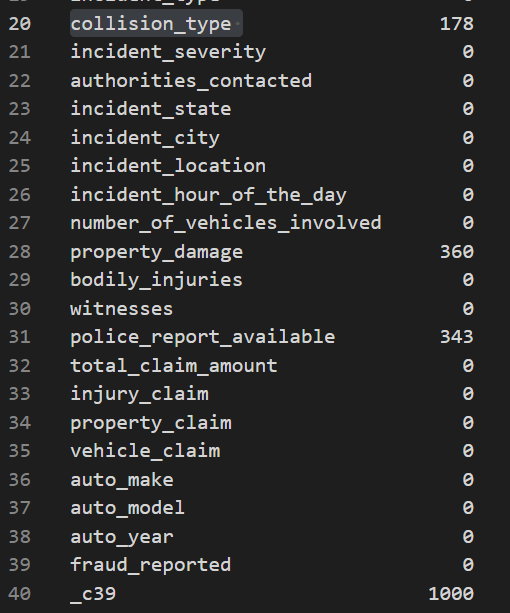
我们可以看到`age`和`months\_as\_customer`之间有很高的相关性。我们将删除“Age”列。 `total\_clam\_amount`，`injury\_claim`，`property\_claim`，`vehicle\_claim`之间也存在高度相关性，因为总索赔是所有其他索赔的总和。 所以我们将删除总索赔栏。

代码要点：

1. 各个模型的测试集准确率：



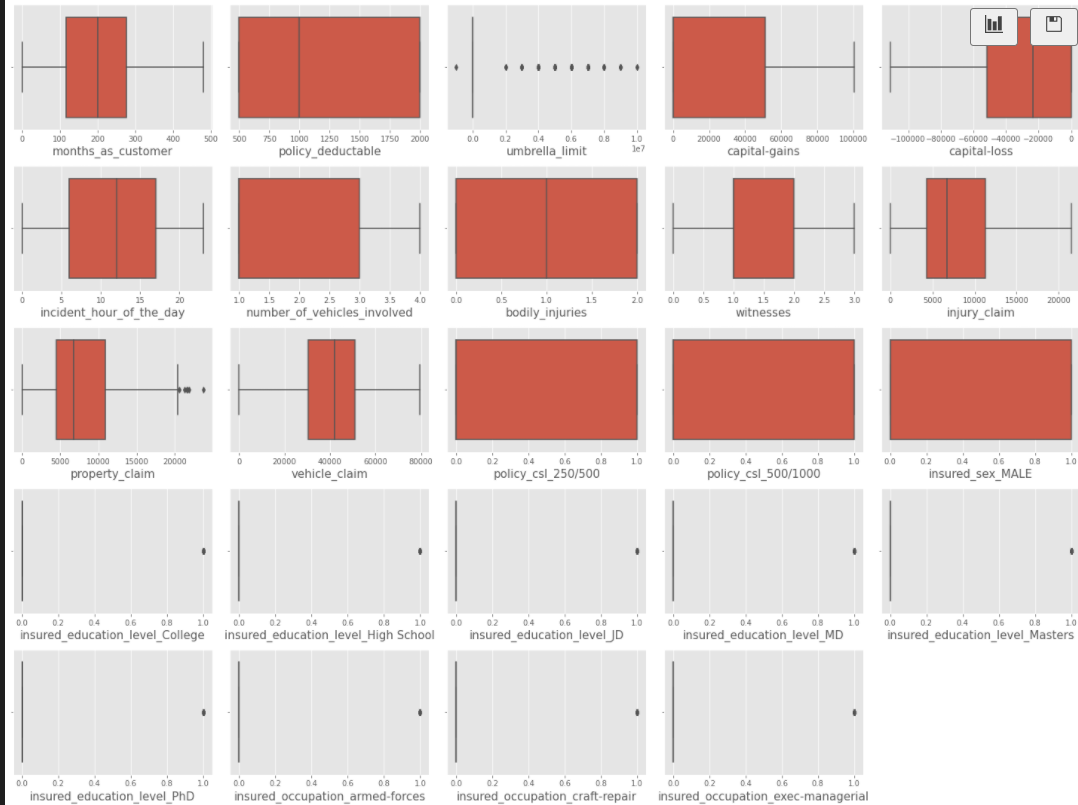
1. 原数据集缺失值情况，代码里面直接用第一个案件的对应值补上了



1. 与预测无关的列：

'policy\_number','policy\_bind\_date','policy\_state','insured\_zip','incident\_location','incident\_date', 'incident\_state','incident\_city','insured\_hobbies','auto\_make','auto\_model','auto\_year', '\_c39'在训练之前就删除了

1. 我们可以看到`age`和`months\_as\_customer`之间有很高的相关性。我们将删除“Age”列。 `total\_clam\_amount`，`injury\_claim`，`property\_claim`，`vehicle\_claim`之间也存在高度相关性，因为总索赔是所有其他索赔的总和。 所以我们将删除总索赔栏。
2. 对于离散型数据列，代码用one hote编码取代了；数值型的特征没有处理
3. 离群值检测：



1. 25%的数据作为测试集
2. 决策树、ada boost、xg boost进行了调参X
3. 组合单一限额 组合单一限额 (CSL) CSL 是一个单一数字，用于描述每次发生或事故的人身伤害责任保险和财产损失责任保险的总和的预定限额。

训练用数据列名：：

['months\_as\_customer',

'policy\_deductable',

'umbrella\_limit',

'capital-gains',

'capital-loss',

'incident\_hour\_of\_the\_day',

'number\_of\_vehicles\_involved',

'bodily\_injuries',

'witnesses',

'injury\_claim',

'property\_claim',

'vehicle\_claim',

'policy\_csl\_250/500',

'policy\_csl\_500/1000',

'insured\_sex\_MALE',

'insured\_education\_level\_College',

'insured\_education\_level\_High School',

'insured\_education\_level\_JD',

'insured\_education\_level\_MD',

'insured\_education\_level\_Masters',

'insured\_education\_level\_PhD',

'insured\_occupation\_armed-forces',

'insured\_occupation\_craft-repair',

'insured\_occupation\_exec-managerial',

'insured\_occupation\_farming-fishing',

'insured\_occupation\_handlers-cleaners',

'insured\_occupation\_machine-op-inspct',

'insured\_occupation\_other-service',

'insured\_occupation\_priv-house-serv',

'insured\_occupation\_prof-specialty',

'insured\_occupation\_protective-serv',

'insured\_occupation\_sales',

'insured\_occupation\_tech-support',

'insured\_occupation\_transport-moving',

'insured\_relationship\_not-in-family',

'insured\_relationship\_other-relative',

'insured\_relationship\_own-child',

'insured\_relationship\_unmarried',

'insured\_relationship\_wife',

'incident\_type\_Parked Car',

'incident\_type\_Single Vehicle Collision',

'incident\_type\_Vehicle Theft',

'collision\_type\_Rear Collision',

'collision\_type\_Side Collision',

'incident\_severity\_Minor Damage',

'incident\_severity\_Total Loss',

'incident\_severity\_Trivial Damage',

'authorities\_contacted\_Fire',

'authorities\_contacted\_None',

'authorities\_contacted\_Other',

'authorities\_contacted\_Police',

'property\_damage\_YES',

'police\_report\_available\_YES']

[“作为客户的月份”，

“保单免赔额”，

“雨伞限制”，

“资本收益”，

“资本损失”，

“一天中的事件”，

“涉及的车辆数量”，

“身体伤害”，

“证人”，

“伤害索赔”，

“财产索赔”，

“车辆索赔”，

“政策csl 250/500”，

“政策\_csl\_500/1000”，

“投保性别男性”，

“受保教育水平大学”，

“受保教育水平高中”，

“被保险人教育水平JD”，

“被保险人教育程度MD”，

“受保教育程度硕士”，

“被保险人教育程度博士”，

“被保险职业武装部队”，

“被保险职业和工艺维修”，

“被保险人职业高管”，

“保险职业”和“农耕渔业”，

“被保险职业”和“清洁工人”，

“被保险人职业机器操作检查”，

“被保险职业和其他服务”，

“被保险人职业私人住宅服务”，

“投保职业专业”，

“投保职业保护性服务”，

“投保、职业、销售”，

“投保职业技术支持”，

“投保职业运输”，

“投保关系”和“非家庭成员”，

“被保险人关系或其他亲属”，

“投保关系”和“亲生子女”，

“投保关系”和“未婚”，

“被保险人与妻子的关系”，

“事故类型停车”，

“事故类型单车碰撞”，

“事故类型车辆盗窃”，

“碰撞类型碰撞”，

“碰撞类型碰撞”，

“事故\_严重性\_轻微损坏”，

“事故\_严重程度\_总损失”，

“事故\_严重性\_轻微损害”，

“当局联系了火灾”，

“当局没有联系”，

“当局联系了其他人”，

“当局联系了警察”，

“财产损失是的”，

“警察报告可用是”]