Machine Learning-aided Demographic-Humour type analysis

Name: Iyngkarran Kumar
Banner ID: *qclx31*
Module Name: Artificial Intelligence

Abstract—The following report presents an investigation into the relationship between (self-judged) humour types and demographic information, with the assistance of predictive machine learning algorithms. We used data from an online humour styles questionnaire that assigns participants scores for four humour types - ”affiliative”, ”self-enhancing”, ”aggressive” and ”self-defeating”. From these scores we aimed to determine if there was a relationship between a participant’s humour types, and their age and gender. Following transformations and encoding of features and targets (section 2), training on a range of linear, non-linear and ensemble models (section 3) and hyperparameter tuning (section 4), we found that few models achieved better than random prediction accuracy (17% in our case). From this evidence, we conclude in section 5 that there is likely no relationship between humour-type and demographic information *that is tractable by the machine learning methods used in this report*.

Index Terms—Demographics, Humour types, Machine Learning, Classification

1 Introduction

It has been observed that an individual’s use of humour can have noticeable effects on both their mental and physical wellbeing [1]. As a result, there may be great value in determining the relationship between humour use and key demographic characteristics, such as age and gender. The appearance of any strong correlations between these variables could then inform social policy decisions, and lead to an increase in the welfare of the general population.

In this study we used the results of the survey conducted in [3] to analyse relationships between self-judged humour-type and demographic data, namely, age and gender. To do so, machine learning *classification* algorithms were utilised, implemented using the Python library scikit learn. Participants in the survey was asked to answer 32 questions relating to their own use of humour (giving a value from 1-5), and from these responses, scores for four ”humour-types” were be calculated [1]. These humour-types scores (from now on referred to as H-types) measured how strongly the following four adjectives matched the participants use of humour: ”affiliative”, ”self-enhancing”, ”aggressive” and ”self-defeating”. A four-dimensional vector composed of the participant’s score on each of these H-types formed our feature vector, with the output of our machine learning models being one of eight possible age group and gender classes, encoded as an integer. The eight classes were those in the Cartesian product of the two sets G = {”male”,”female”} and A = {”10-20”,”20-30”,”30-40”,”40-70”}. We will begin by discussing in further detail the dataset used for this study, as well as the transformations applied to prepare it for insertion into a predictive model.

2 Exploratory Data Analysis and Transformation

The dataset used in this study consisted of responses from 1071 participants. Each participant gave a self-judged accuracy score from 0-100 in addition to their 32 questions responses, with 100 indicating that the participant was completely confident with their answers, and 0 the opposite case. We begun by removing rows where age was given as greater than 120 (on the basis that these were unrealistic) or gender given as 0 (participants were asked to respond with 1,2 or 3). Furthermore, we removed rows in which gender was recorded as 3 (corresponding to ”other”); less than 1% of partipants recorded this response yet it would have increased the number of classes by 50%, significantly increasing the degree of imbalance of the dataset.

After our initial revision of the dataset we were left with 1055 responses; the mean age of respondents was 26 (ranging from 14 to 70), with females accounting for 45% of responses and males 55%. It should stated that in the original survey, if participants *did not* select an answer for any of the 32 questions, a -1 was entered instead. To make sure that this did not affect our results we replaced all -1’s given for question X with the modal response to question X, and *recalculated humour type scores*. Given that -1’s were entered rather infrequently (the question with the most -1’s entered was Q7, with 13), we were not concerned about ‘flooding’ the dataset with the modal class.

As stated above, the input of our machine learning models was a four-dimensional vector of participant humour types. The dimensionality of our feature vector was relatively low, therefore we decided against the incorporation of dimensionality-reduction techniques. Indeed, as can be seen from Figure 1, the fourth principal component accounted for over ten percent of the variance across the humour types dataset, which we considered to be non-negligible.

[1] These were linear combinations of the responses to eight questions, and were also in range of 1 to 5.

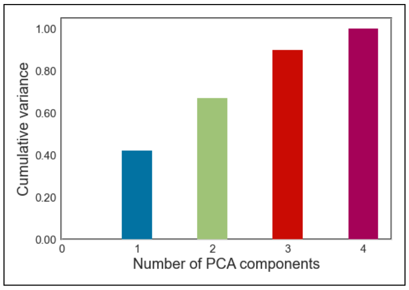

Fig. 1: Cumulative variance of the humour types dataset against number of principal components.

To gauge the distribution of H-types across the 1-5 range, we used figure 2. We observe that the ”self-enhancing”, ”affiliative” and ”self-defeating” H-types are all relatively well distributed over the range 1.0 to 5.0, whereas there seems to be a disproportionately large number of respondents with H-type ”aggressive” centred around 3.0. From the range of the distributions in 2, we also decided against feature scaling. All features are in the same range and have relatively small absolute values, thus we were not concerned about problems of convergence that may arise from unscaled features.

It is important to note the degree of imbalance across the dataset. Figure 3 makes explicit the oversampling of the ”10-20” and ”20-30” age groups across both genders in the initial study; we see that these age groups have roughly three to five times more data points than the ”30-40” and ”40-70” groups. If we had not grouped the ”40-50”,”50-60” and ”60-70” groups together, the imbalance would have been even more exaggerated.

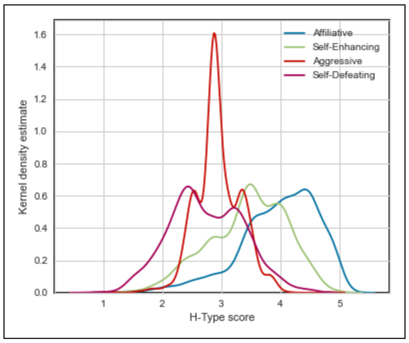

Fig. 2: Kernel density estimates for each of the four H-types.

To resolve the matter of data imbalance we weighted each training vector as follows. Letting X_i be the ith training feature vector, n_{class_i} the count number of the *class* to which X_i belongs, and a_i be the (self-judged) accuracy of the respondent that accounts for X_i, we assigned a weight $w(X_i)$ where:

$$w(X_i) = \frac{a_i}{n_{class_i}} \tag{1}$$

In doing so we hoped to allow the sparsely populated classes greater influence over the cost function when training takes place, and incorporate the (self-judged) accuracy of participants’ answers. However more data should be collected from the 30-70 subset of the population to increase the accuracy of trained models.

Encoding of the target variable was also required, given the categorical nature of the class labels. The labels were of the form ”age-bin, gender”, for example, ”20-30, female”. Age bins may be considered an ordinal categorical variable whereas gender is a nominal categorical variable, thus we tried both integer and one-hot encoding. After plotting Figure 4 (to be discussed in the next section) for both types of encoding and seeing little difference in model performance, we opted for integer encoding due to its simplicity.

Finally, we split the dataset into training and testing subsets. We did so with an 80:20 split, ensuring that the training and test data both had similar distributions over classes.

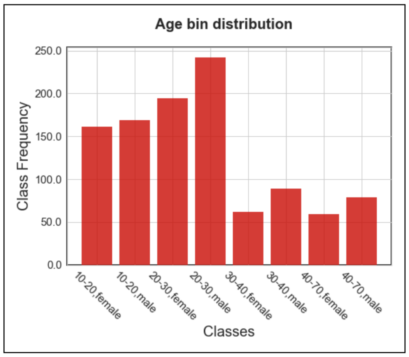

Fig. 3: Class frequency distribution

3 Learning Algorithm Selection

Given the large number of machine learning classification algorithms provided by the *sci-kit learn* library, we began the selection process by filtering classification algorithms to determine the three models that would be most suitable for our study. We did so in the following method: Nine classification algorithms were trained on the dataset discussed above, and were evaluated based on their **accuracy** and **weighted F1 score**[2]. We used the F1 metric as it is particularly suited to measuring performance of models trained on an imbalanced dataset. The results of the evaluation were used to determine whether our problem favours low bias (and high variance) complex models or high bias (and lower variance) simple models. It should be noted that these models were trained with their default hyperparameters, as we deemed it to be impractical to tune the hyperparameters for nine different models. The results for the accuracy and weighted F1 metrics are plotted in Figure 4.

The first thing that can be observed is that performance seems to be relatively poor across the full range of models (the maximum accuracy and F1 score is ~ 20%) , even after extensive feature engineering and target encoding. Indeed, the average performance of the models is in the 15-20% range, which is what we would expect to observe if the models were randomly selecting one of the six classes(16.67%). The reasons for this could be many-fold; perhaps the target variables are not strongly dependent on the features (from observing human behaviour this seems valid - a particular age group is often not associated with using humour in a certain ”style”). Further transformations of the dataset could

[2] Weighted with respect to class frequency.

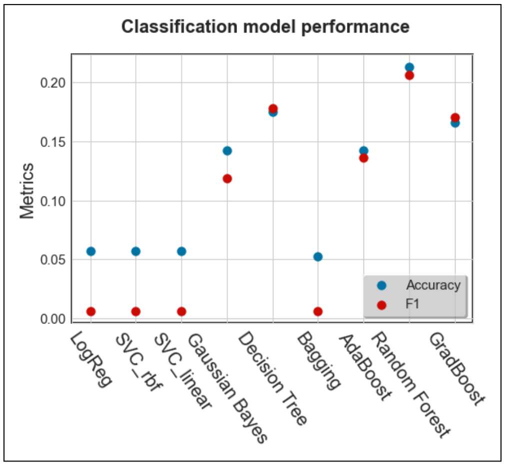

Fig. 4: Accuracy and F1 metrics for a wide range of models. A note on labels on the x axis - LogReg refers to Logistic Regression, SVC_rbf is a Support Vector Classifier with a radial basis function as the kernel.

also be required.
Nevertheless, we continued our study with the top performing models from this evaluation. We believed that it would still be insightful to discern whether or not humour type scores can be used to predict age and gender, and the remainder of this report will look to address this. The Adaptive Boost, Random Forest and Gradient Boost methods were selected for hyperparameter optimisation, due to their *relatively* high accuracy and F1 scores.

4 Model Training and Evaluation

The next step in our study was hyperparameter optimisation of the promising models. To optimize model hyperparameters, we used the *Random Search*[3] functionality provided by *sci-kit learn*, which chooses sets of hyperparameters randomly from a pre-defined search space, and outputs the optimal hyperparameter configuration. By optimal we refer to the hyperparameter configuration that returns the highest **F1** score, chosen due to our unbalanced dataset. The candidate values for each model are given in the tables below, with the hyperparameters obtained through RandomSearch highlighted in red.

- AdaBoost Classifier with Decision Tree weak learner [4]

Hyperparameter	**Candidate Values**
Learning rate (10^x)	-5 ,-4,-3,-2,-1,0,1,2
#Estimators	30,40,50,60,70,80,90, 100

- Gradient Boosting Classifier

Learning rate (10^x)	-5 ,-4,-3,-2,-1,0,1,2
#Estimators	30,40,50,60, 70 ,80,90,100
(Tree) max depth	1,2, 3

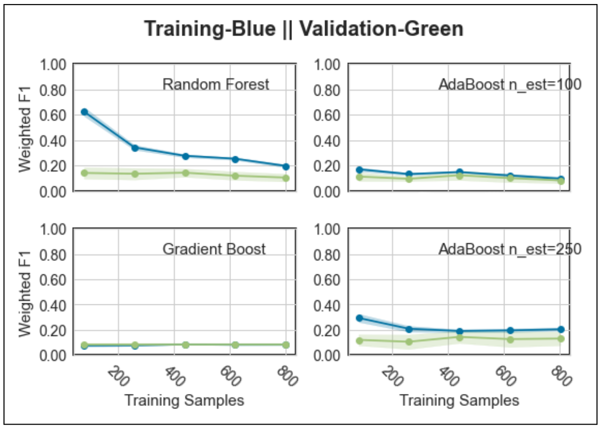

Fig. 5: Learning curves for classification algorithms (Bottom right
is AdaBoost but with a different number of estimators).

- Random Forest Classifier

Split Criterion	Gini , Entropy
(Tree) max depth	1,2, 3
#Estimators	30,40,50,60,70,80, 90 ,100

Using these parameter values we fitted the models to our training data, which yielded the learning curves shown in 5. To obtain the validation dataset we used 10-fold stratified cross validation, with the number of folds chosen rather arbitrarily after plotting validation accuracy against folds and finding no optimal fold number.

The apparent difference in the learning curve shape will be discussed in the next section (5); here, we instead focus on questions of underfitting and convergence. All three model record poor final values of accuracy for **both** curves (training and validation) once convergence has taken place. Such learning curves are the hallmark of underfitting - the classification models used cannot seem to find a suitable hypothesis function that links the H-type vector and age bin and gender. To further check if our models were suffering from underfitting, we plotted learning curves for Adaptive Boost algorithms with 100 and 250 base estimators (increasing estimators for boosting ensemble methods is known to reduce bias at the expense of increasing variance). However, we saw little change in accuracy, as the two right hand side plots of Figure 5 show. This evidence further suggests that there is simply no (linear or non-linear) relationship between humour type and demographic information. On the matter of convergence, all three plots show near to absolute convergence of the training and validation curves, indicating little tendency to overfit and that the training set data is sufficient in number.
The Precision-Recall curves in Figure 6 also prove useful for evaluating model skill level. We justify the use of the Precision-Recall curve rather than other means of evaluation such as the receiver operating characteristic (ROC) curve, primarily due to the misleading view that the latter can give for a classifier that has been trained on an imbalanced dataset [6] (as was the case in this study). An ideal model would have precision = recall = 1.0, which is clearly not the case in 6. Here instead we see the opposite, that is, poor

[3] Given the relatively small search spaces we found Random Search a better option than Grid Search or Bayesian methods.
[4] Decision Tree weak learner chosen due to its performance in Figure 4

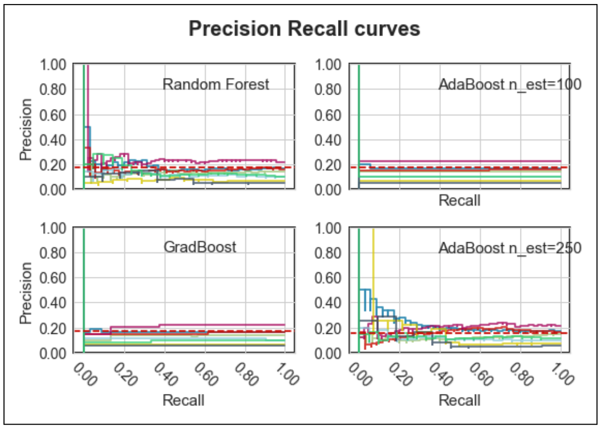

Fig. 6: The Precision-Recall curves for the models, where precision measures the reliability of the positive classifications made by the model, and recall is the ability for a model to identify actual positive results. Each coloured line represents the Precision-Recall curve for a single class.

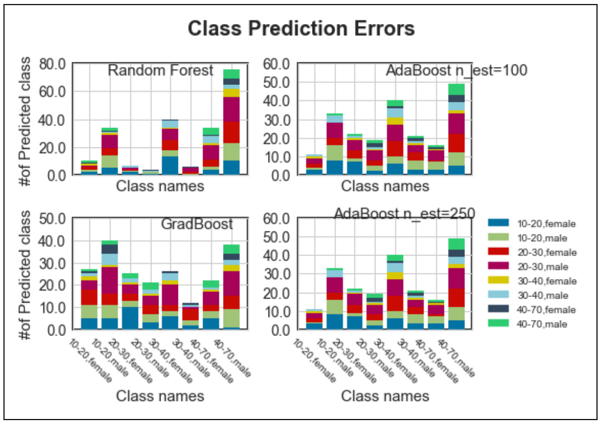

Fig. 7: Class prediction errors. Note the difference in Y axis scales.

precision for all values of recall, once again demonstrating the poor skill of the models in predicting age bin and gender type.

5 Model Comparison

We draw the study to a close by considering methods by which the models could be compared. Classification error plots are useful in diagnosing exactly where the models are making incorrect predictions, and are given in Figure 7. The height of each column indicates the *number of predictions* that the model made of Class X, with the stacked bar showing the *true class* distribution. Recall the *expected* distribution across classes (Figure 3). We see, with respect to this distribution, that the predicted class distributions below are roughly distributed across all possible classes (with the exception of the Random Forest), rather than concentrated in the ”10-20” and ”20-30” groups. Such a distribution implies that the trained models may be engaging in random selection, which we have come to expect given the evidence so far that there is no suitable hypothesis function mapping H-types to age and gender. The tendency for the Random Forest to overestimate the ”40-70” class was a suprise to us, and we cannot offer an explanation for this result at present.

Finally, we can refer to Figure 6 to further compare model performance. The Random Forest and AdaBoost Classifier (with 250 weak learners) have marginally higher precisions across a range of recall values than the other models trained. If one was to pursue this study further, these may be the most promising lines to follow.

6 Conclusion

This report aimed to establish whether there was a tractable hypothesis function that would predict age bin and gender given a four dimensional feature vector consisting of humour type scores found in [3]. After initial feature engineering and target integer encoding, we trained nine classification algorithms on our dataset; the three best performers then underwent hyperparameter tuning and more rigorous evaluation. The majority of evaluation and optimisation was done with respect to weighted F1 scores, due to the imbalance of the dataset. Our results suggest that there is little to no relationship between humour type and the age and gender. Evaluation metrics rarely exceeded 20% for both training and validation curves, and methods to increase model bias (such as raising the number of estimators for Boosting algorithms) did little to change this. Given the importance of feature engineering in machine learning experiments, there is always a chance that we may have overlooked a particularly important transformation, but we believe this is not the case.

One technique that could be used to improve the methodology of our future work is that of finding the Pearson correlation coefficients (PCC) between the feature vectors and target variables during EDA. Upon doing this midway through our study, we found the following:

	SE	Aggressive	SD	Affiliative
PCC	0.03	0.04	0.05	0.02

where we have used SE and SD to refer to the self-enhancing and self-defeating humour types. Thus, it was shown that there was practically no *linear* relationship between the H-types and age and gender. Of course, no linear relationship does not necessarily imply that a non-linear relationship will not be found, but this result would have given an early indication that *humour-types and age and gender are likely not linked*.

References

[1] Earleywine. M, *Humour and Psychological Well-being*, (2010), Humor 101

[2] Pedregosa et al., *Sci-kit learn: Machine Learning in Python*, (2011), Journal of Machine Learing Research

[3] Rod A. Martin et al., *Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire*, (2003), Journal of Research in Personality

[4] T Franssen et al. , *Age differences in demographic, social and health-related factors associated with loneliness across the adult life span (19–65 years): a cross-sectional study in the Netherlands*, (2020), BMC Public Health 20

[5] M. Matud et al., *International Journal of Environmental Research and Public Health*, (2019), Joru, International Journal of Environmental Research and Public Health

[6] Davis et al., *The Relationship between Precision-Recall and ROC Curves*, (2006), Association for Computing Machinery