实验 4-2 基于模型辅助海关决策生成监测报表

建议课时：20 分钟

一、 实验目的

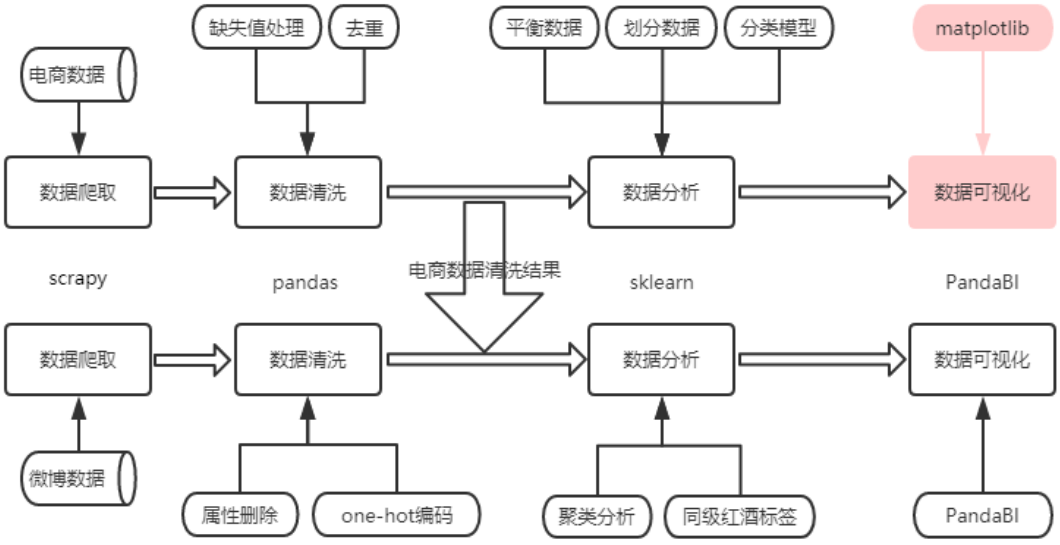
- 了解数据预测的处理流程
- 了解大屏展示与业务相结合的方法

二、 实验环境

Python3 开发环境，第三方包有 pandas，sklearn，xgboost，pickle

三、 实验步骤

本节处理的内容有：



1. 基于模型预测

本节讲述的是如何拿现有模型对单条数据进行预测，有此基础就可以对海关的红酒数据进行价格区间预测，跟报关单上的价格做对比审核。

- 保存上一节的模型

```
pickle.dump(model, open("pima.pickle.dat", "wb"))
```

- 读取模型，并对数据进行预测

注意此处需要对数据进行转换，利用上一个小节保存下来的 encoders.dict 文件对数据做相同的转换。

```python
# 加载模型
import xgboost as xgb
# bst = xgb.Booster({'nthread':4}) #init model
# bst.load_model("xgb1.model") # load data
bst = pickle.load(open("pima.pickle.dat", "rb"))

# 载入测试数据
test_data = df.iloc[[0]]
test = test_data[test_data.columns.difference(['price'])]
Y = test_data['price']
print("the true result: ",Y[0])

# 转换数据
coders = {}
with open("encoders.dict", "rb") as f:
    coders = pickle.load(f)
aa = LabelEncoder()
for x,y in coders.items():
    aa.classes_ = y
    test[x] = aa.transform(test[x])
print(test)

# 预测
p = bst.predict(test[choose])
print(p)
```

结果如下：

```
the true result:  0-50
/usr/local/python3/lib/python3.6/site-packages/ipykernel_launcher.py:20: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-
   alcohol  keyword   year  产区  仙粉黛（Zinfandel）  佳美娜（Carmenere）  佳美（Gamay）  其它  \
0     12.0      189  99999   2              0              0          0   1

   内比奥罗（Nebbiolo）  冰酒/贵腐/甜酒       ...         蛇龙珠（Cabernet Gernischt）  \
0              0        0       ...                                0

   西拉/设拉子（Syrah/Shiraz）  赤霞珠（Cabernet Sauvignon）  起泡酒/香槟  \
0                     0                         0       0

   长相思（Sauvignon Blanc）  雷司令（Riesling）  霞多丽（Chardonnay）  颜色  马尔贝克（Malbec）  \
0                     0             0               0   0            0

   黑皮诺（Pinot Noir）
0               0

[1 rows x 33 columns]
the prediction of test:  ['0-50']
```

2. 生成监测报表

我们得到了如何预测红酒价格区间的方法，本节讲述的是如何展示结果预测之后的效果，不仅仅是从数据的维度，从业务角度出发提供更多的展现效果和形态。

此处介绍一个概念：风险分析——结合业务将“风险”两字体现出来，因此设立了以下几个风险评估的维度。

将通过算法模型找出的有问题的报关单以矩阵的形式展现出来，总体将风险分为三个等级：

- 蓝色：风险等级较低，可以暂时观察；
- 黄色：风险等级中等，建议采取行动；
- 红色：风险等级较高，建议立即采取行动。

风险等级定义方法：

- 风险频率等级：

按企业的报关单数量统计，计算方法为统计该企业在一定时间段内有风险的报关单数除以该企业在这个时间段内所有的报关单数；

用 A、B、C、D、E 来表示等级，每个等级的区间为 20%，即：

A：该企业在这个时间段内有 20%或以下的报关单存在风险；

E：该企业在这个时间段内有 80%或以上的报关单存在风险

- 风险后果等级：

用数字I-V表示等级，I为最轻，V为最严重；

- 申报不规范类型风险：

以缺少关键申报要素数量为划分标准：

如缺少一项关键申报信息，风险等级为II；缺少两项为III；缺少 3 项为IV；

- 低报价风险类型风险：

以该报关单申报单价与价格区间的最低值范围差为标准划分：

如I为申报单价与最低值的差距在 5%以内；

V为申报单价与最低值的差距在 20%以上；

项目上的展示效果如下：


海关税管风险研判平台——风险分析
按 F11 即可退出全屏模式
admin
一级行业：食品
二级行业：葡萄酒
开始时间：2017-11
结束时间：2017-12
风险类型：低价风险
查询
未知风险矩阵
已知风险分布（涉案报关单）
E 2 4 4 5 149
D 8 12 8 14 130
C 11 13 9 13 166
B 21 20 29 20 351
A 64 66 63 57 684
I II III IV V
报关单号 企业名称 商品税号 商品名称 风险说明 风险详情 申报货值(元) 是否被查验 是否被估
名字 税收(百万) 排名 占比
外高桥关 46.236 1 19.754%
新港海关 18.724 2 8.000%
洋山海关（港区） 12.654 3 5.406%
北仑海关 12.092 4 5.166%
烟台海关 11.558 5 4.938%
青岛前湾保税港区 9.482 6 4.051%
内港湾心 7.724 7 3.300%
天津东疆保税港区海... 7.615 8 3.253%
泰贸海关 7.061 9 3.017%
外港海关 6.653 10 2.842%
厦门海沧保税港区港... 6.110 11 2.610%
黄埔新港海关 6.060 12 2.589%
上海西北物流园区 5.385 13 2.301%
大鹏海关 5.122 14 2.189%
深圳盐田综合保税区 4.431 15 1.893%
洋山海关（保税） 4.386 16 1.874%
连大码湾 4.249 17 1.815%
文锦渡关 3.394 18 1.450%