Rossmann 店铺销售预测

项目背景：

Rossmann 在 7 个欧洲国家运营有超过 3000 家门店. 本项目的目的是预测 Rossmann 所有门店在未来 6 周的销售额，根据他们现在所拥有的数据及市场数据进行数据分析。一个稳定且准确率高的预测分析，能帮助门店经理合理的调配资源包括货物及人力，从而将精力放在提高用户体验及满意度上面来提升品牌价值。通过数据预测数据值，本是一个回归预测任务，用到的模型包括 Linear Regression Models[1], Support Vector Machines(SVM) [2], Regression Tree[3], Gaussain Process Regression Models[4], Ensembles of Trees[5]等回归预测模型。

问题描述：

本项目需要进行对未来 6 周 Rossman 各门店的销售数据预测，通过日期, 门店性质(门店类型，竞品性质, 门店推广参数)及节假日因素来对数据进行回归预测。是一个 supervised regression 的问题。通过回归模型可以通过 Mean Absolute Error[6], Mean Square Error[7] 或者 R^2 Score[8]的回归评测方法来判定预测值与真实值之间的差距从而得出该回归模型的表现情况.

数据集及输入描述:

本项目数据来自 Kaggle 的的竞赛项目 Rossmann Store Sales. 分别有

train.csv, test.csv 还有 stroe.csv. 其中 train.csv 里面包含了各个店铺在过去三年的销售的额及客户总体数，是否店铺正常营业，店铺是否在做促销活动及是否公众假日，和学校假日数据。而 test.csv 测试集。训练数据有 1017209 条销售数据，测试集有 41088 测试数据，占训练数据 4%。店铺数据 store.cvs 有 1115 家店铺属性数据其中包括竞争对手属性，包括开业时间及与本店铺距离，是否参与促销活动，促销活动频次及商店总类。一般地，销售数据会受到促销，商店总类，及节假日等因素影响，另外还会受到大经济环境及外部消费数据等因素所影响。

	Store	DayOfWeek	Sales	Customers	Open	\
count	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	
mean	5.584297e+02	3.998341e+00	5.773819e+03	6.331459e+02	8.301067e-01	
std	3.219087e+02	1.997391e+00	3.849926e+03	4.644117e+02	3.755392e-01	
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
25%	2.800000e+02	2.000000e+00	3.727000e+03	4.050000e+02	1.000000e+00	
50%	5.580000e+02	4.000000e+00	5.744000e+03	6.090000e+02	1.000000e+00	
75%	8.380000e+02	6.000000e+00	7.856000e+03	8.370000e+02	1.000000e+00	
max	1.115000e+03	7.000000e+00	4.155100e+04	7.388000e+03	1.000000e+00	

	Promo	SchoolHoliday
count	1.017209e+06	1.017209e+06
mean	3.815145e-01	1.786467e-01
std	4.857586e-01	3.830564e-01
min	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00
75%	1.000000e+00	0.000000e+00
max	1.000000e+00	1.000000e+00

训练数据概述

```
-----------------
              Store  CompetitionDistance  CompetitionOpenSinceMonth  \
count   1115.00000          1112.000000                 761.000000
mean     558.00000             0.071003                   7.224704
std      322.01708             0.101044                   3.212348
min        1.00000             0.000000                   1.000000
25%      279.50000             0.009197                   4.000000
50%      558.00000             0.030393                   8.000000
75%      836.50000             0.090487                  10.000000
max     1115.00000             1.000000                  12.000000

       CompetitionOpenSinceYear       Promo2  Promo2SinceWeek  Promo2SinceYear
count                761.000000  1115.000000       571.000000       571.000000
mean                2008.668857     0.512108        23.595447      2011.763573
std                    6.195983     0.500078        14.141984         1.674935
min                 1900.000000     0.000000         1.000000      2009.000000
25%                 2006.000000     0.000000        13.000000      2011.000000
50%                 2010.000000     1.000000        22.000000      2012.000000
75%                 2013.000000     1.000000        37.000000      2013.000000
max                 2015.000000     1.000000        50.000000      2015.000000
```

商店数据概述

解决方案

本项目目标是预测 Rossmann 未来 6 周销售数据, 销售数据属于连续的数据。本项目为回归预测类问题。应才有回归预测模型进行数据预测。数据及输入中有 1017209 条销售数据，及 1115 家店的性质数据，可通过数据的整合，来建立每一条销售数据的特征，而每一条观测数据的销售值就是目标标签数据，为训练模型作预测值比对。通过回归预测的评审方式，来计算测量模型的表现，从而通过评分来反向推导特征的选择，变化及模型的调节上，最终使评分达到理想状态。

基准模型

本项目我会选择线性回归[9]为其基准模型，线性回归模型简单容易理解，并且能快速训练模型。而通过线性回归得出一个通过 R2 方法来计算的结果约为 0.24 (1 为效果最好)。线性回归是回归预测中最简单最直接的学习算法，对于少量特征的预测有较好效果，但对于本项目多特征值情况，应采用对多维预测有较好效果的算法如随机森林或者梯度决策树模型。

评估方式

该项目的评估方式才有 Root Mean Square Percentage Error(RMSPE)[10]进行评估，此评估公式为一损失函数。换而言之损失约小，模型表现越好，越强的预测能力。

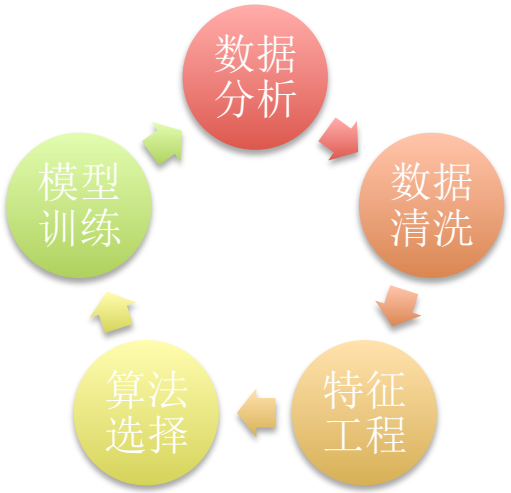
$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2},$$

yhat_i 为预测值，yi 为真实值，通过计算计算差值百分比来评判模型表现能力。

项目设计

本项目为数据分析及预测项目。会涉及到 5 个基本步骤：

数据分析，**数据清洗，特征工程，算法选择，模型训练**



项目流程

1. 数据分析：

数据分析就是先确立本项目的目的，从目的出发进行数据选择及采集，分析数据类型，数据种类及该数据与本项目预测目标是否存在基本联系关系。采集上会选择多少个样本作为模型的训练样本，会选择多少个特征，预测目标数据是否足够及清晰等。

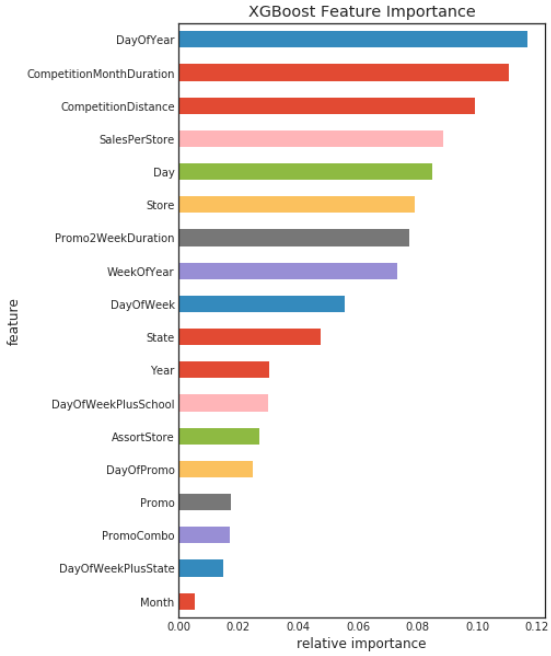
2. 数据清洗

本阶段为检查训练及测试数据中是否存在空值或者异常值，处理重复输入数据及类似数据合并，数据处理的目的是为了使模型能更好地对真实数据的认知。

3. 特征工程

特征工程的目的是基于现有特征组建新的特征，这个组建或者转化的过程需要对数据及该项目知识领域有基本的认识及见解，从而更好地去挑选作为模型训练的特征集，该过程为整个数据预测分析项目的重要环节，对后期模型训练有重要作用。这里还包括数据集的划分，训练集应划分部分数据样本作为验证集，此验证集的目的是作为验证模型选择及参数配置对结果影响的验证测试数据，而不需要用到真实测试数据进行从而有数据泄漏到模型训练过程，影响模型对新数据的预测能力。而这里验证集的划分应按照时间序列划分，从而能反映模型对未来数据即本项目目标的预测能力。

根据我们选取的模型，进行特征工程分析，分析其重要性并对潜在特征进行融合或者筛选工作。下图为 XGboost 模型的特征重要性图表。



4. 模型选择

根据项目特征规模来选择适合的模型进行训练，回归模型有很多，应该从效果及效率上进行取舍。在这里我会选择决策树类模型，对多参数，及二分数据特征有较好效果，特别是 boosting 类的决策树类模型，如 xgboost[11]及 lightgbm[12] 提高了性能及结果表现。

5. 模型训练

模型训练过程包括数据集的划分，包括训练集，验证集及测试集，过程中还要进行模型的调整，例如会用动交叉验证来最大化模型的准确预测训练。还会设计到模型参数调整。最终通过模型表现来回到步骤一进行循环迭代更新，直到达到理想目标为止。

Reference

[1] Linear Regression Models http://www.statisticssolutions.com/what-is-linear-regression/

[2] Support Vector Machine (SVM) https://en.wikipedia.org/wiki/Support_vector_machine

[3] Regression Tree https://en.wikipedia.org/wiki/Decision_tree

[4] Gaussain Process Regression Models https://en.wikipedia.org/wiki/Gaussian_process

[5] Ensembles of Trees https://en.wikipedia.org/wiki/Ensemble_learning

[6] Mean Absolute Error https://en.wikipedia.org/wiki/Mean_absolute_error

[7] Mean Square Error https://en.wikipedia.org/wiki/Mean_squared_error

[8] R^2 Score https://en.wikipedia.org/wiki/Coefficient_of_determination

[9] 线性回归 http://www.statisticssolutions.com/what-is-linear-regression/

[10] RMSPE https://www.kaggle.com/c/rossmann-store-sales#evaluation

[11] xgboost https://arxiv.org/abs/1603.02754

[12] lightgbm
https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf