关于决定数据分析师薪资水平的各要素的分析

---以上海地区为例

by蒋超迪

内容

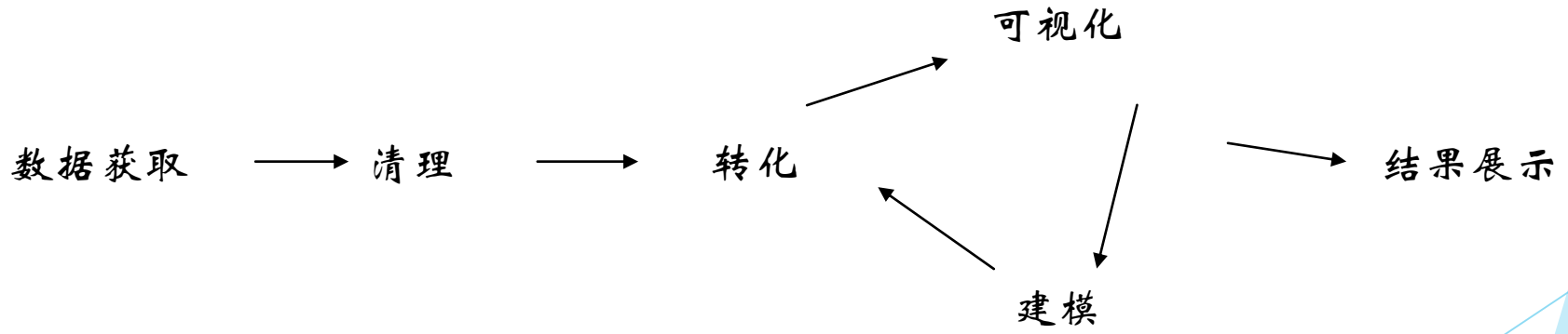
1 项目概述
2 数据的爬取（从智联招聘）
3 总体特征（data summary）
4 建模，预测及可视化分析
5 小结

项目概述

首先我用python写了一个网页爬虫爬取了智联招聘上的带有“数据分析”关键字的职位，以及与之相关的各项数据（包括薪水，工作经验等等）。以爬取的数据作为训练集（train set），用了多种监督学习模型进行建模（用R完成）来预测在沪的数据分析的期望工资。

并且测试和比较了各种模型的预测误差水平。同时，我为了更好地展示结果，进行了大量的可视化处理。

数据分析的基本流程：



数据的爬取（数据获取）

智联招聘上显示的带有“数据分析”关键字的信息：

职位名称	反馈率	公司名称	职位月薪
数据分析		上海中智项目外包咨询服务有限公司	面议
运营数据分析		北京字节跳动科技有限公司	10001-15000
数据分析员（财务）	100%	南极电商(上海)有限公司	6001-8000
联通数据分析专员	100%	上海启彧企业管理咨询有限公司	4001-6000
NO.405 数据分析产品经理	96%	上海KT人才服务有限公司	20001-30000
数据分析师（派驻非洲加纳）	89%	大连飞鹿贸易有限公司	10001-15000
数据分析员	82%	上海杨普申通快递有限公司	2001-4000
数据分析	81%	驴妈妈旅游网	4001-6000
数据分析经理	75%	上海南柏文化传播有限公司	10001-15000
运营数据分析师	73%	驴妈妈旅游网	10000-20000

选中其中一项:

运营数据分析(职位编号：10001241)
北京字节跳动科技有限公司

职位月薪：10001-15000元/月
工作地点：上海
发布日期：2017-01-21
工作性质：全职
工作经验：3-5年
最低学历：大专
招聘人数：2人
职位类别：网络运营专员/助理

职位描述　公司介绍　举报　收藏

岗位职责：
1、监控各类数据，分析比对数据变动；
2、定期分析各维度数据，提炼核心要点并制定建议；
3、了解各位广告产品特点及对应售卖情况，建议较优产品组合；
4、定期输出优秀案例及行业投放通案，指导广告主/代理商更好投放广告。

任职要求：
1、2年以上数据分析经验，互联网广告相关行业者优先；

这里面有我比较感兴趣的信息，包括响应变量“职位月薪”，预测变量“最低学历”等等。在“职位描述”栏，罗列了用人单位的各项招聘要求。对其中的一些关键字进行提取和给分，比如出现“数学”字样,就给该职位的math变量一处分数1,否则给分数0.这些变量按大类分，大致可分为专业类（数学，金融等）/office技能类/统计软件类/开发语言类/大数据技能等等。

爬取方式具体见代码

Data summary

爬得的数据，通过python的pymysql模块被我存贮在了本地的mysql数据库里。用R的数据库接口DBI对该数据进行提取，并进行了一系列的清洗。

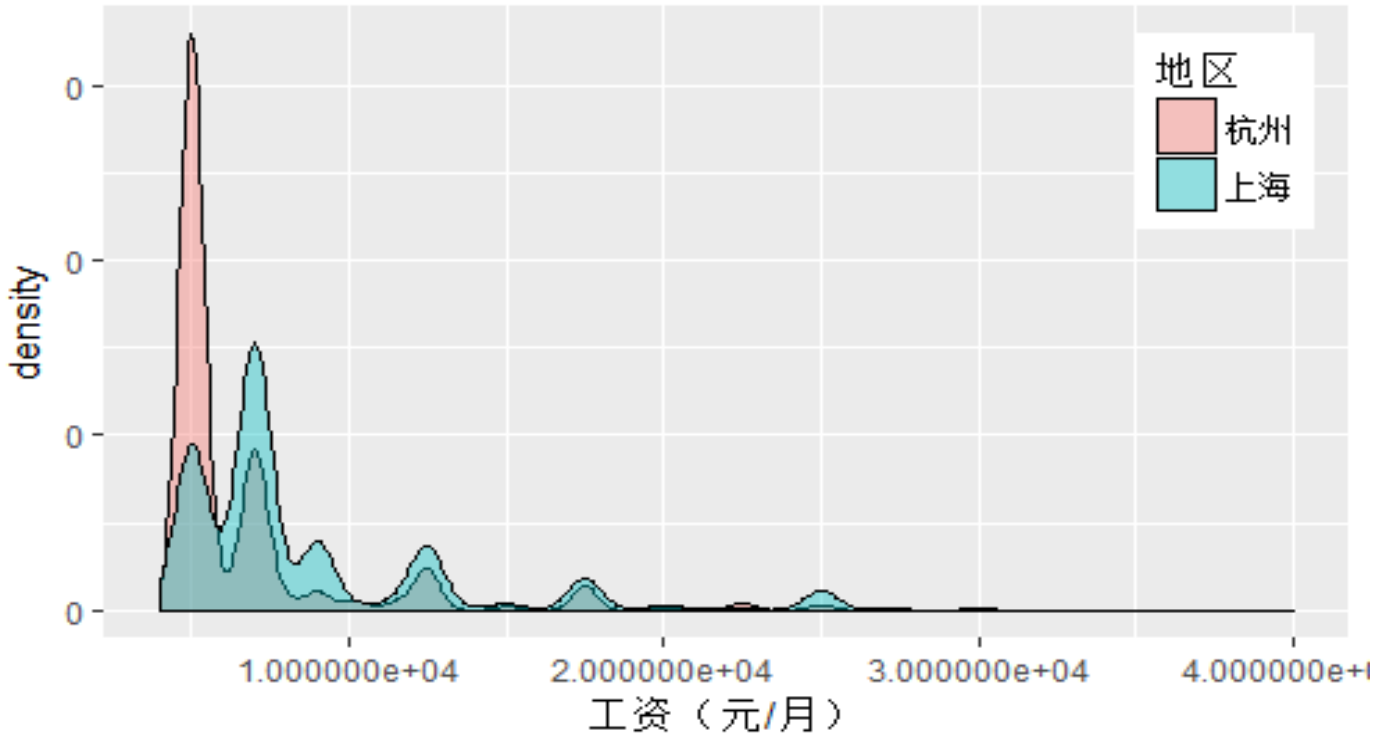
清洗后的数据（一部分）：

company <chr>	salary <dbl>	exp <int>	degree <chr>	location <chr>	postdate <chr>
上海上有资产管理有限公司	7001	2	本科	上海	2017-01-12
中国电信集团号百信息服务...	12501	2	本科	上海	2017-01-11
人人公司（Renren Inc.）	5001	0	本科	上海	2017-01-13
智易(上海)汽车技术咨询有限...	7001	2	本科	上海-浦东...	2017-01-13
上海诺亚金融服务有限公司	9001	0	硕士	上海-杨浦区	2017-01-13
东方财富信息股份有限公司	0	0	不限	上海-徐汇区	2017-01-13
上海恒寿堂健康食品股份有...	4500	0	不限	上海	2017-01-13
上海驰驭信息咨询有限公司	5500	1	本科	上海-嘉定区	2017-01-13
浙江核新同花顺网络信息股...	22500	4	本科	杭州	2017-01-13
上海连图信息科技有限公司	5001	0	大专	杭州-西湖区	2017-01-13
爱奇艺（www.iqiyi.com）	9001	2	本科	上海	2017-01-13
迅付信息科技有限公司	2750	0	本科	上海-徐汇区	2017-01-13
瑞庭网络技术（上海）有限...	12501	2	本科	上海	2017-01-13
迪亚天天(上海)管理咨询服务...	5500	0	不限	上海	2017-01-13
上海找钢网信息科技股份有...	32500	8	本科	上海-杨浦区	2017-01-13

为了方便比较，我也爬取了杭州地区的数据，可以看到上海的数据分析职位的平均月薪要比杭州高2000左右。

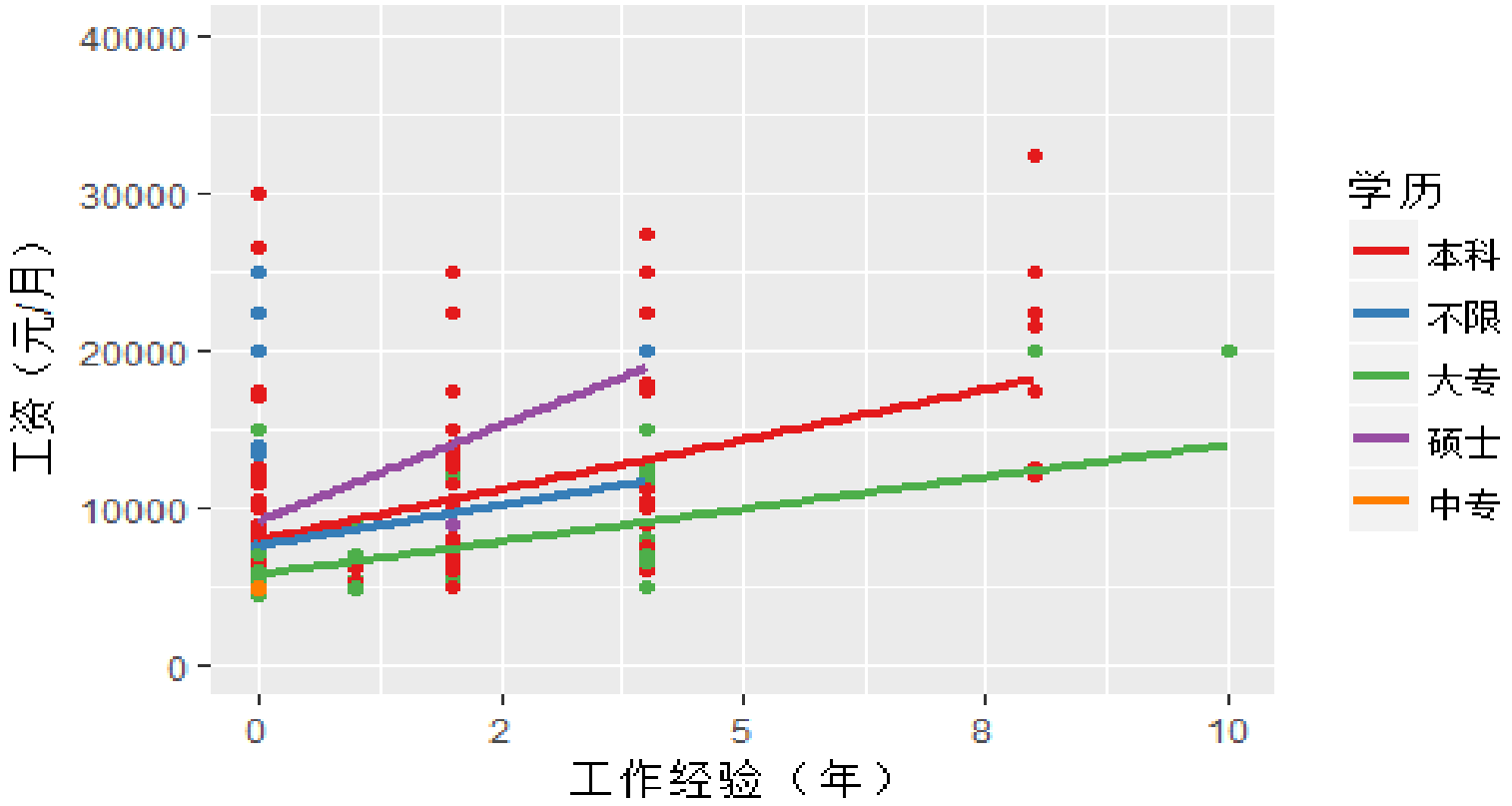
地区	数量	平均工资
<chr>	<int>	<dbl>
杭州	506	6714
上海	1723	8697

两地区数据分析职位月薪的密度分布图



之后的分析将只使用上海地区的数据！

月薪与工作经验及学历的关系



可以看到在上海，数据分析师的薪水和工作经验呈现明显的正相关。同时，工资与学历也有明显的正相关。另外，从回归线的走势来看，薪水和，学历与工作经验的交互项也有关系（工作经验大的地方，工资随学历的变化更明显）

建模及预测

首先我们使用OLS方法，选取所有的预测变量来拟合模型，一部分结果如下（按系数的大小，即变量对工资的影响力排列）

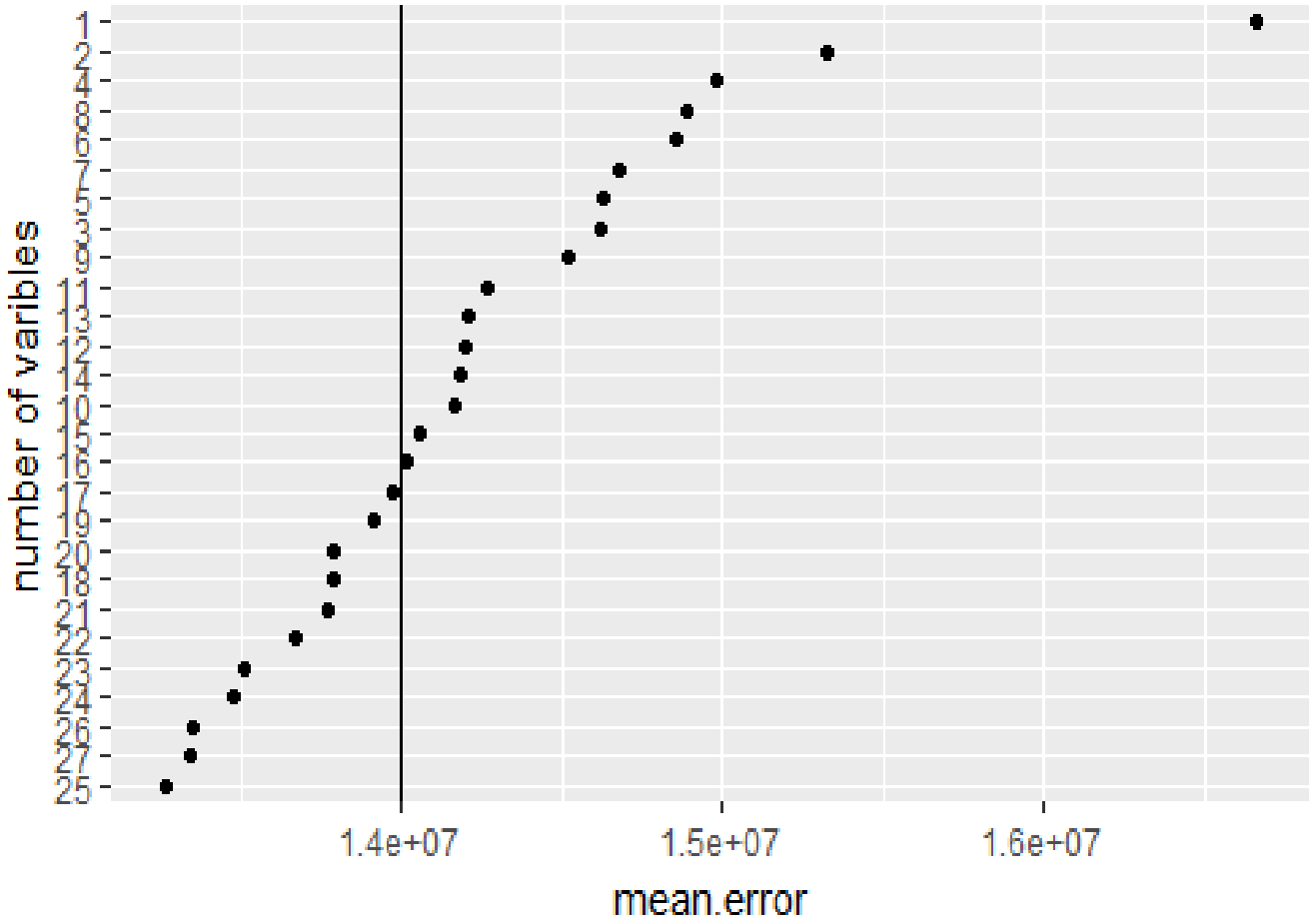
coef <chr>	Estimate <dbl>	Pr(>|t|) <dbl>
(Intercept)	7561.3	1.79e-38
spark	3642.1	3.52e-03
degree硕士	1954.9	1.05e-02
sql	1608.2	5.56e-06
exp	1182.8	2.21e-44
bigdata	876.3	2.23e-02
sas	764.2	8.77e-02
finance	713.4	3.77e-02
r	697.9	2.90e-02
hive	648.2	5.18e-01
math	626.9	6.88e-02
python	588.0	2.35e-01
information	583.2	6.50e-01
dataminning	563.6	5.73e-02
hadoop	268.2	8.01e-01
monitoring	-26.0	9.63e-01

可以看到用最小二乘法拟合模型，spark/硕士/sql技能以及经验（exp）对薪水影响较大（比如在控制其它变量不变的情况下，有spark技能可以多获得3642元/月），而且它们所对应的P值（Pr一栏）均显示这些变量都是统计有意的。

用OLS进行拟合，往往会伴随过度拟合（overfitting）的问题。毕竟我们的目的不是减少train-error（训练模型时的误差），而是test-error即模型应用在新数据集上的误差。

其中的一种改良方法是交叉验证，即把训练数集分成k份，把其中的k-1份数据作为训练模型，并在剩下的一份中验证误差。在此基础上，用全子集回归的方法来进行变量的选择。

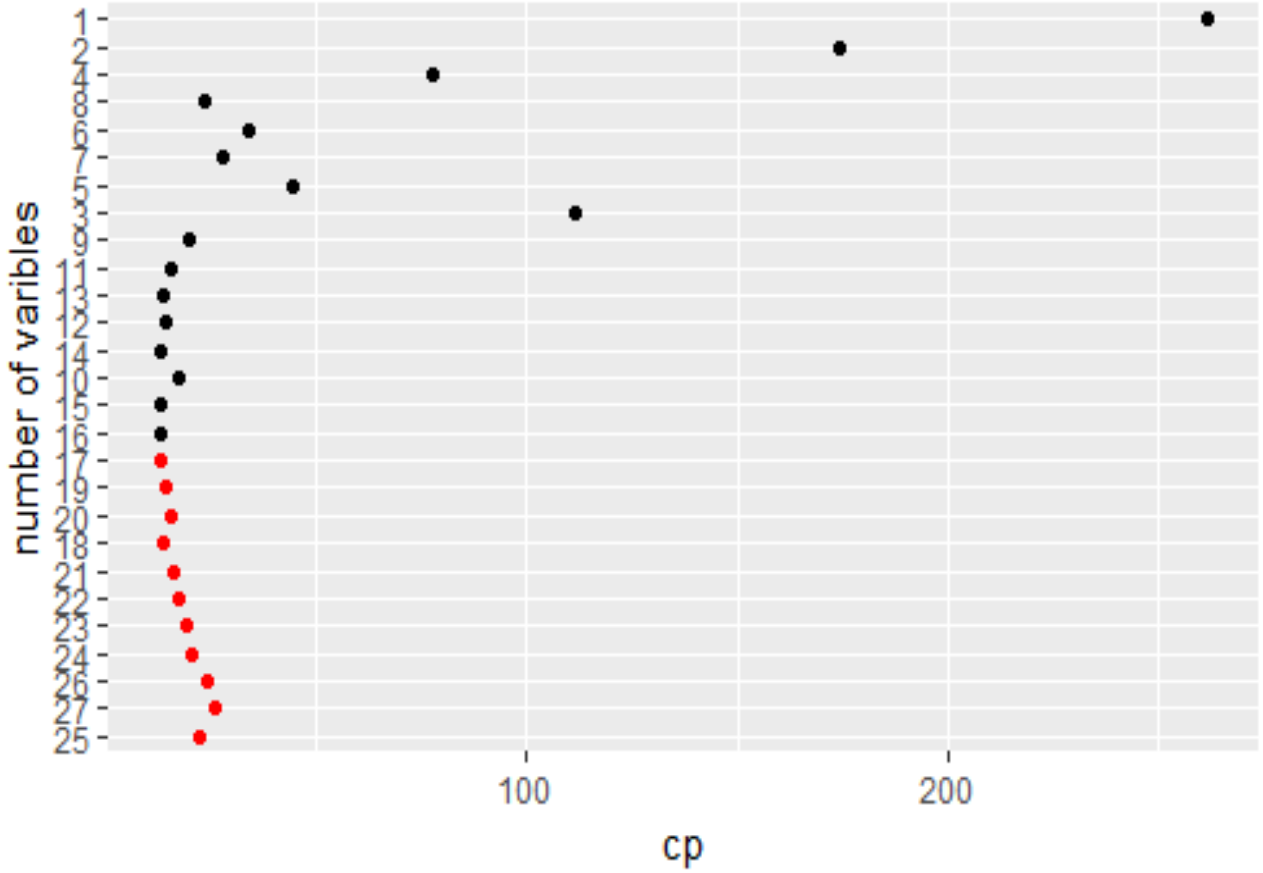
在使用交叉验证基础上的含各变量数的模型的平均误差



可以看到带有25个变量的模型的误差最小，但是和其它模型的差距不显著。

应用一个标准差（one standard-error rule）原则（离最小值一个标准差以内），选择其中11个模型作为候选（黑色实线的左侧）

含各变量数的模型的Cp统计量（越小越好）



varible	estimates
<chr>	<dbl>
(Intercept)	7057
hive	3238
degree硕士	2188
r	1250
exp	1231
sas	1199
finance	959
bigdata	873
math	607
dataminning	517
stata	0
communication	-384
excel	-721
ppt	-724
spss	-914
office	-1052
degree大专	-1695
customerbehavi	-2450

在11个候选模型中，变量个数为17的Cp统计量最小。另外结合奥姆剃刀原则，在模型表现差不多的情况下，尽量选择变量数较少的模型，所以我选择了变量数为17的模型。各变量对工资的影响效果见右图。

可见，Hive技能/硕士/R语言/经验/sas的影响比较大。

Ridge/Lasso 回归

另外一种改良过度拟合的方法就是Ridge回归。

OLS方法用最小二乘法让RSS（Residual sum of squares）最小。在此基础上，Ridge regression加上了一个penalty项来控制变量，以减轻Overfitting问题。

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso regression 与 Ridge regression类似，不过它的penalty项略有不同，可以起到变量选择的效果

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

注：用交叉验证法来决定系数lambda的值

Ridge回归的各变量的系数

varible	estimates
(Intercept)	7585
spark	3323
degree硕士	1820
sql	1498
exp	1112
bigdata	785
sas	779
hive	724
finance	704
r	683
information	630
math	613
dataminning	553
hadoop	552
python	546
monitoring	3
analysis	-58
computer	-222
statistics	-445
communicat	-452
java	-470
degree不限	-478
spss	-701
ppt	-763
office	-814
excel	-1037
degree大专	-1733
customerbeh	-1992
degree中专	-2002

Lasso回归的各变量的系数

varible	estimates
(Intercept)	7213
spark	3535
degree硕士	1478
sql	1426
exp	1161
finance	598
hadoop	557
hive	499
sas	455
r	453
math	435
bigdata	424
python	345
dataminning	279
statistics	-39
spss	-69
communicati	-171
customerbeh	-402
degree中专	-496
ppt	-533
office	-540
excel	-1035
degree大专	-1554

正如我前面提到的那样，lasso回归有很大的机会让一部分变量的系数变为零，所以它的非零系数的变量要少于Ridge回归。

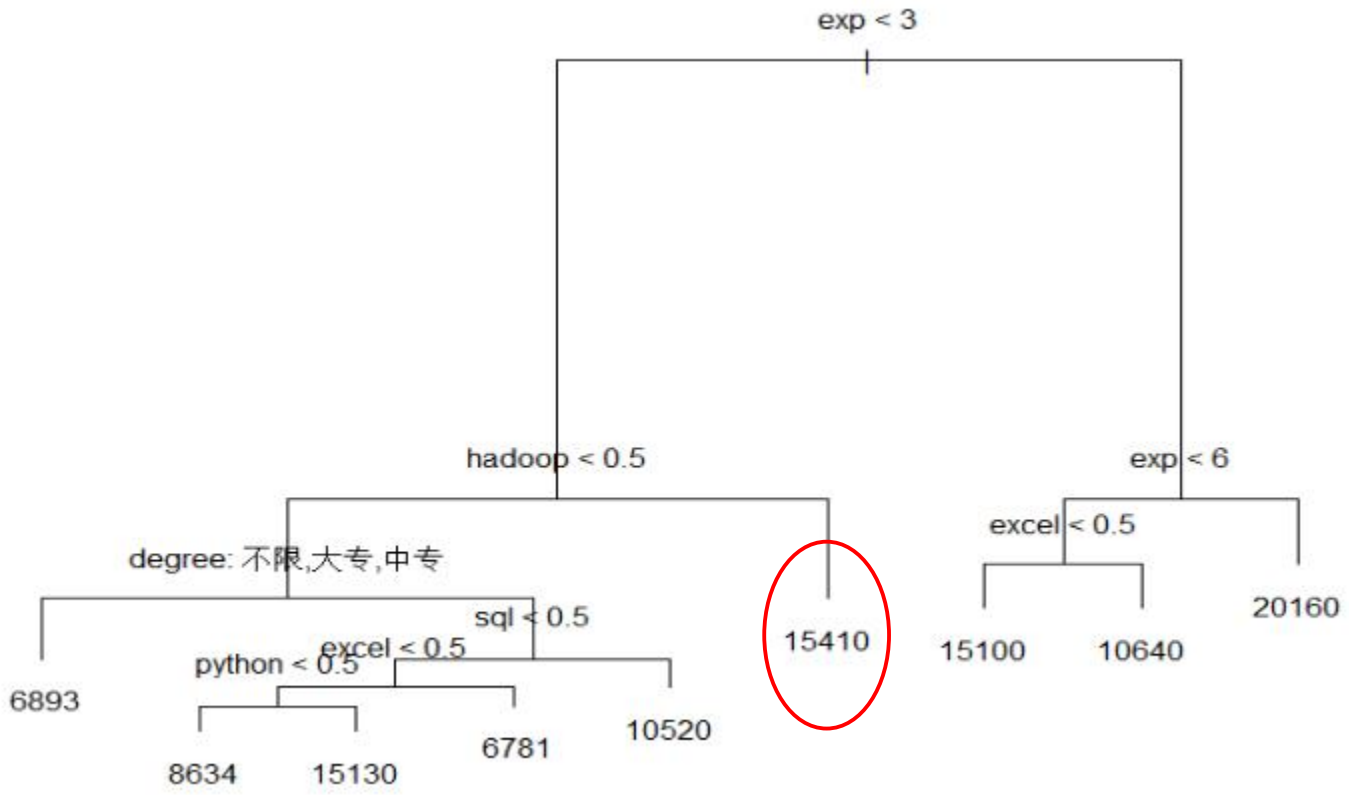
不论是Ridge还是Lasso回归，spark/硕士/sql/exp都是比较中要的因素，这些结果大致与前面的OLS方法类似。

决策树

由于我的数据集的特殊性（大多数变量都只有两个值，0和1），所以使用线性的拟合方法很难获得很好的预测效果（秩缺乏）。

然后，我用决策树方法进行拟合。决策树把数据分成若干个区域，每个区域的预测值为所在区域的数据的平均值决定。

应用了决策树之后的树状模型图



变量的位置越高，说明该变量在减少RSS的重要性越大。

如果我想得到红圈所显示的工资，那么我的经验可以小于3年，但是必须要会Hadoop。

改良决策树的一些方法

Tree pruning:俗称修枝。在决策数的基础上，增加一个控制变量数的penalty项(类似前面的ridge/lasso regression)。

Bagging:使用bootstrap方法，通俗讲就是用很多棵树（比如5000棵）进行拟合。

Random forest:在bagging的基础上，缩减变量的数量进行拟合。可以使用交叉验证法来确定投入模型的变量个数。

Boosting:用极少数量的变量进行逐次拟合，在更新前一次拟合的残差的基础上进行新的拟合。可以使用交叉验证发来确定后拟合模型的系数lambda。

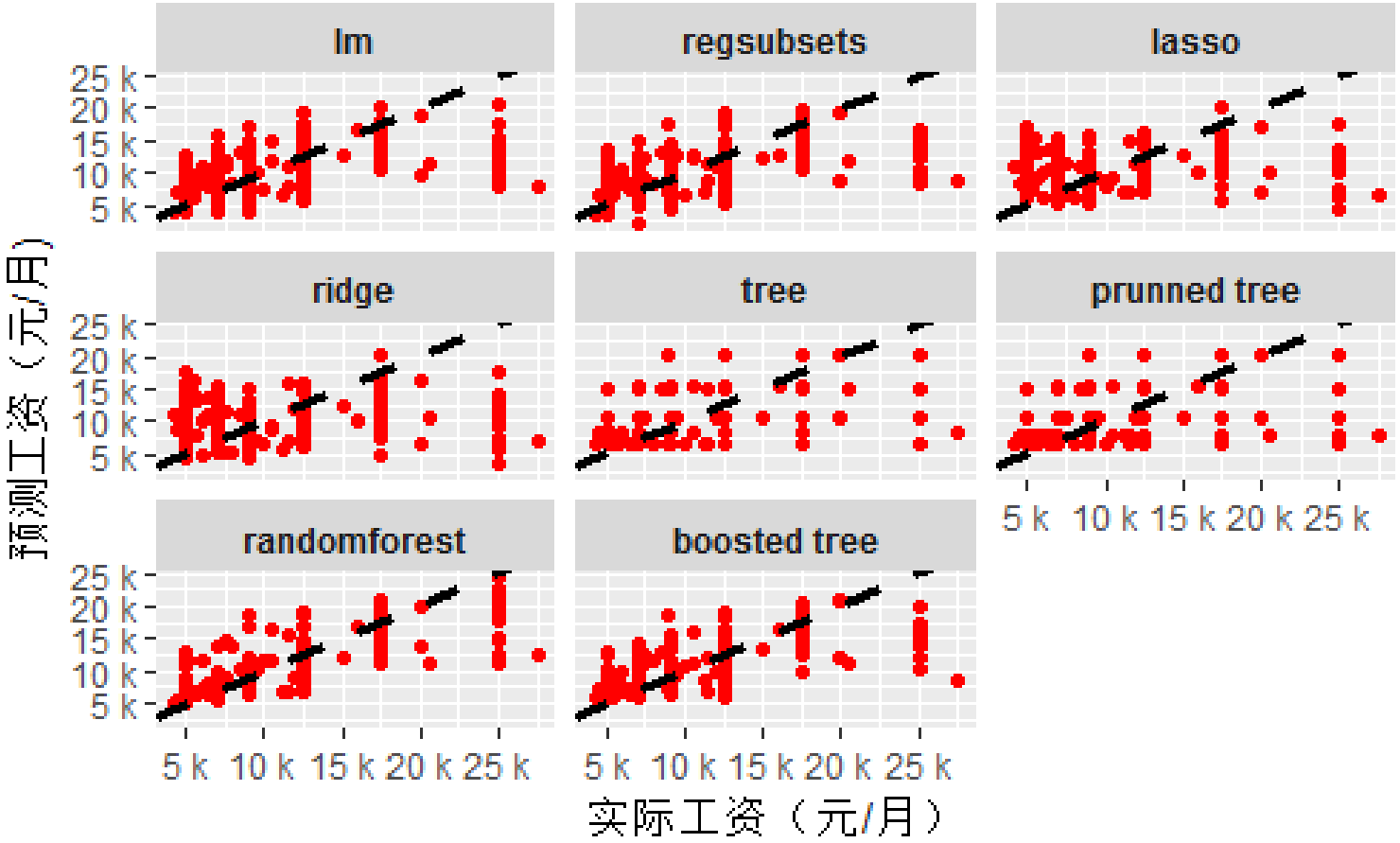
各种模型的拟合效果的比较

为了验证各种模型的拟合效果，我又一次爬取了智联招聘。把各个模型应用在新的数据集（新数据，703个数据）上，比较各个模型的预测误差（test error）。

model <fctr>	mean.error <dbl>
random forest	1059
boostes tree	1725
decision tree	2068
lm	2087
regsubset	2126
prunned tree	2130
lasso	4319
ridge	4529

可以看到各种决策树方法的拟合效果较好，尤其是随机森林，平均误差只有1059元/月。考虑到数据分析师的工资的范围比较大，这可以说是一个非常好的预测。

各种模型的预测工资与实际工资的比较



可以看到各个模型包括随机森林在工资比较高的位置，误差都比较大（虚线的坡度为45度，离它越远意味着误差越大，反之亦然）。意味着模型还有可改进的余地:

1 增加新的预测变量。

2 按工资大小进行分段地建模。

小结

1 由于数据集的特殊性，OLS及它的各种改进方法的拟合效果不佳。而随机森林的拟合效果较为出色。

2 我用随机森林预测了一些自己的期望工资，结果约为11400。结合各个模型的变量系数，如果我想提高自己的薪水，增加自己的经验，学习一些spark/Hadoop等大数据技能会比较关键。

3 尽管随机森林的预测效果不错，但仍然有改进的空间。可能我在下一次预测的时候需要加入新的变量，并对高薪水的数据进行另外地拟合。