**AC209a project workflow checklist**

*This link is helpful, please take a look:*

<http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html>

**Workflow:**

*The following lists work flow that we need to complete for the project.*

1. *Please code everything in ipynb for easy interpretation instead of py.*
2. *At the top of each ipynb, please add remark to briefly describe what the file is for (see ipynb that I reorganized).*
3. *Please add some text to each code block when you code ipynb so that we can quickly put code and text in the website later.*

**Load data (01\_load\_data.ipynb)**

* Load data do minimal data cleaning
* Fill missing data with mean and mode just for now
* Deliver ***[listings\_cleaned]***
* Deliver ***[calendar\_cleaned]***
* *Deliver* ***[reviews\_cleaned]***

**Data exploration**

**(021\_visualize\_listings.ipynb)**

* Visualize listings

**(022\_visualize\_calendar.ipynb)**

* Visualize calendar

**(023\_visualize\_reviews.ipynb)**

* Visualize reviews (by NLTK)

**Baseline model (03\_baseline.ipynb)**

* Fit OLS ***[listings\_cleaned]*** using listing data

**Data preprocessing and reduction**

**(041\_preprocessing.ipynb)**

* Do log(price)
* Do KNN clustering of location to replace zip code, longitude and latitude
* Add prior to the review scores
* Add calendar and review data to listing
* Deliver a dataset ***[data\_w\_calendar\_revew]***

**(042\_predictor\_selection.ipynb)**

* Do model selection using data\_w\_calendar\_revew
* Do OLS lasso using data\_w\_calendar\_revew (rank parameter)
* Compare the above two results and select most important predictors (decide a rule)
* Fill missing data again
  + Drop entry if missing important predictor like location
  + Try fill using interpolation?
  + Try fill using ‘drawn from distribution’?
  + Try KNN of most related predictors?
* Deliver a dataset for regression ***[data\_select]***

**(043\_reduce\_dim.ipynb)**

* Apply PCA to ***[data\_select]*** (keep 95% of explained variance)
* Deliver a dataset for regression ***[data\_pca]***

**Data modeling**

* Use both ***[data\_select]*** and ***[data\_pca]***

**(051\_linear\_regressoin.ipynb)**

* **A.** Try OLS / OLS Lasso / OLS Ridge / Elastic Net

**(052\_quadratic\_regressoin.ipynb)**

* **B.** Try quadratic regression + Lasso + Ridge + Elastic net

**(053\_tree\_regression.ipynb)**

* **C.** Try tree regressors (decision tree + bagging + boosting – AdaBoost and Gradient Boost)

**(054\_SVR.ipynb)**

* **D.** Try SVR

**(055\_ensemble.ipynb)**

* Try ensemble from best of **(A, B, C, D)** with weightings from training scores

**(056\_neural\_network.ipynb)**

* Try neural network
* **Models not covered in class:** elastic net, gradient boost, neural network

**(057\_chosen\_model.ipynb)**

* Do cross validation and fine tuning of the above models
* Can we calculate prediction interval by bootstrapping?

**(058\_to\_classification.ipynb)**

* Try put price to bucket according to quantile, this transforms the problem to be a classification.
* Try KNN, LDA, QDA, decision tree, random forest, SVC
* Do cross validation and fine tuning of the above models.
* Briefly compare results to the chosen regression model.

**Construct a website:**

* Detail above process, put ipynb for download
* Design an interface for users to input (including location, physical criteria, date, and whether review matters)
* Do soft recommendation, vary location, criteria, and date and give predicted price for user reference.

**To do list:**

**Load data**

(done)

* Minimum cleaning to load the data correctly
* Non-English reviews have been cleaned out

(to do)

* (Necessary?) Double check cleaned data to make sure data are loaded correctly

**Exploration**

(done)

* Plotted some visuals for listings and calendar, but not reviews

(to do)

* Plot price spike of 365 days from calendar\_grid
* How to visualize text?

**Preprocessing**

(done)

* Log(price) transformation
* Predictor\_selection for OLS
* Location clustering by KNN to replace zipcode, longitude and latitude
* Fill missing values (numerical – mean, categorical – mode)

(to do)

* Try bucketing price (become a classification problem)
* Location clustering by KNN – overlay results on a NY map
* Try take out entries with missing values (fit only complete data)
* Try interpolation to fill missing values
* Try draw from distribution to fill missing values
* Systematically do PCA (dimension reduction) check variance explained
* Check if standardization makes any difference (especially after PCA)
* Do stepwise model selection
* Review comments analysis by NTLK or other package, extract reviews as features
* Check correlation between review comment features and review scores

**Modeling**

(done)

* OLS - before and after log(price)
* Random forest
* Neural network (4 layers)

(to do)

* Check residual vs fitted value in all models
* Shrinkage - try regularization (Lasso and Ridge) for OLS
* Try quadratic linear regression
* Try bagging and boosting decision trees
* Try gradient boosting decision trees
* Try ensemble regressors
* Try neural network (4 layers) with holiday coded data (more entries now)
* Try SVR with reduced predictors (from preprocessing)

**Validation**

(done)

* Cross validation in OLS

(to do)

* Cross validation in all models