12/11 Team Meeting

Agenda:

1) Presentation 3 rework: shrink down to fit 15 mins including demo
    a) Go through comments
    b) Evenly divided up the slides or keep owners the same?
2) Prepare for showcase presentation:
    a) https://www.ischool.berkeley.edu/projects/2017/ml-based-investment-analytical-tool:
        i) Needs review and update https://docs.google.com/document/d/1PBTOCym6mOBZtBSmwQux6caig0Dp1eQwhjRXUCsEO44/edit.
            (1) Everyone: please review and edit. Feel free to reword!
            (2) Any pics you want to link to this doc?
            (3) Please update by 12/19 so I can update our project page
        ii) Just Zhongqiao, Thong, Sarah
    b) 10-min presentation:
        i) Evenly divided up the slides? Zhongqiao will take over Shankar's slides
        ii) Meet next Mon night (?)
        iii) Rework to fit 10 mins?
        iv) Practice date & time: 12/21? Time?

11/30 Meeting

Grade: 92
Instructor Feedback

Focus on one component only rather than searching for new data which can turn into a rabbit hole at this point. Since there has been a long bull market since 2009, be wary of generalizing too much from strategies learned just from data in that period, or at least compare it to a good baseline (maybe **Bollinger Bands**, ARIMA, random walk, index funds?).
In final presentation/web based deliverable, please be sure to motivate what characteristics of LSTM make it a good match for price prediction. 30-60 days is a very short time horizon for most investors (who are not traders) because commissions will likely eat up most of the gain - a year or several years if more typical.

Questions for OH:

1. [shankar] Please clarify:"focus on one component only"
    a. New data source for news
2. How to avoid generalizing too much since 2009?

    a. We tried to do that by including data since 2000 (before 2007 crash) but we didn't have access to news and financial data before 2010
        i. [shankar] By only predicting over a 60 day term, our model is not intended to predict long term. By being sentiment sensitive to news that precedes the 60 day prediction, this model can predict short term ups and downs.
3. About your baseline suggestions, can you clarify on what you mean about Bollinger Bands, etc.?
    a. Bollinger Bands: is the prediction within the bands?
        i. [shankar] Thong, we should think about prioritizing stocks that are closer to the lower end of the Bollinger band in the prediction term.
    b. ARIMA?
    c. Random walk?
    d. Use Index fund as baseline
    e. [sarah]: we should pick one core metric and focus on this right?

   What is the baseline to choose? (Thong did LSTM without timeseries. Should we show that?)
   Communicate what user gets out of this application

4. "30-60 days are very short time horizon":
    a. But this is the question/scope of this project: find the stocks that give the best % gain in the next 30, 45, 60 days
    b. Should we add 90 day prediction?
5. [sarah] "Not clear on the value add" - Is there anything we can do about this now? Other than try to show how this is a start to a new approach to prediction, we are not expecting this to be a final polished model.

11/20 status meeting

- Check in latest code before meeting
- https://github.com/thongnbui/MIDS_capstone/commits/master

Individual update: 5 mins
- Things done in past week?
- What are you working on?
- Behind schedule? Anything blocked? Need help?

Issues:
- UI: amount needed?

11/6 status meeting

- Check in latest code before meeting

- https://github.com/thongnbui/MIDS_capstone/commits/master

Individual update: 5 mins

- Zhongquiao: use package to generate sentimental score
- Sarah: incorporated Zhongqiao's code (Details?)
- Shankar: parsing json into csv with headers. Will commit 200 tonight; 100/night. By Thursday will have complete 500 dataset.
- Thong: committed work on lstm model predictions for 500 stocks based only on stock price, volume

- Update schedule: https://docs.google.com/spreadsheets/d/1cM4LH8UaCRbIxjURvF70A0_OpFfCy96yMQdAZWcPzM0/edit#gid=0
- Things done in past week?
- What are you working on?
- Behind schedule? Anything blocked? Need help?

Next steps:

- Thong: to integrate sentiment data and revenue (blocked by issues below)
- Sarah:
    - Resolve blocking issues
    - Continue UI work (Flask)
- Shankar, Zhongqiao:
    - Resolve blocking issues
    - Pick top 5 features
    - Start working with Sarah on UI

Issues:

1) 2 options for predictions: 30 and 60 days?
2) NLP and revenue data only started from 2015? (missing data from before)
    a) Need exact dates for earliest dates: Zhongqiao, Shankar to research
    b)
3) These currently are blocking Thong from integrating these dataset into LSTM models to train and predict

- Sarah:
    - missing other sentiment data for other stocks
    - 1 row for each date: avg(sentiment scores)? Add counts for news items per day (helps distinguish days with no news)?
    - Boolean -> 0/1
    - ETA?
- Shankar/Zhongqiao: metadata
    - Is it possible to convert Q1, Q2, Q3, Y2Y to a date?
    - ETA?

3) Others?

Team Discussion:

- Please make sure data checked in with format: [Sarah: uploaded individual company level csvs for basic NLP features]
    - data/<type>/AAPL.csv:
        - Heading row
        - date, <feature1>, <feature2>, etc.
- Pick 4 stocks for samples for NLP, metadata: [Sarah: I really think we need to use all the stock ultimately]
    - ?
    - ETA to be available in github:
        - NLP: [Sarah: Very basic features (sentiment, polarity, certain keywords) already up; we should build on this for the MVP and can improve this after we have everything at MVP stage)]
        - Metadata: Q1, Q2, Q3, Y2Y. What date should we use?
            - Zhongqiao: Quarterly reports available on the daily news?
            - Shankar: try to find some specific dates
    - Vote for top 5 features to use [Sarah: I think we definitely need to decide this empirically using one of the many feature selection techniques we have learned. I do not think we should decide this arbitrarily]
- Thong:
    - to integrate top 5 features of NLP, metadata into the existing model
        - Train
        - Test
        - predict
- UI (Flask) started by Sarah:
    - ETA for basic framework check-in? [Basic framework will be checked in by next Monday; already have a Flask/html integration running. We can demo the basic web interface in our second presentation if that's appropriate].
    - Thong (anyone else?) will work on next steps

Demo:

- Each team member will show their works

10/30 status meeting

We only have 6 weeks left
Code check in:
https://github.com/thongnbui/MIDS_capstone/commits/master

Individual update: 5 mins

- Update schedule: https://docs.google.com/spreadsheets/d/1cM4LH8UaCRbIxjURvF70A0_OpFfCy96yMQdAZWcPzM0/edit#gid=0
- Things done in past week?
- What are you working on?
- Behind schedule? Anything blocked? Need help?
- What's next?

Team Discussion:

- NLP work:
  - Format of NLP output? data/nlp/AAPL.csv: date, <some NLP/sentiment scores>?
  - ETA for samples? Next week
- Metadata work:
  - Using similar format as above: data/metadata/AAPL.csv: date, revenue, PE?
- UI using Django: Sarah + team
- Data since 2000 (~4465 rows):
  - Not enough data -> overfitting with current stock price
    - Do nothing and make a note?
    - Get more data (?)
    - Finetuning later to overcome this issue
  - More sparse data from news: 0 for dates without news?

10/16 status meeting

Individual update: 5 mins
- Things done in past week?
- What are you working on?
- Behind schedule? Anything blocked? Need help?
- What's next?

Team schedule - task list
- Are we on-track with schedule?
https://docs.google.com/spreadsheets/d/1cM4LH8UaCRbIxjURvF70A0_OpFfCy96yMQdAZWcPzM0/edit#gid=0

Discussion:
- NLP: what to do next? Research
- Scope of project: 500 stocks too many? Yes, will use 500 stocks
- Baseline Model? Random forest

10/6 - Team meeting

- Present: Shankar, Thong, Zhongqiao
- Sarah is at a conference

Agenda:

1) Presentation1: 15 mins
    a) Slides:Shankar, Zhongqiao sign up for slides.
    b) Finish all by Monday
    c) Monday night meeting for dry-run practice?
2) Project description:
    a) Sarah: where is the doc

Before next class: Share on wall (for each project)

- Problem Statement
- Value Proposition
- Use Case / Customer Segment
- Minimum Viable Product
- Data sets
- Project management:
- Technical approach and other considerations
- What kind of expert input would be most helpful?

3) Project status: https://docs.google.com/spreadsheets/d/1cM4LH8UaCRbIxjURvF70A0_OpFfCy96yMQdAZWcPzM0/edit#gid=0
    a) Review schedule timeline.
    b) Individual status update:
        i) Are we behind schedule? Yes
        ii) Blocking issues? Getting historical data for non-price data
        iii) Need help? Not yet. Owner continues to work on issues
    c) Weekly project meeting on Sat around 11:15am PT? Or Sun at 10am PT?
        i) Shankar
        ii) Thong: ok both, prefer Sun at 10am
        iii) Zhongqiao
        iv) Sarah
4) Next steps:
    a) Sign up for next works?
        i) Thong: try baseline model
        ii) Rest of team continues to work on unfinished tasks
    b) NLP implementation: conflicting ideas sentiment analysis vs NLP?
        i) Zhongqiao agreed that sentiment analysis is part of NLP

ii) What are the output? We think that something like the sentiment score but still TBD
iii) When can we get examples of output?

c) UI exploration?

10/1- Feedback from Coco and David on new proposal

- Problems are too challenging?
- Domain expertise:
  - We may have to ask someone knowledgeable in finance (Shuang Chan?)
- Who the audience is?
  - Users
  - Trading for short term, long-term? How long (1-month, 2 months)?
- NLP:
  - Sentiment analysis: lots of noise from Twitter. Maybe focus on news only (more credible, regular in structure)?
    - Earning surprises -> stock prices
    - Merger acquisitions:
      - management changes: CFO resigns -> negative impacts
    -

Action items: None