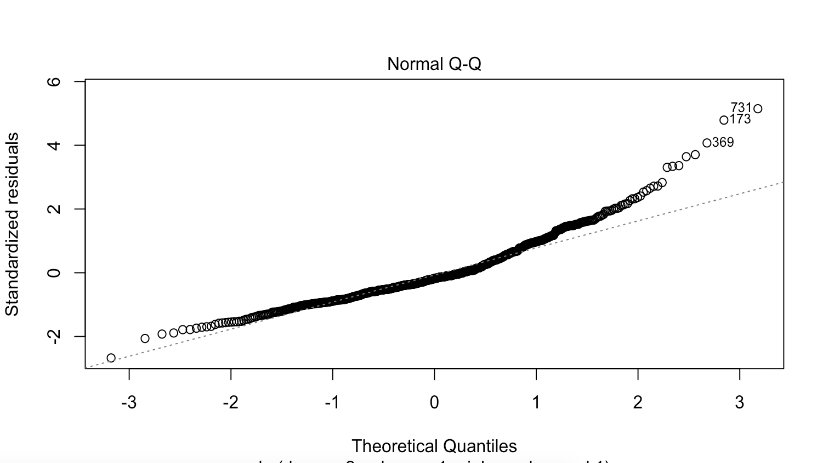
**Figure 1.31:** *Linear Regression Summary*

The slope of *depress1* is about 0.43 which reveals that *depress2* does not rise as quickly as *depress1.* This indicates that the treatment program, is effective in some capacity. However, the existence of a control group does affect this data. The ratio of control group to treatment group is 1:2. Therefore, the effect of the slope is not as significant, but there is some change indicated. The true slope when taking out the control groups would be about the result of 0.57 - 0.33 which is 0.24.



***Figure 1.32*:** *Residuals vs Fitted Graph*

Next, looking at the Residuals vs Fitted graph, we can see that the data is relatively shapeless which is good. However, there are areas of higher density, and a couple extreme outliers.. Despite the slight obstacles, this graph shows that there are probably no non-linear relationships in the data.



***Figure 1.33*:** *Q-Q Plot*

Lastly, looking at the QQ-Plot, due to the amount of curve, I would say that this plot is not very normally distributed and has a right skew. Also, for the outliers, I consider them important because of how the direction of the QQ-Plot was skewing that way anyways, which validates them more.

Compared to analysis 1, even though there there are only three predictor variables in this analysis, the MSE goes down, showing that the use of the other additional variables can be detrimental to the data.

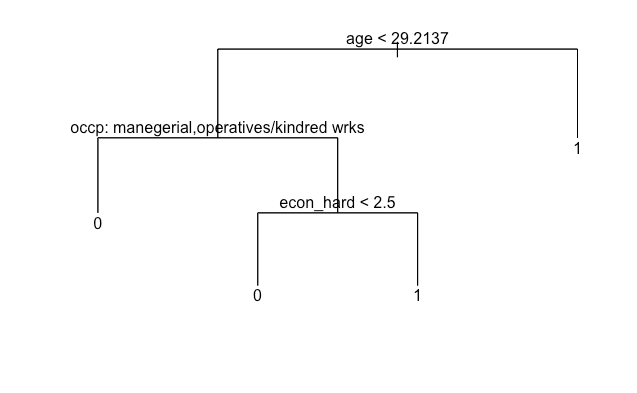
Classification

For classification, the *comply* feature was chosen as the response variable with all other applicable variables used as the predictors.  Only the rows with *treat* = 1 were chosen as they are the only rows where comply can be 1.  Three models were trained on the same subset and benchmarked on the same testing set. They were then compared to each other by examining precision, recall, F1 Score, specificity and accuracy from the test set. However, the primary attributes being examined are accuracy, as well as any inference they can be used for. Accuracy is chosen because it is the primary goal of these models. F1 score is chosen because it is on one of the best ways to measure a balanced, in addition to this, the test set is slightly unbalanced towards the negative side which makes F1 score a good of ensuring the models aren’t overly biased towards either case.

*Classification Tree*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Precision: 0.415  Recall: 0.379  F1 Score: 0.396  Specificity: 0.663  Accuracy: 0.553 | |  |  |  | | --- | --- | --- | |  | *Compliant* | *non-compliant* | | *Predicted Compliant* | *22* | *31* | | *Predicted non-compliant* | *36* | *61* |   **Figure 2.11:** *Confusion Matrix for Classification Tree* |

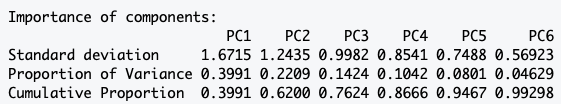
The  Classification Tree was k-fold cross validated at k = 10,20,50,100 and 450(Leave one out CV) and then pruned to 4 terminal nodes based on the results. Having been trimmed down to 4 leaves, many of the 14 initial parameters are made meaningless which resulted in the Tree in Figure 2.12.  The only parameters remaining are *oocp*, *econ\_hard* and *age*.  This makes the Tree relatively easy to interpret but its testing metrics are quite low across the board.

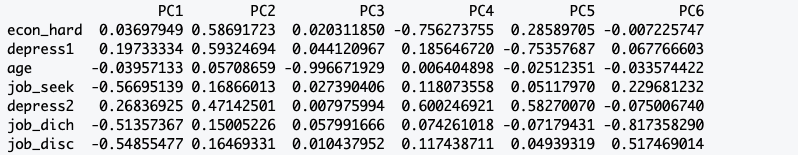
**

**Figure 2.12:** *Classification Tree with 4 leaves*

*Principle Components*

Three components were chosen using Kaiser Criterion, other amounts were tested but 3 gave the best results. Together they accounted for 75% of the variance. The first component was largely related to *job\_disc*, *job\_dich* and *job\_seek* making it seem like anything to do with job search efforts. The second is most related to *depress1*, *depress2* and *econ\_hard* making it seems like anything to do with depression (*econ\_hard* is self-evaluated, making it somewhat tied to depression). The third component was almost entirely tied to *age* with a negative correlation. It is also worth noting that the PCA uses only the numeric variables in the dataset giving it only 7 predictors instead of 14.



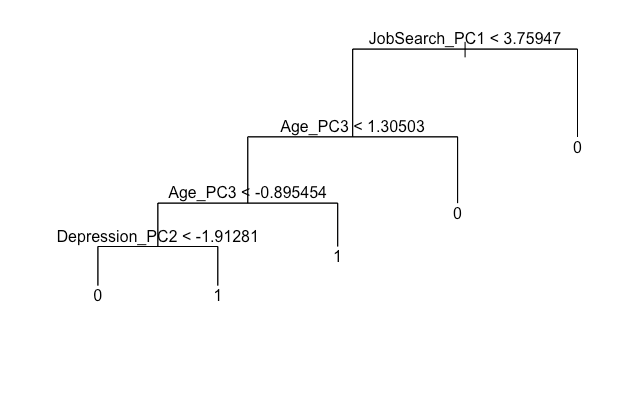


**Fig 2.21:**  First 6 Principal Component Loadings

*PCA Tree*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Precision: 0.588  Recall: 0.172  F1 Score: 0.267  Specificity: 0.924  Accuracy: 0.633 | |  |  |  | | --- | --- | --- | |  | *Compliant* | *non-compliant* | | *Predicted Compliant* | *27* | *18* | | *Predicted non-compliant* | *31* | *74* |   **Figure 2.22:** *Confusion Matrix for PCA Tree* |

The PCA Tree outperformed the regular Tree in accuracy but had a terrible F1\_score of 0.267. It ended up mostly guessing the majority class making the model mostly useless in terms of classifying. The inference should also be used carefully and only referenced with less biased models. The inference of variable importance and variance distribution acquired from the PCA itself is quite useful.

**

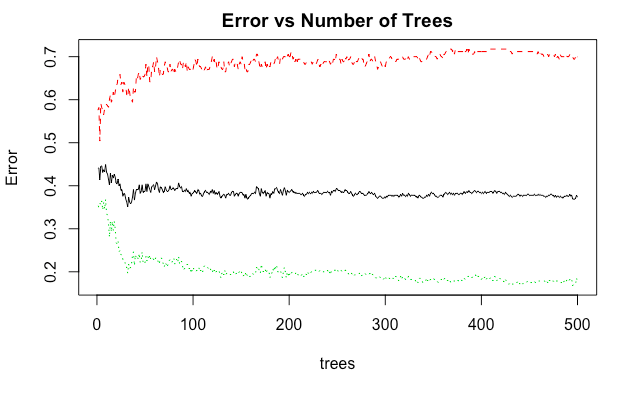
**Figure 2.23:** *PCA Tree*

*Classification Random forest*

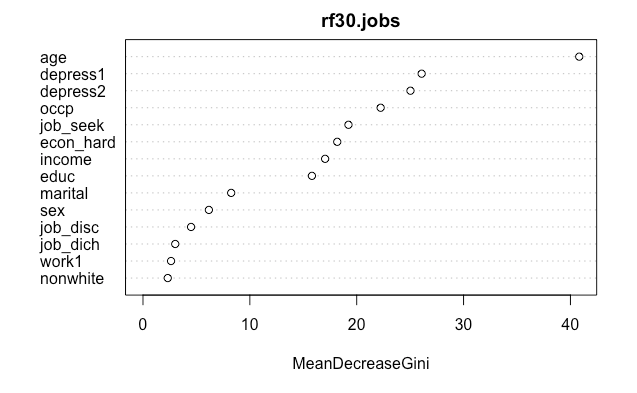
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Precision: 0.600  Recall: 0.466  F1 Score: 0. 524  Specificity: 0.804  Accuracy: 0.673 | |  |  |  | | --- | --- | --- | |  | *Compliant* | *non-compliant* | | *Predicted Compliant* | *10* | *7* | | *Predicted non-compliant* | *48* | *85* | |

**Figure 2.31:** *Confusion Matrix for Random Forest with 30 Trees*

Random forest was cross validated using bagging and then compared different mtry numbers(see Regression Random Forest for mtry explanation) to each other and 6 was chosen as the best amount. The forest was then plotted with out of bag error vs number of Trees (shown in fig 2.32) and 30 Trees was chosen as the optimal amount. The RF has a fairly high benchmark scores on the test set and lends itself fairly well to interpretation thanks to its variable selection shown in fig 2.33. The variable importance also matches with what we saw with the Tree where *age* was the most important variable.

**

**Figure 2.32:** *Random Forest Trees decision plot*

****

***Figure 2.33****: Variable Importance Plot for Random Forest*

Comparison

The Classification Tree is a reasonable model but with an accuracy of 55.3% and an F-score of 0.396 it’s lacking heavily in terms of prediction and isn’t accurate enough to be very confident in its interpretation. The PCA Tree guessed mostly the majority class and as such does not give a very reliable interpretation or classification results on its own. However, the PCA results give some interesting variable importance results and the PCA Tree can be compared with the other models for inference..  On the other hand, Random Forest outperforms the other 2 models in all measurements used and does very well in interpretability, giving us a fairly clear picture of variable importance in figure 2.31. With an accuracy of 67% and an F1 score of 0.524, it does the best job for predicting among the 3 models tested as well as the most reliable interpretation. The PCA and variable importance plot from Random Forest also gave similar results, showing *age*, *depress1* and *depress2* to be some of the most important variables. Using the loadings from PCA in figure 2.21 and the PCA Tree in figure 2.23 it is found that the older you are, the more likely you are to comply(see node 2 and 3 in figure 2.23, keeping in mind the negative correlation with *age*). Furthermore, the higher PC2, the more likely you are to comply. The main reason for this is theorized to be due to the *econ\_hard* level being high, making the subject want to comply with the treatment programs more to improve their economic standing. This also matches the last node on the normal Classification Tree in figure 2.12 involving *econ\_hard*. These findings match up with the Random Forest Variable importance plot. This could help future studies in similar fields pick candidates that are more likely to comply with the program. These findings show how models with some weaknesses or faults can be used together to draw some interesting and convincing conclusions on data.

Clustering

*Hierarchical Clustering*

A distance matrix is calculated using Euclidean Distance as the measure for each predictor values. With this distance matrix, a number of clustering with different linkage types are made.

1. Single Linkage: After applying this linkage type, 2 groups are found. (Figure 3.4)
2. Average Linkage: After applying this linkage type, 2-3 groups are found. There is a mix of very small and big groups. (Figure 3.1)
3. Complete Linkage: After applying this linkage type, 4-6 groups are found. This includes 4 major groups and few minor groups. (Figure 3.5)
4. Ward.D Linkage: After applying this linkage type, 2 distinct groups are found with few smaller groups. It was densely packed. (Figure 3.3)
5. Ward.D2 Linkage: After applying this linkage type, 2-4 groups are found. (Figure 3.2)

After plotting each dendrogram, the Tree is cut several times so there can be 2,3,4 or 5 clusters. Since Ward.D Linkage gives the best result in terms of distinction of clusters, this linkage type is preferred.

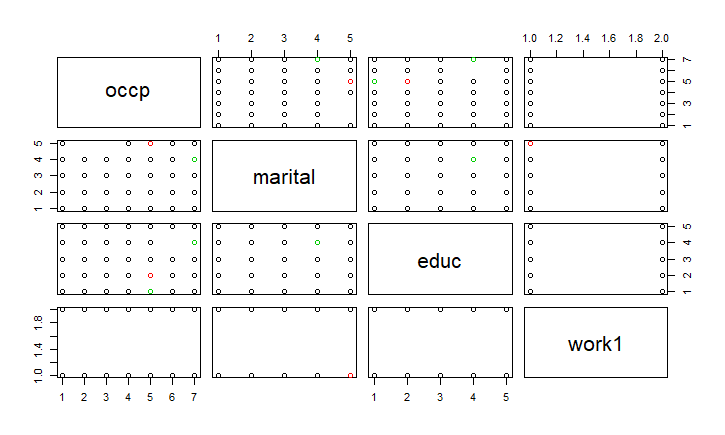
|  |  |
| --- | --- |
| https://lh4.googleusercontent.com/9kqaHVWhY8cxa5vgbNy1g2eqU1nnDAK1_HGWHHw6fcRCre-TWZH9v4TZXLPkpoGwPC__OlEs1QrxRrW5WErrU_Ok1N4IP3KqS3P2ddNQyRPyc-Ck2qoZ1Nb72fq2BOPUO5wWF9CK  **Figure 3.1:** Average Linkage | https://lh5.googleusercontent.com/lOUTKfArxXaoTfCXlaUtLG_ycKXO5gydLRFAbO5eNTgmrUJ6mvG7cJAZTHinlOSAhRzwRD8r_uGGKCH4R6tkvcOvIE0p8S_lOaKVLqGe3E4_KPkFew9FW3x3a5sbgOFmgoIeH2Qg **Figure 3.2:** Ward.D2 Linkage |
| https://lh5.googleusercontent.com/s2DdWB9IzgzewFHY8DEux-6Y9TEjfZoyBcsjeACOyiJi4lk8tlLjmCPQe8H_rhFvnFUbtLdFm0GUtbNjGIaQ0H-jM-yWdMUUP__rglI3TAMq-_QmjbSre5-N-VG2VksYylHeYZwb  **Figure 3.3:** Ward.D Linkage | https://lh6.googleusercontent.com/iOmi4UEUWE1PFyid4WPMG3cCrP-FuZfUY6c7Z4cUt2tI1mB7l1rlylTRxcZCLboN718-IjZCaXz_Ap1l0cGlB8uDqTqDKRAueW1ICU4M9LNjrwr2WiAWLcKwufg19af685dGODum  **Figure 3.4:** Single Linkage |
| https://lh5.googleusercontent.com/kMBRWc71_QAIUY1RWwDf8a3qhBv-OIfzlo2WUFC7I3m2vTFY1dfL8_VmYguhT062RjM_PnUKmkPjDcDmIK4Y-oGWtfl4iaJbPSFOt0tcDFmO1mBqVQvzDuJx2jtTPkAzykEl4U2y  **Figure 3.5:** Complete Linkage |  |

*K-Means Clustering*

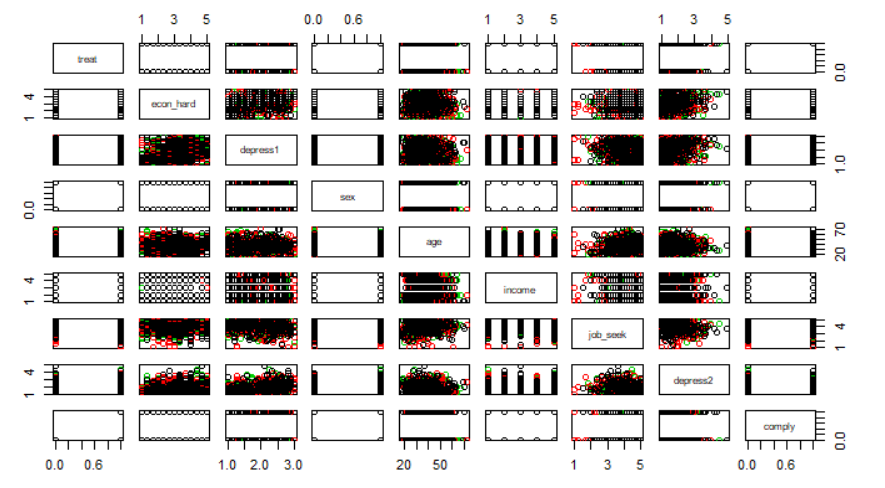
Several sets of variables are plotted while performing K-Means Clustering and none of them shows any distinct group. The variables were mixed and matched for several plots but all of them failed to show any interesting results. None of them had clusters that could be used for predicting the relationship between the variables in the dataset. Within cluster sum of squares by cluster is: [1] 2227709 2250137 2250139

 (between\_SS / total\_SS =  88.9 %)

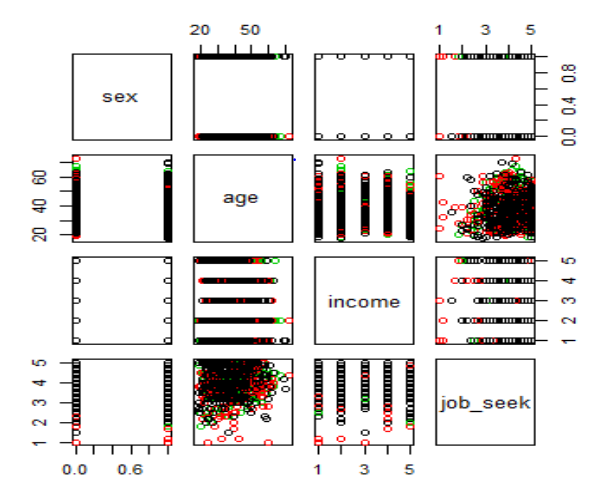
The plot (Figure 4.1) below compares the variables: sex, age, income and job seek, but none of them show any distinct group. This indicates that they are not connected.Similarly, the next plot (Figure 4.2) compares variables that include almost half the dataset including *job\_seek, age, sex, income*, etc. It shows no distinct group either. One more plot (Figure 4.3) using the variables occupation, marital status, education and *work1* is plotted and once again it shows no clusters, indicating almost no correlation between the variables.



**Figure 4.1:** K-means on variables that include sex, age, income and job seek



**Figure 4.2:** K-means on variables that include almost half the dataset



**Figure 4.3:** K-means on occupation, marital status, education and work1

Conclusion

From the analysis of the Job Search Intervention Study, the following information was gathered. A Pruned Regression Tree demonstrated that the three most important predictors for the response variable *depress2*, are *depress\_1*, *job\_seek* and *econ\_hard/comply*. This was further verified by the variable important plot from the Random Forest Regression and also it’s more accurate prediction. Linear Regression further improved prediction accuracy with predictor variables: *depress1, job\_seek* and *work1*. The MSE is substantially lower than the Test MSE of Tree and Random Forest. Thereby providing more accurate prediction. From the above models, it is apparent that making an effort to seek employment and willingness to comply with the treatment process responds to lower depression levels(depress2), therefore implying that it is not necessarily the treatment itself, that helps improve subjects’ metal state but their effort in seeking reemployment.

Classification on the comply feature resulted in Random Forest being the best model to predict compliance. Through PCA and the PCA Tree, it was found that the older you are, the more likely you are to comply with the treatment, which is useful information for future treatments. Lastly for clustering, both hierarchical and k-means clustering were performed. When using Hierarchical Clustering, Ward D. linkage showed the most distinct groups compared to the other 4 types used in the data and so Ward.D Linkage is considered the most preferred linkage for this dataset. When performing k-means clustering, no distinct groups were identified. This indicated that k-means clustering is not ideal for this dataset.

Reference

Vinokur, Amiram, and Yaacov Schul. "Mastery And Inoculation Against Setbacks As Active Ingredients In The JOBS Intervention For The Unemployed". *Citeseerx.Ist.Psu.Edu*, 2019, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.583.4019&rep=rep1&type=pdf. Accessed 1997.