AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING אפקה המכללה האקדמית להנדסה בתל-אביב

Machine Learning 2022 Exercise 4

For this exercises you will participate in a Kaggle competition – "House Prices - Advanced Regression Techniques".

You will perform a regression task.

You are required to house prices in the city Ames, Iowa, USA

The competition dataset provide a large amount of features for you to play with. You are welcome to remove features you deem irrelevant, or calculate new features.

The data description is available at: https://www.kaggle.com/c/house-prices-advancedregression-techniques/data . Please make sure to read it carefully, make sure you understand the meaning of each column and how each feature might affect the price of a house.

If you have question about the competition or about the data, you are advised to refer to the discussion section of the competition: https://www.kaggle.com/c/house-pricesadvanced-regression-techniques/discussion.

You may take inspiration from the work of other participators, in the notebook section of the competition: https://www.kaggle.com/c/house-prices-advanced-regressiontechniques/notebooks.

Your predictions to the competition will be measured using the RMSE metric.

Requirements:

1. Your entire solution has to be written in a single python notebook (ipynb file)
2. Your submission to the Moodle has to include your solution file as an ipynb and as an html notebook (please refer to the attached guidance how to download your jupyter notebook as html) zipped into a single zip file.
3. The name of your files should begin with your full name and ID number.

Written by: Ido Meroz, IdoMe@afeka.ac.il Based on the work of Mr. Aviad Malachi

4. The first MD –Mark Down cell of your submitted notebook has to include your full name, id number and a link to your Kaggle account.
5. You don't have to repeat the EDA- Essential Data Analysis phase. Use the EDA from exercise 2.
6. This exercise should be a continuation of exercise 2. Add a level 1 title (using a single #): "Exercise 4" and present your work in the cells underneath this title.
7. You may rearrange and improve work you have done in exercise 2.
8. Use random_state = 42.
9. For your regression task you need to use **all** the following algorithms:
    - LWLR - Locally Weighted Linear Regression.
    - KNN – K nearest Neighbors
    - Decision Trees
    - SVM – Support Vector Machines with one kernel or more.
10. Explain the differences between the algorithms, analyze the results, and present the algorithm that achieved the best results.
11. Use PCA for dimensionality reduction.
12. Try different Hyper Parameters.
13. You are advised to use various Ensembles methods with different models.
14. Make sure to explain and analyze your results.
15. Attach a screenshot of up to your 10 most recent submissions. Emphasize your best submission. In addition attach a picture of your location and score on the leaderboard to your notebook.
16. Use a MD cell to summarize your work (5-20 lines), explain your work, what worked well and what did not work well.
17. Write a short TL;DR – To Long Didn't Read (5-10 lines) at the second MD cell of your notebook (under your full name and your ID number). Use the TL;DR to explain what you will be doing make, mainly what worked well.
18. Your last cell should hold references to resources you used including notebooks you took inspiration from, links, books etc.

Written by: Ido Meroz, IdoMe@afeka.ac.il Based on the work of Mr. Aviad Malachi

AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING אפקה המכללה האקדמית להנדסה בתל-אביב

Notebook Structure:

1. Your name ID Number and link to your Kaggle account.
2. Follow the Structure from exercise 2.
3. "#" title: "Exercise 4"
4. TL;DR for Exercise 4
5. Experiments you made with PCA, different models, ensembles and hyper parameters search.
6. Graphs and results analysis.
7. Screenshots of submission and place in the leaderboard as mentioned above.
8. Summary
9. References

Grade Structure

1. Simple, organized explained and clean code 10%
2. Organized, understandable and explained notebook 10%
3. Effort and self-learning 10%
4. Correct implementation of the requirements and valid notebook structure 70%
5. 10% bonus for extreme effort.

Remarks

1. Show and analyze the effect of PCA on the models use chose.
2. Chose a model and an Ensemble method you deem right for the data and explain your choice.
3. You should implement function in order to avoid code duplication.
4. Functions you implement should contain explanations of what they do.
5. You are advised to use meaningful names for variables and functions.
6. Make sure to show your understanding, what worked and what didn't work and try to explain it.

GOOD LUCK!

Written by: Ido Meroz, IdoMe@afeka.ac.il Based on the work of Mr. Aviad Malachi