Regresja zmian cen akcji

Karol Oleszek

16 maja 2020

Spis treści

1 Wstęp

Przewidywanie zmian cen akcji oraz innych instrumentów finansowych znajduje się w centrum zainteresowania inwestorów. Zmiany cen są podstawowym zjawiskiem powodującym bogacenie się lub ubożenie inwestora indywidualnego bądź instytucjonalnego, dlatego też próby zrozumienia i opisania reguł rządzących tym zjawiskiem są kluczowe dla podejmowania skutecznych decyzji o alokacji kapitału.

Teoria rynków kapitałowych proponuje wiele różnych wyjaśnień zmienności cen: hipoteza rynku efektywnego (Bachelier [1]) zakłada, że ceny rynkowe akcji w danej chwili odzwierciedlają wszystkie dostępne informacje o spółce; autorzy i zwolennicy hipotezy krótkoterminową zmienność cen opisują jako losowy ruch wokół efektywnej wartości. Hipoteza rynku efektywnego znalazła wielu zwolenników, którzy poddawali w wątpliwość samą zasadność przewidywania cen (Cowles [4]), jak również zainspirowała powstanie indeksowych funduszy inwestycyjnych.

Hipoteza rynku efektywngeo spotkała się z szeroką krytyką ze strony ekonomistów i inwestorów giełdowych, którzy wskazywali na kontrprzykłady obalającę hipotezę. Współcześnie właściwie wszystkie duże organizacje finansowe używają różnego rodzaju systematycznych narzędzi do analizy i prognozy zmian cen na rynkach kapitałowych (Graham Capital Management [5]). Duże oraz wciąż rosnące znaczenie ma też algorytmiczny handel (Capgemini [2]).

Do złożonego problemu jakim jest symulacja i prognostyka zachowania rynków kapitałowych stosuje się bardzo szeroki wachlarz metod statystycznych, algorytmicznych i ekonometrycznych. Duże zastosowanie mają metody uczenia maszynowego (Shunrong Shen [7]), w tym głebokie sieci neuronowe o niekonwencjonalnych architekturach. Ponadto do prognostyki coraz częściej używa się analizy języka naturalnego (Zhaoxia Wang [8]).

Poniższa praca zawiera przekrojową regresję zmian cen akcji na rynku amerykańskim w 2019 z wykorzystaniem standardowych narzędzi ekonometrycznych. Zbiór danych służący do konstrukcji modelu zawiera dane z roku 2018, dotyczące sytuacji finansowych, kapitałowych i operacyjnych spółek, zawarte w formie wskaźników i pozycji ze sprawozdań finansowych.

2 Cel projektu

2.1 Model wyboru akcji do celów inwestycyjnych

Celem projektu jest wyznaczenie bazowego poziomu efektywności wyboru spółek, których akcje w nadchodzącym roku zyskają na wartości. Za wybór odpowiadał będzie model, który powstanie przy użyciu metody najmniejszych kwadratów i który będzie mógł służyć jako punkt odniesienia do badania efektywności innych metod predykcji.

Efektywność prognostyczna modelu zostanie zbadana przy użyciu *średniego błędu prognozy ex post*, danego wzorem:

$$ME = \frac{1}{s}\sum_{t=1}^{s}(y_t - y_t^P)$$

,

Gdzie:
s - ilość obserwacji w testowym zbiorze danych
y_t - prawdziwa wartość zmiennej objaśnianej
y_t^P - prognozowana wartość zmiennej objaśnianej

2.2 Zbadanie zależności pomiędzy zmianą cen, a informacjami finansowymi

Ponadto model posłuży do oceny wpływu informacji finansowych zawartych w publicznie dostepnych źródłach na przyszłą wartość spółek giełdowych. Ocena ta może być użyteczna przy podejmowaniu decyzji o tym, jakie dane zbierać na temat spółek w celu skutecznego przewidywania ich przyszłej wyceny.

Miarą tej oceny będzie współczynnik determinacji R^2, dany wzorem:

$$R^2 = \frac{\sum_{t=1}^{n}(y_t^P - \overline{y})^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}$$

,

Gdzie:
n - ilość obserwacji w uczącym zbiorze danych
y_t - prawdziwa wartość zmiennej objaśnianej
y_t^P - prognozowana wartość zmiennej objaśnianej
$\overline{y}$ - średnia arytmetyczna zmiennej objaśnianej

3 Opis danych

3.1 Zbiór danych

Zbiór danych użyty w projekcie pochodzi z internetowej platformy Kaggle (Carbone [3]). Zawiera on zmienną objaśnianą Y - procentową zmianę ceny akcji danej spółki w 2019 roku, oraz zmienne objaśniające $X_i, i = 1 \dots k$ - k-1 wskaźników finansowych i pozycji z formularza *10-K*[1], a także zmiennej kategorycznej oznaczającej sektor gospodarki rozważanej spółki.

3.2 Usuwanie braków danych

Dane zostały zebrane przy użyciu interfejsu programistycznego *Financial Modeling Prep API* i zawierały pewne braki wynikające z różnic w dokumentach źródłowych. Dla celów analizy usunięte zostały wszystkie obserwacje, w których brakowało więcej niż 50 wartości oraz wszystkie zmienne, w których co najmniej 10% obserwacji nie miało przypisanej wartości. Po tej transformacji, w zbiorze danych pozostało 4122 obserwacji oraz 179 zmiennych (4392 x 222, 9,98% braków przed transformacją). Wciąż brakujące 0,82% wartości zostało zastąpionych średnimi arytmetycznymi odpowiednich zmiennych.

3.3 Transformacja zmiennej kategorycznej

Kategoryczna zmienna objaśniająca *Sector*, która przyjmowałą 11 różnych wartości (Consumer Cyclical, Energy, Technology, Industrials, Financial Services, Basic Materials, Communication Services, Consumer Defensive, Healthcare, Real Estate, Utilities) została przekształcona na 10 zmiennych zerojedynkowych. Po tej operacji zbiór danych składał się ze 188 zmiennych.

3.4 Zmienne objaśniające

Zbiór danych składa się z grupy zmiennych opisujących realne wielkości ze sprawozdań finansowych wyrażone w dolarach amerykańskich np. zysk brutto, wydatki na badania i rozwój. Na drugą grupę zmiennych składają się wskaźniki finansowe, będące często przekształconymi zmiennymi z pierwszej grupy, wyrażone jako stosunek różnych wielkości np. zysk na akcję, wzrost zysku w ciągu roku. Trzecia grupa zmiennych to przekształcona zmienna *Sector*. Ze względu na liczbę zmiennych pełna lista zmiennych wraz ze statystykami znajduje się w załączniku do pracy.

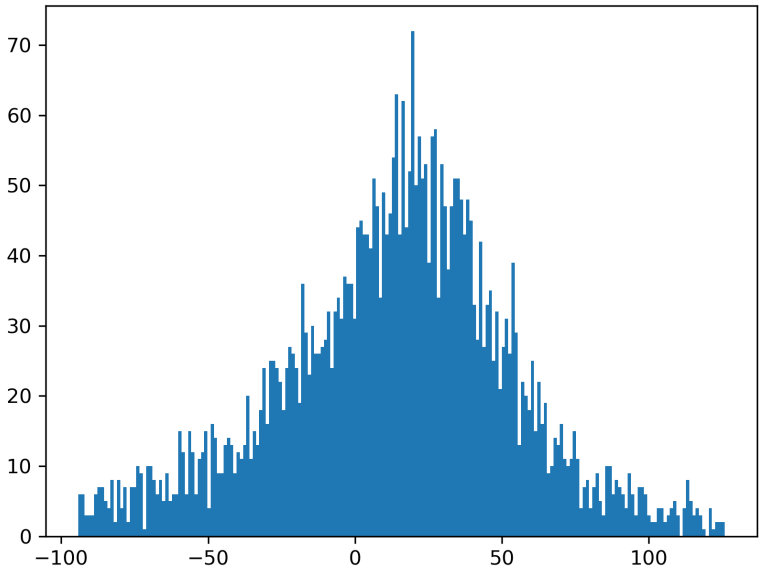
[1] *Form 10-K* jest to coroczne podsumowanie finansowe składane przez amerykańskie spółki giełdowe do *U.S. Securities and Exchange Commission*, federalnej agencji nadzoru finansowego.

3.5 Rozkład zmiennej objaśnianej

Zmienna objaśniana Y to wyrażona w procentach zmiana ceny akcji spółki w roku 2019. Przyjmuje ona wartości -99,86% do +3756,72%. Większa część wartości jest większa od zera co wskazuje na pozytywną bazową efektywność tzw. strategii *buy and hold*, która polega na zakupie akcji, a następnie oczekiwaniu na wzrost jej wartości.

Statystyka	Wartość
Średnia arytmetyczna	21,15
Odchylenie standardowe	84,93
Kwartyl dolny	-9,30
Mediana	17,83
Kwartyl górny	40,92
Wartość najmniejsza	-99,86
Wartość największa	3756,71

Tabela 1: Statystyki - zmienna objaśniana



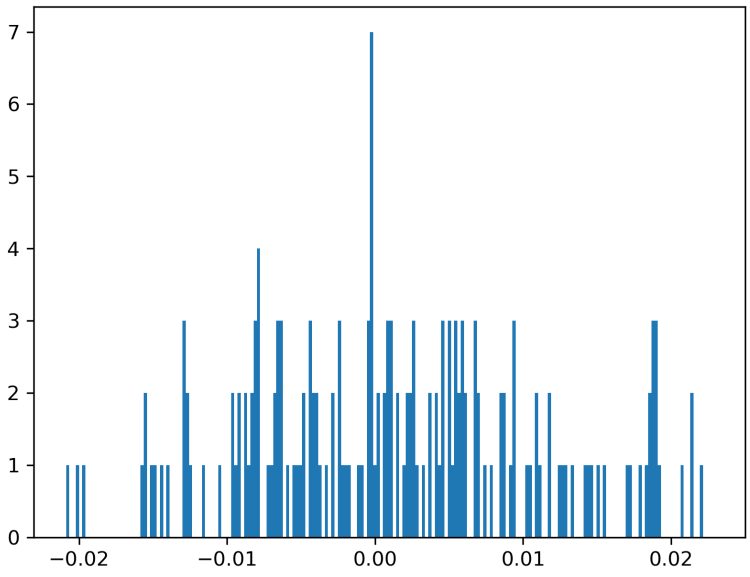
Rysunek 1: Histogram zmiennej objaśnianej

3.6 Korelacja

Korelacja zmiennej objaśnianej ze zmiennymi objaśnianymi jest bardzo słaba, co wskazuje na potencjalnie silnie nieliniowy charakter zachodzących zależności.

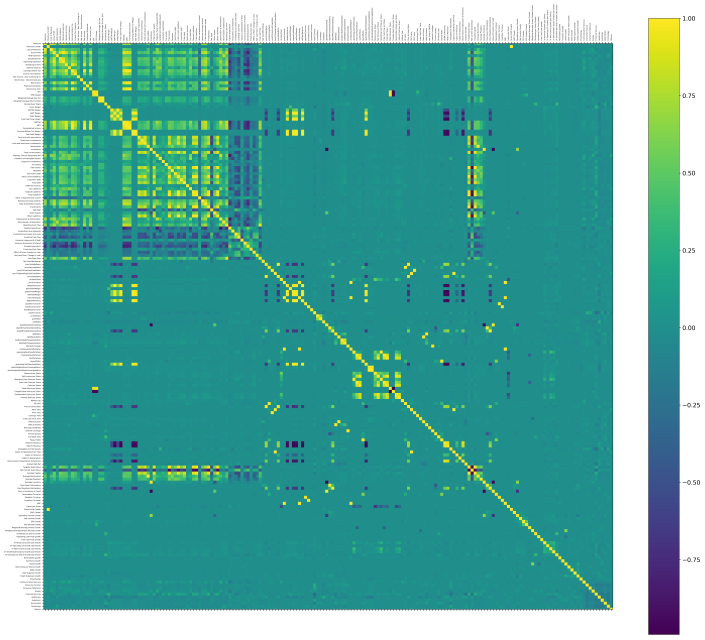
Statystyka	Wartość
Średnia arytmetyczna	0.0006535276131026212
Odchylenie standardowe	0.01430737685985159
Kwartyl dolny	-0.006560508005214499
Mediana	0.0009029552423272379
Kwartyl górny	0.008516164727314403
Wartość najmniejsza	-0.07707504123521973
Wartość największa	0.04058496958769639

Tabela 2: Korelacja zmiennej objaśnianej ze zmiennymi objaśnianymi



Rysunek 2: Histogram korelacji zmiennej objaśnianej ze zmiennymi objaśnianymi

Korelacja pomiędzy niektórymi zmiennymi objaśniającymi jest silna, co można wyjaśnić zależnościami pomiędzy rozmiarem spółki, a wielkościami w *10-K Form*. Nie wszystkie bazowe zmienne objaśniające mogą znaleźć się w modelu, ponieważ spowodowałoby to wystąpienie zjawiska współliniowości.



Rysunek 3: Macierz korelacji

3.7 Podział zbioru danych

Zbiór danych podzielony jest na dane treningowe(90%) użyte do estymacji parametrów modelów oraz dane testowe(10%) służące do oceny prognozy *ex post*.

4 Wybór postaci modelu oraz dobór zmiennych do modelu

4.1 Specyfikacja kryteriów

Praca odpowiada na dwa pytania:

1. Jaka jest bazowa efektywność predykcji?
2. Jaki jest wpływ informacji zawartych w rozważanych danych na zmiany cen akcji?

Odpowiedzią na pierwsze pytanie jest efektywność modelu o najlepszych właściwościach prognostycznych.

Odpowiedzią na drugie pytanie jest ocena współczynnika determinacji poprawnie zweryfikowanego modelu ekonometrycznego.

4.2 Rozważana klasa funkcji

Modele wybrane są z klasy funkcji dopuszczalnych $\mathscr{F} : \mathbb{R}^{\mathbb{K}} \to \mathbb{R}$, gdzie K: liczba zmiennych w zbiorze danych.

Funkcje przyjmują analityczną postać:

$\hat{y} = a_0 + a_1 * X_1 + ... + a_k * X_k$;

parametry szacowane są przy użyciu metody najmniejszych kwadratów.

Ze względu na potencjalnie nielinowy charakter zależności rozważony został zbiór danych rozszerzony o przetransformowane zmienne: $e^{X_i}, X_i^2, \frac{1}{1+X_i}$ o łącznym rozmiarze 4392 x 888. Dalsze rozszerzanie zbioru danych osłabiło by stabilność numeryczną modeli i nadmiernie zwiększyło by wymiar Wapnika-Czerwonienisa, co osłabiło by możliwości generalizacji przez modele (Kamiński [6]).

4.3 Rozważany podzbiór funkcji

Ze względu na ilość zmiennych, użycie metody o wykładniczej złożoności obliczeniowej, np. metody Hellwiga, byłoby nieefektywne. W związku z tym rozważany jest K-elementowy zbiór funkcji wyłoniony przy użyciu uproszczonej *metody krokowej wstecz*. Rozważany jest model ze wszystkimi zmiennymi, a następnie z modelu usuwana jest najmniej istotna statystycznie zmienna. Procedura jest powtarzana dopóki w modelu jest więcej niż jedna zmienna. Dla celów stworzenia podzbioru funkcji nie jest sprawdzana normalność rozkładu resztowego.

4.4 Kryterium wyboru modelu prognostycznego

Spośród rozważanego podzbioru funkcji wybrana jest funkcja z najmniejszym *absolutnym błędem prognozy ex post* oszacowanym przy użyciu sprawdzianu krzyżowego na danych treningowych.

Sprawdzian krzyżowy polega na podziale zboriu danych na J podzbiorów, a następnie wykonaniu J ocen błędu prognozy(za każdym razem inny podzbiór jest uznawany za testowy, J-1 podzbiorów treningowych) i obliczeniu ich średniej.

$$\hat{ME} = \frac{1}{J}\sum_{t=1}^{J}(ME_i)$$

4.5 Kryterium wyboru modelu analitycznego

Spośród rozważanego podzbioru funkcji wybrana jest funkcja z najwyższym współczynnikiem determinacji R^2, która została poprawnie zweryfikowana. W przypadku wystąpienia niepożadanych zjawisk w modelu, rozważany podzbiór funkcji powiększony zostaje o model z zastosowanymi stosownymi korektami.

5 Weryfikacja poprawności modelu

Poniżej opisane są procedury weryfikacji poprawności modelu oszacowanego metodą najmniejszych kwadratów. Przyjęty poziom istotności $\alpha = 0,05$.

5.1 Współliniowość

Wyznaczamy k modeli MNK, gdzie zmienną objaśnianą jest jedna ze zmiennych objaśniających, a zmiennymi objaśniającymi pozostałe zmienne. Wyznaczamy k współczynników determinacji; jeżeli którykolwiek z nich większy jest niż 0,9 , to w modelu występuje niepożądane zjawisko współliniowości zmiennych.

5.2 Koincydencja

Jeżeli dla każdego i=1..k:

$$sgn(r_i) = sgn(a_i)$$

,

gdzie:

r_i: korelacja pomiędzy zmienną objaśnianą, a i-tą zmienną objaśniającą.

a_I: oszacowany i-ty parametr modelu.

to model jest koincydentny. Koincydencja modelu jest pożądaną cechą.

5.3 Efekt katalizy

Niech (X_i, X_j) będzie regularną parą korelacyjną. Wówczas jeżeli $r_{ij} < 0$ lub $r_{ij} > \frac{r_i}{r_j}$, gdzie:

r_{ij}: korelacja między i-tą, a j-tą zmienną objaśniającą.

to zmienna X_i jest katalizatorem. Poprawny model nie zawiera katalizatorów.

5.4 Normalność rozkładu reszt

Normalność rozkładu składnika resztowego modelu jest cechą niezbędną do poprawnej interpretacji m.in. testów istotności parametrów modelu. Normalność rozkładu można zbadać przy użyciu testu Jarque-Bera.

H_0: składnik losowy modelu ma rozkład normalny.

H_1: składnik losowy modelu nie ma rozkładu normalnego.

Statystyka testowa JB ma rozkład χ^2 o dwóch stopniach swobody.

$$JB = \frac{n}{6}(A^2 + \frac{1}{4}(K-3)^2)$$

,

gdzie:

Współczynnik skośności:

$$A = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^3}{(\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2)^{\frac{3}{2}}}$$

Kurtoza:

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^4}{(\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2)^2}$$

5.5 Istotność zmiennych objaśniających

Zmienne w poprawnym modelu są statystycznie istotne. Do badania istotności zmiennych służy test t-Studenta.

H_0: $a_i = 0$.

H_1: $a_i \neq 0$.

Statystyka testowa t_{a_i} ma rozkład t-Studenta o n-(k+1) stopniach swobody.

$$t_{a_i} = \frac{a_i}{D(a_i)}$$

Dyspersja estymatora i-tego parametru modelu:

$$D(a_i) = \sqrt{d_{ii}}$$

Macierz kowariancji estymatora a:

$$\hat{D}^2(a) = S^2(X^T X)^{-1}$$

Estymator wariancji s^2 składnika losowego:

$$S^2 = \frac{e^T e}{n-(k+1)}$$

5.6 Istotność współczynnika determinacji

Istotność współczynnika determinacji(inaczej istotność wszystkich zmiennych naraz) jest pożądaną cechą modelu i może być zbadana za pomocą testu F.

H_0: $a_1 = a_2 = ... = a_k = 0$.

H_1: $a_1 \neq 0 \vee a_2 \neq 0 \vee ... \vee a_k \neq 0$.

Statystyka testowa F ma rozkład F-Snedecora-Fishera z $r_1 = k$ i $r_2 = n - (k+1)$ stopniami swobody.

$$F = \frac{R^2}{(1-R^2)} \frac{n-(k+1)}{k}$$

5.7 Liniowość postaci modelu

Do badania liniowości modelu ekonometrycznego służy test serii.

H_0: model hipotetyczny jest liniowy.

H_1: model nie jest liniowy.

Statystyka testowa r ma rozkład serii z parametrami N_1 oraz N_2.

r - liczba serii w wektorze resz modelu(uporządkowanych wg. wartości zmiennej objaśnianej).

N_1 , N_2 - liczba dodatnich i ujemnych reszt modelu.

5.8 Homoskedastyczność

Homoskedastyczność jest pożądaną cechą modelu. Ze względu na wysoką proporcję ilości zmiennych objaśniających do ilości obserwacji homoskedastyczność najlepiej sprawdzić jest przy użyciu testu Goldfielda-Quandta.

H_0: $s_1^2 = s_2^2$. Homoskedastyczność.

H_1: $s_2^2 \neq s_2^2$. Heteroskedastyczność.

Statystyka testowa F ma rozkład F-Snedecora-Fishera z $r_1 = n_1 - (k+1)$ i $r_2 = n_2 - (k+1)$ stopniami swobody.

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$$

Próba podzielona jest na dwa zbiory i badana jest równość wariancji w podpróbach.

$$\hat{s}_1^2 = \frac{{e_1}^T e_1}{n_1 - (k+1)}$$

$$\hat{s}_2^2 = \frac{{e_2}^T e_2}{n_2 - (k+1)}$$

5.9 Stabilność parametrów modelu

Stabilność parametrów modelu, która jest pożądana cechą, może zostać zweryfikowana przy pomocy testu Chowa.

H_0: $\alpha = \beta = \gamma$. Paremetry modelu są stabilne.

H_1: Parametry modelu nie są stabilne.

Statystyka testowa F ma rozkład F-Snedecora-Fishera z $r_1 = k+1$ i $r_2 = n-2(k+1)$ stopniami swobody.

$$F = \frac{RSK - (RSK_1 + RSK_2)}{RSK_1 + RSK_2} \frac{n-2(k+1)}{k+1}$$

RSK - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + ... + \alpha_k x_{kt} + \epsilon_t, t = 1, 2, ..., n$$

RSK_1 - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + \epsilon_{1t}, t = 1, 2, ..., n_1$$

RSK_2 - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \gamma_0 + \gamma_1 x_{1t} + ... + \gamma_k x_{kt} + \epsilon_{2t}, t = n_1 + 1, n_1 + 2, ..., n$$

5.10 Autokorelacja składnika losowego

Autokorelacja składnika losowego jest niepożądaną cechą modelu ekonometrycznego. Występowanie zjawiska autokorelacji pierwszego rzędu można badać przy założeniu, że:

$$e_t = re_{t-1} + h_t$$

Służy do tego test Durbina-Watsona.

H_0: $r = 0$. Zjawisko autokorelacji I rzędu nie występuje.

H_1: $r \neq 0$. Zjawisko autokorelacji I rzędu występuje.

Jeżeli statystyka testowa DW znajduje się w przedziale $(1,5; 2,5)$, to można przyjąć, że zjawisko autokorelacji I rzędu nie występuje.

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1}^2)}{\sum_{t=1}^{n} e_t^2}$$

6 Korekty

7 Wybrane modele

8 Prognoza

9 Interpretacja

10 Podsumowanie

11 Spis tabel

12 Spis rysunków

13 Literatura

[1] Louis Bachelier. *Théorie de la Spéculation*. 1900.

[2] Capgemini. *High Frequency Trading: Evolution and the Future*. 2012.

[3] Nicolas Carbone. *200+ Financial Indicators of US stocks (2014-2018)*. 2019.

[4] Alfred Cowles. *Can stock market forecasters forecast?* 1932.

[5] L.P. Graham Capital Management. *Systematic Global Macro: Performance, Risk and Correlation Characteristics*. 2013.

[6] Bogumił Kamiński. *Teoria uczenia statystycznego z perspektywy ekonometryka*. 2017.

[7] Tongda Zhang Shunrong Shen Haomiao Jiang. *Stock Market Forecasting Using Machine Learning Algorithms*. 2012.

[8] Zhiping Lin Zhaoxia Wang Seng-Beng Ho. *Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment*. 2018.