2019 CCF 大数据与计算智能大赛

乘用车细分市场销量预测模型

《大数据分析》实验报告

目录

1 研究背景与最终排名

1.1 研究背景

近几年来，国内汽车市场由增量市场逐步进入存量市场阶段，2018 年整体市场销量首次同比下降。在市场整体趋势逐步改变的环境下，消费者购车决策的过程也正在从线下向线上转移，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。汽车销量的预测在宏观上有利于汽车行业产能监控、避免产能过剩、把控行业成长趋势、促进行业更好发展，微观上有利于车企制定生产营销策略、平衡供需、优化供应链。目前，基于网络搜索数据的汽车销量预测研究较少，并且传统的汽车销量预测方法有着预测精度低、预测粒度大等不足。本文就此问题将网络搜索数据与时间序列结合建立预测精度更高的混合模型。根据 60 款车型在 22 个细分市场（省份）的销量连续 24 个月（从 2016 年 1 月至 2018 年 12 月）的销量数据和同时期的用户互联网行为统计数据（各细分市场每个车型名称的互联网搜索量数据、主流汽车垂直媒体用户活跃数据等），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。

1.2 最终排名

考虑到提高课程项目报告说服度和课程项目打分需要，就把竞赛最终成绩和排名放到最前面。最终，我们小组的销量预测模型在 2019CCF 大数据与智能计算大赛乘用车细分市场销量预测项目中最终

成绩如下：

173

当前排名： 第 259 名　　最高得分：0.61758220　　排名变化： -

由于第一名成绩是 0.66（准确率 66%）左右，其实我们的成绩已经算是不错，仅差了不到 5 个百分点。

2 研究方法及创新

2.1 研究方法

从目前存在的预测方法入手分析，选取适用于当前数据的销量预测方法；对数据进行处理，与实际相结合初步建立基于规则预测模型；提取数据特征，训练 LightGBM 预测模型；时间序列乘法分解销量数据，训练 ARIMA 预测模型；结合三者提出新的混合预测模型，调整参数权重以达到最优结果。

2.1.1 LightGBM 预测模型

LightGBM 是一个梯度 Boosting 框架，分布式的，高效的，使用基于决策树的学习算法。其树生长算法直接选择最大收益的节点来展开，在更小的计算代价上去选择我们需要的决策树，控制树的深度和每个叶子节点的数据量，能减少过拟合。

2.1.2 ARIMA 预测模型

Autoregressive Integrated Moving Average model，差分整合移动平均自回归模型，又称整合移动平均自回归模型，时间序列预测

分析方法之一。ARIMA（p，d，q）中，AR 是"自回归"，p 为自回归项数；MA 为"滑动平均"，q 为滑动平均项数，d 为使之成为平稳序列所做的差分次数（阶数）。

2.2 创新点

(1)模型滚动预测，模型对销量数据每个月拟合一次，用新加入的数据拟合下一个月的数据。采用滚动预测的模型相比原模型线上测试准确率提升 7%。

(2)时间序列乘法分解数据，并将季节因素添加权重，结合实际调整权重参数。采用带权重的季节因素预测模型相比未使用时间序列分解的预测模型线上预测准确率提升 18%。

(3)混合模型预测，将基于规则的预测模型、基于 LightGBM 的预测模型和基于 ARIMA 的预测模型相融合，混合模型相比单一模型线上预测准确率提升 8%。

3 数据集分析及处理

3.1 数据集分析

比赛提供的数据集一共包括三个训练数据集和一个测试数据集：

[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。

首先分析字段我们发现，训练集中历史销量数据和测试集中字段基本一致，可以说是相关性最高的数据集。车型搜索数据中，只有销量字段换成了搜索量，那么销量和搜索量之间的关系也应该可以挖掘出来。而媒体评论和车型评论训练集中，就和测试集差别很大，并且只有评论数量数据，而好评差评等数据并没有说明，实际和销量是否有相关性，或相关性的正负都不好说，可以说是价值最低的数据集。

3.2 数据预处理

3.2.1 数据相关性分析

数据共包含销量数据、搜索量数据和评论量数据，对子数据集分别进行皮尔森相关性系数计算，结果显示其数据相关性较低。由于搜索、购买和评论行为存在先后顺序，数据存在领先和滞后性，因此对皮尔森相关性计算中加入时差系数，即按月份移动数据，以较早的搜索数据、适中的销售数据和较晚的评论数据整合计算，以获得最优的相关性，并分别按其滞后阶数调整子数据集。

3.2.2 异常点处理

根据实际与常识，将整体的训练数据按省份和车型划分为 1320 个子数据集，分别讨论其异常点的取值范围。

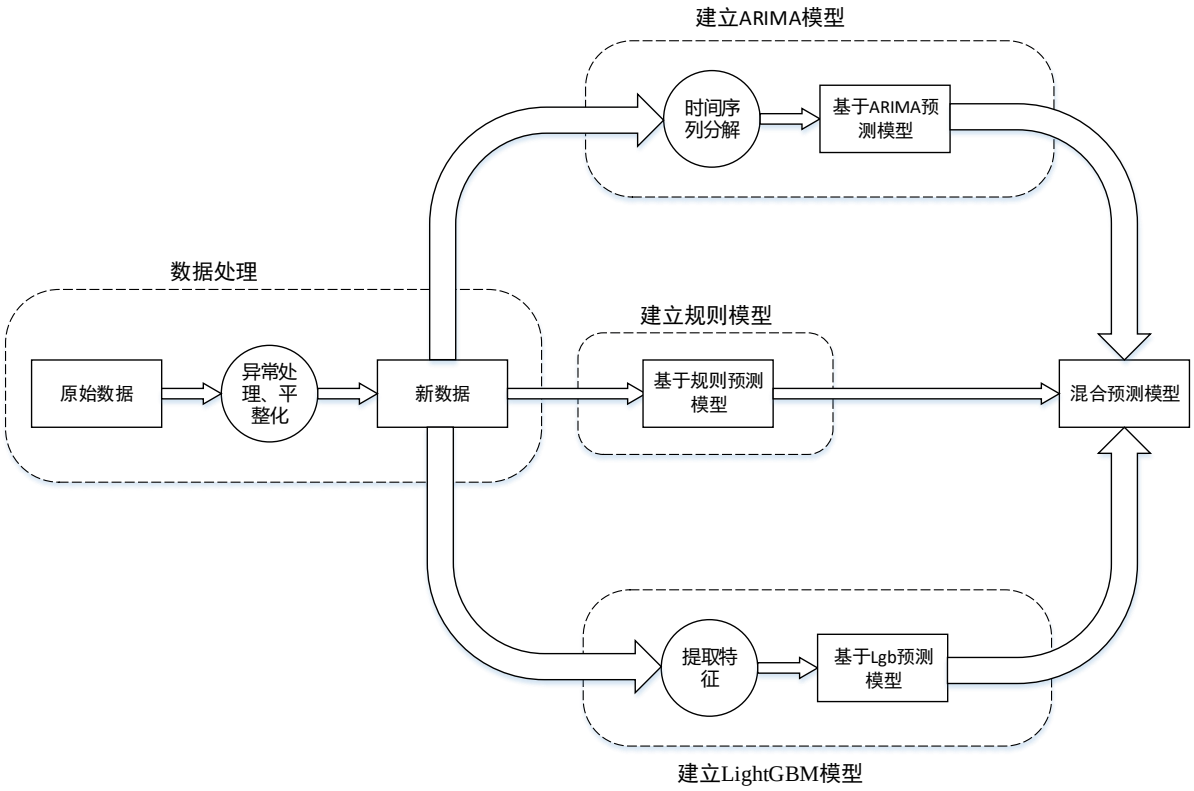
由于其受季节因素的影响，同年中数据的较大波动为正常现象，所以该数据中其异常点为不同年同月份数值比差异较大的数据。通过线下测试集的实验，显示以每个子数据集月比值的 2.5 倍一三分位距作为离群点的检测间距时可以取得最优的结果。确定异常点后，将每个异常点的数值修正为同分月数据的几何平均值。

3.2.3 数据平稳性处理

通过 ADF（Augmented Dickey-Fuller test）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。

4 研究过程和预测模型

整个项目的研究过程分为基于规则预测模型、基于 ARIMA 预测模型、基于 LightGBM 预测模型和混合预测模型四个阶段。



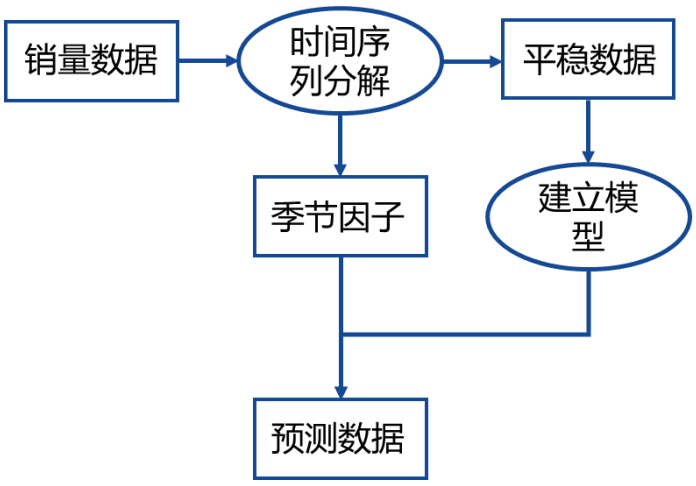
4.1 基于规则预测模型

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。

将历史销量数据子数据集的 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。

4.2 基于 ARIMA 预测模型

由于销售数据是完全以时间轴的为线索的观测数据，并且该数据集具有显著的季节特征和周期变化，因此选取时间序列预测分析方法的差分整合移动平均自回归模型(ARIMA)模型作为该数据预测的主模型。



4.2.1 时间序列分解

由于销量数据受季节变化较大，易造成模型拟合偏差，为消除此影响，在平稳化的子数据集上利用 STL 算法分解出季节因子，使得子数据集数据更加平稳，消除周期化特征。2018 年预测数据的季节因子为 2016 年与 2017 年同月份季节因子的加权和，在线下测试集实验得出最优权重。

4.2.2 Auto_arima 方法

由于 ARIMA 算法中 p（自回归项数），q（滑动平均项数），d（差分次数）参数的取值对模型的拟合有着重要的影响，而针对大量的子集循环判断取值建模复杂度较高，并且难以还原数据，因此调用 pmdarima.arima 中的 auto_arima 算法自适应的选取最优 p、q、d 参数取值，以达到最佳的模型拟合效果。

4.2.3 滚动预测

由于实际情况中不同省份不同型号的车辆销售存在相对的独立性，针对此情况，对每个子数据集单独处理共建立 1320 个 ARIMA 模型，在模型的拟合过程中，采用滚动迭代的方式，每个月以新数据重

新拟合一次模型用来预测下个月的销售数值。最终以此方法取得了较原 ARIMA 模型线上更高的准确率。

4.3 基于 LightGBM 预测模型

4.3.1 提取特征

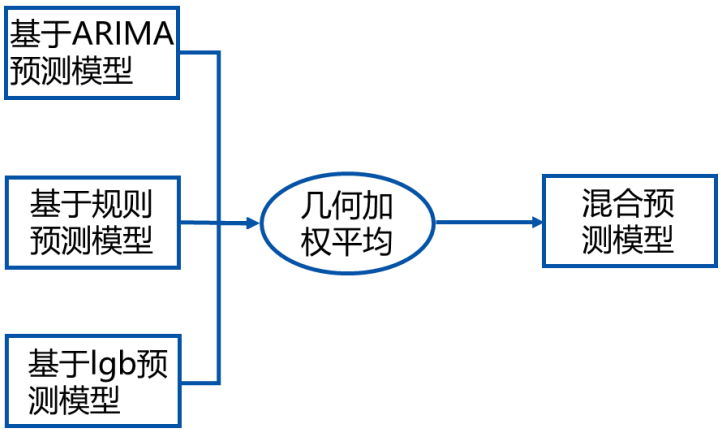
由于 LightGBM 是基于决策树的算法，因此其特征是影响模型预测效果的重要因素。基于相关性滞后调整的子数据，取评论数、回复数、搜索数、同车型年均销量、同车种类年均销量、同省份年均销量和每月份权重作为拟合模型的特征，每月权重根据当年新年等节假日所在月份确定。

4.3.2 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。

4.4 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。



5 总结

该销量预测模型结合了 ARIMA 与 LightGBM，并在长预测中补足了实际规则，采用了滚动预测模型、时间序列季节因素分解和混合模型等创新点，在预测的准确度上取得了一定的提高。但仍存在以下不足：在规则模型中对影响销量的特征针对性不强，规则过于片面；ARIMA 模型时间复杂度过高，调试较困难；LightGBM 模型特征较少，缺乏影响大的强特。