Task 2 coding is fully done.

1. Data Cleaning:

Imputation

Out put: Clean and consistence data set.

Combined data set -- Thursday . csv format

1. Data Exploration: ignore or not
2. Model building: feature selection technique, algorithm selection, K-folder technique
3. Prediction: average accuracy comparison

No need to use PCA for this case, but worth to have a try.

|  |  |  |  |
| --- | --- | --- | --- |
| Model 1: Regression Model | Model 2: Classification model | Due |  |
| Output: Highest demand for one day  Features: All features  Combine 2 data set together. | Output: Highest price category for each day  Features: All features  Combine 2 dataset together | Wed | 1. Fill them by some judgement (correlation)-Sean 2. Simplefill 3. Delete Null rows (6 rows) |
| Feature selection | Feature selection |  |  |
| Linear regression: | KNN or Decision tree for classification |  |  |
| Split the model, K-fold method(parameter) | Split the dataset, K-fold method(parameter) |  |  |
| Run to see the average accuracy. | Run to see the average accuracy |  |  |

Clean method:

Model: Play around based on Tuesday’s workshop. Let’s touch up on Thur evening

Feature selection

KNN or Decision tree for classification

Split the dataset, K-fold method

Run to see the average accuracy

featrues select:

1. chi2

2. MI

Liner reggress NUMBER

Classification

Discion tree

Knn

evaluate the following:

Liner: mean2 r2

Classification: accuracy score

Visualization!

Report

Greedy Ins only if we have time

K fold: split data if we have time