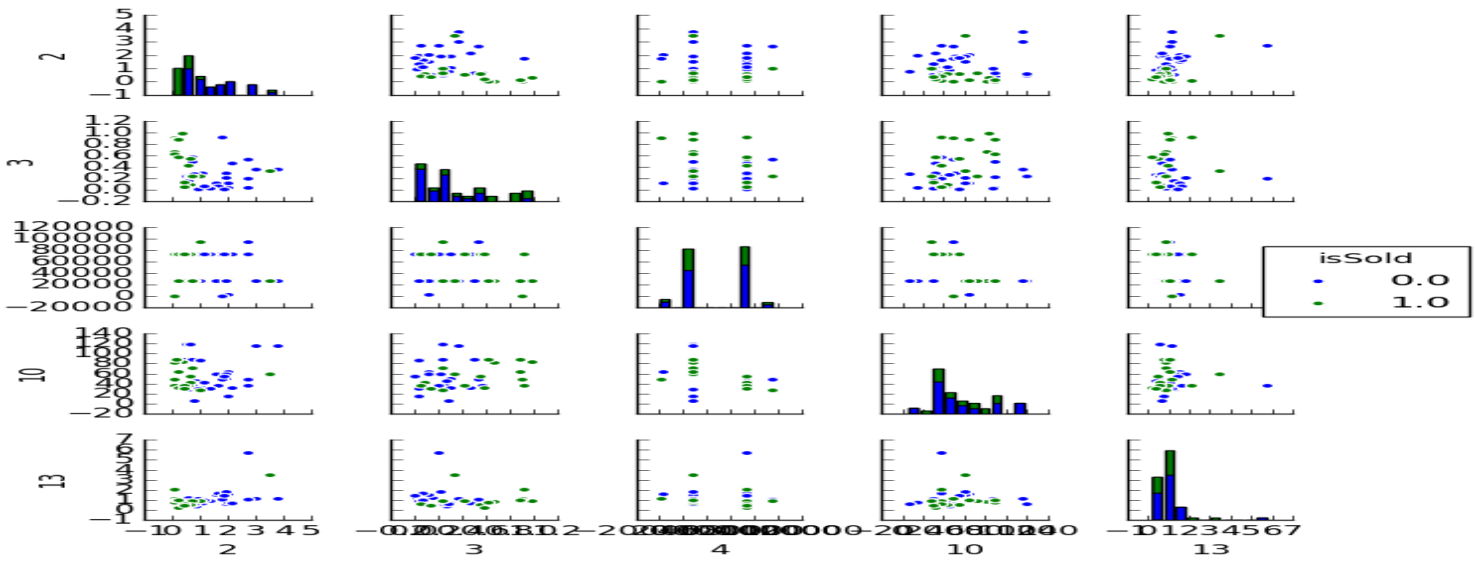
ebay 在线数据分析及预测

1 数据描述

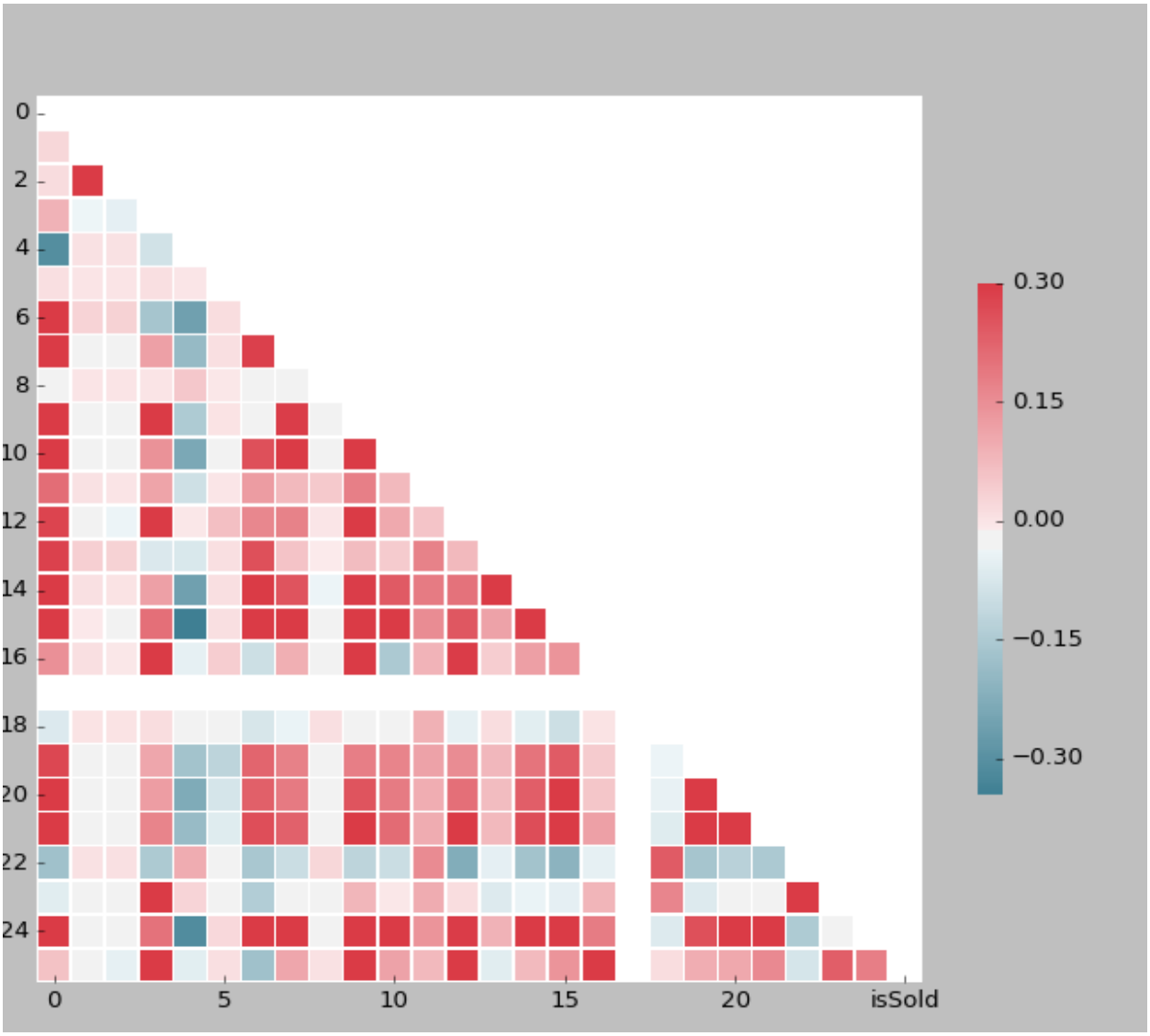
此次我们运用 ebay 在线拍卖数据的数据集，这个数据集中包括训练集，测试集，训练子集，测试子集，下表中给出了这四个文件的内容简介:

Training Set	2013 年 4 月的所有拍卖	25.8588
Test Set	2013 年 5 月第一个周的所有拍卖	3.7460
Training Subset	2013 年 4 月成功交易的所有拍卖	7.9732
Test Subset	2013 年 5 月第一周成功交易的所有拍卖	9392

数据可视化



从 2,3,10 维特征的散列图及柱状图可看出，这几个唯独并不是有很好的区分度，横纵坐标的值分别代表不同唯独之间的正负相关性。为了查看数据特征之间的相关性，及不同特征与 isSold 之间的关系，我们生成热度图来显示其两两组队之间的相关性



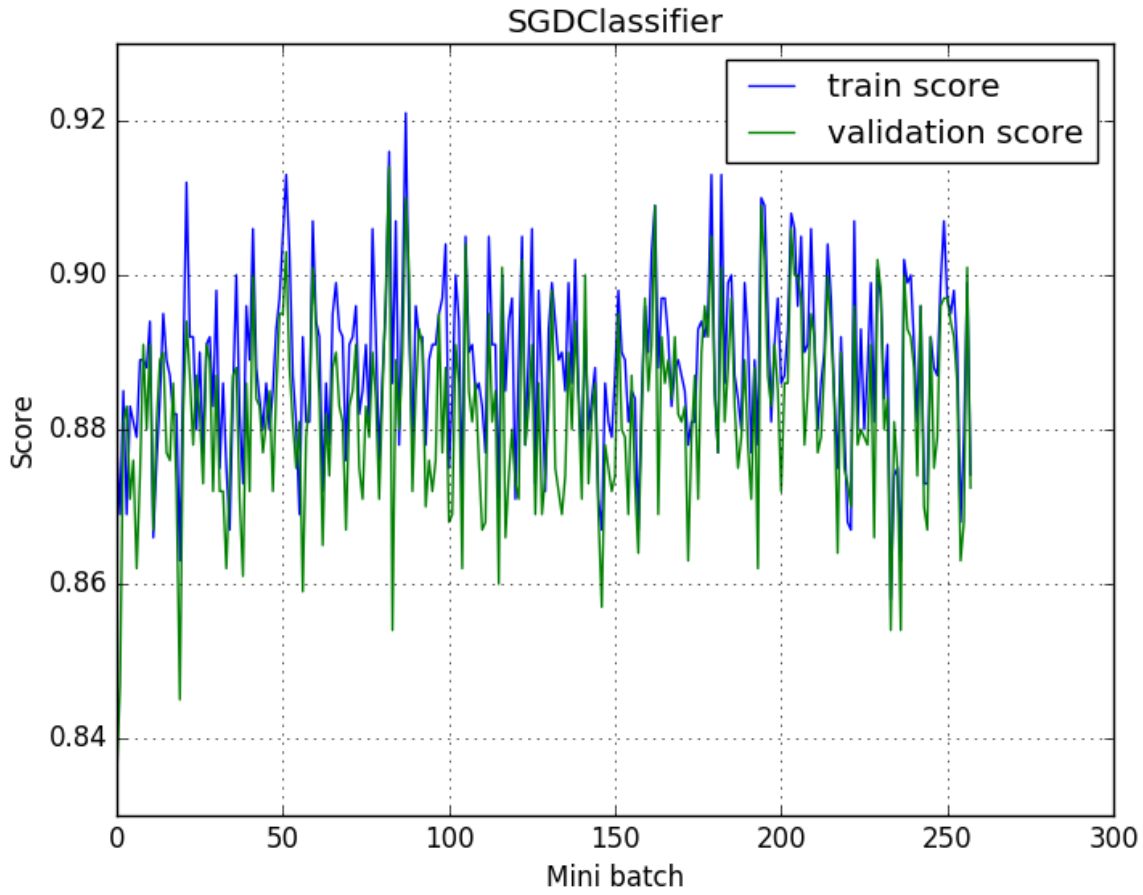
上幅图中，颜色越偏向红色的相关性越大，越偏向蓝色的相关性越小且负相关。白色即两个特征之间没有太大的相关性。通过最后一行可以看出，不同维的属性与类别 isSold 之间的关系，其中第３，９，１２，１６维特征与拍卖是否成功有很强的正相关性，其中３，９，１２，１６分别代表属性 SellerClosePercent，HitCount，SellerAuctionSaleCount，BestOffer，表示当这些属性的值越大时越有可能使拍卖成功，其中第６维特征 StartingBid 与成功拍卖 isSold 之间有较大的负相关性，可看出当拍卖投标的底线越高则这项拍卖的成功性越低。

通过这幅热度图的第一列我们还可以看出不同特征与价格 Price 之间的相关性，同样的我们可以根据这些相关性，选出比较有利于我们实现本次研究的第二个任务——拍卖价格预测的特征

2 利用数据预测拍卖是否会成功

由于我们的数据量比较大，且特征维度也不是特别少，因此我们利用机器学习提供的 SGDclassifier 来进行预测，通过梯度下降法在训练过程中优化目标函数使得预测与真实值之间的误差 loss 最小化。SGDclassifier 每次训练过程没有用到所有的训练样本，而是随机的从训练样本中选取一部分进行预测，而且 SGD 对特征值的大小比较敏感，而通过上面的数据显示，可以知道在我们的数据集里有数值较大的数据，如 Category，因此我们对数据进行预处理，使得每个属性的波动幅度不要太大，有助于训练时函数收敛。

训练结果如下图，由于 SGDclassifier 是在所有的训练样本中抽取一部分作为本次的训练集，因此在这里不适用交叉验证。



可看到 SGDclassifier 的训练效果还不错，我们也可以通过一些降维方法将数据可视化，这里不再演示。分类器训练结束后，可查看分类器在测试集上的测试效果。

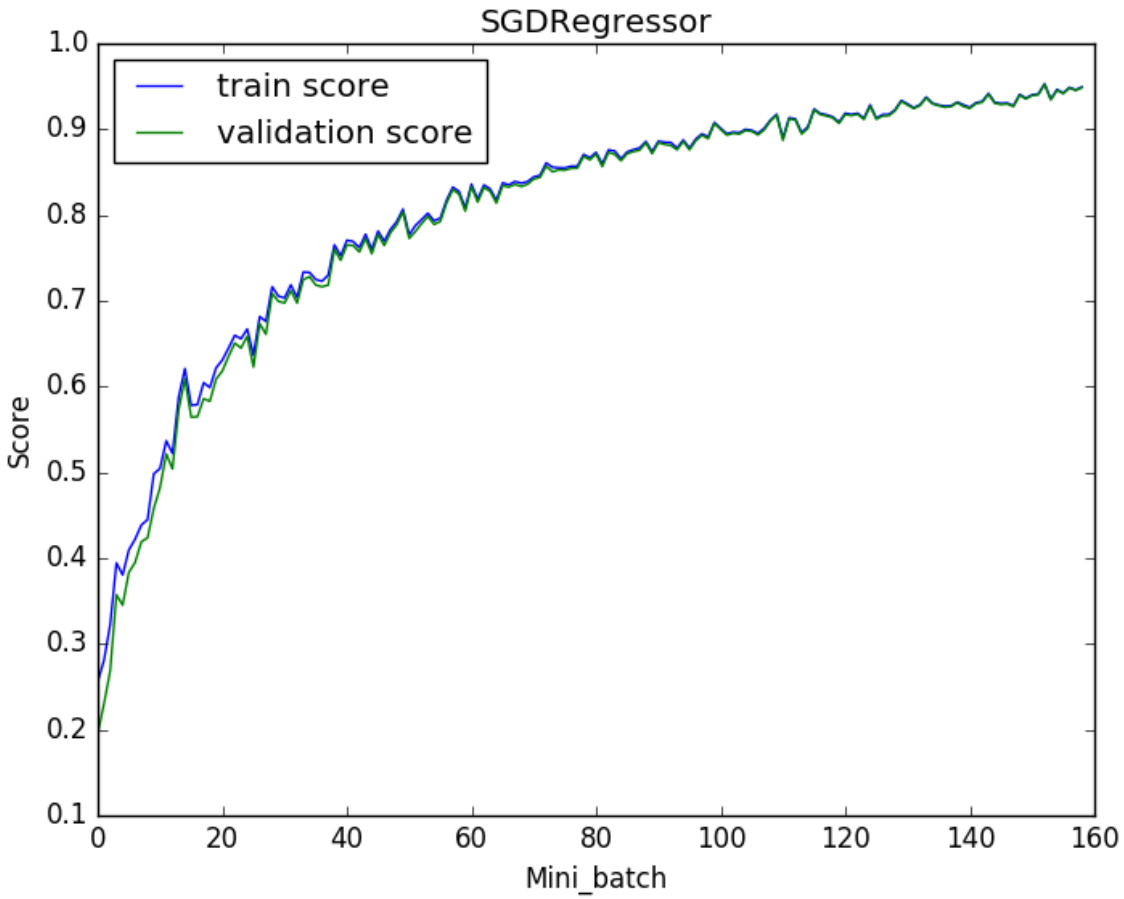
测试结果:

```
SGDClassifier training performance on testing dataset:
    Precision: 0.819
    recall: 0.737
    F1: 0.776
```

3 预测拍卖最终成交价格

由于价格的分布是一个连续的区间，因此这与预测拍卖成功是不同的，**预测价格是一个回归预测，而判断拍卖是否会成功是一个分类任务。在这里我们采取 SGDRegressor 进行回归预测。**

训练过程：



在测试集上的测试结果：

```
SGD regressor prediction result on testing data: 0.665
```

从训练过程可以看出，SGDRegressor 的回归效果还不错，因此我们没有在进一步的选择其他的模型进行尝试