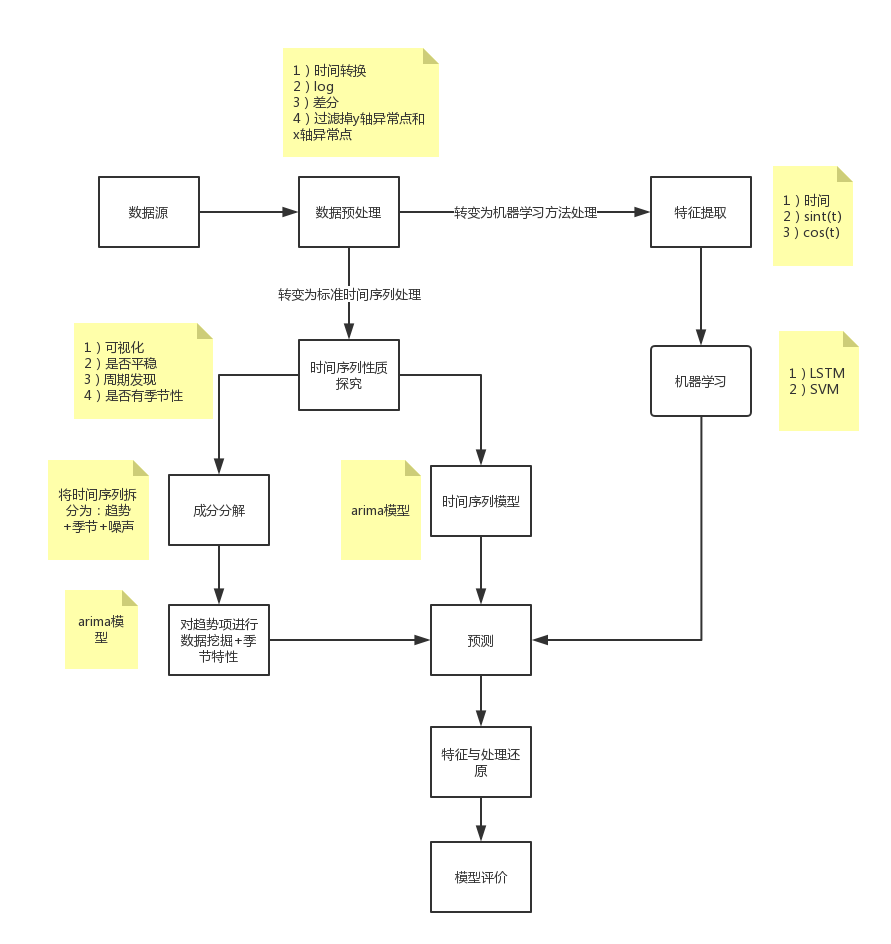
时序数据挖掘

# 时序数据挖掘示意图



# 流程注意事项

## 数据预处理

### 时间字符串转时间戳 + 排序

SQL算子就可以做到，如果是毫秒时间则需要写一个udf

### 时间分箱 + SQL过滤频数过低的箱子

### 过滤极端值

根据分位数，过滤掉前后若干百分比的数据。

### 数量数据预处理

功能简介：将输入数据进行预处理并输出

输入：DataFrame，假定时间列已经被转为了毫秒级的时间戳，且已经过滤了极端值

功能点：

1）log变换

2）diff

3）模型需要有变换和反向变换两种操作

## 基于时间序列的数据挖掘方法

arima

## 模型评价

### 模型评价标准

模型评价要数据拆分，评价标准和实际应用场景有关系。

有两种方式：

1）一种场景是找规律，一次性找到大致的规律，之后规律不变一直用，以这个为标准可以预测今后相当长的时间内的数据，且预测差异和时间无关。

2）另一种场景是利用当前的数据预测一个近期内的数据，数据还会一直来，只是当前不知道，需要预测而已。当下一批数据来了之后就可以继续预测再下一阶段的数据。此时的预测和时间远近有关，越近越准确，越远越不准确。

### 数据拆分

由于数据的顺序对模型有关系，因此不能随机拆分。