Индустриальный семинар

Прогноз метрик

Элен Теванян

Руководитель направления алгоритмического анализа, X5 Tech

x5retail.tech


Technical interview
The actual job

Вы не можете управлять тем, что не измеряете

Иерархия метрик

Метрики обучения моделей

Функции, которые оптимизируем математическими методами, для решения какой-то задачи

Иерархия метрик

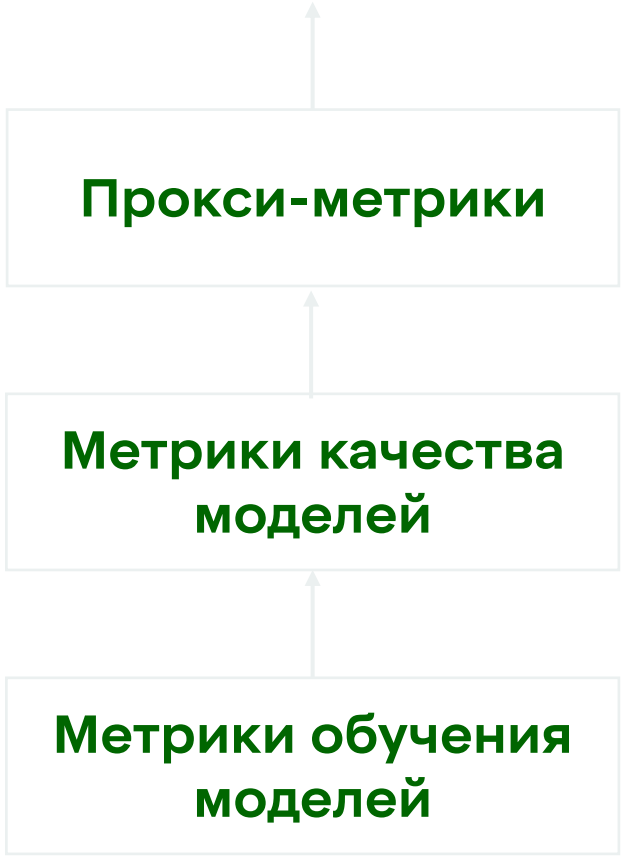
Метрики качества моделей

Метрики для сравнения моделей между собой

Метрики обучения моделей

Функции, которые оптимизируем математическими методами, для решения какой-то задачи

Иерархия метрик



Напрямую связаны с бизнес-метриками или влияют на них. Измеряются быстро.

Метрики для сравнения моделей между собой

Функции, которые оптимизируем математическими методами, для решения какой-то задачи

Иерархия метрик

Бизнес-метрики	Деньги или все, что напрямую связано с деньгами. Измеряются поздно и постфактум.
Прокси-метрики	Напрямую связаны с бизнес-метриками или влияют на них. Измеряются быстро, в пилотах.
Метрики качества моделей	Метрики для сравнения моделей между собой
Метрики обучения моделей	Функции, которые оптимизируем математическими методами, для решения какой-то задачи

Другая классификация

Целевые

Показатели, на которое направлено изменение

Опережающие

Показатели, хорошо коррелируемые с целевой,

Guardrail

Показатели, на которые направленно влияет изменение, но не являющиеся целевыми. Рекомендуется за ними наблюдать и на их основе в том

Другая классификация

Целевые

Средний чек

Опережающие

Добавление товара в корзину

Guardrail

Время от входа в корзину до ее прохождения

Метрики тщеславия

Метрики качества

Иерархия метрик

Бизнес-метрики

Деньги или все, что напрямую связано с деньгами. Измеряются поздно и постфактум.

Прокси-метрики

Напрямую связаны с бизнес-метриками или влияют на них. Измеряются быстро, в пилотах.

Метрики качества моделей

Метрики для сравнения моделей между собой

Метрики обучения моделей

Функции, которые оптимизируем математическими методами, для решения какой-то задачи

Что измеряют в ритейле?

Что измеряют в ритейле?
Примеры из области CVM

- Средний PTO на клиента
- Частота покупок клиента
- LTV
  customer retention rate
- PTO в период действия кампании
- PTO при совершении целевого действия в период действия кампании
- Кол-во чеков в период действия кампании
- Средний чек в период действия кампании
- доп PTO на человека в кампании
- доп PTO по кампании
- чистый отклик в покупку в период действия кампании
- чистый отклик в целевое действие в период действия кампании
- валовый доход по кампании

Как же прогнозируются?

Как же прогнозируются?

- Есть здоровый массив исторический данных

Как же прогнозируются?

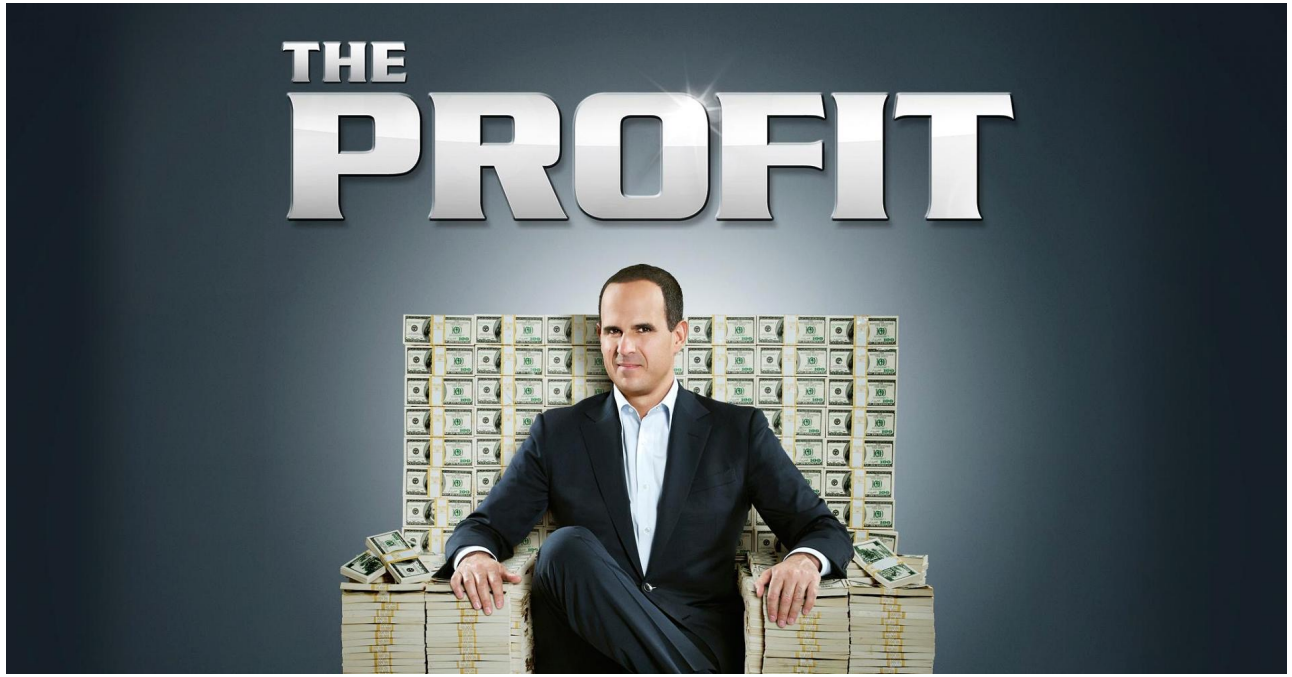
- Есть здоровый массив исторический данных
- Формализуются метрика и методология ее расчета

Как же прогнозируются?

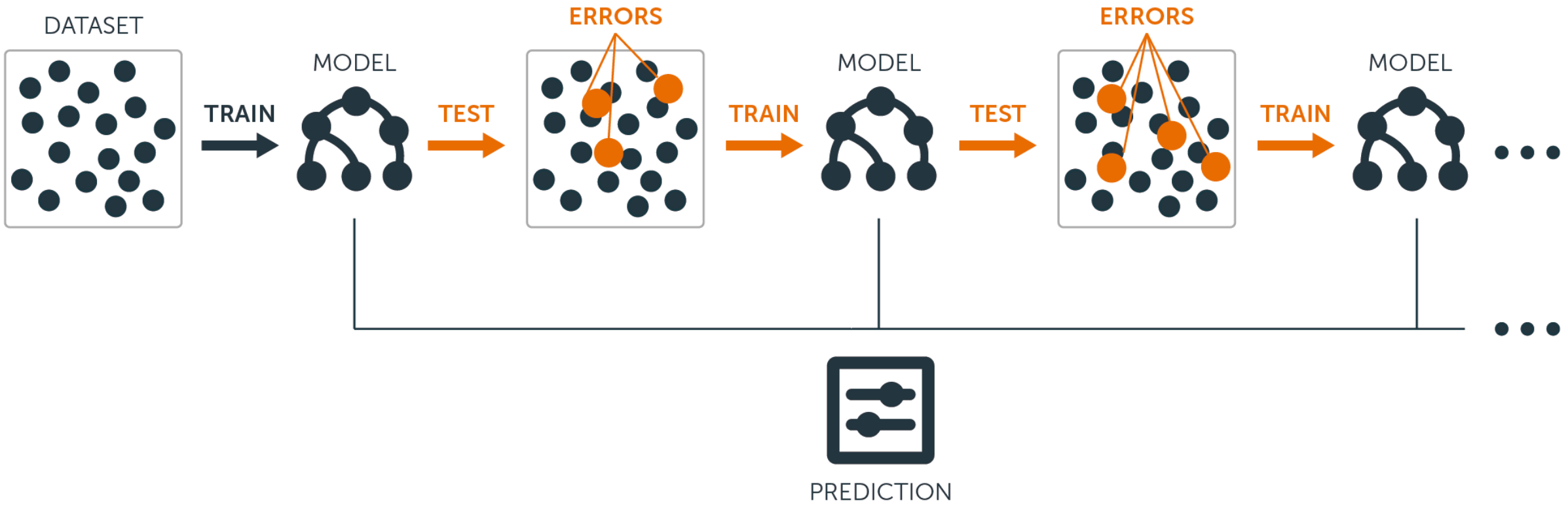
- Есть здоровый массив исторический данных
- Формализуются метрика и методология ее расчета
- Собирается выборка

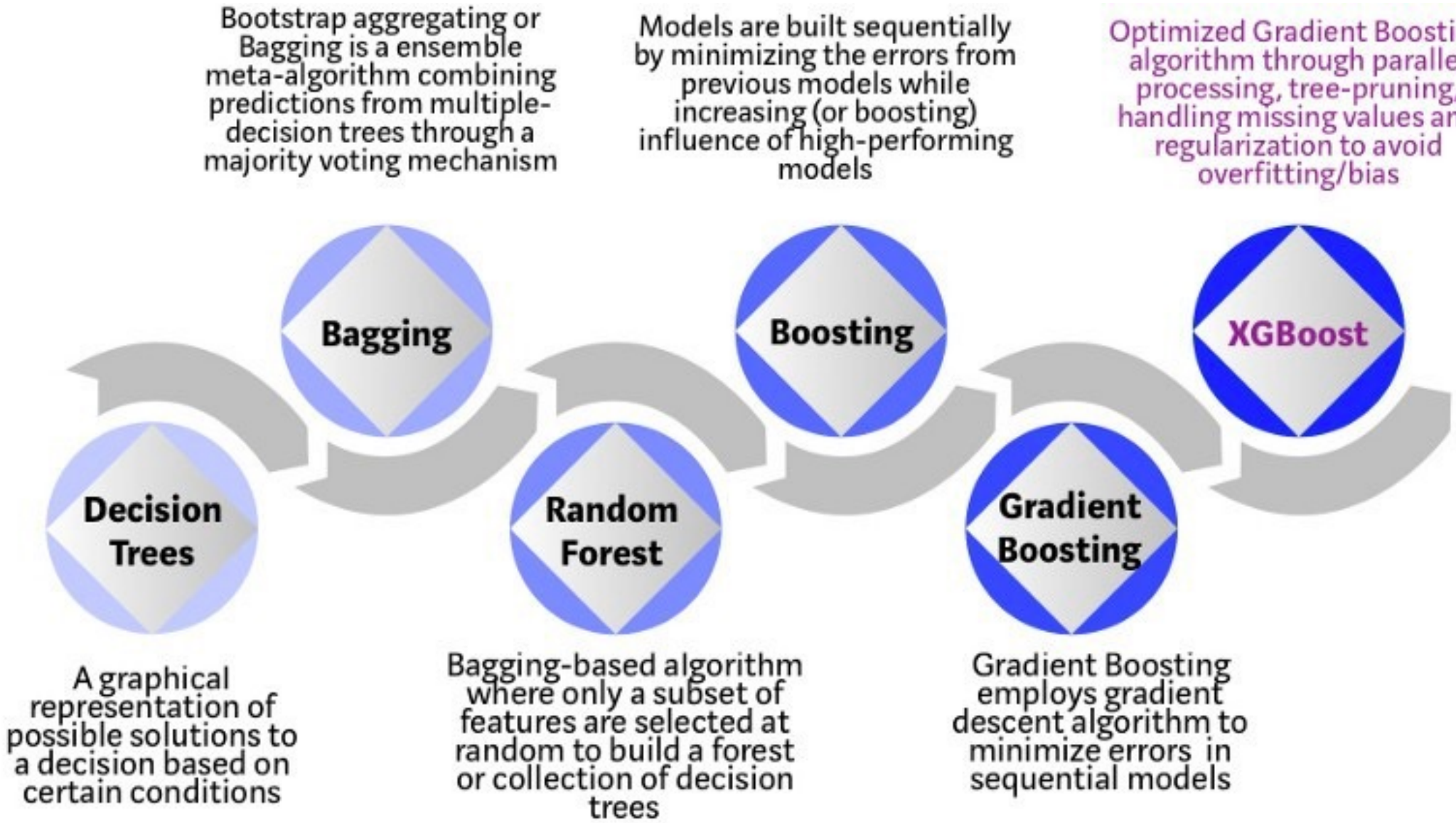
Как же прогнозируются?

- Есть здоровый массив исторический данных
- Формализуются метрика и методология ее расчета
- Собирается выборка
- Обучается модель

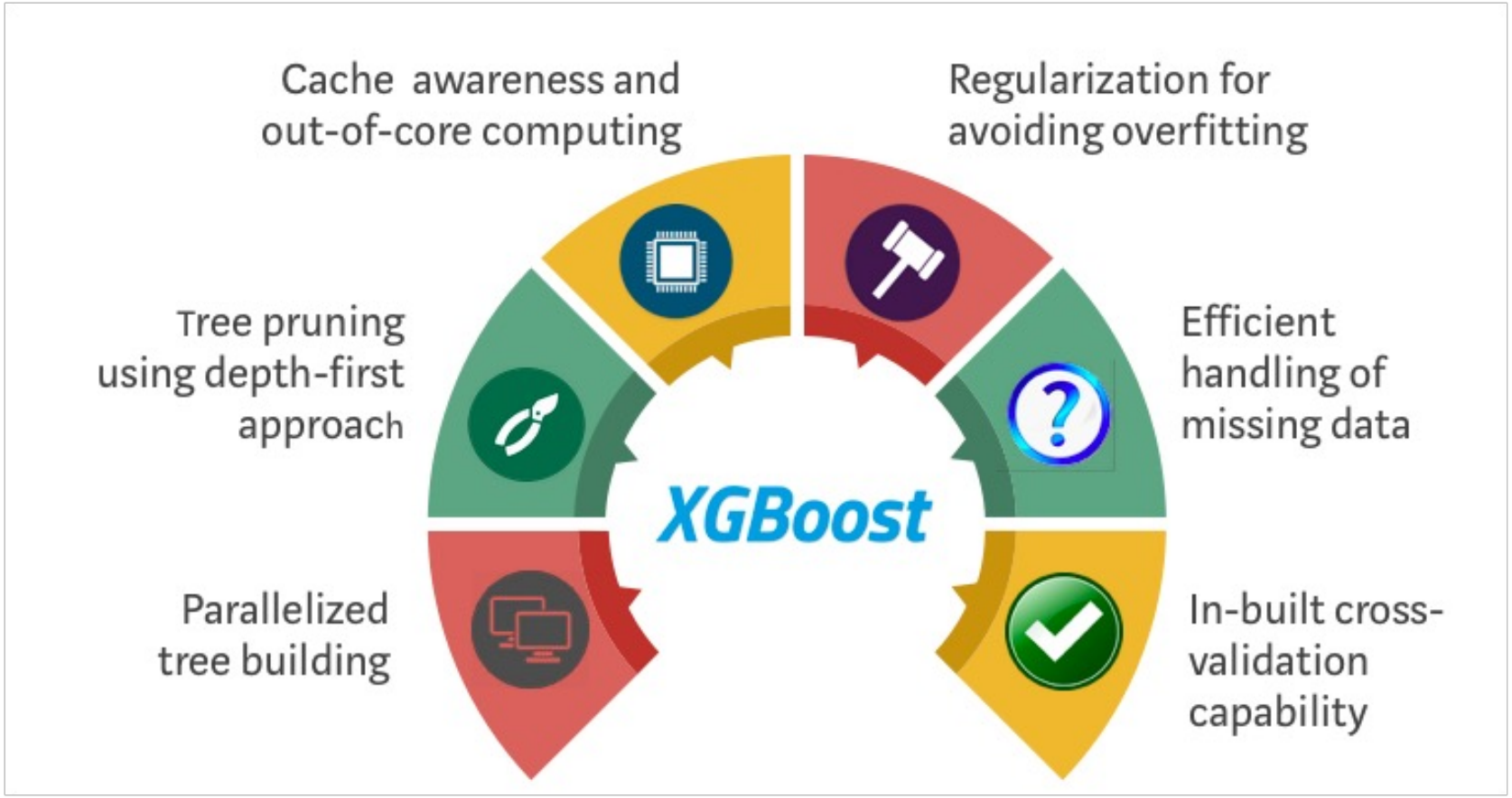


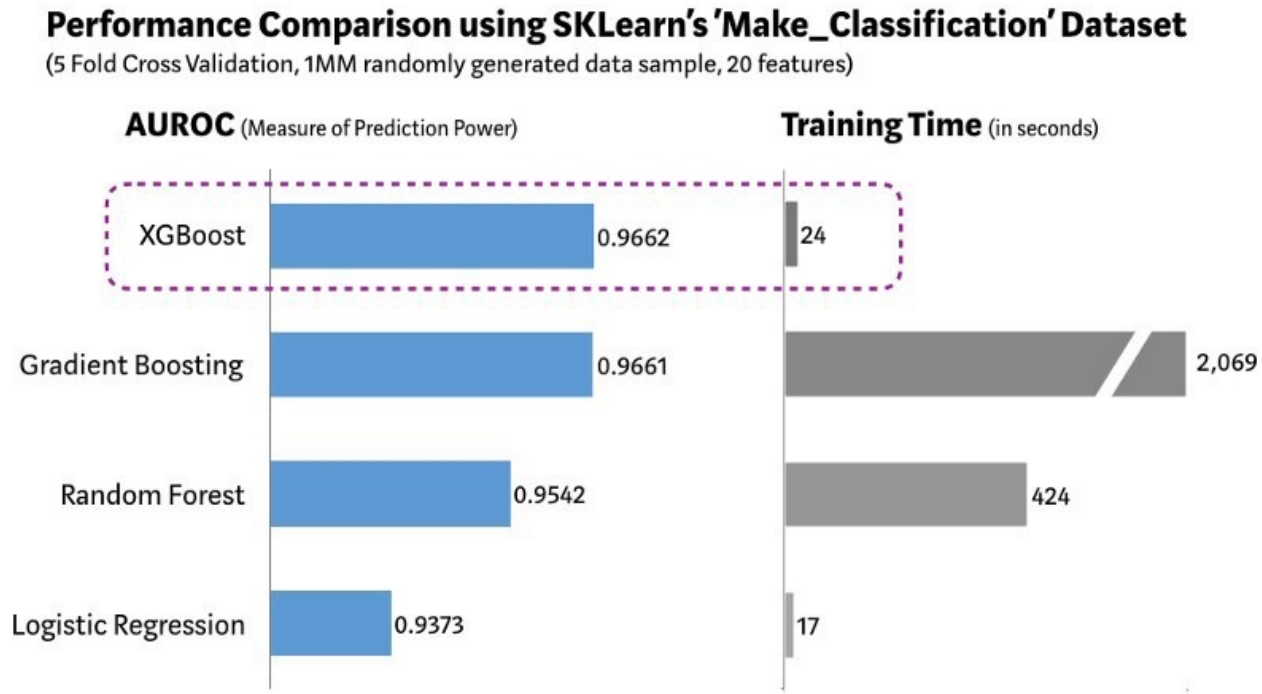
Что под капотом?
Модель




Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism
Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models
Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias
Bagging
Boosting
XGBoost
Decision Trees
Random Forest
Gradient Boosting
A graphical representation of possible solutions to a decision based on certain conditions
Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees
Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models

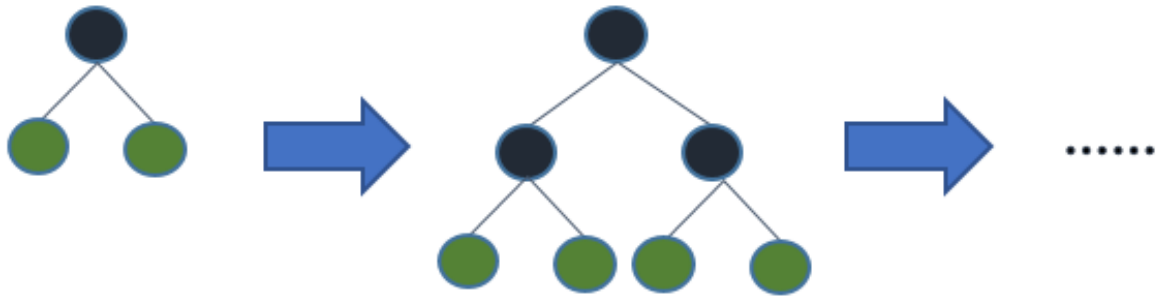
XGBoost


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability


Performance Comparison using SKLearn's 'Make_Classification' Dataset
(5 Fold Cross Validation, 1MM randomly generated data sample, 20 features)
AUROC (Measure of Prediction Power)
Training Time (in seconds)
XGBoost
0.9662
24
Gradient Boosting
0.9661
2,069
Random Forest
0.9542
424
Logistic Regression
0.9373
17

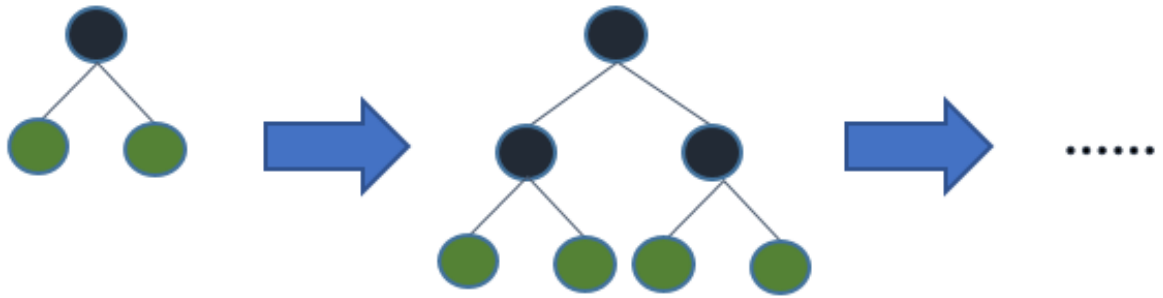
LightGBM

- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина



LightGBM

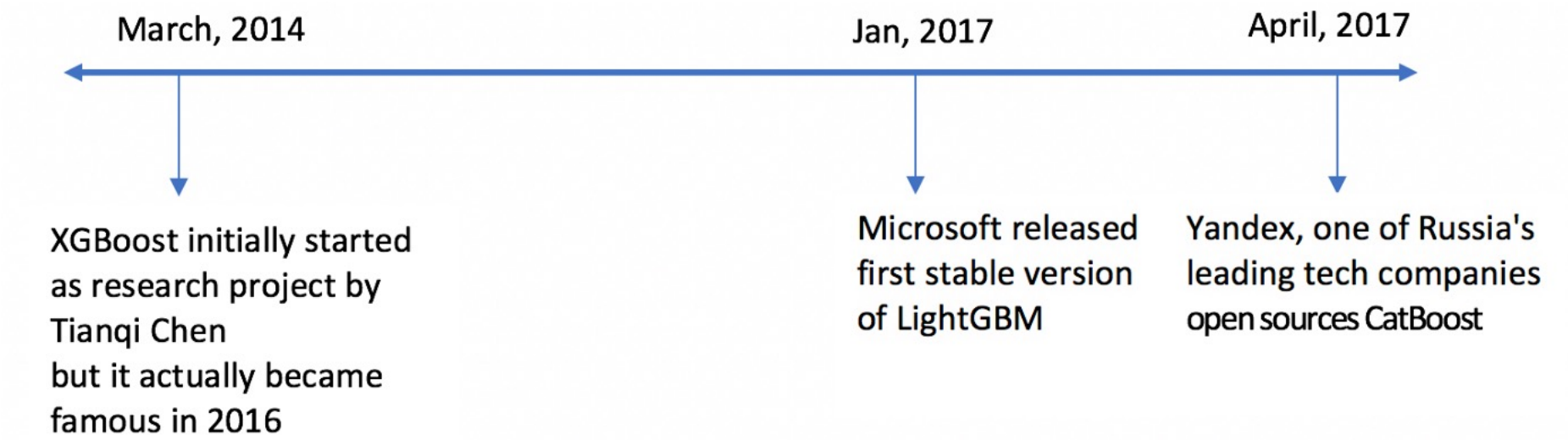
- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина
- Leaf-wise: среди текущих листьев выбирается тот, чьё разбиение сильнее всего уменьшает ошибкv



Level-wise tree growth

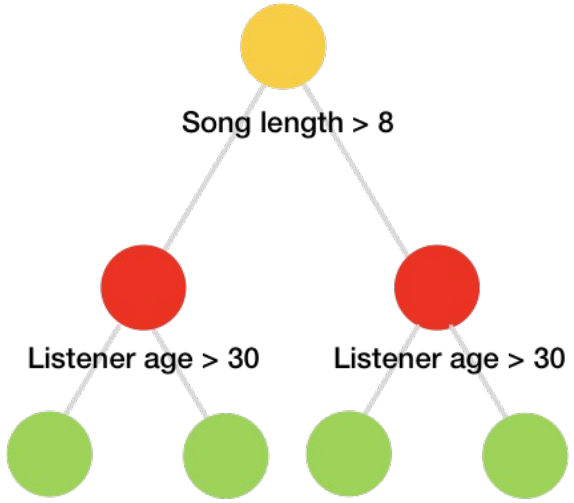
XGBoost vs LightGBM

- XGBoost разветвляет один уровень одновременно, LightGBM – одну вершину
- Разработчики XGBoost добавили эту опцию в свою реализацию, но XGBoost LightGBM быстрее в 1.3 – 1.5 раза, чем XGB.


March, 2014
Jan, 2017
April, 2017
XGBoost initially started as research project by Tianqi Chen but it actually became famous in 2016
Microsoft released first stable version of LightGBM
Yandex, one of Russia's leading tech companies open sources CatBoost

CatBoost

- Oblivious decision trees
- Ограничение: на одном уровне дерева используется один и тот же предикат



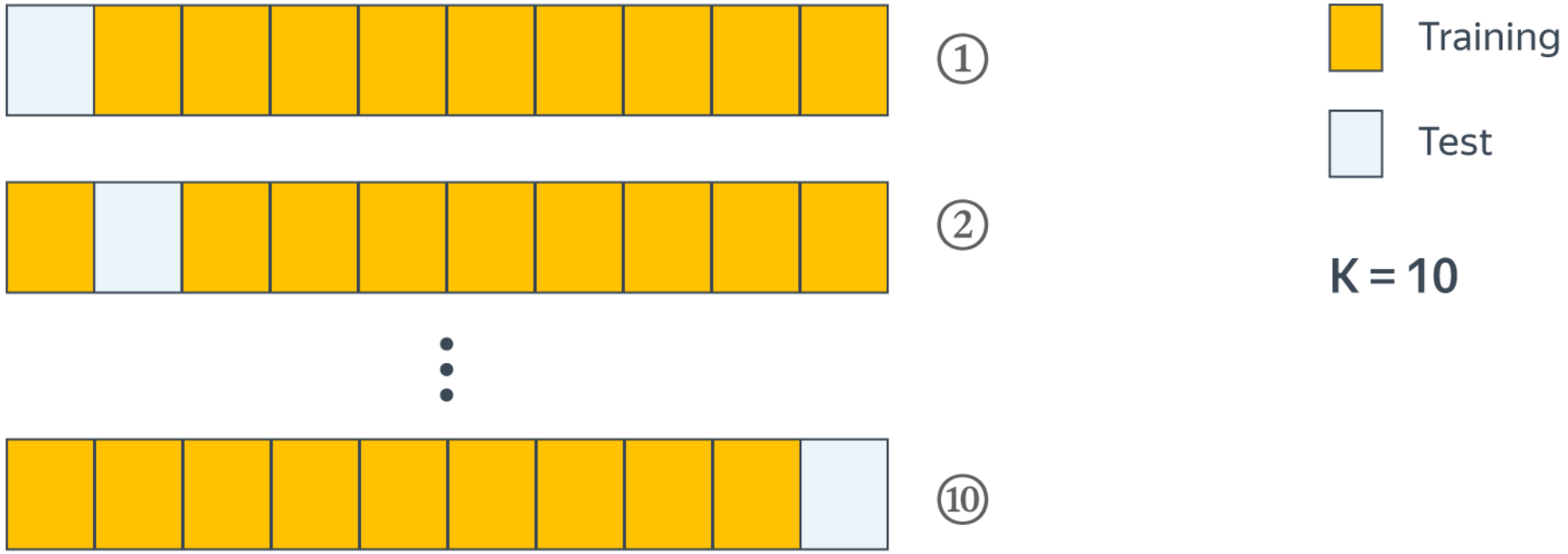
	CatBoost	LightGBM	XGBoost
Developer	Yandex	Microsoft	DMLC
Release Year	2017	2016	2014
Tree Symmetry	Symmetric	Asymmetric Leaf-wise tree growth	Asymmetric Level-wise tree growth
Splitting Method	Greedy method	Gradient-based One-Side Sampling (GOSS)	Pre-sorted and histogram-based algorithm
Type of Boosting	Ordered	-	-
Numerical Columns	Support	Support	Support
Categorical Columns	Support Perform one-hot encoding (default) Transforming categorical to numerical columns by border, bucket, binarized target mean value, counter methods available	Support, but must use numerical columns Can interpret ordinal category	Supports, but must use numerical columns Cannot interpret ordinal category, users must convert to one-hot encoding, label encoding or mean encoding
Text Columns	Support Support Bag-of-Words, Naïve-Bayes or BM-25 to calculate numerical features from text data	Do not support	Do not support
Missing values	Handle missing value Interpret as NaN (default) Possible to interpret as error, or processed as minimum or maximum values	Handle missing value Interpret as NaN (default) or zero Assign missing values to side that reduces loss the most in each split	Handle missing value Interpret as NaN (tree booster) or zero (linear booster) Assign missing values to side that reduces loss the most in each split

Function	XGBoost	CatBoost	Light GBM
Important parameters which control overfitting	1. **learning_rate or eta** – optimal values lie between 0.01-0.2 2. **max_depth** 3. **min_child_weight:** similar to min_child leaf; default is 1	1. **Learning_rate** 2. **Depth** - value can be any integer up to 16. Recommended - [1 to 10] 3. No such feature like min_child_weight 4. **l2-leaf-reg**: L2 regularization coefficient. Used for leaf value calculation (any positive integer allowed)	1. **learning_rate** 2. **max_depth**: default is 20. Important to note that tree still grows leaf-wise. Hence it is important to tune **num_leaves** (number of leaves in a tree) which should be smaller than 2^(max_depth). It is a very important parameter for LGBM 3. **min_data_in_leaf**: default=20, alias= min_data, min_child_samples
Parameters for categorical values	Not Available	1. **cat_features**: It denotes the index of categorical features 2. **one_hot_max_size**: Use one-hot encoding for all features with number of different values less than or equal to the given parameter value (max – 255)	1. **categorical_feature:** specify the categorical features we want to use for training our model
Parameters for controlling speed	1. **colsample_bytree**: subsample ratio of columns 2. **subsample**: subsample ratio of the training instance 3. **n_estimators**: maximum number of decision trees; high value can lead to overfitting	1. **rsm:** Random subspace method. The percentage of features to use at each split selection 2. No such parameter to subset data 3. **iterations**: maximum number of trees that can be built; high value can lead to overfitting	1. **feature_fraction:** fraction of features to be taken for each iteration 2. **bagging_fraction:** data to be used for each iteration and is generally used to speed up the training and avoid overfitting 3. **num_iterations:** number of boosting iterations to be performed; default=100

Что под капотом?

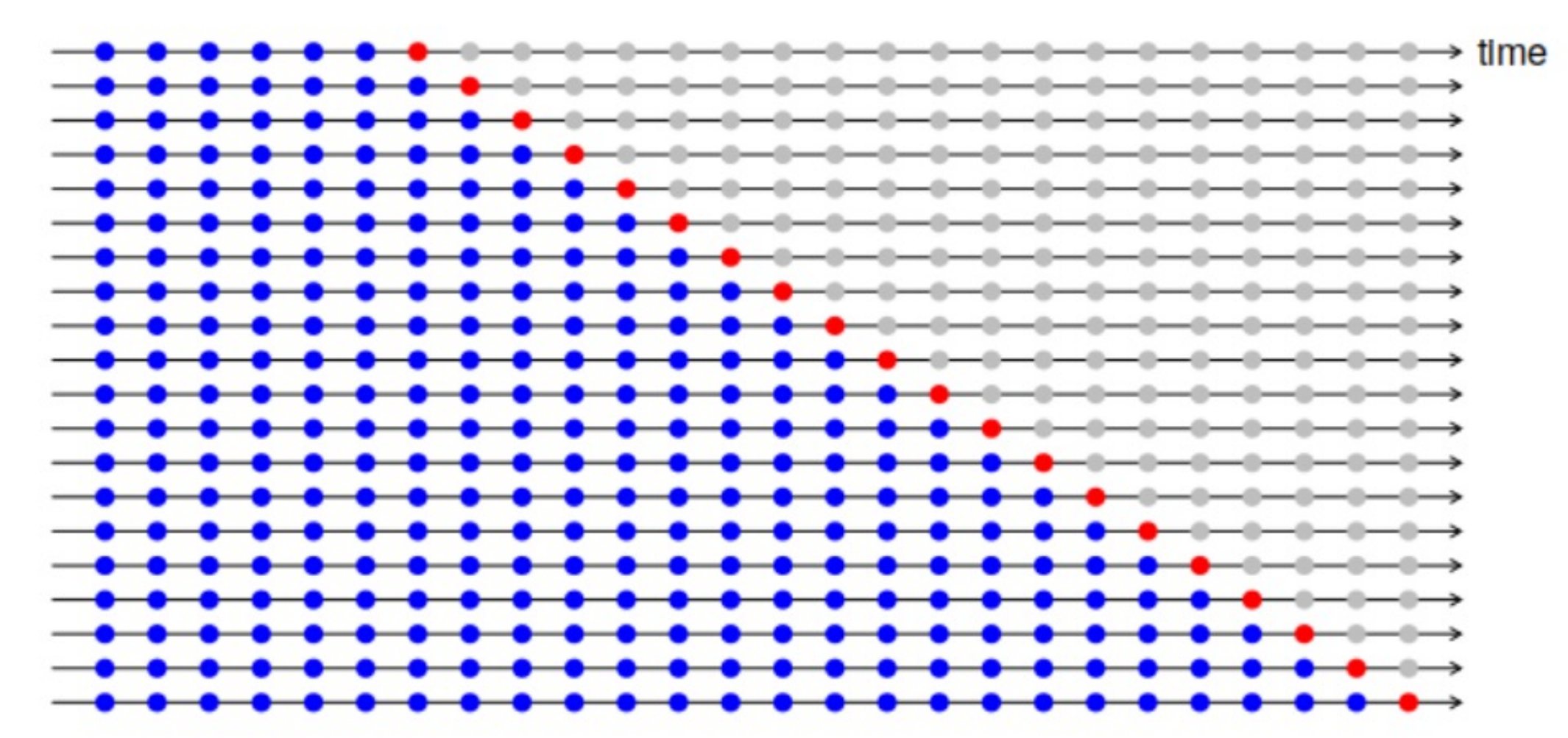
Валидация

Что под капотом?
Валидация. K-Fold

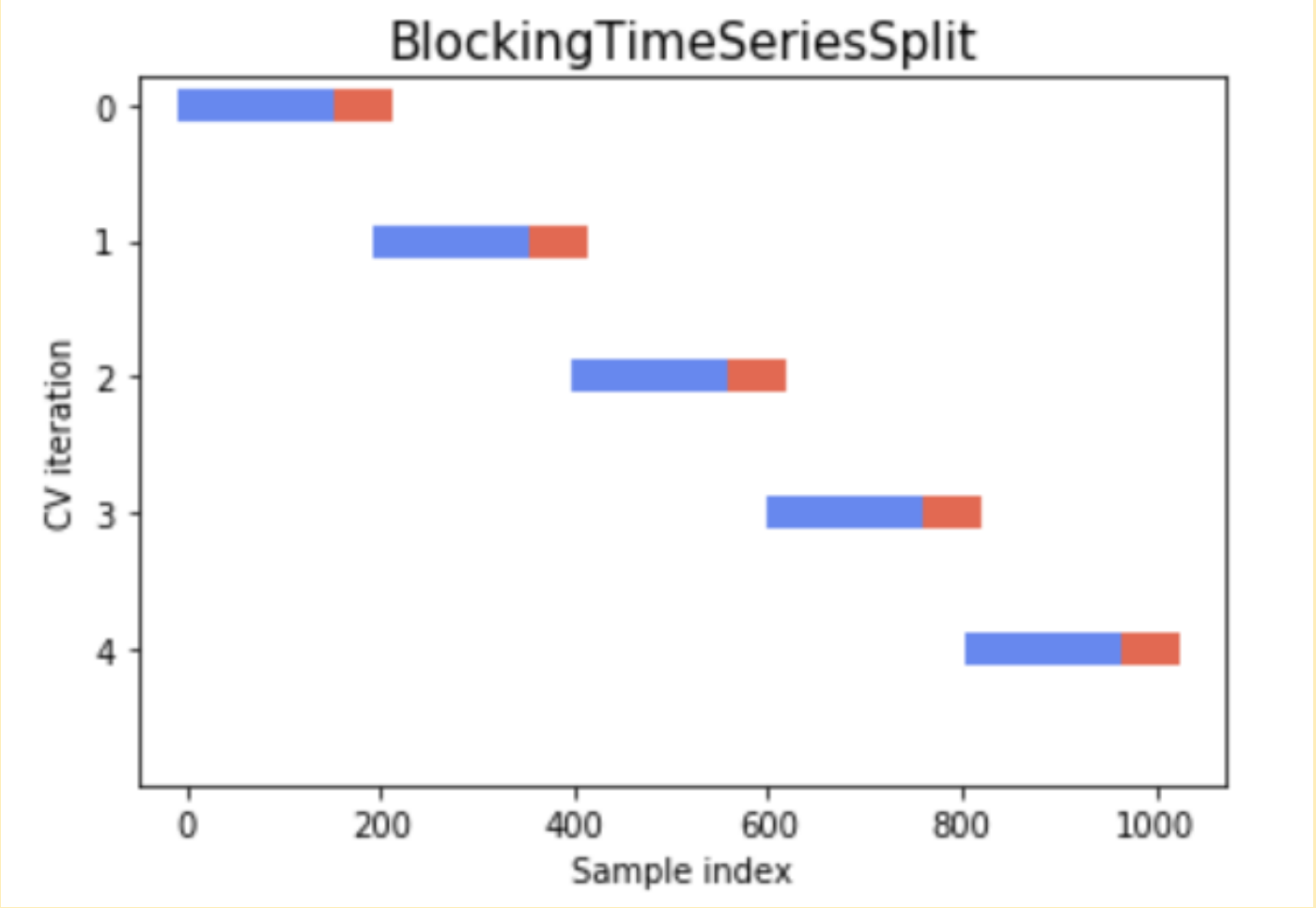


Что под капотом?

Валидация. Walk Forward



Что под капотом?
Blocked Cross Validation



Что под капотом?
Неявное про метрику

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Что под капотом?
Неявное про метрику

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$\text{WAPE} = \frac{\sum_{i,t}|y_{i,t} - \hat{y}_{i,t}|}{\sum_{i,t}|y_{i,t}|}$$

X5 Tech

20 августа 2021