Assessing Biases in Teacher Evaluations Dataset

Nikolaos Tsikouras

April 2021

Motivation for the Project

For this Project we are going to be using the Teacher Evaluations dataset [5]. As we will see in later sections the dataset consists of 21 variables, with the variable of interest being "score", which takes values from 1 (very unsatisfactory) to 5 (excellent). Using the rest of the feature variables we are going to train a model which will predict scores of professors based on a vector of his/her characteristics, and by a vector of the course's characteristics.

Being interested in social psychology as well as behavioral economics and having recently finished reading the inspiring book Thinking, Fast and Slow [6] by Dr. Kahneman, we chose this dataset since we would like to investigate if there is a link between a professor's score and his/her looks. Additional motivation for this project was that economists have considered the effects of beauty in the labor market, more specifically how earnings are affected by it [10]. Thus, since teaching quality, or in our context teacher evaluations, is a factor taken into account for salaries [12], this is a good opportunity to see if teaching quality, as perceived by students, is solely dependent on a professor's ability or if external characteristics, such as beauty, ethnicity or gender lead students to evaluate teachers more or less favourably.

An interesting question is by how much does a teacher impact students' learning outcomes. Although this is a questions which is feasibly impossible to answer with the given dataset, we are going to provide a possible solution in the final part of this Project.

Literature Review on the Problem and Plan

Questions related to teacher effectiveness have a long history in the literature within the broader field of research on teaching and teacher education, as well as research on school effectiveness. [16]. The image of a professor has been found to be a critical factor in determining how effectively he or she is perceived to be able to teach [4]. Additionally, Anderson [2] concluded that attractiveness is positively associated with perceptions of success when people meet for the first time. Dion et al. [8] demonstrated that individuals found attractive people higher on multiple dimensions, including professional success. Finally, Landy et al. [14] found that attractive people were perceived as more talented. These articles are indications that a person's looks is, in some way, affecting how he/she is perceived.

The first study to look at teaching evaluations was by Ambady et al [1], who conducted a study where they showed individuals a video tape including 30 seconds from a class for 13 different teachers. They found that these silent video clips were enough to significantly predict global end-of-semester students evaluations of teachers. Although this study focused on nonverbal behavior, it demonstrates that teacher evaluations are not solely based on teacher performance, since if they were, having a video sample of 30 seconds should not have any predictive power of the final teacher scores.

An indicator of students' bias regarding task evaluation is demonstrated by Biddle and Hamermesh [3]. In this study male college subjects read an essay that was supposedly written by a college freshman. They then were asked to evaluate the quality of the essay. On the essay there was a photo attached, a third of the students were led to believe that the writer was physically attractive, a third of them that she was physically unattractive and the final third had no information on the appearance. The results showed that the subjects evaluated favorably the essay when they knew that the writer was attractive,

less favorably when she was unattractive and intermediately when they had no information. Again this is in indication that students' judgement can be affected by external characteristics.

Regarding the model we are going to train in the following sections, Kotsiantis et al. [13] have proposed the use of Decision Trees to answer a similar question; "predicting students' performance in distance learning" and they have achieved a high accuracy, thus taking this into consideration we are going to use a random forest model which is a method that combines several decision trees.

Introducing the Dataset

The dataset consists of end semester student evaluations for a large sample of professors from the University of Texas at Austin. Additionally, there exist ratings for the physical appearance of the professors by six students. Furthermore, there is a variable for number of students in class which ranges from 8 to 581, while the number of students who completed the survey ranges from 5 to 380.

The dataset also contains information about the faculty member's gender, whether on the tenure track or not, minority status and whether he/she was educated in an English-speaking country or not.

The result is a data frame where each row contains a different course, 463 in total, and each column has information on either the course or the professor.

The dataset is balanced in gender, a variable which we are going to further investigate later, with roughly 42% (195) females and 58% (268) males. Roughly 85% of the faculty members are in the minority category while 5% were not educated in an English-speaking country.

The ratings for the physical appearance had a 10 (highest) to 1 rating scale. As mentioned before the ratings come from six students; three women and three men, with one of each gender being a lower-division and two of each gender being upper-division students. Using these 6 values, we created a new column with the average physical appearance rating. Table 1 presents the Mean and Standard Deviation of the ratings of the professors' beauty. Although, 5 should have been the average mark, it seems that this is not the case for most of these ratings since most of them seem to be skewed to the right. Additionally, we calculated the fifteen Pearson pairwise correlations of the beauty ratings and they ranged from 0.51 to 0.69 with an average of 0.60. Therefore, we can conclude that beauty ratings are consistent across the raters as demonstrated by the high correlations.

	Mean	Std. Dev.
Male, Upper	4.14	2.11
Male, Upper	4.75	1.57
Male, Lower	3.41	1.63
Female, Upper	5.01	1.93
Female, Upper	5.21	2.01
Female, Lower	3.96	1.87
Average	4.41	1.52

Table 1: Mean and Standard Deviation of the ratings of the professors' beauty.

Finally, Figure 1 shows the density plots of the score for the two genders with the mean indicated by the vertical dashed lines. The distributions seem quite similar but the main difference is that the mean score for the female faculty members (4.09) is lower compared to the mean score for the male faculty members (4.23). We are going to investigate this further later.

Introducing the Model and Results

For the problem in hand we are going to train a Random Forest model. The feature variables for the full model are Rank (teaching, tenure track, tenured), Ethnicity (not minority, minority), Gender (female, male), Language of school where professor received education (English or non-English), Class Level (lower or upper), Class Credits (one credit, multi credit), Average Beauty (1-10). One of the reasons we picked these 7 variables is that, using any of the other variables did not make any logical

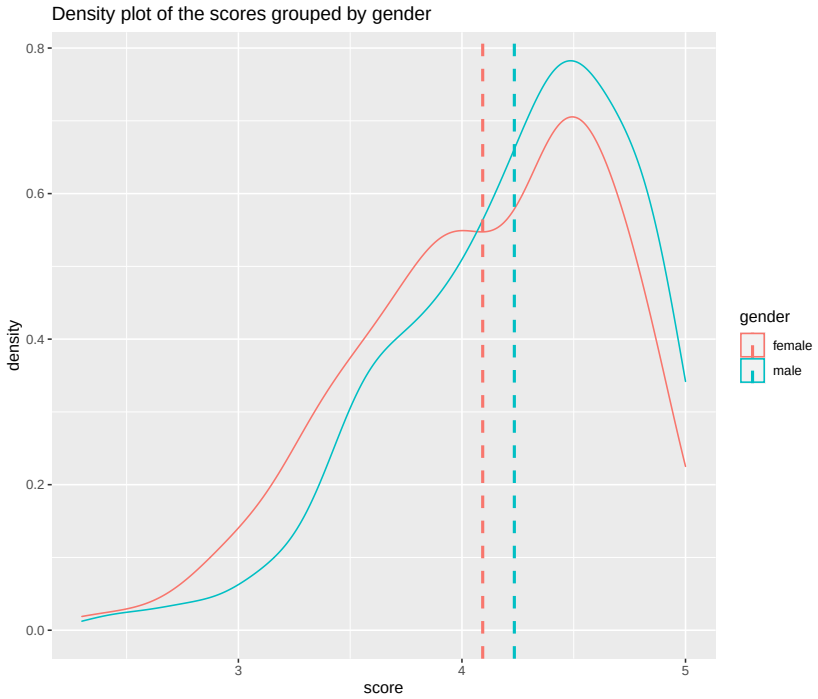

Figure 1: Density plots of the score for the two genders. The dashed vertical lines correspond to the mean of each distribution.

sense. Additionally, we tested fitting a model using all the variables and the MSE was the same, we also tried removing variables one by one and did not notice a major change in MSE, thus using Occam's razor principle, we will pick the model that uses less variables.

A random forest is an ensemble learning method and it works by constructing multiple trees at training time and it outputs the average prediction of the individual trees. To create one tree the algorithm first needs to create bootstrap samples, i.e. samples with replacement of the same size as the original dataset, and then each tree is fitted on the resampled data. It is clear that for each tree the algorithm is not using all of the data points, these (unused) data points are called Out Of Bag (OOB) data points and are used to estimate the test error.

With that being said, we are going to explain the training process. First, we are splitting the dataset into a training and test set using a 70-30 split. The test set is going to be used for the final prediction. Now, to train the model we are going to use we are going to use the OOB error estimates to identify hyperparameters. We are creating a grid of hyperparameter combinations and we are going to find which of those combinations return the smallest OOB error, more specifically we are using as number of variables randomly sampled as candidates at each split "mtry": (3, 5 by 2), maximum number of terminal nodes "maxnodes": (10, 100 by 5), minimum size of terminal nodes "nodesize": (1, 21 by 2) and number of trees to grow "ntree": (100, 1000 by 50). Then for each of these combinations we are training a random forest for the training set and we pick the set of parameters that return the lowest OOB MSE. The best hyperparameters are mtry: 5, maxnodes: 90, nodesize: 5, ntree: 350. The MSE we got in the training set is 0.113, while the MSE we got in the training set using the aggregated predictions from the trees that did not see the examples during training (OOB) is 0.205, which is expected to be higher by definition. Finally, the MSE on the test set with OOB-tuned hyperparameters is 0.167, thus there is no sign of overfitting, which is expected since Random Forests are not known to overfit.

Figure 2 shows the true values of y against their predictions by the full Random Forest model. We have also plotted the $y = x$ line for reference. The trained model seems to predict the scores quite well, as indicated by the low MSE.

To explore if the trained model is biased we are going to measure which variables have the most predictive power, i.e. we are going to calculate variable importance for the variables in our model. Variables with high importance are crucial and their values affect significantly the outcome values.

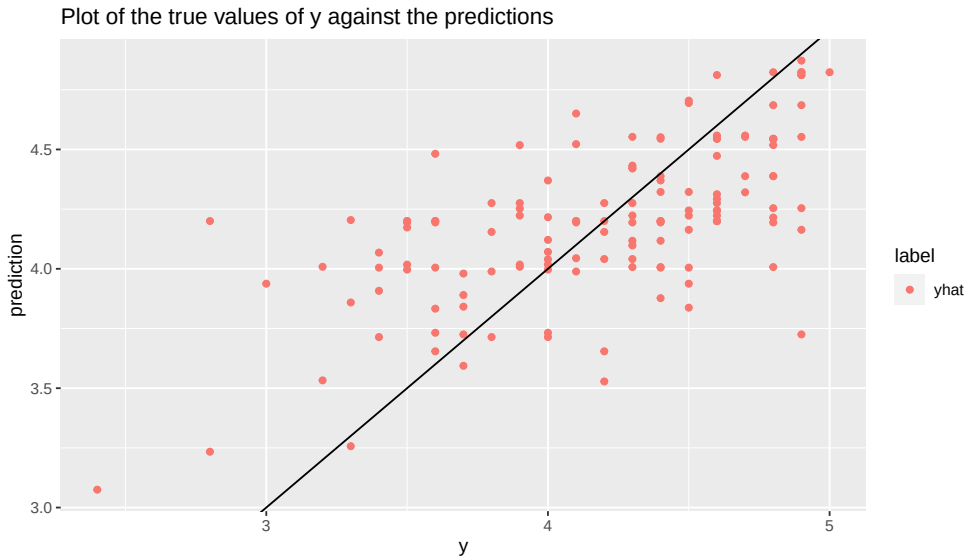

Figure 2: Plot of the true values of y against their predictions by the full model. The black line is the $y = x$ line.

On the other hand, variables with low importance can be omitted from a model, making it simpler and faster to fit and predict. We are going to calculate two measures; percentage increase in MSE, which is calculated by permuting randomly the values of the OOB samples, and Increase in Node Purity, which is calculated based on the reduction of Sum of Squared Errors whenever a variable is chosen to split. [9]

Figure 3 is a plot of Percentage Increase in MSE and Node Purity; since these two sets of measures are on different scales, we have normalised them to $[0, 1]$. The Spearman correlation between the two sets of variable importances is 0.82, so the two sets of rankings are very similar in this case. Naturally, we would not expect high values for the importance of Beauty, Ethnicity or Gender, since no study shows that Beauty, Ethnicity or Gender is indicative of a professors' teaching ability. However, Figure 3 surprisingly indicates that Beauty is the most important variable in our model followed by Gender (if using the %Increase in MSE).

To measure the magnitude of Beauty we are going to remove it from the model and calculate the MSE. We need to tune the hyperparameters of this new model using the same method as before and the same grid of values. The best hyperparameters are mtry: 3, maxnodes: 90, nodesize: 15, ntree: 150. The MSE on the test set with OOB-tuned hyperparameters is 0.269 which compared to the MSE on the test set in the full model (0.167) is obviously smaller. This difference clearly shows what we have argued before; Beauty is a variable that has a substantial predictive power, and this indicates that the model is biased. Figure 4 is a plot of Percentage Increase in MSE and Node Purity for this new "reduced" model; again we have normalised these two measures to $[0, 1]$. The Spearman correlation between the two sets of variable importances is 0.82, so again the two sets of rankings are very similar. This plot suggests that every variable is almost equally important (except Language), with notably Gender and Ethnicity being quite high. Perhaps, we should investigate further, if these variables bias the model in any way.

Recommendations for Further Work

Using the full Random Forest model specified above we achieve an MSE on the test set of 0.167. The results we got, through the analysis we conducted, are conclusive; the feature variable Beauty has a major role in predicting the score of a professor's evaluation. However, care must be taken to how we interpret our results so as to not put our preconceived ideas or biases forward. One of the ways to interpret this is that good looking professors are treated more favourably by students. Another way to interpret this is, good looking professors make students pay more attention to the class, since for

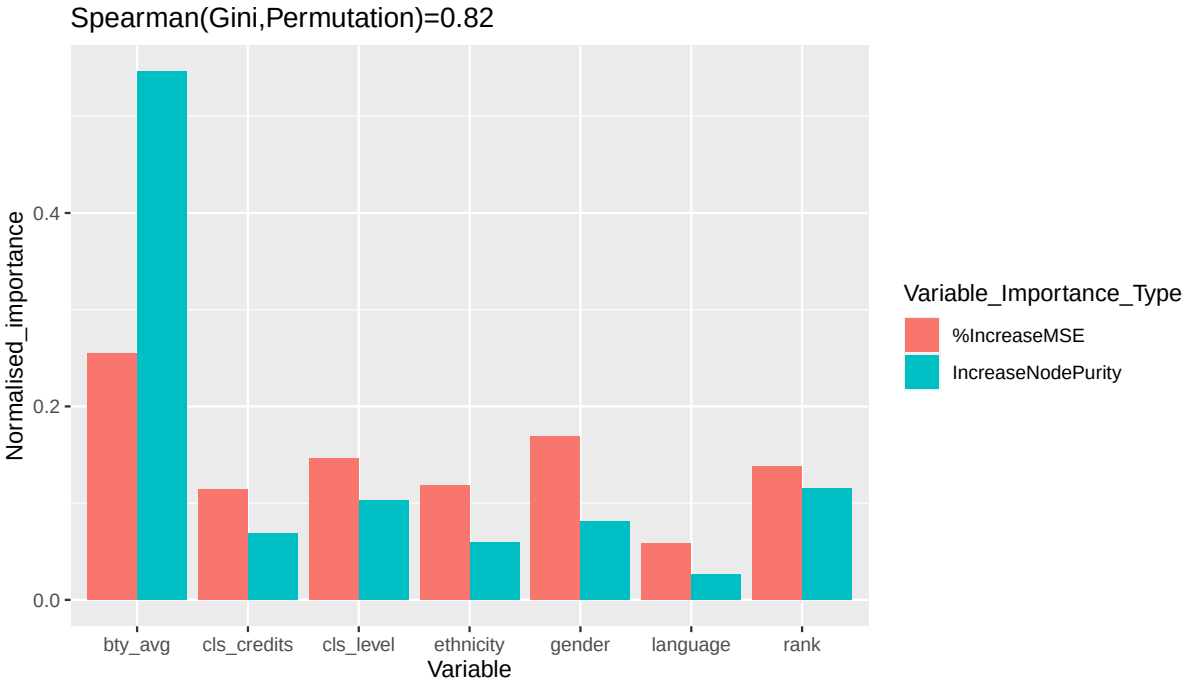

Figure 3: Plot of Gini and Permutation importance normalised to $[0, 1]$ for the feature variables in the full Random Forest model. The Spearman Correlation between the two sets of Variable Importances is 0.82.

example they might have more confidence, thus implicitly helping the students actually learn more. To find the answer to such a question, we need to quantitatively measure how much does a teacher influence students' learning outcomes. As we have already mentioned with this dataset we can not feasibly find which interpretation is correct.

To answer this we are going to present a new framework, the idea for which is inspired by the so called Value Added Models (VAM) which were developed by Hanushek [11] and Murnane [15] who measured the added value that a school teacher has on the students. We note here that we are not going provide full details of a complete model, this is just an idea we had when doing literature review and could have been implemented given more time and data. VAM is a model that includes, all the non-school factors that contribute to growth in student achievement which is a lengthy list [7]. The concept of a VAM can be demonstrated using a two-level model of student achievement. The first level of the model represents the influences of the characteristics of the individual student and his/her family on growth in student achievement (i.e. home and community supports or challenges). The second level represents the effect of the professor on growth in student achievement.

With these in mind, the model would require grades from the previous year and grades from the current year as well as information about the first level we have mentioned above. Then taking these into account and incorporating the information we have gathered regarding the teacher scores we should statistically isolate the contribution of the first level to student achievement from all other sources of student achievement.

Other things we could have implemented given more time so as to further improve the predictions would include to check several other models. Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Multilayer Perceptron (MLP) or Naive Bayes are some of the options we could have tried. Kotsiantis et al. [13] have shown that the Naive Bayes algorithm was the best (by a small margin) for predicting students' performance in distance learning. Perhaps, similar dynamics might lead this algorithm to perform well in the concept we are studying.

As far as the data is concerned, although the 7 variables we are using is not a big number, we could try and remove extra features since they can decrease performance by preventing the model from learning the actual relationships present. We have already conducted feature importance analysis (Figure 3, Figure 4) and thus we could try removing the least important variable(s) in our

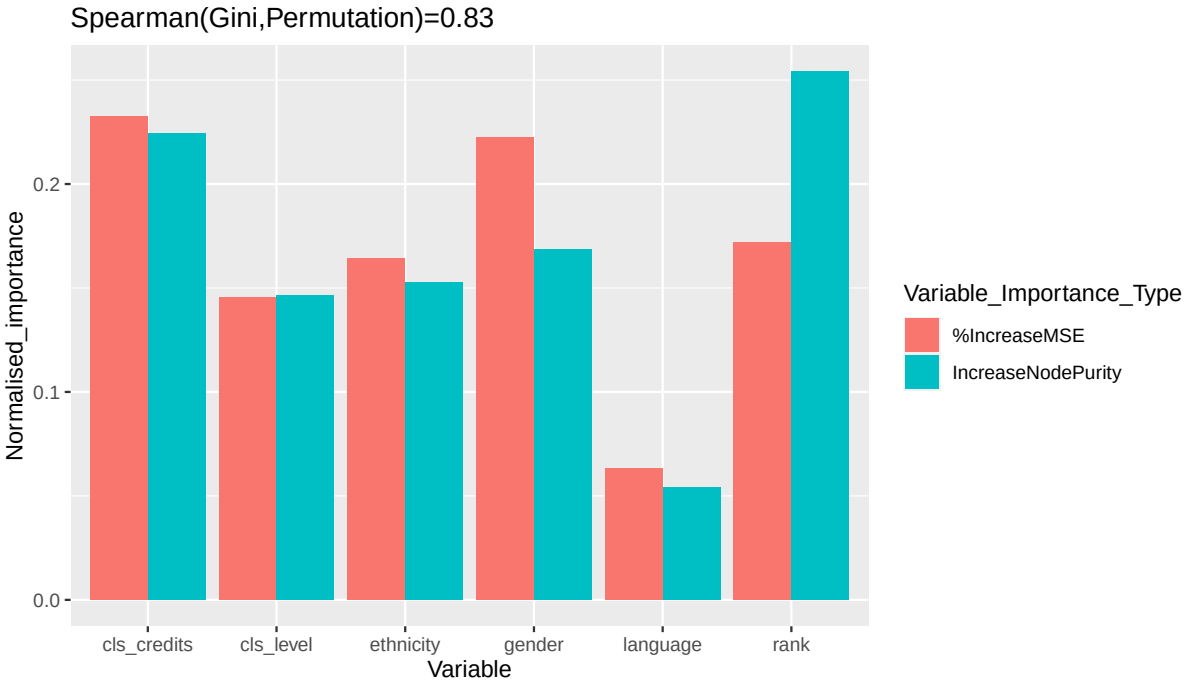

Figure 4: Plot of Gini and Permutation importance normalised to $[0, 1]$ for the feature variables in the Random Forest model without Beauty. The Spearman Correlation between the two sets of Variable Importances is 0.82.

model (language) train the model using the same procedure as before, and then test how the model performs. Other methods we could try for dimensionality reduction include Principal Component Analysis (PCA) or Independent Component Analysis (ICA), however these methods would make the model less interpretable since they transform the features. These methods can potentially increase performance as well as shorten the run time of our model.

Additionally, by the way this sample was gathered we can conclude that it is not truly independent. This is because the researchers first sampled 94 professors and then gathered data over a 2 year period (2000-2002), this sampling scheme resulted in 463 classes, with the number of classes taught by a unique professor in the sample ranging from 1–13. Therefore, it is clear that each class is not independent from every other class since a professor might be teaching in more than one class and in multiple years. This certainly introduces some sort of bias in our model, thus to further improve we would want to gather more high quality data.

References

[1] Nalini Ambady and Robert Rosenthal. "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness." In: *Journal of personality and social psychology* 64.3 (1993), p. 431.

[2] Norman H Anderson. "Primacy effects in personality impression formation using a generalized order effect paradigm." In: *Journal of personality and social psychology* 2.1 (1965), p. 1.

[3] Jeff E Biddle and Daniel S Hamermesh. "Beauty, productivity, and discrimination: Lawyers' looks and lucre". In: *Journal of labor Economics* 16.1 (1998), pp. 172–201.

[4] Stephen Buck and Drew Tiene. "The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations". In: *The Journal of Educational Research* 82.3 (1989), pp. 172–177.

[5] *CHANCE*. `https://chance.amstat.org/2013/04/looking-good/`.

[6] Kahneman Daniel. *Thinking, fast and slow*. 2017.

[7] Linda Darling-Hammond et al. "Evaluating teacher evaluation". In: *Phi Delta Kappan* 93.6 (2012), pp. 8–15.

[8] Karen Dion, Ellen Berscheid, and Elaine Walster. "What is beautiful is good." In: *Journal of personality and social psychology* 24.3 (1972), p. 285.

[9] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning.* Vol. 1. 10. Springer series in statistics New York, 2001.

[10] DS Hamermesh and AM Parker. *Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity (No. NBER Working Paper No. w9853).* 2003.

[11] Eric Hanushek. "Teacher characteristics and gains in student achievement: Estimation using micro data". In: *The American Economic Review* 61.2 (1971), pp. 280–288.

[12] David A Katz. "Faculty salaries, promotions, and productivity at a large university". In: *The American Economic Review* (1973), pp. 469–477.

[13] Sotiris Kotsiantis, Christos Pierrakeas, and Panagiotis Pintelas. "PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES". In: *Applied Artificial Intelligence* 18.5 (2004), pp. 411–426.

[14] David Landy and Harold Sigall. "Beauty is talent: Task evaluation as a function of the performer's physical attractiveness." In: *Journal of Personality and Social Psychology* 29.3 (1974), p. 299.

[15] Richard J Murnane. "The impact of school resources on the learning of inner city children." In: (1975).

[16] Xiaoxia A Newton et al. "Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts". In: *education policy analysis archives* 18 (2010), p. 23.