**Twitter上的情绪如何预测股价走势**

**正文**

贪婪和恐惧是股市的两大驱动力。事实证明，社交媒体信息中的积极和消极情绪，比如Twitter，可用于预测股票价格的日常变动或走势。

尽管新闻肯定会影响股市价格，但公众情绪状态也可能发挥同样重要的作用。**我们从心理学研究中得知，情感和信息一样，在人类的决策过程中扮演着重要的角色。行为金融学进一步证明，金融决策在很大程度上是由情绪驱动的。**因此我们有理由假设，公众情绪能够像新闻一样推动股市的价格。

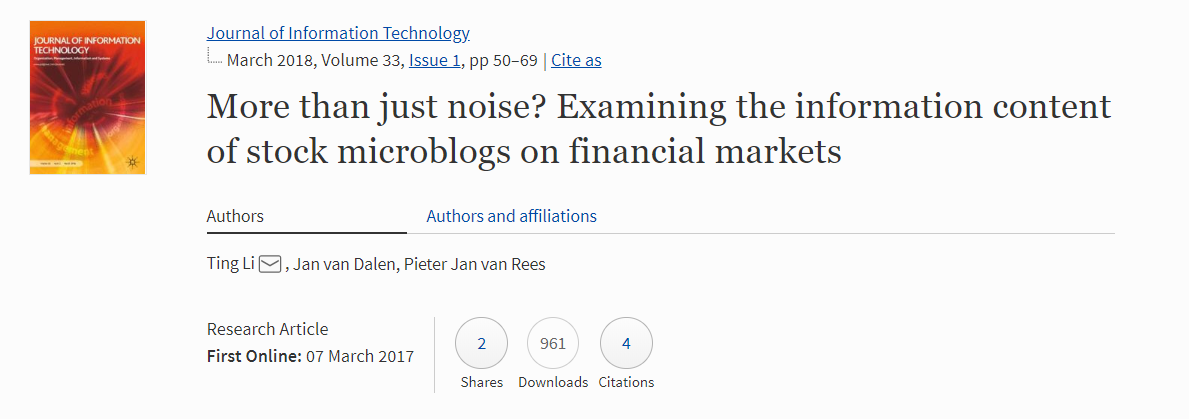
这里有一些研究可供大家参考：

*论文地址：*[*https://arxiv.org/pdf/1010.3003.pdf*](https://arxiv.org/pdf/1010.3003.pdf)



*论文地址：*

*https://link.springer.com/article/10.1057/s41265-016-0034-2*



*论文地址：http://blueanalysis.com/iulianserban/Files/twitter\_report.pdf*



*论文地址：http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf*



本文整个分析过程都是基于**Python编写**。

普及一个知识：

1、Twitter（推特）：是国外的一个社交网络及微博客服务的网站。

2、Tweet：是用户发到Twitter上的信息，为了接收或者发送Tweets首先要注册一个免费的Twitter帐号。

3、微博（MicroBlog）：是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过WEB、WAP以及各种客户端组件个人社区，以140字左右的文字更新信息，并实现即时分享。

**假设**

今天的Tweet带有正面或负面情绪，并包含一个或几个cashtags可以影响股票明天的走势。如果今天负面情绪占主导地位，那么明天的股票价格预计会下跌，反之则会上涨。Twitter账户的粉丝数量也是一个主要因素。一个账户的关注者越多，推文的影响力就越大，他们的情绪对股价的影响也越大。

**cashtags是什么?**

Twitter的一项功能允许用户点击股票代码，看看“Twitterverse”在说些什么，，比如$GOOG、$AAPL或$FB。该系统的工作方式Twitter众所周知的#hashtags相同。Cashtags要求“$”后面跟着股票代码。

**公众号补充：**

一个通用标准$符号被纳入了twitter的官方标记（cashtag），Twitter宣布这是包含了股票跟踪链接，用户点击股票信息便会显示到搜索页面上。

国内的雪球早已将$标记融入自己的微博服务中，且这些投资社区对$标记利用得更好。点击$标记后可显示出对应公司/股票的实时股价等交易信息及其他投资者对于这支股票的讨论。

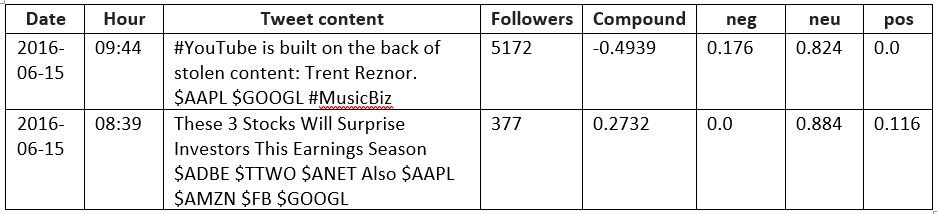
**数据集**

从2016年3月28日到2016年6月15日，79天内收集了大约100万条推文，其中提到了纳斯达克100指数成分股公司的cashtags。这些数据由followthehashtag.com提供，这是一个Twitter搜索分析和商业智能工具。

<https://www.followthehashtag.com/datasets/nasdaq-100-companies-free-twitter-dataset/>

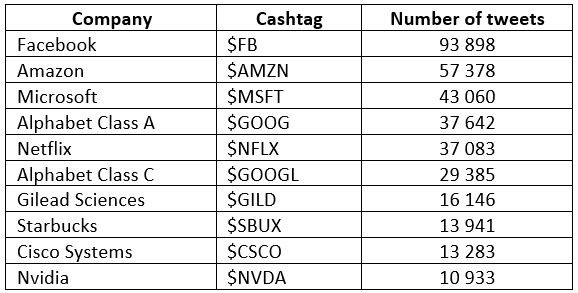


这里有两个带有cashtags的负面和正面推文的例子，分别代表苹果、谷歌和其他少数公司。



在数据中的100只原始股票中，不得不因为各种数据特定的原因而减了15只，比如日期上的不一致，或者仅仅是因为关于cashtags的推文太少，也就是说，甚至连每天的推文都没有。排除在外的人包括Apple，Tesla和Yahoo。

最终分析中包含推文最多的cashtags是（前12名）：

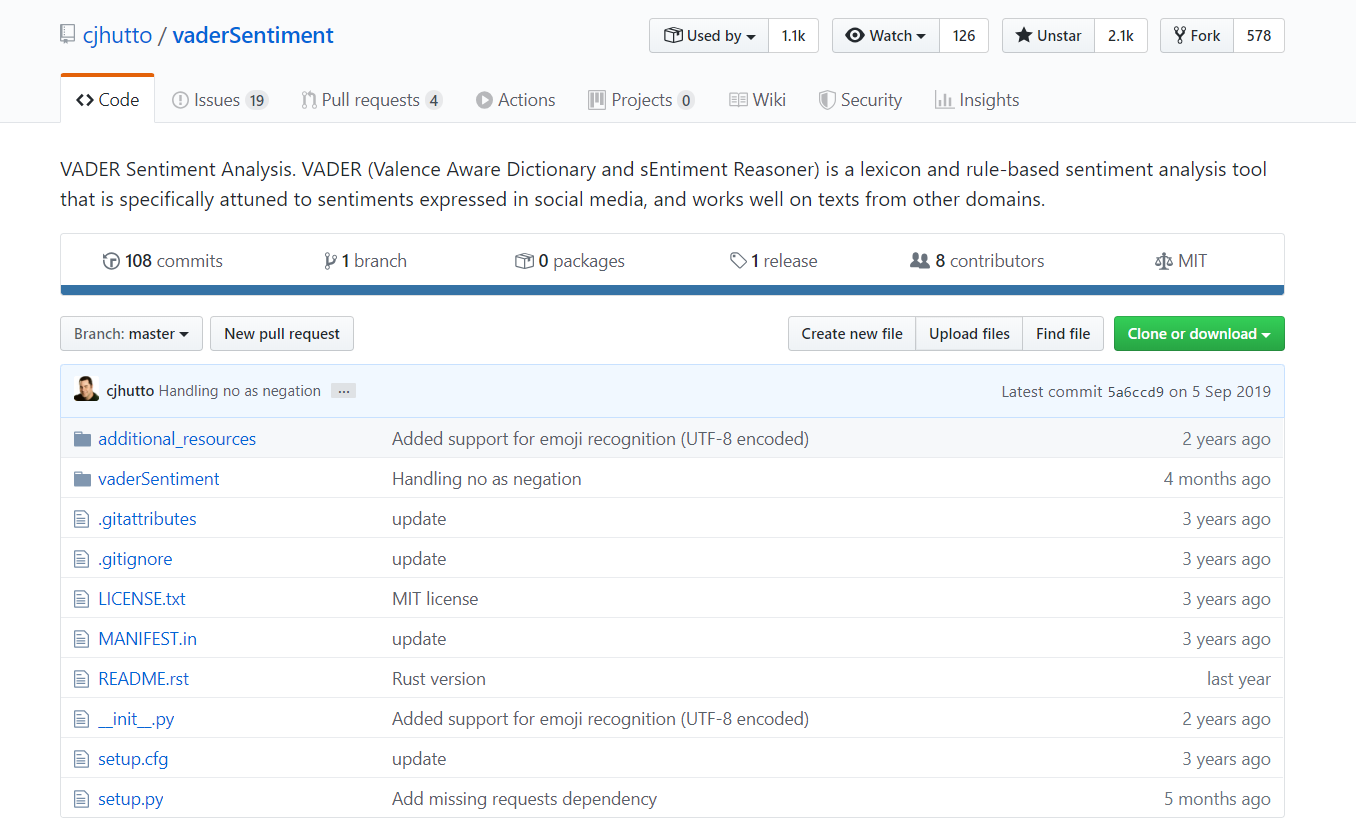


在这79天的时间里，100只股票cashtags的平均推文数为6446条，即每只股票/cashtags每天有81条推文。

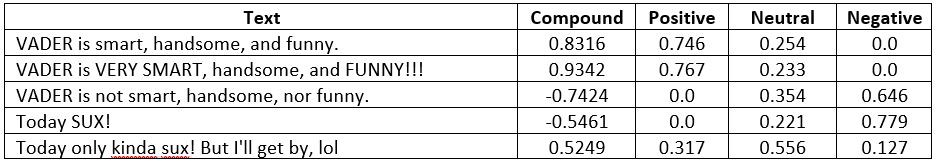
**衡量 tweets上的情绪**

为了提取每条 tweets的情绪，我们使用了**VADER**，这是一个现成的Python机器学习库，用于自然语言处理，特别适合阅读tweets的情绪。

*地址：https://github.com/cjhutto/vaderSentiment*



VADER更注重大写字母的识别，还能识别俚语、感叹号和最常见的表情符号。情绪得分从极负（-1）到极正（+1），中性为0。比如：



**为tweet数据创建每日平均值**

**在将每条推文与其情绪相结合后，将其乘以该帐户的关注者数量。这样，在最终的模型中，更多“有影响力”账户的推文情绪将得到了更多的权重。在此之后，这些推文（平均每条cashtags有6500条）被压缩到75行，其中包括每条情绪的每日平均值，然后将其与相关股票的每日价格变化进行比较。**

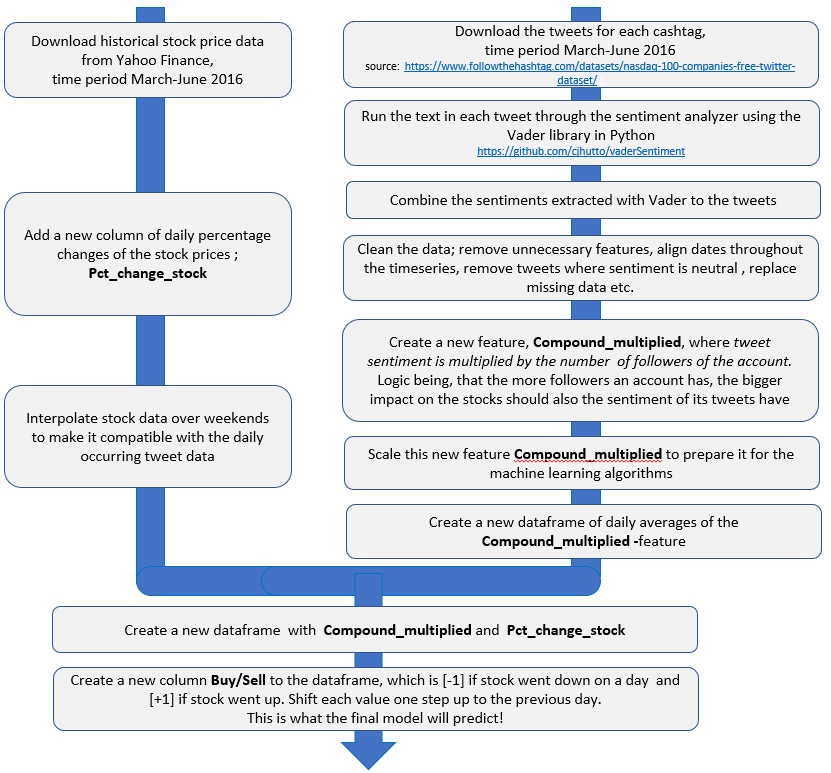
**收集股票数据**

使用Python的pandas-datareader库，从Yahoo Finance下载股票的每日数据。在股票数据中添加每日百分比变化列，并对周末缺失的数据进行插值之后，现在可以合并这两个数据集，即推文的情绪和股票的每日变化。

一个具有“Pct\_change\_stock”和“compound\_multiply”两个特征的新dataframe，以及一个添加标签数据列“Buy/Sell”，现在已经准备好在训练中使用。

**第一部分流程图分析**

股票数据（左箭头）Twitter数据（右箭头）



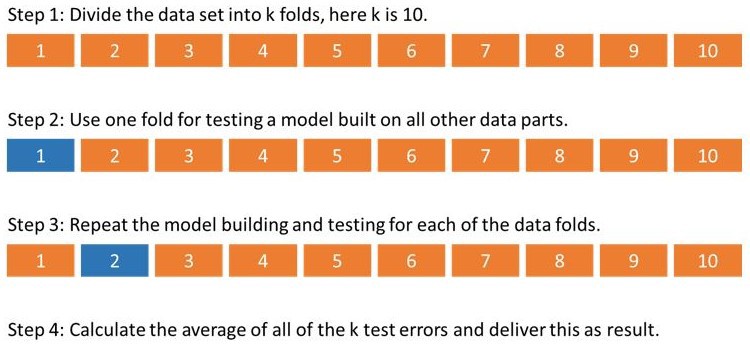
**机器学习分类器**

由于这是一个二元分类任务，即结果要么是“买入”，要么是“卖出”，因此我们使用了6种这样的算法：

* **KNN**
* **Logistic回归**
* **支持向量机（SVM）**
* **朴素贝叶斯**
* **决策树**
* **随机森林**

**训练/测试数据分割**

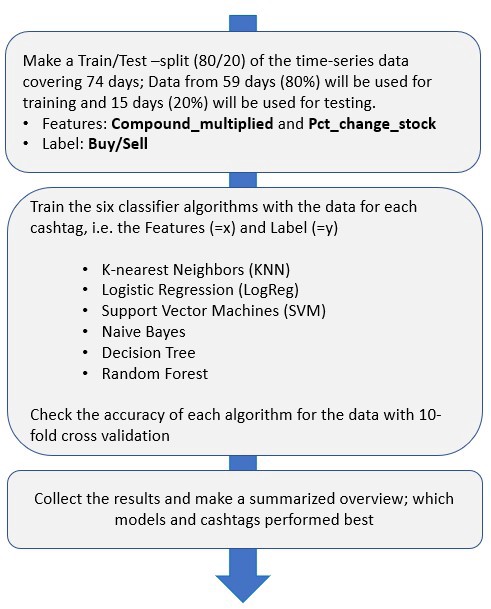
在74天可用的数据中，每只股票59天（80%）的数据用于训练，15天（20%）的数据用于测试每种算法的准确性。



**交叉验证**

由于数据量有限，仅使用20％的数据（15天）和80％的训练数据（59天）进行测试可能不够有代表性。为了避免训练/测试分割不完全随机的可能性，对数据进行交叉验证，这样得到每个算法精度更具代表性的结果。训练数据进一步分成10个子集，每个子集都与其他9个子集进行测试。

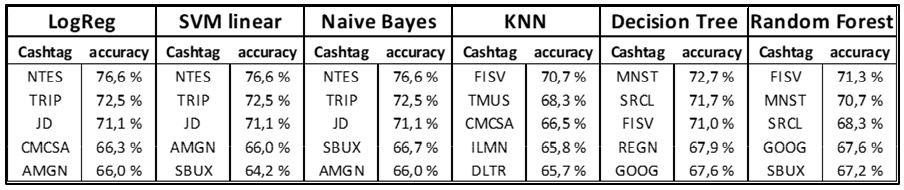
**第二部流程图分析**



**结果**

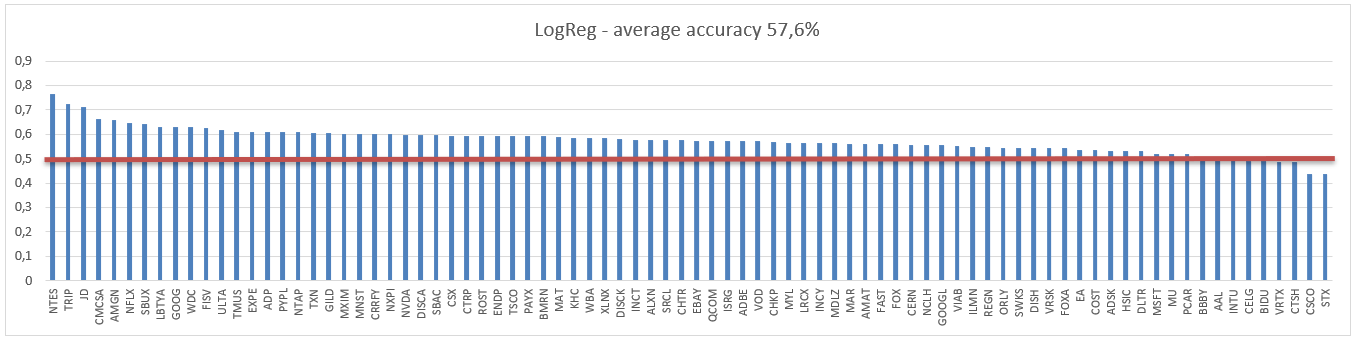
将85只股票分别通过6个二元分类器和10倍交叉验证后，结果如下。**平均每个分类器的准确率都在50%以上**。**这意味着，推特上的情绪具有预测力，至少比抛硬币强。抛硬币的平均准确率为50%，所以准确率超过50%在一定程度上证明了模型获得“非凡”收益的能力。更重要的是，对于许多股票，模型的准确性/预测能力在65-75%之间！**

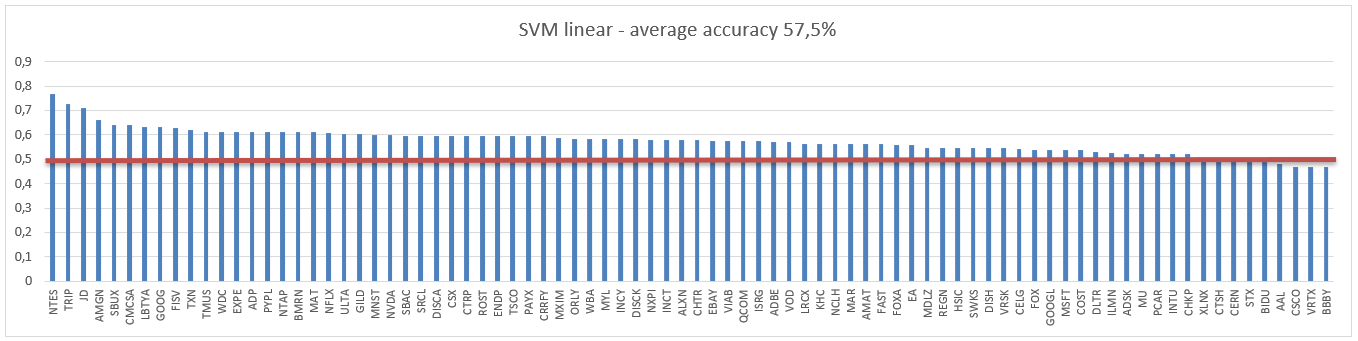
每个cashtag分类器的准确率（前5名）：



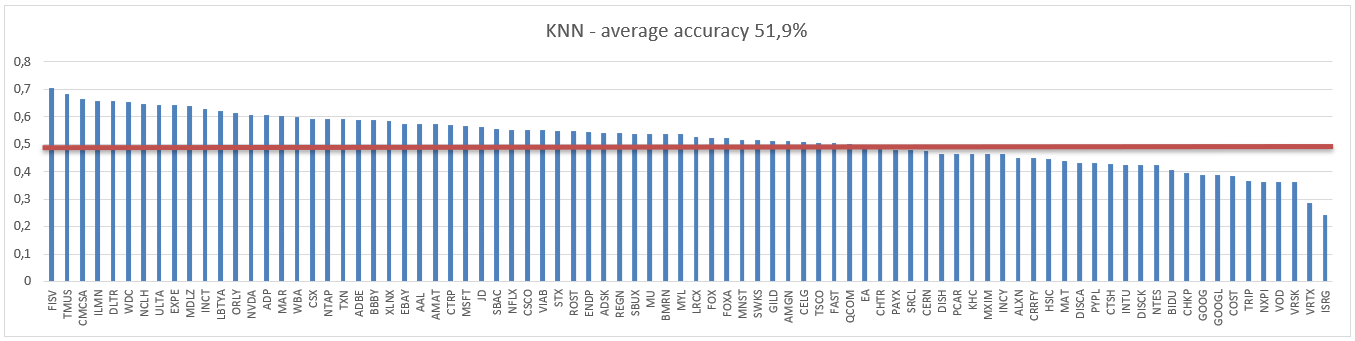
在下面的图表中，红线表示50%的准确度限制。

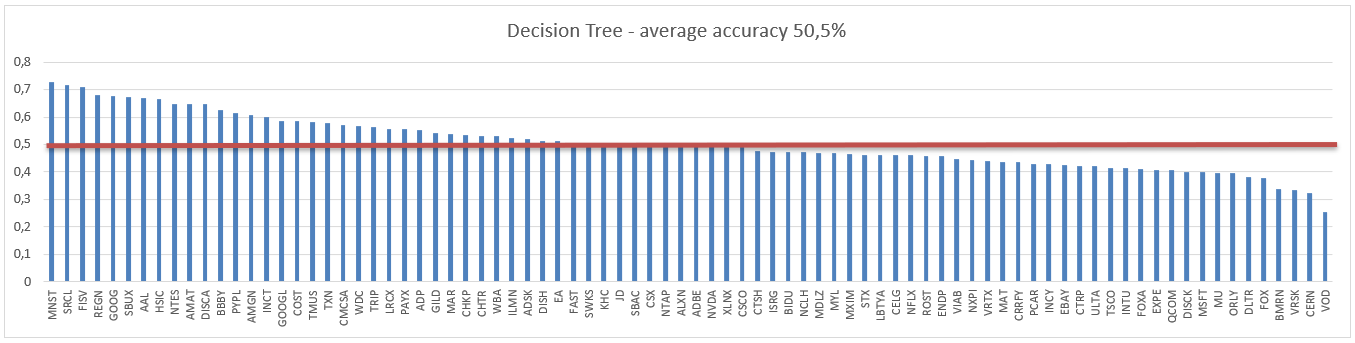
以下是所有分类器的平均准确率：

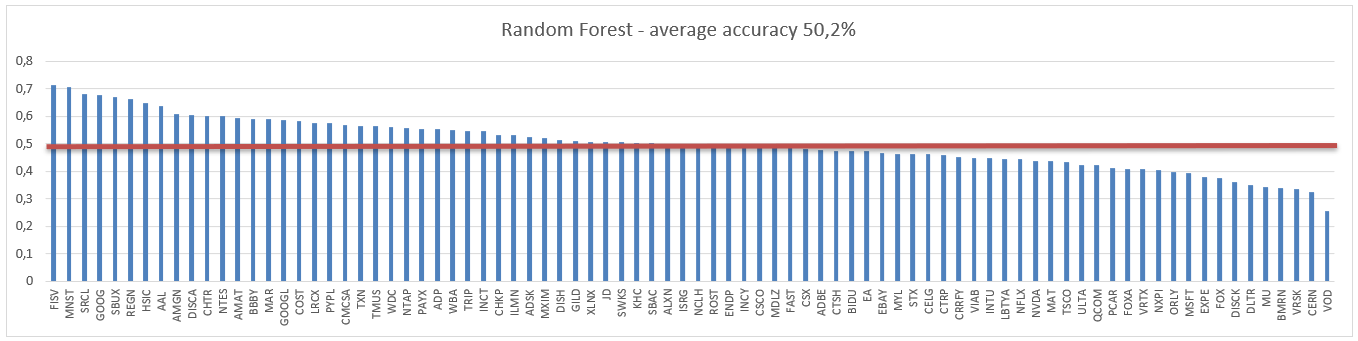








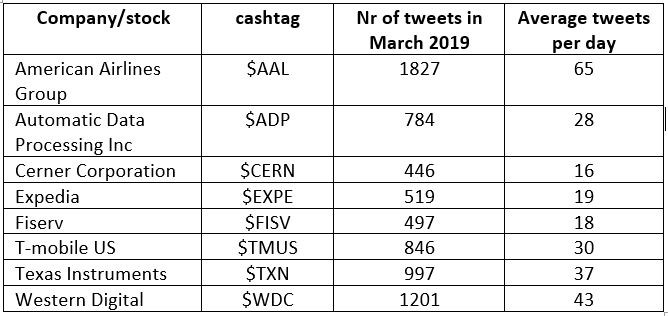




接下来，我们将简单买入持有策略的盈亏与使用模型实现的盈亏进行了比较。令我们惊讶的是，在为期四周的模拟交易中，大多数模型的利润都远超我们的预期！

**下载tweets**

我们选择了纳斯达克的8只股票进行模拟，三月模拟交易的推文总数接近7200，平均大约800每条股票的推文。



tweet数据是通过使用其Developer API“抓取”Twitter而收集的。我们在2016年3月下载了所有包含cashtags $AAL、$ADP、$CERN、$EXPE、$FISV、$TMUS、$TXN和$WDC的tweets。

**下载和准备其余的数据**

前面我们详细解释了后续步骤的过程，下面简要的做一个回顾：

1、推文通过情绪分析算法运行，每个推文都有一个情绪；积极的，中性的或消极的。

2、每条推文都乘以该账户的关注者数量。这样，在最终的模型中，更“有影响力”账户的推文情绪就会得到更多的权重。

3、Tweet数据被压缩到28行，包含每一个情绪的日平均，并与同期相关股票的日价格变化进行比较。

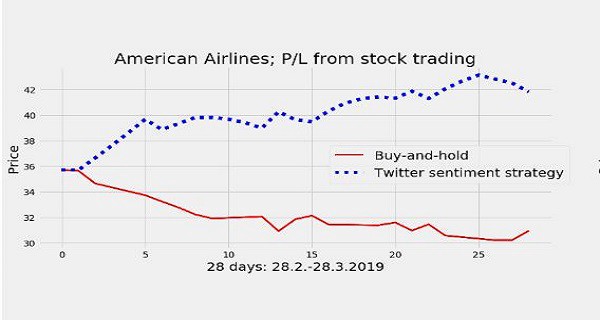
4、股票数据下载并添加“每日变化百分比”列中。

5、Tweet和股票数据相结合，并添加一个标签列，即“买进或卖出”。这就是模型试图预测的内容。换句话说，基于今日推特情绪的预测值，预测一只股票应该在明天买进还是卖出?

然后通过比较买入持有策略与六种不同模型来使用这些数据集，每个每日预期的每日股票价格变动是使用模型预测的。

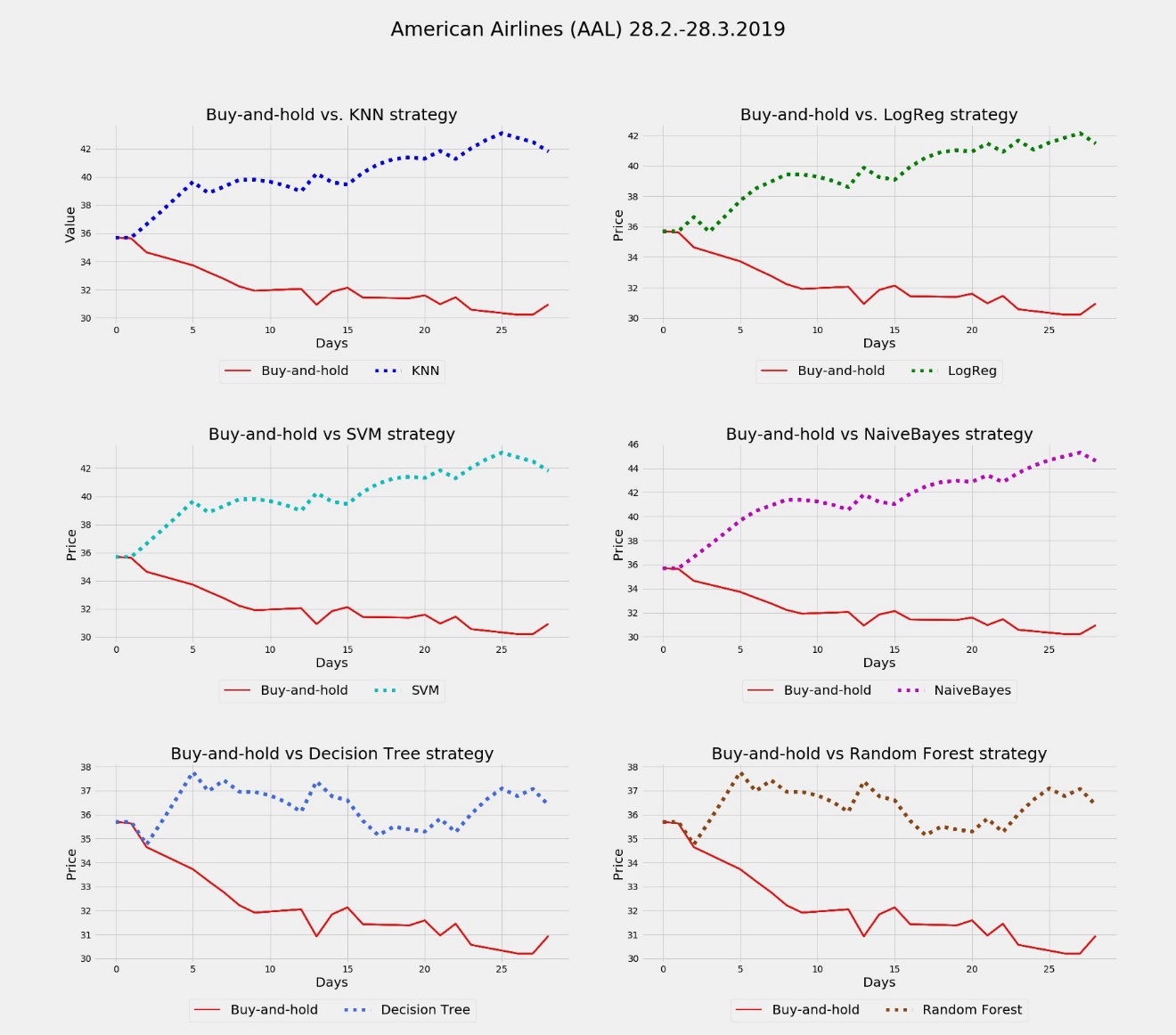
**进行模拟交易2019年3月**

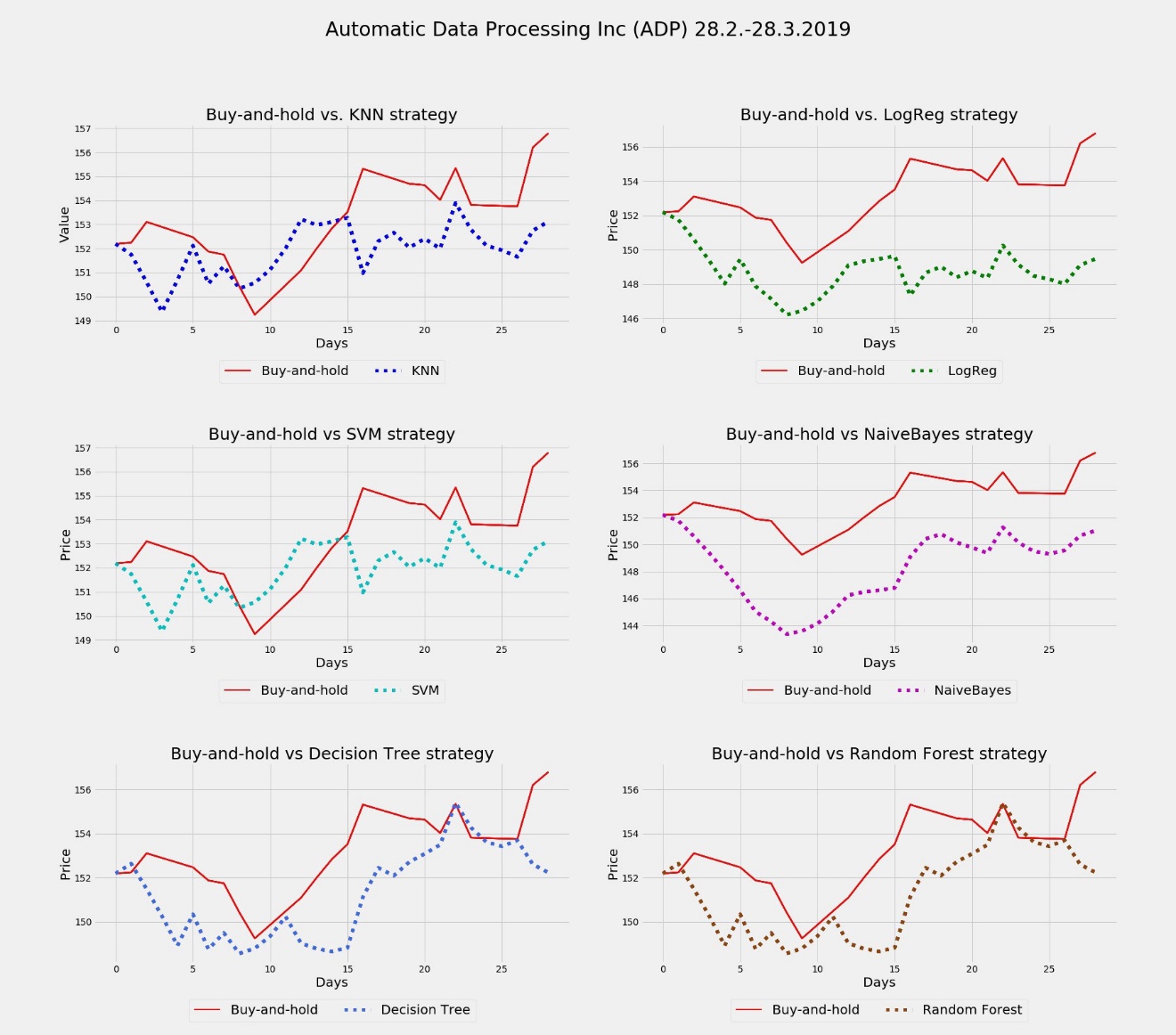
对8只股票分别采用买入并持有策略，与其他6种基于二分类算法策略进行比较。

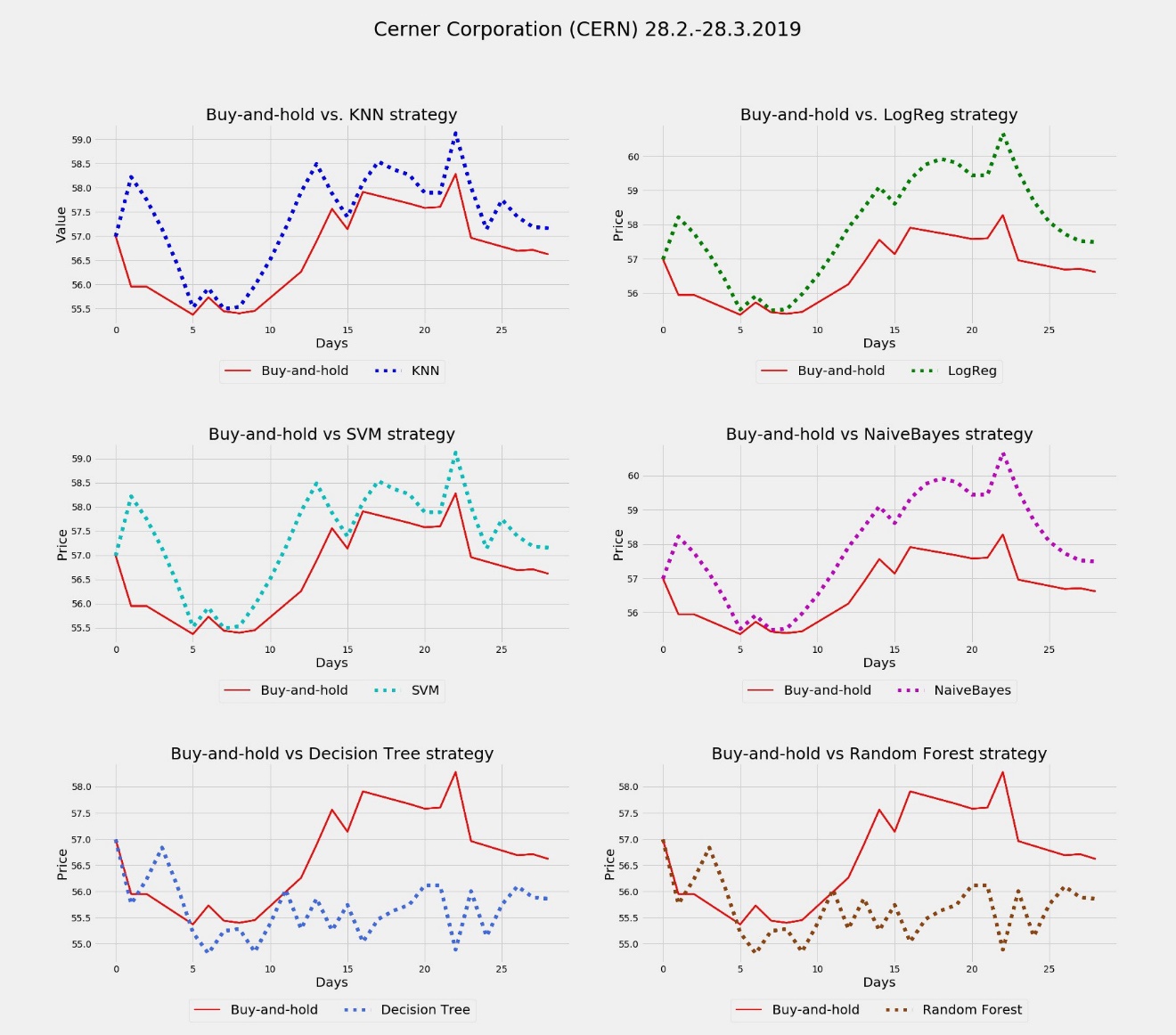
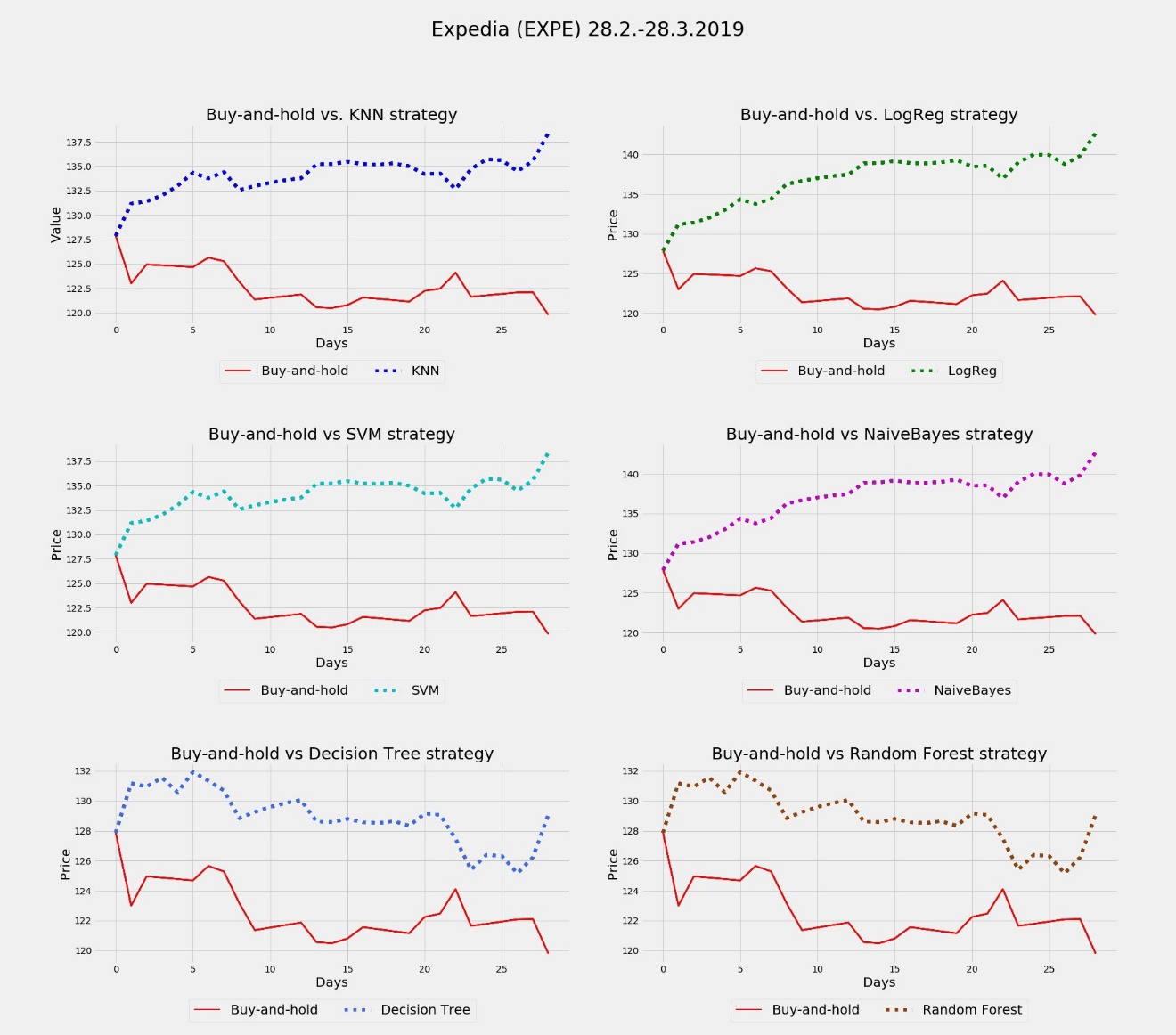


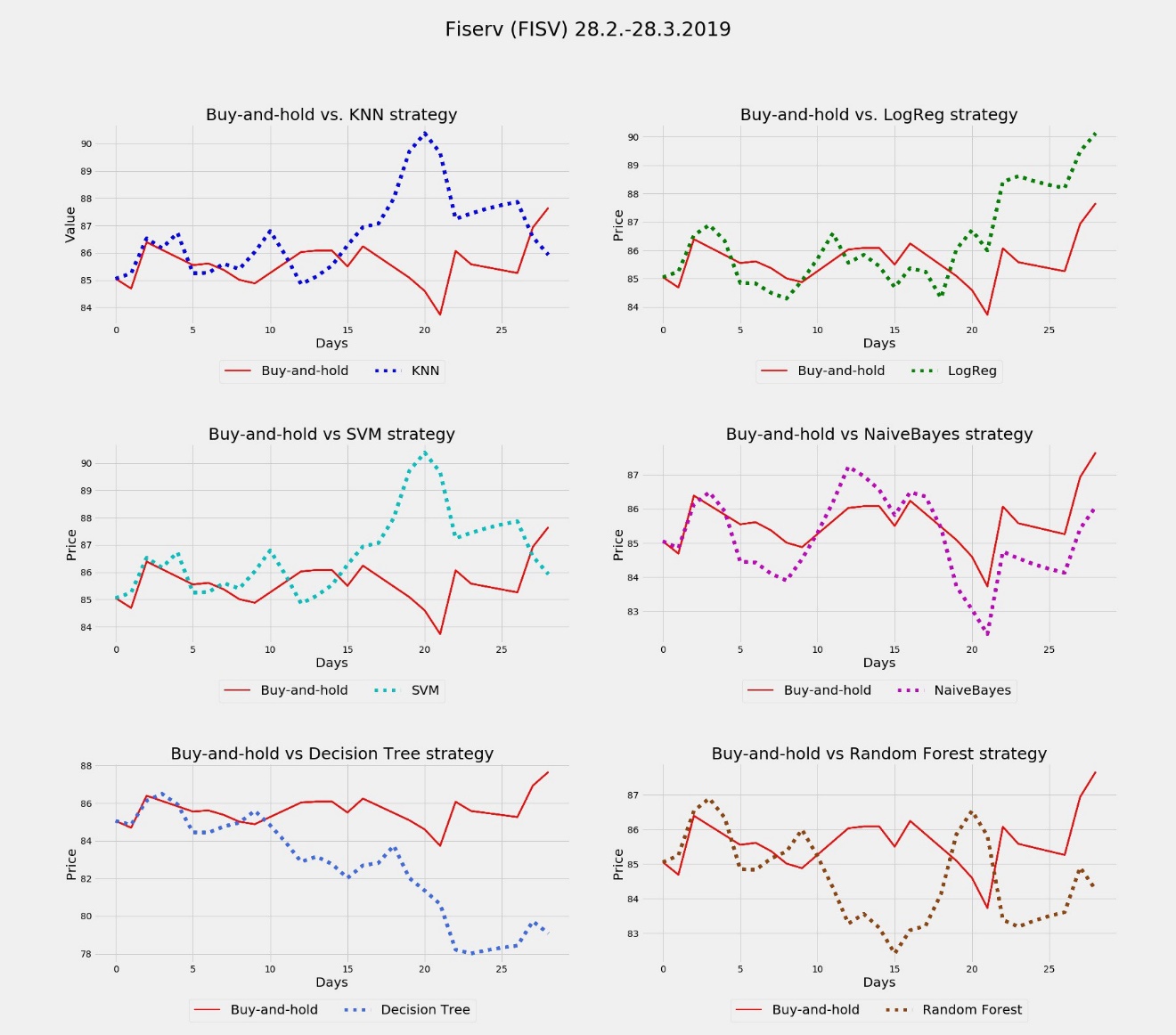
每个模型都使用2016年的原始推文进行训练。然后给出了每日建议：明天开盘时买入或卖出，收盘时卖出或买入。

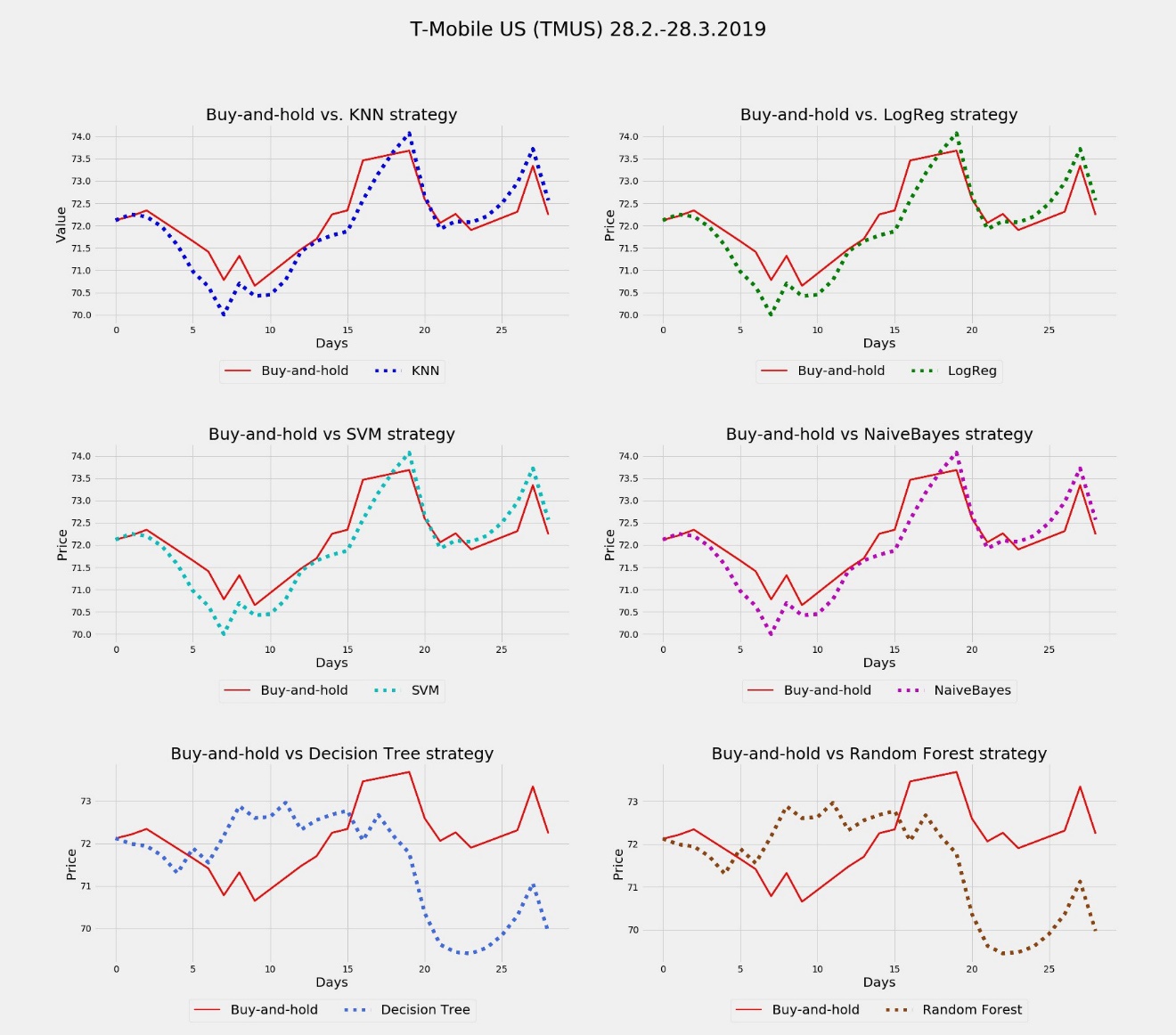
看下图的策略结果：

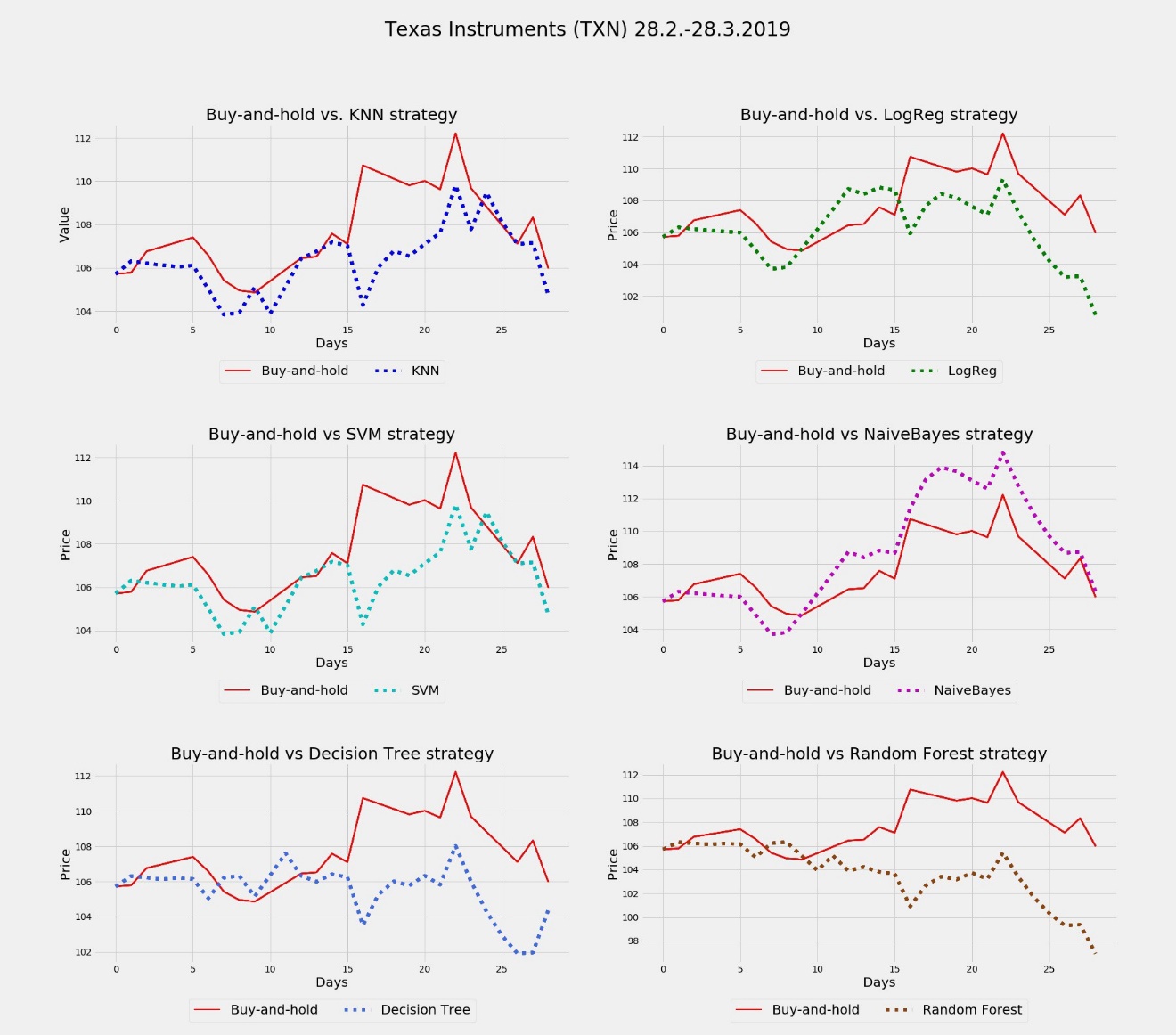


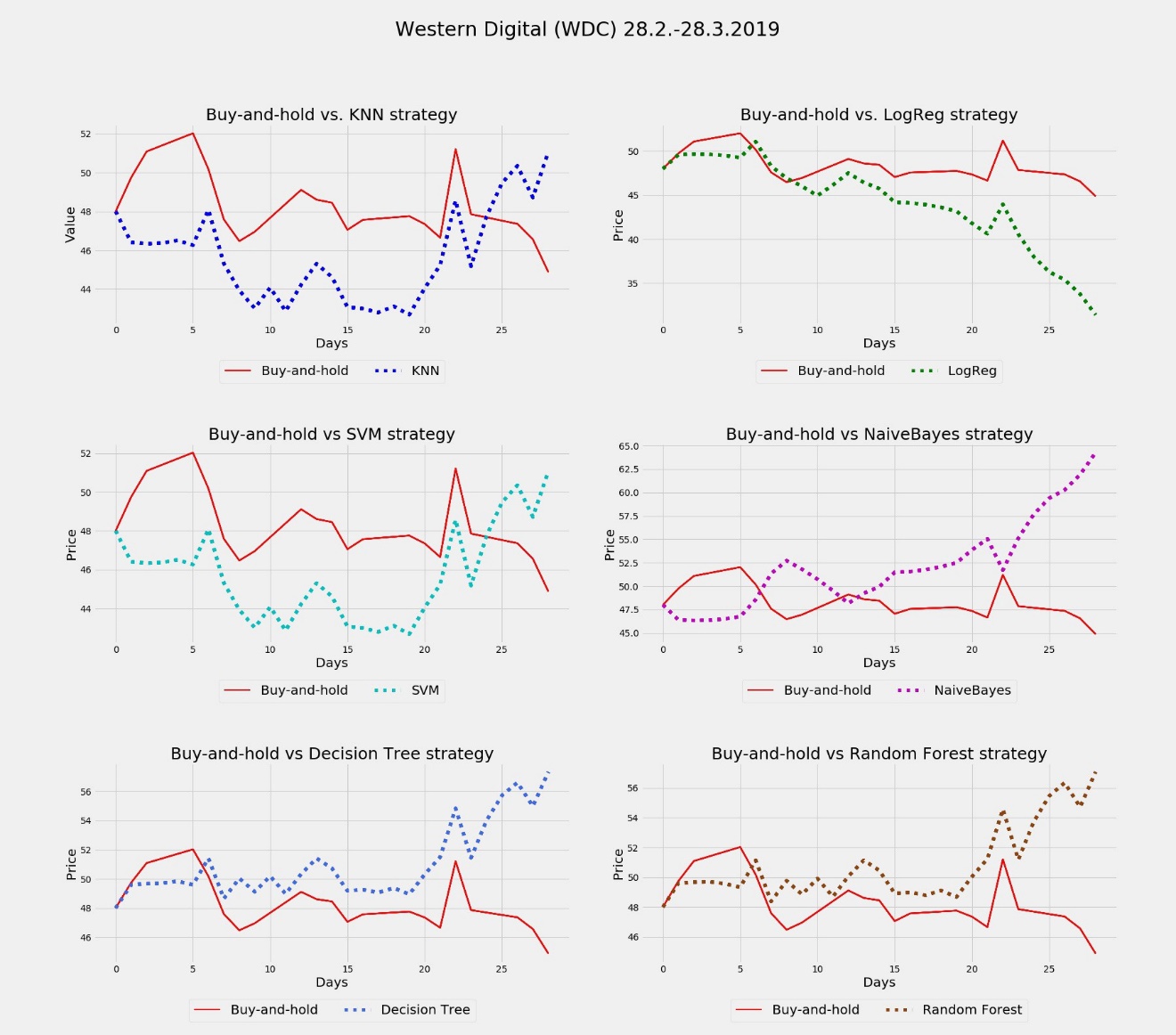




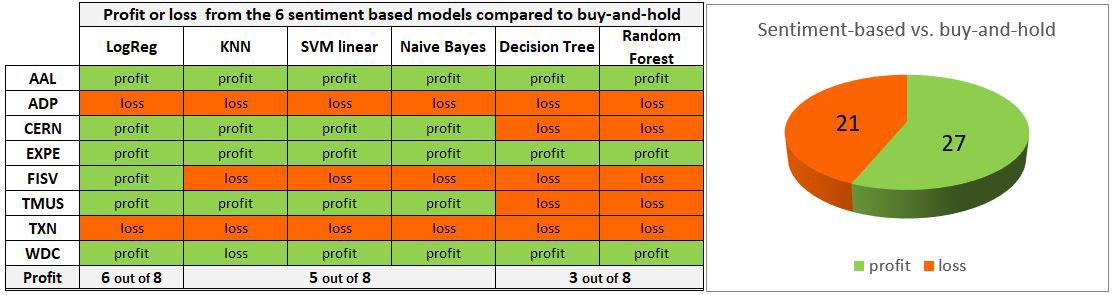






**总结**

总体而言，**基于情感的twitter策略在近60%的模拟案例中击败了买入并持有策略。**



除去两个表现最差的模型，决策树和随机森林，结果得到了进一步的改进。在三分之二的案例中，“买入并持有”不理想。



**如果只遵循表现最好的模型Logistic回归，那么在4只股票中每3只股票就会盈利！**

**进一步完善模型思路**

1、模型只有75天的数据用于训练和测试。如果情绪真的具有预测能力，那么从**更长的、甚至更近的时期添加更多数据，可能会显著改善结果**。

2、**为了使每周仅5天的股票数据与每周7天的twitter数据相吻合，需要对周末调整后的收盘价进行插值**。虽然考虑了特征工程，但周末创建的股票价格是人为的，可能会扭曲结果。**考虑到推文对周一股市走势的影响，或许周五到周日的推文应该以某种方式组合在一起。**

3、可以考虑将推特情绪的结果与其他技术结合使用，**比如LSTM神经网络进行时间序列分析，总是提前一天做出预测**。

4、尝试使用其他一些现成的模型，比如**TextBlob**，而不是VADER来提取tweet情绪。或者更好的方法是，通过建立一个神经网络来训练你的情绪分类器，然后用你自己的数据来训练它，比如这里的数据；1.6mio将每一行标记为0=负，2=中性，4=正。

5、**时间对最终结果的影响有多大？**在模拟中，最终的P/L取决于周期的长度。在某些情况下，交易期越长，利润就会变成亏损，反之亦然。

6、**模拟中没有考虑交易成本**。至少在最终利润相当微薄的情况下，交易成本可以将利润变成亏损。

7、**能否在特定业务领域的特定股票中发现模式？**在这项分析中，美国航空和Expedia这两家旅游公司的股票收益最高。这仅仅是个巧合，还是某些企业的股票走势更容易引发推特情绪？