Risk Prediction Using Quantified News

Vu Anh Huynh[1] Liang Zhang[2]

[1]Department of Aeronautics and Astronautics, MIT
[2]Department of Nuclear Science and Engineering, MIT

Acknowledgement: Dan diBartolomeo, Anish Shah, Louis Scott

Northfields 18th Annual Summer Seminar

June 7, 2013

Outline

Problem Definition

AAPL

Problem Definition

AAPL

Risk Prediction Using Quantified News

- Given time-series of prices and intraday volatility of a security,
- Given time-series of quantified news for the same security,
- Build a model to predict the next intraday volatility.

Mathematical Formulation

Risk Prediction Using Quantified News

- $\{P_t\}_{0 \leq t}$: Price time-series,
- $\{\sigma_t\}_{0 \leq t}$: Intraday volatility time-series:

$$\sigma_t = \sqrt{252\pi/8} \log(P_{high}/P_{low})$$

- $\{n_t\}_{0 \leq t}$ where $n_t \in \mathbb{R}^K$: News time-series where $n_t = 0$ for dates without news.

Mathematical Formulation

Risk Prediction Using Quantified News

- $\{P_t\}_{0 \leq t}$: Price time-series,
- $\{\sigma_t\}_{0 \leq t}$: Intraday volatility time-series:

$$\sigma_t = \sqrt{252\pi/8} \log(P_{high}/P_{low})$$

- $\{n_t\}_{0 \leq t}$ where $n_t \in \mathbb{R}^K$: News time-series where $n_t = 0$ for dates without news.

Find a model $\mathcal{F}$ such that:

$$g(\sigma_{t+1}) = \mathcal{F}\Big(\{P_\tau\}_{0 \leq \tau \leq t}, \{\sigma_\tau\}_{0 \leq \tau \leq t}, \{n_\tau\}_{0 \leq \tau \leq t}\Big) + \epsilon_{t+1}, \qquad (1)$$

where $g : \mathbb{R} \to \mathbb{R}$ is a function that describes a property of σ_{t+1}.
The choice of the function g is important.

Approach: Using regression and model selection

- There are many approaches to construct such a model $\mathcal{F}$.
- Each approach is chosen based on the question that we pose about the prediction ability.

Approach: Using regression and model selection

- There are many approaches to construct such a model $\mathcal{F}$.
- Each approach is chosen based on the question that we pose about the prediction ability.
- In this work, we use regression and model selection to answer the question:
  What is the numerical value of tomorrow's intraday volatility? (i.e. g(x)=x)

$$\sigma_{t+1} = \mathcal{F}\left(\{P_\tau\}_{0\leq\tau\leq t}, \{\sigma_\tau\}_{0\leq\tau\leq t}, \{n_\tau\}_{0\leq\tau\leq t}\right) + \epsilon_{t+1}.$$

Approach: Using regression and model selection

- There are many approaches to construct such a model $\mathcal{F}$.
- Each approach is chosen based on the question that we pose about the prediction ability.
- In this work, we use regression and model selection to answer the question:
  What is the numerical value of tomorrow's intraday volatility? (i.e. g(x)=x)

$$\sigma_{t+1} = \mathcal{F}\left(\{P_\tau\}_{0\leq\tau\leq t}, \{\sigma_\tau\}_{0\leq\tau\leq t}, \{n_\tau\}_{0\leq\tau\leq t}\right) + \epsilon_{t+1}.$$

- We will compare performance against a base model, *which is a random walk model*.

Data: RavenPack – analytics engine converting news text to quantitative data

News data provided by Ravenpack $\{n_t\}_{0 \le t}$

- More than half a million pieces of news per month (nearly 1000 news per hour).
- For each news, detailed information is provided, such as type, relevance, event sentiment, novelty, etc. (see next slide)

	Entity	**Relevance**	**ESS**	**CSS**	**AEV**	**NIP**	**AEV**
t_1	company A	100	85	55	77	12	68
t_2	company B	...	...	...	...	...	...
t_3	company A	...	...	...	...	...	...
t_4	company C	...	...	...	...	...	...
t_5	company B	...	...	...	...	...	...

Data

News Data Field Descriptions (highlights)

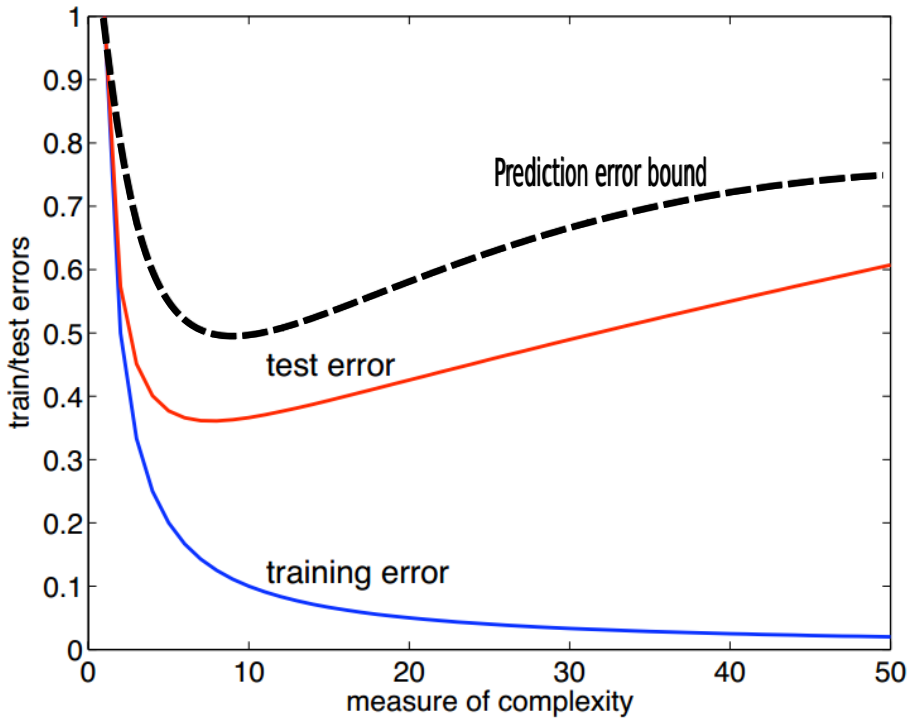
- **Entity**: company names.
- **RELEVANCE(N)**: 0-100, indicates how closely related the entity is to the underlying news, can be converged to number of news per day.
- **ESS**: 0-100, represents the news sentiment, i.e. higher values indicate postive sentiment, while lower values below 50 show negative sentiment.
- **AES**: 0-100, the percentage of positive events measured over a rolling 90 day window.
- **AEV**: the count of events measured over a rolling 90 day window.
- **ENS**: 0-100, represents how "new" or novel a news is.
- **CSS**: 0-100, sentiment score combined various analysis techniques.
- **NIP**: 0-100, the degree of impact a news flash has on the market over the following 2-hr period.

Regression Model Selection: General Result and Fundamental Limit

- Consider nested models: $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq ...$ used for training to estimate parameters of the models,
- Split data into training set and testing set,
- Report training error and testing error for each model $\mathcal{F}_i$:

Regression Model Selection: General Result and Fundamental Limit

- Consider nested models: $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \ldots$ used for training to estimate parameters of the models,
- Split data into training set and testing set,
- Report training error and testing error for each model $\mathcal{F}_i$:



Implication: Balance between training error and complexity of a model.

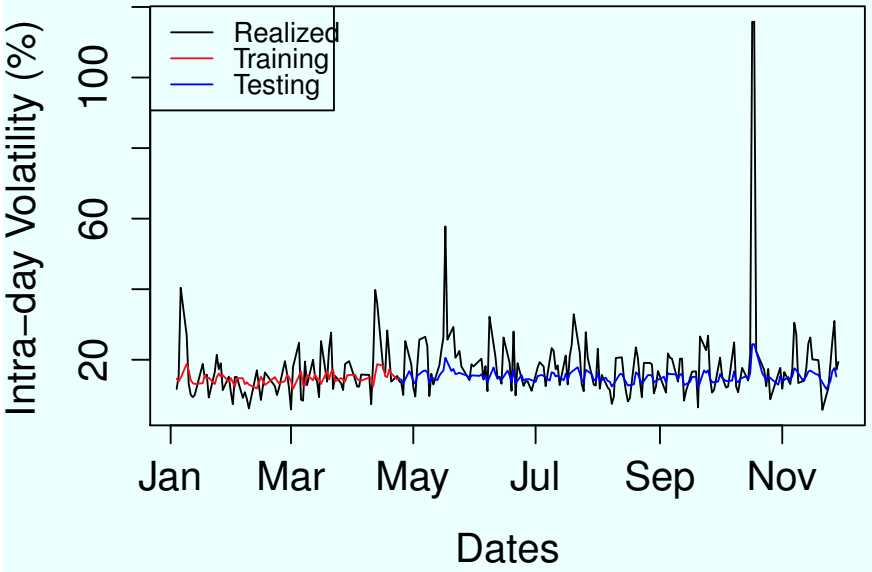
An Example of Nested Models

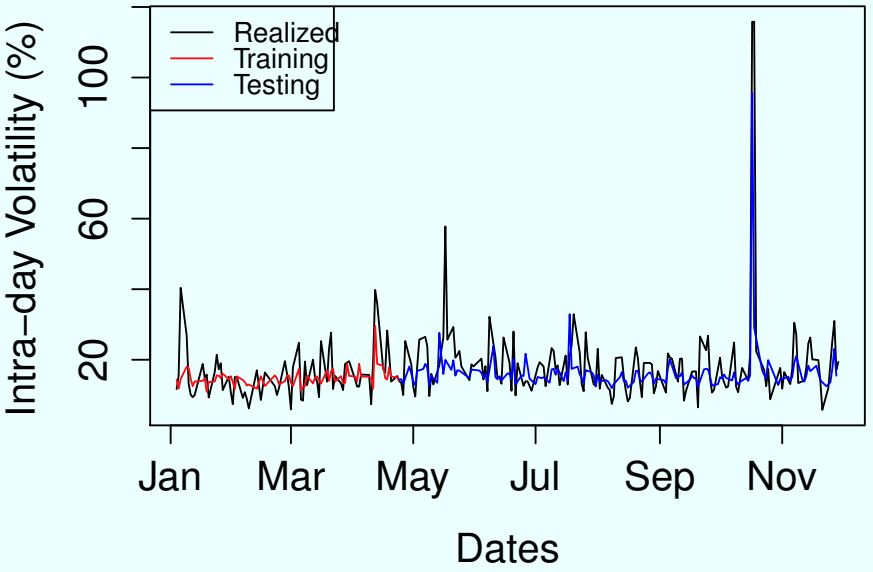
1. $log(\sigma_{t+1}) = log(\sigma_t) + \epsilon_{t+1}$
2. $log(\sigma_{t+1}) = log(\sigma_t) + N_t + \epsilon_{t+1}$
3. $log(\sigma_{t+1}) = log(\sigma_t) + N_t + CSS_t + \epsilon_{t+1}$
4. $log(\sigma_{t+1}) = log(\sigma_t) * N_t + CSS_t + \epsilon_{t+1}$
5. $log(\sigma_{t+1}) = \left(log(\sigma_t) + CSS_t\right) * N_t + \epsilon_{t+1}$
6. $log(\sigma_{t+1}) = \left(log(\sigma_t) + CSS_t\right) * N_t + ENS_t + \epsilon_{t+1}$
7. $log(\sigma_{t+1}) = \left(log(\sigma_t) + CSS_t\right) * N_t + ENS_t + NIP_t + \epsilon_{t+1}$
8. $log(\sigma_{t+1}) = \left(log(\sigma_t) + CSS_t\right) * N_t + ENS_t * NIP_t + AES_t + AEV_t + \epsilon_{t+1}$
9. $log(\sigma_{t+1}) =$
   $\left(log(\sigma_t) + CSS_t + ENS_t\right) * N_t + ENS_t * NIP_t + AES_t * AEV_t + \epsilon_{t+1}$
10. $log(\sigma_{t+1}) =$
    $\left(log(\sigma_t) + CSS_t + ENS_t + ESS_t\right) * N_t + ENS_t * NIP_t + AES_t * AEV_t + \epsilon_{t+1}$

Note: $A * B$ means $A + B + AxB$

Results

Google (NASDAQ: GOOG)

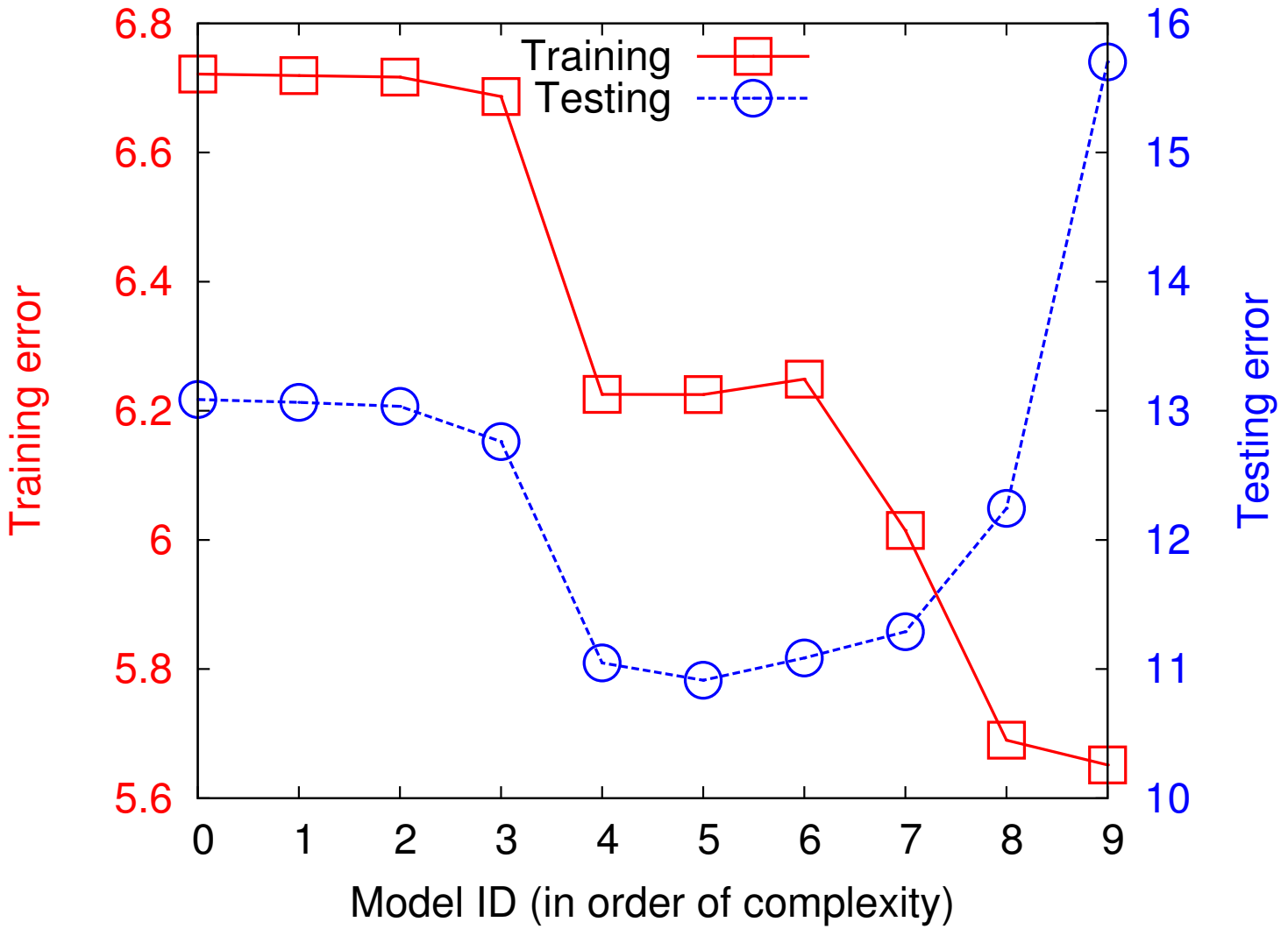




$log(\sigma_{t+1}) = log(\sigma_t) + \epsilon_{t+1}$	$log(\sigma_{t+1}) = \Big(log(\sigma_t) + CSS_t\Big) * N_t + ENS_t + NIP_t + \epsilon_{t+1}$
$R^2 = 1.3\%$	$R^2 = 31.4\%$

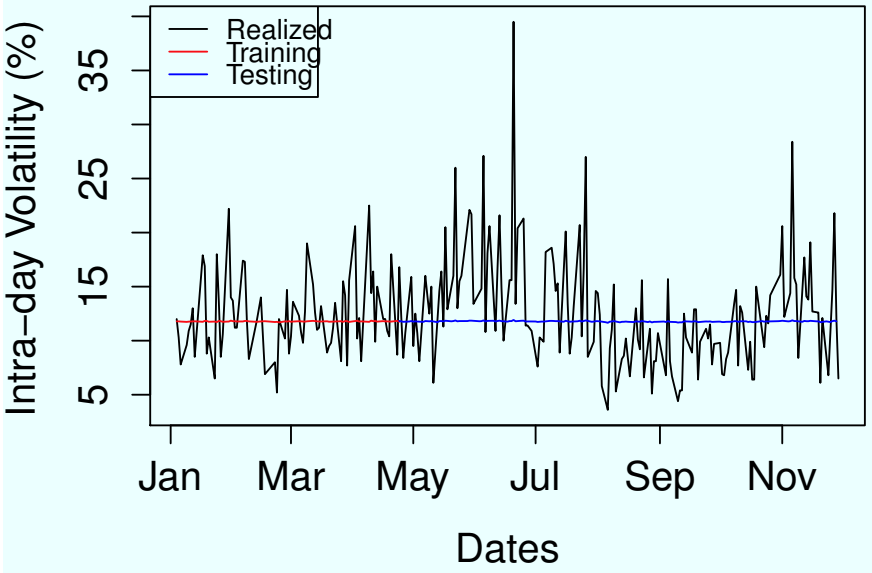
Results

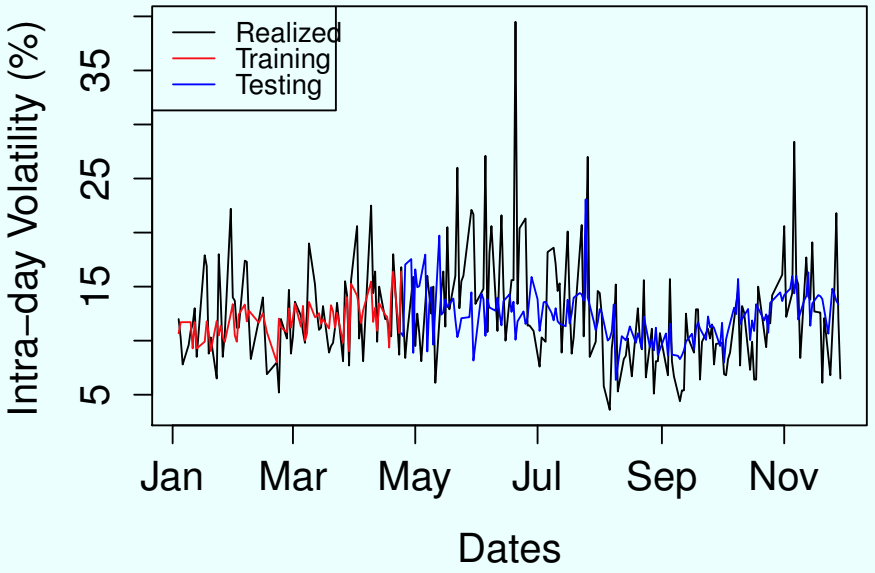
Google (NASDAQ: GOOG)



Results

Exxon Mobil (NASDAQ: XOM)

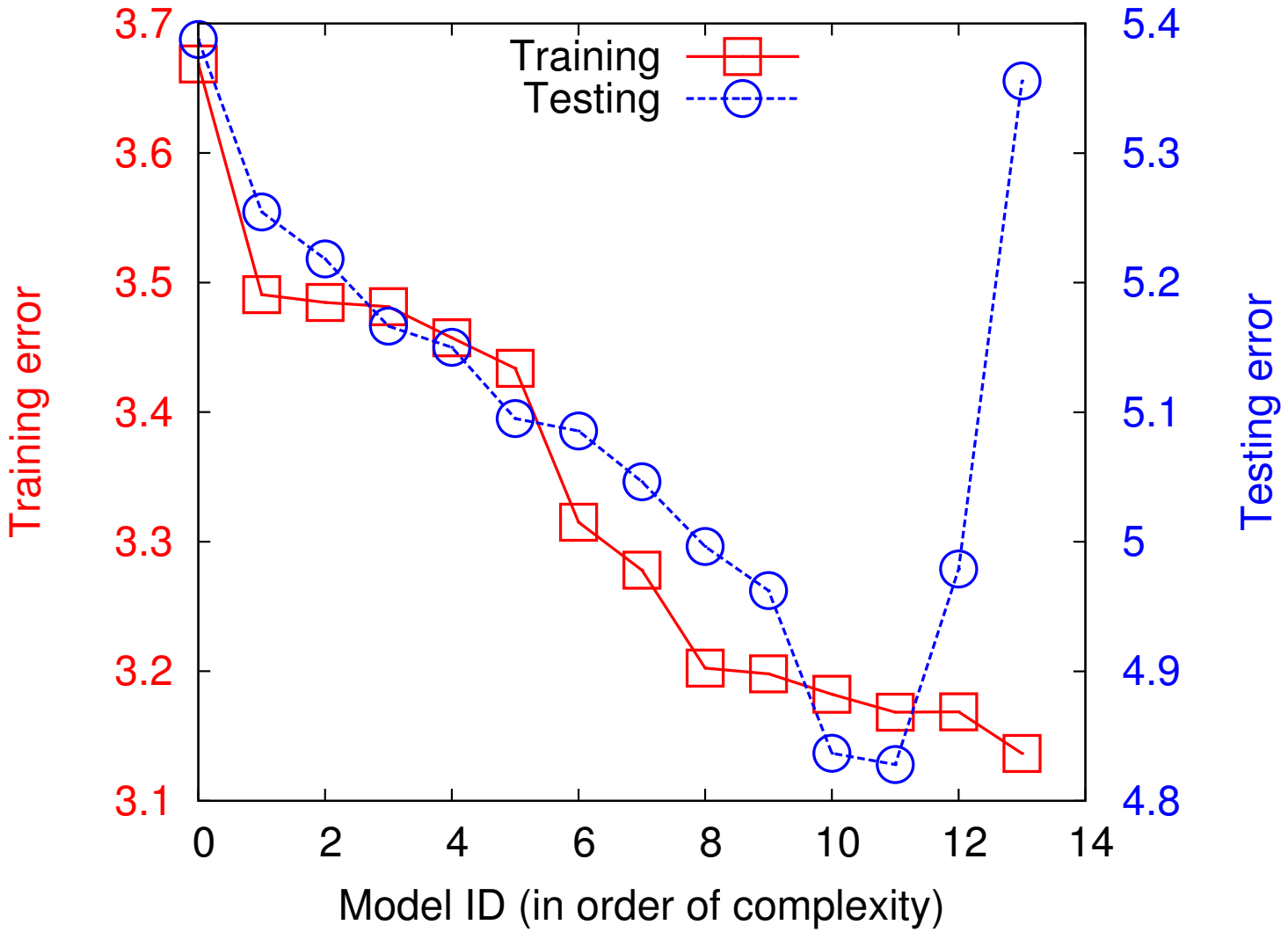




$log(\sigma_{t+1}) = log(\sigma_t) + \epsilon_{t+1}$	$log(\sigma_{t+1}) = \Big(log(\sigma_t) + AES_t + CSS_t + ENS_t\Big) * AEV_t + \Big(CSS_t + ESS_t\Big) * N_t + \Big(NIP_t + ESS_t\Big) * ENS_t + \epsilon_{t+1}$
$R^2 = -1.8\%$	$R^2 = 18.4\%$

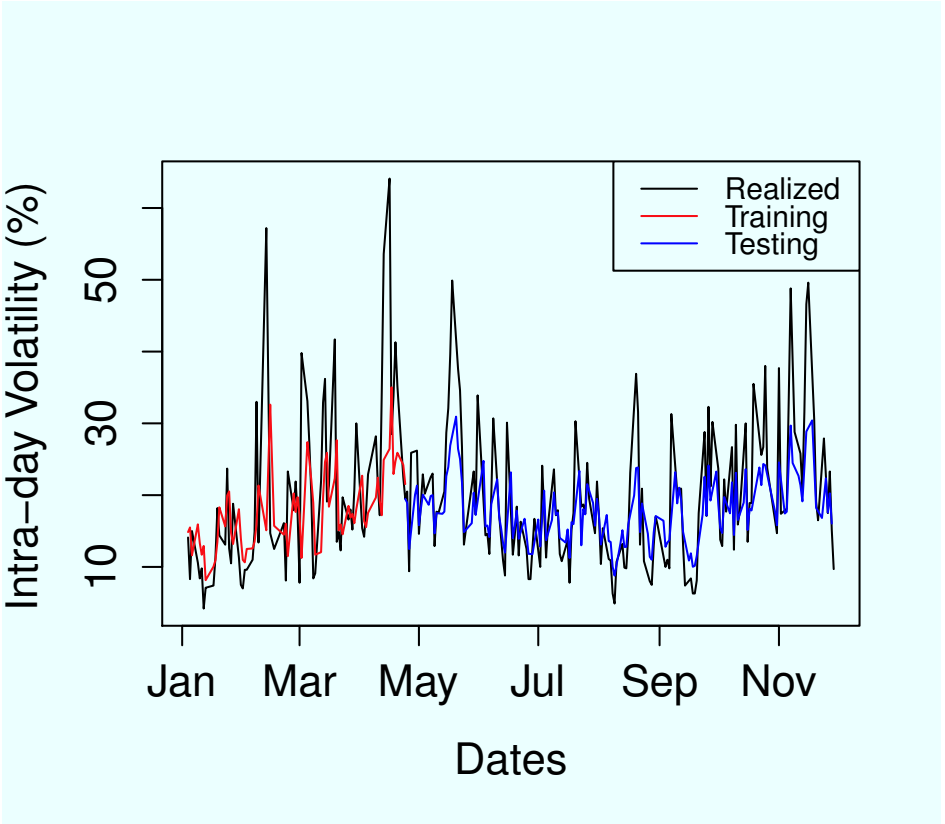
Results

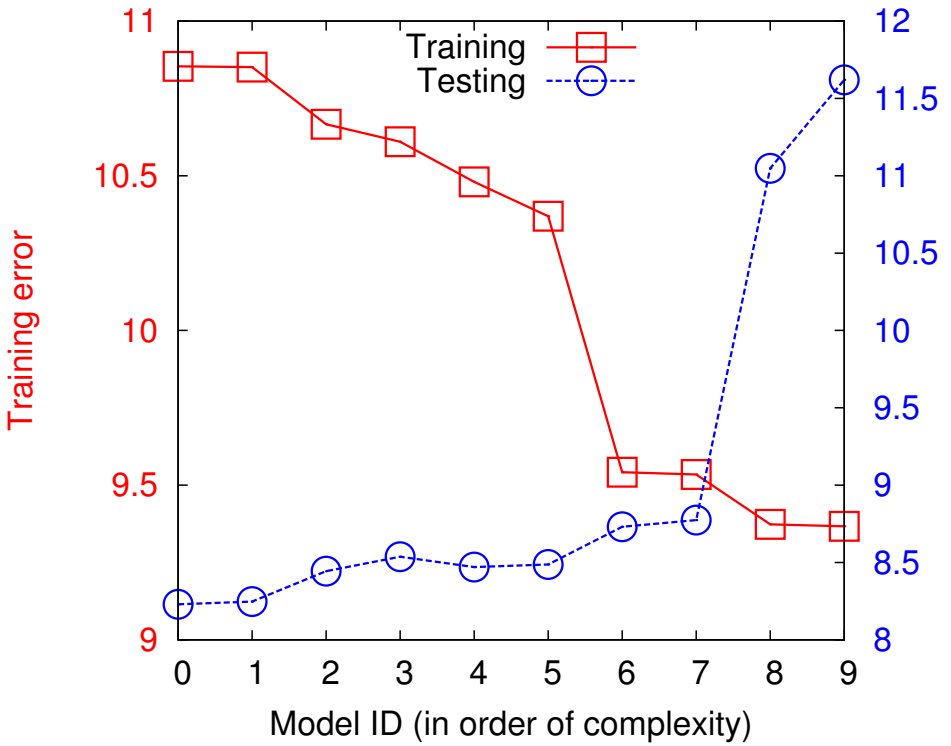
Exxon Mobil (NASDAQ: XOM)



Results

Apple (NASDAQ: AAPL)

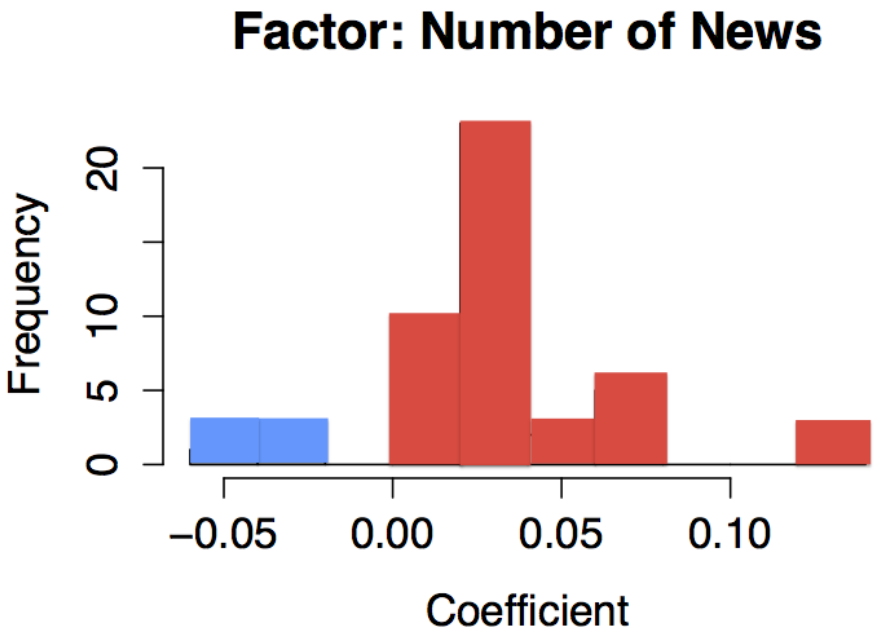


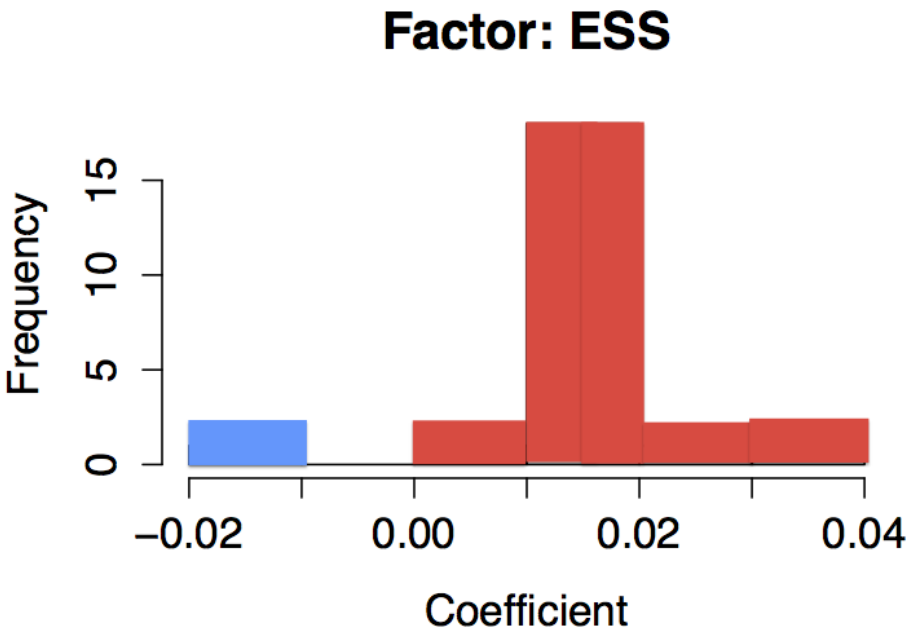


No model can be better than the simple random walk model ($R^2 = 24.4\%$)

Statistics of regression models

- We investigated different regression models for over 150 stocks in the year of 2012
- There are many stocks that are *news neutral*, in the sense that volatility is indifferent to any factors from news analytics.
- For stocks that are response to news factor, volatility is consistently correlated with certain news factors (as shown in the figures below).





Conclusions and Future work

- We have investigated the regression approach to model how news affects intraday volatity of securities.
- For certain stocks, we built up nested regression model to significantly improve the prediction of intra-volatility .
- There are *news neutral* stocks for which news-incorporated models do not outperform the random walk model.
- For many stocks, volatility is positively related to the number of news as well as the news sentiment.
- We are building statistics on a larger dataset to support our hypothesis.