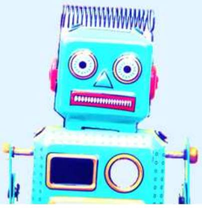
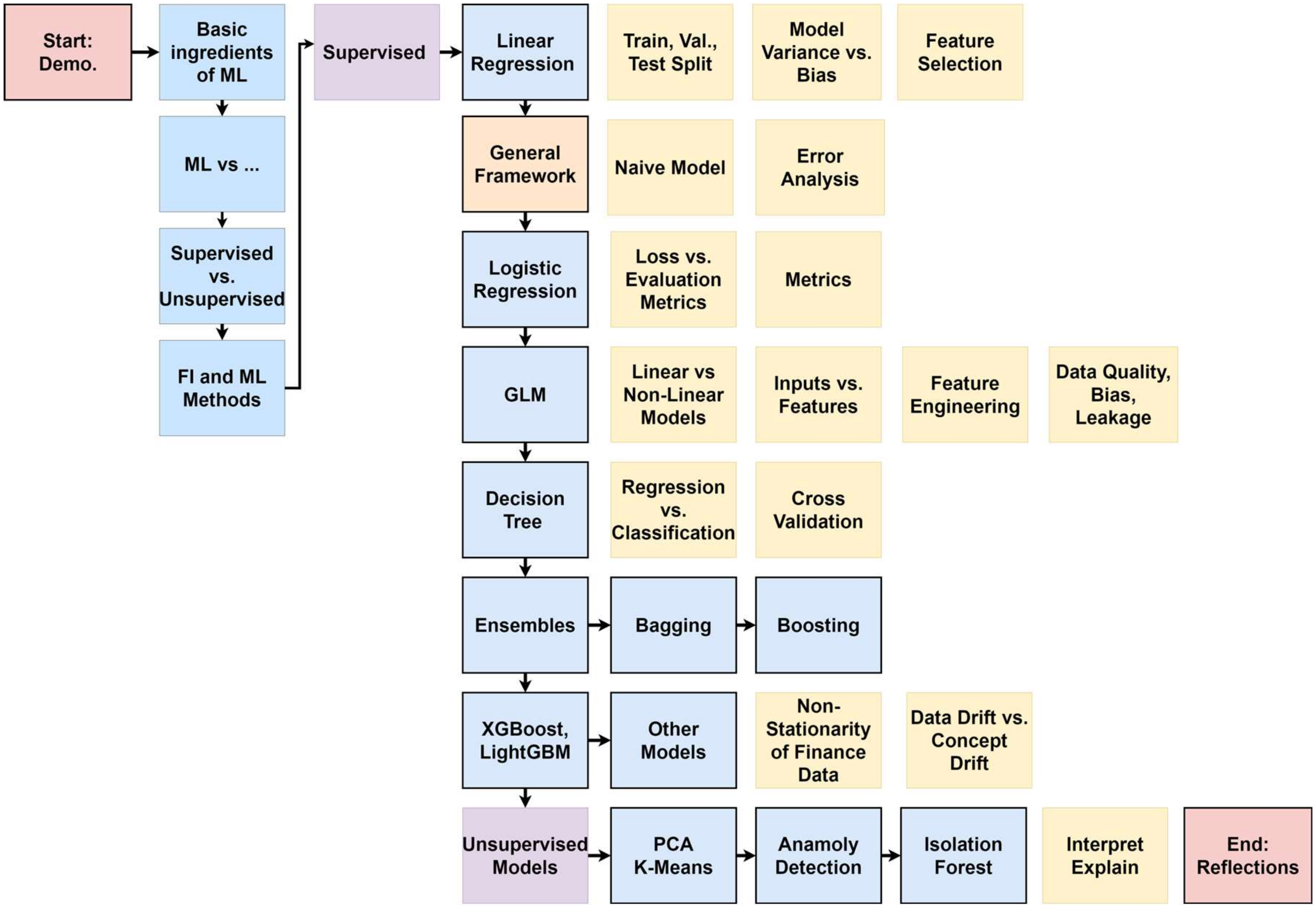
Machine and Deep Learning Fundamentals & Applications

Gary Ang



Overview



Start: Demo.
Basic ingredients of ML
ML vs ...
Supervised vs. Unsupervised
FI and ML Methods
Supervised
Linear Regression
Train, Val., Test Split
Model Variance vs. Bias
Feature Selection
General Framework
Naive Model
Error Analysis
Logistic Regression
Loss vs. Evaluation Metrics
Metrics
GLM
Linear vs Non-Linear Models
Inputs vs. Features
Feature Engineering
Data Quality, Bias, Leakage
Decision Tree
Regression vs. Classification
Cross Validation
Ensembles
Bagging
Boosting
XGBoost, LightGBM
Other Models
Non-Stationarity of Finance Data
Data Drift vs. Concept Drift
Unsupervised Models
PCA K-Means
Anamoly Detection
Isolation Forest
Interpret Explain
End: Reflections

What we will focus on

- Intuition
- Mental models
- Patterns
- Concepts

Demonstration

Twitter handle:

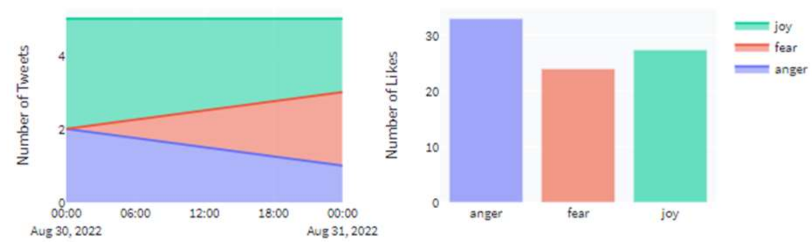
crypto

Number of tweets:

10

Get tweets!

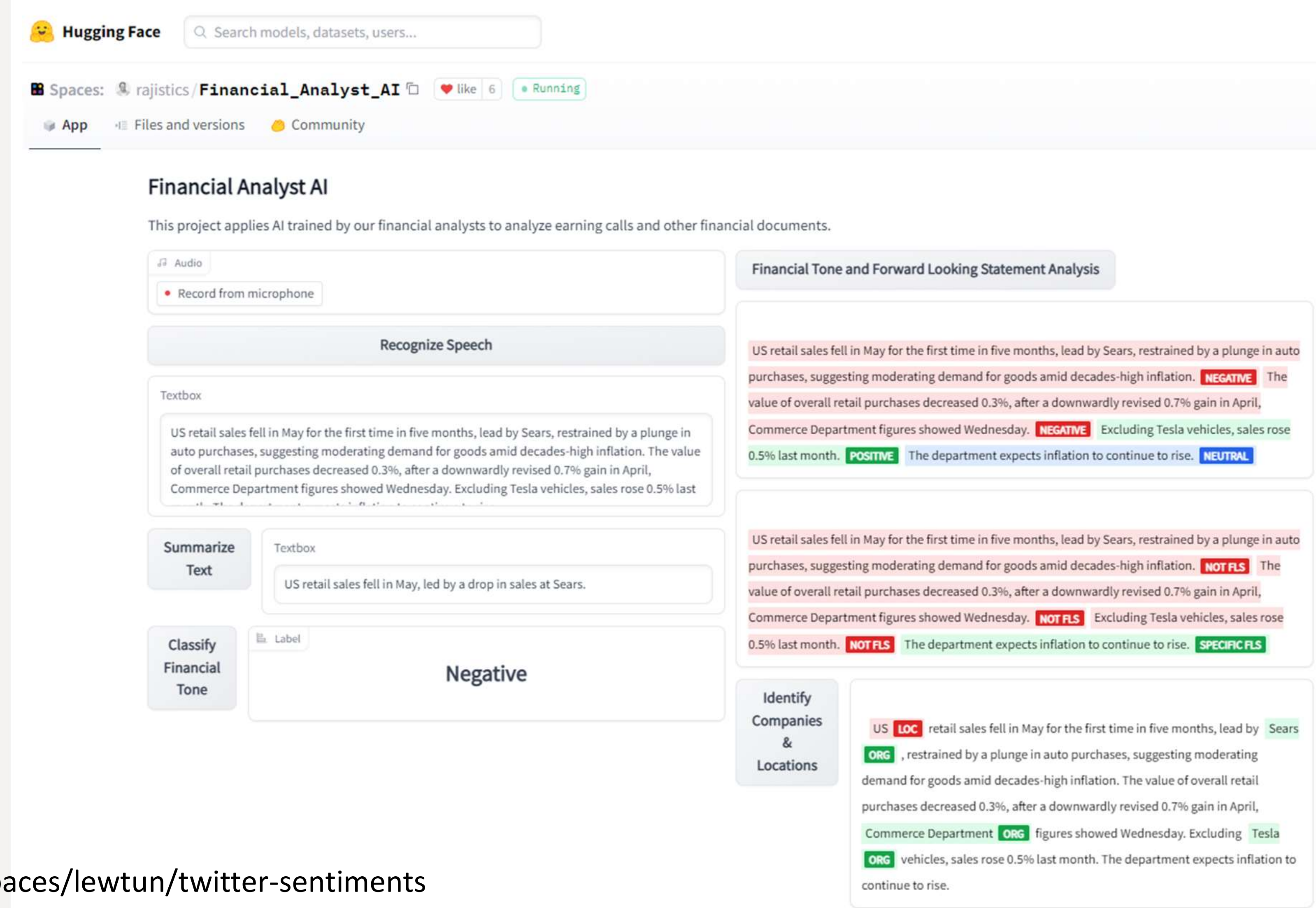
Twitter Emotion Analyser



	tweets	timestamps	retweets	likes	labels	scores
3	jpmorgan asset management's david kelly has some words of advice for investors rattled by a hawkish fed: forget about short-term direction and focus on valuations https://t.co/bxa8lzwi6z	2022-08-31 02:02:39+00:00	17	32	fear	0.986367
5	"the idea of millions of people in a fully customized live environment doing whatever we want significantly outstrips our computational capabilities today," venture capitalist matthew ball said https://t.co/aezregzthm	2022-08-30 23:51:40+00:00	7	21	joy	0.971417

https://huggingface.co/spaces/lewtun/twitter-sentiments

Demonstration

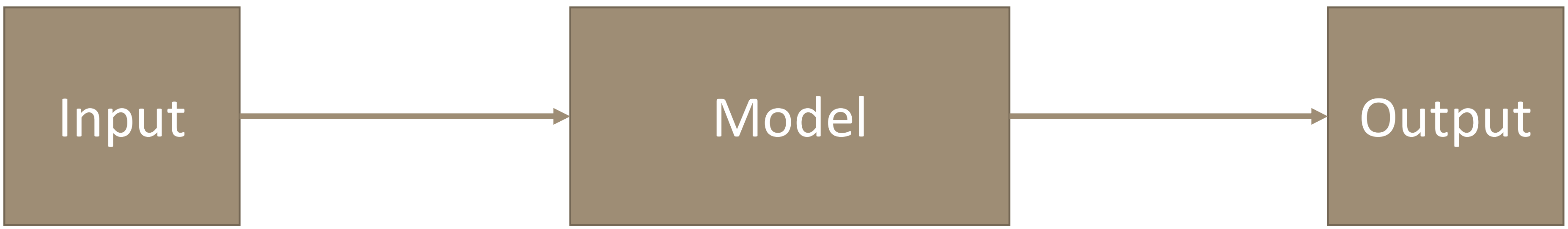


https://bit.ly/3AHNPRo

https://huggingface.co/spaces/lewtun/twitter-sentiments

Let’s jump in …

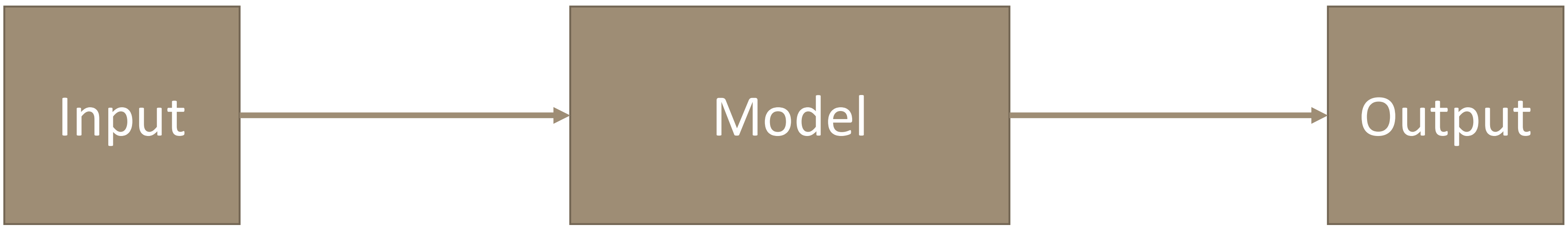
Take ‘Classify Financial Tone’ as the task we are interested in …



What kind of data was used?

Let's jump in …

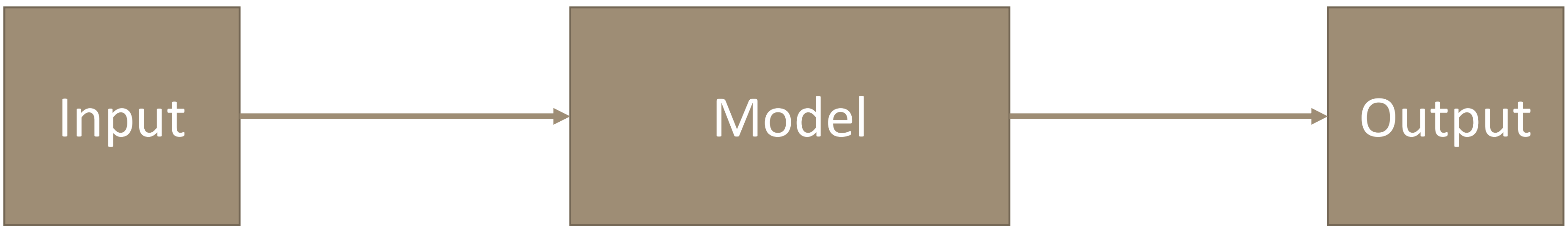
Take 'Twitter Emotion Analyser' as the task we are interested in …



What are the special characteristics of the data that you need to consider when designing the model?

Let’s jump in …

Take ‘Twitter Emotion Analyser’ as the task we are interested in …



What’s special about the outputs?

Machine Learning vs.

Artificial intelligence

Deep learning

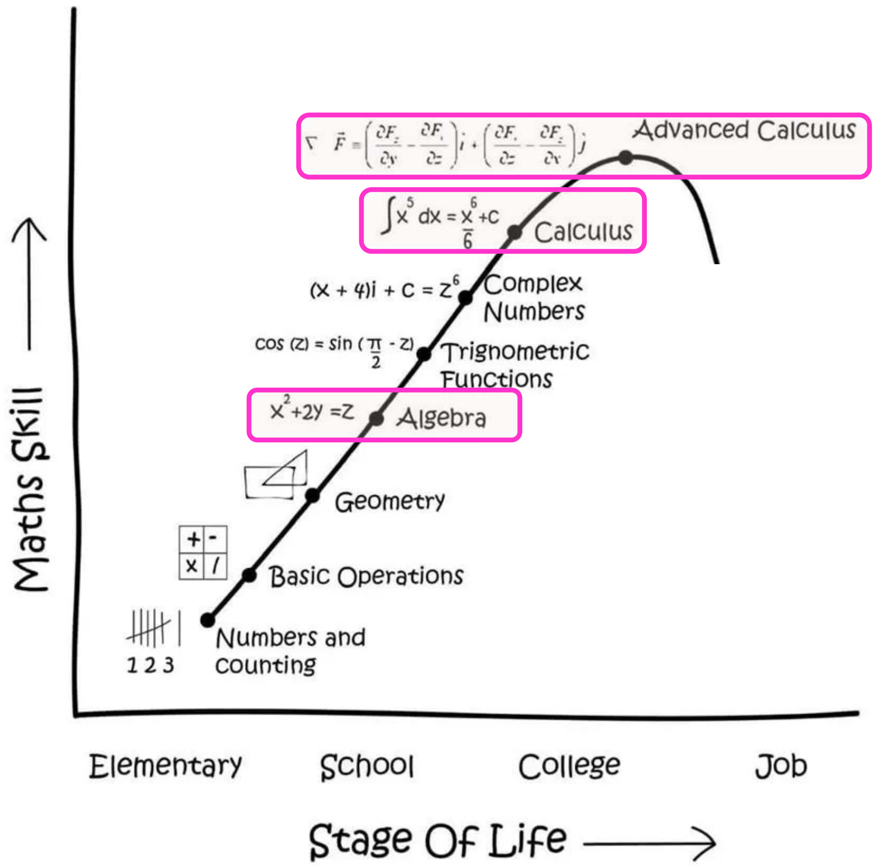
Statistics

Financial engineering

Econometrics

The line separating these areas is very thin and porous

Significant overlaps!



Maths Skill
Stage Of Life
Elementary
School
College
Job
1 2 3
Numbers and counting
Basic Operations
Geometry
$x^2+2y=Z$
Algebra
$\cos(Z) = \sin(\frac{\pi}{2} - Z)$
Trignometric Functions
$(X+4)i + C = Z^6$
Complex Numbers
$\int x^5\, dx = \frac{x^6}{6} + C$
Calculus
Advanced Calculus

We learnt this in primary school, remember?

Different focus, nomenclature, techniques but is it that different?

How do I arrive at the formulation?

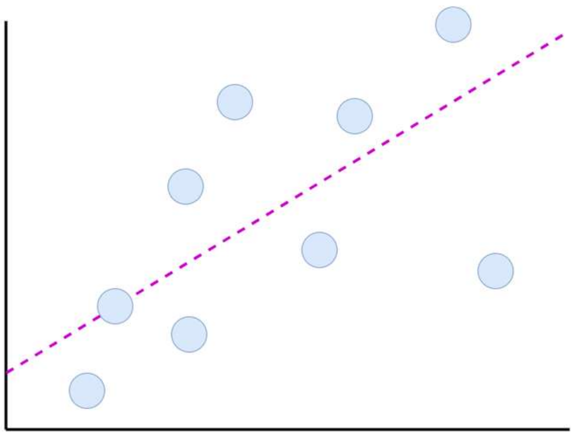
$$Y = AX + B$$

How good are the predictions, responses, targets, or dependent variables?

Are the coefficients, weights, or parameters statistically significant?

How do I find, learn, or train, these coefficients, weights, or parameters

What are the characteristics of the inputs, features, or independent variables?



In machine learning, we typically …

Personally, don't think defining what is *machine learning* and what isn't is important, but compared to:

- *Statistics*, we focus a little less on the statistical significance of coefficients in the model
- *Financial engineering and econometrics*, we focus less on defining a specific form that matches a theory (we let the data determine that for us)

3 major approaches

Imagine ….you as a baby, learning to navigate this world.

1. You touch a cup with steam rising from it. It scalds you. You learn to never touch a cup with steam rising from it.

2. You see a dog and a cat. You point to the dog and say ‘cat. Your mom corrects you.

3. You see a group of red and green blocks. You use the colors to organize them into 2 groups.

Supervised Learning

Unsupervised Learning

Reinforcement Learning

3 major approaches

Not all there is, there are others, and also combinations

Imagine ….you as a baby, learning to navigate this world.

1. You touch a cup with steam rising from it. It scalds you. You learn to never touch a cup with steam rising from it.

2. You see a dog and a cat. You point to the dog and say ‘cat. Your mom corrects you.

3. You see a group of red and green blocks. You use the colors to organize them into 2 groups.

Supervised Learning

2. You see a dog and a cat. You point to the dog and say ‘cat. Your mom corrects you.

Unsupervised Learning

3. You see a group of red and green blocks. You use the colors to organize them into 2 groups.

Reinforcement Learning

1. You touch a cup with steam rising from it. It scalds you. You learn to never touch a cup with steam rising from it.

Supervised vs. Unsupervised (or self-supervised)

Line between supervised and unsupervised learning can also be thin, but let's leave that aside for now

Based on a common-sense understanding, classify the following tasks: You have …

Data on past financial statements inc. credit scores of companies, *want to predict credit scores of companies based on their financials*

Supervised vs. Unsupervised (or self-supervised)

Line between supervised and unsupervised learning can also be thin, but let's leave that aside for now

Based on a common-sense understanding, classify the following tasks: You have …

Past interest rates time-series, *want to group interest rates that behave similarly*

Supervised vs. Unsupervised (or self-supervised)

Line between supervised and unsupervised learning can also be thin, but let's leave that aside for now

Based on a common-sense understanding, classify the following tasks: You have …

Data on past credit-card transactions (e.g., amounts, nature, fraudulent), *want to predict whether a current transaction is fraudulent or not*

Supervised vs. Unsupervised (or self-supervised)

Line between supervised and unsupervised learning can also be thin, but let's leave that aside for now

Based on a common-sense understanding, classify the following tasks: You have …

Data on past transactions on the block-chain, *want to detect future transactions that are anomalous (and possibly illicit) for further investigation*

Common Models

Supervised Learning

- K-nearest neighbors
- Linear discriminant analysis
- Linear regression
- Logistic regression
- Support vector machines
- Decision trees

Unsupervised Learning

- Clustering - K-mean, DBSCAN
- Dimensionality reduction - Principal component analysis, Singular Value Decomposition, T-distributed stochastic neighbor embedding
- Latent Dirichlet Allocation

Neural Networks

- Multilayer Perceptron/Dense Neural Network
- Recurrent Neural Network
- Convolutional Neural Network
- Transformer
- Graph Neural Network

Common Models

Supervised Learning

- K-nearest neighbors
- Linear discriminant analysis
- Linear regression
- Logistic regression
- Support vector machines
- Decision trees

K-nearest neighbours to detect fraud based on top-K similar data points

Linear discriminant analysis for credit default classification by separating default and non-defaults

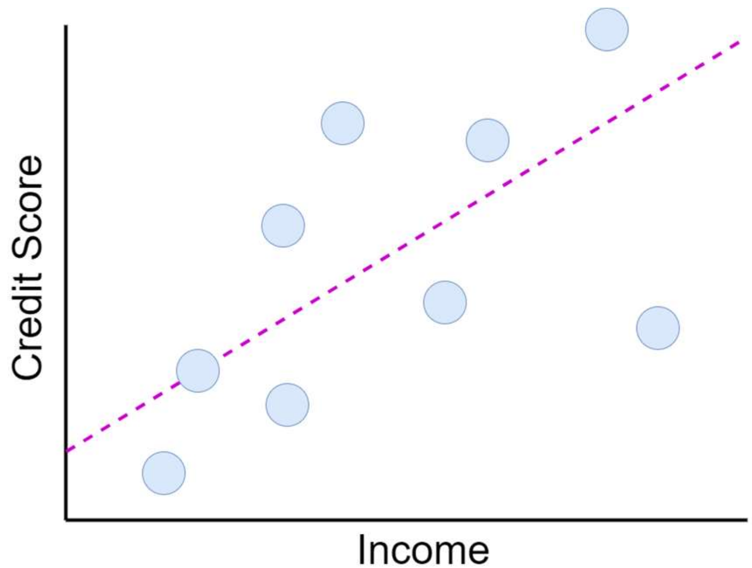
Credit scoring or default prediction use **logistic regression**, often with LASSO regularization

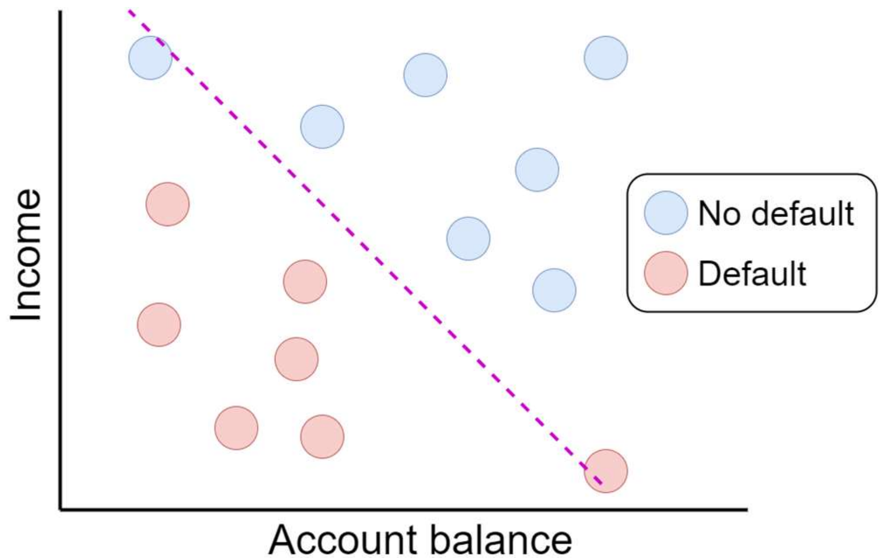
Linear and logistic regression from same family as Tweedie model used for insurance claims modelling – GLM

Decision tree-based models (CART, XGBoost, LightGBM), for tabular data, e.g., credit scoring, fraud detection, AML detection, buy/sell recommendations

Support vector machines for textual data, e.g., financial sentiment analysis

Some basic questions on supervised learning





- What are the labels and inputs? (Indicate on axis or chart)
- Which figure shows a regression, which shows a classification? (Indicate on chart)
- How would you describe the relationship between inputs and outputs?

Let's discuss

Regression	Classification
Predict numbers	*Predict categories*

Are the following tasks regression or classification? (Indicate with R or C)
What are the unique aspects of these tasks?

- **Forecast stock prices**
- **Nowcast GDP**
- **Predict cashtag (e.g., $AAPL) associated with a news article**
- **Predict bank's customer satisfaction level**
- **Predict top financial news that a customer may be interested in**
- **Recommend finance products of interest to a customer**

Common Models

Unsupervised Learning

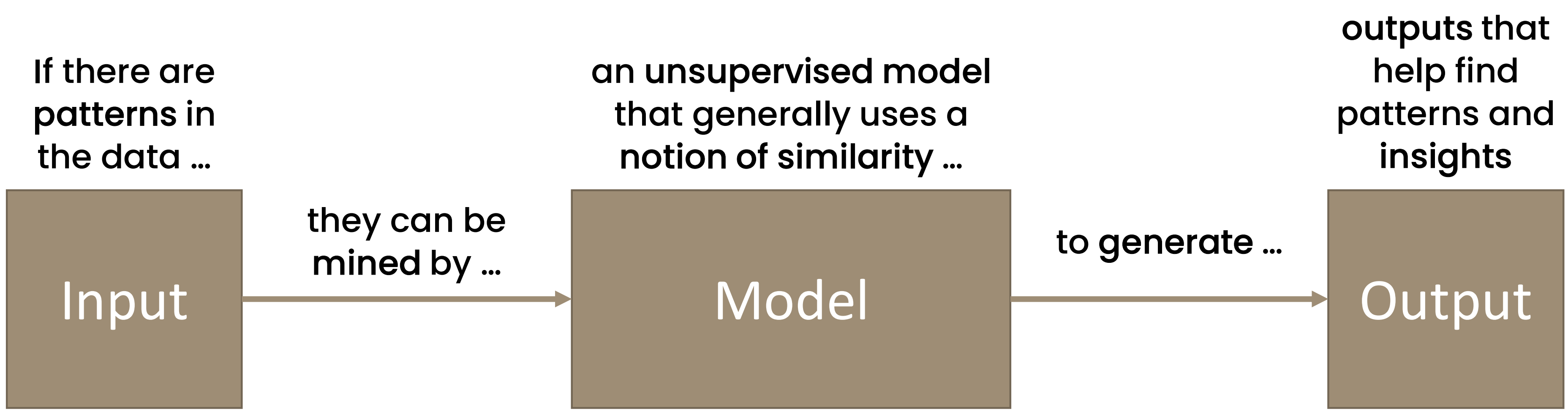
- Clustering - K-means, DBSCAN
- Dimensionality reduction - Principal component analysis, Singular Value Decomposition, T-distributed stochastic neighbor embedding
- Latent Dirichlet Allocation

Clustering to find groups of related bank customers, detect anomalous actors or transactions,

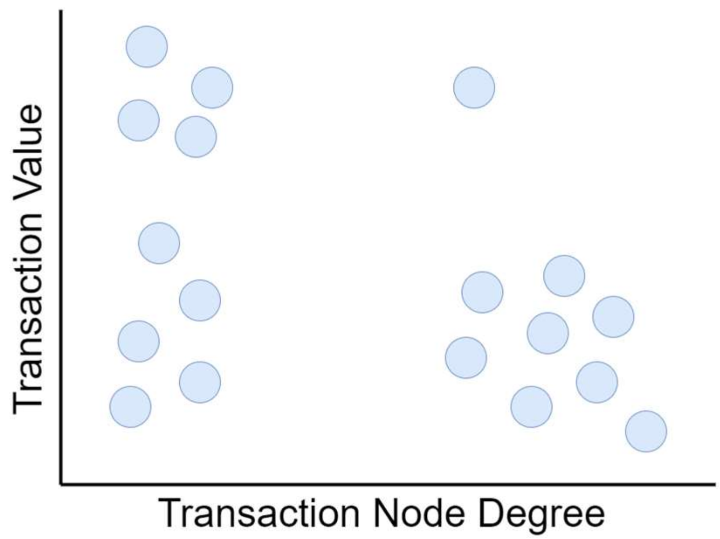
Dimensionality reduction to find key drivers of asset movements (e.g. slope, curvature, twist), visualize similar transactions or customers

Latent Dirichlet Allocation to find underlying topics in financial news

Unsupervised learning …



Some basic questions on unsupervised learning



- What are the potential inputs and outputs?
- What patterns or groups can you recognize?
- What is the notion of similarity that you are using to recognize these patterns?

Common Models

Supervised Learning | Unsupervised Learning

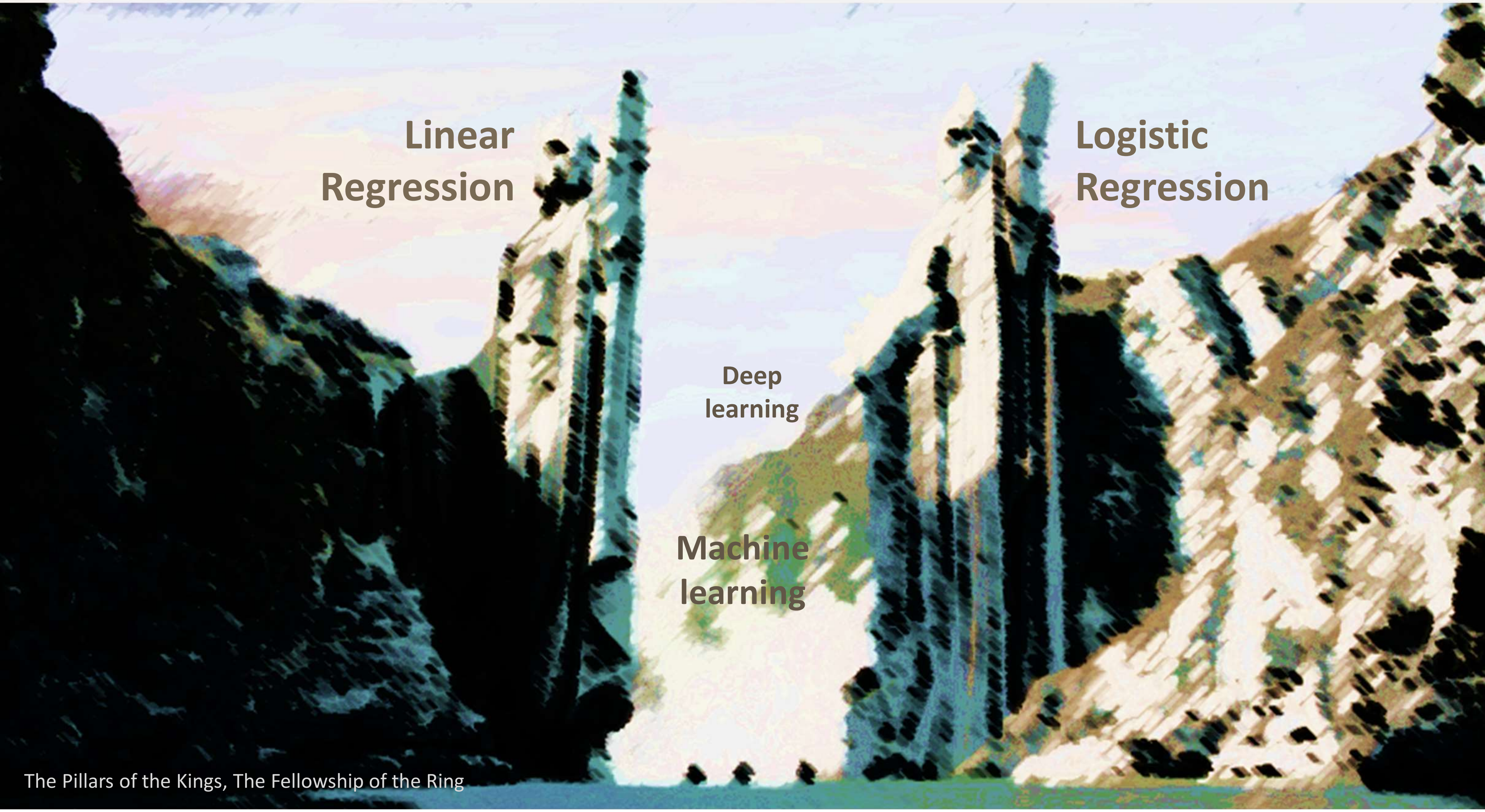
Neural Networks

- Multilayer Perceptron/Dense Neural Network
- Recurrent Neural Network
- Convolutional Neural Network
- Transformer
- Graph Neural Network

- Expressive, flexible, can be adapted for different types of tasks
- Different architectures for different data-types

Types of data

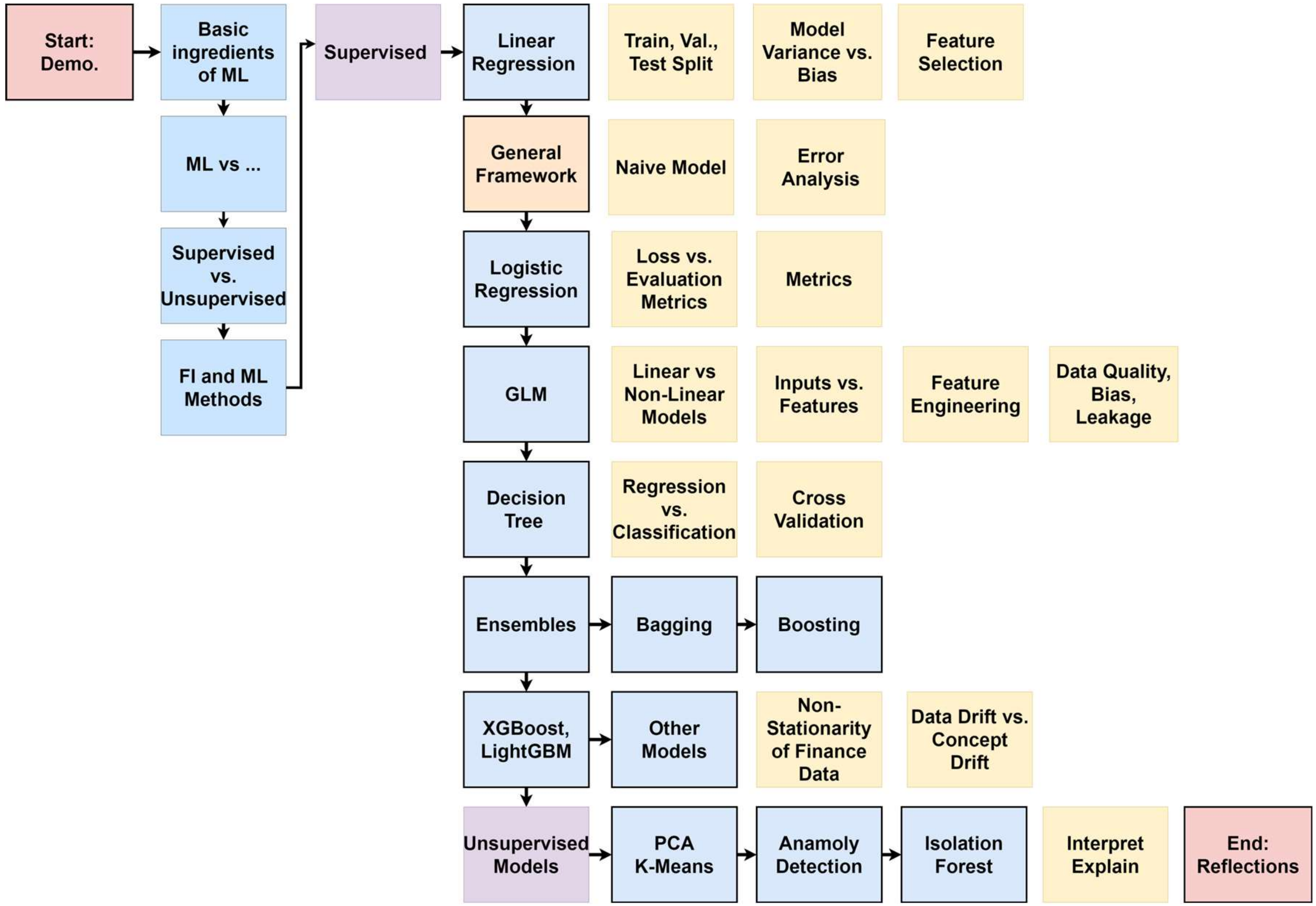
- What are these common data-types?
- What are the fundamental differences between these data-types?
  - Think about volume, stationarity, structure
- Name an example of each of these data-types in FIs



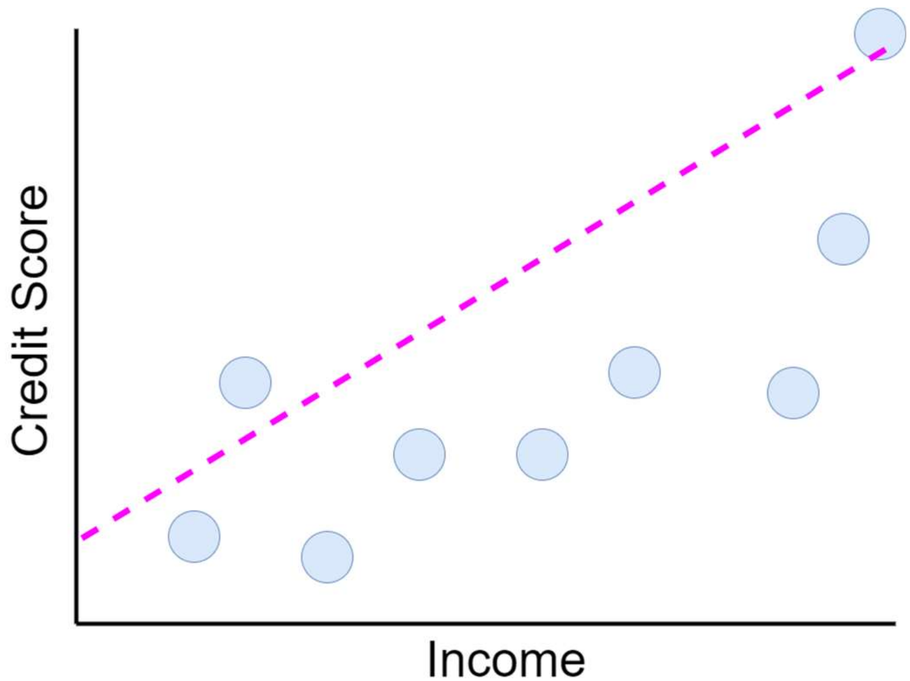
The Pillars of the Kings, The Fellowship of the Ring

Our Journey Today

Key Models & Concepts

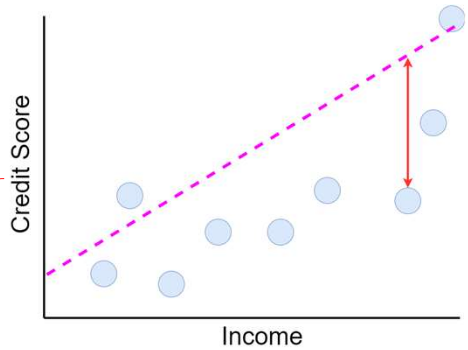


Linear regression



Framework: Linear Regression

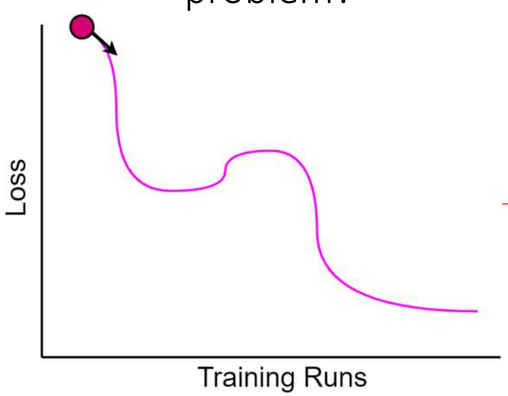
$$y = AX + B$$



Form	Loss
Train	Evaluate

Gradient descent, one of many ways to train/optimize.

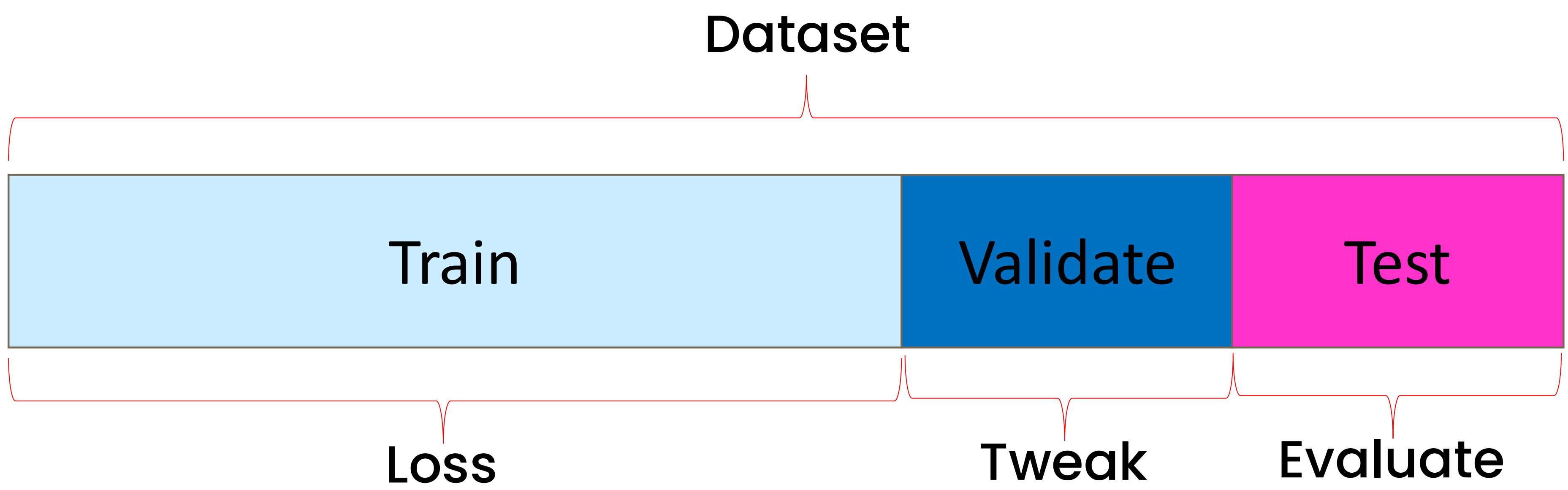
Can you spot a common problem?



Do these have to be the same?

Squared or Absolute or %?

Train, Validate, Test Split



Think of how a school examination is designed …

Same principles when selecting data for the splits

Sometimes, we are more interested in the parameters than the performance …

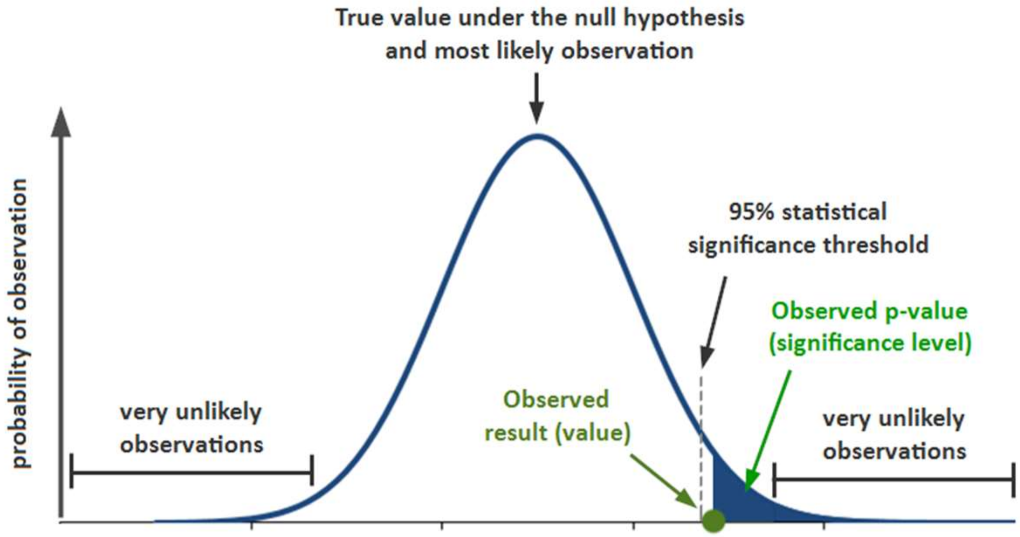
(usually not the focus of machine learning, but good to know)

$$y = 0.3X_1 + 0.5X_2 + 0.1$$

What is the null hypothesis?

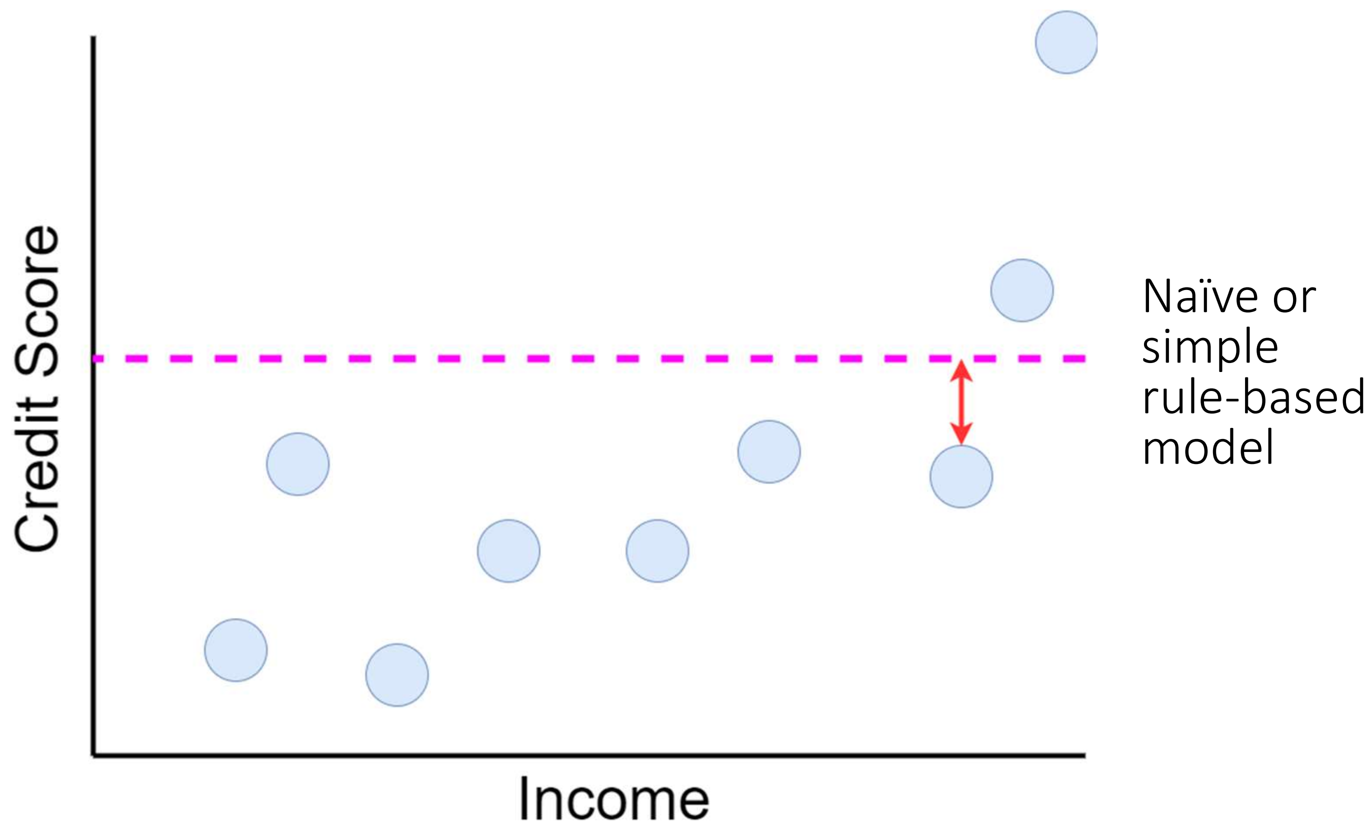
If the p-value associated with 0.3 is <0.05, what does it potentially imply?

Multi-collinearity (e.g., where 2 features are so closely related that you could potentially derive one from the other) can mess this up!

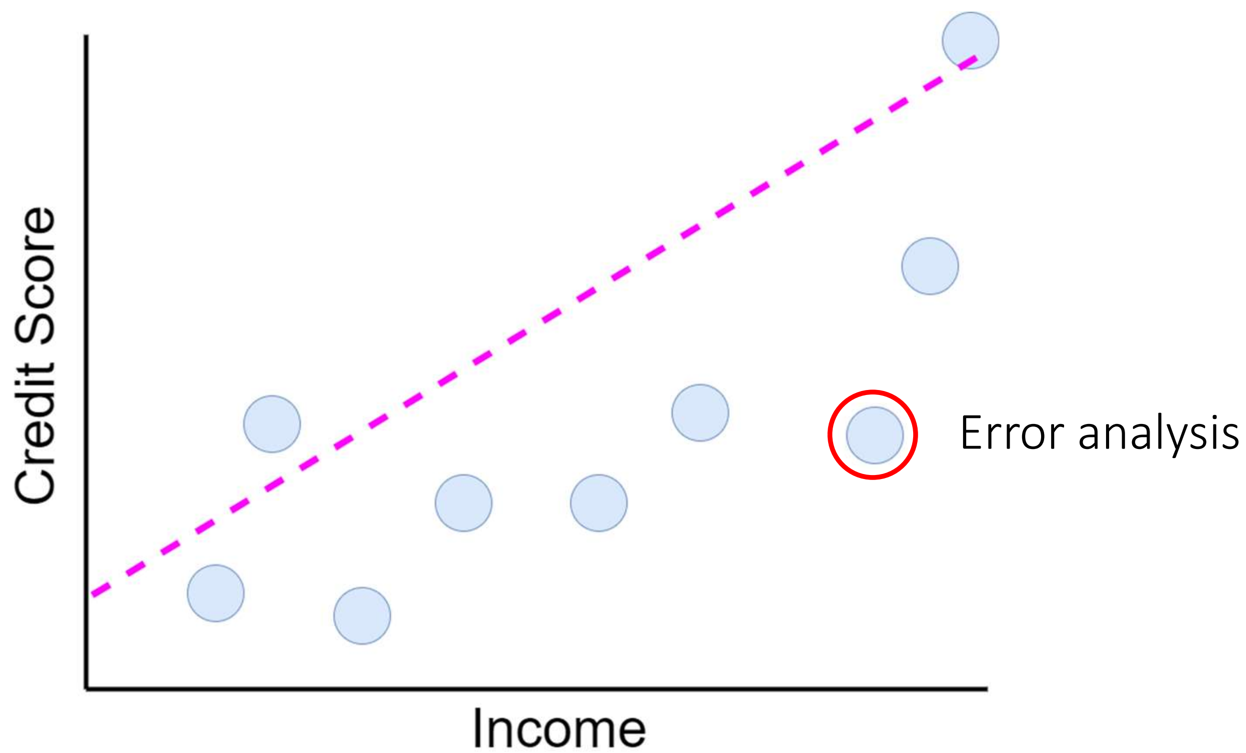


https://www.gigacalculator.com/calculators/p-value-significance-calculator.php

What is a good sanity check?

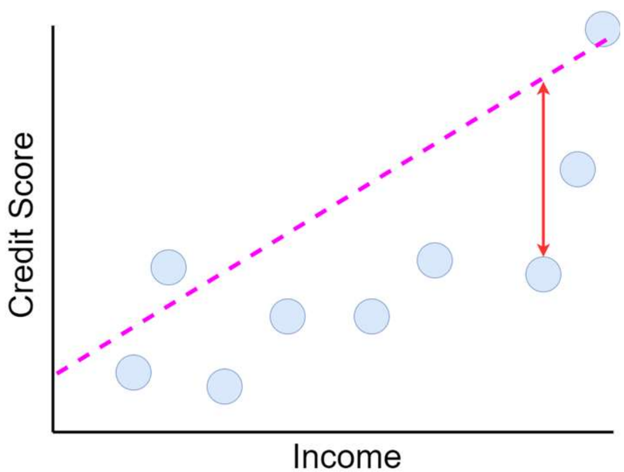


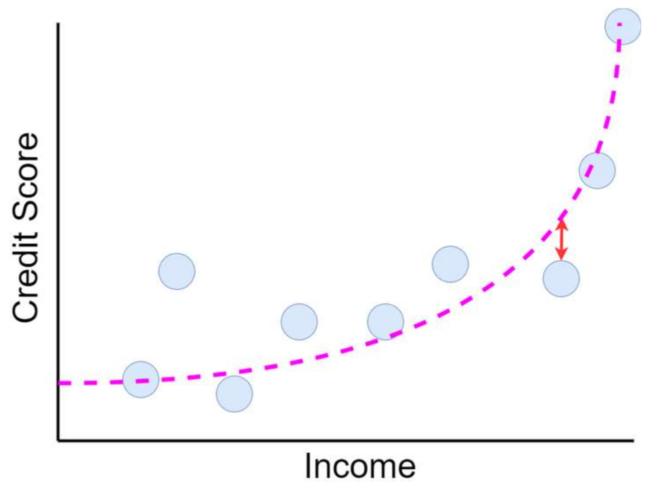
What is a good sanity check?

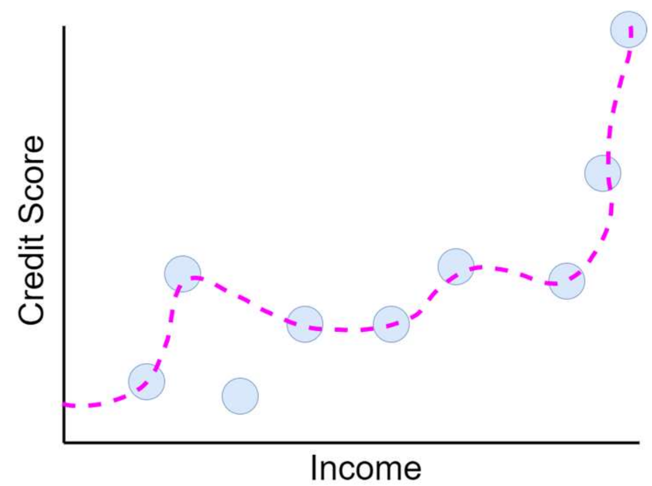


Model variance and bias

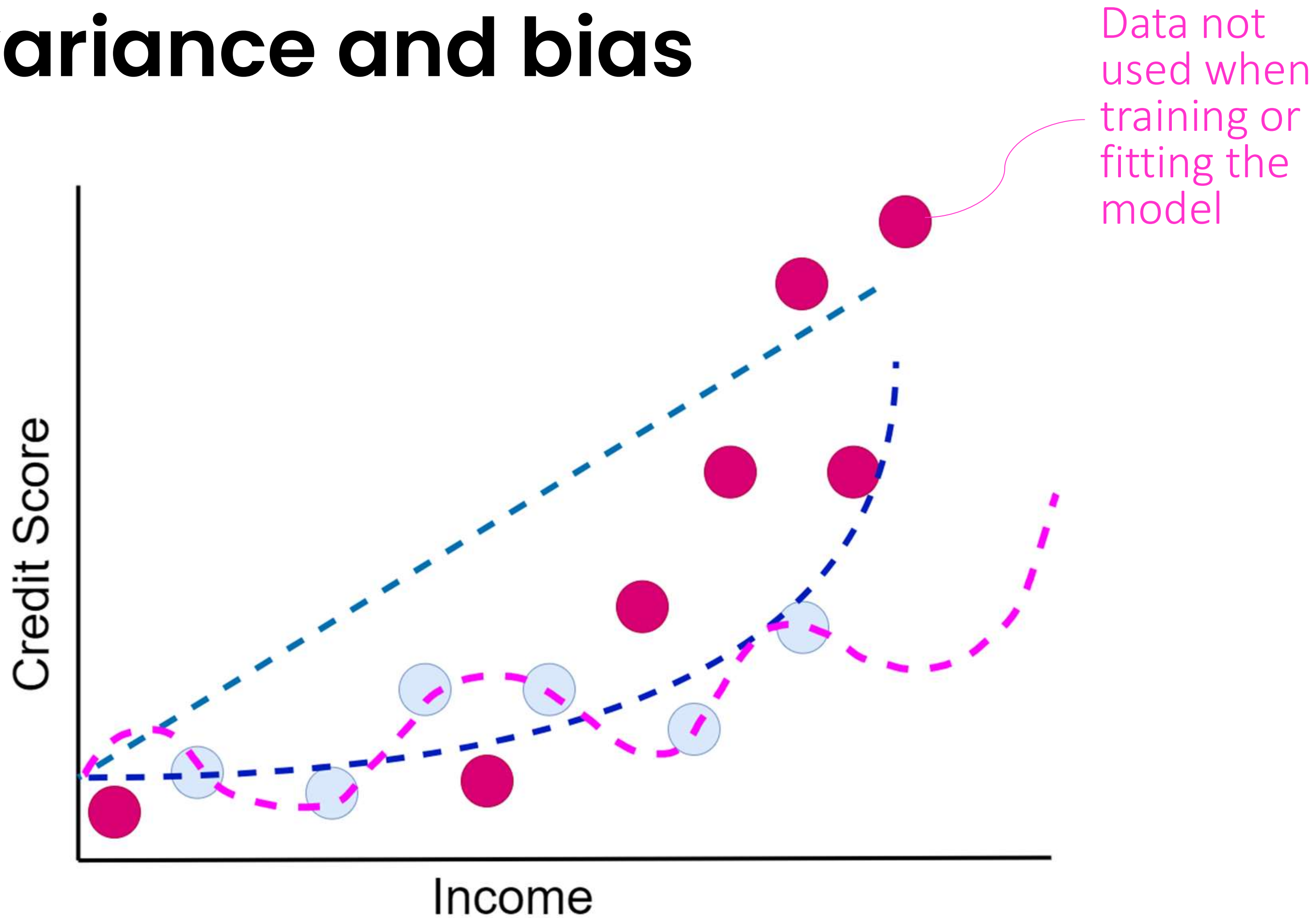
What is the issue with this model?







Model variance and bias



Model variance and bias

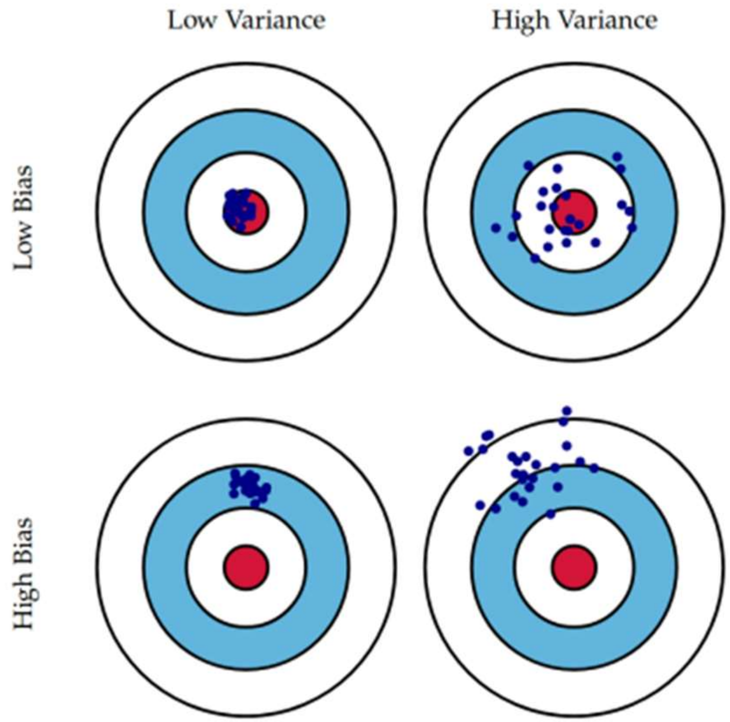


Fig. 1 Graphical illustration of bias and variance.
http://scott.fortmann-roe.com/docs/BiasVariance.html

Match them

The prediction error of a model comprises a (1) part that can be minimized and a (2) part that cannot be minimized.

The **reducible** part comprises model bias , which can lead to a model being (3); and model variance which can lead to a model being (4).

Overfit
Underfit
Reducible
Irreducible

Input or Feature selection and regularization

- Forward selection
- Backward selection
- LASSO

We usually deal with multi-collinear inputs here

What is LASSO? What is regularization?

Recall A, B in the linear equation → model parameters or weights w

What if we include in the loss?

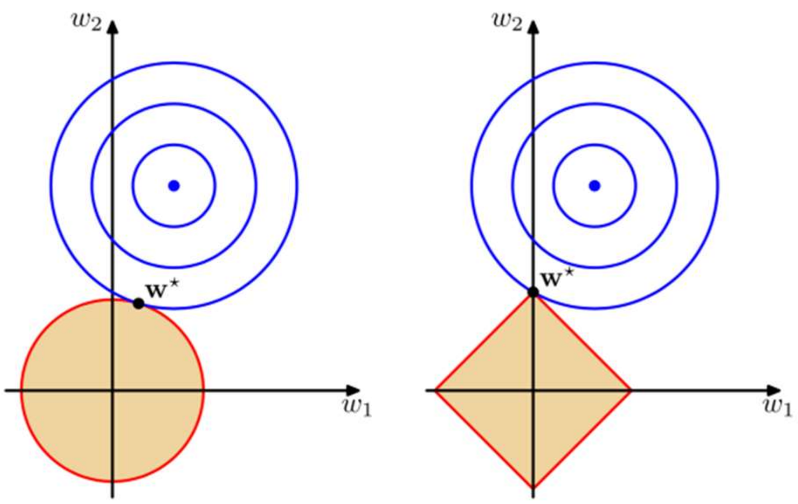


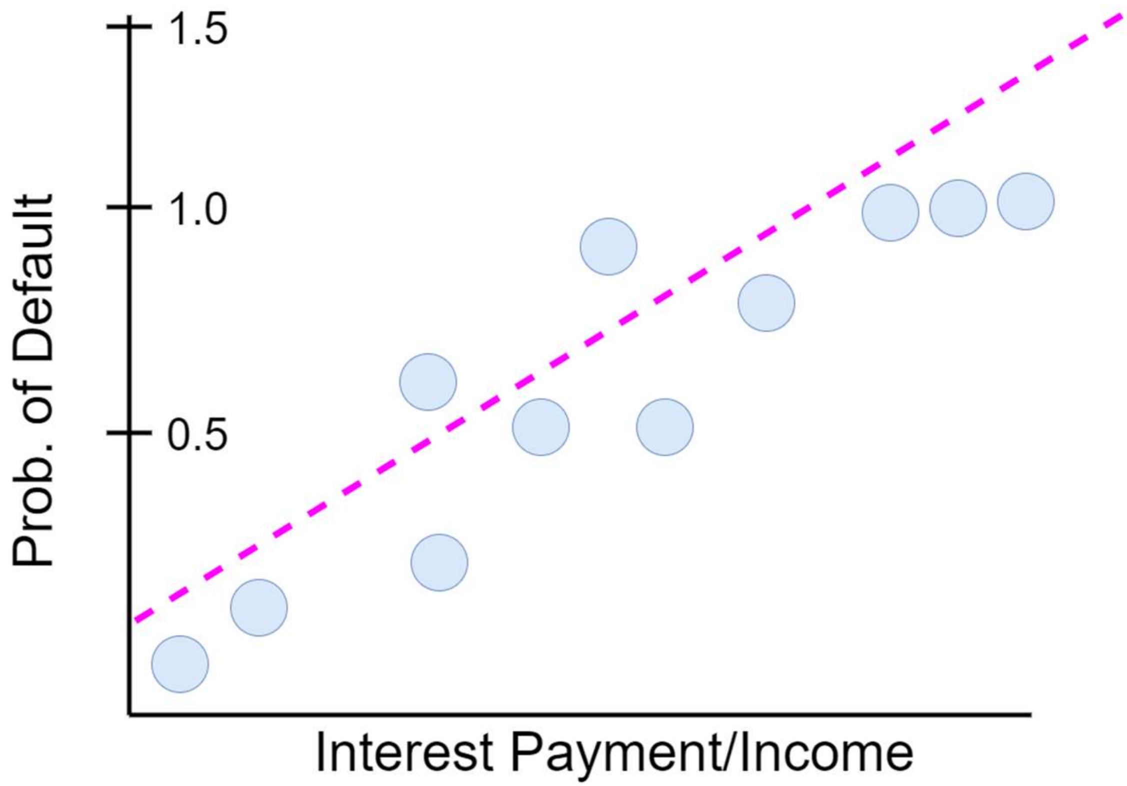
Figure from Machine learning: a probabilistic perspective, Kevin Murphy

- What if we wanted to model default probabilities instead of a credit score?

- What is the fundamental difference between the two quantities?

Logistic regression

$$Z = AX + B$$

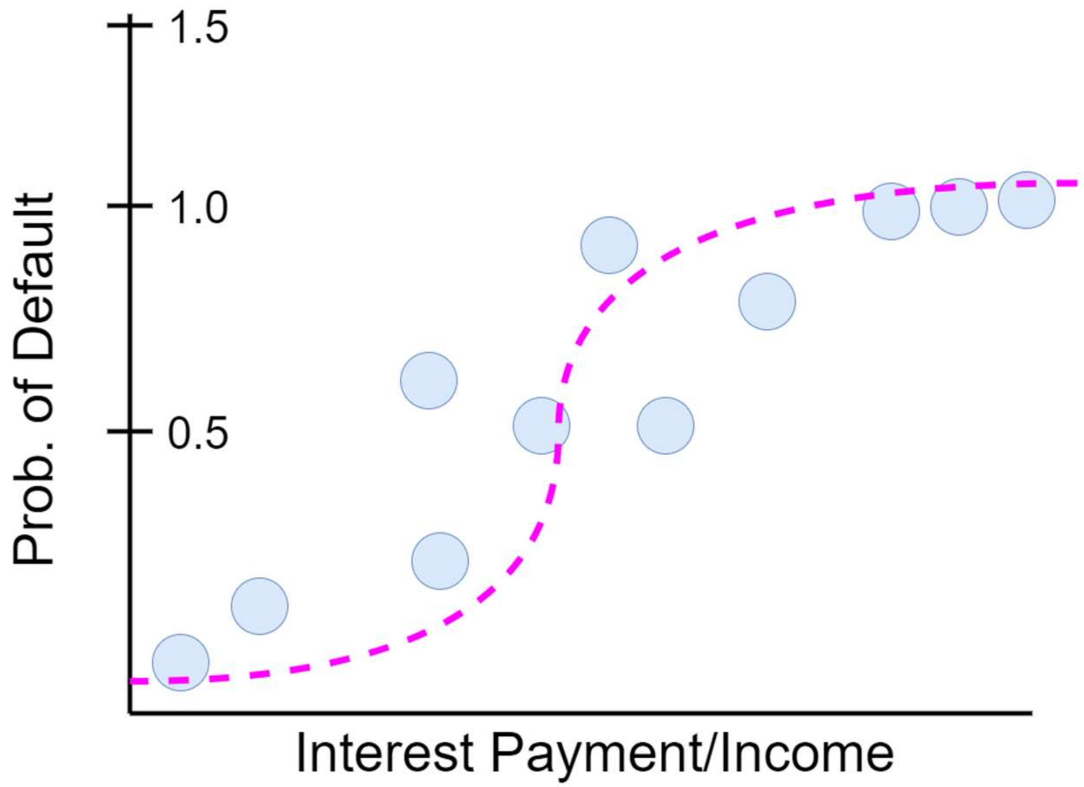


Logistic regression

$$Z = AX + B$$

$$P = \sigma(Z)$$

$$P = \frac{1}{1 + e^{-(AX+B)}}$$

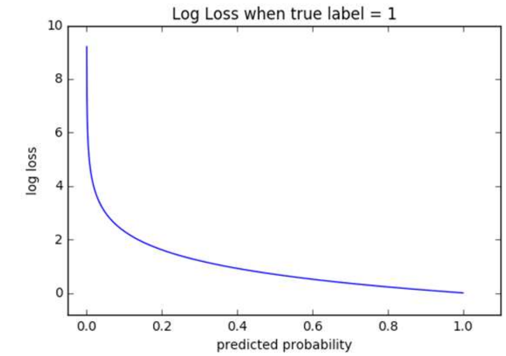


Framework: Logistic Regression

$$P = \frac{1}{1 + e^{-(AX+B)}}$$

Cross-Entropy

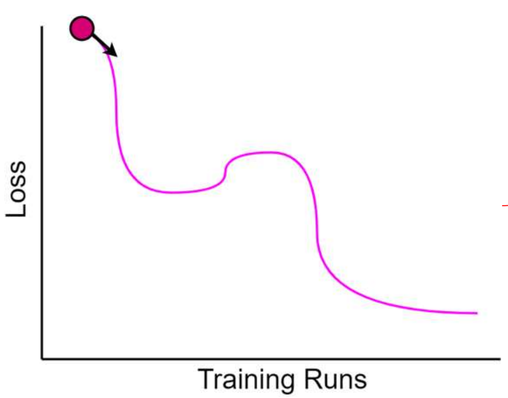
$$-Ylog(P) - (1 - Y)\log(1 - P)$$



Form	Loss
Train	Evaluate

Gradient descent not the only way

Also Max. Likelihood Est.

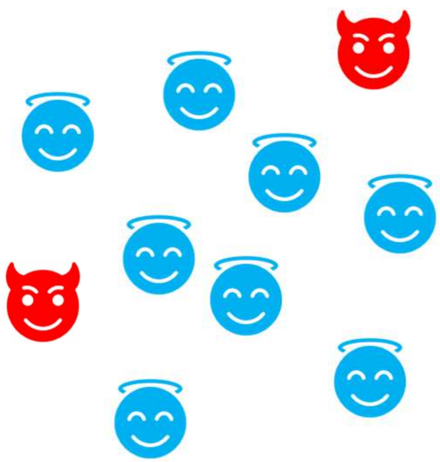


Accuracy, Recall, Precision, F1

Confusion matrix

What is a True Positive, False Positive, True Negative, False Negative?

(Indicate with TP, FP, TN, FN)

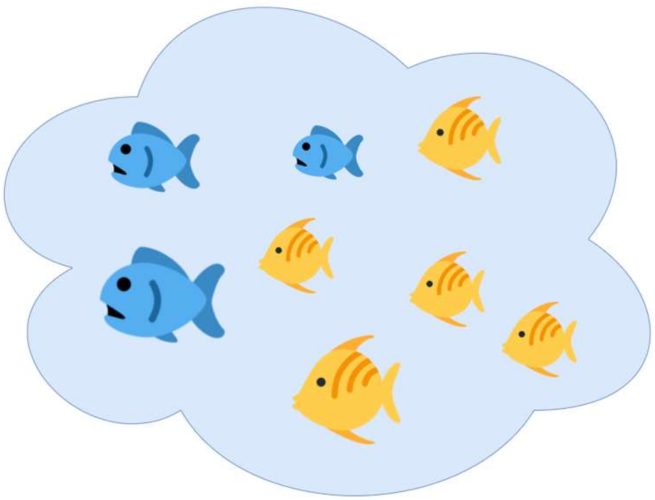


		Actual	
		Positive	Negative
Predicted	Positive		
	Negative		

Now, let’s assume a legitimate transaction is ‘1’ or ‘positive’, and assume you just naively predict ‘1’ all the time. Fill in the confusion matrix. What is the accuracy?

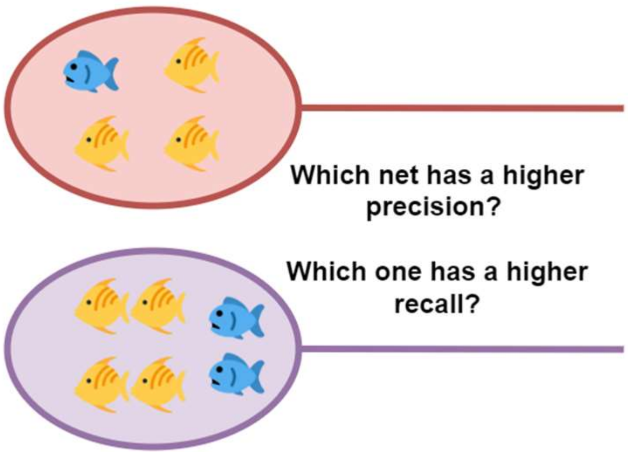
Accuracy is straightforward ….

Treat yellow fish as a '1'



What is the % accuracy for this dataset that any model needs to at least be better than?

Net = What you predict as '1'



$$\text{Precision} = \frac{tp}{tp + fp}$$
$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

If one is interested in predicting defaults (assume 1 for defaults), recall or precision?

Can you live with letting a fish get away?

If one is interested in predicting good leads (good leads as '1') for selling financial products to, recall or precision?

Can you live with letting a fish get away?

If we regard the detection of illicit entities or transactions as a '1', precision or recall?

Can you live with letting a fish get away?

From Linear and Logistic Regression to Generalised Linear Models (GLM)

$$Y = AX + B$$

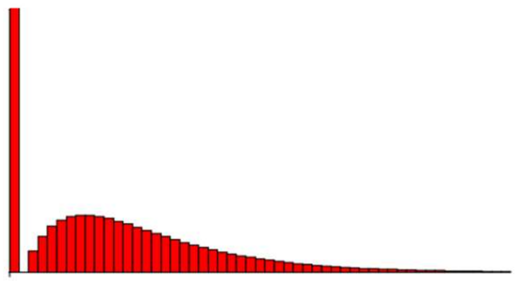
Random component that is normally distributed with some mean and variance

$$Y = \underbrace{LinkFunction}(Expected[Y]) + \varepsilon$$

Recall the σ function that we used when we went from linear to logistic regression

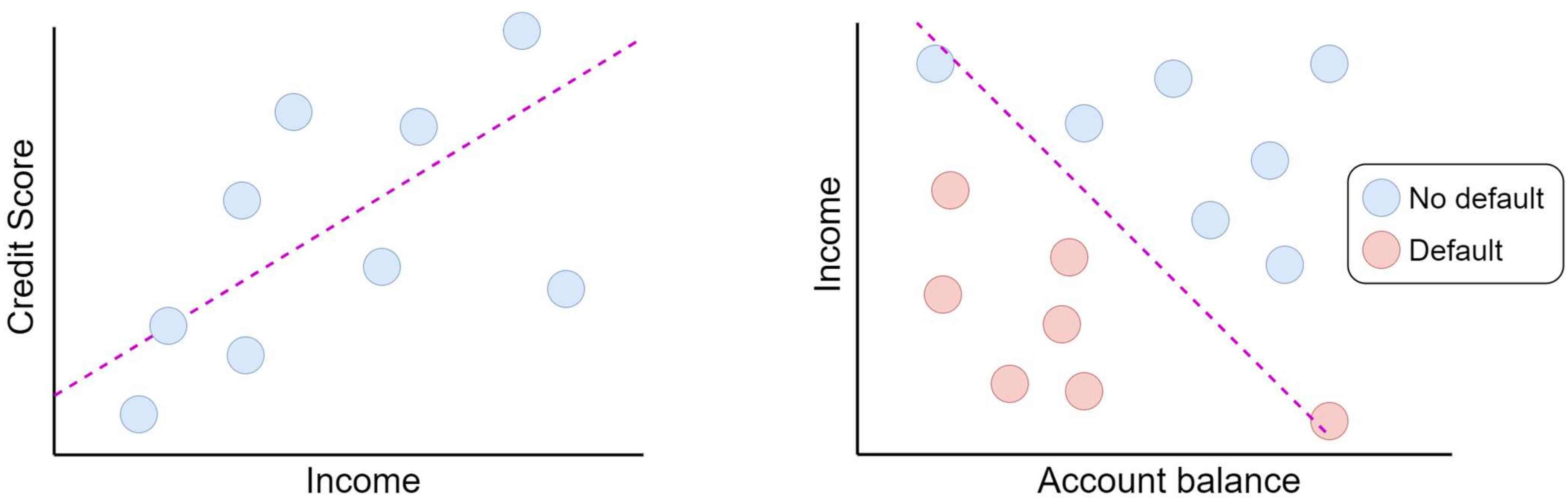
Other functions can be used to allow for non-linearities

E.g., Tweedie GLM used for modelling insurance claims experience

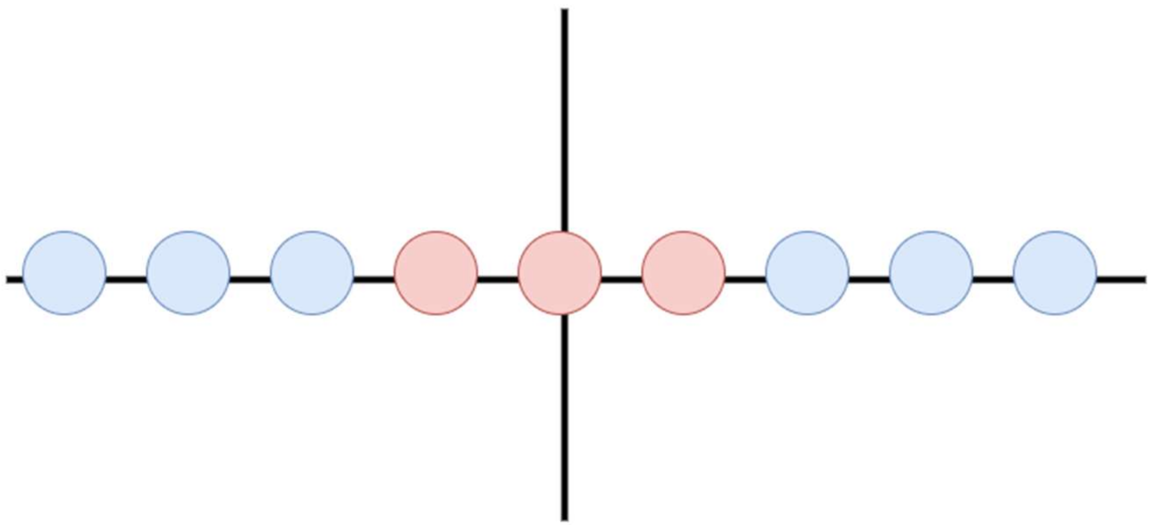


$$Y = LinkFunction(Expected[Y]) + \varepsilon$$

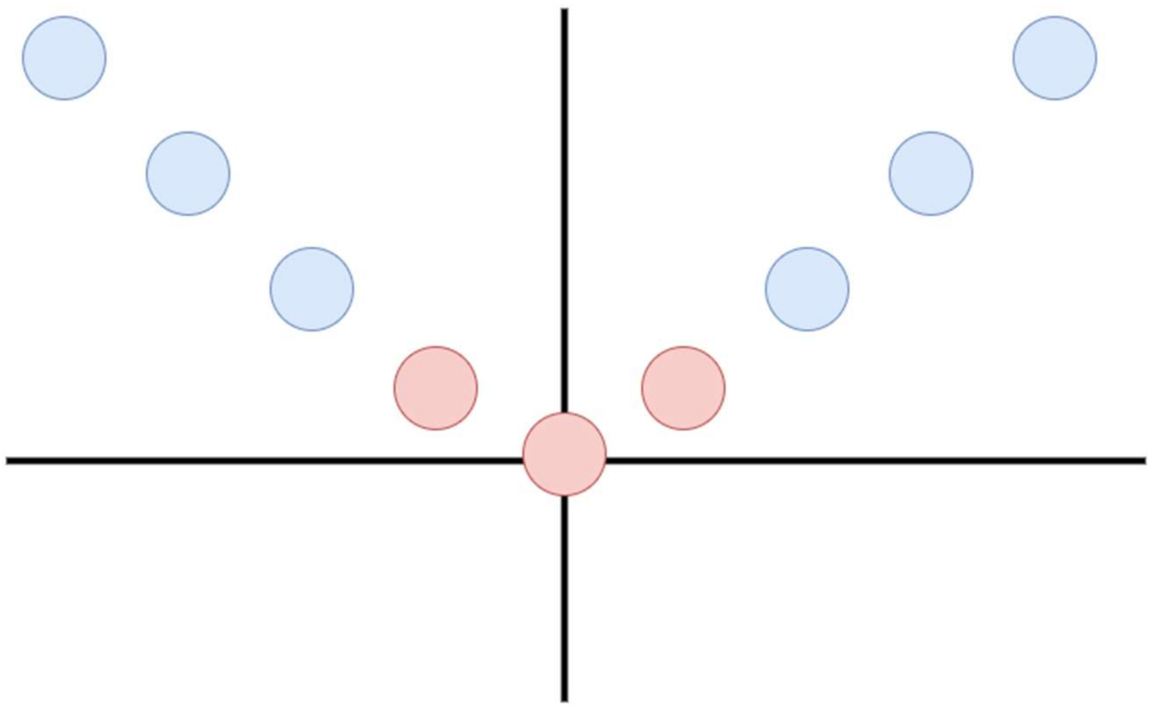
Regression vs. Classification Redux



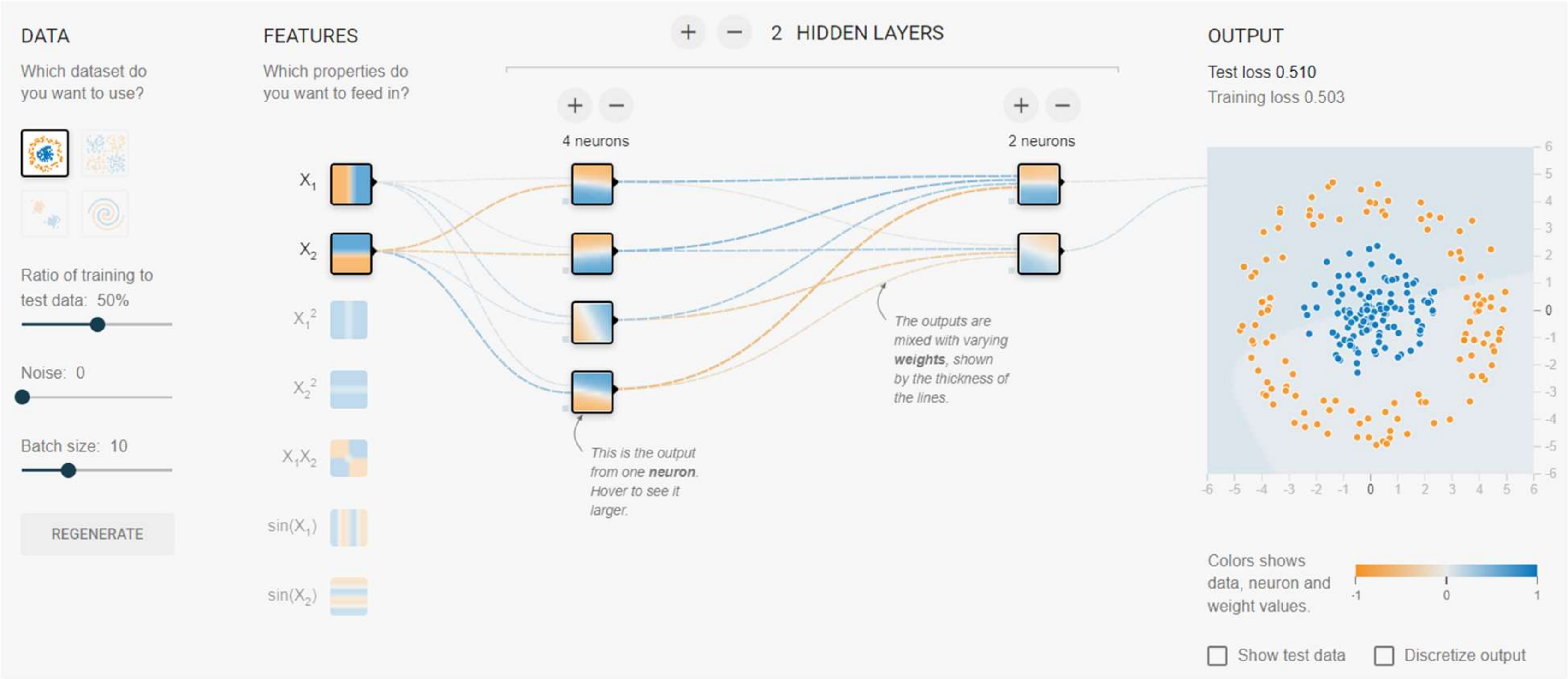
What if the data for a classification task looks like this?



What if the data for a classification task looks like this?



Linear vs. Non-linear Problems



https://playground.tensorflow.org/

Feature Engineering

Not only about selecting right inputs:

- Transform non-linear to linear problem (we saw this)
- Capture interactions
    - *E.g., think about BMI and TDSR vs their constituents*
- Utilize unstructured inputs
    - *E.g., think about tabular information vs. images, text, audio, networks*

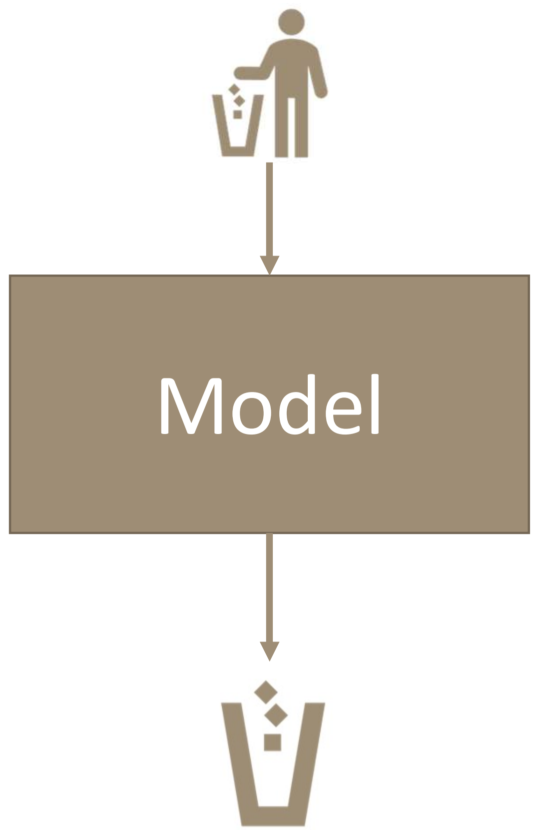
Data quality

Addressing all of these is a pipe-dream.

But important to know if these exist in our data.

- **Accuracy**
  - E.g., mis-labelled illicit transactions
- **Completeness**
  - E.g., omission of transactions stored in another banking system
- **Consistency**
  - E.g., unclear instructions when designating a loan as defaulted
- **Currency**
  - E.g., characteristics/distribution of fraudulent transactions changing over time due to change in tech. and consumer behaviour

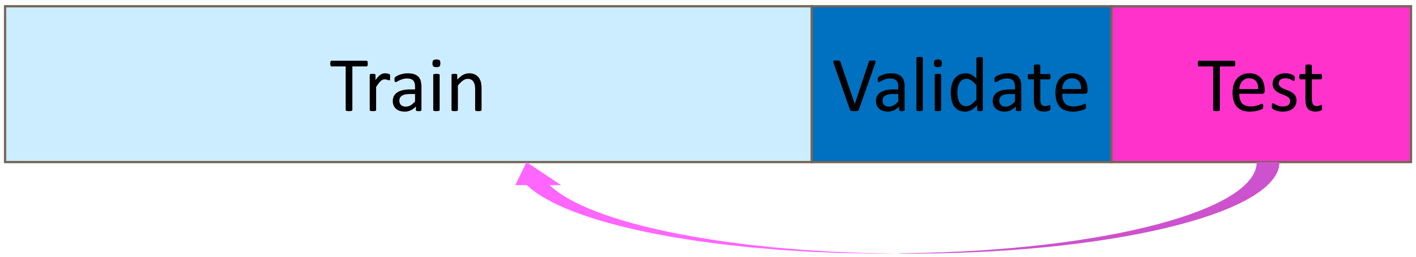
Data Bias



Examples of harmful data bias

- **Distribution** bias
  - Personal attributes
- **Representation** bias
  - Match general population but under-represent certain segments
- **Implicit** bias
  - Not all bias are obvious, e.g., gender vis-à-vis income vis-à-vis location vis-à-vis race
- **Labelling** bias
  - Even experts label things differently, e.g., 2nd opinions?

Data leakage



- Very common even for random train, validation test splits
  - Using total time customer spent in bank to predict customer purchase intent so as to act on it while customer still in bank
  - Use data before t to predict prob. of default at $t + 1$, data before t includes post-default adjustments
  - Predicting illicit transactions using complete incident reports
  - Using mean and standard deviation of entire dataset to scale/normalize data

There is a whole paper on this at https://reproducible.cs.princeton.edu/

Data leakage

Paper	Muchlinski et al.	Colaresi and Mahmood	Wang	Kaufman et al.
Claim	Random Forests model drastically outperforms Logistic regression models	Random Forests models drastically outperform Logistic regression model	Adaboost and Gradient Boosted Trees (GBT) drastically outperform other models	Adaboost outperforms other models
Error	**[L1.2] Pre-proc. on train-test** (Incorrect imputation)	**[L1.2] Pre-proc. on train-test** (Incorrect reuse of an imputed dataset)	**[L1.2] Pre-proc. on train-test**. (Incorrect reuse of an imputed dataset) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)	**[L2] Illegitimate features** (Data leakage due to proxy variables) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)
Impact	Random Forests perform no better than Logistic Regression	Random Forests perform no better than Logistic Regression	Difference in AUC between Adaboost and Logistic Regression drops from 0.14 to 0.01	Adaboost no longer outperforms Logistic Regression. None of the models outperform a baseline model that predicts the outcome of the previous year
Discussion	Impact of the incorrect imputation is severe since 95% of the out-of-sample dataset is missing and is filled in using the incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Use several proxy variables for the outcome as predictors (e.g., *colwars, cowwars, sdwars*, all proxies for civil war), leading to near perfect accuracy

https://reproducible.cs.princeton.edu/

Data leakage

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Medicine	Bouwmeester et al.	2012	71	27	No train-test split
Neuroimaging	Whelan et al.	2014	—	14	No train-test split; Feature selection on train and test set
Autism Diagnostics	Bone et al.	2015	—	3	Duplicates across train-test split; Sampling bias
Bioinformatics	Blagus et al.	2015	—	6	Pre-processing on train and test sets together
Nutrition research	Ivanescu et al.	2016	—	4	No train-test split
Software engineering	Tu et al.	2018	58	11	Temporal leakage
Toxicology	Alves et al.	2019	—	1	Duplicates across train-test split
Satelitte imaging	Nalepa et al.	2019	17	17	Non-independence between train and test sets
Clinical epidemiology	Christodoulou et al.	2019	71	48	Feature selection on train and test set
Tractography	Poulin et al.	2019	4	2	No train-test split
Brain-computer interfaces	Nakanishi et al.	2020	—	1	No train-test split
Histopathology	Oner et al.	2020	—	1	Non independence between train and test sets
Computer security	Arp et al.	2020	30	30	No train-test split; Pre-processing on train and test sets together; Illegitimate features; others

https://reproducible.cs.princeton.edu/

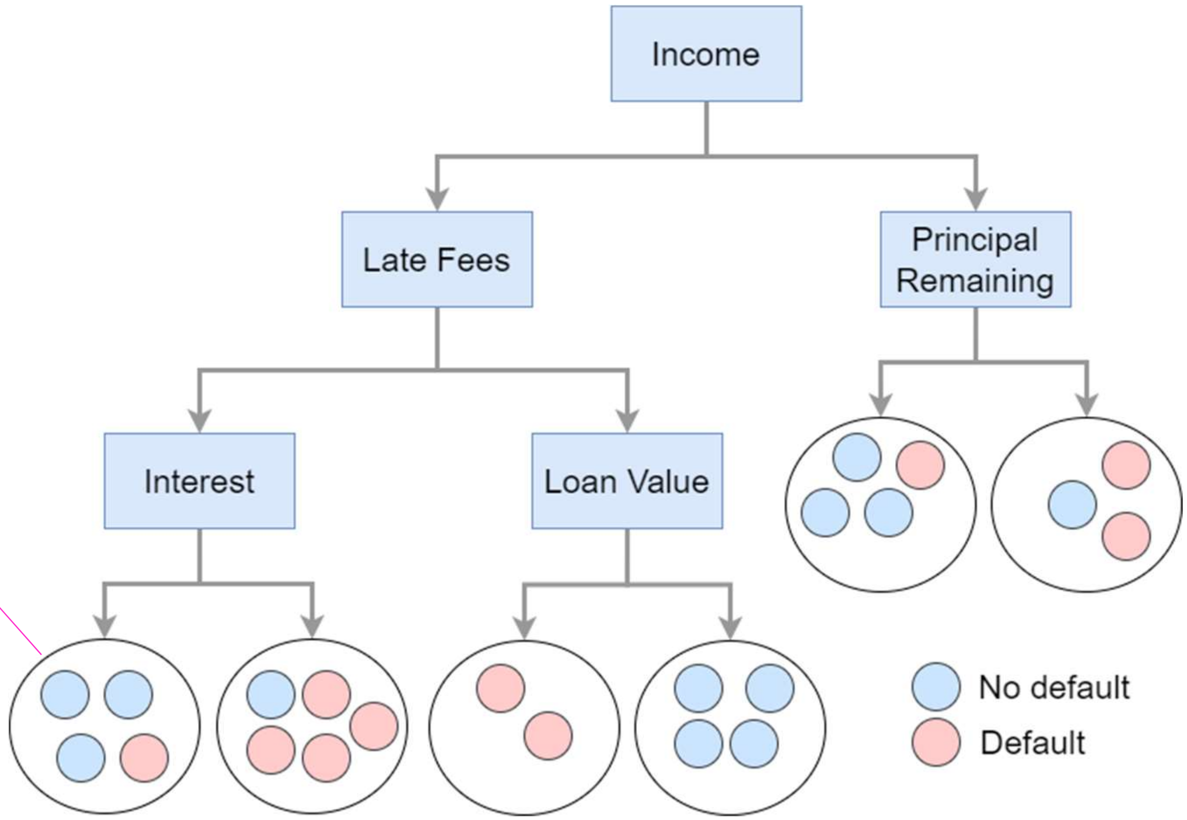
What is a good indication of data issues?

- Extreme results
  - Especially when compared to a naïve model

- Both ways
  - Too good – data leakage, evaluation errors ….
  - Too bad – dirty data, evaluation errors …

Decision Tree

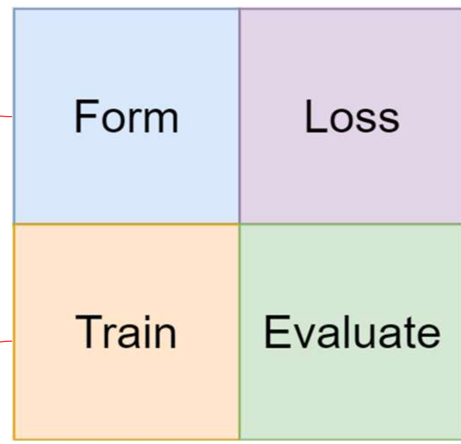
What is the class?



Framework: Decision Tree

What would be a good criteria to split?

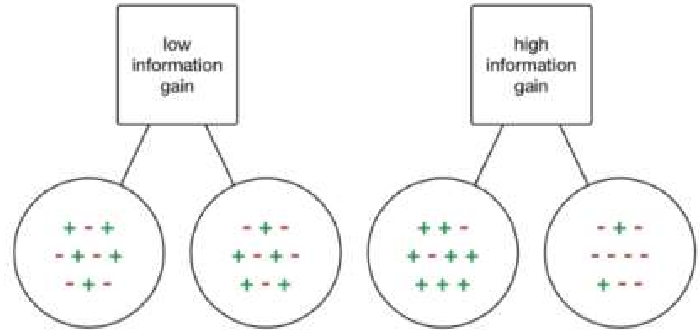
What is the form?

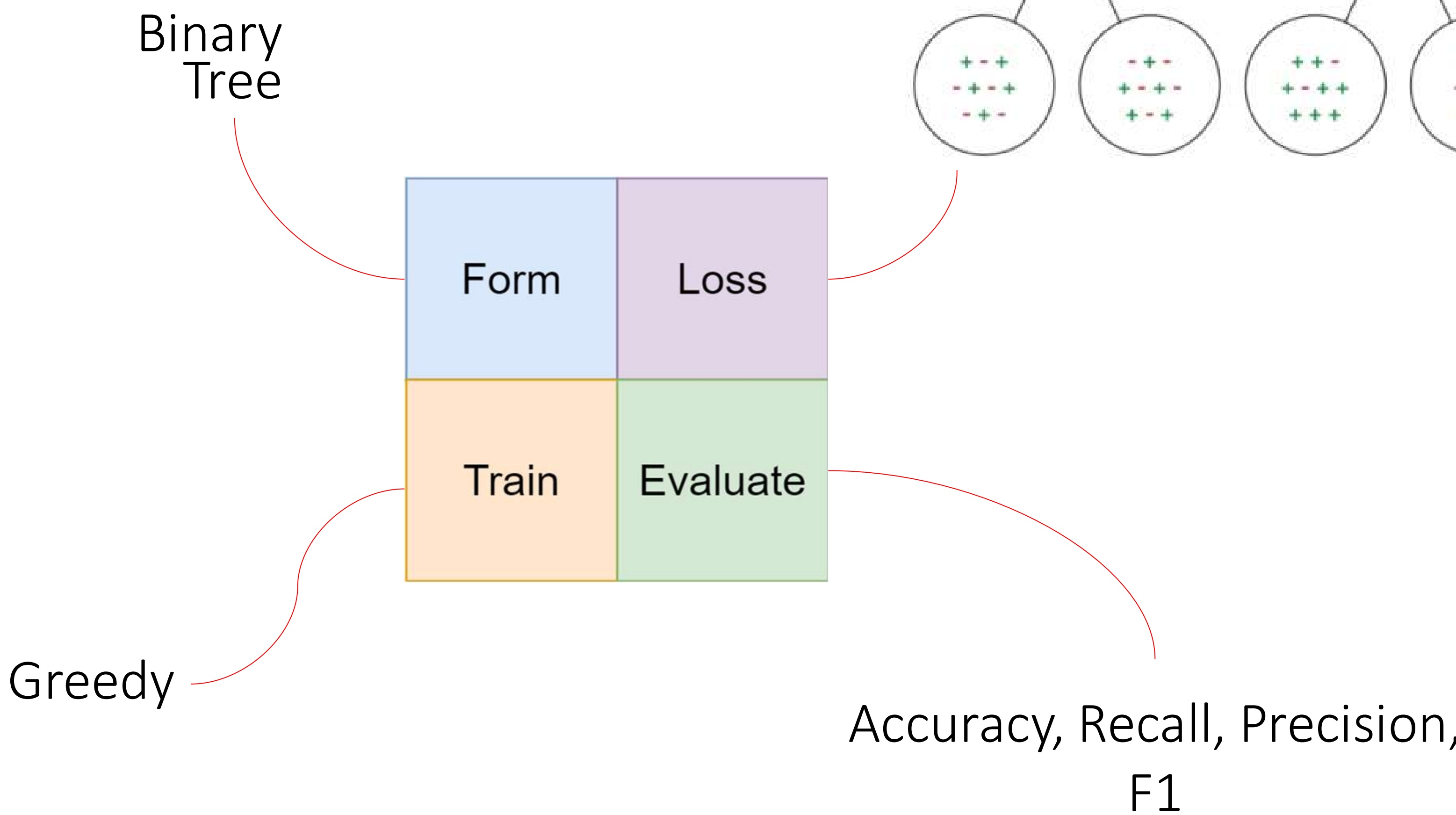


Given the splitting criteria, how could one build the tree?

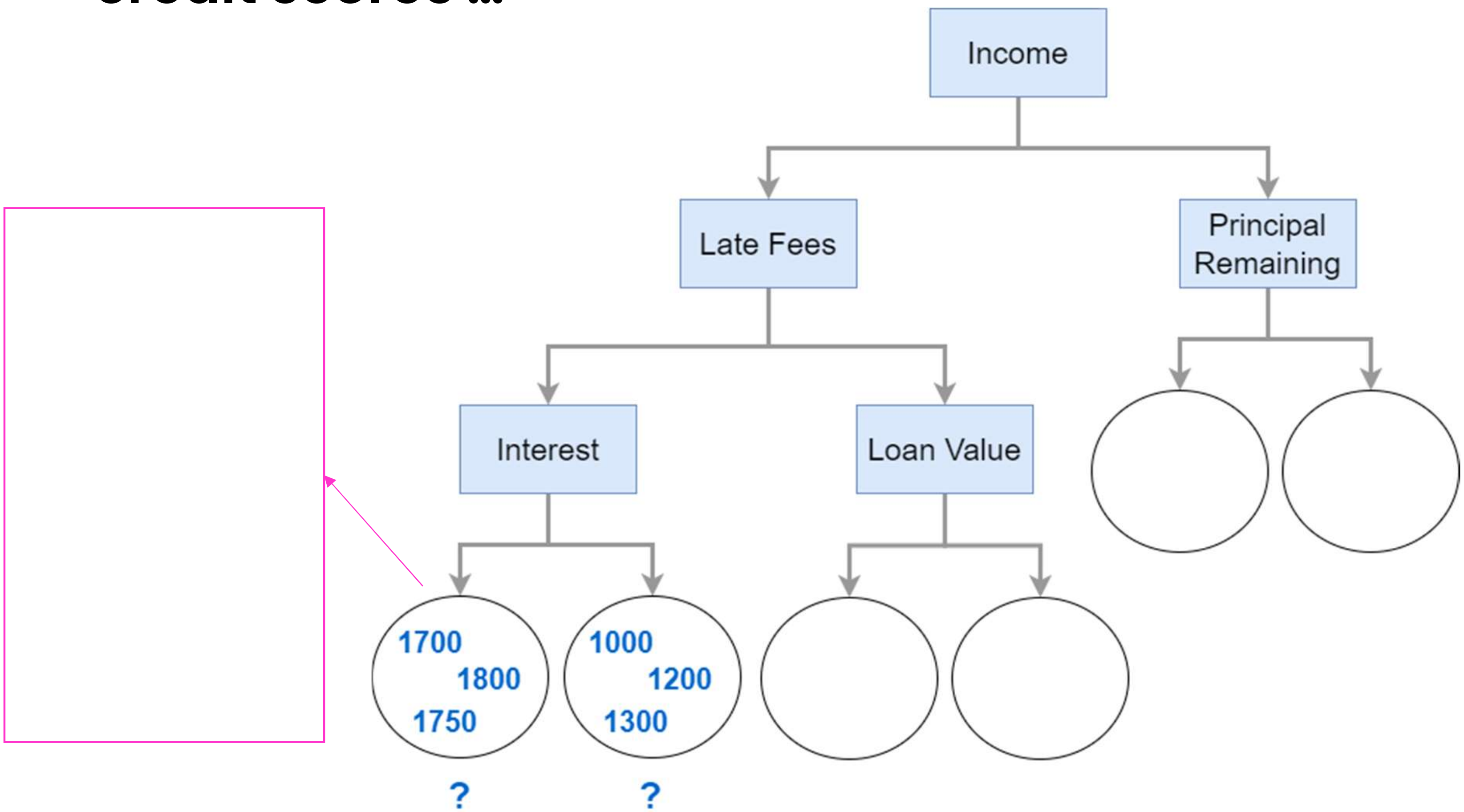
Recall what we used for logistic regression

Framework: Decision Tree

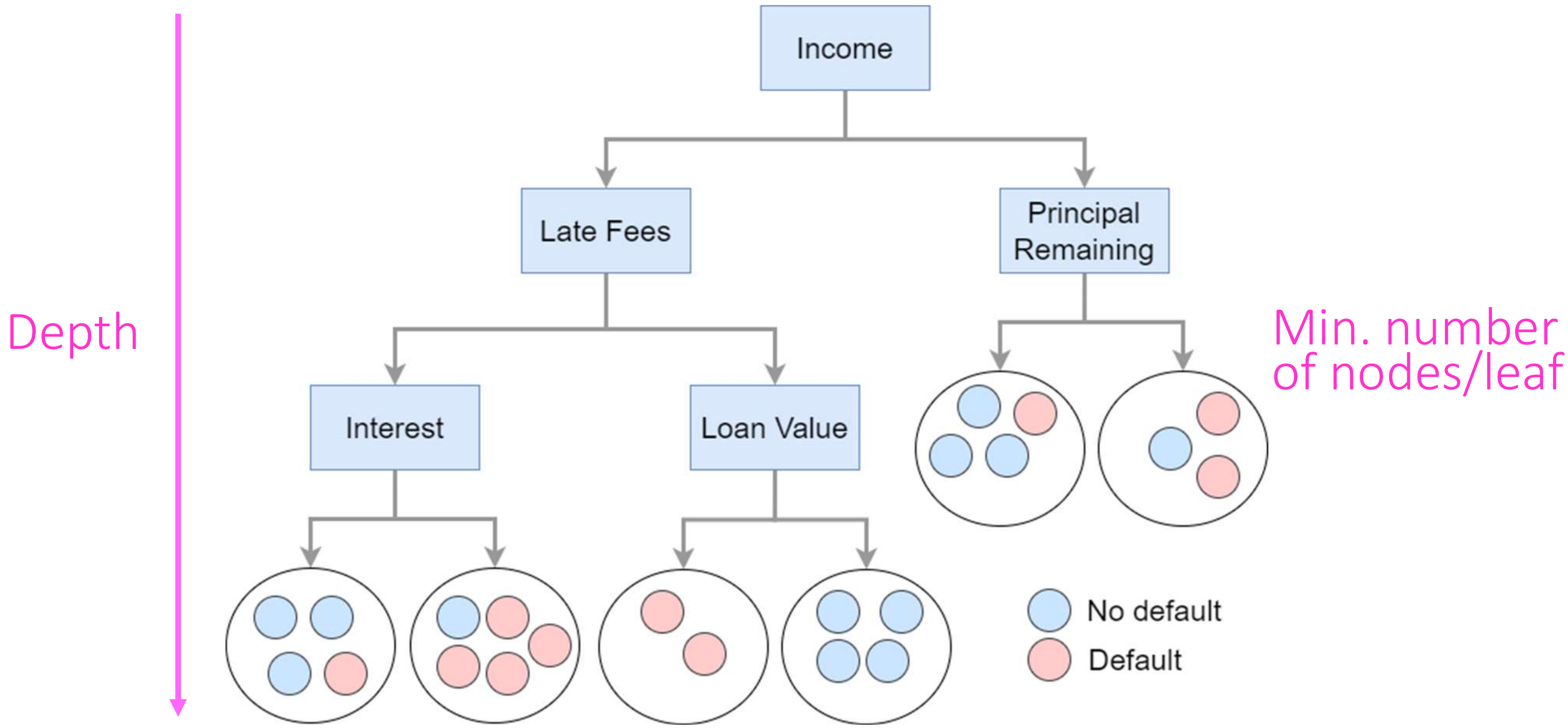




Now think about how this would work for regression on credit scores ...



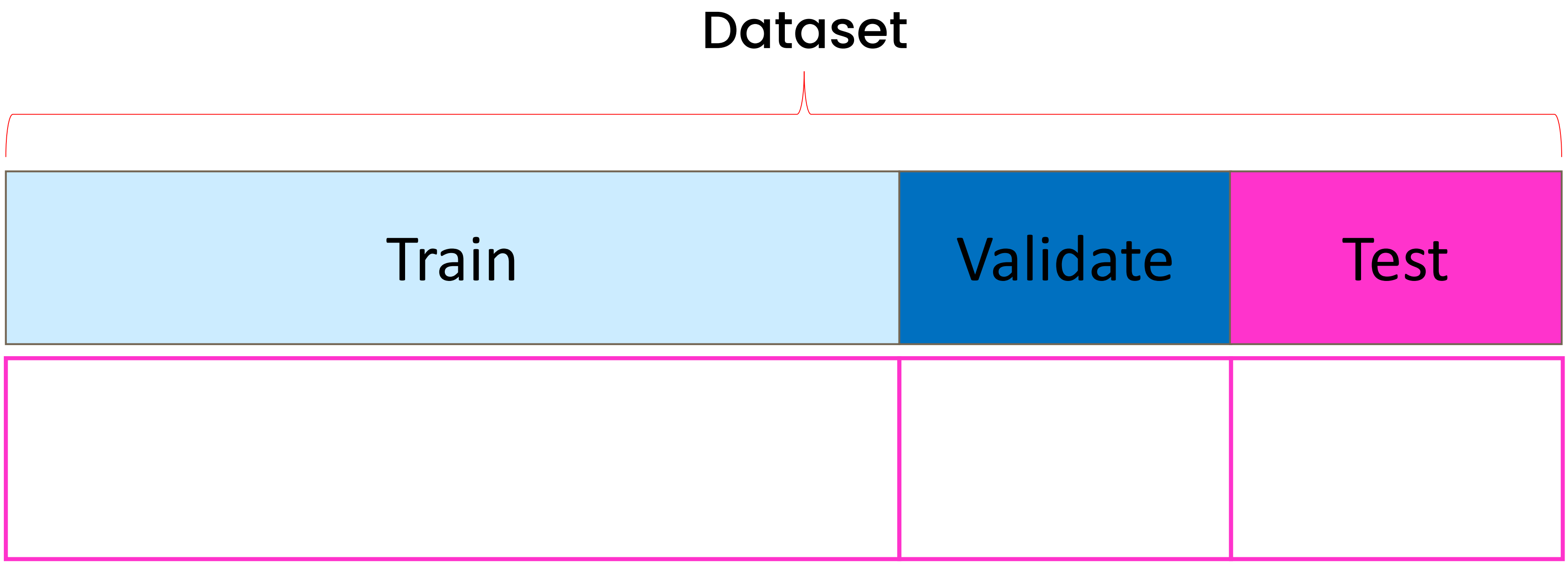
Hyperparameters



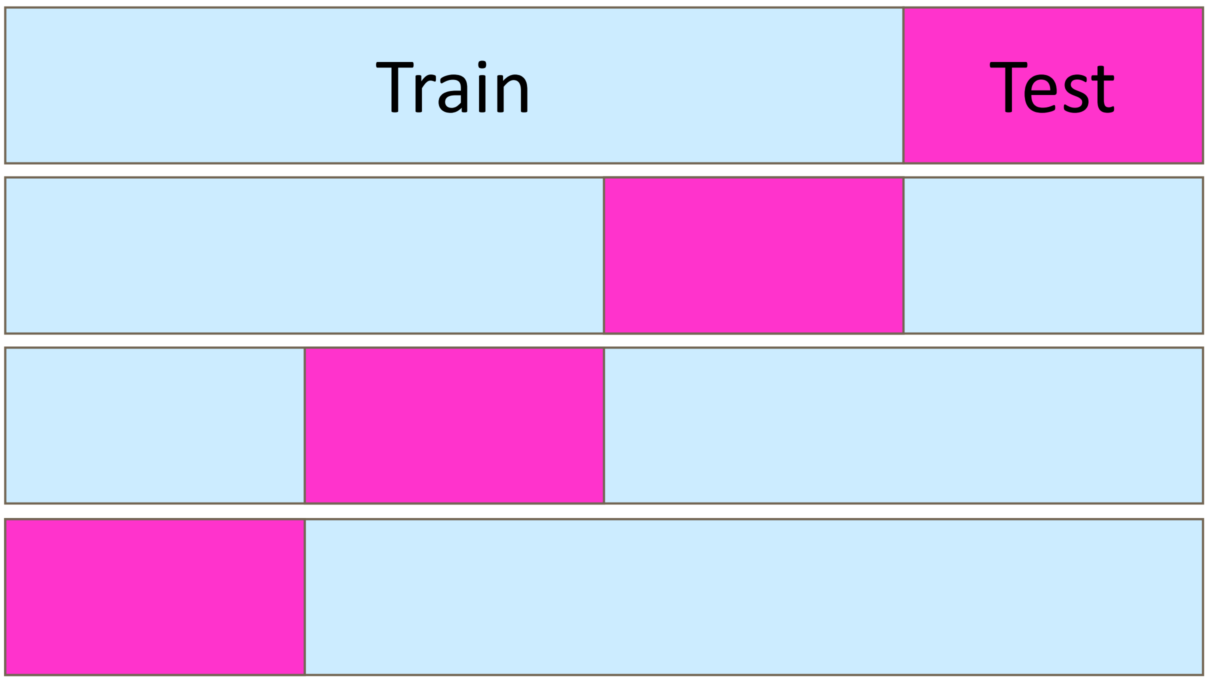
class `sklearn.tree.`**`DecisionTreeClassifier`**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0*) ¶

Discuss:
What determines parameters?
What determines hyperparameters?

Where should you tune hyper-parameters?

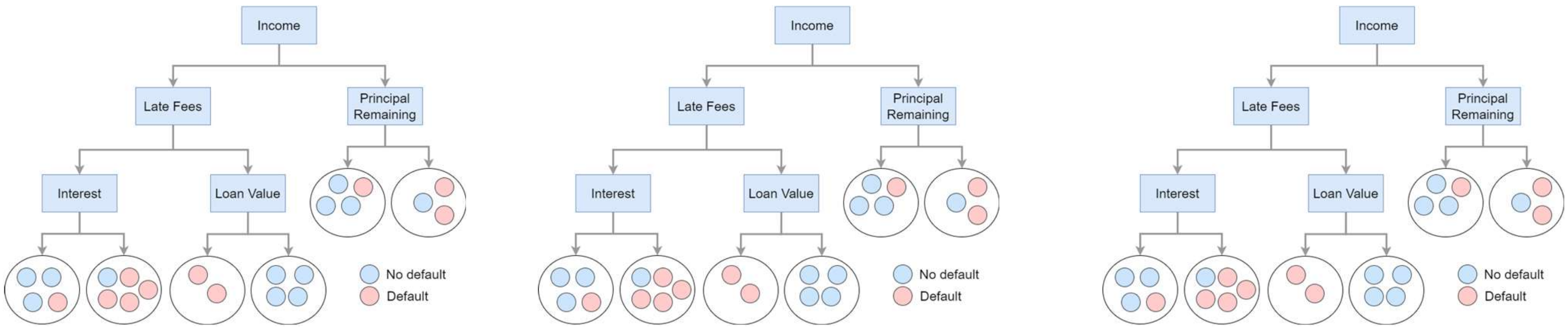


Cross Validation

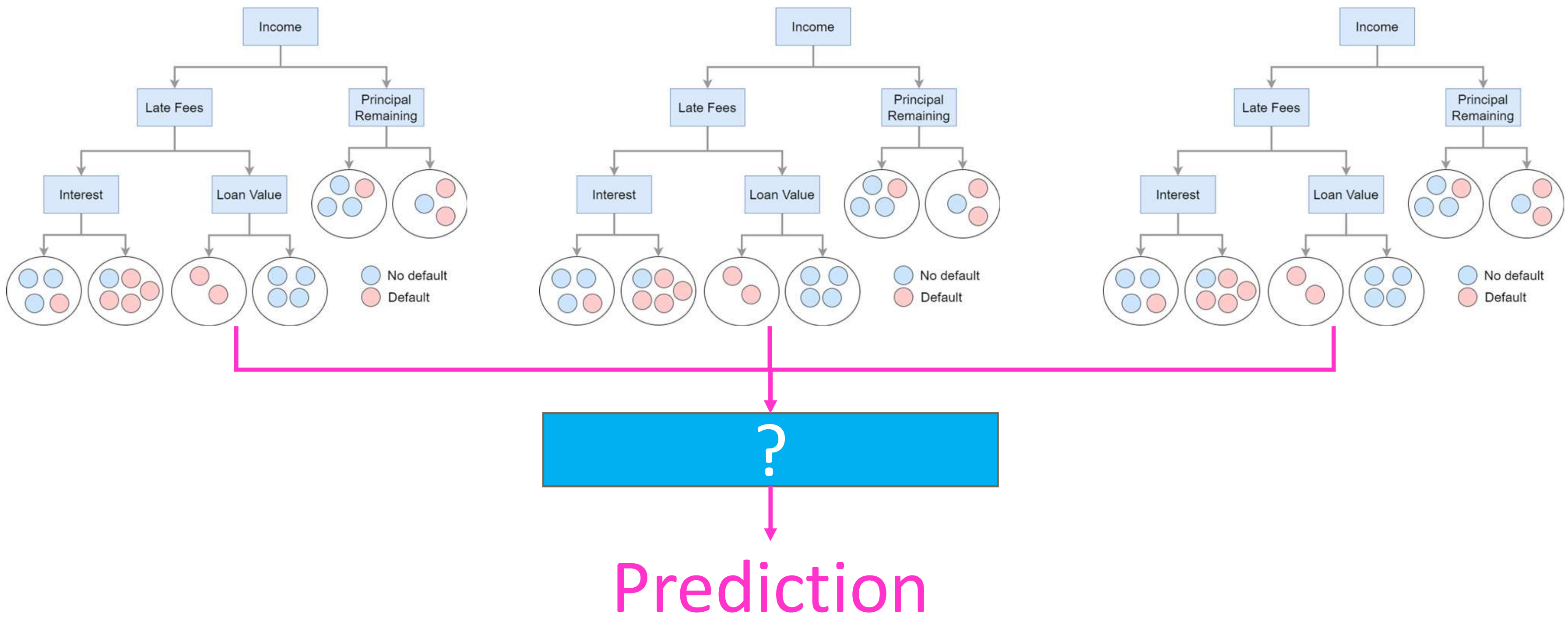


What are the advantages and disadvantages, compared with a simple train, validate, test split?

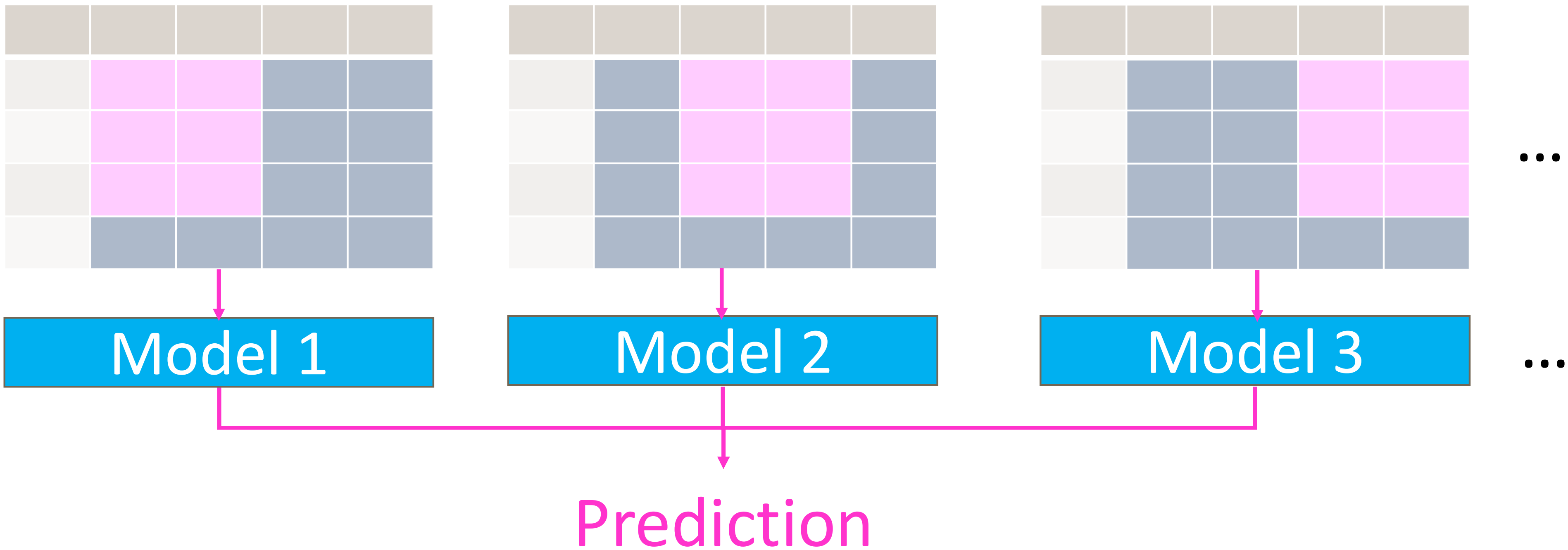
Why stop at one decision tree?



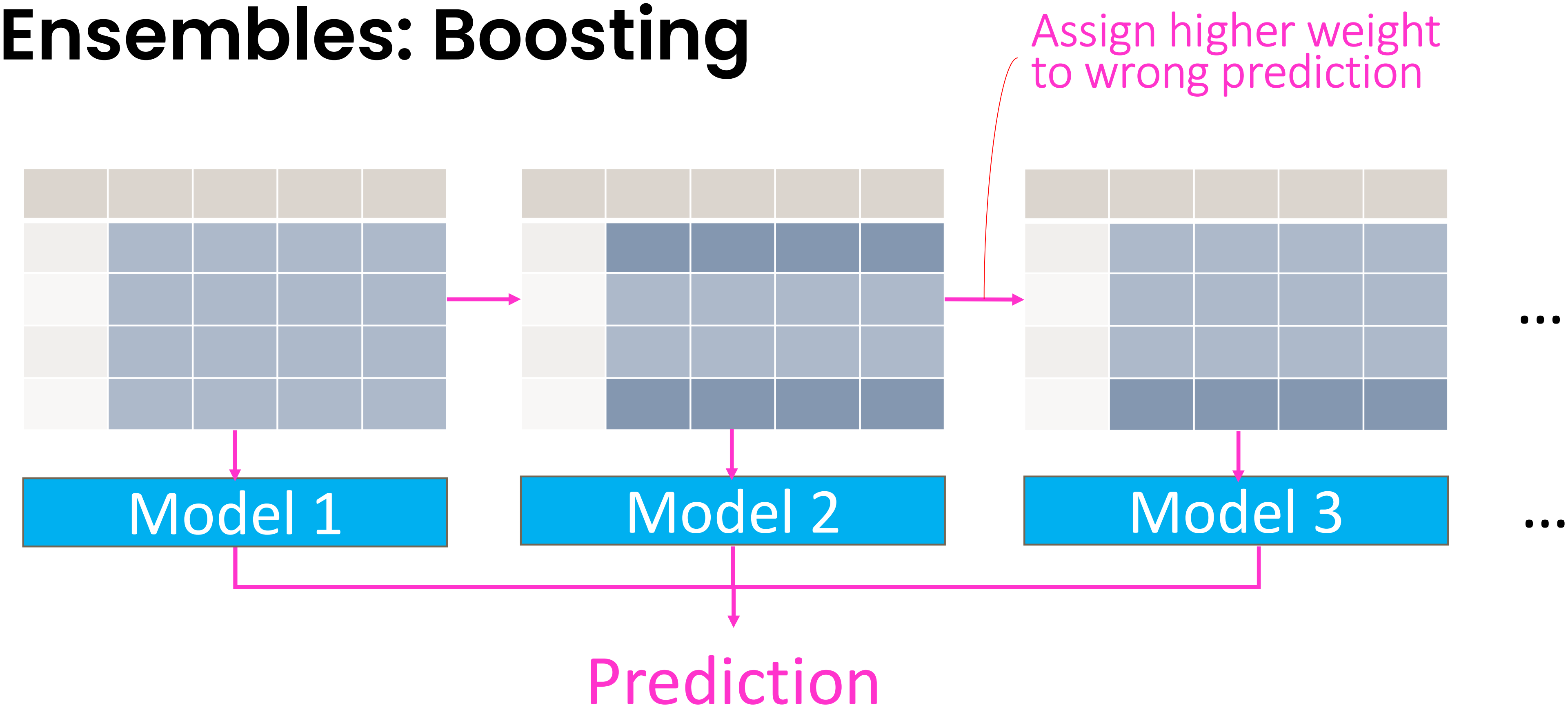
Ensembles


Income
Late Fees
Principal Remaining
Interest
Loan Value
No default
Default
Income
Late Fees
Principal Remaining
Interest
Loan Value
No default
Default
Income
Late Fees
Principal Remaining
Interest
Loan Value
No default
Default
?
Prediction

Ensembles: Bagging



Ensembles: Boosting



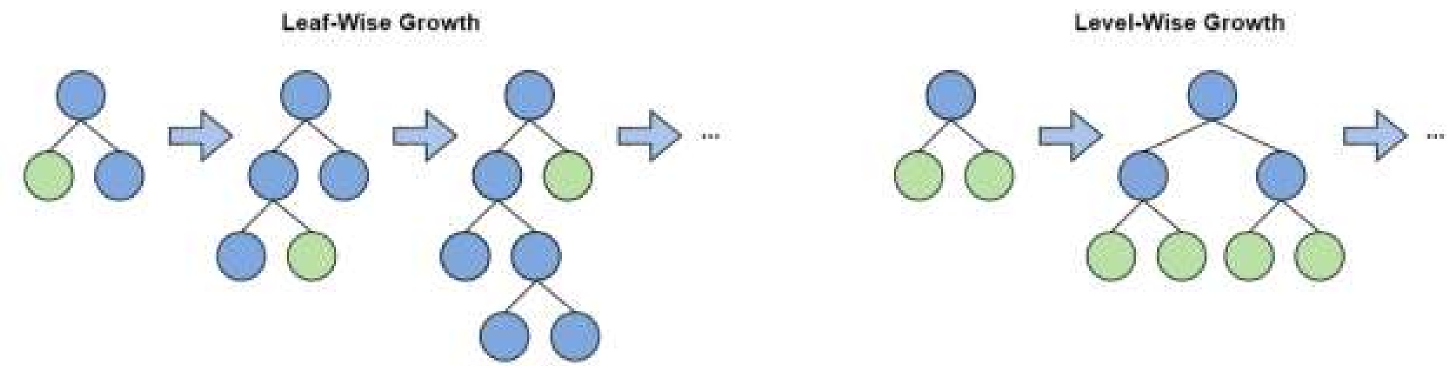
XGBoost

XGBoost is basically based on the idea of boosting, but with some additional math and optimization

For the curious, more details available at https://xgboost.readthedocs.io/en/stable/tutorials/model.html

XGBoost vs. LightGBM

LightGBM grows leaf-wise (horizontally) while XGBoost grows level-wise (vertically)



LightGBM

XGBoost

For the curious, more details available at https://towardsdatascience.com/catboost-vs-lightgbm-vs-xgboost-c80f40662924

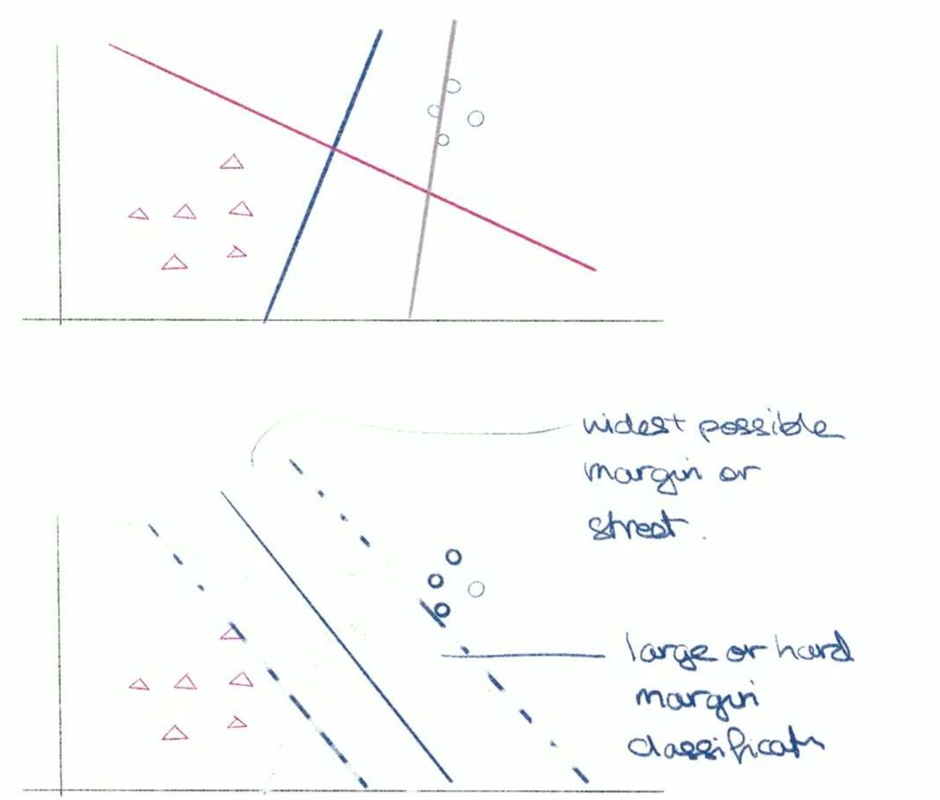
Many other models

Support Vector Machines

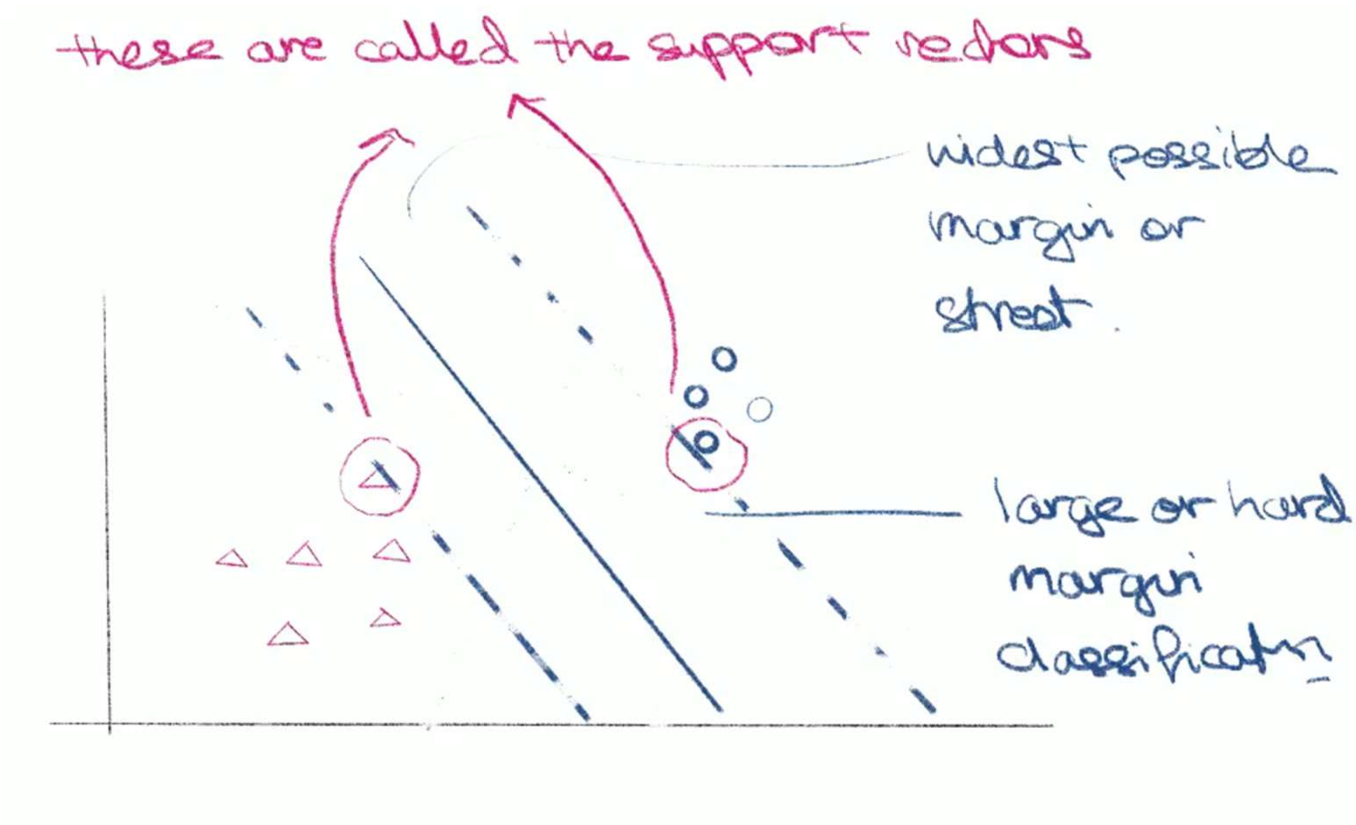
what does a SVM look like?

$$y = \begin{cases} 0 & \text{if } wx + b < 0 \\ 1 & \text{if } wx + b > 0 \end{cases}$$

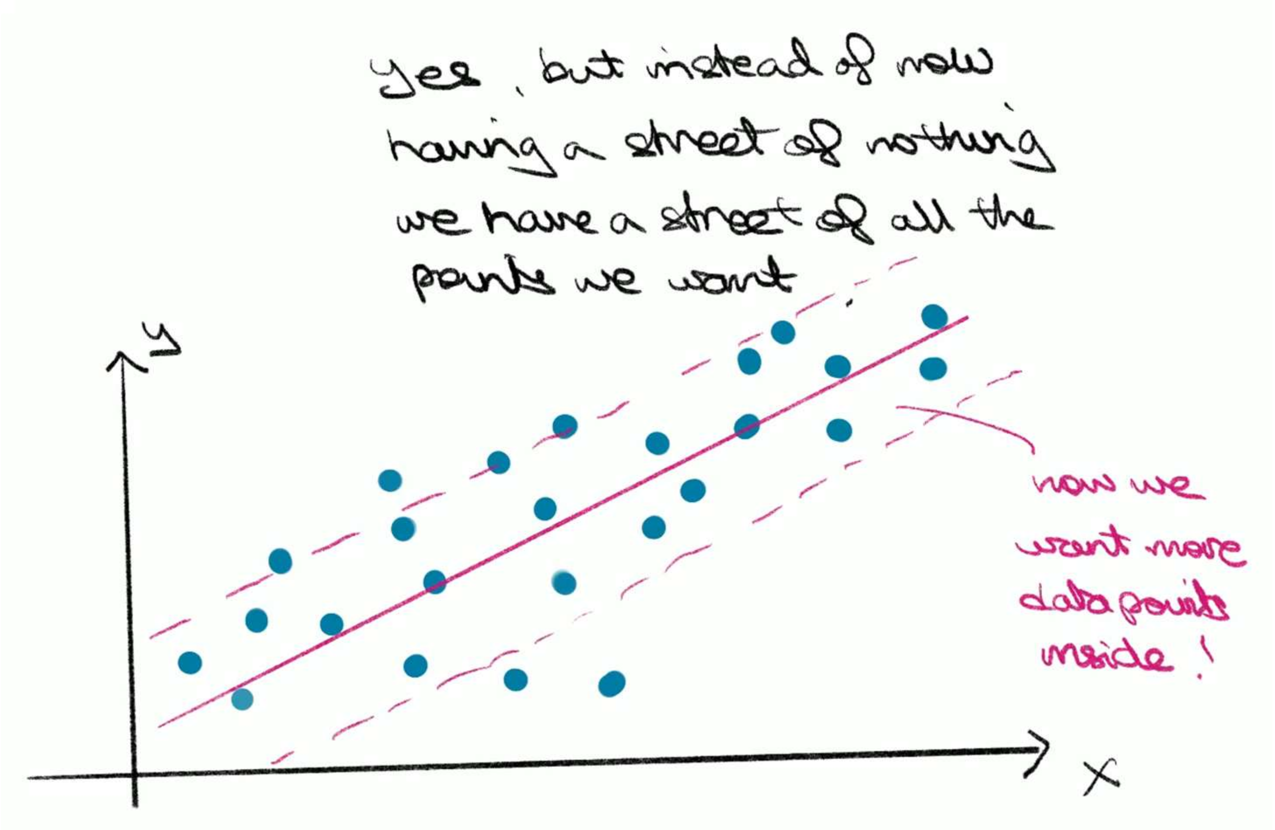
simplest linear SVM.



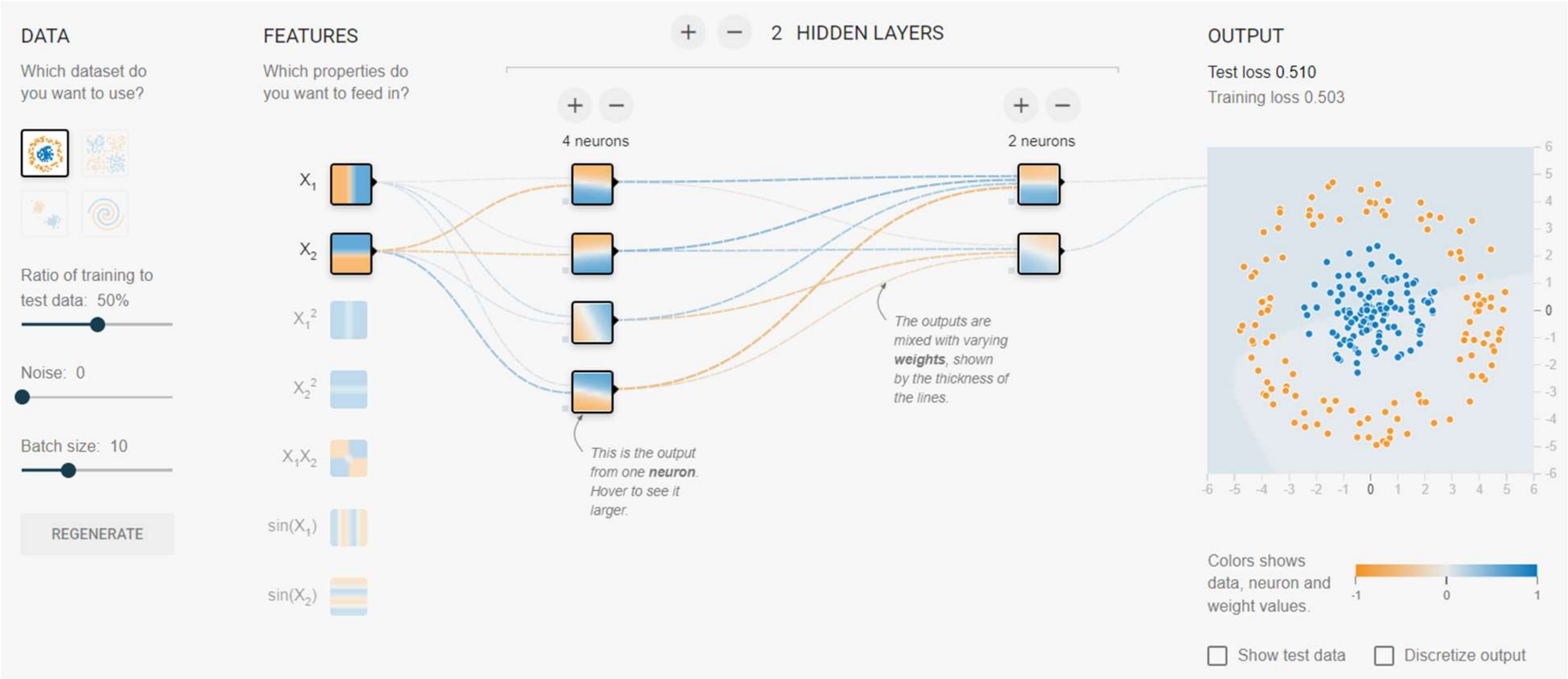
Support Vector Machines



Support Vector Machines For Regression?

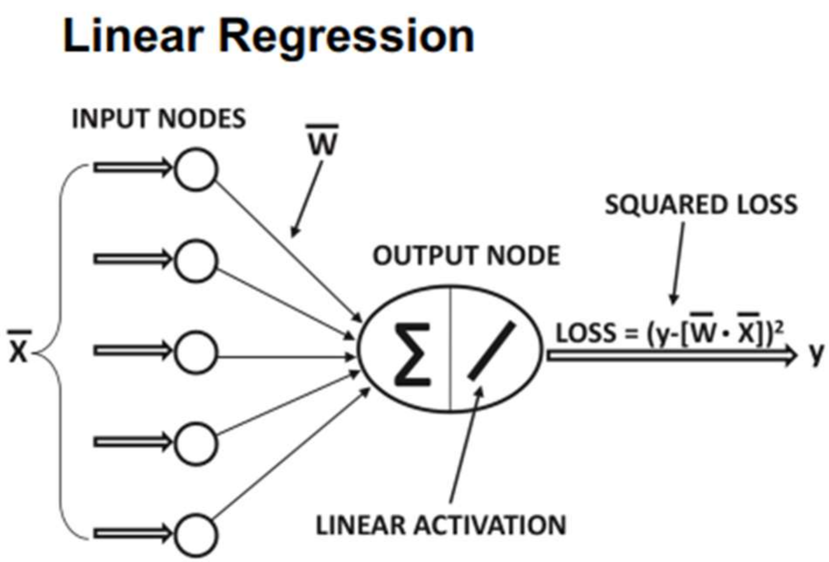


Neural Networks



https://playground.tensorflow.org/

Neural Networks



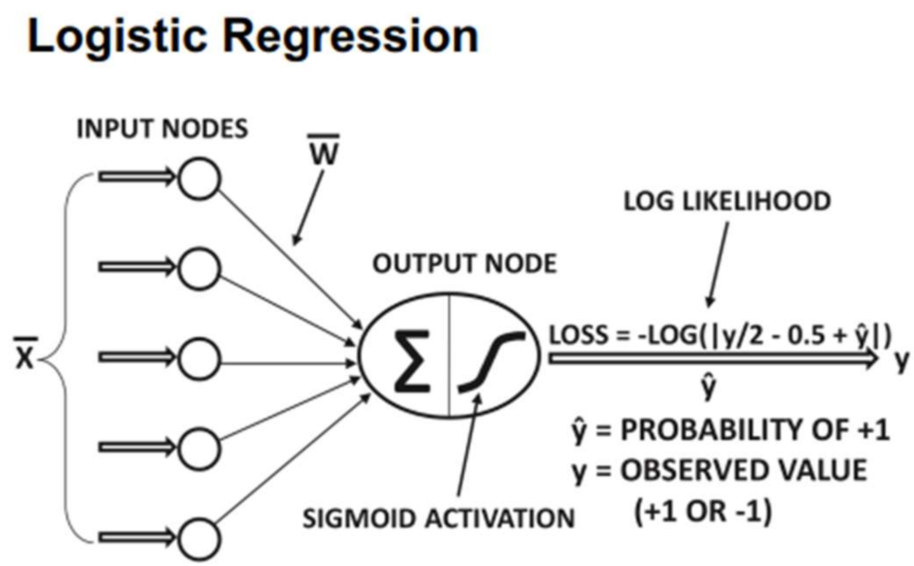


Figure from Neural Networks and Deep Learning, Charu Aggarwal

Neural Networks

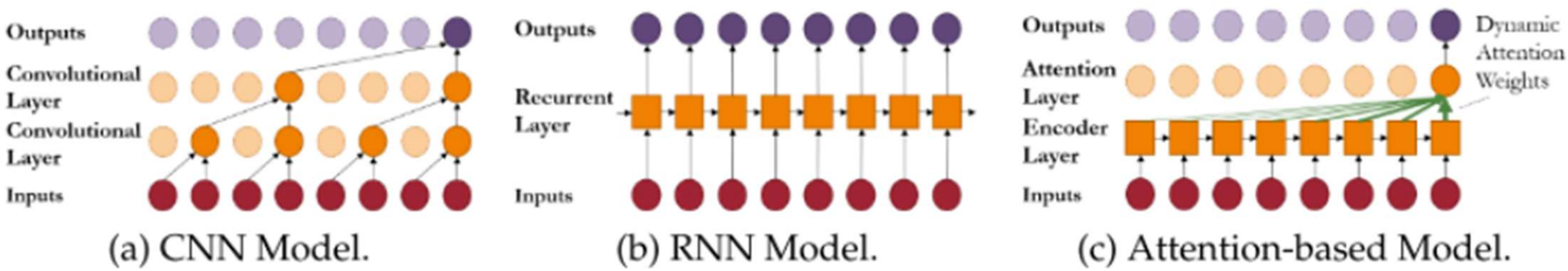
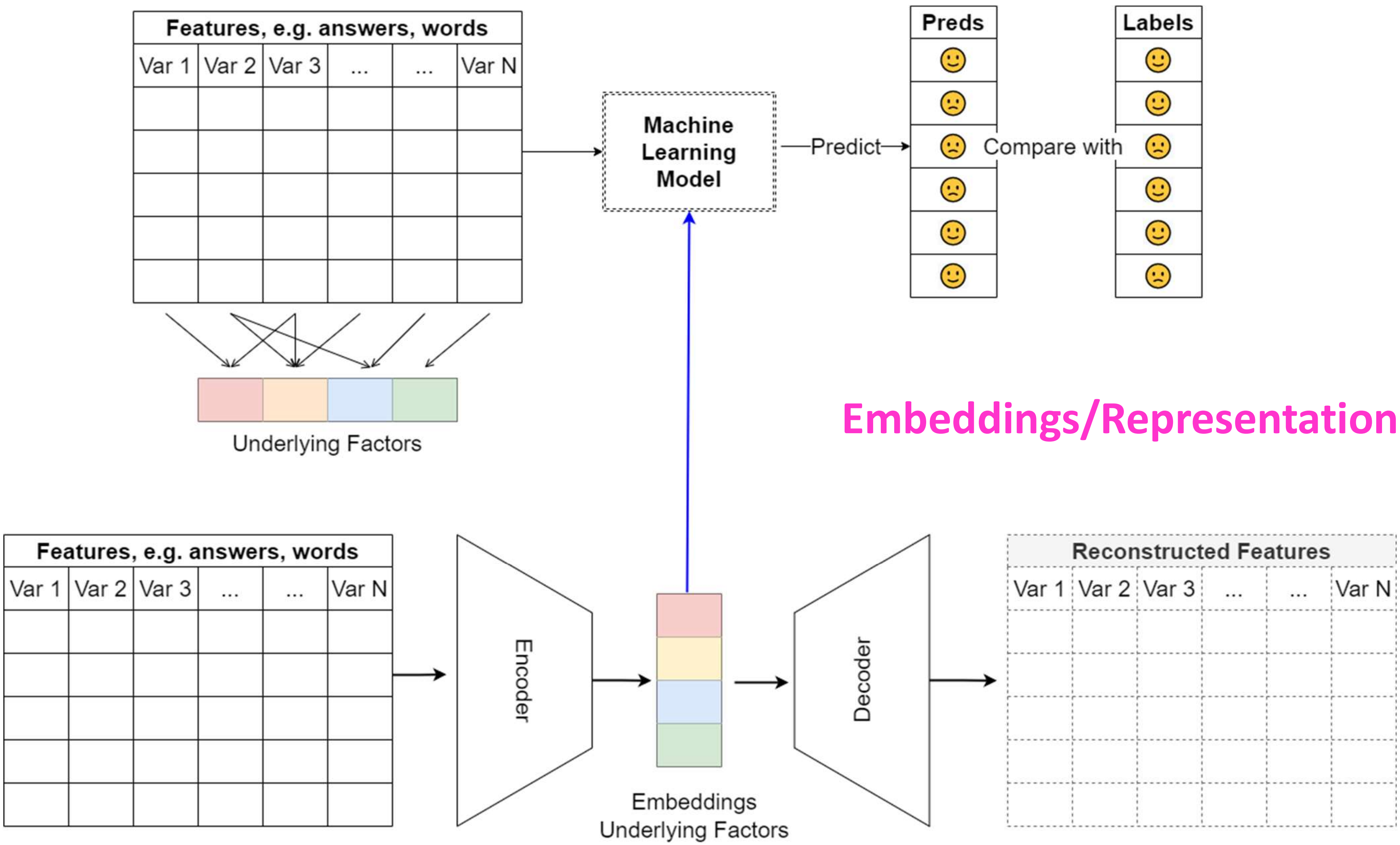
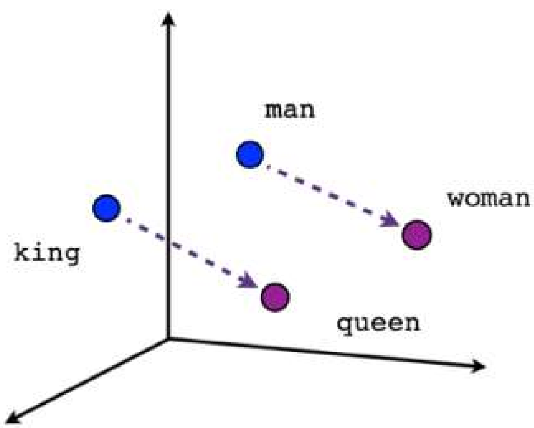


Figure 1: Incorporating temporal information using different encoder architectures.

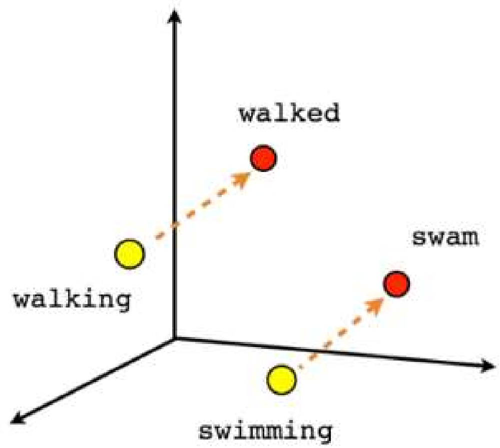
From Time Series Forecasting With Deep Learning: A Survey, Lim et al., 2020

Embeddings/Representations

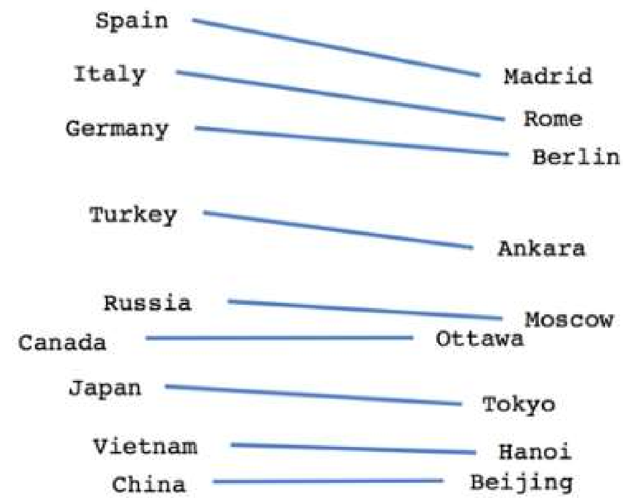




Male-Female



Verb tense



Country-Capital

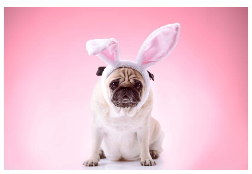
https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

Model Selection: Considerations?

- What are the **unique characteristics** of your data?
- What has **worked** for my problem or task?
    - If problem or task is unique, what is a **similar** problem or task?
- Get a **naïve baseline**
    - Choose a simple model, e.g., simple logistic regression
- Consider a few **challenger models**
- **Train and evaluate**

Some other concepts

Which task is easier?



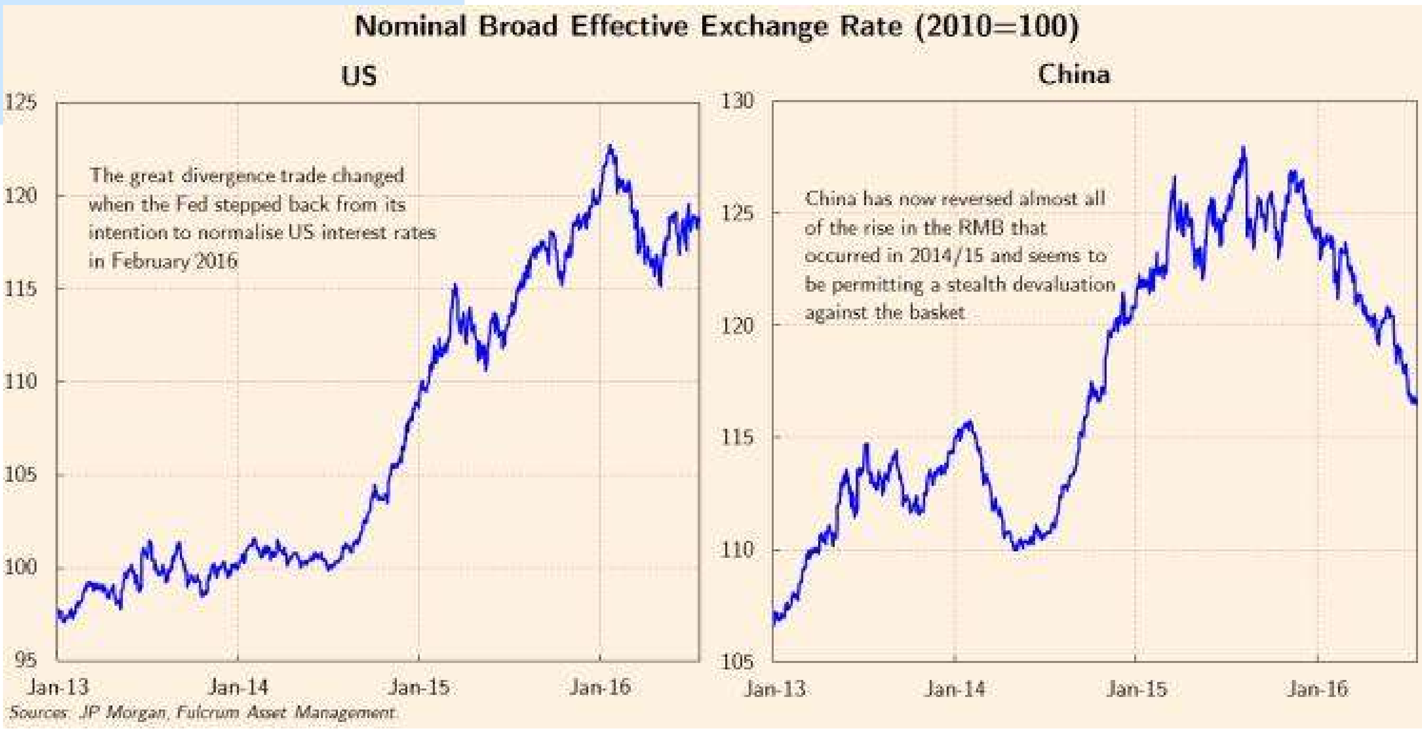
[Cat, Dog]

$1,500

Financial Data

- Machine and deep learning seem to be more successful on computer vision and natural language tasks. What are some key differences between data in computer vision and natural language domains compared to finance?

Signal-to Noise & Non-Stationarity



https://on.ft.com/3z3LU8J

Drift

- **Data drift**: Using credit transaction data before *chip and pin* to train a model for data after *chip and pin*

- **Concept drift**: Using the same model to detect fraud after it becomes known that your model depends on a specific network centrality measure to detect fraud

Financial Data

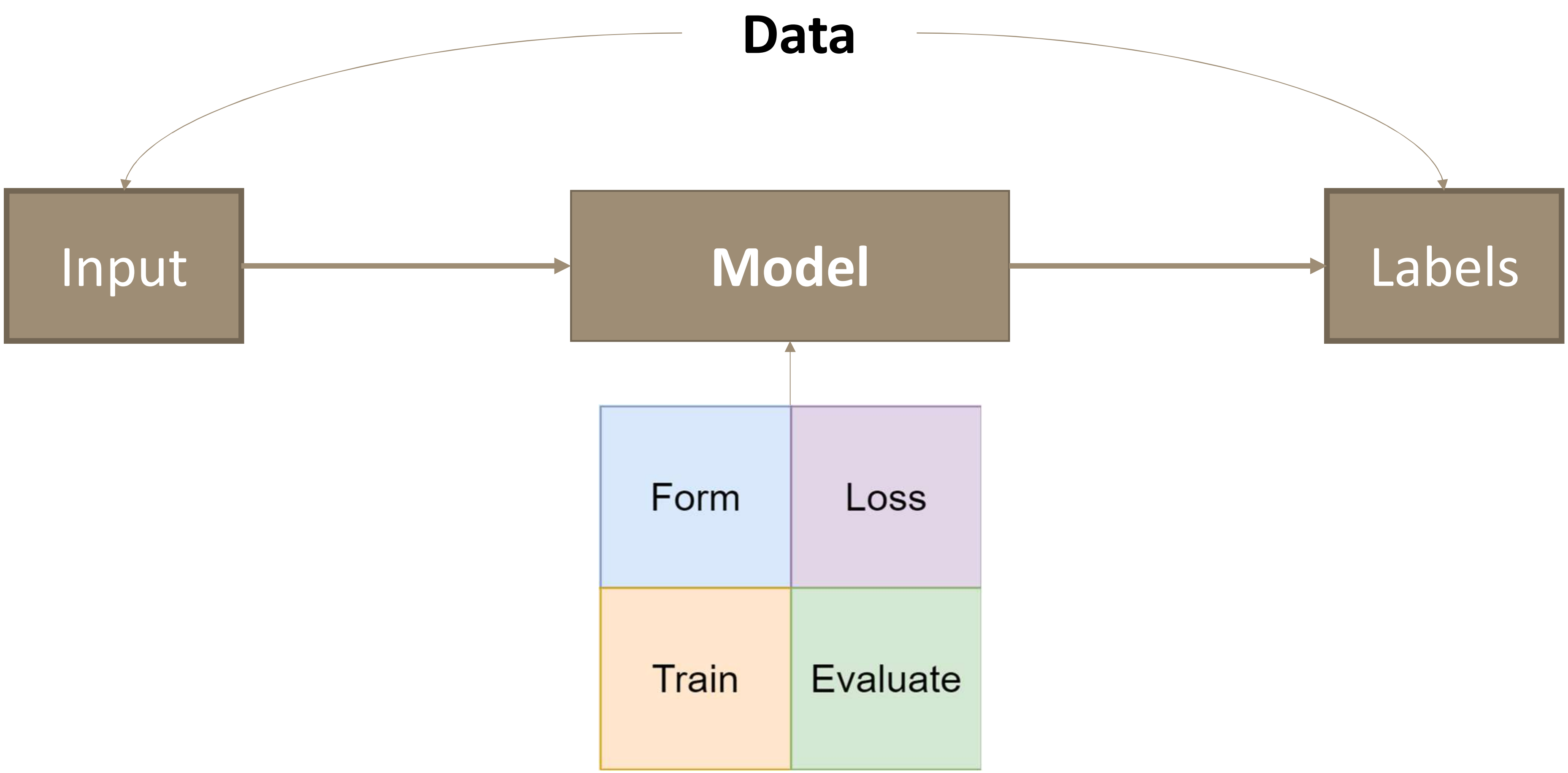
- What implications does this have for models in FIs?

Deep learning today

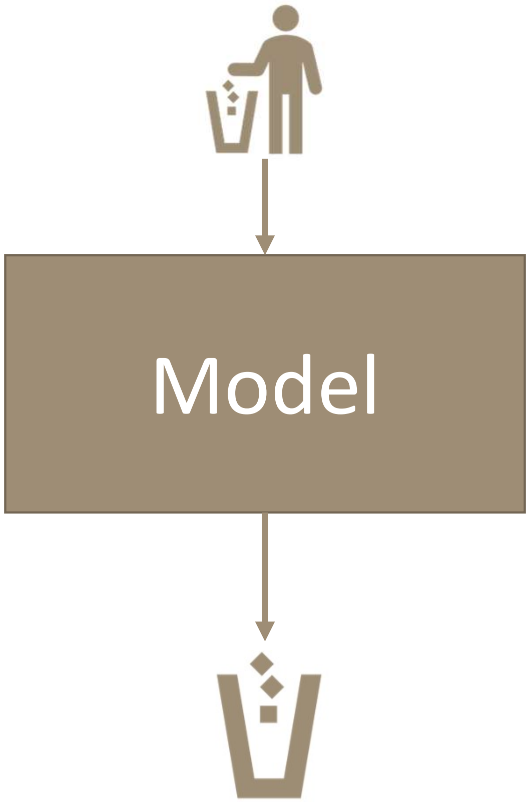
Prompt: Singapore 50 dollar note, Ghibli style

https://huggingface.co/spaces/huggingface/diffuse-the-rest

Data and Models For Supervised Models



Data: Rubbish in Rubbish Out



- If data **quality** is poor or **biased**, then unlikely to get good results

- Possibilities:
  - Data available, but processed poorly
  - Data available, but poor quality or biased
  - Data available, but not useful for task
  - Insufficient data for task

Data: Feature Engineering

Not only about **selecting right inputs**:

- **Transform non-linear to linear** problem (we saw this)
- Capture **interactions**
  - *E.g., think about BMI and TDSR vs their constituents*
- Utilize **unstructured inputs**
  - *E.g., think about tabular information vs. images, text, audio, networks*

Data quality

- **Accuracy**
  - E.g., mis-labelled illicit transactions
- **Completeness**
  - E.g., omission of transactions stored in another banking system
- **Consistency**
  - E.g., unclear instructions when designating a loan as defaulted
- **Currency**
  - E.g., characteristics/distribution of fraudulent transactions changing over time due to change in tech. and consumer behaviour

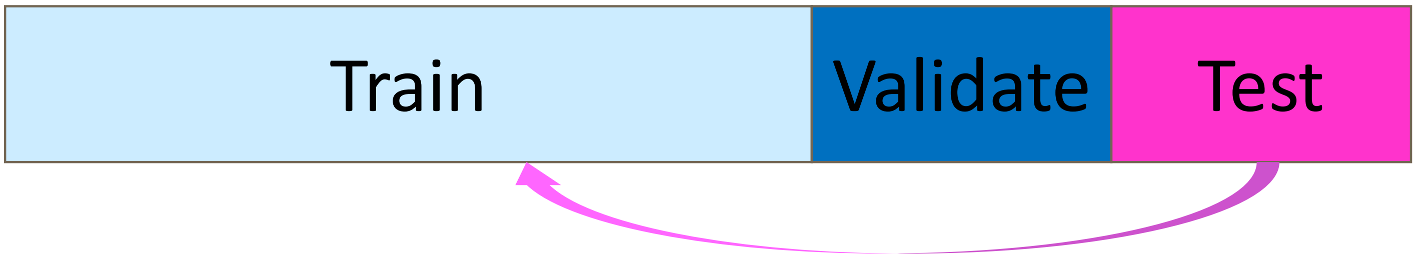
Addressing all of these is a pipe-dream.

But important to know if these exist in our data.

Data bias

- **Distribution** bias
  - Personal attributes
- **Representation** bias
  - Match general population but under-represent certain segments
- **Implicit** bias
  - Not all bias are obvious, e.g., gender vis-à-vis income vis-à-vis location vis-à-vis race
- **Labelling** bias
  - Even experts label things differently, e.g., 2nd opinions?

Data leakage



- Very common even for random train, validation test splits
  - Using total time customer spent in bank to predict customer purchase intent so as to act on it while customer still in bank
  - Use data before t to predict prob. of default at $t + 1$, data before t includes post-default adjustments
  - Predicting illicit transactions using complete incident reports
  - Using mean and standard deviation of entire dataset to scale/normalize data

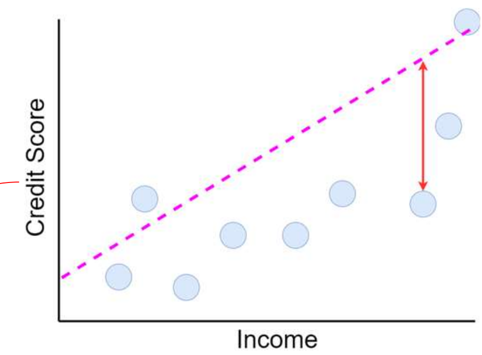
Drift

- **Data drift**: Using credit transaction data before *chip and pin* to train a model for data after *chip and pin*

- **Concept drift**: Using the same model to detect fraud after it becomes known that your model depends on a specific network centrality measure to detect fraud

Framework: Linear Regression

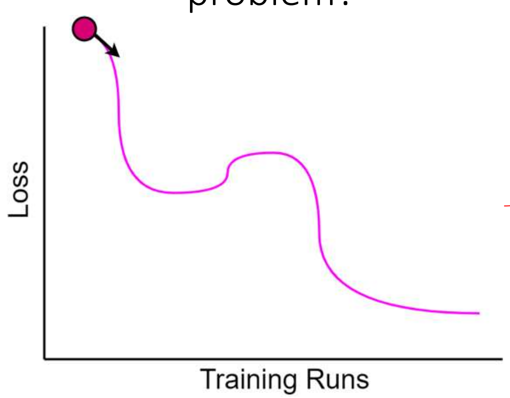
$$y = AX + B$$



Form	Loss
Train	Evaluate

Gradient descent, one of many ways to train/optimize.

Can you spot a common problem?



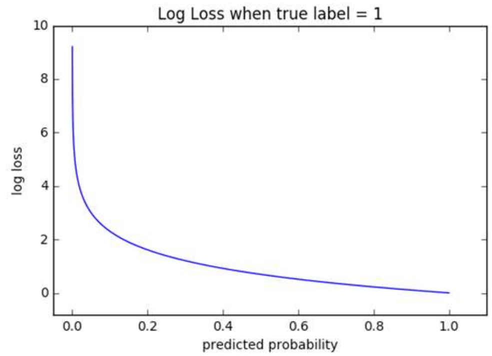
Root Mean Squared Error, Mean Abs. Error, Mean Abs. Percentage Error

Framework: Logistic Regression

$$P = \frac{1}{1 + e^{-(AX+B)}}$$

Cross-Entropy

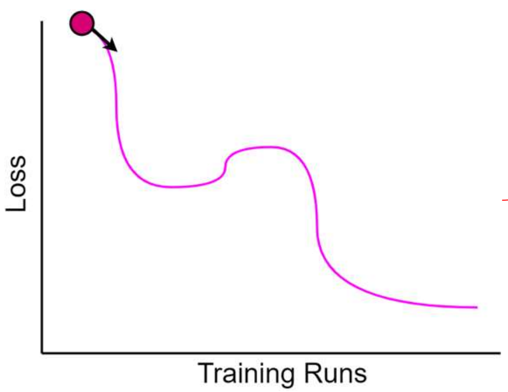
$$-Ylog(P) - (1 - Y)\log(1 - P)$$



Form	Loss
Train	Evaluate

Gradient descent not the only way

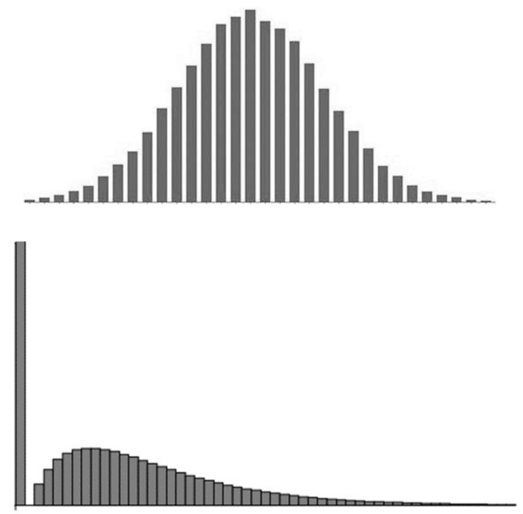
Also Max. Likelihood Est.



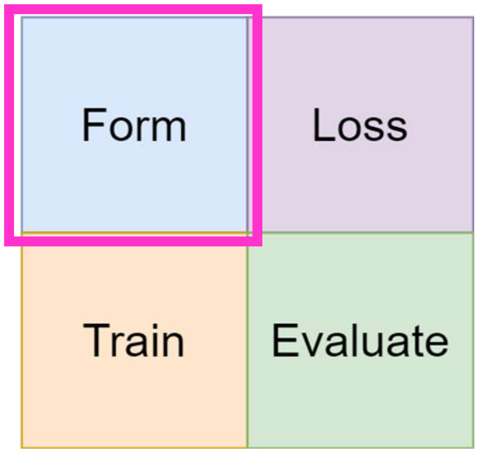
Accuracy, Recall, Precision, F1

Linear vs. Non-Linear

$$\boldsymbol{Y = LinkFunction(Expected[Y]) + \varepsilon}$$



$$Y = AX + B \qquad\qquad P = \sigma(Y) = \sigma(AX + B)$$

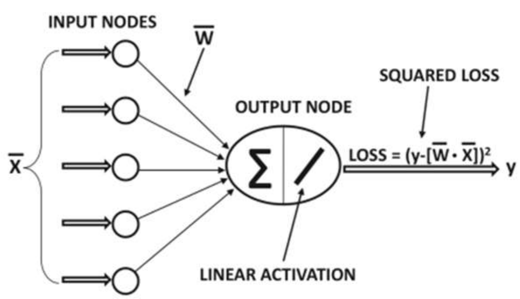


$$\boldsymbol{Y = AX^2 + B}$$

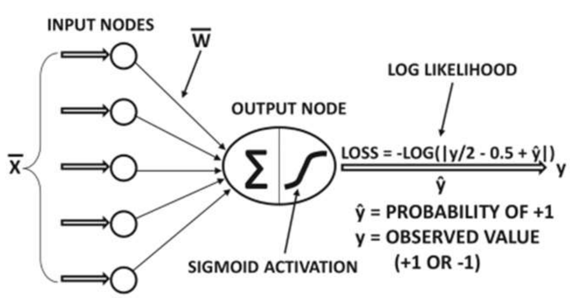
$$\boldsymbol{Y = AX_1^2 + BX_2^2 + CX_1X_2 + D}$$

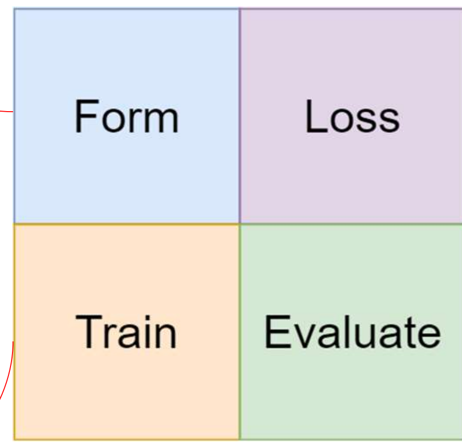
Framework: Neural Networks

Linear Regression



Logistic Regression



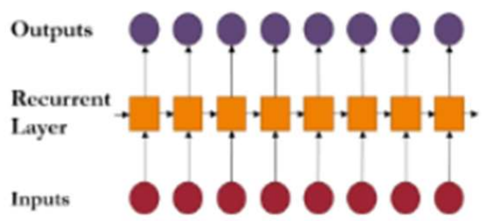


Cross-Entropy Loss, Squared Error Loss

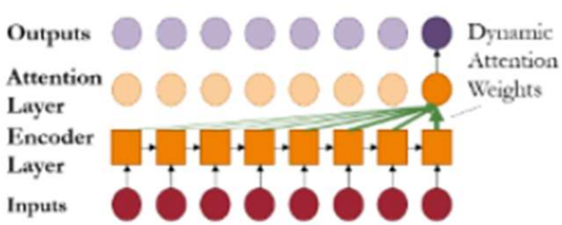
Gradient descent

Accuracy, Recall, Precision, F1, Root Mean Squared Error, Mean Abs. Error, Mean Abs. Percentage Error

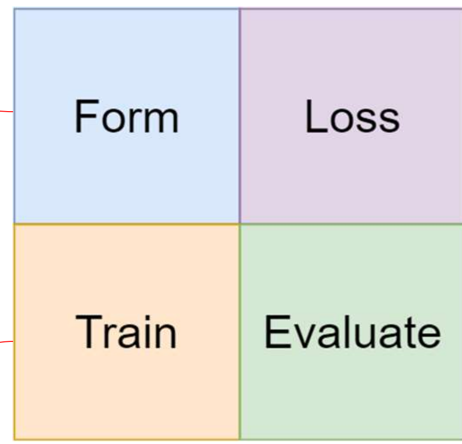
Framework: Neural Networks



(b) RNN Model.



(c) Attention-based Model.



Cross-Entropy Loss, Squared Error Loss

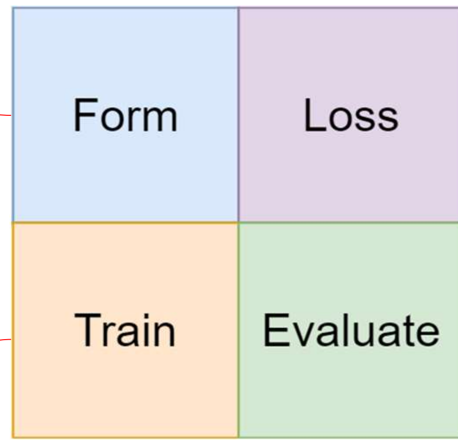
Gradient descent

Accuracy, Recall, Precision, F1, Root Mean Squared Error, Mean Abs. Error, Mean Abs. Percentage Error

Framework: Decision Tree

What would be a good criteria to split?

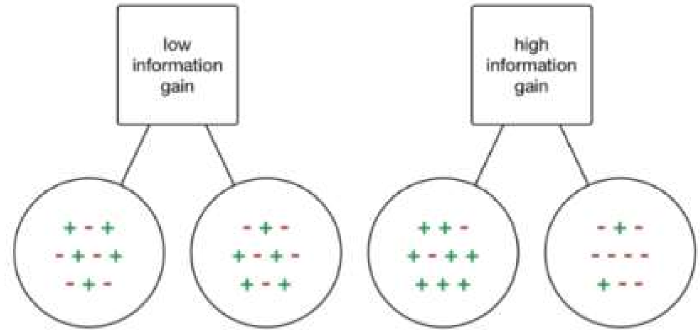
What is the form?



Given the splitting criteria, how could one build the tree?

Recall what we used for logistic regression

Framework: Decision Tree



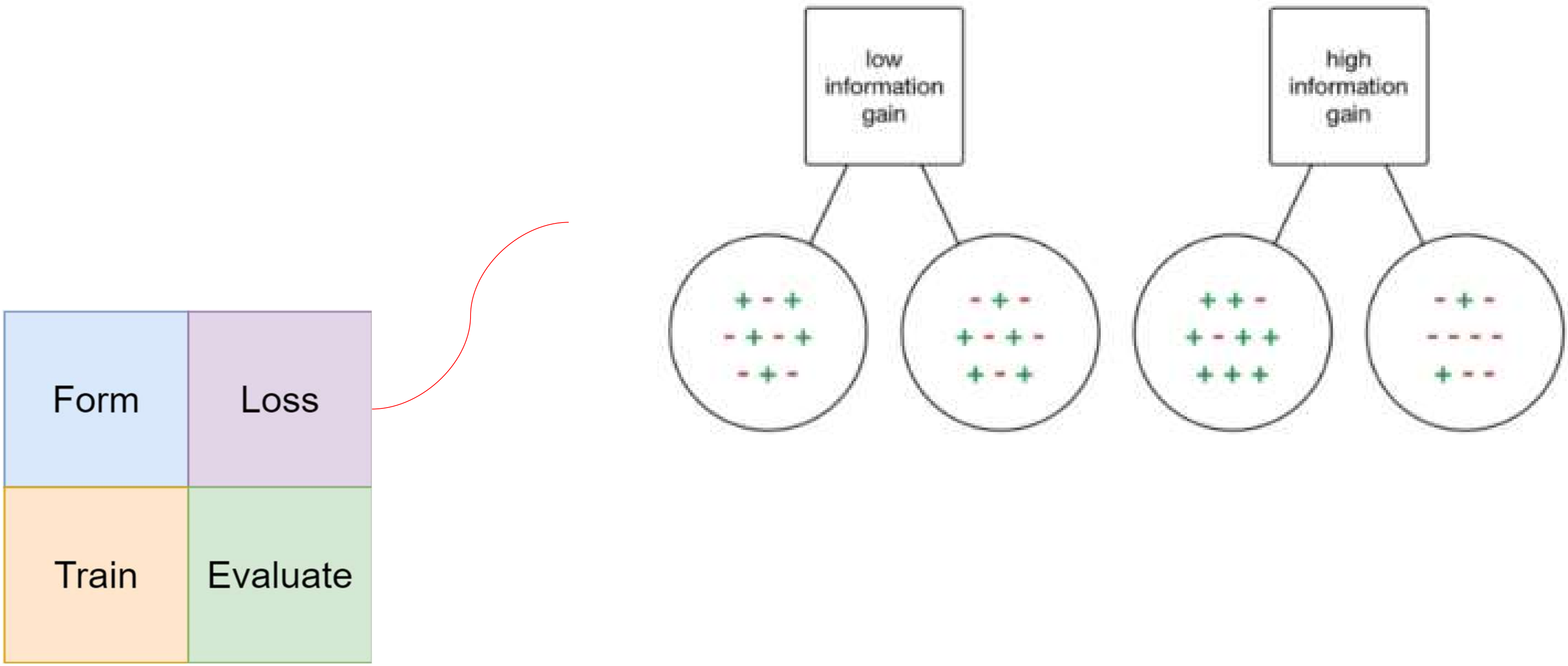
Binary Tree

Form	Loss
Train	Evaluate

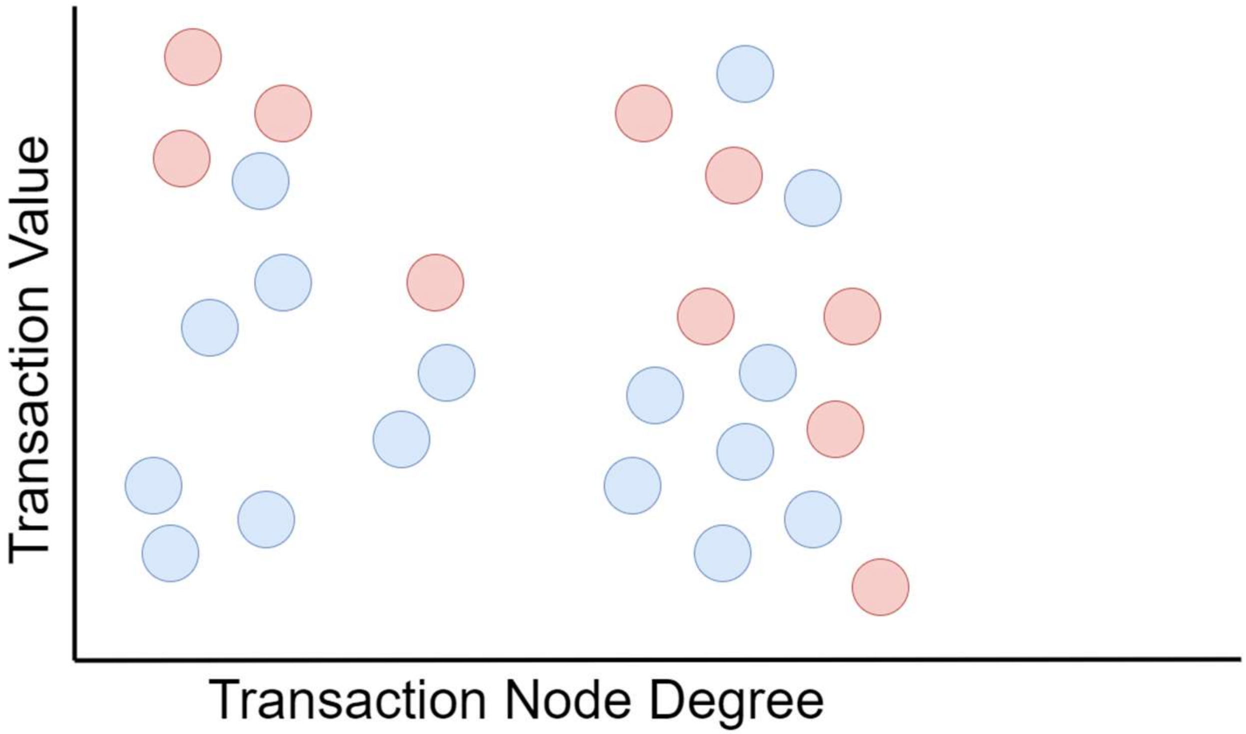
Greedy

Accuracy, Recall, Precision, F1

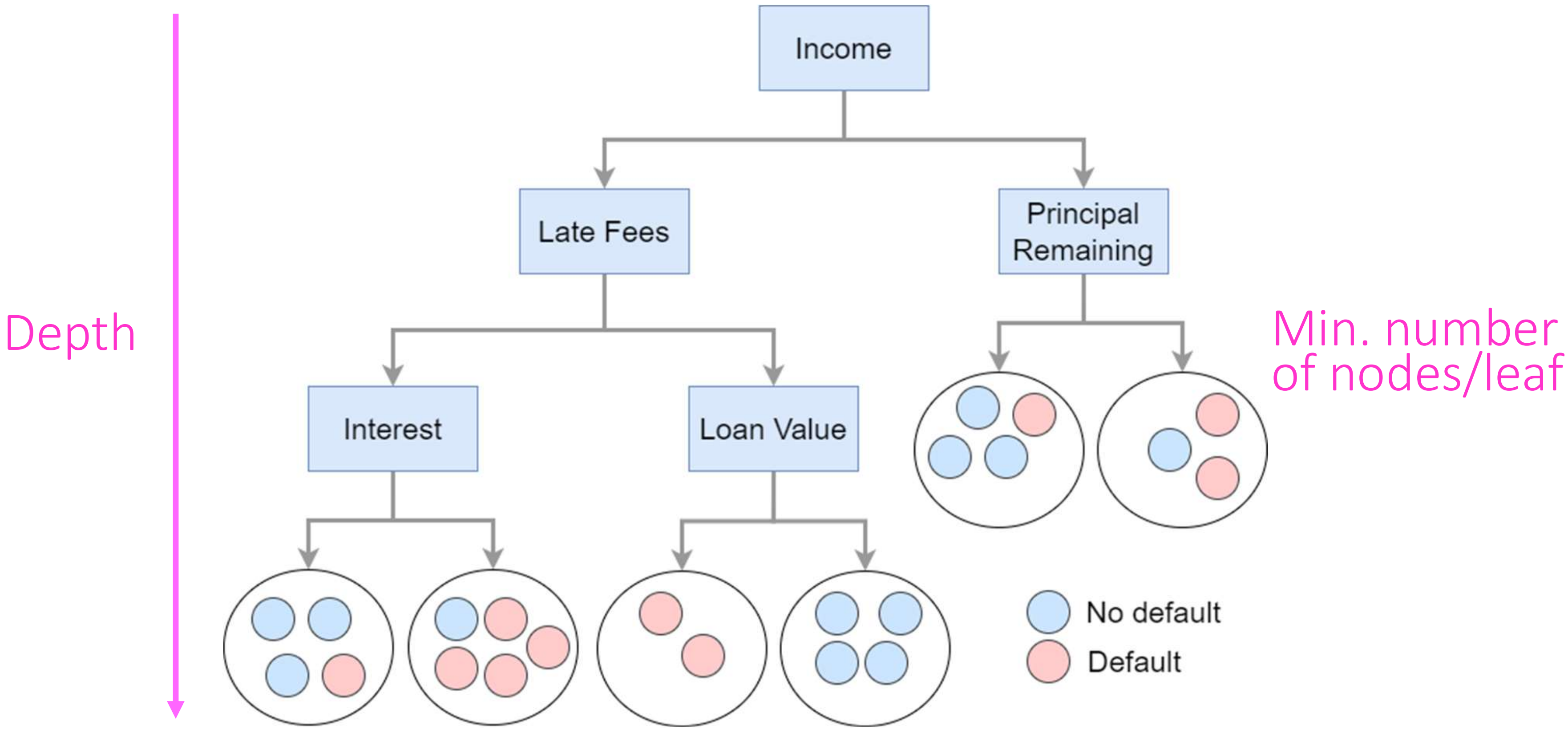
Framework: Decision Tree



Decision Trees - Recap

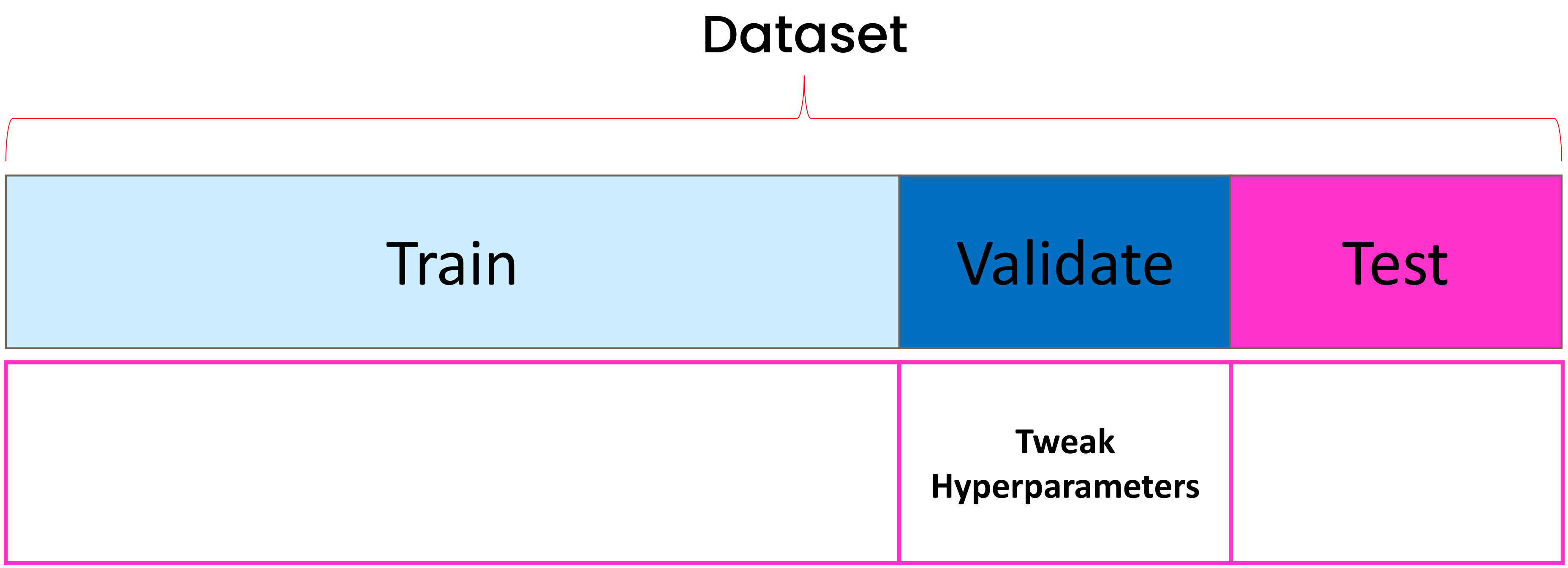


Recall difference between parameters and hyperparameters



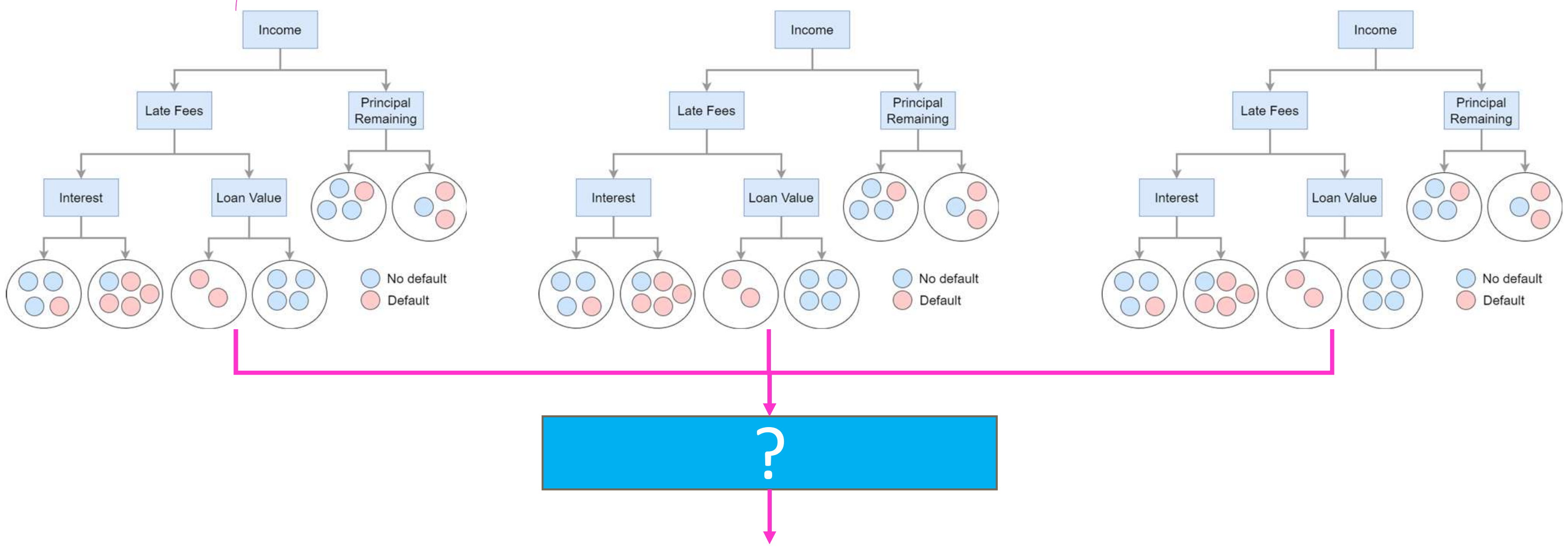
class `sklearn.tree.`**`DecisionTreeClassifier`**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0*) ¶

Recall train-validation-test splits and purpose?



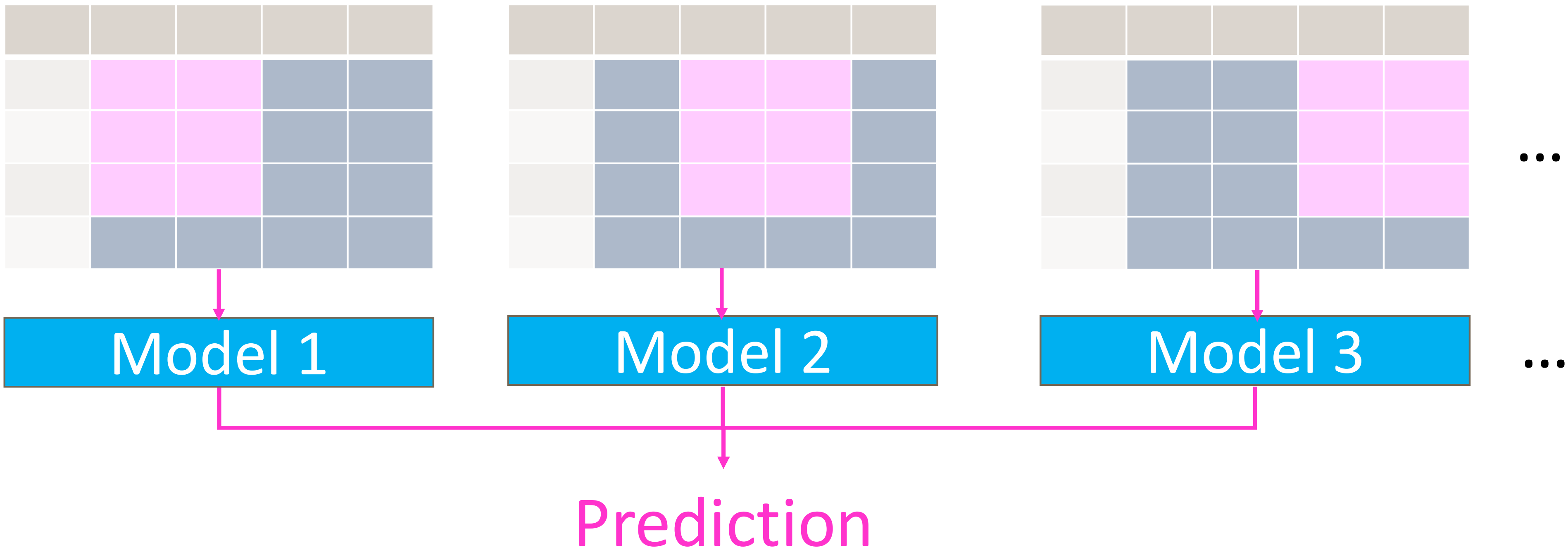
Ensembles

A single tree tends to overfit

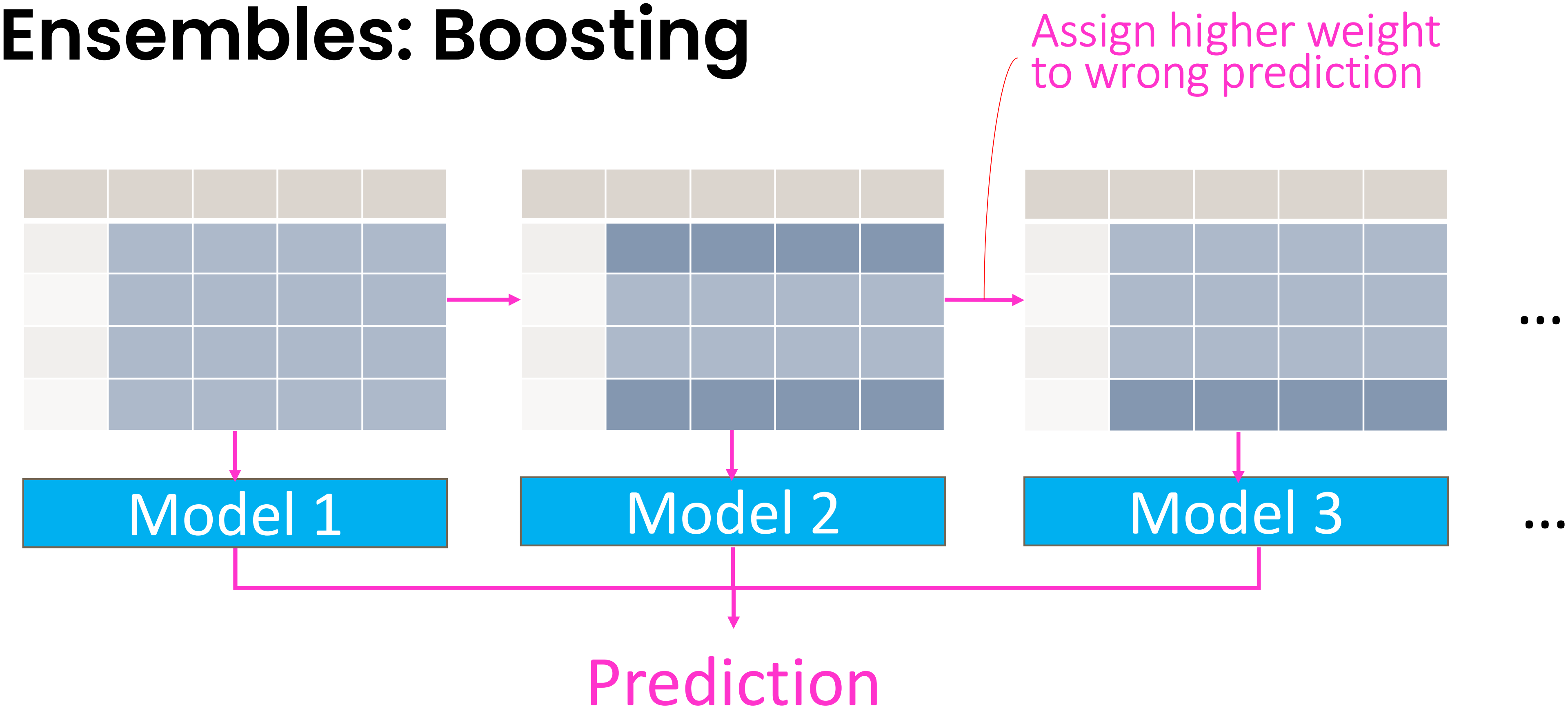


Prediction

Ensembles: Bagging



Ensembles: Boosting



XGBoost

XGBoost is basically based on the idea of boosting, but with some additional math and optimization

For the curious, more details available at https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Usually, trees in ensembles are weak learners

Weak learners means we typically constrain their hyper-parameters, which makes them over- or underfit?

An ensemble of weak learners can potentially help us achieve both low ______ and _______.

Let’s move on to unsupervised models

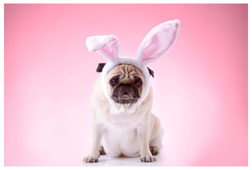
Supervised learning

$$Y = AX + B$$

Input → Model → Output

X A,B Y

Data Labels



→ [Cat, Dog]

Supervised learning

$$Y = AX + B$$

Input → Model → Output

X A,B Y

Data Labels

→ $1,500

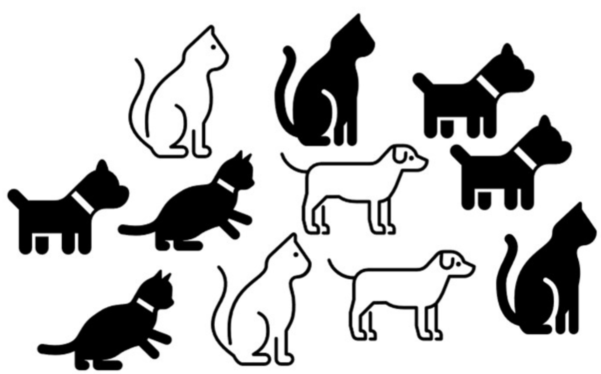
Unsupervised learning

$$Y = AX + B$$

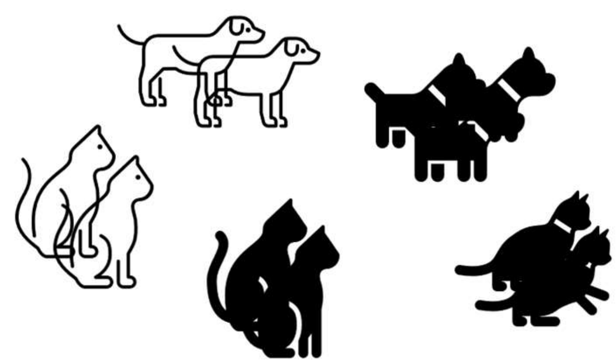
Input → Model → Output

X A,B Y

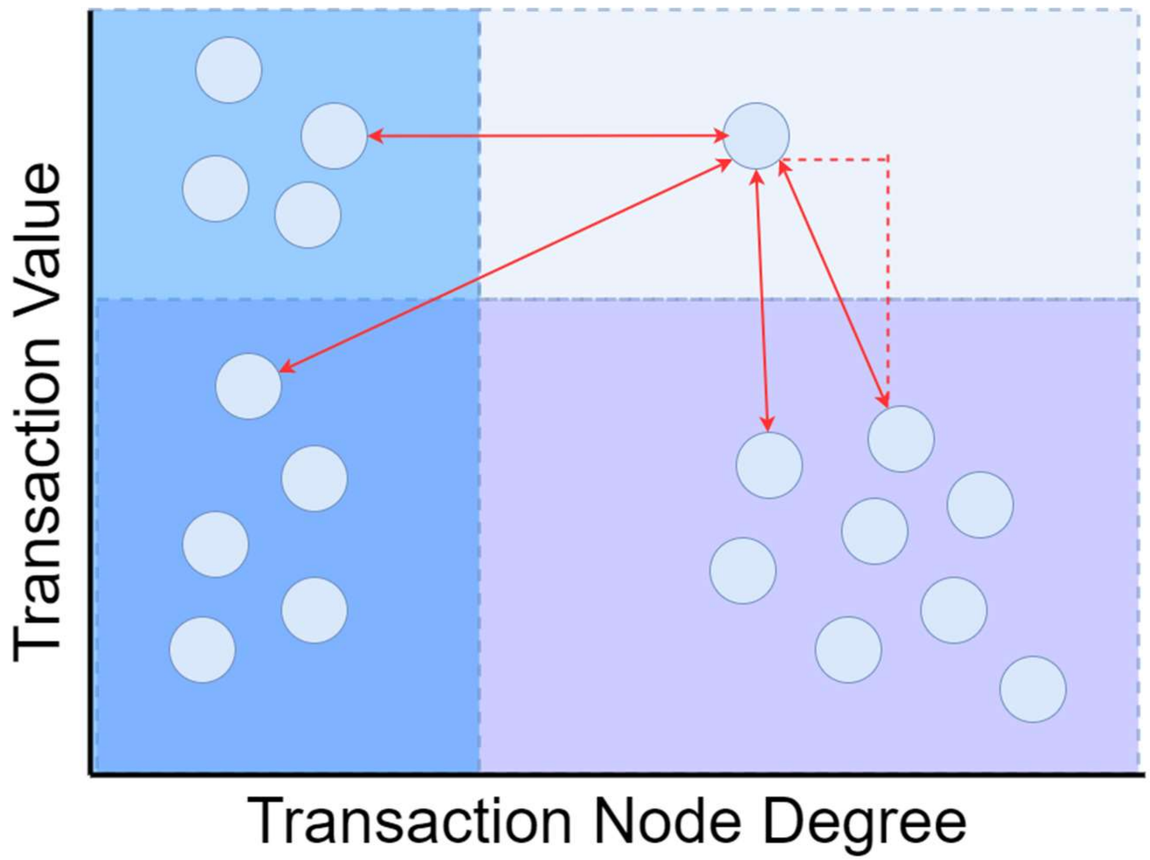
Data



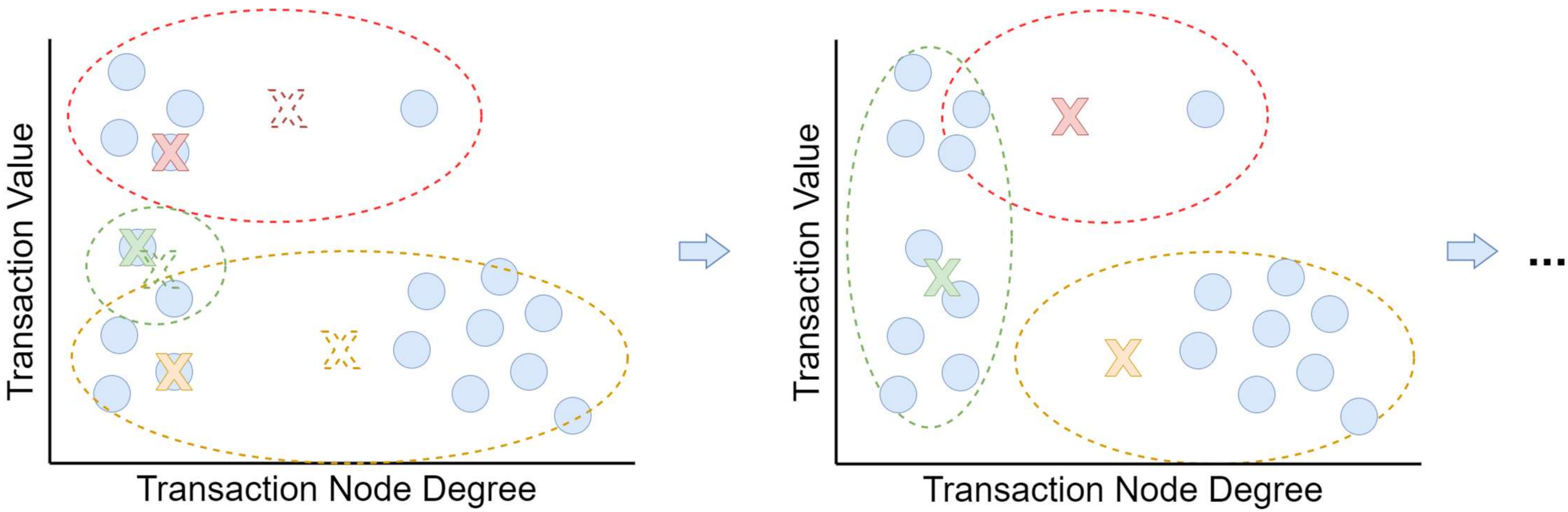
Groupings



Clustering: K-Means

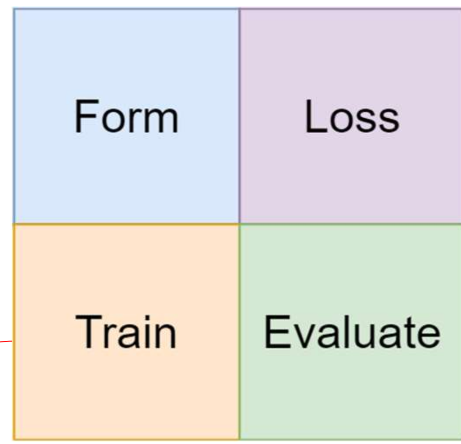


Clustering: K-Means



Framework: K-Means

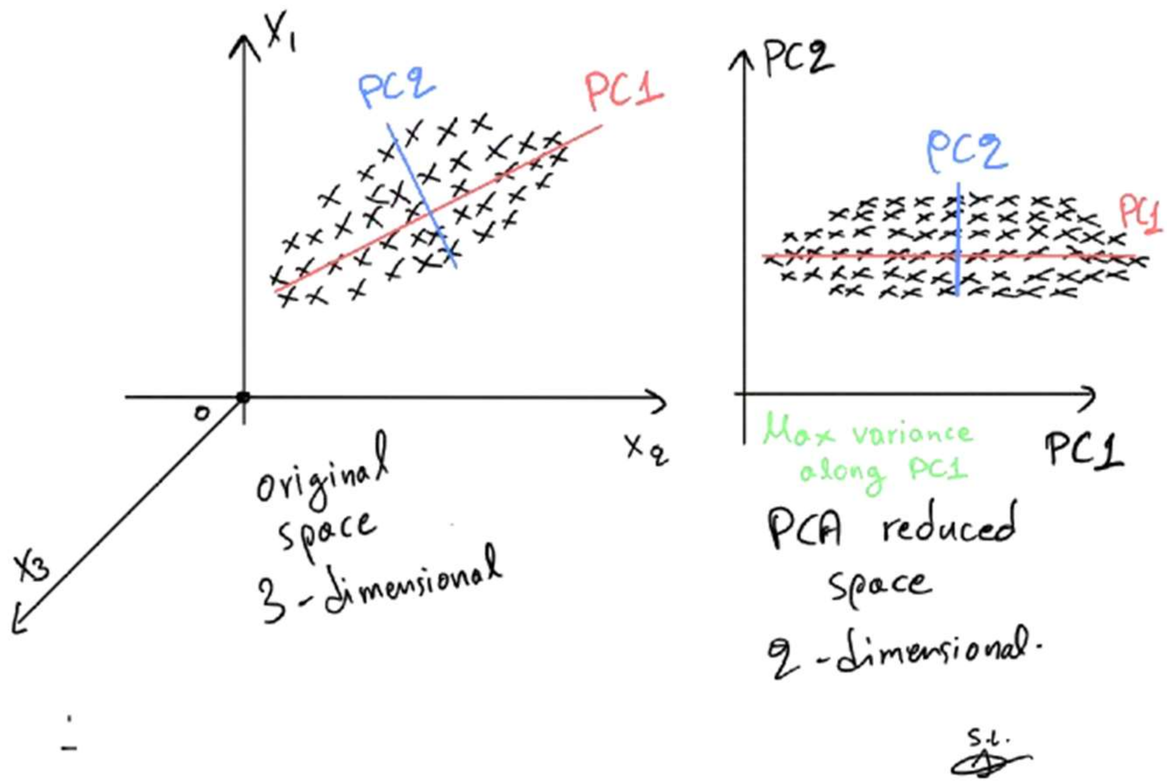
What is the loss?



How do you train the model?

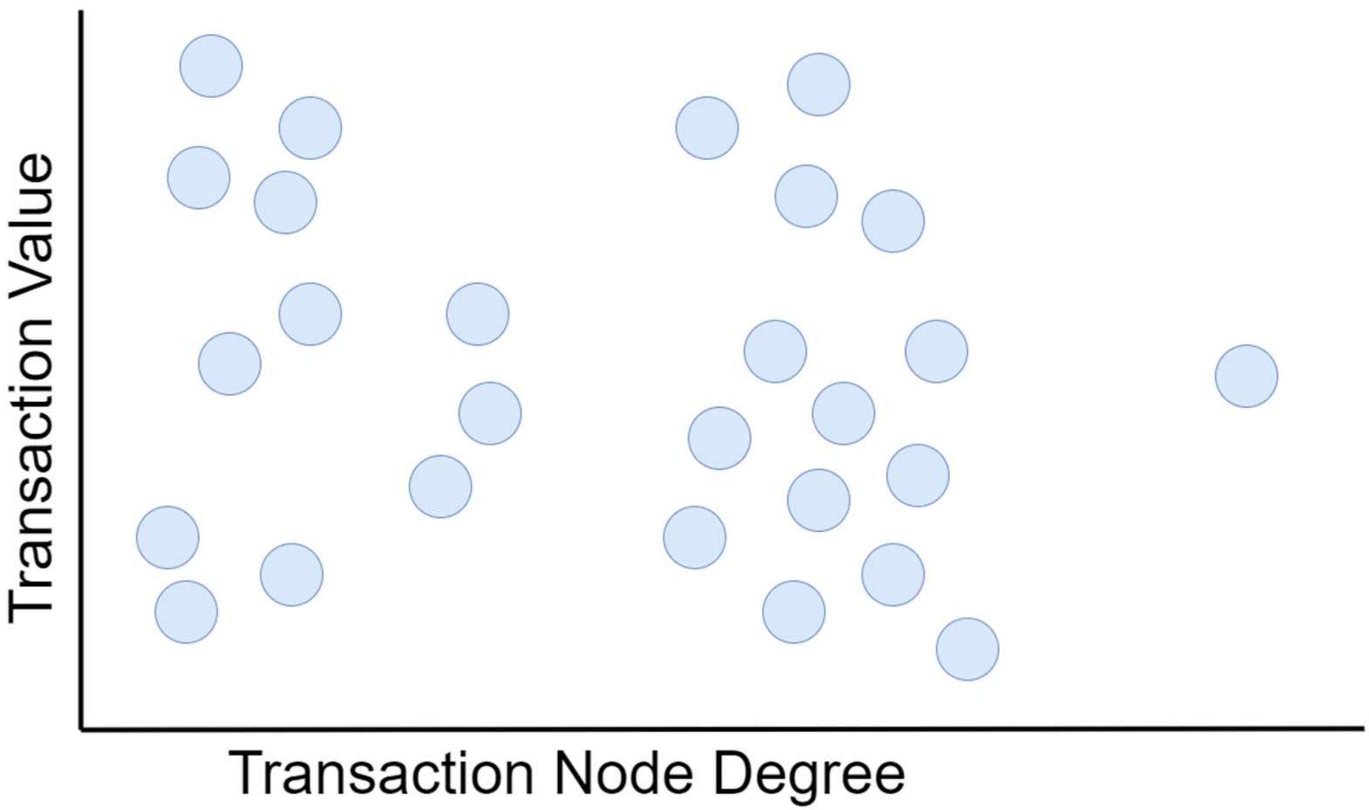
How do you evaluate?

Dimensionality Reduction



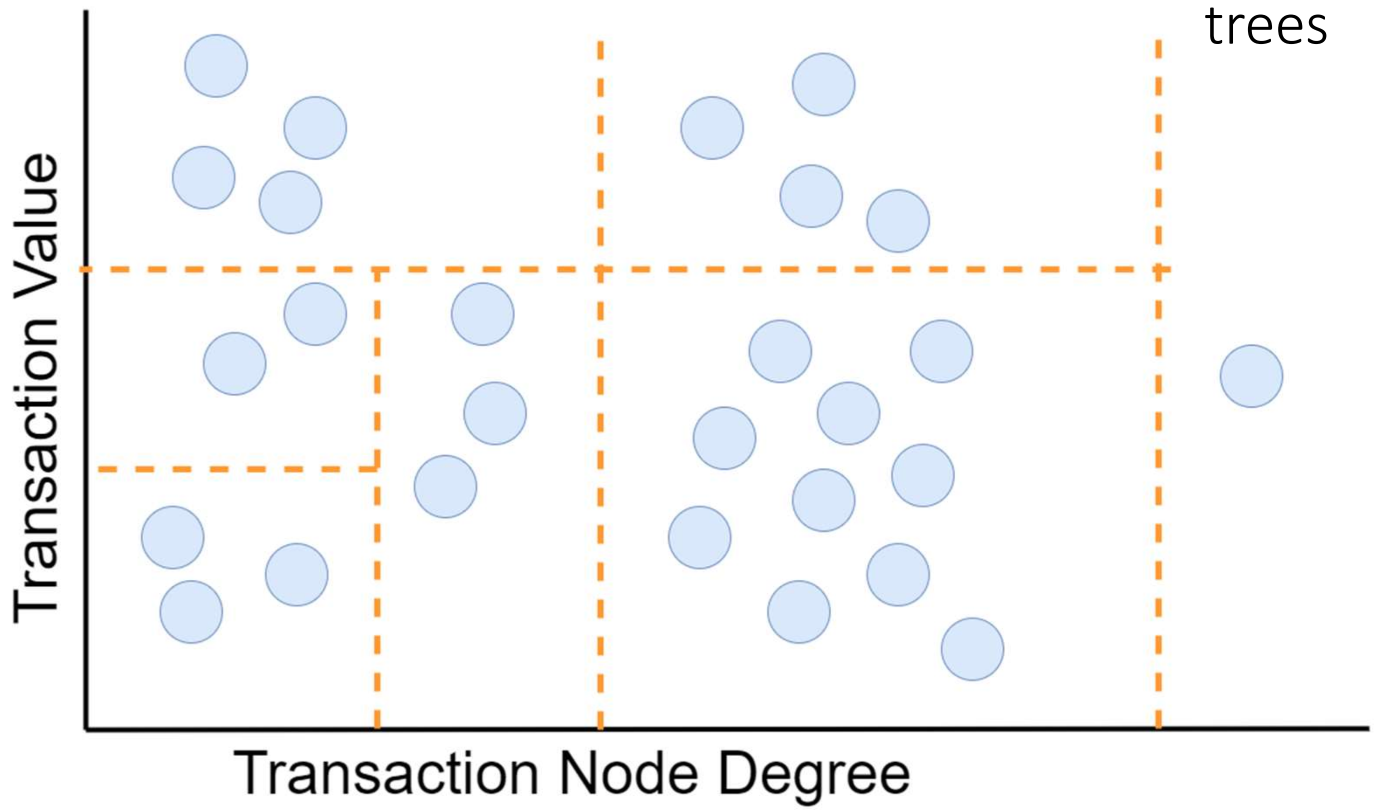
From https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

Spot the anomaly

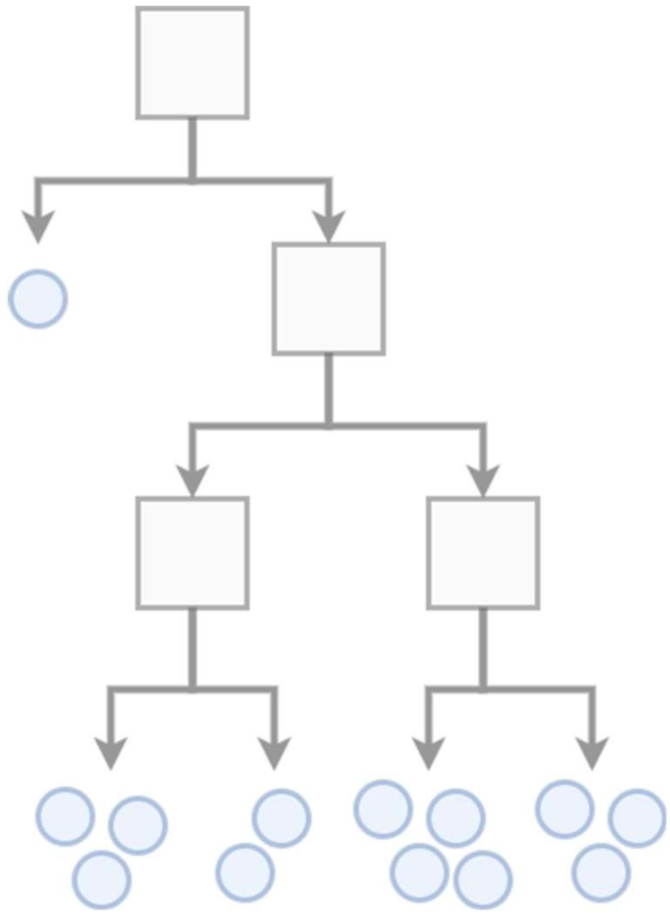


Isolation Forest

Recall decision trees



Isolation Forest



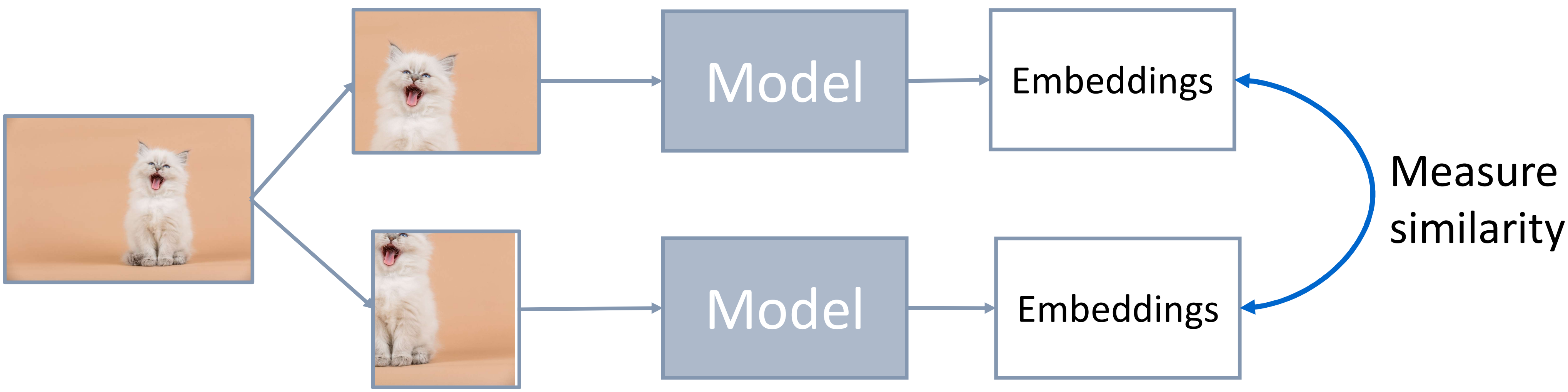
Which path leads to an anomalous instance?

But what is an anomaly?
Is it always this clear?

For an illicit transaction detection problem?
For a fraudulent credit card transaction?

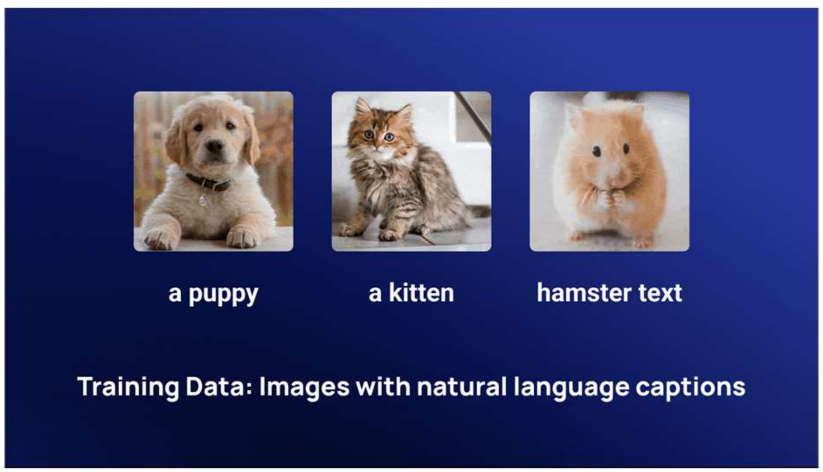
Think about the characteristics of your data.

Another perspective: Self-Supervised



In deep learning, unsupervised models can be very powerful!

Recall the demonstration earlier



Interpretability and Explanability

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు)
@rao2z

There seems to be an almost willful confusion about the need and role for explainability of #AI systems on #AI twitter.

Contrary to the often polarizing positions, it is neither the case that we always need explanations nor is it the case that we never need explanations. 🧵 1/

10:22 AM · Jul 11, 2022 · Twitter Web App

27 Retweets 6 Quote Tweets 102 Likes

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11
Replying to @rao2z

We look for explanations of high level decisions of (what for us are) explicit knowledge tasks; and where contestability and collaboration are important.

We rarely look for explanations of tacit knowledge/low level control decisions. 2/

1 2 8

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

I don't need explanation on why you see a dog in a picture; why you put your left foot 3 mm ahead of your left, or why facebook recommends me yet another page.

I do want one if am denied a loan, or I need a better model of you so I can coordinate with you. 3/

1 11

Interpretability and Explanability

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

Trust can reduce the need for explanations, but trust has to be earned and can't always be legislated. I ask explanations from my doctor despite the many impressive degrees on his wall. 4/

1 3 8

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

Explainability doesn't necessarily require full understanding on the part of the receiver. I ask my doctor for explanations of his diagnosis, despite the fact that I don't quite understand all the details. 5/

1 3

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

Explanations are always about the receiver's mind (..and vocabulary and reasoning..). After all, a doctor explains her decision in different ways to the patient and her colleague..

(This also explains the popularity of those viral @wired 5 levels of explanations videos!) 6/

1 1 8

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

We should stop conflating Interpretability and Explainability. Interpretability is w.r.t. a large population (e.g. human race) and explainability is w.r.t. individuals/specific groups.

The interpreter does the heavy lifting in the former; the explainer does it in the latter. 8/

1 3 8

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

The Rosetta Stone is interpretable to the human race (thanks to the heavy lifting by Young and Champollion).

I want my loan approval decision be explainable to me without me having to break my back.. 9/

1 6

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

An autonomous car's control decisions need to be interpretable to the (determined) investigators after a crash, its decision to take a particular route should be easily explainable to the rider in the car. 10/

1 2 7

Interpretability and Explanability

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

Explanations are sought after the fact--and thus do allow post hoc rationalizations.

Fear of this possibility shouldn't make them irrelevant!

We should design #AI systems to provide "truthful" explanations, but the comprehensibility may well necessitate approximation. 11/

3 3

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

Explanations are most definitely not a soliloquy. A direct trace of your or the #AI system's internal reasoning rarely makes for a good explanation--given the differing mental models, vocabulary and inferential abilities of the humans receiving the explanation. 12/

2 3 4

Subbarao Kambhampati (కంభంపాటి సుబ్బారావు) @rao2z · Jul 11

[fwiw, a lot of the above is related to our ongoing research on explainable human-#AI interaction. If you are interested:

Still an open research area!!

Interpretability and Explanability

- Not straight-forward
- **Trade-offs** – performance vs. interpretability
- **Good for trust, fairness, checks on model robustness**
- But …
  - Is it really needed, e.g., **impact**, **well studied**?
  - Could it have unintended effects, e.g., **adversarial** attacks?

Interpretability and Explanability

- Even interpreting linear regression is not straightforward
- Why?

$$Y = AX_1X_2X_3 + BX_1X_3 + CX_2X_3 + DX_1^3 + \ \ldots$$

Interpretability and Explanability

- Intrinsic or post-hoc
  - Logistic regression vs. **LIME**
- Model-specific or agnostic
  - Decision trees vs. **SHAP**
- Local or global
  - **Instance or class**

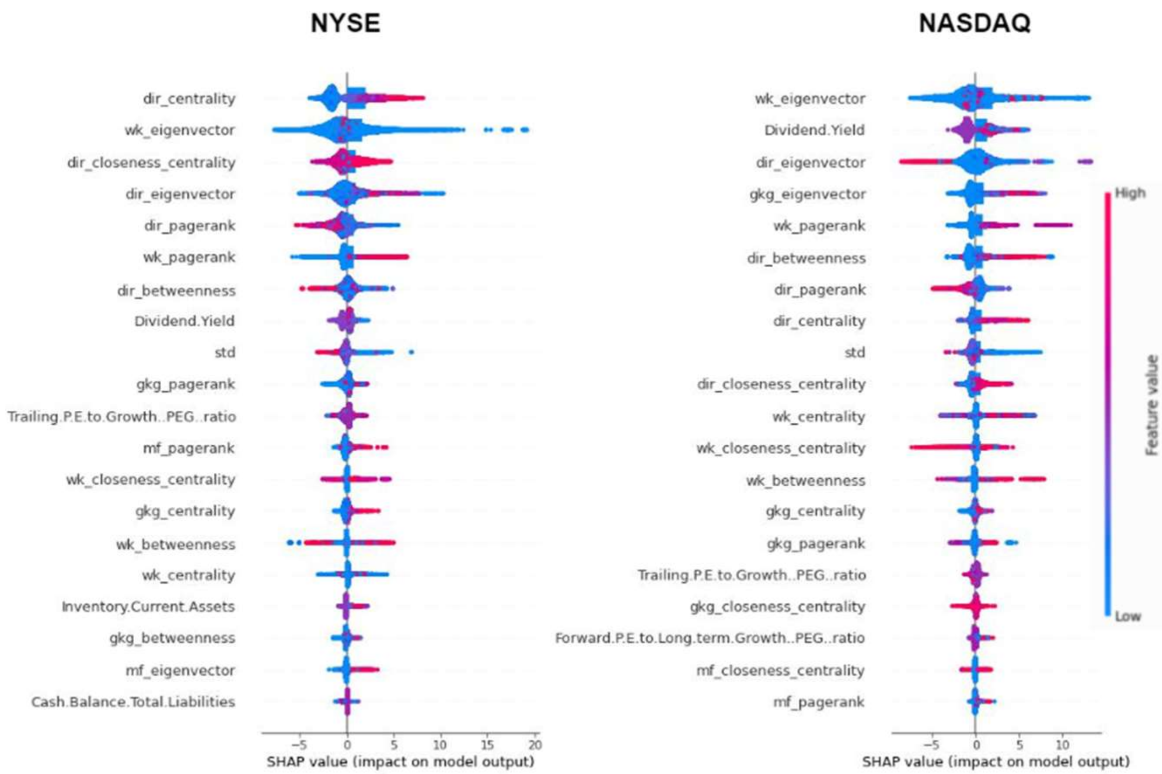
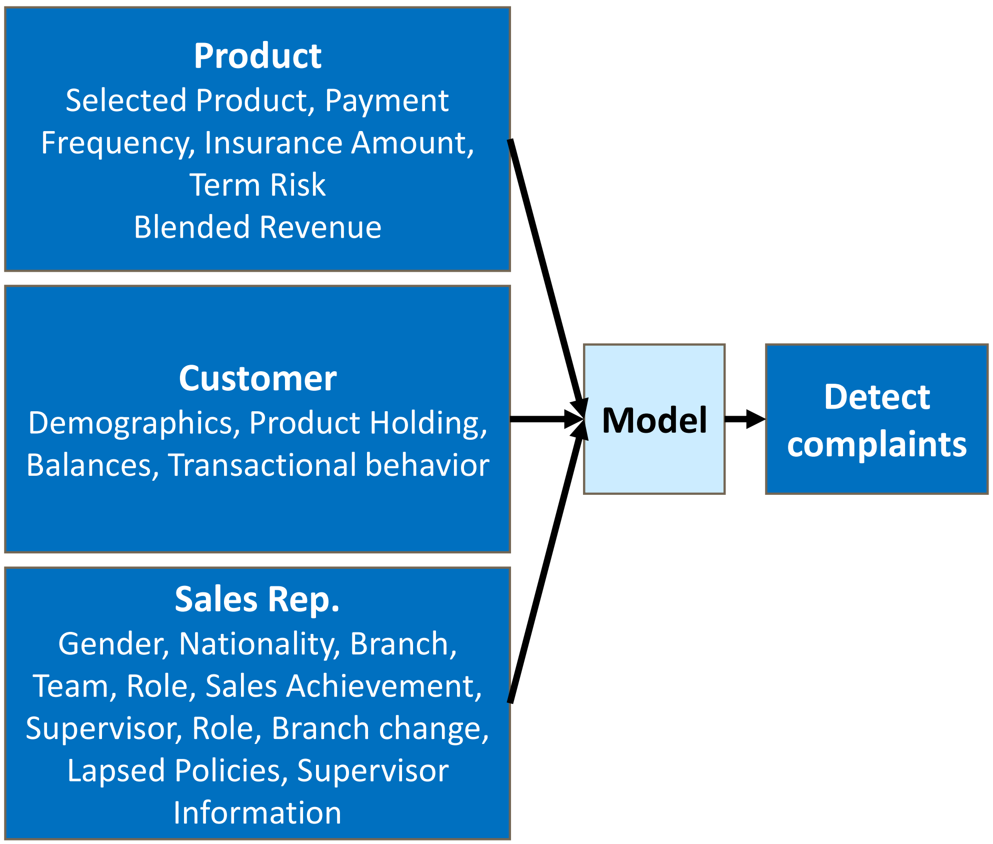


Figure 3: SHAP Importances. Figure shows the beeswarm plots for Tot. ESG Ratings for NYSE and NASDAQ datasets. Beeswarm plots show all the SHAP values, grouped by the features on the y-axis, with the SHAP values on the x-axis. The SHAP values indicate how much each factor contributed to the model's prediction when compared to the mean prediction. More positive or more negative SHAP values indicate that the feature had a significant positive or negative impact on the model's prediction. For each group, the colour of the points is determined by the value of the feature. For example, for dir_centrality for NYSE, we see that higher values of the dir_centrality features (more reddish shade) correspond to more positive SHAP values, while lower values of the dir_centrality features (more blueish shade) correspond to more negative SHAP values. The features are ordered by the mean SHAP values, i.e. more important features are at the top.

Round-up and Reflections

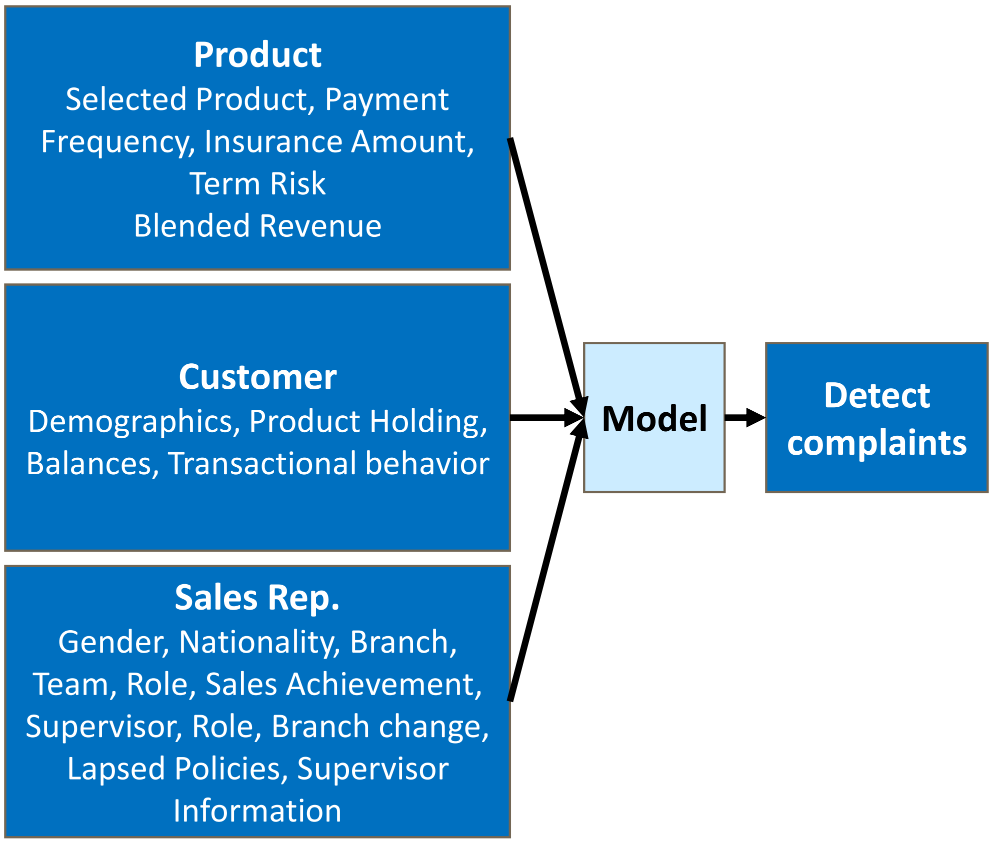
Let’s discuss



Scenario: Insurer collects >5000 attributes to pre-emptively detect complaint cases

- What is the issue with simply dumping all attributes in the model?
- What are some techniques that can address this?

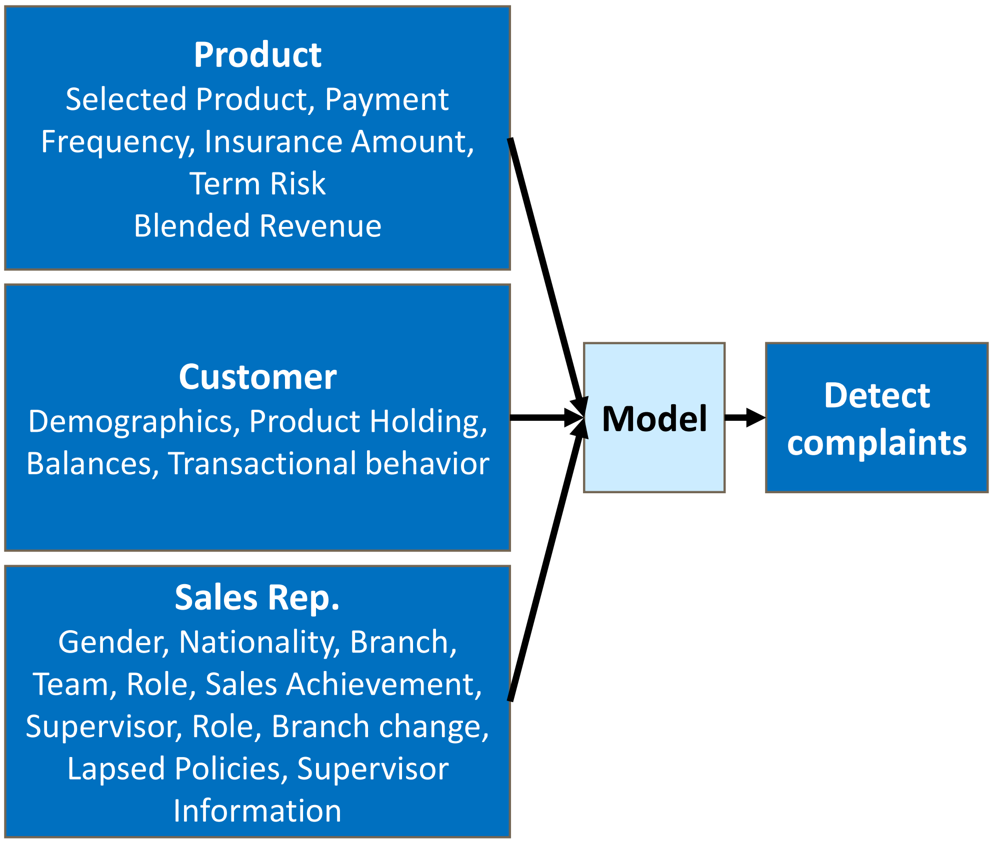
Let’s discuss



Scenario: Insurer collects >5000 attributes to pre-emptively detect complaint cases

- What type of learning problem is this?
- What are some models that are suitable?

Let’s discuss



Scenario: Insurer collects >5000 attributes to pre-emptively detect complaint cases

- What evaluation metrics would you choose?
- What other methods could you apply to understand whether the model is working well?