FINANCIAL DATA FORECASTER

Shubhankar Agrawal  3035345306
Karan Mahajan  3035346025
BACKGROUND AND MOTIVATION

The growing importance of technology in the world of finance has created several opportunities for its exploitation to create value from the international markets. One such area of computing, machine learning has created a significant impact by creating potential to forecast financial markets.

BACKGROUND

The study aims to glean insights into whether traditional machine learning models could successfully forecast financial time series data. It also aims to discover whether there is a significant relation between markets contingent upon sentiment and market indices that may be of interest.

MOTIVATION
This project aimed to collect financial and social media data over the past 10 years for its purpose.

A comprehensive modelling approach with detailed experiments to select the optimal modelling scenarios.

Evaluation of the forecasted values against the market to verify if the predictions are financially viable.

These were some of the key objectives we aimed to complete as part of building this financial data forecaster for the project.
LITERATURE REVIEW

Having read through several previous papers on similar topics, we decided to follow a cross-domain approach using forex and market indices to forecast the prices of the time series.
Financial Data was collected using the InvestPy API on Python. This enabled to query Forex and Market index prices over a historical period of 10 years from 2010-2020, which was then stored in a spreadsheet. The data collected in the OHLCV format was stored in a CSV to ease further processing steps as part of the project.

Social Media data was collected from Twitter and Reddit. Both these social media sites were queried for the countries selected, their capitals and all their heads of state during the query period. Scripts were written for this purpose and the data was stored as JSON objects in a data repository on the server.
To canonicalize textual data, it was processed through BERT. BERT analyses each textual input and predicts whether the input is positive or negative.

It was hypothesized that market sentiment can be derived by analysing human sentiment. Therefore, the results were aggregated over the time period on a daily basis and statistical features like mean, count, variance, etc were derived to act as input along with the textual data.

Gathering Market Sentiment
EXPLORATORY DATA ANALYSIS

**AUGMENTED DICKEY FULLER TEST**

The ADF test helped identify that the raw values weren’t stationary which led to the requirement for the returns calculated for the prices.

**SEASONALITY**

Seasonal decomposition helped identify the seasonal component in the data to use in classical time series models.

**CORRELATION**

Correlation metrics were obtained between target variables and lagged features to identify cross domain pairs that would help obtain better predictions.
DATA PREPROCESSING

Feature Engineering was performed simultaneously with the Exploratory Data Analysis to generate features as well as transform the existing data to be better suited for our machine learning models.

**STATIONARY DATA**
The returns of the prices were calculated to be used as feature and target values in the approach to ensure stationarity.

**ADDITIONAL FEATURES**
Additional features such as intraday returns, inter-day close/open difference and moving averages were generated.

**LAGGED VALUES**
The target value was lagged and added to the features to act as an autoregressive variable.
EXPERIMENTATION
A company is an association or collection of individuals, whether natural persons, legal persons, or a mixture of both. Company members share a common purpose and unite in order to focus.

The first approach was to use regression models to forecast the true values/returns of the data using the provided features. This did not lead to good results.

Finally, a binary classification approach was adopted to predict whether positive or negative returns would be seen for the forecasted period.

To get better results, a bins classification approach was adopted with the returns split into bins followed by a multi-class classification.
REGRESSION ANALYSIS
Machine Learning Methods

RAW VALUES
The raw values led to unfavourable accuracy measurements as the returns over a day were of a very small range

RETURNS
The returns were a more favourable variable to predict, although not high accuracy.

R2 Score < 0.10
Seasonal ARIMA was used to regress the raw values and the returns into the future. Diagnostics show that the returns are more favourable to the ARIMA modelling as they more suited to the normal distribution.

ARIMA works better than Prophet with the raw values as it applies a rolling moving average throughout the test period whereas Prophet applies a precalculated trend from the training data on the testing period.

Facebook’s Prophet model was tested with the data. The seasonal components were automatically picked by the Prophet model. The prophet model did not provide any favourable results for predictions too.

Prophet provided better results when compared to ARIMA when considering the returns. This can be attributed to the automatic selection of seasonality inferred by the Prophet model.
The returns were split into bins (brackets) based on the quantiles of the data. Following this, a multiclass classification was applied to this data. Low precision and recall scores for the class 1 bring down the overall performance of this model. Investigation showed that the margin for more misclassified bins were narrower.

To aim for better results, the binary classification was applied on the returns by splitting them into 2 bins on the fact that whether they were positive or negative. This approach had a better performance compared to the previous model owing to the narrowed scope of prediction (binary scale).
FEATURES
Lagged Returns of the raw features were to be used as the features of the prediction models.

TRANSFORMATIONS
Using the returns, feature scaling methods were rendered void, thus none were used further.

TARGET
Returns of the forex and index market prices on a binary scale were the values to be predicted.

CROSS-VALIDATION
Cross-validation did not show any improvements in performance in experiments thus it was not used.
Principal components analysis was used to reduce dimensionality of the feature space and to make it more interpretable for models to run. The PCA results attributed most of the explained variance to less than half of the features. This helped us reduce the number of features, as PCA provided better results with the model.
Sentiment scores did not have any significant influence on the accuracy of the models. Moreover, there was no correlation between the sentiment scores and the values of the returns.

A closer analysis revealed that the sudden peaks and falls in the market indices showed only a later follow-up in sentiment scores, if not no relation, in accordance with the efficient markets hypothesis.
The experimentation phase led to the selection of some forex and market indices, along with some models to be taken forward in the optimization process.

**INDICES SELECTED**

2 forex indices and 2 market indices were selected after having experimented over all of the available pairs.

**CROSS-DOMAIN MARKETS**

The 2 most strongly correlated APAC markets for each of these indices were identified to be used as exogenous features.

**MODELS SELECTED**

The 2 models with the best performance from the binary classification phase were selected.

### Forex Index

- BDT (Bangladeshi Taka)
- MNT (Mongolian Togrog)
- VND (Vietnamese Dong)
- IDX Composite (Indonesian Stock Exchange)
- LKR (Sri Lankan Rupee)
- NZX MidCap (New Zealand Exchange)

### Market Index

- Karachi 100 (Karachi Stock Exchange)
- CSE All-Share (Colombo Stock Exchange)
- INR (Indian Rupee)
- Nikkei 225 (Tokyo Stock Exchange)
- IDR (Indonesian Rupiah)
- MNE Top 20 (Mongolia Stock Exchange)

- RidgeClassifier
- SVC
OPTIMIZATION
Although a powerful tool, the amount of customization required for this project rendered it inadequate to effectively compare performance across hyperparameters.

Built using techniques like multithreading, the aim here was to compare performance across thousands of combinations of hyperparameters, for both regular approach and walk-forward approach.

GridSearchCV  Our Routine
## HYPERPARAMETER SELECTION

<table>
<thead>
<tr>
<th></th>
<th>Ridge</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularization Parameter</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>Tolerance</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>Cross Market Feature</td>
<td>VND</td>
<td>IDX Composite</td>
</tr>
<tr>
<td>Accuracy</td>
<td>61.68</td>
<td>61.1</td>
</tr>
</tbody>
</table>

Ridge Classifier BDT Optimization

- **VND**
- **IDX Composite**
WALK FORWARD APPROACH

STEP 1
Take data for time period 0 to n and predict n+1 and record the prediction.

STEP 2
Slide window by one step. Train model again on data from time period 1 to n+1 and predict for day n+2. Repeat this process for the entire data set.

RESULTS
Significantly improves predictive performance of the algorithm for forex. However, the same is not reflected in Index predictions.
ENSEMBLE VOTING APPROACH

Take a majority vote of best performing models. An ensemble voting model reduces the randomness of a model and increases variance. We can see that this approach did increase performance albeit marginally.
COMING TOGETHER

Build Model -> Walk Forward -> Hyperparameter Optimization -> Ensemble Voting

Repeat
Major Conclusion

The optimization process highlighted several shortcomings of our initial approach. Initial approach built models using default parameters and just using the hyperparameter optimization technique improved performance by 5%. Walk forward approach showed an 10% gain in accuracy for forex.

It showed us that an, ensemble voting performed the best across the board. Walk forward approach was seen as the best way to predict forex and a traditional 80-20 split was found meaningful for Indices.
Instead of computing moving averages over market value, the strategy developed computes a moving average of the predictions.

Since the aim of backtesting is to analyse the quality of predictions, it was hypothesised that if the general trend predicted by the model was in line with the actual market trend, the strategy would perform well and the quality of predictions would be deemed high.
Intraday Trading Strategy

Analysing the quality of predictions and whether it can home in on major downfalls and shocks and help save losses in equity. The trading strategy was based on the strategies of “short selling” and “long buying”.

Predictions for the testing period

Look at prediction of time t+1

Pred [t+1] == 1?

True

Buy @ t+1 Open
Sell @ t+1 Close

False

Sell @ t+1 Open
Buy @ t+1 Close
RESULTS
With aim to continue this project beyond the scope of this course, here are some areas the team wanted to continue development on, with respect to this financial data forecaster.

The current economic climate owing to the pandemic acts completely different to the normal economic periods. Testing our model in these uncertain times for the economy would help understand better model performance. It would also be impactful to utilize this current time period as data to train the model as well.

The walk forward models currently developed can be analysed on out of box data by adding support for online learning and prediction. The team believes that there is potential to generate relatively accurate predictions which could be used in trading. Therefore, we aim to package this application better for financial use.
THANK YOU

Bibliography


