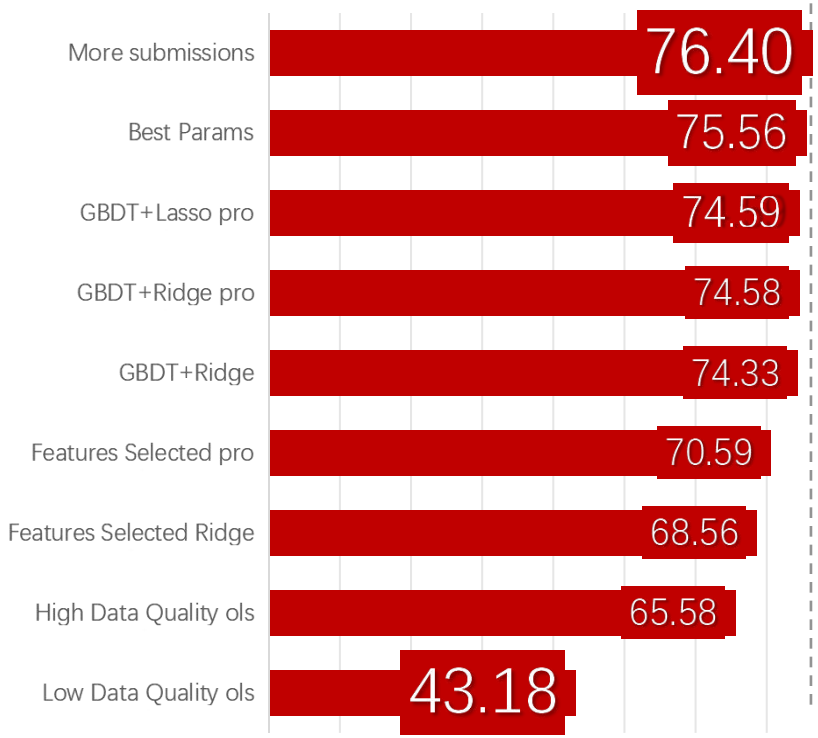
3	Wuzhaoyi_0104	76.40	5
5	Calero Taydus	75.56	35

Metrics	In	out	Cv
Ridge (Sales)	**203,146.54**	***215,200.00***	***214,800.00***
Lasso (Sales)	***210,400.00***	***220,100.00***	***220,500.00***
Ridge (Rent)	**60,590.60**	***65,300.00***	***65,120.00***
Lasso (Rent)	**67,804.75**	***72,100.00***	***71,950.00***

Combined Application of Various Models and Improvement of Data Quality is also important



Gradient Boosting Decision Tree (GBDT) Feature Transformation with Regularized Linear Models

Entire Workflow

Basic

Key points should be carefully extracted from every column

High Data Quality ⟫ ⟫ ⟫ ***Excellent feature engineering***

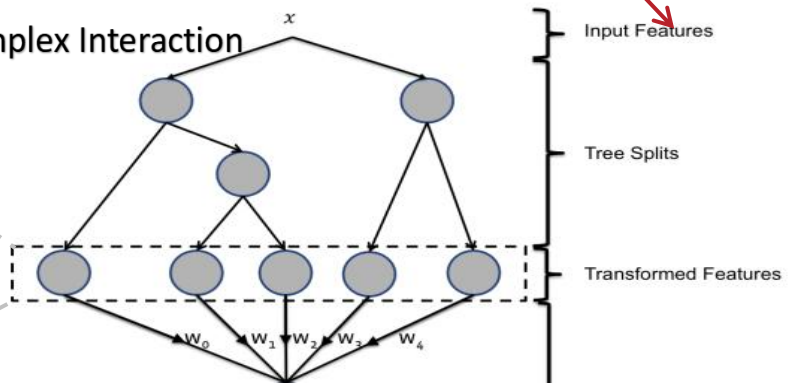
Leaf Encoding

Construct secondary features --- More thorough Input Features

How Nonlinear and Complex Interaction

One Hot Metrics

Index	0	1	2	3	4
0	0	3	0	0	0
1	22	0	0	0	17
2	7	5	0	1	0
3	0	0	0	0	0
4	0	0	14	0	8



Selected Top_n Origin Features Concat

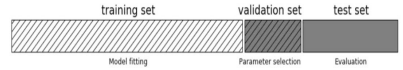
Objective

MAE

X_train_final
X_test_final

Linear Models

Parameters Tuning



Models

76.40 is the score of the Linear Model . ***Non-linear Models may be higher***

Source: *"Practical Lessons from Predicting Clicks on Ads at Facebook" - Xinran He, Junfeng Pan, Ou Jin*

模型结果汇报与亮点——魏旎

- 结果汇报
- 模型亮点
  - 数据提取
    - 运用正则表达式匹配、字符串匹配、字典映射，完成数据提取和汉字转换
    - Sales房屋优势、Rent配套设施等列具有自然语言特征，对关键词编码
  - 特征工程
    - 独热编码：对分类型变量、无序数特征（城市/区域/板块）的变量进行独热编码
    - 引入交互项：楼层比、套内面积比、梯户比、楼房比、绿化容积比
  - 缺失值处理
    - 删除缺失值过多的特征和样本
    - 数值型特征：InterativeImputer迭代填充
    - 分类型特征：基于数值型特征KNN，在最相近的K个样本中取众数
  - 异常值发现与处理：3σ原则，删除异常样本（Sales: 88688 Rent: 84358）
  - 目标变量转化：Price -> Price/Area

Metrics	Type	In sample	Out of sample	Cross validation	Kaggle
OLS	Sales	0.199	0.2011	0.1998	61.51
	Rent	0.2283	0.2278	0.2289	
Lasso	Sales	0.1988	0.2006	0.1996	61.96
	Rent	0.2282	0.2277	0.2288	
Ridge	Sales	0.1984	0.2004	0.1992	61.79
	Rent	0.2283	0.2278	0.2289	
ElasticNet	Sales	0.1991	0.2008	0.1999	61.64
	Rent	0.2283	0.2277	0.2289	
Average	Sales	-	-	-	61.89
	Rent				
Best linear model	Sales	0.1984	0.2004	0.1992	61.79
	Rent	0.2283	0.2278	0.2289	