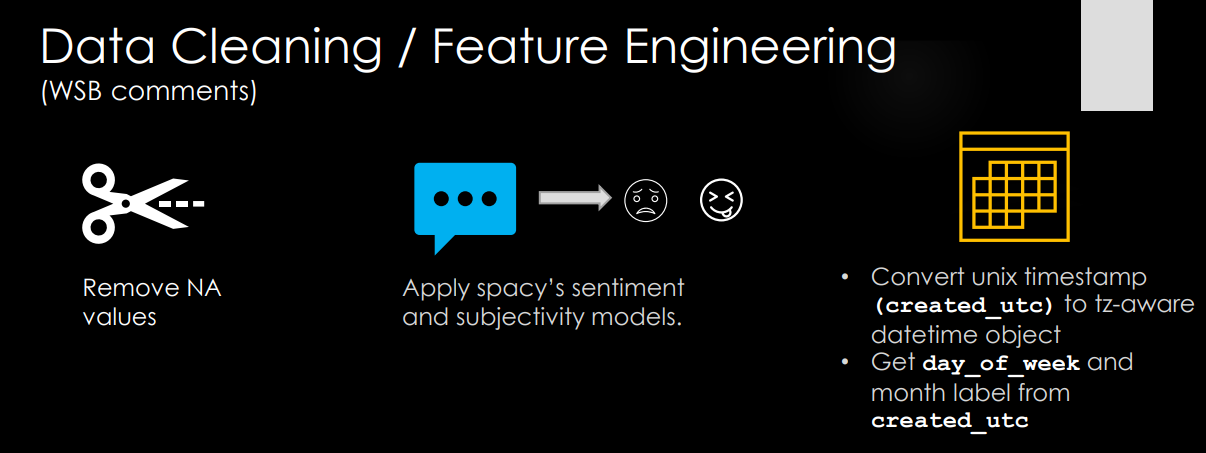
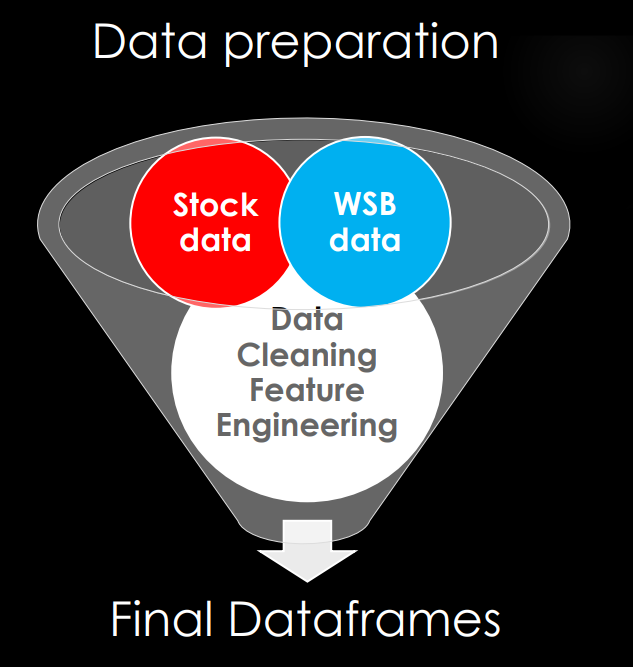
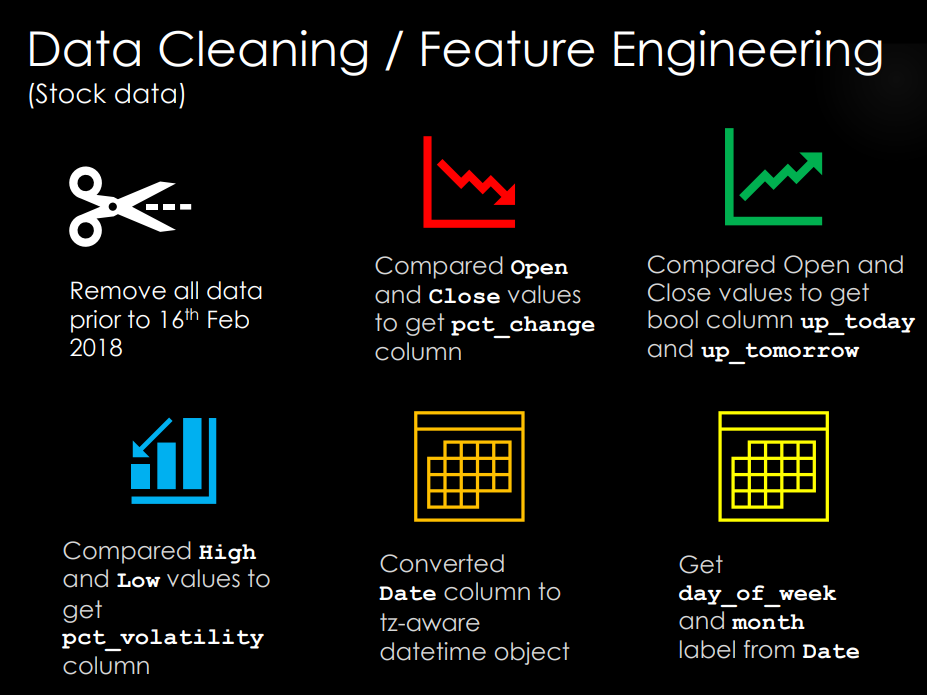
**Capstone Project**

**Document Skeleton**

Process overview





Environment:

OS: Win 10  
Python: 3.8.8  
Libraries:  
Package Version

---------------------- -------------------

absl-py 0.12.0

argon2-cffi 20.1.0

astunparse 1.6.3

async-generator 1.10

attrs 20.3.0

backcall 0.2.0

bleach 3.3.0

blis 0.7.4

cachetools 4.2.1

catalogue 1.0.0

certifi 2020.12.5

cffi 1.14.5

chardet 4.0.0

click 7.1.2

colorama 0.4.4

cycler 0.10.0

cymem 2.0.5

Cython 0.29.21

decorator 4.4.2

defusedxml 0.7.1

en-core-web-md 2.3.1

entrypoints 0.3

filelock 3.0.12

flatbuffers 1.12

gast 0.3.3

gensim 4.0.0

google-auth 1.28.0

google-auth-oauthlib 0.4.4

google-pasta 0.2.0

grpcio 1.32.0

h5py 2.10.0

idna 2.10

importlib-metadata 3.7.3

ipykernel 5.3.4

ipython 7.21.0

ipython-genutils 0.2.0

jedi 0.17.0

Jinja2 2.11.3

joblib 1.0.1

jsonschema 3.2.0

jupyter-client 6.1.7

jupyter-core 4.7.1

jupyterlab-pygments 0.1.2

Keras-Preprocessing 1.1.2

kiwisolver 1.3.1

Markdown 3.3.4

MarkupSafe 1.1.1

matplotlib 3.4.0

mistune 0.8.4

murmurhash 1.0.5

nbclient 0.5.3

nbconvert 6.0.7

nbformat 5.1.2

nest-asyncio 1.5.1

nltk 3.5

notebook 6.3.0

numpy 1.19.5

oauthlib 3.1.0

opt-einsum 3.3.0

packaging 20.9

pandas 1.2.3

pandocfilters 1.4.3

parso 0.8.1

pathy 0.4.0

pickleshare 0.7.5

Pillow 8.1.2

pip 21.0.1

plac 1.1.3

preshed 3.0.5

prometheus-client 0.9.0

prompt-toolkit 3.0.17

protobuf 3.15.7

pyasn1 0.4.8

pyasn1-modules 0.2.8

pycparser 2.20

pydantic 1.7.3

Pygments 2.8.1

pyparsing 2.4.7

pyrsistent 0.17.3

python-dateutil 2.8.1

python-Levenshtein 0.12.2

pytz 2021.1

pywin32 227

pywinpty 0.5.7

pyzmq 20.0.0

regex 2021.3.17

requests 2.25.1

requests-oauthlib 1.3.0

rsa 4.7.2

sacremoses 0.0.44

scikit-learn 0.24.1

scipy 1.6.2

seaborn 0.11.1

Send2Trash 1.5.0

setuptools 52.0.0.post20210125

six 1.15.0

smart-open 3.0.0

spacy 2.3.5

spacy-legacy 3.0.1

spacytextblob 0.1.7

srsly 1.0.5

tensorboard 2.4.1

tensorboard-plugin-wit 1.8.0

tensorflow 2.4.1

tensorflow-estimator 2.4.0

termcolor 1.1.0

terminado 0.9.3

testpath 0.4.4

textblob 0.15.3

thinc 7.4.5

threadpoolctl 2.1.0

tokenizers 0.10.2

tornado 6.1

tqdm 4.59.0

traitlets 5.0.5

transformers 4.5.0

typer 0.3.2

typing-extensions 3.7.4.3

urllib3 1.26.4

wasabi 0.8.2

wcwidth 0.2.5

webencodings 0.5.1

Werkzeug 1.0.1

wheel 0.36.2

wincertstore 0.2

wrapt 1.12.1

xgboost 1.3.3

zipp 3.4.1

Problem statement

Business problem statement:

* Bank wants to use WSB activity to give stock trading department an edge.

Data problems:

* Is there a relationship between the sentiment or number of comments on WSB and stock prices?
* Can we use stock data and WSB comments data to predict the next day’s price movement?

Industry/ domain

* This is using sentiment analysis to predict stock price movement

Stakeholders

* MD of Foreign bank

Data

* Data sources: Yahoo Finance daily price movement data for the following stocks:
  + GME, TSLA, PLTR, NOK, BB, SP500, AMC
* Kaggle: WSB comments from Feb-2018 to Feb-2021
  + 13m data points, contains comments and unix timestamp

Data science process

Data analysis

* Used pandas to remove NA values on text column
* Manually removed text such as ‘[deleted]’ or ‘[removed’] as it is not detected by pandas
* Convert unix timestamp to timezone aware datetime object

Feature engineering – WSB dataframe

* Obtained sentiment and subjectivity score by applying spacy model on WSB text – **This part took 2 days for the spacy model to obtain**
* Obtained labels for day\_of\_week and month from datetime object

Feature engineering – stock dataframe

* Remove dates prior to 16th Feb 2018
* Obtain pct\_change from Open and Close values
* Obtain up\_today and up\_tomorrow columns from from Open and Close values
* Obtain pct\_volatility from High and Low values
* Convert Date column to tz aware datetime object
* Obtained labels for day\_of\_week and month from Date column

Combining dataframes

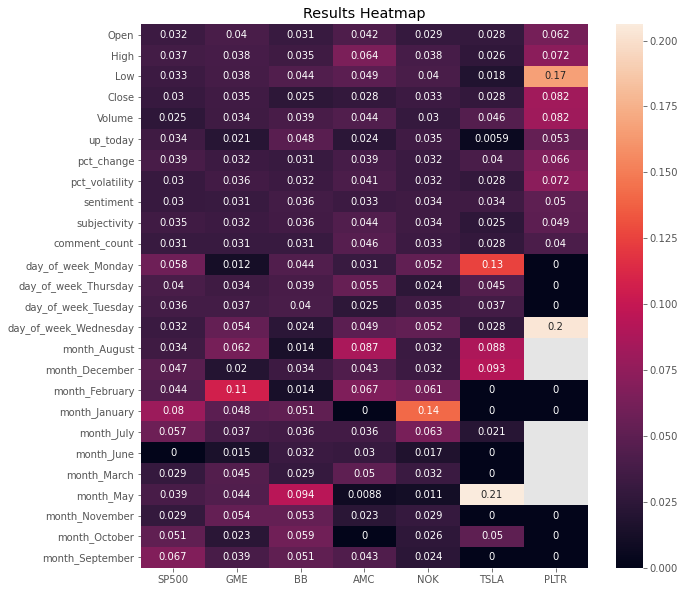
* Group wsb dataframe by date with the following data:
  + Mean subjectivity
  + Mean sentiment
  + Count of comments
* Left join wsb data on stock data resulting in 7 different dataframes
* Target variable is up\_tomorrow

Modelling

* Main features:
  + Open : float64
  + High : float64
  + Low : float64
  + Close : float64
  + Volume : float64
  + up\_today : bool
  + pct\_change : float64
  + sentiment : float64
  + subjectivity : float64
  + comment\_count : int64
  + day\_of\_week : category
  + month : category
* Target variable:
  + up\_tomorrow : bool
* Model used is default XGBoost classifier and trained on local machine.
* Model is evaluated with accuracy score
* 80% : 20% train\_test\_split was used

Outcomes

* Baseline prediction is to predict that the stock will go up every day
* Best performing stock is PLTR with 60% model prediction accuracy and an improvement of over 25% from the baseline prediction
* PLTR’s stock data is from 2nd Oct 2020 onwards while all other stocks had data from 16th Feb 2018
* PLTR could have more accurately represented r/wallstreetbets effect on the stock price as its membership growth started in Aug2020
* r/wallstreetbets could have initially have a low level of influence due to its membership count initially.
* Today r/wallstreetbets has over 9m members
* The stock that is most affected by sentiment and subjectivity is PLTR
* The stock that is most affected by comment\_count is AMC



Future work

* While the model shows that r/wallstreetbets may have some influence on the stock prices recently, more data and future work is needed
* Spacy’s sentiment and subjectivity models don’t fully capture the context of the language used in r/wallstreetbets
* In the future, BERT could be trained on labelled data to better understand wallstreetbets langauge

Conclusion

* The results of PLTR suggest that WSB could be starting to have influence on the stock market
  + The cult-like behavior in WSB is contagious and irrational
  + WSB encourages other people to hold stocks thorough memes – something never done before and goes against the oldschool investing advice
* The investing landscape has changed significantly since the Warren Buffet days
  + People were expecting a crash at the end of a 10 year economic cycle but no crash came
  + S&P broke 4000 in the middle of a pandemic
  + It seems like the economy and stock market movement ad decoupled
* Wallstreetbets could be a new way to identify very specific trading and investing opporunities