Rossmann商场销售数据预测项目

开题报告

程雪颖 June 4th,2019

1.问题描述

这个问题源自于KAGGLE的一个竞赛项目,项目的情况是根据ROSSMANN连锁店的1115个店铺的销售数据和店铺推广，竞争对手等信息的数据，这些数据是1115个店铺在2015年8月1日之前的1017209条销售数据，来预测未来六周的销售情况。

本问题是一个典型的监督学习的回归问题，项目中给出了一组关于ROSSMANN商店在两年时间的销售数据和一些特征数据，其中特征数据如商店的种类，竞争对手的开业时间和距离，是否推广等等信息属于自变量，具体的销售数据属于因变量。

2.项目背景

有监督学习通过训练样本数据得到一个模型，然后用这个模型进行推理。如果推理预测的是一个离散值，则称为分类问题，比如银行通过用户数据判断是够给一个申请人批放贷款;如果预测的是一个连续值，则被称为是回归问题，例如气象台通过空气湿度等参数预测降雨量。

监督学习是机器学习中最早的一类，例如线性判别分析（LDA）[1]，逻辑回归[2]，感知器模型[3]（1958年），kNN算法[7]（1967年）在有机器学习这个概念之前就已经有了，并且在很多领域都有了广泛的应用。

从1980年开始，机器学习的发展带动了大量的算法，这些算法也被分为监督学习，非监督学习和强化学习。能够应用与监督学习的回归算法有非常多，比如线性回归，决策树，逻辑回归模型等等，还有近年在竞赛中屡屡夺冠的XGBOOST,LIGHTGBM等基于决策树的算法。

数据和输入

数据集来自于KAGGLE,这个数据集是ROSSMANN店的销售数据和店铺信息的数据集，目标是预测未来六周的销售情况。 提供的销售数据有'推广情况'，'竞争对手情况','假期','季节'和地址。 原始数据表有： train.csv test.csv store.csv

这个项目来自于KAGGLE竞赛平台， 提供了1115个ROSSMANN商店的销量数据。 文件

- train.csv - 包含销量的历史数据 • 字段:
    - Store: 每个店的ID
    - Open:是否 开业
    - Sales:营业额
    - Customers：顾客数目
    - Date:日期
    - DayOfWeek：星期几
    - Promo:指示商店当天是否正在运行促销
    - StateHoliday:国家假日
    - SchoolHoliday:学校假日
- test.csv - 用于预测的数据

- store.csv - 店铺的相关信息
"Store","StoreType","Assortment","CompetitionDistance","CompetitionOpenSinceMonth","CompetitionOpenS

    - Store:每个店的 ID
    - StoreType:店的类型
    - Assortment: 分类级别:a = 基本,b = 额外,c = 扩展
    - CompetitionDistance:到最近的竞争对手商店的距离(以米为单位)CompetitionOpenSinceMonth:竞争对手开业的时间
    - CompetitonOpenSinceYear:竞争对手开业的年份
    - Promo2:是一些商店的连续和连续促销:0 = 商店不参与,1 = 商店参与
    - Promo2SinceWeek:描述商店开始参与促销2 的星期
    - Promo2SinceYear:描述商店开始参与促销2 的年份
    - PromoInterval:促销间隔 - 描述开始促销2的连续间隔,命名重新启动促销的月份。例如,"2 月、5 月、8 月、11 月"是指该商店的任意年份的 2 月、5 月、8 月、11 月的每一轮

- Train.csv和store.csv两个数据可以被结合起来，以id为索引，形成一个大数据表，基于这个大的数据集，去掉customers这个test数据集中没有 的参数，对其他的特征进行选择和筛选，给予这个数据集进行分析，以sales为结果变量。

In [4]:

```
import pandas as pd
import numpy as np
train = pd.read_csv('train.csv',parse_dates=[2],low_memory=False)
test = pd.read_csv('test.csv',parse_dates=[3],low_memory=False)
store = pd.read_csv("store.csv")
sample_submission = pd.read_csv("sample_submission.csv")
```

In [2]:

```
train.shape
```

Out[2]:

```
(1017209, 9)
```

In [3]:

```
test.shape
```

Out[3]:

```
(41088, 8)
```

In [4]:

```
store.shape
```

Out[4]:

```
(1115, 10)
```

In [5]:

```
store.columns
```

Out[5]:

```
Index(['Store', 'StoreType', 'Assortment', 'CompetitionDistance',
       'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2',
       'Promo2SinceWeek', 'Promo2SinceYear', 'PromoInterval'],
      dtype='object')
```

In [6]:

```
#查看测试集
test.head().append(test.tail())
```

Out[6]:

	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
0	1	1	4	2015-09-17	1.0	1	0	0
1	2	3	4	2015-09-17	1.0	1	0	0
2	3	7	4	2015-09-17	1.0	1	0	0
3	4	8	4	2015-09-17	1.0	1	0	0
4	5	9	4	2015-09-17	1.0	1	0	0
41083	41084	1111	6	2015-08-01	1.0	0	0	0
41084	41085	1112	6	2015-08-01	1.0	0	0	0
41085	41086	1113	6	2015-08-01	1.0	0	0	0
41086	41087	1114	6	2015-08-01	1.0	0	0	0
41087	41088	1115	6	2015-08-01	1.0	0	0	1

评估指标

这个KAGGLE项目要求检验模型和解决方案采用RMPSE这个评估算法： RMPSE这个评估指标计算的图的结果:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i-y}{y_i})^2}$$

这个指标计算的是被预测结果和真实数据之间的差异相对与真实数据的比例，也就可以表现出衡量值和真实值之间的误差。这个评估算法的数值越接近于0，说明误差越小。

基准模型

基准模型选择决策树模型和随机数森林模型。 之所以选择决策树模型，是因为现在在KAGGLE竞赛中表现最好的算法XGBOOST和LIGHTGBM等都是基于回归树算法的，而随机森林模型也是基于决策树的一种模型，选择这两个模型作为基准模型，可以对比出使用xgboost模型的效果。

决策树出现与1980年[3]，决策树是一种基本的分类与回归方法，回归决策树主要指CART(classification and regression tree)算法[4]。

随机森林[5]出现于2001年，属于集成算法，现在还在被大规模使用。

回归树就是将特征空间的划分，采用启发式方法，每次划分逐一考察当前集合中所有特征的所有取值，根据平方误差最小化准则选择其中最优的一个作为切分点，依次将输入空间划分为两个区域，对每个区域重复划分过程，直到满足停止条件，这样就生成了一棵回归树。

项目设计

在本项目中，可以按照监督学习回归分析的方式进行：

- 1.数据探索，对数据进行研究，特别是了解数据的特征值。 通过PYTHON中的一些方法，结合可视化展示，对数据进行探索，目的是对要分析的数据有个大概的了解。弄清数据集大小，特征和样本数量，数据类型，数据的概率分布等。
- 2.数据清洗 检查数据的完整性和一致性，一般来说需要去重和处理空值。数据清洗的效果会极大的影响结果。
- 3.特征工程 合并train,csv和sales.csv数据集，对数据的特征进行分析，可以根据数据特征的状态进行ONE-HOT编码，或者是对日期信息等不能直接被模型处理的信息进行处理，尽可能提取出能够被模型有效抽取的特征，根据特征的数目，考虑是够要进行主成分分析。
- 4.数据拆分为训练集和验证集 一般讲训练集的20%提取出来作为验证集，这样可以验证模型的效果。
- 5.应用基准模型 应用随机森林和决策树模型，并且在验证集上测试得分，作为之后的参考。
- 6.应用xgboost模型，提交基准模型得到的结果到KAGGLE上作为对比。 构建默认参数的xgboost模型，作为之后调整参数的基础。
- 7.根据模型在验证集上的表现进行参数调整 对XGBOOST模型进行调参，XGBOOST模型的参数较多，在调整上有一个先后顺序，一般是按照1、最佳迭代次数：n_estimators;2.min_child_weight以及max_depth：3.调试参数：gamma：4.subsample以及colsample_bytree：5.Learning_rate。
- 8.提交KAGGLE查看模型结果

参考文献

[1] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 7 (2): 179–188.

[2] Cox, DR (1958). The regression analysis of binary sequences (with discussion). J Roy Stat Soc B. 20 (2): 215–242.

[3] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[4] Breiman, L., Friedman, J. Olshen, R. and Stone C. Classification and Regression Trees, Wadsworth, 1984.

[5] Breiman, Leo. Random Forests. Machine Learning 45 (1), 5-32, 2001.

[6] 如何对 XGBoost模型进行参数调优 https://zhidao.baidu.com/question/493269744162619412.html (https://zhidao.baidu.com/question/493269744162619412.html)