Panel Cross-Validation in ML

Augusto Gonzalez-Bonorino

Zillow Housing Value Dataset

- Housing price and valuation panel dataset
  - Monthly dates (time unit) - 04/1996 to 12/2017
  - States (entity unit)
- 13212 observations
- 82 features (16 with less than %50 of NaN)
- 1 target (ZHVIPerSqft_AllHomes)
  - Zillow Home Value Index x ft^2 for all homes by state

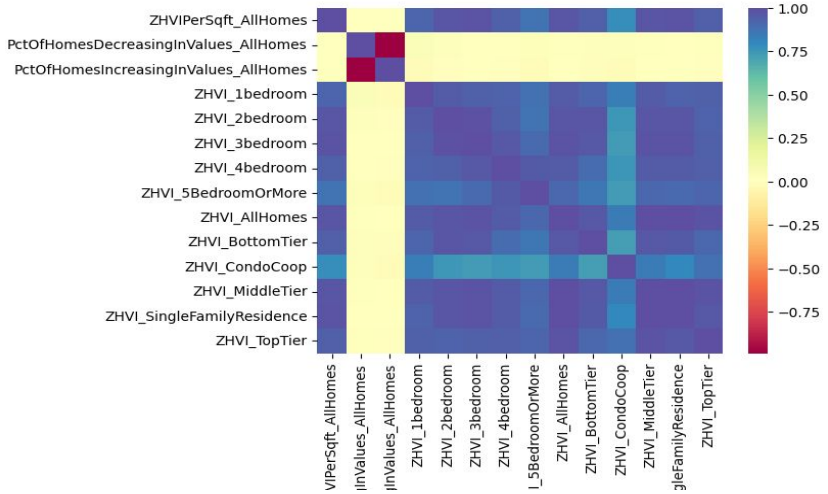
Preprocessing

- Drop features with %50 > NaN
- Drop states with missing dates
  - To balance groups
- Group based mean-imputation
- If not enough data, impute with feature's mean

- Group-based scaling for NN dataset
- No transformation for tree models
- Engineered Month, Day, Year

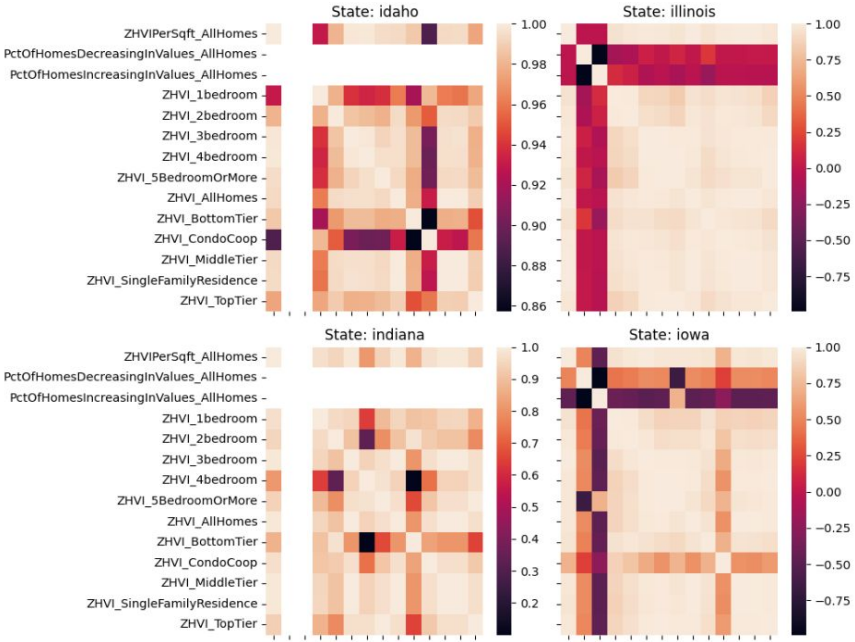
- 12004 obs, 18 vars for training

```
((12004, 16),
 Date                                      0.000000
 RegionName                                0.000000
 ZHVIPerSqft_AllHomes                      0.042069
 PctOfHomesDecreasingInValues_AllHomes     0.276325
 PctOfHomesIncreasingInValues_AllHomes     0.276325
 ZHVI_1bedroom                             0.174608
 ZHVI_2bedroom                             0.099300
 ZHVI_3bedroom                             0.006664
 ZHVI_4bedroom                             0.061480
 ZHVI_5BedroomOrMore                       0.089220
 ZHVI_AllHomes                             0.054898
 ZHVI_BottomTier                           0.065062
 ZHVI_CondoCoop                            0.117877
 ZHVI_MiddleTier                           0.054898
 ZHVI_SingleFamilyResidence                0.054898
 ZHVI_TopTier                              0.033239
```

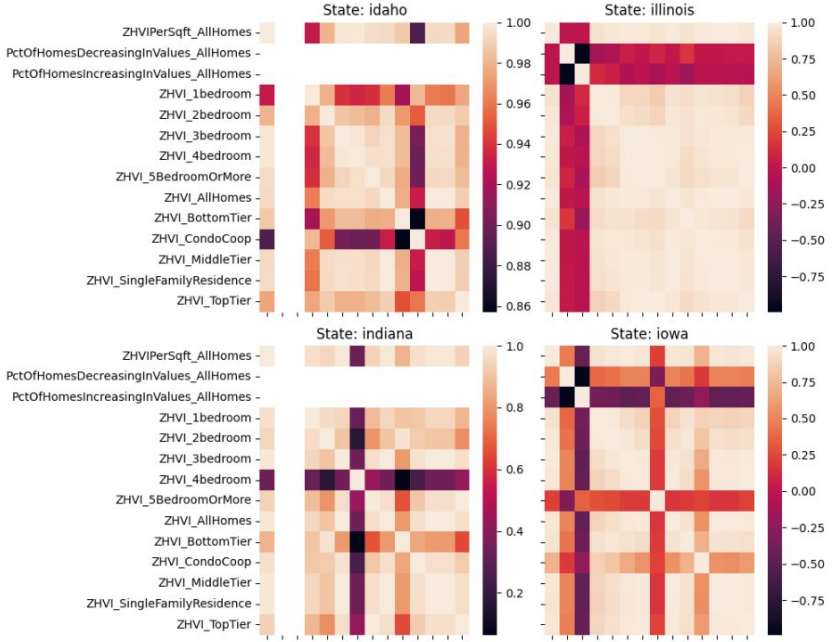


Check your groups...

Pre-Imputation



Post-Imputation



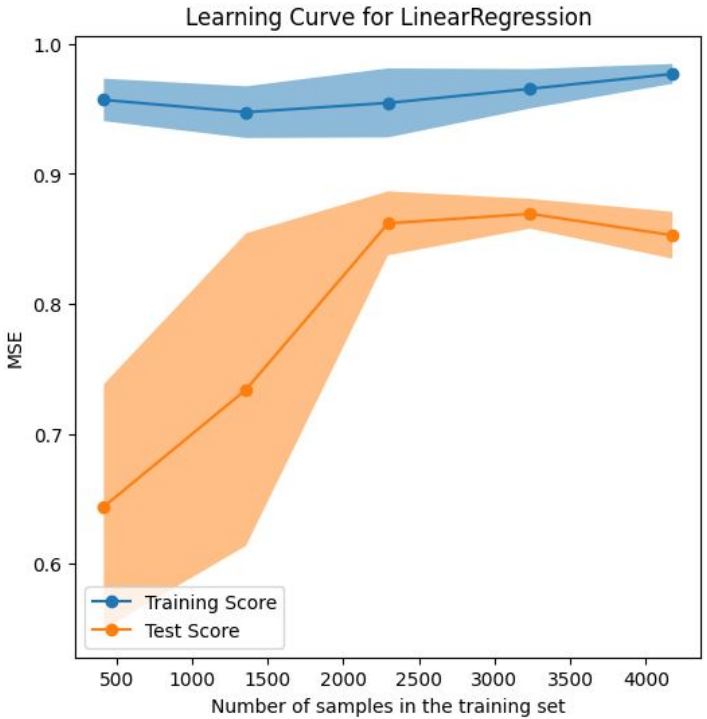
Cross-Validation

- **Goal:** get disjoint and sequential sets
- K-Fold(sklearn's default for regressors)
- TimeSeries Split (sklearn's CV generator for ts)
- Stratified K-Fold(sklearn's default for classifiers)
- Panel (not implemented)

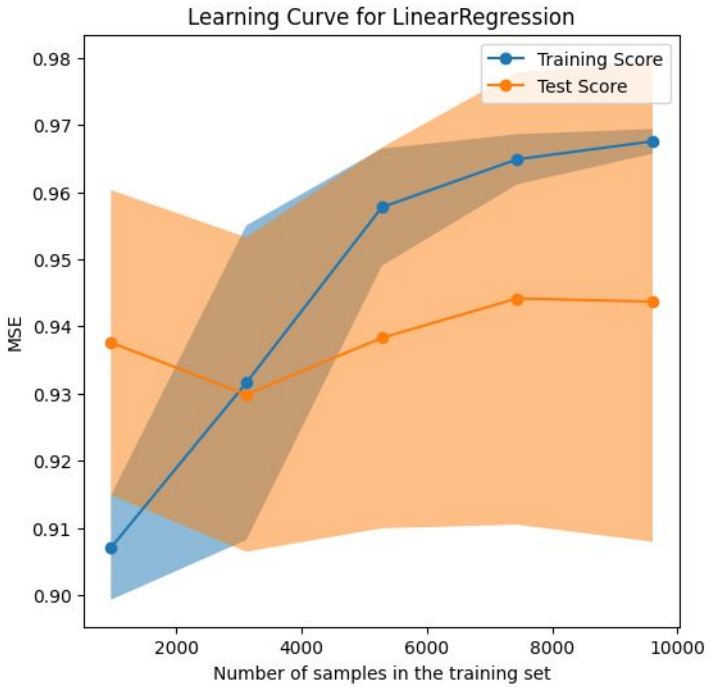
```
Split 1
X train: (4176, 64),     X test: (3915, 64),
Y train: (4176,),        Y test: (3915,)
States in train set: [ 1  4  5  6  8  9 10 12 13 15 16 17 18 19 20 21]
States in test set: [22 23 24 25 26 27 28 29 31 32 33 34 35 36 37]

Split 2
X train: (8091, 64),     X test: (3913, 64),
Y train: (8091,),        Y test: (3913,)
States in train set: [ 1  4  5  6  8  9 10 12 13 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
 31 32 33 34 35 36 37]
States in test set: [39 40 41 42 44 45 47 48 49 51 53 54 55 46 50]
```

Results (linear regression)

Panel 2 Splits*

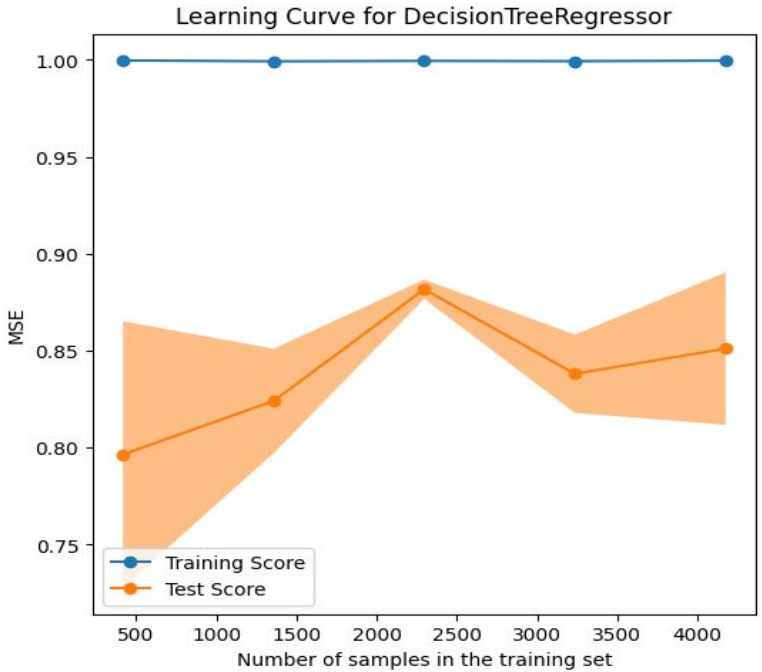


Sklearn cv=5

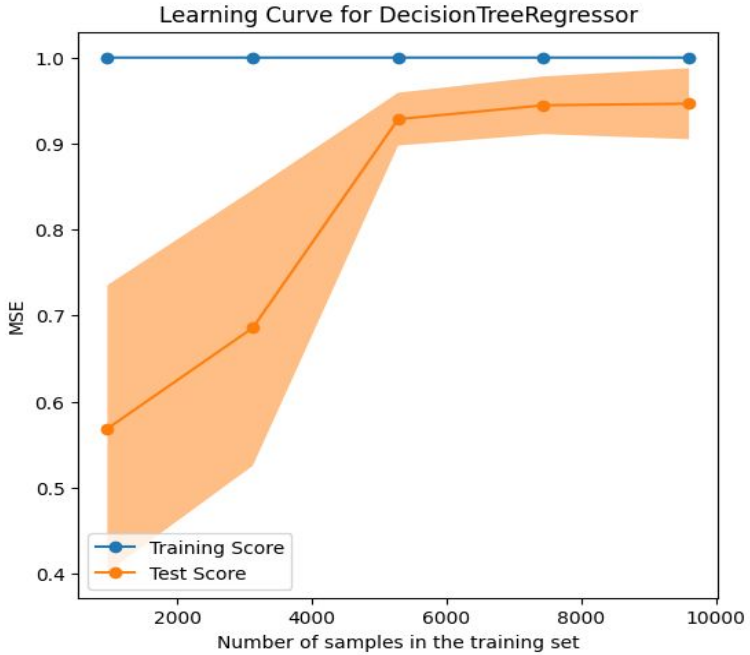


Results (decision tree)

Panel 2 Splits

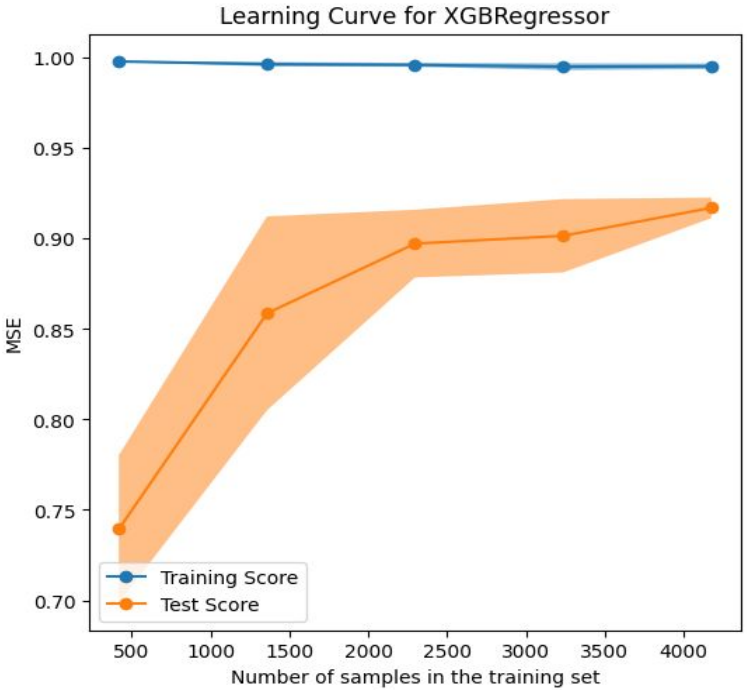


Sklearn cv=5

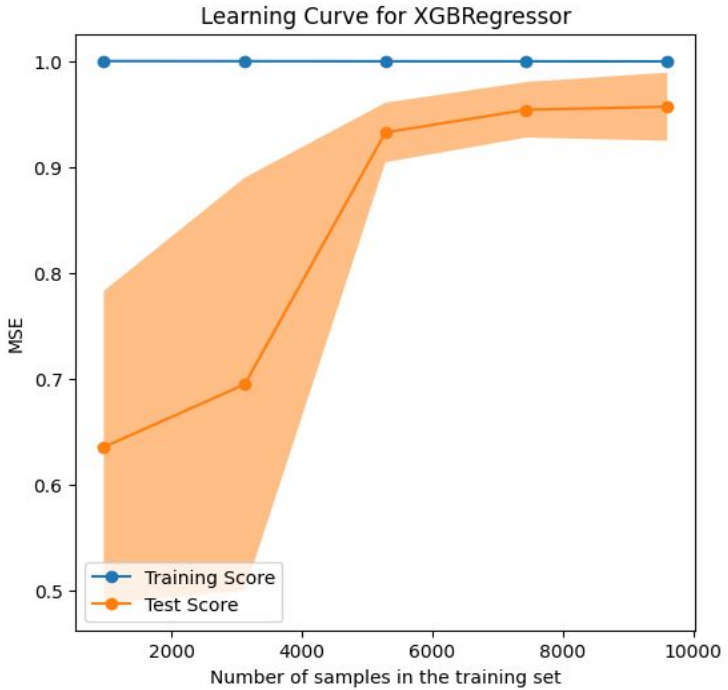


Results (cont'd)

Panel 2 Splits

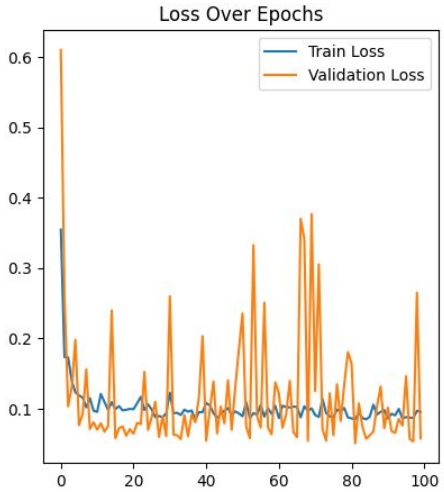


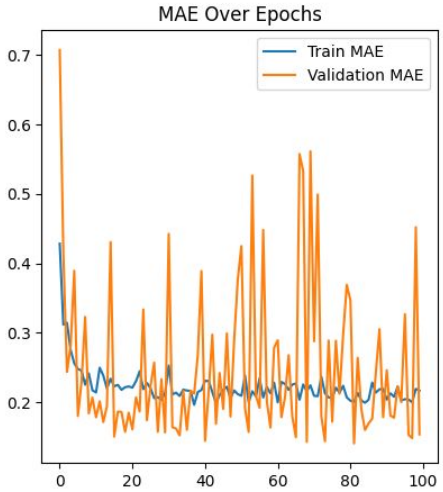
Sklearn cv=5



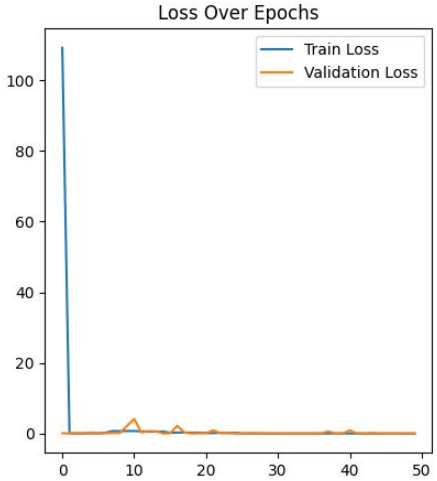
Results (neural networks)

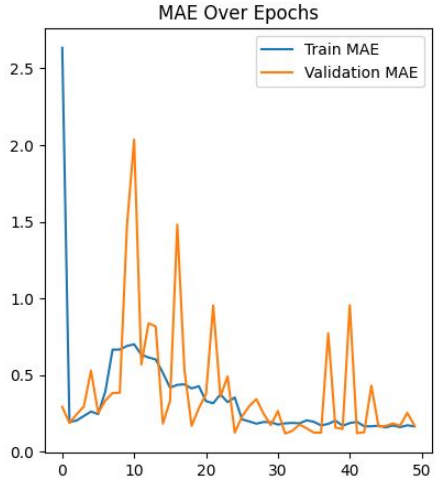
TanH - 100 epochs - batch size 16 - Adam





ReLU - 50 epochs batch size 16 - Adam





- 3 hidden layers, fully connected, regular train-test split, FIPS one-hot encoded

Conclusion

- Thinking about the dataset structure matters
- Panel CV split:
  - Demands larger datasets to address curse of dimensionality
  - More interpretable results but less predictive power
  - Group based predictions allow for easier comparative analysis
  - Stabilizes linear regression training
- Tree models handle panel data well
- Could neural networks benefit from a panel CV split?
- If dataset structure is respected, can this approach help bridge ML and Econometrics?

Thank you!

Results (Encoded FIPS)