COMP 468 Machine Learning in Python, Spring 2023
Instructor: Zafer Aydın
Homework 1

1. Work on the exercise notebook of Lesson 7 of Intro to Machine Learning course. In that notebook, we used the following features only

```
features = ['LotArea', 'YearBuilt', '1stFlrSF', '2ndFlrSF',
'FullBath', 'BedroomAbvGr', 'TotRmsAbvGrd']
```

Re-select features from the list of the following features

- 'MSSubClass'
- 'LotArea'
- 'OverallQual'
- 'OverallCond'
- 'YearBuilt'
- 'YearRemodAdd'
- '1stFlrSF'
- '2ndFlrSF'
- 'LowQualFinSF'
- 'GrLivArea'
- 'FullBath'
- 'HalfBath'
- 'BedroomAbvGr'
- 'KitchenAbvGr'
- 'TotRmsAbvGrd'
- 'Fireplaces'
- 'WoodDeckSF'
- 'OpenPorchSF'
- 'EnclosedPorch'
- '3SsnPorch'
- 'ScreenPorch'
- 'PoolArea'
- 'MiscVal'
- 'MoSold'
- 'YrSold'

Your feature set should satisfy the following properties

- Read the feature descriptions on this page

  https://www.kaggle.com/c/home-data-for-ml-course/data

  and select features that could be most useful for predicting house price.

COMP 468 Machine Learning in Python, Spring 2023
Instructor: Zafer Aydın
Homework 1

- Update the exercise notebook of lesson 7 of Intro to Machine Learning course using the features you selected. Do not use any imputer (i.e. for filling missing values since this should not be necessary at all) or encoder (e.g. for converting categorical features to numeric) in your code.
- Run the notebook, call the train_test_split with random_state set to 1 as we did in class to generate validation set. Compute the MAE on validation set by training the random forest regressor using X_train, y_train.
- Repeat selecting features (by including or removing features) and minimize the MAE on validation set. Make sure it is lower than 17300.
- Submit your predictions on test set to Kaggle challenge as we did in class (by training the random forest regressor on full version of training set, i.e. X, y). Compute the leaderboard score.

(a) What is the MAE on validation set and on test set (i.e. leaderboard score) if you use the original feature set (with 7 features) given in the exercise notebook of lesson 7 (also given at the beginning of this homework)?

(b) Which feature set gave the best MAE on validation set after you modified the feature set?

(c) What is the best MAE value you obtained on validation set using the feature set you reported in part (b)? What is the MAE on test set (i.e. leaderboard score) using the same feature set? Do you get improvement in MAE scores obtained in part (a)?

Submission

Once you finish, click File and Download Notebook. Submit your notebook with .ipynb extension to Canvas. Your notebook should include the best performing feature set you found based on the MAE on validation set. Submit your answers to parts (a) and (b) as a text or Word document to Canvas.