Jayoung Kang

Minilab 4

Exercise 1.1. Import the sklearn.svm package. Using the documentation of the svc (link given above), answer the following questions:

a. *What are the kernels supported by svc in sklearn?*

- The kernels supported by svc in sklearn are ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’ or a callable. If none are specified, ‘rbf’ will be used. The kernel is the function that quantifies the similarity of two observations. A linear kernel quantifies the similarity of a pair of observations using Pearson standard correlation and the svc becomes linear in the features/

b. *What is the default value of the Regularization parameter C?*

- The default value of the regularization parameter is 1.0. The strength of regularization is inversely proportional to C.

c. *What can you explain about the degree parameter?*

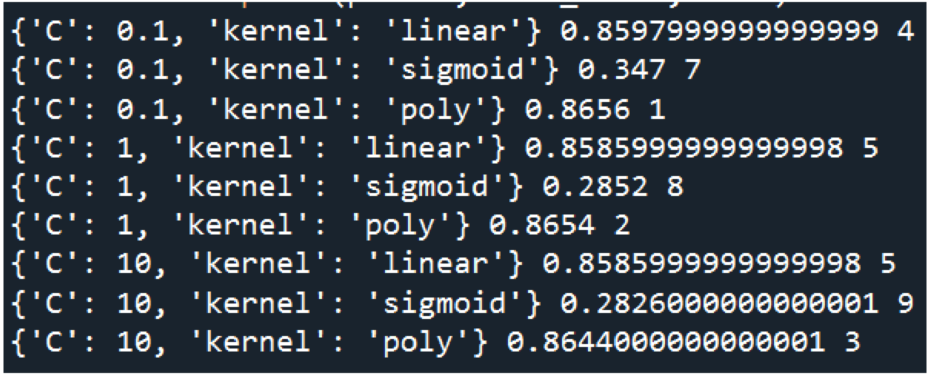
- The degree parameter states the degree of the polynomial kernel function ‘poly’ and the default value is set to 3. It refers to the degree of the polynomial used to find the hyperplane to split the data. When using the ‘poly’ kernel with d > 1, the svc algorithm gives a more flexible decision boundary than the ‘linear’ kernel. It fits the svc in a higher-dimensional space involving polynomials of degree d.

Exercise 2.1. Finish cleaning the datasets by

a. *repeating the cleaning process above for all the categorical variables in vote and work, not just prcitshp, and*
b. *converting the target variable of vote to binary values 0 and 1 as done for work. Ensure that the core variables have the same structure between the vote and work datasets*

Exercise 2.2. Consider 4 values of the C: 0.1, 1, 5 and 10. Consider three kernels: “linear”, “poly” and “sigmoid.” Report the cross-validation error rates of all 12 SVM models. Then, pick and report the C and kernel that maximizes the 5-fold cross-validation. Use this model for the rest of the exercises.

- The third set (C = 0.1, Kernel = 'poly') gives us the best result



```
{'C': 0.1, 'kernel': 'linear'} 0.8597999999999999 4
{'C': 0.1, 'kernel': 'sigmoid'} 0.347 7
{'C': 0.1, 'kernel': 'poly'} 0.8656 1
{'C': 1, 'kernel': 'linear'} 0.8585999999999998 5
{'C': 1, 'kernel': 'sigmoid'} 0.2852 8
{'C': 1, 'kernel': 'poly'} 0.8654 2
{'C': 10, 'kernel': 'linear'} 0.8585999999999998 5
{'C': 10, 'kernel': 'sigmoid'} 0.2826000000000001 9
{'C': 10, 'kernel': 'poly'} 0.8644000000000001 3
```

Exercise 2.3. What is the accuracy score of the model that you decided from Exercise 2.2 when fitting it to work_df?

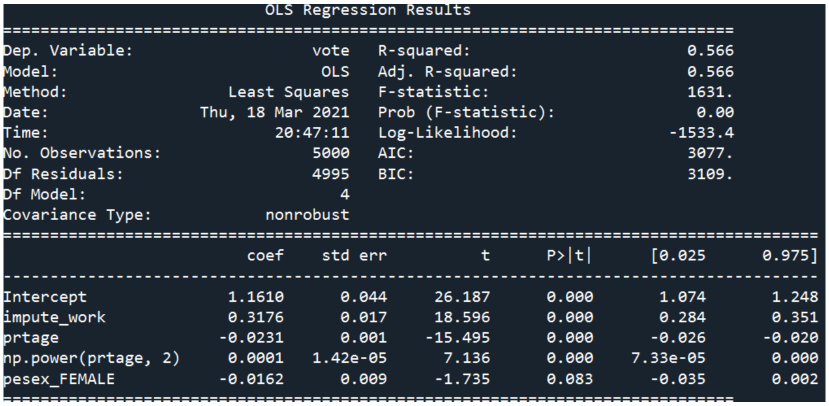
- Accuracy score = 0.8652

Exercise 2.4. Replicate the above code chunk [that predicts work status in the voting data] for the model that you fit in Exercise 2.3. The result should be the imputed work schedules and answer if accuracy score is applicable or not in this case. If not, explain why.

- 0.5084
- The accuracy score is approximately 50%, meaning that the imputed values predicted work schedule did just about as good as random guessing.
- The accuracy score is not applicable in this case because the imputed work schedule is from the vote data and so we don't actually observe the work schedules in the vote data.

Exercise 2.5. Regress the voting status on imputed work schedules. Use age, squared age, and sex as regressors in addition to the imputed work schedule. Be sure to convert the variables to an appropriate format. Interpret and discuss the results.

- The coefficient for impute_work is 0.3176 and statistically significant, meaning that there is a positive correlation between flexible working schedule and voting. Since the regression is not a logistic regression we can only comment on the direction of the relationship
- Pesex_Female has a negative coefficient but it is not statistically significant at a significance level of 0.05.
- Prtage and prtage^2 coefficients are statistically significant. Taking the first derivative shows that depending on the prtage value, there is a concave relationship between prtage and vote. However for most ages (below 115.5 – the maximum point) the relationship is positive.
- However since this data is semi-synthetic, so these results shouldn't be treated as entirely credible.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   vote   R-squared:                       0.566
Model:                            OLS   Adj. R-squared:                  0.566
Method:                 Least Squares   F-statistic:                     1631.
Date:                Thu, 18 Mar 2021   Prob (F-statistic):               0.00
Time:                        20:47:11   Log-Likelihood:                -1533.4
No. Observations:                5000   AIC:                             3077.
Df Residuals:                    4995   BIC:                             3109.
Df Model:                           4
Covariance Type:            nonrobust
=========================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept              1.1610      0.044     26.187      0.000       1.074       1.248
impute_work            0.3176      0.017     18.596      0.000       0.284       0.351
prtage                -0.0231      0.001    -15.495      0.000      -0.026      -0.020
np.power(prtage, 2)    0.0001   1.42e-05      7.136      0.000    7.33e-05       0.000
pesex_FEMALE          -0.0162      0.009     -1.735      0.083      -0.035       0.002
==============================================================================
```

Exercise 2.6. Correct for the attenuation bias in your results from Exercise 2.5. Is the bias corrected version larger or smaller? Does the bias-correction change your results from earlier?

- 0.44363979629861533
- The bias corrected version is larger but it does not change my results substantially from 2.5. The direction of the relationship between vote and flexible work schedule is still positive. The larger coefficient can be indicative of a larger magnitude of the relationship.