**Data Analytics II:**

**Assignment #2**

1. Diabetes Linear Regression example
   1. Carry out the Linear Regression example on the SKLearn website at <http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#example-linear-model-plot-ols-py>
   2. Using the same data source, calculate the coefficients, residual sum of squares, and variance score for all features in the diabetes dataset. Compare the variance explained by the model in part a to that in part b and comment on how much additional predictive power you get by adding extra explanatory variables.

For part a:

Coefficient: 938.23786125

Mean Squared error: 2548.07

Variance Score: 0.47

For part b:

Coefficient:

[ 3.03499549e-01 -2.37639315e+02 5.10530605e+02 3.27736980e+02

-8.14131709e+02 4.92814588e+02 1.02848452e+02 1.84606489e+02

7.43519617e+02 7.60951722e+01]

Mean Squared error: 2004.57

Variance score: 0.59

(0.59-0.47)/0.47= 0.2553

Which means that around 25.53% additional predictive power is gained by adding extra explanatory variables.

By adding extra explanatory variables, the R^2 score rose from 0.47 to 0.59. and the mean squared error decreased from 2548.07 to 2004.57.

* 1. Split the data into 10 “folds” of equal size. For each fold, calculate the coefficients, residual sum of squares, and variance score for all features.

I did this part in Jupyter Notebook

* 1. What is the range of variance score generated using 10-fold cross-validation? What is the average variance score generated? What is the standard deviation of the variance scores? Fit all of the data using a regression model, and calculate the resulting coefficients, mean-squared error, and variance score How does the average variance score across 10-fold cross-validation compare to the score generated when fitting all of the data? What can we conclude about the data based on these results?

According to the 10-fold cross-validation I generated, the variance scores are: 0.36, 0.27, 0.55, 0.37, 0.59, 0.47, 0.52, 0.51, 0.66, 0.55. Therefore, the range of variance score generated is [0.27, 0.66]

The average variance score is 0.48387

The standard deviation of the variance scores is 0.1130

Already fitted all the data using a regression model in part b, so I will use the results from part b in addition to part c to answer the following comparison questions.

1. How does the average variance score across 10-fold cross-validation compare to the score generated when fitting all of the data?

As we know from part b, the variance score generated when fitting all the data is 0.59, while the average variance score across 10-fold cross-validation is 0.48387, which is smaller than the variance score generated when fitting all the data.

1. What can we conclude about the data based on these results?

The variances in the dataset is not distributed equally or randomly among all the data, otherwise the average variance score across 10-fold cross validation should be very close to the variance score generated when fitting all the data

1. Carry out the Underfitting/Overfitting tutorial at: <http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html#example-model-selection-plot-underfitting-overfitting-py>
   1. Write comments in your code explaining the function of each line of code

I did this part in Jupyter Notebook

* 1. Which of the plots in the tutorial is underfit? Define “underfitting” in the context of linear regression. Does the error in this model constitute bias or variance? Explain.

The first plot is underfit. As we can see, the model cannot really represent the data that we have. As we can see, the first plot is a very simple straight line and a lot of samples are not on the model line. The error in this model constitute bias.

* 1. Which of the plots in the tutorial is overfit? Define “overfitting” in the context of linear regression. Does the error in this model constitute bias or variance? Explain.

The third plot is overfit. We can see that in the third plot, the model is trying to fit every data point in our sample, so the line has a really unique pattern, so it’s likely to fail to in prediction. The error in this model constitute variance.

* 1. Is R2 a better metric for model error? Why or why not?

No. Because R^2 is always going to increase when you include more independent variables, while model error might increase as well.

* 1. Calculate the AIC and BIC of each of the models in the tutorial. Which one would you select based on each criterion?

Calculated the AIC and BIC of each of the models in Jupyter Notebook

AIC for the 1 model is: 572.7196292626445

BIC for the 1 model is: 574.1208266443067

AIC for the 2 model is: 578.7196292626445

BIC for the 2 model is: 584.3244187892931

AIC for the 3 model is: 600.7196292626445

BIC for the 3 model is: 621.7375899875768

I would choose the first model based on AIC/BIC but I would choose the second model because the first model is underfitting and the third model is overfitting. I would choose the third model based on R^2 score.

1. Using SKLearn, build an OLS regression model for the Auto MPG Data Set including all explanatory variables (except car name).—linear regression
   1. Clean the data, removing samples with empty entries and scaling each feature to have zero mean and unit variance
2. What are the best fitting regression coefficients?

Regression Coefficients: [[ 0.13883145 -0.72761561 1.6752486 -0.58538002 2.63785346 1.15375847 -5.27368628]]

1. How much variance in the data does this model explain?

0.36

1. What is the residual sum of squares?

381.386

1. Calculate the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for two models – one with all coefficients, and one with the largest magnitude coefficient. According to each of these criteria, which one of these two is a better model?

The model with ALL coefficients is a better model according to AIC&BIC

1. Using the AIC or BIC criteria, find a better model for this dataset

Using the AIC/BIC criteria, a better model for this dataset is with ["displacement","horsepower","weight","acceleration","model year","origin"]

As independent variables which has an AIC of 1169.65 and BIC of 1197.45

1. Use 10-fold cross validation and MSE as a metric to select among these three models. How are the results different from using AIC or BIC?

The MSE For the model with all coefficients: 0.64

The MSE for the model with the largest coefficient: 0.36

The MSE for the model with the better AIC/BIC: 0.64

The results are pretty similar to the results I got from using AIC/BIC

1. Allow for the use of interaction terms (but not quadratic terms) in the MPG dataset. Show the coefficients. How would you interpret the results of this model? Is this a better or worse model than the best model in #3?

Regression Coefficients: [[ 4.58388525 -34.52559606 11.30133173 1.59541636 -15.68524825

1.49643914 -16.07459277 -1.56073878 6.97965649 0.49752305

6.72750518 -14.990987 1.26984398 -2.39596685 14.71422849

-4.11548098 26.30717675 2.62664441 -6.07113681 -2.8683943

-7.30805571 -0.67766814 2.87926056 -9.57023841 -1.59152114

11.6191628 6.83191683 8.82963216]]

The R^2 of the linear regression model with interaction terms allowed is better but may be because the new linear regression model has more independent variables, so R^2 value is not dependable here. However the mean squared error and residual sum of squares are both smaller, which might indicate that this linear regression model is better

1. Download data from: <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.csv>

These data list passengers on the *Titanic* and whether they survived the sinking of the ship.

1. Import the data to SKLearn, retaining the following data fields: pclass, survived (the target variable), sex, age, sibsp, parch, fare, embarked. (descriptions of these fields are provided here: <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/Ctitanic3.html#pclass> ). Segment the data by those who survived and those who did not survive. For those who survived, indicate the mean of each continuous variable and the mode of each categorical variable.
2. What is the mean of each continuous variable and the mode of each categorical variable for each sex? What is the observed probability of survival for each sex?

In Jupyter notebook

1. What is the mean of each continuous variable and the mode of each categorical variable for each passenger class? What is the observed probability of survival for each passenger class?

In Jupyter notebook

1. Fit a logistic regression model to the data using all variables from part 1. Using 10-fold cross-validation, what is the model’s average prediction accuracy?

0.77

1. Now that you’ve fit a model using logistic regression, calculate the AIC and BIC of this model on all of the data.

AIC:861.998

BIC:29.595

1. Using only these measures, what is a model that is more parsimonious according to both AIC and BIC?

Predicting survival rate based on fare

1. What is the accuracy of this model?

64%

1. How does it compare to a random coin flip?

better

1. How does it compare to a model that constantly predicts one outcome?

Worse, because if a model constantly predicts survival, the accuracy of that model will be around 0.6877, and 64% is smaller than 68.77%

1. How does it compare to the “saturated” model that includes all predictors in the dataset?

The accuracy is not as good

1. Report the coefficients of the parsimonious model and interpret, in words, what they mean

Coefficient:[0.01202229]

Meaning that the survival rate=0.01202229\*fare+intercept+error

Which means that the higher someone paid for his/her fare is, the more likely that person would survive. For every one more dollar spent on the fare, the probability of that person surviving increases by 0.01202229

1. Download the breast cancer Wisconsin dataset from the UCI Machine learning repository: <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/> -- specifically, use the WDBC dataset.
   1. Fit a univariate OLS linear regression model to predict whether a given sample is benign or malignant using the Standard Error of the area as the independent variable. Specifically, using 0.5 as your cutoff criterion, “classify” samples in the test set by whether they are predicted to be malignant. Using 10-fold cross validation, what is the accuracy of this approach?

The accuracy is 0.09

* 1. Does linear regression perform well at this classification task? How does it compare to a model that consistently estimates “benign”? How does it compare to the logistic regression model calculated in #1?

No. It is worse than the model consistently estimates “benign” because the percentage of benign in the dataset is 62.85% and the accuracy for the linear model is 0.09

* 1. Explain the difference in performance between the logistic and linear regression models.

The accuracy for logistic regression is 0.87. The difference is because linear regression treated the target values as continuous values, so there are a lot of factors that are going to deviate the accuracy of linear regression, like the balance of data. We can see that benign has a bigger percentage so the linear model will be dragged more towards 0. And so on.

1. Download data from: https://github.com/ga-students/sf-dat-21/blob/master/unit-projects/dataset/admissions.csv

This data indicates whether students were admitted to graduate school based on their GRE scores, their GPA, and the prestige of their undergraduate institution (data are described at (https://github.com/ga-students/sf-dat-21/blob/master/unit-projects/dataset/README.md)

1. Import the data to SKLearn, treating academic rank as a categorical variable. Do some preliminary data exploration – what are the mean and median GRE, GPA of students who were admitted? What about students who were not admitted? What is the modal academic prestige rank among students who were admitted? What about students who were not admitted?

For admitted students:

GRE mean=618.57

GRE median=620.00

GPA mean=3.49

GPA median=3.54

Modal academic prestige rank: 2.0

For non\_admitted students:

GRE mean=573.58

GRE median=580.00

GPA mean=3.35

GPA median=3.34

Modal academic prestige rank: 2.0

1. What is the mean and median GRE score and GPA for each academic rank? What is the observed probability of admission for each academic rank?

1:

GRE mean:611.80, median: 600.00

GPA mean:3.45, median:3.53

2:

GRE mean:596.62, median: 600.00

GPA mean:3.37, median:3.38

3:

GRE mean:574.88, median: 580.00

GPA mean:3.43, median:3.43

4:

GRE mean:570.15, median: 560.00

GPA mean:3.32, median:3.33

The observed probability of admission for rank 1 is: 1.1785714285714286

The observed probability of admission for rank 1 is: 0.5578947368421052

The observed probability of admission for rank 1 is: 0.3010752688172043

The observed probability of admission for rank 1 is: 0.21818181818181817

1. Fit a logistic regression model to the data using all variables. Using 10-fold cross-validation, what is the model’s average prediction accuracy?

Average accuracy: 0.71

1. Now that you’ve fit a model using logistic regression, calculate the AIC and BIC of this model on all of the data. Using only these measures, what is the model that best predicts the likelihood of admission?

AIC:259.924

BIC:14.2638

The model with gpa and prestige as independent variables best predicts the likelihood of admission

1. What is the accuracy of this best-fitting model?

0.72

1. How does it compare to a random coin flip?

BETTER, because the accuracy is greater than 50%

1. How does it compare to a model that constantly predicts one outcome?

Better, because the accuracy is greater than always predicting 1(0.4649) or always predict 0(0.5351)

1. How does it compare to the “saturated” model that includes all predictors in the dataset?

Better, because accuracy is slightly higher. However, AIC/BIC is also slightly higher

1. Interpret, in words, the meaning of the best-fitting model that you generated in part e.

The best-fitting models uses gpa and prestige as independent variables to predict the admission result. It has 72% accuracy, which may indicate that the admission result is strongly related to gpa and their undergraduate school’s academic ranking

1. EXTRA CREDIT: Use SKLearn to fit several binary logistic regression models to the Flags dataset from the UCI repository (using religion as the dependent variable). Next, fit a multinomial logistic regression model. Use 10-fold cross-validation to test the accuracy of each model. Which one is better, and why?

Using the better-fitting model, predict the religions of the 34 new countries created since 1990 found at <http://geography.about.com/cs/countries/a/newcountries.htm> What is the accuracy on this new dataset? What is the accuracy of the worse-fitting model in the new dataset? If there is a difference, explain the difference.

1. FINAL PROJECT Module 2: Fit a logistic regression model or a linear regression model to the data source you identified in Module #1. Use the AIC and/or BIC criteria, and K-Fold cross validation to determine a best-fitting model. Interpret the coefficients of the model in the context of your chosen dataset and write a short paragraph describing your findings.

the best-fitting model’s data columns are:

'sex', 'cp', 'restecg', 'thalach','exang', 'oldpeak', ‘slope’, 'ca'

And the coefficients are: [[-0.73186402 0.7980464 -0.95463918 -0.60733082 0.46638956 -1.36670591 0.44611106 0.0199979 ]]

Which means that, when males are more prone to have heart diseases. The third chest pain type and the first resting electrocardiographic results are the most related to heart diseases, when the max heart rate reached is lower, the person is more likely to have heart diseases, when exercise induced angina is 1, the person is more likely to get heart diseases, ST depression induced by exercise relative to rest is negatively related to heart diseases, the bigger the slope of the peak exercise ST segment and number of major vessels colored by flourosopy are, the more likely the person has heart disease