Qingyu Zhang / Andy

N-number: N19903322

Prof. Pascal Wallisch

Fundamentals of Machine Learning

HW2 Report: Salaries

Question 1:

1）What was done:

Predictors were selected as requested (excluding variables 1-3 as well as other qualitative variables, and excluding variables 5-7. One of the five dummy variables for Education was dropped according to hints, as well as one of the five dummy variables for Race). Dropna operations were performed according to predictors, outcome variables, and values in the Race and Education columns.

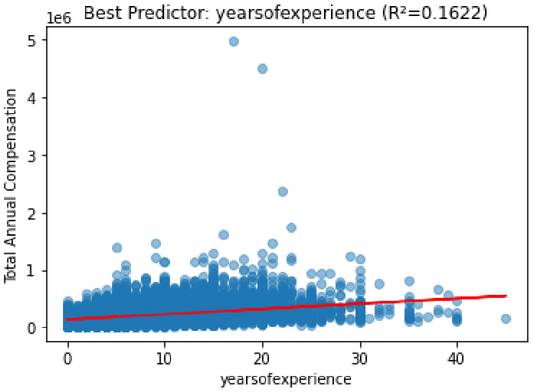
Separate ordinary linear regressions were created using each predictor and the corresponding R-squared values were recorded. Find the best predictor based on the highest R-squared value. Created a multiple linear regression using all predictors and recorded the R-squared value. The regression plot for the best predictor and the actual vs. predicted plot for the multiple linear regression were drawn.

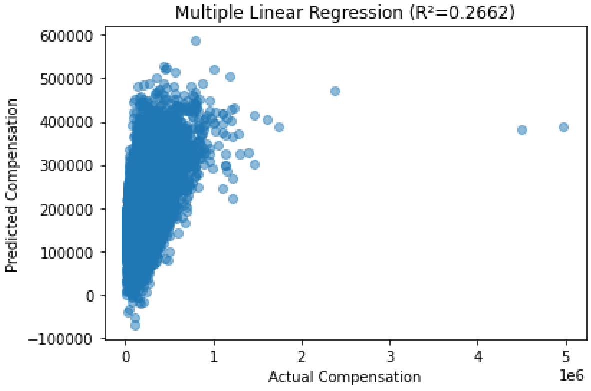
2）Why this was done:

Data preprocessing is basically done based on hints. For example, the reason for removing one of each of the dummy variables in Education and Race is because after NaN is dropped, knowing four dummy variables makes the fifth 100% certain. One is deleted to prevent the model from being overdetermined. For the modeling aspect, the R-squared is the best indicator to evaluate the effectiveness of a predictor. So, we can use this to select the best predictor. Drawing plots can visualize the fitting effect.

3）Findings:

The best predictor of total annual compensation is “years of relevant experience” with R^2 of 0.1622. The R^2 of the multiple linear regression model that uses all predictors is 0.2662. Followings are plots:





4）Interpretation:

Years of relevant experience is the strongest individual factor affecting total annual compensation, with R^2 of 0.1622. This means this best predictor explains about 16.22% of variance in total annual compensation. The full multiple linear regression model's R^2 is 0.2662, meaning about 26.62% of the compensation variance is explained by the selected predictors. The full model performs a bit better, but its R^2 just increases a little bit and is still low, indicating that most of the variance is explained by Year of Relevant Experience. Variables other than it are not good linear predictors of total annual compensation.

Introducing them does not improve the model performance much, but it raises the risk of overfitting.

Question 2:

1) What was done:

   Standardize on all predictors. Ridge regression was performed on total annual compensation using all predictors. Using the train test split, the model was fitted on the training set using the RidgeCV() function and the hyperparameter lambda was automatically tuned. This function cross-validated each of the 100 different lambda values and the lambda with the smallest MSE was selected as the optimal hyperparameter and the Ridge Regression model was automatically fitted with it. R-squared value was calculated on the test set to evaluate the model. The plot of actual vs. predicted values was also drawn. Access the alpha_ parameter of the model to access and record the best lambda.

2) Why it was done:

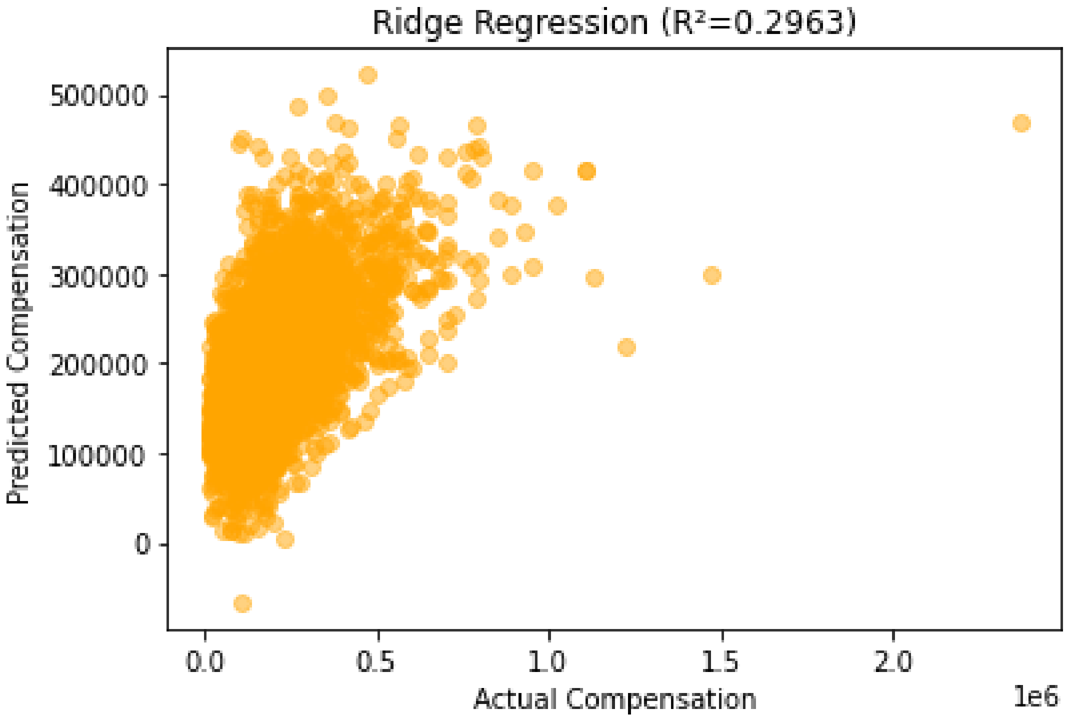
   Ridge Regression is a regularized regression designed to suppress overfitting and multicollinearity due to too many predictors. So I think the question "do the same thing as in question 1" means only build a Ridge regression with all predictors. After all, it doesn't make much sense to use Ridge regression for a single predictor. Using the

RidgeCV() function makes it easy to simultaneously perform hyperparameter tuning and fitting the model with the selected optimal lambda. This function uses k-fold cross-validation, where the training set is divided into smaller training and validation sets a few times over, yielding a better and more convenient optimal lambda than manual tuning. In addition, the R-squared is also the best metric for evaluating Ridge regression. The point of standardization is to prevent the effect of different scales of predictors on the penalty term (e.g., predictors with too large magnitude of values may be penalized more).

3) Findings:

The Optimal Lambda 46.4159. R^2 of the Ridge regression using this Lambda is 0.2963.

Following is the plot:



4) Interpretation:

Ridge Regression (R^2 0.2963) slightly improved R^2 compared to multiple linear regression in Question 1 (R^2 0.2662). This suggests that the predictive effectiveness of Ridge, an L2 regularization, is not significantly improved compared to multiple linear regression. The large lambda value 46.4159 suggests strong regularization was needed due to multicollinearity between predictors. Ridge needs to use a larger lambda to stabilize parameter estimates and avoid overfitting. Of course, the larger lambda could also be due to noisier data, so Ridge needs stronger regularization to suppress the fit to the noise.

Question 3:

1) What was done:

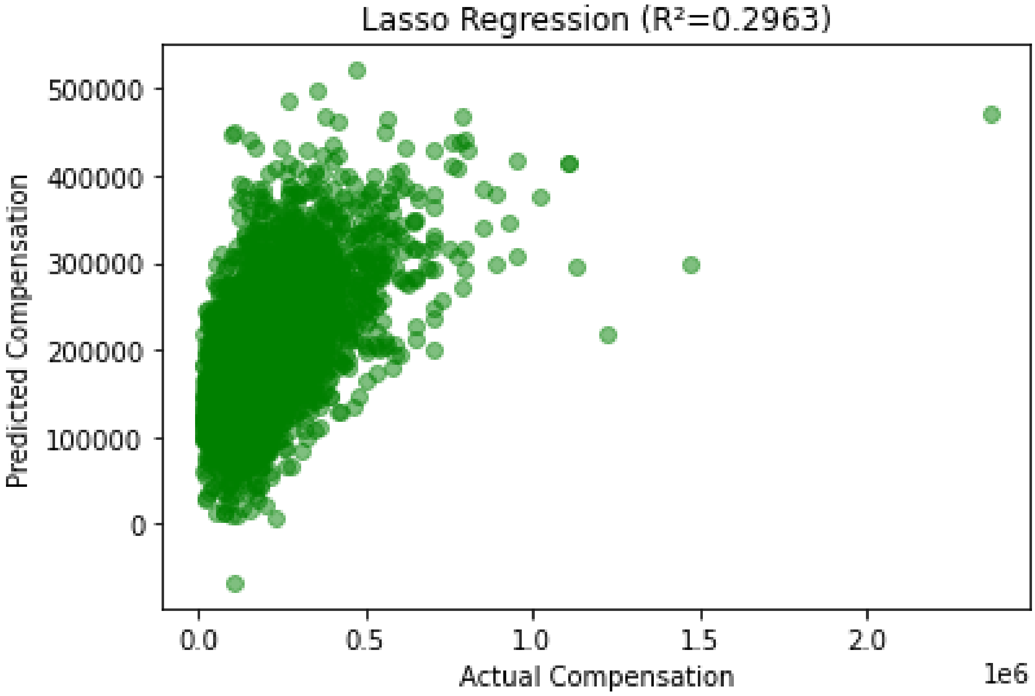
Similar to problem 2, except this time Lasso Regression, the L1 regularization, was used. Lasso regression was performed on total annual compensation using all predictors. Using the train test split, the LassoCV() function was used on the training set to fit the model and auto-tune the hyperparameter lambda. This function is similar to the RidgeCV() function in Problem 2. R-squared value was computed on the test set to evaluate the model. The plot of actual vs. predicted values was also drawn. The coef_ parameter of the fitted Lasso model was called to see how many regression coefficients beta were shrunk to 0. Access the alpha_ parameter of the model to access and record the best lambda.

2) Why it was done:

All parts are similar to Problem 2. Lasso Regression is also a regularized regression designed to suppress overfitting and multicollinearity due to too many predictors. The LassoCV() function is also used because of its convenient automatic hyperparameter tuning and fitting the model with optimal hyperparameter. In addition, the R-square is likewise the best metric for evaluating Lasso regression. The point of standardization is to prevent the effect of different scales of predictors on the penalty term (e.g., predictors with too large magnitude of values may be penalized more).

3) Findings:

The Optimal Lambda 107.2267. R^2 of the Lasso Regression using this Lambda is 0.2963. Number of coefficients shrunk to Zero: 2. Following is the plot:



4) Interpretation:

R^2 of Lasso remains similar to Ridge. Lasso Regression(R^2 0.2963) slightly improved R^2 compared to multiple linear regression in Question 1 (R^2 0.2662). This suggests that the predictive effectiveness of Lasso, an L1 regularization, is not significantly improved compared to multiple linear regression. Similar conclusion as Question 2, the large lambda value 107.2267 again suggests strong regularization was needed due to multicollinearity between predictors. After trying both Ridge and Lasso regularizations, it was found that neither R-squared improved significantly, suggesting that perhaps we need non-linear methods. In addition, Lasso Regression removed 2 variables (shrunk 2 betas to 0), indicating that some predictors were redundant or not useful. While most features still contribute to some extent.

Question 4:

1) What was done:

Filters and retains only those rows/samples with a value of Male or Female in the gender column. Create a new column with a 0/1 binary value for gender (mark those with a gender of Male as 1 and Female as 0). Standardize total annual compensation data. Using train test split, build a logistic regression model on the training set to predict the 0/1 binary value gender classification using total annual compensation. Calculate AUROC and AP scores on the test set for model evaluation. Draw ROC curve and PR curve to visualize the model classification effect. Calculate precision, recall, and accuracy using probability 0.5 as the classification threshold as a reference.

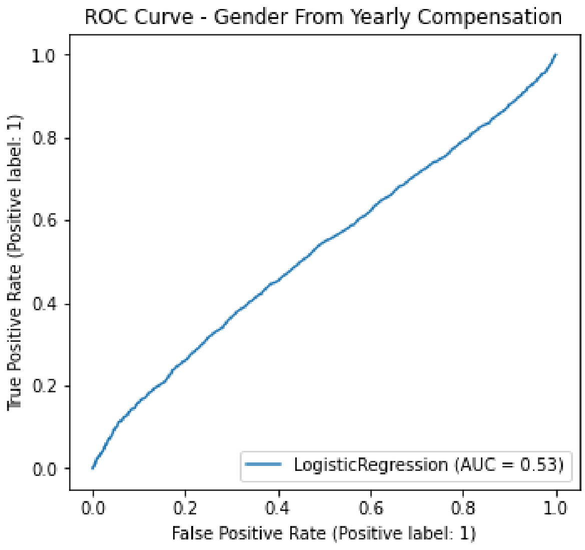
2) Why it was done:

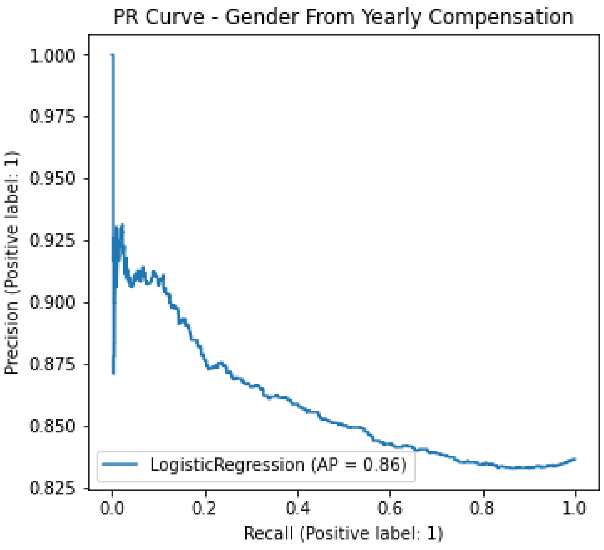
Logistic regression classification models only accept binary values like 0/1 as outcome variable, so I convert Male and Female markers to 1/0 binary values. Standardizing the predictor improves the performance and stability of the model. Use train test split to prevent overfitting and test the generalization ability of the model. AUROC score and ROC curve are the best metrics for evaluating the logistic regression classification model. PR curve and AP score help us to check whether the dataset has data imbalance or not. Precision, recall, and accuracy calculated with a threshold of 0.5 are for reference only.

3) Findings:

AUROC: 0.5290 AP: 0.8587

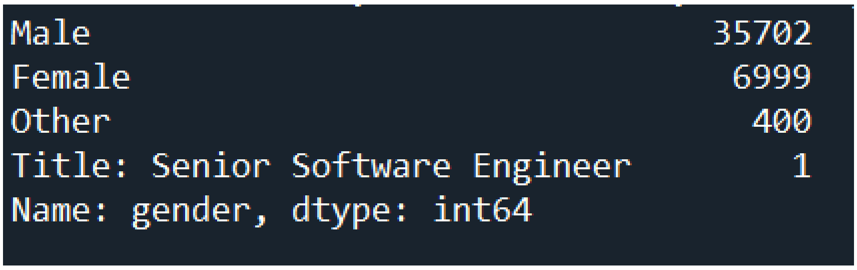
Precision: 0.8364, Recall: 1.0000, Accuracy: 0.8364 (Threshold of 0.5)


ROC Curve - Gender From Yearly Compensation


PR Curve - Gender From Yearly Compensation

4) Interpretation:

The low AUROC (0.5290) suggests that total annual compensation alone is not a good predictor of gender. AUROC scores generally take values between 0.5 and 1, with larger values being better. An AUROC score close to 0.53 suggests that the model works only a little better than randomly guessing gender (AUROC of 0.5). So, there is no data evidence to show that there is an income difference between men and women (and we can't use income to predict gender categorization very well). Also, AP = 0.86 indicates that Precision stays high in most Recall positions. This may be due to a significant imbalance in the number of male and female genders. The model may have overfitted to the majority class. (After checking the value count, I found that there are indeed far more males than females in the dataset.)



```
Male                                35702
Female                               6999
Other                                 400
Title: Senior Software Engineer         1
Name: gender, dtype: int64
```

(83.6% of male + female people are male. And we have accuracy of 0.836 with threshold of 0.5. This means the model with threshold of 0.5 just predict everyone as male and blindly get accuracy of 0.836)

Question 5:

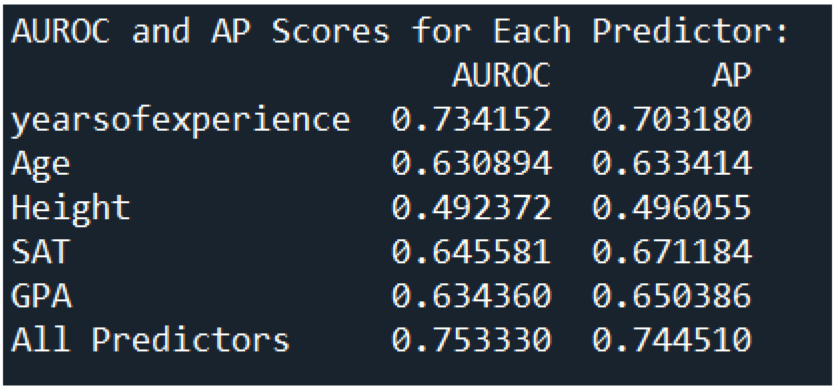
1) What was done:

   Select the five required features as predictors. Calculate the median of total annual compensation. Mark samples with annual compensation larger than this median as 1 (high income) and vice versa as 0 (low income). Standardize predictors data. Use a train test split. First create logistic regression to predict income categories using each individual predictor separately, and record AUROC and AP scores. The predictor with the highest AUROC score was selected as the best predictor. After that all predictors were used to build logistic regression to predict the income category jointly and AUROC and AP scores were recorded. Draw ROC and PR curves for single best predictor model and all predictors model.

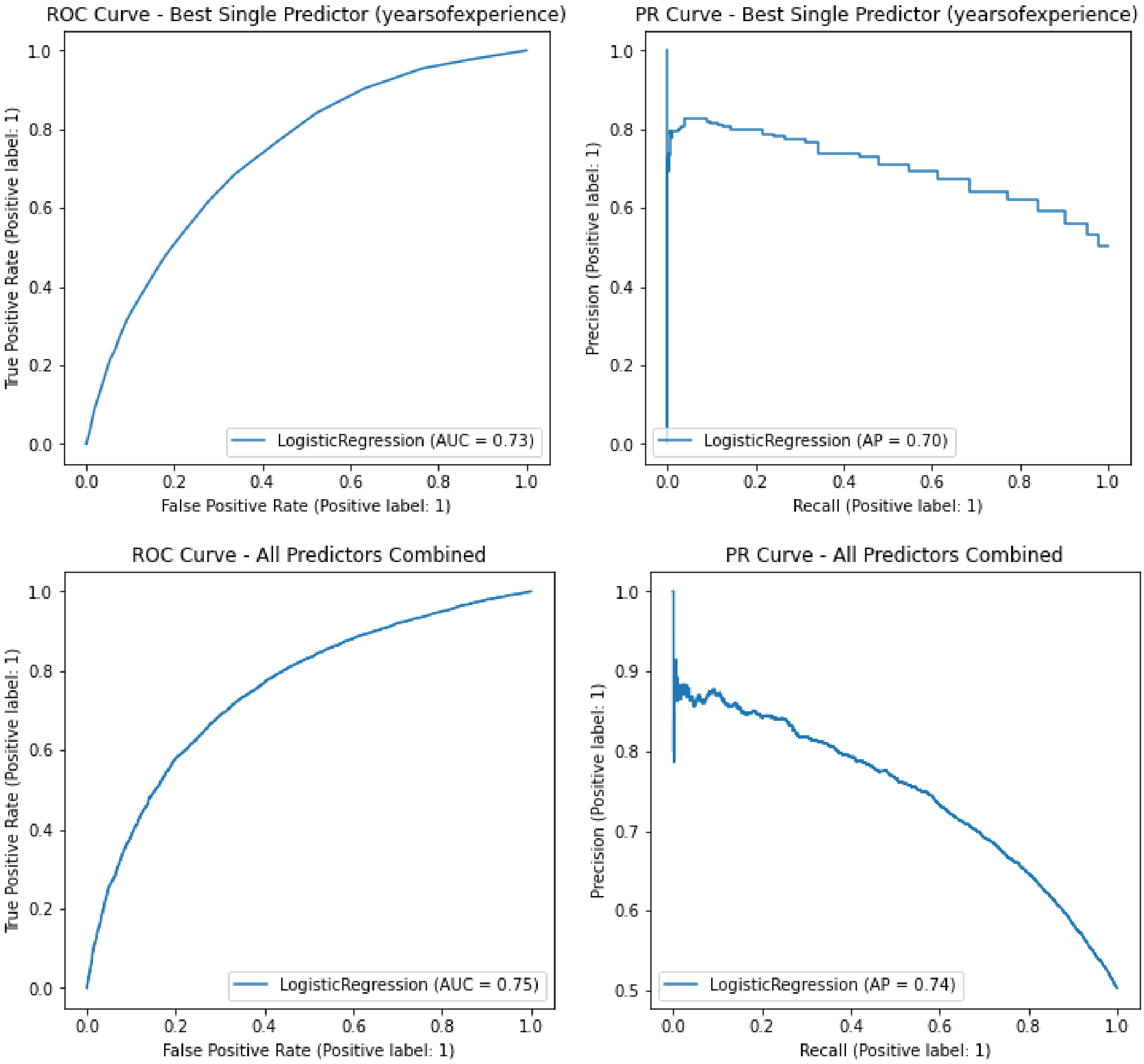
2) Why it was done:

   Similar to Problem 4, we categorize annual income by median into high-income and low-income categories in order to use logistic regression classification, a binary classification model. Standardizing the predictor improves the performance and stability of the model. Use train test split to prevent overfitting and test the generalization ability of the model. AUROC score and ROC curve are the best metrics for evaluating logistic regression classification models.PR curve and AP score help us to check the dataset for data imbalance. So, we also use AUROC to determine which is the best predictor. Since I wasn't sure if the question wanted us to build univariate models for each predictor or use all predictors together, I did both. This way I can compare them as well.

3) Findings:



```
AUROC and AP Scores for Each Predictor:
                        AUROC        AP
yearsofexperience    0.734152  0.703180
Age                  0.630894  0.633414
Height               0.492372  0.496055
SAT                  0.645581  0.671184
GPA                  0.634360  0.650386
All Predictors       0.753330  0.744510
```



4) Interpretation:

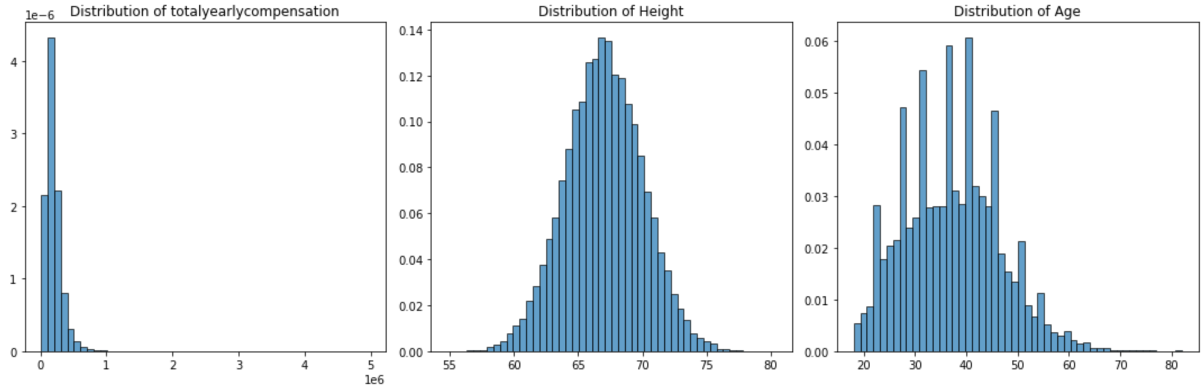
Years of relevant experience is the best single predictor of high/low pay (AUROC of 0.73). Such AUROC value and ROC curve suggest that Years of relevant experience is a good predictor of income levels. Age, SAT, and GPA (all with AUROC scores of about 0.63-0.65) are OK predictors, not great, but better than random guessing. Height has no predictive power for income levels, and has the same effect as random guessing (which is reasonable). The AUROC score for all predictors together is about 0.75, which is only a little higher than that of years of relevant experience. This suggests that the performance improvement is moderate. Years of relevant experience plays a major role, and the other variables may be redundant.

Extra Credit Problems:

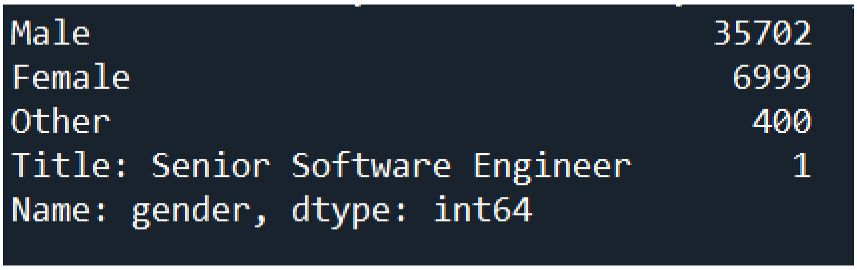
(a) Plotted histograms of the distributions of 3 variables. It seems that Height is distributed normally, which is what I expected since theoretically medium height people should be the most numerous and extreme height people should be less numerous.

The other 2 are not normally distributed. Salary distribution is what I expected. Most people should be low and middle income. Then there is a very small percentage of people with very high incomes as outliers.

The age distribution surprised me a bit. Because I don't understand why there are a few specific ages that appear so prominently and frequently. If you ignore that, the distribution makes sense because most of the people working are between 30-45 years old.



(b) I think the grossly disproportionate ratio of men to women in this dataset that I found in question 4 is an interesting finding.



```
Male                                     35702
Female                                    6999
Other                                      400
Title: Senior Software Engineer              1
Name: gender, dtype: int64
```