Predicting Self-enhancing Humour Scores by Machine Learning

Module Name: COMP2261
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Self-enhancing is a benign type of humour derived from the Humor Styles Questionnaire(HSQ). In order to find a model to predict an individual's self-enhancing degree based on two detrimental humour types and their age and gender, this study displays the detailed procedure of creating the best machine learning model for the mission including data preprocessing, algorithm selection, model training, evaluation and comparison. The Ridge algorithm is selected for the final model. After model training and evaluation, this study suggests that the correlations between self-enhancing humour scores and the chosen features are unpredictable. However, applying machine learning still paves new paths for psychological research.

Index Terms—humor styles questionnaire, hyperparameter tuning, machine learning, regression

1 Introduction

With enough data and specific valid correlations between features of the data, we can apply machine learning to predict one of the features by other features. According to the research on Development of the Humor Styles Questionnaire (HSQ)[1], the usage of humour can be divided into four different dimensions - affiliative, aggressive, self-enhancing and self-defeating, in which affiliative and self-enhancing are recognised as benign types of humour. In contrast, the other two types of humour are considered detrimental to people's well-being. This study mainly focuses on predicting the scores of self-enhancing humour by the scores of negative scales of humour as well as the value of age and gender applying machine learning. Instead of analysing the final prediction, this study mainly focuses on displaying the procedure of forming the model. Modeling self-enhancing humour type using the aforementioned features offers better understanding on whether an individual is likely to use self-enhancing humour to cope with the incongruities of life based on their age, gender and frequency of using negative types of humour.

2 Methodology

This section presents the process of machine learning, which includes data preprocessing, algorithm selection, model training and evaluation.

2.1 Data Preprocessing

The dataset was collected using an interactive online version of the Humor Styles Questionnaire from [1]. Three types of data in the dataset need cleaning, which are:

- invalid data
- imbalanced data
- reversed data

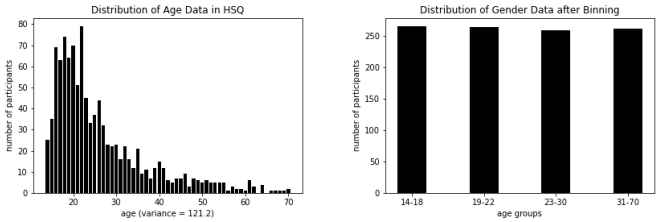
Invalid data includes **massive values in the age column**, which are unrealistic for a human's life expectancy [2]. We can easily find out and delete the rows consisting of invalid ages by filtering out rows that have an invalid number in age. Apart from that, invalid data can also be seen in the **gender column with input 0**, the same method is applied to target the invalid rows by filtering those rows out and deleting them. Another type of invalid data is the **empty inputs by the participants** resulting in -1 in question fields. A '-1' input means the participant did not answer this question, which will affect the final result of the score of the type of humour affected by the question giving that the score of the question is the mean value of the related questions. Deletion of those rows with -1 inputs is not ideal because it will cause a loss of a great number of rows for modeling. So after filtering rows that have more than 1/3 of empty inputs (see table 1), and deleting them, replacement of all the rest of the -1 inputs with the mean value of each question from all its valid inputs is performed.

TABLE 1
Participants with more than 1/3 of empty inputs

Q1	...	Q32	empty_count
2.0	...	3.0	13
1.0	...	NaN	19
1.0	...	NaN	15
NaN	...	NaN	32

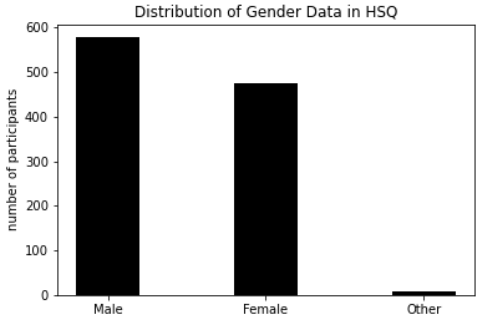
Imbalanced data can be seen in age and gender. **Age data** has a right skewed distribution (table 2, left), in which case, the tail region may act as an outlier for the statistical model and is not good for the model's performance. So data quantile binning is implemented to form a better distribution of the data (table 2, right). Applying data binning is reasonable because we only want to see whether age impacts the tendency to use self-enhancing humour; after data binning, the age groups are still ordered. In terms of

TABLE 2
Age Data Preprocessing



gender data(table 3), the amount of participants who select 'Other' in gender is too little. I can't apply the data binning method because the gender data is not continuous. So I use the down-sampling method by deleting all rows which selected 'Other' in the gender column.

TABLE 3
Gender Data



According to the HSQ research paper[1], some of the questions have **reversed values**, which are not accounted for according to the codebook of the HSQ dataset. Only reversed values related to affiliative humour have been handled. So reversing the data of other questions with reversed values is necessary and will lead to the better accuracy of the prediction.

The main characteristic of the dataset is a low correlation between data. As can be seen from the scatter plots (table 4) below, there isn't any distinctive correlation between different humour types. The mean value of self-enhancing between females and males is almost the same.

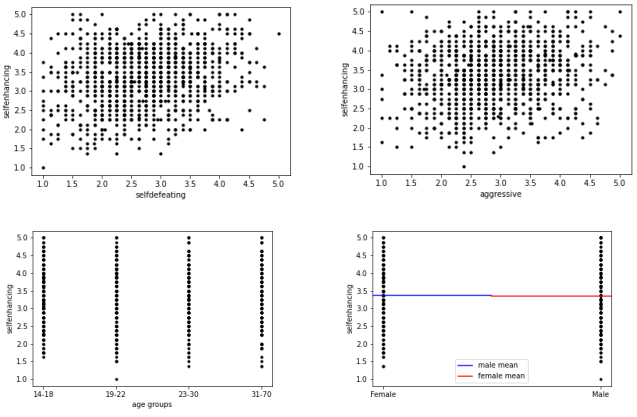
2.2 Algorithm Selection

Considering it is a supervised machine learning task, regression algorithms are chosen for model training because the target output of the model is the self-enhancing score, which is a continuous quantity. Therefore, I chose LinearRegression, Lasso and Ridge from all regression algorithms.

LinearRegression is chosen because it is the most basic regression model in the regression family, it is a stable algorithm.

In order to reduce model complexity and prevent overfitting, which may result from LinearRegression, Lasso regression and Ridge regression are selected. Ridge shrinks the coefficients and it helps to reduce the model complexity

TABLE 4
Correlation Between Data



and multi-collinearity. Lasso also shrinks the regression coefficients, but it often results in some parameters being zero, which can be used for feature selection.

2.3 Model Training

After splitting the data into training set and test set, a pipeline is created to compare each model with different hyperparameters and find the best combination for the model using GridSearchCV. Compared to RandomSearchCV, GridSearchCV is more computationally expensive but considering the dataset is not too large for training, the trade-off for optimal combination of parameters supplied is worthy.

For each loop in the pipeline, we first rescale the features and centralise data with unit variance using StandardScalar in case the variance of data is too big, allowing the preprocessed data to conform to a standard normal distribution.

Then we set the value of hyperparameters to pick from. We set α as a hyperparameter for regularized regressions like Lasso and Ridge. Apart from that, to overcome underfitting, I set different polynomial degrees as hyperparameters to see whether the model performs better by increasing its complexity. I set up to degree five because there are five features in the model. As for α, I set 0.01,0.1,1,10,100,1000 at first to see which fits the best and gradually narrowed down the range.

Before model training, considering the dataset for training is small, I used Repeated K-fold cross-validation to tackle data over-fitting. It will give a more precise estimate of performance without being too computationally expensive.

Luckily, the built-in ConvergenceWarnings from sklearn will help us detect convergence problems, and there are not any.

2.4 Model Evaluation

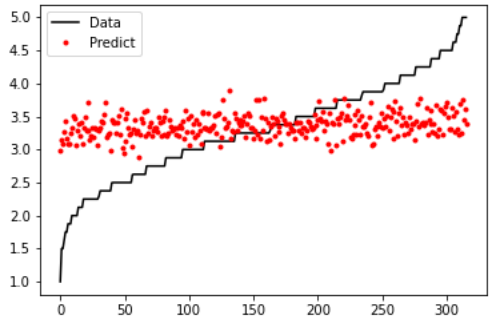
After having the best estimator from GridSearchCV for the model, mean squared error (MSE), mean absolute scaled error (MASE) and R squared are applied as evaluation metrics. Using MSE and MASE we can check how well the model performs compared to the actual data. MASE is used apart from MSE so that we can have more valid validation in case the model has a lot of outliers. By using R squared,

we can see whether the model correctly represents enough of the data variance. After implementing three metrics for evaluation, we can get these results:

- MSE = 0.49
- MASE = 0.57
- R^2 = 0.0793

All three evaluation metrics show that the model's performance is not ideal. Table 5 is the visualisation for the test data, it can be clearly seen that the predicted values concentrated between 2.5 to 4 such that a lot of to-be-predicted values outside this range are lost.

TABLE 5
Prediction for Test Data after Model Training



2.5 Model Comparison

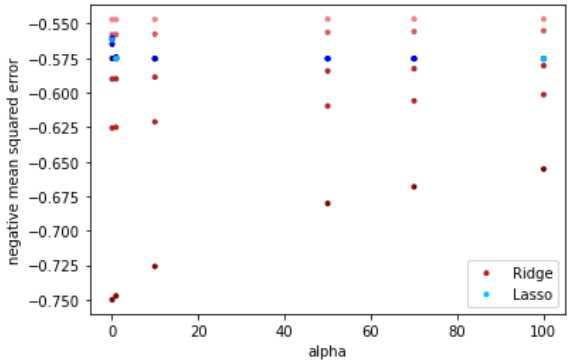
Using GridSearchCV enables us to easily compare each model with different hyperparameters by its attribute 'cv_results_', table 6 shows the ranking of each combination of an algorithm with different polynomial degrees and alphas. We can see that Ridge has taken up a lot of the top spaces in the ranking. Table 7 is the visualisation of how hyperparameters affect the performance of the model. The dots with shades of blue represent Lasso-based models, and the dots with shades of red represent Ridge-based models. The lower the polynomial degree of a model is, the lighter the colour of the dots on the graph is. For example, the dots on the top of the graph with light pink represent Ridge-based models with degree 1 and those models with degree 1 perform better than the models with other degrees. As for α, it's unclear whether it has a positive correlation with MSE when the polynomial degree of Ridge is 1 or 2. However, when the degree gets larger, the larger α is, the better the model performs.

In terms of Lasso, we can't see any correlations between different hyperparameters and the performance of the model. It is reasonable that some of the parameters disappear in Lasso because with Lasso, the model will be penalized for the sum of absolute values of the weights. So some unimportant coefficients of features will be shrunk to zero.

TABLE 6
Model Comparison

Rank	Algorithm	MSE	R^2	α	Degree
1	Ridge	0.546675	0.038751	70	1
2	Ridge	0.546708	0.038640	50	1
3	Ridge	0.546711	0.038763	100	1
4	Ridge	0.546952	0.038092	10	1
5	LinearRegression	0.547046	0.037895	1	1
...	...	...	...	...	...
65	LinearRegression	0.754426	-0.326316	NaN	5

TABLE 7
Correlations of Hyperparameter and Model Performance



3 Conclusion and Discussion

In conclusion, this study is predicting the scores of self-enhancing humour based on the scores of self-defeating humour, aggressive humour, the value of age and the value of gender by Ridge-based models. Unfortunately, the final prediction is not ideal because the correlation between different features is very small. The major limitation of the approach I used for achieving my goal is that I only consider three regression algorithms, which is not enough. Apart from that, I would need a better dataset for prediction, which will have the characteristic of having more data and a better correlation between features. From this study, I recognise the convenience and power of machine learning for research.

References

[1] Martin, R., *Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire.*, Journal of Research in Personality, 37, pp. 48-75, 2003.
[2] *GRG World Supercentenarian Ranking List*, Gerontology Research Group, https://grg.org/WSRL/TableE.aspx, Accessed 14 January 2022