Social Network Analysis

Stefan Dimitrov
School of Computer Science
McGill University
Montreal, Quebec, Canada
stefan.dimitrov@mail.mcgill.ca

***Abstract*—This dataset includes data for an exchange-traded security linked to natural gas; and weather measurements. My goal is to evaluate if daily temperature measurements can contribute to predicting the direction of daily returns for the security. The results show that, when used with the Logistic Regression classifier, the dataset performs slightly better than the baseline classification of the majority result.**

The dataset is available here: http://cs.mcgill.ca/~sdimit8/comp598a1.zip

***Index Terms*—Gas, equities, returns, weather, prediction**

I. Problem description

This work is motivated by the UGAZ and DGAZ ETNs (Exchange Traded Notes) which are linked to the S&P GSCI Natural Gas Index ER and track 3 times the daily performance of the index[1][2]. UGAZ will normally rise with the increase of gas prices, while DGAZ will fall by the same amount. In addition to gas contracts, trader speculation about the future direction of gas prices also affects the returns of UGAZ and DGAZ. Thus, the two products are highly volatile and offer traders an opportunity for a substantial intraday profit (or loss). By predicting the direction of gas contracts at market open, a trader can decide whether to trade DGAZ or UGAZ and hence profit in both cases, when gas prices are declining or rising.

As the UGAZ & DGAZ ETNs were created in 2012, there is a very limited amount of data available for training. Instead, I am using the UNG ETF (Exchange Traded Fund) end of day data. This instrument is similar to UGAZ as it also tracks the daily changes in natural gas prices, but it is not triple leveraged.

The prediction question is then a binary classification one, namely: given the daily temperature in New York City, the open price of UNG for the day, the change from the last closing price (overnight change) and the previous day's intraday return, predict whether today's intraday return will be a positive or a negative one. With this information a trader can profit by buying UGAZ at market open (if return is positive) or DGAZ (if return is predicted negative) and selling before market close.

II. Related work

Weather has been linked with feature prices in the past, in particular with natural gas. The reason is because natural gas is used for heating in the winter and for energy generation which powers air conditioning systems in the summer. Thus, short term temperature variations can cause volatility in the demand for natural gas. As Mu (2007) found, up to 50% of the US demand for natural gas is affected by weather. [3]

Furthermore, weather has been found to influence the intraday trading behaviour of traders by affecting their mood [4]. Chang et al. (2008) look into the overall stock market, but not gas equities in particular.

I have not found any prior study attempting to predict the direction of daily returns for a gas equity based on weather measurements and deviations from the mean temperature for a given day. While Mu (2007) looks into the effect of weather shocks on gas features returns, he does not look into the psychological influence of weather variations on traders.

III. Dataset description

The dataset consists of 1847 examples with 8 features which may be related to the direction of daily returns of a gas security, and a result column which indicates if the returns were negative (0) or positive (1). The features I have chosen for this dataset are: day of year (mapped to mean temperature), season, day of week, the price of the security at market open, its change overnight, the direction of return for the previous day of trading, the mean temperature for all years in NYC on the current day of the year and the actual average temperature for the day.
The three seasons used in this dataset are as per Mu (2007): November, December, January, February, March: 0 (winter); June, July, August: 1 (summer); April, May, September, October: 2 (shoulder) [3]

IV. Methods

In preparing the dataset I first used the Google Finance feature of Google Spreadsheets to obtain end of day data for UNG since inception (2007-04-18) until 2014-09-17[1].

[1]https://support.google.com/docs/answer/3093281?hl=en

Then I processed the data to generate the Return, Day of Year, Season, Day of Week, Open, Overnight Change and Previous Return columns. To calculate mean temperature I downloaded all daily weather data for the New York Central Park weather station from 1763 until 2014. [2] Originally I intended to use the observed temperature at 12:00 AM, however this data was not available for recent years. Thus, to calculate the mean I averaged all minimum and maximum temperature measurements for a given day of the year. The observed temperature column I populated with the average of minimum and maximum temperature on the specific date.

Once I had the dataset, I attempted to implement logistic regression. Unfortunately, I was not successful. I encountered issues with float precision in Python and probably other implementation problems. I did implement a Naive Bayes classifier successfully, however it is very slow. In order to be able to present this report on time I resorted to developing a Scikit-learn based solution.[5] Thus, the rest of this section and the results discuss my findings with the aforementioned solution.

I used leave-one-out cross-validation to test the performance of my dataset with logistic regression, LDA and Naive Bayes classifiers, as well as to identify the best model dimension. I repeated the cross-validation experiment starting with only two dimensions (Day of Year and Season) and adding one dimension on every run until all 8 dimensions were included in the model, for every classifier. I also implemented a "baseline" classifier which always chooses the predominant result from the test set.

I repeated all tests with half of the training data in order to evaluate how changing the data size affects the performance of the classifiers.

V. Results

A. Logistic Regression

I first performed logistic regression with half of the dataset (923 samples):

TABLE I
Logistic Regression (923 samples)

d	$\text{Error}_{\text{train}}$	$\text{Error}_{\text{valid}}$
Baseline	0.4702	0.4702
2	0.4702	0.4702
3	0.4605	0.5439
4	0.4504	0.4680
5	0.4568	0.4637
6	0.4529	0.4670
7	0.4155	0.4323
8	0.4131	0.4301

[2] http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/

As demonstrated in table 1, the model performs slightly better than the baseline (always choosing the most popular result). It appears that adding weather measurements (mean and observed temperature) helps reduce both the training error rate and the estimated true prediction error.

Next, I repeated the test with the full dataset (1847 samples), as can be seen on table 2. The results are very much consistent and show that the model of all 8 dimensions performs better than average and better than models of lower dimensions. However, my results reveal that increasing the amount of training data decreases the performance of the model, which is unexpected. It could be explained by the fact that the dataset was not randomized. Thus, there appear to be other factors which may help in predicting the direction of gas equities, and they could be an interesting topic for further research.

A slight overfitting can be observed when adding the overnight return variable to the larger dataset. What is interesting is that adding mean temperature alone does not improve performance, but adding the observed temperature does.

TABLE II
Logistic Regression (1847 samples)

d	$\text{Error}_{\text{train}}$	$\text{Error}_{\text{valid}}$
Baseline	0.4694	0.4694
2	0.4694	0.4694
3	0.4694	0.4694
4	0.4613	0.4613
5	0.4578	0.4694
6	0.4438	0.4510
7	0.4450	0.4564
8	0.4387	0.4499

B. LDA

Table 3 summarizes the results of the LDA classifier. Similar results were obtained as with Logistic Regression. LDA performs even better on the smaller dataset, but performs worse than Logistic Regression on the full dataset.

C. Naive Bayes

Table 4 shows the results of running the scikit-learn Naive Bayes classifier. The Naive Bayes classifier performs the worst of all three evaluated in this paper. A possible reason would be that the dataset dimensions are not truly independent: season for example can be inferred from day of year.

I also implemented a Naive Bayes classifier, but it is much slower than the Scikit-learn version. Yet, the results are the same (apart for some very slight differences on the smaller dataset which I attribute to including the header line in the

TABLE III
LDA

	923 samples		1847 samples	
d	$Error_{train}$	$Error_{valid}$	$Error_{train}$	$Error_{valid}$
Baseline	0.4702	0.4702	0.4694	0.4694
2	0.4702	0.4702	0.4694	0.4694
3	0.4603	0.5439	0.4694	0.4694
4	0.4504	0.4659	0.4613	0.4613
5	0.4567	0.4637	0.4579	0.4700
6	0.4526	0.4702	0.4496	0.4591
7	0.4138	0.4312	0.4528	0.4683
8	0.4139	0.4280	0.4473	0.4586

row count and rounding). The results of my classifier can be seen on Table 5.

TABLE IV
NAIVE BAYES - SCIKIT-LEARN

	923 samples		1847 samples	
d	$Error_{train}$	$Error_{valid}$	$Error_{train}$	$Error_{valid}$
Baseline	0.4702	0.4702	0.4694	0.4694
2	0.4702	0.4702	0.4694	0.4694
3	0.4561	0.4561	0.4654	0.5912
4	0.4605	0.4800	0.4613	0.4613
5	0.4460	0.4615	0.4574	0.4802
6	0.4556	0.4724	0.4709	0.4727
7	0.4140	0.4301	0.4629	0.4651
8	0.4058	0.4334	0.4608	0.4656

TABLE V
NAIVE BAYES - PURE PYTHON

	924 samples		1847 samples	
d	$Error_{train}$	$Error_{valid}$	$Error_{train}$	$Error_{valid}$
Baseline	0.4702	0.4702	0.4694	0.4694
2	0.4708	0.4708	0.4694	0.4694
3	0.4567	0.4567	0.4654	0.5912
4	0.4611	0.4805	0.4613	0.4613
5	0.4440	0.4665	0.4574	0.4802
6	0.4521	0.4773	0.4709	0.4727
7	0.4144	0.4307	0.4629	0.4651
8	0.4054	0.4351	0.4608	0.4651

While not shown, I did try various combinations of dimensions with all three classifiers and experienced similar or worse results. Therefore, based on the results above, the model with all 8 dimensions is able to predict the direction of daily returns for natural gas securities slightly better than the baseline, when used with Logistic Regression classifier (see fig.1) .

VI. DISCUSSION

This is a first attempt at creating a dataset linking weather variations to intraday trading returns for a gas security. I chose the dimensions for this dataset based primarily on empirical observations, but also validated their relevance using the leave-one-out cross-validation method. Predicting intraday returns for any security, especially one as volatile as UNG, is a very challenging task. While the proposed dataset performs less than perfectly, it can be used as a good starting point for further research.

Gas supply is relatively fixed in the short term, because building wells and pipes takes time. On the other hand, temperature fluctuations affect gas demand as consumers and industries buy more gas for heating/cooling. Thus, the changes in demand with fixed supply lead to volatility in the equilibrium price. Traders of gas equities analyse supply and demand data to gain insight into the market direction. However, supply, storage and demand numbers are available only on the day when a report is released. Yet, traders need to make buy/sell decisions every day the market is open. Hence, weather appears to be the one input factor which is continuously available and affects prices.
Normally gas features are very volatile, still there are 36 days in the dataset for which the intraday return is 0.00% (open price is equal to close price). For these days I assign the intraday return to 0 (negative).

A possible future direction for research would be to consider temperature measurements throughout the day, as Chang et al. (2008) do for the overall market[4], but particularly with respect to gas equities. If enough DGAZ/UGAZ data is available, it may provide better insight into the psychology of intraday traders, due to the triple returns offered by these instruments. Other variables, such as the market demand/supply numbers, can help predict the trend for gas prices. At the same time, the day when these numbers are reported experiences unusual returns and volatility [3]. Augmenting the daily returns prediction model with the aforementioned extra dimensions could improve its performance, hence it would be a worthwhile direction of further research.

VII. APPENDIX

Data Dictionary:

Column	Units	Description
Return	N/A	Intraday return: 0 - negative; 1 - positive
Day of Year	N/A	The day of year from 1 to 366. This is used for temperature.
Season	N/A	0 - winter; 1 - summer; 2 - shoulder (fall & spring)
Day of Week	N/A	0 - Monday; 1 - Tuesday; 2 - Wednesday; 3 - Thursday; 4 - Friday
Open	USD	Price per share of UNG for the first trade during market hours of the day.
Overnight Change	%	Change in price per share of UNG between the last trade during market hours of the previous day and the first trade during market hours of the current day
Previous Return	%	Intraday return for the previous trading day
Mean Temperature	Degrees Celsius	Mean temperature in NYC for the current day of the year
Observed Temperature	Degrees Celsius	Average temperature in NYC on the current day

ACKNOWLEDGMENT

"We hereby state that all the work presented in this report is that of the authors."

REFERENCES

[1] VelocityShares (2012, February 8). 3x Long Natural Gas ETN [Factsheet]. Available: http://etfdb.com/factsheets/UGAZ/

[2] VelocityShares (2012, February 8). 3x Inverse Natural Gas ETN [Factsheet]. Available: http://etfdb.com/factsheets/DGAZ/

[3] Xiaoyi Mu, "Weather, storage, and natural gas price dynamics: Fundamentals and volatility," Energy Economics, Volume 29, Issue 1, January 2007, Pages 46-63.

[4] Shao-Chi Chang, Sheng-Syan Chen, Robin K. Chou, Yueh-Hsiang Lin, "Weather and intraday patterns in stock returns and trading activity," Journal of Banking & Finance, Volume 32, Issue 9, September 2008, Pages 1754-1766.

[5] Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR 12, pp. 2825-2830, 2011.