Minutes\_0 (June 2nd) :

* Task: [Problem Overview](http://cikm2017.org/CIKM_AnalytiCup_task3.html)
  + Target\_0: IsClear: Binary Classification, Regression
  + Target\_1: IsConcise: Binary Classification, Regression
* [Dataset](https://drive.google.com/drive/folders/0B2ikXQtYRMajYzhqc0h0OUVMbUk?usp=sharing), and [Schema](https://competitions.codalab.org/competitions/16652#learn_the_details-evaluation)
* External dataset: [Official External Data Thread](https://competitions.codalab.org/forums/13392/)
* Competition Ends July 31, 2017, 11:59 p.m.
* Steps:
  + Cleansing
    - Noises
      * Html tags in *‘short\_description’* (%94)
    - Missing values (NA, NULL)
      * ‘*product\_type‘* (less than %1)
      * *‘category\_lvl\_3’* (about %6)→ assign *‘category\_lvl\_2’*
      * *‘description’* (less than %1)
    - Outliers
      * *‘price’* {-1, 999999, 9999999}, Normalization based on country
  + Exploration (Flirting with data)
    - Data and targets distribution
      * Only 3 combinations for (IsClear, IsConcise): (1,0), (1,1), (0,0)
      * P(1,0) = 0.2580,P(1,1) = 0.6853,P(0,0) = 0.0566,P(0,1) = 0.0000
      * **P(1|0) = 0.8200**,P(1|1) = 1.0000,P(0|0) = 0.1800,P(0|1) = 0.0000
    - Data and [targets correlation](http://scikit-learn.org/stable/modules/tree.html#multi-output-problems) ([wiki](https://en.wikipedia.org/wiki/Multi-label_classification))
      * Count(~IsClear & IsConcise) = 0 → if ~IsClear then ~IsConcise
      * If IsConcise then IsClear
    - Data Augmentation (generating pos and neg sample from real data)
      * Training data
      * Test/Validation data: average over the expanded and original data point
  + **Feature engineering:** 
    - Extraction
      * [excel sheet](https://docs.google.com/spreadsheets/d/1P1KaeQ_XlvzRQpRyMxmRChfGcReODUrp9BQXx58XvzU/edit?usp=sharing)
    - Selection
      * variance/chi2/f-test/mutual-information
      * stability selection
      * recursive feature elimination and cross-validation
  + [Model Engineering](http://scikit-learn.org/stable/_images/sphx_glr_plot_classifier_comparison_001.png) ([Estimators](http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py))
    - [Decision Tree](https://competitions.codalab.org/forums/13392/2089/) (DT)
    - Random Forest (RF)
    - [Logistic Regression](https://competitions.codalab.org/forums/13392/2087/) (LR)
    - ​[Naive Bayes](https://competitions.codalab.org/forums/13392/2088/) (NB)
    - SVD
    - Stacked Autoencoders
    - NNet (DL) ([Keras](https://keras.io/))
  + Model Tuning
    - Grid Search
    - Randomized
  + Model Evaluation
    - RMSE

Minutes\_1(June 6th):

* Extracting features discussed in [excel sheet](https://docs.google.com/spreadsheets/d/1P1KaeQ_XlvzRQpRyMxmRChfGcReODUrp9BQXx58XvzU/edit?usp=sharing)
* Hossein does the coding in python, others provide helping artifacts to him
* The first submission due date is ?
  + Submission specification
* Next session will be on the submission specification just before the submission

Further Reading/Lesson Learned

* [A Review of Multi-Label Classification Methods](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.6415&rep=rep1&type=pdf) ([codes](http://cse.seu.edu.cn/PersonalPage/zhangml/)) ([java](http://mulan.sourceforge.net/index.html))
* out-of-core/online learning
* <http://nbviewer.jupyter.org/github/donnemartin/data-science-ipython-notebooks/blob/master/scikit-learn/scikit-learn-intro.ipynb>
* <http://nbviewer.jupyter.org/github/donnemartin/data-science-ipython-notebooks/blob/master/kaggle/titanic.ipynb>

# [Vowpal Wabbit (Fast Learning)](http://hunch.net/~vw/)