- Homework 2 Spec sheet -

Please load and use the “techSalaries2017.csv” data file. This dataset contains self-reported information on over 62,000 workers in the US tech industry in 2017.

The first row represents the column header. Each row after that represents the information of one person.

Columns represent (in order):

1) Company where they work
2) Job title
3) Office location
4) Total annual compensation (in $)
5) Base salary (in $)
6) Value of stock grants (in $)
7) Bonus payments (in $)
8) Years of relevant experience (in years)
9) Time with this company (in years)
10) Gender (self-reported)
11) Terminal Degree is Masters (1 = yes)
12) Terminal Degree is Bachelors (1 = yes)
13) Terminal Degree is Doctorate (1 = yes)
14) Terminal Degree is High School (1 = yes)
15) Terminal Degree is some college (1 = yes)
16) Self-identifies as Asian (1 = yes)
17) Self-identifies as White (1 = yes)
18) Self-identifies as Multi-Racial (1 = yes)
19) Self-identifies as Black (1 = yes)
20) Self-identifies as Hispanic (1 = yes)
21) Race as a qualitative variable
22) Education as a qualitative variable
23) Age (in years)
24) Height (in inches)
25) Zodiac sign (Tropical calendar, 1 = Aries, 12 = Pisces, with everything else in between)
26) SAT score
27) GPA

We/you will want to use most of these variables in prediction models.

This data is self-reported, so sometimes it will be missing if the person (for whatever reason) did not provide this information. For instance, the information on education (variables 11-15) is only meaningfully interpretable for any given row, if the corresponding value in variable 22 is not “NA”. NA indicates missing data. The same is true for variables 16 to 21.

Mission command approach: As per §4.5 of the Sittyba, we will tell you what to do ("answer these questions"), not how to do it. That is up to you. However, we want you to:

a) Do the homework yourself. Do not copy answers from someone else.
b) Restrict your methods (for now) to what was covered in the lecture/lab (in other words, linear regression, regularized regression and logistic regression)
c) Include the following elements in your answer (so we can grade consistently):

Each answer should contain these elements:

1) A brief statement (~paragraph) of what was done to answer the question (narratively explaining what you did in code to answer the question, at a high level).
2) A brief statement (~paragraph) as to why this was done (why the question was answered in this way, not by doing something else. Some kind of rationale as to why you did x and not y or z to answer the question – why is what you did a suitable approach?).
3) A brief statement (~paragraph) as to what was found. This should be as objective and specific as possible – just the results/facts. Do make sure to include numbers and a figure (=a graph or plot) in your statement, to substantiate and illustrate it, respectively.
4) A brief statement (~paragraph) as to what you think the findings mean. This is your interpretation of your findings and should answer the original question.

Note: Brief actually means "brief". There is no need to write a dissertation. There is value to being concise. A couple of pages should be sufficient for the entire report. Do – however – write a report. A data and code-dump is not very useful or valuable in practice. People who pay you so they can ask you questions usually want them answered.

Please answer the following questions in your report:

1. Using multiple linear regression: What is the best predictor of total annual compensation, how much variance is explained by this predictor vs. the full multiple regression model?

2. Using ridge regression to do the same as in 1): How does the model change or improve compared to OLS? What is the optimal lambda?

3. Using Lasso regression to do the same as in 1): How does the model change now? How many of the predictor betas are shrunk to exactly 0? What is the optimal lambda now?

4. There is controversy as to the existence of a male/female gender pay gap in tech job compensation. Build a logistic regression model (with gender as the outcome variable) to see if there is an appreciable beta associated with total annual compensation with and without controlling for other factors.

5. Build a logistic regression model to see if you can predict high and low pay from years of relevant experience, age, height, SAT score and GPA, respectively.

Extra credit:

a) Is salary, height or age normally distributed? Does this surprise you? Why or why not?
b) Tell us something interesting about this dataset that is not already covered by the questions above and that is not obvious.

Hints / Suggestions / Clarifications

1) The qualitative variables (1-3, like “company”) are included for the sake of completeness and do not have to be included in prediction models. Restrict those models to quantitative data only.
2) Do *not* include variables 5, 6 or 7 in the prediction model of total annual compensation. Obviously, you can predict total annual compensation from a linear combination of base salary, stock options and bonuses. Including it would be like predicting age in years from age in seconds. The reason these variables are included here is because this is a primary target for interesting findings for the 2nd extra credit question.
3) Variables 11 to 15 decode variable 22 into binary “dummy variables” to save you some time. You can include these dummy variables in your prediction model like any other variable. The associated beta would then – for instance – the linear effect of having a PhD as a terminal degree (vs. not). You don’t separately have to include variable 22 in your prediction model (as this information is already contained in variables 11 to 15). However, there are two important things to consider: If variables 11 to 15 are all 0, it does not mean that this person has no education at all. It means that no information on education was provided (“NA” in variable 22). However, Python doesn’t know that, so if you include these dummy variables in your model, you need to restrict yourself to datasets that provide this information (otherwise, Python will not interpret the missing information as “no information”, but as “no education”, which is misleading). Also, note that if you do restrict yourself to cases where this information exists, you might want to leave one dummy variable out, as the model is otherwise overdetermined – for instance, if variables 11 to 14 are all 0, variable value 15 has to be 1, otherwise the data point wouldn’t even be included in the analysis. All of these points about the education variables and dummy variables also apply variables 16 to 21, but for race.
4) In order to answer question 4, you will need to code the qualitative gender identities as numerical categories (e.g. 0 and 1 or 1 and 2). As the controversy pertains to male/female gender disparities specifically, it is ok to restrict your analysis to these two categories for the purposes of answering this question.
5) To answer question 5, you will need to create a new outcome variable (high vs. low earner) by median split (1 = worker makes more than the median annual salary, 0 = worker makes less than the median annual salary), so that your logistic regression model has a categorical outcome variable. You can use the predictors as is, as they are already quantitative.
6) For questions 2 and 3, make sure to suitably split your data into training and test sets, in particular for the hyperparameter tuning.
7) Make sure to actually answer the questions in your report. Contrary to popular belief, numbers do not speak for themselves (data or otherwise, e.g. parameters)
8) Feel free to decode the 3 categorical variables into quantitative categories to look for interesting insights for the 2nd extra credit question, but be mindful that this will be challenging, unless you can restrict your analysis to a few categories that have a large number of data points associated with them (e.g. “software engineer” or large tech companies with many employees).