Stock Behaviour Prediction based on the news

Xiangwei Peng
Xiaokang Wang

Introduction

Motivation:

- Stock price is influenced by both market and other factors

Research Question:

- Can we also using news information to forecast stock price?

Subtasks and KPIs:

- Forecast the stock returns to minimize the prediction errors (MSE)
- Capture the daily stock movement to maximize the accuracy

Model Framework:

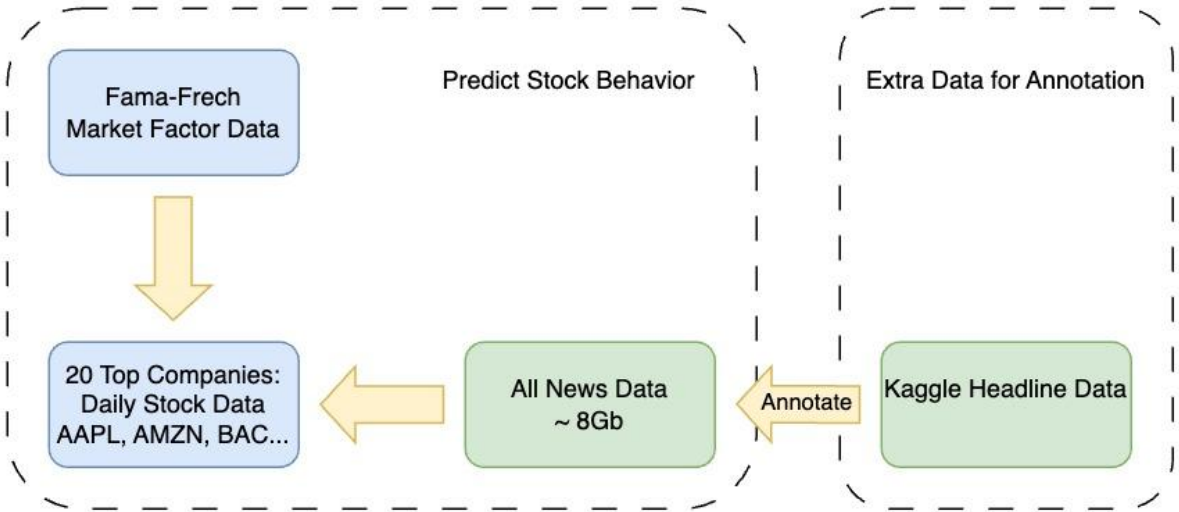
- Additive structure: market factor model and news topic model

$$r(t) = \hat{r}(t, \text{Market}) + f(t, \text{News}) + \epsilon$$

Dataset

Time Period: 2016 ~ 2019

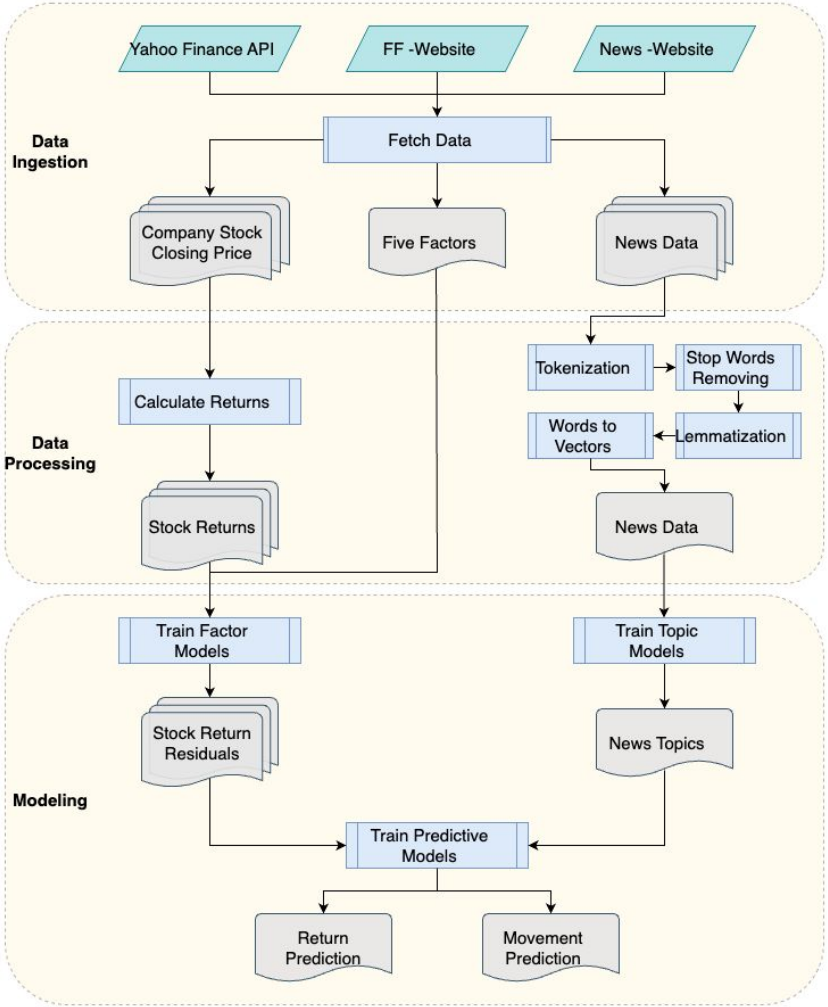
- Training period: 2016 ~ 2018 (3 years)
- Testing period: 2019 (1 year)



Pipeline

Build an automatic pipeline (engine) to:

- Data ingestion
  - stock price
  - market factors
  - news data
- Data process
  - stock closing price to returns
  - news text data to numeric vectors
- Train models
  - classifier to annotate news data
  - topic model to extract news topics
  - factor model and predictive model to forecast stock behaviors



News Data Process and Topic Model

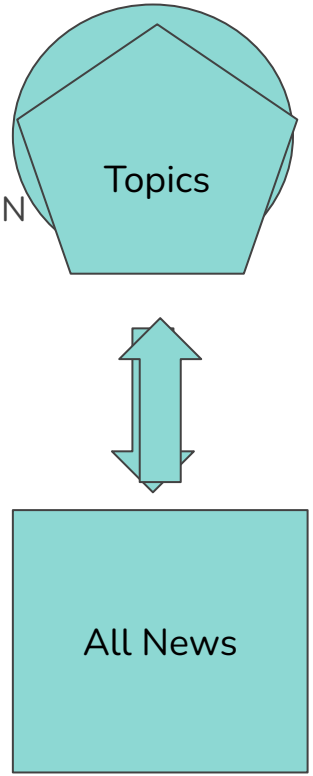
Classification model:

- Classify news category based on headline
- Ensemble soft voting models of Logistic, Random Forest, XGBoost, CNN
- Using this to label the all news data set

Topic model:

- Clustering news based on the similar semantic themes (topics)
- Hierarchical Dirichlet Process (HDP) to inference 500 topics
- Normalized daily topic counts as features

Keywords of the Topic 74:
[Theme: Social Chaos] crackdown, uprising, troop ...
[Theme: Tension Feeling] anxietry, upset, misunderstanding ...
[Theme: Conflict Issue] 911, bombshell, restraining, toxic ...



Factor Model and Predictive Model

Factor Model

- Linear regression on market factors

$$\hat{r}(t) = \beta_0 + \beta_1 MER_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t$$

Regularized Predictive Models:

- Predictors: daily news topic counts
- Candidate models:
  - Linear model: Ridge, Lasso
  - Nonlinear model: Random Forest, XGBoost
- XGBoost outperforms

Results

Stock Ticker: Mobil Corp (XOM) as an example

Choose two baseline models:

- The first day's (Jan 1, 2019) price
- tomorrow's price as today's price

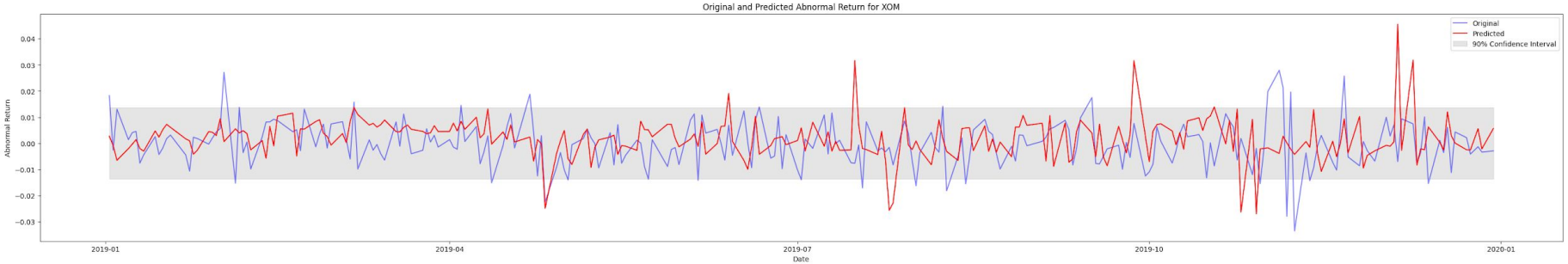
Stock Return Prediction

Model	Baseline 1	Baseline 2	Factor model	Predictive model
MSE	1.976	2.096	1.243	1.241

Results

Stock Movement Prediction

- Test Accuracy 68.9%



Discussion

Conclusion:

- Predicting stock return is relative difficult
- The high accuracy suggests we can use the news information to predict stock price movement

Future work:

- Data limitation: stock related news for more detailed information
- Sentiment Analysis: model the influence direction of the news
- Dynamic Assumption: continuous update model for the online stream data
- Trading Strategy: leverage forecast power to seek capital gain
- ...

Thank you!