San Diego House Price Predicting

**Wen Yan, Mengting Wang, Salah Ahmad, Xia Song**

### Validation methodology

Again, our modeling is to predict the a house’s sold price under current market. Since it only makes sense to train the model from previous transaction data and predict the transaction to happen, traditional K-fold cross validation won’t work since its division on training and testing data is orderless. Generally, we’ll use walk-forward validation method with RMSE as metrics for our model evaluation.

**Evaluation Metrics**: RMSE on target variable ‘sold\_price’

**Walk-forward validation**:

It’s good for time-series prediction model evaluation because it’s checking the stability of the model performance over time. While the traditional K-fold cross validation is not appropriate because it’s not reasonable to train the data from later time for prediction on previous time. Figure 1 below explains how walk-forward validation works, it’s still doing X-fold validation with each parameter combination from a parameter grid and select the optimized parameter set based on the average validation score.

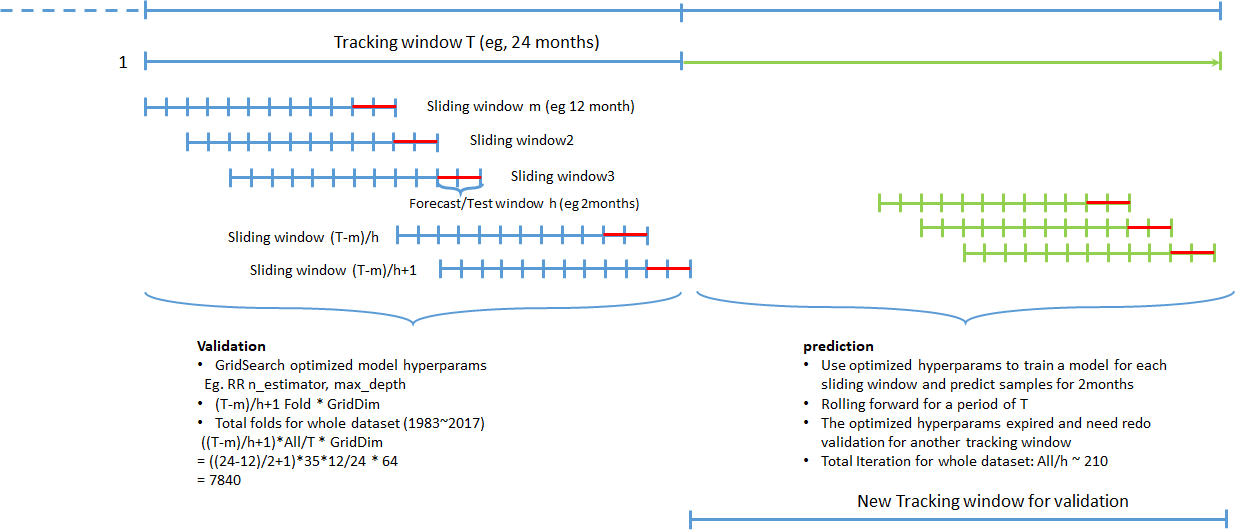


Figure 1. Walk-forward validation with sliding window

**Time windows:**

* Forcast/Testing window (h): the lifetime that trained model can be used for prediction before retraining.
* Sliding/Rolling window (m): the time length that data will be used for training plus the testing window
* Tracking window (T): the whole time range of data that can be used for validation, the data before this range will be considered too old to be relevant.

**Validation process:**

* With some baseline model, we lock down a grid of parameter sets for each model type which we want the validation process help to optimize. Eg RandomForestRegressor: n\_estimators = (10,20,30,40 ..), max\_depth = (4,5,6,7,8,9,10)
* Starting from the 1st tracking window T(eg 1 year), we do a full walk-forward validation for a grid search on the parameter ranges to get the best model parameters.
* With the selected model parameters, we train a model with data sliding window m(eg. 6 months) and use it to predict forecast window h(eg. 2 month), and roll forward for next 2 month prediction, until we predict T (1 year) new data.
* Go back to step2 to redo walk-forward validation for next T (1 year) model selection and prediction.

**Iteration needed:**

Similar to K fold cross validation which need K iterations to evaluate one config in the grid of modeling hyperparameters. Walk-forward validation need (T-m)/h + 1 iterations for one config.

Suppose the model hyper parameter grid to evaluate has dimension of D, and whole data history length is A, the total iteration for validation will be ((T-m)/h+1)\*D\*A/T, plus A/h of final model training for prediction.

Eg. With 1month as timestep, T = 12, m=6, h=2, D=64, A=(2017-1983)\*12, total iteration number is 8960 + 420

### Sliding window selection

So walk-forward validation helps on optimizing the model’s hyperparameters, but it also introduce 3 new parameters of time windows. We need determine the each window size. The bigger window the validation and prediction use, the more data the modeling can use, but the more possible the time factor can be involved. And very likely, the optimized window size can vary along time.

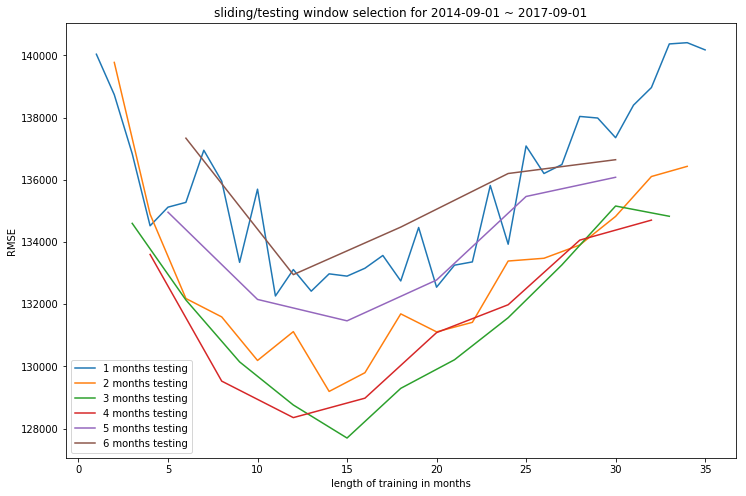
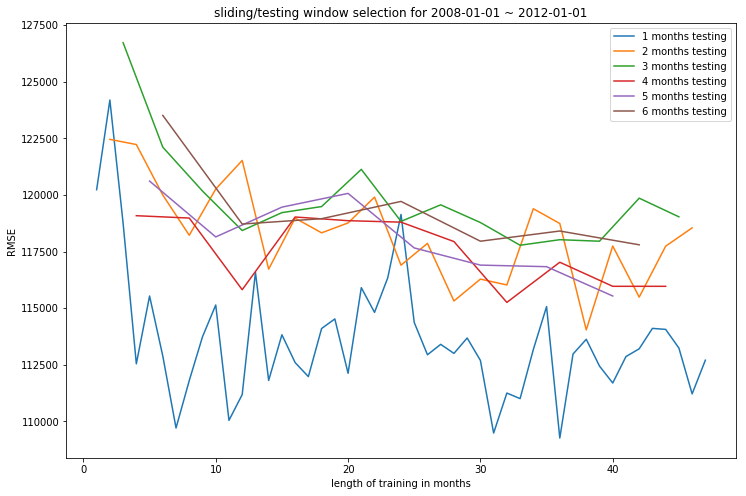
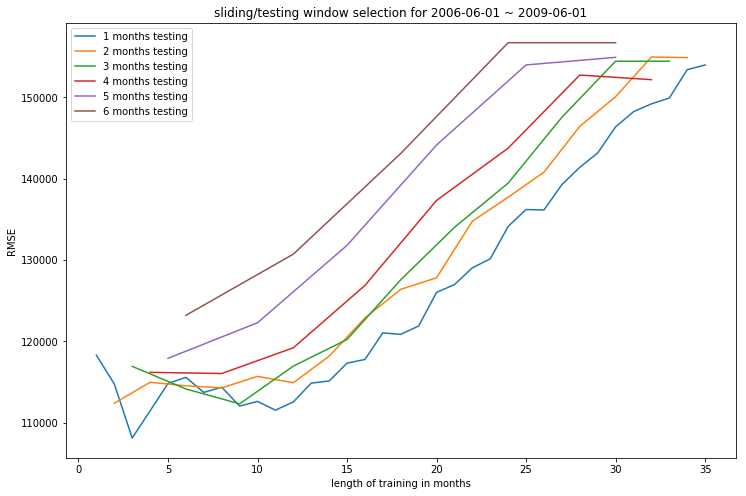
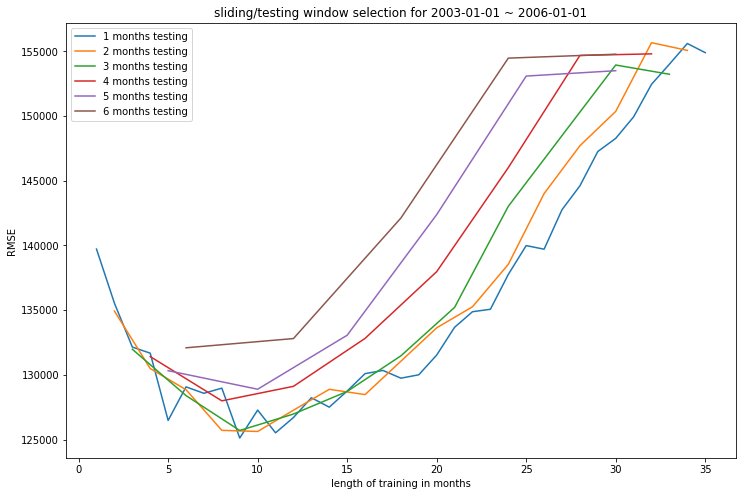


Figure 2. RMSE of various sliding/testing window

at different time (2006-01, 2009-06, 2012-01, 2017-09)

From figure 2, we can see under rapid changing market, the shorter training period the better performance the model can achieve (eg, 2009-06), while under stable market, training window size doesn’t matter much (eg, 2012-01), and in other cases, the model get best performance at some adjacent point.

In general, at different market environment, the window size need be validated to achieve optimized modeling performance.

### Baseline Model Result

### Next steps for improvements

### More data preprocessing for outliers

xxx

### Multiple Segmentation and modeling

xxx

### More feature engineering

Year\_built

School score, distance

### External data