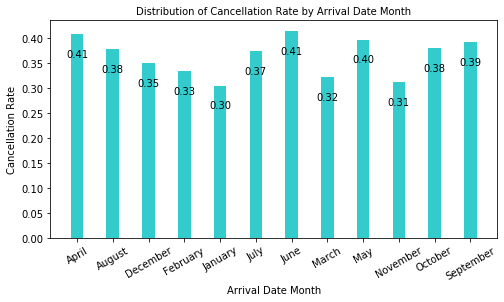
1. 先踢掉所有没有明确条目说明的项目；日期（可以查看seasonality，但是在prediction中不考虑）

Check Seasonality of Month:



基本看不出有seasonality，这个variable可以踢掉了

去掉的variables：

'ArrivalDateYear', 'ArrivalDateMonth', 'ArrivalDateWeekNumber', 'ArrivalDateDayOfMonth',

“ Country”, 'Agent','Company', 指代不明

'ReservationStatus', 'ReservationStatusDate' 这两个都在cancel之后发生，无法用于预测，删除

增加variables：

‘Type’: Resort – 1, City – 0

PreCancelRate = PreviousCancellations/(df['PreviousCancellations']+df['PreviousBookingsNotCanceled']+1)

BookAssignDiff: 分配的房间和预约的房间不一样 – 1， 分配的房间和预约房间一样 - 0

1. One-hot 处理 Categorical data

'Meal' : 'BB ', 'FB ', 'HB ', 'SC ', 'Undefined'

'MarketSegment': 'Direct', 'Corporate', 'Online TA', 'Offline TA/TO', 'Complementary', 'Groups', 'Undefined', 'Aviation'

'DistributionChannel': 'Direct', 'Corporate', 'TA/TO', 'Undefined', 'GDS'

'DepositType': 'No Deposit ', 'Refundable ', 'Non Refund '

'CustomerType’: 'Transient', 'Contract', 'Transient-Party', 'Group'

删掉原来的列和base列：

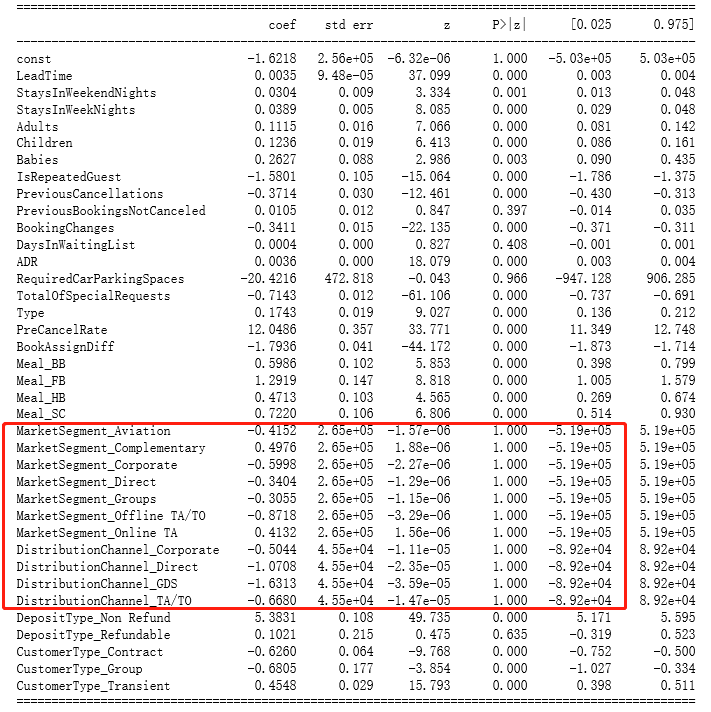
base = ['Meal\_Undefined',

'MarketSegment\_Undefined',

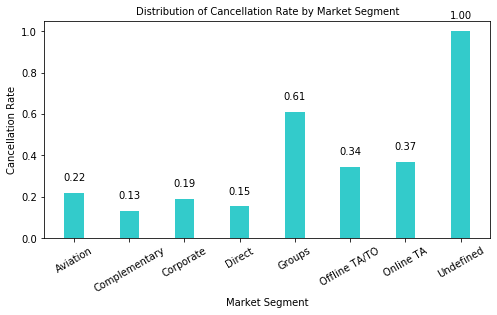
'DistributionChannel\_Undefined',

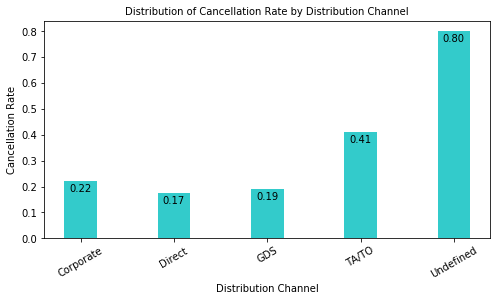
'DepositType\_No Deposit ',

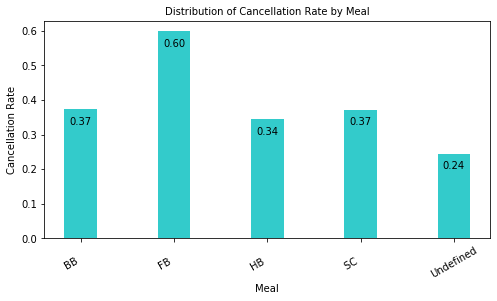
'CustomerType\_Transient-Party']



结果很奇怪，把这两个列取出来，check一下







Undefined占的数量比较少，但是很影响结果，所以把Undefined observation去掉，并且重新设置base

新的base

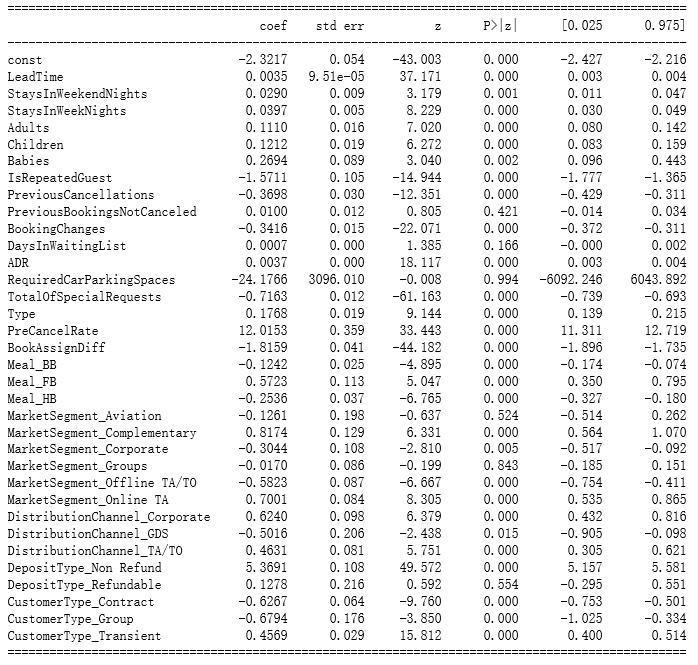
base = ['Meal\_SC ',

'MarketSegment\_Direct',

'DistributionChannel\_Direct',

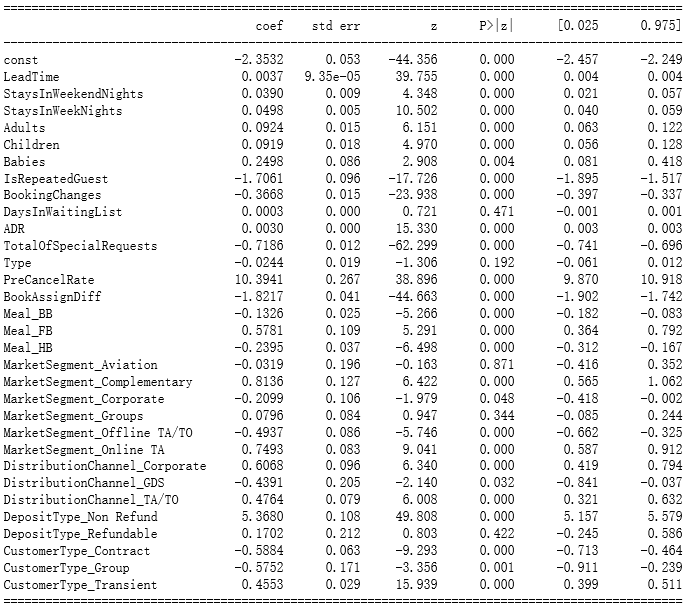
'DepositType\_No Deposit ',

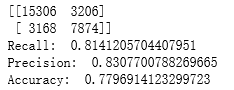
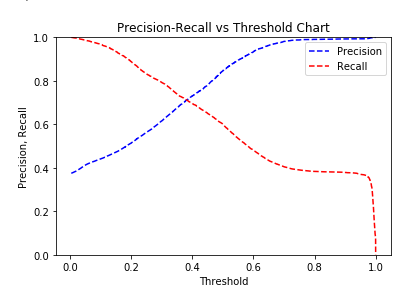
'CustomerType\_Transient-Party']



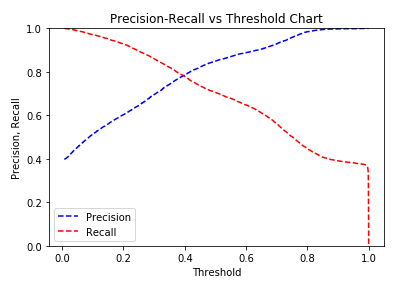
因为PreCancelRate是用PreviousCancellations , PreviousBookingsNotCanceled 算出来的，所以把这两项删掉。

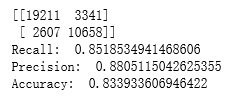
RequiredCarParkingSpaces不significant，所以删掉。

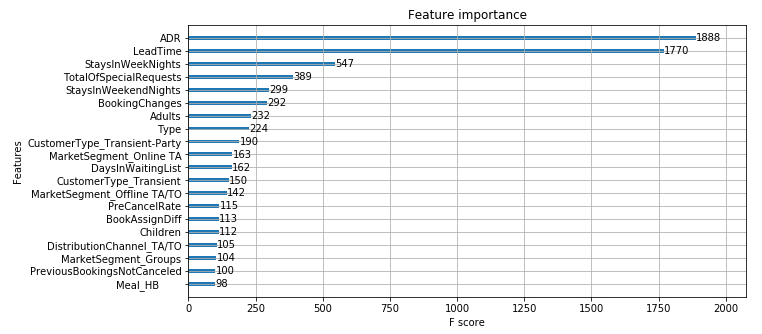




1. 用XGboost试一下，XGBoost的好处在于可以处理missing value，自动crossvalidation



结果比Logistic更好



**这个地方需要Design一个cost function，因为要考虑“把不会cancel的人误认为会cancel”和“把会cancel的人误认为不会cancel”哪个损失更大一点。（overbooking损失更大还是空房的损失更大）**