BTC price calculation Machine learning project

Manuel Carlos Cabanillas - The Bridge

07/05/2022

Project Presentation Agenda

07/05/2022

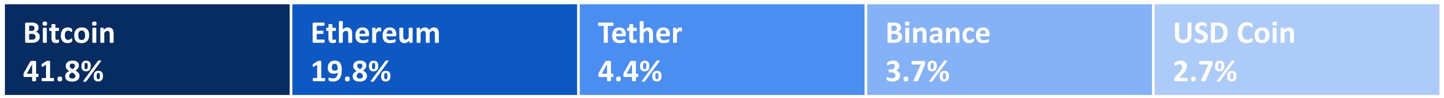
Approx. 15 minutes

Bitcoin introduction

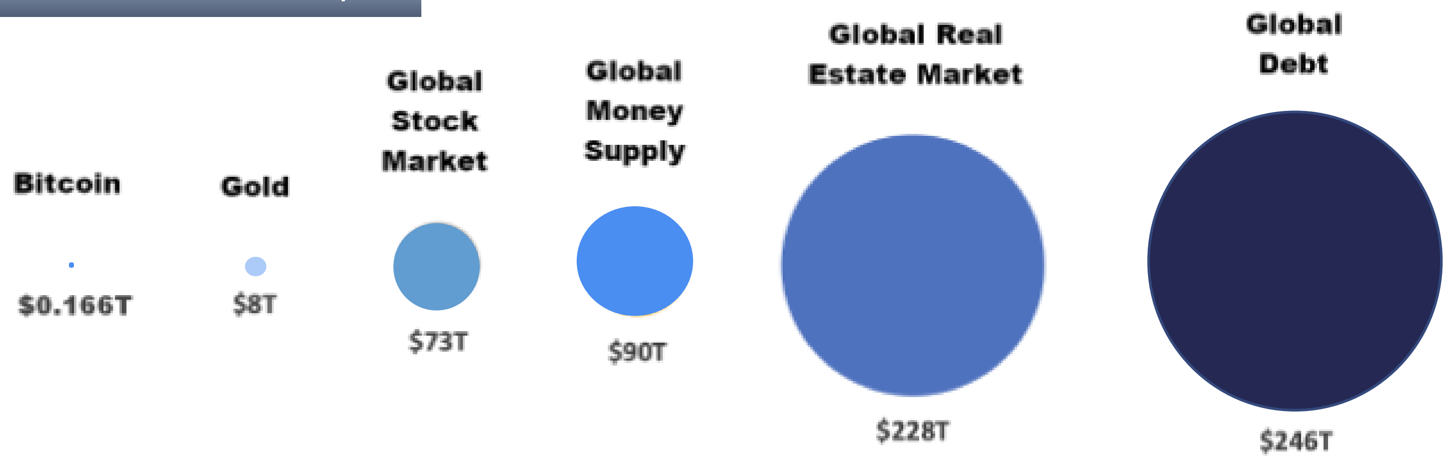
Bitcoin quick snapshot

- Bitcoin is a digital currency, introduced in 2008 by Satoshi Nakamoto (probably)
- It is enabled by the blockchain technology and allows for peer-to-peer transactions secured by cryptography
- Despite being the dominant cryptocurrency, it is small compared to investment market overall.
- Is a very volatile asset, its price almost doubling and halving again in the last year

BTC vs. Altcoins in %

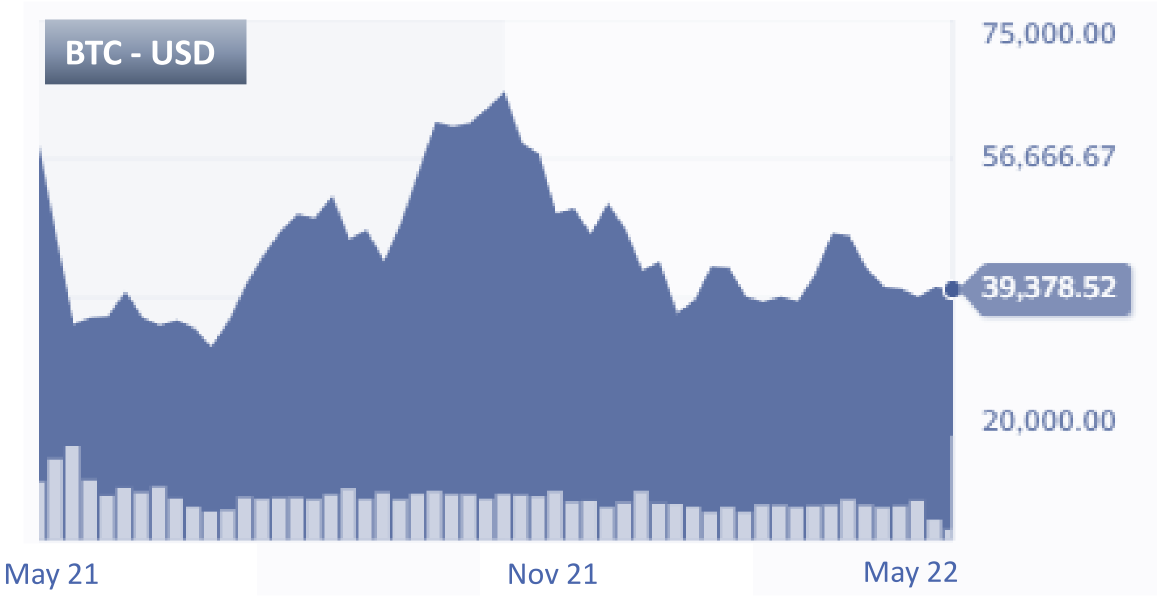


BTC vs. other assets in $T



Bitcoin quick snapshot

- Bitcoin is a digital currency, introduced in 2008 by Satoshi Nakamoto (probably)
- It is enabled by the blockchain technology and allows for peer-to-peer transactions secured by cryptography
- Despite being the dominant cryptocurrency, it is small compared to investment market overall.
- Is a very volatile asset, its price almost doubling and halving again in the last year



Blockchain technology

1- What it is, it is safe?

Database shared among many individuals, each one has a copy of it in its computer

Online-book format for the register of buy/sell/other txs

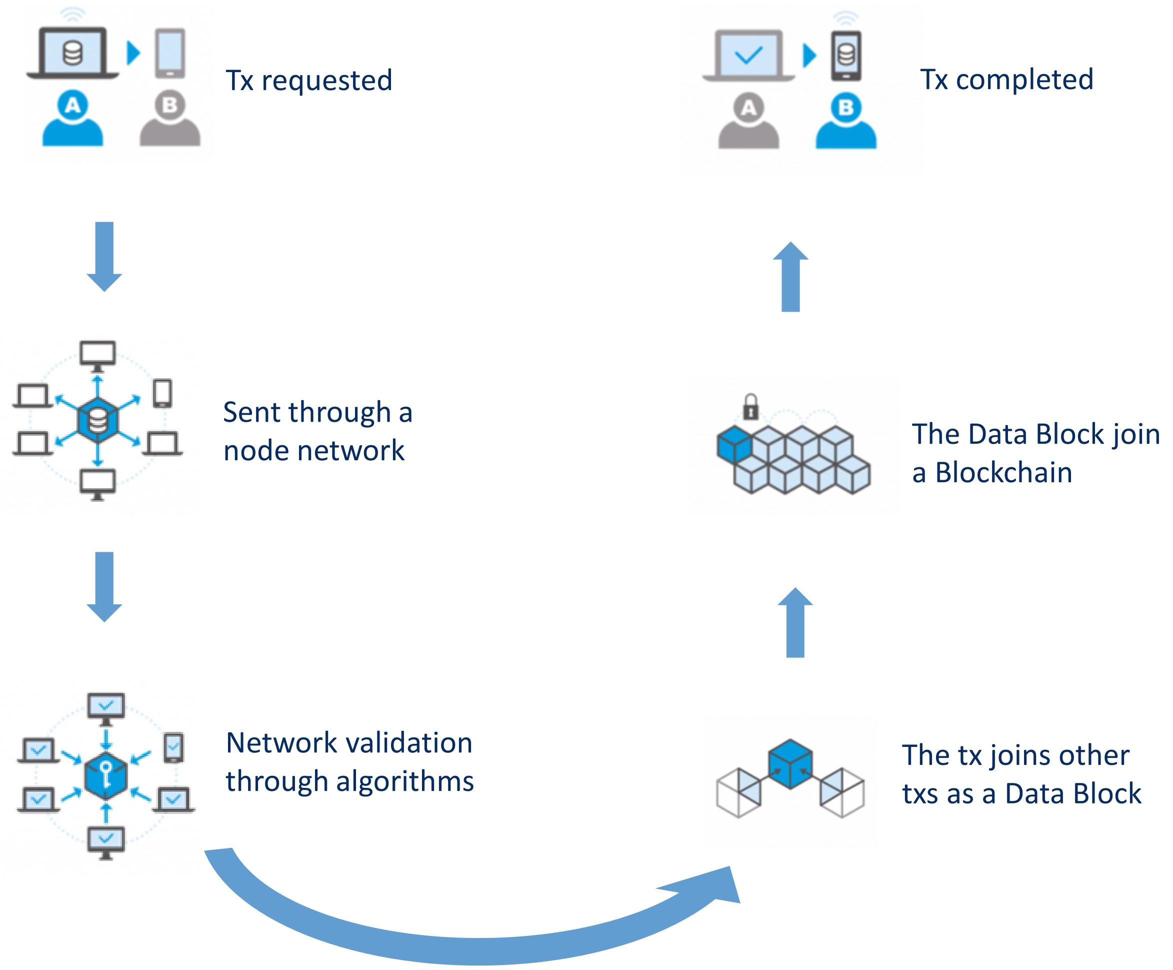
Txs codes, amounts, dates, participants – all registered

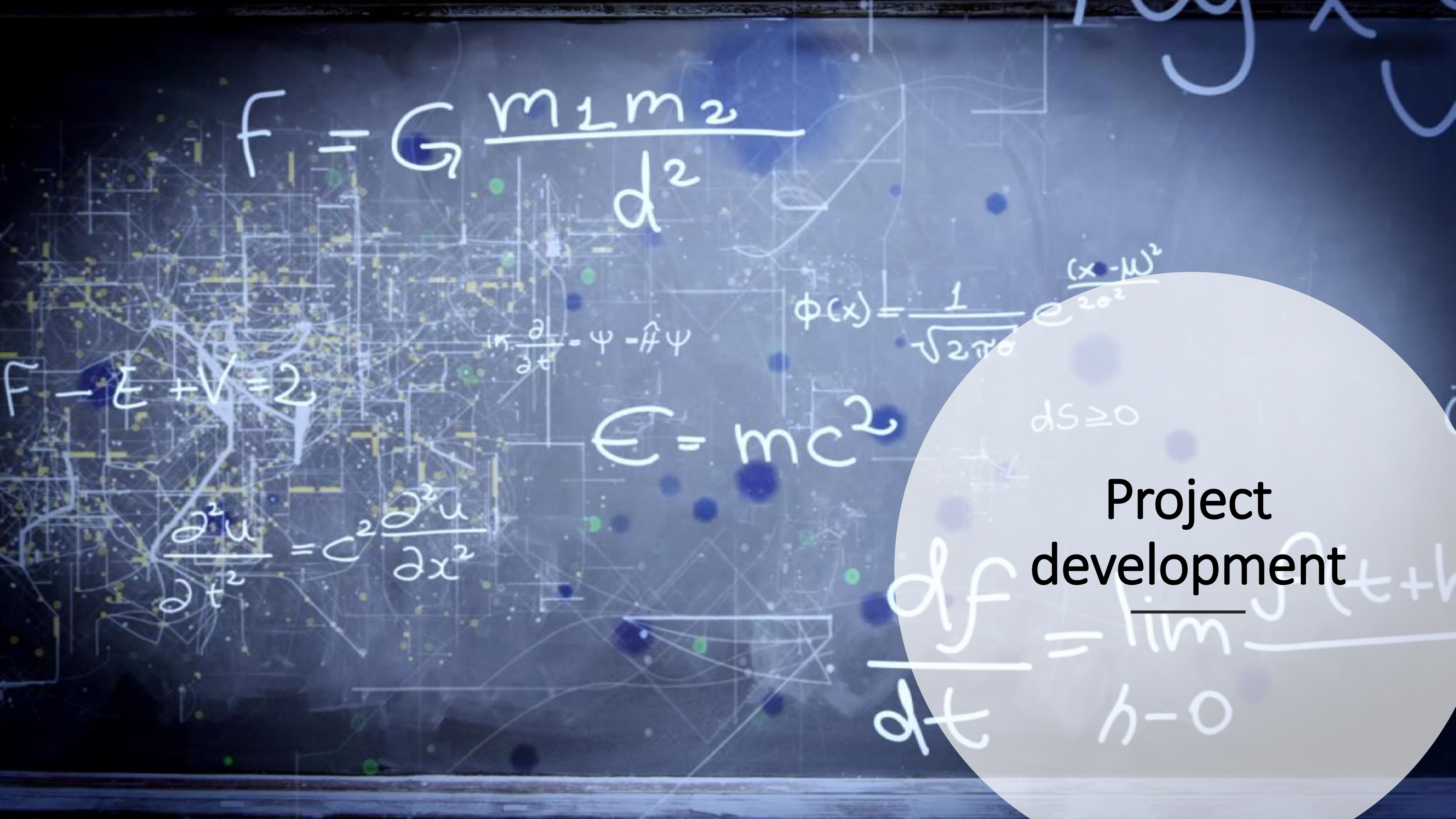
Hard to manipulate as it is behind a multi password code distributed among many users

For modifying the database it is needed a modification in several copies from different users, not just in one

The above described system is what the system is earning worldwide acceptance

2- Functioning





Project development

Project summary

DATA COLLECTION

- 58 features
- 9Y - From 01/2013 to 03/2022

DATA CLEANING & EXPLORATION

- Missing values imputation
- Correlation with target
- Inter - correlation

FEATURE ENGINEERING

- Ratios
- Technical indicators
- 4,973 transformed features

FEATURE SELECTION

- 1st step: 115 features based on correlation parameters
- 2nd step: Final selection of 18 features (Random forest)

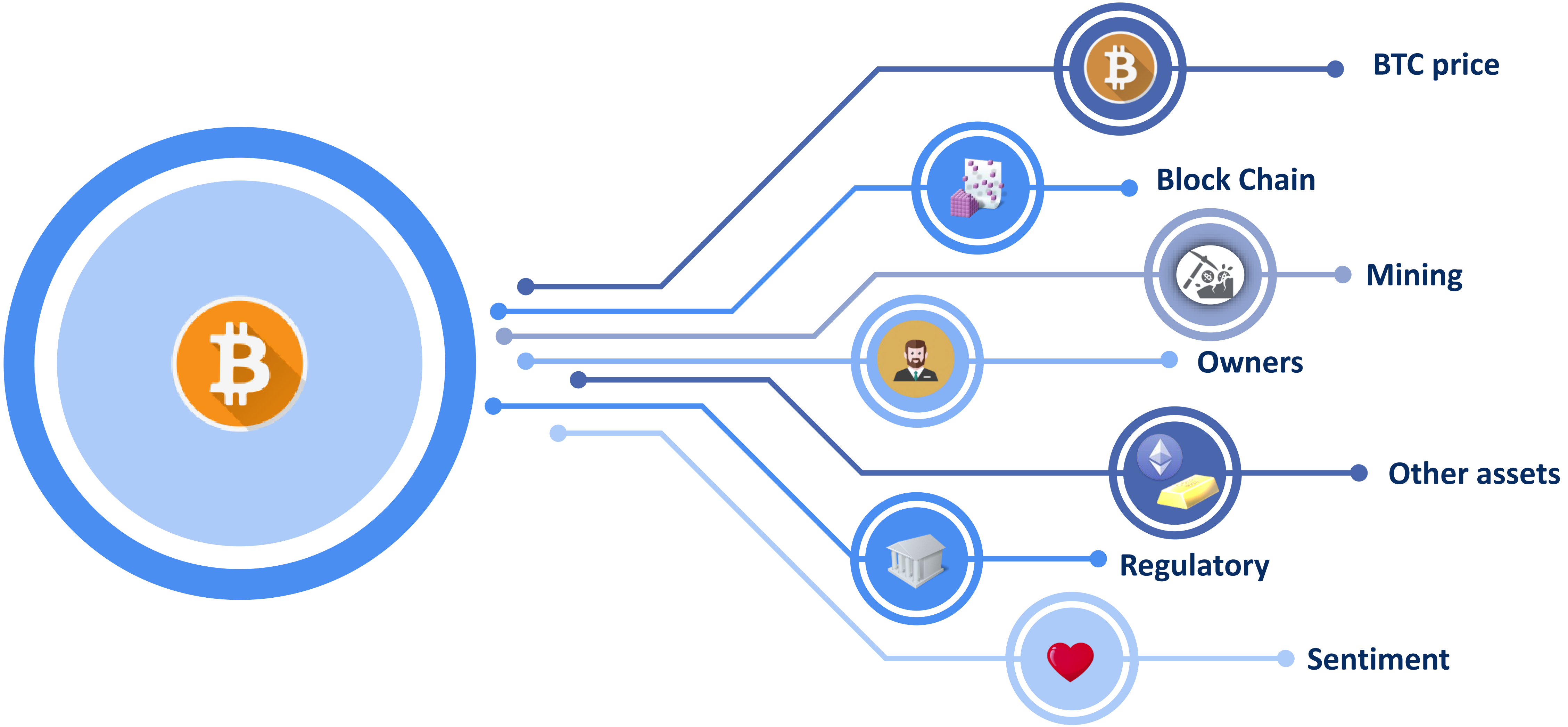
MODEL SELECTION

- 1st step: XG Boost, LSTM and Logistic Regression
- 2nd step: Different LRs contest
- Further features filtering based on contribution to model

PRODUCTION PHASE

- Automatization of final features for quick BTC daily price prediction 24 hours in advance

Data collection:
Influencing factors in Bitcoin price



Data collection:
Influencing factors in Bitcoin price

- BTC Closing Price
- Highest Price
- Lowest Price
- Avg. Price
- Market cap USD
- Volatility

- Nr. BlockChain tx
- Avg. block size
- Nr. addresses
- Sent coins in USD
- Avg. tx fee
- Median tx fee
- Average block time
- Total number coins
- Transactions_volume
- UTXOs (age)

- Mining Profitability
- Avg. mining difficulty
- Average hashrate
- Avg Fee/Block Reward
- Miners revenue

- 100 Richest/Total coins
- Bulls and Bears
- Break even Price
- Large holders Net Flow
- In /out the money
- Daily active adresses
- Avg. Transaction Value
- Large Transactions
- Med. Transaction Value

- DXY
- lbma_gold
- SP500
- MSCI ACWI Index
- WTI USD
- CBOE VIX
- FUTDXY
- FUTMSCIACWI
- FUT WTI
- FUT 500
- FUT GOLD
- FUT VIX
- FUT BTC
- ETH
- BDM ex. MegaCap

- Tweets per day
- Google Trends
- twitter sentiment
- github sentiment
- telegram sentiment

Data collection: Sources

Financial data

investpy
API

yahoo! finance
API

Nasdaq Quandl
API

Block Chain & Sentiment

BitInfoCharts
Scrap

csv

intotheblock
csv

Indexes & Futures

Cboe
Scrap

S&P Global
csv

FTX
Scrap

Data cleaning: Missing values

- **24/58 variables have missing values:**

Features	Missing values
FUTALT	2391
BDM_spot	2049
FUTMSCIACWI_	1430
ETH_spot	1164
lbma_gold	1055
DXY_spot	1049
SP500_spot	1048
VIX_spot	1048
FUTDXY	988
MSCIACW_spot	964
FUTVIX	953
WTI_spot	949
FUTWTI	949
FUTGOLD	941
FUT500	938
tweets_day	519
FUTBTC_	485
Weighted sentiment	197
Bull_Bear_Diff	101
active_addresses	22
top100_to_total_percentage	6
in_out_ratio_adj	3
profit_losses_ratio_adj	1
avg_block_time	1

Missing first years, holidays & weekends

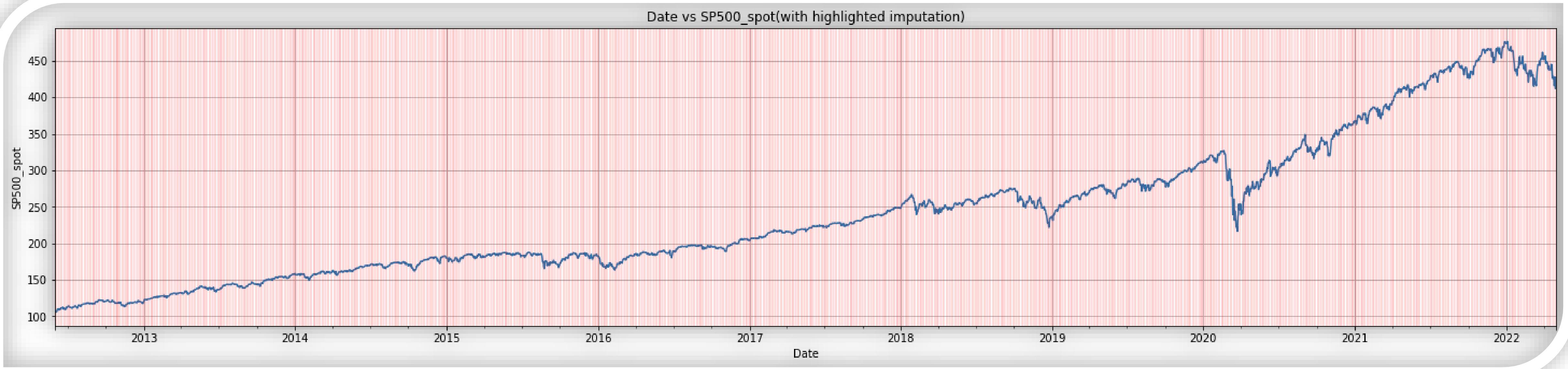
Missing holidays & Weekends

Missing first years data & other (random)

Missing other (random)

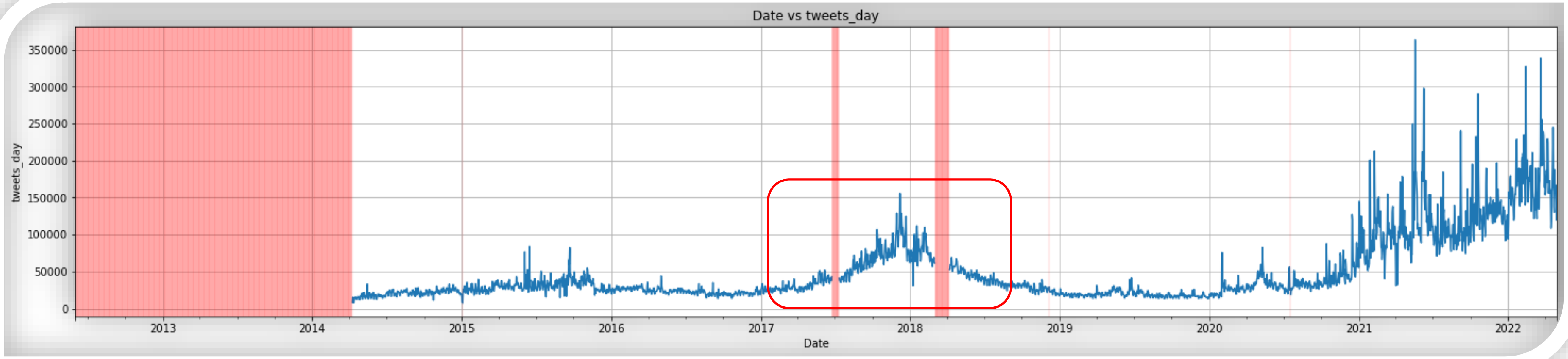
Data cleaning:
Missing holidays & weekends

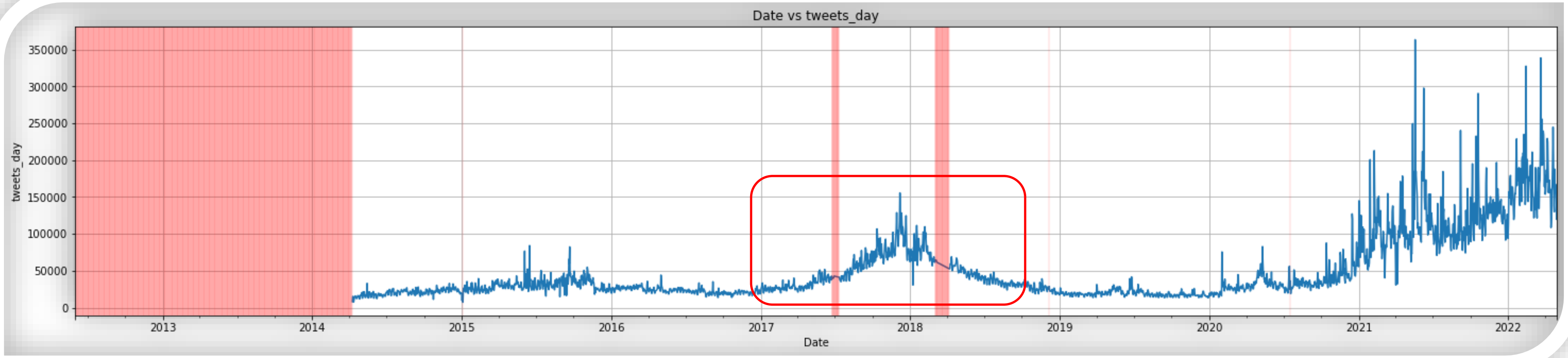
- For missing holidays & weekends: forwardfill, remembering last value



Data cleaning:
Missing values filling by type of miss

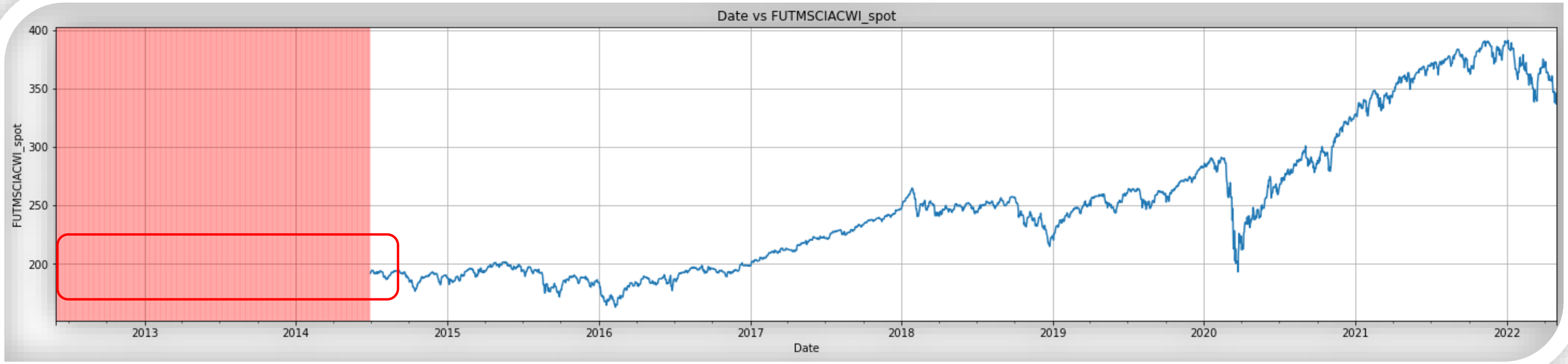
- For missing random values: interpolate

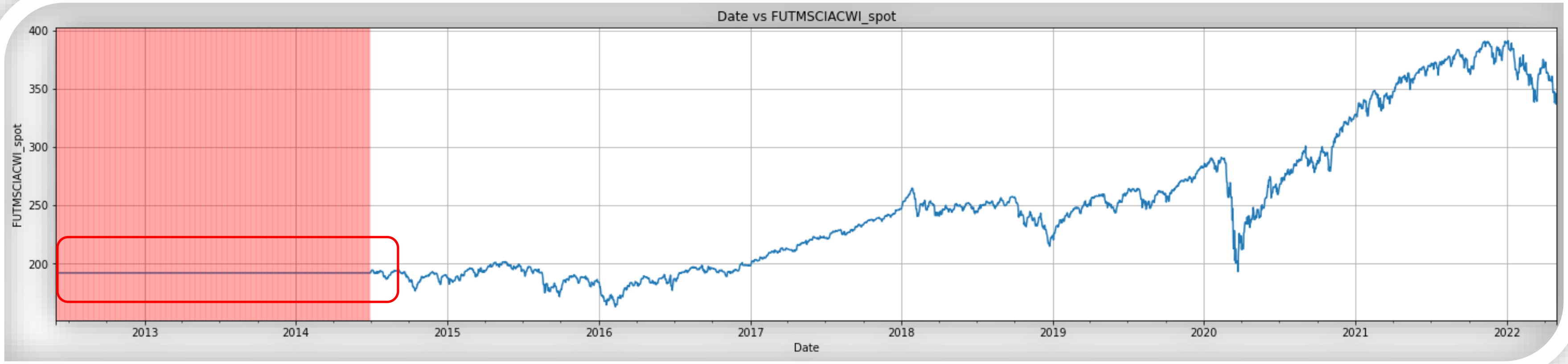




Data cleaning:
Missing values filling by type of miss

- For first years data missing values: Backfill, not optimal but aceptable given BTC low Price fluctuation for first years

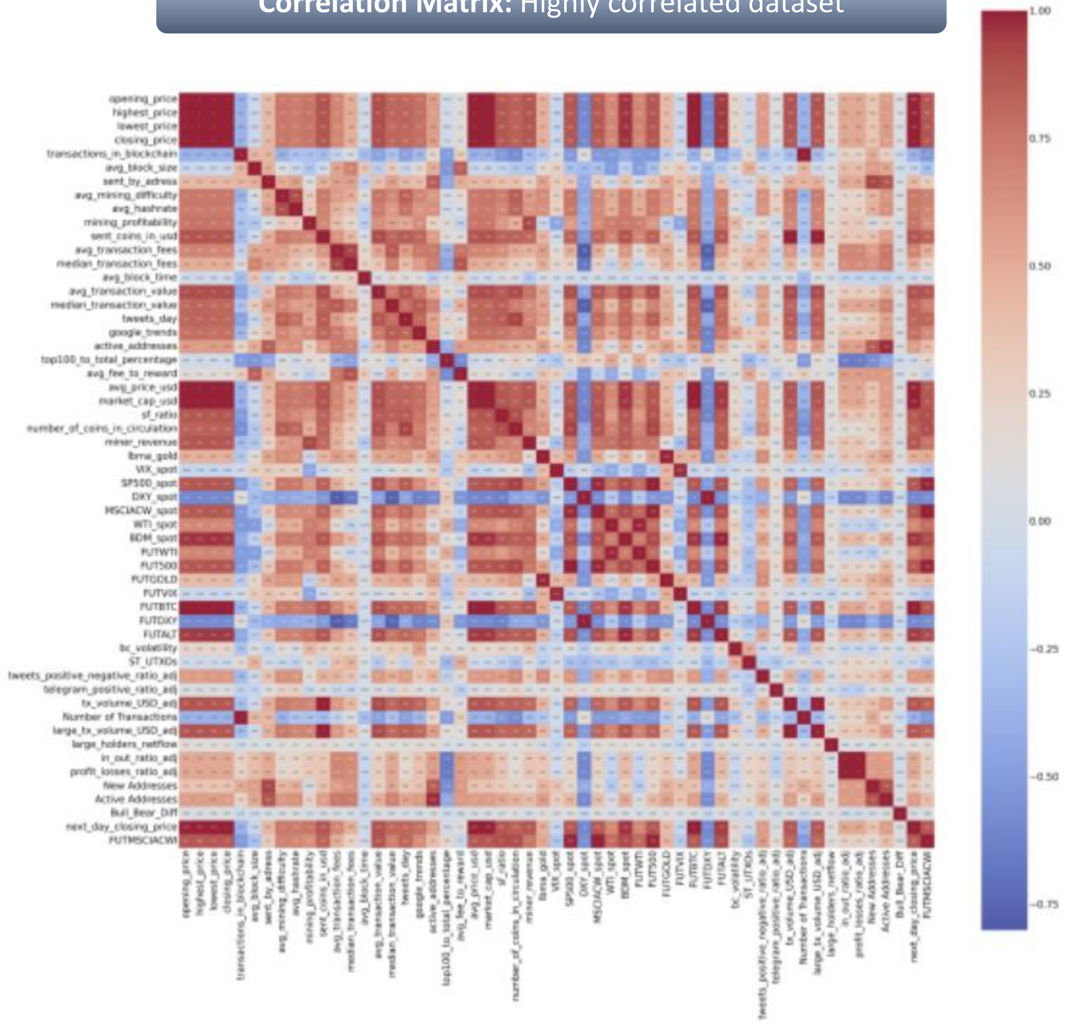




Features exploration:
Correlation Matrix & with target

- 58 features
- Approx. 3,376 datapoints

Correlation Matrix: Highly correlated dataset



Correlation vs. Target > 0.7 → 25 features: all categories represented

Feature	Correlation Coefficient
Closing_price	**0.998**
FUTBTC_spot	**0.998**
avg_price_usd	0.998
highest_price	0.998
market_cap_usd	0.998
lowest_price	0.998
opening_price	0.997
miner_revenue	0.951
ETH_spot	0.939
FUTMSCIACWI_spot	0.919
MSCIACW_spot	0.911
FUT500	0.894
SP500_spot	0.891
BDM_spot	0.890
sf_ratio	0.880593
avg_transaction_value	0.839
avg_hashrate	0.837
avg_mining_difficulty	0.836
tweets_day	0.815
Social_Volume	0.761
lbma_gold	0.729
FUTGOLD	0.718
sent_coins_in_usd	0.709
tx_volume_USD_adj	0.708
Social_Volume_AI	0.705

BTC Price spot & FUT

Other assets

Concentration

Mining

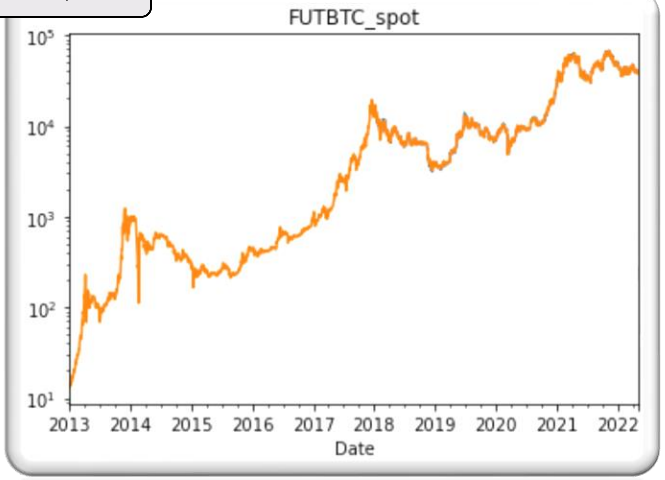
Sentiment

Block Chain

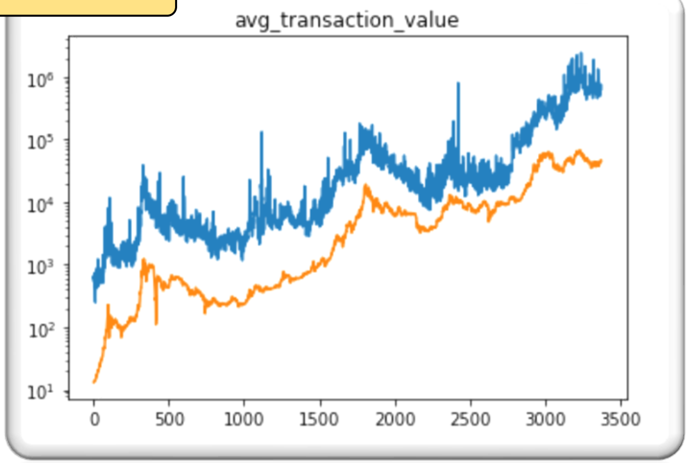
Feature exploration: Correlations with Target – log scale

- Target
- Feature

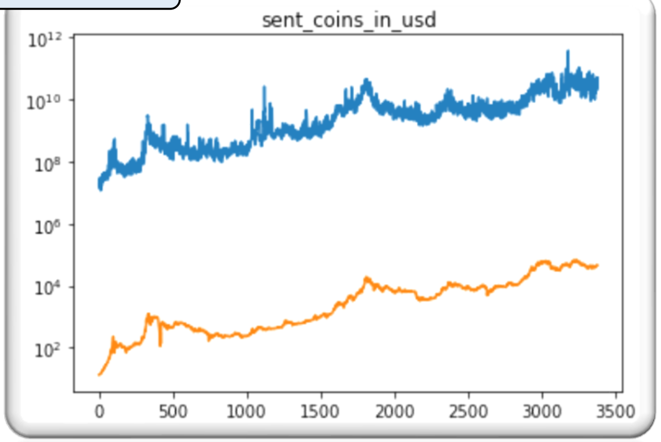
BTC Price / FUT



Concentration

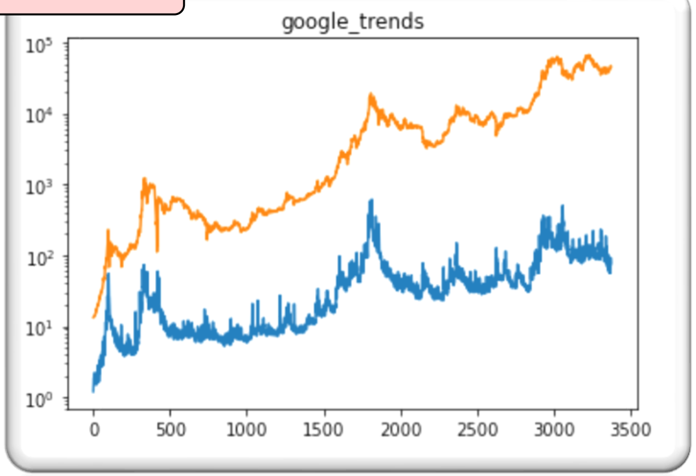


Block Chain



Mining

Sentiment



Other assets

Feature Engineering before selection

1. Ratios calculation done already before feature exploration:

- **Stock Flow ratio = Number_of_coins_in_circulation / Annual_flow (est.)**
- Tweets positive ratio = (Positive tweets - Negative tweets) / Total tweets
- Telegram positive ratio = idem.

- In/out the money = nr. Of owners in the money / nr. Of owners out the money
- Break_even = nr. Of owners with realized gains / nr . Ow owners with realized losses

2. Popular technical indicators calcuation:

- **Over bitcoin prices:**
  - Ichimoku clouds: ISA_9, ISB_26, ITS_9, IKS_26, ~~ICS_26~~ *

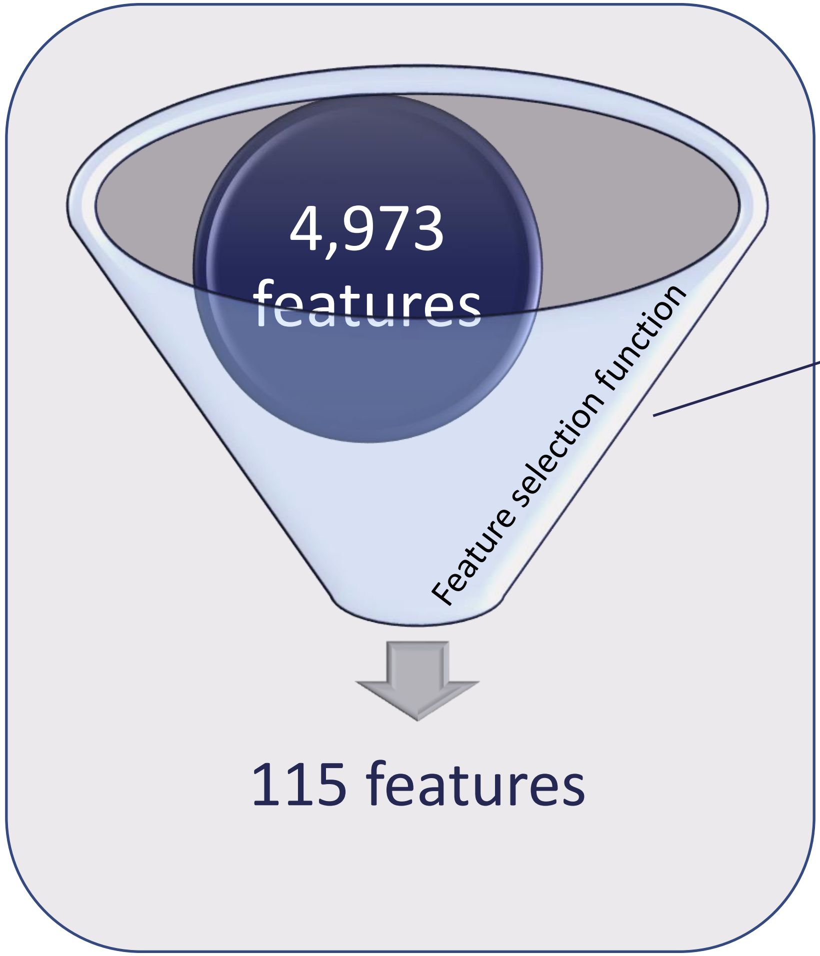
- **Over all variables** (for 3,5,7,10,15,30,60 and 90 days):
  - Moving averages: Sma, wma, ema, dema, tema, MACD
  - Volatility: Standard deviation & Variance
  - Intervals measuring :RSI, Bollinger Bands
  - Trends: Rate of Change

- 4,973 transformed variables

- Approx. 90 sub-features per initial feature

** ICS_26 Removed as it is calculated with future data*

Feature Selection: First filtering before running random forest



- For avoiding having an extremely correlated dataset, and reduce computational costs, for each initial feature we extract the 2 features most correlated with our target but avoiding correlation > 0.9 between the subfeatures, see example below:

Rank	Sub - feature	Correlation with target	Correlation with closing price
1	closing_price	0.999	1
2	closing_price tema3	0.999	0.999
3	closing_price dema3	0.999	0.999
…			
3,640	closing_price stdev60	0.898	0.897

Selected features

- Resulting 115 features

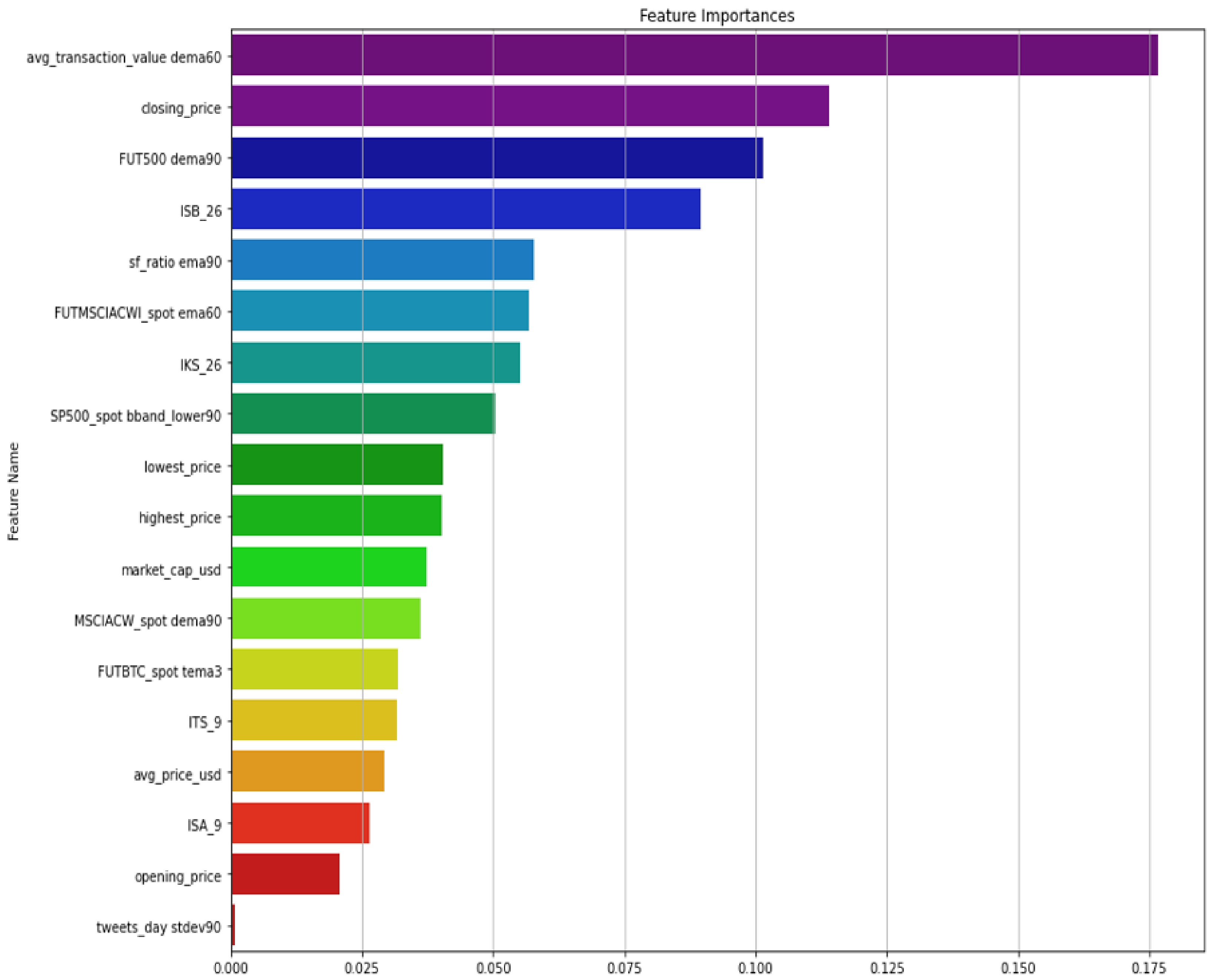
- Data scaling is applied after the filtering to all features except for ratios between 0 and 1

- Random forest regressor is applied afterwards with 2,000 n_estimators

Feature selection:

Random forest results

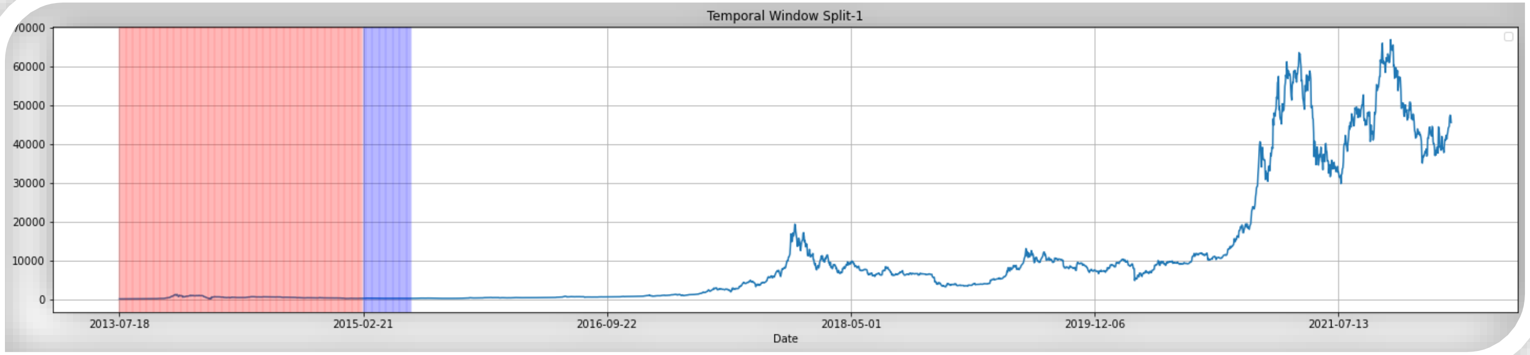
18 features contributing to the model



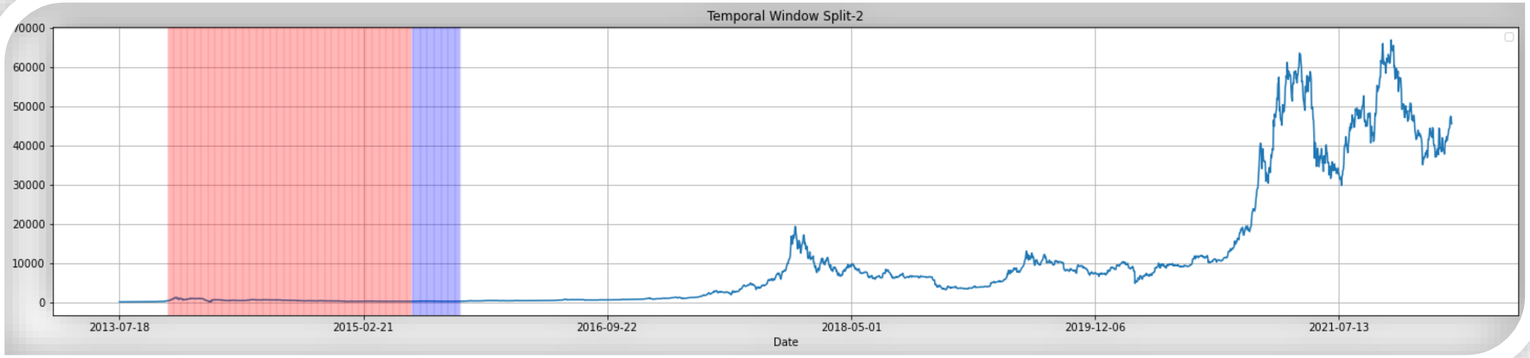
Model selection: multiple training windows

- Given Bitcoin recent incorporation and its volatility we Will take different training Windows for each prediction.
- Optimal value after testing has been fixed at 500 samples training window for 100 samples prediction.
- Below first 2 training(red)/testing(blue) windows of a total of 30.

Window 1



Window 2



Model selection:
First contest

- Metrics: RMSE , MAE , Pearson Correlation

- First iteration with different models: Linear Regression, LSTM, Random forest

Metric	XG Boost	LR	LSTM
mae_train	150	242	2,873
rmse_train	231	379	3,967
mae_test	1,695	500	6,418
rmse_test	2,143	665	6,840
Perason_corr	0.718	0.998	0.206

- Best model is Linear Regression so lets focus in regressions

Machine Learning models:

Regression selection

Full period ~ 3000 days

	LR	Ridge LR	Bayessian Ridge	LRSGD	ARD	ARD (params)
rmse_train	379	553	381	479	384	385
rmse_test	665	1043	560	754	522	503 ✓
Perason corr.	0.998	0.995	0.999	0.998	0.999	0.999

Last 900 days

	LR	Ridge LR	Bayessian Ridge	LRSGD	ARD	ARD (params)
rmse_train	850	1052	854	972	862	862
rmse_test	1541	1955	1289	1487	1190	1166 ✓
Perason corr.	0.996	0.991	0.997	0.995	0.997	0.997

Last 300 days

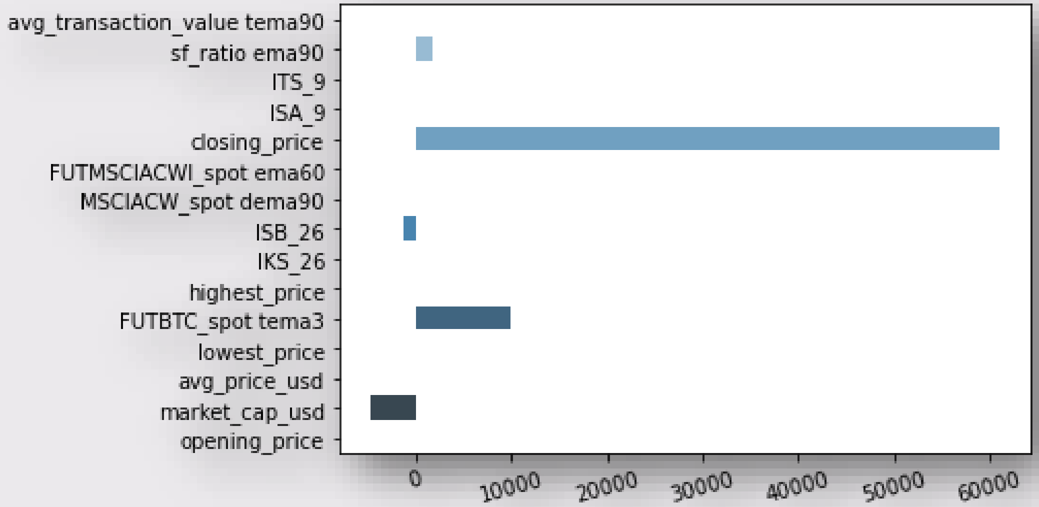
	LR	Ridge LR	Bayessian Ridge	LRSGD	ARD	ARD (params)
rmse_train	1594	1775	1602	1787	1617 !	1615
rmse_test	1744	1558	1533	1530	1407	1411 ✗
Perason corr.	0.978	0.978	0.980	0.977	0.982	0.982

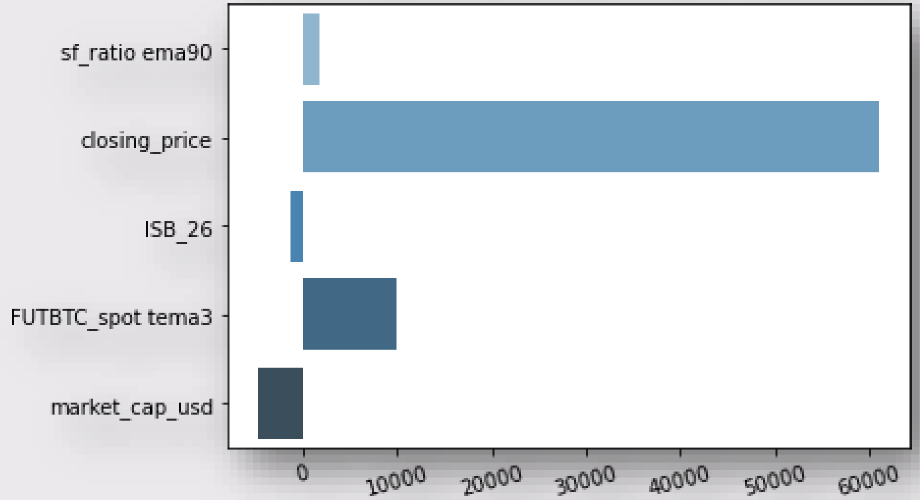
- **Model performance decreases the shorter the period** as: i) recent volatility, ii) lower difference in lowest and highest price
- **ARD and Bayessian models offered the best results**:
  - ! RMSE test is lower than train for less than 300 days (easier targets to predict tan in training)
  - ARD is chosen as it gives lower rmse in test data for all time periods and has les overfitting
  - Gridsearch CV have been applied to ARD algorithm but results are not better for all time periods

Machine Learning models:
Features contribution - Further Adjustment of ARD model

- Some features selected by the random forest are not contributing to the model.

- In a second run with only the key features results were very similar but Pearson correlation slightly improved from 0.982 to 0.983 for the last 300 days

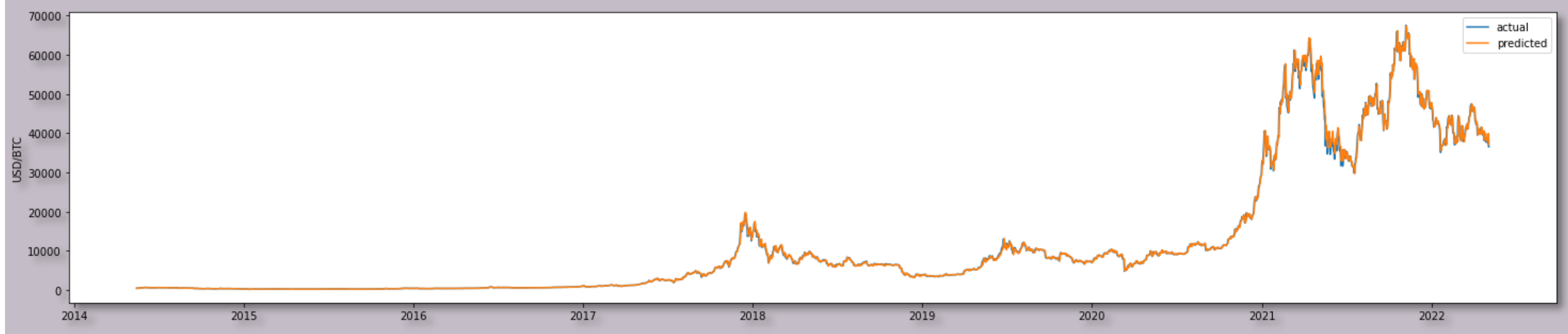




Model ready for production

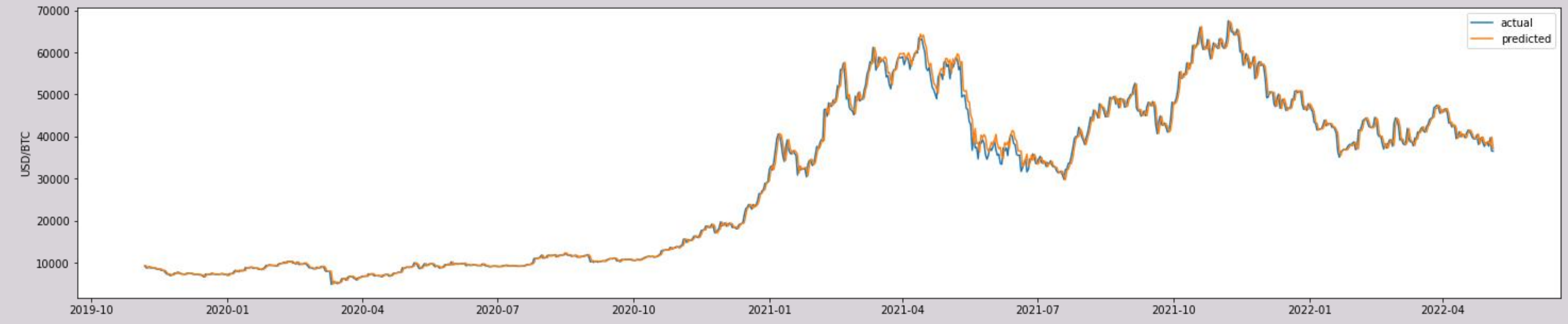
Full period
~ 3000 days

0.999%



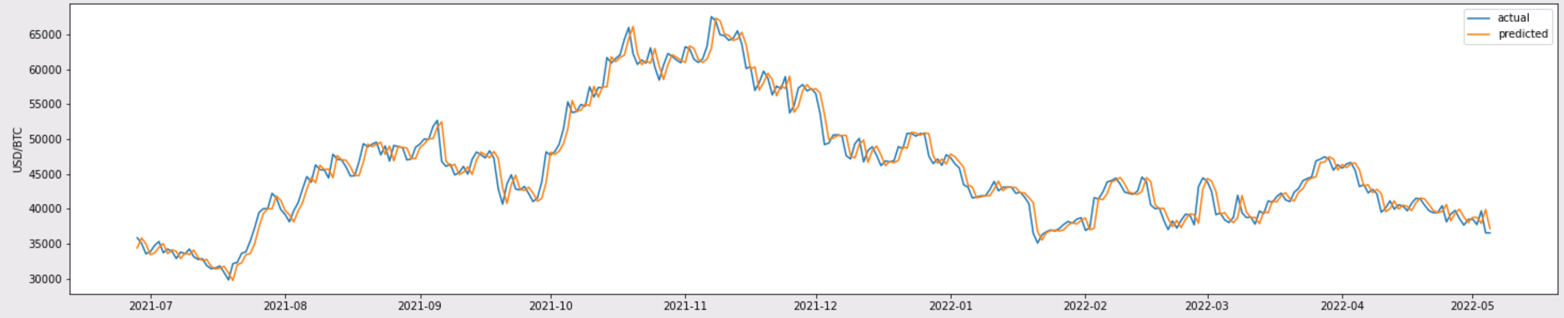
Last 900 days

0.997%



Last 300 days

0.982%

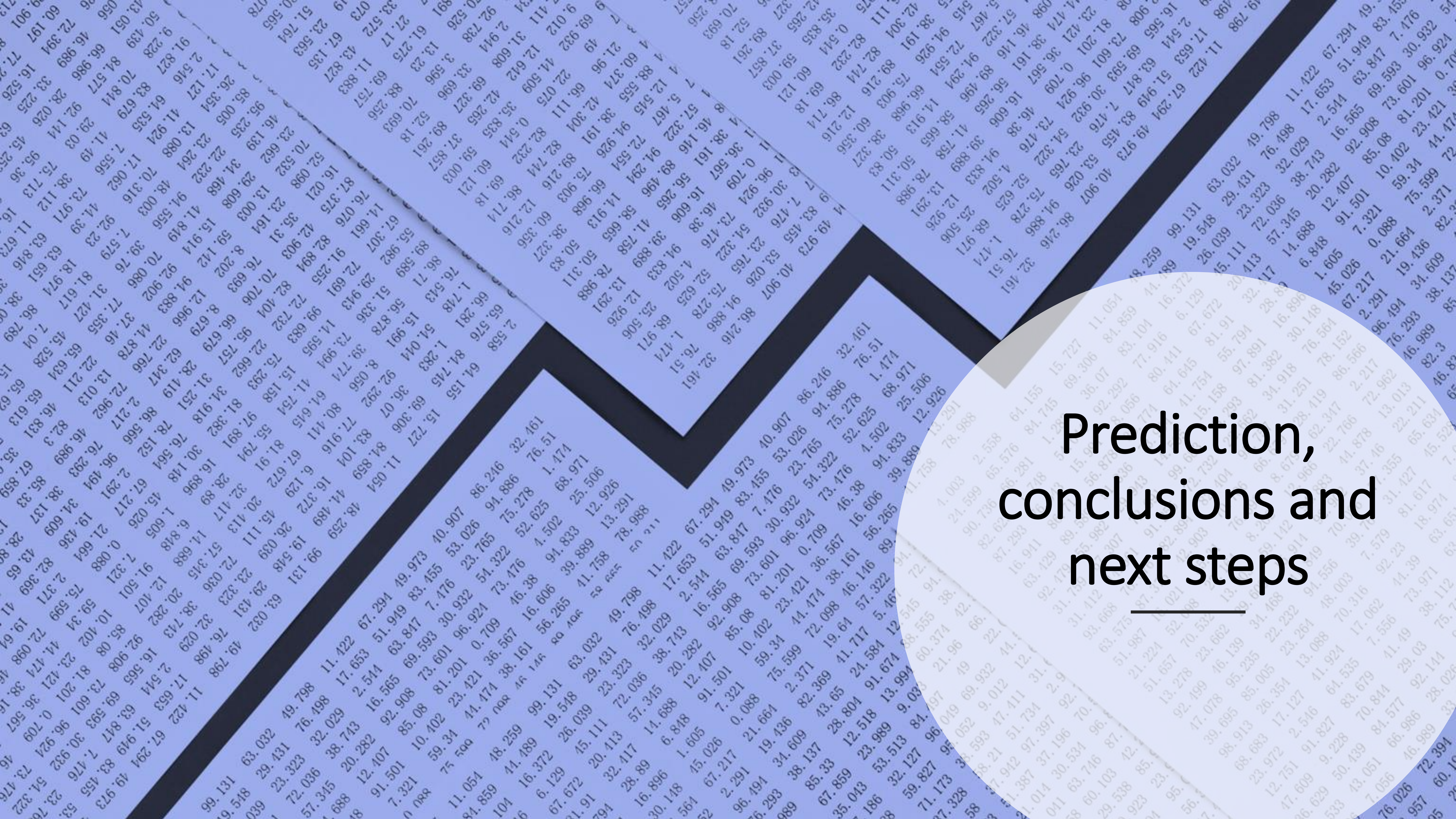


At what time can we make predictions? Can we get all features automatically?

- Need to take into consideration at what time the features are available from its respective sources:

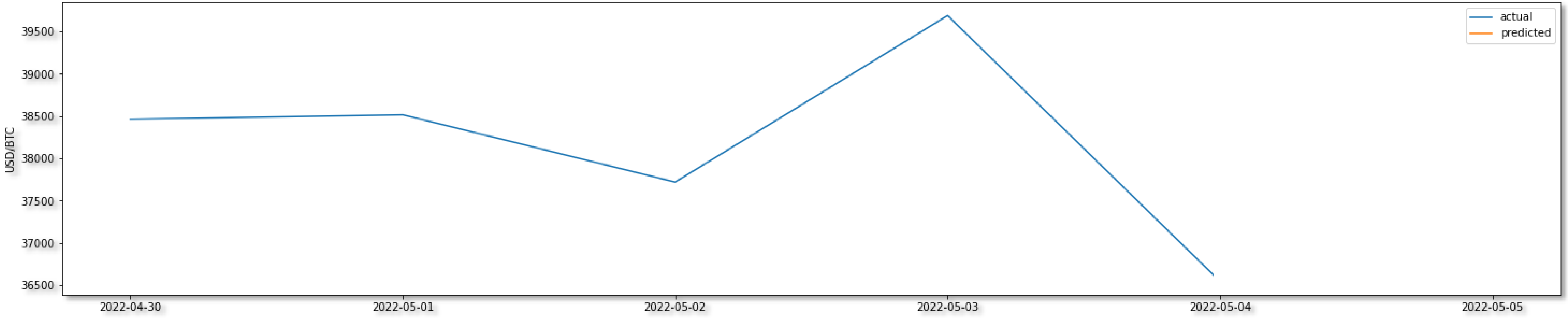
Feature	Source	UCT	CT	ET
BTC closing price	Investpy (API)	23:59	18:59	**01:59 (+1)**
ISB_26	BTC closing price	23:59	18:59	**01:59 (+1)**
Market cap USD	bitinfocharts.com (Scrap)	23:59	18:59	**01:59 (+1)**
Sf_ratio	Number of coins in circulation – Quandl (API)	23:59	18:59	**01:59 (+1)**
FUTBTC	Yahoo Finance (API)	--	16:00	**23:00**

- ✓ All features are available at ET 2:00 for make the prediction in 24 hours for the next day BTC closing Price at 2:00 next day
- ✓ All features can be obtained either from APIs or scratching with no manual input
- ✓ Model Output: `06/05/2022 predicted closing price :  36987.62 USD`

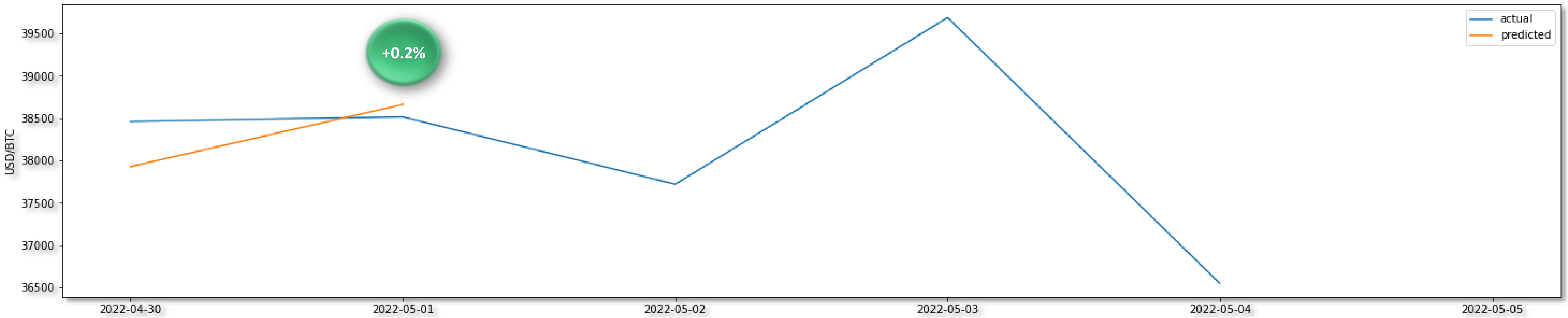


Prediction, conclusions and next steps

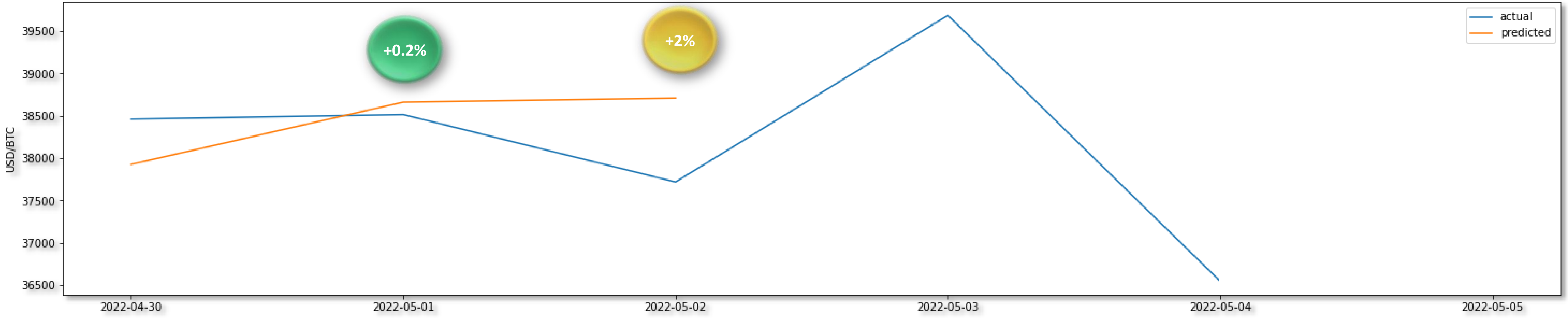
Zoom on last days predictions



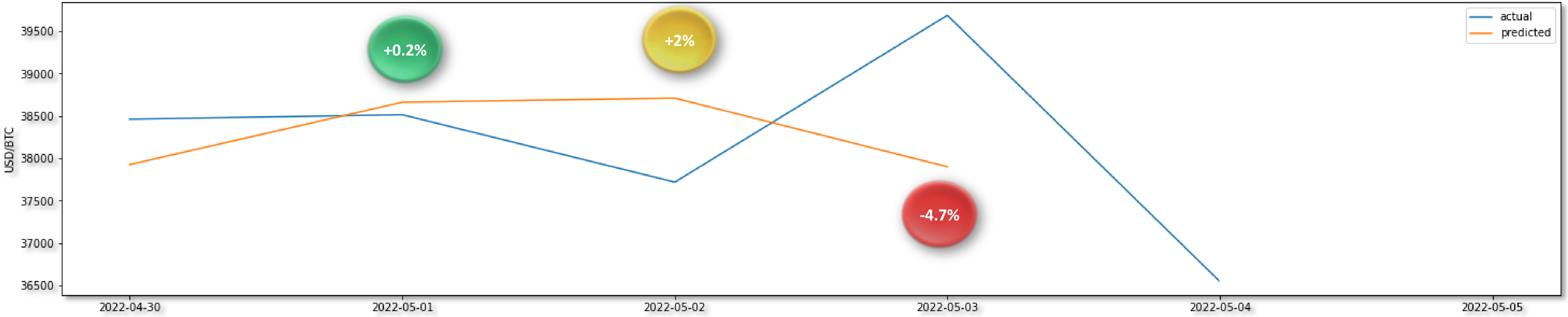
Zoom on last days predictions



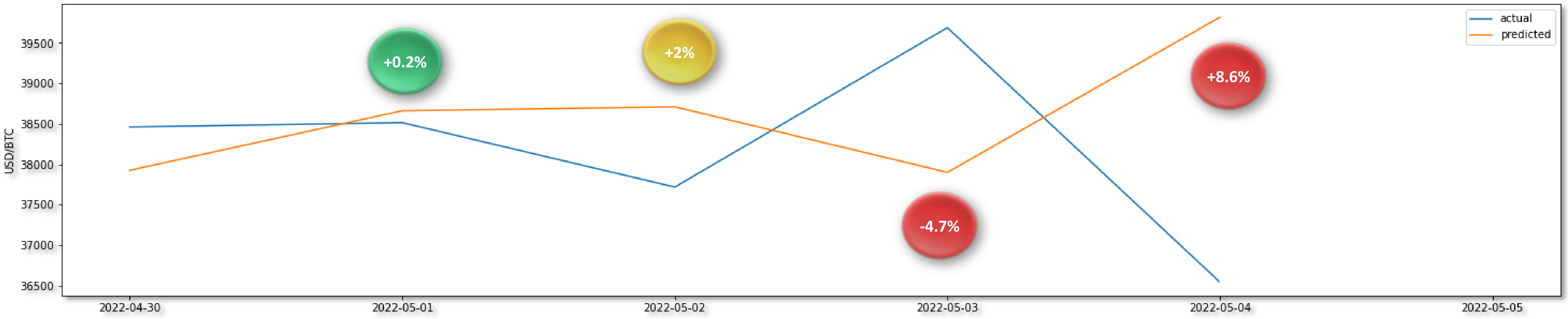
Zoom on last days predictions



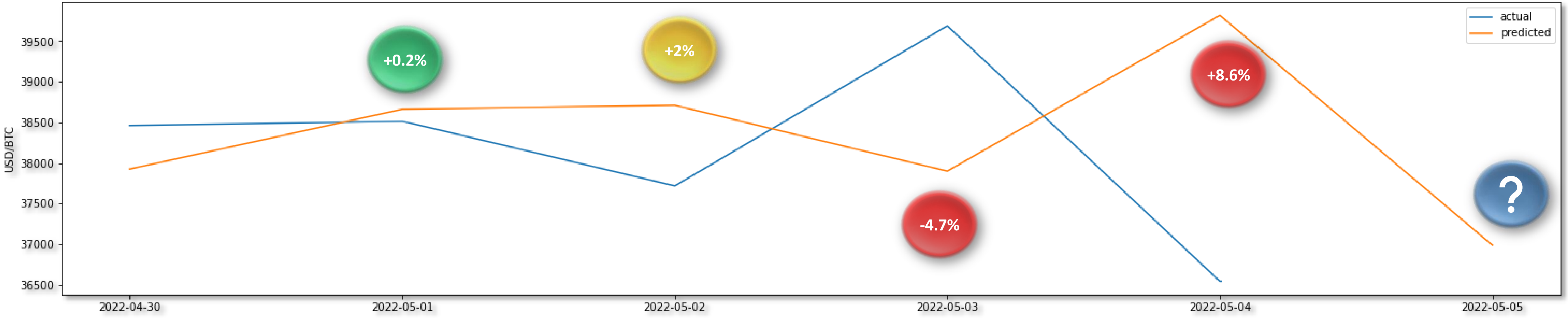
Zoom on last days predictions



Zoom on last days predictions

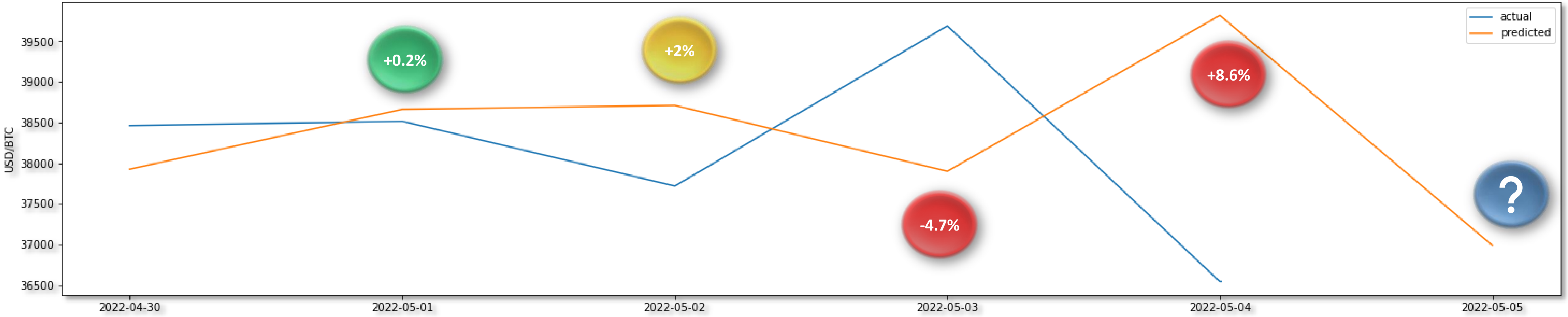


Zoom on last days predictions




```
06/05/2022 predicted closing price :  36987.62 USD
```

Zoom on last days predictions



- The model is not valid for trading confidently and results seems to be impacted by the intraday volatility

Date	Difference test/pred	Intraday volatility
2022-05-02	0.2%	2,8%
2022-05-03	2%	3%
2022-05-04	-4.7%	6%
2022-05-05	8.6%	10.1%

Conclusions and Next steps

Negative & Autocritic

- **Closing Price feature has a very high weight in the model** which makes the predictions to follow the previous day closing price
- This is partially due to intraday BTC **extreme volatility** which is highly correlated to the model error

Autocritic:

- **Bulky feature engineering**, withouth deep technical knowledge
- **Focus on Regressions due to apparently good results** and not enough time employed in time series approach with embedings

Positive

- **Strong data set** is already created with many features detected to have a strong impact on target
- **The framework to make quick predictions** and extract the data manually is already build
- **Error** in the model is **lower than** the daily **volatility** so if we can reduce its impact we might get better results

Next Steps

1. Review of feature engineering step trying to reduce nr. of resulting features

2. Focus on data trends more than in raw data:

- Try **LSTM** time series models with embeddings
- Transform data set to profitabilities (train and target)

3. Try intraday predictions :

- The shorter the term the lower the impact of volatility
- Sentiment variables will probably have a greater impact

Questions