Structure

1- Overview
   1.1 - Stocks of Interest
   1.2 - Previous Work

2- An Unsuccessful Attempt to Scrape Text Data Directly
   2.1 - Sina Weibo Scrawling
   2.2 - Twitter Scrawling

3- News Data Collection and Processing
   3.1 - News Collection and Aggregation - Using Tushare
   3.2 - Data Processing
      3.2.1 - Filtering (Only Keeping News related to This Stock)
      3.2.2 - Sort According to Date
      3.2.3 - Sentimental Analysis
      3.2.4 - Kernel Smooth Method to Get Score in Each Day

4- Market Data Collection and Input Preparation

5- Building and Applying the Data Preparation Pipeline

6- Training and Testing Linear Regression Models
   6.1 - Universal Model
   6.2 - Individual Models
      6.2.1 - Percentage Change Prediction Results
      6.2.2 - Price Movement Prediction Results

7- Using Model to Actually Trade

8- Include History Market Data

9- Training and Testing Neural Networks

10- Rolling Window Framework

11- Where More Studies can be Done
   11.1 - Data Processing Stage
   11.2 - Input preparation Stage
   11.3 - Model Training Stage

1 - Overview

This project aims to collect text data related to certain stocks from social media and official news and then perform sentimental analysis and other kinds of data analysis on the collected data.
Such information is then further filtered and leveraged (probably together with the history stock data), to try to predict the stock index movement of the selected stocks using machine learning techniques.

1.1 - Stocks of Interest

As required, we only consider the current FTSE China A50 index constituents. We use python lists to store the information like codes, names and industries of these fifty stocks.

	序号	证券代码	公司名称	Company Name	行业
0	1	000002	万科	China Vanke A	房地产
1	2	601360	三六零	SJEC Corp	信息技术
2	3	600104	上汽集团	SAIC Motor Corp	汽车
3	4	600018	上港集团	Shanghai International Port	交通运输
4	5	600030	中信证券	CITIC Securities	金融业
5	6	601998	中信银行	China Citic Bank A	银行业
6	7	601766	中国中车	CRRC A	交通运输
7	8	601390	中国中铁	China Railway A	重工
8	9	601800	中国交建	China Communications Construction	重工
9	10	601628	中国人寿	China Life Insurance A	保险业

1.2 - Previous Work

After some research, I find that the majority of work already done on this topic focuses solely on collecting text data without the search of a particular keyword (just collect all twitters / weibos from the Internet and perform some rudimentary filtering).
Then, the text data is used mainly to predict the movement of some **index** as a whole, rather than the prices of some individual stock. In other words, previous work is mostly done on the graininess of the entire market, seldomly so specific as to study any individual stock.

Representative:

- Stanford: Using Twitter to predict Dow Jones Industrial Average
  http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf
- Shanxi University: Using Weibo to predict average of stock prices
  http://cdmd.cnki.com.cn/Article/CDMD-10108-1018312077.htm

2 - An Unsuccessful Attempt to Scrape Text Data Directly

My first experiment to obtain data is to manually collect text data from two sources. This experiment turns out to be unfruitful.

2.1 - Sina Weibo Scrawling

The first source I choose is Sina Weibo, the major social media platform where the general public in Mainland China express their thoughts and views casually.

Although Sina Weibo has its own API, it is heavily restricted and does not support the function of "search by keyword" very well. Therefore, I wrote my own scrawler to go through pages one by one and collect Weibo live.

For example, We can collect one page of Weibo related to the first stock 万科 and sort them into one pandas dataframe to have a casual look.

mid	comments	likes	reposts	source	text	time	user_gender	user_verified
4413883316134334	0	0	0	微博 weibo.com	<a href="https://m.weibo.cn/search?containeri...	刚刚	m	True
4413881541356044	0	0	0	iPhone客户端	记昨天。图二万科后面的骨汤麻辣烫。	7分钟前	f	False
4413880648347677	17	15	27	石家庄超话	#正定文化村#这个中秋你想怎么过？不如来万科&园博园中秋嘉年华！这里有奇幻的水幕电影、震撼的...	11分钟前	m	True
4413880136434605	4	4	4		【深圳回迁房"灰色"生意：违建变身拿"红本" 旧改3年房价翻5倍】短短5个月，凭借一套位于深...	13分钟前	m	True
4413879897164478	0	0	0	iPhone客户端	🎉🎉🎉良渚万科未来城二期小高层稀缺三房两卫带车位月租只要3900 只要3900 👏👏 <...	14分钟前	f	False
4413876831816569	0	0	0	iPhone客户端	合景退房成功 林给我买万科大都会滨江	26分钟前	m	False
4413874516763746	0	3	1	微博 weibo.com	上市企业员工都坚决不买自家上市企业股票，包括贵州茅台，格力电器，万科，其中原理实际也不难分析，	35分钟前	m	True
4413871097596495	0	0	0	荣耀V20 4800万3D相机	万科翡翠云图市集——红孩儿也来出集啦	48分钟前	m	True
4413870988330221	0	0	0	HUAWEI Mate 20	这是包手啦继续添加<img alt=[good...	49分钟前	m	True
4413870850223413	0	0	0	iPhone客户端	万科可以啊，给业主转基因大豆油！怎么想的？！ <...	49分钟前	f	False

As one can see, the above information, if in large scale, has quite a lot potential to exploit.

Unfortunately, further collection presents a somewhat insuperable obstacle.

Due to the limits imposed by Weibo, the best I can achieve is collecting around 300 weibos for each of the 50 stocks. (Demo).


```
"mid": "4412748823707166",
"time": "4小时前",
"source": "微博 weibo.com",
"user_verified": true,
"user_gender": "m",
"reposts": 0,
"comments": 0,
"likes": 0,
"text": "中国股市：周三持有以下票的朋友，有拿不准或被套的下方评论或留言，盘中精选强势股，尽在蔚
    <br />(600018)上港集团(600019)宝钢股份(600020)中原高速(600021)上海电力<br
    />(600022)山东钢铁(600023)浙能电力(600025)华能水电(600026)中远海能 "
```


300 Weibos cover a period of roughly 4-7 days, which is far from enough for extracting features, combining with daily stock prices and then training for models.

2.2 - Twitter Scrawling

Adapted from my own previous work at Berkeley, simply using Tweepy API (instead of manually scraping) to collect tweets related to certain keyword. Demo Below.

Note: Here, we search by the English name of the corresponding stock rather than the Chinese name.


```
{"created_at": "Tue Sep 03 05:33:34 +0000 2019", "id": 1168758916760555520, "id_str": "1168758916760555520",
   "full_text": "All of a sudden the Yen-basis widening is making sense. First GPIF &amp;amp; now
   this:\n\nDai-ichi Life Insurance —~$3.7t in assets- \"cut holdings of stocks and increased currency
   hedging on foreign bonds as U.S.-China trade frictions have escalated\" \n\nhttps://t.co/E6RXicWB9y",
   "truncated": false, "display_text_range": [0, 270], "entities": {"hashtags": [], "symbols": [],
   "user_mentions": [], "urls": [{"url": "https://t.co/E6RXicWB9y", "expanded_url":
   "https://www.theguardian.pe.ca/business/reuters/
```


Similar to the situation with Weibo, the best I can achieve is collecting around 500 tweets for each of the 50 stocks, which again, covers a period that is way too short to conduct any meaningful research.

3 - News Data Collection and Processing

3.1 News Collection and Aggregation - Using Tushare

After some research, I decided to use the API of **Tushare Pro**, an open source platform designed for financial big data focused on Mainland China.

Using this API, I can directly download news within a given period, with little to no constraints on the request frequencies. This platform allows me to access news from five sources, namely, 新浪财经，华尔街日报，同花顺，东方财富，云财经.

After some efforts and using some tricks, we can successfully collect all news from all of these five sources during the period **2019.01.01-2019.8.31** and save them locally for future accesses.

We can combine all news we get from these five different sources to form one single giant pandas Dataframe with 380k news inside.

```
news_aggregated = pd.concat(news_by_source, sort=True).fillna(value=0).iloc[:,1:]
news_aggregated.head()
```

	content	datetime	source	title
0	市场消息：Uber的估值促使零工经济面临审查。	2019-01-01 23:39:43	sina	0
1	【ST慧球：上海高院一审宣判 公司无需对顾国平债务承担担保责任】ST慧球(600556)1月...	2019-01-01 23:25:29	sina	0
2	【FAANG股分析之奈飞：2018年完胜FAANG股同行 但继续烧钱或使其2019年的股价承...	2019-01-01 23:06:07	sina	0
3	【FAANG股分析之苹果：苹果股价走低可能持续到2019年】苹果在2018年全年累跌近7%，...	2019-01-01 22:52:53	sina	0
4	【陕西省"民参军"企业已达589家】陕西省近日出台了一揽子优惠政策扶助民营企业参加军工生产，...	2019-01-01 22:42:33	sina	0

```
news_aggregated.shape
```

```
(379598, 4)
```

3.2 - Data Processing

Though at first it may appear natural to process these 380k news altogether in several batches, it turns out impossible to carry out certain operations like sentimental analysis in such a large scale.

Therefore, I decided to perform data analysis for each stock individually.

Basically, a pipeline to process these text data from news is established and then later applied to each of the fifty stocks of interest.

For illustrative purpose, I choose a specific stock, 万科 000001, and build the pipeline step by step using this particular stock as an example.

3.2.1 - Filtering (Only Keeping News related to This Stock)

What comes first is a somewhat bold filtering. We only keep those news that are related to our stock.

The word "related" is defined by specifying the following two rules:

1) The name of this stock appears in this piece of news, in this case, 万科

2) The industry of this stock appears in this piece of news, in this case, real estate, or, 房地产

```
name_appeared = news_aggregated[news_aggregated['content'].str.contains(stock_names_chn[0]).fillna(value=False)]

print(name_appeared.shape)
name_appeared.head()
```

```
(624, 4)
```

	content	datetime	source	title
177	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	2019-01-01 09:02:56	sina	0
1758	【在港上市地产股集体大涨】长实集团在香港一度升5.4%，创下2016年3月来最大上涨。恒基地...	2019-01-04 15:24:33	sina	0
2002	【万科：2018年实现合同销售金额6069.5亿元】万科早间公告称，2018年12月份公司实...	2019-01-04 07:45:01	sina	0
2003	【万科：2018年实现合同销售金额6069.5亿元】 万科A早间公告：2018年12月份公司...	2019-01-04 07:43:54	sina	0
2131	【杭州万科回应作家投诉：两年来进行十余次沟通 已诉诸法律途径】针对作家张艳华投诉项目漏水一事...	2019-01-05 15:04:03	sina	0

```
industry_appeared = news_aggregated[news_aggregated['content'].str.contains(stock_industries[0]).fillna(value=False)]

print(industry_appeared.shape)
industry_appeared.head()
```

```
(4440, 4)
```

	content	datetime	source	title
10	【重庆调整房产税起征点 专家：这是例行调整】1月1日起，重庆市主城个人新购高档住房房产税起征...	2019-01-01 22:23:09	sina	0
69	【方正固收提出10个论断：经济增长拐点不会在2019出现】1、2019年的信用扩张很难起来。...	2019-01-01 17:09:56	sina	0
105	【王府井：长春赛特奥莱MALL对外营业】王府井公告，2018年12月30日，公司旗下长春赛特...	2019-01-01 15:38:20	sina	0
112	【天津：深化房地产市场调控 2018年住房交易量价稳定】去年以来，我市认真贯彻党中央、国务院...	2019-01-01 15:23:31	sina	0
120	【中国银行与浦发银行展开海南FT账户体系下多个品种业务办理】海南自由贸易账户(FT账户)体系...	2019-01-01 14:49:16	sina	0

Now, we can see that the major problem, which is the difficulty of scraping the Internet with particular keyword is solved by simply retrieving a huge amount of news with no keyword and bring the filtering offline to our own local machines.

3.2.2 - Sort According to Date

Currently, the news data is aggregated from different sources. We want to sort it according to date so that it can be in a "time-series" form.

Currently, the news data is aggregated from different sources. We want to sort it according to date so that it can be in a "time-series" form.

To do this, we first need to extract the date and time information separately from the 'datetime' column and add a new column to the dataframe. For the time information, we extract the hour (in 24-hour format) and minute separately.

	content	datetime	source	title	date	hour	minute
177	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	2019-01-01 09:02:56	sina	0	2019-01-01	09	02
1758	【在港上市地产股集体大涨】长实集团在香港一度升5.4%，创下2016年3月来最大上涨。恒基地...	2019-01-04 15:24:33	sina	0	2019-01-04	15	24
2002	【万科：2018年实现合同销售金额6069.5亿元】万科早间公告称，2018年12月份公司实...	2019-01-04 07:45:01	sina	0	2019-01-04	07	45
2003	【万科：2018年实现合同销售金额6069.5亿元】 万科A早间公告：2018年12月份公司...	2019-01-04 07:43:54	sina	0	2019-01-04	07	43
2131	【杭州万科回应作家投诉：两年来进行十余次沟通 已诉诸法律途径】针对作家张艳华投诉项目漏水一事...	2019-01-05 15:04:03	sina	0	2019-01-05	15	04

Special Note:

In later sections, we will trade "close-to-close" (long / short stocks approximately at the market close time of each trading day).

Therefore, we have to adjust the date and time information accordingly here.

Specifically, take June the 6th as an example, if we want to trade at some time point before 3:00 PM this day, all news that are published after 3:00 PM may not be utilized today to make predictions (In face, they can only be used tomorrow and hence should be considered useful for the prediction and trading in June the 7th).

Additionally, to account for the fact in reality that the last 30min of the market is usaually very volatile, we further assume that we can only get and process the news data by 2:30 PM each day(so that we can finish trading by 3:00 PM).

Thus, we manually define the time span of "June the 6th" to be "Beginning from 14:30 of June the 5th, To 14:30 of June the 6th", and create a new column in our dataframe to represent the actual date under our specific definition.

```
def adjusted_date(row):
    if int(row['hour']) < 14:
        return row['date']
    elif int(row['hour']) == 14:
        if int(row['minute']) <= 30:
            return row['date']
        else:
            return next_day(row['date'])
    else:
        return next_day(row['date'])
```

	content	datetime	source	title	date	hour	minute	real_date
177	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	2019-01-01 09:02:56	sina	0	2019-01-01	09	02	2019-01-01
1758	【在港上市地产股集体大涨】长实集团在香港一度升5.4%，创下2016年3月来最大上涨。恒基地...	2019-01-04 15:24:33	sina	0	2019-01-04	15	24	2019-01-05
2002	【万科：2018年实现合同销售金额6069.5亿元】万科早间公告称，2018年12月份公司实...	2019-01-04 07:45:01	sina	0	2019-01-04	07	45	2019-01-04
2003	【万科：2018年实现合同销售金额6069.5亿元】 万科A早间公告：2018年12月份公司...	2019-01-04 07:43:54	sina	0	2019-01-04	07	43	2019-01-04
2131	【杭州万科回应作家投诉：两年来进行十余次沟通 已诉诸法律途径】针对作家张艳华投诉项目漏水一事...	2019-01-05 15:04:03	sina	0	2019-01-05	15	04	2019-01-06

	content	source	title	date	real_date
177	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	sina	0	2019-01-01	2019-01-01
166	【2018年千亿房企达30家，恒大夺销售权益榜首位】据2018年12月31日克而瑞发布的《2...	wallstreetcn	0	2019-01-01	2019-01-01
167	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	wallstreetcn	0	2019-01-01	2019-01-01
2003	【万科：2018年实现合同销售金额6069.5亿元】 万科A早间公告：2018年12月份公司...	sina	0	2019-01-04	2019-01-04
2002	【万科：2018年实现合同销售金额6069.5亿元】万科早间公告称，2018年12月份公司实...	sina	0	2019-01-04	2019-01-04

3.2.3 - Sentimental Analysis

The step is where a sentimental score is assigned to each and every piece of news we collected, indicating how positive / negative this new is.

We use the API from Baidu AI Lab to accomplish this task, the result we get from Baidu indicates the probability that this news is positive.

	content	source	title	date	sentiment_score
177	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	sina	0	2019-01-01	0.839714
166	【2018年千亿房企达30家，恒大夺销售权益榜首位】据2018年12月31日克而瑞发布的《2...	wallstreetcn	0	2019-01-01	0.914582
167	【2018年千亿房企达30家 恒大夺销售权益榜首位】克而瑞发布的《2018年度中国房地产企业...	wallstreetcn	0	2019-01-01	0.839714
1257	云财经讯，万科A公告：2018年12月份公司实现合同销售面积438.7万平方米，合同销售金额...	yuncaijing	万科A公告：2018年12月份公司实现合同销售面积438.7万平方米	2019-01-04	0.163503
1254	云财经讯，万科A：2018年1-12月累计合同销售额6069.5亿元，合同销售面积4037....	yuncaijing	万科A：2018年1-12月累计合同销售额6069.5亿元	2019-01-04	0.366212

After obtaining the sentiment scores from the collected and filtered news, we can actually dump the news itself, only keeping the useful information and form a new pandas Dataframe.
We save the dataframe locally with proper naming to save us from pinging Baidu AI each time we want such data in the future.

	date	sentiment_score	source
177	2019-01-01	0.839714	sina
166	2019-01-01	0.914582	wallstreetcn
167	2019-01-01	0.839714	wallstreetcn
1257	2019-01-04	0.163503	yuncaijing
1254	2019-01-04	0.366212	yuncaijing

We do exactly the same thing with the news related to industry that we filtered in the last step.
One thing to note is that local saving is actually of much more importance here since we can avoid the extra work for future stocks in the same industry, in this case, real estate.

3.2.4 - Kernel Smooth Method to Get Score in Each Day

Now that we have obtained the sentimental scores for all news that are directly or indirectly related to the stock that we are studying, we further process these data.
First, we group by date, and then for each day, simply take the average of sentiment scores for all news that are published on that day, regardless of its source.

date	sentiment_score
2019-01-01	0.864670
2019-01-04	0.311315
2019-01-05	0.332797
2019-01-07	0.596884
2019-01-08	0.584160

Next, we create a dictionary mapping date to its sentiment score.

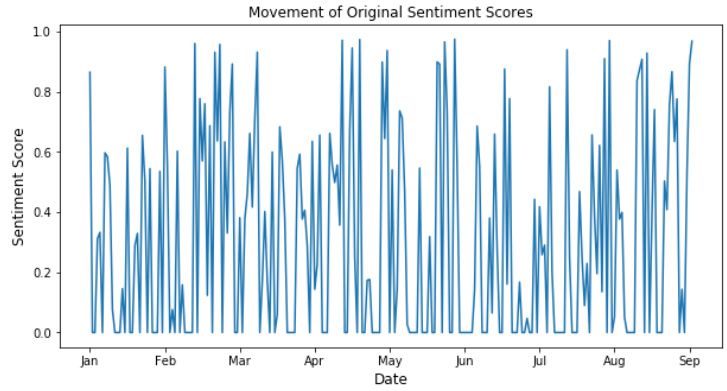
```
date_to_score_name = list(grouped_name.to_dict().values())[0]
date_to_score_name
```

```
{'2019-01-01': 0.8646699999999999,
 '2019-01-04': 0.31131512727272725,
 '2019-01-05': 0.33279713999999994,
 '2019-01-07': 0.596884,
```

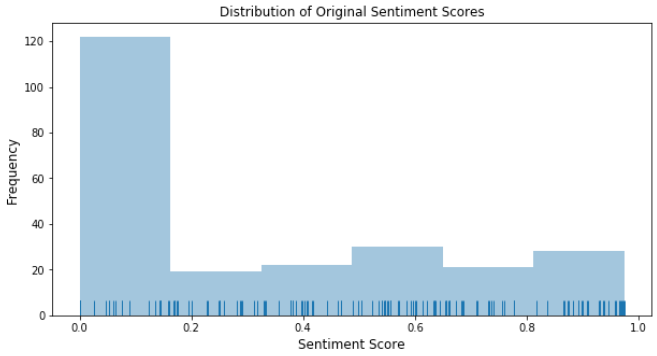
At this stage, not every single day has a sentimental score. Only those days when at least one piece of news is published has a score, which is not very sensible.

We can plot the scores for each and every day, and it is easy to observe that there are many days with score 0 and the curve is extremely volatile, not something desirable.

We can also take a look at the distribution at all the scores. There are roughly half of days that do not have a sentiment score.



We can also take a look at the distribution at all the scores. There are roughly half of days that do not have a sentiment score.



Here, I use the “one-sided” kernel smooth method to work around this issue and to comply with the actual situation.

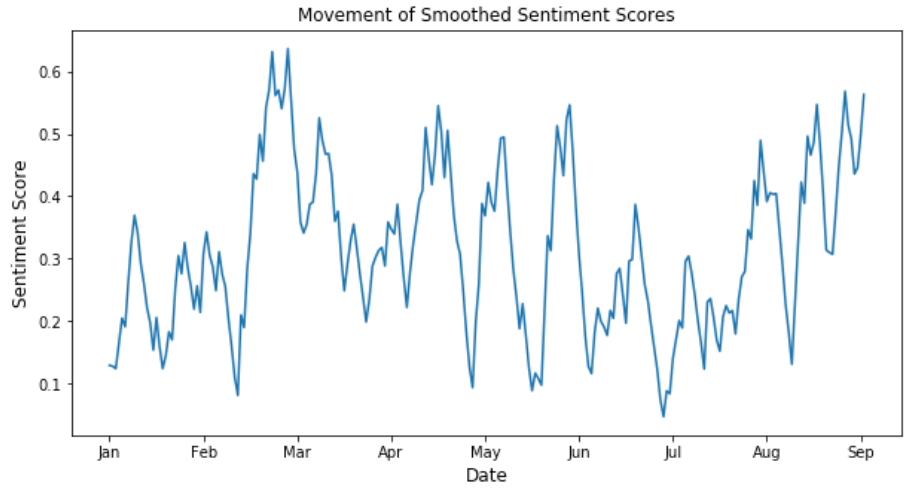
In real life, it is not the case that a news only has effect to the performance of the stock on the single particular day that this news is published. Instead, its influence should last for a couple of days and gradually decay to zero.

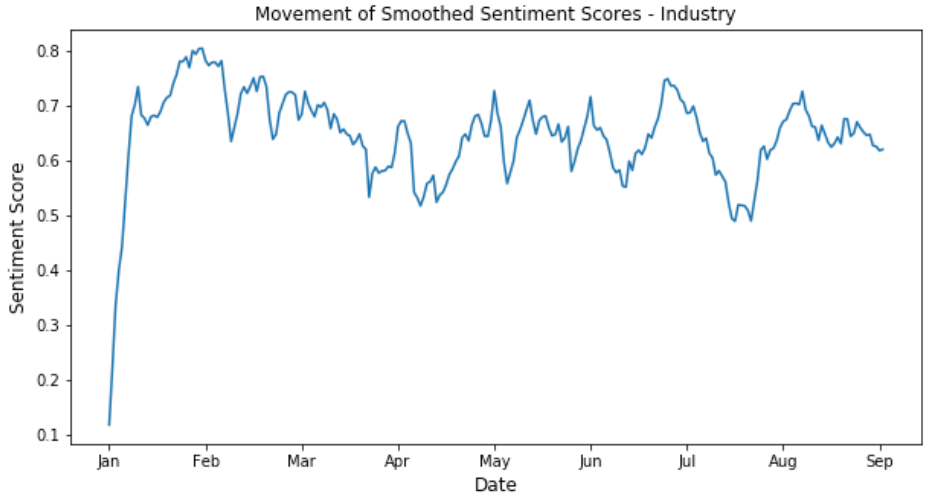
Three different kinds of kernels are implemented and are ready to be further tested upon.

One thing that is obviously different from the traditional kernel smooth method is that the kernel I implemented here is single sided rather than two sided.

```
def kernel_density(kind, width, distance):
    if distance > width:
        return 0
    if kind == 'uniform':
        return 1 / width
    if kind == 'square':
        return 1.5 * (1 - math.pow((distance/width) ,2)) / width
    if kind == 'triangular':
        return 2 * (1 - (distance/width))/ width
```

Width of the kernel is set to 10 currently, subject to future fine tuning.





We can plot the smoothed scores and the result is not only much more desirable but also more realistic when considering real life situation.

One observation from the above plot is that the curve experiences a somewhat peculiar sharp increase at the beginning. This can be explained since the for first 10 (or whatever the width we set) days, their scores should come from the 10 days prior to the starting date of our entire news data set, which is of course, missing. Therefore, this sudden increase is actually quite reasonable and suggest that we should exclude the first 10 days from our training.

4 - Market Data Collection and Input Preparation

Still using the first stock in our list, 000001, Vanke, as an example, we retrieve the daily market data (open prices, close prices, high, low, volume...) for the concerned period.

	ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
0	000001.SZ	20190830	14.29	14.39	14.10	14.16	14.13	0.03	0.2123	798500.57	1136119.659
1	000001.SZ	20190829	14.22	14.24	14.08	14.13	14.27	-0.14	-0.9811	609034.34	861177.482
2	000001.SZ	20190828	14.26	14.28	14.05	14.27	14.31	-0.04	-0.2795	829657.68	1176329.095
3	000001.SZ	20190827	14.36	14.48	14.24	14.31	14.25	0.06	0.4211	1356713.37	1948127.123
4	000001.SZ	20190826	14.42	14.50	14.15	14.25	14.65	-0.40	-2.7304	1415122.54	2018498.039

At this stage, for this particular stock, we have actually acquired all input data. All there left to do is to sort them through and build a dataframe for further training and testing.

However, before doing that, several important simplifications and assumptions need to be made:

1) What we are trying to predict here is only the percentage change of some particular stock. We consider this as a key measurement for how well the stock is performing and as an indicator for our trading strategy in later sections.
Everything else including open price, close price and trading volume is regarded as side information and not predicted. (Some of them might be used for prediction, though.)

Special Note:
As mentioned before, since we intend to trade "close to close", the real percentage change that we should focus on is the close price to close price percentage change.
Fortunately, if we take an in-depth look at the dataframe we obtained using Tushare, we can actually find out that the "pct_chg" column given in the table is already what we want.
For instance, from 20190829 to 20190830, the price of Vanke A goes up from 14.13 to 14.16 (close to close), and the percentage change given, 0.2123 is precisely 0.03 divided by 14.13.

	ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
0	000001.SZ	20190830	14.29	14.39	14.10	14.16	14.13	0.03	0.2123	798500.57	1136119.659
1	000001.SZ	20190829	14.22	14.24	14.08	14.13	14.27	-0.14	-0.9811	609034.34	861177.482
2	000001.SZ	20190828	14.26	14.28	14.05	14.27	14.31	-0.04	-0.2795	829657.68	1176329.095
3	000001.SZ	20190827	14.36	14.48	14.24	14.31	14.25	0.06	0.4211	1356713.37	1948127.123
4	000001.SZ	20190826	14.42	14.50	14.15	14.25	14.65	-0.40	-2.7304	1415122.54	2018498.039

Therefore, we can directly use data in the percentage change column given in the table and there is no need to manually calculate our own percentage change.

2) We assume that the performance of a stock is related to only the recent news and recent stock market data.
For now, we define the word "recent" as "last three trading days" for market data and "last three days" for news.
Of course, such assumption and definition are extremely arbitrary and is subject to possible great future changes to improve the performance of models.

3) We further simplify the market data at each day to three numbers: close price, percentage change and trading volume.
Here, as mentioned in the second point, the percentage change here is the ratio of the close price today over the close price of the last trading day.

To begin with, we start off simple and easy, we only use recent news sentiment scores (and not include recent market data) to predict the percent change of trading days.

Nevertheless, as other market data might be used for future model complication, we prepare 16 numbers (namely the close prices, percent changes and trading volumes of the previous three trading days and sentiment scores of the news in the past three days from both name-specific news and industry-specific news, plus 1 for output, i.e., percentage change on that day) for each and every trading day from the beginning of January to the end of August.

```
print(final_dataframe.shape)
final_dataframe.head()
```

```
(157, 16)
```

	close-1	close-2	close-3	pctchg-1	pctchg-2	pctchg-3	vol-1	vol-2	vol-3	name-1	name-2	name-3	industry-1	indutrys-2	industry-3	pctchg-0
0	9.94	9.66	9.74	2.8986	-0.8214	-0.1026	1233486.36	402388.11	865687.66	0.369590	0.327061	0.262958	0.701542	0.680709	0.607741	1.6097
1	10.10	9.94	9.66	1.6097	2.8986	-0.8214	1071817.66	1233486.36	402388.11	0.341526	0.369590	0.327061	0.734110	0.701542	0.680709	0.9901
4	10.20	10.10	9.94	0.9901	1.6097	2.8986	696364.55	1071817.66	1233486.36	0.221808	0.260669	0.292360	0.664182	0.676530	0.682409	-0.8824
5	10.11	10.20	10.10	-0.8824	0.9901	1.6097	500443.59	696364.55	1071817.66	0.197657	0.221808	0.260669	0.679139	0.664182	0.676530	1.2859
6	10.24	10.11	10.20	1.2859	-0.8824	0.9901	542160.55	500443.59	696364.55	0.154045	0.197657	0.221808	0.681913	0.679139	0.664182	2.3438

5 - Building and Applying the Data Preparation Pipeline

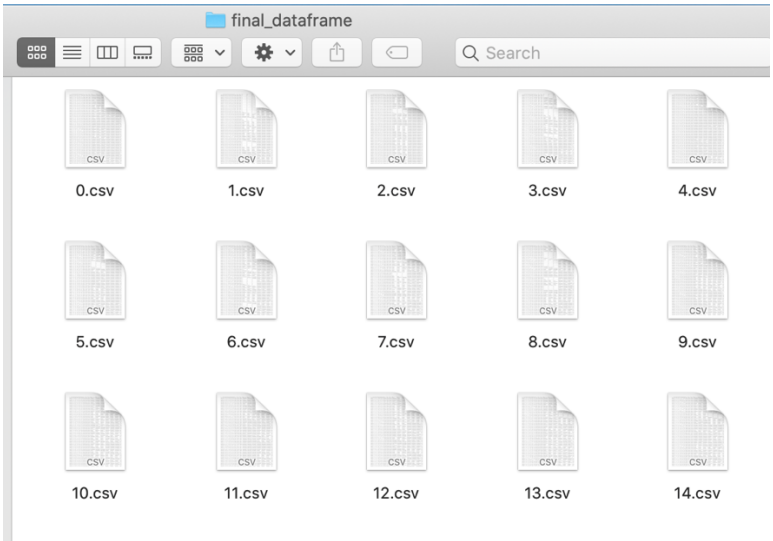
Now, we have gone through the entire process of preparing the input and output data for one stock.

Since eventually we will need to prepare such data for each and every stock from the fifty stocks of interest, it would greatly ease our life if we define several functions to modularize the whole process and build a pipeline for this data preparation process.

After all functions been defined for each step along the way, we can simply define another function to summarize the whole process, within which we call those functions defined above and meanwhile deal with some tiny details that may cause error.

```
def build_dataframe(i):

    # Filter the 380k news and extract those related to either this particular stock or industry
    name_appeared = filter_news(stock_names_chn[i])
    industry_appeared = filter_news(stock_industries[i])

    # Sort According to Date
    name_appeared = sort_by_date(name_appeared)
    industry_appeared = sort_by_date(industry_appeared)

    # Get sentiment scores for name appeared news and industry appeared news
    name_sentiment_result = get_sentiment_score(name_appeared, stock_names_chn[i])
    industry_sentiment_result = get_sentiment_score(industry_appeared, stock_industries[i])

    # Get smoothed scores
    smoothed_scores_name = get_smoothed_score(name_sentiment_result)
    smoothed_scores_industry = get_smoothed_score(industry_sentiment_result)

    # Get Five Dicts and Build Dataframe
    market_data = pro.daily(ts_code=stock_codes[i]+'.SZ', start_date='20190101', end_date='20190831')
    if market_data.shape[0] == 0:
        market_data = pro.daily(ts_code=stock_codes[i]+'.SH', start_date='20190101', end_date='20190831')
    final_dataframe = build_final_dataframe(*build_five_dicts(market_data, smoothed_scores_name, smoothed_scores_ind

    final_dataframe.to_csv('final_dataframe/' + str(i) + '.csv')

    return final_dataframe
```

Finally, we call the above function for each stock and get a data matrix for each stock that we save locally for the model training and testing in the next section.



6 - Training and Testing Linear Regression Models

As mentioned in section 4, the very first thing we try is using recent news sentiment scores (6 numbers only) to try to predict the percent change of trading days.
To start with, we use the most simple and fundamental model: linear regression.

6.1 - Universal Model

Based on the thought that the underlying logic of using recent news and market data to predict current stock performance should be homogenous rather than heterogenous for different stocks, the first attempt is aggregating all data from all stocks first and then training a universal model for prediction.

```
l = []
for i in range(50):
    l.append(pd.read_csv('final_dataframe/'+str(i)+'.csv'))
all_data = pd.concat(l)
print(all_data.shape)
all_data.head()
```

```
(7828, 17)
```

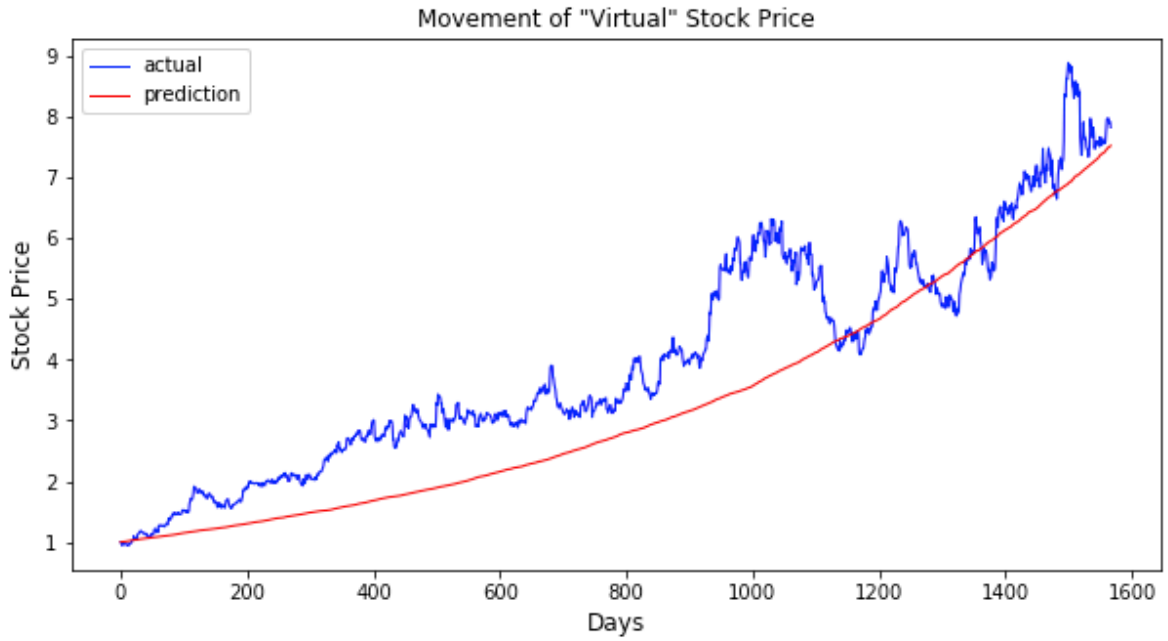
	Unnamed: 0	close-1	close-2	close-3	pctchg-1	pctchg-2	pctchg-3	vol-1	vol-2	vol-3	name-1	name-2	name-3	industry-1	industry-2	industry-3
0	20190110	25.33	25.00	25.05	1.3200	-0.1996	0.4813	340140.62	214382.02	427154.85	0.447473	0.335565	0.310806	0.718453	0.675728	0.603760
1	20190111	25.11	25.33	25.00	-0.8685	1.3200	-0.1996	224649.20	340140.62	214382.02	0.478322	0.447473	0.335565	0.729811	0.718453	0.675728
2	20190114	25.30	25.11	25.33	0.7567	-0.8685	1.3200	240694.90	224649.20	340140.62	0.345607	0.395677	0.435481	0.664394	0.678656	0.686718
3	20190115	25.02	25.30	25.11	-1.1067	0.7567	-0.8685	183281.34	240694.90	224649.20	0.286875	0.345607	0.395677	0.671603	0.664394	0.678656
4	20190116	25.16	25.02	25.30	0.5596	-1.1067	0.7567	368758.85	183281.34	240694.90	0.255294	0.286875	0.345607	0.688667	0.671603	0.664394

There are 7828 data points in total, we shuffle the data and randomly split into training set and testing set with the ratio of 4:1.
Then, a linear regression is fitted on the training data and tested on the testing data.

```
x = 1

p_actual = [1]
p_pred = [1]

for y_t in y_test:
    p_actual.append(p_actual[-1] * (1+(0.01*y_t)))
for y_p in y_pred:
    p_pred.append(p_pred[-1] * ((0.01*y_p)+1))

plt.figure(figsize=(10,5))
plt.title("Movement of \"Virtual\" Stock Price")
plt.xlabel('Days', fontsize=12)
plt.ylabel('Stock Price', fontsize=12)
plt.plot(p_actual, linewidth=1, color='blue', label='actual')
plt.plot(p_pred, linewidth=1, color='red', label='prediction')
plt.legend();
```

Here, to visualize how accurate our prediction after the entire dataset being randomly shuffled, we create a "virtual stock" by simulating its actual price using actual percentage change data (which comes from different stocks and is not sorted in chronological order) and simulating its predicted price using corresponding predicted percentage changes.

The result looks like the following:



We can see that although the prediction can somewhat capture the main trend (going up) of the actual movement, clearly it appears way too stable compared to the actual fluctuation of stock prices.
The reason is probably that we have used 6k~7k data points to train a linear regression with only 6 features. It is thus reasonable that we end up with the prediction algorithm being inordinately "conservative", in the sense that it will always predict a gentle, steady, slow increase no matter what input is given.

6.2 - Individual Models

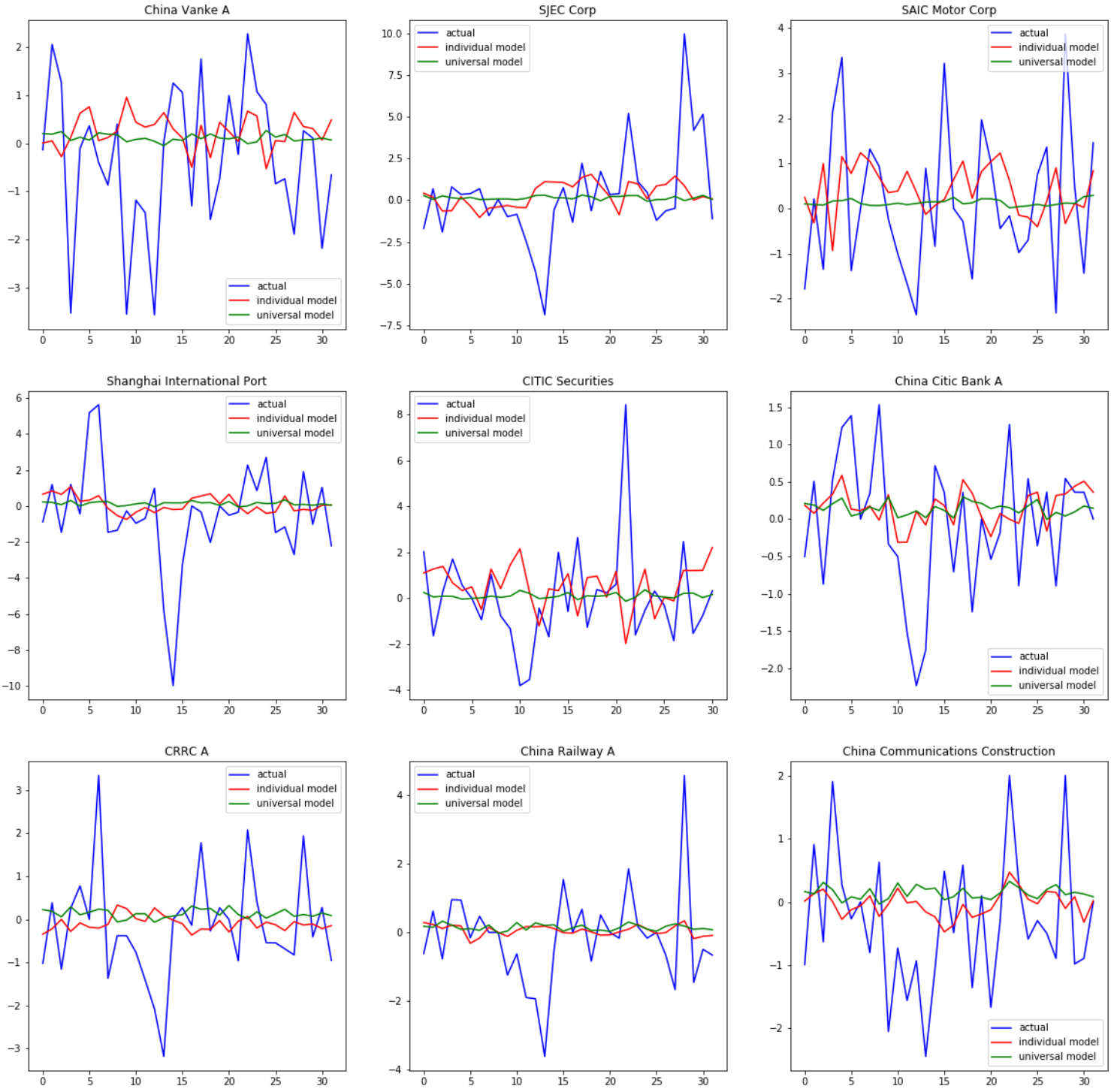
6.2.1 – Percentage Change Prediction Results

Seeing that the result of training one universal model after aggregating data from all stocks is not that desirable, we try a different approach and train individual models for each stock.
Again, we use the simplest linear regression model, but instead of training one model over 7828 data points, we train 50 models, each over around 130 data points.
Here, since the testing set contains only around 30 data points, we can actually plot the predicted percentage changes and compare to the actual numbers. (Note that this is not possible in the previous section since 1000 data points would be too versatile can no meaningful conclusion can be drawn).

So, we compare the actual percentage changes, the predicted percentage changes using universal model and the percentage changes using the individual model.

The result looks like the following (only the results for the first 9 stocks are displayed):

We can observe that the individual model is clearly better than the universal model generally speaking. It is indeed capable of capturing some of those fluctuations.

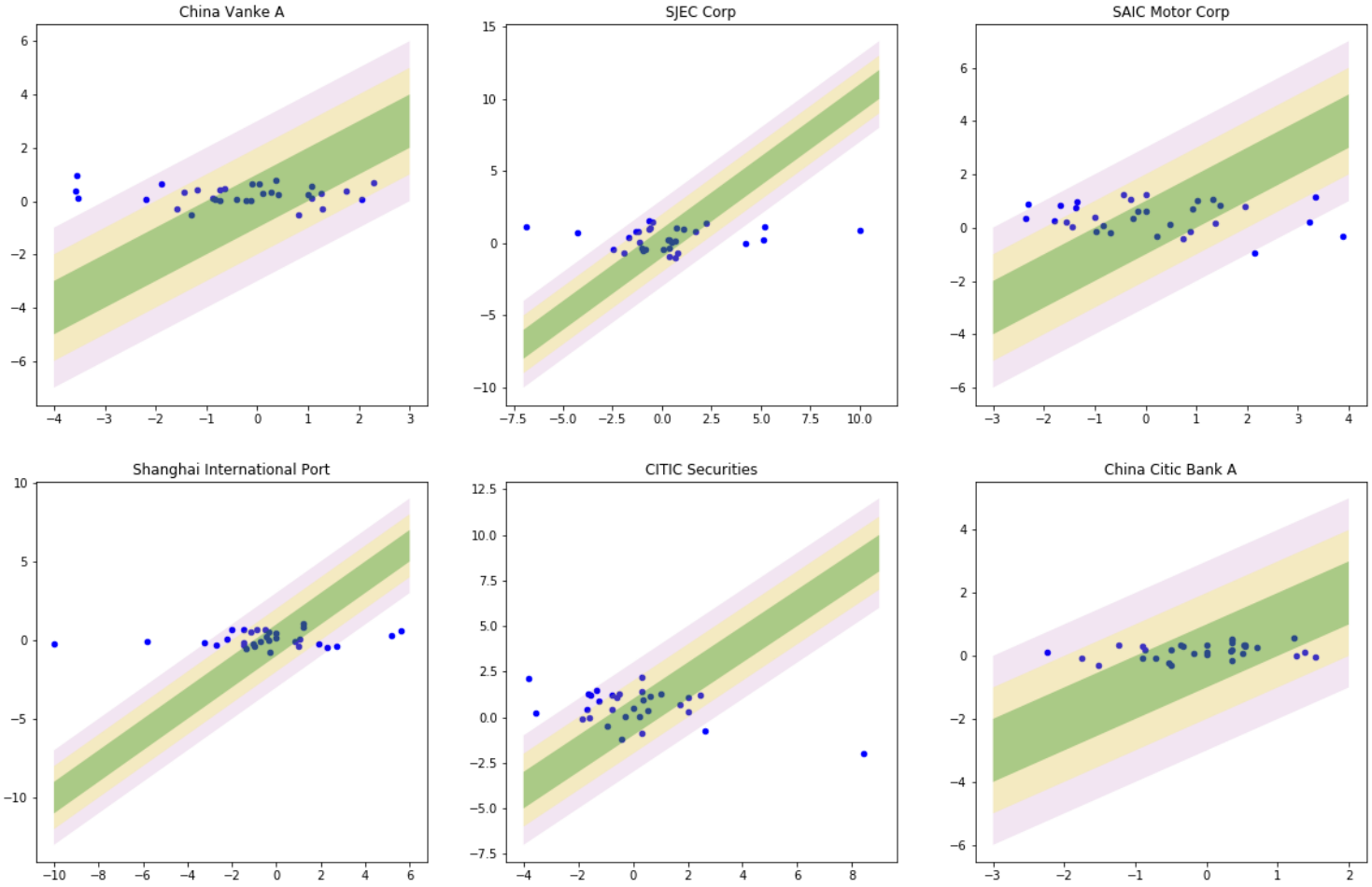


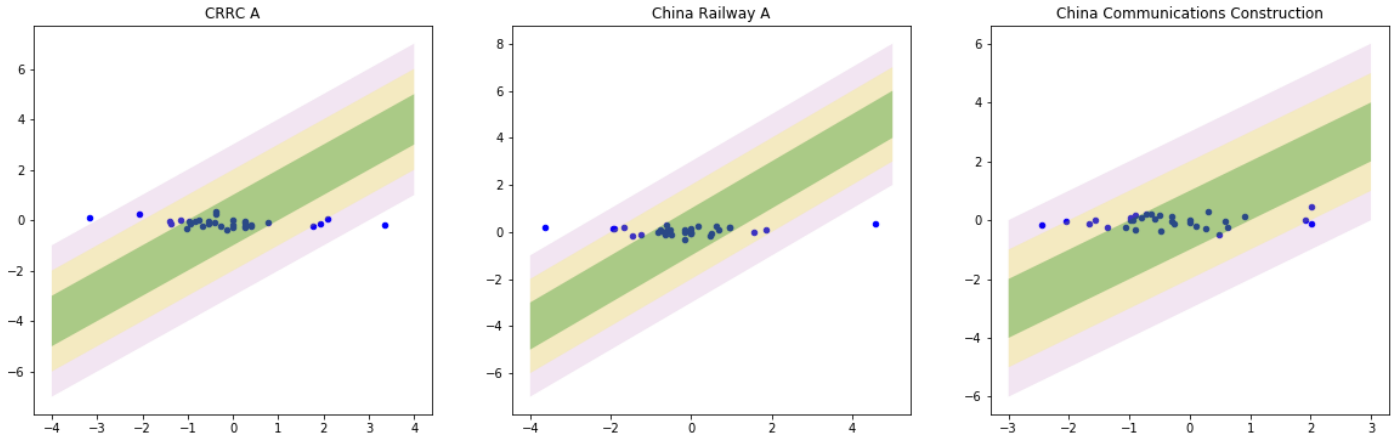
However, both predicted results are too “stable”, there are way less fluctuations and variance in the predicted percentage changes compared to the actual situation.

Another thing is that the changes predicted using individual model appears to be one or two days "later" than actual changes, which is also reasonable due to the untimeliness nature of nature.
The stock may respond immediately to events, market emotions and public opinions, but it would take time for the corresponding news to be created, edited, published and then used to predict prices.
Given that our model only uses news in the previous three days (not inclusive of today) to predict the percentage change today (which is sensible since you cannot know the news today in advance), this result is anticipated.

We can also plot the results using scatter plots to get a sense of how accurate our results are compared to the actual situation.

Below:
X-axis is the actual percentage changes; Y-axis is the predicted percentage changes individual models).
Green: +/- 1%, Yellow: +/- 2%, Purple: +/- 3%





We can see that while almost all points fall into the purple region (except some outliers), there are not so many points in the green region, meaning that the predicted percentage changes and the actual values indeed have some differences.

6.2.2 – Price Movement Prediction Results

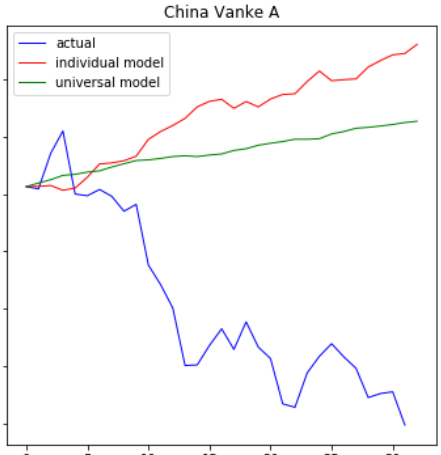
Here, because during the training of individual model, no random shuffling is performed on the dataset and chronological order is reserved, the testing data is actually the set of some consecutive trading days. Therefore, we can actually compare the "calculated price using predicted percentage change" to the actual price movements of any particular stock.

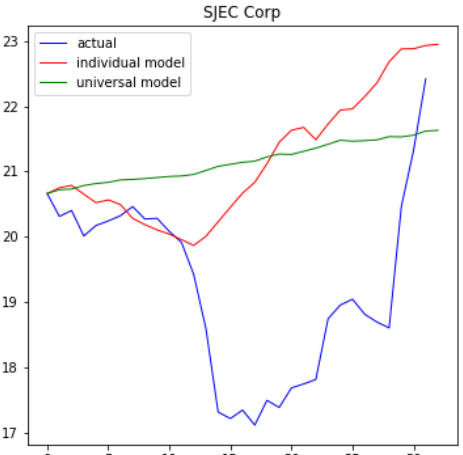
Result is shown below.

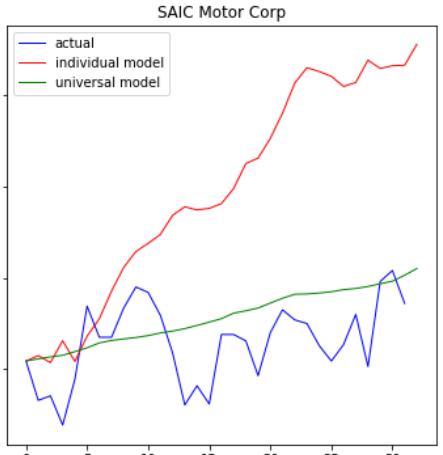
Note that here the actual price movement curve is plotted by simply linking (connecting by straight line) the dots of actual close prices each day and is hence not so accurate compared to the real-life situation where the price changes every second.

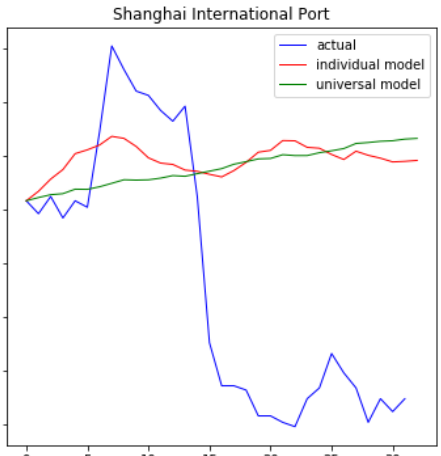
We can see that though the individual model may be slightly better, both models actually perform quite badly when we draw out the price movement curve.

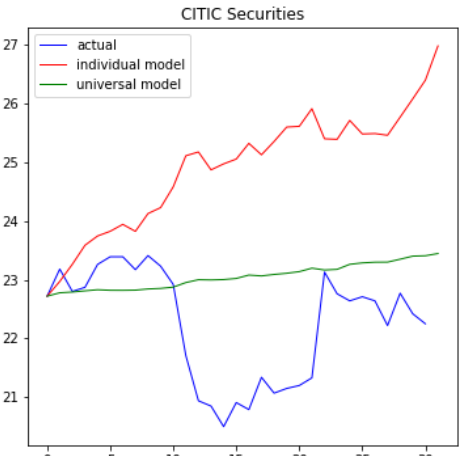
One explanation of this would be that our prediction solely focuses on percentage changes. And any error, any deviation from the actual changes would be accumulated and then "amplified" in a sort of exponential way.

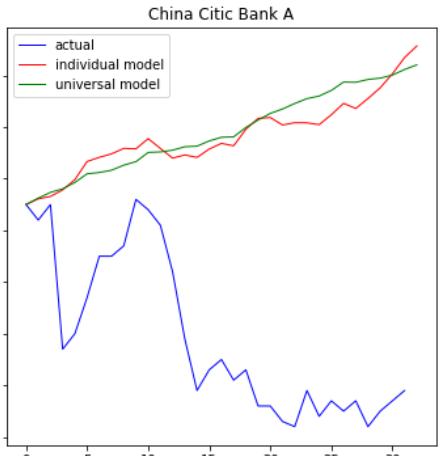

China Vanke A
actual
individual model
universal model

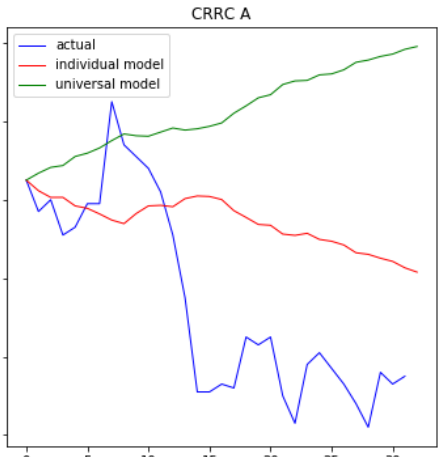

SJEC Corp
actual
individual model
universal model

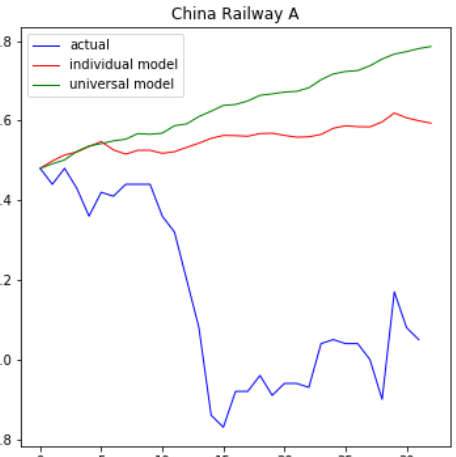

SAIC Motor Corp
actual
individual model
universal model

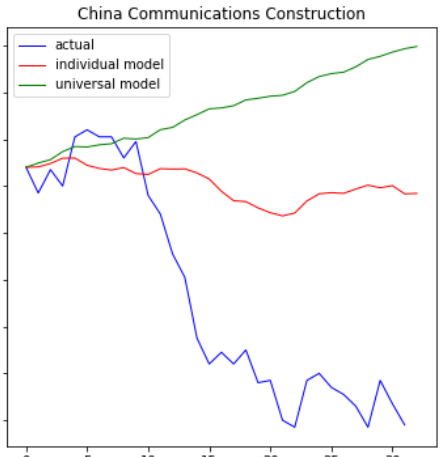

Shanghai International Port
actual
individual model
universal model


CITIC Securities
actual
individual model
universal model


China Citic Bank A
actual
individual model
universal model


CRRC A
actual
individual model
universal model


China Railway A
actual
individual model
universal model


China Communications Construction
actual
individual model
universal model

7 – Using Model to Actually Trade

Although our model is very naïve, simple and inaccurate, we implement the trading simulation and test its performance anyways.

A model may perform poorly in terms of giving inaccurate predictions of the absolute values it intend to predict (in this case, percentage changes), nevertheless, it may be a different story as how it performs in terms of capturing the differences, or relativity across different inputs (in this case, how predictions of different stocks compare to each other). In other words, though the absolute value prediction may be poor, it is still likely for the model to be powerful in term of distinguishing those stocks with potential and high probability to increase in prices.

The trading simulation is pretty straight-forward and goes like the following:

Since the entire August is well within the testing dataset of individual models for all stocks, we can use the model to trade daily in August.

We start with 1 million RMB, for each day, we predict the percentage changes for all 50 stocks, long the top **N** stocks (by top, we mean highest predicted percentage changes), and short the bottom N stocks. This way, the market risk may be perfectly hedged in some way.

Furthermore, assume that we always all-in with whatever the amount of fund we get left each day and we always invest half of our money to long and the other half to short. Besides, assume that our fund is split equally on the **N** stocks that we long / short each day.

In the end, we check how much money we have left in our fund.

As a reference, for each N, we also randomly long N stocks, randomly short N stocks and repeat such random strategy three times and take average to see its performance.

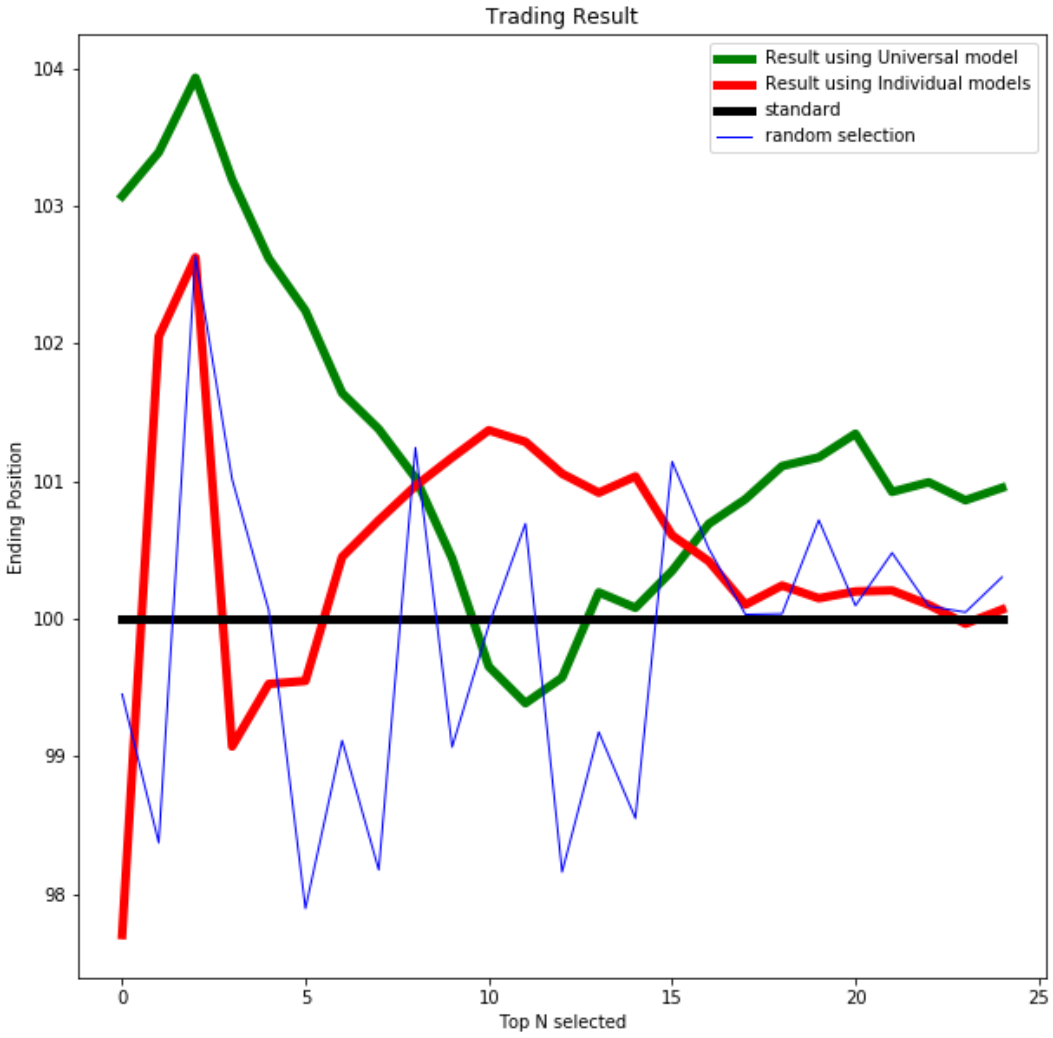
Following is the result of testing N from 1 all the way up to 25 (values above 25 introduces overlap of long and short and is hence not that meaningful to be tested):

The first thing we can observe is that both the green line and the red line experience much less fluctuations going up and down around the standard and appear much more stable compared to the blue line (the average of random selection trading).

Secondly, although sometimes the green and the red fall below the 100-standard line, generally speaking, there is no huge problem stating that they both outperform the blue line, in the sense that are above the blue line in most cases.

Thirdly, if we were to extract the average height of the curve (make positive and negative cancel out), (or more mathematically integrate along the curve), it is obvious that we would definitely get a positive result for green and red and perhaps around zero for blue.

In conclusion, our prediction does seem to have a little bit edge over the dummy trading and "no prediction at all" strategy.



However, there are indeed some peculiar data points in the above plot that worth a digging deeper and take a second look.

For example, the red line starts at somewhere quite low (below 98) and we may draw out its entire performance curve during this month to try to figure out the reason of this situation.

After plotting out the curve, we see that the portfolio goes down sharply at the very beginning and never successfully to goes up beyond the standard ever since.

A further examination reveals that this is caused by the algorithm constantly picks the stock CITC Securities at the beginning few days (which is a huge mistake) and it basically affects the performance in so large a way that no future attempts to bring the performance back up can remedy this mistake.

What this suggests is that when N is extremely small, the overall performance of the portfolio constructed daily by our algorithm can be seriously affected by its tendency of making mistakes on one or two particular stocks and its prediction on other stocks, however accurate, cannot help to reverse the situation due to the small N setting.

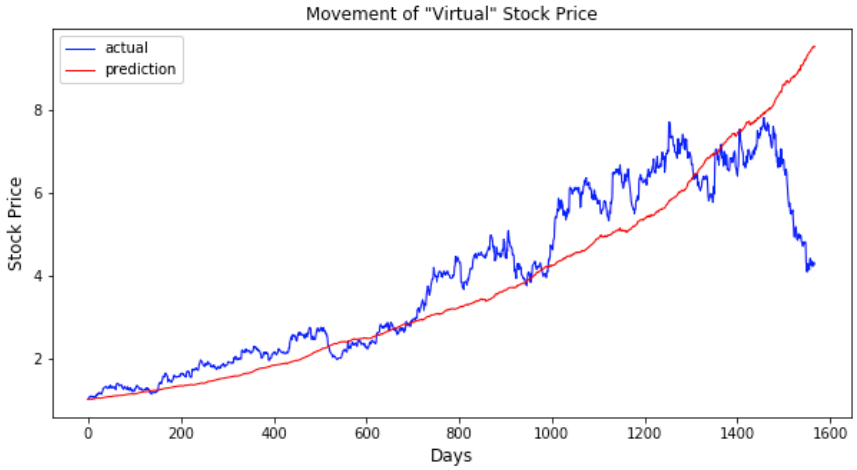
8 – Include History Market Data

Sticking with the simplest linear regression model, we try to include more features from the history market data for prediction this time.

This time, instead of only 6 sentiment scores, we use 15 numerical numbers for prediction. Namely, we use the open prices, percentage changes and trading volumes of previous three *trading days* and the six numbers already in use (news sentiment scores) to predict the percentage change today.

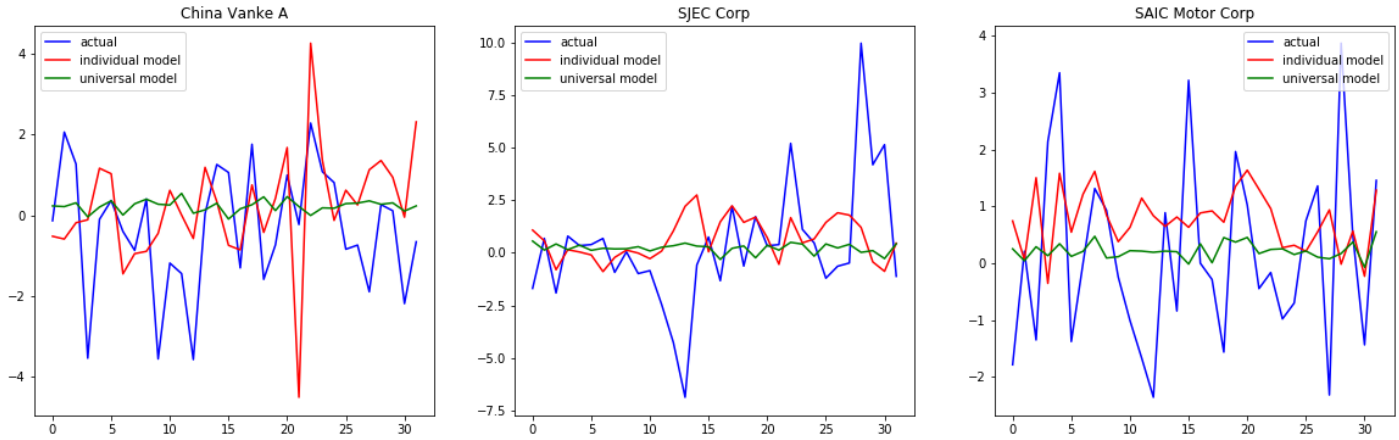
Basically, we use everything we prepared in the final dataframe (previously in section 5) to train and test a new linear regression model.

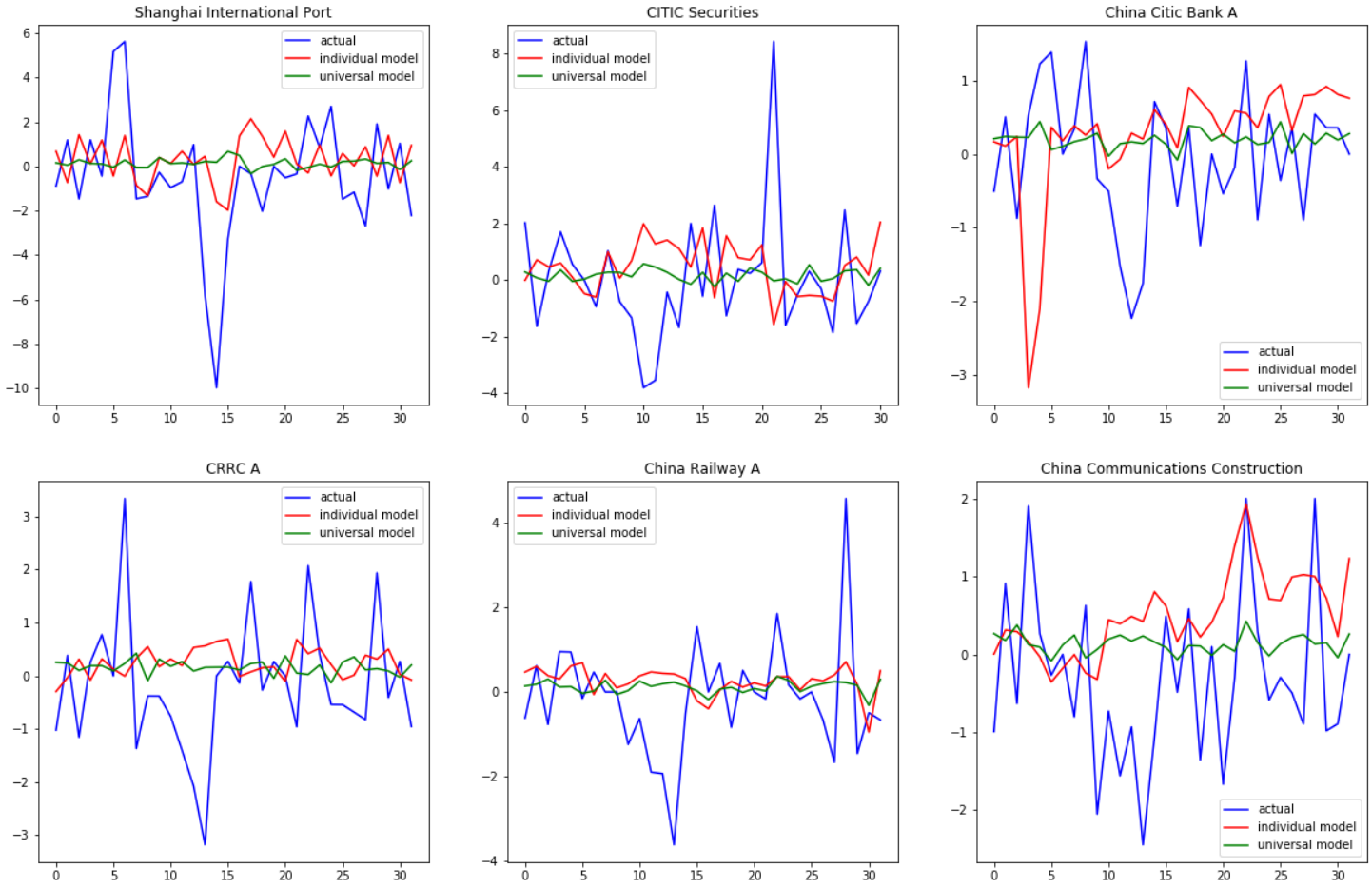
We display the results like section 6.



The result using universal model is largely the same as before. (6k-7k points is still way too much for 15 features).

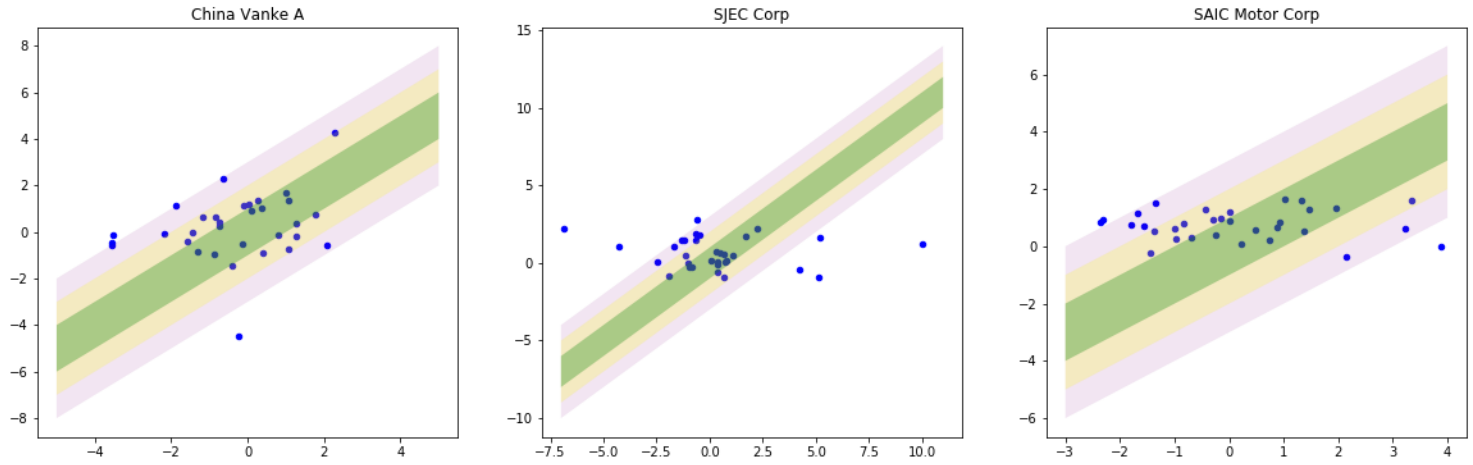
Next, the prediction of percentage changes.

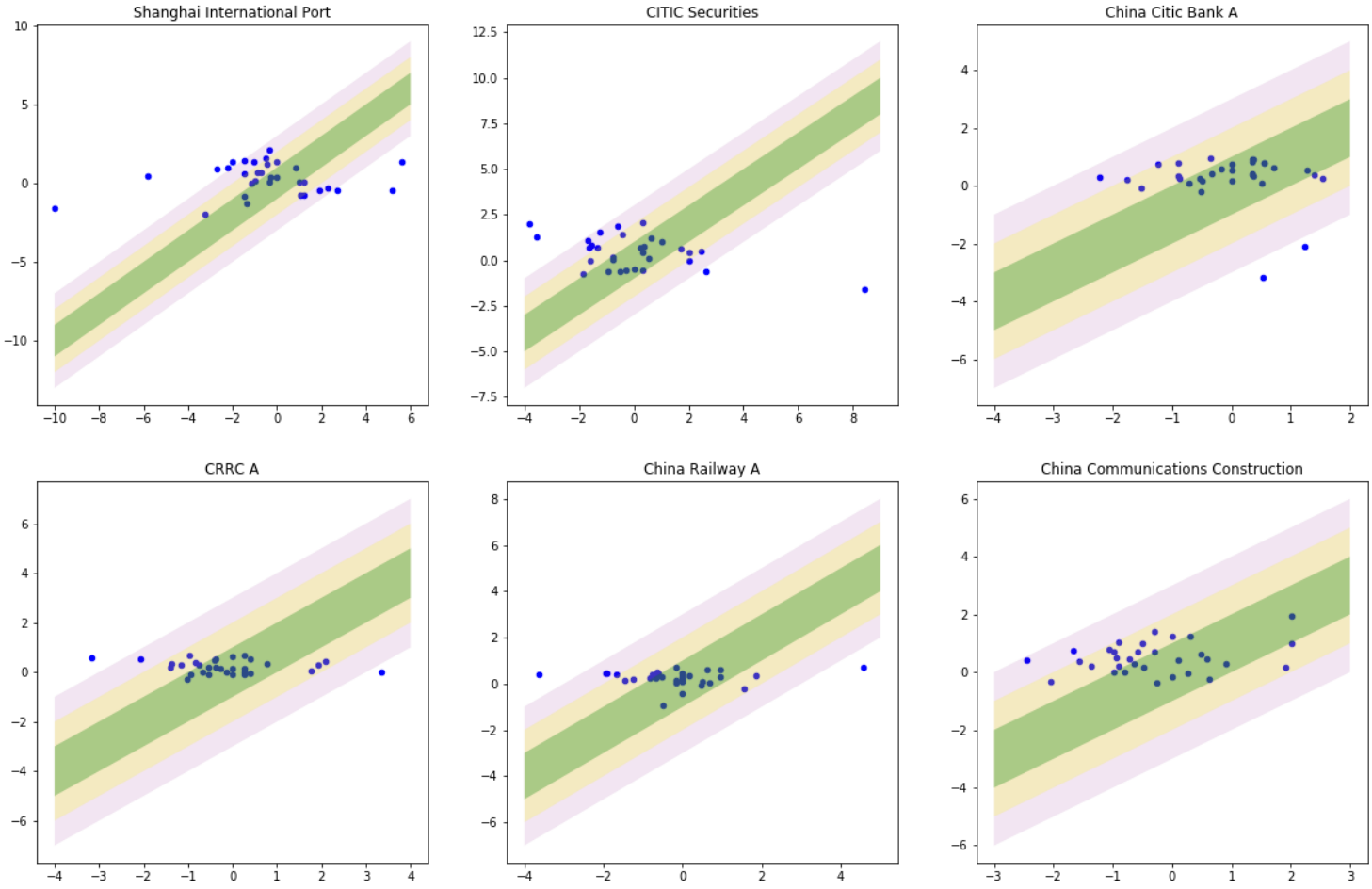




It seems with the addition of more features, the prediction of individual models become slightly more volatile and probably as a result of this, somewhat more capable of capturing the ups and downs of the actual movements.

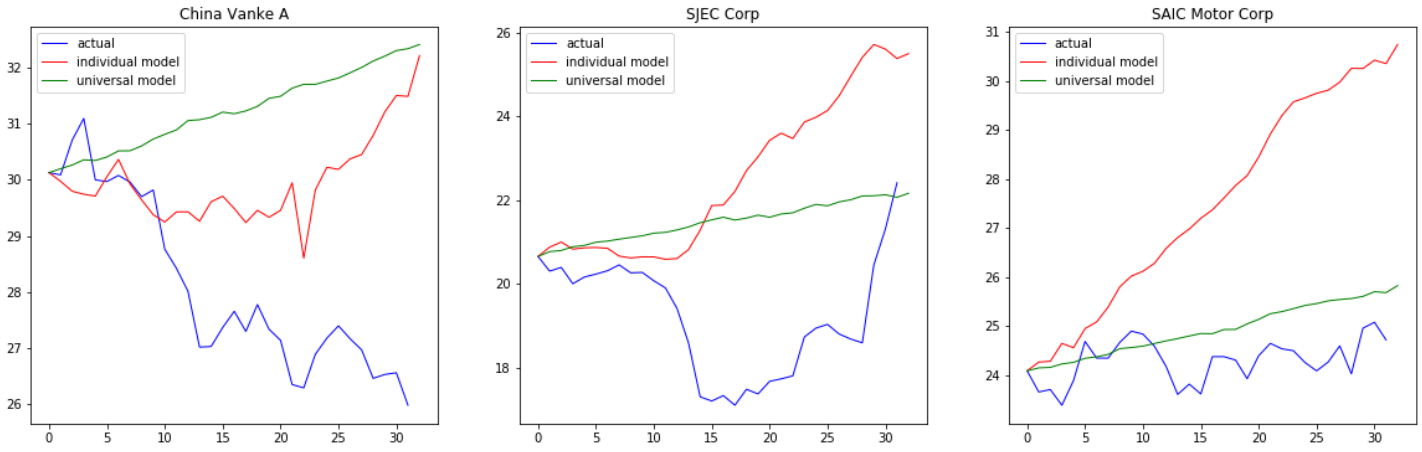
We then plot the scatter plot like previously.

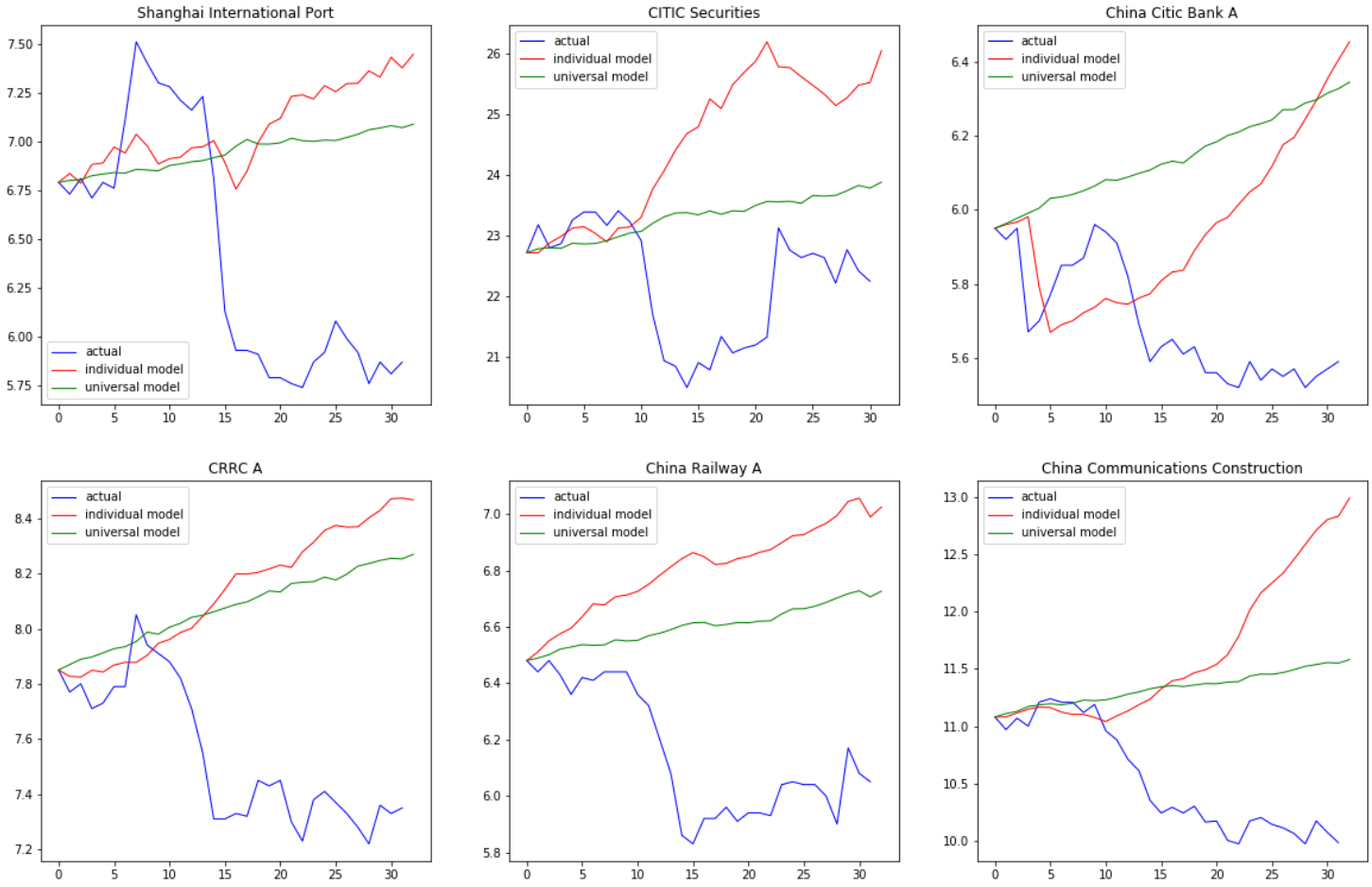




In general, there does not exist an obvious change that the dots converge more to the green region (which should be the case if the model is greatly improved by the addition of new features).

Next, we plot the prices calculated using the predicted percentage changes.





Again, there does not seem to be any discernable improvement. The predicted model, whether it is the universal one or the individual one, always tends to be over-optimistic and give a price that constant increases steadily.

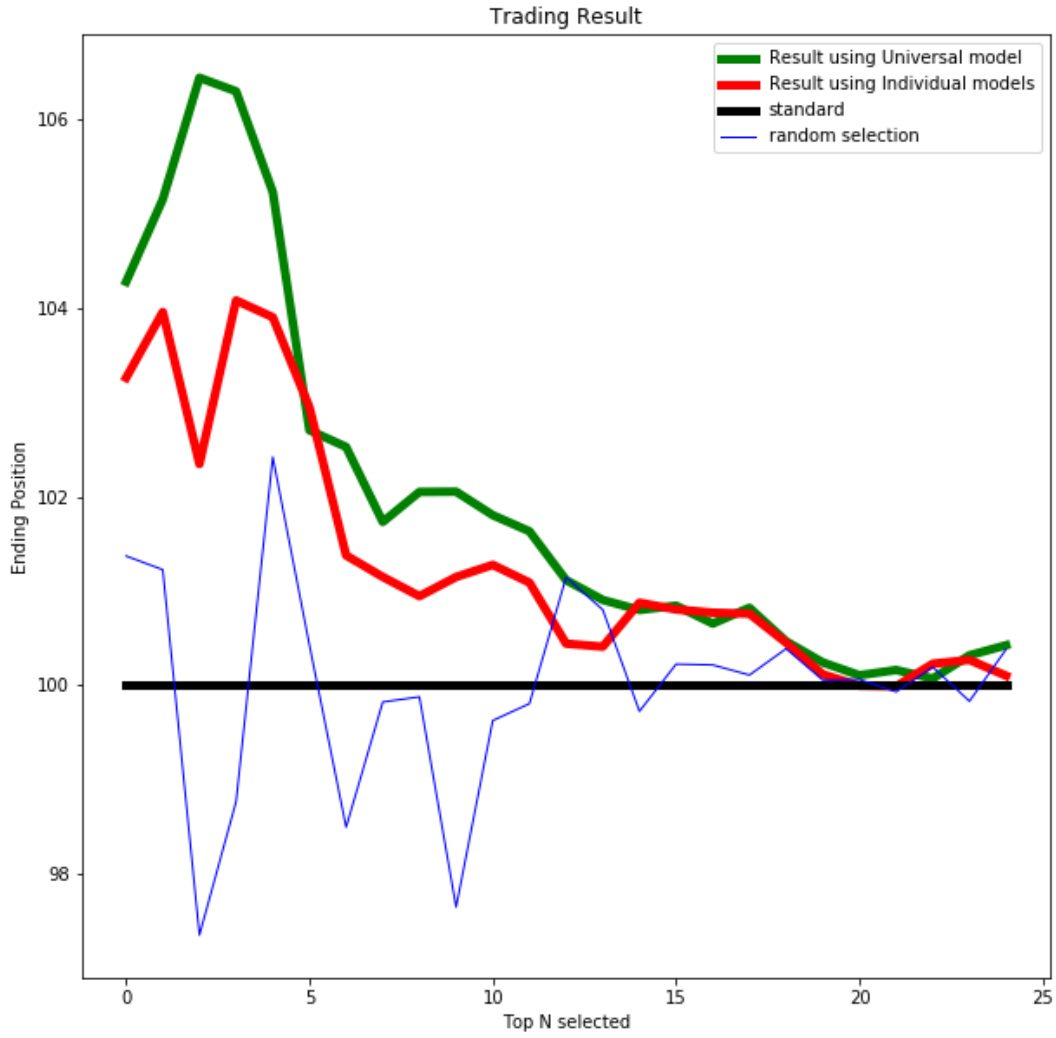
We put the models into trading and long short stocks with equal amount of money on each day to see our results like section 7.

This time, the green line and the red line both, not only absolutely outperform the blue line in almost all cases, but also constantly stay above the 100-standard line, which is clearly an improvement over the previous section where both curves will fall below 100 at some point.

To sum up, with the introduction of new features like market data, though the model may not have improved greatly in terms of predicting absolute values of percentage changes, its ability to identify those promising stock does seem to increase in a noticeable way.

However, though the performance of simulated trading is indeed improved, what role do these extra new features play in the training and prediction process and how significant are each of the new added features remain unclear and something definitely worth further investigation.

One possible explanation to the improved performance of universal model is that with the increased number of features, the problem of the data points being over-abundant become less severe relatively, which enables the prediction to be more “bold” and less “conservative” (as can be observed from the percentage change prediction graph itself) and thus more heterogenous across different stocks.

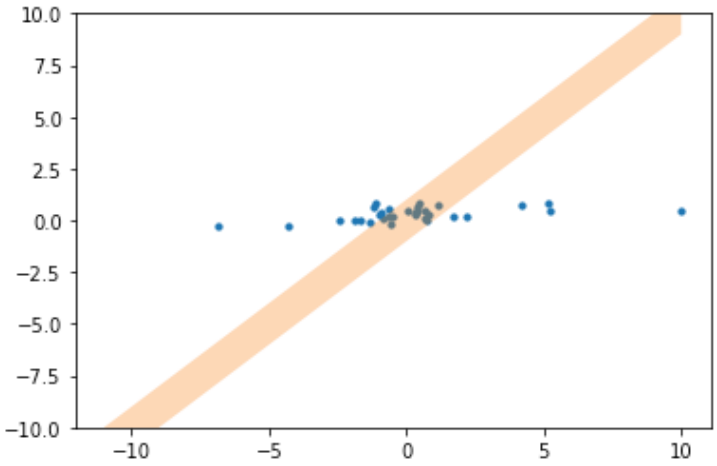


9 – Training and Testing Neural Networks

Now that we have explored much using the simplest linear regression model, we try a basic neural network this time, which should be more sensible since the relationship between news sentiment scores and stock performance is probably non-linear.

Quite strangely, once I started to implement the neural network, the output converges to constant zero in an exponential speed. This is the case with both universal model and individual models.

Following is the scatter plot using the same logic as before.



The basic neural network structure (using RELU as activation function, one hidden layer with 5 neurons), after training, tends to always give a prediction of zero, regardless of the input given.

The reason of such peculiar behavior is yet to be investigated, but this is clearly not what we desired.

10 – Rolling Window Framework

As mentioned in previous sections, one of the major problems with the universal model is that the simple model tends to be "overwhelmed" by a relatively large amount of data points (which inevitably contains way too much noise). This is also evidenced by the obvious improvement after introducing more features and thus complicating the model.

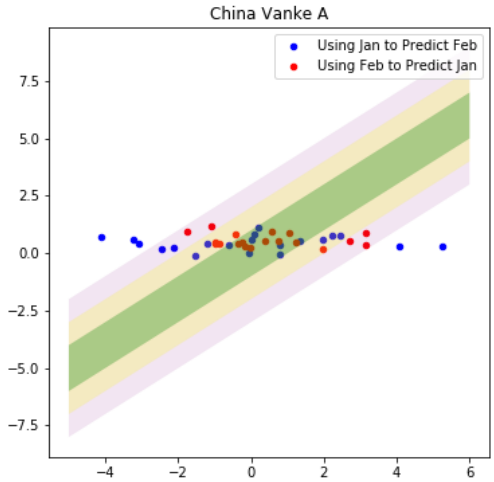
In this section, we try to introduce the dynamic rolling window of training and testing our model, which might in some way solve the above-mentioned problem

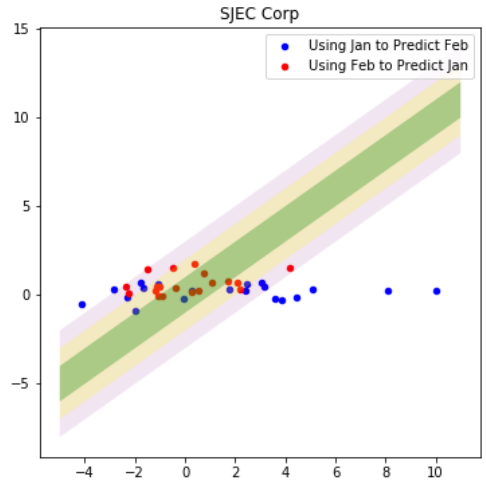
Basically, instead of using around 6k data points from 6 months all at once, we only use around 1k data points from one month to train a linear regression model at first. Then, at the end of each one-month period, we review our model and make changes to its parameters so that hopefully it can make better predictions.

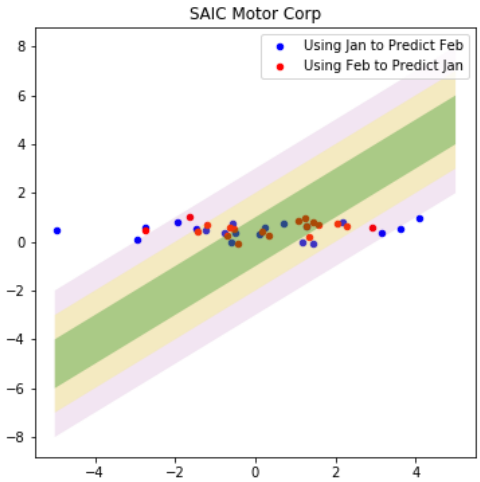
This way, the entire dataset is fed into the linear regression in a streaming way instead of a "flooding in" way. Besides, this approach also better mimics the real-life situation where a machine learning model is first implemented, applied and then constantly and periodically tested, reviewed, amended and improved with new data being obtained new results become available daily.

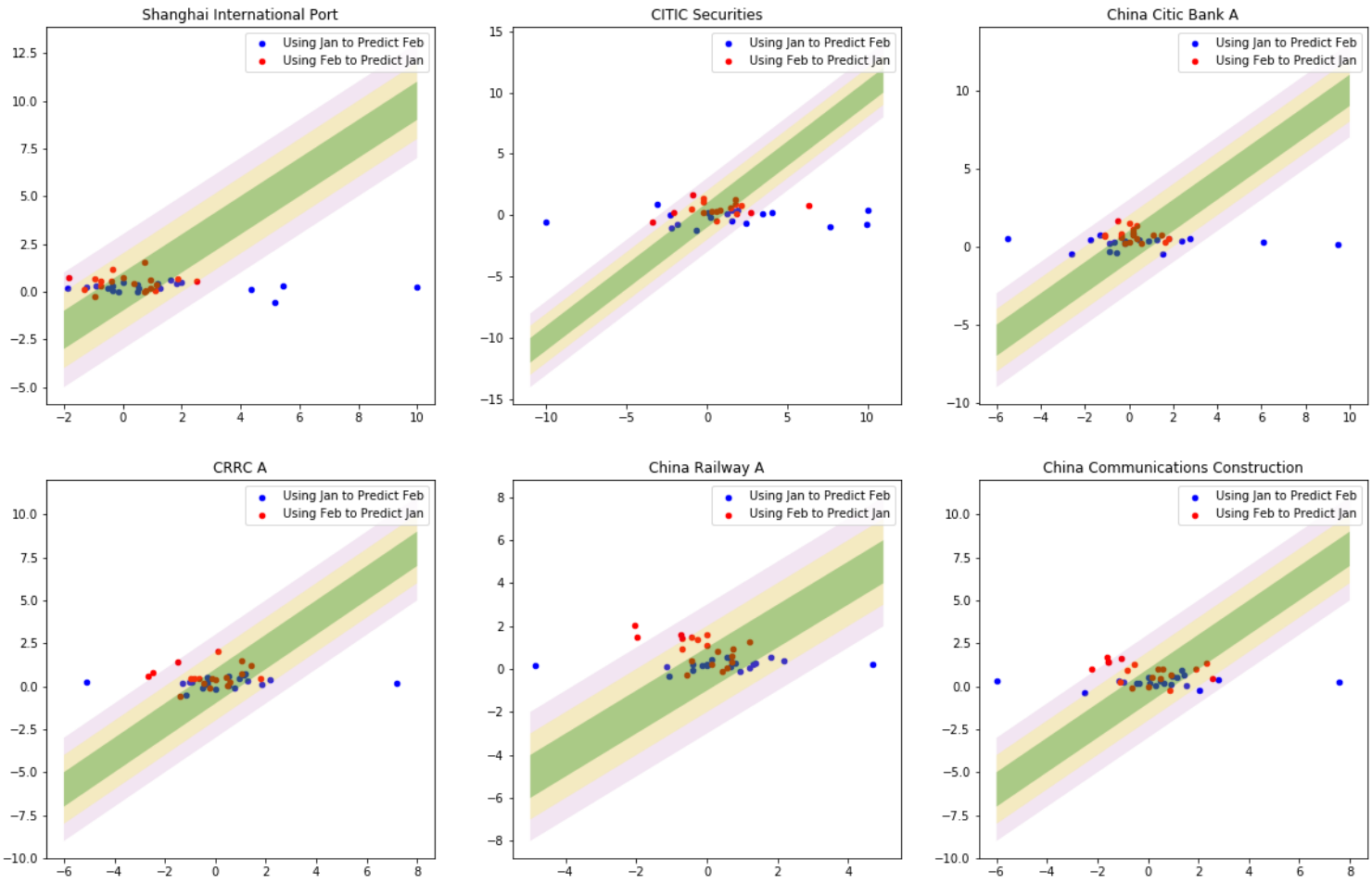
So, we first limit ourselves to the January data (from 20190101 to 20190131) aggregated from all fifty stocks and try to fit a linear regression model. Then, at the end of each of the following one-month period, we fit a new model using the newly obtained one-month data, compare it to the one currently in use and select the superior one out of two.

For example, at the end of February, we can compare how the February model performs in January to how the January model performs in February and select the model with better performance to use in March.









Note, we define January and February not strictly following the literal meaning of month, but in such a way January is the first 1/8 portion of the entire dataset for each stock and February is the second 1/8 portion. This way, data points across different stay the same and it is meaningful to compare the mean squared error of using the model trained on one month to predict another month.

Continuing the above example, we can get an intuition from the plot that the red dots are somewhat more converged to the green and yellow area compared to the blue dots, meaning the model trained on February actually does a better job. This reconciles with the result derived from calculating the mean squared error

```
mean_squared_error(list(feb_data_i.iloc[:,-1]), regr_jan.predict(feb_data_i.iloc[:,1:-1]))
```

```
5.84972099272089
```

```
mean_squared_error(list(jan_data_i.iloc[:,-1]), regr_feb.predict(jan_data_i.iloc[:,1:-1]))
```

```
3.354897160364451
```

Then, we can use the February model in March for further comparison and repeats such procedure until the end of August.

In the meantime, starting from February, we can also use the selected model to trade just like before and get a seven-month trading result, with potentially different models used in each month during this process.

We implement the entire process in detail and first look at what models win in each month.

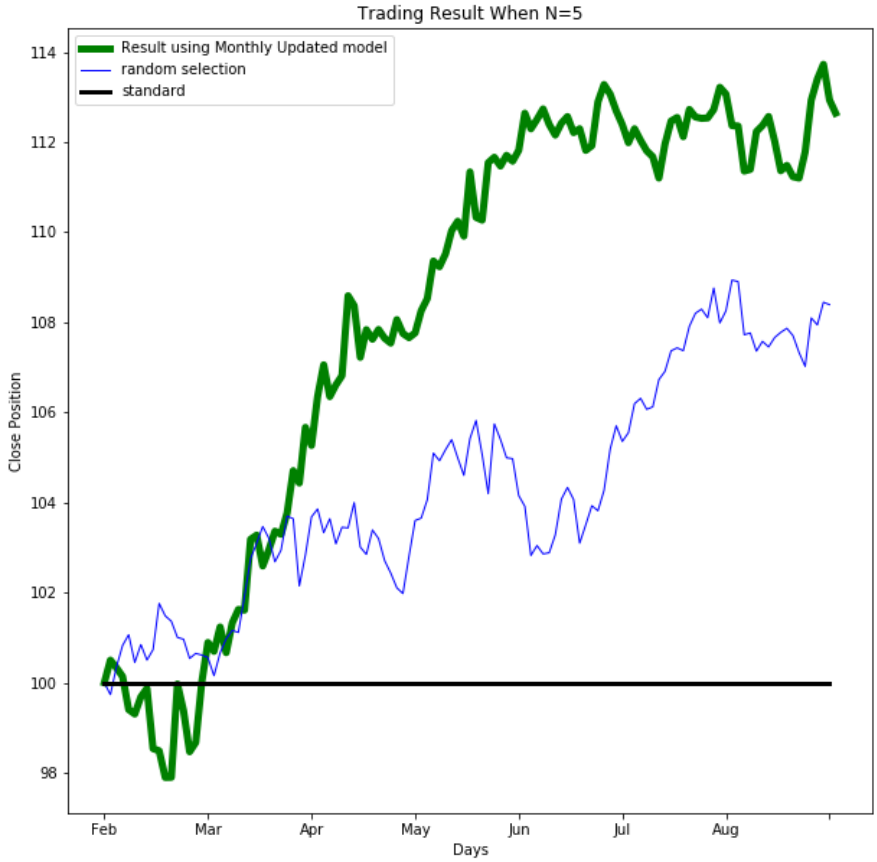
```
models_in_use = [regr_jan]
models_in_use_index = [1]

for m in range(2, 8):
    this_month_regr = linear_model.LinearRegression()
    this_month_regr.fit(data_months[m-1].iloc[:,1:-1], data_months[m-1].iloc[:,-1])
    this_pred_prev = mean_squared_error(list(data_months[m-2].iloc[:,-1]), this_month_regr.predict(data_months[m-2].
    prev_pred_this = mean_squared_error(list(data_months[m-1].iloc[:,-1]), models_in_use[-1].predict(data_months[m-1
    if this_pred_prev <= prev_pred_this:
        models_in_use.append(this_month_regr)
        models_in_use_index.append(m)
    else:
        models_in_use.append(models_in_use[-1])
        models_in_use_index.append(models_in_use_index[-1])
```

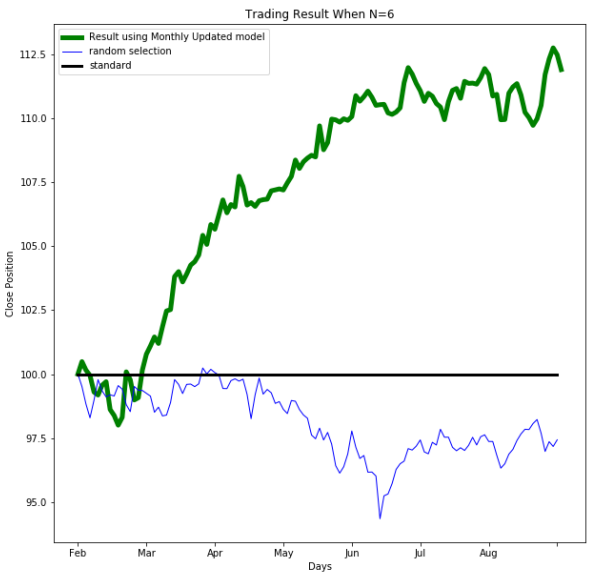
```
models_in_use_index
```

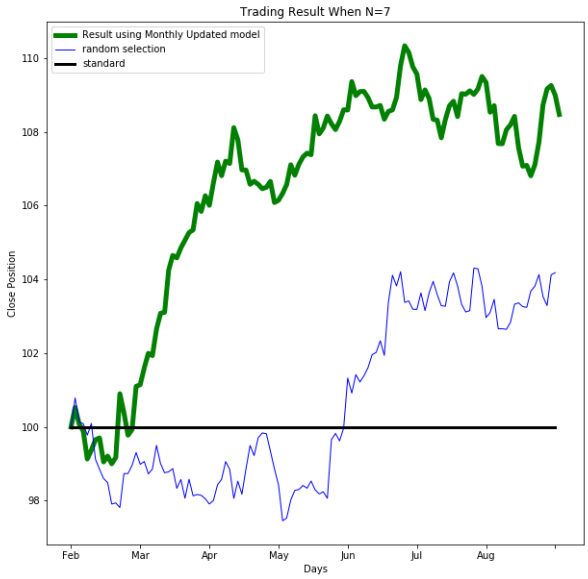
```
[1, 2, 2, 4, 4, 4, 4]
```

This means that the January model is used in February (without a second choice), the February model is used in March (example) and April, and the April model won all the way from May to August.

We then look at how this dynamic model will actually work when we trade from beginning of February to the end of August using selected models in each month.







Above, we plot three case where N is set to 5, 6 and 7 respectively and compare them to the random selection strategy.

In almost all cases, the performance in February is quite poor and is below the standard-100 line, which suggests that the model used at the beginning (January model) is quite inaccurate. However, as the March discards this terrible model and begins using February model and May begins using April model, the overall performance soon goes up rather quickly.

Generally speaking, the above results are indeed somewhat satisfactory. We did actually see the success from model selection observingh that the green line well outperforms the blue line in almost all times.

11 – Where More Studies can be Done

Basically, we have finished going through a shallow and experimental exploration of using text data to predict stock performance. Due to the time and efforts limitation, there are lots of places where we make casual, arbitrary assumptions and set some hyperparameters without any careful investigation.

This section summarizes those places where we skip a detailed look and where more studies can definitely be done to improve the performance of models or generate new ideas.

11.1 - Data Processing Stage

- When aggregating news on each single day, we simply discard the source information and take the average of all news published that days regardless of its source. We can distinguish news from different sources by one hot encoding since the reliability and the target readers of different websites can be quite different and this can have effect on the final prediction.

- We use the square kernel for the decay of the effect of news, there are various other kernels in the statistics field that may better model the real-life situation. I have implemented three simplest kernels but do not yet have time to test out all of them.

- In terms of the width (how long at maximum would the effect of a news last), we arbitrarily set this to be 10 days. Different settings can definitely be tested to see the changes.

11.2 – Input Preparation Stage

- For each piece of news, only its sentiment sore is used for prediction. Other information may have value as well, like the length of the text, appearance of names of other stocks, etc., may also be computed and used for prediction.

- When we say using “recent” news sentiment scores (and market data), we arbitrarily set “recent” to be three days. This can be adjusted to 5 days, one week, one fortnight or even one month to include more data and perhaps perform other sorts of data aggregation.

- Another fairly important point is that we do not distinguish news published in weekends / holidays and news published in trading days. While it is the situation in real life that the news on Saturday and Sunday will affect the market on Monday, in a quite different way compare with how the news in Monday will affect the market on Tuesday, this is not taken into consideration currently while implementing our model.

11.3 – Model Training Stage

- Given the nature of the data being “time-series”, recurrent neural network structures and Long Short Term Memory can be built and tested to exploit the underlying relationship of the news sentiment scores and stock market data between consecutive days.

- Given the nature of the stock market being similar to fuzzy logic, networks like SOFNN (Self Organizing Fuzzy Neural Networks) can be implemented and tested.