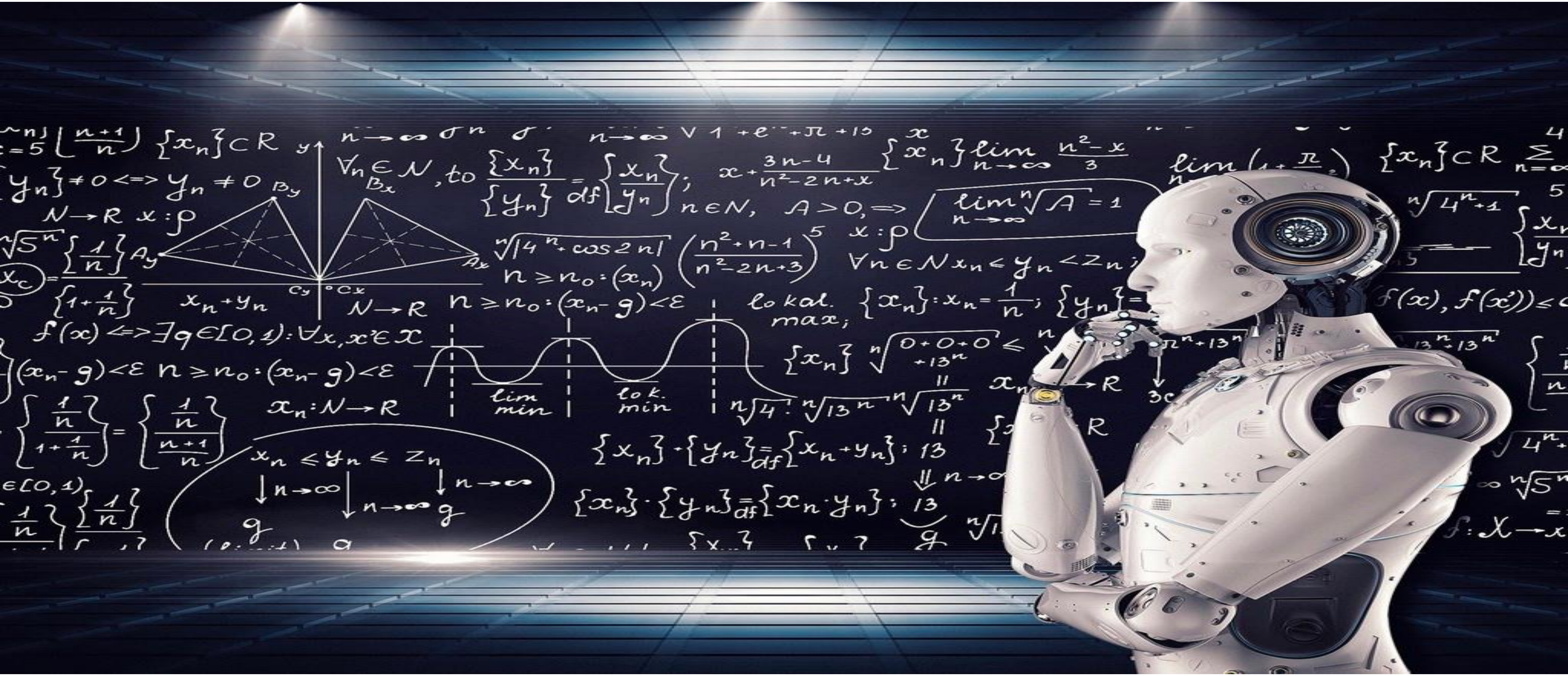
Performance Evaluation



Agenda

- Confusion Matrix
- KPIs
- ROC, PR

Confusion Matrix

True / Predicted	positive	negative
positive		
negative		

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

IDC HERZLIYA

Confusion Matrix

True / Predicted	positive	negative
positive	TPs	FNs
negative	FPs	TNs

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

Confusion Matrix

True / Predicted	positive	negative
positive	TPs=3	FNs
negative	FPs	TNs

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

Confusion Matrix

True / Predicted	positive	negative
positive	**TPs=3**	**FNs=**
negative	**FPs=2**	**TNs=**

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

Confusion Matrix

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

Confusion Matrix

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

Sample ID	Predicted class	True (concept) class
1	T	T
2	T	T
3	T	T
4	T	F
5	T	F
6	F	F
7	F	F
8	F	F
9	F	F
10	F	F
11	F	F
12	F	T
13	F	T
14	F	T

IDC HERZLIYA

Agenda

- Confusion Matrix
- KPIs
- ROC, PR

Confusion Matrix - KPIs

- Accuracy
- True positive rate (Recall)
- True negative rate (specificity)
- Precision
- F-score
- Alert rate

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

Confusion Matrix - KPIs

- Accuracy = what is the precentage of samples that are correctly labelled?
- $Accuracy = \frac{TP+TN}{P+N}$
  - In our example Accuracy=9/14=64.2%

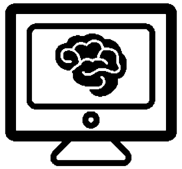
True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

Confusion Matrix - KPIs

- Accuracy = what is the precentage of samples that are correctly labelled?
- $Accuracy = \frac{TP+TN}{P+N}$
  - In our example Accuracy=9/14=64.2%
- Q: Is a classifier with 90% accuracy a good classifier?

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

Confusion Matrix - KPIs



True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

- Accuracy = what is the precentage of samples that are correctly labelled?
- $Accuracy = \frac{TP+TN}{P+N}$
  - In our example Accuracy=9/14=64.2%
- Q: Is a classifier with 90% accuracy a good classifier?
- A: Depending on the dataset!

True / Predicted	positive	negative
positive	TPs=30	FNs=20
negative	FPs=80	TNs=870

True / Predicted	positive	negative
positive	TPs=250	FNs=50
negative	FPs=50	TNs=650

IDC HERZLIYA

Confusion Matrix - KPIs

- Accuracy = what is the precentage of samples that are correctly labelled?
- $Accuracy = \frac{TP+TN}{P+N}$
  - In our example Accuracy=9/14=64.2%
- Q: Is a classifier with 90% accuracy a good classifier?
- A: Depending on the dataset!

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

True / Predicted	positive	negative
positive	TPs=30	FNs=20
negative	FPs=80	TNs=870

T = 50, F=950 : 5% positive rate

True / Predicted	positive	negative
positive	TPs=250	FNs=50
negative	FPs=50	TNs=650

T = 300, F=700 : 30% positive rate

Confusion Matrix - KPIs

- Q: Is a classifier with 90% accuracy a good classifier?
- A: Depending on the dataset!

True / Predicted	positive	negative
positive	TPs=30	FNs=20
negative	FPs=80	TNs=870

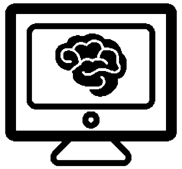
T = 50, F=950 : 5% positives

True / Predicted	positive	negative
positive	TPs=250	FNs=50
negative	FPs=50	TNs=650

T = 300, F=700 : 30% positives

- Let's compare to a Majority classifier –
  - Predict "False" for all samples
  - In the left dataset we'll get 95% accuracy
  - In the right dataset we'll get 70% accuracy
- $accuracy\ lift\ over\ majority = \frac{classifier\_accuracy}{mojority\_accuracy}$

Confusion Matrix - KPIs



- Accuracy
- True positive rate (Recall)
- False positive rate (FPR)
- Precision
- Alert rate

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

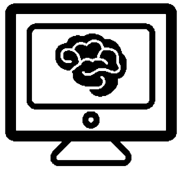
Confusion Matrix - KPIs

- True positive rate (Recall)
  - Out of all the positive samples, how much did the classifier "catch"
  - $TPR = Recall = \frac{TP}{P}$
- False positive rate (FPR)
  - Out of all the negative samples, how much did the classifier mark as positive
  - $FPR = specificity = \frac{FP}{N}$

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

**we'll use those KPIs for the ROC curve*

Confusion Matrix - KPIs

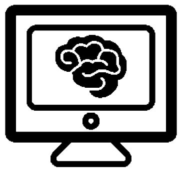


True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

- True positive rate (Recall)
  - Out of all the positive samples, how much did the classifier “catch”
  - $TPR = Recall = \frac{TP}{P}$
- Precision (positive predictive value - PPV)
  - Out of all the samples the classifier marked as True, on what percent was it correct
  - $PPV = precision = \frac{TP}{TP+FP}$

**we’ll use those KPIs for the PR curve*

Confusion Matrix - KPIs



True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

- F1 Score
  - Harmonic mean between precision and recall
  - $F_1 = 2 * \frac{recall * precision}{recall + precision}$
  - Much more affected by the lower value

Recall, precision	Average	F1
R=1, P=0	0.5	**0**
R=0.9, P=0	0.45	**0**

Confusion Matrix - KPIs

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

- F1 Score
  - Harmonic mean between precision and recall
  - $F_1 = 2 * \frac{recall * precision}{recall + precision}$
  - Much more affected by the lower value

Recall, precision	Average	F1
R=1, P=0	0.5	**0**
R=0.9, P=0	0.45	**0**
R=0.8, P=0.3	0.55	0.436

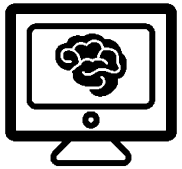
Confusion Matrix - KPIs

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

- F1 Score
  - Harmonic mean between precision and recall
  - $F_1 = 2 * \frac{recall*precision}{recall+precision}$
  - Much more affected by the lower value

Recall, precision	Average	F1
R=1, P=0	0.5	0
R=0.9, P=0	0.45	0
R=0.8, P=0.3	0.55	0.436
R=0.7, P=0.7	0.7	0.7

Confusion Matrix - KPIs



- "Alert rate" (non formal name)
  - % of samples marked as True
  - $alert\_rate = \frac{TP+FP}{T+F}$

- Very useful where part of the dataset is ***not*** labeled

True / Predicted	positive	negative
positive	TPs=3	FNs=3
negative	FPs=2	TNs=6

Agenda

- Confusion Matrix
- KPIs
- ROC, PR

Probability Based Classifier

- Many classifiers predict the *probability* of sample to belong to the Positive class
- Choosing different thresholds yeilds different confusion matrices

Sample ID	Predicted probability	True (concept) class
1	0.95 T	T
2	0.9 T	T
3	0.89 T	T
4	0.85 T	F
5	0.8 T	F
6	0.65 T	T
7	0.6 T	F
8	0.6 T	F
9	0.58 T	F
10	0.3 F	F
11	0.21 F	T
12	0.13 F	F
13	0.01 F	T
14	0.01 F	F

Probability Based Classifier

- Many classifiers predict the *probability* of sample to belong to the Positive class
- Choosing different thresholds yeilds different confusion matrices

Sample ID	Predicted probability	True (concept) class
1	0.95 T	T
2	0.9 T	T
3	0.89 T	T
4	0.85 T	F
5	0.8 T	F
6	0.65 T	T
7	0.6 T	F
8	0.6 T	F
9	0.58 T	F
10	0.3 F	F
11	0.21 F	T
12	0.13 F	F
13	0.01 F	T
14	0.01 F	F

IDC HERZLIYA

Probability Based Classifier

- Many classifiers predict the *probability* of sample to belong to the Positive class
- Choosing different thresholds yeilds different confusion matrices
- Thershold = 0.7
  - Precision = 3/5
  - Recall = 3/6
  - FPR =2/8

Sample ID	Predicted probability	True (concept) class
1	0.95 T	T
2	0.9 T	T
3	0.89 T	T
4	0.85 T	F
5	0.8 T	F
6	0.65 T	T
7	0.6 T	F
8	0.6 T	F
9	0.58 T	F
10	0.3 F	F
11	0.21 F	T
12	0.13 F	F
13	0.01 F	T
14	0.01 F	F

IDC HERZLIYA

Probability Based Classifier

- Many classifiers predict the *probability* of sample to belong to the Positive class
- Choosing different thresholds yeilds different confusion matrices
- Thershold = 0.7
  - Precision = 3/5
  - Recall = 3/6
  - FPR =2/8
- Thershold = 0.5
  - Precision = 3/9
  - Recall = 3/6
  - FPR =5/8

Sample ID	Predicted probability	True (concept) class
1	0.95 T	T
2	0.9 T	T
3	0.89 T	T
4	0.85 T	F
5	0.8 T	F
6	0.65 T	T
7	0.6 T	F
8	0.6 T	F
9	0.58 T	F
10	0.3 F	F
11	0.21 F	T
12	0.13 F	F
13	0.01 F	T
14	0.01 F	F

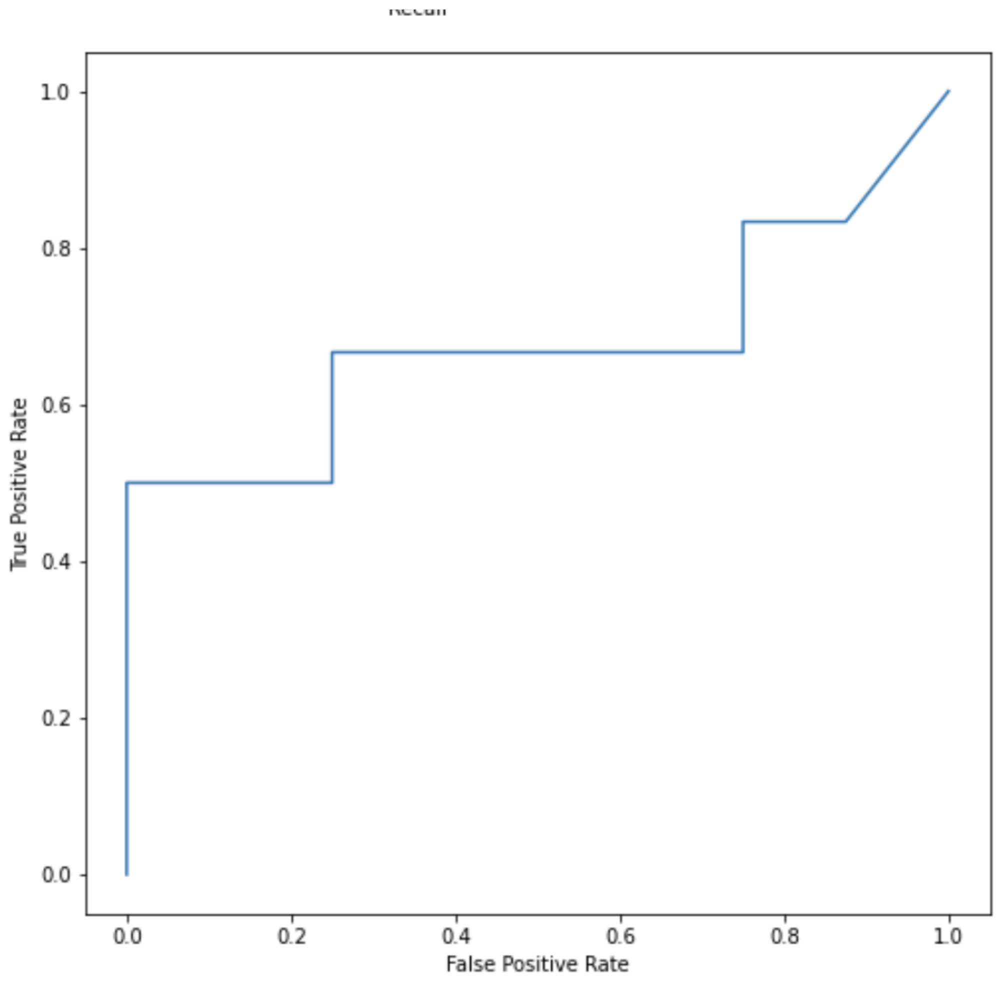
PR, ROC curves

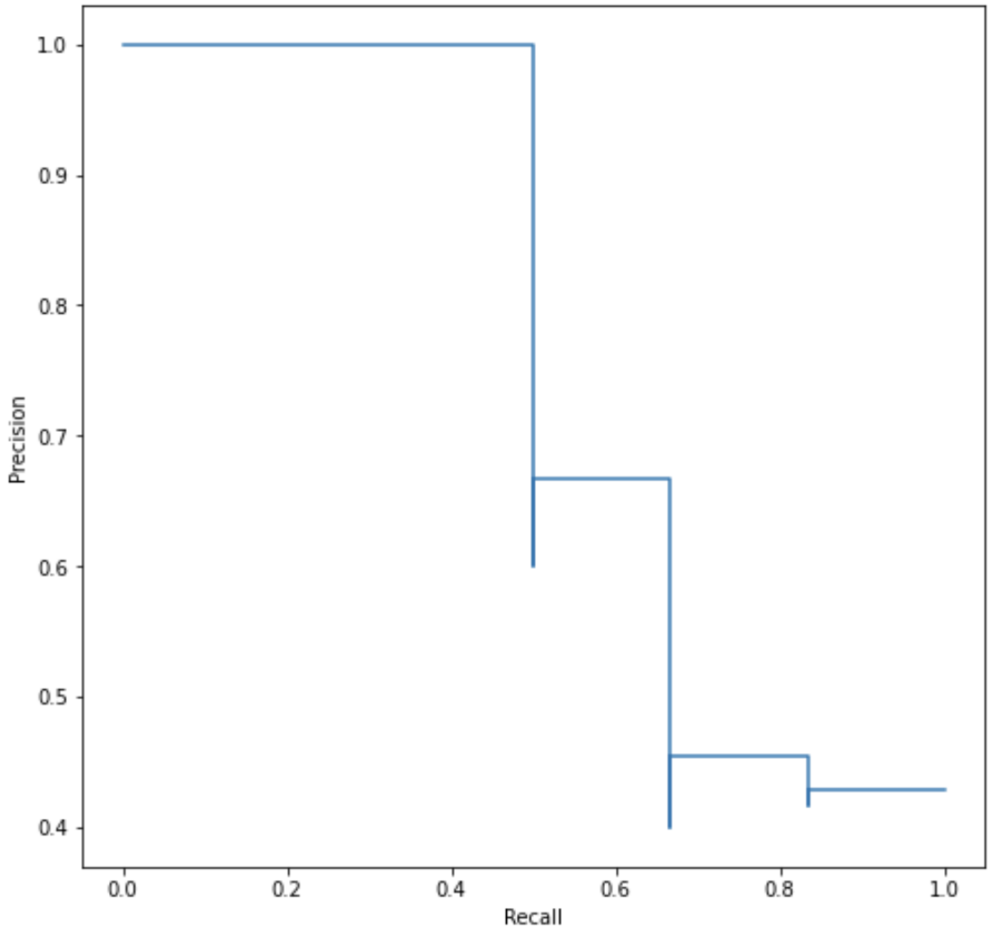
- Each point in the ROC curve fits a different threshold

Thresh	Precision	Recall	FPR=FP/N
0	0.43	1	1
0.05	0.42	0.83	0.875
0.2	0.45	0.83	0.75
0.25	0.4	0.67	0.75
0.4	0.44	0.67	0.625
0.59	0.5	0.67	0.5
0.62	0.67	0.67	0.25
0.7	0.6	0.5	0.25
0.82	0.75	0.5	0.167
0.87	1	0.5	0
0.89	1	0.33	0
0.92	1	0.17	0

Sample ID	Predicted probability	True (concept) class
1	0.95 T	T
2	0.9 T	T
3	0.89 T	T
4	0.85 T	F
5	0.8 T	F
6	0.65 T	T
7	0.6 T	F
8	0.6 T	F
9	0.58 T	F
10	0.3 F	F
11	0.21 F	T
12	0.13 F	F
13	0.01 F	T
14	0.01 F	F

PR, ROC curves

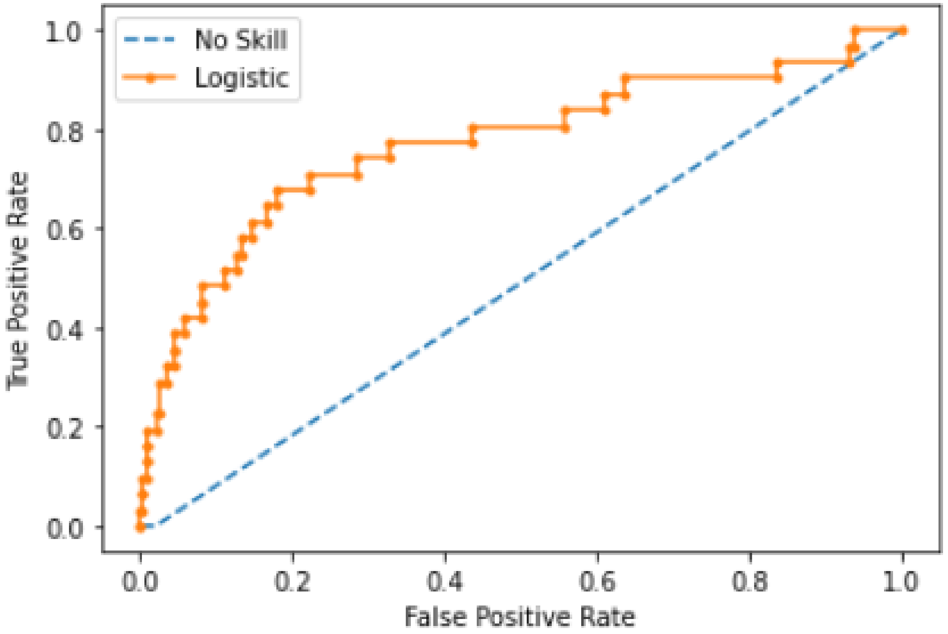




PR and ROC curves for imbalanced data

- A random dataset of 4000 samples , 1% positive class, 99% negative class

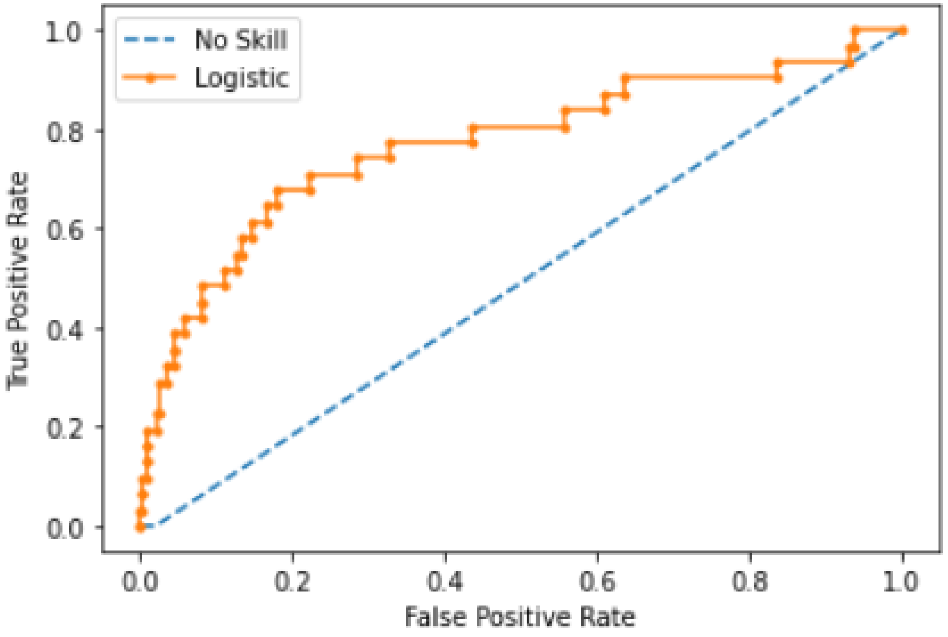
```
DummyClassifier(strategy='stratified')
No Skill ROC AUC 0.490
LogisticRegression()
Logistic ROC AUC 0.771
```



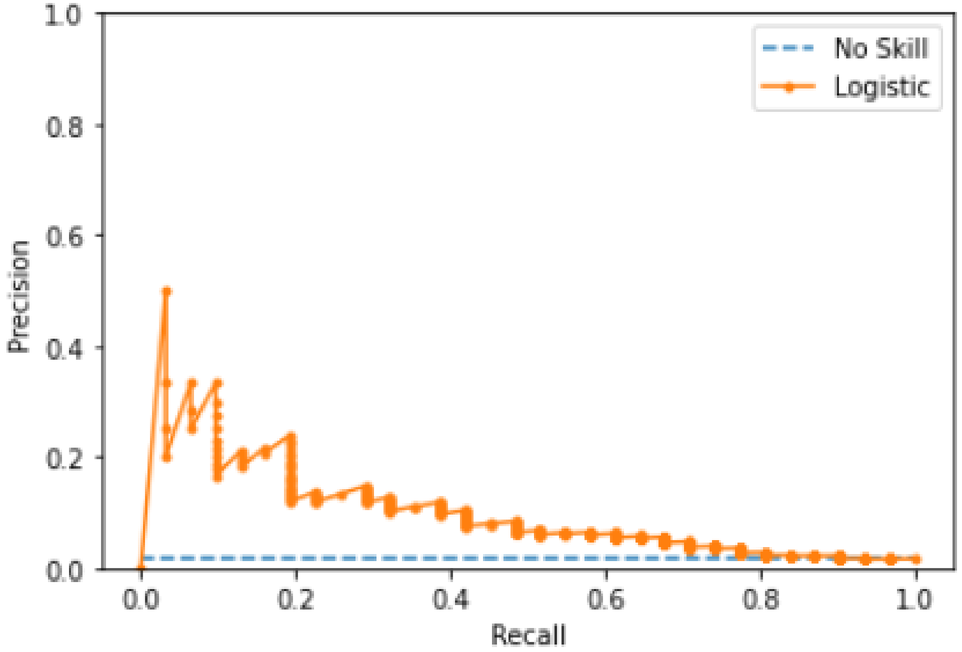
PR and ROC curves for imbalanced data

- A random dataset of 4000 samples , 1% positive class, 99% negative class

```
DummyClassifier(strategy='stratified')
No Skill ROC AUC 0.490
LogisticRegression()
Logistic ROC AUC 0.771
```




```
DummyClassifier(strategy='stratified')
No Skill PR AUC: 0.036
LogisticRegression()
Logistic PR AUC: 0.098
```

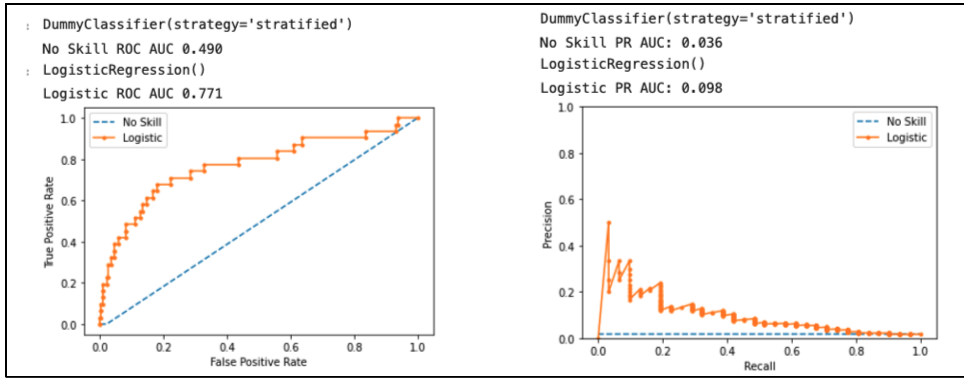


Agenda

- Confusion Matrix
- KPIs
- ROC, PR
- Comparing Models

Comparing models

- In many cases – this comes from a business requirement
- AUC
  - Area under the ROC/PR curve
  - Can help to choose the best model by the highest AUC
  - In practice – can help reject model with low AUC

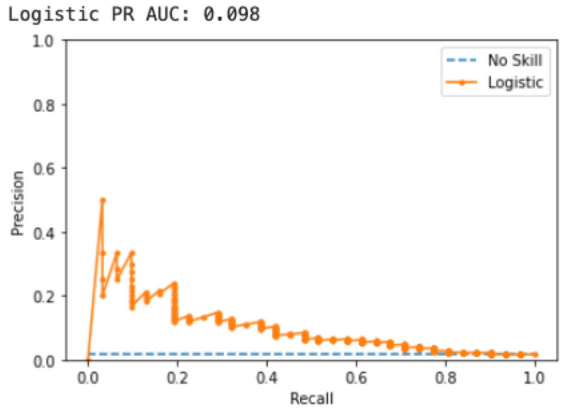


- Required recall , Required precision
  - We can only advertise to customers with conversion rate above 20%
  - We need to catch >90% of the frauds above 1000$
- Precision @ K
  - We can classify 100 cases/day, we need the highest precision
- Max F1
  - We need both good recall and good precision

PR curves – some practical aspects

- The first samples that have the highest score have a large impact on the way the graph looks like
  - If we replace the label of the first couple of samples (highest score) the graph may look very different
  - AUC can dramatically change

	prob	label
1670	0.405384	0
1991	0.360221	1
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0

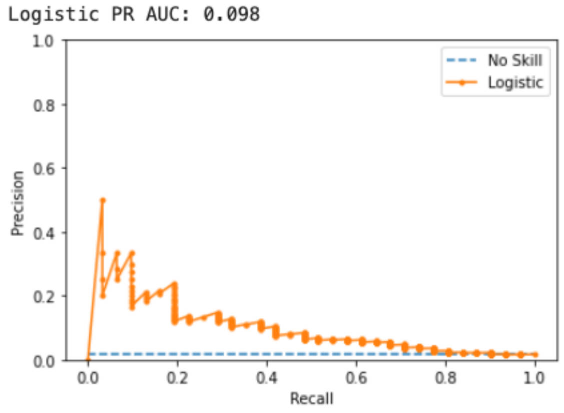


Baseline

PR curves – some practical aspects

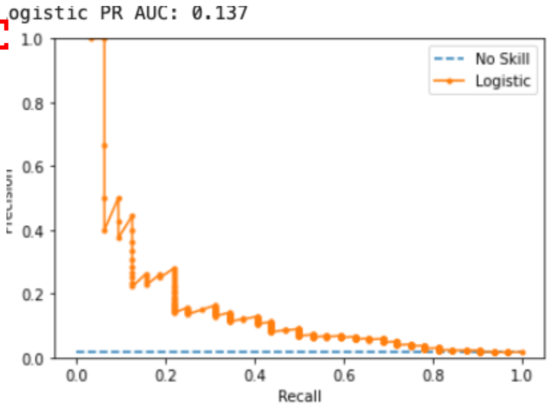
- The first samples that have the highest score have a large impact on the way the graph looks like
  - If we replace the label of the first couple of samples (highest score) the graph may look very different
  - AUC can dramatically change

	prob	label
1670	0.405384	0
1991	0.360221	1
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0



Baseline

	prob	label
1670	0.405384	1
1991	0.360221	1
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0

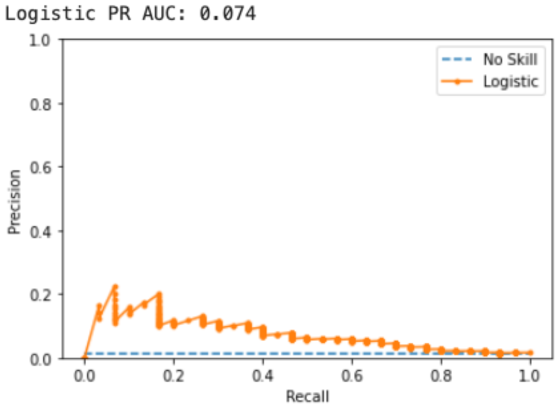


Change sample #1670 from False to True

PR curves – some practical aspects

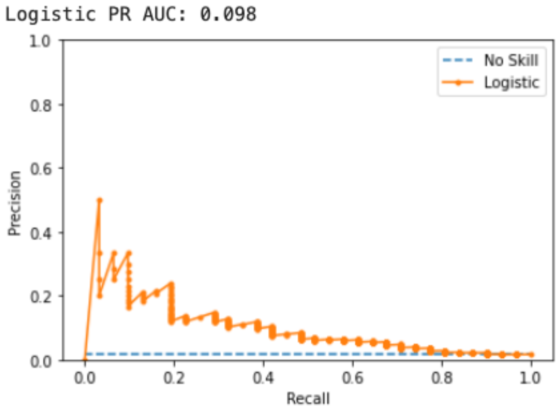
- The first samples that have the highest score have a large impact on the way the graph looks like
  - If we replace the label of the first couple of samples (highest score) the graph may look very different
  - AUC can dramatically change

	prob	label
1670	0.405384	0
1991	0.360221	0
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0



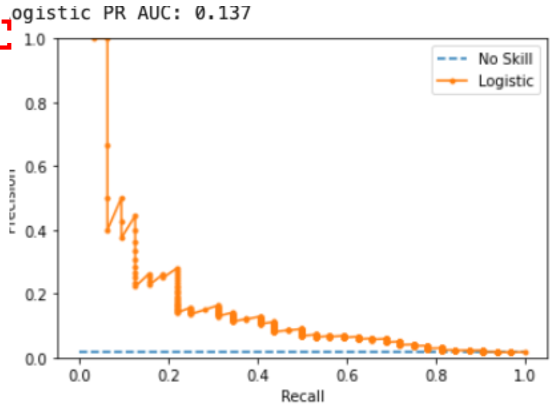
Change sample #1991 from True to False

	prob	label
1670	0.405384	0
1991	0.360221	1
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0



Baseline

	prob	label
1670	0.405384	1
1991	0.360221	1
1363	0.252155	0
1635	0.241596	0
693	0.218432	0
1205	0.187401	1
980	0.172855	0
1928	0.168705	0
1137	0.160115	1
1332	0.154576	0

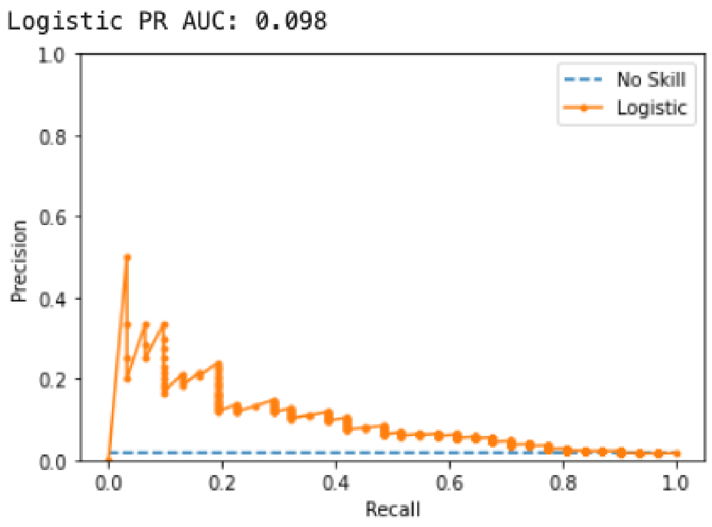


Change sample #1670 from False to True

PR curves – some practical aspects

A “bumpy”/”noisy” PR curve is less indicative of model’s performance

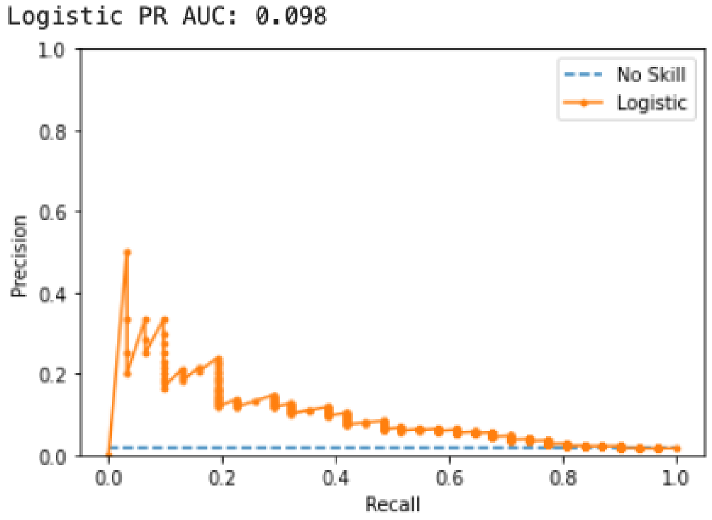
- Apply binning

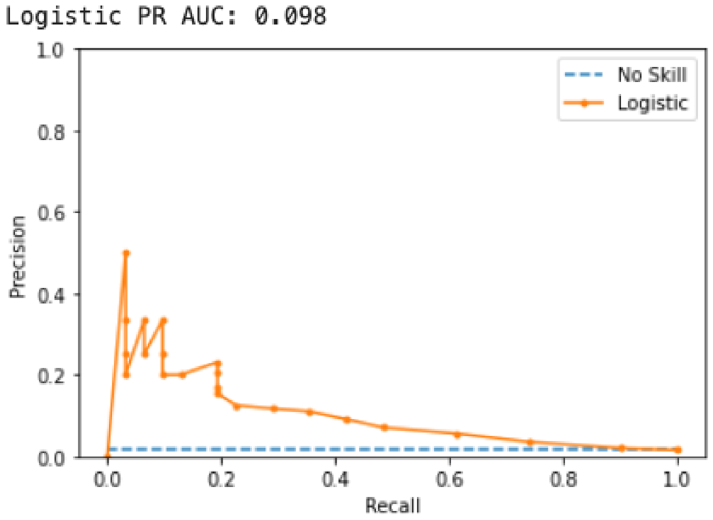


PR curves – some practical aspects

A “bumpy”/”noisy” PR curve is less indicative of model’s performance

- Apply binning



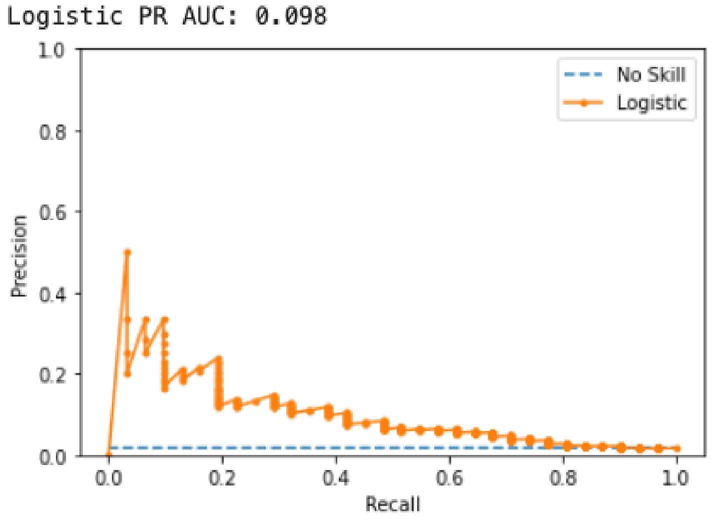


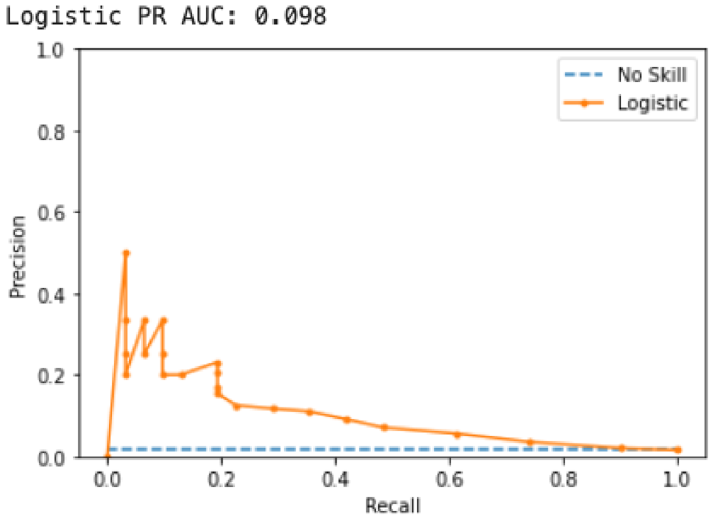
Round 2 digits
(0.218→0.22)

PR curves – some practical aspects

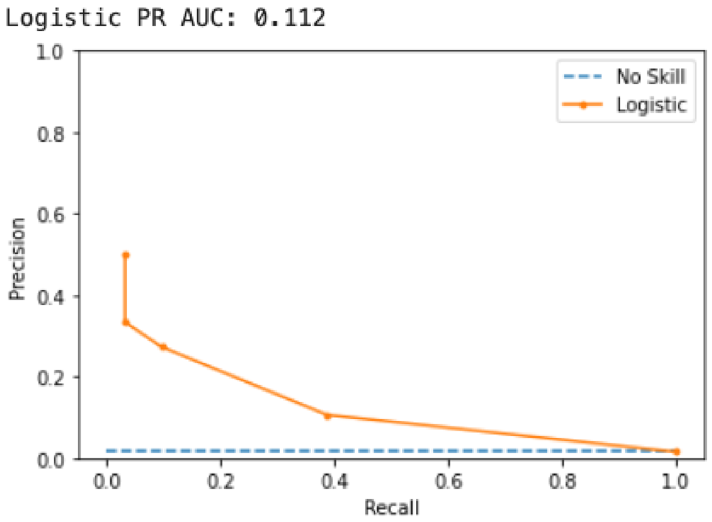
A “bumpy”/”noisy” PR curve is less indicative of model’s performance

- Apply binning





Round 2 digits
(0.218→0.22)



Round 1 digit
(0.218→0.2)