**Fundamentals of Machine Learning: Homework 2**

Before answering the questions, it was important to clean the dataset and make sure there was no covariance. Hence, the education and race columns were dropped, and the dummy columns were used. Furthermore, from the dummy variables, the two\_or\_more\_races and some\_college columns were dropped to reduce covariance amongst the data columns. At last, all the na columns were dropped to make the results more accurate and representing.

**Question 1**

To answer the question the data was divided into X and Y data frames, where X is all of the independent variables, and Y is the compensation. Then, using the stats model library, an OLS model was fitted using the X and Y data. Then using the independent variable with the highest coefficient and the lowest p-value the best predictor was found. At last, an OLS model was fitted using the best predictor only as the X data, and the results were plotted along with the R^2 values.

The advantage of using OLS is that it provides unbiased estimates of the regression coefficients. Other methods require several parameters that can be tougher to calculate and provide.

The data found through both of the OLS models led to:

The best predictor of total annual compensation is yearsofexperience

R-squared for full model: 0.26351507839527455

R-squared for best predictor model: 0.16017546046861586

Chart, bar chart

Description automatically generated

As the years of experience data has the highest beta coefficient value, it is the best predictor of the yearly compensation. The whole model accounts for a variance of 26% in the yearly compensation while the best predictor accounts for 16%. These aren’t the highest of variance values and wouldn’t be strongly reliable estimates for the yearly income.

**Question 2**

Using the X and Y data from the previous question, a Ridge regression model was fitted with 10 folds to prevent over fitting and ensure hyperparameter tuning. The lamba value for this model was calculated. Following this, Y values were predicted using the test data and the R^2 values were calculated. Next, all the coefficient values were looked at and the the highest value was established as the best predictor. Finally, the same Ridge model and the steps that follow were fitted using the best predictor value. The R^2 values were compared and the coefficients and variance values were plotted.

Ridge regression is a type of regularized regression that adds a penalty to the sum of squared coefficients in the model, which helps to prevent overfitting and improve generalization to new data. Ridge regression can help to address this problem by shrinking the regression coefficients towards zero, which can reduce the variance of the estimates and improve the accuracy of the model.

The data found through both of the Ridge models led to:

Regression Coefficients: [ 0.00000000e+00 8.87856920e+03 -2.27831181e+03 -4.83153644e+03

-4.48166627e+03 -2.66531262e+04 6.67154387e+04 -3.88676753e+04

-4.74102135e+01 3.33010255e+03 -1.48303008e+03 -3.34344728e+02

-6.48226762e+01 -2.61473714e+02 2.89735949e+02 7.67826073e+03]

R-squared (full model): 0.26351056772120696

Best Predictor: Doctorate\_Degree

R-squared (reduced model): 0.02037232137775502

R-squared Difference: 0.24313824634345194

Full Lambda: 10.0

Reduced Lambda: 1.0

Chart, bar chart

Description automatically generatedChart

Description automatically generated

The R^2 value of the full prediction model stays the same as the OLS at 26%. However, the best predictor is the Doctorate degree rather than the years of experience. The doctorate degree predicts 2% of the income. The lambda value for the full model is 10. In this case, the lambda value of 10 suggests that the model used a relatively strong regularization penalty to prevent overfitting of the model to the training data.

**Question 3**

Using the X and Y data from the previous question, a Lasso regression model was fitted with 10 folds to prevent over fitting and ensure hyperparameter tuning. The lamba value for this model was calculated. Following this, Y values were predicted using the test data and the R^2 values were calculated. Next, all the coefficient values were looked at and the highest value was established as the best predictor. Finally, the same Lasso model and the steps that follow were fitted using the best predictor value. The R^2 values were compared and the coefficients and variance values were plotted.

Lasso is preferred for feature selection in multiple linear regression because it can effectively handle high-dimensional datasets with irrelevant variables, leading to more parsimonious and interpretable models. Other methods can also be used for feature selection, but they may not be as effective in removing irrelevant variables as Lasso, particularly when the dataset has a large number of features.

The data found through both of the Lasso models led to:

Regression Coefficients: [ 0. 8514.21149681 -1324.54999283 -0.

0. -6441.20450362 0. -0.

0. 0. -0. -0.

0. -0. 317.04427456 0. ]

R-squared (full model): 0.24366514272229978

Best Predictor: yearsofexperience

R-squared (reduced model): 0.16017460566035913

R-squared Difference: 0.08349053706194065

Full Lambda: 5163.99254606459

Reduced Lambda: 715.065589458419

**Chart, bar chart

Description automatically generatedChart, waterfall chart

Description automatically generated**

The R^2 value of the full prediction model reduced to 24. However, the best predictor is the the same as the OLS – yearsofexperience. The years of experience predicts 16% of the income. The lambda value for the full model is 5163. In this case, the lambda value of 5163 suggests that the model used a relatively strong regularization penalty to prevent overfitting of the model to the training data. The lambda value of 5163 suggests that the Lasso algorithm applied relatively strong regularization to the model, which likely resulted in some of the predictor coefficients being shrunk towards zero or completely removed from the model. A total of 12 predicors are shrunk to 0 as an effect of the regularization.

**Question 4**

The code uses logistic regression to predict gender based on total yearly compensation. It splits the data into training and test sets, fits the logistic regression model to the training data, and evaluates the accuracy of the model using cross-validation. It then predicts the probabilities of each gender for the test set, calculates the ROC curve and AUC score, and plots the ROC curve. The code also predicts the gender for the test set using a threshold of 0.5, calculates the confusion matrix and accuracy score, and plots the confusion matrix. Finally, it prints out the accuracy score.

The code does it for both, with controlling for other factors and without.

Logistic regression is a commonly used method for modeling binary outcomes, such as whether a certain event occurs. In the case of finding an appreciable beta associated with a predictor, logistic regression can be a good choice because it estimates the probability of an event occurring based on the predictor variable. Other methods, such as linear regression, may not be appropriate because they do not account for the fact that the outcome is restricted to two possible values.

The method followed above produced the following data and results for **without** controlling for other factors:

Variable Beta

0 totalyearlycompensation -6.967237e-07

1 yearsofexperience -6.689760e-03

2 yearsatcompany -3.074862e-03

3 Masters\_Degree 2.066259e-04

4 Bachelors\_Degree -1.135169e-04

5 Doctorate\_Degree -1.026866e-04

6 Highschool -1.309442e-04

7 Race\_Asian 8.842013e-06

8 Race\_White -2.839398e-04

9 Race\_Black 2.076693e-04

10 Race\_Hispanic -7.140988e-05

11 Age -1.143473e-02

12 Height -8.347280e-03

13 SAT -3.331105e-04

14 GPA -6.958041e-05

Cross-validation Accuracy: 0.82 (+/- 0.00)

Chart, line chart

Description automatically generatedChart

Description automatically generated

The method followed above produced the following data and results for **with** controlling for other factors:

Variable Beta

0 totalyearlycompensation -0.000007

Cross-validation Accuracy: 0.82 (+/- 0.00)

Chart, line chart

Description automatically generatedChart

Description automatically generated

The logistic regression model had an 82% accuracy in predicting whether someone was male or female, for both based on both with and without considering other predictors along with the income. The largest beta value having an impact on predicting the gender was the race\_asian.

**Question 5**

In the code the independent variables and dependent variable are defined by the programmer. The data is then split into training and testing sets. A logistic regression model is fit to the training data, and predictions are made on the testing data. Accuracy, precision, recall, and F1 score are then calculated and printed out. Finally, a confusion matrix is plotted to visualize the performance of the model on the testing data.

A logistic regression model is well-suited for predicting high and low pay from years of relevant experience, age, height, SAT score, and GPA because it can handle binary outcomes, which is appropriate for the two possible outcomes in this case. Other methods may not be appropriate for this scenario because they may not be able to handle binary outcomes, or they may not be able to handle a combination of continuous and categorical predictors.

The data from the above code provided the results:

Accuracy: 0.6723549488054608

Precision: 0.6945532217871129

Recall: 0.6275329688002573

F1 Score: 0.6593443730990199

Chart, treemap chart

Description automatically generated

An accuracy value of 0.67 indicates that the model is correct in its predictions about 67% of the time. The precision value of 0.69 indicates that the model correctly predicted 69% of the true positive cases, while the recall value of 0.63 indicates that the model correctly identified 63% of the actual positive cases. The F1 score, which is a harmonic mean of precision and recall, is 0.66. To some degree, the model does a good job of predicting low and high paying years given the factors: relevant experience, age, height, SAT score and GPA.

**Extra Credit**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated** 