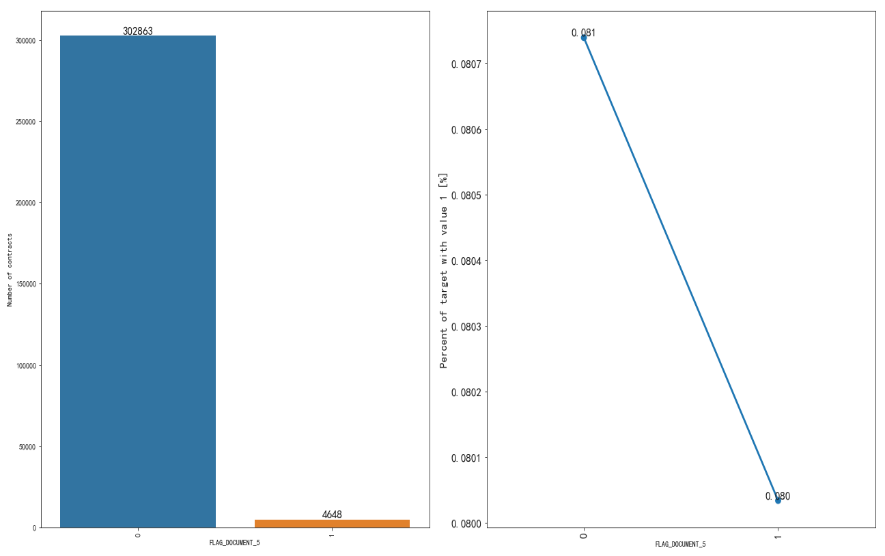
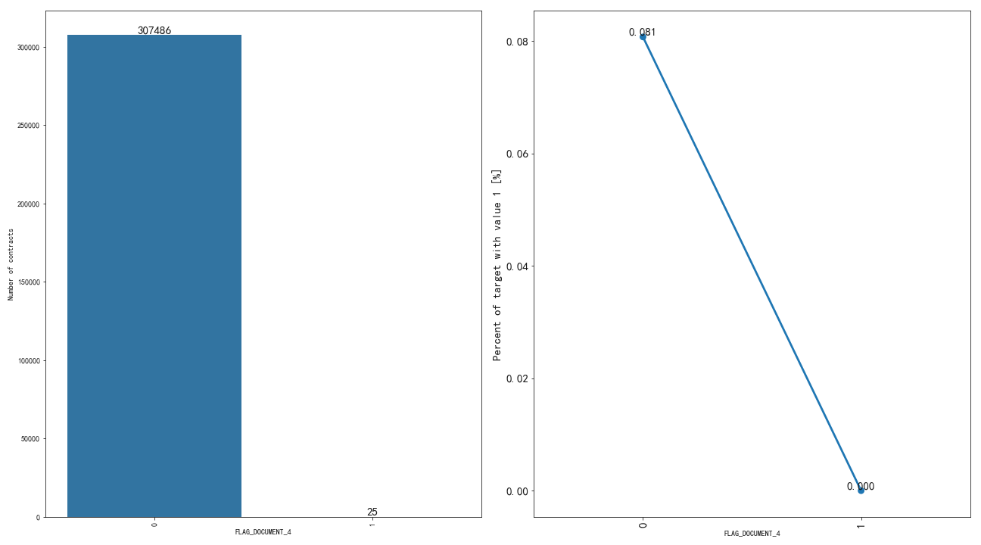
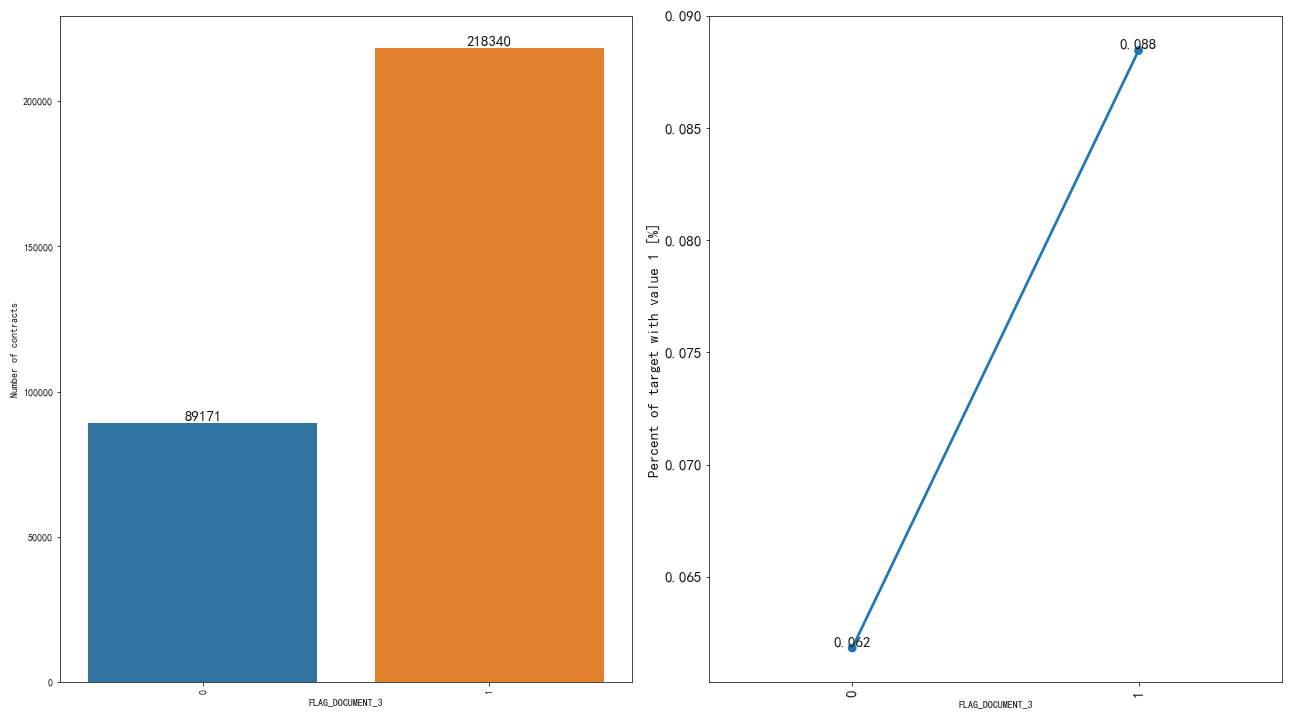
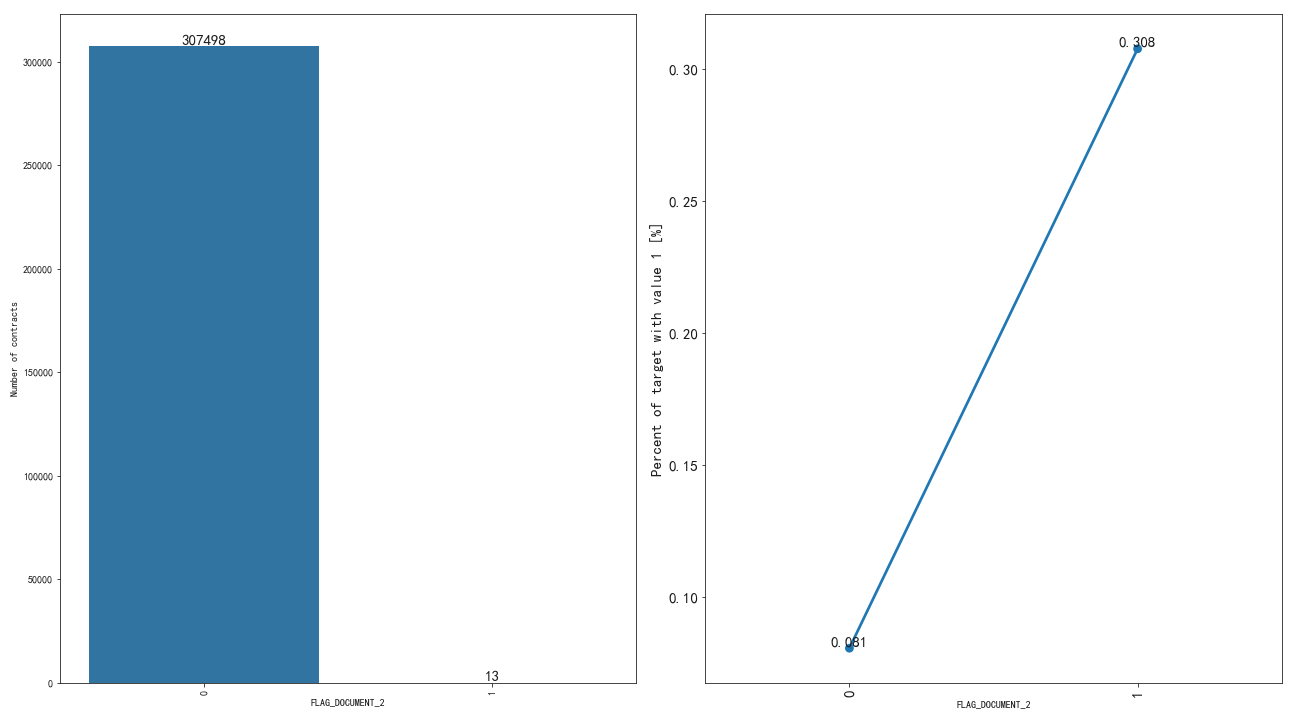
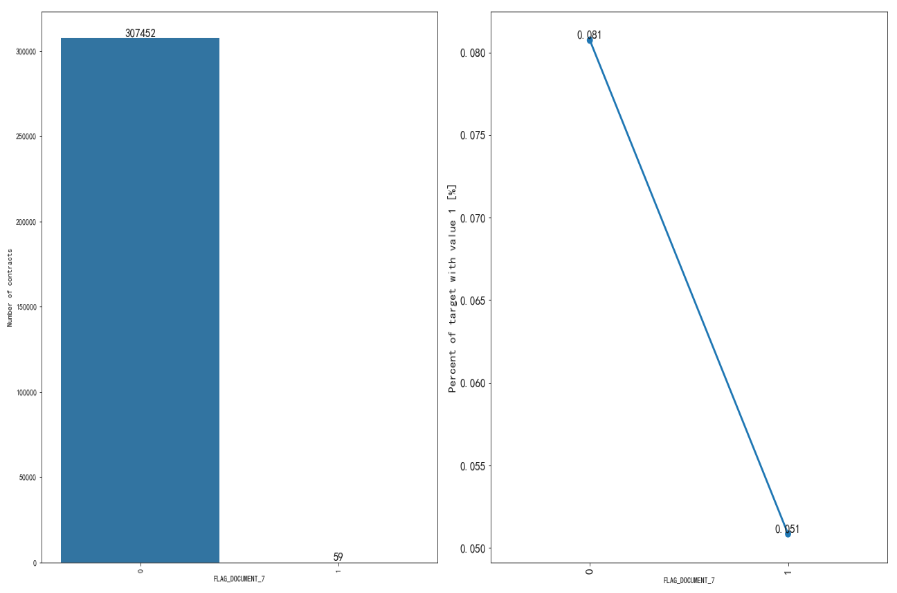
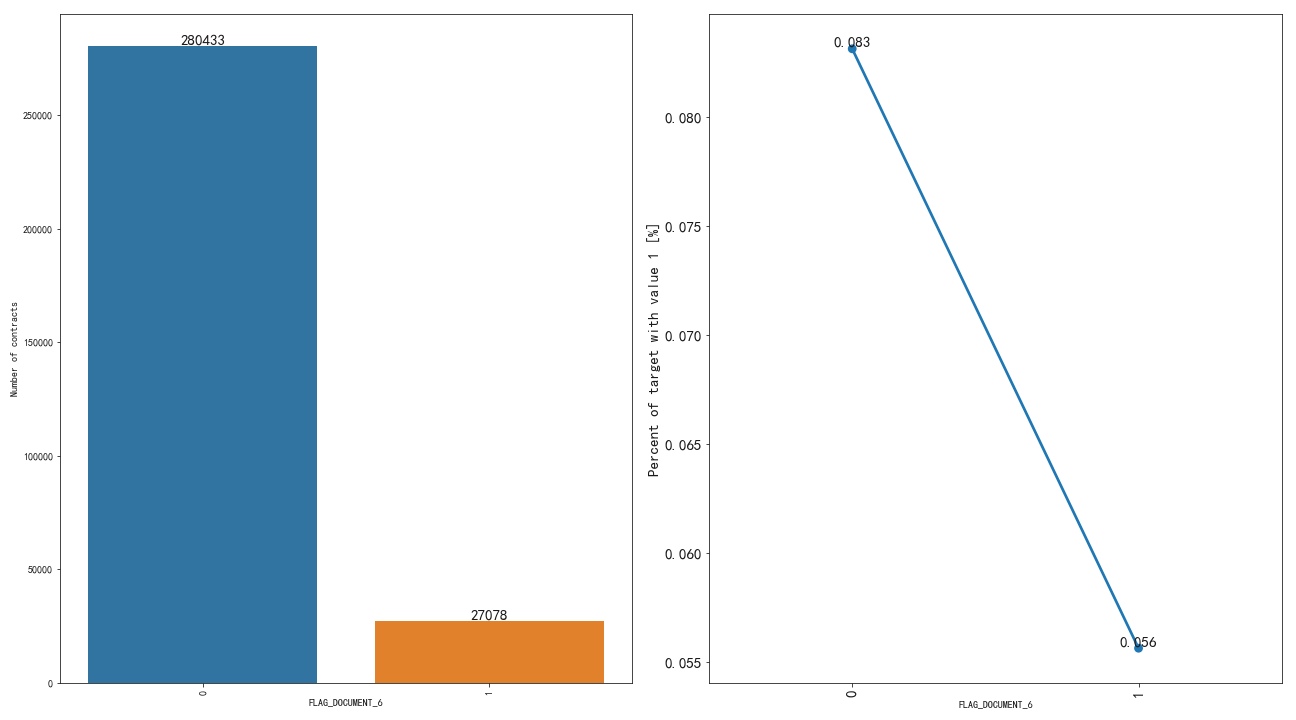
Main table 平均违约率为0.0807

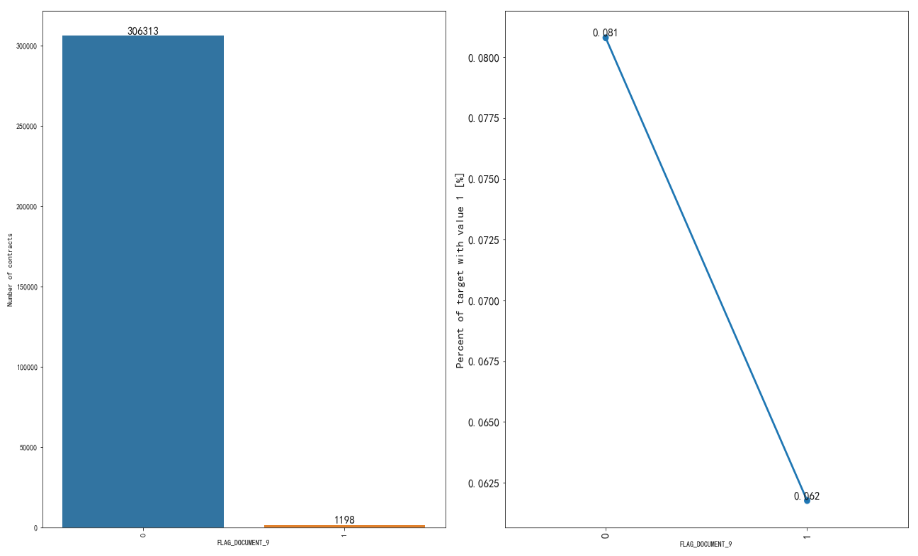
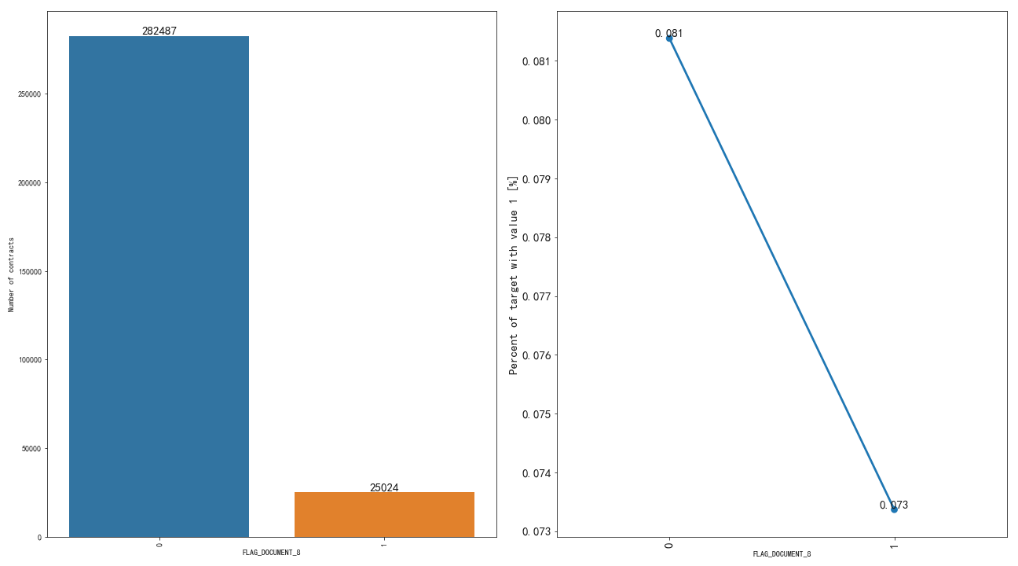
# 贷款人信息完整程度

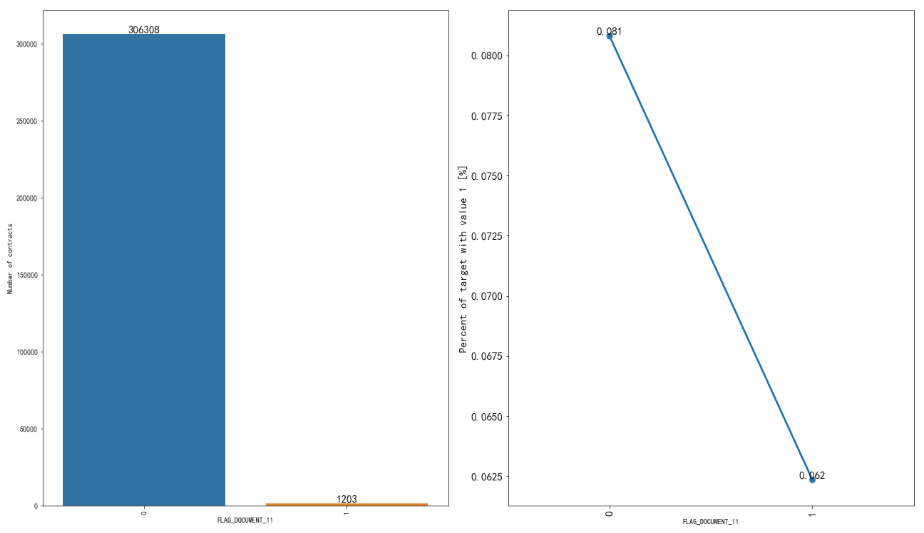
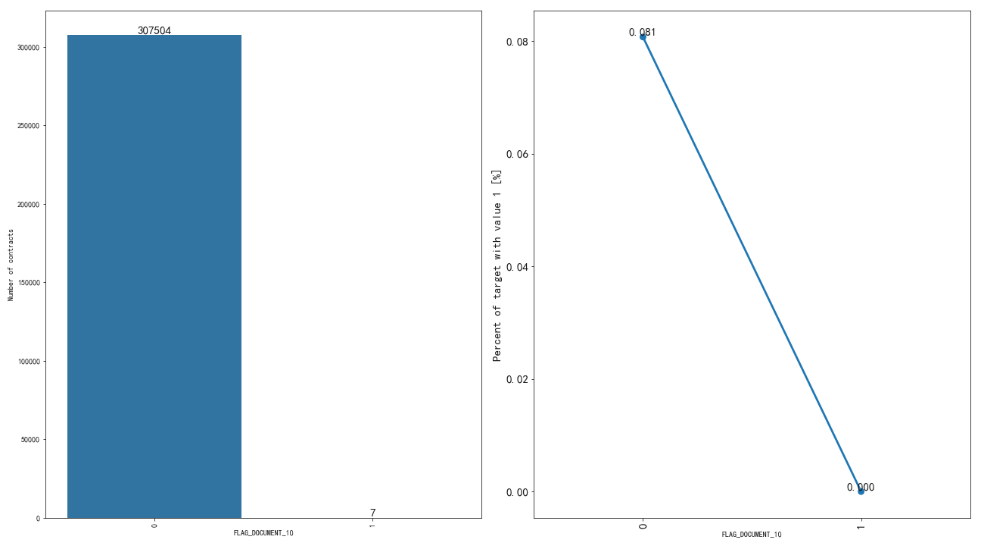
## 提交文件完整程度

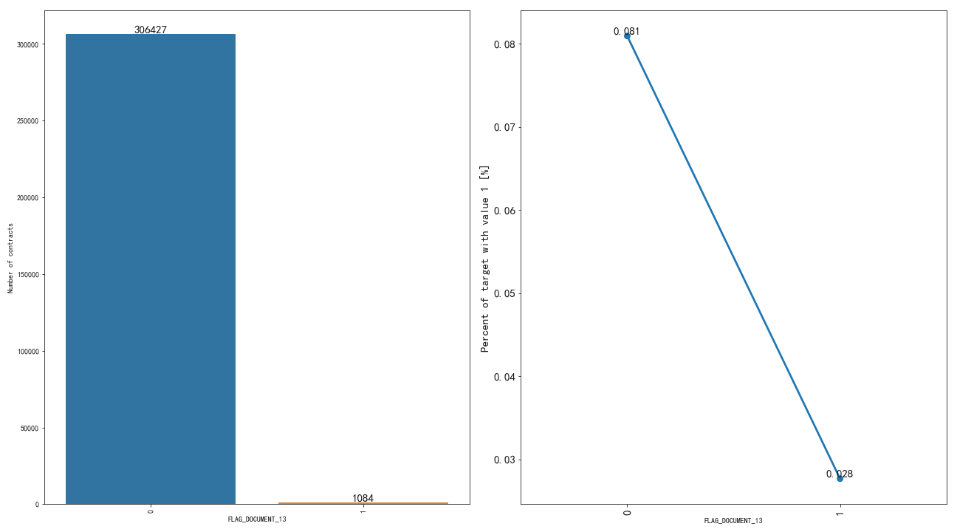
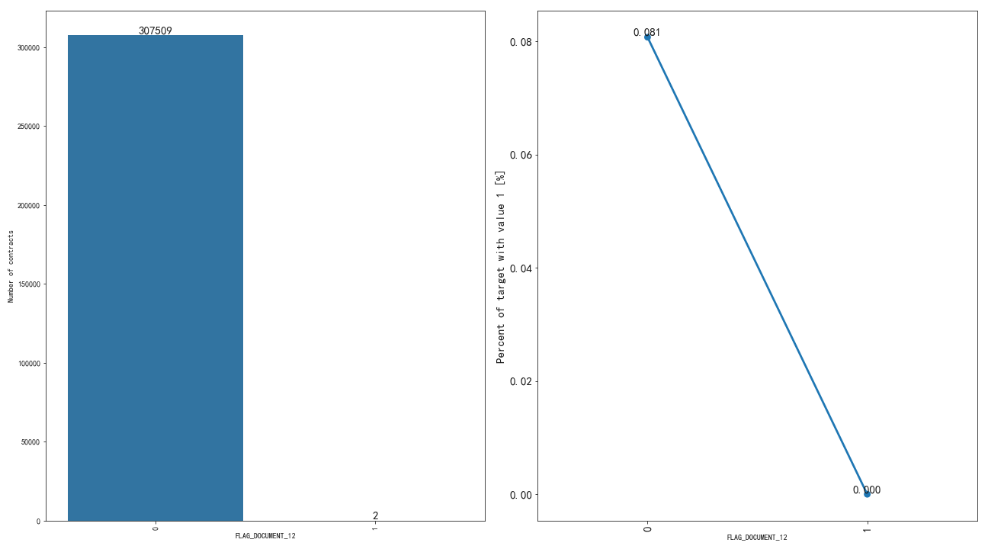
### FLAG\_DOCUMENT\_X(x=(2-21)

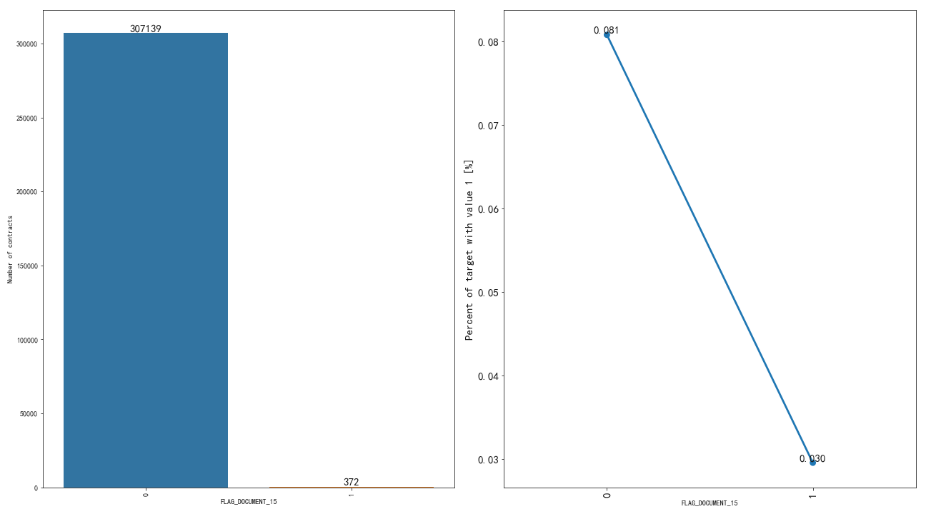
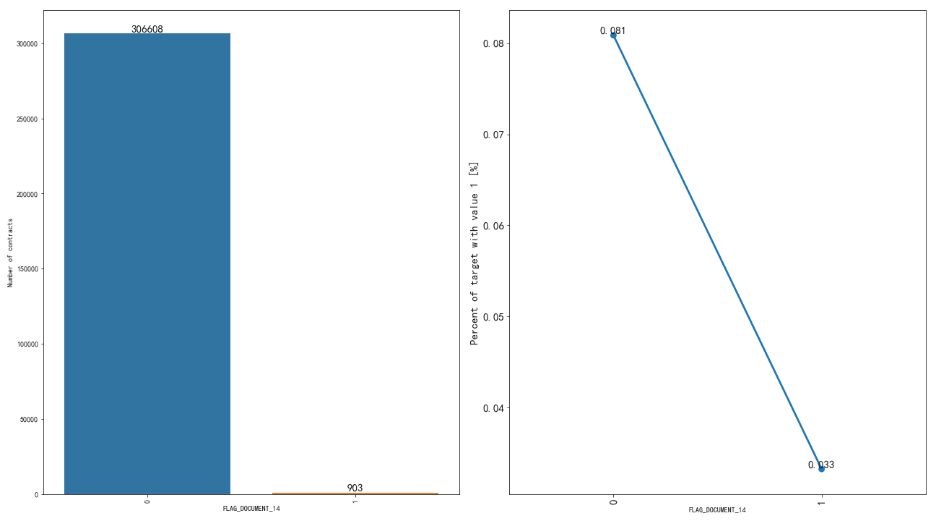


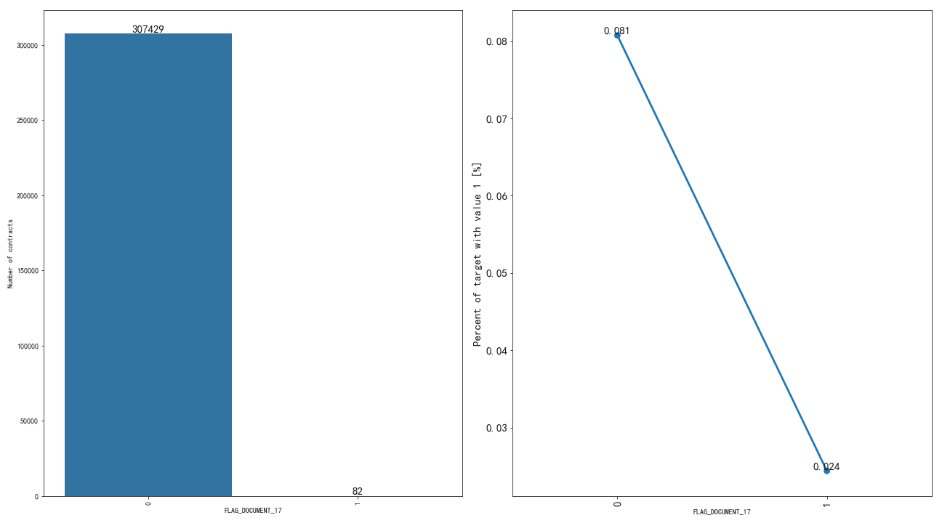
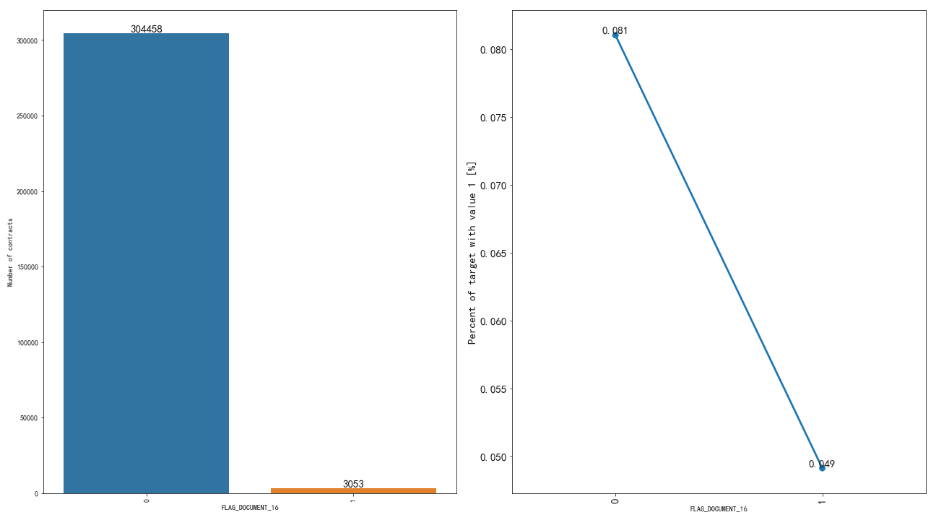


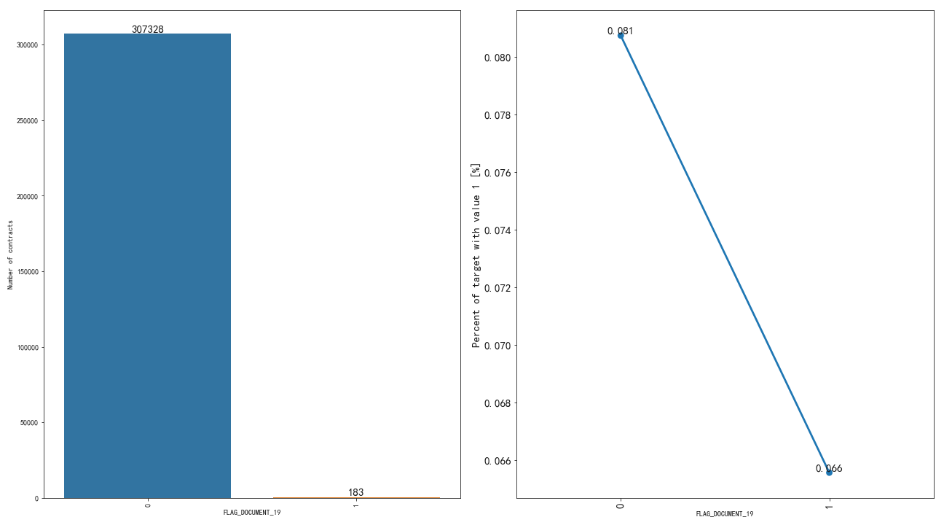
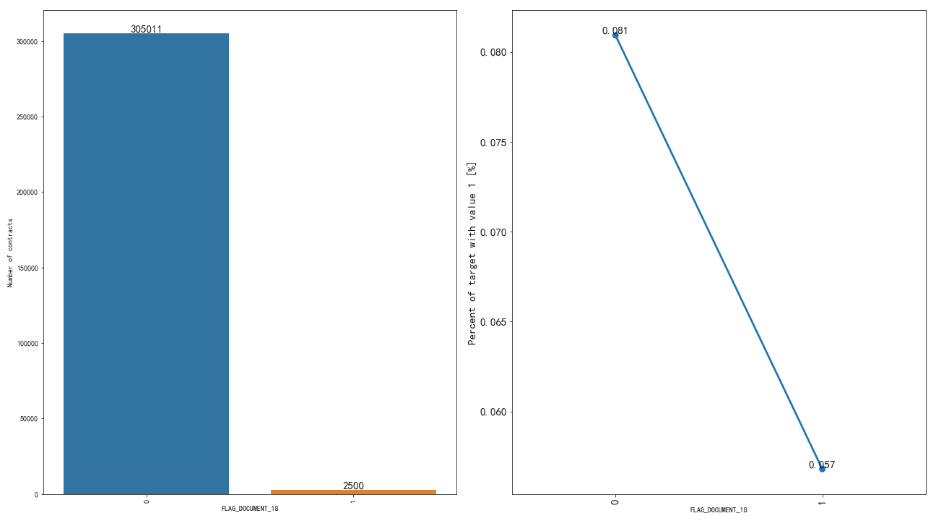


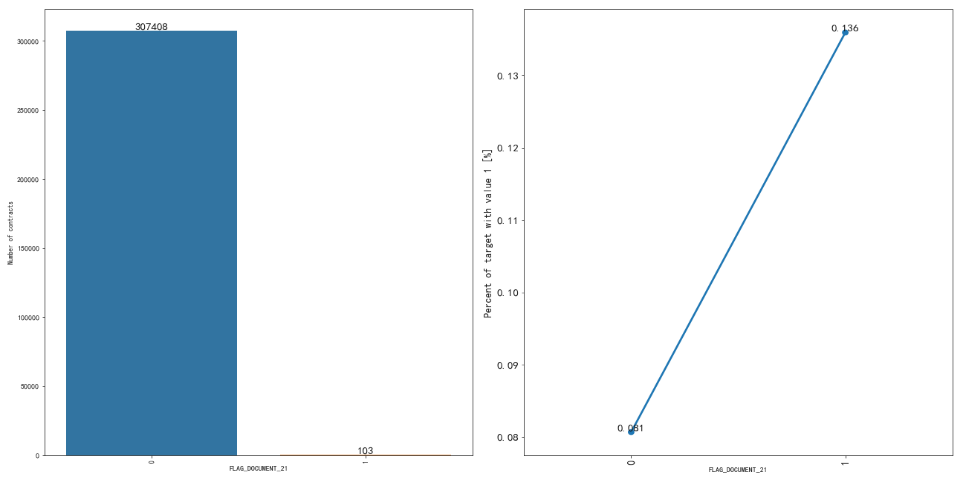
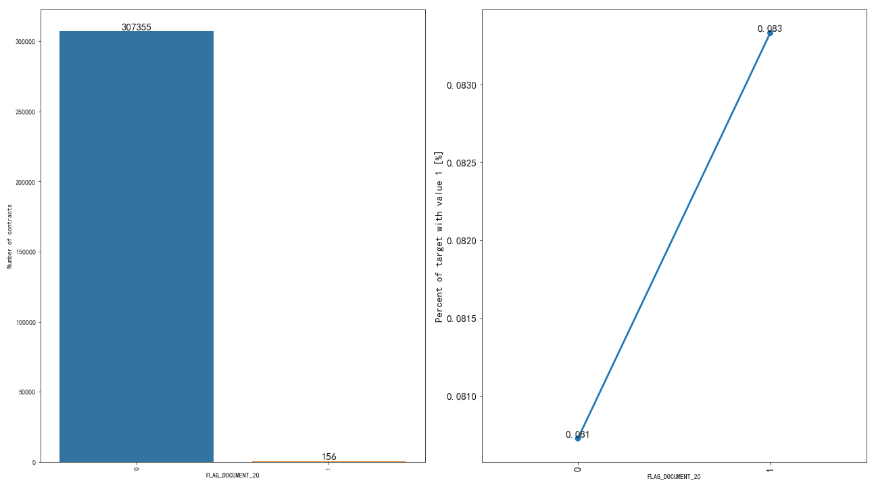






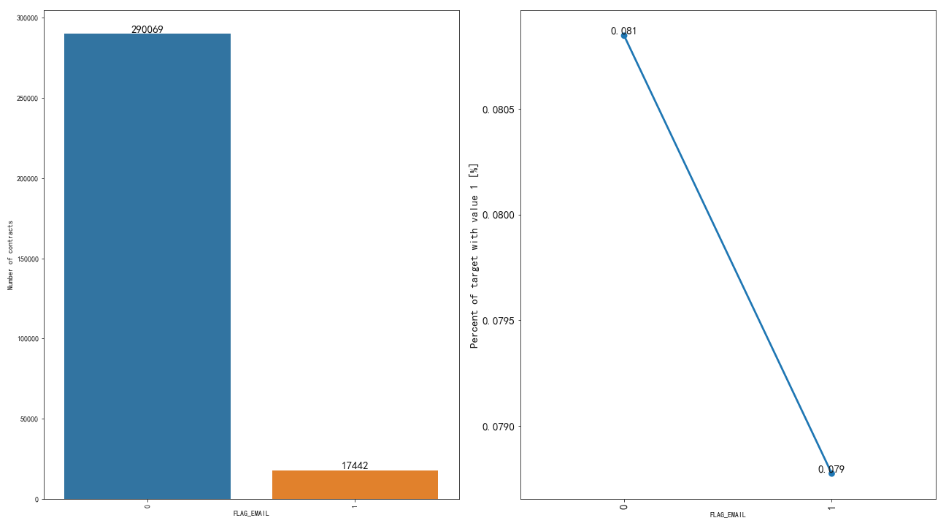
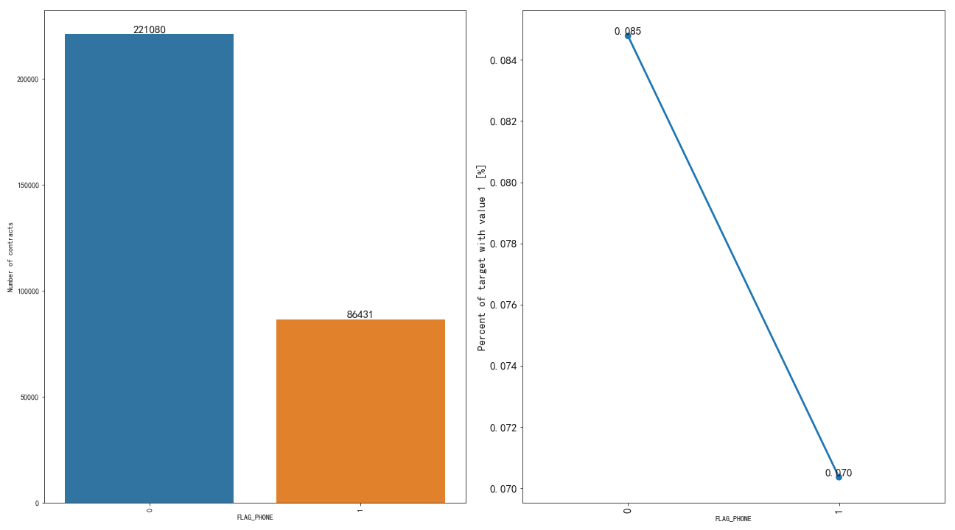
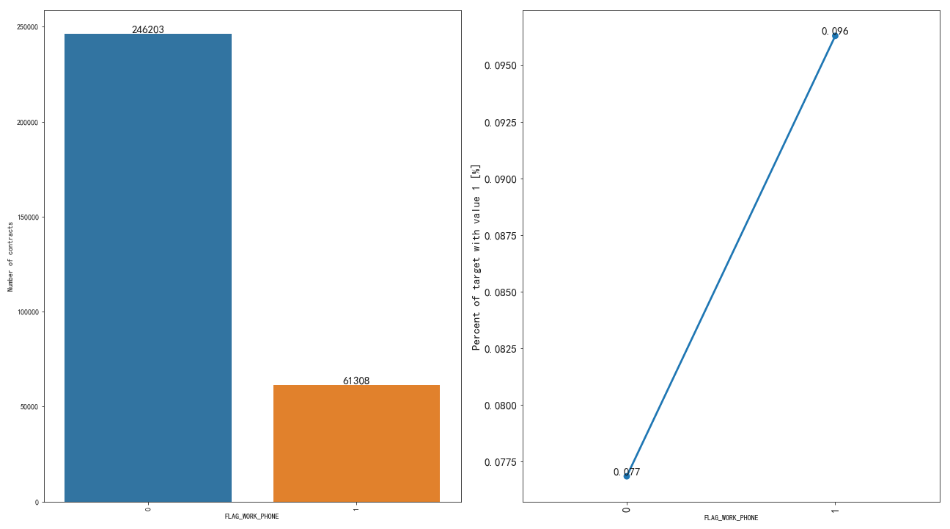
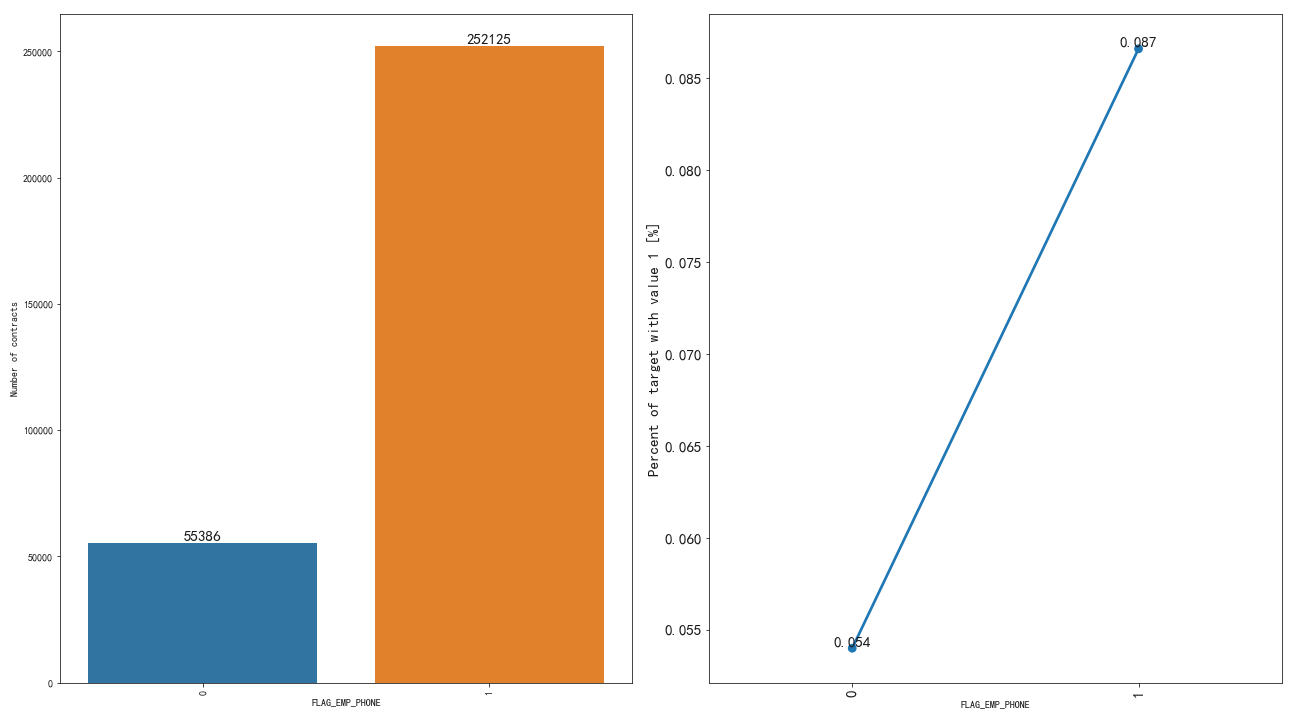
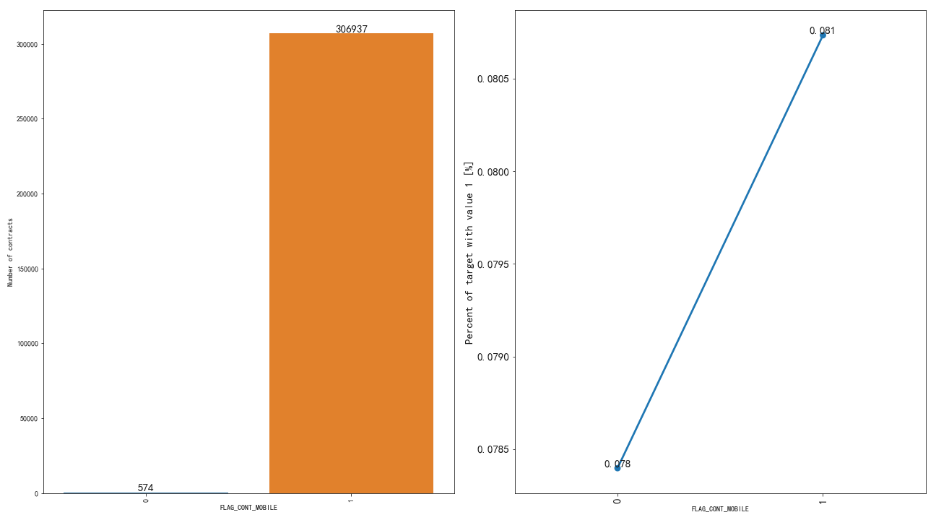
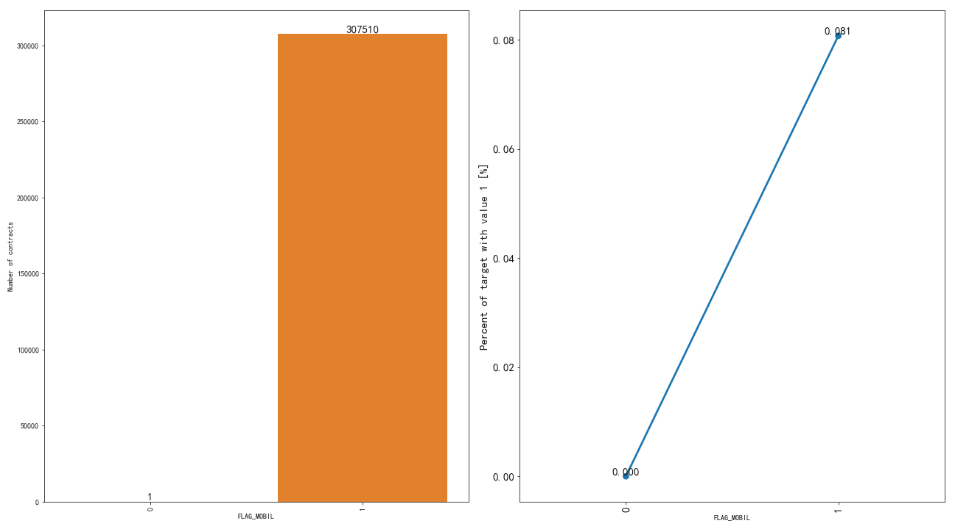






* 20个字段无缺失，为布尔型变量。
* 提供文件4-19的贷款记录占比非常非常小，flag=0的贷款记录的违约率接近于整体违约率，flag=1的违约率略小于整体违约率。
* 提供20文件的人也很少，flag=0与1时违约率很接近，可以说该字段对影响很小
* 原始特征包含的信息较少，可以提取特征后剔除
* 分析每条记录提供文件4至20的个数与违约率的关系
* 提供2，21文件的人也很少，但是提供了这个文件却违约的贷款记录占比较高，违约率达到了0.3和0.13。远远高于整体违约率。
* 与上述合并 同处理
* 提供文件3的人较多，但提供了该文件的贷款记录违约率略高于为提供的。
* 不处理

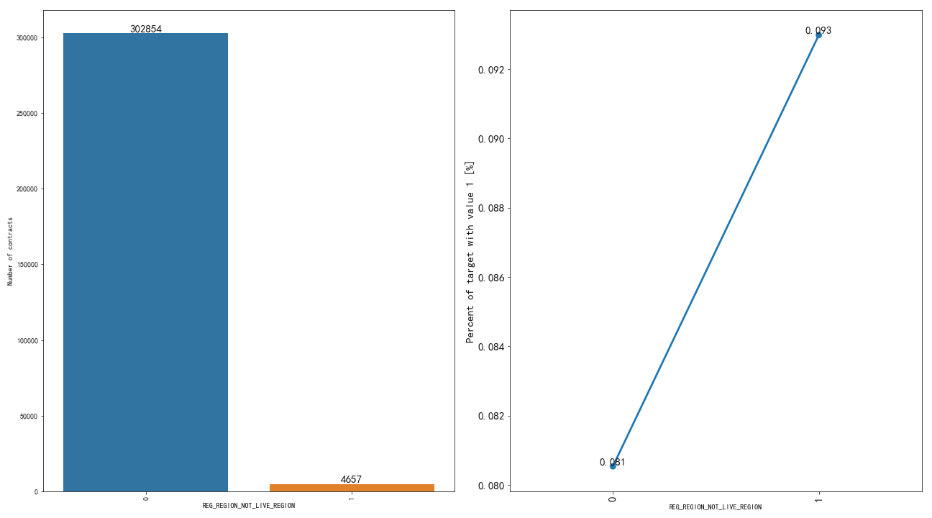
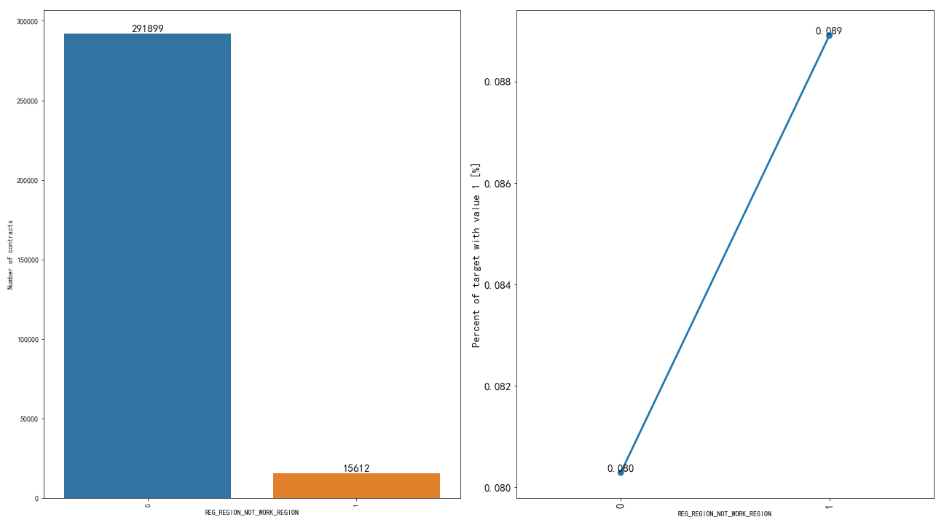
### 电话邮件信息完整程度

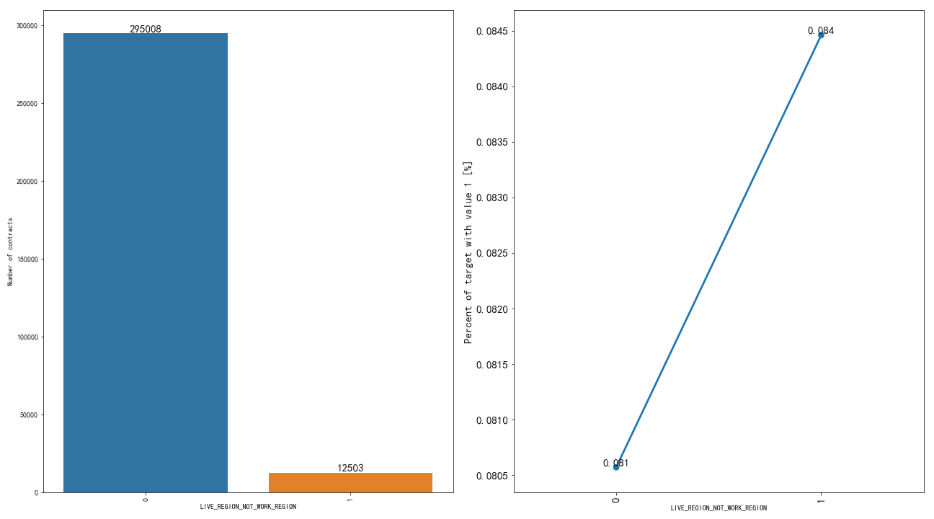


* 20个字段无缺失，为布尔型变量。
* 只有1条贷款记录没有提供手机。极少数贷款记录的手机号打不通。
* 提供家庭电话的贷款记录较少，违约率较小
* 提供邮件的贷款记录较少，违约率接近。
* ★提供雇主电话较多，提供自己工作电话的人较少。提供者比未提供者的违约率更高。
* FLAG\_MOBIL，FLAG\_CONT\_MOBILE，FLAG\_EMAIL原始特征包含的信息较少，可以提取特征后剔除
* 分析每条记录提供电话和邮件地址个数与违约率的关系
* 分析提供老板电话和工作电话与违约率关系

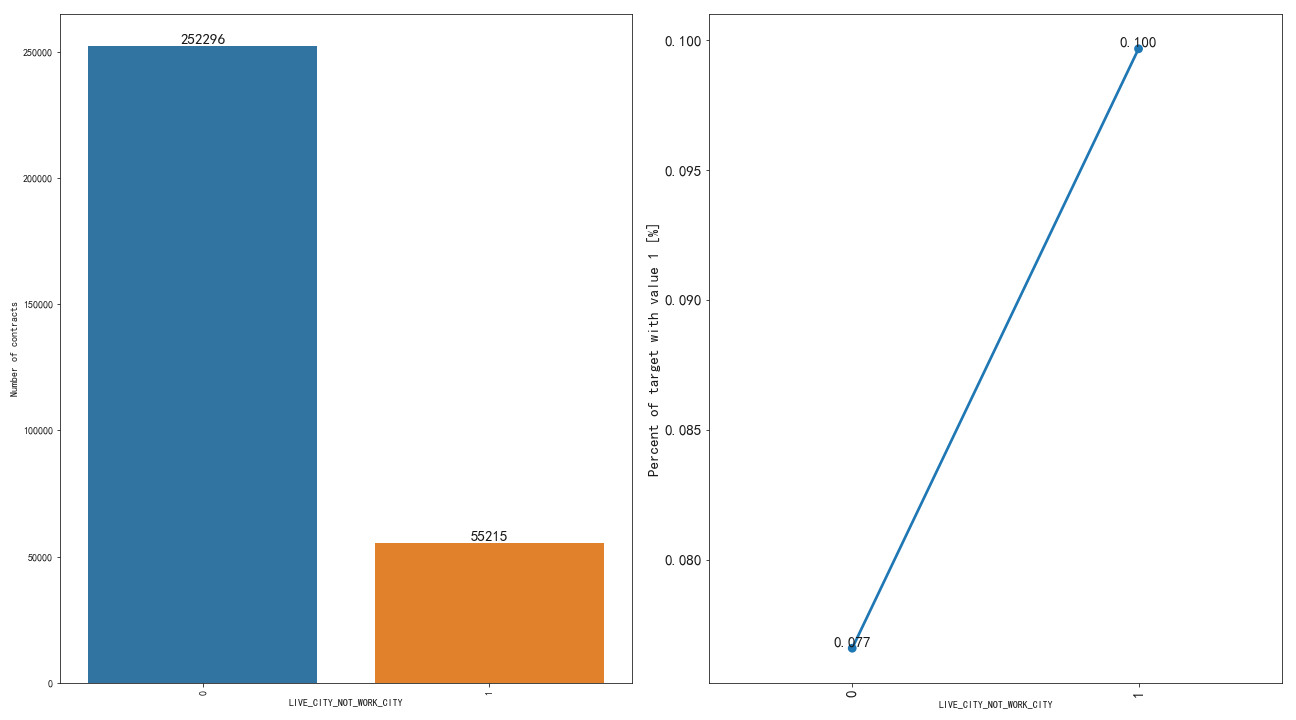
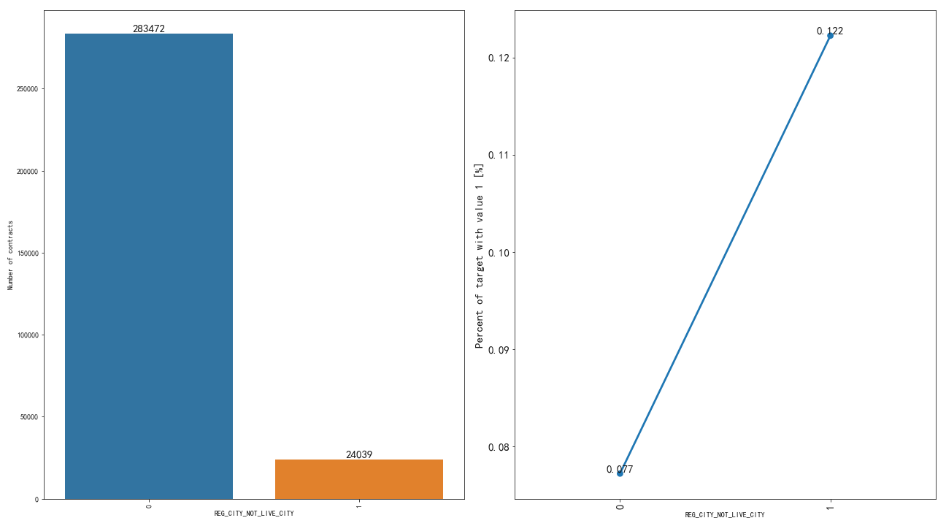
# 地址信息（工作地址，生活地址，户籍地址三者是否相同）

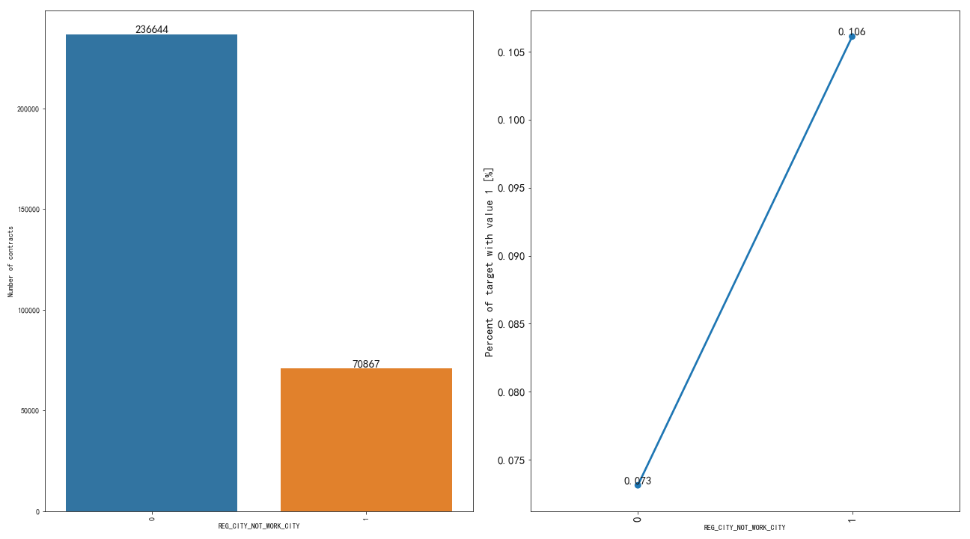
## 从REGION层面

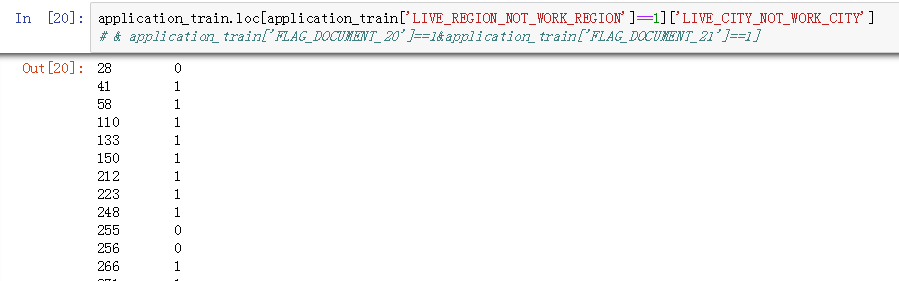


## 从city层面



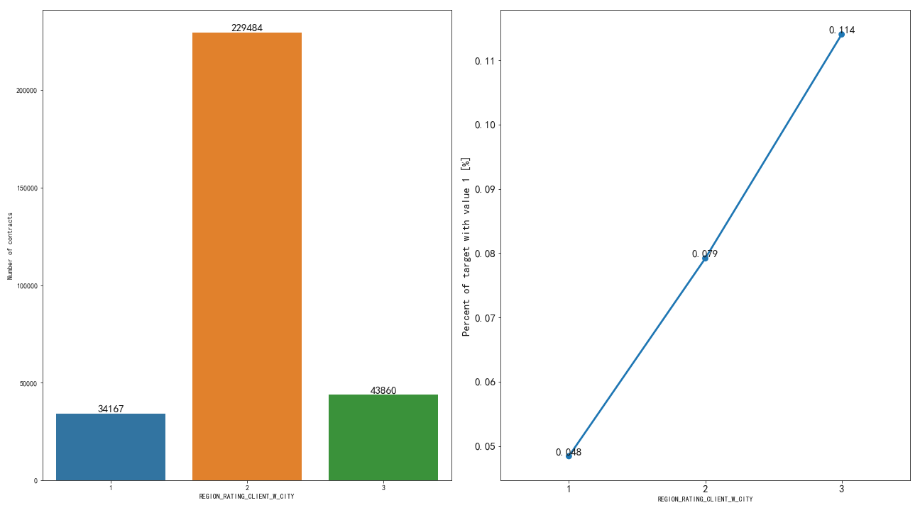
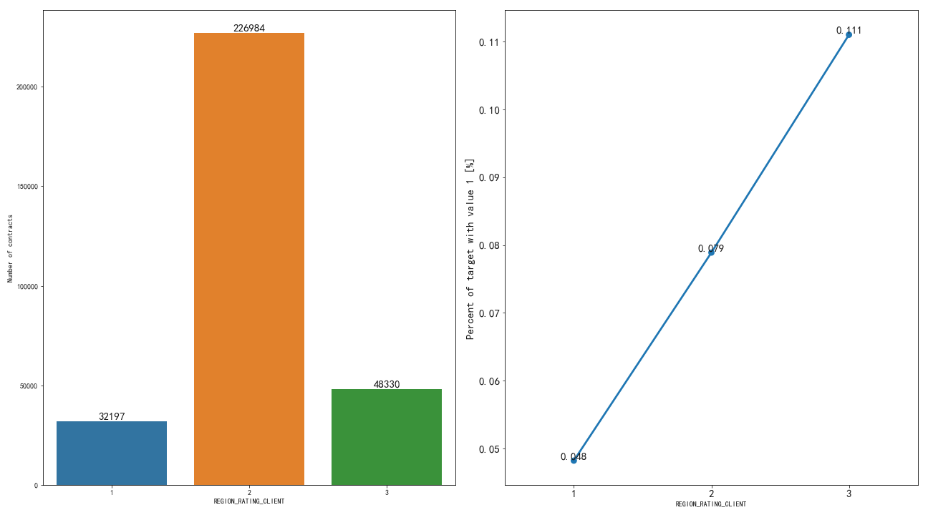
* 20个字段无缺失，为布尔型变量，1代表不同，0代表相同。
* 在city层面地址不同的记录多于在region层面，说明city的范围比region小，按常理说在大范围上region层面地址不同的，在city层面肯定是不同的。然后在实际数据中，在region层面地址不同的，在city层面上却是相同的。这是为什么？城市同名



* 共有5中情况，户籍所在城市与工作城市不同， 户籍所在城市与生活城市不同，工作城市与生活城市不同，三者都不相同说明该申请人生活非常不稳定，三者都相同说明该申请人生活较为稳定。不同地址越多，违约可能越大。
  + 提取特征 不同地址个数
  + 分析当有两个地址不同时，违约率高低排名
  + 比较相关性高低，哪个高就用哪个层面

# 生活环境(归一化后)

## 生活环境评级



# 社交圈信用水平

