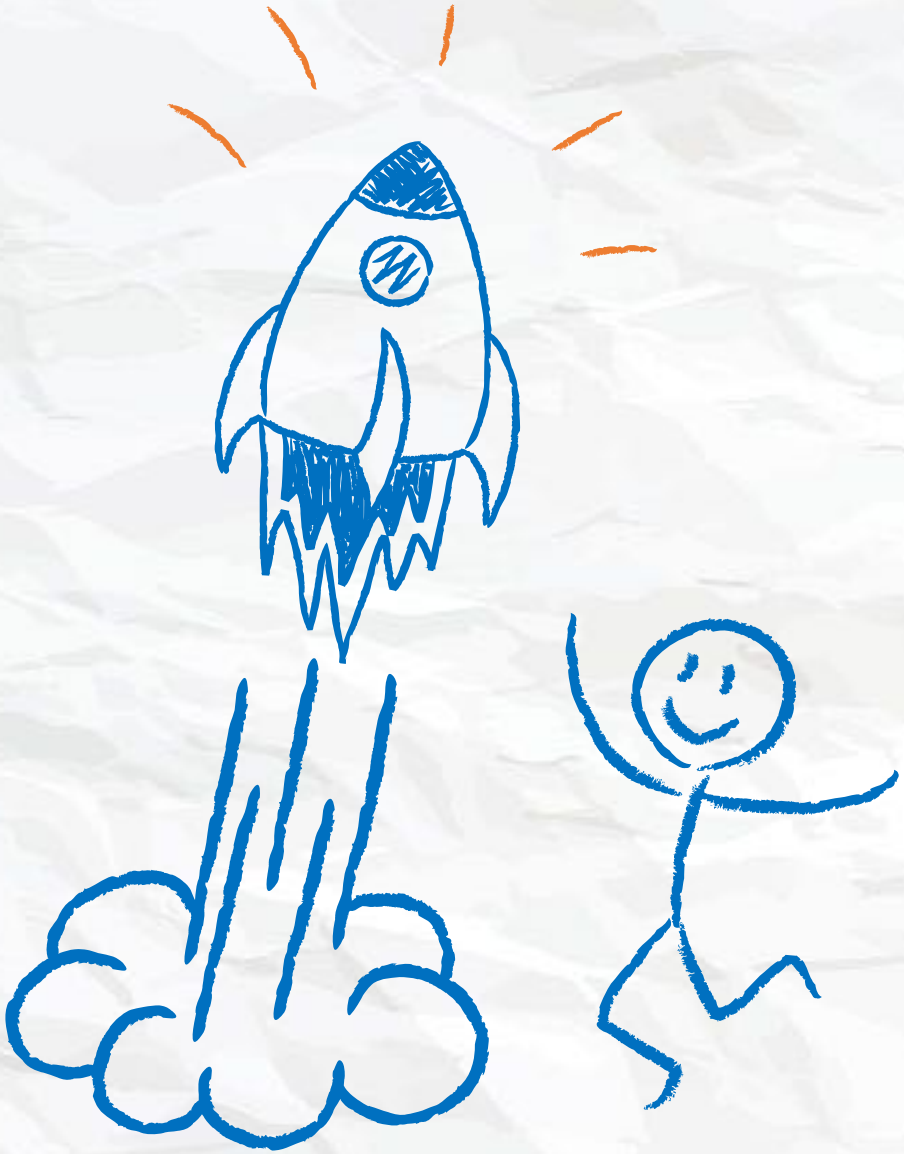
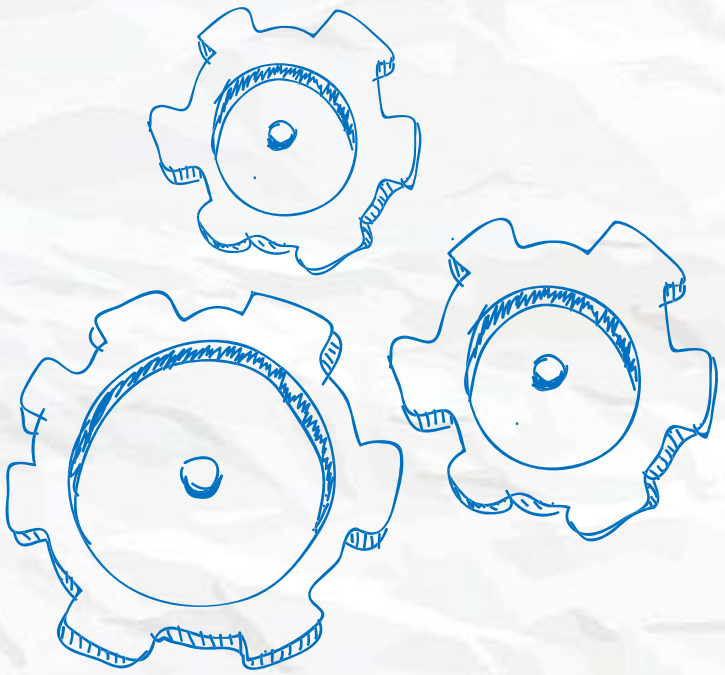


2018 金融数据分析与数据挖掘

Kaggle 房价预测问题

By 王丹妤 高宇馨 李莉欣

ABOUT US



王丹妤：模型设计+代码编写+润色报告PPT

高宇馨：PPT 制作+报告(概况、数据预处理等其他部分)

李莉欣：报告(模型特点整理、图)

CONTENTS

PART 01

问题概述

房价预测问题

基本信息

- 来自Kaggle
- 79个解释变量（爱荷华州艾姆斯住宅）
- 目的：预测房屋的最终价格

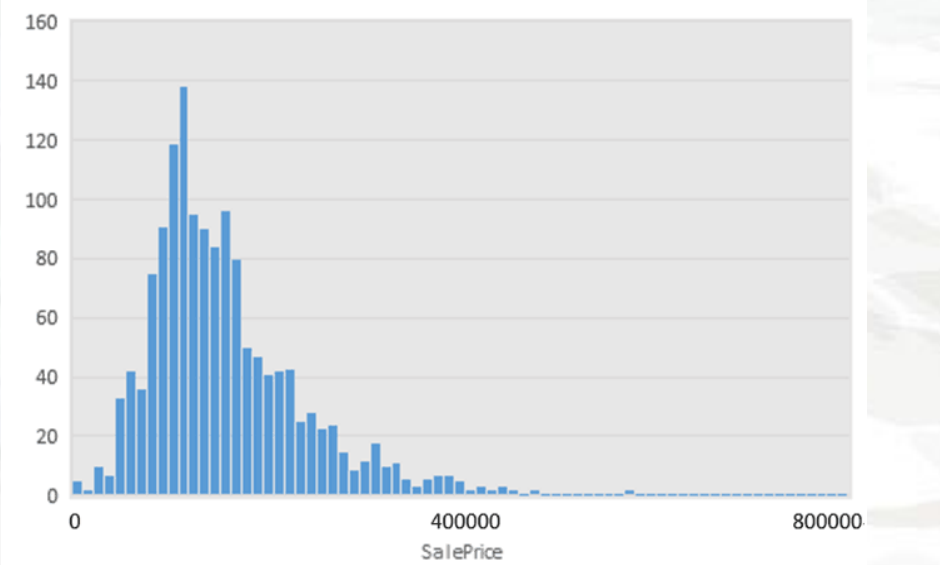
任务

1. 已有的数据分析
2. 数据预处理-包括特征转换、数据类型转换、异常值检测和估算缺失值。
3. 基准建模-利用基础模型建模。
4. 模型改进-合并，调参，得到最优结果

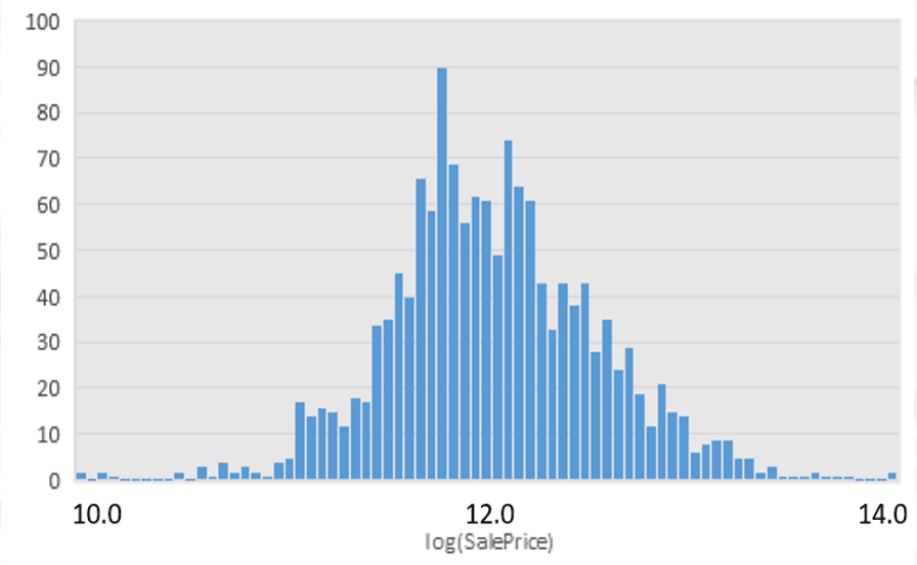
PART 02

问题分析

数据探索



原始数据：房价的分布



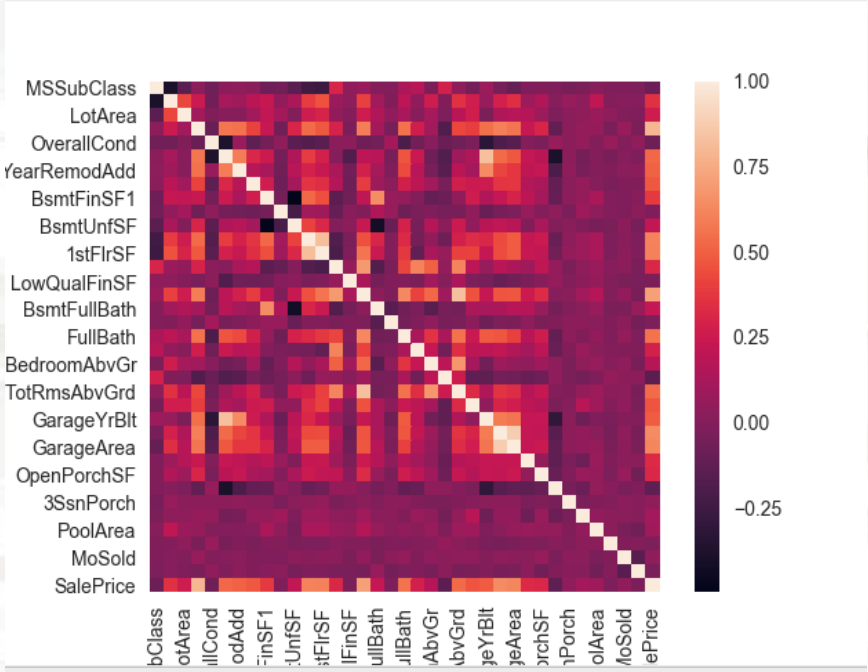
线性回归后不符合正态分布

进行平滑处理，取log，符合高斯分布

数据探索

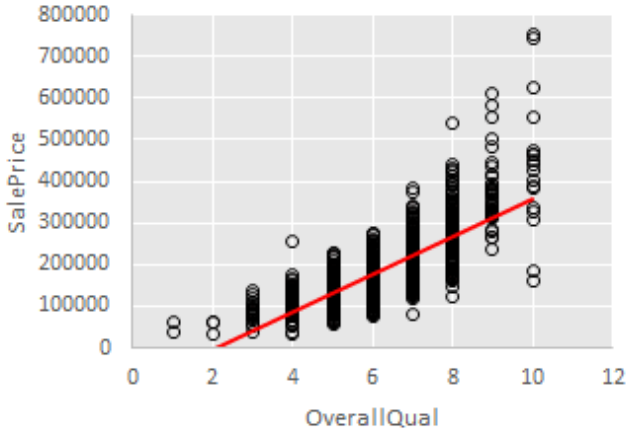
OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852

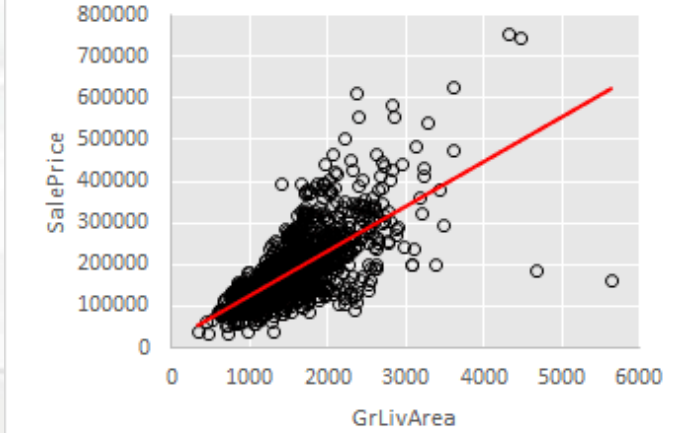
相关系数最高的前六位

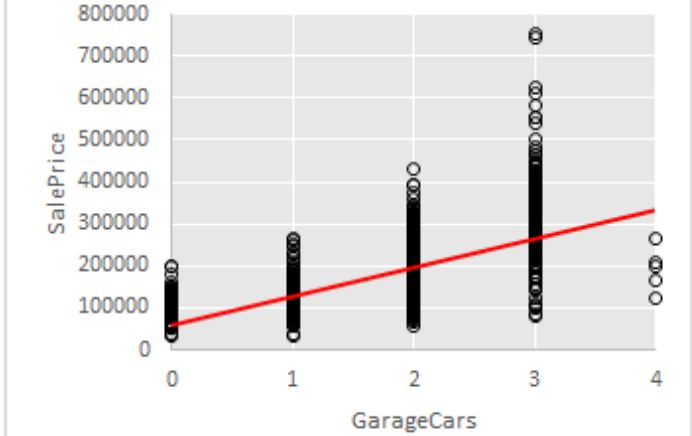


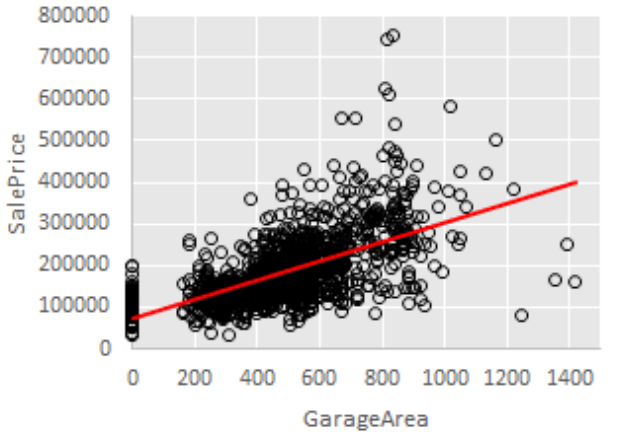
数值变量和房屋销售之间的相关系数矩阵

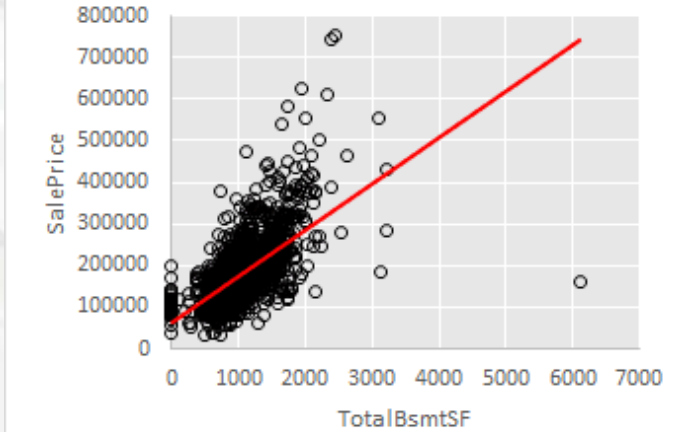
数据探索

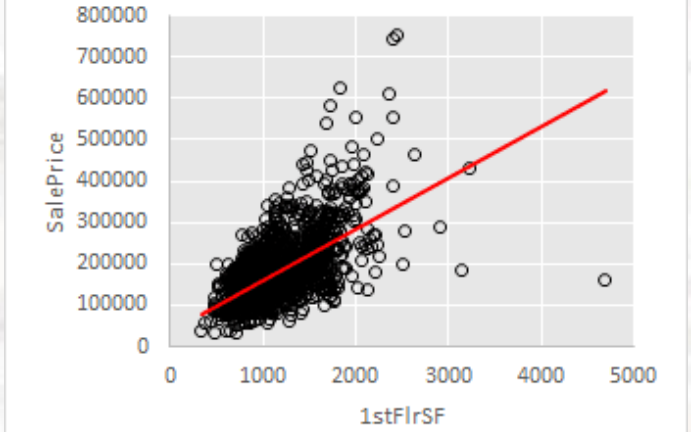

SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
2
4
6
8
10
12
OverallQual


SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
6000
GrLivArea


SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
1
2
3
4
GarageCars

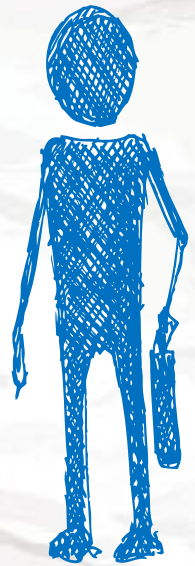

SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400
GarageArea


SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
6000
7000
TotalBsmtSF

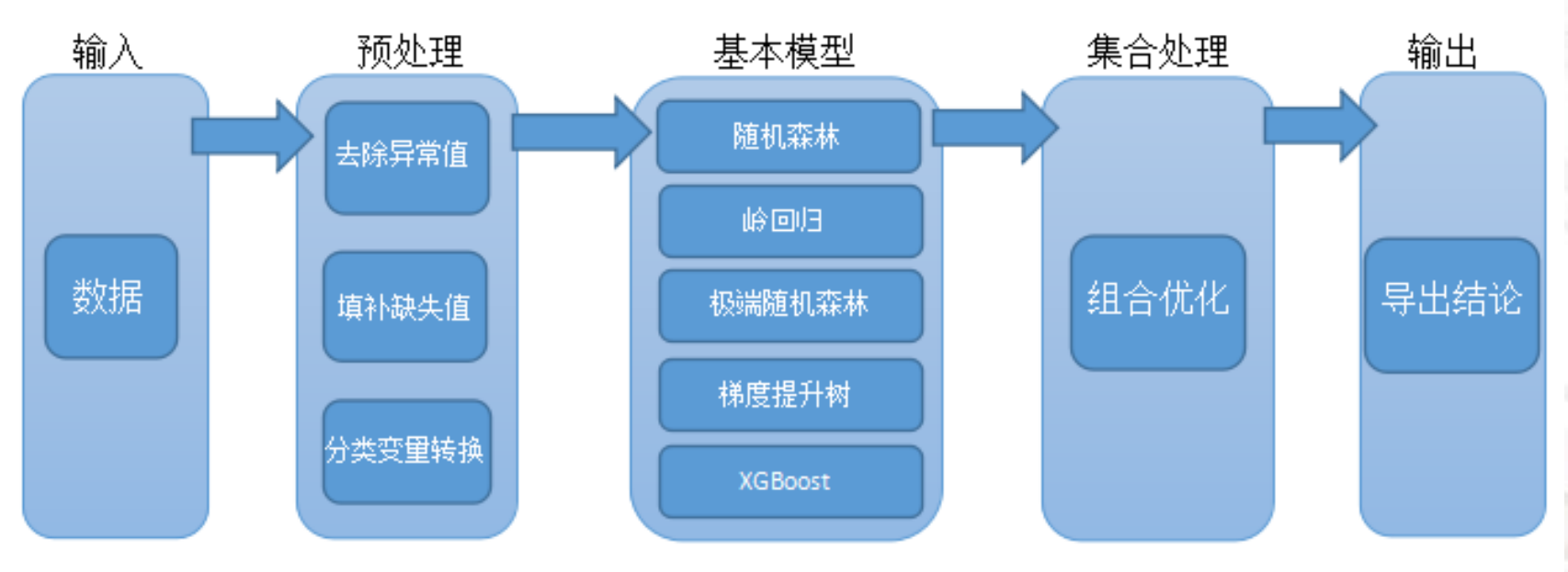

SalePrice
800000
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
1stFlrSF

异常数据信息

PoolQC	7 non-null object
Fence	281 non-null object
MiscFeature	54 non-null object
Alley	91 non-null object
FireplaceQu	770 non-null object
LotFrontage	1201 non-null float64
MasVnrType	1452 non-null object
MasVnrArea	1452 non-null float64
BsmtQual	1423 non-null object
BsmtCond	1423 non-null object
BsmtExposure	1422 non-null object
BsmtFinType1	1423 non-null object
BsmtFinType2	1422 non-null object
Electrical	1459 non-null object
GarageType	1379 non-null object
GarageYrBlt	1379 non-null float64
GarageFinish	1379 non null object
GarageQual	1379 non-null object
GarageCond	1379 non-null object



思维导图


输入
预处理
基本模型
集合处理
输出
数据
去除异常值
填补缺失值
分类变量转换
随机森林
岭回归
极端随机森林
梯度提升树
XGBoost
组合优化
导出结论

PART 03

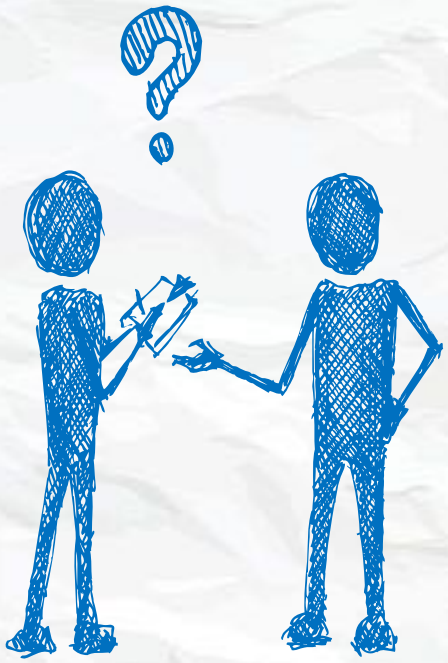
数据预处理

数据预处理

删除变量

所有变量中Alley变量有1369个缺失值，FireplaceQu变量有690个缺失值，PoolQC变量有1453个缺失值，Fence变量有1179个缺失值，MiscFeature变量有1406个缺失值。由于缺失值过多，我们将删掉这五个相关变量。

数据预处理



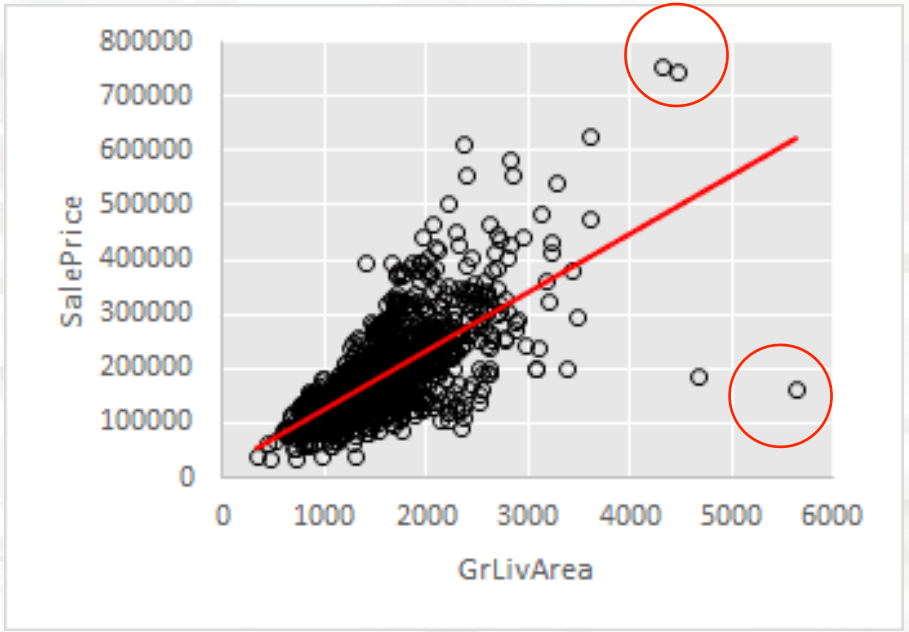
填补缺失值

然后对有缺失值的 14 个变量 LFage，GarageType等进行平均值填充。

数据预处理

删除outlier

针对之前六个相关度大的属性，根据图中的离散点进行删除。



数据预处理

其他

•对于分类变量，我们采取将分类变量转换为虚拟变量或指示变量的方式。

•特别的，如果某个数据的‘Sale Condition’属性值若为‘Abnormal‘，则将所对应数据删除。

PART 04

模型建立

五个基本模型

随机森林

岭回归

Extra trees

梯度提升树

XGBoost

五个基本模型

1

调用Python中的
机器学习库

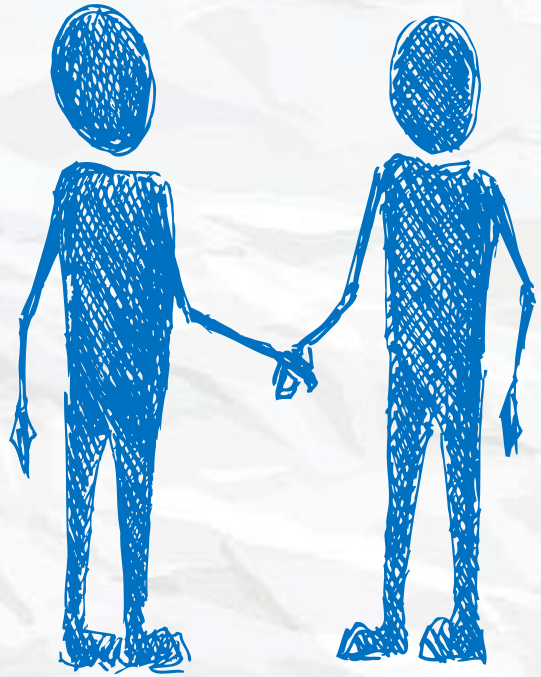
2

运用GridSearchCV
自动调参，产生模型

3

用RMSE评价模型

合并基本模型



运用bagging的方法
将随机森林和岭回归
模型结果合并

基本模型比较

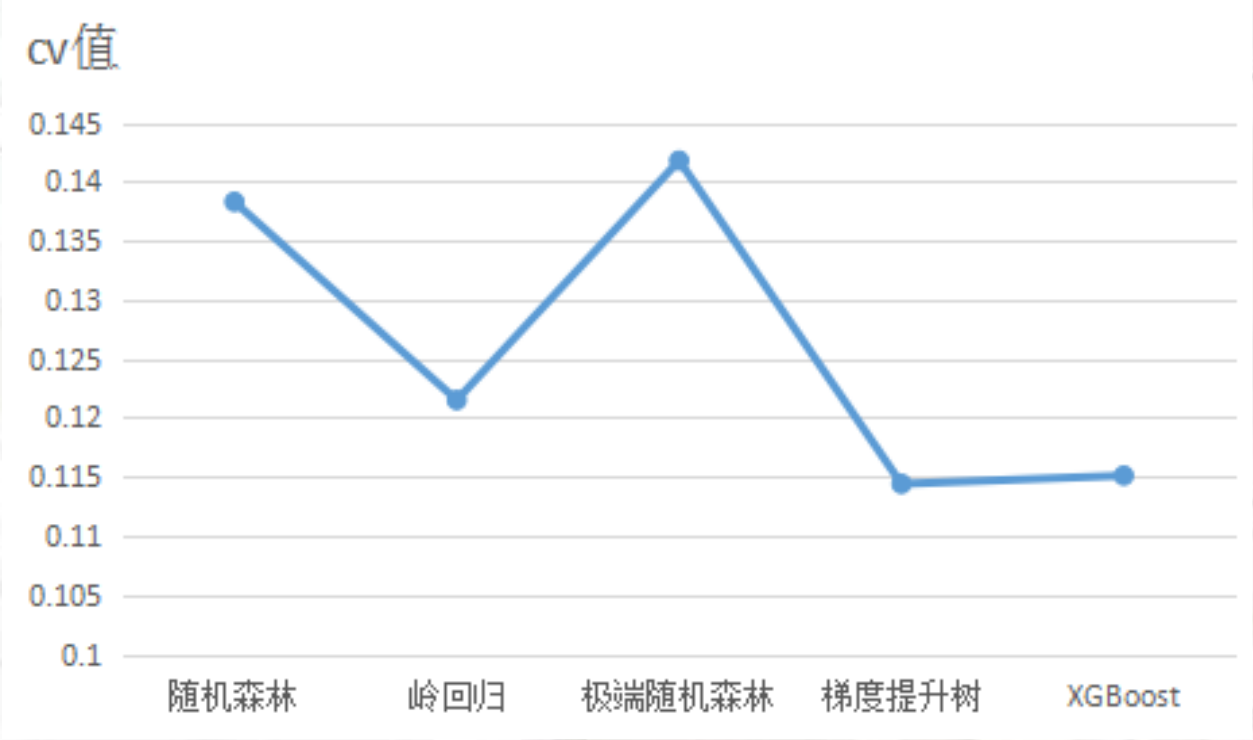
	优点	缺点
随机森林	实现简单；不容易过度拟合；能处理高纬度数据；平衡数据误差；可以选出重要特征	数据噪声过大时还是容易过度拟合；其随机性对于模型难以进行解释
岭回归	改良的OLS估计，在存在共线性问题和病态数据偏多的研究中有较大的实用价值。	结果是有偏的，降低了精度
Extra trees（极端随机森林）	随机性比随机森林更强，模型的方差相对于RF进一步减少，在某些时候，extra trees的泛化能力比RF更好。	但是偏倚相对于RF进一步增大
Gradient boosted tree（梯度提升树）	适用面广，几乎可用于所有回归问题（线性/非线性），亦可用于二分类问题	对异常值非常敏感
XGBoost（Extreme Gradient Boosting）	有效防止过度拟合；精度高；可适应纬度高的情况；	调参复杂

PART 05

结果评价与分析

模型自我评估

```
Random forest regression:
Best CV Score:
0.13825001670785134
Gradient boosted tree regression:
Best CV Score:
0.11444568720434202
Extreme Gradient Boosting regression:
Best CV Score:
0.1150576611704134
Extra trees regression:
Best CV Score:
0.14182317000239522
Ridge Regression:
Best CV score:
0.12154299504355613
```



Kaggle评价结果

Top 20%

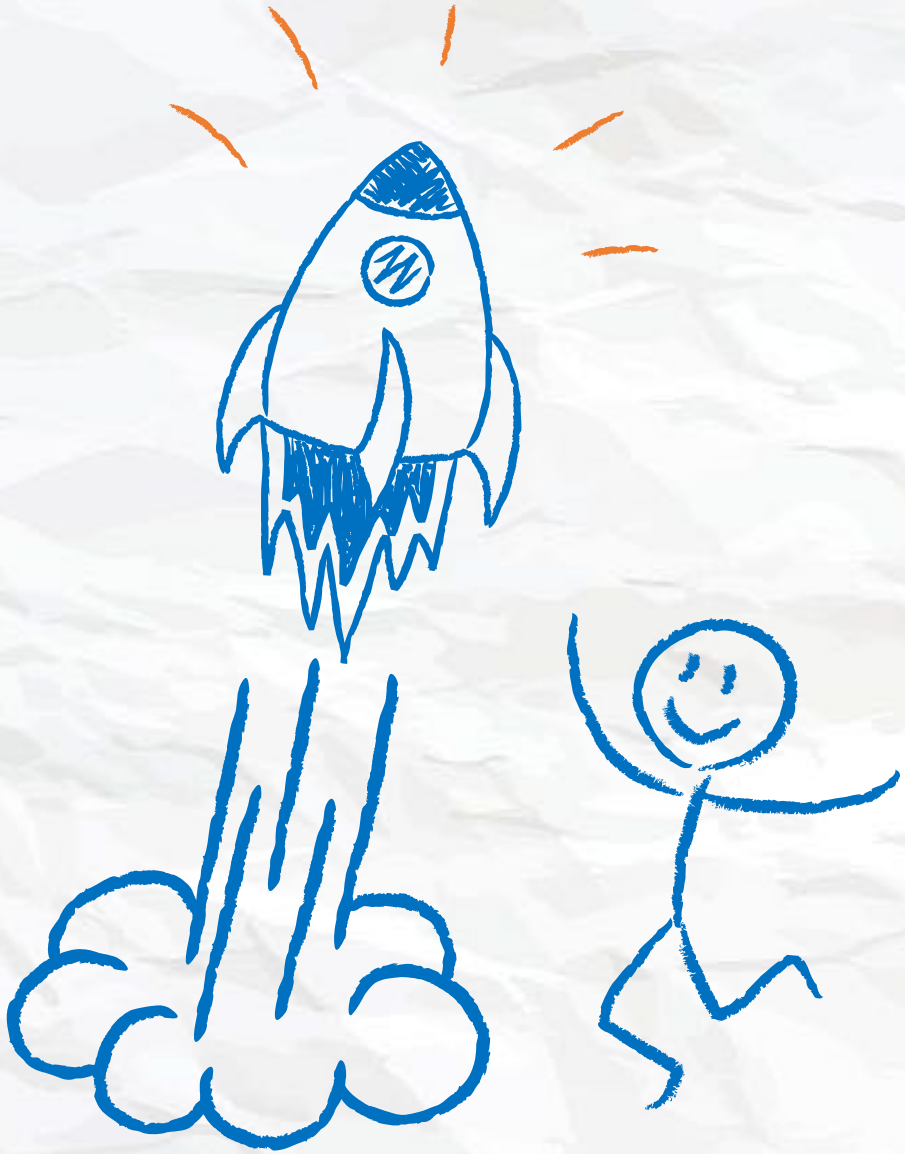
结果分析

从Kaggle提供的结果来看，合并模型结果最优，其次是XGBoost模型。这与自己预估的结果相类似。

反思与改进

极端随机森林在Kaggle的模型评价比在交叉验证环节中要好；相反，梯度提升树的情况刚好与之相反，在实际评价中表现较差，可能还是存在overfitting的情况。

- 变量分析不够细致，以后可以进行主成分分析。
- 对参数的调整范围比较宽泛，没有精确估计。



THANK YOU