How to Apply Machine Learning for Forecasting & Anomaly Detection At Scale?

Ella Hilal

Head of Data Science and Engineering, Revenue and Growth, Shopify.

@a_hilal
@ShopifyData

shopify

At a Glance

Shopify is a provider of essential internet infrastructure for commerce.

2006
Platform Released

$4.2B
Revenue
(Last 12 months)

7,000+
Employees
(December 31, 2020)

>1.7 Million
Merchants
— in —
>175 Countries

~$400B
Total sales by merchants on Shopify (Cumulative as of beginning of October)

4.9 million
Merchant Staff Accounts
(December 31, 2020)

>43,000
Partners who have referred at least one merchant to Shopify in the last 12 months
(September 30, 2021)

>7,000
Apps in our App Store
(September 30, 2021)

$230M
Paid out in 2020 to partners by Shopify for apps benefiting our merchants

Commerce: A force for Good

We have a Global Mission

* Captured as of December 31, 2019

$307+ billion

Global economic impact from merchants on Shopify

Together, Shopify merchants would make up the 7th largest company in the world in terms of revenue, above Apple, BP, and Volkswagen.

3.6 million

Jobs created by Shopify merchants

Worldwide, 1 in 1,000 employed adults is directly or indirectly employed by a Shopify merchant.

Collectively, Shopify merchants support the largest workforce in the world.

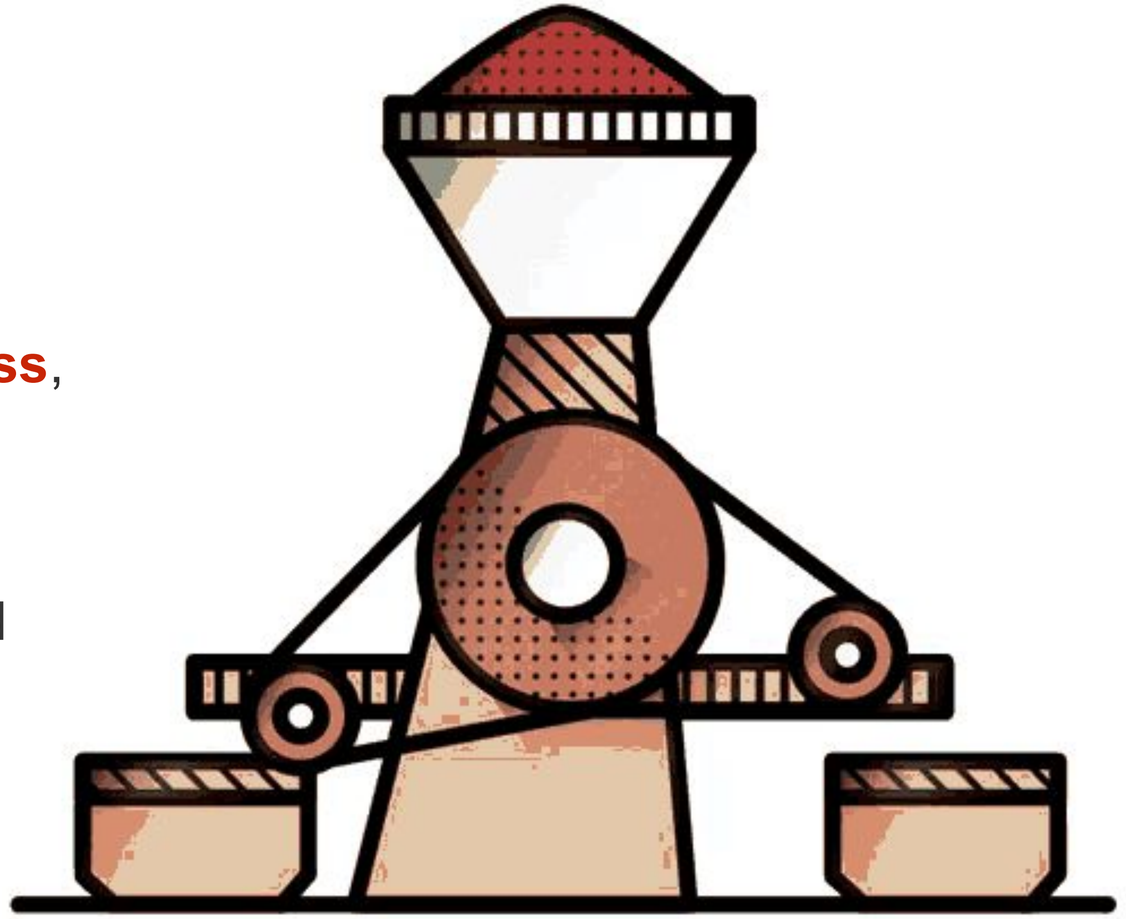
Forecasting

Business Forecasting

Retail is changing.

Businesses are part of an ecosystem that is changing rapidly

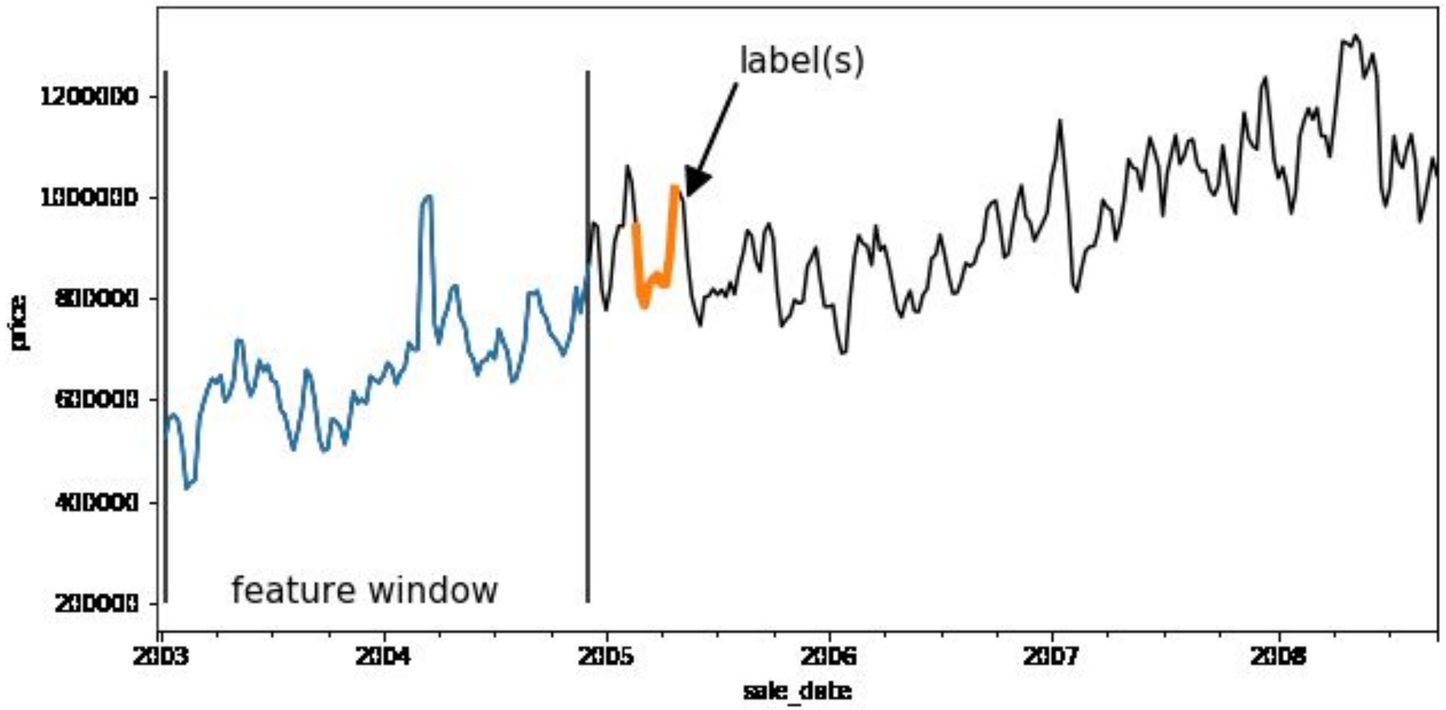
- Business forecasting consists of tools and techniques used to **predict changes in business**, such as sales, expenditures, profits and losses.
- The goal of business forecasting is to **develop better strategies** based on these data-informed predictions
- Financial and operational decisions are made based on current **market conditions** and predictions on **how the future looks**.



Business Forecasting is HARD

Dynamic market conditions

Social and Economical Variability



Underlying biases

Implicit/explicit assumptions, e.g.: Products and services launches or sunsetting

Seasonality

Need to forecast for a long time horizons

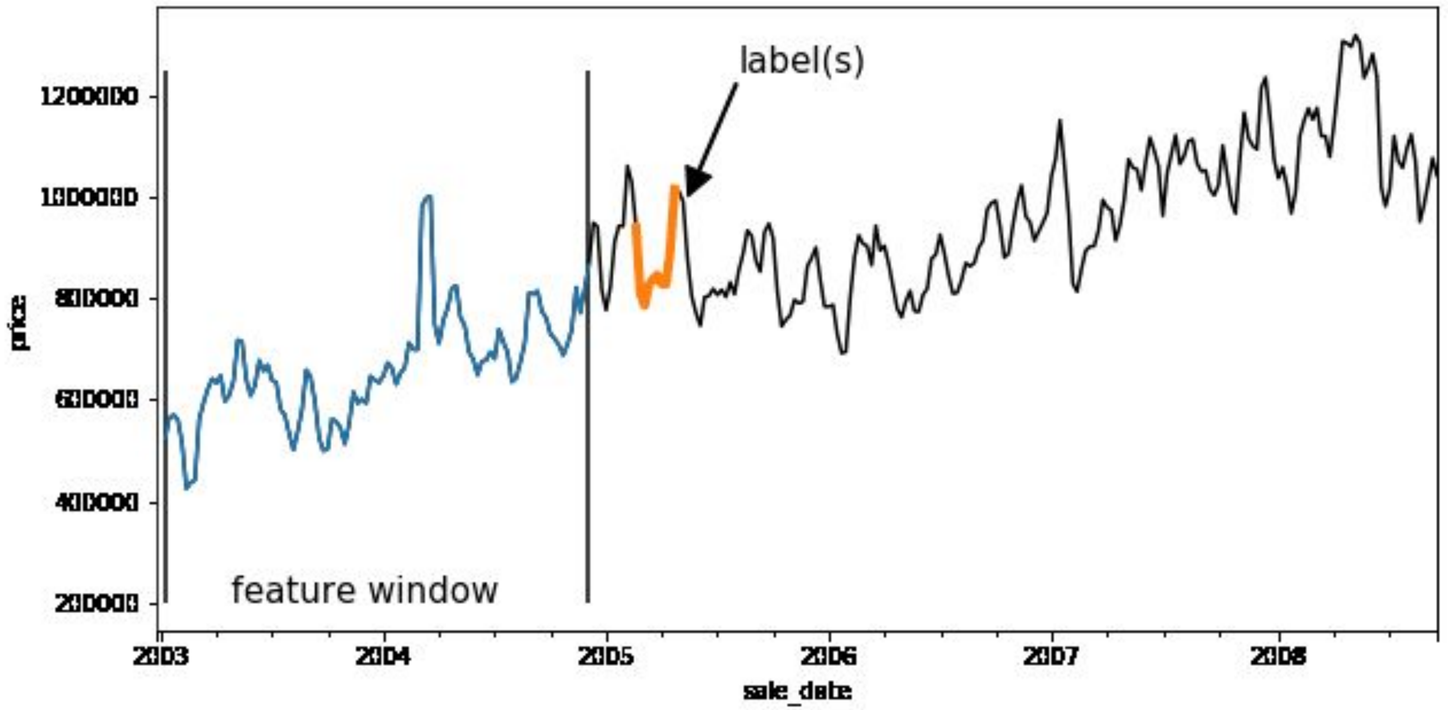
Regulatory changes

2020

Business Forecasting is HARD

Dynamic market conditions

Social and Economical Variability



Underlying biases

Implicit/explicit assumptions, e.g.: Products and services launches or sunsetting

Seasonality

Need to forecast for a long time horizons

2020

Regulatory changes


#SchittsCreek
I'VE WOKEN UP IN A
BLACK MIRROR EPISODE

Business Forecasting *is* not an Easy task

5 key steps to

Apply Machine Learning for Forecasting & Anomaly Detection At Scale

Step 1: Be clear on your Forecasting Needs

Step 1: Be clear on your Forecasting Needs

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Low latency constraints

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Low latency constraints

How far to forecast in the future?

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Low latency constraints

How far to forecast in the future?

Time period to be forecasted

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Low latency constraints

How far to forecast in the future?

Time period to be forecasted

How often are you making this decisions?

Step 1: Be clear on your Forecasting Needs

Have the real talk and do not assume

What decisions will be informed by this information?

Choose the metrics to forecast effectively

When are you making this decisions?

Low latency constraints

How far to forecast in the future?

Time period to be forecasted

How often are you making this decisions?

Daily, weekly, or monthly cadence

Step 1: Be clear on your Forecasting Needs

Have the **real talk** and do not assume

What decisions will be informed by this information?
Choose the metrics to forecast effectively

When are you making this decisions?
Low latency constraints

How far to forecast in the future?
Time period to be forecasted

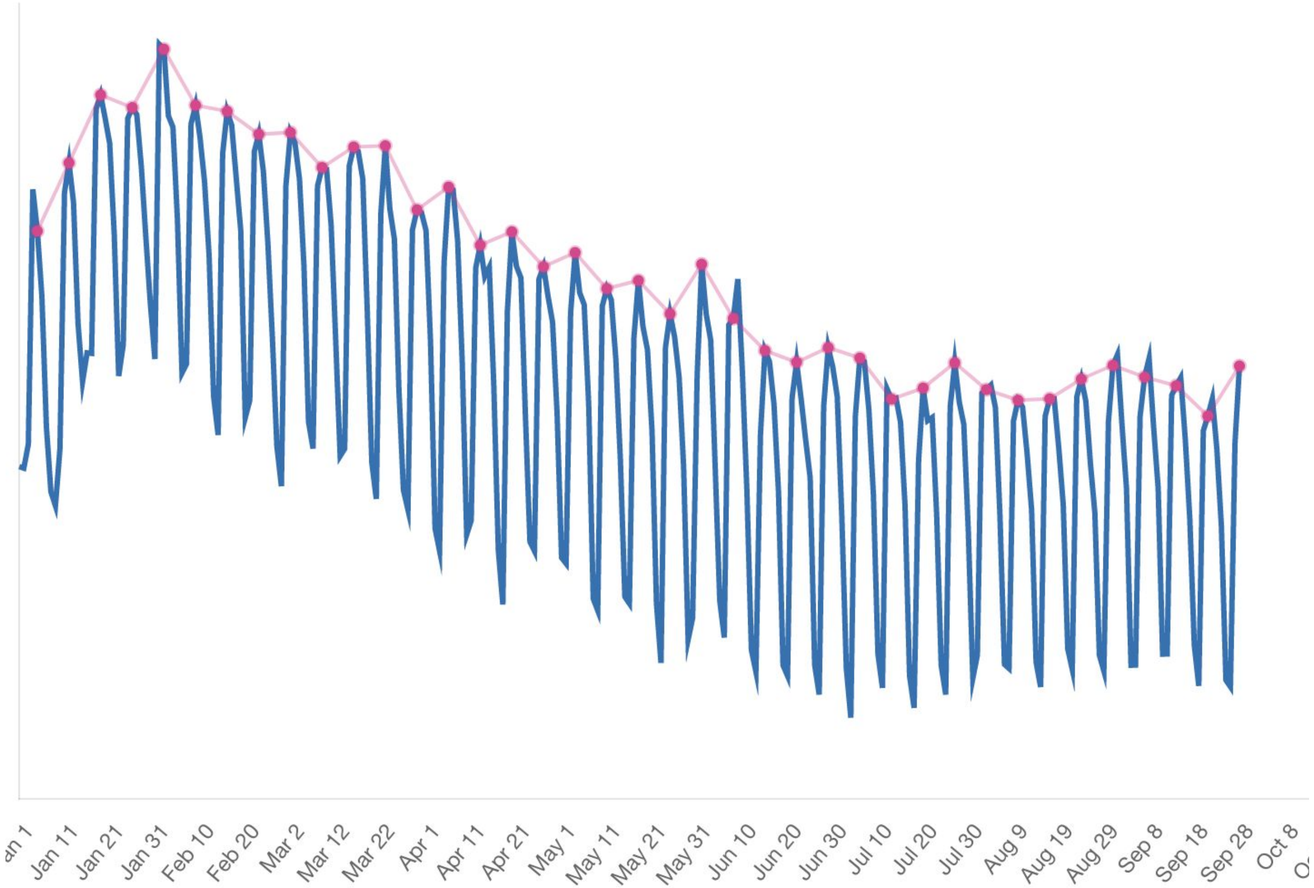
How often are you making this decisions?
Daily, weekly, or monthly cadence

Step 2: Understand your Data

Step 2: Understand your Data

What data is available? and its properties?

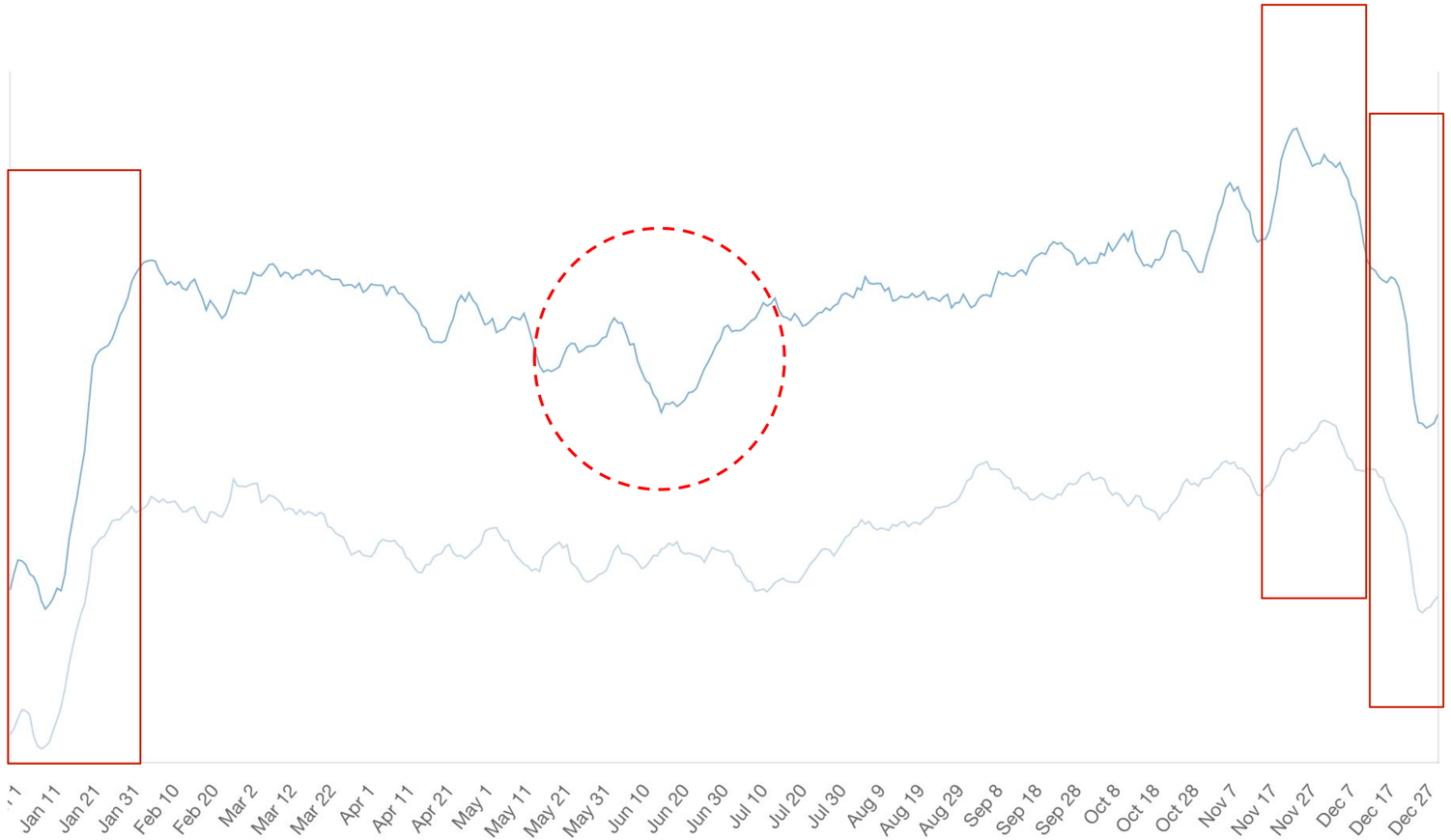
- Understand the properties of the time-series data, e.g.: *univariate or multivariate?*
- Detect and eliminate non-stationary behavior in the time series, e.g.: *moving average*



Step 2: Understand your Data

What data is available? and its properties?

- Seasonality or event versus anomalies and the reasons that drive it



Step 3: Choose the right forecasting method

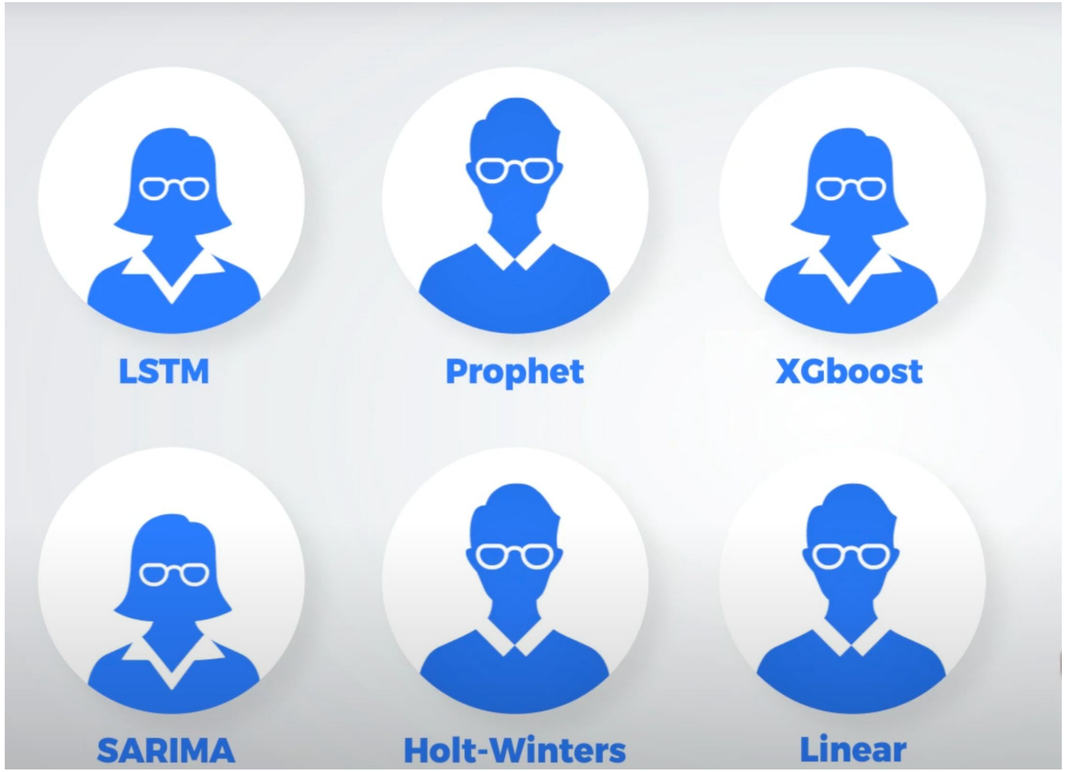
Step 3: Choose the right forecasting method

- Define what do you need from the model.

- To produce effective business forecasts at scale, **a forecasting model must**:
  - Be easy to interpret,
  - Possess the ability to be tuned easily,
  - Be relatively fast,
  - Provide automated forecasts (no manual intervention)
  - Can manage multiple types of seasonality and recurring events
  - Can manage missing observations or large outliers
  - Can manage historical trend changes, for instance due to customer growth

Step 3: Choose the right forecasting method

- Machine Learning can provide effective forecasting at a pretty large scale, and can work effectively on time series data with varying properties.



Step 3: Choose the right forecasting method

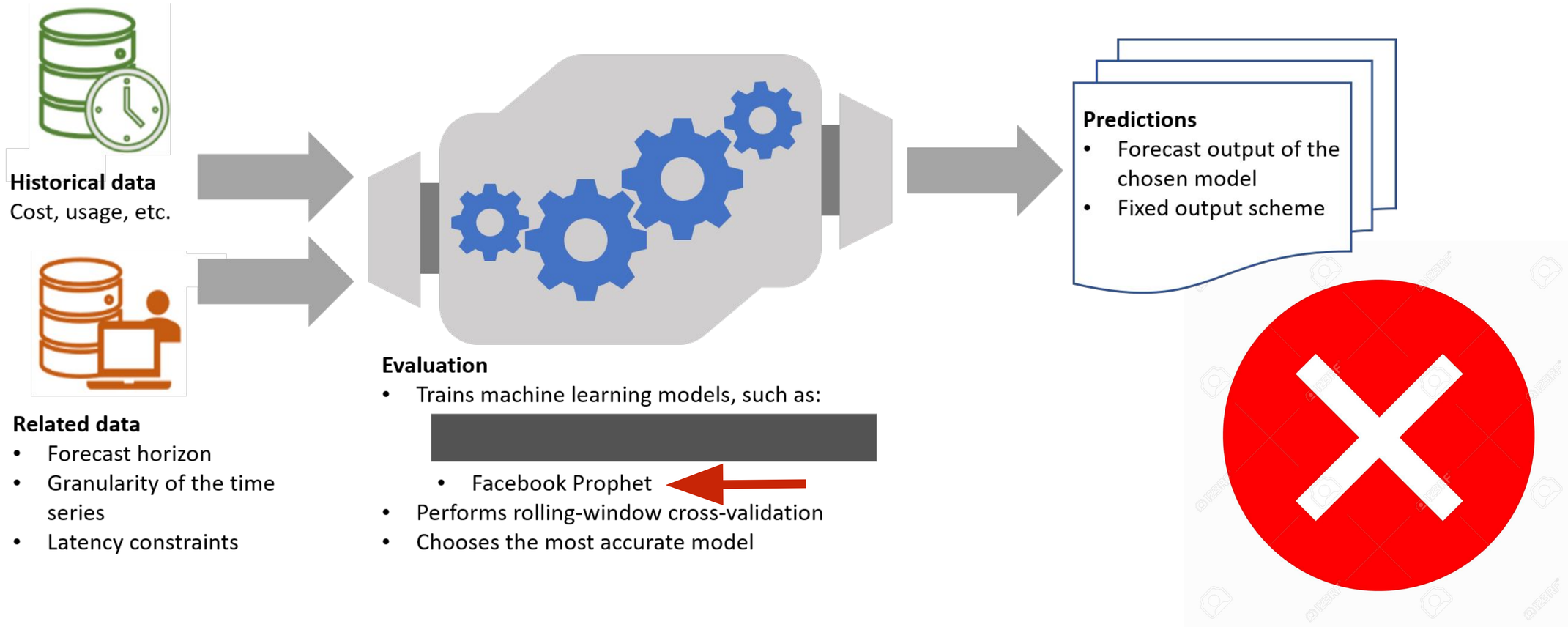
- For our use case, we selected to start with the **Facebook Prophet**,
  - Prophet is an additive regression model
  - It is extremely scalable and is able to generate forecasts quickly over millions of data points is key.
  - Ease to mark periods of exception (known as “holidays”)

If we check our list of what do you need from the model:

✅ Be easy to interpret,
✅ Possess the ability to be tuned easily,
✅ Be relatively fast,
✅ Provide automated forecasts (no manual intervention)
✅ Can manage multiple types of seasonality and recurring events
✅ Can manage missing observations or large outliers
✅ Can manage historical trend changes, for instance due to customer growth

Step 3: Choose the right forecasting method

- Applied Prophet to Forecast at the aggregate topline metric trends
- We initially manage 2020 as an anomalies year and marked it as a period of exception



Step 3: Choose the right forecasting method

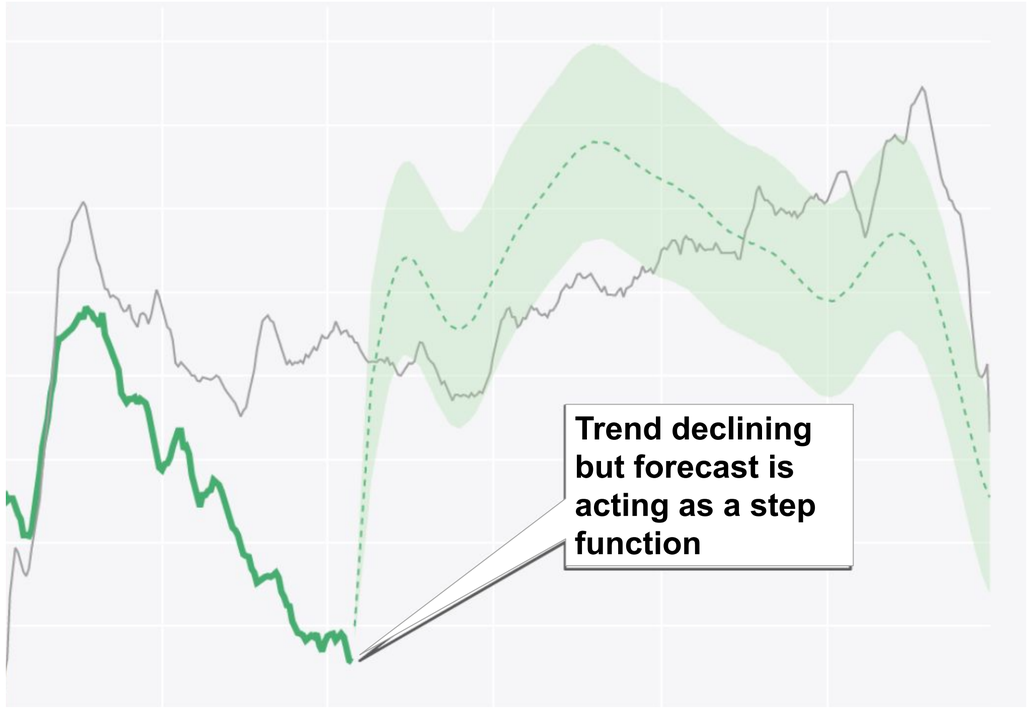
Decisions that we missed:

- What is the degree of accuracy desirable?
- What is the granularity needed to better understand the forecast?

➔ Top down approach or bottom-up?
➔ Single model or ensemble?
➔ How often do you plan to retrain your model?
➔ Did you spend enough time tuning? Did you overfit?

Step 3: Choose the right forecasting method

Bottom-up	Gives better visibility on the drivers of the Forecast
Single model	Keeping it simple first
Retrained daily	Adjusting your forecasts based on actual results
Spend time tuning and avoided overfitting	“Yes” ...



Step 4: Manage Anomaly Bias

Step 4: Manage Anomaly Bias

Anomalies is not always a bad thing.

If effectively explained and maybe recurring, you many want to *Amplify* the anomaly to learn from it. If not, you many want to under weigh it in your learning.

Enforcing a *Linear* trend instead of logarithmic corrected for anomalous trend in the previous year and resulted in a forecast that was a continuation of recent trends.

What timeframe is most informative of future behaviour

We understood that although we have clear annual cycles and annual seasonality, the *last 3 months* trend is very indicative for the next performance.

2020 is an anomalous year but we also needed to account for *2021's unique trends*

Using external data to augment the context

To inform the model with current data, you can incorporate other *external signals* that are drivers of the future trend.

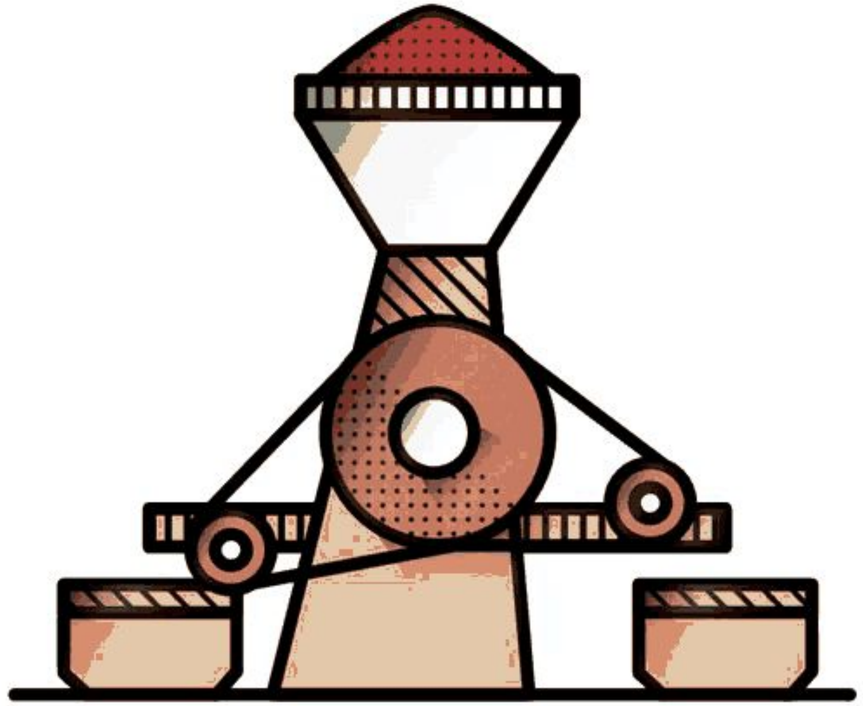
Step 5: Productionize the Model

Step 5: Productionize the Model

Similar to any machine learning model in production, the forecasting is effective when it is fully automated.

This entails:

1. How is your training data stored? raw or transformed?
2. How will you retrieve the data for training?
3. How will you retrieve data for prediction?
4. How do we get feedback from a model in production? Any systematic alerts?
5. How are you tracking drift?



Voila

Step 1: Be clear on your Forecasting Needs

Step 2: Understand your Data

Step 3: Choose the right forecasting method

Step 4: Manage Anomaly Bias

Step 5: Productionize the Model

Thanks!

Ella Hilal

Head of Data Science and Engineering, Revenue and Growth, Shopify.

@a_hilal