**House Prices: Advanced Regression Techniques Project**

“Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.”

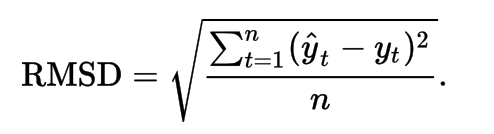
Link: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

**Introduction**

This is a Kaggle Knowledge competition that started on August 30, 2016 and will end on March 1, 2017. Consider the dataset as an augmented (and more modern) version of the widely used Boston Housing data set used in regression research. There are already many people on the Kaggle forums posting ideas and writing scripts with findings and so this will be a very good time for you to get involved! The data set features nominal, ordinal, and categorical variables that could impact a home’s sales price.

**Objective**

The great thing about Kaggle competitions is that the final objective is always very clear – predict the final price of each home with the least RMSE.



In addition to the accuracy, you and your teammates will be drafting an ipython notebook detailing your data cleaning, exploratory analysis, and modeling methods.

**Logistics**

Project submissions will be due by Friday midnight 9/30. Since this is not an in-class Kaggle competition exclusive to us, your team’s submission will be on the actual public leaderboard. Project notebooks are due before the start of class on Monday 10/3. Make sure your notebook is saved in your Docker folder somewhere for us to view.

**Notebook Guidelines**

1. Follow the general structure of notebooks you see on Kaggle:
   1. Basic data exploration with graphs, summary statistics, and outlier detections
   2. Data cleaning/Feature Engineering
   3. Model building
   4. Model evaluation
   5. Model interpretation
2. If you are in part of a team, you can split the work per section – but the best way to learn is to help do a little of everything. There are a lot of features and no person will produce the exact same analysis.
   1. Tip: You will only learn 2 types of models in this module (1 of which is most relevant) but **explore the others**.Even though we won’t get to some of them that appear in the kernels (e.g. xgboost), it is a great way to have the team try out different things that others have used even if you don’t immediately understand what it does. Plus, it will boost your accuracy over everyone else ☺
3. You may have noticed that some kernels are really detailed already. It is completely fine to show some of the same things. In fact, I learned Kaggle by copying what others did and then investigating other variables that may have been overlooked.
4. Get in the habit of asking how each variable relates to each other! Use tools like correlation and time series graphs to spot patterns and trends. These will help you understand the data set as well as get a good sense of what your model should be considering important.

**Grading**

We will set our own benchmark accuracy score and your model MUST beat it to pass the class. ☺

Points are distributed uniformly per team as follows:

70 points – Benchmark = 0.135 (RMSE)

.

.

100 points – Top team model.

The notebook will also be pseudo-graded for effort and quality on a 0-50 scale (10 points X 5 sections in the notebook structure). The first two project notebooks **will not** count to your grade but your final two project notebooks will. We give a score so you can gauge what our expectations are.

For each section, point values are outlined below:

1. Data Exploration

2 points for graphs of any kind

2 points for statistics of any kind

3 points for thoroughness (did you analyze enough to get a good understanding of the data?)

3 points for quality/creativity

1. Data Cleaning & Feature Engineering

5 points for cleaning the data of the most apparent problems (e.g. missing values, outliers, dummy variables, etc.)

2 points for feature engineering any variable

3 points for “creative/unique” feature engineering

1. Model building

5 points for setting up the train and test sets correctly

2 points for successfully running a model

3 points for successfully running more than one model

1. Model evaluation

7 points for correct out-of-sample test accuracy to the public leaderboard accuracy (within a reasonable threshold)

3 points for demonstrating that your local test accuracy is correlated with the public leaderboard accuracy

1. Model interpretation

4 points for interpretation of any kind

3 points for thoroughness (did you fully grasp what your model is telling you about the data?)

3 points for comparing different models