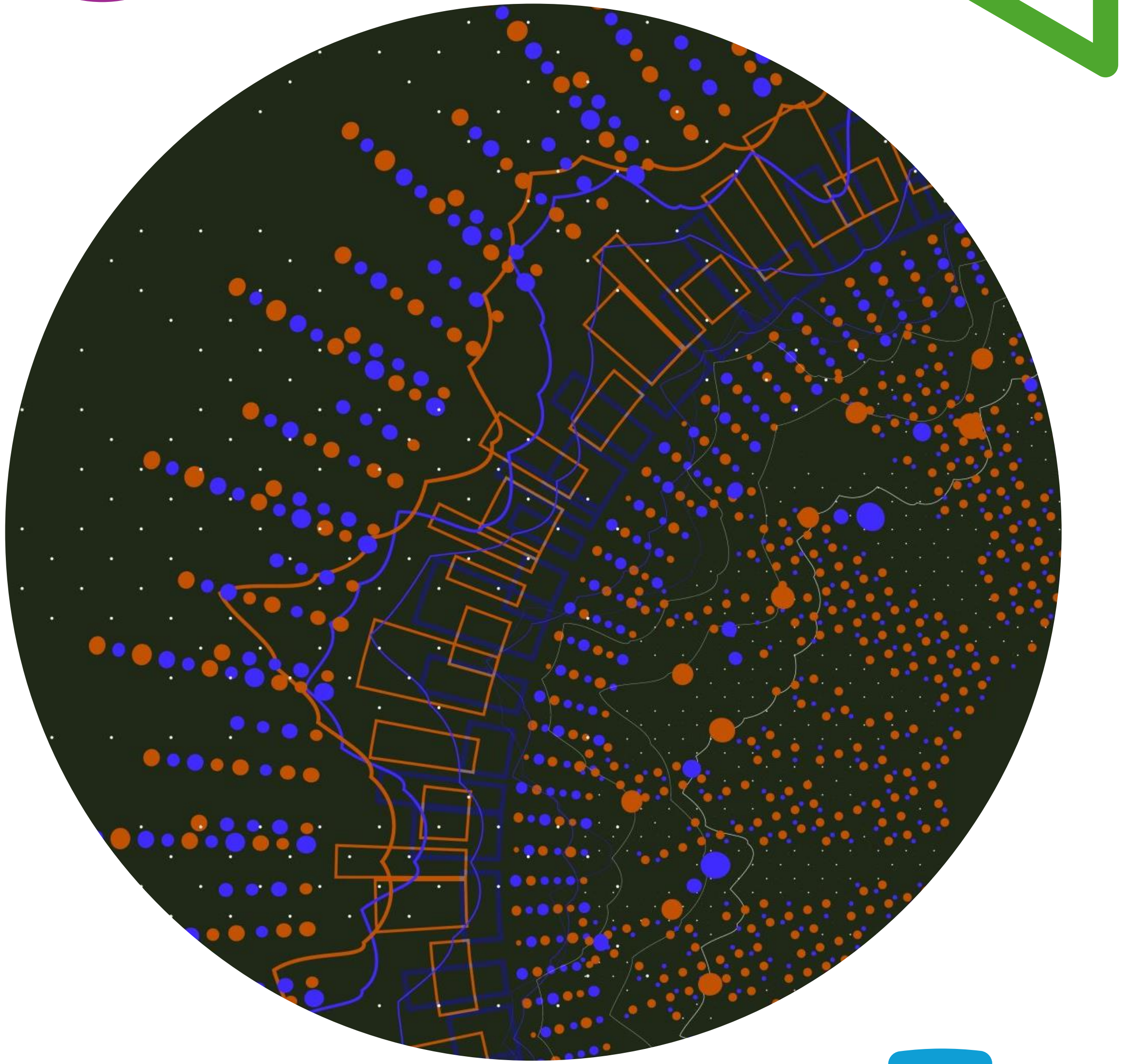


Forecasting Unit Trust Sales and Redemption

Leveraging on Data Science

25th July 2024

AIML Knowledge Sharing

About Me

Aswadi Abdul Rahman

aswadi.abdulrahman@gmail.com

https://www.linkedin.com/in/aswadiabdulrahman/

Head of Analytic Department, ASNB

Other Working Experiences:

time

CIMB

AXIATA DIGITAL

axiata

celcom

HUAWEI

slido

Please download and install the
Slido app on all computers you use

How familiar are you with unit trust sales and redemption?

ⓘ Start presenting to display the poll results on this slide.

slido

Please download and install the Slido app on all computers you use

How would you rate your proficiency with AI and machine learning tools/method?

ⓘ Start presenting to display the poll results on this slide.

Why we need to have unit trust forecast

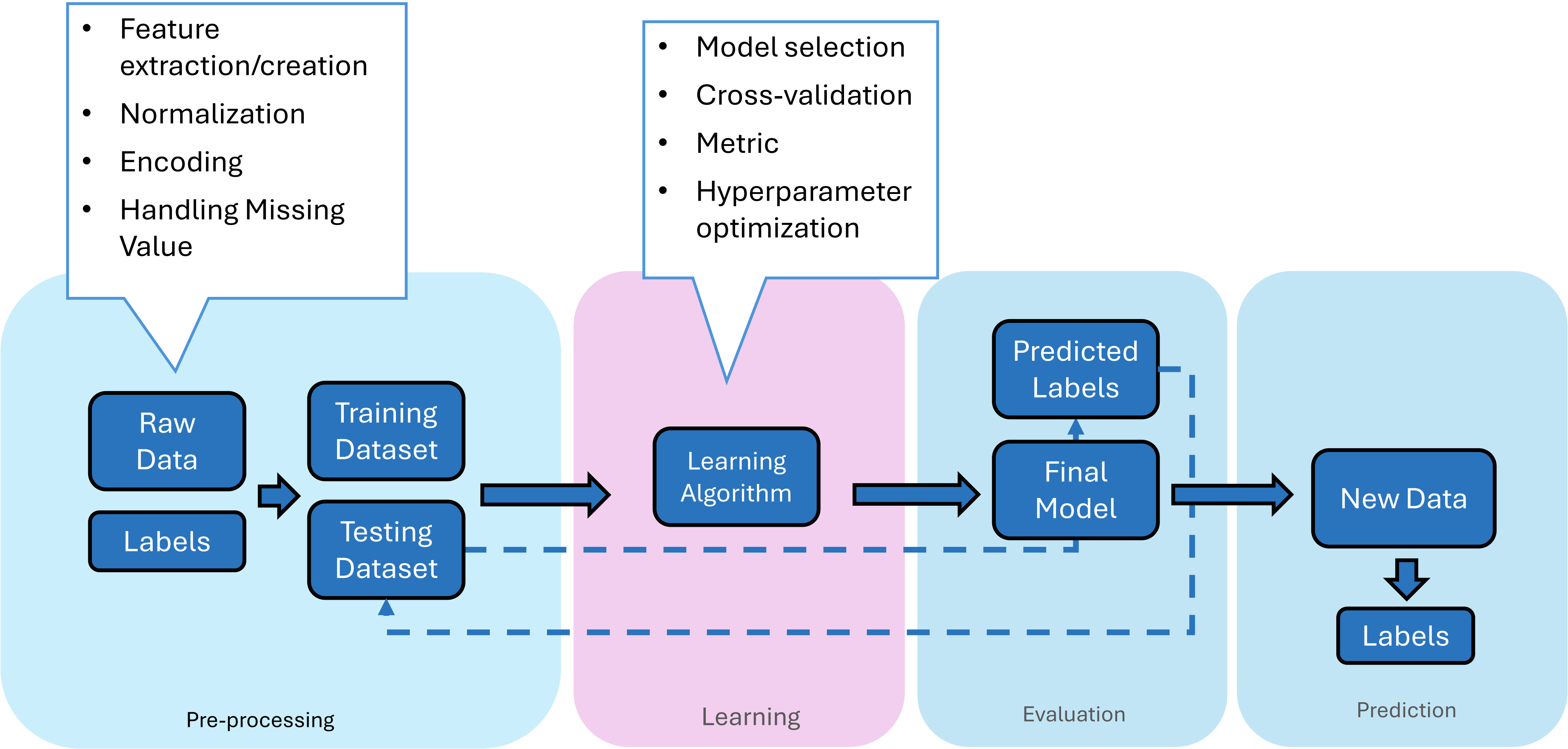
Enhanced Strategic Planning: Prepare for potential sales or redemption trends for next year.

Investment Decision: Enable fund managers to manage effectively.

Risk Management: Identify market trends and risks, allowing proactive measures.

Product Management: Ensure proper allocation of investment products.

Flow Chart of Predictive Modeling

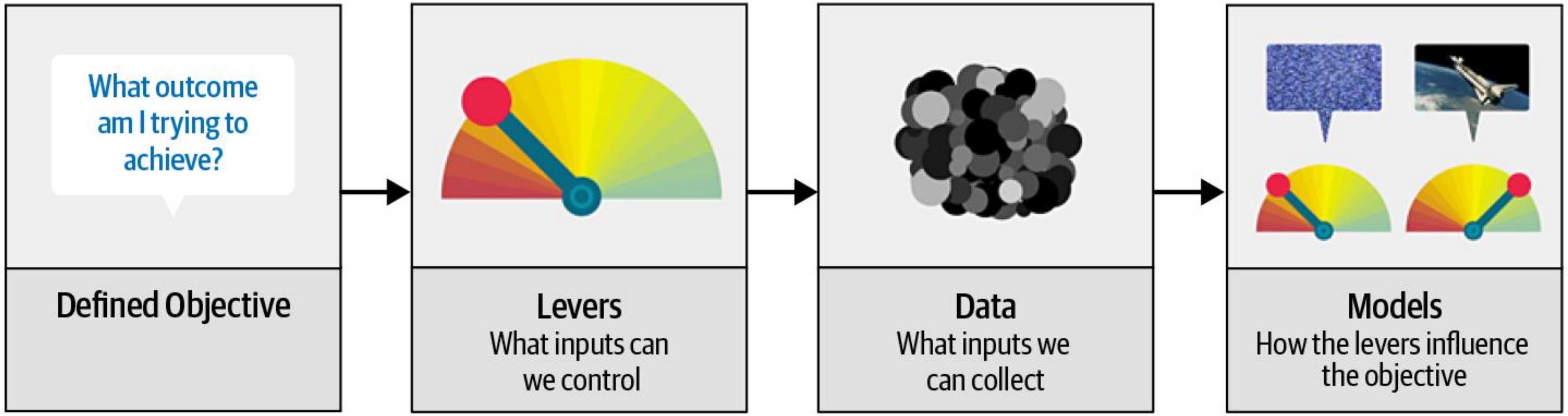


- Feature extraction/creation
- Normalization
- Encoding
- Handling Missing Value

- Model selection
- Cross-validation
- Metric
- Hyperparameter optimization

Details on the Pre-processing

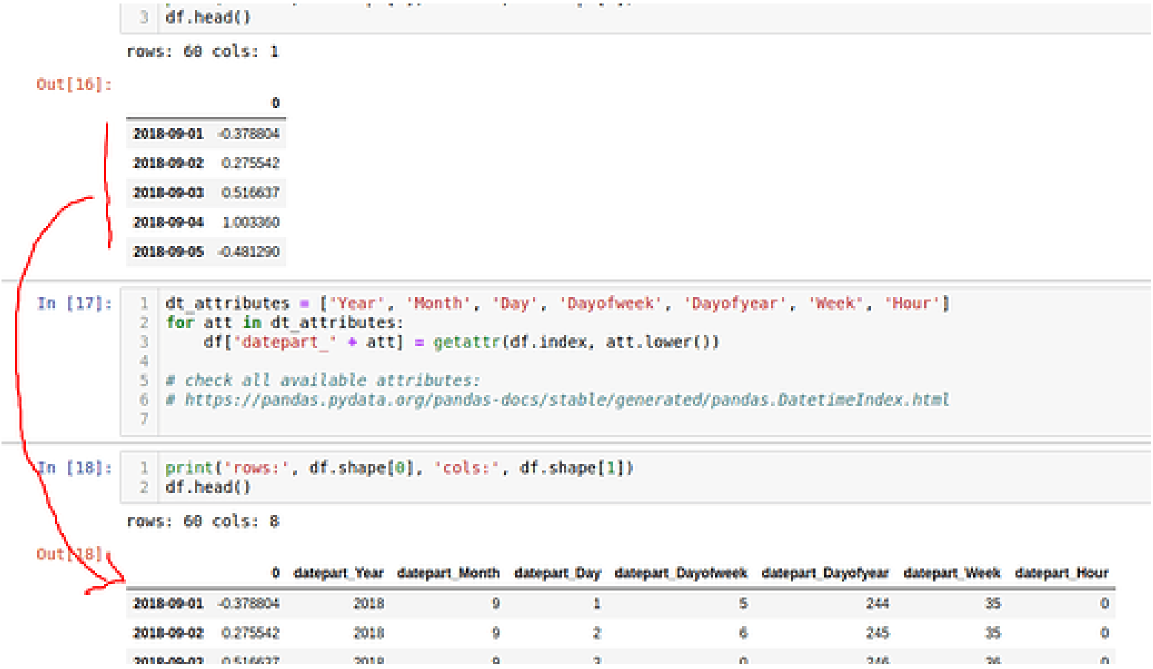
- **Feature Selection:**
  - Training: Include all relevant features you can think of.
  - Forecast: Consider features you might have access to in the future.


What outcome am I trying to achieve?
Defined Objective
Levers
What inputs can we control
Data
What inputs we can collect
Models
How the levers influence the objective

“The Drivetrain Approach” - Designing Great Data Products (Jeremy, Margit, Mike)

Details on the Pre-processing

- **Feature Extraction:**
  - Break down dates into components like year, month, and day.
  - Understanding features that having the same correlation (multi-collinearity) using rank correlation



```
df.head()
```

```
rows: 60 cols: 1
```

Out[16]:

	0
2018-09-01	-0.378804
2018-09-02	0.275542
2018-09-03	0.516637
2018-09-04	1.003360
2018-09-05	-0.481290

In [17]:

```
dt_attributes = ['Year', 'Month', 'Day', 'Dayofweek', 'Dayofyear', 'Week', 'Hour']
for att in dt_attributes:
    df['datepart_' + att] = getattr(df.index, att.lower())

# check all available attributes:
# https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DatetimeIndex.html

```

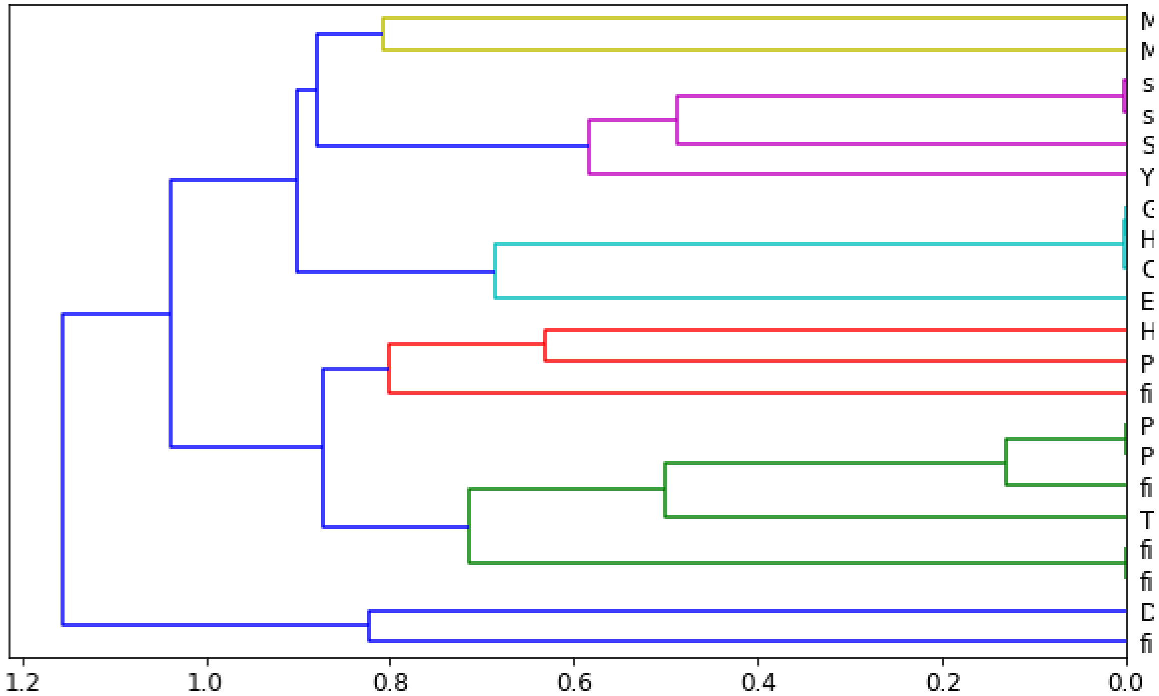
In [18]:

```
print('rows:', df.shape[0], 'cols:', df.shape[1])
df.head()
```

```
rows: 60 cols: 8
```

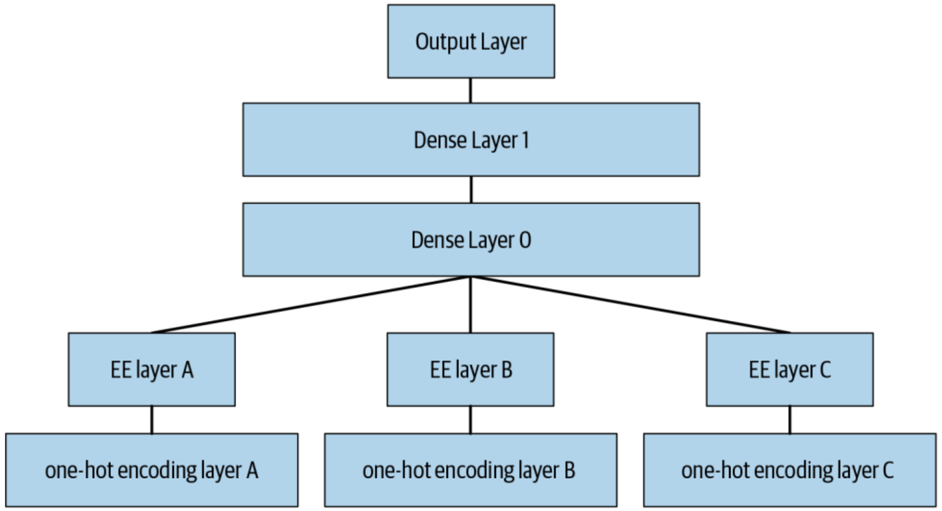
Out[18]:

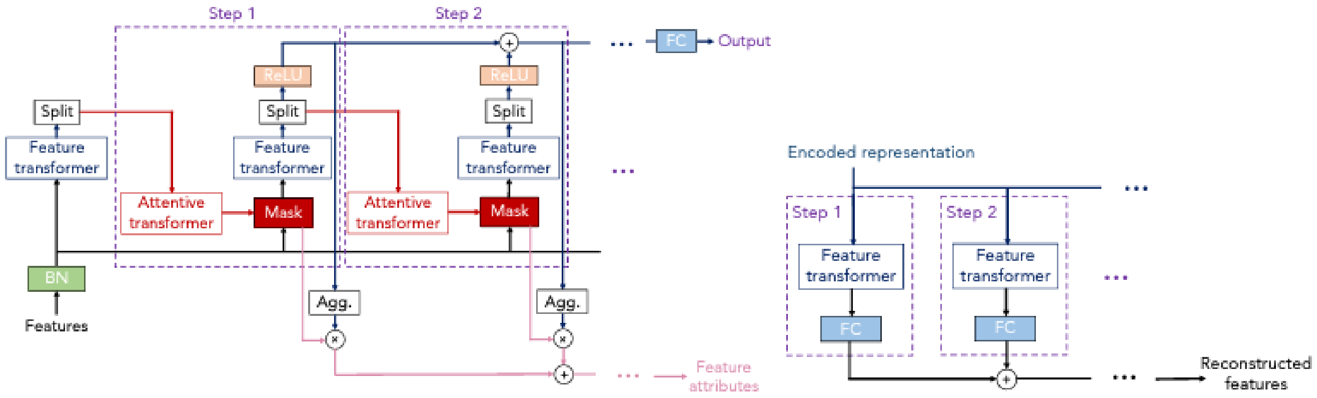
	0	datepart_Year	datepart_Month	datepart_Day	datepart_Dayofweek	datepart_Dayofyear	datepart_Week	datepart_Hour
2018-09-01	-0.378804	2018	9	1	5	244	35	0
2018-09-02	0.275542	2018	9	2	6	245	35	0
2018-09-03	0.516637	2018	9	3	0	246	36	0



Details on the Pre-processing

- **Encoding:**
  - Primarily use label encoding.
  - Occasionally use embedding techniques like entity embedding or unsupervised like TabNet.





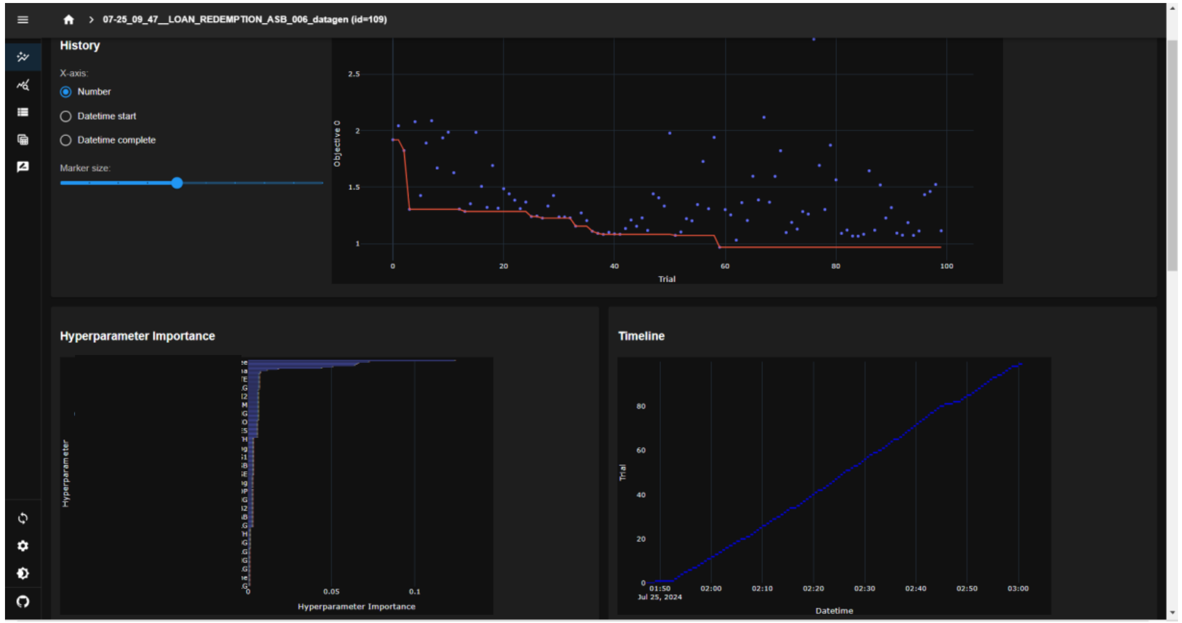
Details on the Pre-processing

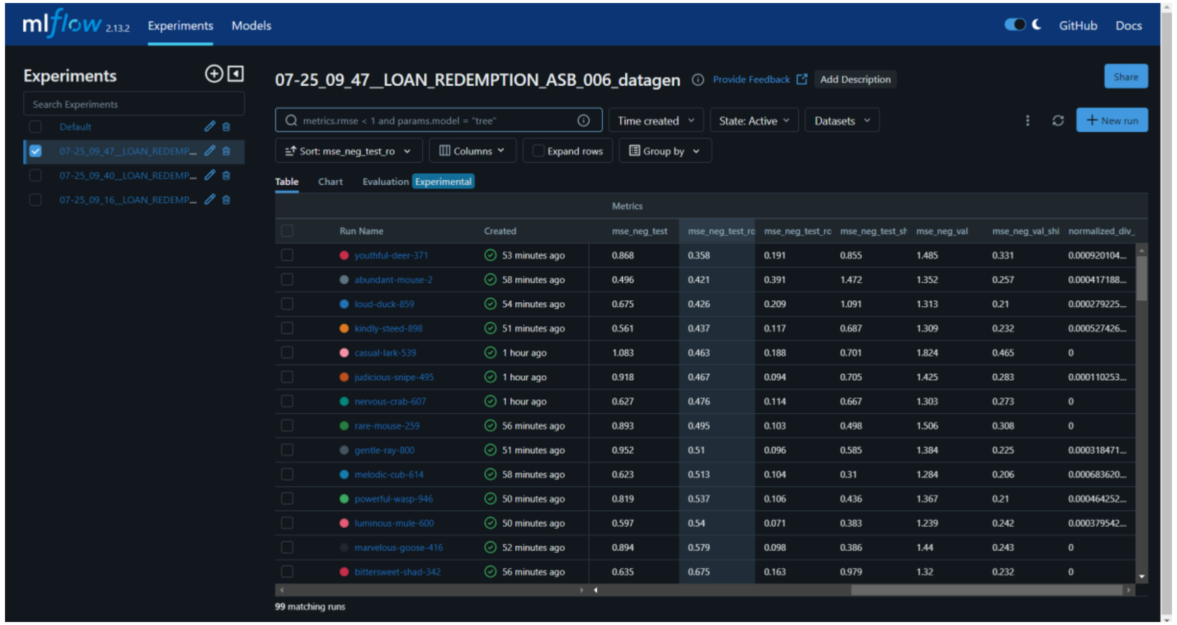
- **Normalization:**
  - Typically, use the log1p transformation for normalization.

- **Handling Missing Values:**
  - Create new features to label missing values.
  - Usually, fill missing values with -1.

Details on the Learning

- **Hyperparameter optimization:**
  - In the beginning we use grid search for random forest.
  - Now we use Optuna using Bayesian optimization on XGBoost.
  - Then store all the experiment in ML Flow





07-25_09_47__LOAN_REDEMPTION_ASB_006_datagen

metrics.rmse < 1 and params.model = "tree"

Sort: mse_neg_test_ro

Metrics

	Run Name	Created	mse_neg_test	mse_neg_test_ro	mse_neg_test_ro	mse_neg_test_sh	mse_neg_val	mse_neg_val_shi	normalized_div_
□	youthful-deer-371	53 minutes ago	0.868	0.358	0.191	0.855	1.485	0.331	0.000920104...
□	abundant-mouse-2	58 minutes ago	0.496	0.421	0.391	1.472	1.352	0.257	0.000417188...
□	loud-duck-859	54 minutes ago	0.675	0.426	0.209	1.091	1.313	0.21	0.000279225...
□	kindly-steed-898	51 minutes ago	0.561	0.437	0.117	0.687	1.309	0.232	0.000527426...
□	casual-lark-539	1 hour ago	1.083	0.463	0.188	0.701	1.824	0.465	0
□	judicious-snipe-495	1 hour ago	0.918	0.467	0.094	0.705	1.425	0.283	0.000110253...
□	nervous-crab-607	1 hour ago	0.627	0.476	0.114	0.667	1.303	0.273	0
□	rare-mouse-259	56 minutes ago	0.893	0.495	0.103	0.498	1.506	0.308	0
□	gentle-ray-800	51 minutes ago	0.952	0.51	0.096	0.585	1.384	0.225	0.000318471...
□	melodic-cub-614	58 minutes ago	0.623	0.513	0.104	0.31	1.284	0.206	0.000683620...
□	powerful-wasp-946	50 minutes ago	0.819	0.537	0.106	0.436	1.367	0.21	0.000464252...
□	luminous-mule-600	50 minutes ago	0.597	0.54	0.071	0.383	1.239	0.242	0.000379542...
□	marvelous-goose-416	52 minutes ago	0.894	0.579	0.098	0.386	1.44	0.243	0
□	bittersweet-shad-342	56 minutes ago	0.635	0.675	0.163	0.979	1.32	0.232	0

99 matching runs

Details on the Learning

- **Evaluation Metric:**

Mean Squared Error (MSE):

Definition: Measures the average squared difference between predicted and actual values.

Use for Subscription forecast

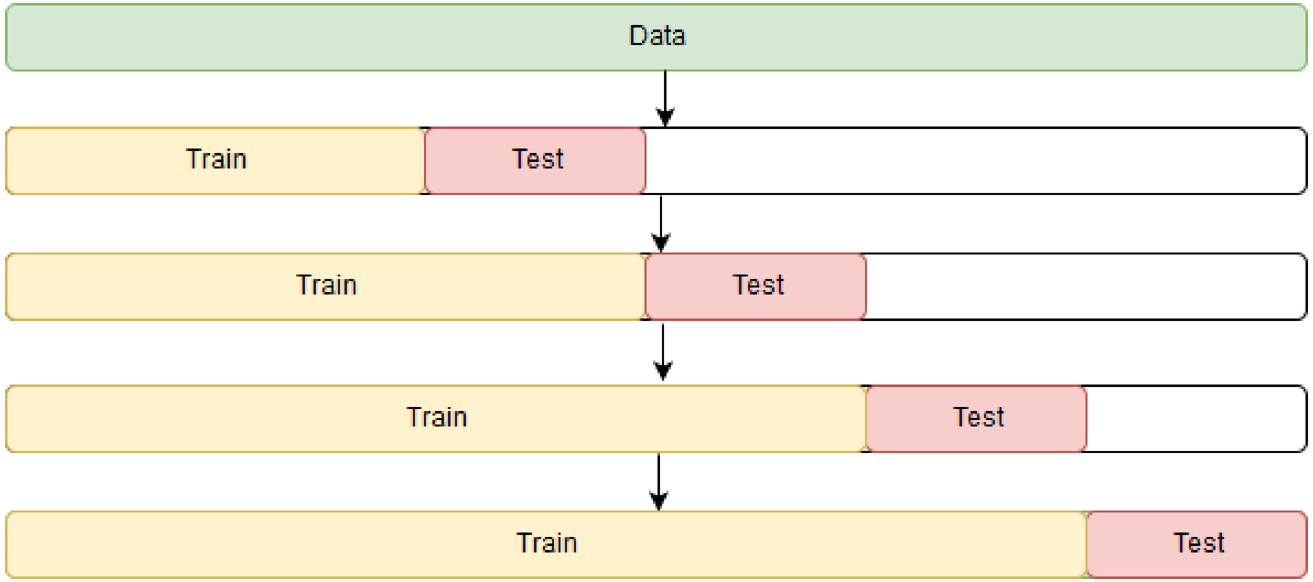
Mean Squared Error (MSE) penalized under Predict:

Definition: Measures the average squared difference between predicted and actual with penalized errors under zero.

Use for Redemption forecast

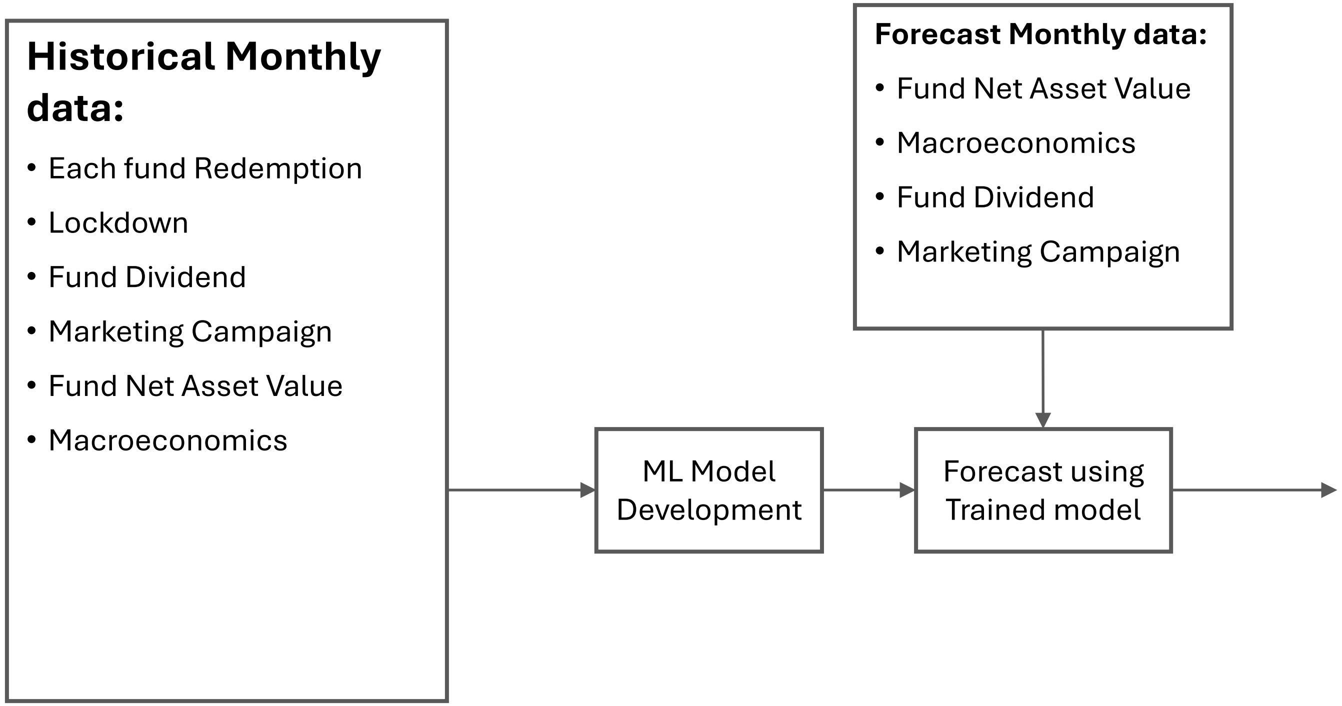
Details on the Learning

- **Model selection:** Our baseline is random forest, then move up to gradient boost tree.
- **Cross-validation:** K-fold rolling windowing



Details on the Learning

By forecasting fund sales and redemptions, we can identify potential risks and take steps to mitigate them.



Historical Monthly data:

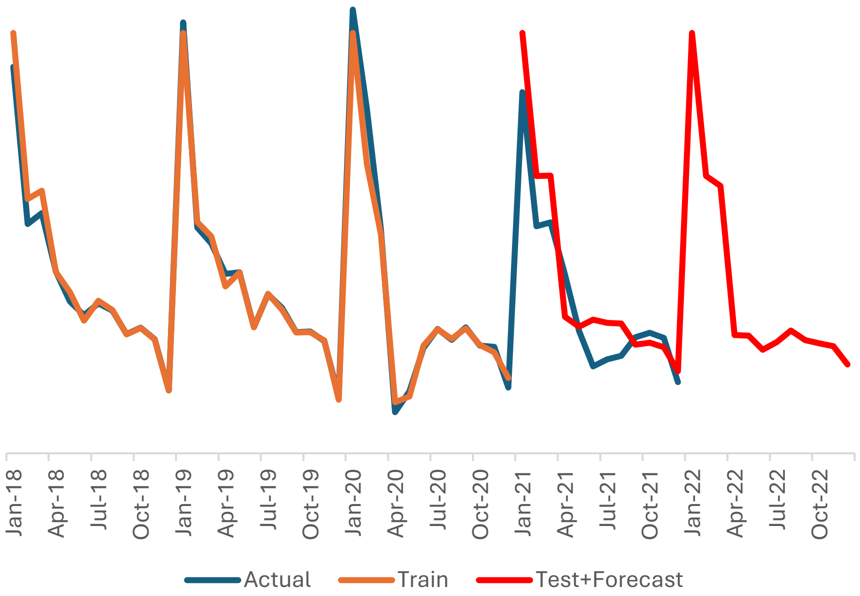
- Each fund Redemption
- Lockdown
- Fund Dividend
- Marketing Campaign
- Fund Net Asset Value
- Macroeconomics

ML Model Development

Forecast Monthly data:

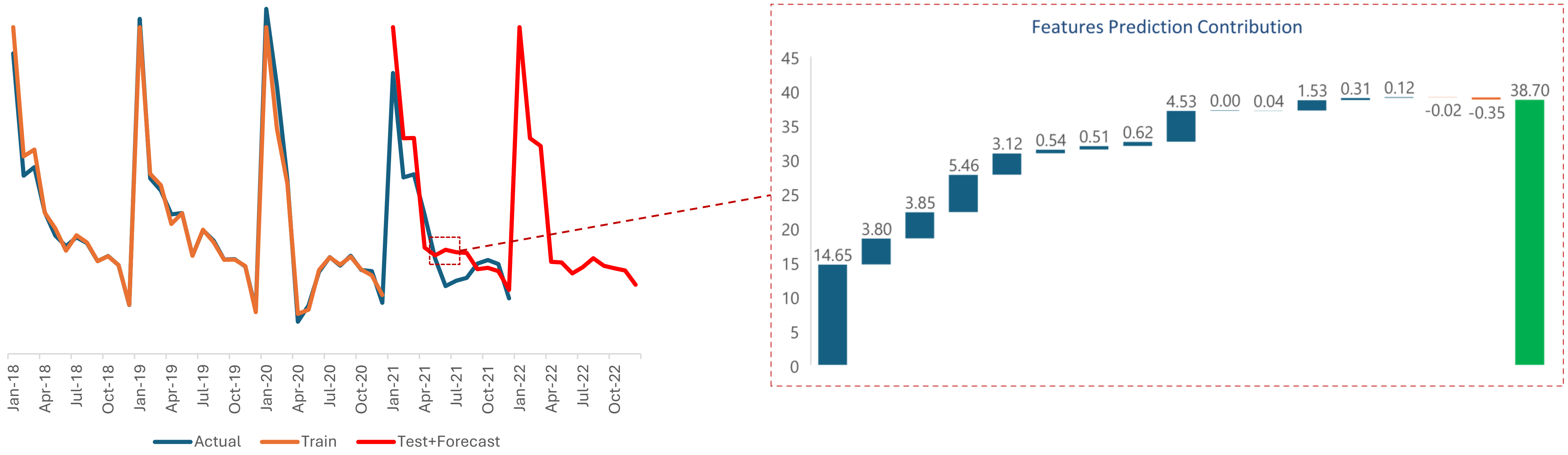
- Fund Net Asset Value
- Macroeconomics
- Fund Dividend
- Marketing Campaign

Forecast using Trained model



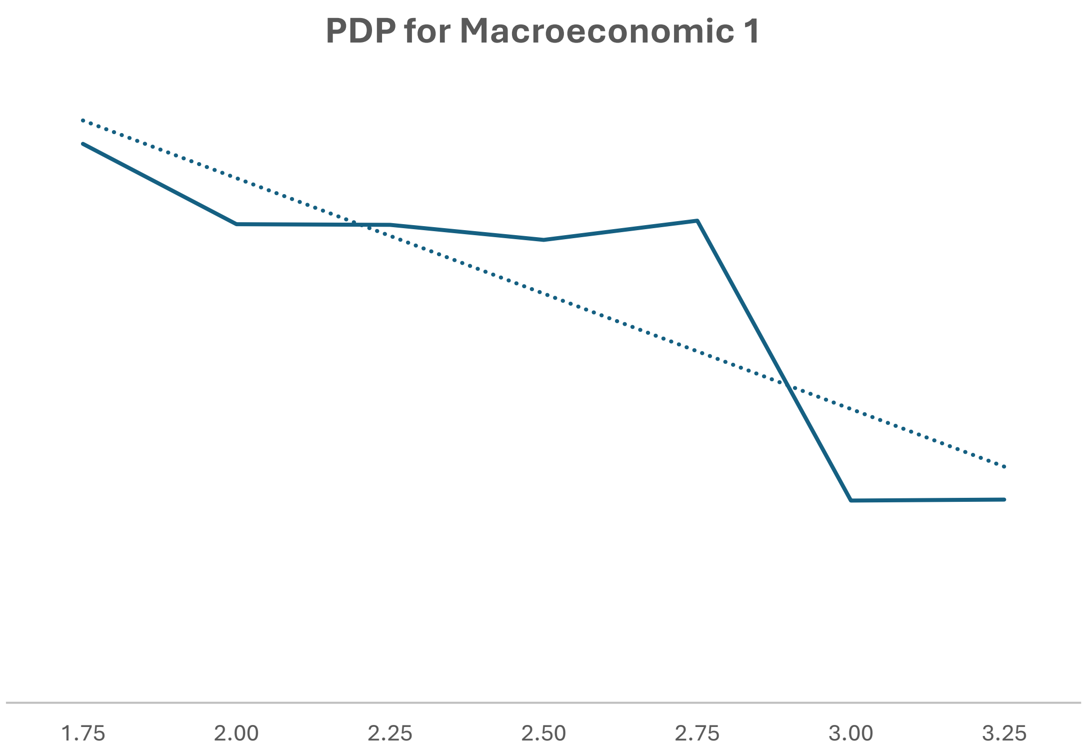
What matters for others?

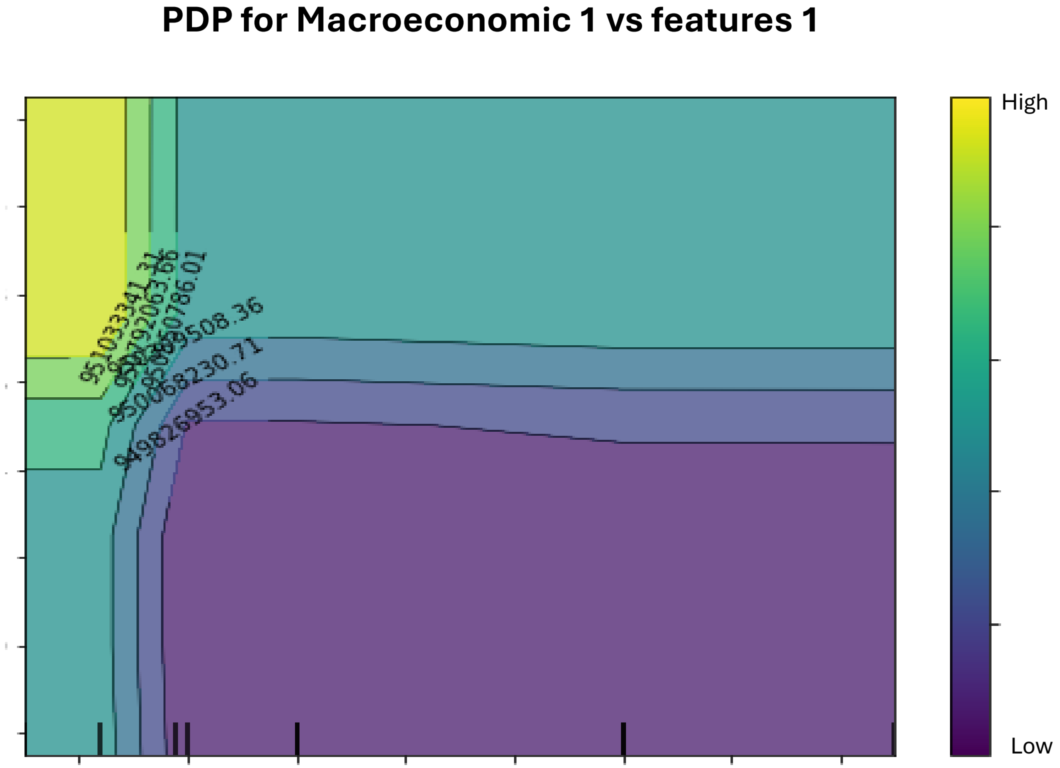
- Each prediction value was determined by its trained features and weight which using interpreter (tree interpreter)
- What happen on each forecasted value? Discuss with subject matter expert



What matters for others?

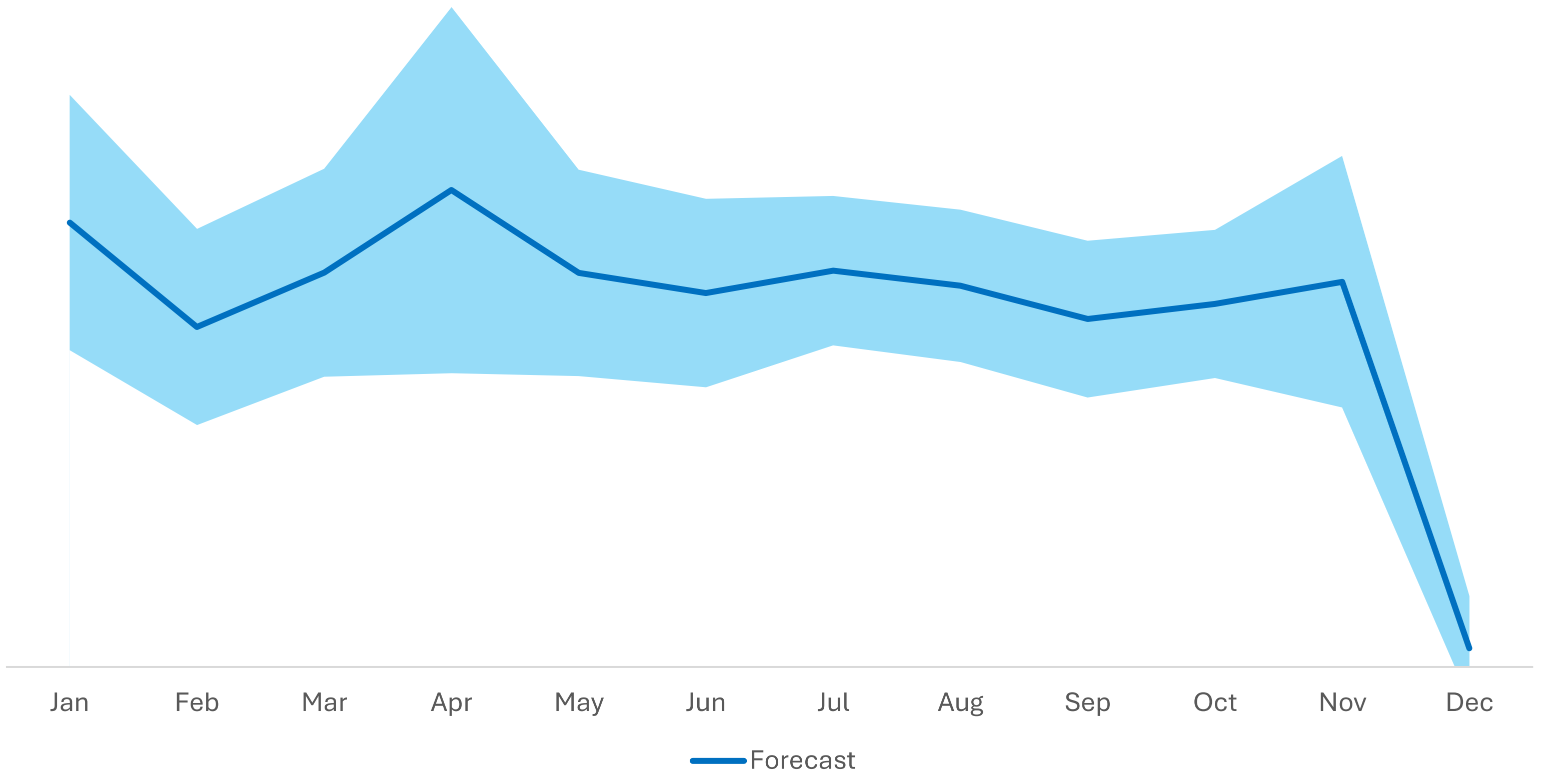
By training the data using the ML models, other than forecast we will also extract the sub output which are features what-if analysis using partial dependent plot (PDP).

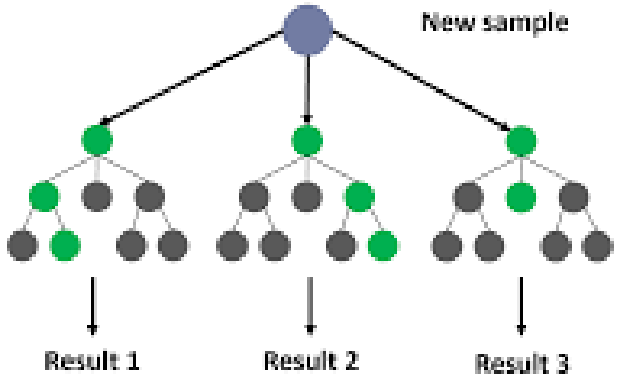

PDP for Macroeconomic 1


PDP for Macroeconomic 1 vs features 1

What matters for others?

Forecast with confidence interval will have affect on their decision. We are using standard deviation of the prediction tree.





Conformal Prediction?

Forecasting Challenges

- **Retrain with new data:**
  - To avoid data drift
  - NEVER train the whole dataset for forecast.
  - Need to consider future data availability especially external data.
- **Strategic consideration:**
  - Some fund that does not have enough data
  - And new initiatives should be done outside the model.
- **Future accuracy expectation:** Find stability instead of accuracy
- **Retrain with new method:** Only when your model is not stable & outside of the confidence interval.

Other Use Cases

- Unsupervised: Topic Modelling Clustering
- Predicting Customer Unitholding
- AI Chatbot – Leverage on OpenAI LLM
- AI Avatar Chatbot
- Lookalike model for Customer Segment
- Instant Analysis using LLM
- AMLA Transaction Classification

We are hiring!!

AMLA Data Analytic Manager

Experience Needed

- >5 years of total working experiences
- >3 years on Compliance and AMLA fields
- >3 years hands on working on ML models in python and ETL

What will you do

- Supervised 2 data analytics/ data science
- Working together with compliances department and security commission for relevant guidelines

Email to aswadi@pnb.com.my

Thank You