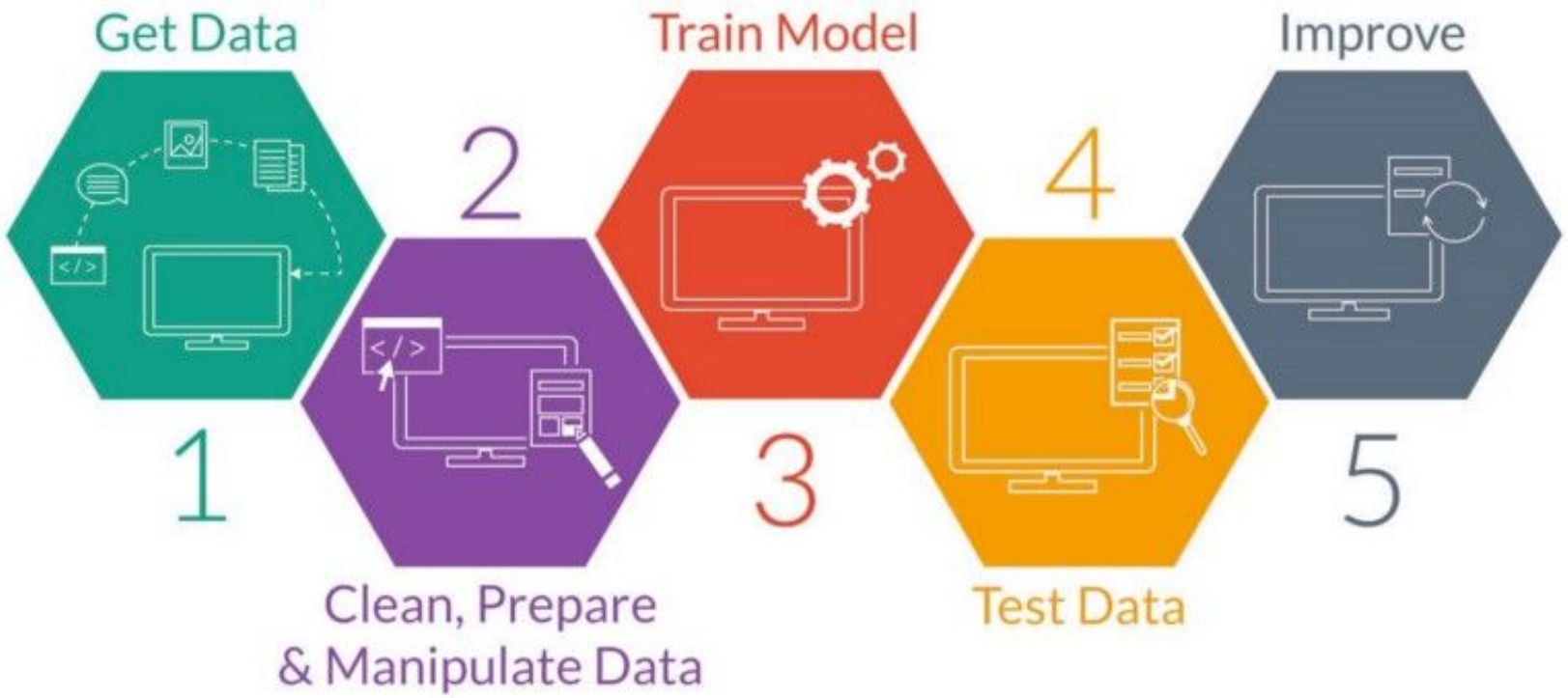
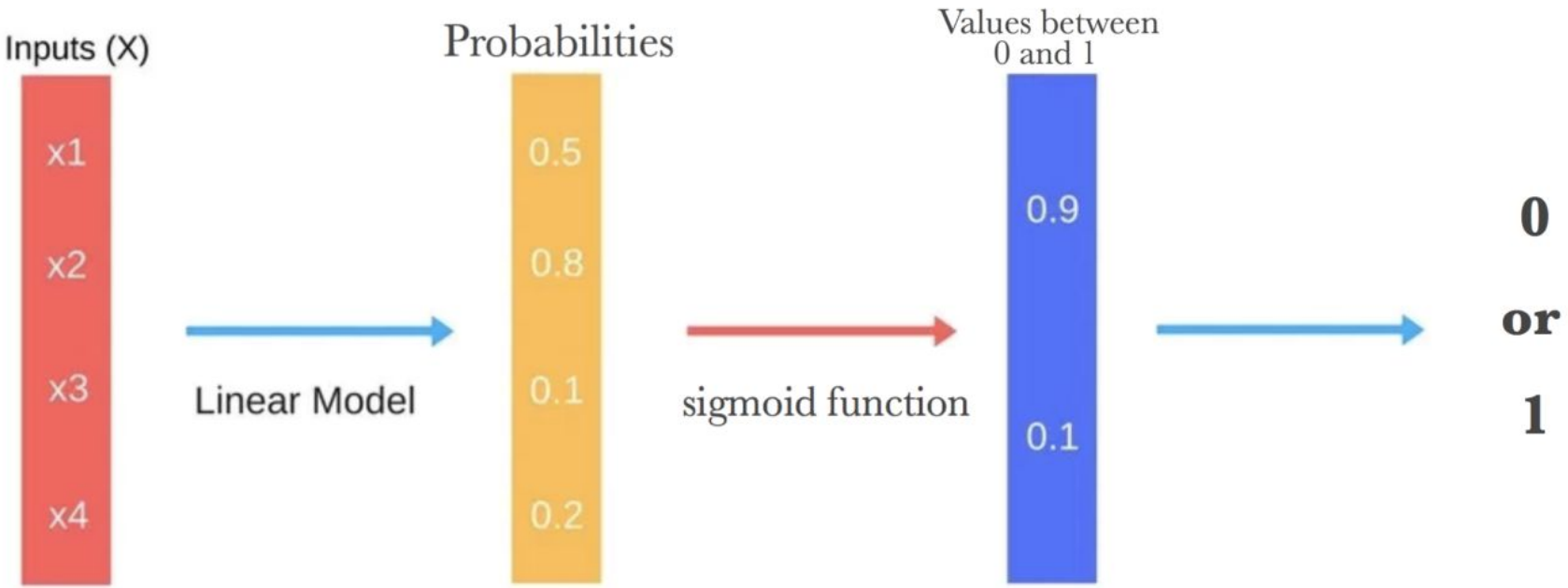
A Brief Tour of Classification Algorithms

Vanessa Rivera Quinones

ML Pipeline


Get Data
1
2
Clean, Prepare
& Manipulate Data
Train Model
3
4
Test Data
Improve
5

Logistic Regression

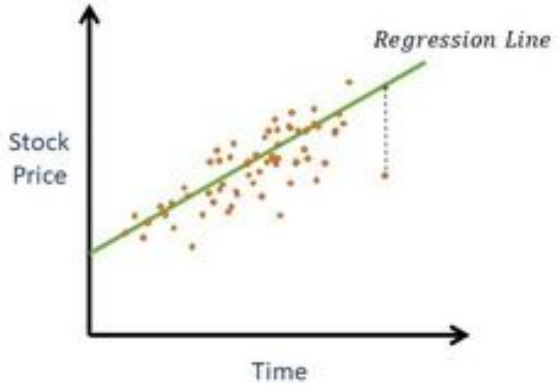


Logistic Regression

Linear Regression Vs Logistic Regression

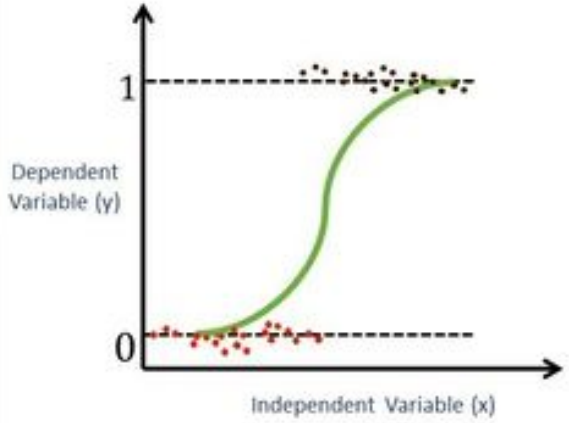
Linear Regression

- Aim is to predict continuous valued output.
- Output value can be any possible integer number.

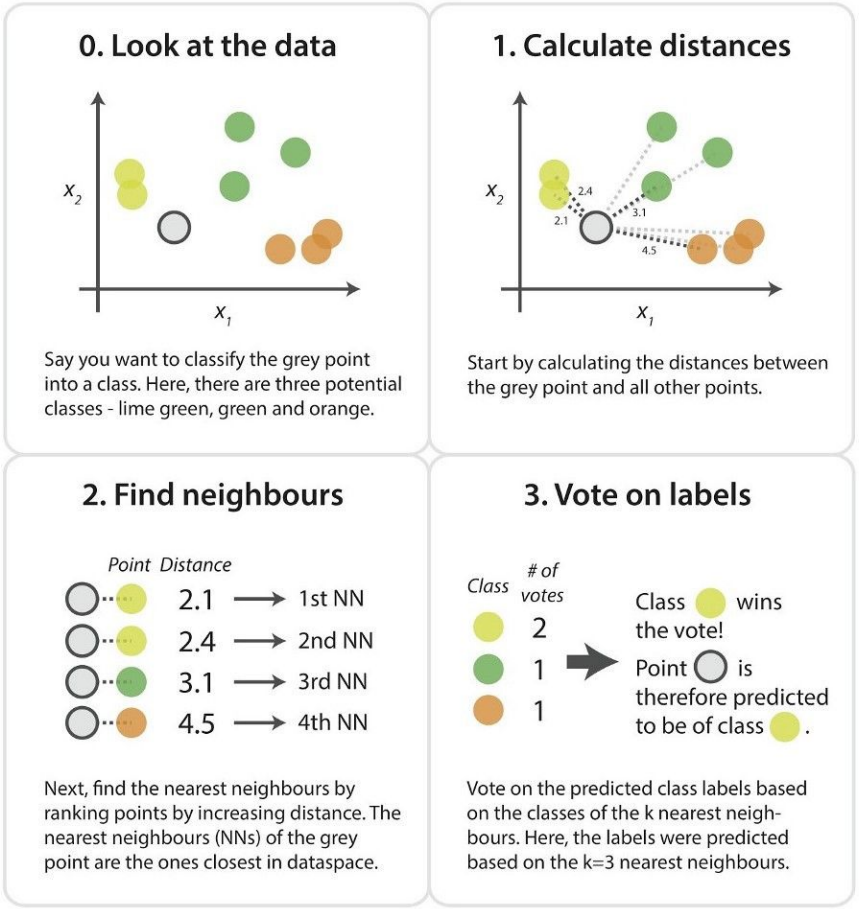


Logistic Regression

- Aim is to predict the label for input data.
- Output is categorical (Binary) i.e. 0/1, True/False, etc.

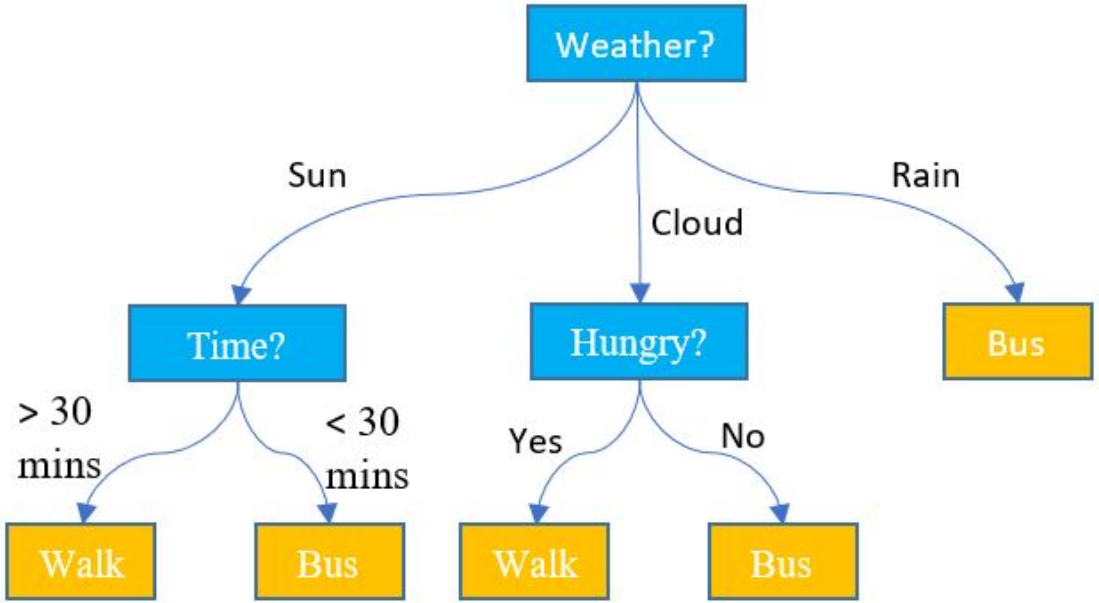


KNN

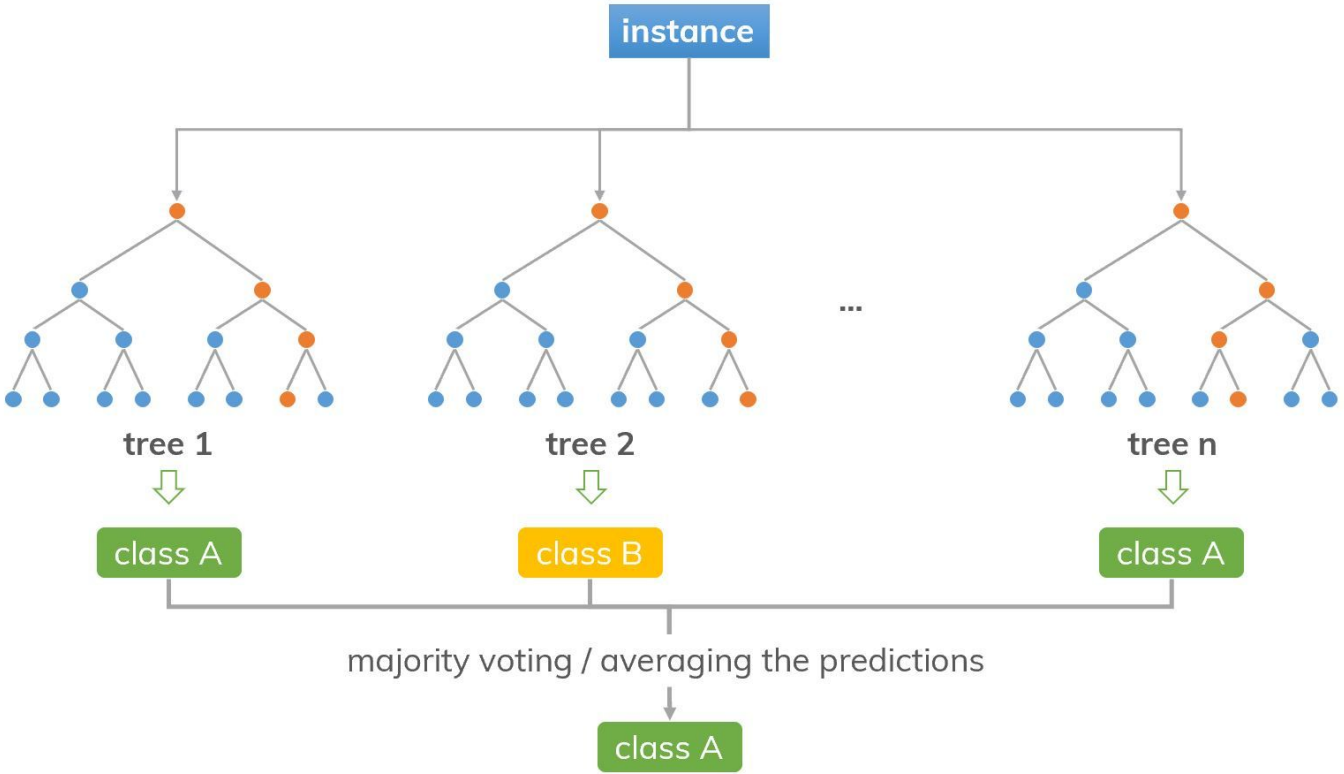


Source: How to find the optimal value of K in KNN?

Decision Trees



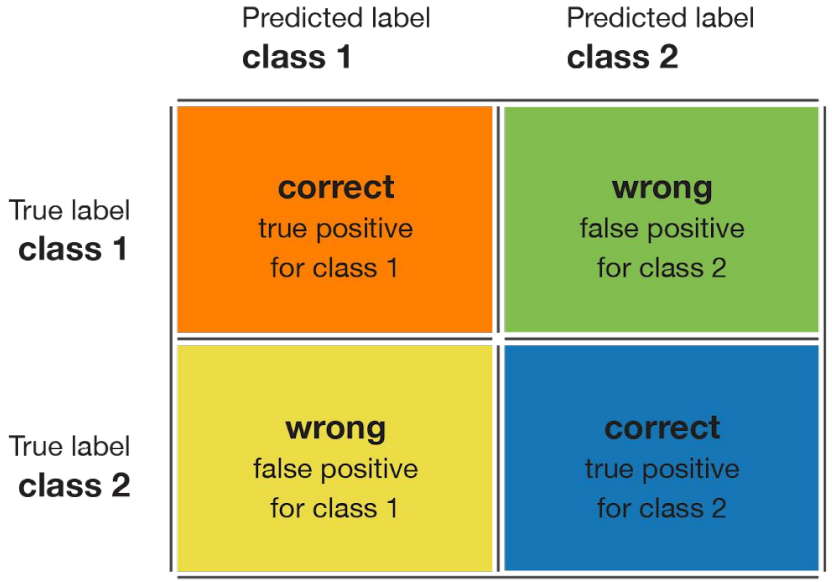
Random Forest

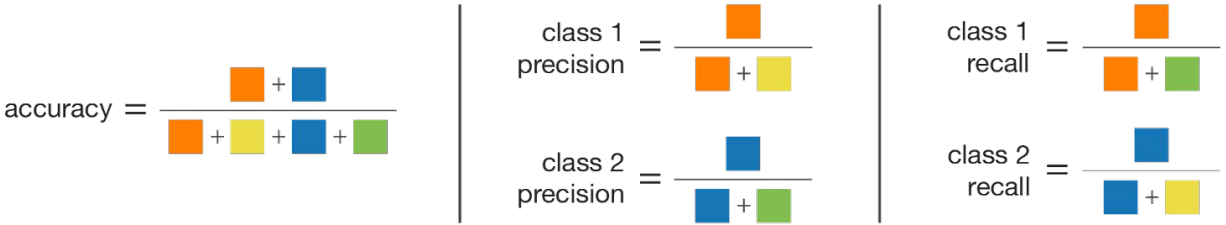


Metrics

Metric	Measures	In Scikit-learn
Precision	How many selected are relevant?	from sklearn.metrics import precision_score
Recall	How many relevant were selected?	from sklearn.metrics import recall_score
F1	Weighted average of precision & recall	from sklearn.metrics import f1_score
Confusion Matrix	True positives, true negatives, false positives, false negatives	from sklearn.metrics import confusion_matrix
ROC	True positive rate vs. false positive rate, as classification threshold varies	from sklearn.metrics import roc
AUC	Aggregate accuracy, as classification threshold varies	from sklearn.metrics import auc

Metrics


Predicted label
class 1
Predicted label
class 2
True label
class 1
True label
class 2
correct
true positive
for class 1
wrong
false positive
for class 2
wrong
false positive
for class 1
correct
true positive
for class 2


accuracy =
class 1
precision =
class 2
precision =
class 1
recall =
class 2
recall =

Metrics in Context

Bad Loan = 1

Good Loan = 0

Cost of **FN** > Cost of **FP**

Actual

Predict		**Bad Loan (1)**	**Good Loan (0)**
	Bad Loan (1)	✔ TP - 559	✘ FP - 0
	Good Loan (0)	✘ FN - 33	✔ TN - 22

Accuracy: Out of the total prediction made, how many did we predict correctly?

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuray = (559+22)/(559+22+33+0) = 95%

Precision: Out of the loan that is predicted as a bad loan, how many did we classify correctly?

$$\textit{Precision} = \frac{TP}{TP + FP}$$

Precision = *559/(559+0) = 100%*

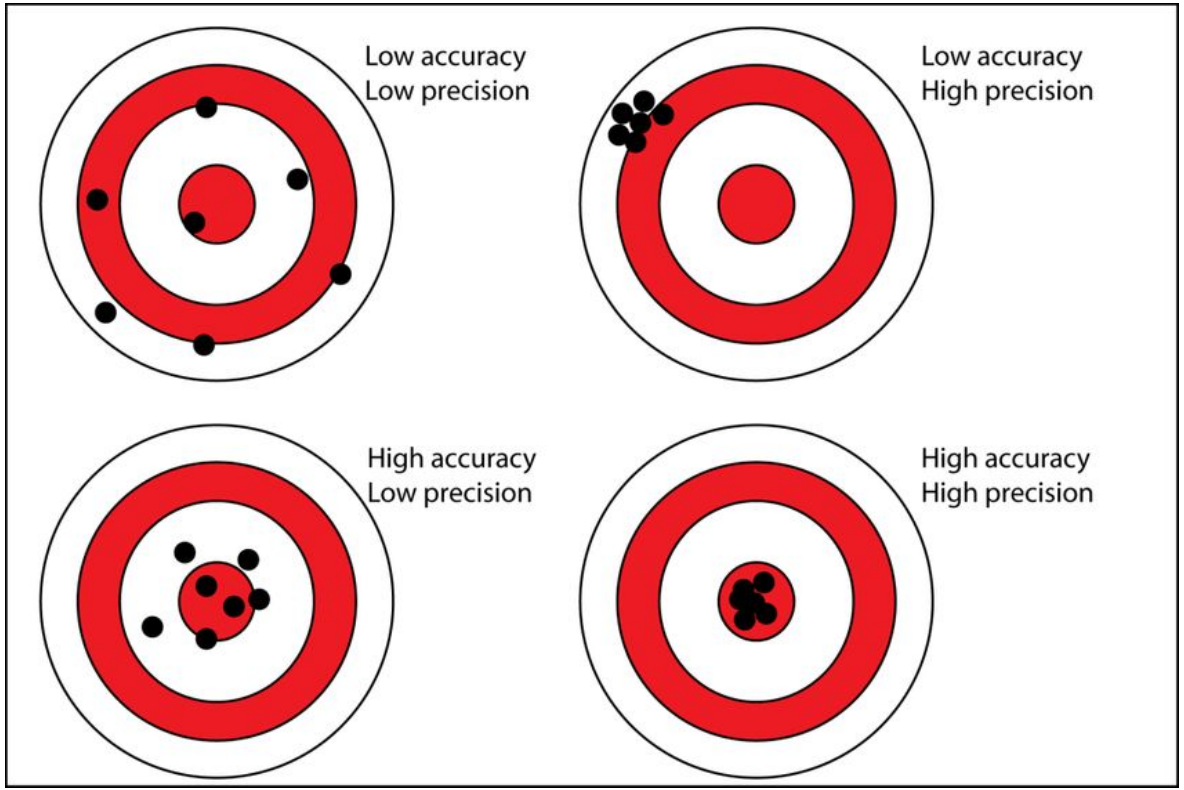
Recall: Out of the **actual** bad loan, how many did we correctly predict as a bad loan?

$$\textit{Recall} = \frac{TP}{TP + FN}$$

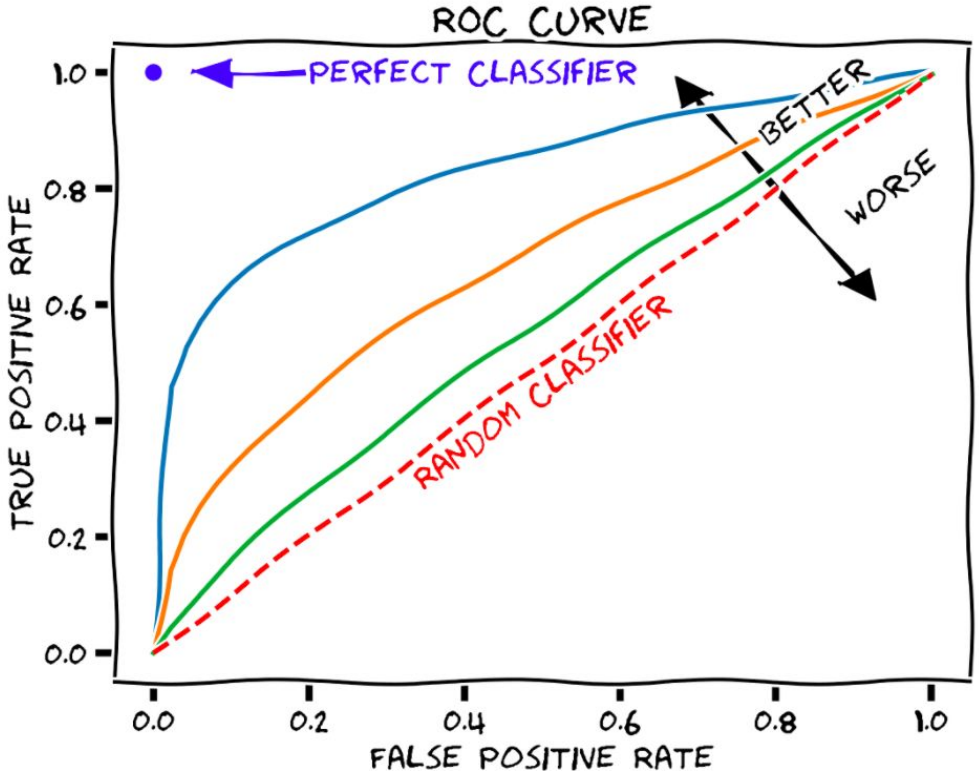
Instead of using **accuracy**,we should evaluate **recall**. If we can decrease **FN**, the recall will increase.

Recall = *559/(559+33) = 94.5%*

Precision vs. Accuracy



ROC



- A perfect classifier has 0 false positives (0.0 rate) and correctly predicts all true positives (1.0 rate)
- If you flip a coin (random classifier), your rates will be roughly the same.

Interesting Resources

- https://www.kaggle.com/
- https://www.drivendata.org/