Exposure Based Face Memory Test Responses
Final Project Report

Hardy Bright

Brown University - Data 1030 - Fall 2020

https://github.com/HardyBright/data1030.git

Dataset: Exposure Based Face Memory Test Responses (EBFMT)

Introduction

Data Science concepts and techniques help us to move toward a fuller understanding of how neurological systems shape our experiences. In return, the pursuit of insight into brain function, informs Data Science. The Open-Source Psychometrics Project developed the combination test and survey to study Prosopagnosia, or “face blindness”. The dataset, Exposure Based Face Memory Test Responses, used in this project was compiled as a part of their ongoing work (Open-Source Psychometrics Project). The dataset was shared to Kaggle by Lucas Greenwell who is listed as it’s owner (Greenwell, Kaggle). Prosopagnosia being a neurological condition is interesting physiologically and its study expands into the psychology of how a person with the condition navigates daily interactions that may go unnoticed by others. This project is motivated by the question of how well can the “score” earned on the Exposure Based Face Memory Test can be predicted using the accumulated data from previous test-takers. The target variable is a continuous numerical value making this a regression problem. This report updates the information offered in my Midterm Project Report (Bright).

The Open-Source Psychometrics Project has continued to administer their test online, https://openpsychometrics.org. At the time of writing this report, there does not appear to be a publicly active data analysis project using this dataset. The dataset was uploaded to Kaggle in May of 2020. There are no associated projects listed by Kaggle

and no discussions regarding the dataset (Greenwell, Kaggle). For background information on what motivated their research, The Open-Source Psychometrics Project references published research on the condition from the Oxford Handbook of Face Perception (Calder, Rhodes, Johnson, Haxby) (Open-Source Psychometrics Project).

Exploratory Data Analysis

The tab delimited CSV data file has 1768 rows with 275 feature columns and is Independent and Identically Distributed (IID). The majority of the values are numerical with one column (IP_country) containing strings. The codebook.txt file, included with the dataset, provides some detail about the features and the numerical translations that were used. The first 75 features are categorical numerical values (denoted Q1 to Q75) which correspond to the answers to face recognition questions for each respondent. Column 275 is denoted as the “score” (Figure 1) and is the continuous target variable. Although the codebook.txt does not explicitly state how the score was achieved, it appears to be a cumulative value of the number of correct answers. During the test, the respondent is asked if the person in the photo being shown to them is a “new” face (encoded as 1) or a “seen” face (encoded as 2).

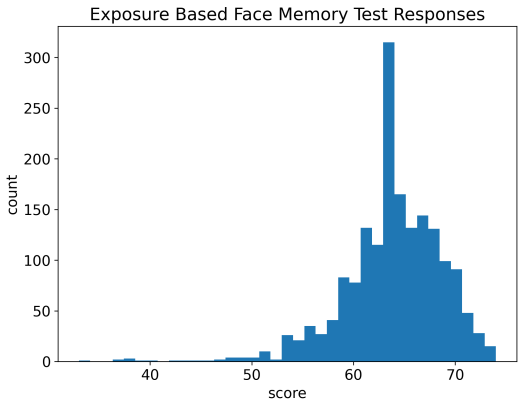

Figure 1 - The “score” (target variable) test result counts.

Going further into organization, it's noted that the developers of the online test collected measurements on the elapsed time each participant. There are 75 features (LAPSE1 to LAPSE75) referencing the time it took to answer the corresponding question (milliseconds). Other seemingly time related metrics (introlapse, testelapse and fastclicks) were not explained in the documentation. The test-takers were asked to participate in an optional survey for demographic information (country, english as native language, age, education, gender, urban, orientation, race, religion, and dominant hand). With the exception of "IP_country" (string abbreviations), the demographics were recorded as numerical categories. Additionally, the survey included questions from the 100-item version of the HEXACO Personality Inventory - Revised (HEXACO-PI-R) (Lee, Ashton). Each person was given a random sample of the 100 test questions and their responses which were encoded ranging from 1 (disagree) to 0 (neutral) to 5 (agree). All the personality test data was recorded into features PQ1-PQ100 and PQI1-PQI11. Therefore, the dataset can be considered to be an organized whole with 3 connected parts; face recognition test data and score, demographic survey data, and personality test data.

Methodology

The data frame was separated with feature columns assigned to "x" and the target variable assigned to "y". The splitting strategy used was 60% for training, 20% for validation and 20% in a test set held out from the other two. Ten numbers that were used as random seeds were stored in an array. This allowed for a new random seed to be used for each iteration through the pipeline. After further investigation following the initial EDA stage, I chose to exclude (as opposed to imputation) two points (indexes 215 and 343) that contained missing values within the 'IP_country" feature. These two points constituted 0.11 percent of the overall and appear to be Missing Completely At Random (MCAR). Two additional points were excluded (indexes 903 and 996), both due to

apparently incorrect “age” data being entered. Before preprocessing, there are 0% missing values in the test set.

The OneHotEncoder was used for all categorical features. This included the numerically encoded values outlined previously. The continuous features were addressed with the StandardScaler. Finally, the MinMaxScaler was used on the “age” category. After preprocessing the feature columns increased to 1043 in the train, validation and test sets for one random state as an example. The datapoints were separated into 1058 for training and 343 for validation and test. Because this is a regression target variable, that column was not preprocessed.

I chose R^2 score (or coefficient of determination) for it’s versatility and the interpretability of the resulting value. As the scikit-learn documentation points out, “it provides a goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model” (scikit-learn developers, 2007-2020). There’s a scale indicting the performance of the model ranging from -1 to 1, and a baseline of 0 with 1 being the best score possible.

The same general pipeline and techniques was used across the three machine learning algorithms for consistency, reproducibility, and comparability of results. A consideration was to build a process where the algorithm in use and its hyperparameters can be interchanged leaving other factors constant. The pipeline makes use of 10 random states that are stored in a array. Dataset splitting, preprocessing, training set fit, model transforms, and metric calculations, are all done through successive stages. R^2 scores are calculated and collected into arrays. Their overall mean and standard deviations are calculated for each algorithm. In cases where the hyperparameters can be adjusted, that is done for each random state.

When fitting Lasso regularization, I used a range of alpha values in logspace from -7 to 29. When the pipeline used for Random Forest Regression, the n_estimators

was kept constant (100) while maximum tree depth used values of 1, 2, 3, 5, 10, 20, and 50.

Results

Linear regression was the first algorithm used with interesting results. The model produced consistently high R^2 scores across all dataset splits and random states. The mean test R^2 score is 0.9996 +/- 0.00117 standard deviation. Given the structure of the metric, this indicates an nearly perfectly predictive model.

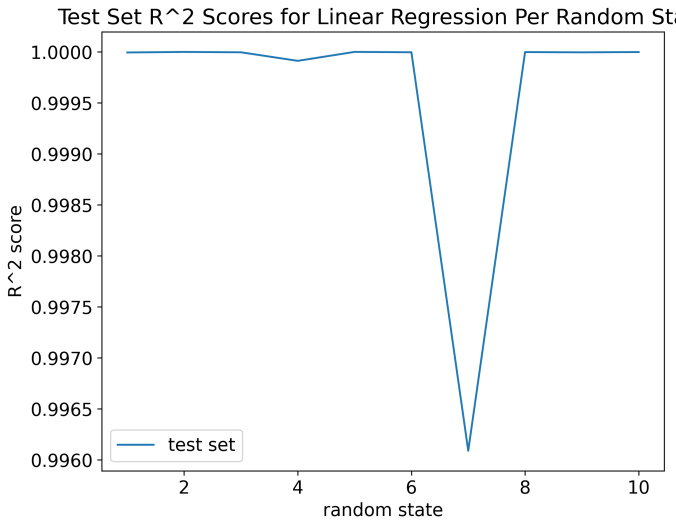

Figure state.

The consistently high scores may be an indication of an issue somewhere within the process. To discover an explanation for the scores, I reexamined to pipeline but did not uncover any errors leading to this. In the next step, I calculated the correlation coefficients for the training set to determine if there are any features that after preprocessing, held very high correlation or anti-correlation relationships with the target variable. There were no values to the high or low extremes indicating that any particular features were causing these scores. Additionally, I altered the preprocessing strategy for a series of categorical features (PQ1 - PQ100). Previous to the change, the

OrdinalEncoder was being employed due to the fact that although they are numerically encoded already, they have an order to them as referenced in the Exploratory Data Analysis section. There was no overall change to metrics with the preprocessing change.

The second algorithm I used was Lasso regularization. My consideration here was to continue to look for answers to the previous scores with an algorithm that is related but has hyperparameter adjustments for more analysis. Here I used the pipeline to vary the alpha parameter as noted in the methodology section. The results are largely the same as for linear regression. For example, Figure 4 below depicts the metric for one random state which is one from the same set used for other algorithms. Again, there is a consistently high R^2 score until near the end of the alpha scale when it declines sharply.

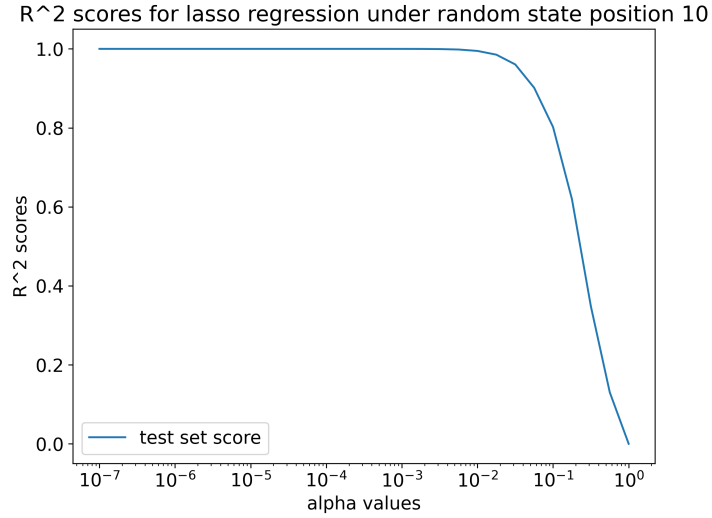

Figure 4 - R^2 scores per Lasso regularization alpha value. Note the decline only after consistently high R^2 scores.

When applying the Random Forest Regressor the story being conveyed through the metrics changed. Random forests consistently performed well (above the 0 baseline) but not has high as linear regression or Lasso. Using the same methodology, but varying the number of trees available, the algorithm produced a overall mean R^2

score of 0.4424 +/- 0.02902. The best metric value, 0.59361 +/- 0.02455, was produced with a max_depth of 20.

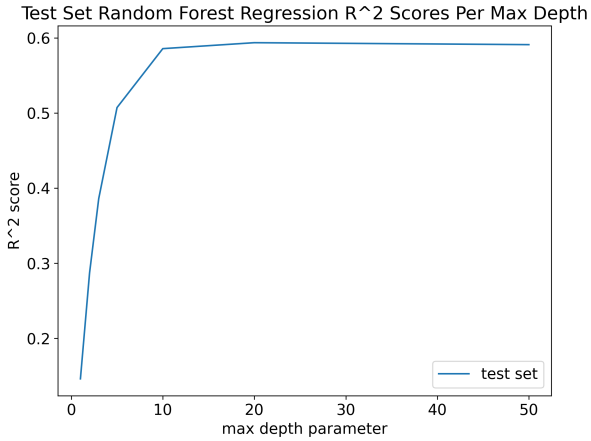

Figure 5 - Test set Random Forest Regressor R^2 score and max_depth parameter.

It's interesting to note that the model's behavior shown in Figure 5 was also present in the training and validation sets. In all cases, there are sharp increases until reaching max_depth of 10. After reaching the peak value, the performance levels off. It's indicative that increasing the max_depth will likely not increase predictive power.

Outlook

I believe there is much more to explore regarding this data set using different techniques. The Fact that the Random Forest scores were marginally lower in every way to the other two techniques. In the future, I would tune the Random Forest Regressor's n_estimators parameter to determine if the algorithm would perform even better given higher values (while keeping the max_depth constant at 20).

To improve the methodology, I would incorporate "pickle", GridSearchCV and ParameterGrid functionality into the workflow. All three of these technologies streamline processes that I coded by hand. Technologies like "pickle" increase reproducibility and interpretation from others working on the project. I also believe that in future versions of

the workflow it will be an advantage to contain the pipeline in a function which may make it more modular and portable to other projects.

Works Cited

Bright, H. . Midterm Project Report, Exposure Based Face Memory Test Responses Project, Brown University, Data 1030, 10-14-20

Exposure Based Face Memory Test (EBFMT), Open-Source Psychometrics Project, https://openpsychometrics.org/tests/EBFMT/, https://openpsychometrics.org, Site updated September 17, 2019 (Accessed 10-12-20)

Greenwell, L. . https://www.kaggle.com/lucasgreenwell/exposure-based-face-memory-test-responses/, Uploaded to Kaggle 05-31-20 (Accessed 10-2-20 and 11-30-20)

Calder, A, Rhodes, Johnson, G, Haxby, J (editors). "Oxford handbook of face perception". New York: Oxford University Press, 2011.

scikit-learn developers (BSD License), “3.3 Metrics and scoring: quantifying the quality of predictions”, https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics, 2007-2020, (Accessed 11-30-20)

scikit-learn developers (BSD License), “sklearn.metrics.r2score”, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html, 2007-2020 (Accessed 11-30-20)

Lee, K., & Ashton, M. C. . http://hexaco.org/hexaco-inventory, 2009, (Accessed 10-12-20)

Lee, K., & Ashton, M. C., http://hexaco.org/downloads/English_self100.doc , 2009, (Accessed 10-12-20)

Lee, K., & Ashton, M. C.. Psychometric properties of the HEXACO-100. *Assessment*, *25*, 543-556. 2018