使用定性变量的代码实现

statsmodels

小胖

目录

初探数据

数据简介

数据来自美国加州大学欧文分校

美国个人收入的普查数据

预测变量是年收入分类

数据的变量说明

变量名	变量类型	说明
age	数值型变量	年龄
workclass	类别型变量	工作类型，如公务员、私企职工等
fnlwgt	数值型变量	抽样权重。（普查时使用的变量，与建模分析无关）
education	类别型变量	学历，如本科、研究生等
education_num	数值型变量	受教育年限
martial-status	类别型变量	婚姻状况
occupation	类别型变量	所在行业
relationship	类别型变量	家庭角色，比如丈夫、妻子等
race	类别型变量	种族
sex	类别型变量	性别
capital_gain	数值型变量	该年度投资收益
capital_loss	数值型变量	该年度投资损失
hours_per_week	数值型变量	每星期工作时间
native_country	类别型变量	出生国家
label	类别型变量	年收入分类，分为两类：“>50K”和“<=50K”

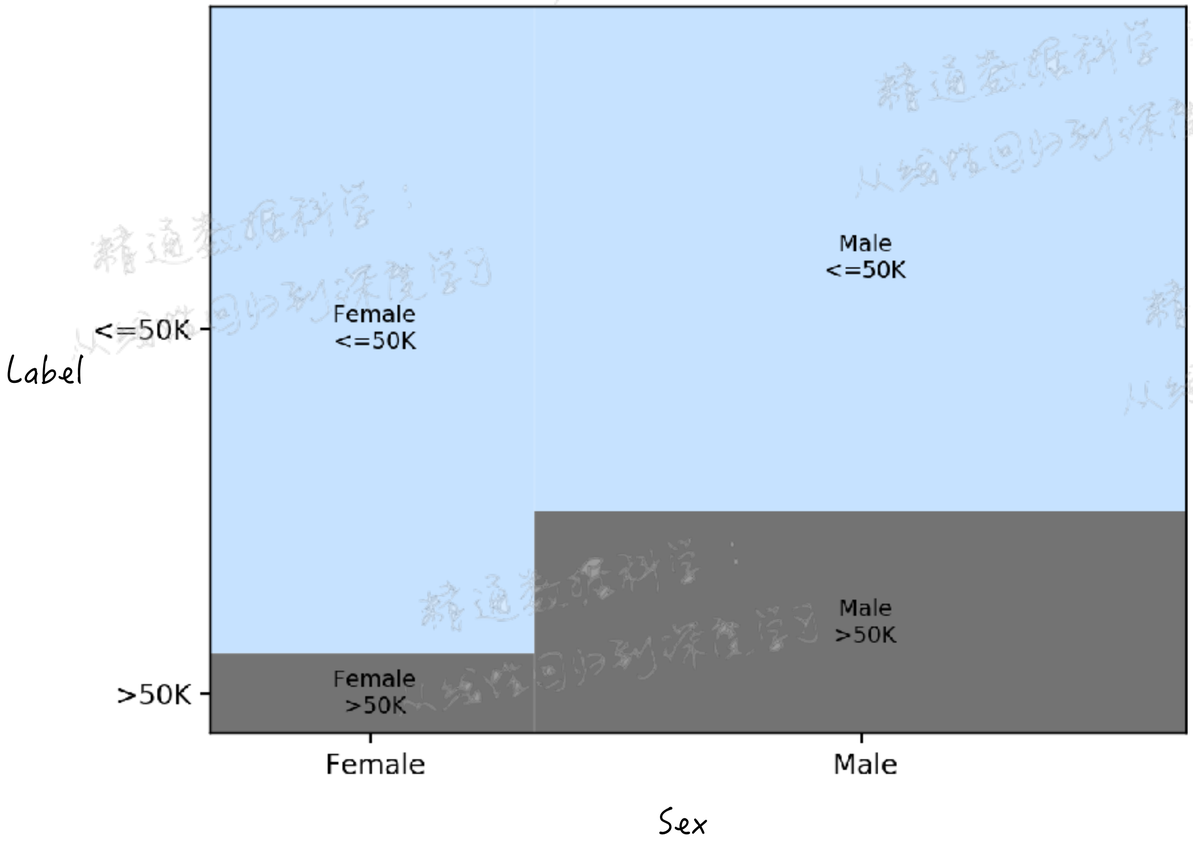
类别变量（定性变量）

预测变量

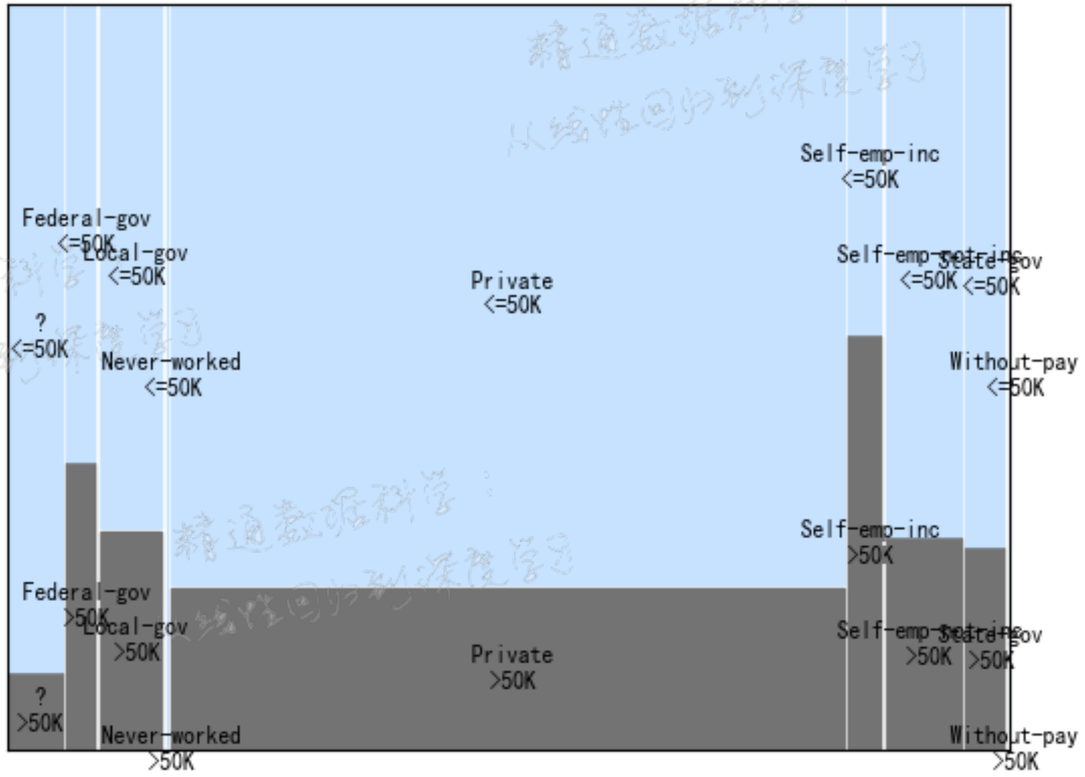
初探数据

这里是副标题文字

性别和年收入分类的交叉报表



工作类型和年收入分类的交叉报表



目录

搭建模型

虚拟变量

将数据分为训练集和测试集

使用statsmodels，转换类别变量

训练模型，得到模型结果

分析模型参数的显著性、假设检验

sex=Male的系数大于0

变量系数不显著

Logit Regression Results

Dep. Variable:	label_code	No. Observations:	6512
Model:	Logit	Df Residuals:	6498
Method:	MLE	Df Model:	13
Date:	Fri, 31 May 2019	Pseudo R-squ.:	0.2732
Time:	12:30:09	Log-Likelihood:	-2611.5
converged:	False	LL-Null:	-3593.4
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.8382	0.305	-25.734	0.000	-8.435	-7.241
C(sex)[T. Male]	1.2566	0.090	13.959	0.000	1.080	1.433
C(workclass)[T. Federal-gov]	1.1871	0.280	4.238	0.000	0.638	1.736
C(workclass)[T. Local-gov]	0.9936	0.255	3.904	0.000	0.495	1.492
C(workclass)[T. Never-worked]	-24.1726	8.42e+05	-2.87e-05	1.000	-1.65e+06	1.65e+06
C(workclass)[T. Private]	0.7343	0.226	3.253	0.001	0.292	1.177
C(workclass)[T. Self-emp-inc]	1.6271	0.280	5.809	0.000	1.078	2.176
C(workclass)[T. Self-emp-not-inc]	0.7105	0.254	2.798	0.005	0.213	1.208
C(workclass)[T. State-gov]	0.7586	0.278	2.732	0.006	0.214	1.303
C(workclass)[T. Without-pay]	-9.1249	714.596	-0.013	0.990	-1409.708	1391.458
education_num	0.3361	0.016	21.347	0.000	0.305	0.367
capital_gain	0.0003	2.13e-05	14.196	0.000	0.000	0.000
capital_loss	0.0009	7.55e-05	12.084	0.000	0.001	0.001
hours_per_week	0.0265	0.003	8.836	0.000	0.021	0.032

搭建模型

剔除不显著类别

将数据分为训练集和测试集

使用statsmodels，剔除不显著类别

训练模型，得到模型结果

分析模型参数的显著性、假设检验

显著的类别

Logit Regression Results

Dep. Variable:	label_code	No. Observations:	6512
Model:	Logit	Df Residuals:	6500
Method:	MLE	Df Model:	11
Date:	Fri, 31 May 2019	Pseudo R-squ.:	0.2732
Time:	12:36:09	Log-Likelihood:	-2611.6
converged:	True	LL-Null:	-3593.4
		LLR p-value:	0.000

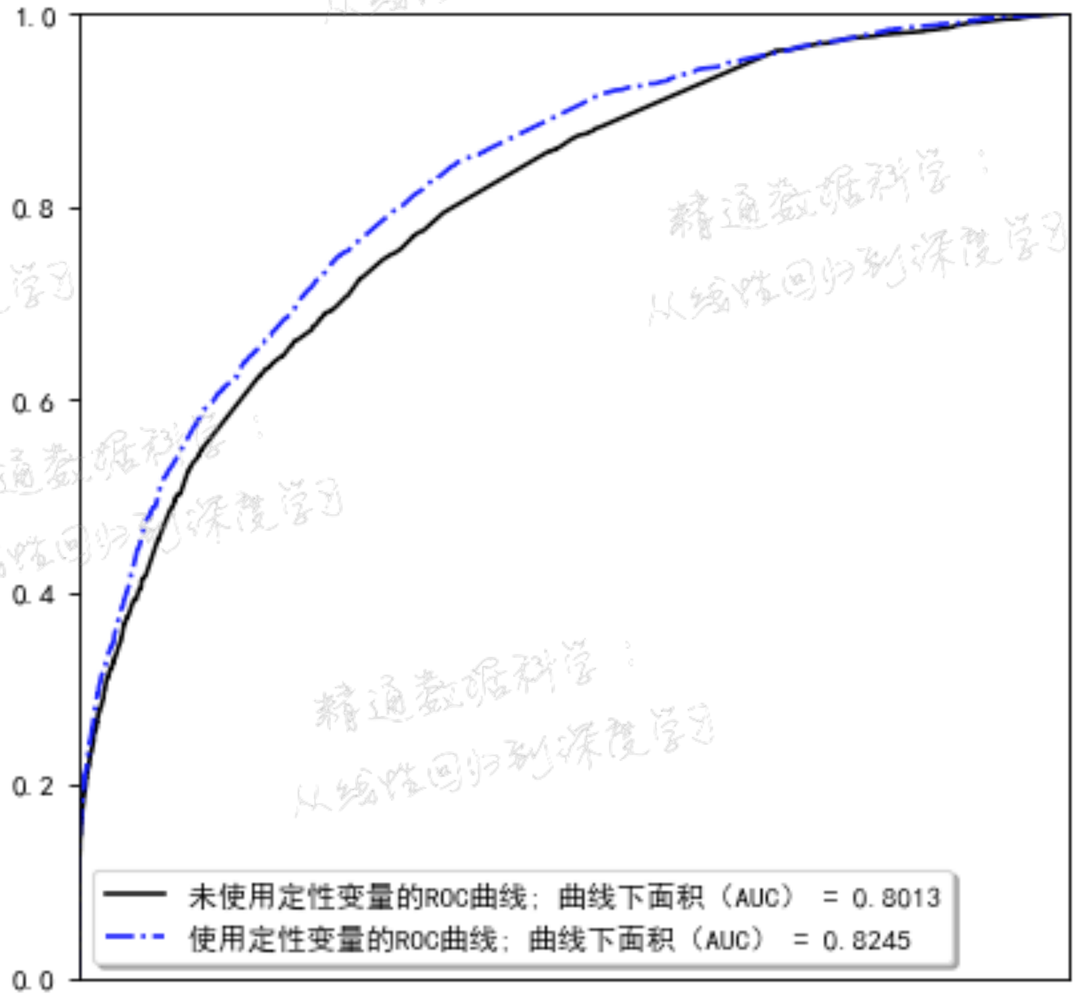
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.8418	0.304	-25.757	0.000	-8.438	-7.245
C(workclass, contrast_mat, levels=l) State-gov	0.7614	0.278	2.743	0.006	0.217	1.306
C(workclass, contrast_mat, levels=l) Self-emp-not-inc	0.7133	0.254	2.810	0.005	0.216	1.211
C(workclass, contrast_mat, levels=l) Private	0.7371	0.226	3.267	0.001	0.295	1.179
C(workclass, contrast_mat, levels=l) Federal-gov	1.1899	0.280	4.249	0.000	0.641	1.739
C(workclass, contrast_mat, levels=l) Local-gov	0.9964	0.254	3.916	0.000	0.498	1.495
C(workclass, contrast_mat, levels=l) Self-emp-inc	1.6298	0.280	5.820	0.000	1.081	2.179
C(sex)[T. Male]	1.2566	0.090	13.959	0.000	1.080	1.433
education_num	0.3361	0.016	21.347	0.000	0.305	0.367
capital_gain	0.0003	2.13e-05	14.196	0.000	0.000	0.000
capital_loss	0.0009	7.55e-05	12.085	0.000	0.001	0.001
hours_per_week	0.0265	0.003	8.841	0.000	0.021	0.032

目录

比较模型结果

AUC

加入定性变量之后，模型的效果显著提升



THANK YOU