Stock prediction leveraging Machine Learning and Deep Learning

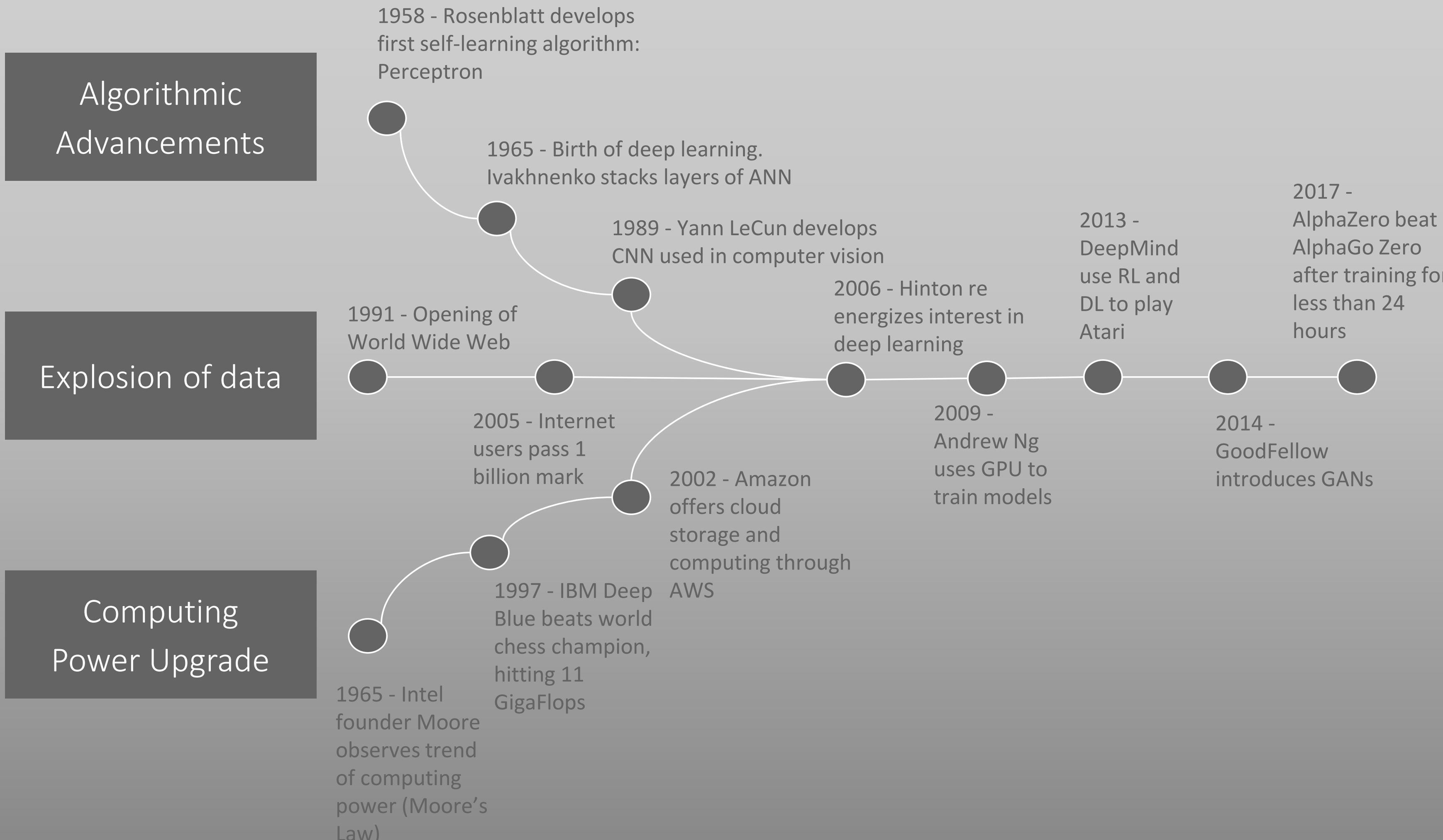
Sydney November 2018

Carolina Hoffmann-Becking

Agenda

1. Introduction: AI drivers and applications in Finance
2. Machine Learning in Finance
    - 2.1 **Development of trading strategy** leveraging data analysis and visualisation with Python libraries
    - 2.2 **Development of predictive model** for a continuous time series data set with -
        - 2.2.1 Machine Learning Algorithms
        - 2.2.2 Recurrent Neural Network Learning Architecture
    - 2.3 **Boosting your performance with Ensemble Learning**
3. Summary

From 1958 to today - Artificial Intelligence Drivers



Intelligence Landscape – Do we need to be scared of AI?

1 **Human Intelligence**
- Reasoning
- Situation assessment
- Decision making

2 **Augmented Intelligence**
Combining human intelligence with machine intelligence

3 **Machine Intelligence**
- Recognition
- Problem solving
- Execution and action

Data + AI = Enhanced alpha?

Business applications of AI in Finance

Research

- ✓ AI-supported access to and analysis of complex alternative data
- ✓ Sentiment Analysis
- ✓ Optical Character Recognition (OCR)

Portfolio Management

Early warning system for risk mitigation

- ✓ AI alerts sent directly to PMs to manage stock specific risks and market risks

AI assisted investment decisions

- ✓ Deep Portfolio Theory using neural networks for portfolio construction

Support

- ✓ Chat Bots
- ✓ Call Center Natural Language Processing
- ✓ Compliance Facial Recognition

Develop predictive model on time series data set

Development of predictive model on time series data set

Data Analytics

Current: IFTT

- ✓ "If value today over historical value then hold"
- ✓ No learning aspect

? How do I incorporate the learning aspect?

Machine Learning

Transform time series data set into non-time series data set to apply machine learning algorithms

- Starting point

Deep Learning

Use Deep Neural Network Learning architectures to predict on a time series data set such as Recurrent Neural Networks

✓ Most accurate results

Development of trading strategy

Crypto Currency Dataset

- ✓ Install and setup of Anaconda
- ✓ Import dataset
- ✓ Understand and clean dataset
- ✓ Develop trading strategies from data insights
  - ✓ Features correlation
  - ✓ Descriptive data points
  - ✓ Visualize data including data normalization
- ✓ Transform time series data set into non-time series for ML application
  - ✓ Feature engineering
  - ✓ Test statistical stationarity
- ✓ Built Predictive Regression model

Python Libraries

Leveraging data analysis and visualization with Python Libraries

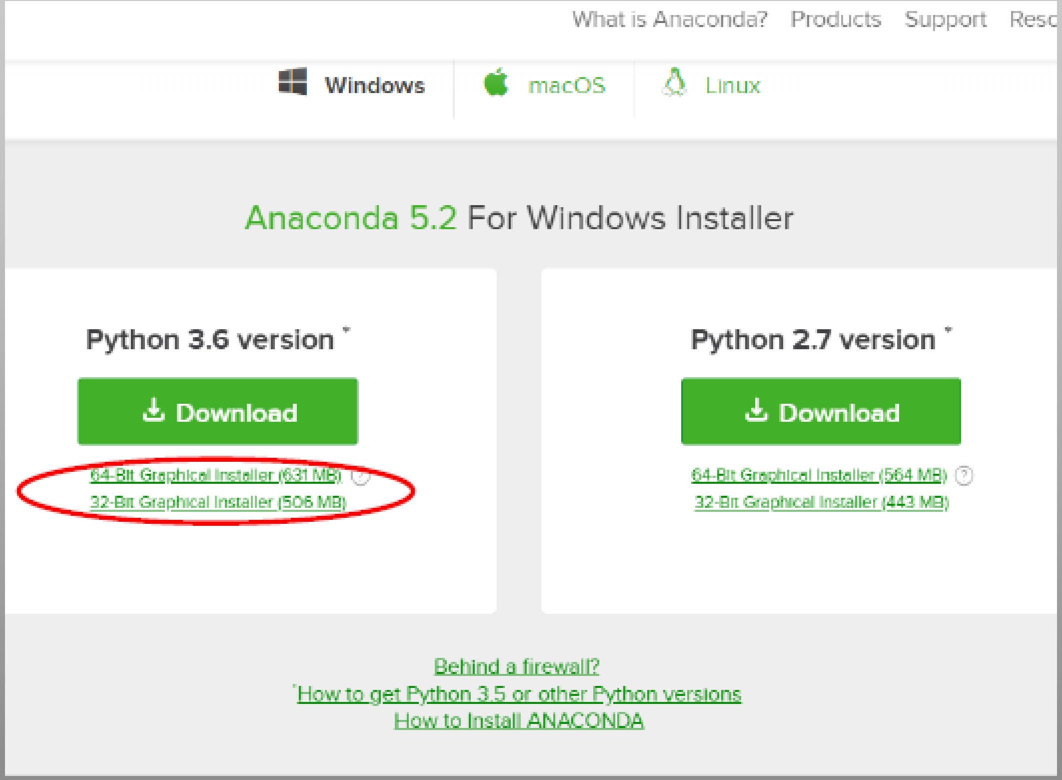
- ✓ NumPy – Numerical Processing
- ✓ Pandas – Data Analysis and Visualisation
- ✓ Matplotlib – Data Visualisation
- ✓ Scikit learn (sklearn) – Preprocessing ML Datasets

Coding exercise

Download Anaconda 5.2

- ✓ Anaconda version: **5.2**
- ✓ Python version: **3.6**
- ✓ 64-Bit Graphic installer

Link: https://www.anaconda.com/download/



Develop predictive model on time series data set

Development of predictive model on time series data set

Data Analytics

Current: IFTT

- ✓ "If value today over historical value then hold"
- ✓ No learning aspect

? How do I incorporate the learning aspect?

Machine Learning

Transform time series data set into non-time series data set to apply machine learning algorithms

- Starting point

Develop predictive model on time series data set

Machine Learning

Time series data set

- ✓ Continuous data
- ✓ Leverage value of historical data by **turning time series data set into non-time series data set**
- ✓ Features Engineering
- ✓ Apply supervised regression algorithm since continuous data
- ✓ Testing with MAE, MSE, RMSE – "On average, how far off you are from the correct continuous value"

Date	Close Price	Volume	P/L	Y = Close Price TMR
2017-01-31	25.3	5200	+2.5	

Date	Last 5 day average close price	Last 5 day average Volume	Last 10 day average P/L	Y = Close Price TMR
2017-01-31	32	12200	+2.5	

*Dates here function as index while time series aspect has been removed

Develop predictive model on time series data set

Development of predictive model on time series data set

Data Analytics

Current: IFTT

- ✓ "If value today over historical value then hold"
- ✓ No learning aspect

? How do I incorporate the learning aspect?

Machine Learning

Transform time series data set into non-time series data set to apply machine learning algorithms

- Starting point

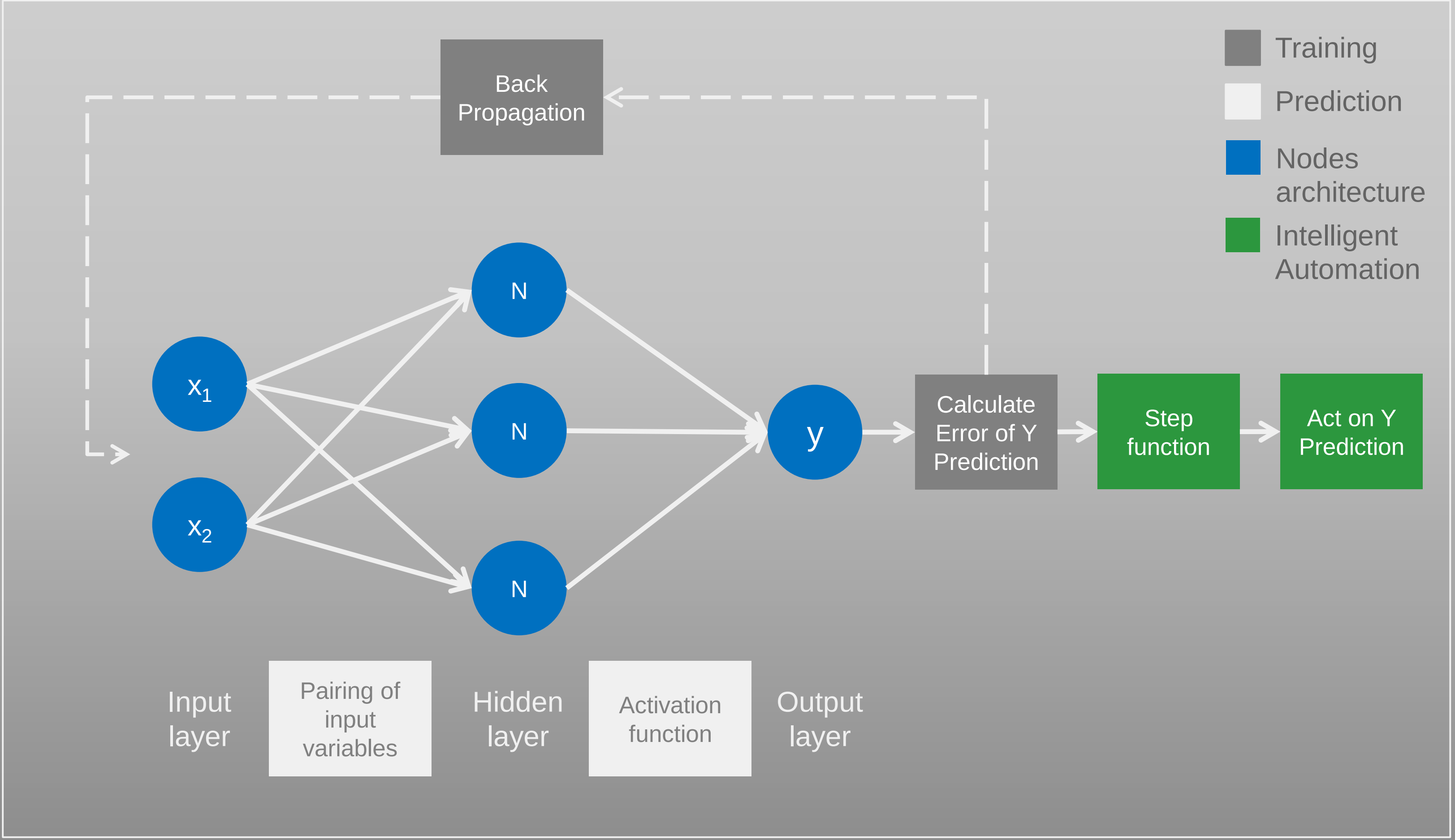
Deep Learning

Use Deep Neural Network Learning architectures to predict on a time series data set such as Recurrent Neural Networks

✓ Most accurate results

Develop predictive model on time series data set

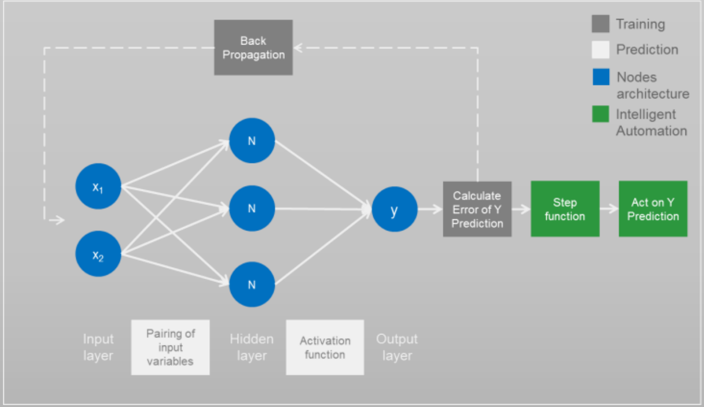
Deep Learning - Artificial Neural Network



Develop predictive model on time series data set

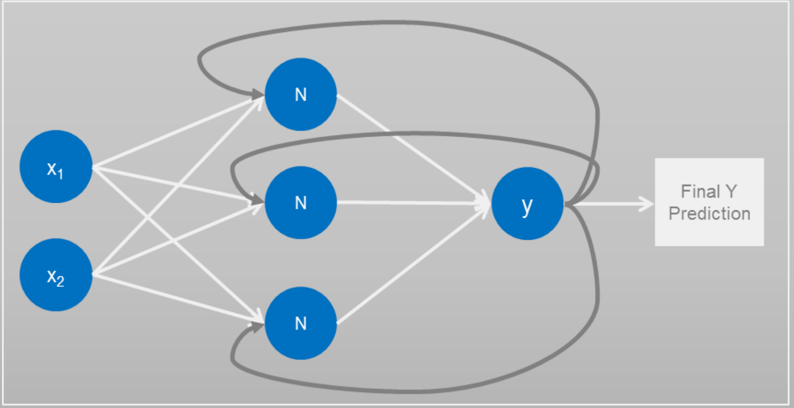
Deep Learning

Artificial Neural Network (ANN)



Basic form of **Artificial Neural Networks** are feedforward ones which can solve complex static problems.

Recurrent Neural Network (RNN)

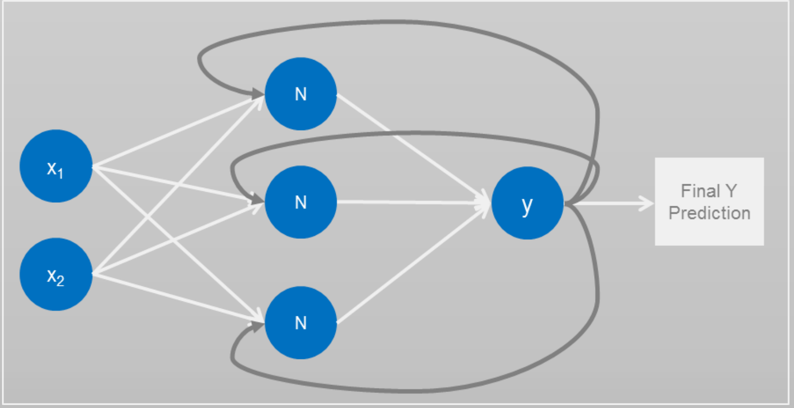


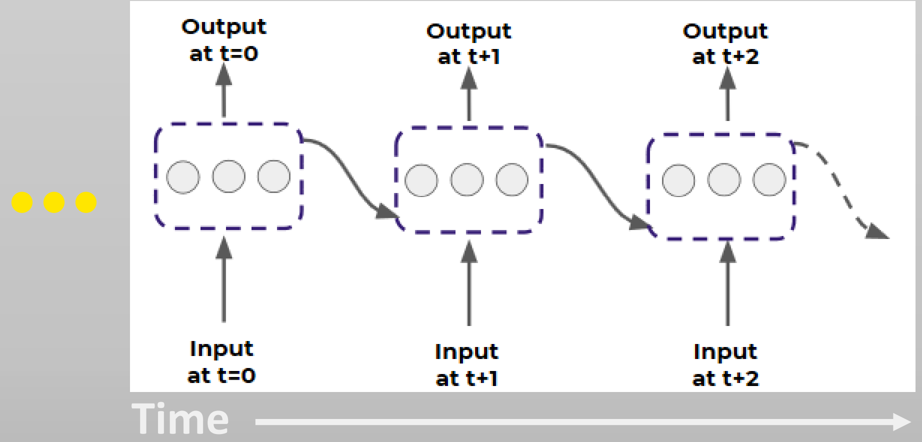
Recurrent Neural Networks (RNNs) are specifically good at dealing with **sequence information**:

- ✓ Time series: Stock price prediction
- ✓ Natural Language Processing: Generate text

Develop predictive model on time series data set

Deep Learning – Recurrent Neural Network





- ✓ Recurrent Neural Networks (RNN) are compared to the short term memory part of the brain
- ✓ Recurrent neurons receive both input from the current time step (i.e. Input at t=0 is x1 and x2 at t=0) and from the previous time step (i.e. Input at t=1 is BOTH output from t=0 that the recurrent neuron sends back to itself AND x1 and x2 values for t=1)
- ✓ "Memory cell": output of the recurrent neurons at a time step t is a function of all the inputs from previous time steps
- ✓ Challenge for RNN: Vanishing Gradient: Backpropagation goes backwards from the output to the input layer. Hence for deeper networks gradients often get smaller, eventually causing weights to never change at lower levels.
- ✓ Solution: GRU (Gated Recurrent unit) and LSTM (Long short term memory)

Boost your performance with Ensemble Learning!

[Ensemble Learning provides a "Prediction Consensus" where multiple models, classifier, repressor and other predictors, are mixed and combined to make better predictions and hence better decisions based on the predictions.]

1 Random Forest

2 Ada Boost

3 Gradient Boost

4 XG Boost

Ensemble Learning

Boost your performance with Ensemble Learning!

1 Random Forest of decision trees boosts prediction accuracy through achieving consensus on predictions from individual decision trees

- ✓ In a forest, n-trees are trained independently and simultaneously on a **random subset of training data**
- ✓ During testing each “new” data point will be pushed through all trees simultaneously to receive individual predicitions
- ✓ Random Forest includes Supervised Machine Learning Algorithms for Classification and Regression tasks: RandomForestClassifier and RandomForestRegressor
- ✓ Voting approach to achieve prediction consensus – “mode of the results” - for **Classification**
- ✓ Average predicted value to achieve consensus – “mean of the results” - for **Regression**
- ✓ Validation: Setting limitations to decision trees’ features to **avoid overfitting**
- ✓ Application example: Predict the expected loss or profit of a specific stock

Boost your performance with Ensemble Learning!

2 **Ada Boost** can boost predictions for datasets incorporating classification algorithms focused on leveraging features with small explanatory power

- ✓ Adaboost trains multiple models sequentially
- ✓ Based on the first core model, sequential models focus on adjusting feature weights for wrongly predicted values enabling features with small explanatory power contribute at lower scale predictions
- ✓ AdaBoost is **adaptive** in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers
- ✓ The output of the other learning algorithms (**'weak learners'**) is combined into a weighted sum that represents the final output of the boosted classifier.
- ✓ from sklearn.ensemble import AdaBoost**Regressor** | AdaBoostRegressor()

Ada Boost

Boost your performance with Ensemble Learning!

3 **Gradient Boosting** can boost predictions minimizing the error in the form of an ensemble of weak prediction models

- ✓ Machine learning technique for regression and classification problems
- ✓ Gradient Boosting fits different models on your training data calculating the cost function (error)
- ✓ It combines the models in the form of an ensemble of weak prediction models by letting them vote on their own goodness of fit and return the model mixture that works best
- ✓ Each Model will output its minimum error
- ✓ The **minimum error tells the optimum weight of X's in the model**
- ✓ By Default the models look at **MSE (Mean squared Error**): Minimization in comparison to other models MSE
- ✓ from sklearn.ensemble import GradientBoostingClassifier, GradientBoostingRegressor

Gradient Boost

Ensemble Learning

Boost your performance with Ensemble Learning!

How eBay predicts your next click with XG Boost!

You may have heard of the Ensemble Theorie Gradient Boosting, which looks for the minimum error in your predictive model to define the optimum weights for the model's Inputs (X). Now Ebay uses **Extreme** Gradient Boosting - "XG Boost".

XG Boost has Regularisation Penalities built in, which will penalise the output of your model the more X Inputs you add. In addition cross validation for each interim model are built in splitting your test data into small slices cross testing as you move on building the model. This approach makes your model more robust and less likely to overfit.

Alan Lu, Director of Engineering & Applied Science at eBay explained to me how eBay utilises Behavioral Data, Query Information, Item Information, Seller & Buyer Information aswell as Categorial Information for Feature Generation before applying XG Boost to predict your next click. This enables eBay to utilise clicks as a proxy for user engagement but also as a Framework for personalisation and co-optimization for other key metrics such as Conversion Rate, Revenue etc.

https://www.tianhui.hk/blog/how-ebay-predicts-your-next-click-with-xg-boost

- ✓ from xgboost import XGBRegressor, XGBCLassifier

4 XG Boost

Summary

What you need in a time series search for alpha!

1 Feature Engineering	2 Recurrent Neural Network	3 Ensemble Learning
4 Multi-collinearity	5 Hetero-scedasticity	6 Statistical stationarity

Building blocks for Time Series Predictive Model

Time Series Special Characteristics