FML Homework 2 - Erin Choi

Data Cleaning and Preprocessing

After loading the data, I checked for abnormal non-null values in each column with value_counts(). One unit had a gender value of "Title: Senior Software Engineer," so I changed this value to null before dropping all rows with null values for race, education, or gender. Null-dropping was done to restrict my analysis to data that provides information for these columns since I believed these features are all important predictors. I dropped the categorical race and education columns, then one-hot encoded the gender and zodiac columns and dropped the original categorical columns as well. I then standardized all the continuous numerical features using StandardScaler() since these features have different scales and are measured in very different units, which may affect how weights are fitted to the features in regression models later on.

Question 1

I created a new dataset based on the cleaned and standardized data from above by dropping qualitative variables as well as predictors that would be overpowering when predicting total annual compensation (base salary, stock grant value, and bonus). Among the remaining predictors, I determined the single best predictor of total yearly compensation by obtaining correlations between the target and each predictor, since correlation quantifies how strong a linear relationship is between two variables. Before performing regressions, I dropped some dummy variables to prevent overdetermination in the models; one variable was dropped for each of the education, race, gender, and zodiac features. I then split the data into training and test sets and fitted two linear regressions.

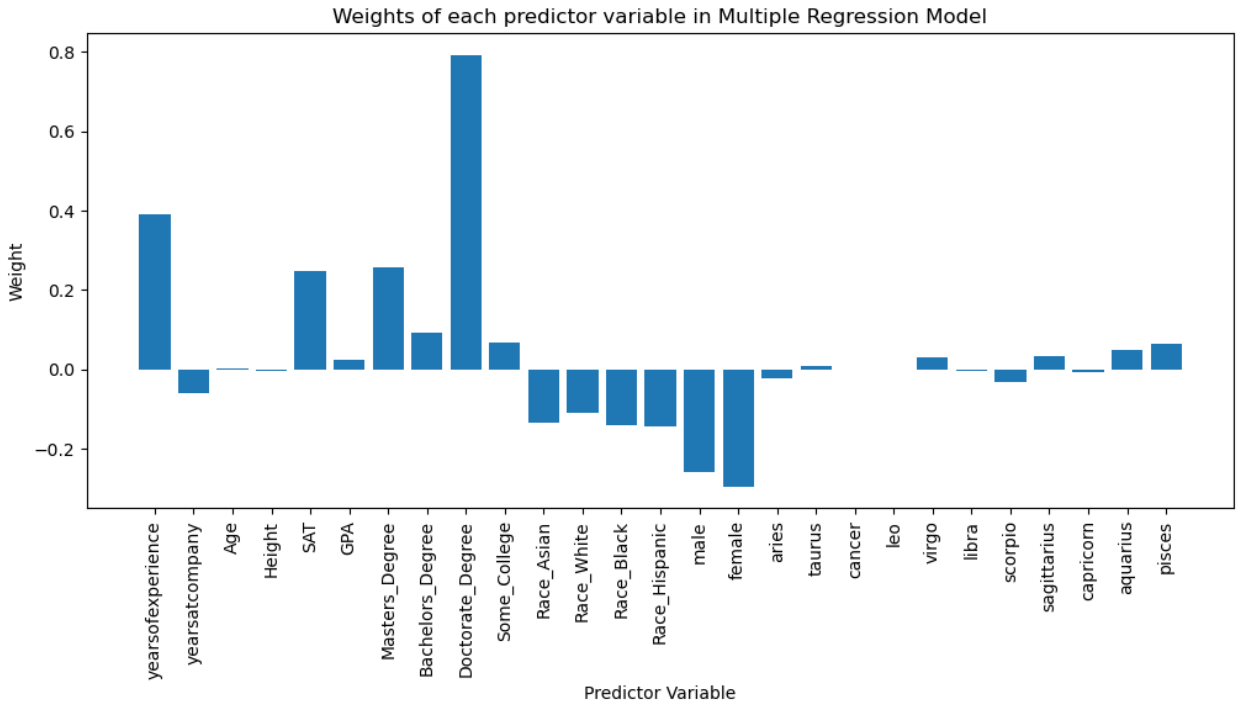
Years of experience was most highly correlated with annual compensation, as the correlation (0.403) had the highest absolute value among all correlations found, so **years of experience is the best predictor.** It makes sense that years of experience is a good predictor; people with more professional experience generally get paid or compensated more for their work, since more time spent doing a certain kind of work leads to greater expertise.

The simple linear regression using just the best predictor (years of experience) yielded an R^2 of 0.1725 and an RMSE of 0.8565. The R^2 value indicates that about **17.25% of the variance in total annual compensation can be explained by variance in years of experience**, which is not a large amount. Meanwhile, the multiple linear regression produced an R^2 of 0.2958 and an RMSE of 0.7901. Again, the R^2 value here indicates that about **29.58% of the variance in total annual compensation can be explained by the variance in all predictors**. This is also not a large amount of variance explained by the predictors, but it is almost twice the proportion explained by the model using the single best predictor.

The RMSE also decreases from the single-predictor linear regression to the multiple regression, showing that the multiple regression's predictions are more accurate than the years-of-experience regression's on average. **Thus the combination of all predictors predicts total annual compensation better than years of experience alone.**

Additionally, I wanted to observe how important years of experience was in the multiple regression model in comparison to the single-predictor model. The weight of years of experience in the initial model was about 0.407, and its weight in the multiple regression was very similar at about 0.390. According to the chart below, it remained an important predictor in the multiple regression model as the

feature with the second-highest coefficient, with the most important predictor being whether a worker has a doctorate degree or not.



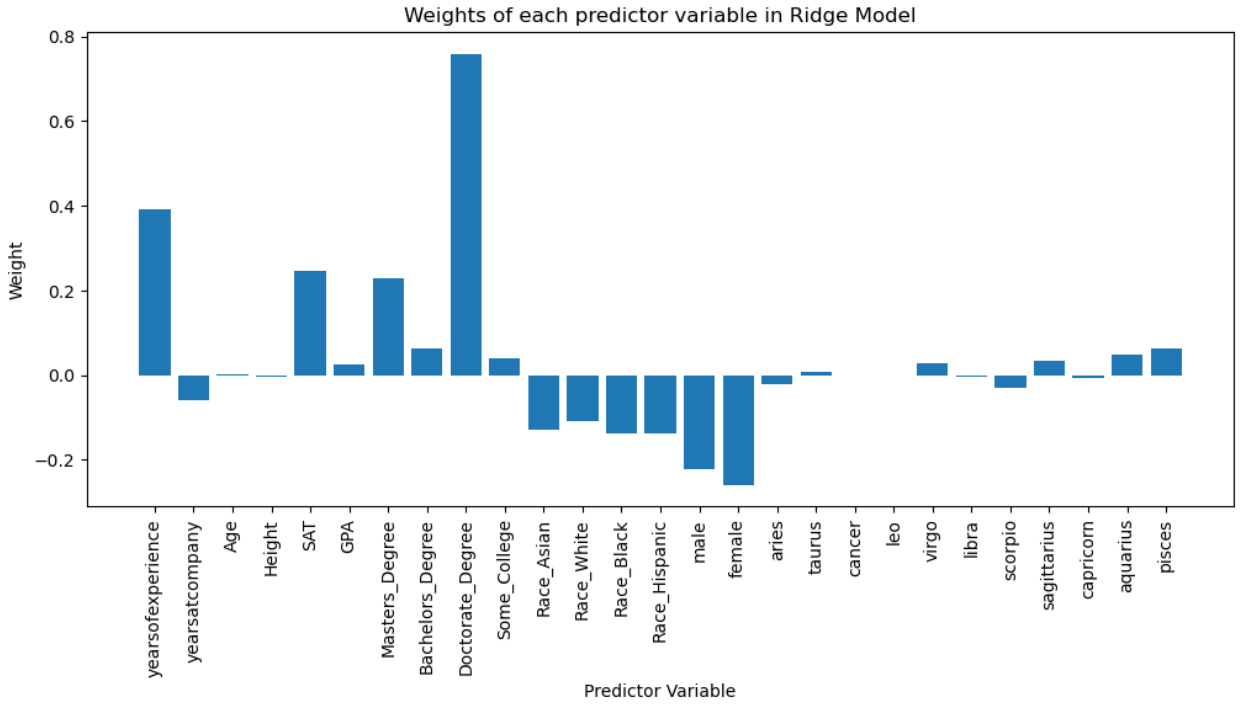
Question 2

Using the same data as in Question 1, I performed a ridge regression, looking for the optimal lambda by using GridSearchCV(). I changed the grid range and step size several times to various values to find the optimal parameter. I also plotted the weights of each predictor in the best model.

I initially used a sci-kit learn implementation of ridge model hyperparameter tuning alongside grid search, which resulted in a different optimal lambda, but I chose to stick with the grid search method as the resulting model's performance was slightly better, based on the R^2 and RMSE metrics.

Grid search found that the **optimal lambda was 5.87**, and the ridge model using this parameter produced an R^2 of 0.2959 and an RMSE of 0.7900. The ridge model's performance was equivalent to that of the multiple regression model, which had an R^2 of 0.2958 and an RMSE of 0.7901.

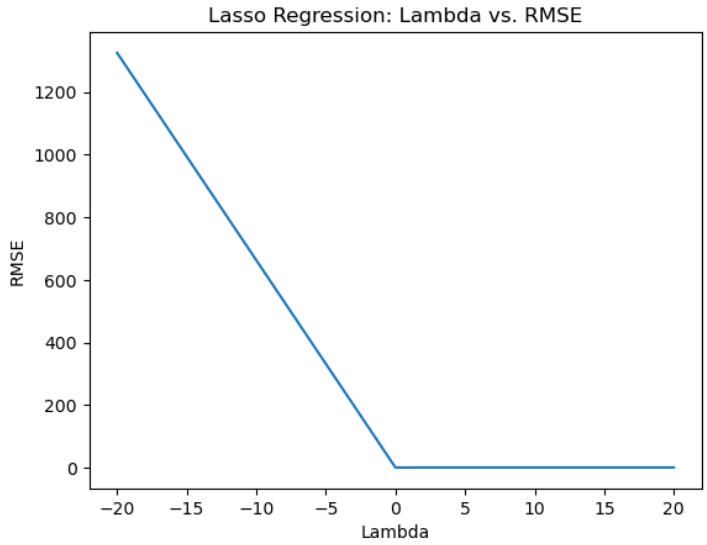
Looking at the weights of the predictors in the ridge model (see the chart below) does show that **the betas of education-related variables have all shrunk a little from OLS** (see the chart in Question 1), but, again, **this did not lead to an improvement from OLS.**



Question 3

Again using the same data, I performed a Lasso regression, looking for the optimal lambda by using GridSearchCV(). I changed the grid multiple times to identify the optimal parameter.

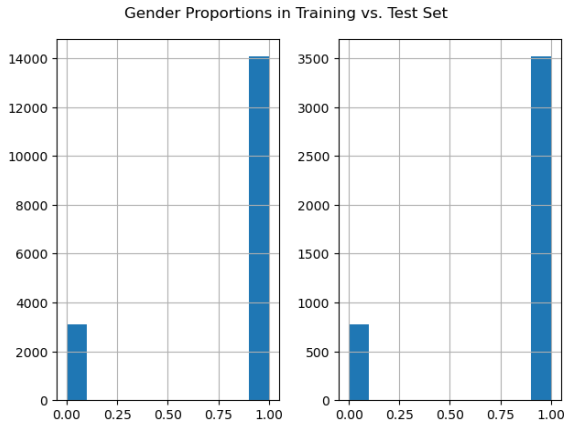
Interestingly, the grid search found that **the optimal lambda was 0**. I used the sci-kit learn implementation of hyperparameter tuning to see if it would result in a different value, but it also outputted an optimal parameter of 0 and a lambda to RMSE graph (see below) that looks like a hinge loss function; after reaching a lambda of 0, all positive lambda values have the same RMSE value.



A lambda value of 0 for Lasso regression is equivalent to using OLS, so there is **no change to the model in comparison to the multiple regression model** in Question 1. Thus the R^2 for this Lasso model is also 0.2958, and the RMSE is also 0.7901. This also means **none of the betas have shrunk to zero** in this "Lasso" model.

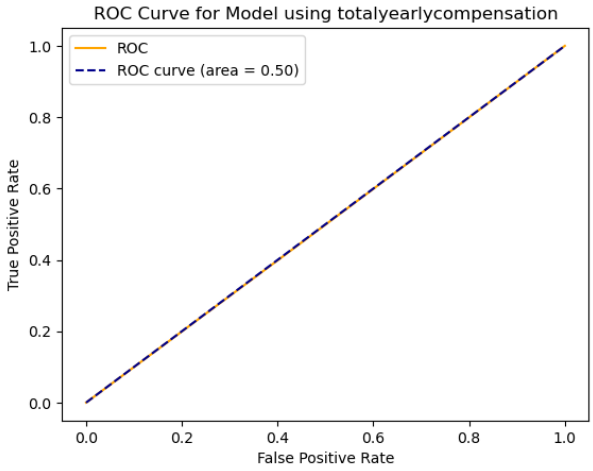
Question 4

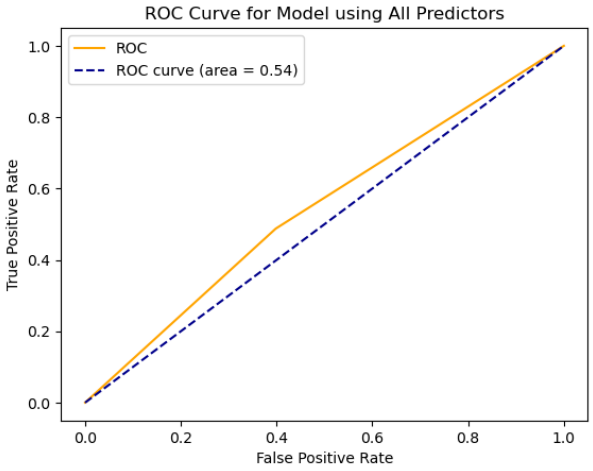
I created a new dataset based on the data used for the above questions by dropping units with a gender of "other" since we are examining disparities between male and female workers. Even though we are no longer predicting total yearly compensation, I still chose not to use the columns that are very highly correlated with that variable due to collinearity concerns. I dropped the "female" dummy variable and used the "male" column to represent the target variable of gender (so male=1, female=0). The classes (male and female) were heavily imbalanced, so instead of train_test_split(), I used StratifiedShuffleSplit() as it preserves class proportions. I verified this by graphing the class counts for the training versus test sets (see right). Because of the class imbalance, I ran logistic regressions with the parameter class_weight = "balanced" to adjust the class weights. I fit a logistic regression model using just total yearly compensation, then fit another model controlling for all other predictors.



The single-predictor model resulted in a beta of 0.1417, but the model was not very accurate (accuracy = 0.4445) and resulted in many false negatives, or incorrect classifications of males as females (recall = 0.4135). In the ROC curve graph below (see the first graph), the AUC of the model was only 0.50, meaning its ability to classify is no better than guessing.

The logistic regression model using all predictors produced a beta of 0.0887 for the total yearly compensation predictor. This model was also not very accurate (accuracy = 0.5083), but it was slightly more accurate than the single-predictor model. It also had a slightly higher recall score than the previous model, at 0.4876. The AUC was improved to 0.54 (see the second graph below), but this still indicates that the model's predictions are similar to randomly guessing.





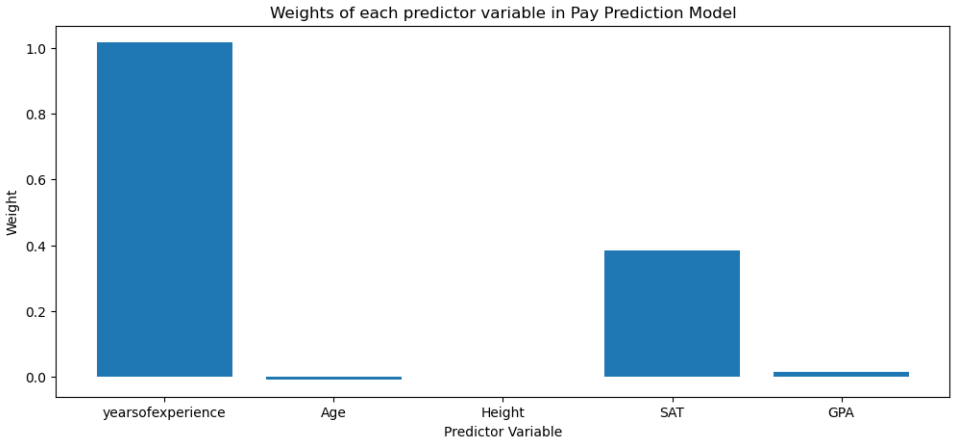
The beta associated with total annual compensation is more appreciable in the model that does not control for other factors, and the beta is closer to 0 when controlling for other factors. However, as found above, both models are basically randomly guessing. The latter model is slightly better

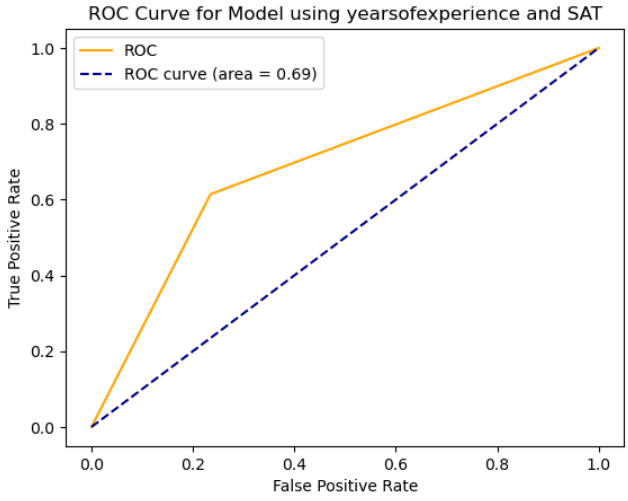
at classifying workers as male or female, but based on the small beta for the multivariate model, total annual compensation is not an important contributor to predicting workers' gender. Unexpectedly, **these results do not include any strong evidence to support the existence of a gender pay gap.**

Question 5

I re-loaded the original dataset to create another dataset for this question because I could use much more of the original data, as it does not involve any of the columns containing many null values (race, education, and gender). I kept only the columns relevant to this problem and created a new outcome variable that labels those with a higher-than-median annual salary as high earners (highearner=1) and those with a lower-than-median salary as low earners (highearner=0). The base salary column was no longer needed, so it was dropped. I standardized the predictors and split the data into training and test sets using train_test_split() and fit a logistic regression using all 5 predictors.

The resulting model performed decently, with an accuracy of 0.6926, precision of 0.7067, recall of 0.6137, and AUC of 0.69. I examined the weights of each predictor variable by plotting the values. As seen below, it appeared that the only important variables were years of experience and SAT score, so I tried fitting another logistic regression model using just these variables.

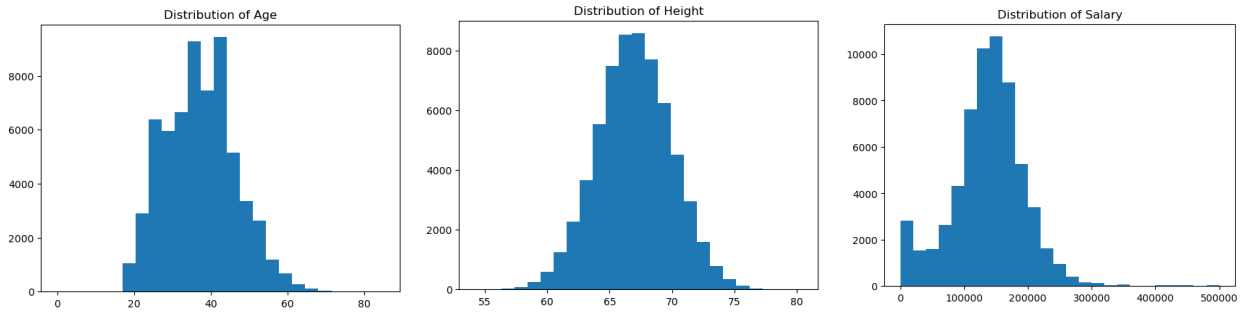




The new model produced essentially the same values as the previous model for accuracy, precision, recall, and AUC. This means using a model with only years of experience and SAT scores as predictors will correctly classify workers as high or low earners just as well as the original five-predictor model. An AUC of 0.69 is not amazing, but **using years of experience and SAT scores in a logistic regression model, you can predict high versus low pay close to acceptably well.**

Extra Credit Part 1

I used all of the original data to plot histograms for age, height, and salary since none of these columns contain null values.



Age is not normally distributed, which makes sense, since people don't work salary-paying jobs until they are adults. Also, workers retire at varying ages, resulting in the tail on the higher end of the distribution.

Height is normally distributed in this data, which is also unsurprising since height is a commonly used example of a normally-distributed human trait. I would have understood if it were not normal, though, due to the data being self-reported; people had the opportunity to report themselves as taller than they actually are.

Salary is not normally distributed, mostly due to the high number of low- or no-salary jobs in the data. This is surprising because the data pertains to workers at tech companies, and I would expect all people working in tech to receive salaries, even if not necessarily high ones. 2304 workers in this dataset receive no salary, and another 500 workers have salaries below $20,000, which means these workers' compensation is given mostly or entirely in bonuses, stock grants, and/or some other unrecorded way.