1. 整体概要
2. 问题阐述

比赛提供了俄罗斯电商公司1C从2013年1月至2015年10月的每日销售数据，包含了商品信息、店铺信息、商品分类等，要求利用这些数据预测下一个月每家商店售卖每种商品的月销量。

1. 研究意义

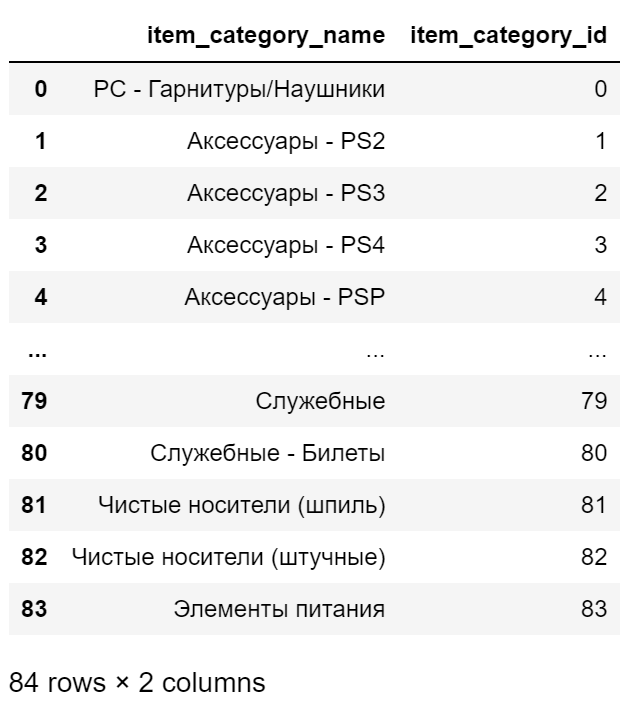
通过数据预处理和模型构建，获取训练集样本中有用的信息，对未来一个月的销售情况做出预测，能够帮助电商更好准备商品和精准营销销售，提高利润空间。

通过不断的尝试和调整，我们最终找到对我们的模型最优的参数设置，在10000余支队伍中获得前7%的成绩。

1. 研究思路

首先，我们对源数据进行预处理，清洗掉和填充异常值，并根据实际意义和信息价值进行分类和添加特征，构造完整和整齐的分类数据和计数值，对不同特征数据进行标准化之后，构建了XGBoost机器学习模型，手动尝试调整参数，根据最小均方 误差值判断预测结果准确度。 最后选取最优参数组合，得到结果。

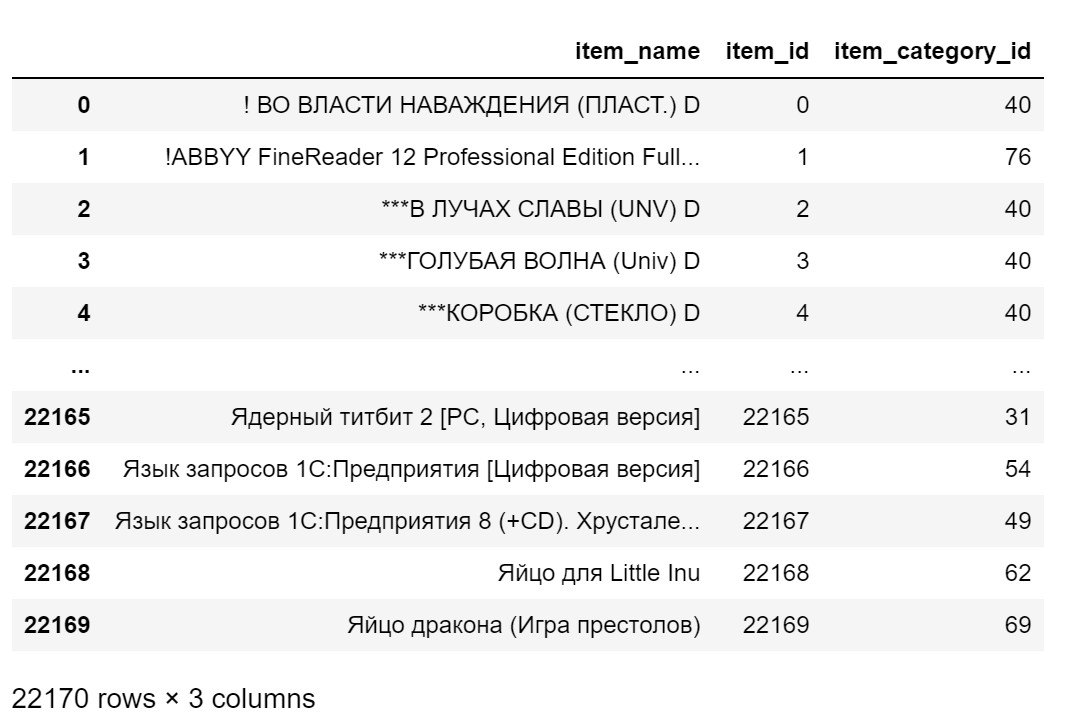
1. 数据概况
2. item\_categories.csv



图表 1

分别描述了商品分类的全称和ID，全程中分隔符“-”前是商品品类，如“书籍”“礼品”等，“-”后为商品的具体型号。共有84个商品类别，根据中文翻译的实际含义，其中存在一些相似的品类命名。

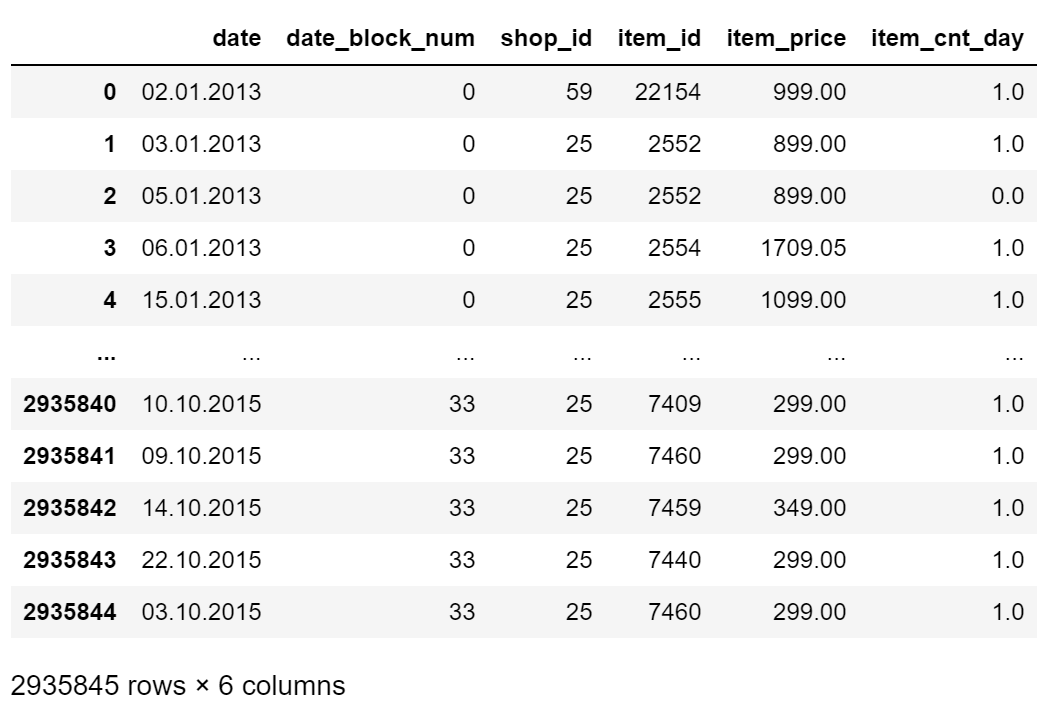
1. item.csv



图表 2

商品全程item\_name中，“！”后跟的是商品的售卖地城市，“[]”中表示的是商店名称，观察发现存在相同商店名称与不同城市的组合，认为是同一家商店在不同城市的分店。

1. salestrain.csv



图表 3

记录了每家商店每天售卖每种商品的信息和数量。

1. shop\_name.csv



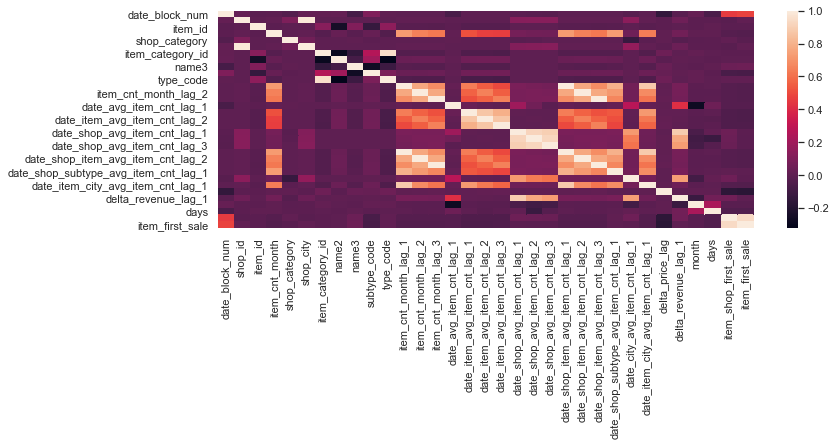
图表 4

上表截取了商品名称存在高度相似的几家店铺信息，并将0号店铺ID更改为57，1号店铺ID更改58，10号店铺ID更改为11。



图表 5

观察发现shop的命名规则为“(!+)城市+商店主营商品大类+后缀”，可借此提取新的特征。



图表 6