|  |  |
| --- | --- |
|  |  |
| 项目报告3 |  |
|  |  |
|  | 课程：数据挖掘姓名：姜其升 |
|  | 学号：1752058指导教师：饶卫雄教授 |

[项目报告3 1](#_Toc16291)

[1场景一 5](#_Toc13504)

[1.1基本思路 5](#_Toc1374)

[1.2数据预处理 5](#_Toc23152)

[1.3 SVM机器学习模型 5](#_Toc5925)

[1.3.1预测结果 5](#_Toc15160)

[1.3.2算法性能 6](#_Toc22176)

[1.3.3 RSE 6](#_Toc27382)

[1.4随机森林机器学习模型 6](#_Toc16889)

[1.4.1预测结果 6](#_Toc27086)

[1.4.2算法性能 7](#_Toc28898)

[1.4.3 RSE 7](#_Toc4193)

[1.5 MLP机器学习模型 7](#_Toc16992)

[1.5.1预测结果 7](#_Toc11998)

[1.5.2算法性能 8](#_Toc1569)

[1.5.3 RSE 8](#_Toc32143)

[2场景二 9](#_Toc25940)

[2.1基本思路 9](#_Toc12999)

[2.2数据预处理 9](#_Toc12117)

[2.3 SVM机器学习模型 9](#_Toc29031)

[2.3.1预测结果 9](#_Toc4272)

[2.3.2算法性能 10](#_Toc7243)

[2.3.3 RSE 10](#_Toc12103)

[2.4随机森林机器学习模型 10](#_Toc15967)

[2.4.1预测结果 10](#_Toc15520)

[2.4.2算法性能 11](#_Toc29748)

[2.4.3 RSE 11](#_Toc22435)

[2.5 MLP机器学习模型 11](#_Toc27457)

[2.5.1预测结果 11](#_Toc8533)

[2.5.2算法性能 12](#_Toc27709)

[2.5.3 RSE 12](#_Toc17486)

[3场景三 13](#_Toc20784)

[3.1基本思路 13](#_Toc27901)

[3.2数据预处理 13](#_Toc3375)

[3.3 SVM机器学习模型 13](#_Toc24825)

[3.3.1预测结果 13](#_Toc23270)

[3.3.2算法性能 14](#_Toc5618)

[3.3.3 RSE 14](#_Toc4171)

[3.4随机森林机器学习模型 14](#_Toc27895)

[3.4.1预测结果 14](#_Toc18751)

[3.4.2算法性能 15](#_Toc7473)

[3.4.3 RSE 15](#_Toc26600)

[3.5 MLP机器学习模型 15](#_Toc29003)

[3.5.1预测结果 15](#_Toc8443)

[3.5.2算法性能 16](#_Toc21775)

[3.5.3 RSE 16](#_Toc15197)

[4场景四 17](#_Toc23814)

[4.1基本思路 17](#_Toc9663)

[4.2数据预处理 17](#_Toc39)

[4.3 SVM机器学习模型 17](#_Toc23073)

[4.3.1预测结果 17](#_Toc15312)

[4.3.2算法性能 18](#_Toc1630)

[4.3.3 RSE 18](#_Toc23703)

[4.4随机森林机器学习模型 18](#_Toc27185)

[4.4.1预测结果 18](#_Toc15044)

[4.4.2算法性能 19](#_Toc7767)

[4.4.3 RSE 19](#_Toc28524)

[4.5 MLP机器学习模型 19](#_Toc21938)

[4.5.1预测结果 19](#_Toc28545)

[4.5.2算法性能 20](#_Toc28831)

[4.5.3 RSE 20](#_Toc9306)

[5对比分析 21](#_Toc15320)

[5.1预测结果对比 21](#_Toc5501)

[5.2性能对比 21](#_Toc16288)

[5.3 RES 22](#_Toc5553)

[5.4对比小结 25](#_Toc26886)

## 1场景一

### 1.1基本思路

在仅使用b.i特征的情况下，针对训练数据中商品每天的当日销量为目标特征、其他特征（即历史信息）均为属性特征，利用SVM、随机森林、MLP等3个方法进行建模，预测测试数据中某商品对应日期当日（标记为d’）至第6日（d’+6）共计7天的每日销量，可考虑如下算法：首先完成商品d’当日的销量预测，然后利用该预测销量更新上述b)的相关特征，继续预测d‘+1当日销量。。。重复该步骤，直至完成第6日（d’+6）当日销量预测。分别选取商品80%和20%的时序数据作为训练和测试数据。

### 1.2数据预处理

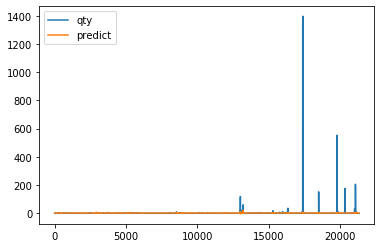
将数据reshape并作标准化处理。

1. x\_i = np.reshape(x\_i, (-1, 15))
2. y\_i = np.reshape(y\_i, (-1, 1))
4. **from** sklearn.preprocessing **import** StandardScaler
5. sc\_X = StandardScaler()
6. sc\_y = StandardScaler()
7. X\_i = sc\_X.fit\_transform(x\_i)
8. Y\_i = sc\_y.fit\_transform(y\_i)

### 1.3 SVM机器学习模型

#### 1.3.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/1/svr.csv中。



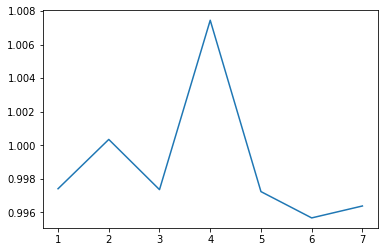
#### 1.3.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 240.44 | 34.77 |

#### 1.3.3 RSE

整体的RSE=0.9976846464361963

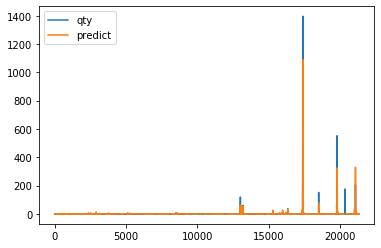
多步预测RSE



### 1.4随机森林机器学习模型

#### 1.4.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/1/rfr.csv中。



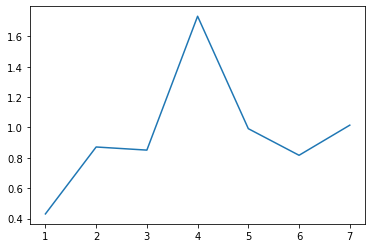
#### 1.4.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 43.09 | 125.18 |

#### 1.4.3 RSE

整体的RSE=0.9009606998114897

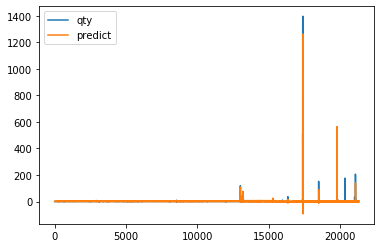
多步预测RSE



### 1.5 MLP机器学习模型

#### 1.5.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/1/mlp.csv中。



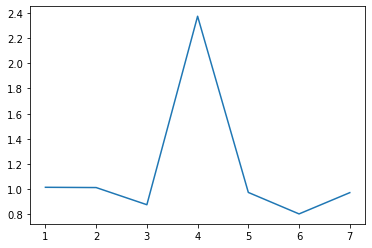
#### 1.5.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 24.76 | 37.07 |

#### 1.5.3 RSE

整体的RSE=1.0703638155102346

多步预测RSE



## 2场景二

### 2.1基本思路

在使用b.i+b.iv特征的情况下，针对训练数据中商品每天的当日销量为目标特征、其他特征（即历史信息）均为属性特征，利用SVM、随机森林、MLP等3个方法进行建模，预测测试数据中某商品对应日期当日（标记为d’）至第6日（d’+6）共计7天的每日销量，可考虑如下算法：首先完成商品d’当日的销量预测，然后利用该预测销量更新上述b)的相关特征，继续预测d‘+1当日销量。。。重复该步骤，直至完成第6日（d’+6）当日销量预测。分别选取商品80%和20%的时序数据作为训练和测试数据。

### 2.2数据预处理

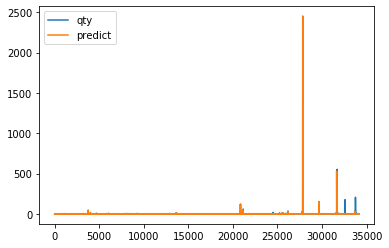
将数据reshape并作标准化处理。

1. x\_i = np.reshape(x\_i, (-1, 15+9))
2. y\_i = np.reshape(y\_i, (-1, 1))
4. **from** sklearn.preprocessing **import** StandardScaler
5. sc\_X = StandardScaler()
6. sc\_y = StandardScaler()
7. X\_i = sc\_X.fit\_transform(x\_i)
8. Y\_i = sc\_y.fit\_transform(y\_i)

### 2.3 SVM机器学习模型

#### 2.3.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/2/svr.csv中。



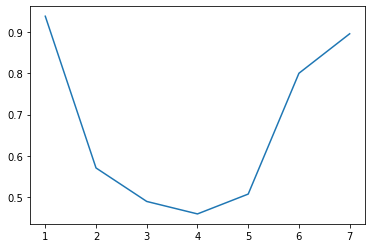
#### 2.3.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 274.45 | 28.80 |

#### 2.3.3 RSE

整体的RSE=0.654334116210174

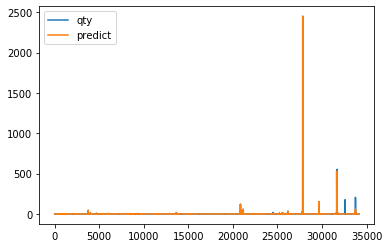
多步预测RSE



### 2.4随机森林机器学习模型

#### 2.4.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/2/rfr.csv中。



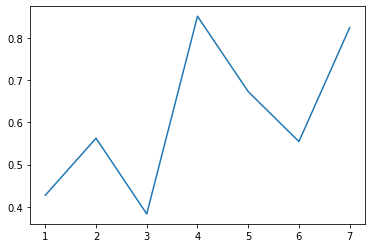
#### 2.4.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 49.77 | 101.11 |

#### 2.4.3 RSE

整体的RSE=0.5729715336751021

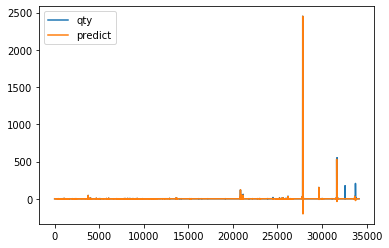
多步预测RSE



### 2.5 MLP机器学习模型

#### 2.5.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/2/mlp.csv中。



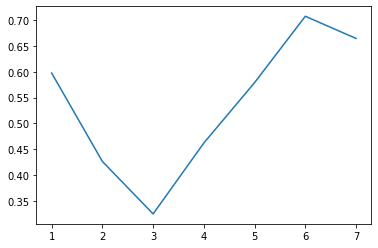
#### 2.5.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 38.89 | 30.44 |

#### 2.5.3 RSE

整体的RSE=0.4981346901772707

多步预测RSE



## 3场景三

### 3.1基本思路

在使用b.i+b.ii+b.iii+b.iv特征的情况下，针对训练数据中商品每天的当日销量为目标特征、其他特征（即历史信息）均为属性特征，利用SVM、随机森林、MLP等3个方法进行建模，预测测试数据中某商品对应日期当日（标记为d’）至第6日（d’+6）共计7天的每日销量，可考虑如下算法：首先完成商品d’当日的销量预测，然后利用该预测销量更新上述b)的相关特征，继续预测d‘+1当日销量。。。重复该步骤，直至完成第6日（d’+6）当日销量预测。分别选取商品80%和20%的时序数据作为训练和测试数据。

### 3.2数据预处理

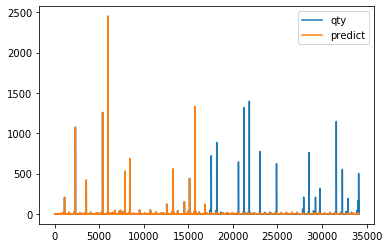
将数据reshape并作标准化处理。

1. x\_i = np.reshape(x\_i, (-1, len(feature\_name)))
2. y\_i = np.reshape(y\_i, (-1, 1))
4. **from** sklearn.preprocessing **import** StandardScaler
5. sc\_X = StandardScaler()
6. sc\_y = StandardScaler()
7. X\_i = sc\_X.fit\_transform(x\_i)
8. Y\_i = sc\_y.fit\_transform(y\_i)

### 3.3 SVM机器学习模型

#### 3.3.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/3/svr.csv中。



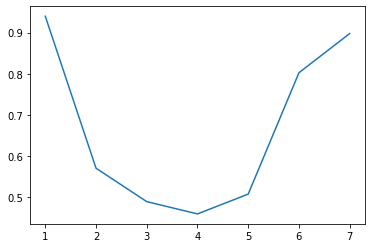
#### 3.3.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 618.97 | 273.02 |

#### 3.3.3 RSE

整体的RSE=0.6552643940758035

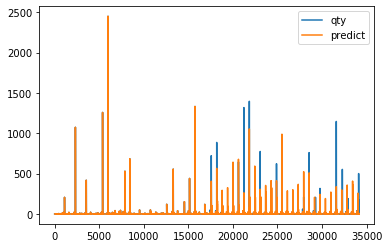
多步预测RSE



### 3.4随机森林机器学习模型

#### 3.4.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/3/rfr.csv中。



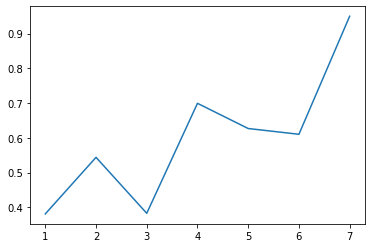
#### 3.4.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 288.49 | 351.06 |

#### 3.4.3 RSE

整体的RSE=0.5719396891740829

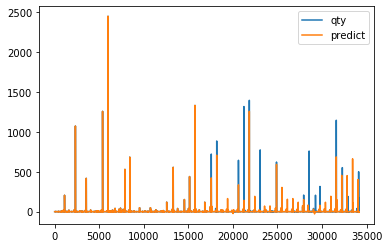
多步预测RSE



### 3.5 MLP机器学习模型

#### 3.5.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/3/mlp.csv中。



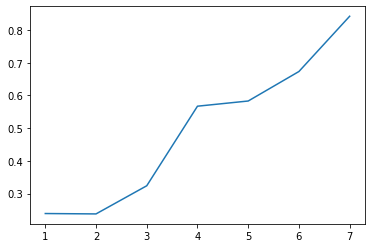
#### 3.5.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 25.33 | 273.12 |

#### 3.5.3 RSE

整体的RSE=0.48361747634130103

多步预测RSE



## 4场景四

### 4.1基本思路

在使用b.i+b.ii+b.iii+b.iv+b.v+b.vi特征的情况下，针对训练数据中商品每天的当日销量为目标特征、其他特征（即历史信息）均为属性特征，利用SVM、随机森林、MLP等3个方法进行建模，预测测试数据中某商品对应日期当日（标记为d’）至第6日（d’+6）共计7天的每日销量，可考虑如下算法：首先完成商品d’当日的销量预测，然后利用该预测销量更新上述b)的相关特征，继续预测d‘+1当日销量。。。重复该步骤，直至完成第6日（d’+6）当日销量预测。分别选取商品80%和20%的时序数据作为训练和测试数据。

### 4.2数据预处理

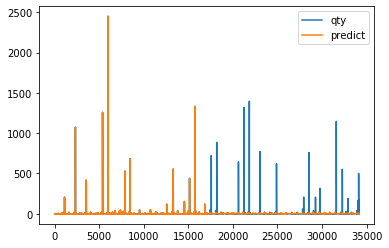
将数据reshape并作标准化处理。

1. x\_i = np.reshape(x\_i, (-1, len(feature\_name)))
2. y\_i = np.reshape(y\_i, (-1, 1))
4. **from** sklearn.preprocessing **import** StandardScaler
5. sc\_X = StandardScaler()
6. sc\_y = StandardScaler()
7. X\_i = sc\_X.fit\_transform(x\_i)
8. Y\_i = sc\_y.fit\_transform(y\_i)

### 4.3 SVM机器学习模型

#### 4.3.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/4/svr.csv中。



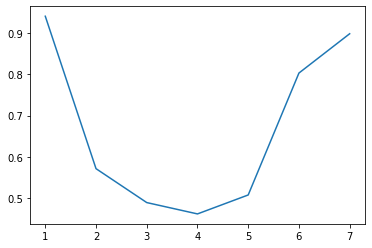
#### 4.3.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 41.75 | 275.83 |

#### 4.3.3 RSE

整体的RSE=0.6556691307190992

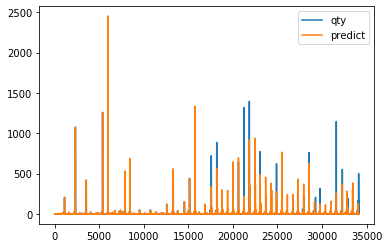
多步预测RSE



### 4.4随机森林机器学习模型

#### 4.4.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/4/rfr.csv中。



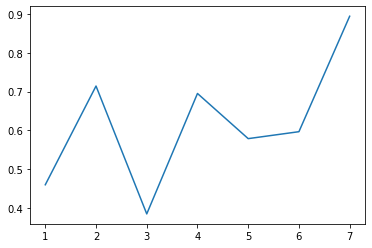
#### 4.4.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 479.85 | 347.86 |

#### 4.4.3 RSE

整体的RSE=0.5830222098913406

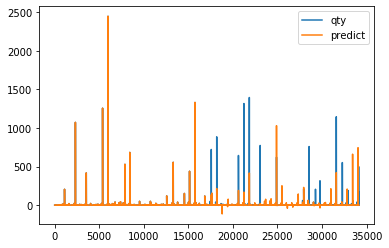
多步预测RSE



### 4.5 MLP机器学习模型

#### 4.5.1预测结果

下图为预测结果中的值的对比图，其中qty为真实值，predict为预测值。预测结果文件在2p3/4/mlp.csv中。



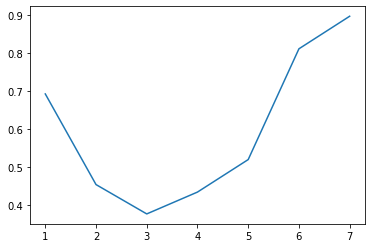
#### 4.5.2算法性能

|  |  |
| --- | --- |
| 训练时间（s） | 测试时间（s） |
| 35.92 | 281.87 |

#### 4.5.3 RSE

整体的RSE=0.5720325517341787

多步预测RSE



## 5对比分析

### 5.1预测结果对比

根据上述内容，我们可以发现在特征组合为b.i+b.iv的情况下，不论哪种模型，其训练效果均好于其余的组合，尤其时对于一些偏离的较大的点。而对于每种组合内部，随机森林和MLP而言其训练效果均要好于SVM。

### 5.2性能对比

|  |  |  |  |
| --- | --- | --- | --- |
| 特征组合 | 机器学习模型 | 训练时间（s） | 测试时间（s） |
| b.i | SVM | 240.44 | 34.77 |
| 随机森林 | 43.09 | 125.18 |
| MLP | 24.76 | 37.07 |
| b.i+b.iv | SVM | 274.45 | 28.80 |
| 随机森林 | 49.77 | 101.11 |
| MLP | 38.89 | 30.44 |
| b.i+b.ii+b.iii+b.iv | SVM | 618.97 | 273.02 |
| 随机森林 | 288.49 | 351.06 |
| MLP | 25.33 | 273.12 |
| b.i+b.ii+b.iii+b.iv+b.v+b.vi | SVM | 41.75 | 275.83 |
| 随机森林 | 479.85 | 347.86 |
| MLP | 479.85 | 347.86 |

如上图所示，我们发现随着特征数量的增加，每个模型的训练时间和测试时间基本都在增加；在特征组合一定的情况下，SVM的时间最长，MLP花费时间最短；对于随机森林和MLP在特数量不大时，训练的时间小于测试的时间；对于第四个场景下的SVM训练很短，可能时由于达到了其他的终止条件，训练过早结束而导致的。

### 5.3 RES

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征组合 | 机器学习模型 | 1步 | 2步 | 3步 | 4步 | 5步 | 6步 | 7步 |
| b.i | SVM | 0.99741122896008 | 1.0003472368488218 | 0.9973578099592385 | 1.007444743667932 | 0.9972405799275884 | 0.99567383623162 | 0.9963844935142236 |
| 随机森林 | 0.43033914434659304 | 0.8711150785892618 | 0.8505587545304785 | 1.7321745015309533 | 0.9916879291683497 | 0.8164856083786179 | 1.0148171013402523 |
| MLP | 1.0158304970355212 | 1.0137003559979647 | 0.877483576253748 | 2.373460402654206 | 0.9749796649623728 | 0.8045087127775142 | 0.9733096485164573 |
| b.i+b.iv | SVM | 0.9387316004595031 | 0.5703268836642078 | 0.4889698649411448 | 0.45870236296358813 | 0.5068874684709798 | 0.800259898764541 | 0.8961485364406993 |
| 随机森林 | 0.4271570728043932 | 0.5623360804181009 | 0.38286829858977556 | 0.8515163894744833 | 0.6728927674953398 | 0.5544649461597351 | 0.8244904419166643 |
| MLP | 0.5973192722491424 | 0.42640744553580623 | 0.32500128265371714 | 0.4621042894953252 | 0.5788238155258085 | 0.7071288008515509 | 0.6641957031767457 |
| b.i+b.ii+b.iii+b.iv | SVM | 0.9395922581834383 | 0.5707198747283626 | 0.48958825716098925 | 0.4597777005416951 | 0.5078967597076346 | 0.802411154908965 | 0.89757874762223 |
| 随机森林 | 0.3811394182703276 | 0.5439162682239238 | 0.3829358734205516 | 0.6993069744517743 | 0.6268075126704337 | 0.6102332254130962 | 0.9498132520298082 |
| MLP | 0.23874166560596946 | 0.2375645149779624 | 0.32365655721103165 | 0.5669952313223444 | 0.5828545739566778 | 0.6730711167347492 | 0.842071555588914 |
| b.i+b.ii+b.iii+b.iv+b.v+b.vi | SVM | 0.9401125345179211 | 0.5716038655961306 | 0.48962426910088036 | 0.4621990134077155 | 0.5080119058679273 | 0.8027667759616205 | 0.8977822788004007 |
| 随机森林 | 0.45987967680544634 | 0.7142064234623329 | 0.38477304234793713 | 0.6950620650948268 | 0.5787319852483758 | 0.5966110880874649 | 0.894029344120625 |
| MLP | 0.6931254388097412 | 0.4549222564570274 | 0.3778125310496247 | 0.4352170100473241 | 0.5206352291605907 | 0.8117228296318183 | 0.8972965611971304 |

如上图所示，对于相同的场景下，基本为RSE随着步数的增加而增加，但是对于SVM而言，单步预测时误差更大，对于仅有特征b.i时随机森林和MLP在第四步时均出现了RSE的突然增加。在同一模型下，RSE的变化趋势基本相同，其中随机森林和MLP在第四步均出现了RSE的突然增加，而对于SVM而言，在多步预测时（大于1步小于7步）其RSE反而小。

### 5.4对比小结

对于特征而言，最好选择b.i+b.iv的特征进行训练，以防特征过少或特征过多而造成的训练结果较差。而对于模型的选择时，则推荐随机森林或者MLP模型，其总体效果会优于SVM，而如果选择特定未来某几天内（如2-6天），则可以尝试使用SVM模型进行训练。