**IJCAI-17口碑商家客流量预测大赛解题报告**

**cGAN队**

# 简介

商家客流量的预测对商家的经营管理至关重要。在口碑平台上，客户流量定义为“单位时间内在商家使用支付宝消费的用户人次”。在这个问题中，口碑提供了用户的浏览和支付历史，以及商家相关信息，我们需要根据这些信息预测2000家商家未来14天的每天的客流量。

随着移动定位服务的流行，阿里巴巴和蚂蚁金服逐渐积累了来自用户和商家的海量线上线下交易数据。蚂蚁金服的O2O平台“口碑”用这些数据为商家提供了包括交易统计，销售分析和销售建议等定制的后端商业智能服务。本次比赛是阿里巴巴集团与IJCAI-17共同主办的，是工业界与学术界的一次深入的交流碰撞。值得创新的是本次比赛不限制使用外部数据，因此预测这个问题同时也引入了先验。

评测标准为sMAPE, 需要预测所提供的的2000家商家2016.11.01-2016.11.14内各自每天的客户流量，其实预测结果为非负整数，同时值得注意的是，官方确保实际预测的商家中未来十四天的销量是严格为正的，这也是一个值得注意的先验。

# Models

宏观问题需要用宏观的解法，微观问题用微观的建模思路，这是在我们考虑数据的时候必须首先明确的一点，显然，按天的预测是属于宏观问题，因此微观、细粒度的数据维度特征是信噪比是小于1的，我们在实际中的验证结果也如此，加入细粒度的特征并不能提高线上得分表现。

我们队从四种不同角度的模型出发，利用了Time Series，Neural Networks，Tree Based Models，Prior Belief Models ,对预测序列进行建模，然后进行了简单的融合。下面将对所使用的模型分别介绍。

## Time Series

本次比赛提供了历史一年内的商家客流量序列，但是其存在严重的缺失，并且序列的时效性依旧是很很重要的，通过线下验证我们使用了窗口为51以及28天的时间序列进行建模。

同时缺省值的处理，我们使用了两种不同的填充方式：

a.按星期就近填充

b.趋势分解后，线性插值填充

Time Series Based Models as below:

1. SVD + stlf/ets - this model applied SVD to the training data as preprocessing, and then forecast each series with stlf(), using an exponential smoothing model (ets) for the non-seasonal forecast.
2. SVD + stlf/arima - the same, but with an arima model for the non-seasonal forecast
3. Standard scaling + stlf/ets + averaging - like (1), but SVD was not used. Instead, the data were standard scaled, and a correlation matrix was computed.
4. Raw Series + stlf/arima - This used auto.arima() from the forecast package.
5. Raw Series + stlf/ets – We set biasadj as TRUE, which makes our model more robust.
6. A simple average of our time series models (totally 10 models).

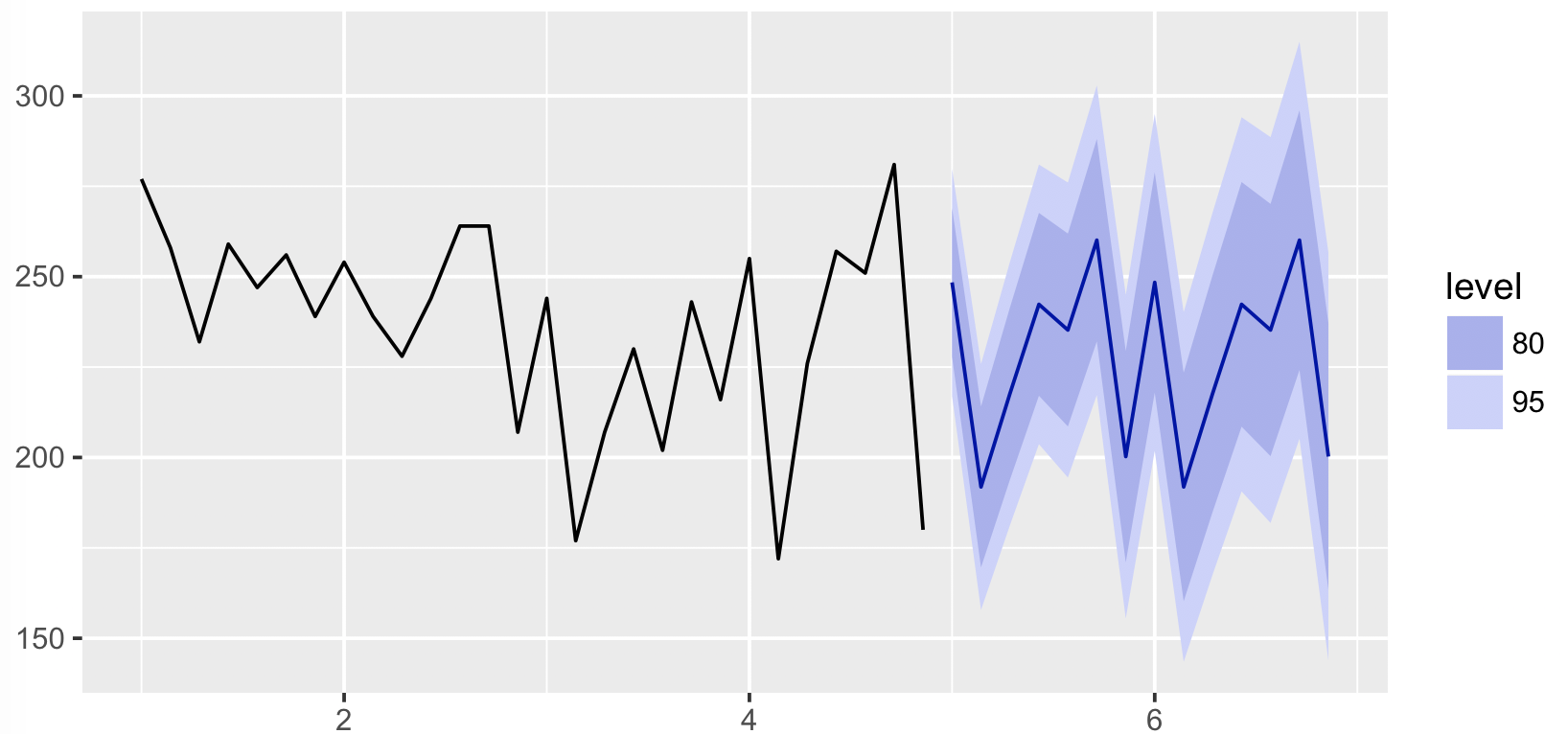
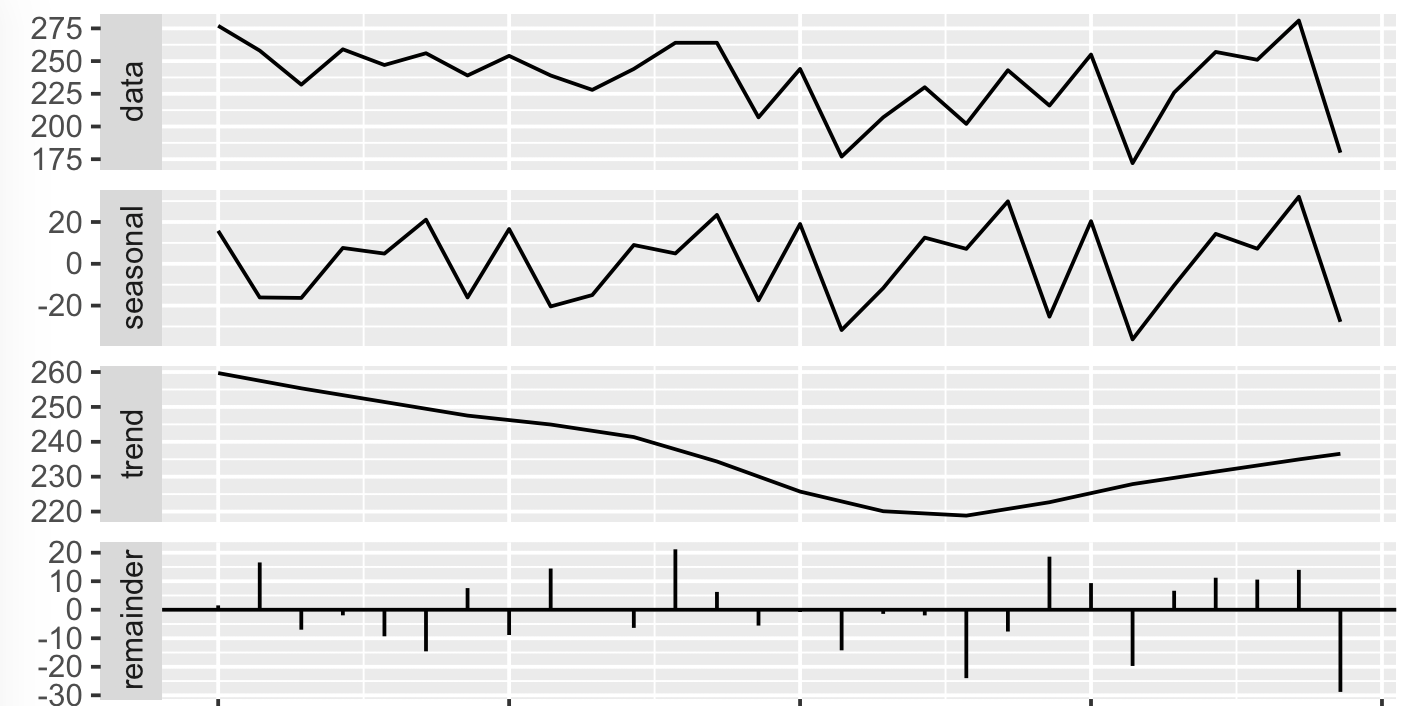


图2-1 stl +ets forecast

Tips:

1. We use 7 as the SVD decomposition component. This is chosen by our validation.
2. Always, we predict 7 days, then make a copy for next week.
3. Our best single time series model get 0.086 in the leaderboard (SVD + stlf/ets), with a simple average of our time series models, we get 0.082 in the leaderboard.
4. We use a powerful time series R packages ‘forecast’.

## Neural Networks

我们使用了stacking of different NNets , 单NN最高分在0.084(LSTM , windows=51)，通过stacking , 我们的NNs可以达到了0.082。

同样，我们仍然没有加入除时间序外的特征。此处对商家进行了一个embedding.

使用的NN models有：

1. LSTM+ embedding of shop\_id with slide windows 51.
2. CNN+ embedding of shop\_id with slide windows 51.
3. MLP+ embedding of shop\_id with slide windows 51.
4. A two level stacking among all NN models.

## Tree Based Models

我们使用了XGBoost, ExtraTrees, RandomForests 三种树模型，其中XGBoost 使用了28天的序列，没有添加任何其他特征，线上得分为0.082。ETs 和 RFs 使用了21天的序列，添加了商家信息的特征，线上得分为0.083。

考虑到序列预测问题本身存在很大的噪声，同时为了追求线上与线下的一致性，我们一直使用最后一周作为验证集，同时模型只预测第一周的商家流量。

我们的基本原则是，在没有引入任何先验的情况下，模型对于下一周的预测一定是偏向上一周的，也就是说预测第二周本身来说在没有先验条件的情况下的最优的做法是直接复制第一周的预测。

## Prior Belief Models

以上模型的假设是存在问题的，理由如下：

1. 序列存在趋势性的问题，直接拷贝第二周丢失了趋势。
2. 第二周时间区间上(11.8-11.14)跨度经过了双11，通过查阅口碑中国的微博，我们发现11.11这一天至少在北京、上海、广州、杭州等城市存在促销活动，这也是先验规则之一。
3. 不同商家间的销售模式是不同的，例如外卖商家与火锅商家在具体销售表现上是不同的，双11当天外卖商家的销量变化是远远大于火锅商家的。
4. 天气数据也存在一定的影响，例如十一月份火锅商家整体上销量是越来越好的。
5. 不同地区存在不同的促销方式，例如杭州地区的促销较多，同时杭州地区的预期增长是较其他城市最高的。

基于以上合理的先验假设，我们对第二周商家销量行为进行了规则上的调整。

# Ensemble

我们采用stacking level 3 的方式融合所有不同模型。最终得分为0.0795，同时通过对于第二周的人为先验规则，使得分提高了0.0015。

整体模型框架如下：

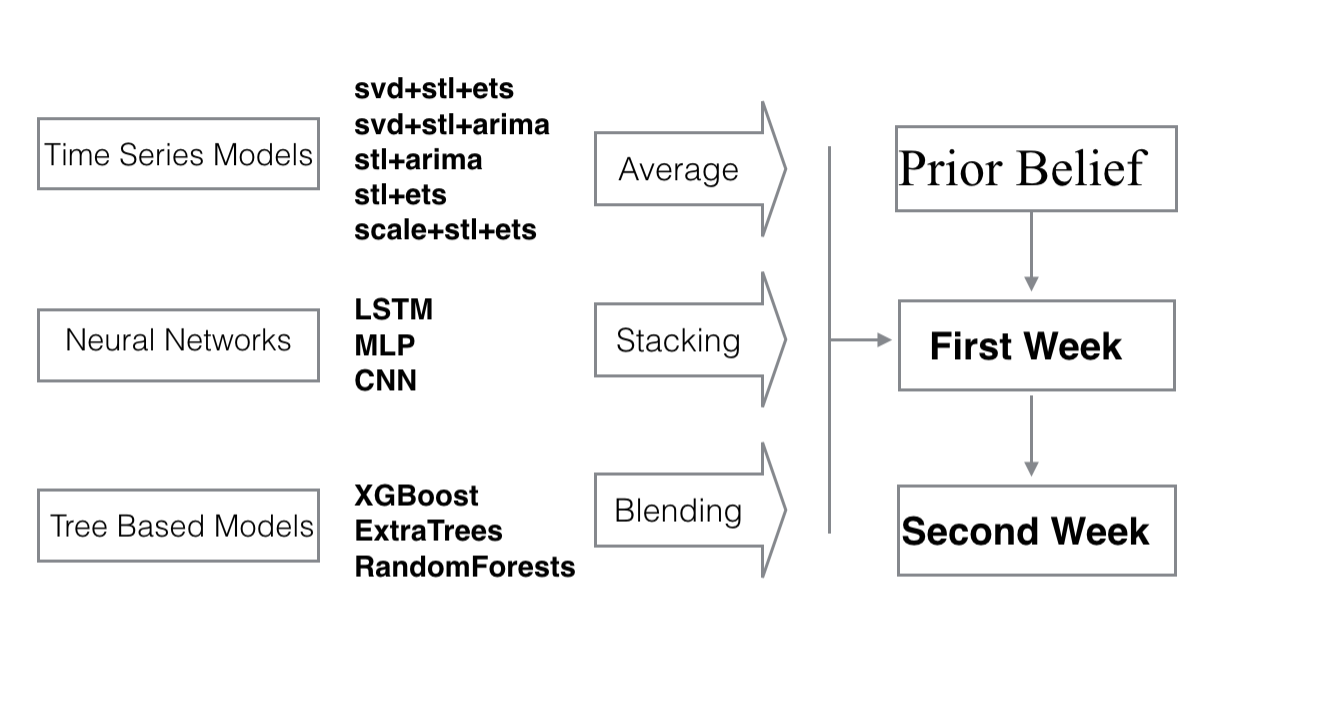


图3-1 整体模型框架

# 总结

感谢阿里巴巴、IJCAI-17共同举办这次比赛，让我们有机会能够接触到真实的数据以及模型的应用场景。在比赛过程中，我们也尝试了各种各样的方法，虽然并不是总能带来提升，但是在对算法的理解上又更近了一步，对自身的能力的提升带来了很大的帮助，期望主办方今后还能举办更多有趣有挑战的竞赛。

# 参考

1. Džeroski S, Ženko B. Stacking with Multi-response Model Trees[C]// Multiple Classifier Systems, Third International Workshop, MCS 2002, Cagliari, Italy, June 24-26, 2002, Proceedings. DBLP, 2002:201-211.
2. Gers F, Eck D, Schmidhuber J, et al. Applying LSTM to Time Series Predictable through Time-Window Approaches[C]// International Conference on Artificial Neural Networks. Springer-Verlag, 2001:669-676.
3. https://www.otexts.org/fpp
4. Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754 .
5. Hamilton J D. Time series analysis[M]// Time Series Analysis. 1994:401-409.
6. Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[J]. 2016.
7. Sollich, P. and Krogh, A., *Learning with ensembles: How overfitting can be useful*, Advances in Neural Information Processing Systems, volume 8, pp. 190-196, 1996.