Stock Volatility Forecasting and Word Features Selection — Based On 10-K Corpus

WANG Zeyu
LIU Chang
TAN Huangao
LI Borui
ZHU Xinqi
GAO Yifeng

March 19, 2025

Contents

Introduction - Goals

Our project aims to:

- Predict the log volatility of the following year via the MD&A section of the Form 10-K;
- Find out the importance of each token and check whether they are similar over the years.

Introduction - Motivation

These two goals are meaningful because:

- The MD&A section is most related to the forecasting in Form 10-K;
- A better predicting model can lead to a better portfolio;
- Such models show the market information transmission mechanisms.

Introduction - Dataset

We use a portion of the data in 10-K Corpus[1], including:

- The tokenized MD&A sections;
  - (e.g. 1996.tok.tgz)
- The log volatility in the related year.
  - (e.g. 1996.logvol.-12.txt, 1996.logvol.+12.txt)

[1] Shimon Kogan et al. (2009). "Predicting risk from financial reports with regression". In: *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280.

Methods - Lemmatization, Stemming and Stop token filter

We use the Natural Language Toolkit (NLTK)[2] as lemmatizer, stemmer and stop token filter.

- Some words are regarded as insignificant in text analysis;
  - (e.g. a, an, the, above, across)
- Words appear in several inflected forms but with same meaning should be considered as the same;
  - (e.g. achieves, achieved, achievement, achievements $\rightarrow$ achiev)

[2] Bird et al. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Methods - Term Frequency–Inverse Document Frequency (TF-IDF)

The term frequency–inverse focument frequency (TF-IDF)[3][4] measures the importance of a word to a document in corpus via:

$$\text{(Term frequency)} \qquad \mathrm{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

$$\text{(Inverse document frequency)} \quad \mathrm{idf}(t, d) = \log_2 \left(\frac{N}{|\{d : d \in D, t \in d\}|} \right),$$

where $f_{t,d}$ is the raw count of a term t in a document d, and N is the total number of documents.

[3] Karen Sparck Jones (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1, pp. 11–21.

[4] Stephen Robertson (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5, pp. 503–520.

Methods - Nonnegative Matrix Factorization (NMF)

Given a matrix $A \in \mathbb{R}^{n \times m}$, the nonnegative matrix factorization (NMF)[5] aims to find two nonnegative matrixs $W \in \mathbb{R}^{n \times d}$, $H \in \mathbb{R}^{d \times m}$, where d is the given number of dimension, such that

$$A \approx WH.$$

The previous research[6] have shown that this method is able to learn the semantic features of text.

[5] Daniel Lee and H Sebastian Seung (2000). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13.

[6] Daniel D Lee and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *nature* 401.6755, pp. 788–791.

Methods - Naive Model (Baseline)

- The naive model (or Persistence model) uses the current (or same period) value as the forecast;
- The previous research[7] has shown that the naive model is better than many other complex models, especially for high-volatility time series;
- It doesn't involve the word features or any assumptions.

[7] Nico Beck, Jonas Dovern, and Stefanie Vogl (2025). "Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability". In: *Applied Intelligence* 55.6, p. 395.

Methods - Lasso/Ridge Regression & Decision Tree

- These models (also their extensions) are powerful in forecast and can have better performance compare with other models like heterogeneous autoregressive (HAR) models[8][9][10];
- These models are inherently interpretable, which can help us extract whether a word implies the increasing or decreasing of volatility.

[8] Chao Liang et al. (2023). "Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors?" In: *International Journal of Finance & Economics* 28.4, pp. 3689–3699.

[9] Xiafei Li, Chao Liang, and Feng Ma (2022). "Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model". In: *Annals of Operations Research*, pp. 1–40.

[10] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev (2023). "A machine learning approach to volatility forecasting". In: *Journal of Financial Econometrics* 21.5, pp. 1680–1727.

Methods - Model Evaluation

We use mean absolute error and mean squared error to evaluate the results:

$$\text{MAE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i|,$$

$$\text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2.$$

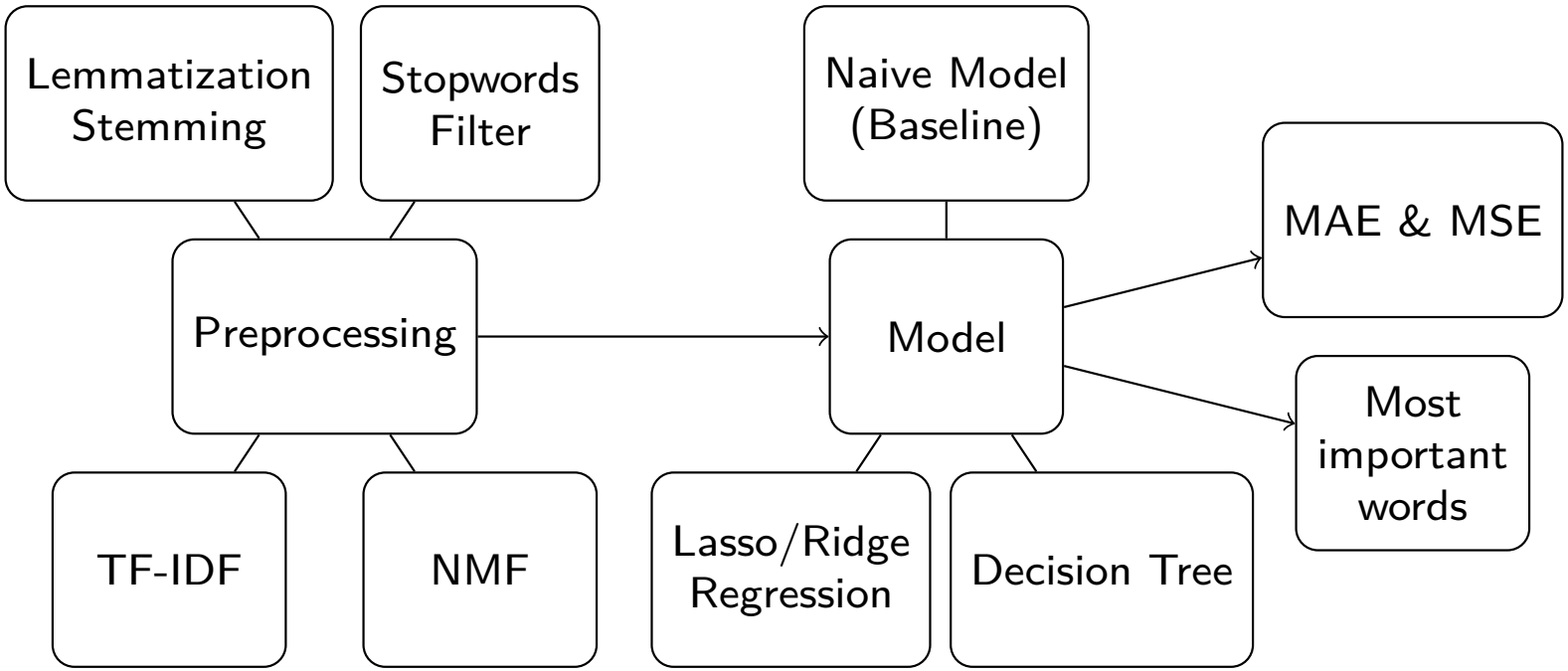
Methods - Model Evaluation

Given the model parameter $\beta \in \mathbb{R}^d$ and NMF components $H \in \mathbb{R}^{d \times m}$, we compute

$$H^+\beta = (V^T\Lambda^+ U)\beta$$

as the weight for each words via the pseudo inverse, where $H = U^T\Lambda V$ is the SVD and Λ^+ is the diagonal matrix consisting of the reciprocals of H's singular values.
We selected the top 1000 words (about $1\% \sim 5\%$ of total) and check whether they are similar between different methods and years.

Methods - Overview Flowchart



We test the model in two approaches:

- Split the data from some years into training dataset (80%) and test dataset (20%), and repeat the experiment to avoid flukes;
- Train the model via the data from some years and forecast the following year.

Result - Error

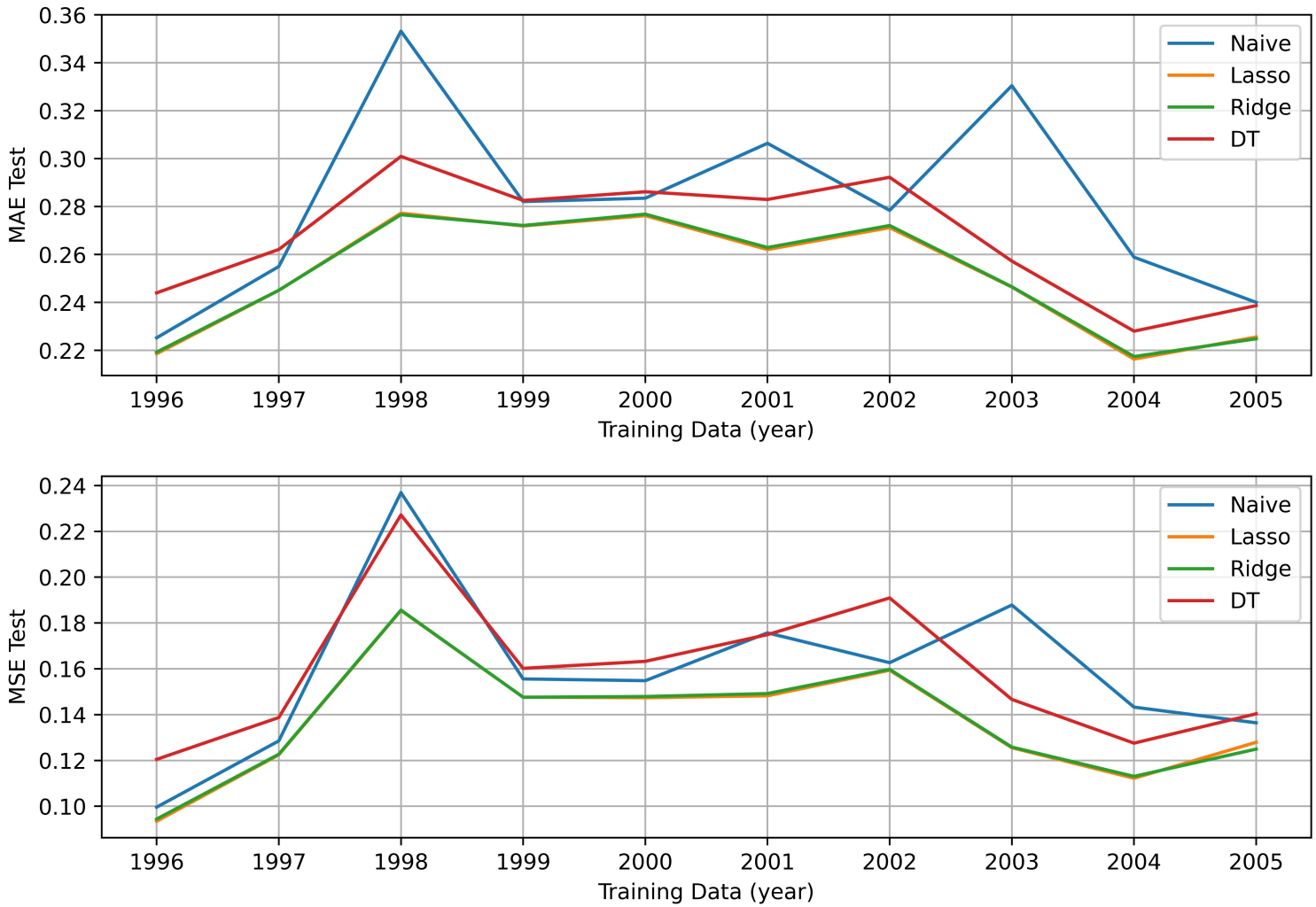


Figure: MAE and MSE for testing 1-year training set.

Result - Error

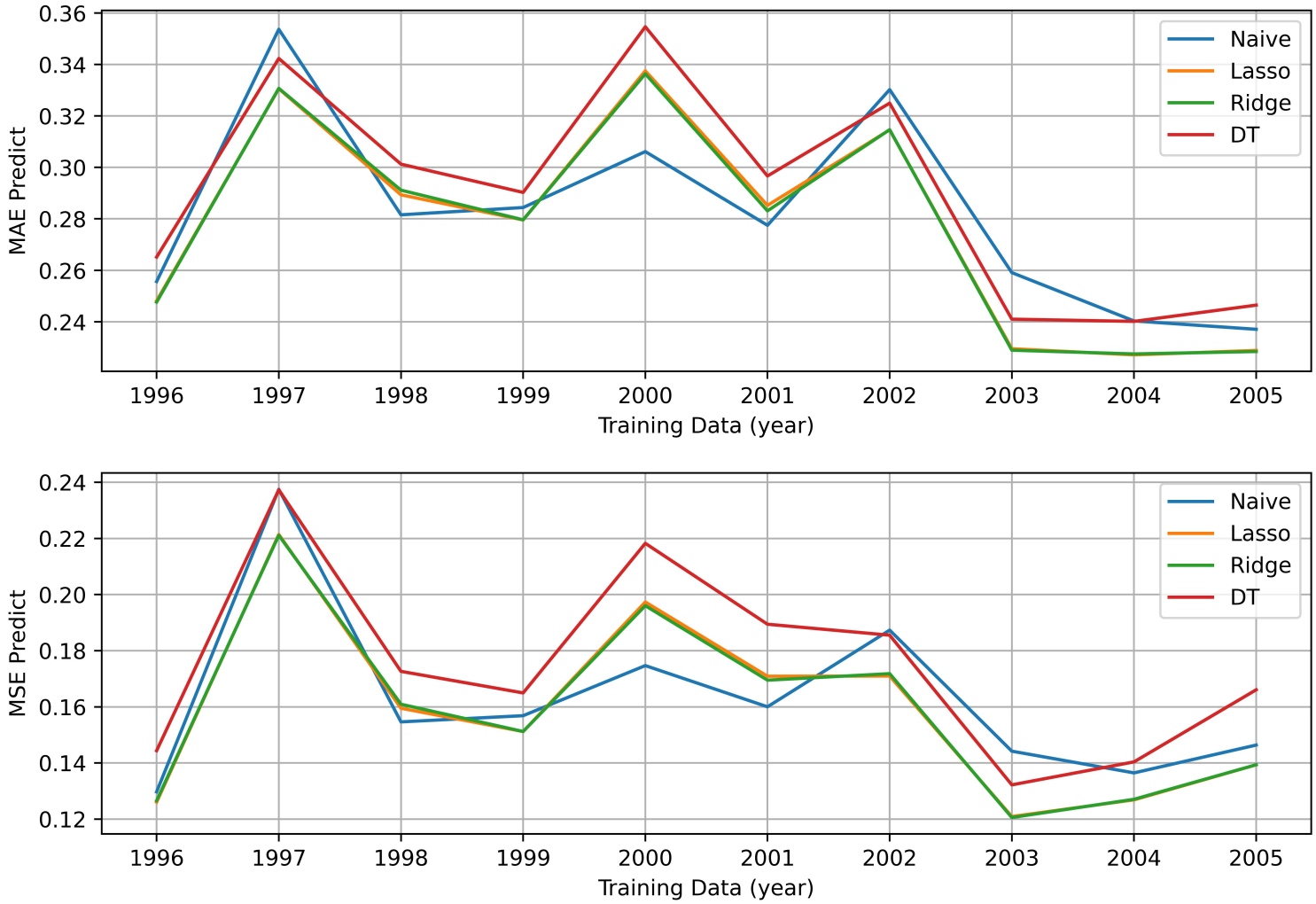


Figure: MAE and MSE for predicting 1-year training set.

Result - Error

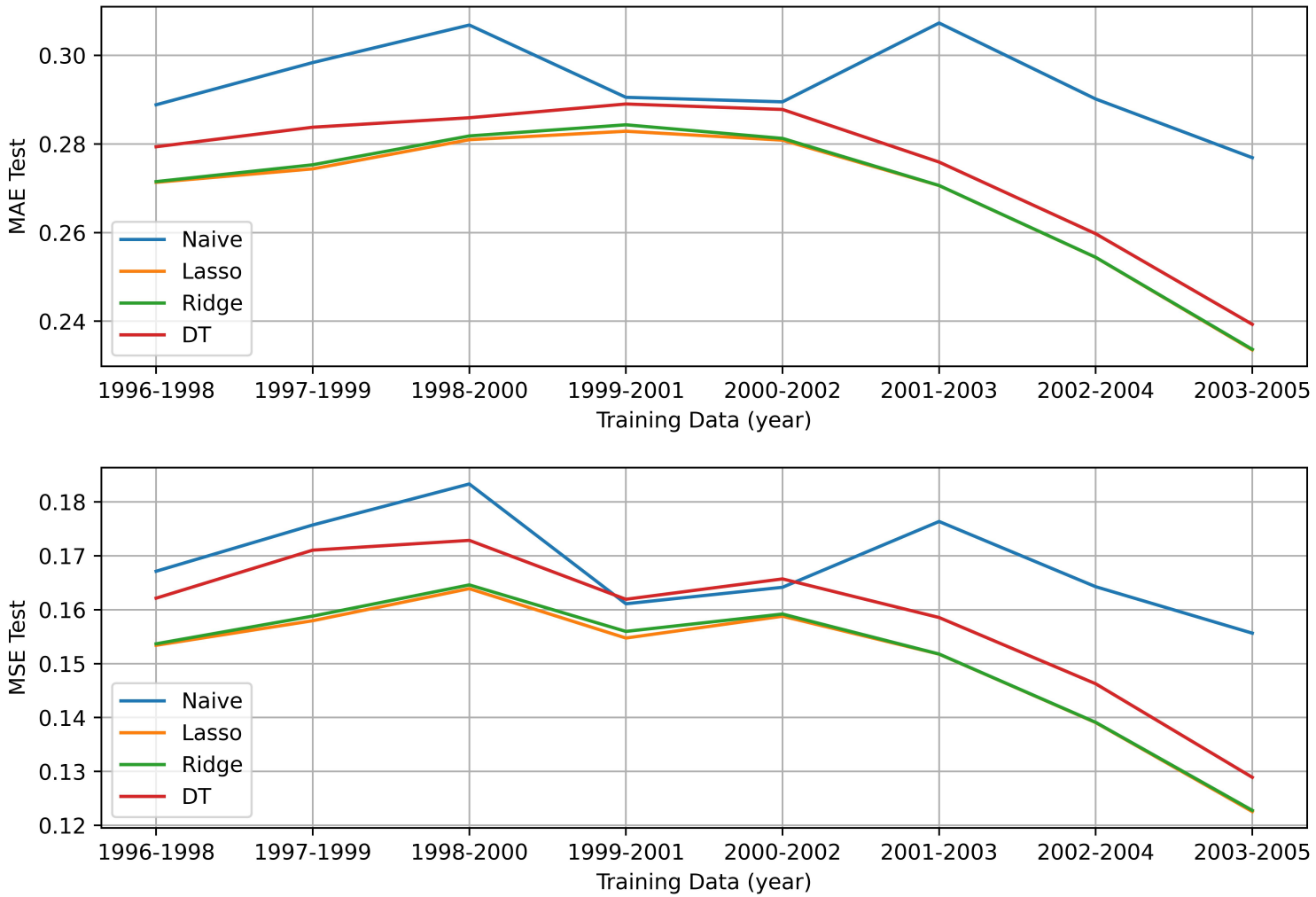


Figure: MAE and MSE for testing 3-year training set.

Result - Error

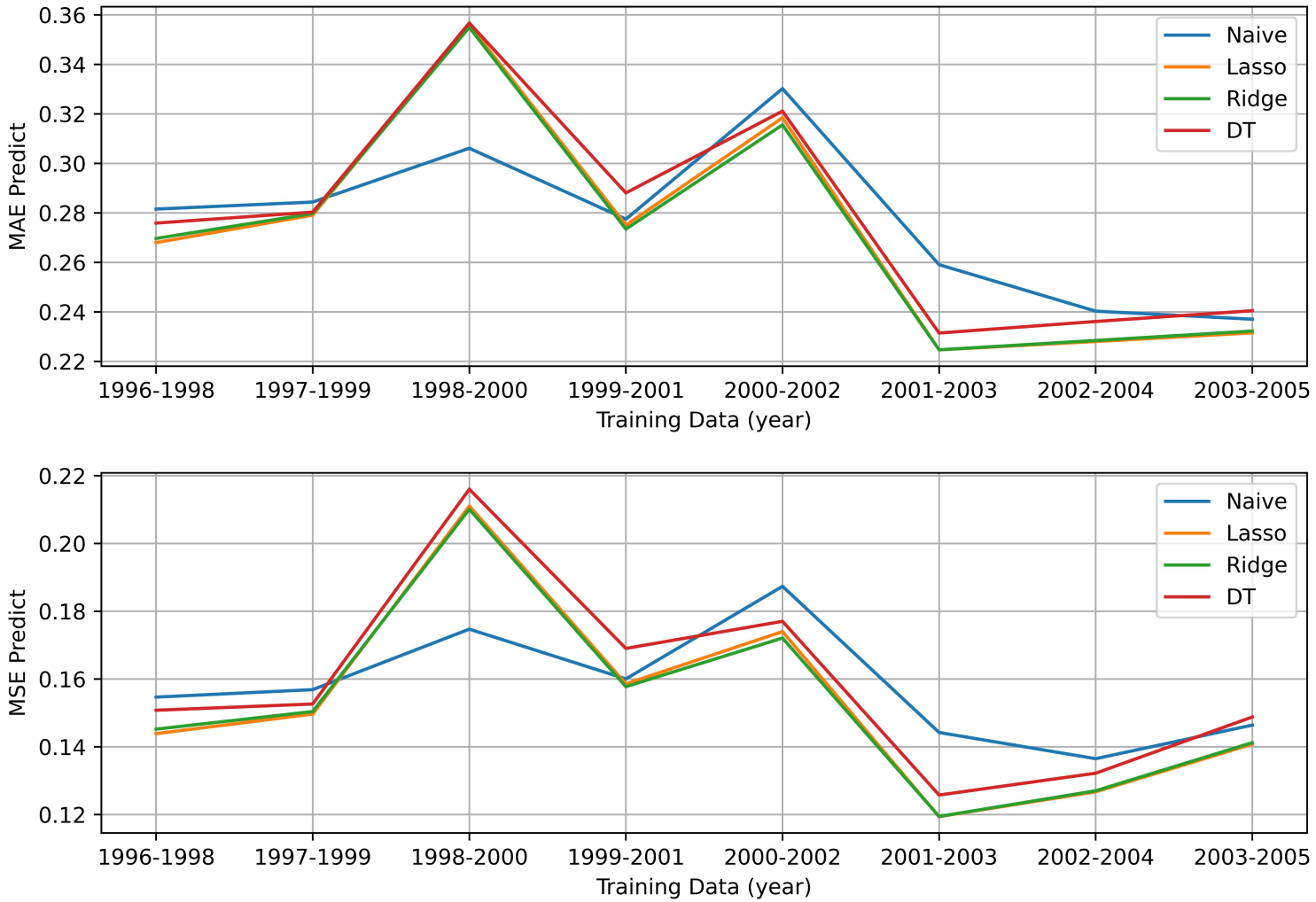


Figure: MAE and MSE for predicting 3-year training set.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	53.9%	60.6%	1.9%
1997-1998	35.1%	39.2%	6.4%
1998-1999	23.6%	31.7%	5.6%
1999-2000	3.8%	27.8%	17.1%
2000-2001	5.1%	24.9%	1.2%
2001-2002	38.0%	30.4%	2.7%
2002-2003	31.7%	31.0%	12.15%
2003-2004	26.2%	58.3%	3.15%
2004-2005	42.2%	38.3%	6.4%

Table: Word features for lasso regression between different years.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	48.3%	55.1%	2.9%
1997-1998	38.1%	36.9%	7.55%
1998-1999	20.2%	32.1%	8.35%
1999-2000	9.9%	38.6%	6.55%
2000-2001	13.2%	18.9%	3.25%
2001-2002	42.2%	35.7%	4.15%
2002-2003	35.8%	36.4%	9.2%
2003-2004	27.3%	44.6%	1.25%
2004-2005	45.5%	43.6%	9.3%

Table: Word features for ridge regression between different years.

Result - Word Features

Year	Increasing Common	Decreasing Common	Different
1996-1997	62.7%	15.8%	3.6%
1997-1998	46.2%	29.4%	6.75%
1998-1999	37.1%	18.8%	8.4%
1999-2000	29.3%	33.2%	4.4%
2000-2001	23.9%	32.7%	6.2%
2001-2002	23.7%	3.6%	10.9%
2002-2003	45.7%	5.8%	4.3%
2003-2004	34.7%	2.0%	11.85%
2004-2005	56.7%	7.3%	2.1%

Table: Word features for decision tree between different years.

Result - Word Features

Year	Lasso Ridge	Lasso DT	Ridge DT
1996	75.2%/71.6%/0.15%	23.2%/7.7%/27.0%	36.1%/25.1%/17.35%
1997	82.0%/87.2%/0.15%	15.4%/7.3%/25.85%	16.6%/12.2%/25.6%
1998	84.4%/85.5%/0.2%	14.1%/15.1%/15.4%	12.2%/15.6%/18.6%
1999	67.9%/83.9%/0.35%	17.8%/23.8%/19.8%	11.5%/23.1%/24.0%
2000	68.1%/55.3%/1.85%	19.0%/25.1%/17.6%	24.1%/9.5%/26.1%
2001	91.7%/92.9%/0.1%	5.5%/0.9%/31.55%	6.8%/1.6%/30.55%
2002	88.6%/79.1%/0.1%	37.7%/49.8%/9.75%	33.8%/39.2%/14.35%
2003	93.6%/89.5%/0.15%	13.4%/10.4%/27.05%	15.2%/13.9%/25.45%
2004	75.2%/73.4%/0.2%	23.5%/24.2%/21.25%	30.4%/20.4%/20.85%
2005	95.6%/90.8%/0.4%	9.9%/2.5%/27.5%	10.2%/5.2%/28.7%

Table: Word features between different models.

Conclusion

Some conclusions are similar to the previous research[11]:

- More training data can lead to higher accuracy and stablity;
- The Sarbanes–Oxley Act (2002) changes the information and distribution of the words;
- The Sarbanes–Oxley Act (2002) makes a higher accuracy in prediction.

[11] Shimon Kogan et al. (2009). "Predicting risk from financial reports with regression". In: *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280.

Conclusion

For the words features:

- There exists the similarity between different years;
- The lasso and ridge regression always give the same features, while the decision tree chooses the different (because the decision tree is more likely to split the data based on the previous volatility);
- The abnormal during 1999 to 2001 may caused by the dot-com bubble;
- Compare with the features of increasing, the of features decreasing are more similar between years.

Reference I

- Beck, Nico, Jonas Dovern, and Stefanie Vogl (2025). "Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability". In: *Applied Intelligence* 55.6, p. 395.
- Bird et al. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- Christensen, Kim, Mathias Siggaard, and Bezirgen Veliyev (2023). "A machine learning approach to volatility forecasting". In: *Journal of Financial Econometrics* 21.5, pp. 1680–1727.
- Kogan, Shimon et al. (2009). "Predicting risk from financial reports with regression". In: *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280.
- Lee, Daniel and H Sebastian Seung (2000). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13.

Reference II

- Lee, Daniel D and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *nature* 401.6755, pp. 788–791.
- Li, Xiafei, Chao Liang, and Feng Ma (2022). "Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model". In: *Annals of Operations Research*, pp. 1–40.
- Liang, Chao et al. (2023). "Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors?" In: *International Journal of Finance & Economics* 28.4, pp. 3689–3699.
- Robertson, Stephen (2004). "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5, pp. 503–520.

Reference III

 Sparck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1, pp. 11–21.