Homework 4 - Linear Regression

NYC property sales + Boston data set (100 points)

Due Date: Monday Feb 22 at 11:59 pm

Instruction:

- This is a group-work assignment!
- You are expected to submit the **.ipynb** file and the exported **.html**.
- Only one member in each group needs to submit the assignment. It will be automatically submitted for the rest of group members.
- This is a long assignment, start early!
- You will be qualified to get full mark if you beat the following performance metrics:
  - Question 1: RMSE_test = 0.6
  - Question 2: RMSE_test = 5.5

Question 1 Linear regression : NYC property sales dataset (60 points)

In this exercise I want you to apply linear regression model to the NYC property sale data set that you cleaned in HW-2 EDA for NYC property sales dataset on Kaggle. You can also use my version of the dataset which is on the GitHub folder for HW4. Import the nyc-rolling-sales_clean.csv as a data frame and call it df. I specifically want you to do the followings:

1. Change the type of the feature variables as you see fit! You can use my answer key for HW2 as a reference. (5 points)
2. Define your target variable as target= log(SALE PRICE) and add it to your data frame. Explain why this transformation would boost the performance of your linear model? (5 points)
3. Define your feature space (X). You can pick as many features as possible! it's your call! (5 points)
4. Use get_dummies(drop_first=True) function from pandas package to make the categorical variables into dummy variables. How many features you have now? wow! welcome to Machine Learning. (5 points)
5. Split the data into test (30%) and train set (70%) (5 points)
6. Use LinearRegression() model from Sklearn package to train the model. Do the followings: (15 points)
    1 Save the predicted values for the test set in y_hat_test. (5 points)
    2 Construct a data frame named log_predictions which has 3 columns: y_test, y_hat_test , resid.(5 points)
    3 Report the RMSE_test (RMSE in the test set) (5 points)
7. Estimate the RMSE_test using K-Fold Cross Validation technique (try K=5 and K=10) and name them as RMSE_CV5 and RMSE_CV10. (15 points)
8. Compare RMSE_CV with RMSE_test from part 3 and explain your observation? (5 points)

Question 2 Polynomial regression: Boston dataset (60 points)

In this exercise, you should work with the boston_polynomial.csv file which is available on the GitHub folder for HW4. Import the boston_polynomial.csv as a data frame and call it df_poly. I specifically want you to do the followings:

1. Define x= np.array(df_poly['LSTAT']) and y= np.array(df_poly['price']). Draw a scatter plot for price vs LSTAT using x and y. (5 points)
2. Import PolynomialFeatures class from sklearn.preprocessing. Now fit_transform your x and call it X_poly. Set polynomial **degree = 5**. (5 points)
3. Split the data into test (30%) and train set (70%) (5 points)
4. Use LinearRegression() model from Sklearn package to train the model. Do the followings: (15 points)
    1 Save the predicted values for the test set in y_hat_test. (5 points)
    2 Construct a data frame named predictions which has 3 columns: y_test, y_hat_test , resid.(5 points)
    3 Report the RMSE_test (RMSE in the test set) (5 points)
5. Estimate the RMSE_test using K-Fold Cross Validation technique (K=5 only) and name it as RMSE_CV5. (10 points)
6. Use my_polynomial_regression() function from the notebook for class 7. With that function, construct a table with 3 columns: Degree (going from 1 to 10), RMSE_train and RMSE_test. (10 points)
7. Use the table from part 6 and plot the RMSE_test and RMSE_train against the Degree on the horizontal axis. (5 points)
8. What is the optimal polynomial degree based on your observations from the above table and chart in part 6 and 7 respectively. Explain your answer (5 points) This is called the **elbow method** by the way!