The Role of Misinformation in Health Discussions: Analyzing Social Media's Impact

I. Abstract

In the era of easy-to-get information floating around, lots of unverified information is being sharing from friends to friends via social media.

II. Introduction

This paper reports the relationship between false or misleading health information can affect people to make decisions from the social media.

III. Methodology

Data Description

The dataset is from HINTS(Health Information National Trend Survey), 2022 survey. Total responds are 6,252 and complete responds are 6,185 and other 67 responds are partially responded. We converted the given R Data(.rda) into CSV file(.csv) for handling with programming on python.

Programming Environment

- Environment: WSL2, Ubuntu 20.04
- Programming Language: Python, version 3.11.8

Variable Descriptions

Table 1

Table 1: Variable Descriptions

Variable	Survey Question	Scale	Description
‘SocMed_MakeDecisions’ (Dependent Variable)	B14 a. I use information from social media to make decisions about my health	Like (Strongly disagree to strongly agree)	
‘MisleadingHealthInfo’ (Main Independent Variable)	B13. How much of the health information that you see on social media do you think is false or misleading?	Like	
‘Age’ (combined as PC)	R1. What is your age	Ratio	
‘IncomeRanges’ (combined as PC)	R13. Thinking about members of your family living in this household, what is your combined annual income, meaning the total pre-tax income from all sources earned in the past year?	Ordinal	
‘MaritalStatus’			

‘BirthGender’			
‘SocMed_DiscussHCP’			
‘SmokeNow’			
‘TimesModerateExercise’			
‘SocMed_SharedPers’			
‘SocMed_SharedGen’			
‘SocMed_Interacted’			
‘SocMed_WatchedVid’			

Note. SocMed stands for Social Media, HCP stands for Health Care Provider.

Table 2

Table 2: Original & Recorded Response Counts

Variable	Original Response	Recorded Response	Original Counts	Recorded Counts
MisleadingHealthInfo	I do not use social media	-1	1211	0
	A little	0	855	855
	Some	1	2256	2256
	A lot	2	1740	1740
BirthGender	Male	0	2307	2051
	Female	1	3535	2800
MaritalStatus	Single, never been married	0	1119	910
	Separated	0	136	102
	Widowed	0	646	378
	Divorced	0	939	673
	Living as married or living with a romantic partner	0	373	346
	Married	1	2624	2119
IncomeFeelings	Finding it very difficult on present income	0	346	277
	Finding it difficult on present income	1	763	605
	Getting by on present income	2	2140	1644
	Living comfortably on present income	3	2518	1933

Note. Invalid data such as missing data, incomplete data, multiple responses selected data and data with technical issues was excluded from the table above.

Table 3

Table 3: Correleation Matrix (Model 1)

	PC	MS	BG	Dis_HCP	SMK	Time_Exc
PC	1.000	-0.181	0.073	0.033	0.087	-0.131
MS		1.000	-0.130	0.002	-0.072	-0.002
BG			1.000	0.029	-0.028	-0.061
Dis_HCP				1.000	0.001	-0.030
SMK					1.000	-0.013
Time_Exc						1.000

Note. PC= Age_Income_PC1, MS=MaritalStatus, BG=BirthGender, Dis_HCP= SocMed_DiscussHCP, SMK=SmokeNow, Time_Exc= TimesModerateExercise

Table 4

Table 4: Correleation Matrix (Model 2)

	PC	MS	BG	Dis_HCP	SMK	Time_Exc	ShrPer	ShrGen	INT	VID
PC	1.000	-0.181	0.073	0.033	0.087	-0.131	0.116	0.072	0.146	0.175
MS		1.000	-0.130	0.002	-0.072	-0.002	-0.030	0.009	-0.023	-0.005

BG			1.000	0.029	-0.028	-0.061	0.014	0.015	0.090	0.021
Dis_HCP				1.000	0.001	-0.030	0.221	0.251	0.256	0.288
SMK					1.000	-0.013	0.038	0.014	0.005	-0.021
Time_Exc						1.000	-0.003	-0.034	-0.027	-0.019
ShrPer							1.000	0.496	0.458	0.258
ShrGen								1.000	0.493	0.365
INT									1.000	0.384
VID										1.000

Note. PC= Age_Income_PC1, MS=MaritalStatus, BG=BirthGender, Dis_HCP= SocMed_DiscussHCP, SMK=SmokeNow, Time_Exc= TimesModerateExercise, ShrPer=SocMed_SharedPers, ShrGen=SocMed_SharedGen, INT=SocMed_Interacted, VID=SocMed_WatchedVid

Table 5

Table 5: Table 4: Correleation Matrix (Model 3)

	PC	MS	BG	Dis_HCP	SMK	Time_Exc	ShrPer	ShrGen	INT	VID	MLHI
PC	1.000	0.181	-0.073	-0.033	-0.087	0.131	-0.116	-0.072	-0.146	-0.175	0.032
MS		1.000	-0.130	0.002	-0.072	-0.002	-0.030	0.009	-0.023	-0.005	0.019
BG			1.000	0.029	-0.028	-0.061	0.014	0.015	0.090	0.021	-0.008

Dis_HCP				1.000	0.001	-0.030	0.221	0.251	0.256	0.288	-0.208
SMK					1.000	-0.013	0.038	0.014	0.005	-0.021	0.013
Time_Exc						1.000	-0.003	-0.034	-0.027	-0.019	0.058
ShrPer							1.000	0.496	0.458	0.258	-0.119
ShrGen								1.000	0.493	0.365	-0.111
INT									1.000	0.384	-0.096
VID										1.000	-0.173
MLHI											1.000

Note. PC= Age_Income_PC1, MS=MaritalStatus, BG=BirthGender, Dis_HCP= SocMed_DiscussHCP, SMK=SmokeNow, Time_Exc= TimesModerateExercise, ShrPer=SocMed_SharedPers, ShrGen=SocMed_SharedGen, INT=SocMed_Interacted, VID=SocMed_WatchedVid, MLHI=MisleadingHealthInfo

Preprocessing

All the variables are numerated for proceeding regression. Responses are converted in numbers zero to (number of valid answers – 1) except the variable named 'MisleadingHealthInfo'. In the middle of converting process, response counts are all checked using 'value_counts()' function to validate the process.

Dependent variable for this research consists of answers from whom not responded “I do not use social media” from survey question B13, variable name ‘MisleadingHealthInfo’. So the response ‘I do not use social media’ was converted into negative number(-1), and changed into NaN(Not a Number) to drop invalid value at once in later step. To secure the completeness of data after dropping the responses, dropping this response proceeded at first. In the same manner, after checking all the valid responses are turned into integers, we dropped the rows with NaN values and all the string values which includes missing data and other invalid data. The codes for typecasting to integer data type to make sure the regression step to be conducted with no errors.

Since the two variables, Age and IncomeFeelings, exhibit high multicollinearity, they were combined into a single component using Principal Component Analysis (PCA). First, the variables were standardized using a StandardScaler, and then PCA was applied to extract one principal component, which is now represented as Age_Income_PC1.

Analysis Methodology

We coded with Python programming language, and converted R Data file into CSV file as noted. Data frame from Pandas stored the CSV file data as a data frame data type. Structure for this research is comparing three models, in terms of performance in R-squared measurements. To see the results at once, iterations are used and the results for each model are saved and print the results after the iteration step.

To dependent variable is fixed, independent variables are added for each step. Model 1 has 6 independent variables(one is combined with two variables) and model 2 has five more independent variables from the same survey question B12, all sub-questions are asking interaction with social media. In the middle, the variable named ‘SocMed_Visited’ was excluded due to the high multicollinearity between other variables. Hence Model 2 has 10 independent variables. Model 3 has one more independent variable, which is our main variable to see the effect on the dependent variable. For each iteration, linear regression model is created and learned for independent variables of model 1, model 2, and model 3 respectively. Performance was measured as R-square score and MSE(Mean Squared Error). Checking VIFs and cross validation process were to check reliability of the models. For the deeper assessment on this model, results from Random Forest Model, and following F1-score was recorded too. After the iteration, model 3 solely used for plotting Residual Histogram, and TensorFlow Model. The accuracy of TensorFlow Model was referenced, and visualized plot from the result of TensorFlow Model for overfitting was also utilized as well as residual histogram.

Models Used

- Linear Regression Model
  - Imported LinearRegression from sklearn.linear_model
  - Test Size: 0.2
  - Random_state: 42
  - Inputs: Independent variables of model 1, model 2, and model 3 respectively, and a dependent variable
  - Outputs: R-squared score(higher the better), MSE score(lower the better)
- VIF
  - Imported variance_inflation_factor from statsmodels.stats.outliers_influence
  - No Parameters
  - Inputs: Independent variables of model 1, model 2, and model 3 respectively, and a dependent variable

    - Outputs: VIFs for each dependent variable
- Random Forest Model
    - Imported RandomForestClassifier from sklearn.ensemble
    - N_estimators: 200
    - Class_weight: Balanced
    - Max_depth: 10
    - Min_samples_split: 4
    - Min_sample_leaf: 2
    - Random_state: 42
    - Inputs: Independent variables of model 1, model 2, and model 3 respectively, and a dependent variable
    - Outputs: Accuracy of prediction
- F1-score
    - Imported f1_score from from sklearn.metrics
    - Results from Random Forest Model was used
    - Average: Micro
    - Inputs: Original and predicted results of Random Forest Model from dependent variable
    - Outputs: Balance between precision and recall(closer to 1, the better)
- TensorFlow Model
    - Scaler: Standard Scaler
    - Model Architecture
        - Layer 1: Dense(32, activation=' relu', L2 Regularized(0.01)
        - Layer 2: Dropout(0.3)
        - Layer 3: Dense(4, activation=‘softmax’)

- Optimizer: Adam (learning rate = 0.05)
- Loss Function: categorical_crossentropy
- Metrics: Accuracy
- Early Stopping
  - Monitor: val_loss
  - Patience: 5
  - Restore Best Weight: True
- Model Hyperparameters
  - Epochs: 20
  - Batch Size: 32
  - Callbacks: early_stopping
- Inputs: Independent variables of model 1, model 2, and model 3 respectively, and a dependent variable
- Outputs: Test Accuracy

IV. Results

Table 5

Table 6: Model Performance Results

Models	Model 1	Model 2	Model 3
R-Squared	0.352362 (-)	0.377237 (5.04%)	0.409439 (8.54%)
MSE	0.401745 (-)	0.386314 (3.84%)	0.366339 (5.17%)

Test results showed improvement in terms of performance on R-squared score and a reduction in errors(MSE) as step goes further. But the low absolute value of the results have potential improvements.

R-squared is also known as Coefficient of Determination, an indicator that shows how the model explains the volatility of dependent variables. The value lies between 0 and 1, the model explains well if the value is closer to 1.

MSE, Mean Squared Error quantifies the prediction failures by getting the results from the difference between the actual and predicted value. The difference is squared to prevent cancellation of errors in the positive and negative directions, and then averaged. A lower MSE indicates better prediction accuracy by the model.

Table 7

Table 7: Other Model Performance Results

Models	Model 1	Model 2	Model 3
F1-score	0.655118 (-)	0.662992 (1.20%)	0.689764 (4.04%)
Random Forest Accuracy	0.655118 (-)	0.662992 (1.20%)	0.689764 (4.04%)
TensorFlow Accuracy*	0.66882 (-)	0.66850 (-0.05%)	0.68488 (2.39%)

Note. The result of TensorFlow model is averaged by 10 times due to its inconsistency.

This model has approximately over 65% of accuracy on prediction, and tendency of improvement over models as well. TensorFlow model was tested for 10 times because of its fluctuation on results(refer to Appendix to see the 10 tests result). In average, model 2 has better performance than other models but the increment is not that significant.

F1-score stands for balance between precision and recall. Precision indicates the ratio of true positives(predicted and actual was true) from model predicted positives. Recall indicates the ratio of true positives from true positives and false negatives which means what was the probability if the model predicted positive. It is better to have value close to 1.

Random Forest is an ensemble learning method based on multiple decision trees. It generates several trees, and each tree makes a prediction. Afterward, the final prediction is obtained by taking the average of the predictions (for regression problems) or the majority vote (for classification problems). Accuracy stands for ratio that model predicted correctly from all the samples.

OLS Regression Results

Table 8

Table 8: Regression Results(Model 1)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.2202***	0.039	5.65	0.000
Age_Income_PC1	-0.0742***	0.018	-4.016	0.000
MaritalStatus	-0.0166	0.026	-0.627	0.531
BrithGender	-0.074	0.027	-0.280	0.779
SocMed_DisscussHCP	0.5079***	0.015	33.159	0.000
SmokeNow	0.0099	0.071	0.140	0.889
TimesModerateExercise	0.0027	0.007	0.370	0.712

Note. *p < .05 *, p < .01**, p <.001 ****

Table 8

Table 9: Regression Results(Model 2)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.1029**	0.040	2.602	0.009
Age_Income_PC1	-0.0349	0.018	-1.900	0.058
MaritalStatus	-0.0257	0.026	-0.995	0.320
BrithGender	-0.0120	0.026	-0.464	0.643
SocMed_DisscussHCP	0.4412***	0.016	27.631	0.000
SmokeNow	0.0318	0.069	0.460	0.645
TimesModerateExercise	0.0017	0.007	0.240	0.810
SocMed_SharedPers	0.0454*	0.022	2.080	0.038
SocMed_SharedGen	0.0592**	0.019	3.093	0.002
SocMed_Interacted	0.0233	0.019	1.214	0.225
SocMed_WatchedVid	0.1027***	0.014	7.557	0.000

Note. *p < .05 *, p < .01**, p <.001 ****

Table 9

Table 10: Regression Results(Model 3)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.3002***	0.046	6.560	0.000
Age_Income_PC1	-0.0357*	0.018	-1.971	0.049
MaritalStatus	-0.0223	0.025	-0.876	0.381
BrithGender	-0.0153	0.026	-0.597	0.551
SocMed_DisscussHCP	0.4198***	0.016	26.280	0.000
SmokeNow	0.0405	0.068	0.594	0.552
TimesModerateExercise	0.0053	0.007	0.756	0.450
SocMed_SharedPers	0.0346	0.022	1.603	0.109
SocMed_SharedGen	0.0576**	0.019	3.049	0.002
SocMed_Interacted	0.0285	0.019	1.503	0.133
SocMed_WatchedVid	0.0903***	0.013	6.694	0.000

MisleadingHealthInfo	0.1506***	0.018	-8.253	0.000

Note. $p < .05$ *, $p < .01$**, $p <.001$ ***

Table11

Table 11: Regression Coefficients for Models 1, 2 and 3

Variables	Model 1	Model 2	Model 3
const	0.2202***	0.1029**	0.3002***
Age_Income_PC1	-0.0742***	-0.0349	-0.0357*
MaritalStatus	-0.0166	-0.0257	-0.0223
BrithGender	-0.074	-0.0120	-0.0153
SocMed_DisscussHCP	0.5079***	0.4412***	0.4198***
SmokeNow	0.0099	0.0318	0.0405
TimesModerateExercise	0.0027	0.0017	0.0053
SocMed_SharedPers		0.0454*	0.0346
SocMed_SharedGen		0.0592**	0.0576**
SocMed_Interacted		0.0233	0.0285
SocMed_WatchedVid		0.1027***	0.0903***
MisleadingHealthInfo			0.1506***

Note. $p < .05$ *, $p < .01$**, $p <.001$ ***

Validation and Reliability

Table 12

Table 12: VIF Results

Varaibles	Model 1	Model 2	Model 3
Age_Income_PC1	1.053364	1.114430	1.118799
MaritalStatus	1.719223	1.762562	1.852129
BirthGender	1.977018	2.028820	2.148691
SocMed_DiscussHCP	1.475907	1.757955	1.767464
SmokeNow	1.038353	1.039084	1.044004
TimesModerateExercise	2.566099	2.825734	3.756909
SocMed_SharedPers		1.715814	1.722602
SocMed_SharedGen		2.245321	2.248258

SocMed_Interacted		1.971016	1.971022
SocMed_WatchedVid		2.728945	2.731656
MisleadingHealthInfo			3.066129

VIF, Variance Inflation Factor, indicates multicollinearity for each independent variable. Correlated variables have high VIF value which can distort the result. Two variables, Age and IncomeFeelings, are combined into one variable as PC(Principle Component) since they had value over 7.

Table 13

Table 13: Cross Validation Check Results

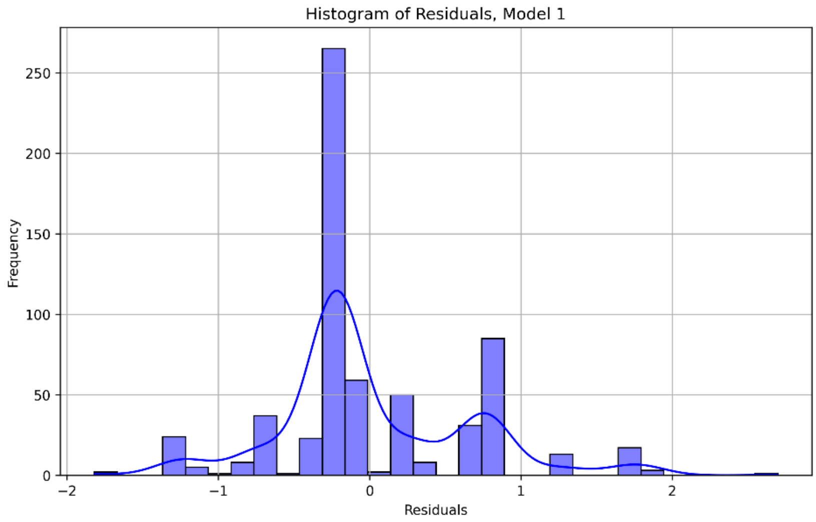
	Model 1	Model 2	Model 3
Standard Deviation of CV scores	0.03451	0.030934	0.030421

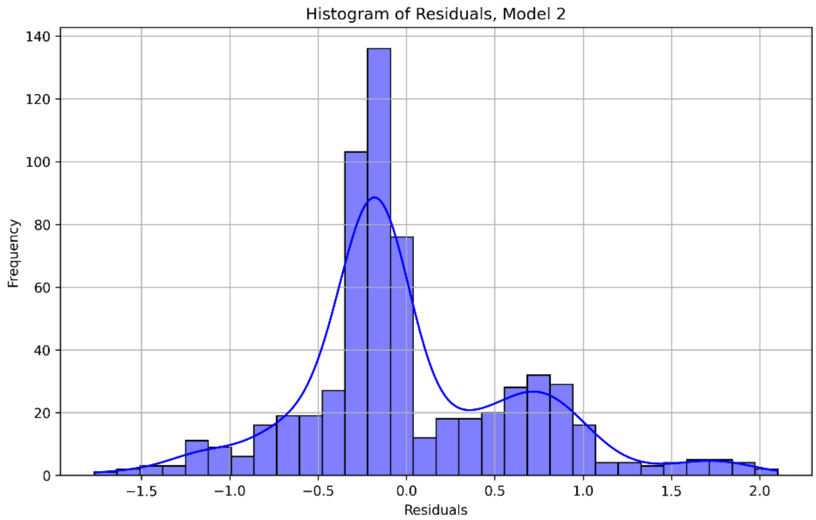
Reliability was also check through cross validation test. The test were conducted five times and reported quite low difference within a trial which can guarantee reliability of this model in some degree. CV scores are from MSE(Mean Squared Error) of the linear regression model.

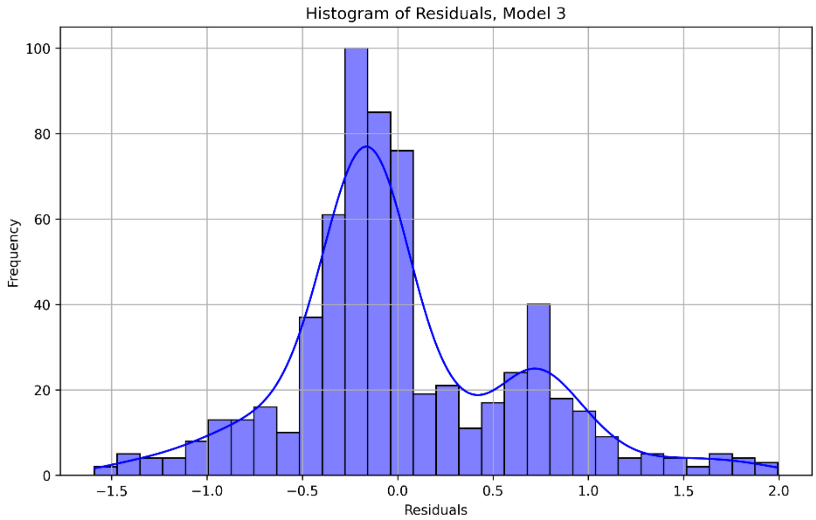
Residual Histogram

Figure 1

Figure 1: Residual Histograms of Linear Regression Model





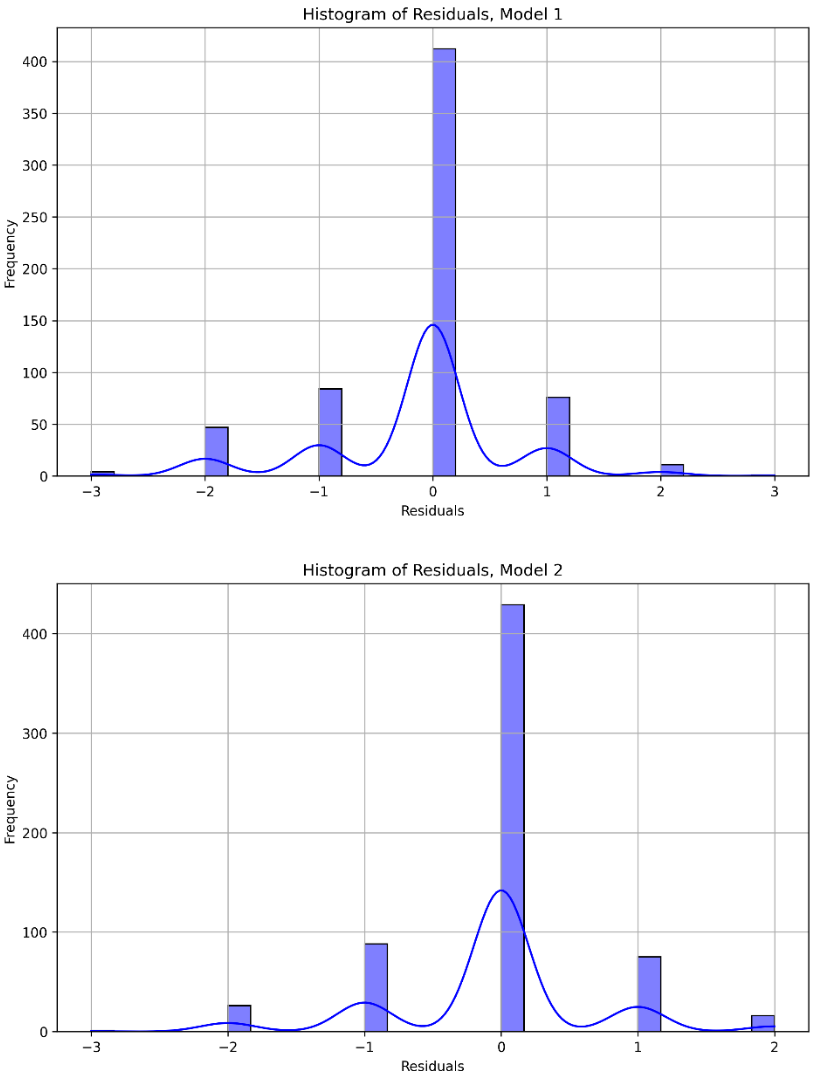


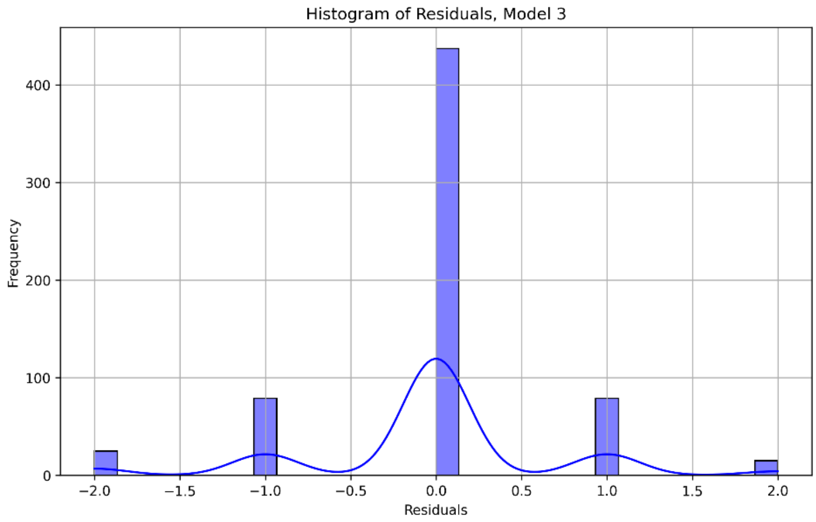
This plot shows residuals in our regression model. It means distribution of the differences between the observed values and predicted values. The residuals are not skewed and seems to follow normal distribution to have bell-shaped curve. This histograms are resulted from linear regression model. The frequency goes lower over the models as illustrated scale of the Y axis got lowered for each step.

Actual vs Predicted Plot

Figure 2

Figure 2: Residual Histograms of Random Forest Model



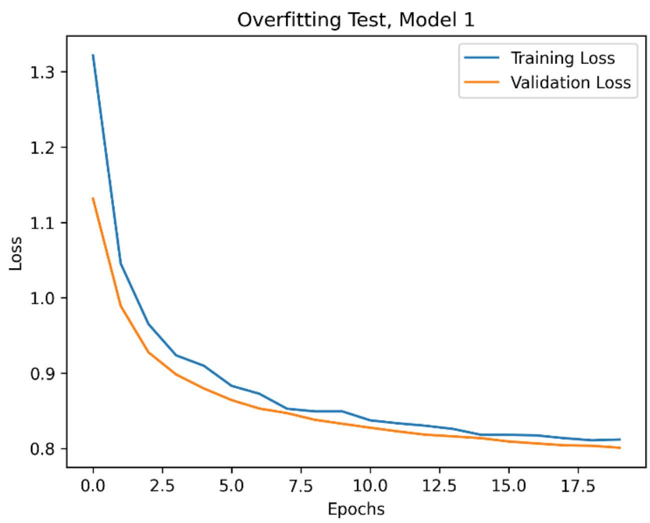


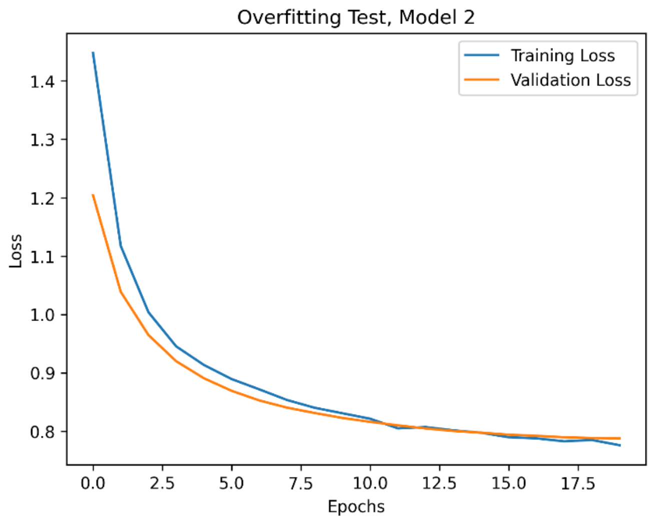
Random Forest Model has discrete outputs, so the graph looks slightly different from the graphs of linear regression model. However, we can say it follows normal distribution.

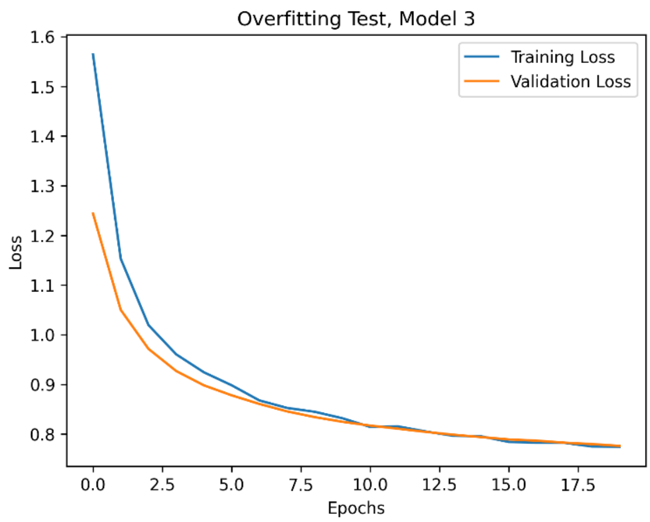
Overfitting Test Results Plot

Figure 3

Figure 3: Overfitting Test Results







This figure illustrates whether the model is overfitted. The results are derived from a TensorFlow model. After the regularization, the curve became smoother and the difference between the graphs are quite little. With the fact that both graphs illustrate decreasing losses and two graphs are almost aligned, this model has less possibilities of overfitting problem.

V. Discussion

CONTENT

VI. Conclusion

CONTENT

VII. Appendix

Appendix A

Data Set Used for This Research

Data set used for this research can be downloaded here. Sign-in is needed. You can use your email address to enter the download page. Or, you can directly download the zip file by clicking here. The data set is R data and supporting documents (ZIP, 16.7 MB) from HINTS 6 (2022) dataset, updated May 2024.

Appendix B

Python Codes

The programming code for this research is uploaded here. Modeling was done with the file 'main.py'. Response counts were executed using the file 'response_counts.py' to store the counts in CSV file.

Appendix C

Tables and Figures

Appendix D

TensorFlow Model Results

	Model 1	Model 2	Model 3
test 1	0.6724	0.6661	0.6835
test 2	0.6724	0.6709	0.685
test 3	0.6693	0.6693	0.6835
test 4	0.6677	0.6693	0.6913
test 5	0.6709	0.6756	0.6866
test 6	0.663	0.6677	0.6866
test 7	0.6646	0.6693	0.685
test 8	0.6677	0.6598	0.6819
test 9	0.6709	0.6693	0.6756
test 10	0.6693	0.6677	0.6898