

乘用车销量混合预测模型



中国科学院大学
University of Chinese Academy of Sciences

目录

01 问题分析

02 主体框架

03 内容阐述

04 预测结果

05 结论

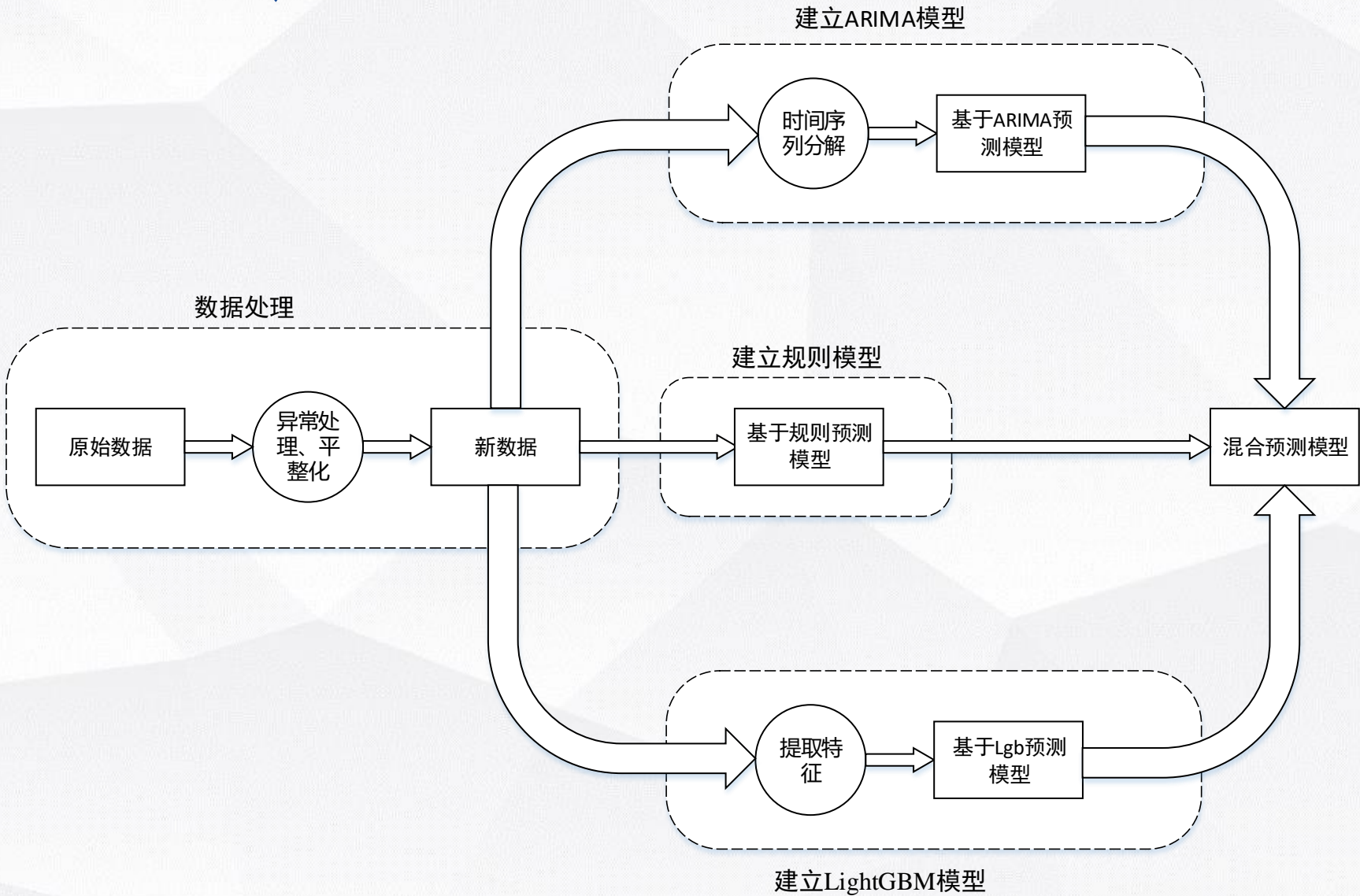
01. 问题分析

在销量数据自身趋势规律的基础上，找到消费者在互联网上的行为数据与销量之间的相关性，为汽车行业带来更准确有效的销量趋势预测。

建立销量预测模型，基于该模型预测同一款车型和相同细分市场在接下来一个季度连续4个月份的销量。

02. 主体框架

乘用车销量预测机制



03. 内容阐述

原始数据

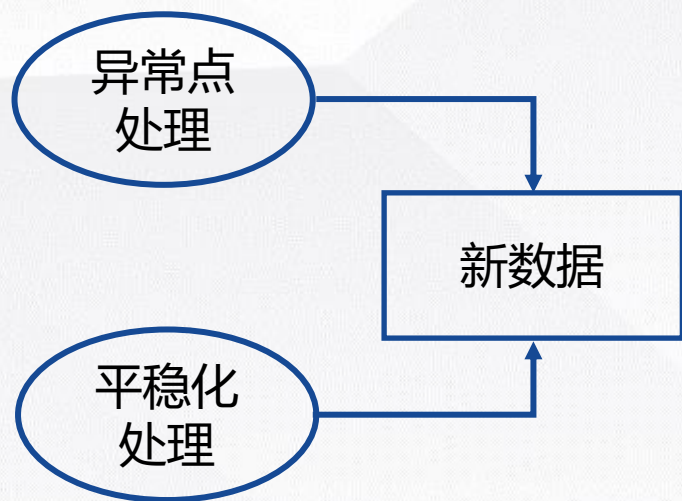
[训练集]**历史销量数据**：包含60个车型在22个省份，从2016年1月至2017年12月的销量。包含字段：省份/省份编码/车型编码/车身类型/年/月/销量。

[训练集]**车型搜索数据**：包含60个车型在22个省份，从2016年1月至2017年12月的搜索量。包含字段：省份/省份编码/车型编码/车身类型/年/月/搜索量。

[训练集]**汽车垂直媒体新闻评论数据和车型评论数据**：包含了垂直媒体中，各车型的每月（不分地域）每月新闻评论数据、车型下的评论数据两部分，这两个数据没有任何包含关系。包含字段：车型编码/年/月/对车型相关新闻文章的评论数量/对车型的评价数量。

[测试集] **2018年1月至4月的各车型各省份销量预测**：包含字段：ID/省份/省份编码/车型编码/年/月/预测销量。

数据处理



将原始的训练数据按省份和车型划分为1320个子数据集，分别在子数据集上进行异常点判断和平稳化处理。

异常点处理

异常点为不同年同月份数值比差异较大的数据，以每个子数据集月比值的2.5倍—三分位距作为离群点的检测间距时可以取得最优的结果。异常点的数值修正为同分月数据的几何平均值。

平稳化处理

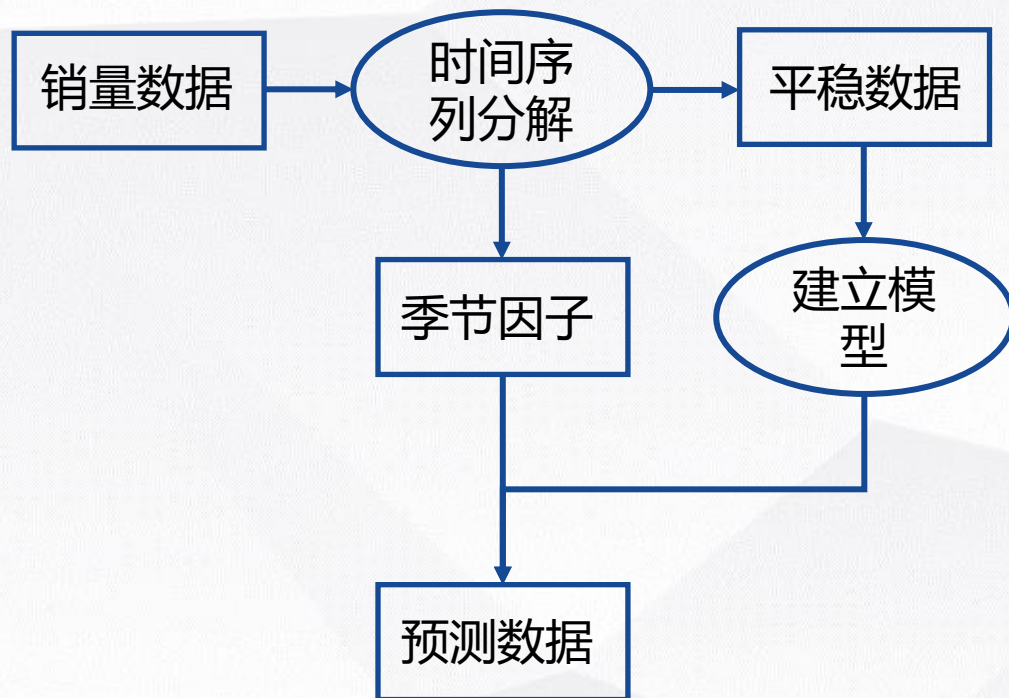
通过ADF检验方法对子数据集进行平稳性检验，检验结果显示其非平稳，因此将数据取log处理，增加其平稳性，以便接下来的模型拟合。对于部分子数据集继续进行差分处理，以达到ADF平稳。

由于销售数据受季节因素影响且随时间变化较大，当预测时间较长时会出现较大的误差，因此引入基于目前销售数据的规则来矫正其他模型的预测偏差。

将子数据集的2017年与2016年销售数据均值的比值作为某省份某车型的销量趋势因子，2017年前后三个月份的加权和作为预测的基础销量，二者相乘为规则下的销量预测。

```
# 17年均值除以16年均值得到趋势因子
df['factor'] = df['17mean'] / df['16mean']
# 取出16年12月, 17年1,2,3,4,5月, 共6个月
df = pd.merge(df, train16[train16['regMonth'] == 12][['adcode', 'model', 'salesVolume']],
              on=['adcode', 'model'], how='left').rename(columns={'salesVolume': 0})
for m in range(1, 6):
    df = pd.merge(df, train17[train17['regMonth'] == m][['adcode', 'model', 'salesVolume']], on=['adcode', 'model'],
                  how='left').rename(columns={'salesVolume': m})
result_df = pd.DataFrame()
temp_df = df[['adcode', 'model']].copy()
# 预测为上一年的一上一个月, 同一个月, 下一个月的加权, 再乘以趋势因子
for m in range(1, 5):
    temp_df['forecastVolum'] = (df[m - 1].values * 0.25 + df[m].values * 0.5 + df[m + 1].values * 0.25) * df['factor']
    temp_df['regMonth'] = m
    result_df = result_df.append(temp_df, ignore_index=True, sort=False)
```


基于ARIMA预测模型

**时间序列分解**

利用时间序列乘法分解分离出季节因子，使得子数据集数据更加平稳，消除周期化特征。

Auto_arima建模

使用auto_arima自适应的选取ARIMA模型中(p,d,q)最优取值。

滚动更新模型

由于子数据集的数据量过小，使用训练集模拟测试并更新模型，逐月预测并为模型添加新数据。

基于LGB预测模型

LightGBM简介

LightGBM是一个用梯度Boosting框架, 基于决策树算法: 直方图算法、直方图做差加速、带深度限制的Leaf-wise的叶子生长策略、直接支持类别特征 (即不需要做one-hot编码)、直接支持高效并行

提取特征

取评论数、回复数、搜索数、同车型年均销量、同车种类年均销量、同省份年均销量、每月份权重等11个参数作为拟合模型的特征。

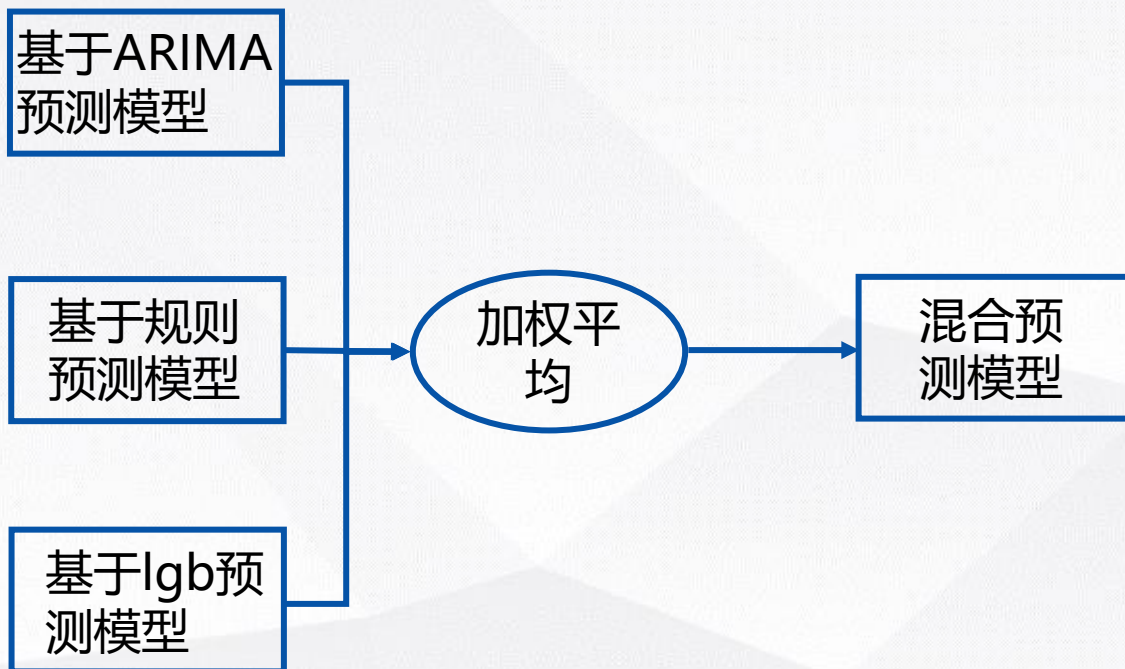
模型调参

通过线下训练集测试确定模型的拟合效果, 调整树深度以及树叶数调整过拟合现象。

模型测试结果:特征重要程度

('model', 31167),
('shift_model_adcode_mt_label_4', 25465),
('popularity', 25055),
('adcode', 23028),
('carCommentVolum', 21617),
('newsReplyVolum', 21041),
('regMonth', 18875),
NRMSE的均值: 0.7105693304304348

混合预测模型



将以上三种模型预测结果，按照加权更新预测，在线下训练集中实验出最优权重，最终得到较单一模型更为精准的结果。

04. 预测结果

初赛排行榜

复赛排行榜

周榜单

A 榜

B 榜




173

当前排名: 第 259 名

最高得分: 0.61758220

排名变化: -

排名	排名变化	队伍名称	最高得分	有效提交次数	最高分提交时间
	-	宝可梦训练师LZA	0.66562545	0	2019-10-29 17:50

05. 结论

创新点

滚动更新模型：相比原ARIMA模型线上测试准确率提升7%。

混合预测模型：混合模型相比单一模型线上预测准确率提升8%。

不足

1

2

3

4

ARIMA模型训练集数据量过少，影响长期预测。

模型混合时缺乏针对性。