December 12, 2022

Exploration of Government's Assessment Policy For Social Welfare Program (SWP)

Si Cheng (1003834158),
Giriraj Heda (1007992415),
Yijun Lin (1003404620),
Rohit Pandit (1009704465)

Instructor: Prof. Tegan Rajkumar Maharaj

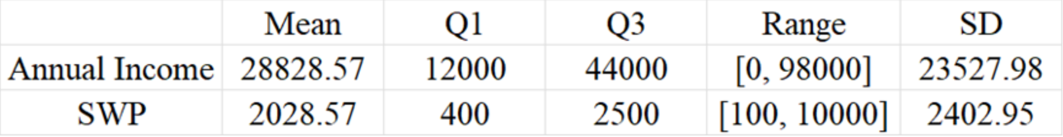
Abstract:

The focus of this research is to explore the US Government's Social Welfare Program (SWP) distribution policy, using a dataset from Kaggle, "Unemployment and mental illness survey". The purpose of selecting this dataset was to understand how SWP can support people during periods of recession, mass layoffs, brutal weather, etc. Logistic regressions were applied to predict the probability of receiving the compensation, followed by predicting the amount that people would receive according to their background, based on the analysis from the decision tree and the random forest. Our study finds that the current assessment policy of the government is not thorough enough as there are people in the social welfare program receiving money when they ideally should not be receiving it or should not be receiving that much.

1. Introduction:

Social welfare refers to various programs set by the government, aiming to help individuals that cannot support themselves. Usually, there exists a list of requirements to check whether a person is eligible for the benefit or not (*CFI Team, 2022*). Although welfare seems to be a big help to society, there are still many issues around it, among which two problems raise some controversies. One is that extremely poor people are usually missed. And the other problem refers to the improper payments and fraud that resulted from the inability of the federal government to supervise the income reported by applicants (*Welfare Issues Page*, n.d.). This leads to a critical question: Is the current eligibility check for receiving money from the Social Welfare Program (SWP) valid enough? It is the desire of the government to make sure the money is allocated to people who really are in need, which inspired us to check if there exists any unfairness in the current SWP policy and distribution by diving into survey data collected by NAMI (National Alliance on Mental Illness) and uploaded on Kaggle (*CORLEY, 2018*). The survey explored the linkage between mental illness and unemployment, where people's income level, mental health, physical conditions, and other basic statuses were asked. It was also indicated in the survey questions whether a person had received certain types of SWP and what benefit they received. The raw dataset generated from the survey has a size of 334 and 31 features in total. We started the data cleaning process by dropping a few highly overlapped features and non-trivial features that had little or no effect on the question of interest. By doing this we narrowed down the scale of features to 14 and then renamed them. In addition, most features are categorical with corresponding levels in text format, we then clubbed those overlapping levels and converted them into numbers that represented different categories under each categorical feature, in order to make the report more clear.

Our analysis started from an overview of the annual income level and the distribution of the amount of SWP compensation, where significance varieties were shown in the descriptive table.



	Mean	Q1	Q3	Range	SD
Annual Income	28828.57	12000	44000	[0, 98000]	23527.98
SWP	2028.57	400	2500	[100, 10000]	2402.95

Table 1.1 Descriptive Table for Annual Income and SWP

Since the annual income level is a critical factor that affects a person's eligibility of receiving the SWP compensation (*About Program Income and Public Assistance, 2022*), we assume that people with a higher annual income would receive less compensation. While the scatter plot regarding these two variables shows a weak positive trend between the Annual Income and SWP with several outliers, which seemed to be opposite to our assumption. It is therefore in the interest of this research to explore the government's current SWP distribution policy in terms of eligibility and the amount of the benefit granted.

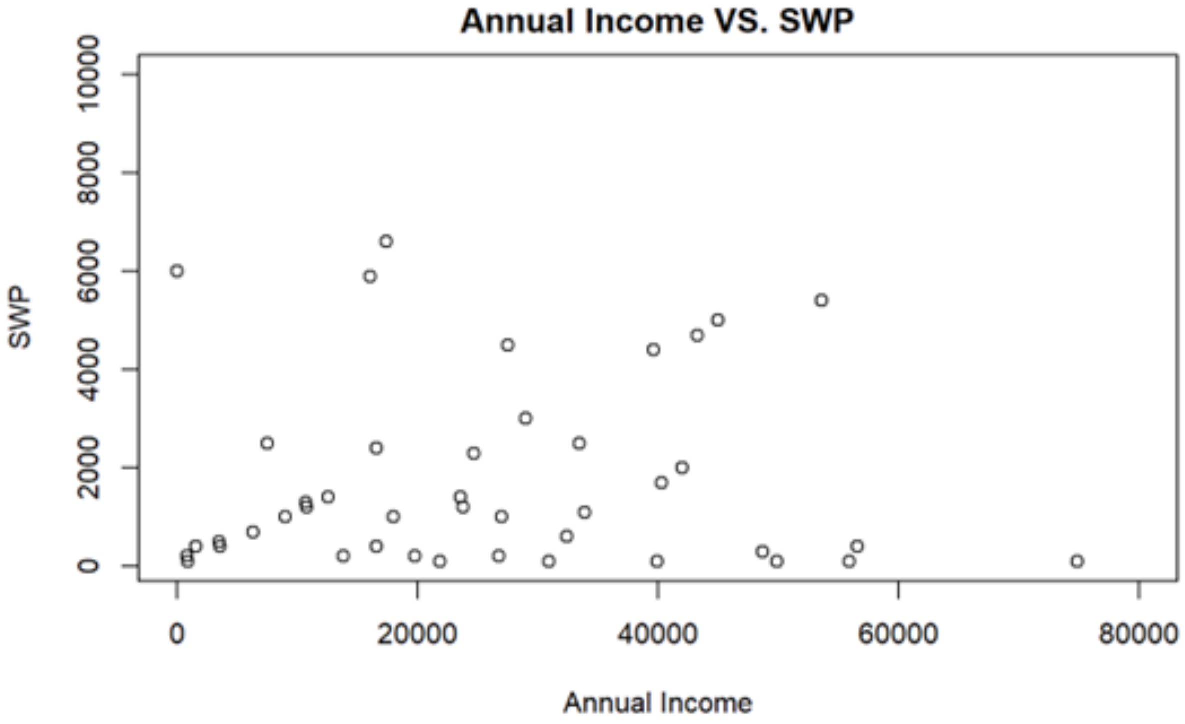


Figure 1.2 Scatter Plot of Annual Income and SWP

2. Methods:

First, the partial variables are eliminated due to their relatively weak linkage with our research goals. Also, dependent variable SWP was mutated with outcomes of 0 (No) and 1 (Yes) to take eligibility at first consideration. Even after the process of elimination, multiple variables (a total of 14 variables) were still not applicable and were not the best fit for logistic regression models (since the outcome for dependent variable is 0 or 1). To find variables that were most suitable, we made use of backward selection process and eliminated the variable step by step by removing variables that had the highest p-value and compared on how the AIC changed between different models. Six different logistic regression models were created and the best performance model with 9 independent variables are selected. After selecting the best logistic regression model, we wanted to check for the current model and its accuracy to see whether the current variables are sufficient to predict the probability of receiving SWP.

As proof of evidence, by using machine learning's Logistic regression algorithm we split data into test and train subsets for the current model and analyzed the AIC result, and model accuracy for test and train data. Also took the ROC-AIC curve as part of the consideration. Based on these three components' results, we reached a proper conclusion for the first part of our question.

Finally, even with the high accuracy of the logistic regression model, we still wondered based on the selected features & individual's background, can we predict using different approaches which algorithm can be used for determining the SWP amount. Making use of Decision Trees and Random Forest algorithms, we were able to plot the Decision Tree and use Random Forest to estimate the amount based on user inputs using the variables selected. By adjusting the probability weights parameter of the decision tree and trying to estimate the amount for different conditions, we were able to compare multiple results and analyzed if the model was reasonable and hence, we were able to make comments and limitations based on the obtained results.

3. Results & Discussion:

By constructing multiple Logistic Regression models with the combination of variables, we were able to compare the AIC scores of these models and select the best one which had variables, Education, Disabled, Internet Access, Annual Income with SWP, Unemployed, Read Books (to get extra skills), Times Hospitalised (indication of the strength of mental illness), Age, Annual Household Income. Now since we had decided on the variables that are most suited for determining SWP, we wanted to check how accurately these variables would identify if the person received SWP. The accuracy for test data was 84% and using the same model, now we were able to run custom inputs and check the probability of how likely the individual would receive SWP.

As the accuracy of the model looks promising, we now were interested in determining how the SWP amount be distributed to individuals by training a model using the same dataset. By adjusting the probability parameters while splitting the dataset into train and test subsets, we analysed different Decision Trees. The results obtained suggest that there are individuals in the SWP program that are receiving high amounts whereas they either should be receiving comparatively less or no benefits. For instance, the model distributed a sum of $1967 to individuals who were hospitalised more than 4 times and no other criteria were taken into consideration. Another such interpretation which can be drawn from the figure below is, individuals who were hospitalised less than 4 times, do not read books, and have a household income greater than or equal to the range of $100,000 - $124,999 were receiving $454 whereas they could be easily supported by their family members. Also, individuals with hospitalisation of less than 4 times, who do not read books, and who have a household income in the range of $50,000 - $99,999 were being allocated a whopping $2529.

These inferences from the model below suggest that the current assessment of the government to identify individuals' eligibility and amount from the SWP program is flawed. The model does not even consider the employment status of the individuals or their disability and defeats the whole purpose of the social welfare program.

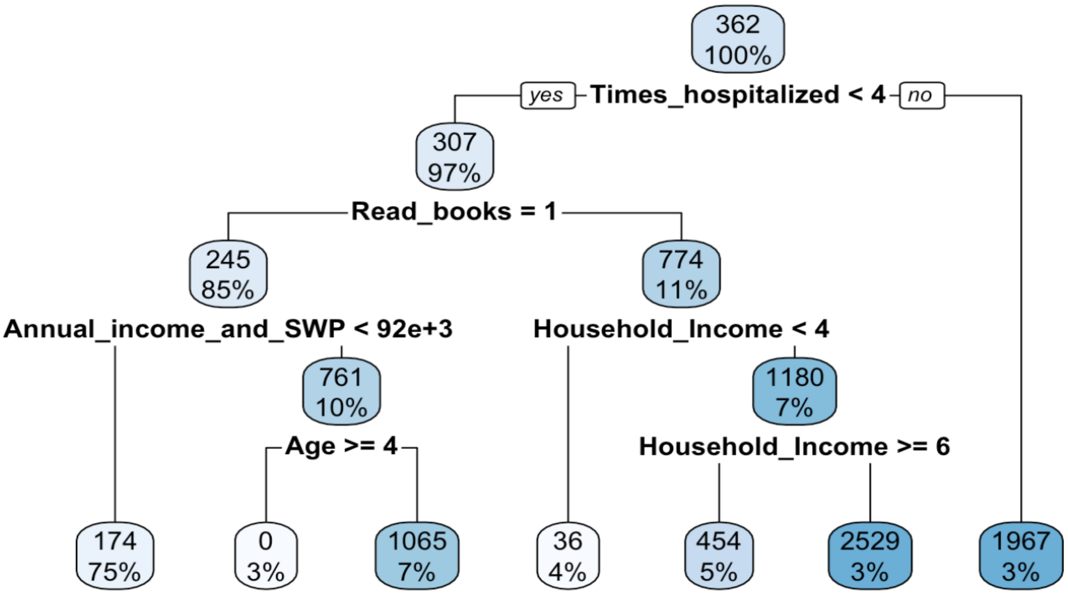
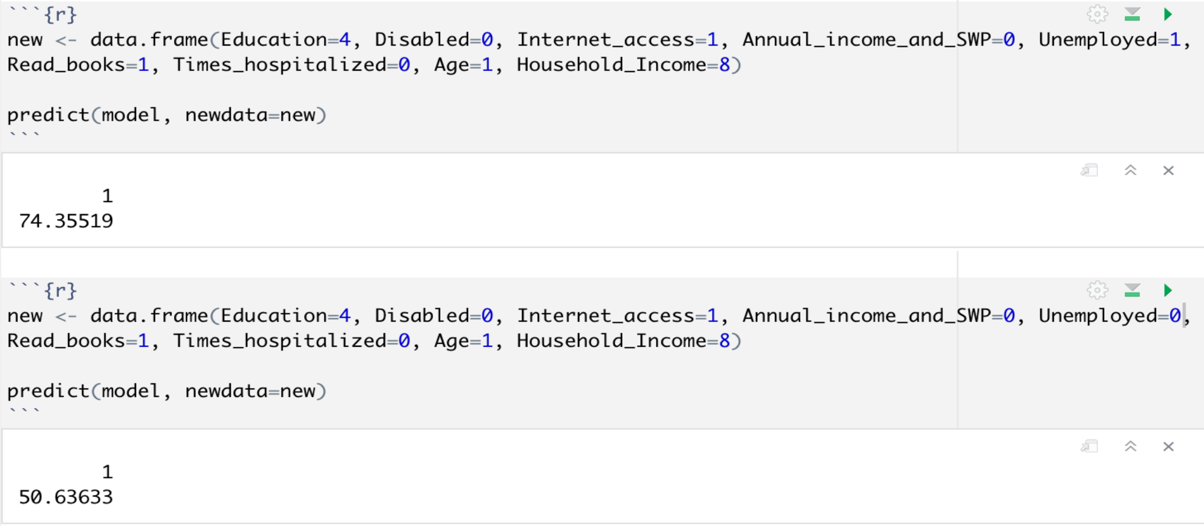


Figure 3.1 Decision Tree for SWP

As the results obtained by using the Decision Tree were surprising, we wanted to corroborate our analysis using another machine learning method and decided to use the Random Forest algorithm to estimate the SWP using the same variables obtained from the Logistic Regression model. Similar results were given by the Random Forest algorithm which allowed us to endorse the results of the Decision Tree. The SWP value did not change by changing the employment status and the amount was being distributed to individuals who can be supported by other factors. Figure 3.2 represents the amount distribution for individuals with high household income ($150,000 - $174,999) and only make changes in employment status.



~~~
```{r}
new <- data.frame(Education=4, Disabled=0, Internet_access=1, Annual_income_and_SWP=0, Unemployed=1,
Read_books=1, Times_hospitalized=0, Age=1, Household_Income=8)

predict(model, newdata=new)
```
~~~

```
       1
74.35519
```

~~~
```{r}
new <- data.frame(Education=4, Disabled=0, Internet_access=1, Annual_income_and_SWP=0, Unemployed=0,
Read_books=1, Times_hospitalized=0, Age=1, Household_Income=8)

predict(model, newdata=new)
```
~~~

```
       1
50.63633
```

Figure 3.2 Output from Random Forest

The inference made from the analysis of the Decision Tree is corroborated by the Random Forest as well. Using Random Forest, we were able to analyse the variable importance using the variable importance plot which suggests that variables, employment status and disability have very low importance and would not affect the result of the models by a great margin.

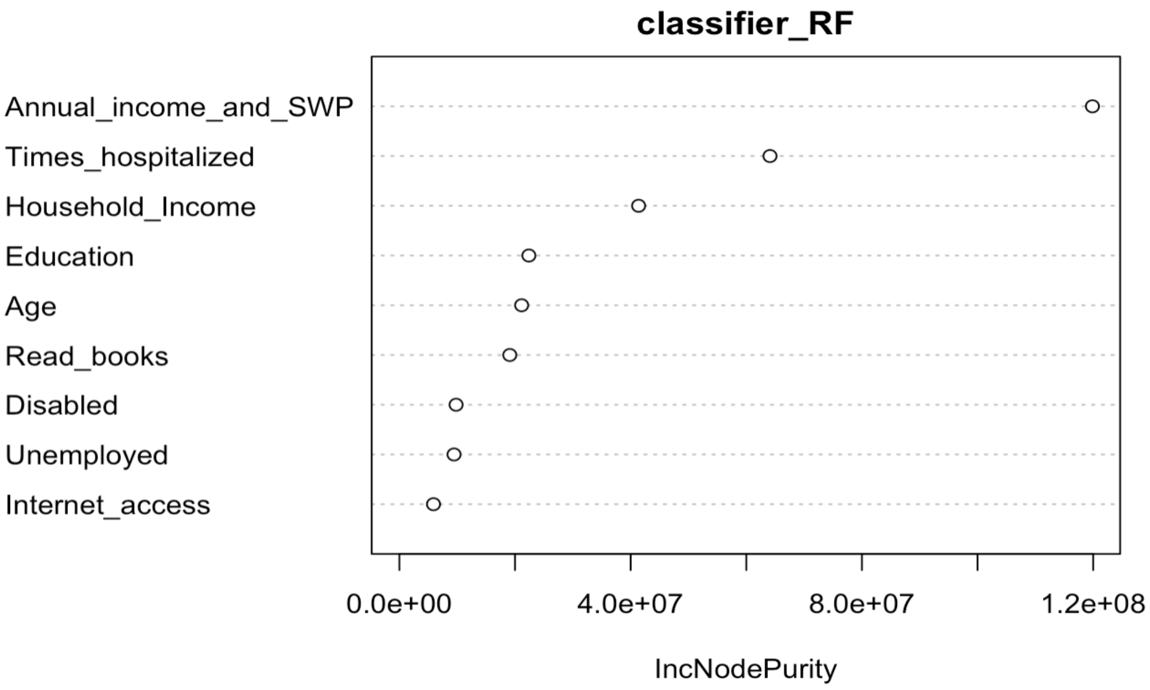


Figure 3.3 Output from variable importance plot

From all the analysis conducted above, it is evident that the current policy of the government to assess individuals and distribute amounts is flawed. Major factors which should be considered for the assessment are not being examined properly and individuals are receiving amounts when they ideally should not be receiving. This results in the negligence of the extremely poor population for whom the policy and the program were created and is serving as no means of upliftment of such individuals.

4. Limitations and Conclusion:

Limitations

The dataset has some restrictions and did not provide us with any insights about the relationship that the individuals with higher household incomes had with their families. We did not have any evidence showcasing they were supported by their families which might result in the individual not receiving the required support. All these cases can be taken into consideration while evaluating the revised welfare program.

Conclusion

On analysing the test results, it can be concluded that the government's current Social Welfare Program (SWP) has some gaps and cannot be solely relied upon for the decision-making processes. There is scope for improving the efficiency of the distribution model further by taking specific factors into consideration such as mental illness, disability status, etc. We came across some unexplainable decisions in our analysis. For instance, individuals hospitalized more than 4 times are given an amount without considering their financial, educational and other relevant variables which should be considered for making decisions for the majority of the cases. Moreover, some individuals with higher household incomes are receiving compensation in comparison to individuals with lower household incomes. Thus, the SWP policy assessment needs to be re-evaluation to eliminate these inconsistencies during the decision-making process. This will help in the efficient distribution of resources while ensuring support for the people in need.

References

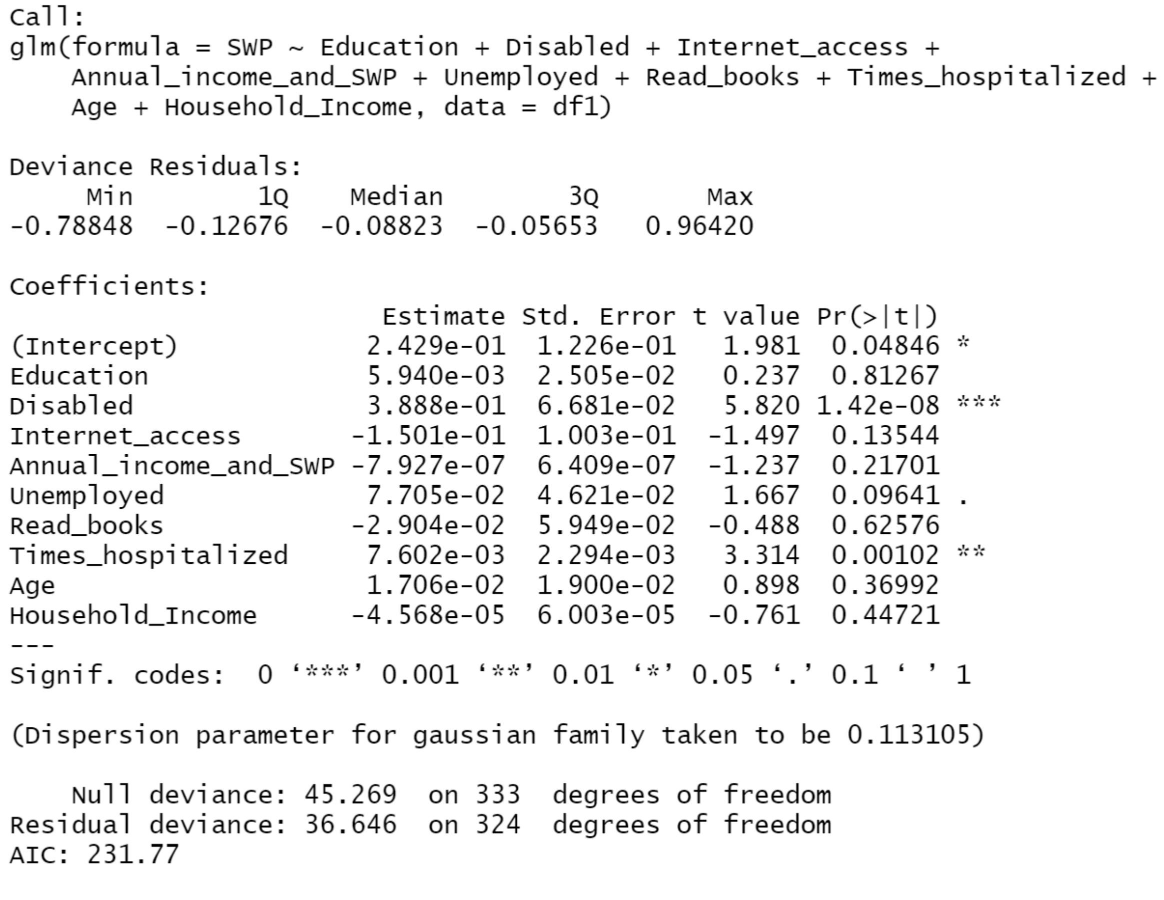
Census Bureau (2022, October 3). About Program Income and Public Assistance. Retrieved December 10, 2022, from https://www.census.gov/topics/income-poverty/public-assistance/about.html

CFI Team. (2022, November 7). Social Welfare System -Overview, How It Works, Examples. Corporate Finance Institute. Retrieved November 27, 2022, from https://corporatefinanceinstitute.com/resources/economics/social-welfare-system/

CORLEY, M. (2018). Unemployment and mental illness survey. Kaggle. Retrieved December 10, 2022, from https://www.kaggle.com/datasets/michaelacorley/unemployment-and-mental-illness-survey

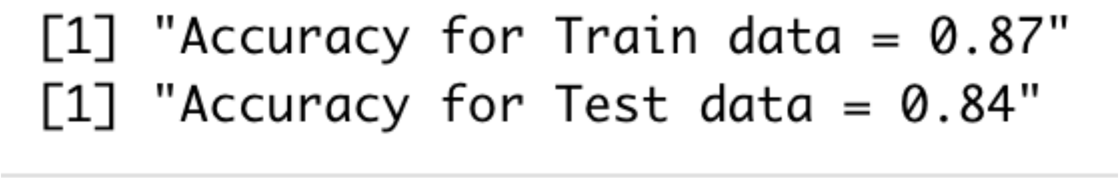
Welfare Issues Page. (n.d.). - *Federal Safety Net. Retrieved November 27, 2022, from* https://federalsafetynet.com/welfare-issues/

Appendix



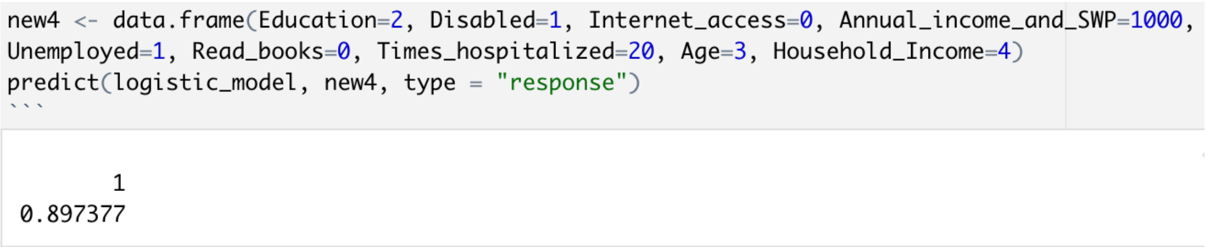
```
Call:
glm(formula = SWP ~ Education + Disabled + Internet_access +
    Annual_income_and_SWP + Unemployed + Read_books + Times_hospitalized +
    Age + Household_Income, data = df1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.78848  -0.12676  -0.08823  -0.05653   0.96420

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.429e-01  1.226e-01   1.981  0.04846 *
Education              5.940e-03  2.505e-02   0.237  0.81267
Disabled               3.888e-01  6.681e-02   5.820 1.42e-08 ***
Internet_access       -1.501e-01  1.003e-01  -1.497  0.13544
Annual_income_and_SWP -7.927e-07  6.409e-07  -1.237  0.21701
Unemployed             7.705e-02  4.621e-02   1.667  0.09641 .
Read_books            -2.904e-02  5.949e-02  -0.488  0.62576
Times_hospitalized     7.602e-03  2.294e-03   3.314  0.00102 **
Age                    1.706e-02  1.900e-02   0.898  0.36992
Household_Income      -4.568e-05  6.003e-05  -0.761  0.44721
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.113105)

    Null deviance: 45.269  on 333  degrees of freedom
Residual deviance: 36.646  on 324  degrees of freedom
AIC: 231.77

Number of Fisher Scoring iterations: 2
```

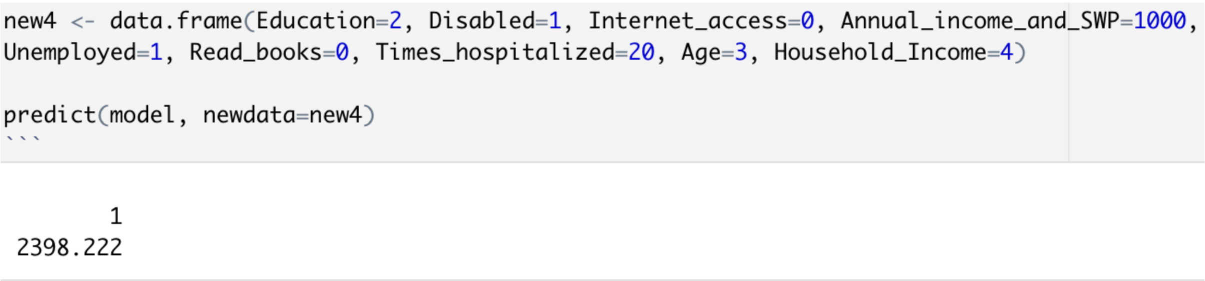
Appendix Figure 1 Output for Logistic Regression with best suited features



~~~~
[1] "Accuracy for Train data = 0.87"
[1] "Accuracy for Test data = 0.84"
~~~~

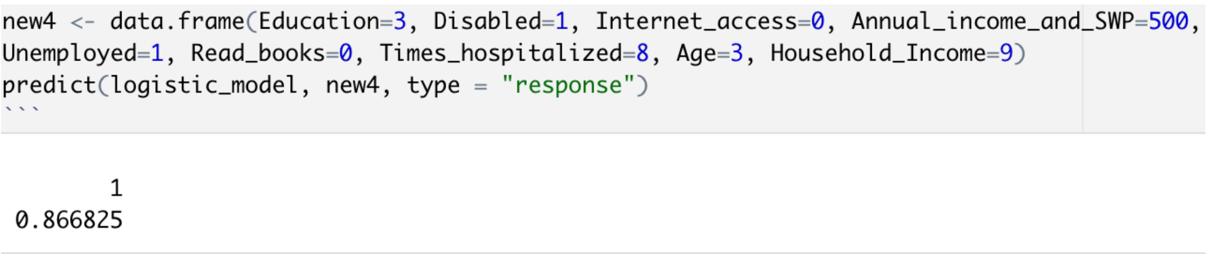
Appendix Figure 2 Accuracy of the train and test model for Logistic Regression

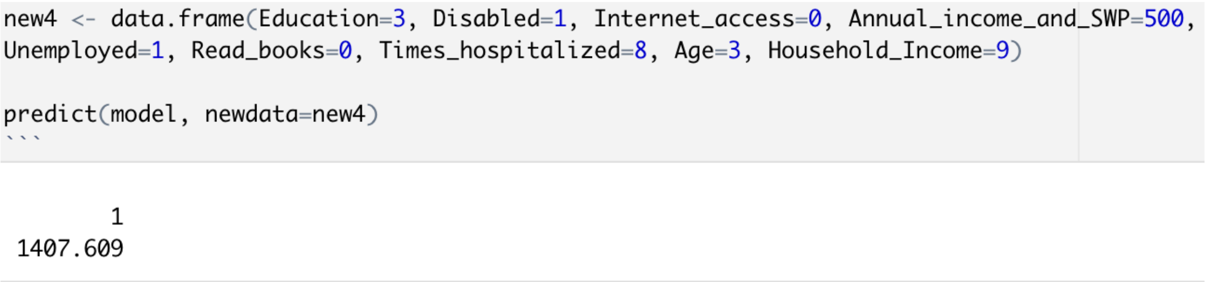


~~~~
new4 <- data.frame(Education=2, Disabled=1, Internet_access=0, Annual_income_and_SWP=1000,
Unemployed=1, Read_books=0, Times_hospitalized=20, Age=3, Household_Income=4)
predict(logistic_model, new4, type = "response")
```

       1
0.897377
~~~~

Appendix Figure 3.1 Probability for the sample 1 inputs(Poor personal conditions)



~~~~
new4 <- data.frame(Education=2, Disabled=1, Internet_access=0, Annual_income_and_SWP=1000,
Unemployed=1, Read_books=0, Times_hospitalized=20, Age=3, Household_Income=4)

predict(model, newdata=new4)
```

       1
2398.222
~~~~

Appendix Figure 3.2 Sample 1 SWP amount predicted by Random Forest



~~~~
new4 <- data.frame(Education=3, Disabled=1, Internet_access=0, Annual_income_and_SWP=500,
Unemployed=1, Read_books=0, Times_hospitalized=8, Age=3, Household_Income=9)
predict(logistic_model, new4, type = "response")
```

       1
0.866825
~~~~

Appendix Figure 4.1 Probability for the sample 2 inputs (Poor personal conditions with high household income)



~~~~
new4 <- data.frame(Education=3, Disabled=1, Internet_access=0, Annual_income_and_SWP=500,
Unemployed=1, Read_books=0, Times_hospitalized=8, Age=3, Household_Income=9)

predict(model, newdata=new4)
```

       1
1407.609
~~~~

Appendix Figure 4.2 Sample 2 SWP amount predicted by Random Forest