Modeling the Economic Impact of News Coverage on Influential Individuals

2025.12.04

7조

2020170837 최원준
2021320328 장지웅
2022320157 노규주

COSE 362
Machine Learning

COSE 362
Machine Learning

Contents

01 Motivation

02 Method

03 Results

04 Future works

7조

2020170837 최원준
2021320328 장지웅
2022320157 노규주

Motivation

Why Financial Markets are Hard to Predict?

- 주식·환율 등 경제지표는 예측이 매우 어려움
- 시장은 다양한 요인들이 동시에 작용하며 비선형적으로 반응함

숫자 지표 이상의 통찰이 필요함

Motivating cases

Ronaldo's Coca Cola gesture followed by $4bn drop in company's market value

Tesla goes bankrupt in Elon Musk April Fools' Day gag

Elon Musk pokes fun at Tesla bankruptcy speculation with a series of tweets and a comical photo of himself passed out against a Model 3.

Bitcoin soars above $36,000 after Elon Musk changes his Twitter bio

Popular tech entrepreneur and Tesla CEO Elon Musk wrote simply '#bitcoin' in his biography on the social media site with no further explanation

Agencies

Updated • 29 Jan 2021, 04:46 PM IST

Initial Approach: The Limits of SNS Data

BBC

Register Sign In

Tesla will no longer accept Bitcoin over climate concerns, says Musk

13 May 2021

Share Save

Elon Musk said cryptocurrency could "not come at great cost to the environment"

Tesla has suspended vehicle purchases using Bitcoin due to climate change concerns, its CEO Elon Musk said in a tweet.

Bitcoin fell by more than 10% after the tweet, while Tesla shares also dipped.

Tesla's announcement in March that it would accept the cryptocurrency was met with an outcry from some environmentalists and investors.

The electric carmaker had in February revealed it had bought $1.5bn (£1bn) of the world's biggest digital currency.

CNBC LIVESTREAM SIGN IN

POLITICS

Trump goes after Amazon over taxes

PUBLISHED THU, MAR 29 2018•8:04 AM EDT UPDATED THU, MAR 29 2018•4:01 PM EDT

Michael Sheetz
@IN/MICHAELJSHEETZ

WATCH LIVE

KE

- T in a tweet, saying the online retailer pays "little governments."
- H after a report said he was "obsessed" with Amazon. The company's shares tanked on that report.
- Several states say Amazon and oth collect sales tax, even in states whe physical presence.
- Amazon does collect sales tax on it District of Columbia.

The Alternative: News Media Instead

🚫 데이터 접근 불가 (크롤링 제한)

⚠️ 높은 표본 편향 (SNS 미사용 인물)

📉 많은 노이즈

The Alternative: News Media Instead

🚫 데이터 접근 불가 (크롤링 제한)

⚠️ 높은 표본 편향 (SNS 미사용 인물)

📉 많은 노이즈

✅ 안정적인 데이터 접근성

📊 정제되고 검증된 정보

📈 풍부한 시장 맥락 포함

뉴스 기사는 시장 예측에 있어 일관성 있고 신뢰도 높은 지표

Problem Definition

- 수치 데이터만으로는 급격한 시장 변동을 온전히 설명하기 어려움.
- 시장은 **영향력 있는 인물들의 발언**에 매우 민감하게 반응함.

Hypothesis

뉴스 텍스트의 의미를 벡터화하여 경제 지표와 결합하면,

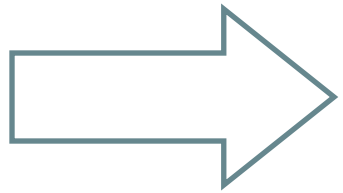
S&P 500의 변동성을 더 정확하게 예측할 수 있을 것이다.

Method

Data Collection → Preprocessing → Feature Engineering → Prediction

Data Collection → Preprocessing → Feature Engineering → Prediction

❑ Input target selection
- 2017~2019년 **TIME 100** (Most Influential People) 선정 인물
- 전체 288명 → Article 수 기준 **top100** 선정
  - Total ***460722*** articles



- Donald Trump
- Ryan Murphy
- Jeff Sessions
- Ryan Reynolds
- Prince Harry
- Kim Jong Un
- ...

❑ Output
- **S&P500**
- 2017~2019년

❑ Time period & Data split
- 2017/01/01 ~ 2019/12/31
- Dataset splitting (4:1:1)

Image generated with Nanobanana 3

Data Collection → **Preprocessing** → Feature Engineering → Prediction

❑ Objective

- 뉴스 기사의 의미(semantic context)를 예측 모델에 반영

❑ **Text Embedding** 활용하여 article text → vector

- Challenge: ***Context Window Limit***

Unit	Embedding model
Headlines	`BAAI/bge-large-en-v1.5`
Chunking+pooling	`BAAI/bge-large-en-v1.5`
Full body text	`jinaai/jina-embeddings-v3`
First + Last paragraphs	`jinaai/jina-embeddings-v3`

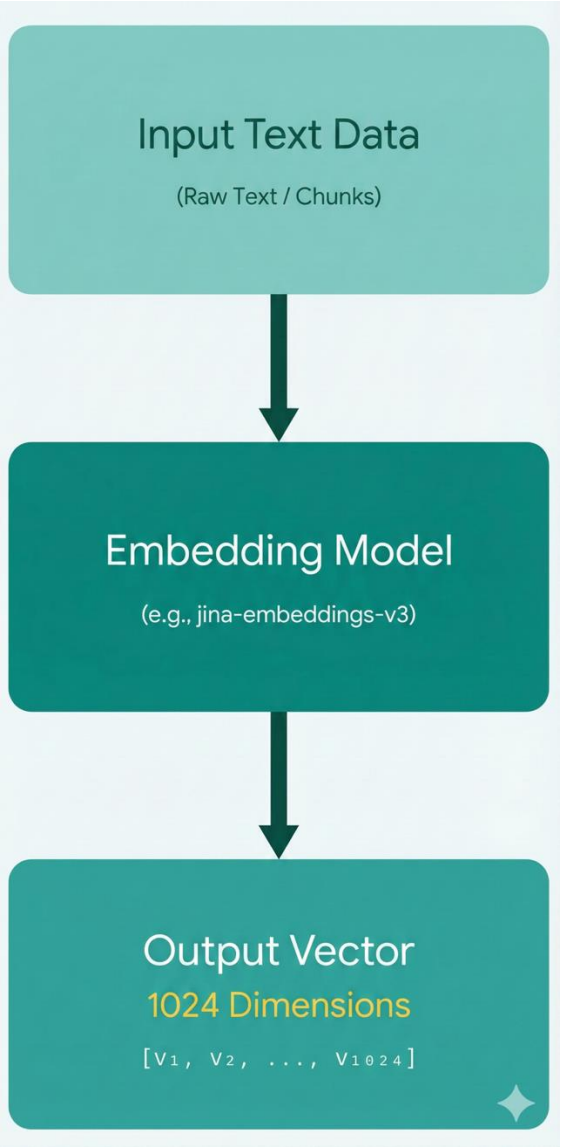


Image generated with Nanobanana 3

Data Collection → Preprocessing → **Feature Engineering** → Prediction

❑ Feature combinations

- A (Baseline): **Economic Indicators** Only
- B: A + **Text Embeddings** (Vector)
- C: B + **Person Metadata** (One-hot)
- D: C + **Fear & Greed Index**

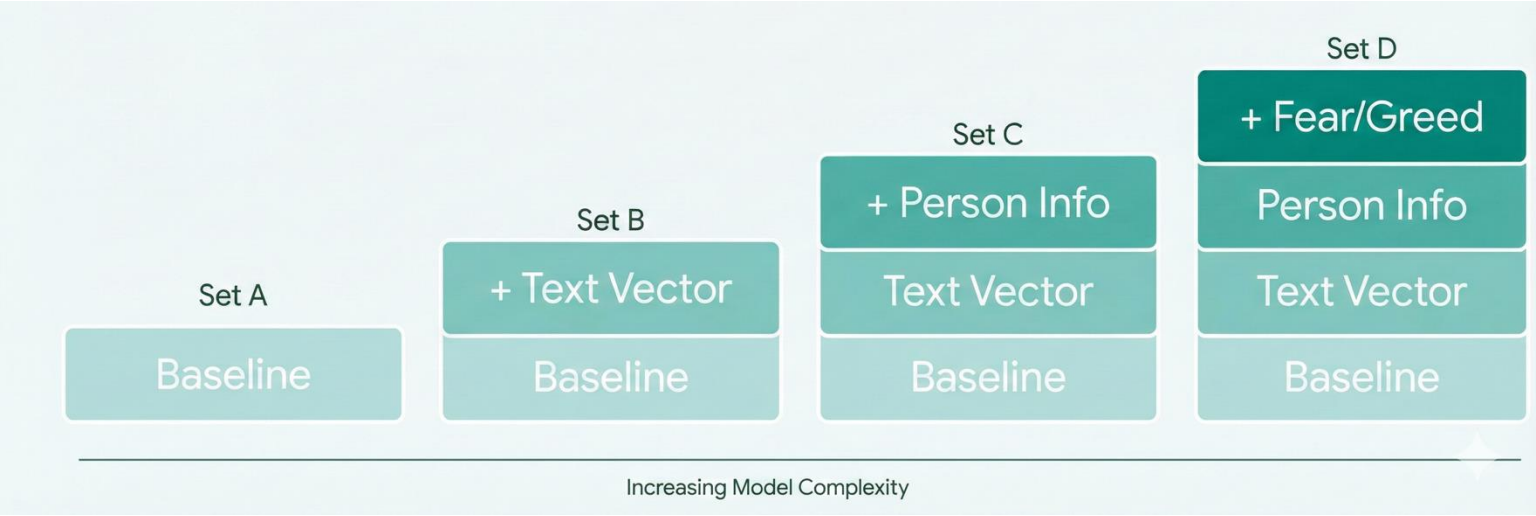
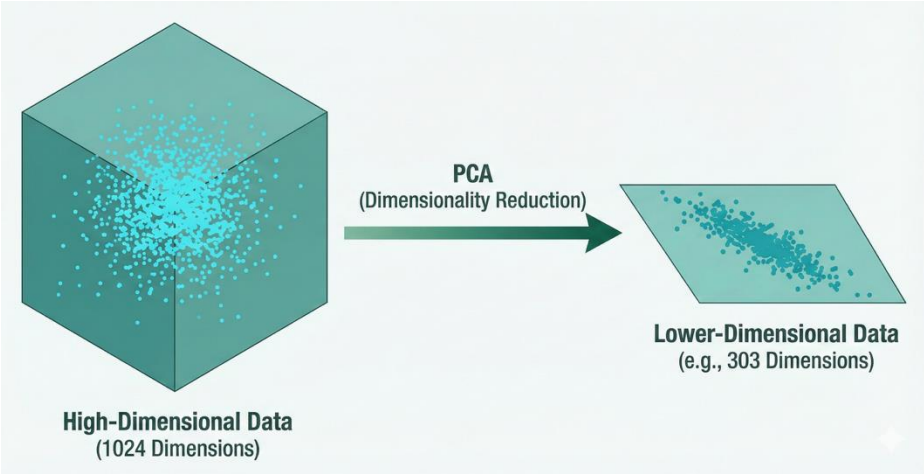


Image generated with Nanobanana 3

Data Collection → Preprocessing → **Feature Engineering** → Prediction

❑ Dimensionality reduction

- Challenge: 고차원 임베딩 (1024 dims) → ***Overfitting Risk***
- Solution: **PCA**
  - pov = 0.9
- 차원축소 결과
  - 1024 → **303** (headline)
  - 1024 → **190** (bodyText)
  - 1024 → **207** (paragraphs)
  - 1024 → **243** (chunking)



❑ set A (Baseline): **Economic Indicators** Only **1**

❑ **Feature combination**

- set B
  - **Text Embeddings**
- set C
  - **Person Metadata**
- set D
  - **Fear & Greed Index**

❑ **Embeddings**

- Headlines
- Full text
- Chunking
- Paragraphs

❑ **PCA**

- Original
- Reduced

3 X 4 X 2 = 24

Data Collection → Preprocessing → Feature Engineering → **Prediction**

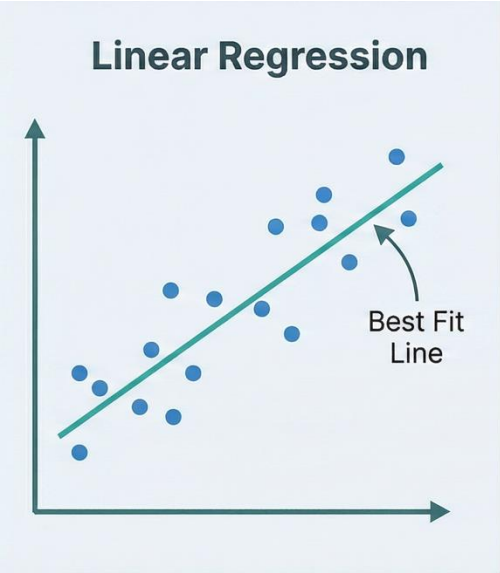
❑ Model Selection

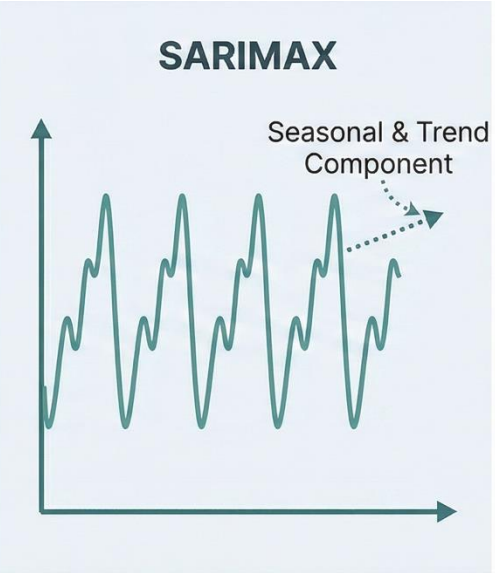
- Linear Regression
- SARIMAX
- LightGBM
- GRU

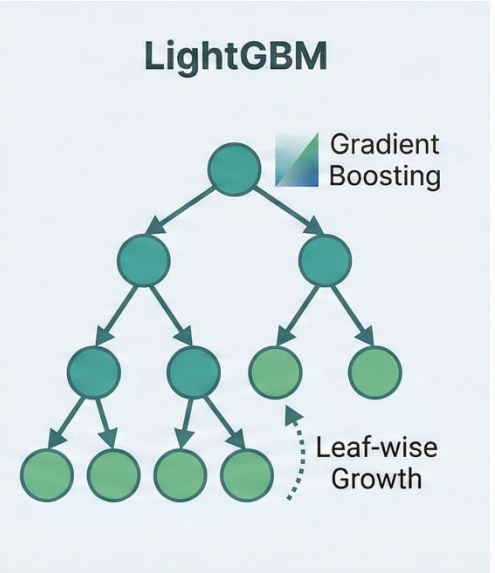
Debiasing

Instance Weighting : $Loss_{new} = Loss \times \frac{1}{N_{articles}}$

Mean Pooling : $X_t = average(v_1, ..., v_n)$







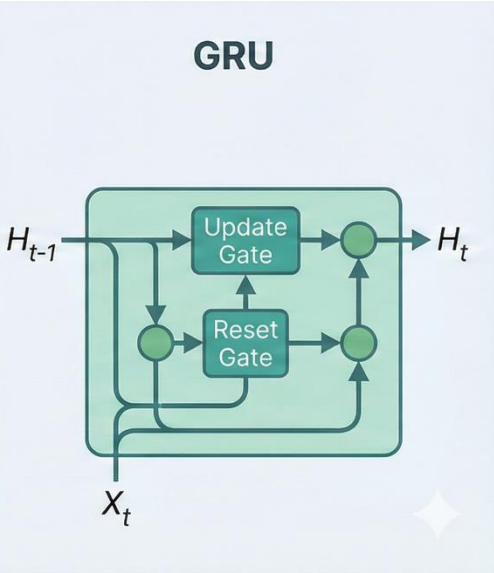


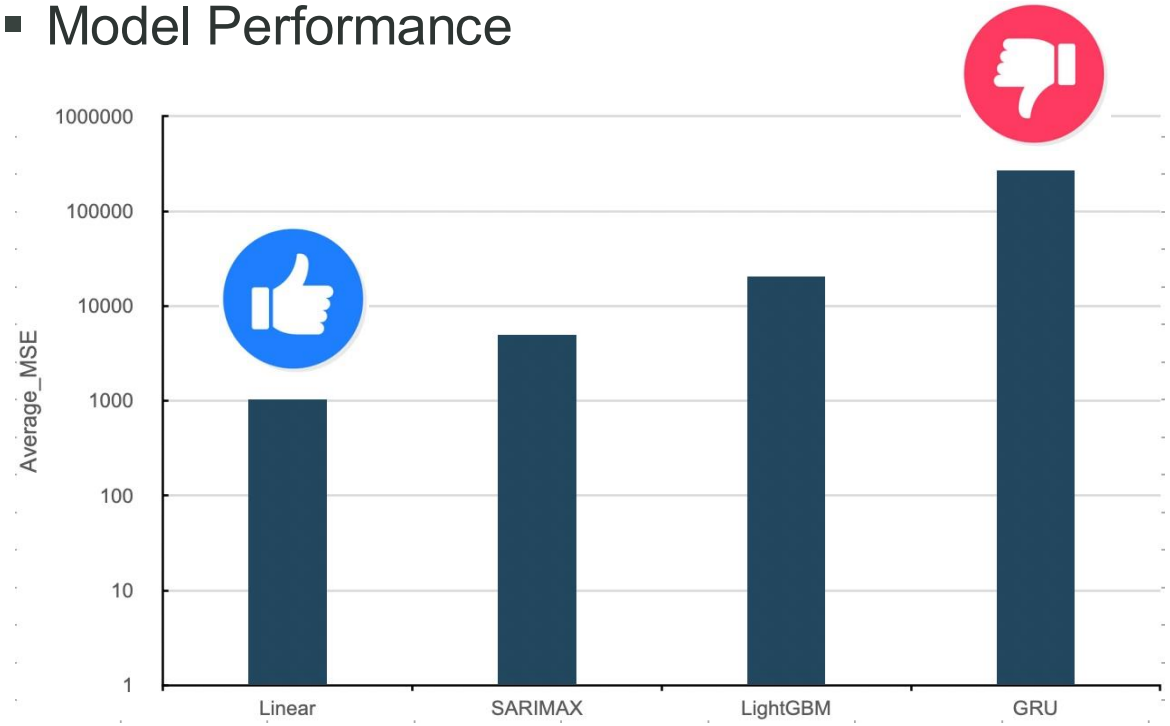
Image generated with Nanobanana 3

Results

Results

❑ Results and Analysis

▪ Model Performance

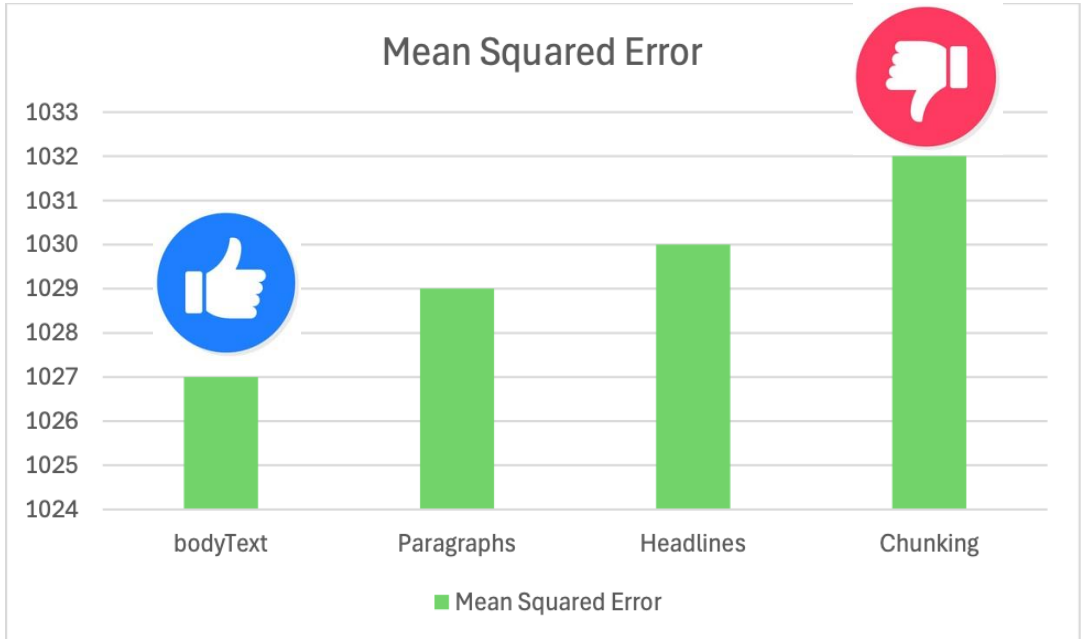


- 금융 데이터 : 시장 참여자의 감정, 루머, 무작위적 거래 등으로 인한 많은 noise + 적은 수의 Sample
- **Linear Regression : 주요 선형관계만을 포착해서 좋은 성능**
- **복잡한 model(LightGBM, GRU) : overfitting**

Results

❑ Results and Analysis

- Embedding Strategy

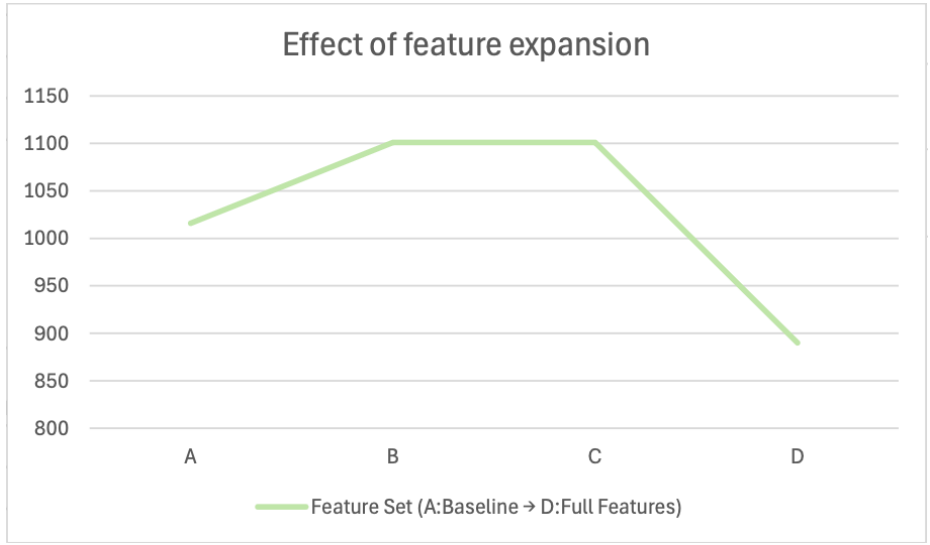


- **BodyText 임베딩이 제일 좋은 성능. 뉴스의 전체적인 흐름을 모델에 전달하는 것이 중요**
- 중요한 내용을 앞뒤에 배치하는 뉴스의 구조를 공략한 Paragraphs 방식도 뛰어난 성능을 보임

Results

❑ Results and Analysis

▪ Feature Effectiveness

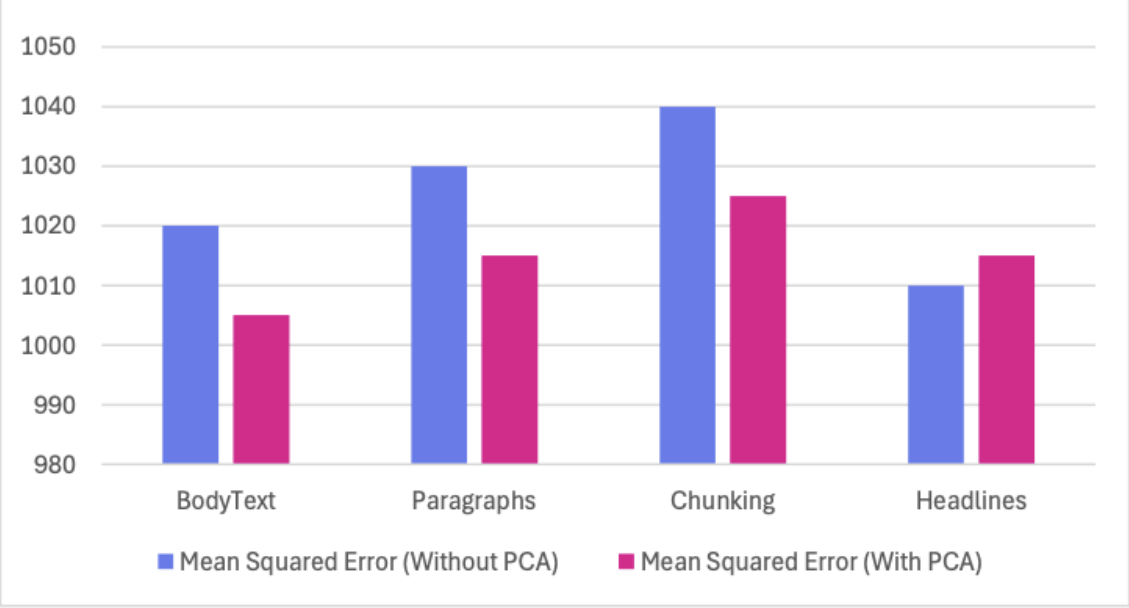


- Baseline(A)에서 텍스트(B)와 인물 정보(C) 추가했을 때, 평균 MSE 1101로 증가
- Fear & Greed(D) 추가하니, 평균 MSE 889를 기록하고 최고 성능 MSE 881까지 달성
- **Dataset D MSE Baseline 대비 13% 감소**
- **뉴스 기사가 적절한 시장 지표와 결합된다면 강력한 예측 시그널로 사용가능**

Results

❑ Results and Analysis

▪ PCA Impact



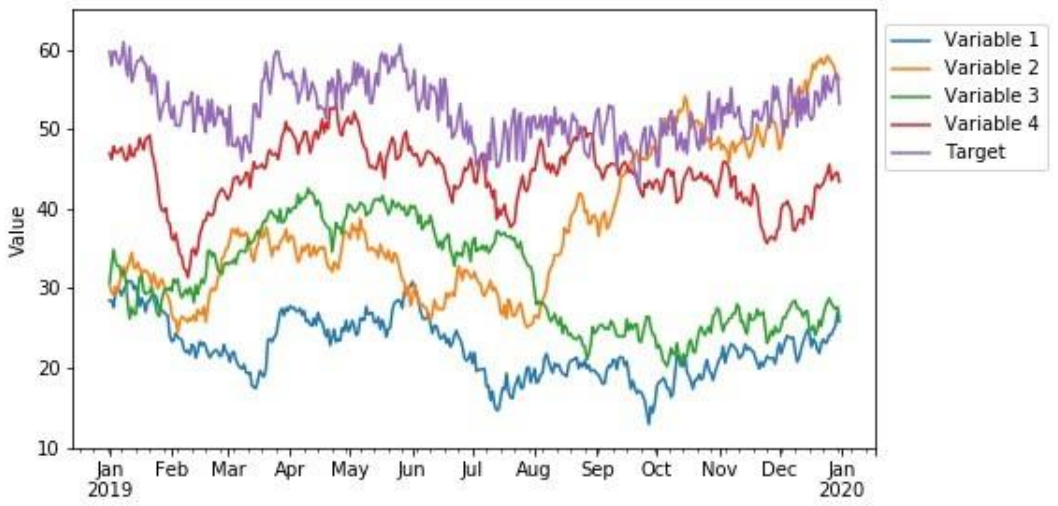
- BodyText, Paragraphs, Chunking의 경우, PCA 적용하면 MSE 감소
- 반면, Headlines의 경우, MSE 증가
- **PCA가 정보의 밀도에 따라서 noise를 제거하는 기능을 수행**

Future works

Improvement Strategy

❑ Model

- Advanced **time-series architectures**
  - Temporal Fusion transformer
  - N-Linear
  - BERT / RoBERTa
- 시퀀스의 맥락을 더 잘 파악



❑ Feature

- 데이터의 노이즈를 줄이기 위해 **거시 경제 지표(VIX, 금리 등)**를 통합
- LLM을 활용한 article 중요 정보 요약

❑ Data

- 더 정밀한 타이밍 포착을 위해 **고빈도 데이터(시간 또는 분 단위)**로 데이터 단위 세분화

Image generated with Nanobanana 3

Toward Real-time Market Intelligence

❑ Dynamic online streaming Inference

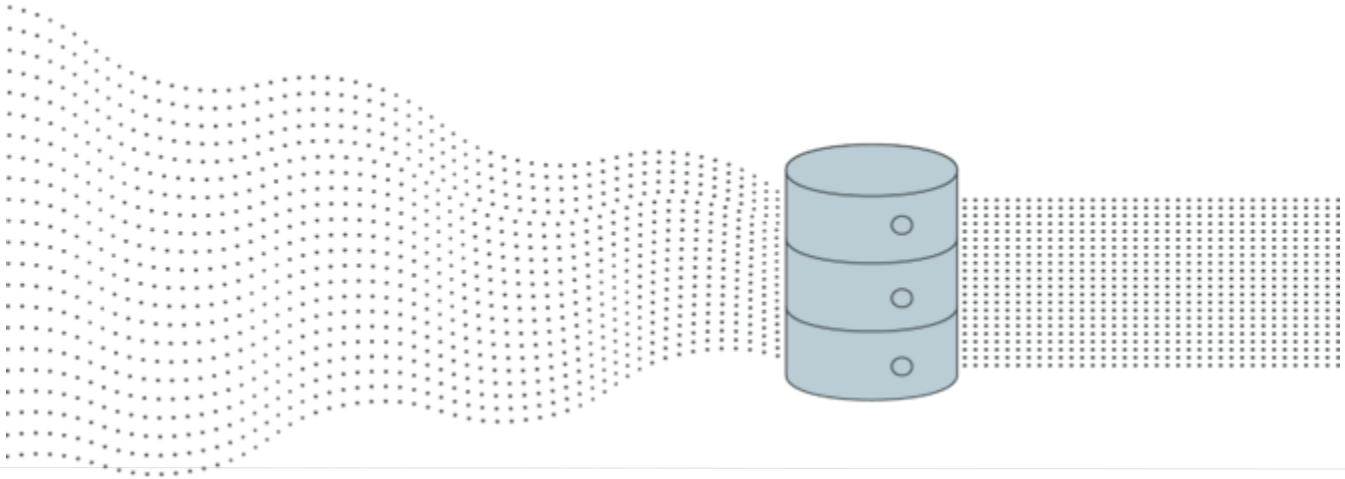
- 실시간 비정형 데이터들을 반영해

  단기 시장 변동성을 예측하는 ***Real-time forecasting model*** 구축

❑ Domain-agnostic Generalization

- 특정 시장에 국한되지 않고

  다양한 경제 시장에도 확장 가능한 ***domain-agnostic forecasting framework***



Q&A