Foundation Project: Final Review

Group 13

Apurva: 12110078 | Gaurav Anand: 12110118 | Manjari Singh: 12110104 | Radha: 12110009 | Ruchi: 12110085

Agenda

Project Overview

Our Approach: CRISP-ML(Q)

- *Business & Data Understanding*
- *Data Preparation*
- *Modelling*
- *Evaluation*
- *Deployment*
- *Monitoring & Maintenance*

Project Overview

Stock Market sentiment analysis model using news articles

In-Scope:

- *Sentiment based analysis*
- *CRISP-ML(Q) methodology*

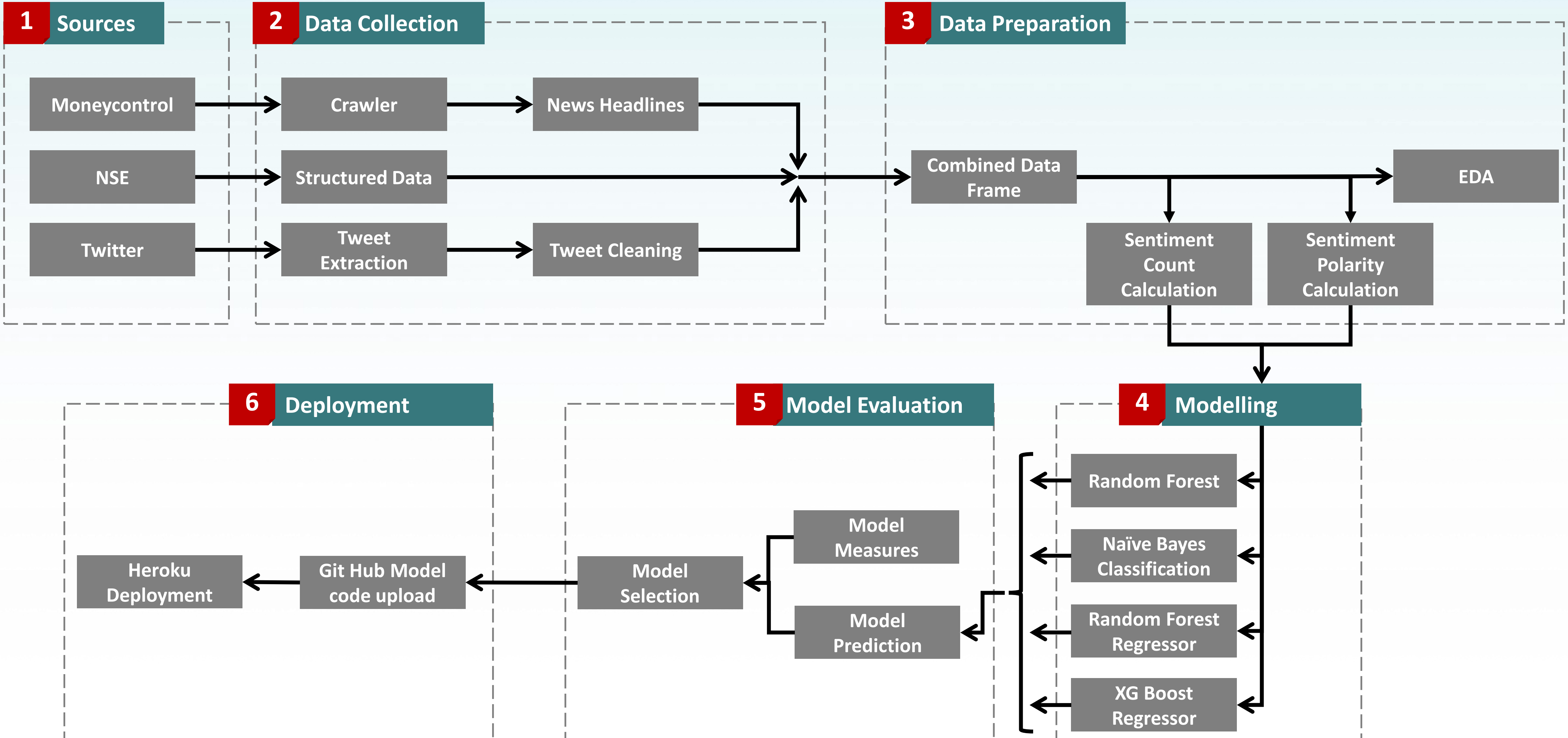
Out of Scope:

- *Fundamental Analysis*
- *Technical Analysis*

Architecture

Access our PROJECT on GIT :
https://github.com/AMPBA-2022S/FP_Group13/tree/master

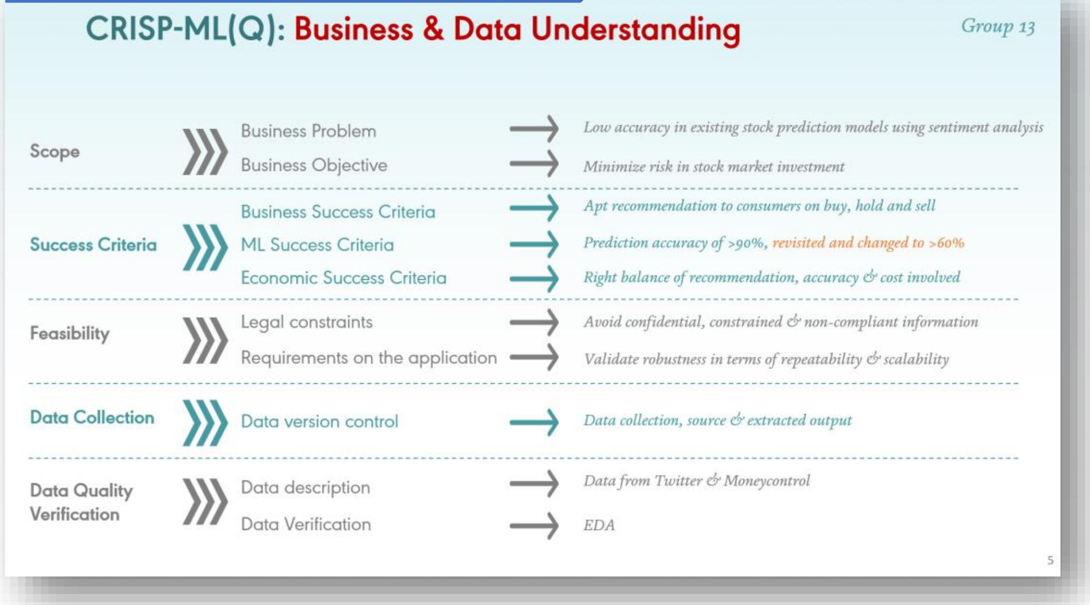
Group 13



CRISP-ML(Q): Business & Data Understanding

Group 13

In mid-review, we committed for..



CRISP-ML(Q): Business & Data Understanding

Group 13

Scope	Business Problem	*Low accuracy in existing stock prediction models using sentiment analysis*
	Business Objective	*Minimize risk in stock market investment*
Success Criteria	Business Success Criteria	*Apt recommendation to consumers on buy, hold and sell*
	ML Success Criteria	*Prediction accuracy of >90%, revisited and changed to >60%*
	Economic Success Criteria	*Right balance of recommendation, accuracy & cost involved*
Feasibility	Legal constraints	*Avoid confidential, constrained & non-compliant information*
	Requirements on the application	*Validate robustness in terms of repeatability & scalability*
Data Collection	Data version control	*Data collection, source & extracted output*
Data Quality Verification	Data description	*Data from Twitter & Moneycontrol*
	Data Verification	*EDA*

5

Our Final delivery..

Data Collection:

- ***MoneyControl:*** *Fetching news via Python web scraping*
- ***Twitter:*** *Capture Tweets with term "SCRIPTNAME" for past 6M using sntwitter.*
- ***NSE:*** *Daily real time prices of "SCRIPTNAME" for the past 6 months*

Version Control

- *Maintaining & refreshing data every week to keep 6M data versions in control*

Data Description

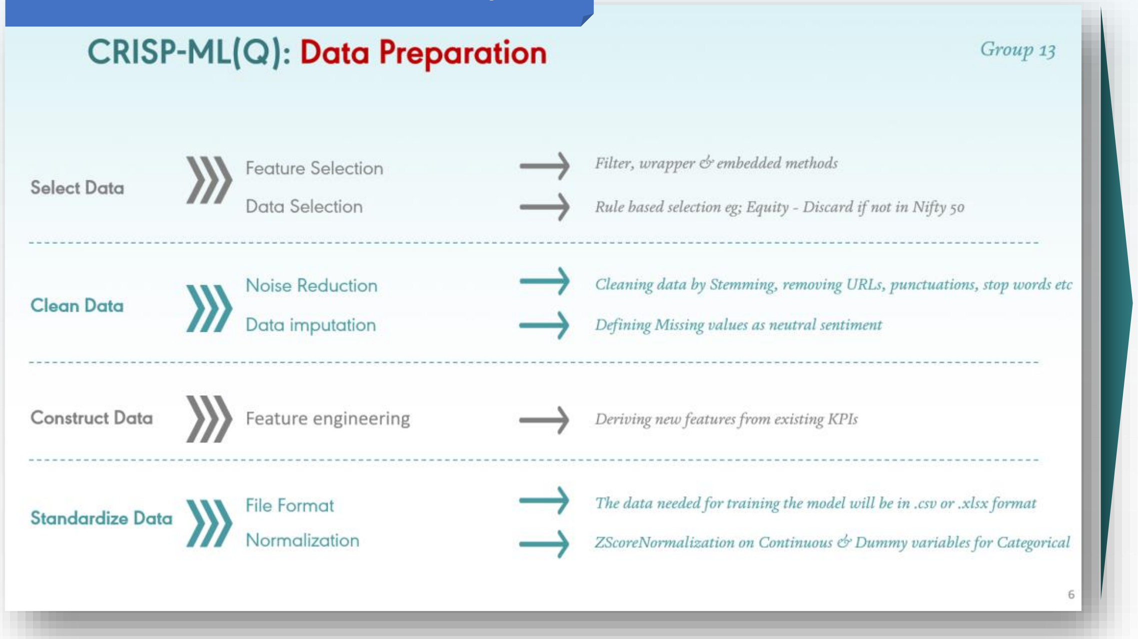
- *MoneyControl - Date of the headline(Date), News Headline for DRREDDY(Text)*
- *Twitter - TweetDate(Date), Tweet content(Text)*
- *NSE - Date for the particular stock(Date), Daily average price(Amount), Highs & Lows of the daily price(Amount)*

Data Verification

- *Sample verification/ exploration done on combined data to keep the requirements in check*

CRISP-ML(Q): Data Preparation

In mid-review, we committed for..



CRISP-ML(Q): Data Preparation

Select Data	Feature Selection	→ *Filter, wrapper & embedded methods*
	Data Selection	→ *Rule based selection eg; Equity - Discard if not in Nifty 50*
Clean Data	Noise Reduction	→ *Cleaning data by Stemming, removing URLs, punctuations, stop words etc*
	Data imputation	→ *Defining Missing values as neutral sentiment*
Construct Data	Feature engineering	→ *Deriving new features from existing KPIs*
Standardize Data	File Format	→ *The data needed for training the model will be in .csv or .xlsx format*
	Normalization	→ *ZScoreNormalization on Continuous & Dummy variables for Categorical*

Our Final delivery..

Feature Selection

- *Features selected using Filter method. For the model, we only need Date of article, text from headlines/tweets for a period of 6M and we created polarity*

Data Selection

- *Textual data dropped from selection, as sentiments are captured for headlines*

Noise Reduction

- *Cleaning tweets & headlines for any unwanted punctuation & removing Hyperlinks. DateTime field cleansed to 'Date' field & unnecessary fields dropped*

Data Imputation

- *Renaming Columns to match indexes while merging the collected data. Replacing NaN values in Polarity field as 'ZERO'.*

Feature Engineering

- *Post sentimental analysis: No. of positive words, No. of negative words, No. of Neutral words. Determining the intensity of the word using SentimentIntensityAnalyzer.*

Data Augmentation

- *Polarity: (No. of positive sentiments - No. of negative sentiments)/(Positive +Negative +Neutral) Count: sum of No of news article, tweets collected for a particular date*

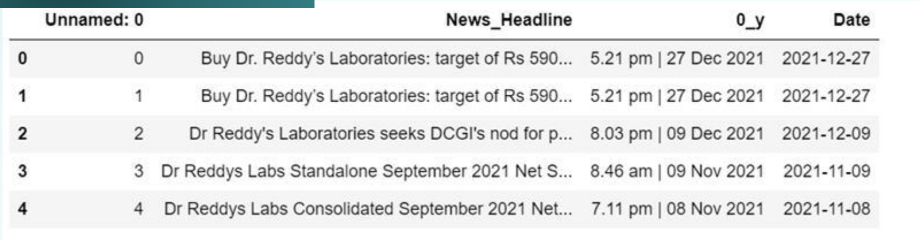
Normalization

- *Polarity is normalized with a value range of -1 to 1. Dummy variables for categorical data has been converted during modeling and as & when required.*

Data Collection & EDA

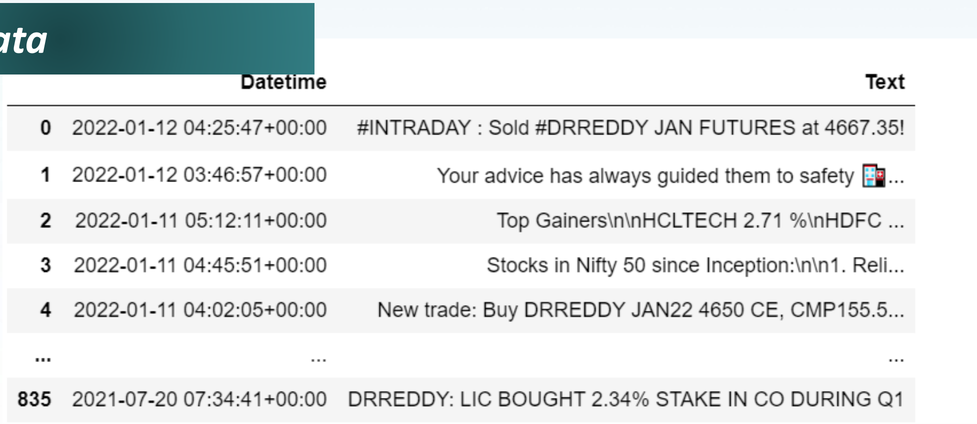
Group 13

Moneycontrol Data



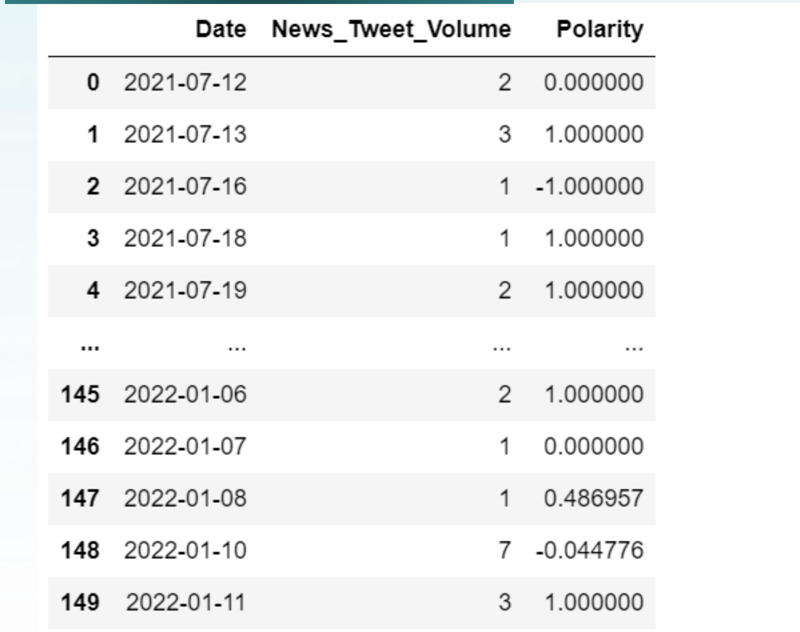
	Unnamed: 0	News_Headline	0_y	Date
0	0	Buy Dr. Reddy's Laboratories: target of Rs 590...	5.21 pm | 27 Dec 2021	2021-12-27
1	1	Buy Dr. Reddy's Laboratories: target of Rs 590...	5.21 pm | 27 Dec 2021	2021-12-27
2	2	Dr Reddy's Laboratories seeks DCGI's nod for p...	8.03 pm | 09 Dec 2021	2021-12-09
3	3	Dr Reddys Labs Standalone September 2021 Net S...	8.46 am | 09 Nov 2021	2021-11-09
4	4	Dr Reddys Labs Consolidated September 2021 Net...	7.11 pm | 08 Nov 2021	2021-11-08

Twitter Data



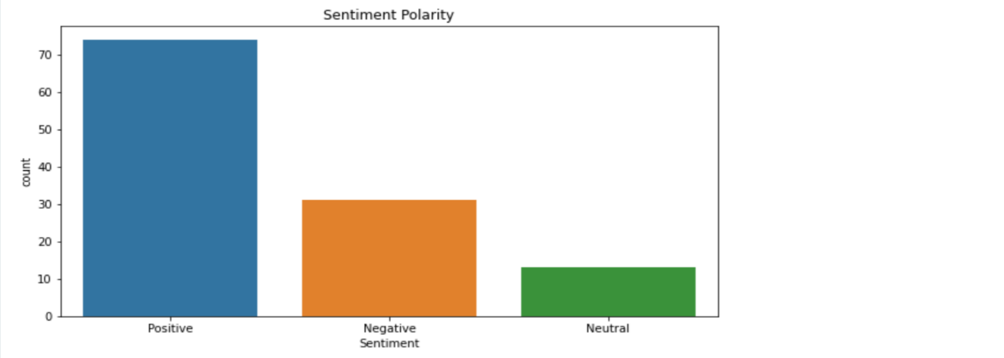
	Datetime	Text
0	2022-01-12 04:25:47+00:00	#INTRADAY : Sold #DRREDDY JAN FUTURES at 4667.35!
1	2022-01-12 03:46:57+00:00	Your advice has always guided them to safety 🏥...
2	2022-01-11 05:12:11+00:00	Top Gainers\n\nHCLTECH 2.71 %\nHDFC ...
3	2022-01-11 04:45:51+00:00	Stocks in Nifty 50 since Inception:\n\n1. Reli...
4	2022-01-11 04:02:05+00:00	New trade: Buy DRREDDY JAN22 4650 CE, CMP155.5...
...	...	...
835	2021-07-20 07:34:41+00:00	DRREDDY: LIC BOUGHT 2.34% STAKE IN CO DURING Q1

Final Output

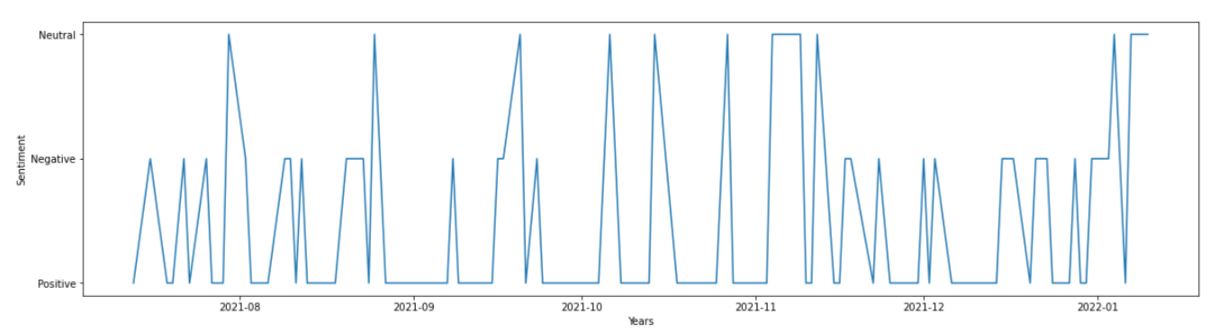


	Date	News_Tweet_Volume	Polarity
0	2021-07-12	2	0.000000
1	2021-07-13	3	1.000000
2	2021-07-16	1	-1.000000
3	2021-07-18	1	1.000000
4	2021-07-19	2	1.000000
...	...	...	...
145	2022-01-06	2	1.000000
146	2022-01-07	1	0.000000
147	2022-01-08	1	0.486957
148	2022-01-10	7	-0.044776
149	2022-01-11	3	1.000000

E
D
A



The graph represents that more tweets with positive sentiment have been twitted which reflects overall trend in stock prices



Sentiment tracing across past 6 months

CRISP-ML(Q): Modelling

Literature Research		*Review existing models available online*
Define quality measures of model	Performance	*Measure of % predictions directionally correct vs. incorrect i.e. Accuracy*
	Robustness	*Variations of results depending on changing market sentiments*
	Scalability	*Ability to scale and accommodate increased workload, volume etc*
	Explainability	*Analytical or visual output of Prediction should be explainable to user*
	Model Complexity	*Model should be able to comprehend complex & additional relationships*
Model Selection		*Model selection basis above defined "Quality Measures"*
Incorporate Domain Knowledge		*EDA basis domain specific features + sentiment analysis*
Model Training		*4 Step process; Split, Transform, Fit & Accuracy*
Assure reproducibility	Result Reproducibility	*K-fold cross validation + random sampling of training data*
	Experimental Documentation	*Maintain document to capture iteration*

MODEL SELECTION

Model assessment:

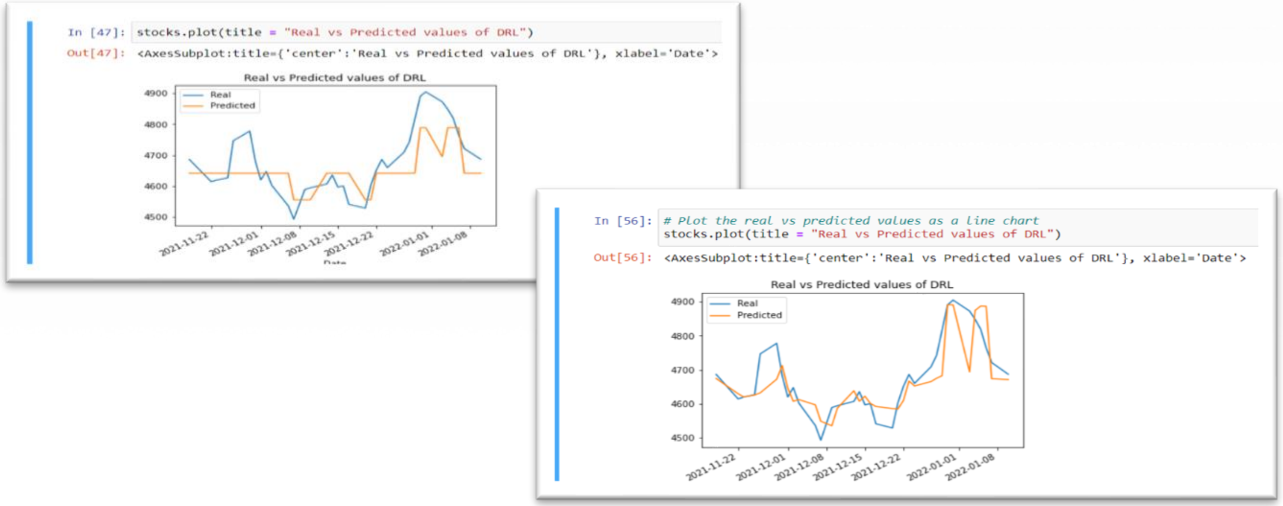
- ❑ Which model to choose?
- ❑ How can we measure it?

	Random Forest Classifier			Naïve Bayes Classifier		
	Actual PASS	Actual FAIL	TOTAL	Actual PASS	Actual FAIL	TOTAL
Predicted Pass	12	6	18	14	4	18
Predicted Fail	8	10	18	15	3	18
TOTAL	20	16	36	29	7	36

	Random Forest	Naïve Bayes Classifier
Accuracy	61.1%	47.2%
Precision	61.1%	46.0%
Recall	61.1%	47.0%
F1	61.1%	42.0%

Measure	Random Forest Regressor	XG Boost Regressor
RMSE	**17.80%**	**14.70%**
R-Sq	**51.40%**	**67.10%**

- *Using regression approach yielded favorable results as against classification.*
- *These models relied heavily on previous prices & as is visible from the trends, Twitter sentiments were not accurately indicative of stock price movements.*
- *The XG Boost regressor was observed to be the optimum model to simulate future prices.*



```
In [47]: stocks.plot(title = "Real vs Predicted values of DRL")
Out[47]: <AxesSubplot:title={'center':'Real vs Predicted values of DRL'}, xlabel='Date'>
```

```
In [56]: # Plot the real vs predicted values as a line chart
         stocks.plot(title = "Real vs Predicted values of DRL")
Out[56]: <AxesSubplot:title={'center':'Real vs Predicted values of DRL'}, xlabel='Date'>
```

CRISP-ML(Q): Evaluation

Validate performance →

- *Split into two parts - Training & Testing data*
- *Training & Testing data to be completely exclusive*

Determine robustness →

- *Measure KPIs like Accuracy %, Precision, Recall, F1 score etc.*
- *Compute Error & other error parameters like RMSE*

Measure	Random Forest	Naïve Bayes
Precision	61%	46%
Recall	61%	47%
F1 Score	61%	42%
Support	36	36

Measure	Random Forest Regressor	XG Boost Regressor
RMSE	17.80%	14.70%
R-Sq	51.40%	67.10%

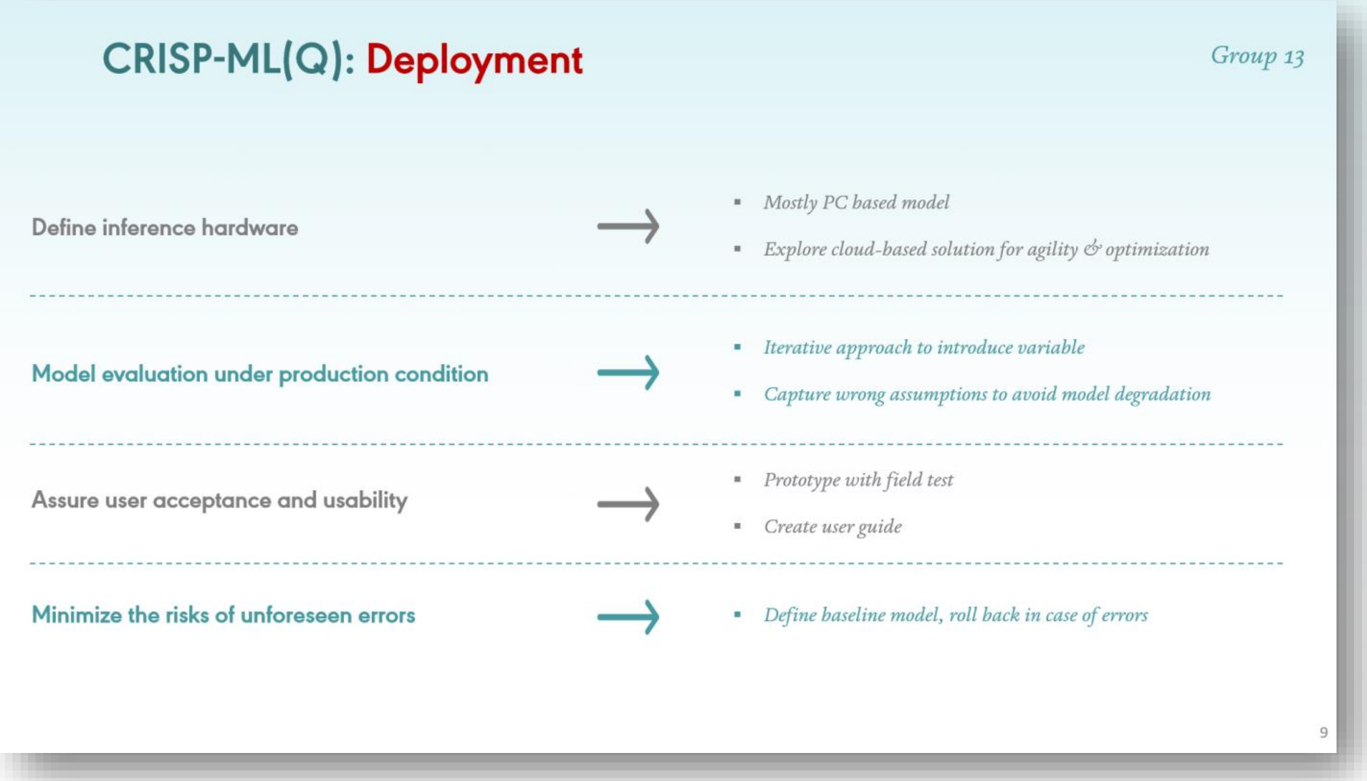
Increase Explainability for ML practitioner & end user

Compare results with defined success criteria

→

- *Predicted result should be explainable to the user/ ML practitioner.*
- *Analytical or visual output of Prediction should be explained to user. Example: Past real vs predicted stock prices.*

CRISP-ML(Q): Deployment



CRISP-ML(Q): Deployment

Group 13

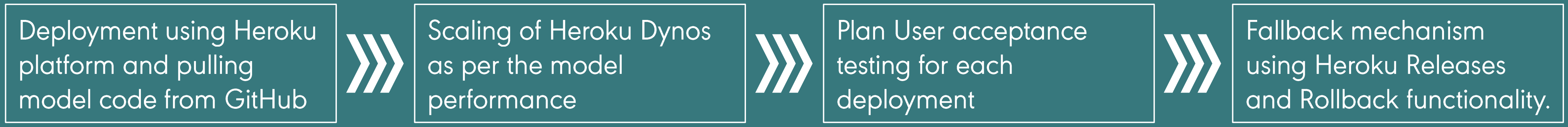
Define inference hardware →	- *Mostly PC based model* - *Explore cloud-based solution for agility & optimization*
Model evaluation under production condition →	- *Iterative approach to introduce variable* - *Capture wrong assumptions to avoid model degradation*
Assure user acceptance and usability →	- *Prototype with field test* - *Create user guide*
Minimize the risks of unforeseen errors →	- *Define baseline model, roll back in case of errors*

Selected Heroku deployment as it provides

- **Dynos:** *Smart containers on a reliable, fully managed runtime environment.*
- **User friendly interface**
- **Direct Deployment** *of Python code from GitHub*
- **Intuitive dashboard** *makes apps management easier*

Model performance is evaluated using a baseline number of Dynos.

Deployment Strategy



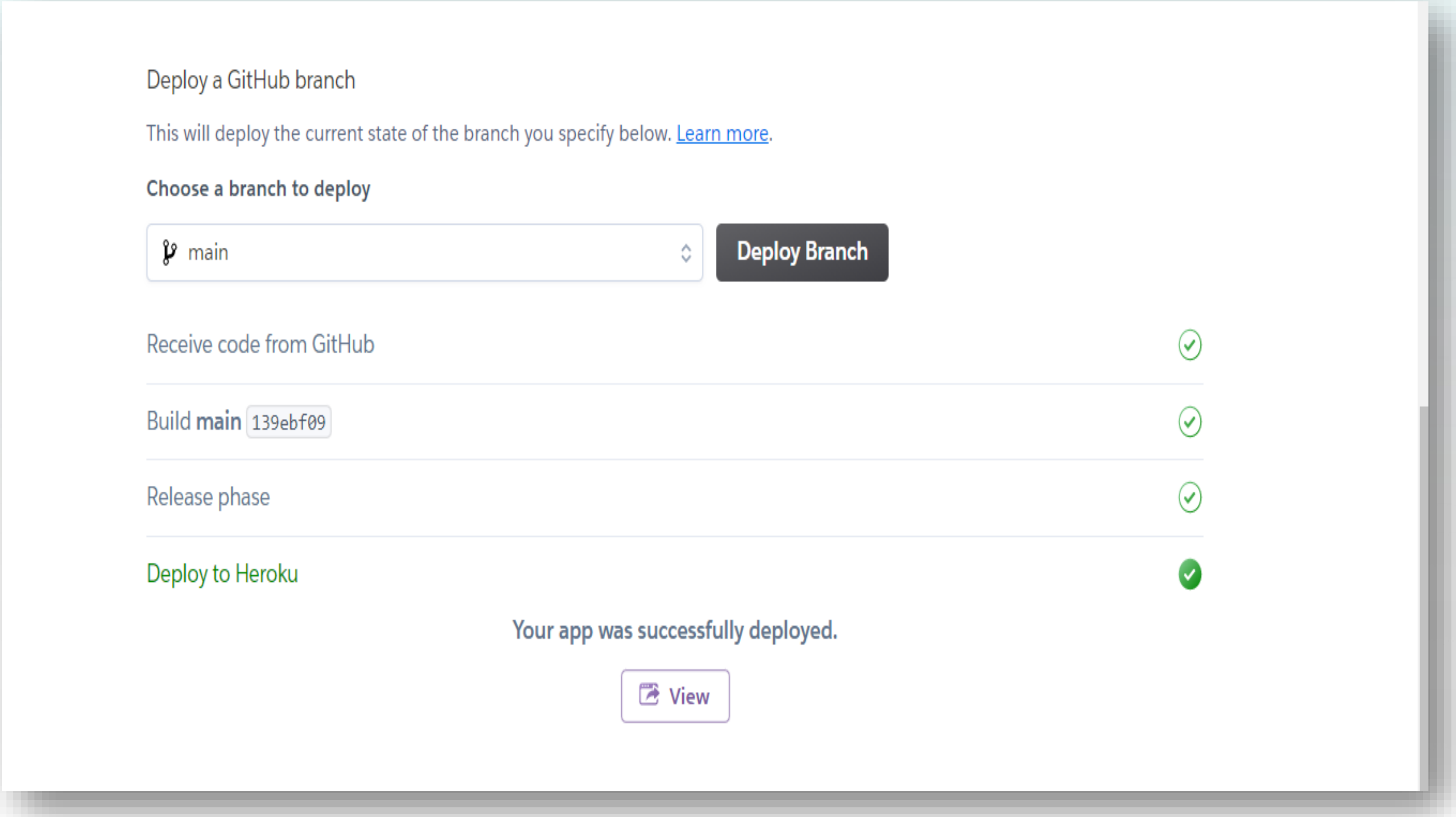
Deployment using Heroku platform and pulling model code from GitHub

Scaling of Heroku Dynos as per the model performance

Plan User acceptance testing for each deployment

Fallback mechanism using Heroku Releases and Rollback functionality.

Successful deployment on ‘Heroku’



Deploy a GitHub branch

This will deploy the current state of the branch you specify below. Learn more.

Choose a branch to deploy

Receive code from GitHub ✓

Build **main** `139ebf09` ✓

Release phase ✓

Deploy to Heroku ✓

Your app was successfully deployed.

CRISP-ML(Q): Monitoring and Maintenance

Non-stationary data distribution	*Data distribution to be upgraded, in case of variation*
Degradation of hardware	*Track technological change which degrades hardware & hence performance*
System updates	*Change in PEST will require shift in strategic update*
Monitor	*Periodic monitoring of model (once in 2 weeks) for consistency & accuracy*
Update	*Archive data before scraping for prediction*

Thanks!!

Group 13