Requires Changes

**还需满足 5 个要求 变化**

作为第一次提交，你仔细回答了问题，也能看得出你认真的思考！  
还有一些小问题需要修改，期待看到你下一次的提交～  
😄

**导入并观察数据**

**学员成功加载波士顿房屋训练数据housedata.csv。**

很棒，使用pandas的read\_csv()函数成功加载了波士顿房屋数据集。

**打印并观察前5条data\_df数据；在data\_df中删除'Id'列数据；使用describe方法观察data\_df各个特征的统计信息。**

很棒，pandas.DataFrame.head()是非常常用的函数，不设置参数的时候会默认打印前5条数据来帮助我们预览数据集。pandas.DataFrame.describe()可以帮助我们观察数据机各个特征的统计信息。

**学员通过观察数据，结合data\_description.txt特征描述，写出与目标变量最相关的5个特征，并进行解释。**

* 不错，你给出了自己的判断，有助于建立自己的数据直觉。这里你也可以通过可视化来支撑你的论点，请参考：
* import seaborn as sns
* sns.heatmap(data\_df.corr(), annot=True)

**数据预处理**

**问题3.1：画出'GrLivArea'和'SalePrice'的关系图，x轴为'GrLivArea'，y轴为'SalePrice'，观察数据**

**通过上图我们可以看到那几个异常值，即'GrLivArea'大于4000，但是'SalePrice'又极低的数据，从data\_df删除这几个异常值，删除后重新绘制'GrLivArea'和'SalePrice'的关系图，确认异常值已删除。**

做得很好👍对Dataframe的使用非常熟练！

**筛选出过多空数据的特征，我们这个项目定为筛选出超过25%的空数据的特征**

正确统计了含有过多空数据的特征👍

**根据data\_description.txt特征描述填充空数据**

也可以这么简单的实现：

data\_df[fill\_Mode] = data\_df[fill\_Mode].fillna(data\_df.mode())

data\_df[fill\_None] = data\_df[fill\_None].fillna('None')

data\_df[fill\_0].fillna(0, inplace=True)

data\_df[fill\_median] = data\_df[fill\_median].fillna(data\_df.median())

**探索性数据分析（EDA）与特征分析**

**绘制'SalePrice'，并说明该直方图属于什么分布**

这里需要回答直方图属于什么分布，属于正偏态和负偏态呢？  
参考链接：[https://baike.baidu.com/item/%E5%81%8F%E6%80%81%E5%88%86%E5%B8%83/445413?fr=aladdin](https://baike.baidu.com/item/%E5%81%8F%E6%80%81%E5%88%86%E5%B8%83/445413?fr=aladdin" \t "_blank)

**绘制相应特征与'SalePrice'的关系图，根据关系图所示进行总结说明问题2.4的所猜测的关系是否正确。**

这里需要使用了plt.scatter绘制散点图才能更好表示相应特征与'SalePrice'的关系，而且不要忘记根据关系图所示进行总结说明问题2.4的所猜测的关系是否正确哦～

**模型实现**

**将data\_df分割为特征和目标变量**

**将features，labels分隔为X\_train, X\_test, y\_train, y\_test**

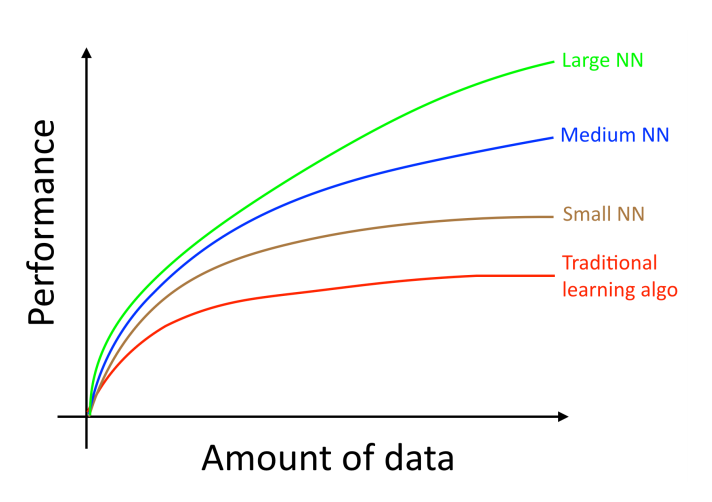
👍对原始数据进行训练集和测试集的划分，能够帮助我们选出泛化能力更佳的模型。

**回答将数据集分为训练数据与测试数据的原因。**

回答的很好！将数据集分成训练集和测试集的好处：既可以用于训练又可以用于测试，而且不会相互干扰，而且可以对训练模型进行有效的验证。用部分训练集进行测试的坏处：模型就是根据训练集得出的，使用训练集进行测试肯定会得出较好的结果，这不能判断训练模型的优劣。

**完成 performance\_metric 函数并回答相关问题。**

**随着训练点的不断增加，学生正确判断图表中训练集和验证集曲线的走向并讨论该模型是否会得益于更多的训练点。**

对训练曲线和测试曲线趋势和意义解释的很好。这里随着数据的增多，max\_depth不变的情况下，模型提升的幅度也越来越小。  
[](https://udacity-reviews-uploads.s3.amazonaws.com/_attachments/3230/1498724415/ml_curve.png)

传统的机器学习算法（又被称为基于统计的机器学习）在数据量达到一定程度后，更多的数据无法提升模型的表现。深度学习的一个优势就是它可以把大量的数据利用起来，提升学习表现。

这里还有更多关于学习曲线的介绍：  
<https://www.coursera.org/learn/machine-learning/lecture/Kont7/learning-curves>  
<http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html>

**学生提供最大深度为 1 和 10 的分析。如果模型偏差或方差较高，请针对每个图形给出合理的理由。**

很好，你选出了最优模型的参数，但是训练集和测试集的得分最接近并不是评判最优模型的标准，我们更应该关注测试集得分的高低。

**学生准确说明网格搜索算法，并简要探讨该算法的用途。**

网格搜索算法不一定要使用交叉验证哦，希望你能清楚～

**学生准确说明如何对模型进行交叉验证，以及它对网格搜索的作用。**

你回答的不错！这里做出一点补充。

* 第三小问，这里补充下cv*results*的详细解释，GridSearchCV 中的cv\_results 返回一个字典，记录了每组网格参数每一次训练/验证对应的训练结果，包括训练/验证时间、训练/验证评估分数以及相关时间和评分的统计信息，从GridSearchCV文档中可以看到具体的参数，在此不一一列举。
* 第四小问，对网格搜索来说，可以将训练数据可以按一定比例分为训练集和验证集，而不使用交叉验证。而如果这一次划分的验证集不具代表性。就会影响模型在未知数据上的表现。而交叉验证很大程度上避免因样本划分不合理导致选择了错误的参数。

**学生在代码中正确实现 fit\_model 函数。**

**进行预测**

**填入最优参数，并评价相关结果。**

你的回答，“获得的结果在0~1之间”，并不能得到“基本满足最优模型的要求”，请再说的准确些～

 重新提交项目

[**下载项目**](https://review-api.udacity.com/api/v1/submissions/1657071/archive)



**重新提交项目的最佳做法**

Ben 与你分享修改和重新提交的 5 个有益的小贴士。

[观看视频](https://review.udacity.com/)(3:01)