机器学习纳米学位报告

毕业项目

林叶

2017 年 02 月 22 日

I. 问题的定义

项目概述

这个项目为 Kaggle 的一个比赛项目。网址为
https://www.kaggle.com/c/rossmann-store-sales

Rossmann 有超过 3000 加分店，遍布欧洲 7 个国家。商店销售受许多因素的影响，包括促销，竞争，学校和州的节假日，季节性和地方性等情况都会影响销售情况的预测。这个项目希望系统可以通过帮助 Rossmann 创建一个强大的预测模型帮助更好的管理药品情况。

项目选择数据集是 Kaggle 的（Forecast Rossmann Store Sales）。采用项目中的 train.csv 作为训练数据。训练完成后可以使用 test.csv 进行预测。并可以提交到 Kaggle 中进行模型评估。

问题陈述

项目是选择 Kaggle 提 Rossmann 药店的供数据集。根据药店信息对药店数据进行分析与预测。原始数据包括内容有:店 ID，周几，日期，客户数量，是否

开业，是否有促销，是否有节假日，是否是学校假日。通过 store.csv 可以获取商店的扩展数据。

预测需要合理根据上述信息进行分析与建模。之后通过训练结果，可以将信息输入到模型中进行预测。看预测结果与实际结果是否接近。

评价指标

通过 Kaggle 官网的 RMPSE 进行评分。

Kaggle 采用 RMPSE 模型进行评分。RMPSE 是 Root Mean Square Percentage Error (RMSPE)。计算公式为：

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

其中 yi 表示单个日子上的单个商店的销售额，y^i 表示相应的预测。 任何有 0 个销售的日子和商店在得分时都会被忽略。

II. 分析

数据的探索

项目中原始内容为 1017209 条。

Sales 最小值：0.00

Sales 最大值：41,551.00

Sales 均值: 5,773.82

Sales 中值 5,744.00

Sales 标准差: 3,849.92

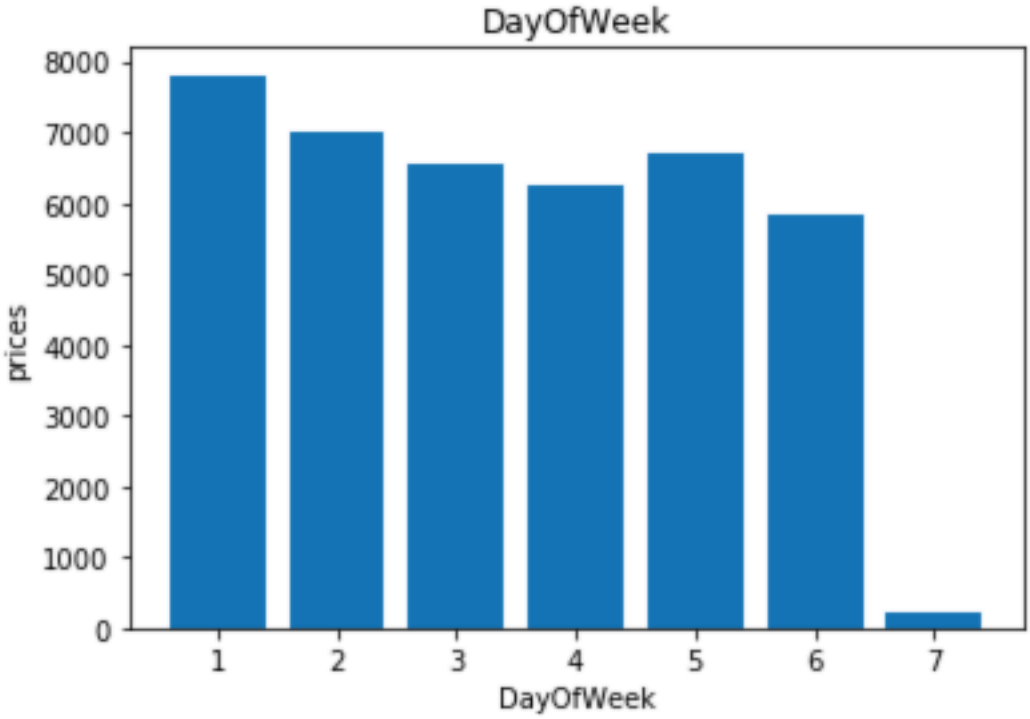
通过对 Sales 的统计，最小值是 0 最大值是 41551。平均值是 5773.82,中值是 5744，标准差是 3849.92。均值与中值非常接近。

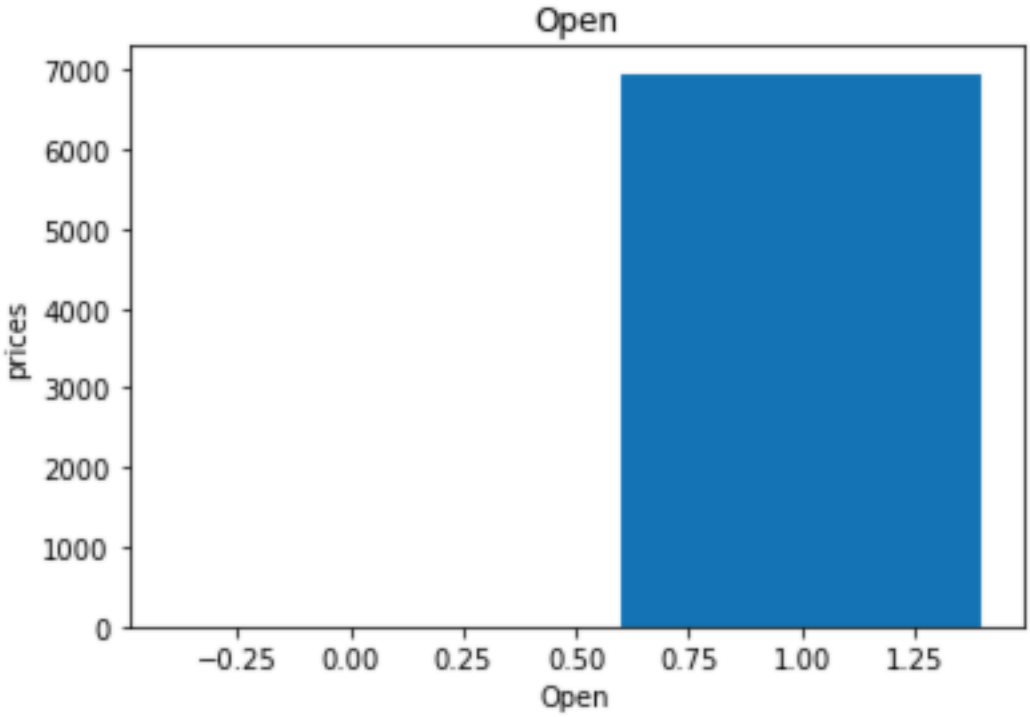
StateHoliday 数据是 0,a,b,c。将数据 a,b,c 转换成 1,2,3。生成根据特征生成图观察数据。

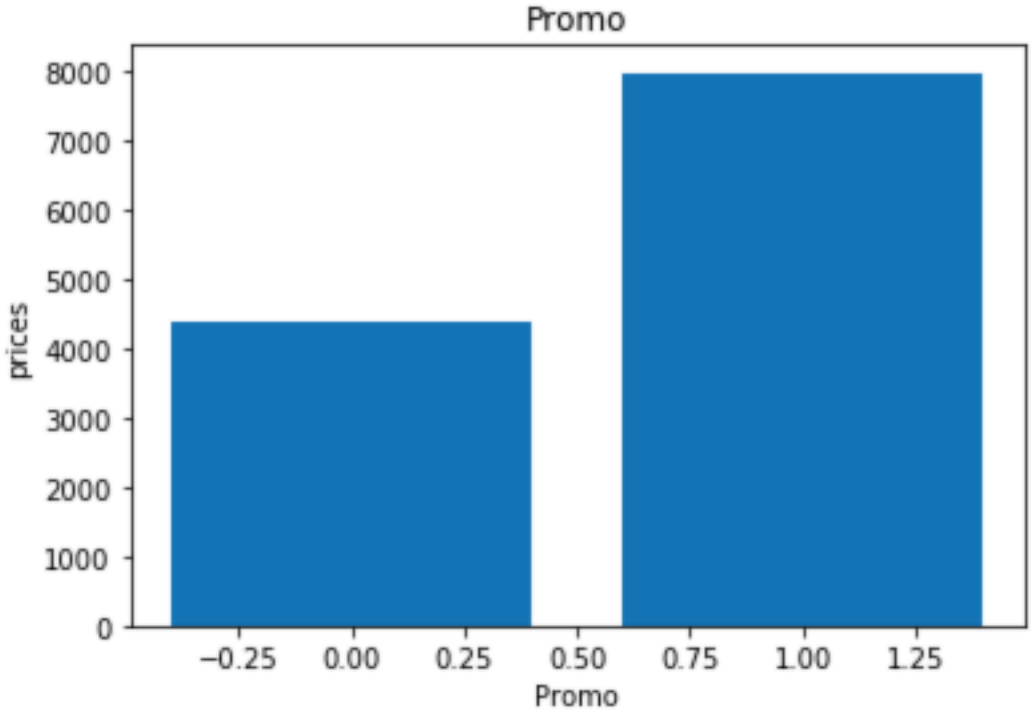
另外观察 store.csv 内容。store 包含信息如下：StoreType、Assortment、CompetitionDistance、CompetitionOpenSinceMonth、CompetitionOpenSinceYear、Promo2、Promo2SinceWeek、Promo2SinceYear、PromoInterval。

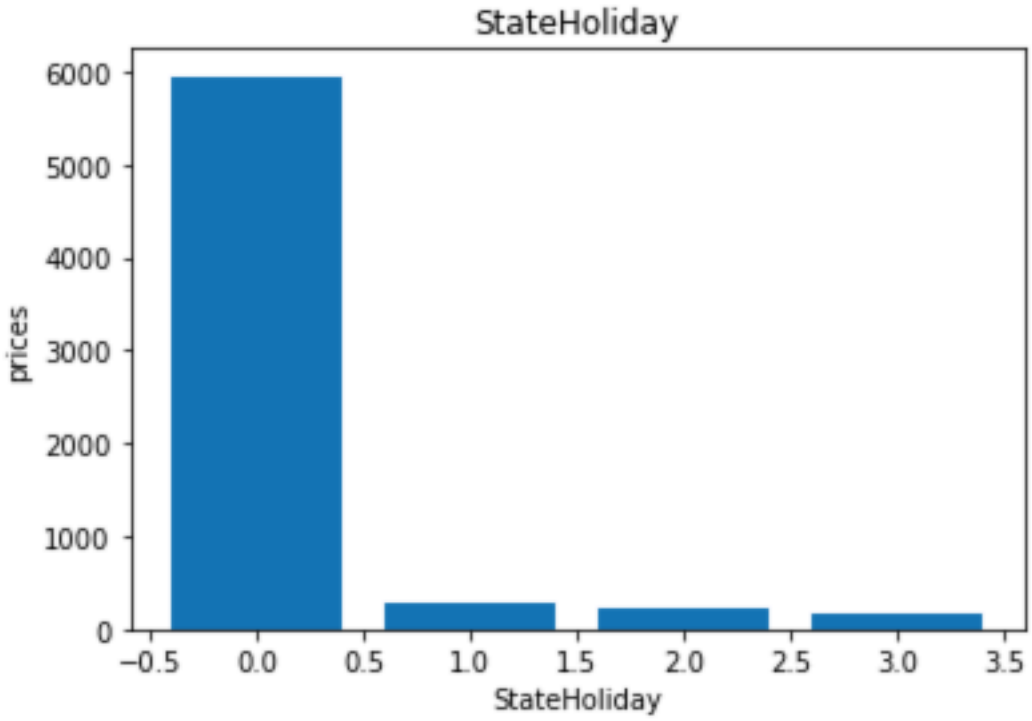
其中 CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceYear,Promo2SinceWeek, CompetitionDistance 数据不全。

观察各个特征与 prices 的均值情况如下：

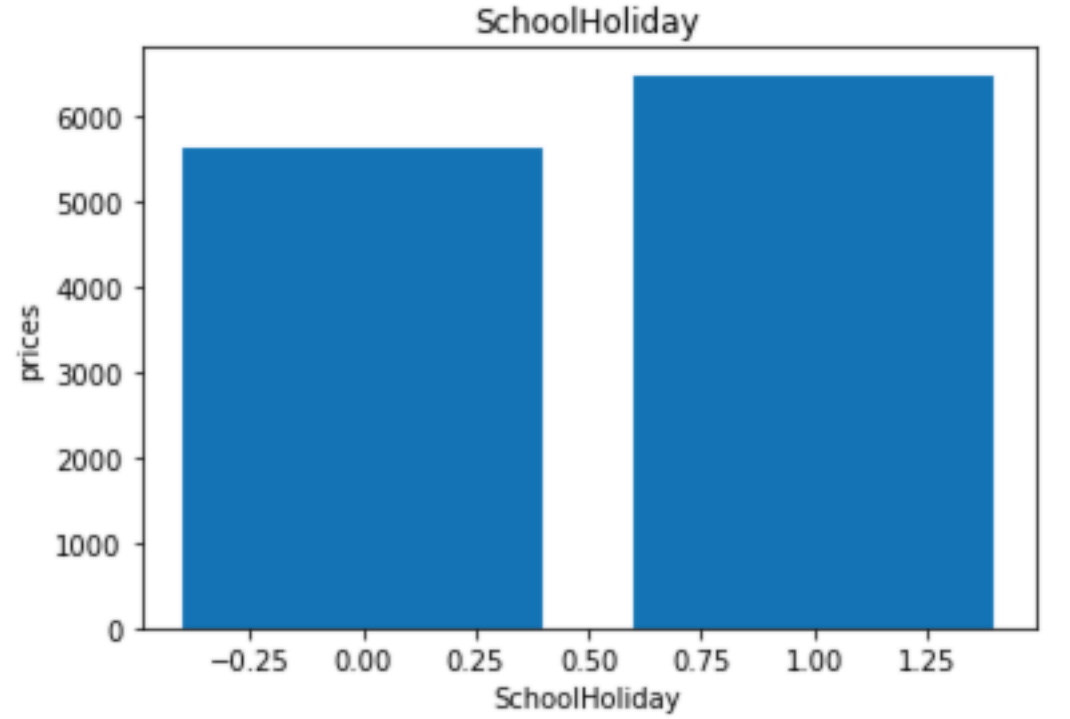

DayOfWeek
8000
7000
6000
5000
4000
3000
2000
1000
0
prices
1
2
3
4
5
6
7
DayOfWeek

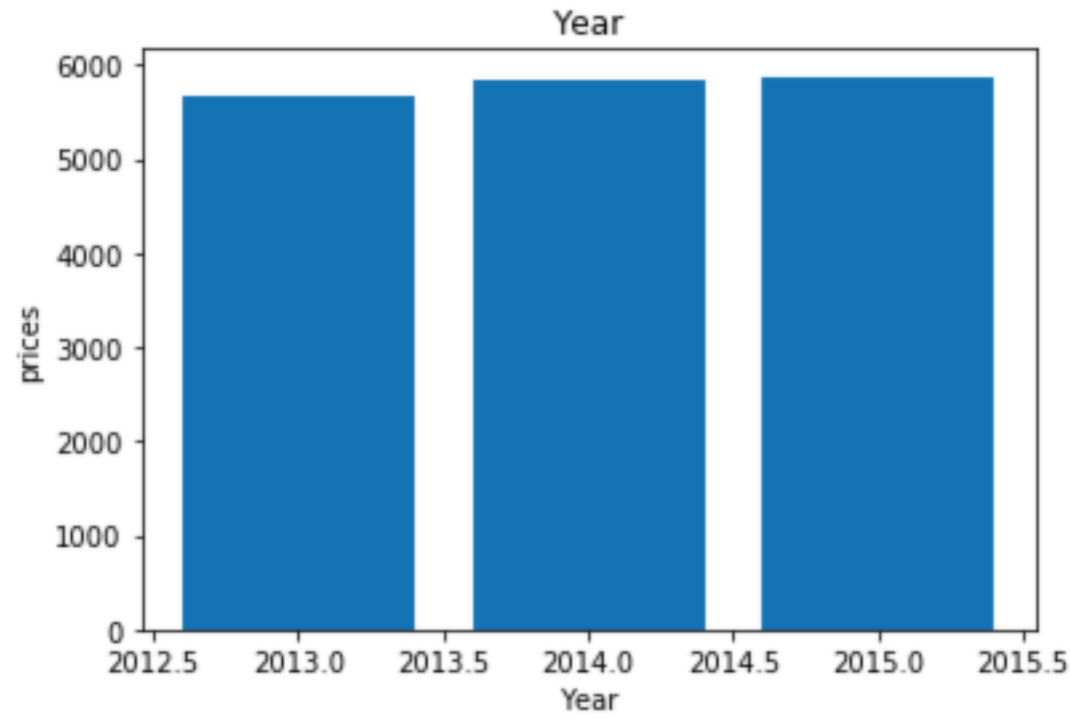

Open
7000
6000
5000
4000
3000
2000
1000
0
prices
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Open

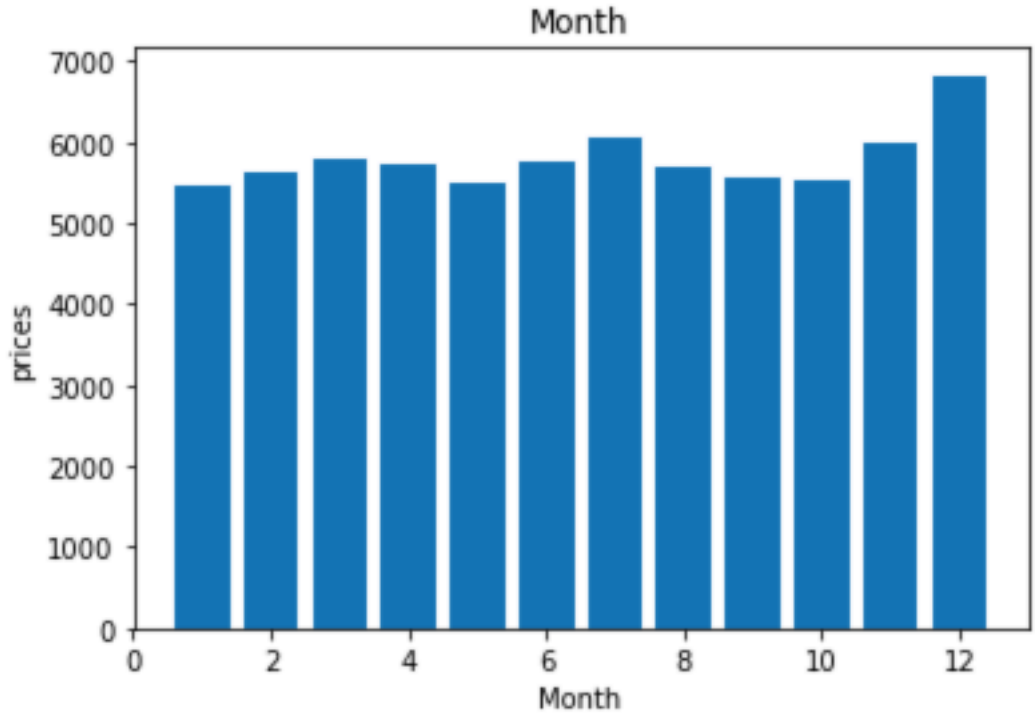

Promo
8000
7000
6000
5000
4000
3000
2000
1000
0
prices
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Promo

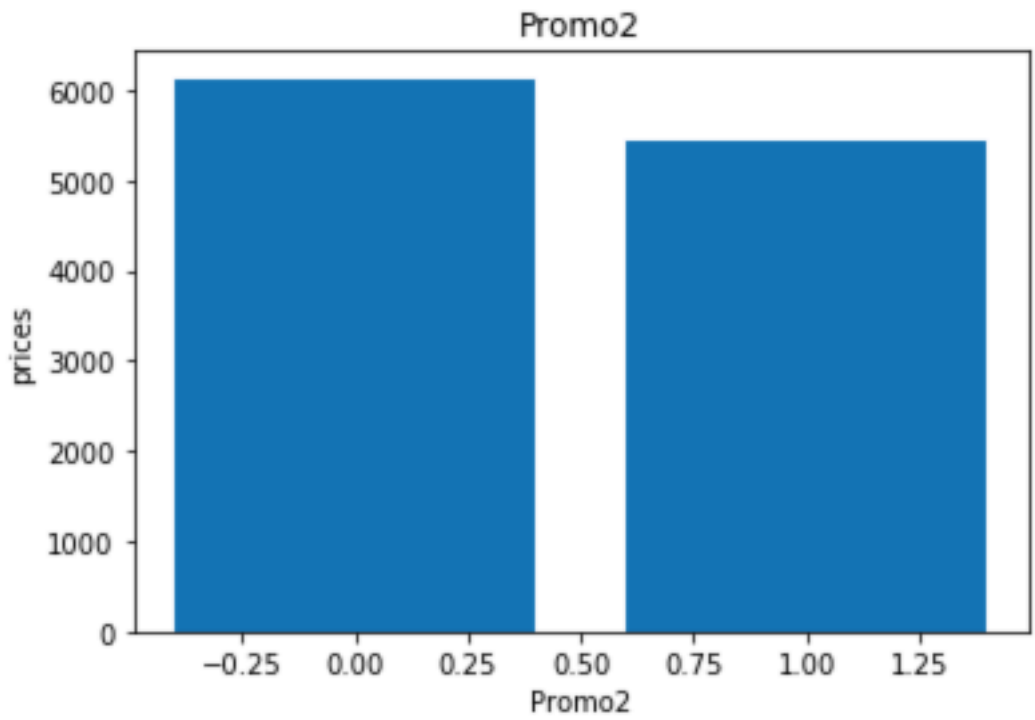

StateHoliday
6000
5000
4000
3000
2000
1000
0
prices
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
StateHoliday

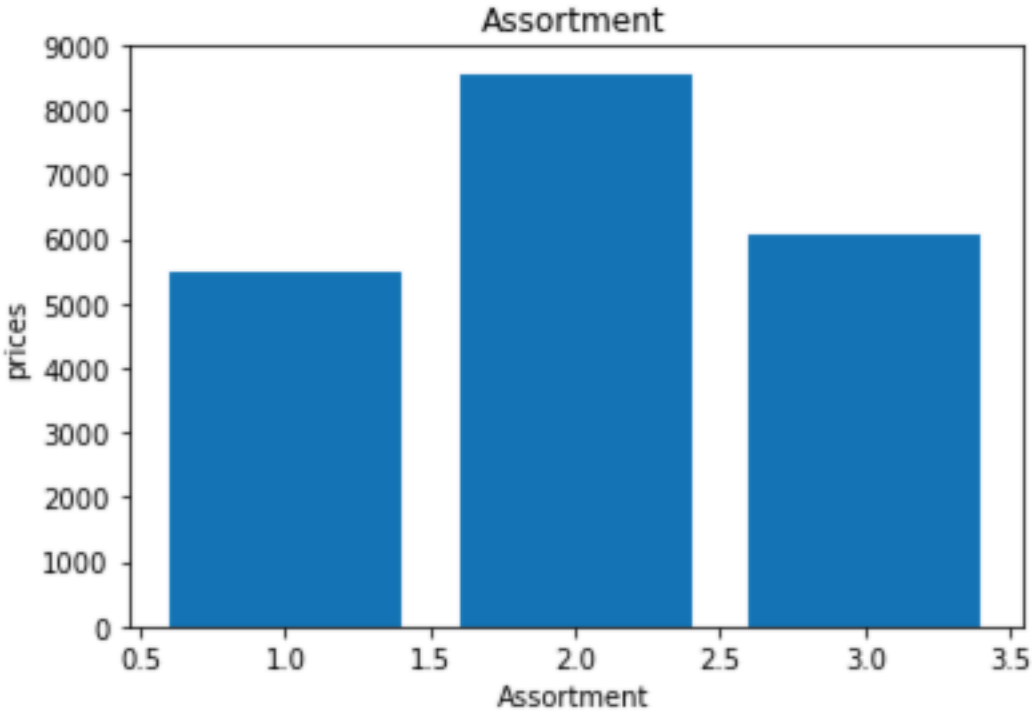
.

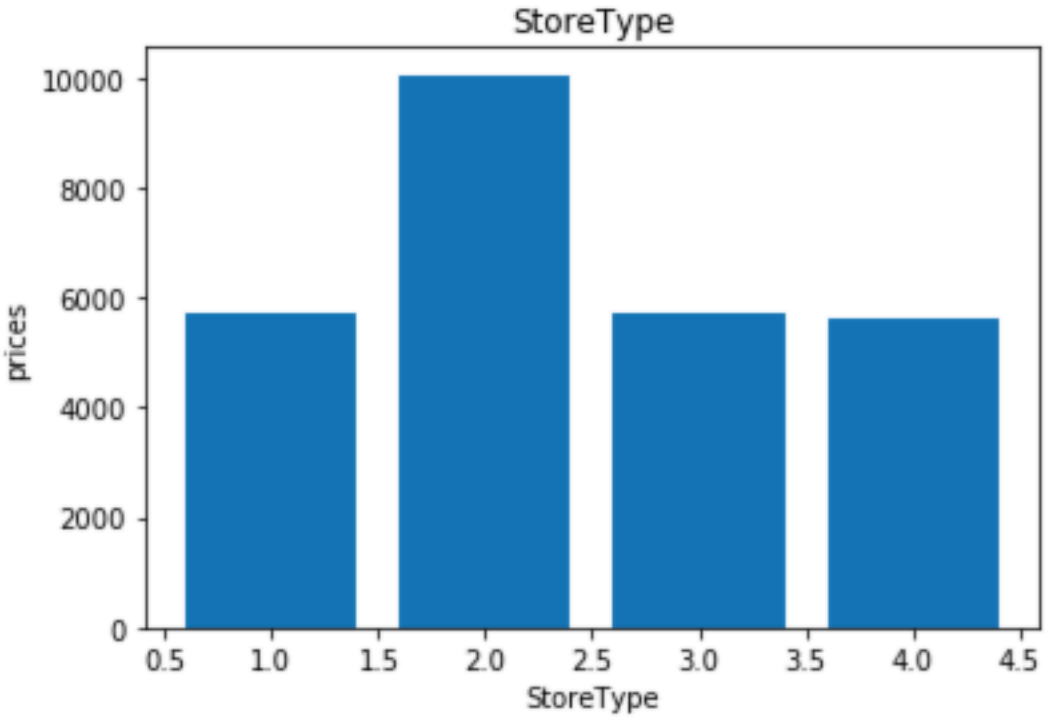

SchoolHoliday
prices
6000
5000
4000
3000
2000
1000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
SchoolHoliday


Year
prices
6000
5000
4000
3000
2000
1000
0
2012.5
2013.0
2013.5
2014.0
2014.5
2015.0
2015.5
Year


Month
7000
6000
5000
4000
3000
2000
1000
0
prices
0
2
4
6
8
10
12
Month


Promo2
6000
5000
4000
3000
2000
1000
0
prices
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Promo2





通过均值来看这些特征与结果都有一定的相关性。尝试以这些特征训练数据。

算法和技术

此项目在属于监督学习。常用的监督学习算法有回归，决策树，神经网络，svm，朴素贝叶斯等算法。而其中逻辑回归，神经网络，svm，朴素贝叶斯多用于分类。预测一般是采用线性回归，决策树回归和随机森林。所以我选用这三种算法进行测试。最终选择效果的好的作为最终模型进行结果进行提交。

线性回归：在统计学中，线性回归（Linear regression）是利用称为线性回归方程的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，大于一个自变量情况的叫做多元回归。（这反过来又应当由多个相关的因变量预测的多元线性回归区别[来源请求]，而不是一个单一的标量变量。）

在线性回归中，数据使用线性预测函数来建模，并且未知的模型参数也是通过数据来估计。这些模型被叫做线性模型。最常用的线性回归建模是给定 X 值的 y 的条件均值是 X 的仿射函数。不太一般的情况，线性回归模型可以是一个中位数或一些其他的给定 X 的条件下 y 的条件分布的分位数作为 X 的线性函数表示。像所有形式的回归分析一样，线性回归也把焦点放在给定 X 值的 y 的条件概率分布，而不是 X 和 y 的联合概率分布（多元分析领域）。

线性回归是回归分析中第一种经过严格研究并在实际应用中广泛使用的类型。这是因为线性依赖于其未知参数的模型比非线性依赖于其位置参数的模型更容易拟合，而且产生的估计的统计特性也更容易确定。

引用自：

https://zh.wikipedia.org/wiki/%E7%B7%9A%E6%80%A7%E5%9B%9E%E6%AD%B8

决策树回归：统计学,数据挖掘和机器学习中的决策树训练,使用决策树作为预测模型来预测样本的类标。这种决策树也称作分类树或回归树。在这些树的结构里，叶子节点给出类标而内部节点代表某个属性。

在决策分析中,一棵决策树可以明确地表达决策的过程。在数据挖掘中,一棵决策树表达的是数据而不是决策。本页的决策树是数据挖掘中的决策树。

引用自：

https://zh.wikipedia.org/wiki/%E5%86%B3%E7%AD%96%E6%A0%91%E5%AD%A6%E4%B9%A0

随机森林：在机器学习中，随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。 Leo Breiman 和 Adele Cutler 发展出推论出随机森林的算法。而"Random Forests"是他们的商标。这个术语是 1995 年由贝尔实验室的 Tin Kam Ho 所提出的随机决策森林（random decision forests）而来的。这个方法则是结合 Breimans 的"Bootstrap aggregating"想法和 Ho 的"random subspace method" 以建造决策树的集合。

引用自：

https://zh.wikipedia.org/wiki/%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97

通过上面几个解释我们可以了解到几个有

基准模型

一般作为预测也可以中值或者均值，这两个值在之前分析中非常接近。分别用这两个值作为结果在 Kaggle 提交测试。

采用所有的 Store 做均值和中值的测试结果分数分别是 0.40815，0.40606。

针对每个 Store 做均值和中值的测试结果分数分别是 0.25278，0.24813。

最终基准值选用对全体 Store 做的均值的评分 0.40815。

III. 方法

数据预处理

销售另外和日期没有直接关系，所以去除了日期。另外查看 test.csv 中没有 Customers 数据，所以去除 Customers 数据。

test.cvs.中 Open 数据可能是空。

	A	B	C	D	E	F	G	H
	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
	480	622	4	2015/9/17		1	0	0
	1336	622	3	2015/9/16		1	0	0
	2192	622	2	2015/9/15		1	0	0
	3048	622	1	2015/9/14		1	0	0
	4760	622	6	2015/9/12		0	0	0
	5616	622	5	2015/9/11		0	0	0
	6472	622	4	2015/9/10		0	0	0
	7328	622	3	2015/9/9		0	0	0
	8184	622	2	2015/9/8		0	0	0
	9040	622	1	2015/9/7		0	0	0
	10752	622	6	2015/9/5		0	0	0

因为 Open 没有准确信息，所以根据 Promo 设置这些数据的 Open，因为 Promo 的时候正常应该是开店。所以 Promo 是 1 就将 Open 设置为 1，Promo 是 0 的时候就将 Open 设置为 0。此部分总共有 11 条数据，而测试数据总量为 41088 条，异常数据量占的比重很小。其实也设置成什么也不会对结果造成大影响，并不用特别在意。

执行过程

- 读取 train.csv。
- 整理数据替换非字符指标。
- 生成训练数据。
- 生成多个回归方法进行测试。使用了 LinearRegression、RandomForestRegressor、DecisionTreeRegressor。
- 采用搜寻网络进行参数的自动优化。
- 根据结果不断对手动对结果进行优化。

完善

第 1 次训练：拿原始数据测试。训练结果如下。

模型名称	Kaggle 得分	本地 RMSPE 得分
LinearRegression	0.76301	0.551757814743
DecisionTreeRegressor	0.40291	0.336798060153
RandomForestRegressor	0.39641	0.336453473775

第 2 次训练：拿全部数据进行学习训练交叉训练。训练结果如下。

模型名称	Kaggle 得分	本地 RMSPE 得分
LinearRegression	0.45592	0.551933640687
DecisionTreeRegressor	0.41490	1.31691570329
RandomForestRegressor	0.53377	0.406010759714

IV. 结果

采用了 LinearRegression、RandomForestRegressor、

DecisionTreeRegressor。三种模型进行预测。三个模型最终的最好成绩如下：

模型名称	Kaggle 得分
LinearRegression	0.76301
DecisionTreeRegressor	0.41490
RandomForestRegressor	0.53377

其中 LinearRegression 速度最快，其次是 DecisionTreeRegressor，

RandomForestRegressor 效率最差。

另外从几个的得分情况分析，LinearRegression 等分最高。但是

LinearRegression 对特征的敏感性比较强，另外从最终数据来说也并符合线性

情况，所以最终随机森林模型。

合理性分析

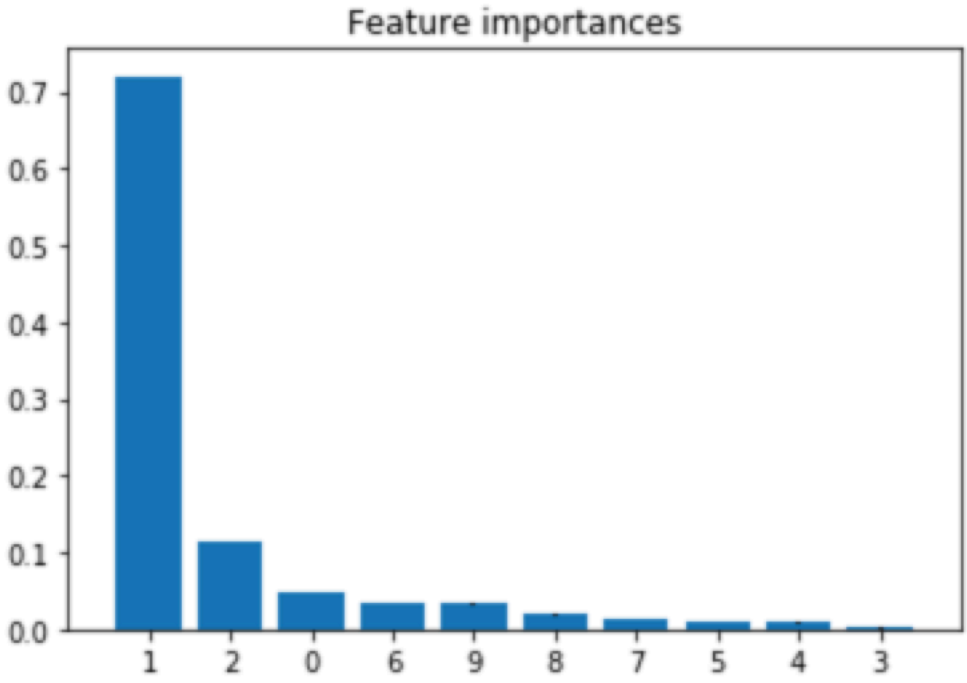
三个模型成绩比较接近。都比均值的基准模型 0.53377 稍微好一些。从目前情况看可以解决一部分问题。

V. 项目结论

结果可视化

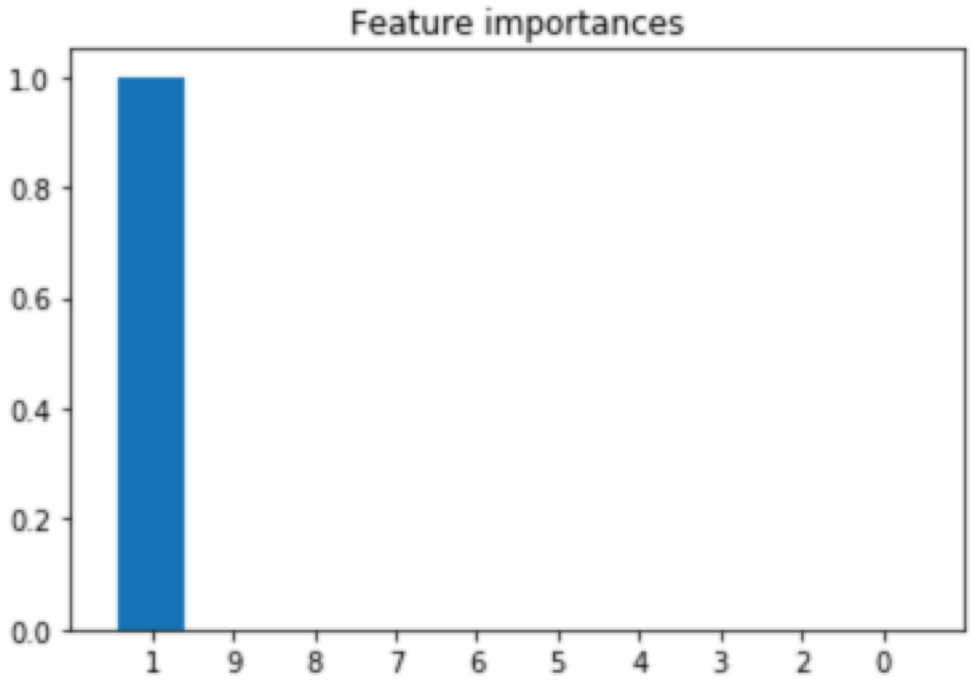
未优化前随机森林的指标重要度柱状图如下。

```
Feature ranking:
1. feature 1 DayOfWeek (0.719652)
2. feature 2 Open (0.113780)
3. feature 0 Promo (0.047766)
4. feature 6 StateHoliday (0.035285)
5. feature 9 SchoolHoliday (0.032428)
6. feature 8 Year (0.018974)
7. feature 7 Month (0.013875)
8. feature 5 Promo2 (0.009048)
9. feature 4 Assortment (0.008060)
10. feature 3 StoreType (0.001132)
```



优化后变为

```
Feature ranking:
1. feature 1 DayOfWeek (1.000000)
2. feature 9 Open (0.000000)
3. feature 8 Promo (0.000000)
4. feature 7 StateHoliday (0.000000)
5. feature 6 SchoolHoliday (0.000000)
6. feature 5 Year (0.000000)
7. feature 4 Month (0.000000)
8. feature 3 Promo2 (0.000000)
9. feature 2 Assortment (0.000000)
10. feature 0 StoreType (0.000000)
```



可以看出主要和 DayOfWeek 有主要关系

对项目的思考

目前来看线性回归算法的和特征选择关系比较大。

随机森林是一种特殊的决策树回归。

在这个项目中随机森林效果比决策树药好。

决策树回归效率比随机森林高。如果效果接近时选择决策树回归会更好一些。

从后面结果来看，应该从需求上更加的了解业务会对需求比较有帮助，另外开始应该对数据进行分析。

需要作出的改进

通过整个项目考虑一些通用方法。在特征不多并且可以方便验证的情况，可以编写程序自动优化选用特征。这样应该可以获取比较好的优化效果与效率。另外可以考虑先 store 进行分类，分类后再进行预测。这样可能准确性会更高。另外，针对各种指标结合实际意义进行整理生成新的指标。另外使用网络搜索时可以结合实际情况使用自己的程序。这样会可能优化效果会更好。