Università degli studi di Milano - Bicocca

Scuola di Economia e Statistica
Corso di laurea Magistrale in
SCIENZE STATISTICHE ED ECONOMICHE

Natural Language Processing in finance

Un'applicazione basata sul modello FinBERT

Relatore: Prof. Matteo Maria Pelagatti
Correlatore: Prof. Antonio Candelieri

Tesi di Laurea di:
Moreno Sanna
Matr. N. 783008

Anno Accademico 2020 - 2021

20/01/2022

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

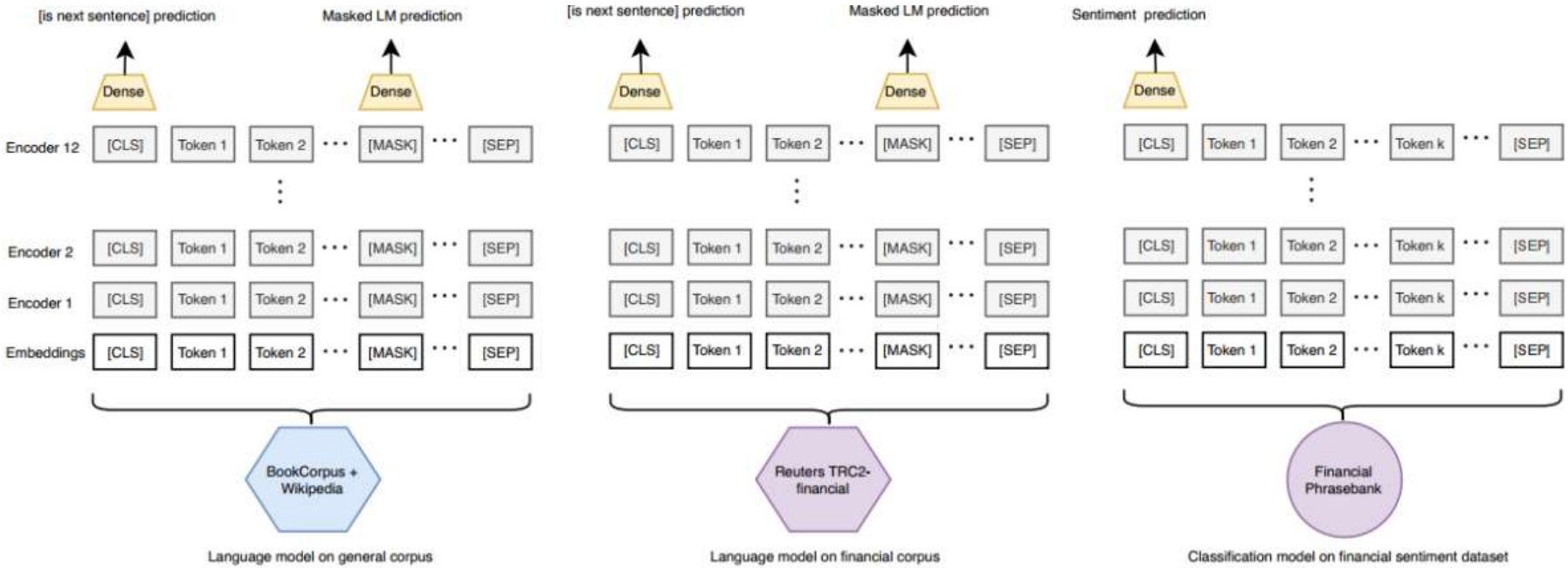
Introduzione e scopo della tesi

- Le opinioni condivise sui social network influiscono sull'andamento dei titoli finanziari?

- Scopo della tesi:
  1) Indagare la relazione fra i messaggi condivisi su Twitter e i risultati degli asset Tesla e BMW.
  2) Provare a costruire una strategia finanziaria basata sul sentimento estratto dai tweet.

- I messaggi sono stati analizzati tramite un modello di Natural Language Processing di nome FinBERT.

Il modello FinBERT



Architettura:

- 12 layer encoder.
- 1 layer finale.
- 110M di parametri.

Tre step di training:

- Pretrainig su corpora generali.
- Pretrainig su corpora finanziario.
- Fine-tuning for Sentiment Analysis.

Sentiment Analysis con FinBERT:

- Classifica ogni frase come positiva, negativa o neutrale, attribuendo a ciascuna etichetta una probabilità.
- Calcola un sentiment score per ogni frase.

Costruzione dataset

1. Analisi dei tweet condivisi nel periodo 01.01.2019-31.03.2020 (12.4M Tesla, 6M BMW).
2. Somma e media degli indicatori per giorno, con e senza *retweet* (*_pond*).
3. Espressione dei risultati dei due asset come:
   - Rendimenti.
   - Extrarendimenti (rendimenti Tesla e BMW meno i rendimenti del fondo CARZ).
   - Segno dei rendimenti e degli extrarendimenti (Var. dicotomica: 1 segno positivo, 0 negativo)
4. Unione degli indicatori e dei risultati, sulla base del giorno di manifestazione.
   - Sono stati considerati anche il primo (*_lag1*) e il secondo ritardo (*_lag2*).

Variabili:

- Indicatori probabilità (es: *sum_positive).*
- Indicatori sentimento (es: *mean_sentiment_score*).
- Conteggio etichette e tweet (es: *n_positive*).

Dim. datasets: 311 righe x 79 colonne.
Training set: 250 righe x 79 colonne.
Test set: 61 righe x 79 colonne.

Modelli e metodi statistici utilizzati

Problemi di regressione:

- ***Metodo backward basato sul criterio di Acaike*** : $AIC = -2loglikelihood(\hat{\beta}, \hat{\sigma}^2) + 2p$
- ***Regressione Lasso***: $\hat{\boldsymbol{y}} = \hat{\mu}_\lambda^L \mathbf{1} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda^L$

$$(\hat{\mu}_\lambda^L, \hat{\boldsymbol{\beta}}_\lambda^L) = argmin_{(\boldsymbol{\mu},\boldsymbol{\beta})\in RxR^p} \frac{1}{2n} \left|\left|\boldsymbol{y} - \mu\mathbf{1} - \boldsymbol{X\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_{\mathbf{1}};\ \lambda \geq \mathbf{0}$$

- ***Support Vector Machine, con kernel polinomiale:***

$$f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i \prec \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x_i}) \succ; \qquad K_p(x, x') = (coef0 + \gamma \prec \boldsymbol{x}, \boldsymbol{x'} \succ)^d$$

- ***Modello AR(1) + eGARCH(0,1):***

$$(1 - \phi_1 B)(Y_t - \boldsymbol{\alpha}^T \boldsymbol{X_t}) = \varepsilon_t\ ;\quad \varepsilon_t = \xi_t \sigma_t;\quad ln(\sigma_t^2) = \omega + \beta_1 ln(\sigma_{t-1}^2);\quad \xi_t \sim ged(0,1,r)\ i.i.d$$

Problemi di classificazione:

- ***Regressione logistica regolarizzata***: $p_i = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta}) = \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{(1+e^{\boldsymbol{X}_i\boldsymbol{\beta}})}$
- ***Support Vector Machine, con kernel radiale e polinomiale:***

$$f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i y_i \prec \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x}_i) \succ + \beta_0;\quad K_r(x, x') = e^{-\gamma \left|\left|\boldsymbol{x}-\boldsymbol{x'}\right|\right|^2};\quad K_p(x, x') = (coef0 + \gamma \prec \boldsymbol{x}, \boldsymbol{x'} \succ)^d$$

- ***Random Forest:*** $\bar{c}(x) = Mode[\hat{c}^b(x), b = 1, \ldots B]$

Metriche utilizzate e valutazione delle performance

Problemi di regressione

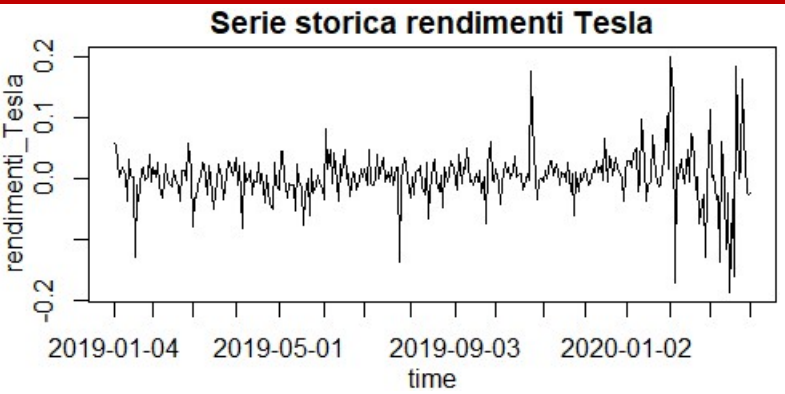
- ***Modello di benchmark:*** Media della serie nel training set.
- ***Indice di correlazione di Pearson***: $\rho = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$
- ***Root Mean Square Error:*** $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- ***RRMSE***: $RRMSE = \frac{RMSE_1}{RMSE_0}$
- $\boldsymbol{R^2}$***:*** $R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}$

Problemi di classificazione

- ***Modello di benchmark:*** Classificatore completamente random.
- ***Indice di correlazione di kendall***: $\tau = \frac{nc - n}{\frac{n(n-1)}{2}}$
- ***Accuracy:*** $ACC = \frac{1}{n}\Sigma_{i=1}^{n} I(y_i = \hat{y}_i)$
- ***AUC***
- ***Matrice di confusione e Curva ROC***
- ***Mean Decrease Accuracy***

Rendimenti Tesla

Serie



$\mu = 0.0026$

$\sigma = 0.043$

$\sigma_{\{02.20-03.20\}} = 0.089$

Variabili più correlate

$\rho \geq 0.2; \rho \leq -0.2$ (Pearson)

	Correlazione
sum_sentiment_score	0.45
mean_sentiment_score	0.41
mean_positive	0.36
mean_sentiment_score_pond	0.35
sum_sentiment_score_pond	0.35
mean_positive_pond	0.25
n_positive_pond	0.22
n_positive	0.21
sum_negative	-0.21
n_negative_pond	-0.22
n_negative	-0.26
mean_negative_pond	-0.30
mean_negative	-0.32

Variable Selection

- 6 variabili selezionate tramite **regressione lasso** con λ=0.0021, (LOOCV):

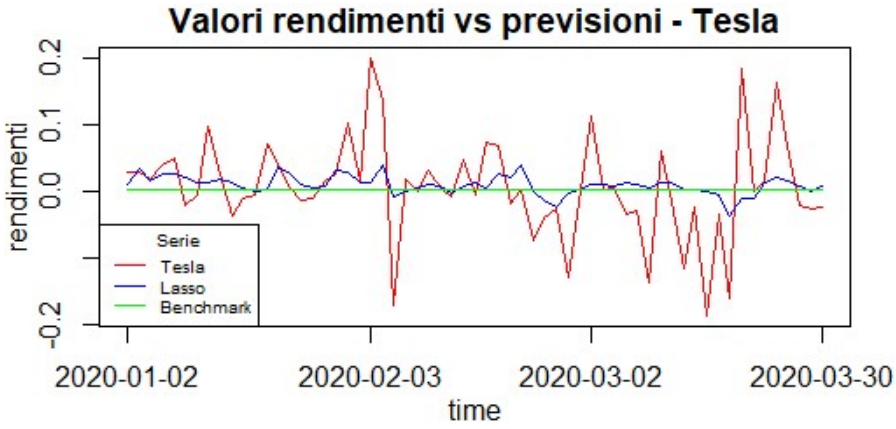
sum_negative, sum_sentiment_score, mean_positive, mean_negative_lag1, mean_neutral_pond_lag1, n_retweet_lag2

Modello Scelto

- **Regressione Lasso** con con λ=0.0021 (LOOCV).

Tabella riepilogativa e grafico

	RMSE	RRMSE	R^2
Lasso	0.0697	0.93	0.13
Benchmark	0.0749	1.00	0.00



Segno rendimenti Tesla

Serie

$S_+ = 163$

$S_- = 148$

Variabili più correlate

$\tau \geq 0.2; \tau \leq -0.2$
(Kendall)

	Correlazione
sum_sentiment_score	0.26
mean_sentiment_score	0.26
mean_sentiment_score_pond	0.22
sum_sentiment_score_pond	0.21
mean_negative	-0.21

Variable Selection

- Una variabile selezionata tramite **regressione logistica regolarizzata** con λ=0.066, (LOOCV):

  sum_sentiment_score

- Le variabili sono state standardizzate.

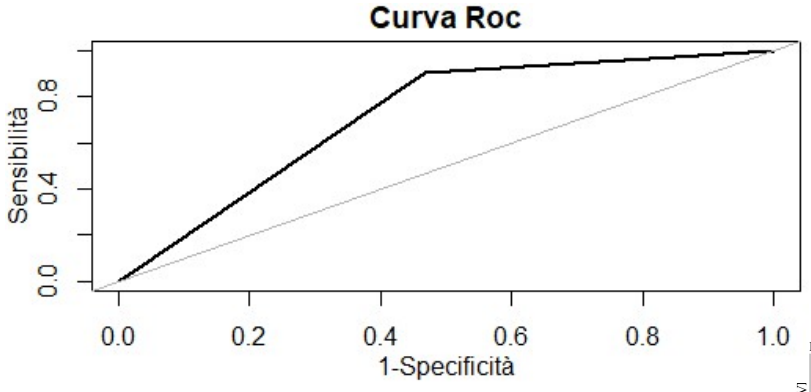
Modello Scelto

- **SVM con kernel radiale** con i seguenti parametri: $\gamma = 1, C = 1$ (CV 10 folds).

Tabella risultati, Matr di confusione e Curva ROC

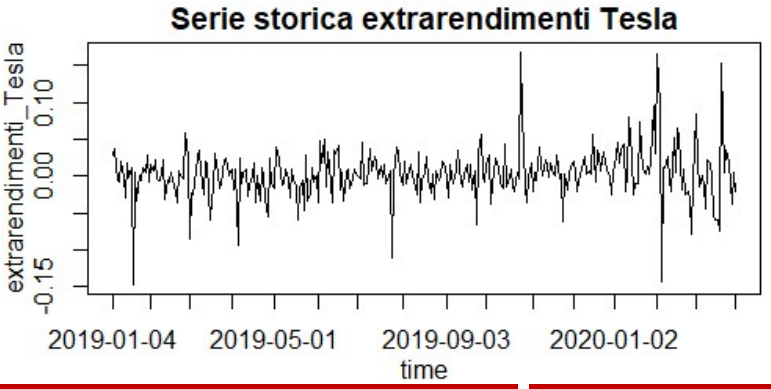
	Accuracy	AUC
SVM	0.721	0.718
SVM vs benchmark	0.221	0.218

	pred_value_0	pred_value_1	Total
True_value_0	16	14	30
True_value_1	3	28	31
Total	19	42	61



Extrarendimenti Tesla

Serie



$\mu = 0.003$

$\sigma = 0.034$

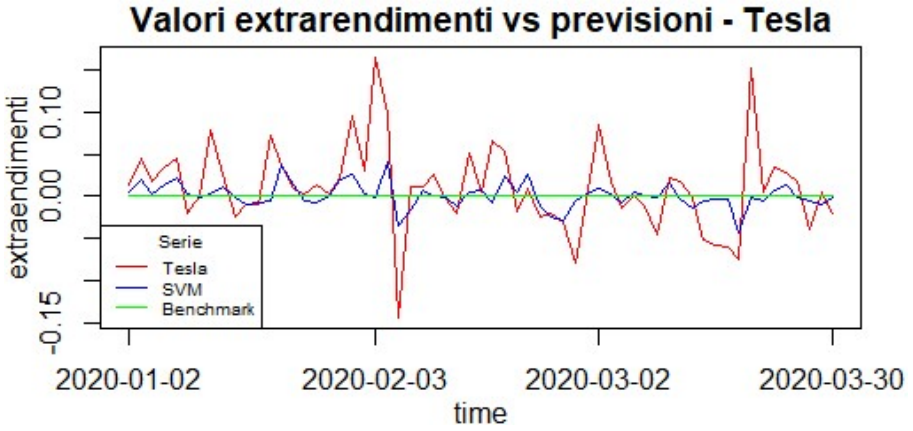
$\sigma_{\{02.20-03.20\}} = 0.058$

Modello Scelto

- **SVM con kernel polinomiale** di parametri: $\gamma = 1.5 * 10^{-4}, coef0 = 3.5,\ degree = 5,$ $C = 1, \epsilon =$ 0.5 (CV 10 folds).

Tabella riepilogativa e grafico

	RMSE	RRMSE	R^2
SVM	0.0453	0.88	0.19
Benchmark	0.0513	1.00	-0.03



Variabili più correlate

$\rho \geq 0.2; \rho \leq -0.2$ (Pearson)

	Correlazione
sum_sentiment_score	0.50
mean_sentiment_score	0.45
mean_positive	0.39
sum_sentiment_score_pond	0.36
mean_sentiment_score_pond	0.36
mean_positive_pond	0.25
n_positive	0.22
n_positive_pond	0.21
sum_negative	-0.25
n_negative_pond	-0.25
n_negative	-0.30
mean_negative_pond	-0.31
mean_negative	-0.35

Variable Selection

- 8 variabili selezionate tramite **regressione lasso** con λ=0.0015, (LOOCV):

  n_neutral, sum_sentiment_score, mean_positive, mean_neutral_pond, mean_negative_lag1, mean_neutral_pond_lag1, n_retweet_lag2, rend_lag2

- Le variabili sono state standardizzate.

Segno extrarendimenti Tesla

Serie

$S_+ = 177$

$S_- = 134$

Variabili più correlate

$\tau \geq 0.2 \; ; \tau \leq -0.2$
(Kendall)

	Correlazione
sum_sentiment_score	0.27
mean_sentiment_score	0.26
sum_sentiment_score_pond	0.22
mean_sentiment_score_pond	0.21
mean_negative	-0.21

Variable Selection

- 6 variabili selezionate tramite **regressione logistica regolarizzata** con λ=0.027, (LOOCV):

  sum_sentiment_score, mean_sentiment_score_lag1, sum_negative_pond_lag2, n_positive_pond_lag2, sgn_lag1, sgn_lag2

- Le variabili sono state standardizzate.

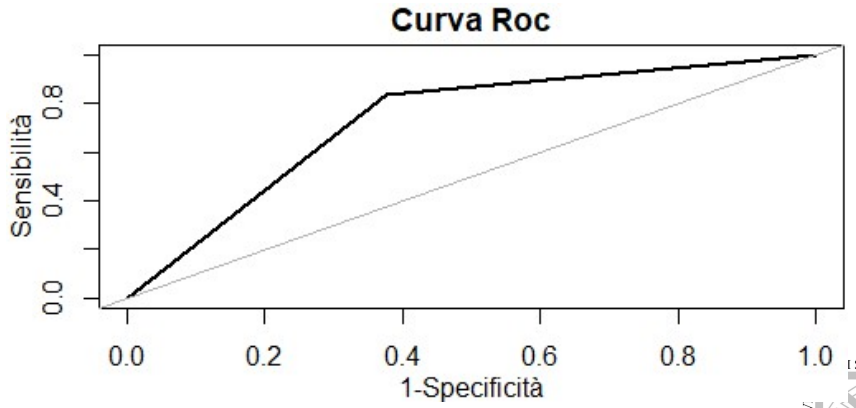
Modello Scelto

- **SVM con kernel radiale** con i seguenti parametri: $\gamma = \frac{1}{6}, C = 1$ (CV 10 folds).

Tabella risultati, Matr di confusione e Curva ROC

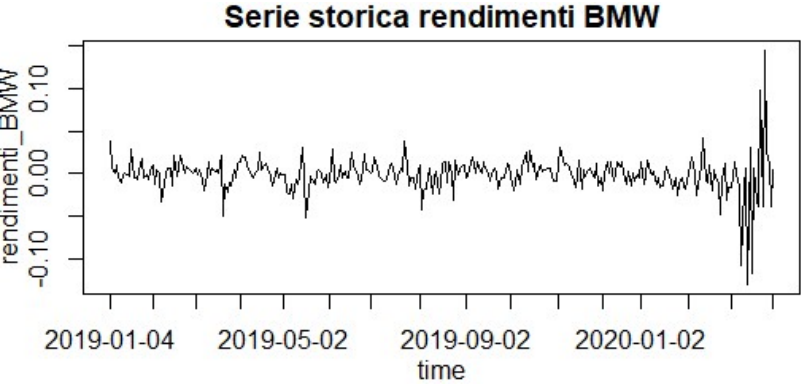
	Accuracy	AUC
SVM	0.754	0.731
SVM vs benchmark	0.254	0.231

	pred_value_0	pred_value_1	Total
True_value_0	15	9	24
True_value_1	6	31	37
Total	21	40	61



Rendimenti BMW

Serie



$$\mu = -0.001$$

$$\sigma = 0.021$$

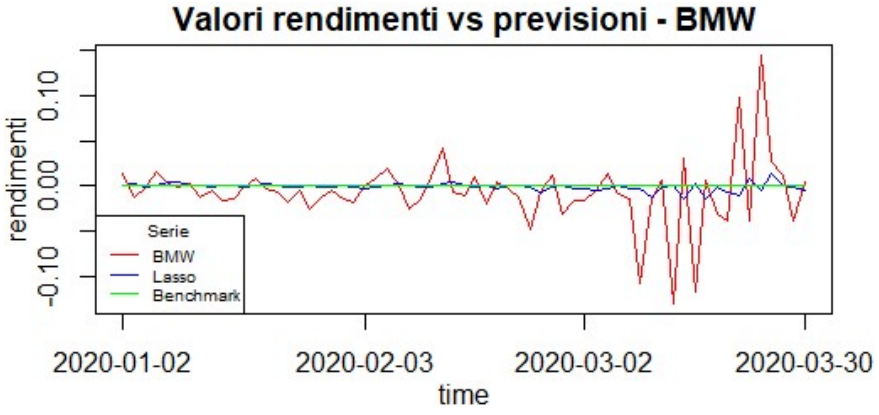
$$\sigma_{\{02.20-03.20\}} = 0.064$$

Modello Scelto

- **Regressione lasso** con λ= 0.0009 (LOOCV).

Tabella riepilogativa e grafico

	RMSE	RRMSE	R^2
Lasso	0.0391	1.02	-0.09
Benchmark	0.0382	1.00	-0.04



Variabili più correlate

$\rho \geq 0.08\,;\rho \leq -0.08$
(Pearson)

	Correlazione
rend_lag2	0.25
mean_sentiment_score_pond	0.10
mean_sentiment_score_pond_lag1	0.10
sum_sentiment_score_pond	0.09
sum_sentiment_score_pond_lag1	0.09
mean_negative_pond_lag1	-0.08

Variable Selection

- 10 variabili selezionate tramite **regressione lasso** con λ=0.0009, (LOOCV):

sum_negative, mean_negative, sum_negative_lag1, mean_negative_lag1, mean_sentiment_score_pond_lag1, rend_lag1, n_positive_pond_lag2, mean_negative_lag2, mean_sentiment_score_lag2, mean_sentiment_score_pond_lag2

Segno rendimenti BMW

Serie

$S_+ = 154$

$S_- = 158$

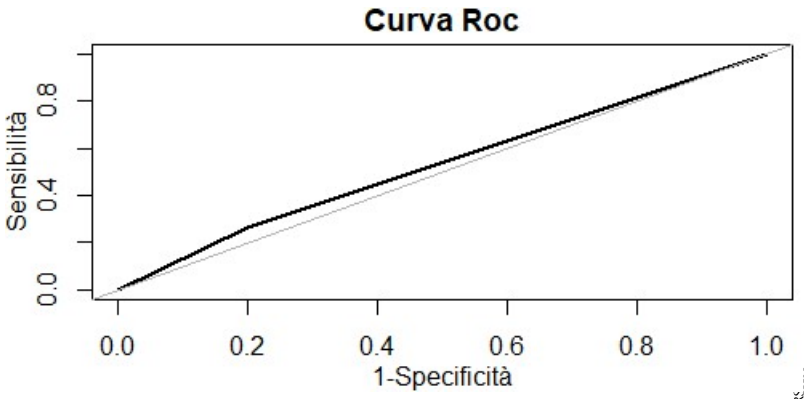
Modello Scelto

- **Random Forest** con m=$\sqrt{79}$ (CV 10 folds)

Tabella risultati, Matr di confusione e Curva ROC

	Accuracy	AUC
Random Forest	0.603	0.53
Random Forest vs benchmark	0.103	0.03

	pred_value_0	pred_value_1	Total
True_value_0	32	8	40
True_value_1	17	6	23
Total	49	14	63



Variabili più correlate

$\tau \geq 0.08 \,; \tau \leq -0.08$
(Kendall)

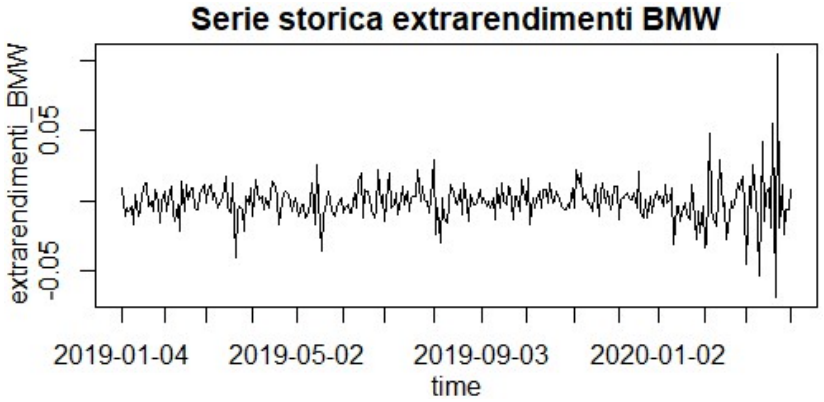
	Correlazione
sgn_lag1	0.18
rend_lag1	0.17
sgn_lag2	0.13
mean_sentiment_score_pond_lag1	0.10
sum_neutral	0.09
n_neutral	0.09
n_Tweet	0.09

Variable importance

	Mean Decrease Accuracy
rend_lag1	6.576250e-03
sum_neutral	4.402423e-03
n_Tweet	3.385141e-03
mean_positive_pond_lag1	3.205175e-03
n_neutral	2.082047e-03
sum_negative_lag1	1.655190e-03
sum_neutral_pond	1.495832e-03
n_neutral_pond	1.330577e-03

Extrarendimenti BMW

Serie



$\mu = -0.001$

$\sigma = 0.014$

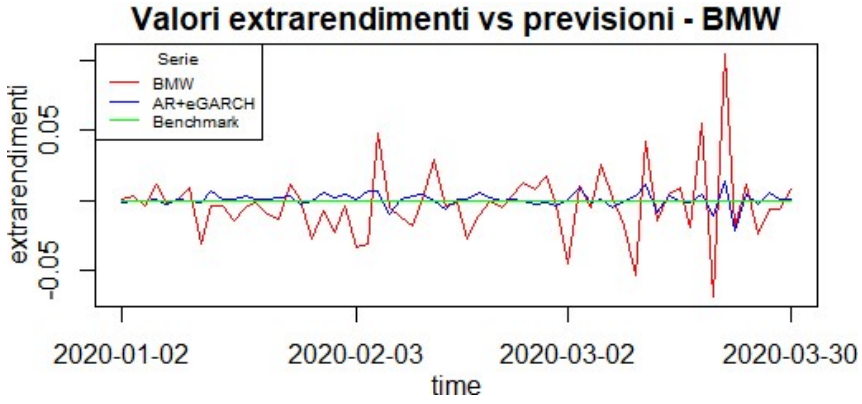
$\sigma_{\{02.20-\quad .20\}} = 0.038$

Modello Scelto

- **Modello AR(1) + eGARCH(0,1), $\xi_t \sim GED(0,1)$**

Tabella riepilogativa e grafico

	RMSE	RRMSE	R^2
AR + eGARCH	0.0235	0.93	0.12
Benchmark	0.0252	1.00	-0.01



Variabili più correlate

$\rho \geq 0.08\,;\rho \leq -0.08$
(Pearson)

	Correlazione
rend_lag2	0.16
mean_sentiment_score_pond_lag1	0.11
mean_negative_lag2	0.11
sum_sentiment_score_pond_lag1	0.09
n_negative_lag2	0.08
mean_sentiment_score_lag2	-0.09
mean_negative_pond_lag1	-0.09
rend_lag1	-0.29

Variable Selection

- Variable selection tramite **metodo backward basato sul criterio di Acaike** con k=1.

Extrarendimenti BMW

Varabili Selezionate e Output AR(1)+eGARCH(0,1)

Robust Standard Errors:

	Estimate	Std. Error	t value	Pr(>|t|)
ar1	-0.205760	0.223575	-9.2032e-01	0.357407
mxreg1	0.157933	0.078909	2.0014e+00	0.045344
mxreg2	-0.128715	0.014072	-9.1471e+00	0.000000
mxreg3	0.083991	0.186960	4.4925e-01	0.653254
mxreg4	-0.015791	0.000010	-1.5266e+03	0.000000
mxreg5	-0.008759	0.000007	-1.2484e+03	0.000000
mxreg6	0.008758	0.000007	1.2564e+03	0.000000
mxreg7	-0.008755	0.000000	-2.3036e+04	0.000000
mxreg8	-0.008751	0.000004	-2.2480e+03	0.000000
mxreg9	-0.007904	0.000007	-1.1504e+03	0.000000
mxreg10	0.007903	0.000006	1.2382e+03	0.000000
mxreg11	0.007892	0.000003	2.2677e+03	0.000000
mxreg12	0.005820	0.000003	1.9758e+03	0.000000
mxreg13	-0.005818	0.000007	-8.9028e+02	0.000000
mxreg14	0.005818	0.000002	2.7783e+03	0.000000
mxreg15	0.005817	0.000003	2.2267e+03	0.000000
mxreg16	-0.005131	0.000008	-6.5766e+02	0.000000
mxreg17	-0.005105	0.000006	-8.5645e+02	0.000000
mxreg18	-0.005085	0.000003	-1.7631e+03	0.000000
mxreg19	-0.000736	0.000000	-6.5121e+03	0.000000
mxreg20	-0.000718	0.000005	-1.4733e+02	0.000000
mxreg21	-0.000676	0.000013	-5.1854e+01	0.000000
mxreg22	-0.002799	0.000001	-2.4872e+03	0.000000
mxreg23	0.002799	0.000001	2.3350e+03	0.000000
mxreg24	-0.002799	0.000001	-2.4734e+03	0.000000
mxreg25	-0.002794	0.000002	-1.4141e+03	0.000000
mxreg26	-0.000046	0.000007	-6.6359e+00	0.000000
mxreg27	0.000027	0.000003	8.8659e+00	0.000000
mxreg28	-0.000012	0.000001	-1.5954e+01	0.000000
mxreg29	0.000008	0.000002	3.2436e+00	0.001180
omega	0.000098	0.018772	5.2370e-03	0.995821
beta1	0.900000	0.002467	3.6487e+02	0.000000
shape	2.000000	0.006736	2.9690e+02	0.000000

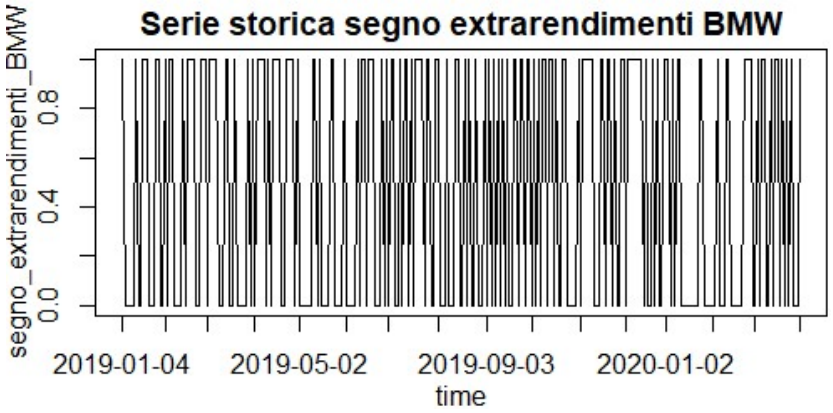
- Vengono selezionate 29 variabili, cioè:

  mean_positive_lag2, mean_negative_pond, rend_lag1, sum_positive_pond, sum_neutral_lag1, n_Tweet_lag1, sum_negative_lag1, sum_positive_lag1, sum_neutral_pond, n_tweet_totali, sum_sentiment_score_pond, n_positive_pond_lag2, n_retweet_lag2, n_neutral_pond_lag2, n_negative_pond_lag2, n_negative_lag2, n_positive_lag2, n_neutral_lag2, sum_neutral_lag2, sum_positive_lag2, sum_negative_lag2, n_positive_pond_lag1, n_tweet_totali_lag1, n_neutral_pond_lag1, n_negative_pond_lag1, sum_negative, n_negative, sum_positive_pond_lag1, sum_sentiment_score_pond_lag1

- Tutti i parametri sono significativi, ad eccezione di:
  1. Omega del modello eGARCH.
  2. Quelli associati al primo lag del rendimento (ar1 & mxreg3).
- $Shape = 2$, GED è uguale ad una normale standard.

Segno extrarendimenti BMW

Serie



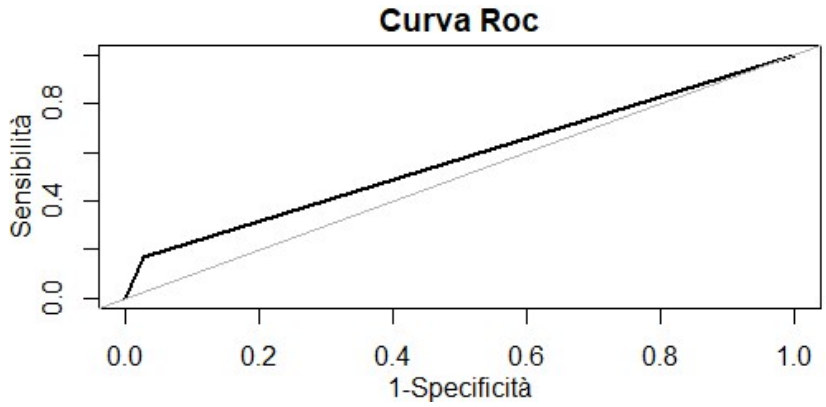
$S_+ = 144$

$S_- = 160$

Tabella risultati, Matr di confusione e Curva ROC

	Accuracy	AUC
SVM	0.656	0.57
SVM vs benchmark	0.156	0.07

	pred_value_0	pred_value_1	Total
True_value_0	36	1	37
True_value_1	20	4	24
Total	56	5	61



Variabili più correlate

$\tau \geq 0.08\,, \tau \leq -0.08$
(Kendall)

	Correlazione
n_negative	0.09
mean_neutral_pond	-0.09

Modello Scelto

- **SVM con kernel polinomiale** con i seguenti parametri:
  $\gamma = \frac{1}{79}, coef0 = 0,\ degree = 3,$
  $C = 1$(CV 10 folds)
- Le variabili sono state standardizzate.

Conclusioni

Osservando i risultati ottenuti dalle analisi precedenti possiamo giungere alle seguenti considerazioni:

1. Le informazioni estratte dai tweet collegati a Tesla sono più correlate con i risultati del titolo, rispetto a quelle relative a BMW.
2. Gli extrarendimenti presentano una correlazione con le informazioni leggermente più forte rispetto ai rendimenti puri, almeno per quanto riguarda l'asset Tesla, e riescono ad essere previsti leggermente meglio.
3. Si ottengono risultati migliori nella previsione dei segni piuttosto che dei valori dei risultati, nonostante la correlazione dei segni con le informazioni sia decisamente minore.
4. Utilizzando le informazioni ricavate dai tweet non si è riusciti in ogni caso ad ottenere dei modelli davvero interessanti per la previsione dell'andamento degli asset collegati.
5. Prendendo in considerazione l'asset Tesla vediamo che le variabili più significative sono tutte non ritardate. Queste:
   - Non possono essere utilizzate per costruire una strategia finanziaria davvero perseguibile.
   - Resta il dubbio di capire effettivamente se si sono manifestati prima i risultati degli asset oppure se si sono condivisi prima i tweet, non avendo un dettaglio temporale maggiore rispetto a quello giornaliero.