The beauty of the teacher more important than the beauty of learning ? The hidden side of Student Evaluations of Teaching

Data Science - Coursework 6

CID 01945214, Imperial College London Spring 2021

Abstract

In this study, we aimed to quantify to which extent Random Forests can predict whether a course is appreciated by the students from characteristics of the teacher and the course. We conducted an investigation suggesting that our predictions may be biased by teachers' beauty. A deeper study of other features did not reveal any additional bias or relationship that deny a bias with respect to teachers' beauty. A suspicious of such bias might lead the university administration not to base their decisions for the courses' organization and choice of the professors on students' evaluation.

1 Introduction

Student evaluations of teaching (SETs) constitute a convenient and cheap way to evaluate teaching quality. Nowadays, they are widely used by universities for promotion, tenure, and teaching award. It has been assumed that students are in the best position to know whether the teaching they receive is adequate and whether they are learning. SETs also make the comparison between teachers very simple by comparing the average rating of any instructor to the average of his department. In this study, we aim to quantify to which extent Random Forests can predict whether a course is well evaluated by students from characteristics of the course and the rank of the associated teacher. A huge matter will be to understand how this machine learning model build his predictions and gain intuition on the criteria influencing the way students evaluate their teachers. This analysis of students behaviours could help administration to be aware of the underlying mechanism using SETs to make decisions for the courses' organization and choice of the professors. We will use a data set gathering student evaluations for a large sample of professors from the University of Texas at Austin [1].

The main objective is to use Random Forests as a predictive model of teacher evaluation by students. Random Forests are an ensemble learning method that consists in constructing many decision tree classifiers on various sub-samples of the dataset. Predictions are made averaging the predictions made by each individual tree. An advantage of using Random Forests is that they are able to capture nonlinear relationships, including interactions among the predictors. Random forests are also interesting as they give a measure of the relative importance of each feature in the prediction process. Moreover, several techniques are known to minimize overfitting putting some constraints on the constructed trees [2]. Such a property is particularly important in our problem as we use a quite small data set.

Some previous studies have demonstrated that student evaluations are biased with respect to the teacher physical appearance ([3] and [4]). We will investigate if the predictions made with random forests are biased by teacher beauty. A key underlying challenge is to find the way to evaluate the beauty of teachers, which is of course a very subjective criterion. To estimate the beauty of a given professor from university of Texas, 6 students who are not in a class taught by this professor have been asked to give him a beauty score based on a picture. Doing so, we avoid asking students to rate both attractiveness and teacher quality at the same time with the risk that the former is impacted by the latter. However, there might be concerns using only a picture to evaluate the global beauty of some teacher. Showing that SETs are biased with respect to teacher beauty should raise concerns about the massive use of them by worldwide universities. In particular, it might lead the universities administration not to base their decisions for the courses' organization and choice of the professors only on these SETs.

The section II summarizes existing bibliography dealing with SETs, the way universities use them and their potential biasedness with respect to teachers' beauty. In Section III we present the data set used to fit Random Forests and explore the impact of teachers' beauty on SETs in the data set. In section IV we discuss and interpret the results of Random Forests. We also seek potential biases present in our model. Section V presents a discussion on the validity of our study and the potential implications for universities board.

2 Litterature review

Student evaluations of teaching have been observed to play a key role in institutional decisions. Some research indicated that teaching evaluation are important for academic administrators in setting salaries (William E. Becker and Michael Watts [5]). Decisions concerning tenure and promotion also often take into account the teaching evaluation (Academic college paper [6]). These evaluations are then particularly important in the teachers career development.

Nevertheless, several studies pointed out that teaching evaluations by students are biased with respect to teachers' physical appearance. In 2002, Feeley found a positive correlation between physical attractiveness and teaching effectiveness ratings [3]. He interpreted this positive correlation as an Halo effect that impacts teachers perception from students. Halo effect is a cognitive bias consisting in an interpretation of information excessively in line with a first impression. A characteristic that is seen as positive about a person tends to make other characteristics of that person more positive (and inversely for a characteristic that is seen as negative). It has been empirically observed by Edward Thorndike in 1920 [7] and was demonstrated by Solomon Asch in 1946 [8]. In the context of studies, the interpretation of the halo effect is the following : students have a first impression of teachers based on its physical appearance, and this first impression interferes when they are asked to judge the quality of teaching. This is to say that more beautiful teachers tend to obtain higher evaluation marks from students.

Some research aimed to quantify more precisely this beauty bias in teaching evaluation from students. For this extent, Daniel S. Hamermesh and Amy M. Parker created a data set gathering a large sample of professors evaluation by students from the University of Texas at Austin [1]. This is the data that we will use in our study. Daniel S. Hamermesh and Amy M. Parker used Multiple linear regression to make predictions from the data. It gave them an interpretable model which highlights biases with respect to teachers' beauty. It will be particularly interesting to compare their conclusions with the one that we will draw form our black box model which is random forests.

Mitchell and Felton [4] also examined the relations between teaching quality and sexiness. They worked on a sample of 3,190 professors at 25 universities taken from RateMyProfessors.com, a web page where students anonymously judge their professors on Quality, Easiness, and Sexiness. They found that "the data reveal that students give sexy-rated professors higher Quality and Easiness scores". They use a much larger data set than us, and it will be interesting to see if conclusions drawn. However, the use of data from RateMyProfessors.com has been criticized (Dennis E. Clayson and Mary Jane Sheffet [9]) because of the impartiality in rating attractiveness and teacher quality at the same time.

As we stated before, teacher evaluations by students are very important in teachers career. Highlighting a bias with respect to the teachers' physical appearance would raise many concerns. We must be very careful on the conclusion we make and ensure that an observation of something we consider as an evidence of biasedness is actually one. In particular, we may wonder whether observing that more attractive teachers get better SET is actually an evidence of a bias. Daniel S. Hamermesh and Jeff E. Biddle [10] indicate that better-looking individuals may have more confidence and self-esteem and that these characteristics may also increase productivity in certain occupations. A legitimate question is : is beauty a real factor when students evaluate teacher or beautiful teachers are simply more productive. At this point, we have no reason to support this last hypothesis, but it just underlines that an observation indicating a bias with respect to teachers' beauty should be analyzed with a lot of precaution.

If teachers' beauty is actually observed to influence SETs, universities may need to rethink the use of student opinion surveys as a valid measure of teaching effectiveness. Other approaches need to be seeked to evaluate teacher quality in a fairer way. A recent article from Toni Feder [11] presents a method initiated by the University of Southern California (USC) called "student learning experience evaluations". USC's new student surveys pose such questions as whether course concepts were well explained, whether the instructor encouraged discussion, whether the instructor was receptive to diverse viewpoints... There are also questions more objective : whether the instructor's handwriting is legible, whether the student could hear the instructor, whether the student understood the textbook. To put in a nutshell, the survey from USC aim to ask students about more concrete and objective characteristics of the teacher than a score out of 5.

3 Data used

The data set used gathers student evaluations for a large sample of professors from the University of Texas at Austin. It was constructed from Daniel S. Hamermesh and Amy M. Parker [1] and firstly released as part of the replication data for the book *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill, 2007 [12]). It can be downloaded for replication at `https://www.openintro.org/book/statdata/?data=evals`. Each row consists of a course with its associated professor, and each column has information on either the course or the professor. The researchers first sampled 94 professors, and then collected data on courses they taught over a two-year period between 2000 and 2002. This sampling scheme resulted in 463 classes, with the number of classes taught by a unique professor in the sample ranging from 1 to 13. The observations are not independent from each other because some courses are attended by the same students and/or taught by the same professor. However, we will assume independency to fit our machine learning model. Finally, we notice that the data set does not present any missing value and no preprocessing is needed before fitting Random Forests (normalization is useless for this model).

The quantity we aim to predict is the professor SET score out of 5. It goes from 1/5 which means that the student is very unsatisfied, to 5/5 which means that the student has an excellent opinion of the course. The data set has 20 features. We make the distinction between two classes of features. On one hand, class I features have information only on the course characteristics and teacher's rank. On the other hand, class II features correspond to characteristics of teachers that can be considered as outside the instructor's control and which should not influence the quality of a course. We will use only the class I features to build our first random forests. The class II features will be used to highlight potential biases in our model predictions.

Class I features	Class II features
Rank of the teacher : teaching, tenure track or tenured	Age of professor
Percent of students in class who completed evaluation	Ethnicity of professor : minority or not minority
Number of students in class who completed evaluation	Gender of professor : female or male
Total number of students in class	Language of school where professor received education : English or non-English
Class level: lower or upper	Beauty rating of professor from lower level female: (1) lowest - (10) highest
Number of professors teaching sections in course in sample: single or multiple	Beauty rating of professor from upper level female : (1) lowest - (10) highest
Number of credits of class: one credit (laboratory class, Physical education class, etc.) or multi credit	Beauty rating of professor from second upper level female : (1) lowest - (10) highest
	Beauty rating of professor from lower level male: (1) lowest - (10) highest
	Beauty rating of professor from upper level male: (1) lowest - (10) highest
	Beauty rating of professor from second upper level male: (1) lowest - (10) highest
	Average beauty rating of professor
	Outfit of professor in picture: not formal, formal

Table 1: Features description.

We split our data set in two parts. The training set contains 80% of the observations and will used to fit Random Forests. The remaining 20% observations compose the test set. The test set is unseen by the model and will only be used at the end to assert the performances of our final model. Our data set contains only 463 observations we will then use the following strategy : the training set will be used for hyperparameters selection by out-of-bag estimate and for fitting Random Forest, the test set will be used for both assessing model performances and looking for bias. The data set was shuffled before the split to ensure that the observations in both the training set and test set are representative of the overall distribution of the data. In the following data exploratory, we then use only the training set.

We first have a look at the distribution of teacher evaluations which is the quantity of interest. The distribution score evaluation is skewed with more evaluation between 4 and 5 (figure 1). Very few teachers obtain evaluation below the average score of 3/5. On figure 2, teacher's rank is not observed to have a noticeable influence on the SET score.

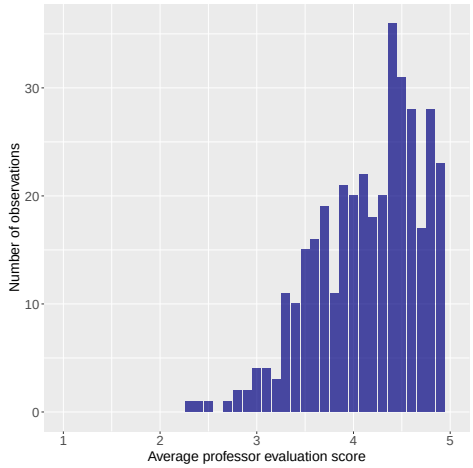

Figure 1: Histogram of SET scores in the training set.

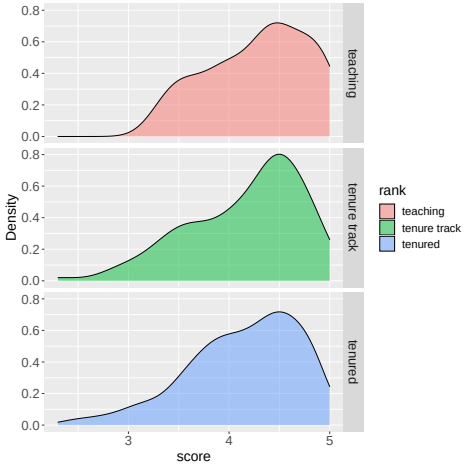

Figure 2: Density of SET scores in the training set conditioning on the rank.

A summary of other features distribution is given table 2 and 3. From this table, it is important to check whether we have a representative sample of students. Looking for the column percentage of students, we can see that on average 74.56% of students filled a SET; and the histogram in figure 3 indicate a strong trend that the majority of students fill the SETs. We can also notice that the number of students varies a lot from 8 students in a small lab session to 579 in a huge amphitheatre. Most courses are multicredits, but the 20 one credit courses all obtain a very high SET score and will be easy to predict for our Random Forests (figure 4).

	SET Score	% students who filled the SET survey	number of students who filled the SET survey	number of students in the class
Min.	2.300	10.42	5.00	8.00
1st Qu.	3.800	62.50	14.50	18.50
Median	4.300	76.92	23.00	29.00
Mean	4.165	74.56	36.56	55.08
3rd Qu.	4.6000	86.96	40.00	59.50
Max	5.000	100.00	372.00	579.00

Table 2: Summary of quantitative features.

Rank of teacher	Class level	Number of professors teaching sections in course in sample	Number of credits of class
teaching : 82 tenure track : 88 tenured : 201	lower : 119 upper : 252	multiple : 238 single : 133	multi credit : 238 one credit : 20

Table 3: Summary of qualitative features.

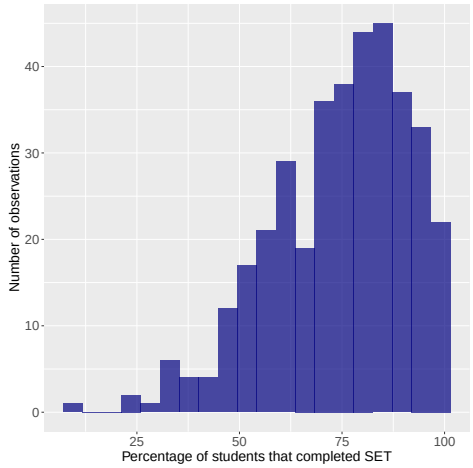

Figure 3: Histogram of percentage of students that completed SET in the training set.

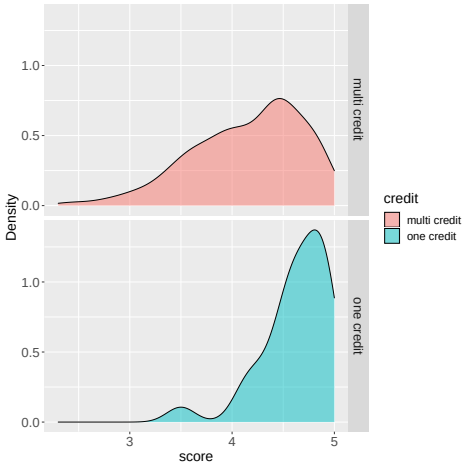

Figure 4: Density of SET scores in the training set conditioning on the class credit.

The Beauty variables are at the heart of our study, it is worth studying it deeper before fitting Random Forests. For a given professor, the beauty score has been collected from 6 students who are not in a class taught by this professor. To make the estimation of the beauty score more accurate, the 6 students sample was composed of 1 female student and 1 male student in lower class as well as 2 female and 2 male students in upper class. Of course this method does not give us an objective beauty score for each teacher, but the advantage is that students are not influenced by the teaching competence and personality of a teacher they already know. The correlation matrix displayed in figure 6 indicates a high correlation above 0.5 for each pairs of beauty score given by students. It suggests that students tend to agree concerning teachers' beauty which in some sense tells us that our measure of beauty should be representative of all students' opinion. Then, we can plot SET scores against the beauty score and observes that a bias appears in favour of the beautiful teachers. This is especially true for male teachers for whom the correlation between SET score and beauty score is higher. The stake of our study will be to determine whether our inference from Random Forests model is influenced by this beauty bias within students.

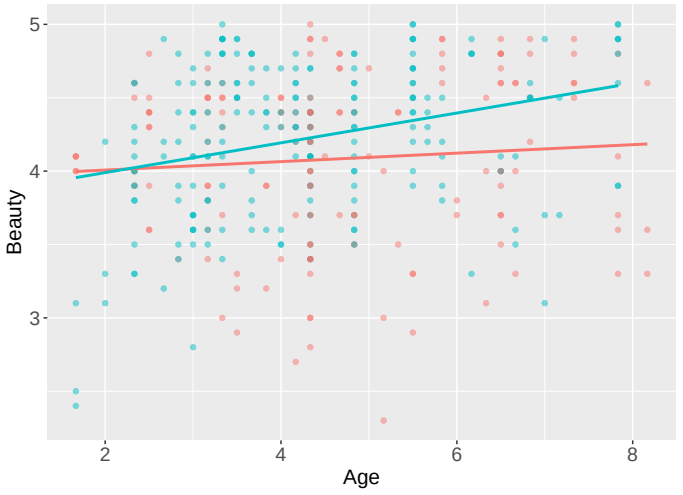

Figure 5: SET scores against the average beauty for teachers in the training set.

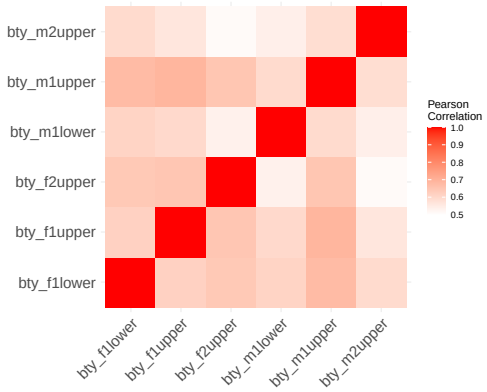

Figure 6: Correlation matrix of the beauty marks given by the 6 students.

4 Results

4.1 Random Forests

We fit Random Forests using as features only the class I features (blue column in table 1). To assert the performances of our model, we evaluate the Mean Absolute Error (MAE) defined by :

$$MAE(\{x_i, y_i\}) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y_i} - y_i|$$

Random Forests allow to minimize overfitting putting some constraint on the decision trees they construct. We will use two different sorts of constraints :

- When splitting a node, only a randomly drawn subset of features of size m_try is considered
- The number of terminal nodes trees in the forest is upper-bounded by an integer max_node

m_try and max_node are hyperparameters that we have to choose. We optimize this choice by doing a cross-validated search over parameter settings. We create a grid with large range of possible values for hyperparameters. We then fit Random Forests for each pair (m_try, max_node) and we compute the out-of-bag error (OOB error). OOB error is the mean prediction error on each training observation using only the trees that did not have this observation in their bootstrap sample. We then select the pair of hyperparameter (m_try, max_node) that obtained the smallest out-of bag error among all fitted model. This method gives us an optimal choice of hyperparameters :

$$m\_try = 2 \qquad max\_node = 60$$

Our final Random Forests obtain a MAE of 0.2945 on the training set and 0.3980 on the test set. we plot the predicted score against the beauty average to investigate a bias with respect to teachers' beauty (figure 7). A simple linear regression makes appear a positive correlation between teacher's average beauty score and SET scores predicted by Random Forests. There is no noticeable distinction between male and female teacher. We can seek a linear relationship between the predicted SET scores and teacher beauty average through a Fisher F-test [13].

Given two quantitative variables X and Y, we consider a linear model of the form $Y = aX + b$. We also consider the null hypothesis $H_0 : a = 0$ against a two-sided alternative $H_1 : a \neq 0$. The Fisher F-test is used to assess whether the model under H_1 fits the data significantly better than the naive model under H_0 by referring the statistic :

$$F = \frac{RSS_0 - RSS_1}{RSS_1} \frac{n-r}{r-s} \tag{4.1}$$

to $F_{1,n-2}$ the F-distributions with degrees of freedom 1 and $n-2$ where n is the number of observations. RSS_0 is the Residual Sum of Squares for the model considered under H_0 ie $Y = b$, and RSS_1 is the Residual Sum of Squares for the linear model considered under H_1, ie. $Y = aX + b$.

We consider the simple linear model $SETscore = a \times BeautyAverage + b$ and apply a Fisher-test. The resulting p-value is equal to 0.039 which means that at a 5% level there is evidence that a is significantly different from 0. It suggests a linear relationship between the beauty average and SET score. However, the p-value is not very significant and we would fail to reject the null hypothesis at a 1% level. We tried restarting the same process of hyperparameters tuning and random forest fitting with several random sampling of training set and test set. Different patterns are obtained when we construct the plot as in figure 7, and the observed positive correlation between Beauty average and SET score may be uniquely due to randomness. A possible explanation is that our training set and test set share some observations (but with possibly different SET scores) as some rows correspond to the same teacher but with different class. Random Forests may overfit these observations, and in this case the model captures mildly the inherent bias with respect to teacher beauty that was present in our initial data. In any case, this is not enough to state that our model is biased by teachers' beauty.

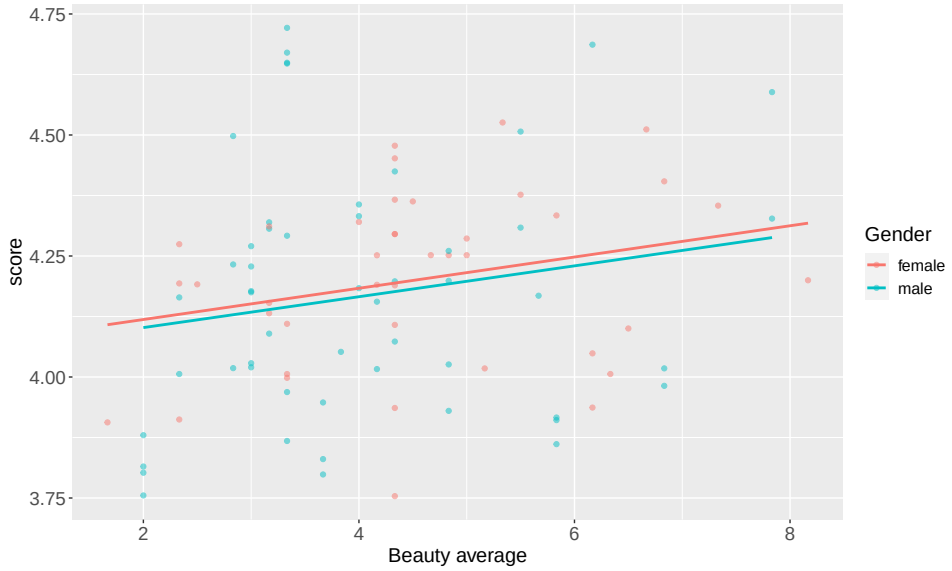

Figure 7: Predicted SET against teacher's beauty for the Random Forest trained on class I features.

We now fit Random Forests adding all features concerning the beauty of teacher : the beauty score given by the 6 students as well as the average of these score. With the same procedure as before, we obtain optimal hyperparameters $m\_try = 2$, $max\_node = 110$. These new variables improve the predictive MAE of 31.6% on the training set and 15.7% on the test set (table 5).

As before, we plot the predicted score against the beauty average to investigate a bias with respect to teachers' beauty. The plot is now very similar to the one obtained on training data (figure 5). A simple linear regression makes appear a positive correlation between beauty average and SET score for the Random Forests predictions.

The correlation is higher for male than for female teachers. A Fisher F-test gives p-values highly significant for the whole sample and when considering only male teachers. This result gives a convincing evidence that beauty score impacts the predicted SET score by our Random Forests.

We may also be interested in Random Forests variables importance to quantify the impact of beauty variables on predictions (figure 9). Feature importance is measured for each feature by collecting how on average it decreases the impurity when splitting a node. Beauty features are observed to be important, and we can especially see that the beauty average.

We expected the beauty variables not to have any impact on random forests predictions for teaching quality. We would also expect these features to have a negligible importance. This is not consistent with our observations, and it is an that Random Forests capture the inherent bias with respect to teachers' beauty that was present in the initial data (figure 5).

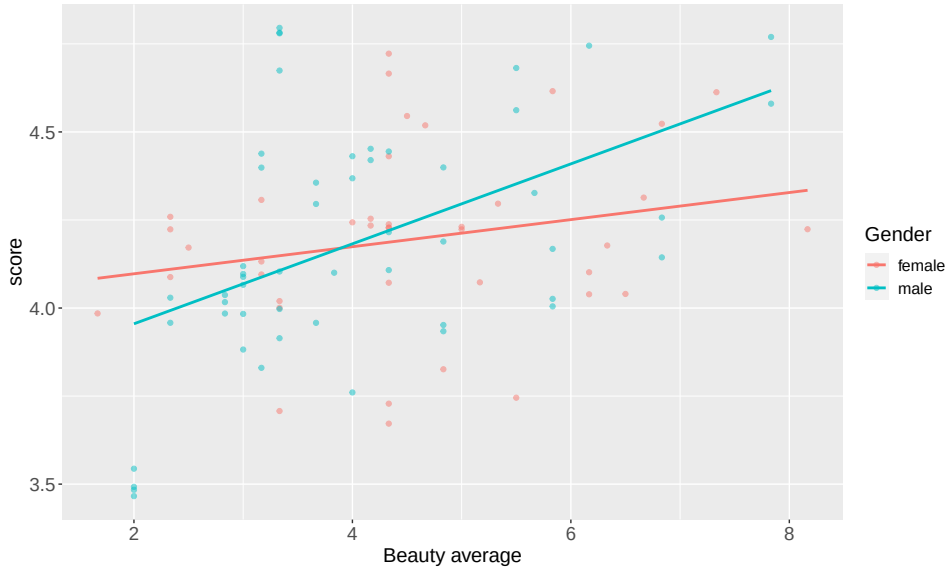

Figure 8: Predicted SET against teacher's beauty for the Random Forest trained on class I and beauty features.

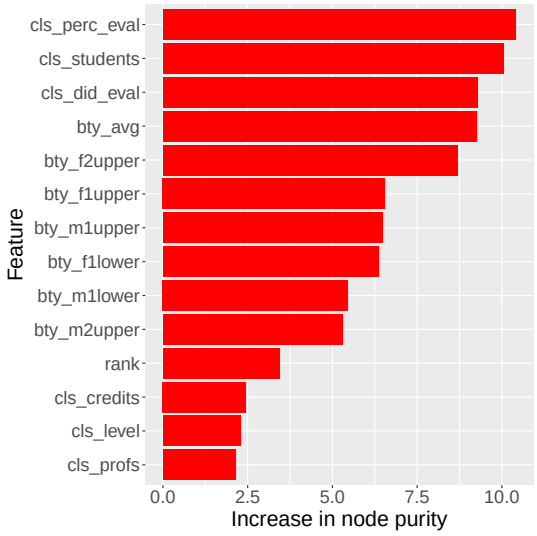

Figure 9: Features importance for Random Forests fitted trained on class I and beauty features.

Model	train MAE	test MAE
Without beauty features	0.2945	0.3980
With beauty features	0.2014	0.3355

Table 4: MAE calculated on training and test set for Random Forests trained on class I and beauty features.

	Global	Male	Female
$\hat{a}$	0.07847	0.11352	0.03845
statistic F	3.891	3.911	1.394
p-value	0.000191	0.000277	0.172

Table 5: Fisher F-test for the predicted SET scores by Random Forest trained on class I and beauty features as a linear function of the average beauty score.

4.2 Correlation between beauty and other variables

We should investigate whether the beauty variable is correlated with other class II features in our data set. It can be imagined a situation where, in our sample, teachers that obtain bad beauty score happen to be the same than the minority part of the population. If these two parts of the population suffer from a bias in their SET

scores, it would be important to nuance the conclusion drawn and not directly conclude that teachers' beauty is directly the source of a manifest bias.

The other class II features (in red column of table 1) except of age are binary categorical variables. For each of these variables, we can divide our sample in two populations conditioning on this variable. We then compare the SET scores distributions for the two populations to determine whether the variable has a non-negligible effect on it (figure 10).

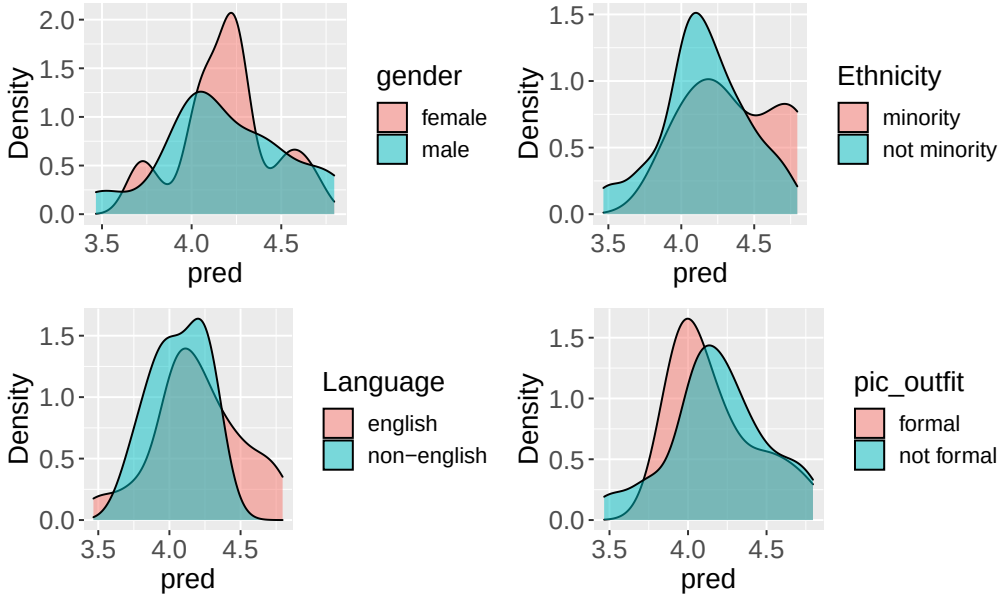

Figure 10: Density curves conditioning on Gender, Ethnicity, Language and Picture Outfit.

We observe slightly different shapes for density curves conditioning on the considered class II variable. It is however globally the same pattern. In statistics, Kolmogorov-Smirnov hypothesis test is used to determine whether two samples are identically distributed [14]. It uses the statistic :

$$D = \sup_x |\hat{F}_1(x) - \hat{F}_2(x)| \tag{4.2}$$

where $\hat{F}_1(x)$ and $\hat{F}_2(x)$ are respectively the empirical distribution functions of the first and the second population. The null hypothesis is rejected at level α if $D > \sqrt{-\log\left(\frac{\alpha}{2}\right)\frac{n+m}{2mn}}$ where n and m are respectively the size of the first population and the second population. We obtain a non-significant p-value at 5 % level when dividing the test set population on each categorical class II features (table 6). It indicates that our predictions do not make appear a bias for these features.

Variable	Gender	Ethnicity	Language	Picture Outfit
statistics D	0.16	0.35	0.35	0.27
p-value	0.566	0.155	0.396	0.428

Table 6: Kolmogorov-Smirnov test to compare distributions of test set observations conditioning on class II features.

It remains to study the dependency between beauty score and teacher age which is quantitative, and whether this latter has an impact on SET score. We can observe a large negative correlation of between the age and the beauty average (figure 11). Students tend to rate older people more severely. Some studies revealed that SET tend to be biased by teachers' age [15]. This is this time a negative correlation : the older the teacher, the lower SET score he obtains. However, we observe only a very slight negative correlation between the age and the predicted SET score with Random Forests with beauty features (figure 12). A Fisher F-test gives p-value equal to 0.369. Then, there is no evidence to say that our predictions are biased with respect to teachers' age.

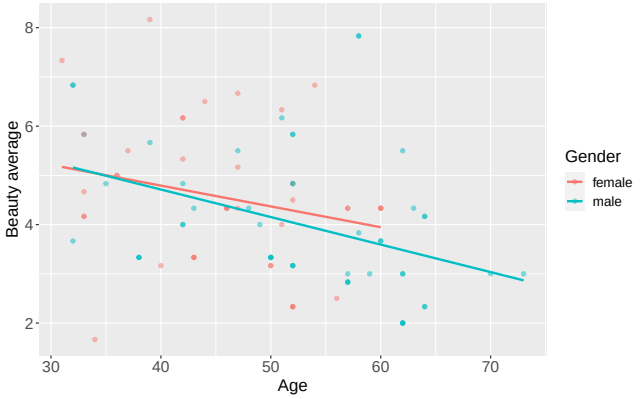

Figure 11: Predicted score of Random Forests trained on class I and beauty features against teacher's age.

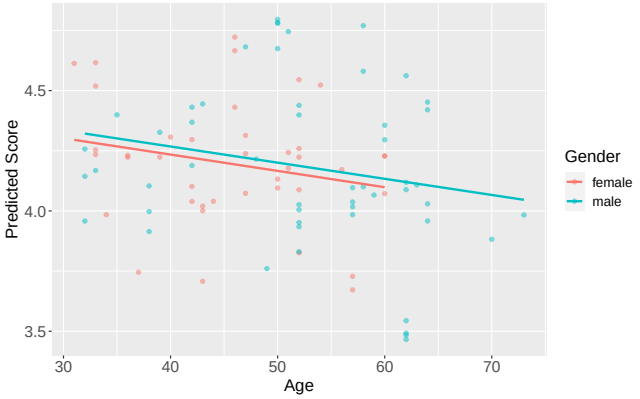

Figure 12: Beauty average against teacher age in training set.

5 Conclusion and discussion

To conclude, we built Random Forests to predict SET scores. We first fitted Random Forests only on features that have information on the course characteristics and on teacher's rank. For this model there was no strong evidence of a bias with respect to teachers' beauty. Adding the beauty features to the training set gave us a model that performed better. The improvement of the model performances, a strictly positive correlation between beauty score and SET score, and the study of variable importance gave us hints that beautiful teachers tend to obtain better SET scores. We also compared the beauty variable with other variables like teacher's age or ethnicity. No argument was found in favour of a bias with respect to one of these variables. Therefore, we detected a potential bias with respect to teachers' beauty but not for other features. However, we cannot deduce a general statement that teachers' beauty has actually an impact on SET scores.

Indeed, there exist many concerns with the beauty variable construction and more globally on the quantification of somebody's beauty. There is not one universal standard of beauty and the students opinion are likely to be different in other university. Using only 6 students to evaluate beauty raises concerns about the generalization of our study. Moreover, beauty is not a fixed factor but can be influenced by other factors such as cosmetics. The use of photographs to evaluate beauty is also questionable because photographs do not represent the full information of attractiveness available to college students enrolled in a class [16]. In particular, rating from photograph gives just an estimation of teachers' *facial* beauty.

We also mention again that our sample has a low size with only information on 94 teachers from the same US university. The generalizability of such a low sample is weak and our model is likely to overfit a lot as some teachers are present in both the training and test set. Another very large issue regarding our data set is the absence of potentially important features that have been observed to be important in teacher evaluations like the subject [17] and the exam difficulty [18].

Concerning the relevancy of our method, let us recall that we restricted ourselves mainly to the study of beauty characteristics and the associated potential bias. A deepest interesting work would consist in adding each feature of class II one-by-one and assess the impact on the model. Illustrate a bias with respect to another feature might change our conclusions. This is something that I would have liked to do if I had more time and a larger number of pages limit. I would also have liked to benefit from a larger data set with more features like the subject, the male/female students percentage and the examination difficulty.

Finally, our study illustrated that it is important to be careful with biases when training predictive models. Data gathered from empirical observations on human population are likely to capture inherent biases that are present in our society. One should be aware of this and make relevant checks before drawing conclusions on human populations from the predictive model. It is especially true when using black box models such as Random Forests or Neural Networks. Such models have become very popular for their capacity to predict accurately, but the lack of transparency on the way they handle features to predict make them hard to interpret. They are susceptible to capture biases without us realizing it, and this should be checked with a careful study after model fitting.

References

[1] Daniel Hamermesh and Amy M Parker. Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. Working Paper 9853, National Bureau of Economic Research, July 2003.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[3] Thomas Feeley. Evidence of halo effects in student evaluations of communication instruction. *Communication Education - COMMUN EDUC*, 51:225–236, 07 2002.

[4] James Felton, John Mitchell, and Michael Stinson. Web-based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. *Assessment Evaluation in Higher Education*, 29:91–108, 06 2003.

[5] William E. Becker and Michael Watts. How departments of economics evaluate teaching. *The American Economic Review*, 89(2):344–349, 1999.

[6] Henry A. Hornstein. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, 4(1):1304016, 2017.

[7] E. L. Thorndike. A constant error in psychological ratings., January 1920.

[8] S.E. Asch. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258–290, 1946.

[9] Dennis Clayson and Mary Sheffet. Personality and the student evaluation of teaching. *Journal of Marketing Education - J Market Educ*, 28:149–160, 08 2006.

[10] Daniel S. Hamermesh and Jeff E. Biddle. Beauty and the labor market. *The American Economic Review*, 84(5):1174–1194, 1994.

[11] Toni Feder. Reevaluating teacher evaluations in higher education. *Physics Today*, 73:24–27, 1 2020.

[12] Andrew. Gelman. *Data analysis using regression and multilevel/hierarchical models*. Analytical methods for social research. Cambridge University Press, Cambridge, 2007.

[13] 1897 Goulden, Cyril Harold. *Methods of statistical analysis*. Wiley, New York, 2nd ed. edition, 1959.

[14] Marvin Karson. Handbook of methods of applied statistics. volume i: Techniques of computation descriptive methods, and statistical inference. volume ii: Planning of surveys and experiments. i. m. chakravarti, r. g. laha, and j. roy, new york, john wiley; 1967, $9.00. *Journal of the American Statistical Association*, 63(323):1047–1049, 1968.

[15] Julianne Arbuckle and Benne Williams. Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, 49:507–516, 11 2003.

[16] Nathan E. Gonyea, Marie Osick, and E. Bradley. An evaluation of the relationship between instructor appearance and college student evaluations of teaching. *Journal of Research in Education*, 28:66–92, 2018.

[17] Michela Ponzo and Vincenzo Scoppa. The good, the bad, and the ugly: Teaching evaluations, beauty and abilities. 01 2012.

[18] Meghan Millea and Paul W. Grimes. Grade expectations and student evaluation of teaching. *College Student Journal*, 36:582+, 2002. M1: 4; M2: 582; 25.

A R code

```
##### Coursework final Data Science #####

setwd("C:/Users/robin/Dropbox/Applications/Overleaf/Coursework_DS_6")

library(ggplot2)
library(corrplot)
library(dbplyr)
library(tidyverse)
library(randomForest)
library(xtable)
library(gridExtra)
library(reshape2)

my_theme =
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        axis.title = element_text(size = 15),
        axis.text = element_text(size = 14),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 12),
        strip.text = element_text(size = 14))

### load and process data ###

data_raw = read.csv("evals.csv")
for (col_name in colnames(data_raw)){
  if (class(data_raw[,col_name]) == "character") {
    data_raw[,col_name] = as.factor(data_raw[,col_name])
  }
}

p = dim(data_raw)[2]
n = dim(data_raw)[1]

### Delete variables bias ###

features_bias = c("age",
                  "gender",
                  "ethnicity",
                  "language",
                  "bty_f1lower", "bty_f1upper", "bty_f2upper", "bty_m1lower", "bty_
                      m1upper", "bty_m2upper", "bty_avg",
                  "pic_outfit", "pic_color")

features_bias_other_bty = c("age",
                            "gender",
                            "ethnicity",
                            "language",
                            "pic_outfit", "pic_color")

data = data_raw[, ! names(data_raw) %in% features_bias, drop = F]
data_bty = data_raw[, ! names(data_raw) %in% features_bias_other_bty, drop = F]
sum(is.na(data))  # No NAs

## split the examples into training and test

set.seed(42)
TEST_SIZE = 0.2

test_indices = sample(1:nrow(data), size=as.integer(TEST_SIZE*nrow(data)), replace=FALSE
    )
```

```
data_train = data[-test_indices,]
data_test = data[test_indices,]
data_train_bty = data_bty[-test_indices,]
data_test_bty = data_bty[test_indices,]
data_train_comp = data_raw[-test_indices,]
data_test_comp = data_raw[test_indices,]

x_train = data_train[,2:dim(data)[2]]
y_train = data_train[,1]
x_test = data_test[,2:dim(data)[2]]
y_test = data_test[,1]

x_train_bty = data_train_bty[,2:dim(data_bty)[2]]
y_train_bty = data_train_bty[,1]
x_test_bty = data_test_bty[,2:dim(data_bty)[2]]
y_test_bty = data_test_bty[,1]

summary(data_train)

##### Data vizualisation #####

# histogram score

pdf("hist_score.pdf")
ggplot(data_train) +
  geom_bar(aes(x = score), fill = 'navyblue', alpha = 0.7) +
  scale_x_continuous(name = "Average professor evaluation score", limits=c(1, 5)) +
  scale_y_continuous(name = "Number of observations") +
  my_theme
dev.off()

# Rank

pdf("rank.pdf")
ggplot(data, aes(x = score, fill = rank)) +
  geom_density(adjust = 1, alpha = 0.5) +
  facet_grid(rank ~ .) +
  labs(fill = "rank",
       y = "Density") +
  my_theme
dev.off()

# histogram percentage

pdf("hist_perc.pdf")
ggplot(data_train) +
  geom_histogram(aes(x = cls_perc_eval), fill = 'navyblue', alpha = 0.7, bins=20) +
  scale_x_continuous(name = "Percentage of students that completed SET") +
  scale_y_continuous(name = "Number of observations") +
  my_theme
dev.off()

# credit

pdf("credit.pdf")
ggplot(data, aes(x = score, fill = cls_credits)) +
  geom_density(adjust = 1, alpha = 0.5) +
  facet_grid(cls_credits ~ .) +
  labs(fill = "credit",
       y = "Density") +
  my_theme
dev.off()

# beauty vs score
```

```
pdf("beauty_age.pdf", width = 8, height = 5)
ggplot(data_train_comp, aes(x = bty_avg, y = score, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "Age",
       y = "Beauty",
       color = "Gender") +
  my_theme
dev.off()


# Students agree on beauty ? #

# male vs female

data_train_comp$bty_mavg = sapply(1:n, function(i) (1/3)*(data_train_comp$bty_m1lower[i]
    + data_train_comp$bty_m1upper[i] + data_train_comp$bty_m2upper[i]))
data_train_comp$bty_favg = sapply(1:n, function(i) (1/3)*(data_train_comp$bty_f1lower[i]
    + data_train_comp$bty_f1upper[i] + data_train_comp$bty_f2upper[i]))

pdf("marks_fem_mal.pdf")
ggplot(data_train_comp, aes(x = bty_mavg, y = bty_favg, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "beauty_score_from_male_students",
       y = "beauty_score_from_female_students",
       color = "Gender") +
  my_theme + geom_abline(intercept = 0, slope = 1, lty = 2)
dev.off()

# every student again each other

cormat = cor(data_raw[,13:18])
melted_cormat <- melt(cormat)
head(melted_cormat)

pdf("cormat.pdf")
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "white", high = "red",
                       midpoint = 0.5, limit = c(0.5,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 15, hjust = 1)) +
  coord_fixed() + ylab("") + xlab("") +
  theme(axis.text=element_text(size=15),
        axis.title=element_text(size=17))
dev.off()


######### Random forests ##########

mae = function(y_true, y_obs) mean(abs(y_true-y_obs))

set.seed(42)
data_forest_test = data

# grid over which we will perform the hyperparameter search:
hparam_grid = as.data.frame(expand.grid(mtry=seq(1, 7, by=1), maxnodes=seq(10, 100, by
    =10)))

# to store the OOB estimates of the MSE
```

```
oob_maes = rep(0.0, nrow(hparam_grid))

# perform the gridsearch
for(hparam_idx in 1:nrow(hparam_grid)) {
  # train candidate model
  this_mtry = hparam_grid[hparam_idx, 1]
  this_maxnodes = hparam_grid[hparam_idx, 2]
  rf = randomForest(x_train,
                    y_train,
                    mtry=this_mtry,
                    maxnodes=this_maxnodes)

  # calculate OOB MAE
  oob_maes[hparam_idx] <- mae(y_train, predict(rf))
}

# select the best model (which has the minimum OOB MSE)
best_hparam_set <- hparam_grid[which.min(oob_maes),]

# train a model on the whole training set with the selected hyperparameters
rf_final <- randomForest(x_train, y_train,
                         mtry=best_hparam_set$mtry,
                         maxnodes=best_hparam_set$maxnodes,
                         importance=TRUE,
                         ntree = 1000)

# the test performance of the final model
yhat_train <- predict(rf_final, newdata=x_train)
yhat_test <- predict(rf_final, newdata=x_test)

# MAE
mae(yhat_train, y_train); mae(yhat_test, y_test)

data_forest_test = data_raw[test_indices,]
data_forest_test$pred = yhat_test

######### Bias beauty random forest ##########

##### beauty #####

pdf("beauty_reg_RF.pdf", width = 10, height = 6)
ggplot(data_forest_test, aes(x = bty_avg, y = pred, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(formula = y ~ x, method = lm, se = FALSE) +
  labs(x = "Beauty_average",
       y = "score",
       color = "Gender") +
  my_theme
dev.off()

data_rf_with_bty = data_forest_test[,c("bty_avg","pred", "gender", "age")]
head(data_rf_with_bty)

mylm = lm(pred~bty_avg, data_forest_test)
summary(mylm)

data_bty_male = data_forest_test[which(data_forest_test$gender=="male"), c("bty_avg", "
    pred", "gender")]
mylm2 = lm(pred ~ bty_avg, data_bty_male)
summary(mylm2)

data_bty_female = data_forest_test[which(data_forest_test$gender=="female"),c("bty_avg",
    "pred", "gender")]
mylm3 = lm(pred~bty_avg, data_bty_female)
```

```
summary(mylm3)


######## Random Forests with beauty features ##########

set.seed(42)

# grid over which we will perform the hyperparameter search:
hparam_grid = as.data.frame(expand.grid(mtry=seq(2, 13, by=1), maxnodes=seq(10, 200, by
    =10)))

# to store the OOB estimates of the MSE
oob_maes = rep(0.0, nrow(hparam_grid))

# perform the gridsearch
for(hparam_idx in 1:nrow(hparam_grid)) {
  # train candidate model
  this_mtry = hparam_grid[hparam_idx, 1]
  this_maxnodes = hparam_grid[hparam_idx, 2]
  rf_bty = randomForest(x_train_bty,
                        y_train_bty,
                        mtry=this_mtry,
                        maxnodes=this_maxnodes)

  # calculate OOB MSE
  oob_maes[hparam_idx] <- mae(y_train_bty, predict(rf_bty))
}

# select the best model (which has the minimum OOB MSE)
best_hparam_set <- hparam_grid[which.min(oob_maes),]

# train a model on the whole training set with the selected hyperparameters
rf_final_bty <- randomForest(x_train_bty, y_train_bty,
                             mtry=best_hparam_set$mtry,
                             maxnodes=best_hparam_set$maxnodes,
                             importance=TRUE,
                             ntree = 2000)

# the test performance of the final model
yhat_train_bty <- predict(rf_final_bty, newdata=x_train_bty)
yhat_test_bty <- predict(rf_final_bty, newdata=x_test_bty)

# MAE
mae(yhat_train_bty, y_train); mae(yhat_test_bty, y_test)

data_bty_forest_test = data_raw[test_indices,]
data_bty_forest_test$pred = yhat_test_bty

## features importance ##

pdf("feat_imp.pdf")
feat_imp = data.frame(rf_final_bty$importance) %>% arrange(desc(IncNodePurity))
feat_imp = subset(feat_imp, select=2)
feat_imp$feature = rownames(feat_imp)
feat_imp$feature = factor(feat_imp$feature, levels = rev(feat_imp$feature))
ggplot(feat_imp, aes(x = feature, y=IncNodePurity)) +
  geom_col(fill='red') + coord_flip() + theme(axis.text=element_text(size=17),
                                              axis.title=element_text(size=20)) +
  ylab("Increase_in_node_purity") + xlab("Feature")
dev.off()


## beauty ##
```

```
pdf("beauty_reg_RF_bty.pdf", width = 10, height = 6)
ggplot(data_bty_forest_test, aes(x = bty_avg, y = pred, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(formula = y ~ x, method = lm, se = FALSE) +
  labs(x = "Beauty_average",
       y = "score",
       color = "Gender") +
  my_theme

dev.off()


data_rf_with_bty = data_bty_forest_test[,c("bty_avg","pred", "gender")]
head(data_rf_with_bty)

mylm = lm(pred~bty_avg,data_rf_with_bty)
summary(mylm)

data_bty_male = data_rf_with_bty[which(data_rf_with_bty$gender=="male"), c("bty_avg", "
    pred", "gender")]
mylm2 = lm(pred ~ bty_avg, data_bty_male)
summary(mylm2)

data_bty_female = data_rf_with_bty[which(data_rf_with_bty$gender=="female"), c("bty_avg"
    ,"pred", "gender")]
mylm3 = lm(pred ~ bty_avg, data_bty_female)
summary(mylm3)



########## Biases study RF final with beauty ############

##### Gender #####

g_gender = ggplot(data_bty_forest_test, aes(x = pred, fill = gender)) +
  geom_density(adjust = 1, alpha = 0.5) +
  labs(fill = "gender",
       y = "Density")  +
  my_theme

ks.test(data_bty_forest_test$pred[which(data_bty_forest_test$gender=="male")],
        data_bty_forest_test$pred[which(data_bty_forest_test$gender=="female")])

##### Ethnicity #####

g_eth = ggplot(data_bty_forest_test, aes(x = pred, fill = ethnicity)) +
  geom_density(adjust = 1, alpha = 0.5) +
  labs(fill = "Ethnicity",
       y = "Density")  +
  my_theme

ks.test(data_bty_forest_test$pred[which(data_bty_forest_test$ethnicity=="minority")],
        data_bty_forest_test$pred[which(data_bty_forest_test$ethnicity=="not_minority")
            ])


##### Language #####

g_lang = ggplot(data_bty_forest_test, aes(x = pred, fill = language)) +
  geom_density(adjust = 1, alpha = 0.5) +
  labs(fill = "Language",
       y = "Density")  +
  my_theme
```

```
ks.test(data_bty_forest_test$pred[which(data_bty_forest_test$language=="english")],
        data_bty_forest_test$pred[which(data_bty_forest_test$language=="non-english")])

##### Picture ######

g_pic = ggplot(data_bty_forest_test, aes(x = pred, fill = pic_outfit)) +
  geom_density(adjust = 1, alpha = 0.5) +
  labs(fill = "pic_outfit",
       y = "Density")  +
  my_theme

ks.test(data_bty_forest_test$pred[which(data_bty_forest_test$pic_outfit=="formal")],
        data_bty_forest_test$pred[which(data_bty_forest_test$pic_outfit=="not_formal")])

# plot

pdf("grid_bias.pdf", width = 10/1.3, height = 6/1.3)
grid.arrange(g_gender, g_eth, g_lang, g_pic, ncol = 2)
dev.off()

##### Age #####

pdf("age_bty_rf_bty.pdf", width = 8, height = 5)
ggplot(data_bty_forest_test, aes(x = age, y = bty_avg, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(formula = y ~ x, method = lm, se = FALSE) +
  labs(x = "Age",
       y = "Beauty_average",
       color = "Gender") +
  my_theme
dev.off()

pdf("age_pred_rf_bty.pdf", width = 8, height = 5)
ggplot(data_bty_forest_test, aes(x = age, y = pred, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(formula = y ~ x, method = lm, se = FALSE) +
  labs(x = "Age",
       y = "Predicted_Score",
       color = "Gender") +
  my_theme
dev.off()

mylm = lm(pred~age,data_rf_with_bty)
summary(mylm)
```