Will the Changes of Information Gathering Behaviors Affect the Financial Market?

Qintai Liu Shuangjin Zhang Yi Deng

Abstract

Nowadays, data analysis becomes a tool for people to predict the movements of human behaviors in different markets. This analysis report focuses on analyzing the changes in searching financial terms online and the price movements on the stock market. In order to get a better estimation, the analysis uses three models including SVM, RandomForest and Logistic Regression to cross-validate the datasets in the past five years in the unit of a week. After comparing three model, the estimator has the highest accuracy rate (mean=0.663, std.=0.072) when Δt=3 and the model is RandomForest (max_features=None, min_samples_split=4). After analyzing five-year datasets, the final result of the evaluation is that the average accuracy is 0.655, the std. is 0.072, and the base rate is 0.60. To conclude, the period between the online key terms searching of investors and the decision makings from investors would cause notable changes in the financial market and this period is 3 weeks.

I. Introduction

As 'Big Data' becomes popular, an increasing number of people pay attention to analyze different data sets to explore different patterns, or even discover the trends in human behaviors. New data sources are collected and updated day by day, and they giving more resources to analysts on analyzing new problems. The financial market is one of the major markets that involves a large amount of human behaviors, especially when people are making trading decisions. This human behavior data analysis project was inspired by the article *Quantifying Trading Behavior in Financial Market Using Google Trends*. During the analysis report, Preis, Moat, and Stanley analyzed the changes in *Google* query volumes for finance-related search terms and raised a new idea related to the behaviors of the players in the financial market (2013). When making trading decisions in the stock market, investors may raise concerns and gather more information by searching related financial terms on the Internet (Preis, Moat, and Stanley, 2013). The collected search data from *Google Trends* was used to analyze this trading decisions behavior. In addition, Preis, Moat, and Stanley found out that some "early warning signs" are helpful to predict future stock market movements (2013).

This project aims to explore the patterns of investor's decisions on stock market based on the previous datasets of finance search terms on *Google Trends* by using different model and parameters. Also, this project will try to use the patterns found out in the analysis to predict the future moves in the stock market.

II. Methodology

Data preparation

1. The datasets used to predict the change of S&P 500 index were downloaded from *Google Trends*. They include total 44 financial terms with specific relative search volume per week in the past 5 years. On the other hand, the weekly historical prices data for S&P 500 index in the past 5 years was downloaded from Yahoo Finance.
2. Read these files into separate data frames and then combine them into one data frame based on time (each week). Calculate the target variable "decrease" (a binary type representing whether the prices for S&P 500 go up or down during a specific week). The increased prices will return "0", while the decreased prices will return "1."
3. Calculate the relative change in search volume for each word in the Δt weeks based on the formula: $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ with $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \ldots + n(t-$

$\Delta t))/\Delta t$, where t is measured in units of weeks, N is the search volume for each word, and Δn is the relative change. *Figure 1* is the result below:

Week	housing:Rel	leverage:Rel	fun:Relative	loss:Relative	debt:Relativ	society:Relat	profit:Relativ	cancer:Relat	growth:Rela	bonds:Relati	headlines:Re	derivatives:F	environmen	credit:Relativ	dow joines:F	marriage:Re	sell:Relative	finance:Rela	color:Relativ	markets:Rel	return:Relati	nasdaq:Rela	gains:Relativ	car:Relative
2012-12-17	-11.6666667	63.3333333	4	-1.3333333	-9.6666667	-13.333333	-22.333333	-10	-17	-27	-4.3333333	-45.666667	-43	-1.6666667	-9	-2.3333333	2	-12.666667	-3.3333333	-2.6666667	2.66666667	-9.3333333	-12.666667	2
2012-12-24	9.33333333	-15	2	15.3333333	12	-1.3333333	-2	-1.3333333	-0.3333333	-10.333333	1	-20.333333	-13.666667	5.33333333	20	2.33333333	1	5	0.33333333	2.66666667	5.66666667	2.66666667	25	1.33333333
2012-12-31	18.6666667	-31.333333	-0.3333333	12	9	9.33333333	15	10.3333333	10	15	4	14	27.3333333	1.66666667	-9.6666667	1.33333333	-2	8.66666667	3	1.66666667	4.66666667	5	-7.6666667	-0.3333333
2013-01-07	9.66666667	-31.666667	-1.6666667	1.66666667	10.6666667	8.33333333	13.3333333	8.33333333	9.66666667	13.3333333	-0.3333333	19.6666667	27.3333333	-1.3333333	-9.6666667	0.33333333	-2.3333333	9	3	3.66666667	3	8.33333333	-11.666667	0.33333333
2013-01-14	4	-9.6666667	-3	-1.3333333	-2.3333333	2.33333333	7.33333333	0.66666667	5.66666667	-0.3333333	0.66666667	8.33333333	10.3333333	-2	-9.6666667	-1.3333333	-1.3333333	1.66666667	2.33333333	3.66666667	6	3.33333333	-10	1.33333333
2013-01-21	1.66666667	-2	-2.6666667	-1.6666667	-3	3.66666667	8	0.33333333	4	9	-0.6666667	16	9	4.33333333	26	0.33333333	-0.3333333	4	2	0.33333333	15.6666667	5.33333333	1.66666667	0
2013-01-28	-0.6666667	-1.3333333	0.66666667	-0.3333333	-4.6666667	1.33333333	1.66666667	0.33333333	1	-0.3333333	2	10.6666667	6.33333333	1.33333333	-8.6666667	1.33333333	-3	-0.6666667	1	-0.6666667	15.6666667	0.66666667	3	-0.3333333
2013-02-04	-3	-0.3333333	19	-3.3333333	-1.6666667	-0.3333333	-4	-0.3333333	1	0.66666667	-2	4.66666667	-2.3333333	0.33333333	-8.6666667	2	-1	0.33333333	-0.6666667	0.33333333	0	-1	1.33333333	0.66666667
2013-02-11	1	0.66666667	0.66666667	2	-2.6666667	0	0.33333333	1.33333333	0.33333333	-6.6666667	-1.3333333	2.66666667	-1.6666667	-0.6666667	-8.6666667	-1	0.66666667	-2.3333333	1.33333333	1.66666667	-6.3333333	-6.3333333	4.33333333	2.66666667
2013-02-18	4.33333333	-0.3333333	-7	1.66666667	2.66666667	2.33333333	4.33333333	2	2.33333333	2	-1	5.66666667	5.33333333	1	0	0	0	4	1	-1	-8.6666667	-0.6666667	0.66666667	0.66666667
2013-02-25	5	2.66666667	-6.3333333	2.33333333	-1	0.66666667	2.33333333	5.33333333	-0.6666667	-0.3333333	1	-3.3333333	1.66666667	-0.3333333	0	0	-0.3333333	2.66666667	0.66666667	1	-7.6666667	3.33333333	3	0
2013-03-04	0.33333333	-0.3333333	1	0.66666667	-1.6666667	-1	-0.6666667	0	-1.6666667	-2.6666667	-0.3333333	-11.333333	-2	-0.3333333	0	1.33333333	0.66666667	1.66666667	1	0.66666667	-3	1	0.33333333	1.66666667
2013-03-11	-0.6666667	-1.3333333	0.66666667	0.33333333	-2.6666667	-2	-1	-1.6666667	-1.3333333	-4.6666667	2.66666667	-5.6666667	-2.3333333	-2.3333333	0	1.66666667	0.66666667	-3	0.66666667	0.66666667	-3.6666667	-3.3333333	-0.6666667	0.33333333
2013-03-18	-0.6666667	-0.3333333	3	-3	-2	-4	-4.3333333	-2.6666667	-2.3333333	-4.3333333	-1.3333333	-9.3333333	-9.6666667	-1.3333333	0	54	0.33333333	-6	1.33333333	2	-1.6666667	-3.6666667	-4	1
2013-03-25	5	-2.6666667	-2.3333333	1	-1	-1.3333333	0.33333333	2	-0.3333333	1	2	-0.6666667	-3	1.33333333	0	-10.333333	-1	-1.6666667	-1.3333333	0.33333333	0	2.33333333	1.66666667	-0.6666667
2013-04-01	3.33333333	0.33333333	-0.6666667	2	1.66666667	5.66666667	5.66666667	2.66666667	4	4.33333333	0	3	12	1.33333333	0	-17	0.33333333	2	0	3	8.66666667	0.33333333	9.66666667	-1.3333333

Figure 1: Relative change in search volume for each word in the Δt weeks (5 rows x 44 columns)

Data Exploration

1. Draw boxplot for 7 words.

From *Figure 2*, there are some outliers existing in each word. The extreme values, such as 60 and -39, do not make any sense because the relative change in the search volume of each word in the Δt=3 weeks is highly unlikely to have such dramatic change. It is concluded that the extreme values and outliers may result from the inaccuracy of the *Google Trend.*

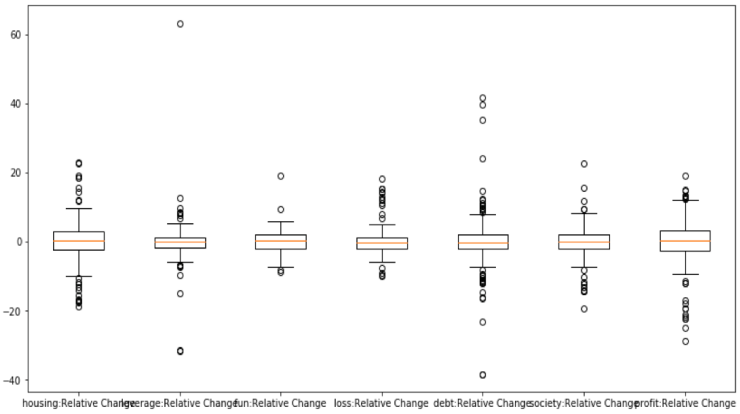


Figure 2: Boxplot for 7 words

2. Draw covariance matrix.

The covariance matrix (*Figure 3*) shows that several words are highly correlated to each other (either positive or negative).

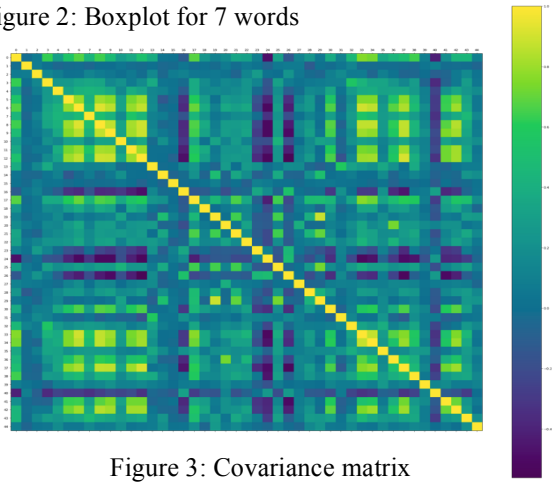


Figure 3: Covariance matrix

Find the best model and Δt

1. Tune parameters for three models including SVM, RandomForest, and Logistic Regression. Get the best combination of parameters for each model based on the mean accuracy rate obtained from cross-validation (cv=5).
2. Find the best model between SVM, RandomForest and Logistic Regression for a specific Δt.
3. Repeat step 1 and 2 for Δt=1, 2, 3 and 4.
4. Compare the selected models for a specific Δt to get the best model and Δt.

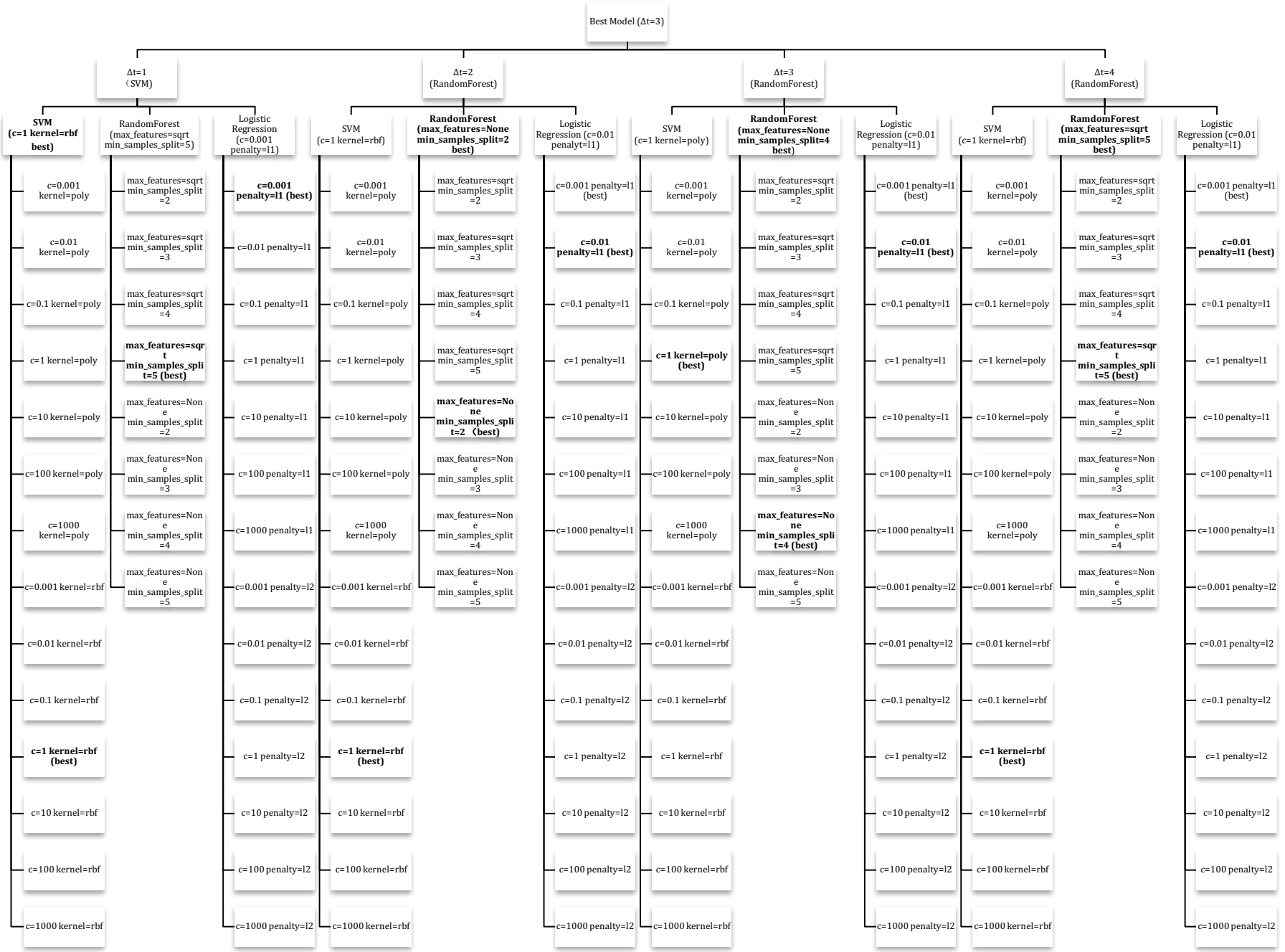


Figure 4: Models and parameters cross validation

III. Results

Figure 5: When Δt=1, SVM (C=1.0, kernel= 'rbf') has the best performance.

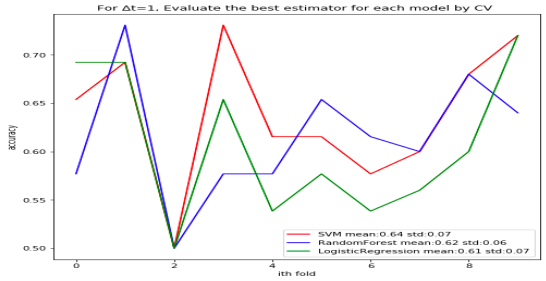


Figure 6: When Δt=2, RandomForest (max_features=None, min_samples_split=2) has the best performance.

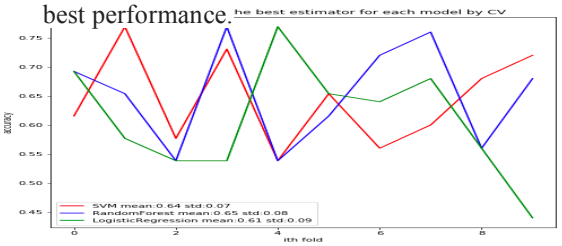


Figure 7: When Δt=3, RandomForest (max_features=None, min_samples_split=4) has the best performance.

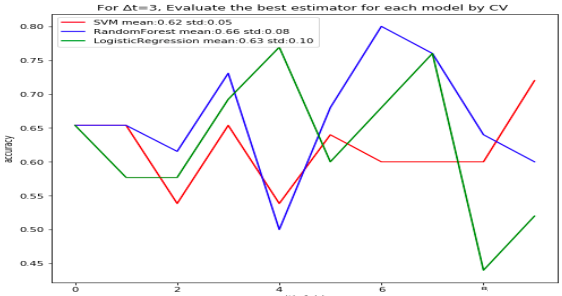


Figure 8: When Δt=4, RandomForest (max_features=sqrt, min_samples_split=5) has the best performance.

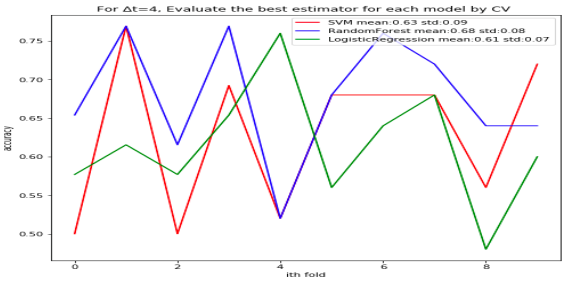
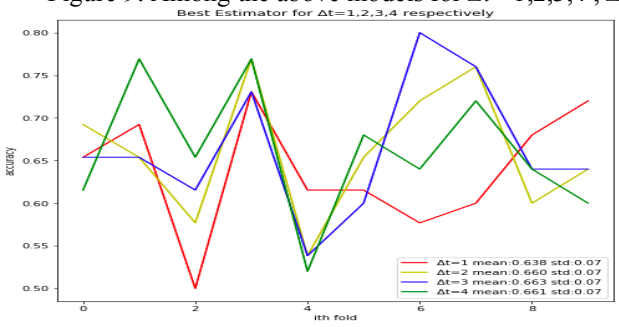


Figure 9: Among the above models for Δt =1,2,3,4 , Δt=3 is the best.



Discussion

When Δt =3 and the model is RandomForest (max_features=None, min_samples_split=4), the estimator has the highest accuracy rate (mean=0.663, std.=0.072). Since there is a limited amount of data available (255 rows), no test data is separated from the original data. The estimator is evaluated on the same dataset as the model fitted due to the limitation. In order to do better analysis, the cross-validation set during training is different from the cross-validation during evaluation based on the number of folds and random_state. The result of the evaluation is that the average accuracy is 0.655, the std. is 0.072, and the base rate is 0.60. Therefore, the accuracy rate of the model built in this analysis is slightly higher than the base rate.

IV. Conclusions

The relative search volume change for some words, which are highly related to investor concern over the past several weeks can reflect investor's decisions on the stock market. It causes the change of the stock price. In conclusion, the period between the online key terms searching of investors and the decision makings from investors would cause notable changes in the financial market and this period is 3 weeks.

References

Preis, T., Moat, H. S., & Stanley, H. E. (2013, April 25). Quantifying Trading Behavior in Financial Markets Using Google Trends. Retrieved November 23, 2017, from https://www.nature.com/articles/srep01684