**Binary Approach:**

After selecting the important variables after the data cleaning process, the natural next step is to study the statistical relationships of these variables to the dependent variable `SCORE\_HAPPY`. As a first step, the SCORE HAPPY variable has been converted into a binary variable as `HAPPY` where people with more happiness score equal to or more than 6 is considered as Happy (1) and with 5 or less as unhappy (0).

Now, Binary model can be easily applied to the dataset and the relationships can be studied in detail which are as follows.

1. **DATA CLEANING:**

First step is to understand the variables and convert categorical variables into appropriate category.

***1.1 AGE:*** This first step is to convert the Age variable from continuously varying integer variable into categorical variable. Age has been converted into 4 groups.

* Less than 25 years = Adults
* Between 26 and 45 years = middle-aged adults
* Between 46 and 60 years = Adults
* More than 60 Years = Older adults

1. **First Iteration:**

First Iteration is to use all the variables in the logistic regression model and predicting the variables that are important for determining the binary Boolean variable `HAPPY`.

**2.1 Data Cleaning:** First step of modeling is to check the null and nan values. Initial inspect of the data for null values shows us that the variables like `HAD\_CHILDREN`, `ASHAMED\_LGBTQ`, `SATISFACTION\_GOVERNMENT` has more null values. Initially the data had 2745 observations. For first iteration, after dropping all the rows with null and nan values, there are 1291 observations which is 47% of total observations.

**2.2 Model fitting:** Next step is to fit the data to the logistic regression. The model converges after 7 iterations with function value 0.26.

**2.3 Model Summary:** After checking the model summary we can see that the first iteration model has converged with pseudo R2 of 0.3926. The log-likelihood of the model is -339.14 and the LL-Null is -558.30.

Checking the P value will show the statistically significant variables. For this statistical study, 95% confidence level is considered thus the variable with P value less than 0.05 is considered as statistically significant variables. The P value, Odd ratio and Coefficients for all the variables for the first iteration is as below. Statistically significant variables that explains the dependent variable `HAPPY` is highlighted in grey with bold fonts.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **coef** | **odds ratio** | **p-value** |
| IMMIGRANTS\_COUNTRY\_BETTER | 0.004309 | 1.004319 | 9.339150e-01 |
| SAYING\_IN\_POLITICS | -0.023057 | 0.977207 | 8.896554e-01 |
| DOMICILE | 0.024119 | 1.024413 | 7.994202e-01 |
| SATISFACTION\_GOVERNMENT | 0.020988 | 1.021210 | 7.213115e-01 |
| SCORE\_PEOPLE\_FAIR | 0.023931 | 1.024220 | 7.012229e-01 |
| EDUCATION\_YEARS | 0.015709 | 1.015833 | 6.081304e-01 |
| INTERNET\_USE | 0.054383 | 1.055889 | 5.554206e-01 |
| IS\_DISCRIMINATED\_GROUP | -0.324630 | 0.722795 | 5.420995e-01 |
| PEOPLE\_TO\_DISCUSS | -0.057877 | 0.943766 | 5.169519e-01 |
| TRUST\_POLITICIANS | -0.040736 | 0.960082 | 5.084245e-01 |
| SCORE\_ATTACHED\_COUNTRY | -0.039196 | 0.961562 | 4.973985e-01 |
| SATISFACTION\_ECONOMY | 0.048106 | 1.049281 | 4.738491e-01 |
| TRUST\_POLICE | -0.045593 | 0.955431 | 4.677887e-01 |
| IS\_FEMALE | 0.160438 | 1.174025 | 4.556584e-01 |
| ASHAMED\_LGBTQ | -0.071256 | 0.931223 | 4.503317e-01 |
| TRUST\_LEGAL\_SYSTEM | 0.046083 | 1.047161 | 4.252148e-01 |
| SCORE\_PEOPLE\_TRUST | 0.061420 | 1.063345 | 3.188539e-01 |
| HAD\_DIVORCED | -0.389104 | 0.677664 | 2.725609e-01 |
| IS\_STUDYING | -0.780440 | 0.458204 | 1.479092e-01 |
| IS\_WORKING | -0.562639 | 0.569704 | 1.392780e-01 |
| IS\_UNEMPLOYED | -0.742119 | 0.476104 | 1.071756e-01 |
| PEOPLE\_IN\_HOUSEHOLD | -0.207431 | 0.812669 | 8.831793e-02 |
| SCORE\_ATTACHED\_EUROPE | 0.084852 | 1.088556 | 7.105072e-02 |
| MEET\_FRIENDS | 0.134813 | 1.144323 | 6.861678e-02 |
| SOCIAL\_ACTIVITIES | 0.329878 | 1.390798 | 5.264214e-02 |
| **FUTURE\_PLANNING** | **-0.074668** | **0.928052** | **4.811126e-02** |
| **IS\_RETIRED** | **-0.744222** | **0.475104** | **3.960174e-02** |
| **SATISFACTION\_EDUCATION** | **0.145056** | **1.156105** | **2.001001e-02** |
| **SATISFACTION\_DEMOCRACY** | **-0.168479** | **0.844949** | **1.434662e-02** |
| **MARITAL\_STATUS** | **-0.136196** | **0.872672** | **1.185047e-02** |
| **HAD\_CHILDREN** | **-0.667284** | **0.513100** | **1.018432e-02** |
| **SCORE\_DISABLE** | **-0.575543** | **0.562399** | **7.560421e-03** |
| **SATISFACTION\_HEALTH** | **0.186807** | **1.205395** | **5.496007e-04** |
| **SCORE\_HEALTHY** | **-0.520303** | **0.594341** | **8.497144e-05** |
| **SATISFACTION\_LIFE** | **0.644713** | **1.905440** | **2.410042e-25** |

**2.4 Takeaway:**  The main takeaway from the first iteration of logit model is to filter the variables that are not statistically significant in explain the dependent variable. According to the model, variables such as future\_planning, is\_retired, satisfaction\_education, satisfaction\_democracy, marital\_status, had\_children, score\_disable, satisfaction\_health, score\_healthy and satisfaction\_life are the variables that are important for our next iteration.

**3. Second Iteration**

After selecting the statistically significant variables, second iteration of model fitting needs to be done.

**3.1 Data Cleaning:** In these new selected variables, HAD\_CHILDREN and MARITAL SATUS have more missing values. In this study, people not responding is also an information. Trying to include this information will help the model to include the hidden information.

HAD\_VARIABLE has two values 0 and 1. This changing all the nan and null values into 2 will help to include the hidden message. Since this is categorical variables, getting dummy after converting the nan values is important.

Similarly, for MARITAL\_STATUS variables, nan values are converted into 7 and dummies are created to include in the model.

After changing the HAD\_CHILDREN and MARITAL\_STATUS nan values, now the data has 2508 observations out of 2745 observations which is 91% of the original dataframe.

**3.2 Model fitting:** After the data cleaning step and dropping the 9% of observations with nan or null values, fitting the data in a logit model gives us function value of 0.27 where the model converges in 8 iterations.

**3.3 Model Summary:** After checking the model summary we can see that the second iteration model has converged with pseudo R2 of 0.568. The log-likelihood of the model is -685.04 and the LL-Null is -1065.0.

Checking the P value will show the statistically significant variables. Similar to first iteration, 95% confidence level is considered thus the variable with P value less than 0.05 is considered as statistically significant variables. The P value, Odd ratio and Coefficients for all the variables for the first iteration is as below. Statistically significant variables that explains the dependent variable `HAPPY` is highlighted in grey with bold fonts.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Coef** | **Odds ratio** | **P-value** |
| SATISFACTION\_DEMOCRACY | 0.008564 | 1.008601 | 8.209693e-01 |
| MARITAL\_STATUS\_2.0 | 1.074456 | 2.928400 | 3.841887e-01 |
| HAD\_CHILDREN\_na | -0.235163 | 0.790442 | 2.279702e-01 |
| IS\_RETIRED | -0.370168 | 0.690618 | 5.993975e-02 |
| **HAD\_CHILDREN\_1.0** | **-0.449466** | **0.637969** | **3.673933e-02** |
| **MARITAL\_STATUS\_3.0** | **-0.827056** | **0.437335** | **2.174503e-02** |
| **MARITAL\_STATUS\_4.0** | **-0.758086** | **0.468562** | **1.402654e-02** |
| **MARITAL\_STATUS\_na** | **-1.604593** | **0.200971** | **6.316612e-03** |
| **SATISFACTION\_EDUCATION** | **0.117368** | **1.124533** | **3.354124e-03** |
| **SCORE\_DISABLE** | **-0.292527** | **0.746375** | **3.184011e-03** |
| **MARITAL\_STATUS\_5.0** | **-0.704927** | **0.494145** | **2.053579e-03** |
| **MARITAL\_STATUS\_6.0** | **-0.700020** | **0.496575** | **9.145595e-04** |
| **SATISFACTION\_HEALTH** | **0.127069** | **1.135495** | **2.965260e-04** |
| **FUTURE\_PLANNING** | **-0.112044** | **0.894005** | **8.558145e-06** |
| **SCORE\_HEALTHY** | **-0.542688** | **0.581184** | **1.691019e-11** |
| **SATISFACTION\_LIFE** | **0.616667** | **1.852742** | **1.279910e-62** |

**3.4 Takeaways:** Variables such as SATISFACTION\_DEMOCRACY, MARITAL\_STATUS\_2.0 with respect to MARITAL\_STATUS\_1.0, HAD\_CHILDREN\_na with respect to HAD\_CHILDREN\_0.0 and IS\_RETIRED are not statistically significant in explaining binary dependent variable HAPPY.

**4. Third Iteration:**

After selecting the variables from second iteration, running a model with only selected variables gives us the following summary.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Coef** | **Odds ratio** | **P value** |
| MARITAL\_STATUS\_3.0 | -0.741600 | 0.476351 | 3.714771e-02 |
| MARITAL\_STATUS\_4.0 | -0.689248 | 0.501953 | 2.437690e-02 |
| HAD\_CHILDREN\_1.0 | -0.411913 | 0.662382 | 2.008467e-02 |
| MARITAL\_STATUS\_na | -1.553205 | 0.211569 | 7.481225e-03 |
| MARITAL\_STATUS\_6.0 | -0.541020 | 0.582154 | 2.995801e-03 |
| SCORE\_DISABLE | -0.297629 | 0.742577 | 1.743756e-03 |
| SATISFACTION\_EDUCATION | 0.122255 | 1.130042 | 1.184476e-03 |
| SATISFACTION\_HEALTH | 0.123855 | 1.131852 | 2.475387e-04 |
| MARITAL\_STATUS\_5.0 | -0.811860 | 0.444031 | 2.300177e-04 |
| FUTURE\_PLANNING | -0.113213 | 0.892961 | 6.406158e-06 |
| SCORE\_HEALTHY | -0.597613 | 0.550123 | 3.962052e-15 |
| SATISFACTION\_LIFE | 0.607351 | 1.835563 | 4.785683e-63 |

All the variables are statistically significant in explaining the binary dependent variable HAPPY.

The model can be defined as follows:

**4.1 Convergence:** Checking the convergence of the logit model gives the following results.

Table

Description automatically generated

The model converges after 7 iterations and the algorithm used is Newton-Raphson algorithm.

**4.2 Odds Ratio Estimates:** Odds Ratio for all the satistically significant variables.

Table

Description automatically generated

**MARITAL\_STATUS\_3.0:** The odds ratio for MARITAL\_STATUS\_3.0 (Legally separated) with respect to MARITAL\_STATUS\_1.0 (Legally married) is 0.476 with CI [0.237,0.957] which means if a person changes from Legally separated to Legally married, his odds of being happy is reduced by half (0.476).

**MARITAL\_STATUS\_4.0:** The odds ratio for MARITAL\_STATUS\_4.0 (Legally divorced/Civil union dissolved) with respect to MARITAL\_STATUS\_1.0 (Legally married) is 0.502 with CI [0.275,0.915] which means if a person changes from Legally divorced/Civil union dissolved to Legally married, his odds of being happy is reduced by half (0.502).

**MARITAL\_STATUS\_na:** The odds ratio for MARITAL\_STATUS\_na (not\_answered) with respect to MARITAL\_STATUS\_1.0 (Legally married) is 0.212 with CI [0.068,0.660] which means if a person does not answer the question changes to Legally married, his odds of being happy is multiplied by 0.212.

**SATISFACTION\_EDUCATION:** Satisfaction education changes from 0 to 10 where 0 being extremely bad and 10 being extremely good. The odds ratio for SATISFACTION\_EDUCATION is 1.130 with CI [1.050,1.217] which means an increase in one score of satisfaction education will increase the odd of being happy by 1.130 times.

**SCORE\_DISABLE:** Score disable is whether people are sick within last days. Value is 0 and 1 where 1 being sick or permanently disabled. The odds ratio for SCORE\_DISABLE is 0.743 with CI [0.616,0.895] which means if people or sick or permanently disabled, odd of being happy by 0.743 times than not being sick.

**MARITAL\_STATUS\_5.0:** The odds ratio for MARITAL\_STATUS\_5.0 (widowed / Civil partner died) with respect to MARITAL\_STATUS\_1.0 (Legally married) is 0.444 with CI [0.288,0.684] which means if a person changes from widowed / Civil partner died to Legally married, his odds of being happy is reduced by half (0.444).

**MARITAL\_STATUS\_6.0:** The odds ratio for MARITAL\_STATUS\_6.0 (Never married or in legally registered civil partnership) with respect to MARITAL\_STATUS\_1.0 (Legally married) is 0.582 with CI [0.407,0.832] which means if a person changes from Never married or in legally registered civil partnership to Legally married, his odds of being happy is reduced by half (0.582).

**SATISFACTION\_HEALTH:** Satisfaction health changes from 0 to 10 where 0 being extremely bad and 10 being extremely good about the state’s health system. The odds ratio for SATISFACTION\_HEALTH is 1.132 with CI [1.059,1.209] which means an increase in one score of satisfaction education will increase the odd of being happy by 1.132 times.

**FUTURE\_PLANNING:** Planning future changes from 0 to 10 where 0 being plan for my future as much as possible and 10 being just take each day as it comes. The odds ratio for FUTURE\_PLANNING is 0.893 with CI [0.850,0.938] which means an increase in one score of future planning will affect the odd of being happy by 0.893 times which means the less you plan for future, you will be less happy comparitively.

**SCORE\_HEALTHY:** Score healthy is general health. Value ranges from 1 to 5 where 1 being very good and 5 being very bad. The odds ratio for SCORE\_HEALTHY is 0.550 with CI [0.474,0.639] which means score healthy increases by 1, the odds is almost reduced by half (0.550) which means if people are being less healthy, the odds of being happy decreases.

**SATISFACTION\_LIFE:** Satisfaction life changes from 0 to 10 where 0 being extremely dissatisfied and 10 being extremely satisfied about one’s life. The odds ratio for SATISFACTION\_LIFE is 1.836 with CI [1.710,1.971] which means an increase in one score of satisfaction life will increase the odd of being happy by 1.836 times.

**4.3 MARGINAL EFFECT:** The marginal effect at mean for the final logit model is as follows: Table

Description automatically generated

HAD\_CHILDREN\_1.0: In average, the difference between the probability to know a person is happy when having children to not having children is -0.0279.

MARITAL\_STATUS\_3.0: In average, the difference between the probability to know a person is happy when they are legally separated to that of legally married is -0.0502.

MARITAL\_STATUS\_4.0: In average, the difference between the probability to know a person is happy when they are legally divorced/civil union dissolved to that of legally married is -0.0467.

MARITAL\_STATUS\_5.0: In average, the difference between the probability to know a person is happy when they are widowed / civil partner died to that of legally married is -0.0550.

MARITAL\_STATUS\_6.0: In average, the difference between the probability to know a person is happy when they are never married or in legally registered civil partnership to that of legally married is -0.0366.

MARITAL\_STATUS\_na: In average, the difference between the probability to know a person is happy when they are not answered to that of legally married is -0.1051.

SATISFACTION\_EDUCATION: In average, if the satisfaction education increases by 1, the probability of being happy increases by 0.0083

SCORE\_DISABLE: In average, the difference between the probability to know a person is happy when they are sick in last 7 days or permanently disabled to that of legally married is -0.0201.

SATISFACTION\_HEALTH: In average, if the satisfaction education increases by 1, the probability of being happy increases by 0.0084

FUTURE\_PLANNING: In average, if the satisfaction education increases by 1, the probability of being happy reduces by 0.0077 (-0.0077)

SCORE\_HEALTHY: In average, if the satisfaction education increases by 1, the probability of being happy reduces by 0.0405 (-0.0405)

SATISFACTION\_LIFE: In average, if the satisfaction education increases by 1, the probability of being happy increases by 0.0411

**4.4 Testing one or multiple parameters:** To test one or multiple parameters about the statistical significance, Likelihood ratio test and Wald test is used. The results are as follows.

Table

Description automatically generated

The p value is 0 for both the tests. The model is globally significant (there is at least one βj different from 0. Our model with selected variables is better than the one with only the intercept.

**4.5 Quality of statistical model:** The quality of the statistical model can be verified using the AIC and BIC values which are as follows.

Table

Description automatically generated

From above results, all three criterion favors model with intercept and covariates than intercept only model since the values are less for intercept and covariates model than the intercept only model.

**4.6 Goodness of fit:** The association pairs calculation will give the measure goodness of fit. The results for Somers’ D (Gini coefficient), Gamma (Goodman-Kruskal Gamma), Kendall’s Tau-a and C are as follows.

If the value of Somers’ D (Gini coefficient), Gamma (Goodman-Kruskal Gamma) and C are closer to 1, then the model is a good fit. If it is close to 0, then we need to add covariates to improve the fitness. For Kendall’s Tau-a, the value changes from -1 to +1 where closer to 1 is a good fit.

Table

Description automatically generated

From above results we can see that the D ≈ 0.8, Γ ≈ 0.8 , τa ≈ 0.2, c ≈ 0.9. The model is a good fit.