```
519120910178
```

王辰君

期末大作业

Problem I

1.1背景与数据分析

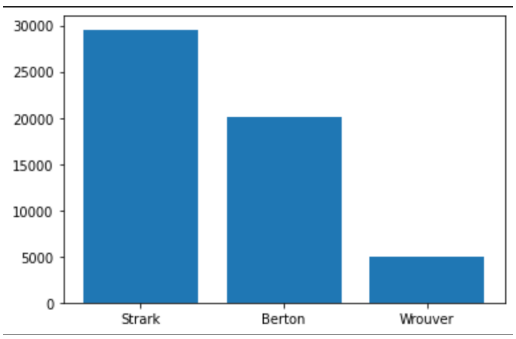
我们所拥有的信息有：

1. **司机的id**

   经观察，我们可以发现司机的id是近乎纯随机，与注册时间、第一次完成订单的时间均无关。我们可以认为公司提供的样本数据集是经过比较严格的随机抽样。id理论上已不具备任何司机的信息，并且也无法参与我们对未来的样本预测。

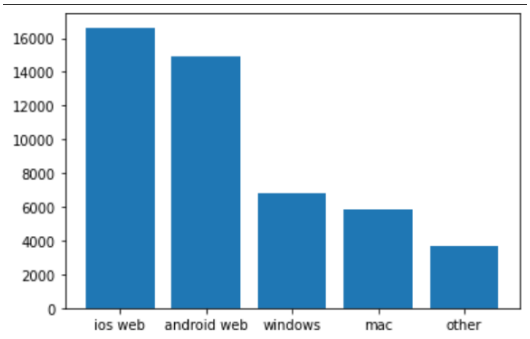
2. **司机的城市信息**

   我们可以发现，样本全部来自于三所城市。城市信息无缺失值。城市信息可能反映出不同的城市的发达程度与交通形式。



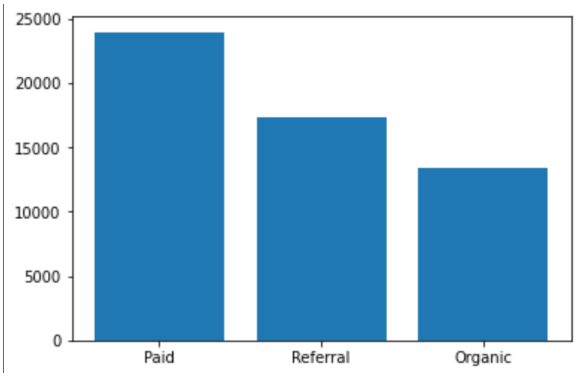
3. **司机的操作系统**

   司机的操作系统，存在一定比例的缺失值（有效值：47824/54681）。数据的分布相对均匀，若使用众数进行弥补则非常容易有失偏颇。在此处我们选择舍弃缺失值，因为缺失值所占样本的比例并不高，缺失值并不会显著缩减样本的规模。



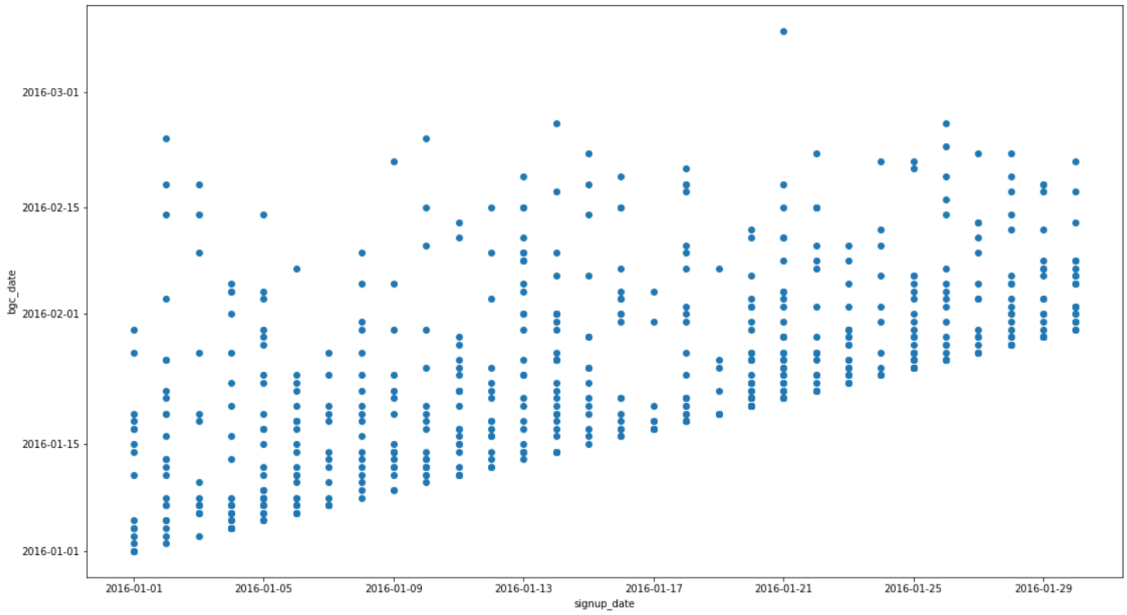
4. **司机的注册渠道**

没有任何缺失值。我们可以认为这个指标可能联系司机对于网约车的了解情况。



5. **司机的注册时间与同意背景调查时间**

此处将两者一起考虑是因为，我认为若只考虑单一的日期，找不到解释意义。我们也无法将之后的日期用于现在这些日期得到的模型的进行预测，故单一的日期我们不纳入模型。但仔细观察会发现同意背景调查的时间与注册时间是明显存在先后顺序的，即先注册再同意背景注册的。两者之差或许包含一定的信息我选择将其纳入模型。



而初步检查相关系数可发现，两者时间差与是否开始驾驶，存在着一定的相关性。

	past_days	y
past_days	1.000000	-0.308949
y	-0.308949	1.000000

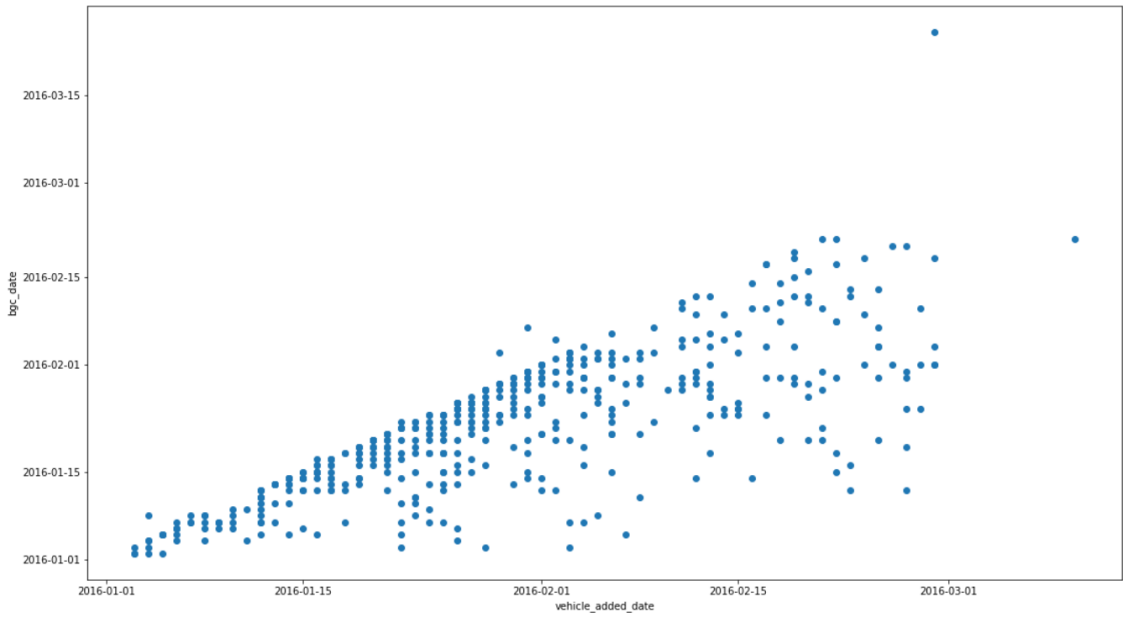
同时，我们还可以发现如果不去进行背景调查是100%没有第一次上路的，所以显然背景调查是上路的必须条件之一。我们可以将未背景调查的样本去除，以研究其他因素的影响。同时，我们还需要另外研究影响背景调查意愿的因素是哪些。

bgc_date	False	True
first_completed_date		
False	21785	26759
True	0	6137

6. **司机的同意背景调查时间与添加车辆的日期**

同样的我们可以发现与背景调查时间相类似，添加车辆的日期也存在着一条边界线，即添加车辆一定是在完成了背景调查之后，并且可以完成添加车辆的大部分司机都是完成了背景调查的(255/12879)。但这张图上我们可以发现一些异常值（即添加车辆在背景调查之前）。事实上，对于所有样本中，这样的情况也是极少数（220/54681)我们有理由认为这是采集时出现的问题，所以可以舍去。同样我们可以认为两者的日期之差也包含着一定的信息。

bgc_date	False	True
vehicle_added_date		
False	21530	20017
True	255	12879



添加车辆的日期与是否上路也存在着相当高的决定性，尽管不如背景调查如此显著。我们也可以去除未上传的车辆样本来观察其他因素的影响。

[26]:

vehicle_added_date	False	True
first_completed_date		
False	41282	7262
True	265	5872

7. **车型与品牌**

车型与品牌经统计，发现在添加车辆有值时，全部均有值。各种司机的车型与品牌首先作为哑变量均纳入模型考虑。

8. **车的年份**

同车型与品牌一样，均有值，初步直接纳入模型。

9. **变量总结与数据筛选结果**

保留的变量为司机所在的城市、司机的操作系统、司机的注册渠道、司机注册与接受背景调查的时间差、司机接受背景调查与提交载具信息的时间差、车辆的品牌与型号、车辆的年份。

数据经处理后，有效值为12659个样本，样本规模依旧很大。并且控制了同意调查与添加了载具之后，样本比起原始来说，上路与否所占比例更加均衡，便于我们更加显著地看出模型预测的效果。因为基准准确率变为54.4%。

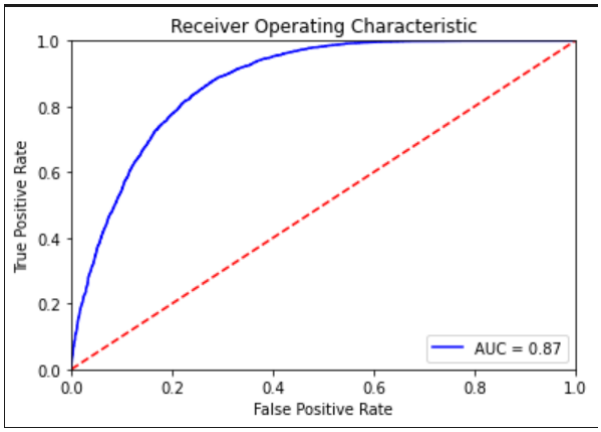
```
first_completed_date
False                  6890
True                   5769
```

1.2初步模型说明

本文中采用逻辑斯蒂回归进行是否会进行第一次上路的预测。因为逻辑斯蒂的回归比较适合进行二分类的问题。同时，本文还将继续探索影响司机是否同意调查以及是否添加载具的因素。

1.3初步模型拟合的结果

1. 以全部的样本、全部的变量进行拟合的结果



可以看到我们的AUC达到了0.87，接近0.9。可见模型的变量抓取还是处于一个比较不错的水平。

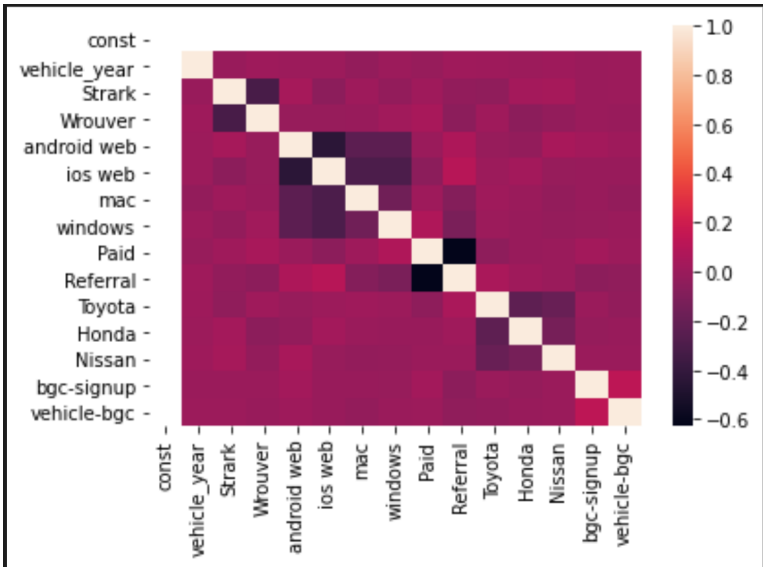
2. 但在解释模型时，我发现了新的问题便是由于车型与品牌的变量过多，而有许多的车型只有单单一个样本。我们对于如此少的样本数的回归是不具有说服力的。并且我们对于这么多的车型，提出优化的措施也是相当不现实的。本文接下来选取样本数足够的车型或品牌作为变量进行研究。

1.4变量改进

我们观察前五热门的品牌的车的数量会发现他们一共占61.8%，前三热门的车牌：Toyota,Honda,Nissan，竟占了50%的样本。这三个品牌的样本规模均达到了1000-3000多个，我认为是足够具有解释性的。我们可以探究这三个品牌的车是否具有比起其他品牌来说更加有做网约车的优势。

车型上，样本规模最大的不过是5%，样本的规模我认为相对较小。并且对于车型的指导意见可能具有时效性，因为车型的更新是相对较快的。

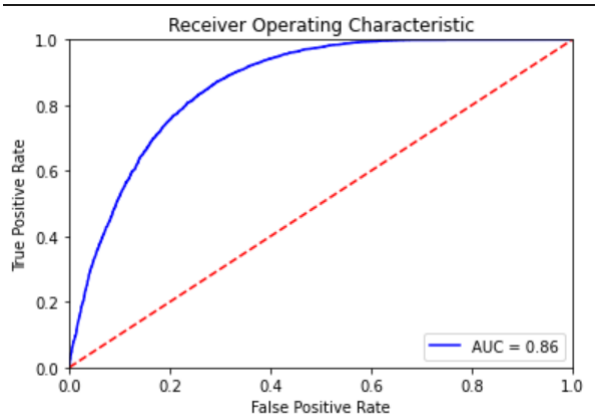
所以，我决定选取那3个品牌重点研究，并且不再考虑车型的影响。



从变量之间的相关性来看，变量之间的共线性并不明显。

1.5第二次模型拟合

1. 选择85%的样本作为训练集，15%的样本作为测试集。ROC曲线如下所示：



可以看到，我们的AUC依旧表现出色，在balanced样本中仍能达到0.86的水准。

2.

	coef	std_err	z	P>|z|	[0.025	0.975]
const	-61.7728	12.632	-4.890	0.000***	-86.532	-37.014
vehicle_year	0.0313	0.006	4.986	0.000***	0.019	0.044
Strark	-0.1117	0.054	-2.072	0.038*	-0.217	-0.006
Wrouver	-0.2137	0.098	-2.180	0.029*	-0.406	-0.022
android web	0.0776	0.091	0.854	0.393	-0.100	0.256
ios web	0.1450	0.087	1.674	0.094	-0.025	0.315
mac	0.5366	0.101	5.301	0.000***	0.338	0.735
windows	0.4672	0.102	4.579	0.000***	0.267	0.667
Paid	0.0818	0.072	1.137	0.255	-0.059	0.223
Referral	0.6575	0.066	9.959	0.000***	0.528	0.787
Toyota	0.0971	0.062	1.559	0.119	-0.025	0.219
Honda	0.0614	0.076	0.806	0.420	-0.088	0.211
Nissan	-0.0250	0.087	-0.287	0.774	-0.196	0.146
bgc-signup	-0.1609	0.005	-34.445	0.000***	-0.170	-0.152
vehicle-bgc	-0.1539	0.005	-33.648	0.000***	-0.163	-0.145

基于以上的结果，我们得到了我们的第一次结论。

1. **车辆生产的年份**

   我们可以看到他的系数是正的，并且通过了显著性检验。这意味着车辆越新，司机上路的可能性也就越大。

2. **城市**

   我们可以看到(Strark,Wrouver)两个城市的系数均为负值，并且显著性均超过了常用的0.05，我们有理由认为这两座城的司机小于Berton。

3. **操作系统**

   我们可以看到四种系统的变量均为正值，但Android和IOS没有通过显著性检验。那我们可以得出Android 和IOS，并没有显著的优势。而mac和 windows则是有相当高的显著性，说明这两个系统中注册的用户有相当的可能性的优势。

4. **注册渠道**

   注册渠道来看，被付费而来司机的显著性并不高，我们有理由认为付费与组织性地加入没有足够大的区别。然而被推荐而来的人的显著性非常高，我们有理由认为这是有优势的。

5. **车的品牌**

   我们选择的三个品牌，显著性均不高。

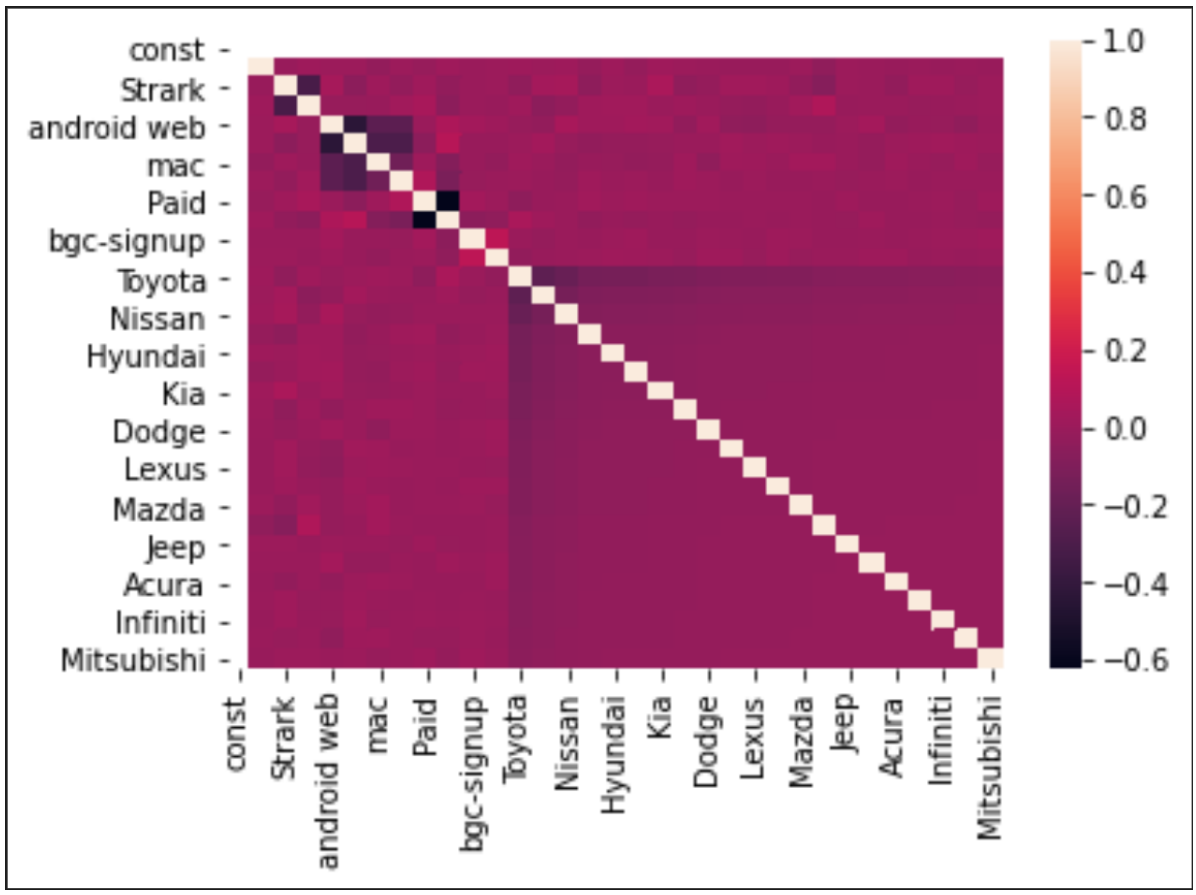
6. **时间点的间隔**

    我们可以看到这两个变量均非常显著，且都为负相关。这说明了两个日期的间隔较短能够有效地帮助我们筛选有可能上路的司机。

我们发现这个模型对于车辆的品牌并没有得到建设性的发现。并且我们的模型只是通过一个训练集和测试集来得到并衡量。下面本文将在变量中加入样本规模达到一定程度的品牌，采用10折的cross validation去选择**lasso** 回归的参数，以此试图找到一个更加符合自然的模型，并且可能找到在我们这一问考虑的前三个品牌以外的可能存在的品牌的影响。

1.6变量再改进

我们选择加入样本规模在100以上的品牌，那样，我们的变量个数增加了18个新的品牌。这样的话我们可能从这18个品牌中，找到某些品牌可能的正面作用。



可以看到我们新添加的变量的共线性没有，所以不影响我们的回归。

1.7Cross Validation Lasso逻辑回归

我们选择使用ROC_AUC作为我们选择的标准。

	coef	feature
0	0.780792	vehicle_year***
1	-0.056275	Strark*
2	-0.045516	Wrouver*
3	0.008734	android web
4	0.040615	ios web*
5	0.157770	mac***
6	0.136721	windows***
7	0.024165	Paid
8	0.297749	Referral***
9	-1.506924	bgc-signup***
10	-1.403595	vehicle-bgc***
11	0.023506	Toyota
12	0.000000	Honda
13	-0.020434	Nissan
14	0.000000	Ford
15	0.000000	Hyundai
16	-0.032806	Chevrolet
17	-0.004582	Kia
18	-0.040930	Volkswagen*
19	-0.029813	Dodge
20	0.006433	Mercedes-Benz
21	0.000000	Lexus
22	0.000000	BMW
23	-0.016445	Mazda
24	-0.037257	Subaru
25	-0.011820	Jeep
26	-0.031294	Chrysler
27	0.041766	Acura*
28	0.013137	Scion
29	-0.025593	Infiniti

	coef	feature
30	0.000000	Audi

回归之后的结果呈现了与我们第二次回归相近的结果。由于程序的输出原因，我们无法画出如之前两个模型的光滑ROC曲线，我们转而选择TPR，来衡量我们的模型。TPR对于我们的问题具有较好的解释性，即我们把可能会上路的司机成功挑选出来的比例。TPR事实上达到了**82%**。我们可以认为具有较好的解释力。

观察模型的各个系数我们会发现其实Lasso得到的结果的显著性方面与我们之前的模型非常相似。也再次印证了我们先前的结论对于抽样的依赖度不高。

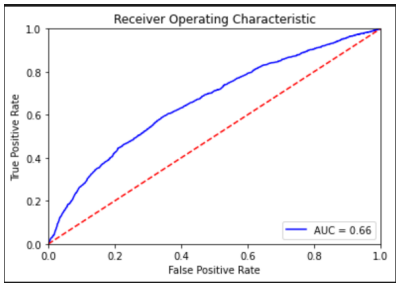
关于品牌方面，我们可以从系数中大致发现，Acura 对比其他的品牌具有一定的显著性为更加优秀，Volkswagen对比其他的品牌具有一定的显著性更加弱势。

1.8对于是否添加载具的回归分析

既然司机上路与否，与他是否添加载具有着极为密切的关系。那我们接下来将研究司机添加载具与否，与其余变量的关系。

我们选择了同意了背景调查后的所有样本进行回归。有效的变量有司机的城市信息、司机的操作系统、司机的推荐渠道、司机同意背景调查与注册的时间差。

同样的我们使用85%作为训练集，15%作为测试集。



可以看到我们的模型在测试集上的效果，并不优秀，ROC只有0.66。那么我们可能需要对模型进行CVlasso回归来提高效果。

经过几次尝试，我们发现如果我们无论用ROC或者accuracy作为选择标准，最后的效果均不好（ROC始终处于0.66之下）。并且TPR的值都相当低，不利于我们发现那些潜在的客户。

所以我们选择直接用TPR也就是Recall进行选择。

最终，我们的模型的TPR可以达到0.5以上。

	COEF	features
0	0.000000	Strark
1	0.000000	Wrouver
2	0.000000	android web
3	0.000000	ios web
4	0.000000	mac
5	0.000000	windows
6	0.000000	Paid
7	0.105918	Referral
8	0.000000	bgc-signup

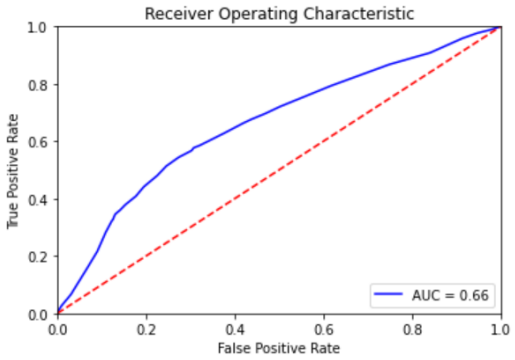
而我们最终选择的变量只有一个Referral。

1.9对于是否同意背景调查的回归分析

同样的司机上路与否，与他是否同意背景调查有着极为密切的关系。那我们接下来将研究司机影响同意背景调查意愿的因素。

我们选择了城市、操作系统、推荐渠道变量均有值的样本进行回归。有效的变量有司机的城市信息、司机的操作系统、司机的推荐渠道。

同样的我们使用85%作为训练集，15%作为测试集。



可见AUC也并不好。同样的，我们使用CVLasso对AUC进行回归。 recall 能够达到91.6%，意味着我们的这个模型能够有效找出可能同意背景调查的人。

	Coef	Features
0	0.036242	Strark
1	-0.085314	Wrouver
2	-0.112912	android web
3	-0.095898	ios web
4	-0.004749	mac
5	-0.010209	windows
6	-0.347040	Paid
7	0.389417	Referral

从上表中，我们可以看出Referral 依旧十分显著好，paid显著不如Oragnic。另外就是城市方面，wrouver 显著 小于Berton。Stark 在此处与Berton并无太大差别。操作系统中，我们可以看到IOS 与 Android 显著较差。

1.10结论分析

我们从以上的对三个指标的回归（两种日期均完整情况下影响用户的正式上路的转化率、同意背景审核后到上传载具信息的转化率以及注册到同意背景审核的转化率），得到了对于三个阶段的优化建议。

1. 从注册到同意背景调查的转化率，我们首先发现wrouver低于另两座城市，那我们可以选择在鼓励新用户注册时，偏重于另两座城市。在操作系统方面，我们发现两个移动端是显著不如其他的系统。那我们在投放广告时，可以将经费多用于电脑端的重点推广。还有就是推荐的注册渠道是效果相当好的，然后付费邀请注册的转换率很低。或许我们可以减少无谓的花钱去让司机注册。我们或许可以鼓励老司机推荐新司机。
2. 从同意背景调查后到上传载具信息的转化率，我们发现显著的只有推荐的注册渠道。这也可以说明推荐的注册渠道的作用十分大，我们或许可以对于这一类人提供一个对上传载具的奖励，让这部分潜在司机转化更好。
3. 从上传完载具信息后到正式上路的转化率中，我们可以关注前两个日期差较短的积极用户，给予他们第一次上路的激励。并且可以在同意背景调查之后，增加一个限时上传载具信息的额外奖励。