Rossmann销售预测报告

# 问题的定义

　　这是Udacity机器学习课程的毕业项目，项目中将应用神经网络求解销售预测问题。

## 销售预测

作为销售计划的中心任务之一，销售预测是指通过一定的手段和方法，对未来特定时间内全部产品或者特定产品的销售数量或销售金额的估计。在操作层面上，销售预测是在充分考虑了未来各种影响因素的基础上，结合本企业的过往销售业绩，运用一定的分析方法提出切实可行的销售目标。

影响销售的因素很多，包括有需求变化、经济变动等成分的外界因素，和有营销策略、生产状况、销售人员等成分的内部因素，但销售预测对于完善客户需求管理、指导运营、优化供应链、提高企业利润方面都具有重大促进作用，降低企业的业务计划的不确定性，因此，提出合理的销售预测一直也是人们孜孜不倦的追求，是企业辅助决策的重要工具。销售预测也几乎成为了所有商业智能（BI）的终极目标。

多年来，人们已经形成了定性分析法和定量分析法两类分析方法。其中，定量分析法中的趋势预测分析法和因果预测分析法在实际应用中也能取得一定的预测效果。他们可以基于历史数据，运用一些直观的算法，如算术平均法、指数平滑法，来进行预测，相比定性分析法，预测效果也有了一定的提高[1]。

但随着经济全球化，商业网络化的进程，市场竞争日趋激烈，业务复杂化，数据规模海量化，传统的预测方法已经越来越力不从心。预测的精确性成为销售预测的核心痛点，人们迫切希望一些性能更高，更智能的预测方法。数据挖掘[2]技术由计算机自动从大量数据集中提取隐含的、事先不知道但有潜在应用价值的信息，可用于学习复杂销售的复杂特征对于销售的影响，从而得到较好的预测效果。本课题将应用机器学习算法来实施销售预测。

## 问题分析

Rossmann是欧洲的一家连锁药店，在欧洲7个国家拥有超过3000家药店。这是一个Kaggle[比赛项目](https://www.kaggle.com/c/rossmann-store-sales)，本课题将按照项目中的说明，需要为Rossmann在德国的1115家药店做出提前6个星期的每日销售预测。

对于Rossmann的销售预测问题将是一个具体领域的销售预测问题，作为药店连锁店，具有一定的行业特征，这些将体现为数据特征。如上节所述，基于机器学习算法的数据挖掘技术会是一种可能的得到一个合适预测模型的办法。

本课题得到的预测模型将用于输出未来6个星期里每天的销售量，预测结果可以和实际销售情况对比，从而衡量预测效果。在实际使用中，还可以随着时间的推移，不断学习和预测。

## 评价指标

本项目将使用Kaggle项目中的评估函数RMSPE（Root Mean Square Percentage Error）作为评估指标：

其中，为一个门店某天的销售量，为相应的预测值。

# 研究分析

## 数据特征

作为一个Kaggle比赛项目，Kaggle提供了项目数据，包括训练数据train.csv，测试数据test.csv，已经商店信息数据store.csv。训练集数据的全部字段说明参见：<https://www.kaggle.com/c/rossmann-store-sales/data>，包括两部分：1，与销售日期相关的日期、促销、节假日特征值，2，回归值：销售额和顾客数量。测试数据并不包含回归值，只适合Kaggle在线测试，本项目采用离线实验的方式，故不使用。商品信息数据store.csv包括与每个门店相关的门店类别、促销情况、竞争门店信息。

另外考虑了，竞争门店的作用，并试图通过*CompetitionOpenSinceYear*和*CompetitionOpenSinceMonth*得到当日门店是否存在竞争门店，以及CompetitionDistance得到的竞争门店的距离，report2.html就是以CompetitionDistance训练的模型，在测试集上的RMSPE错误率约为12.6%，仍然有一定的改进空间。

为了获得更简洁的有效特征值，后来在Kaggle上面发现了门店的[State属性数据集](https://www.kaggle.com/c/rossmann-store-sales/discussion/17048)，通过引入State特征值，使用Entity Embedding算法的神经网络，得到了较好的预测效果。

所以，本项目将选取如下特征值用于销售预测：

* Store：商店编号
* Date：数据统计日期
* Open：当天门店是否开放，节假日商品可能不开放。
* Promo：当天门店是否进行促销活动
* DayOfWeek：一周的星期几编号
* State：门店所属的地区，德国各地区的节假日各不相同，所以可以认为State也表达了StateHoliday和SchoolHoliday的属性。

而回归值，即要预测的数据为：

* Sales：当天销售额

本项目将以离线的方式训练模型，并作出预测，以比赛项目的训练数据作为项目数据，分割出部分（0.5%）作为测试数据。

## 数据探索

本项目做了以下几种数据可视化探索：

1. 按月、按年对比：

随机选取5个store的销售总额趋势图：

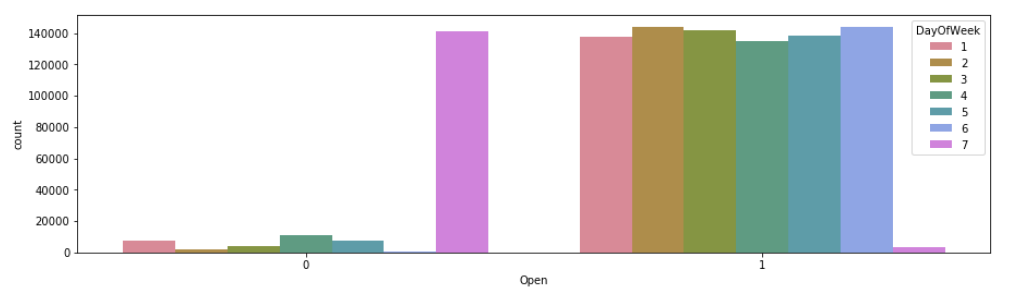


随机选取5个store的销售平均额趋势图：



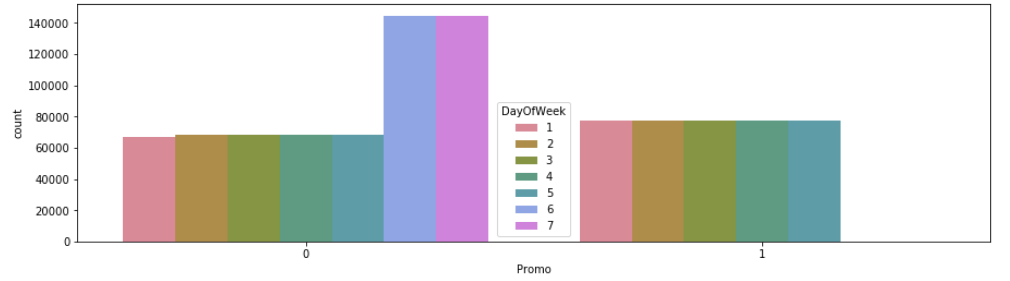
时间上的可视化总结：

1. 各个门店销售趋势走势不尽相同
2. 总体销售额呈现上升趋势
3. Open与销售额



Open特征值的可视化总结：

1. Open对销售额有明显的区分作用，当open=0时，销售额几乎为0，没有很好的区分度
2. Open对销售额的影响是显著的，也是固定的。
3. Prom与销售额



## 算法分析

神经网络

本项目研究了[Entity Embedding算法](https://arxiv.org/abs/1604.06737)，通过实验表明，变量压缩的做法不仅减少了内存的使用，而且相比one-hot encoding提高了算法的效率。

本项目采用[Keras](https://keras.io/)框架来实现具有Entity Embedding功能的神经网络。Keras是一个Python的深度学习库，提供了高层次的API，以便方便快捷地构建（深度）神经网络。

## 基准模型

本项目使用了两种Linear Regression（线性回归）作为基准模型，详细结果参考report.html “3.基准模型”：

1. 总体线性回归

总体线性回归，是试图在整个数据集上训练模型，特征值选择了与Embedding网络模型同样的特征值。测试中取得的RMSPE的错误率不是很理想，只有45%左右。

1. 基于门店的线性回归

考虑到，各个门店的差异较大，又尝试了基于门店的线性回归模型，即对每一个门店训练一个模型，测试取得的RMSPE错误率降到了16%左右。

# 研究方法

## 数据处理

数据处理通过prepare\_data.py实现，主要包含以下内容：

* 销售数据按日期排序，原始数据为时间降序排列，处理为升序排列
* 数据聚合：将State信息整合到Store信息中
* NAN处理：将nan数据用“0”填充
* 数据序列化：将原始3个数据集序列化为train\_data.pickle和store\_data.pickle。
* LabelEncoder：对选取的特征值进行LabelEncoder编码，并保持编码方案到les.pickle一边解码使用。Date也拆分为year，month，day三项。
* 类型一致化：将编码后的特征值一致化为int类型。
* 数据标准化：在训练Keras模型的时候，对回归值Sales进行了log百分比的标准化，减少了数据稀疏性。

全部过程通过prepare\_data.py的prepare\_data()实现。

## 算法实现

　　Entity Embedding神经网络由network.py实现，项目中使用Keras框架构建神经网络，具体做法为：

* 特征值：模型训练使用了Store ID，Day of week, promo，Date，Stat组成的7项特征值，其中Date由Year，Month，Day三个子项组成。
* 网络模型: Sequential
* 网络层：
  + 每个特征值应用一个Embedding层
  + 全连接层：1000维输出
  + 激活函数：relu
  + 全连接层：500维输出
  + 激活函数：relu
  + 全连接层：1维输出
  + 激活函数：sigmoid

网络层总体结构如图：

Dense

lrelu

Dense

lrelu

sigmoid

Dense

* Optimizer和Loss函数
  + Optimizer = adam
  + Loss = mean\_absolute\_error

## 算法改进

本项目尝试了以下几种算法调整：

* **特征值的选取变换：从CompetitionDistance到State：**

就如同开题报告中提到了，首先尝试了使用CompetitionDistance作为一个特征值，具体做法为：

1. 提取每天的CompetitionDistance值：由Store信息中的CompetitionOpenSinceYear和CompetitionOpenSinceMonth组成CompetitionDate和当前时间比较，如果当前时间小于CompetitionDate，以0为Distance，如果大于CompetitionDate，则以Store信息中的CompetitionDistance为当天的CompetitionDistance。
2. 对CompetitionDistance编码：由第一步得到的CompetitionDistance，定义一个距离类别变量：<=50为close，50到200为near，200到1000为normal，大于1000为far（测试过程中进行了不同的调整）。最后再以其他特征值同样的LabelEncoder处理，并在Entity Embedding网络中用Embedding层处理。

最后的测试结果为report2.html所示，测试集上的RMSPE误差为0.126，仍然有改进的空间，所以找到了State数据集，并取得了更好的结果。

* **Optimizer选取：从sgd到adam**

试验中，尝试了多种Optimizer，如标准sgd和adam，统计结果如下：

|  |  |  |
| --- | --- | --- |
| **测试项目** | **sgd** | **adam** |
| 1个模型 | 0.29 | 0.12 |
| 5个模型 | 0.29 | 0.098 |

所以，无论是一个还是多个模型，adam的效果都比sgd要好。

* **网络层选择：**

在全连接层的选择上，也做出了一些调整与尝试，如开始使用了1个全连接层，得到的最好的测试结果仅为0.11，而选择2个全连接层时，可以达到0.098。再次尝试添加全连接层（3层）的时候，测试结果并没有明显改善，故最终使用了2个全连接层。

* **训练多个模型取均值**

试验中，如果每次只训练一个模型，并在测试集上做预测，所得到的结果并不是很理想，比如在采用adam作为Optimizer的时候，一个模型上得到的RMSPE误差大约为0.12，并不是特别理想。但如果训练5个模型，分别训练模型，并对各个模型的预测结果取平均值，所得到的RMSPE误差约为0.098，就能得到了更理想的结果。

# 结果总结

## 模型评估

在选取了[算法实现](#_算法实现)中的各项参数后，模型在测试集上取得了RMSPE误差0.098的结果，预测效果较为理想。

以下是项目中相关参数的总结，进一步从理论上说明了模型的可靠性：

1. adam Optimizer[3]

adam优化方法可以动态调整每个参数的学习率，在于经过偏置校正后，每一次迭代学习率都有个确定范围，使得参数比较平稳。确保了一定的鲁棒性和稳定性，为模型的稳健型提供了一定的保障。

1. 输入层完备性

由[算法实现](#_算法实现)中可以看到，本模型为选取的7个特征值分别建立了tensor，这也为模型的训练提供了更好的学习方向，另外采取了Embedding层，减少内存使用的同时，也提高了算法的效率，但是相关参数需要一定的调整时间。

## 基准模型比较

基准模型（Linear Regression）与Entity Embedding神经网络的结果总结如下：

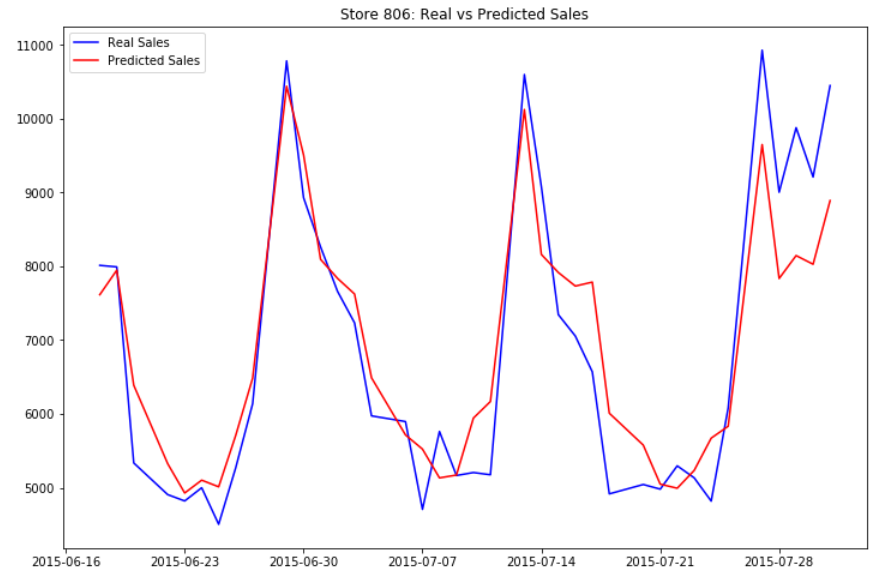
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | 总体线性回归 | 基于门店的线性回归 | 1个模型神经网络 | 5个模型神经网络 |
| **错误率** | 0.458 | 0.162 | 0.12 | 0.098 |

所以，本项目得到的模型，可以较好地拟合Rossmann的销售情况，预测效果也较为满意。

# 结论分析

## 预测结果可视化

　　report.html中可以看到，在测试集中随机选取一个门店执行销售预测，并与真实数据一起做可视化分析统计。如门店806的2015-06-16到2015-07-28区间内的销售预测情况如下：



## 项目思考

本项目在实现过程中有如下两点思考：

1. 使用Embedding网络

根据本项目的特点，各个特征值均为，或者可以转换为，分类变量，适合Embedding网络。

由测试可知，Embedding网络的效率较高，而且只需要选择相对简单的特征值就能取得较好的效果。由参考文献1可知，Entity Embedding避免使用one-hot encoding，由于one-hot encoding将占用更大的空间，使得神经网络变得更复杂。而Entity Embedding压缩了类型的维度，也达到类似编码的效果，是个有趣的处理分类特征值的方法，值得进一步研究。

1. 可视化的挑战与魅力

本项目开始阶段就试图对数据进行一些可视化，report.html中销售额按年、月趋势的可视化操作就是一个例子。

由于之前没有完整的python用于可视化的数据处理和可视化操作，花费了不少时间。遇到的难点有：1，如何做数据的roll up？基于门店的按年、按月统计，最终通过play\_ground的notebook不断尝试，总结了如何分组数据，如何执行resample操作，从而达到想要的效果。2，何种方式执行可配置的可视化？为了达到较理想的可视化效果，同时支持可配置操作，也需要不断查找资料和尝试达到。在对Open和Promo的可视化时，使用了seaborn库，seaborn库是一个很方便的可视化类库。

但是，可视化实现以后，数据就可以通过直观的方式显示，从而快速了解数据特征，进而方便决策。可视化是个强大的工具。

## 项目改进

本项目在开题报告中就打算使用AR模型作为基准模型，相比而言，AR模型会比linear regression对于时间序列的拟合有更好的效果。但是实验中通过SARIMAX尝试建立了模型，但是没有得到按天的预测数据，时间紧迫，所以只好放弃。

如果说有什么改进，现在想到的，也许就是实现AR模型的基准模型，但这也不是本项目算法的改进。

# 参考文献

本项目参考的主要文献有：

【1】Entity Embedding of Categorical Variables: <https://arxiv.org/abs/1604.06737>

【2】Rossmann Store State辅助数据：<https://www.kaggle.com/c/rossmann-store-sales/discussion/17048>

【3】Optimizer比较：<https://zhuanlan.zhihu.com/p/22252270>

【4】Keras文档：<https://faroit.github.io/keras-docs/1.2.2/>