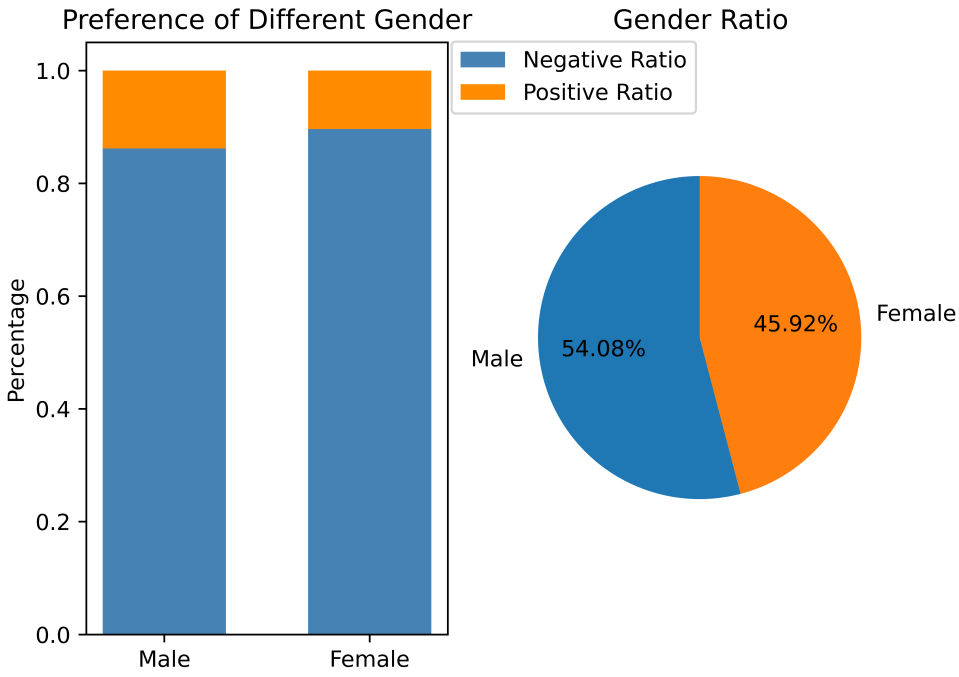
第二次上机作业

姓名：王崇斌；学号：1800011716

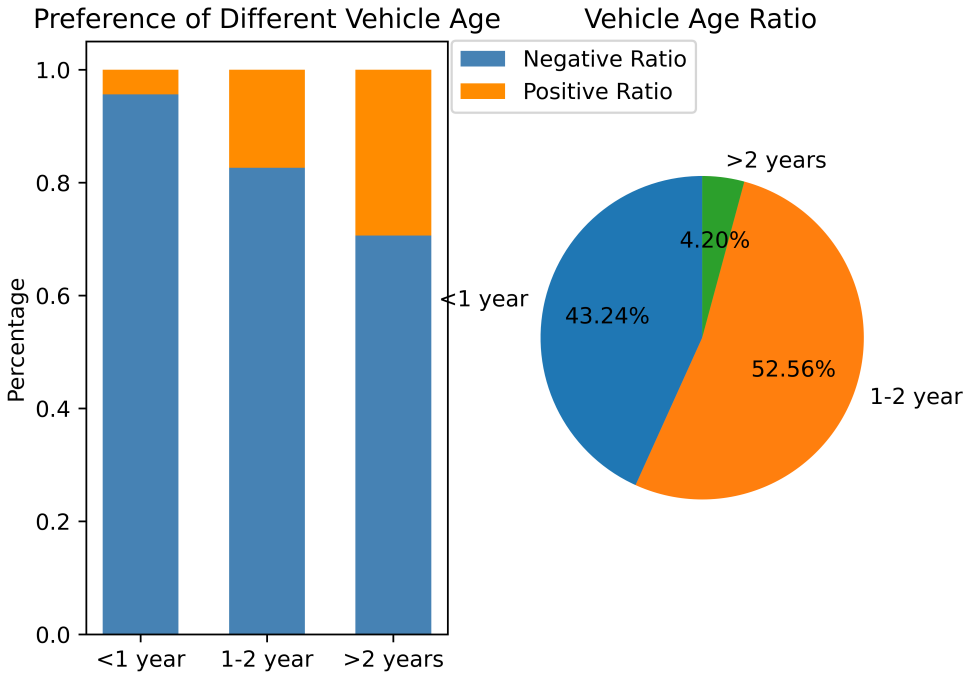
(1)

为了分析题中所述的四种因素与用户购买保险倾向之间的关系，我们使用柱状图+扇形图。以“汽车年限”这一类别举例，通过柱状图反映出不同汽车年限的用户购买保险的比例，再通过饼状图反映出不同汽车年限用户的比例，这样就可以完整的反映整个数据集的特征。（为了清晰起见，我没有将画图的代码写在`logistic_classfication.py`文件中，而是创建了另外的画图代码文件`plot_1.py`）

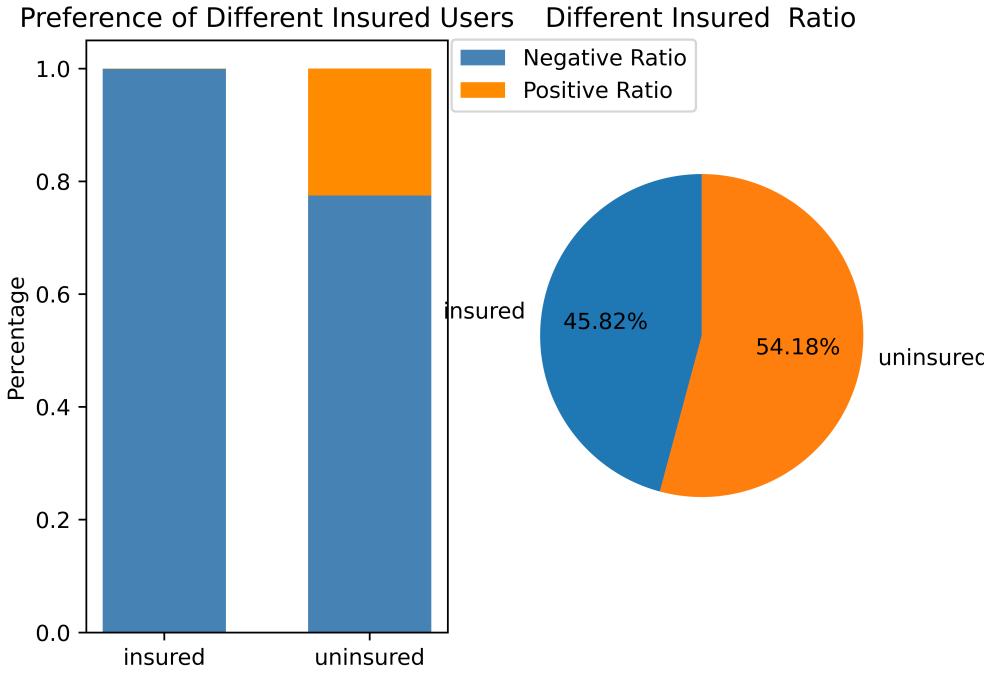
不同性别的用户购买保险的意愿：



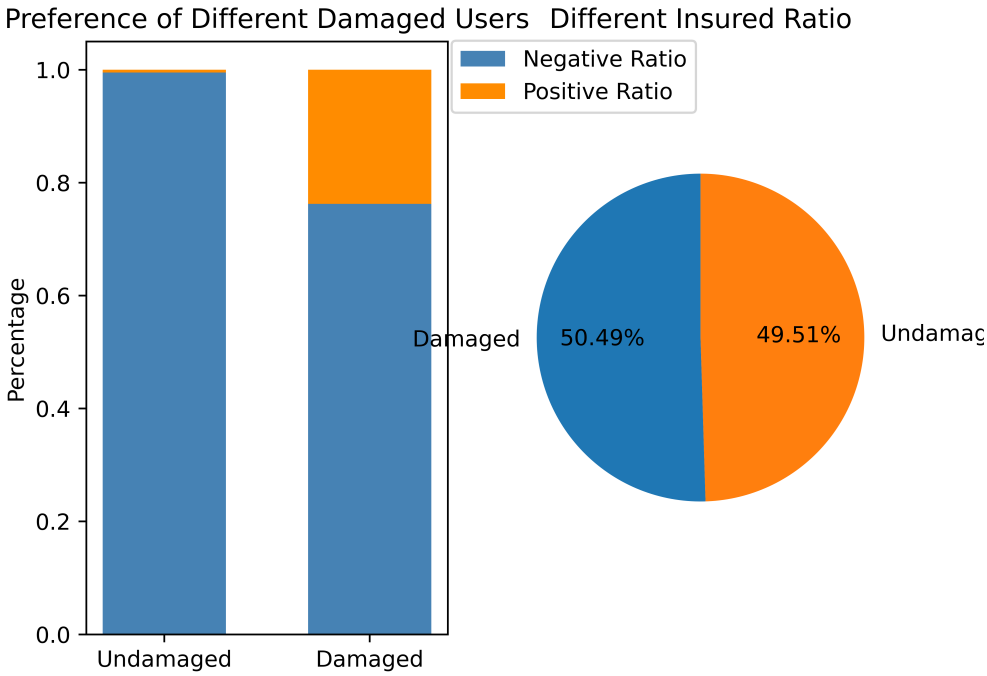
不同汽车使用年限用户的购买保险意愿：



之前购买保险情况不同的用户购买保险的意愿：

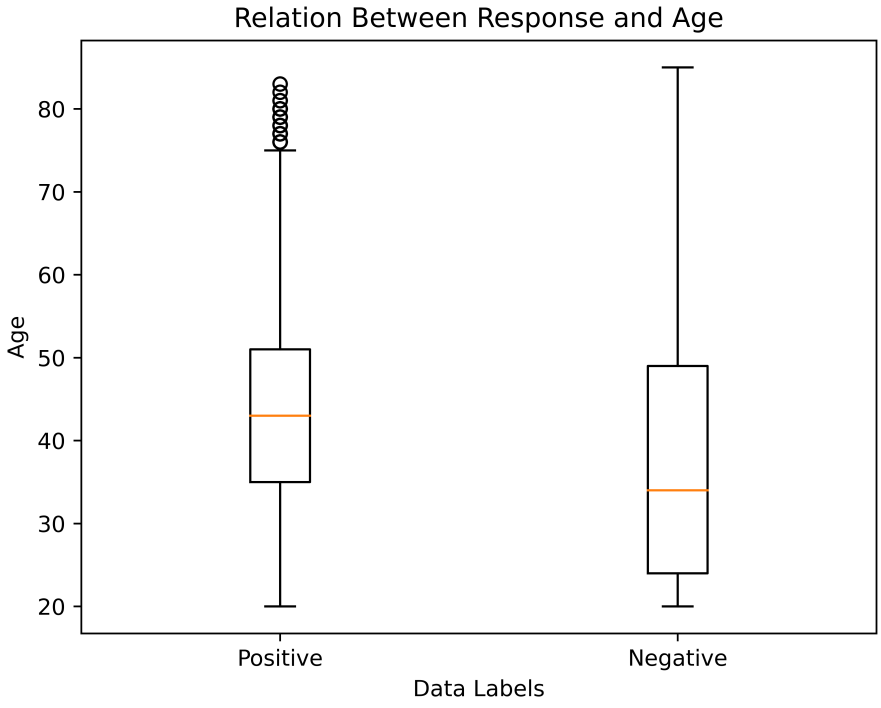


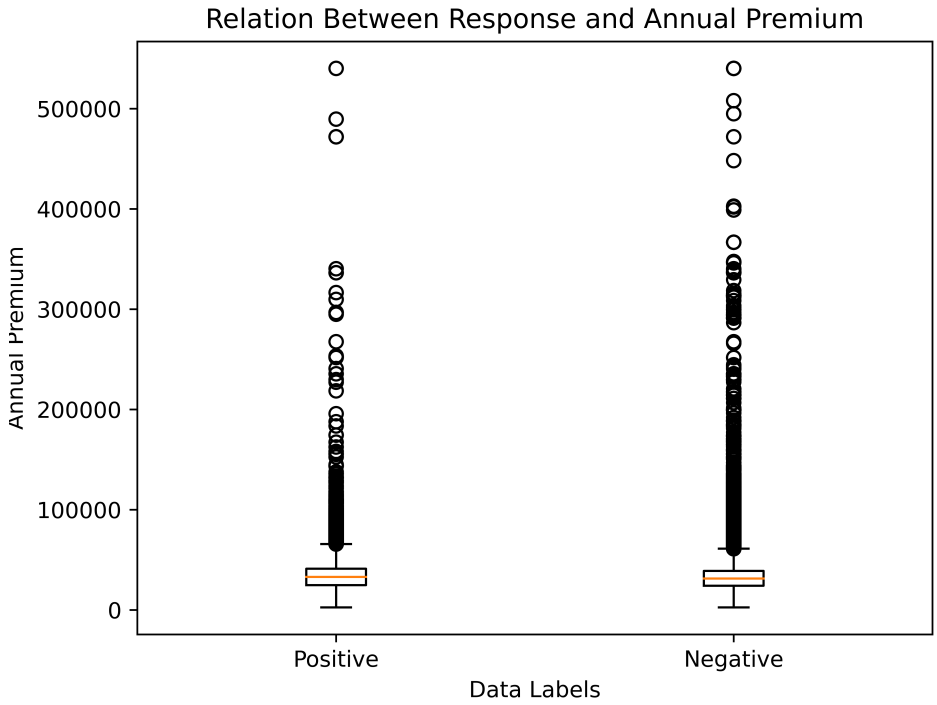
不同汽车损坏情况的用户购买保险的意愿：



(2)

为了通过箱线图分析年龄、保险年费这两个因素与购买保险倾向之间的关系，我们统计不同购买意愿的人群的年龄分布和保险年费分布。相关的绘图代码在`plot_2.py`中给出。（感觉这类问题更应该用频数直方图来分析，而不是箱线图）





(3)

划分数据集的代码在`divid.py`中，计算题中的数据所需要的代码在`cal_3.py`。

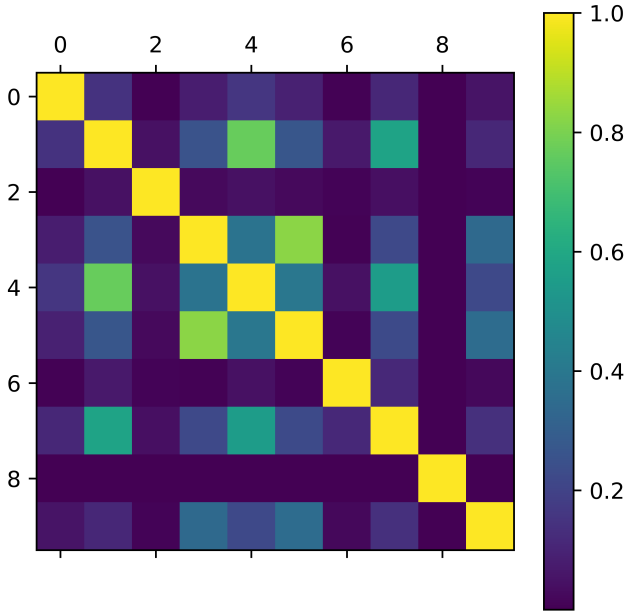
- 男女比例：
    - 训练集：1.172
    - 验证集：1.193
- 年龄的相关数据：
    - 训练集：
        - 最小值：20
        - 最大值：85
        - 平均值：38.80
        - 中位数：36.0
    - 验证集：
        - 最小值：20
        - 最大值：85
        - 平均值：38.89
        - 中位数：36.0
- 驾照比例：
    - 训练集：0.9978
    - 验证集：0.9980

- 之前购买保险的比例：
    - 训练集：0.4577
    - 验证集：0.4598
- 汽车年限的比例：
    - 训练集：
        - < 1: 0.4334
        - 1-2: 0.5247
        - > 2: 0.0419
    - 测试集：
        - < 1: 0.4294
        - 1-2: 0.5283
        - > 2: 0.0423
- 汽车曾经损坏的比例：
    - 训练集：0.5054
    - 测试集：0.5033
- 年保险费：
    - 训练集：
        - 最大值：540165.00
        - 最小值：2630.00
        - 平均值：30580.06
        - 中位数：31686.00
    - 测试集：
        - 最大值：495106.00
        - 最小值：2630.00
        - 平均值：30517.37
        - 中位数：31618.00

(4)

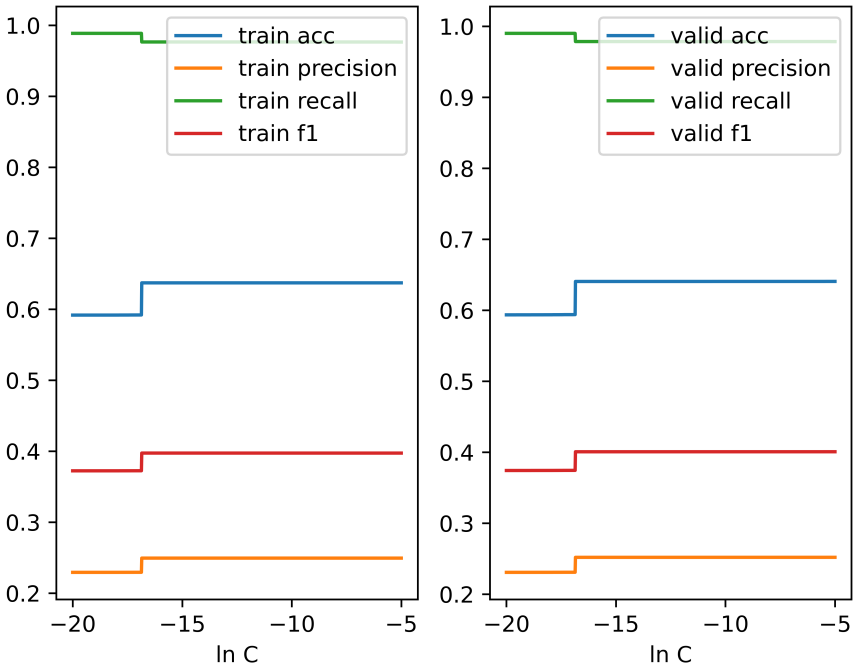
特征的选取：首先根据经验排除一些特征，由于序号与客户本身的情况无关，因此排除。再根据方差检验排除一些方差极小（或者认为这个特征在所有样本上都有着类似的表现）的特征，由之前计算，可以知道所有用户基本上都持有驾照，因此忽略这个特征。

剩下特征的选取计划通过计算Pearson相关系数来实现，相应的代码在`selection.py`中。下面给出Pearson相关系数的图：



其中从0-9代表的特征为`"Gender", "Age", "Region_Code", "Previously_Insured", "Vehicle_Age", "Vehicle_Damage", "Annual_Premium", "Policy_Sales_Channel", "Vintage", "Response"`，可以看出`Reigon_Code`，`Annual_Premium`，`Vintage`这三个特征与其他特征相关性极小，因此忽略。

为了选择出最优的正则化参数，我决定仿照第一次作业中的方法，将正则化参数在对数坐标下扫描一遍，绘制出对应在训练集和验证集上accuracy、precision、recall、f1的变化图象，从图像中选择一个表现较好的正则化参数。



从图中可以看到，recall一直比较高，为了获得更好的accuracy和precision，可以选择相应地牺牲一点点recall，选择正则化参数为$C = \exp(-15)$，可以得到在训练集和验证集上的准确率如下：

- ACCURACY_train = 0.6373
- PRECISION_train = 0.2495
- RECALL_train = 0.9766
- F1_train = 0.3975
- ACCURACY_valid = 0.6406
- PRECISION_valid = 0.2518
- RECALL_valid = 0.9786
- F1_valid = 0.4006

可以看到对于这个问题，我们主要感兴趣潜在的用户，因此只要recall比较高就行，这样我们就基本找全了潜在用户，precision比较低只是说明我们筛选出来的候选人中有意愿购买保险的比例不高。

(5)

confusion matrix of train data is:
[[147956 102859]
[818 34198]]
confusion matrix of valid data is:
[[49587 33997]
[250 11444]]
AUC of train data is 0.7833
AUC of valid data is 0.7859

其余的数据可以参见上一问。