|  |  |  |
| --- | --- | --- |
| **AIN 380** | **Homework #6**  **Multiple Linear Regression ML** | **Due:** |

The application of different environmental factors has an effect on design patterns. The desirability factor of various design patterns has been scored along with a measure of the environmental factors present at the time the design was acquired.

**Multiple Linear Regression ML Models – using** *statsmodels.api*

1. Download the dataset named *Environs.csv* into a Pandas DataFrame. Confirm that all of the data is numeric and that there are no missing data values. **<3pts>**
2. Knowing that R2 is a measure of the goodness of fit of a linear regression model, determine what combination of environmental factors is most predictive of a design pattern, given the data in the file named Environs.csv. To assist you in this effort, a chart has been prepared that you need to complete below. **<7pts>**

|  |  |
| --- | --- |
| **Linear Regression ML Model using these independent variable(s) to predict Design** | **R2 value from a *statsmodels* OLS fit ML** |
| Pressure | 0.013 |
| Time | 0.010 |
| Heat Factor | .721 |
| Time and Pressure | .034 |
| Time and Heat Factor | .721 |
| Pressure and Heat Factor | .859 |
| Pressure and Time and Heat Factor | .874 |

1. Print the linear equation for predicting Design for the model using all three of the features (i.e. Pressure, Time and Heat Factor) as independent variables. Make sure this equation is well labeled in your output and is of the form y = a\* pressure + b \* time + c \* heat\_factor + y-intercept, where a, b, c, and y-intercept are values obtained from the model. **<4pts>**
2. In your Python code, calculate predictions for the Design value using the same model as in #3 for the following Pressure, Time, and Heat Factor values, respectively. The output must be well labeled. **<3pts>**

|  |  |  |
| --- | --- | --- |
| **Pressure** | **Time** | **Heat Factor** |
| 34.85 | 34.6 | 128.48 |
| 31.84 | 35.56 | 132.08 |
| 38 | 33.99 | 133 |

1. If management asks “How much will the predictive quality of the ML model be reduced if only two independent features are used rather than three, **and** which two independent features should be used as the predictor?”, what will be your answer? Explain here. **<4pts>**

\_\_\_\_After looking back at the r-squared values in question three, the r-squared value for Pressure and Heat Factor is very similar to the r-squared value for all three predictors, so I would say that the predictive quality would be lowered, but not by that much. The two independent features that should be used as the predictor should be Pressure and Heat Factor. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Some AI specialists attempt to create a machine learning model using Linear Regression but quit when they fail to find a single feature that yields a significant R2 as an independent variable. Why are they wrong to quit at that point? Explain here. **<3pts>**

\_\_\_They are wrong to quit because Linear Regression is not the only Machine Learning model that they could use, as they just need to find the model that works the best for their dataset. They are also wrong to quit because there may be other reasons that the model isn’t working, such as missing or incorrect data which can heavily throw off the accuracy of the model. The final reason that they are wrong to quit is because there is a chance that one SINGLE feature isn’t a good predictor, but maybe a combination of different features is needed to make a better predictor.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Print your Python code file and output and submit along with this sheet for credit. In addition, upload your Python code in a zip file with the data file to the course Canvas site.