House Prices: Advanced Regression Techniques

**Predict sales prices and practice feature engineering, RFs, and gradient boosting**

**Description**

### 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.

要求购房者描述他们的梦想之家，他们可能不会从地下室天花板的高度或与东西方铁路的接近度开始。 但是这个游乐场比赛的数据集证明，对价格谈判的影响远远超过卧室或白色栅栏的数量。

有79个解释变量描述（几乎）爱荷华州埃姆斯的住宅房屋的每个方面，这个竞赛挑战你预测每个房屋的最终价格。

### Practice Skills

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

## •创意特色工程

## •先进的回归技术，如随机森林和渐变增强

## 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.

Ames Housing数据集由Dean De Cock编制，用于数据科学教育。 对于那些寻找经常被引用的Boston Housing数据集的现代化和扩展版本的数据科学家来说，这是一个不可思议的选择。

Evaluation

### 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.

预测每栋房屋的销售价格是你的工作。 对于测试集中的每个Id，您必须预测SalePrice变量的值。

### 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.)

提交数据是根据预测值的对数与观察到的销售价格的对数之间的均方根误差（RMSE）来评估的。 （记录日志意味着预测昂贵房屋和廉价房屋的错误会同样影响结果。）

### 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)

Frequently Asked Questions

### What is a Getting Started competition?

Getting Started competitions were created by Kaggle data scientists for people who have little to no machine learning background. They are a great place to begin if you are new to data science or just finished a MOOC and want to get involved in Kaggle.

Getting Started competitions are a non-competitive way to get familiar with Kaggle’s platform, learn basic machine learning concepts, and start meeting people in the community. They have no cash prize and are on a rolling timeline.

Kaggle数据科学家为几乎没有机器学习背景的人创建了入门级比赛。 如果您是数据科学新手，或刚完成MOOC并希望参与Kaggle，那么他们是一个很好的开始。

开始比赛是熟悉Kaggle平台，学习基本机器学习概念并开始与社区人员会面的非竞争性方式。 他们没有现金奖励，并且处于一个滚动的时间表上。

### What’s the difference between a private and public leaderboard?

The Kaggle leaderboard has a public and private component to prevent participants from “overfitting” to the leaderboard. If your model is “overfit” to a dataset then it is not generalizable outside of the dataset you trained it on. This means that your model would have low accuracy on another sample of data taken from a similar dataset.

Public Leaderboard

For all participants, the same 50% of predictions from the test set are assigned to the public leaderboard. The score you see on the public leaderboard reflects your model’s accuracy on this portion of the test set.

Private Leaderboard

The other 50% of predictions from the test set are assigned to the private leaderboard. The private leaderboard is not visible to participants until the competition has concluded. At the end of a competition, we will reveal the private leaderboard so you can see your score on the other 50% of the test data. The scores on the private leaderboard are used to determine the competition winners. Getting Started competitions are run on a rolling timeline so the private leaderboard is never revealed.

Kaggle排行榜有一个公共和私人组件，以防止参与者“过度拟合”到排行榜。如果你的模型对数据集“过度拟合”，那么它在你训练的数据集之外是不可推广的。这意味着您的模型对从相似数据集中获取的另一个数据样本的准确度较低。

公共排行榜

对于所有参与者，测试集中相同的50％预测被分配到公共​​排行榜。您在公开排行榜上看到的分数反映了您的模型在这部分测试集上的准确性。

私人排行榜

测试集中另外50％的预测被分配到私人排行榜。在比赛结束之前，参赛者无法看到私人排行榜。在比赛结束时，我们将显示私人排行榜，以便您可以在其他50％的测试数据中看到您的分数。私人排行榜上的分数用于确定竞赛获胜者。入门竞赛在滚动时间轴上运行，因此私人排行榜永远不会透露。

### How do I create and manage a team?

When you accept the competition rules, a team will be created for you. You can invite others to your team, accept a merger with another team, and update basic information like team name by going to the More < [Team](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/team) page.

We've heard from many Kagglers that teaming up is the best way to learn new skills AND have fun. If you don't have a teammate already, consider asking if anyone wants to team up in the [discussion forum](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion).

当您接受比赛规则时，将为您创建一个团队。 您可以邀请他人加入您的团队，接受与其他团队的合并，并通过转到更多<团队页面更新团队名称等基本信息。

我们从很多Kagglers那里听说，团队合作是学习新技能并获得乐趣的最佳途径。 如果你还没有队友，可以考虑询问是否有人想参加讨论论坛。

### What are kernels?

Kaggle Kernels is a cloud computational environment that enables reproducible and collaborative analysis. Kernels supports scripts in R and Python, Jupyter Notebooks, and RMarkdown reports. Go to the [Kernels tab](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/kernels) to view all of the publicly shared code on this competition. For more on how to use Kernels to learn data science, visit the [Read more about our decision to implement a rolling leaderboard on getting started competitions](https://www.kaggle.com/c/house-prices-advanced-regression-techniques#tutorials)[here](https://www.kaggle.com/c/titanic/discussion/6240).

Kaggle Kernels是一种云计算环境，可实现可重复的协作分析。 内核支持R和Python中的脚本，Jupyter笔记本和RMarkdown报告。 转到内核标签查看本次比赛的所有公开共享代码。 有关如何使用Kernels学习数据科学的更多信息，请参阅阅读更多关于我们决定在此处开始入门比赛的滚动排行榜。

### How do I contact Support?

Kaggle does not have a dedicated support team so you’ll typically find that you receive a response more quickly by asking your question in the appropriate forum. (For this competition, you’ll want to use the [House Prices discussion forum](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion)).

Support is only able to help with issues that are being experienced by all participants. Before contacting support, please check the discussion forum for information on your problem. If you can’t find it, you can post your problem in the forum so a fellow participant or a Kaggle team member can provide help. The forums are full of useful information on the data, metric, and different approaches. We encourage you to use the forums often. If you share your knowledge, you'll find that others will share a lot in turn!

If your problem persists or it seems to be affecting all participants then please [contact us](https://www.kaggle.com/contact).

Kaggle没有专门的支持团队，因此您通常会在适当的论坛中询问您的问题，从而更快地收到回复。 （对于本次比赛，您需要使用House Prices讨论区）。

支持只能帮助解决所有参与者正在经历的问题。 在联系支持之前，请查看论坛以获取有关您的问题的信息。 如果你找不到它，你可以在论坛上发布你的问题，这样一个参与者或Kaggle团队成员可以提供帮助。 论坛充满了有关数据，指标和不同方法的有用信息。 我们鼓励您经常使用论坛。 如果你分享你的知识，你会发现其他人将依次分享！

如果问题仍然存在或似乎影响所有参与者，请联系我们。

Tutorials

### Learning with Kaggle Kernels

Kaggle Kernels is an in-browser computational environment that is fully integrated with most competition datasets. Kernels is preloaded with most data science packages and libraries. It supports scripts and Jupyter Notebooks in R and Python, as well as RMarkdown reports. You can create submission files with Kernels and also use it to explore the competition data.

To get started with Kernels you can either:

1. Create a new script or notebook on the [Kernels](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/kernels) tab or
2. “Fork” any kernel to create an editable copy for you to experiment with

We've selected some of the best kernels to help you get started with the competition. You can use the below kernels to create a submission file or to explore the data. Simply open the script or notebook and click "fork" to create an editable copy.

Kaggle Kernels是一种浏览器内计算环境，与大多数竞争数据集完全集成。 内核预装了大多数数据科学包和库。 它支持R和Python中的脚本和Jupyter笔记本以及RMarkdown报告。 您可以使用内核创建提交文件，也可以使用它来浏览竞争数据。

要开始使用内核，您可以：

1.在Kernels选项卡或者上创建一个新的脚本或笔记本

2.“分叉”任何内核以创建一个可编辑的副本供您试验

我们选择了一些最好的内核来帮助您开始竞争。 您可以使用以下内核创建提交文件或浏览数据。 只需打开脚本或笔记本，然后单击“fork”即可创建可编辑的副本。

### Getting started with R

[Detailed Exploratory Data Analysis Using R](https://www.kaggle.com/notaapple/house-prices-advanced-regression-techniques/detailed-exploratory-data-analysis-using-r)

* Use [RMarkdown](http://rmarkdown.rstudio.com/) and popular R packages like [data.table](https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html), [dplyr](https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html), and [ggplot2](http://docs.ggplot2.org/current/)
* Take an in-depth look at missing values, distributions, and correlations
* Click on the “Code” tab to see the underlying code which combines markdown text and R scripts

[Fun with Real Estate Data](https://www.kaggle.com/skirmer/house-prices-advanced-regression-techniques/fun-with-real-estate-data)

* Use Rmarkdown to learn advanced regression techniques like random forests and [XGBoost](https://github.com/dmlc/xgboost)

[XGBoost with Parameter Tuning](https://www.kaggle.com/jiashenliu/house-prices-advanced-regression-techniques/updated-xgboost-with-parameter-tuning/run/362252)

* Implement LASSO regression to avoid multicollinearity
* Includes linear regression, random forest, and [XGBoost](https://github.com/dmlc/xgboost) models as well

[Ensemble Modeling: Stack Model Example](https://www.kaggle.com/jimthompson/house-prices-advanced-regression-techniques/ensemble-model-stacked-model-example)

* Use "[ensembling](http://mlwave.com/kaggle-ensembling-guide/)" to combine the predictions of several models
* Includes GBM (gradient boosting machine), [XGBoost](https://github.com/dmlc/xgboost), ranger, and neural net using the [caret package](https://cran.r-project.org/web/packages/caret/index.html)

[A Clear Example of Overfitting](https://www.kaggle.com/ozagordi/house-prices-advanced-regression-techniques/a-clear-example-of-overfitting)

* Learn about the dreaded consequences of overfitting data

### Getting started with 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/)

用Python进行全面的数据挖掘

•了解变量的分布方式及其交互方式

•在训练机器学习模型之前应用不同的转换

•涵盖单变量和多变量方法

•包含使用matplotlib和seaborn的可视化

房价EDA

•学习使用可视化技术来研究缺失的数据和分布

•涵盖连续和分类数据

•包括相关热图，对偶图和t-SNE，以帮助将适当的输入通知给线性模型

回归分析在艾姆斯数据集中的应用研究

•展示特色工程的有效策略

•包括简化和组合现有功能

•使用scikit-learn，使用不同的正则化方法（包括Ridge，LASSO和ElasticNet）探索线性回归

正则化线性模型

•建立一个基本的线性模型

•使用Keras尝试更高级的算法，包括XGBoost和神经网络