电力窃漏电用户自动识别实验报告

## 背景、

略

## 挖掘目标：

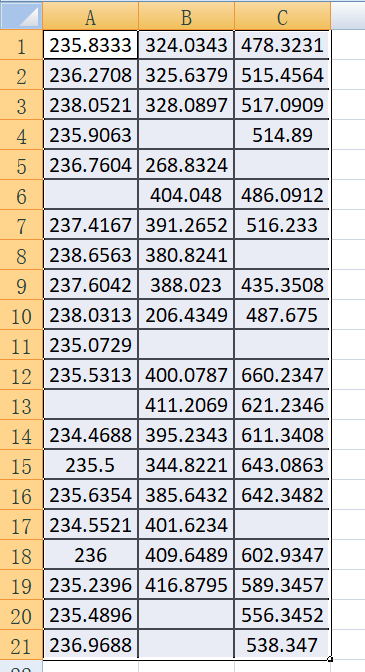
1. 归纳出窃漏电用户的关键特征，构建窃漏电用户的识别模型
2. 利用实时检测数据，调用窃漏电用户识别模型实现实时诊断

## 分析方法与过程

1. 需要剔除非居民类别的数据，如银行、税务、学校和工商等用户
2. 用电负荷不能直接体现出用户的窃漏电行为，终端报警存在很多误报和漏报的情况，故需要进行数据探索和预处理，总结窃漏电用户的行为规律，再从数据中提炼出描述窃漏电用户的特征指标
3. 窃漏电用户识别流程主要包括一下步骤：
   1. 从电力计量自动化系统、营销系统有选择性地抽取部分大用户用电负荷、终端报警及违约窃电出发信息等原始数据
   2. 队样本数据探索分析，剔除不可能存在窃漏电行为的行业，即白名单用户，初步审视正常用户和窃漏电用户的用电特征。
   3. 对样本进行预处理，包括数据清洗、缺失值处理和数据变换
   4. 构建专家样本集
   5. 构建窃漏电用户识别模型
   6. 在线检测用户用电负荷及终端报警，调用模型实现实时诊断。

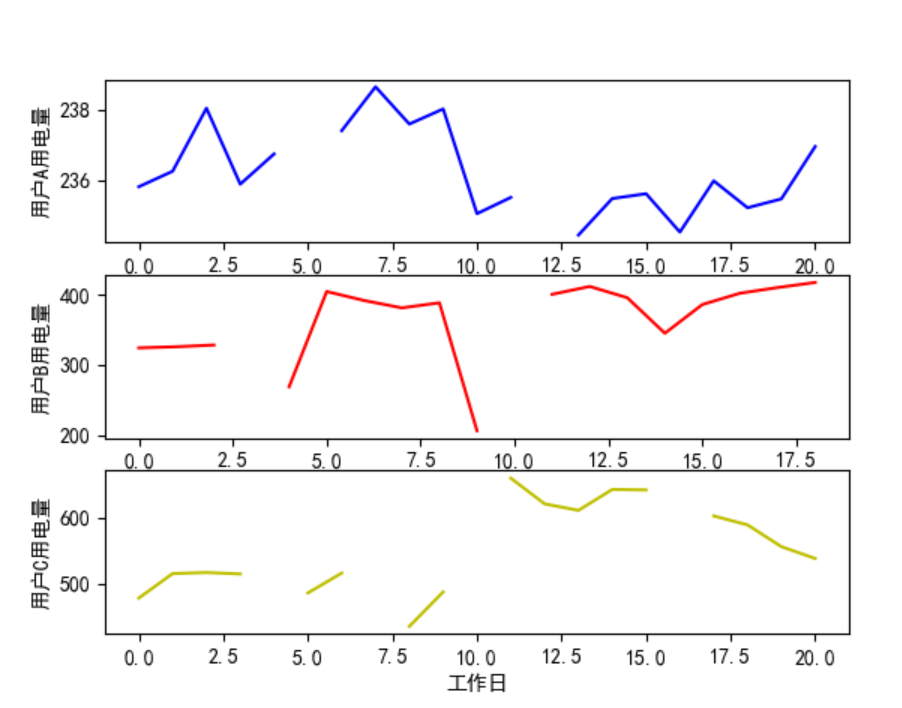
## 数据抽取

本实验中用到的数据如下



为用户A，用户B，用户C，一个月工作日用电量的数据

折线图如下所示：

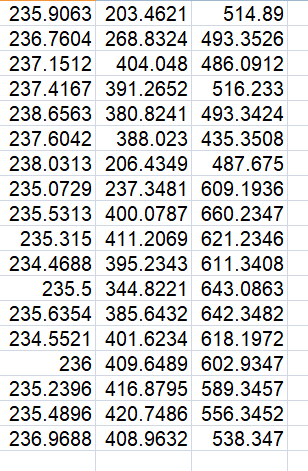


## 数据预处理

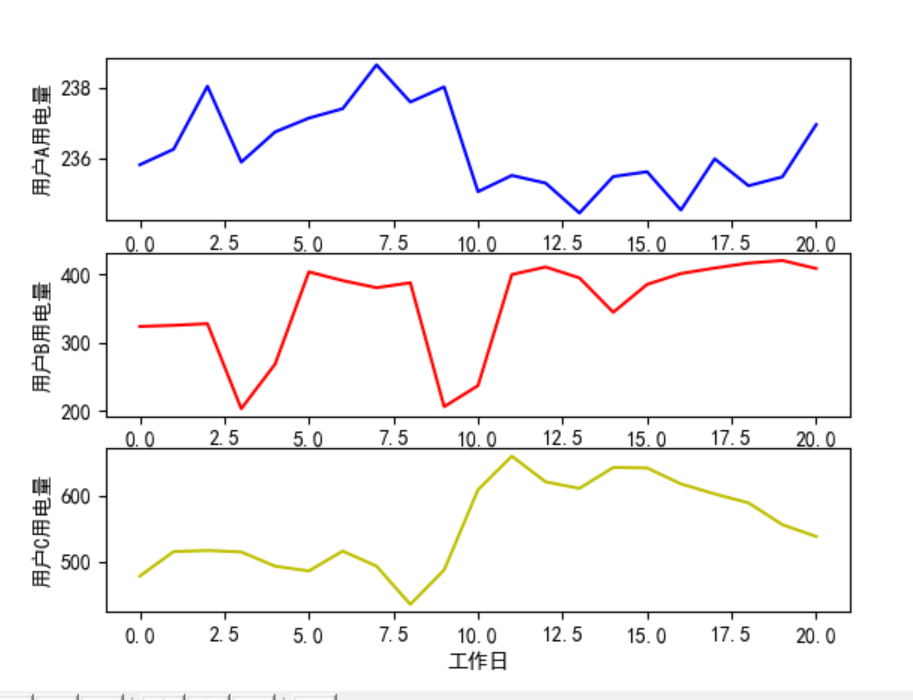
1. 数据清洗
   1. 通过分析过滤掉不可能存在窃漏电的现象的类别
   2. 剔除一些特殊时期的数据，以保证数据的普遍性，如本例中尽量过滤掉节假日的用电数据
2. 缺失值处理

用拉格朗日插值法补值

差值之后的数据如下：



折线图如下：



1. 数据变换

通过电力计算系统采集的电量、负荷，虽然在一定程度上能反映用户窃漏电行为的某些规律，但是要作为构建模型的专家样本，特征不够明显，需要进行重新构造。

* 1. 电量趋势下降指标

利用电量做直线拟合得到的斜率作为衡量，如果斜率随时间不断下降，那该用户的窃漏电可能性就很大。

* 1. 线损指标

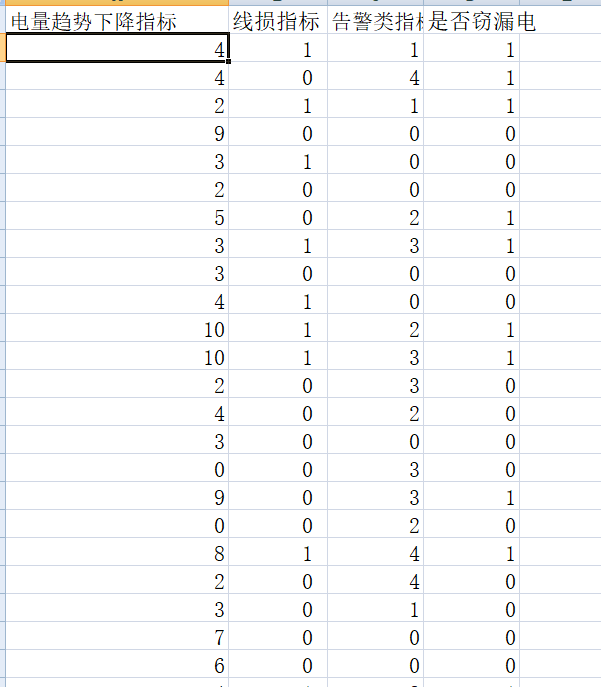
线损率是用于衡量供电线路的损失比例，同时可结合线户拓扑关系，计算出用户所属线路在当天的线损率，单纯以当天的线损率上升了作为窃漏电特征则误差过大，所以考虑前后几天线损率平均值，判断其增长率是否大于1%，若线损率的增长率大于1%则具有窃漏电的可能性。

* 1. 告警类指标

计算终端发生与窃漏电相关的终端报警的次数总和。

## 构建专家样本

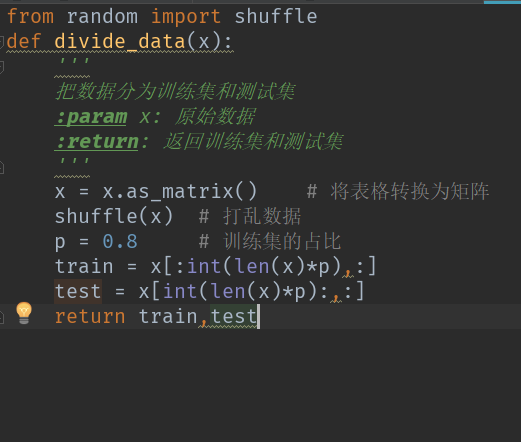
专家样本部分数据如下所示：



其中，电量下降趋势指标越大，代表下降趋势越显著；线损指标取值为0、1，当取值为0是，平均线损率的增长率不大于1%，当取值为1是，平均线损率的增长率大于1%；告警类指标为与窃漏电有关的终端报警次数；是否窃漏电的取值为0、1，为0是没有窃漏电行为，为1则是由窃漏电行为。

## 模型构建

1. 划分训练集和测试集，前80%为训练集，后20%为测试集。划分代码如下：



代码详见：电力窃漏电用户自动识别/divide.py

1. 窃漏电用户识别可通过构建分类预测模型来实现，比较常用的分类预测模型由LM神经网络和CART决策树：
   1. LM神经网络

因为专家样本中由三个指标，有一个分类结果，所以LM神经网络的输入结点由3个，输出结点由一个，隐藏层结点数为10个，优化方法用adma方法。对于激活函数，在隐藏层使用Relu(x) = max(x,0)作为激活函数，输出层使用sigmoid激活函数。

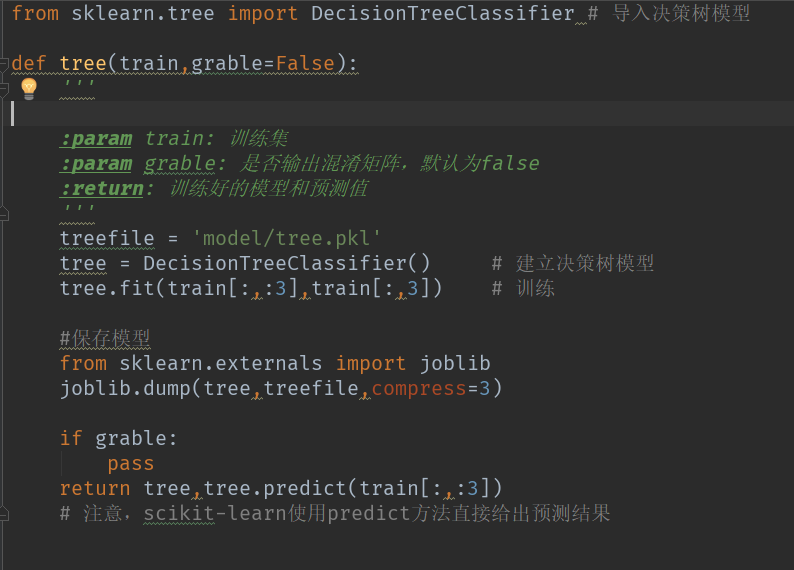
LM神经网络训练代码如下：



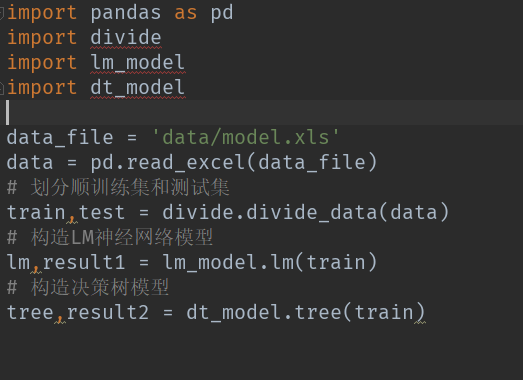
代码详细见lm\_model.py

* 1. CART决策树:

代码如下：



LM神经网络和CART决策树的训练代码如下:

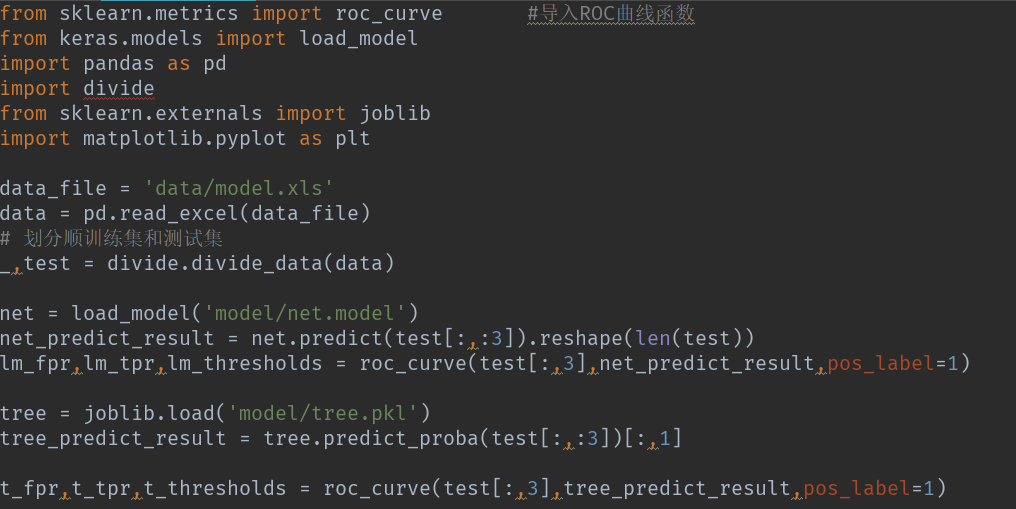


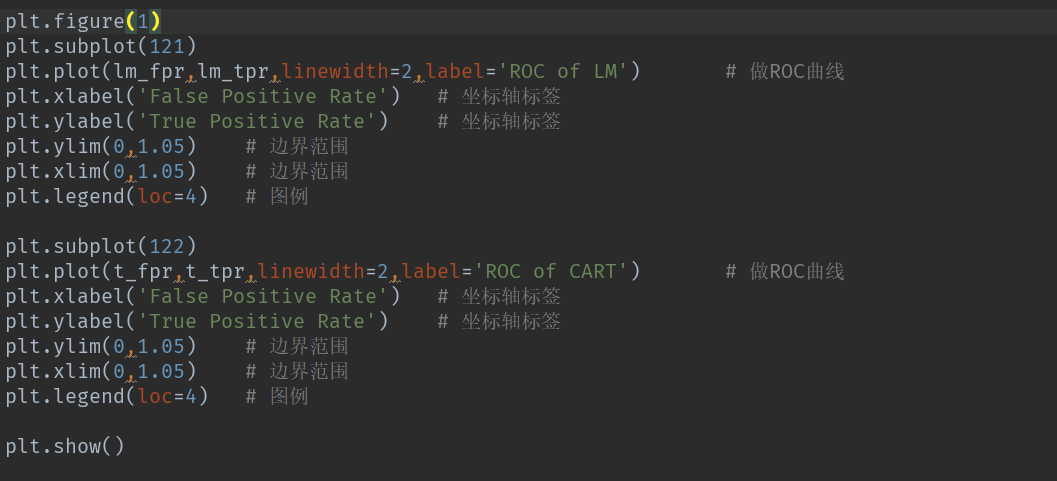
代码详细见train\_model.py

训练过后的模板保存在model文件夹中，net.model时LM神经网络的模板，tree.pkl时CART决策树的模板。

1. 模型评价

用ROC曲线来评价模型，曲线中与x轴围城的面积大的效果较好，代码如下：





代码详细见predict.py

ROC曲线图像如下：

