# House Prices: Advanced Regression Techniques

**Start here if...**

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

**Competition Description**



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.

**Practice Skills**

* Creative feature engineering
* Advanced regression techniques like random forest and gradient boosting

**Acknowledgments**

The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

**Goal**

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable.

**Metric**

Submissions are evaluated on [Root-Mean-Squared-Error (RMSE)](https://en.wikipedia.org/wiki/Root-mean-square_deviation) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

**Submission File Format**

The file should contain a header and have the following format:

Id,SalePrice  
1461,169000.1  
1462,187724.1233  
1463,175221  
etc.

You can download an example submission file (sample\_submission.csv) on the [Data page.](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)

**File descriptions**

* **train.csv** - the training set
* **test.csv** - the test set
* **data\_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
* **sample\_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

**Data fields**

Here's a brief version of what you'll find in the data description file. (red: Continuous category/ blue: numeric/ green: discontinue numeric/ black: category)

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* Id
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* ~~Alley: Type of alley access~~
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to main road or railroad
* Condition2: Proximity to main road or railroad (if a second is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Overall material and finish quality
* OverallCond: Overall condition rating
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 1stFlrSF: First Floor square feet
* 2ndFlrSF: Second floor square feet
* LowQualFinSF: Low quality finished square feet (all floors)
* GrLivArea: Above grade (ground) living area square feet
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* WoodDeckSF: Wood deck area in square feet
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* ~~PoolQC: Pool quality~~
* ~~Fence: Fence quality~~
* ~~MiscFeature: Miscellaneous feature not covered in other categories~~
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

**Other Tutorials in Python**

[**Comprehensive Data Exploration with Python**](https://www.kaggle.com/pmarcelino/house-prices-advanced-regression-techniques/comprehensive-data-exploration-with-python)

* Understand how variables are distributed and how they interact
* Apply different transformations before training machine learning models
* Covers both univariate and multivariate approaches
* Includes visualizations using [matplotlib](http://matplotlib.org/) and [seaborn](http://seaborn.pydata.org/)

[**House Prices EDA**](https://www.kaggle.com/dgawlik/house-prices-advanced-regression-techniques/house-prices-eda)

* Learn to use visualization techniques to study missing data and distributions
* Covers both continuous and categorical data
* Includes correlation heatmaps, pairplots, and t-SNE to help inform appropriate inputs to a linear model

[**A Study on Regression Applied to the Ames Dataset**](https://www.kaggle.com/juliencs/house-prices-advanced-regression-techniques/a-study-on-regression-applied-to-the-ames-dataset)

* Demonstrate effective tactics for feature engineering
* Includes simplifying and combining existing features
* Explore linear regression with different regularization methods including ridge, LASSO, and ElasticNet using [scikit-learn](http://blog.kaggle.com/author/kevin-markham/)

[**Regularized Linear Models**](https://www.kaggle.com/apapiu/house-prices-advanced-regression-techniques/regularized-linear-models)

* Build a basic linear model
* Try more advanced algorithms including [XGBoost](https://github.com/dmlc/xgboost) and neural nets using [Keras](https://keras.io/)

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