Titanic AI project

Session 3

Team:

Yuliia Nikolaenko

Eduardo Bonnefemne

Gabriel Pérez García

Elena-Mihaela Grigore

Questions to answer:

Effect of sample size on model performances

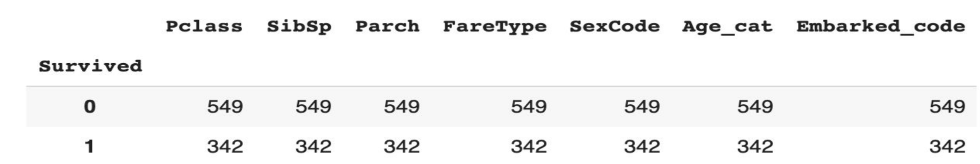
- choose performances indicators for the Titanic predictor;
- try to build and assess performances using different divisions of learning/evaluation/testing sets;
- discuss the results (on kNN and linear models).

Effect of the choice of attributes build model using various subsets of attributes;

- try to automate the generation of subsets, construction (learning) of models and evaluation;
- discuss the impact of attributes: what are the most (or least) importants ?

Work in Progress

Metrics Analyzed in this Notebook	**Further analysis to perform**
Confusion Matrix	Error rate = 1 - Accuracy
Accuracy	R Learning R Val
Balanced Accuracy	F1 scores for each attributes
Matthews correlation coefficient	
ROC	



	Pclass	SibSp	Parch	FareType	SexCode	Age_cat	Embarked_code
Survived							
0	549	549	549	549	549	549	549
1	342	342	342	342	342	342	342

Discussion

- Accuracy and F1 are sensible to imbalanced classes
- F1 score not suitable → dependency with the threshold “s” used in the decision function

Performance Indicators

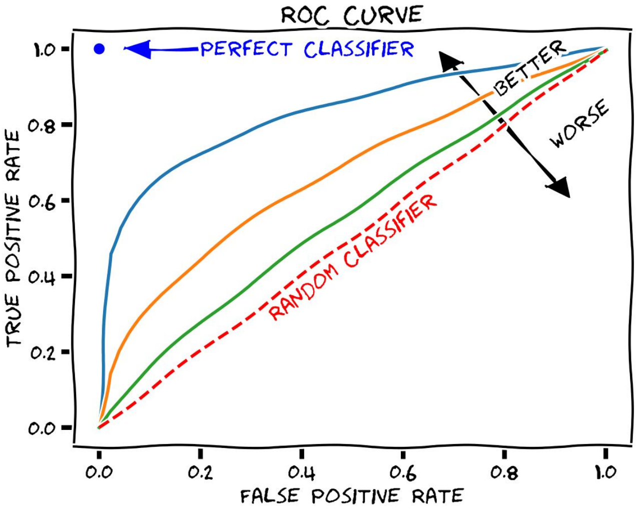
Matthews correlation coefficient

The coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

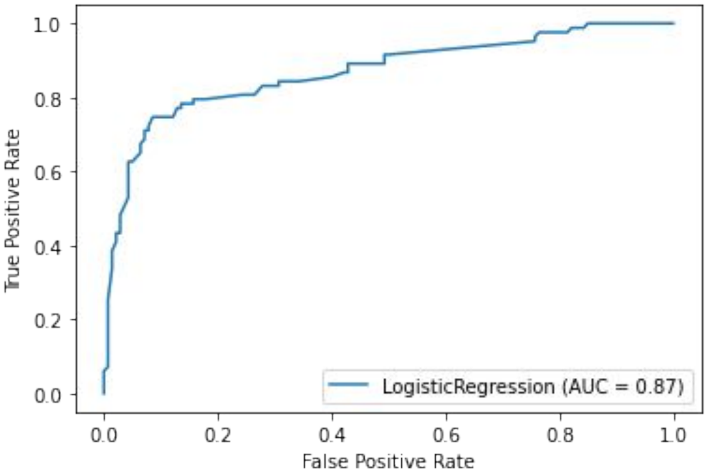
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

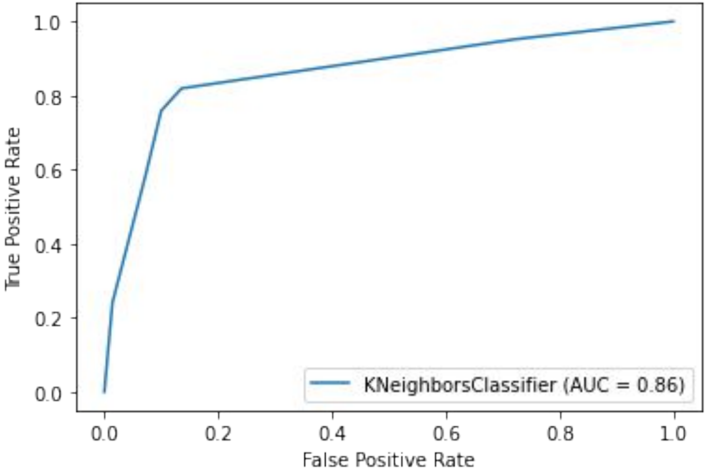
Receiver Operating Characteristic curve (ROC curve)

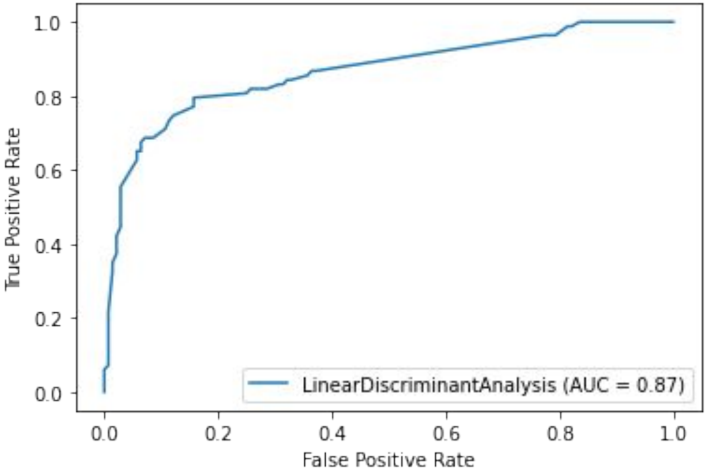
The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



ROC curve results







Part II: Effect of the choice of attributes

Overview

Initial features set (after data wrangling):

Pclass
SibSp
Parch
FareType
SexCode
Age_cat
Embarked_code

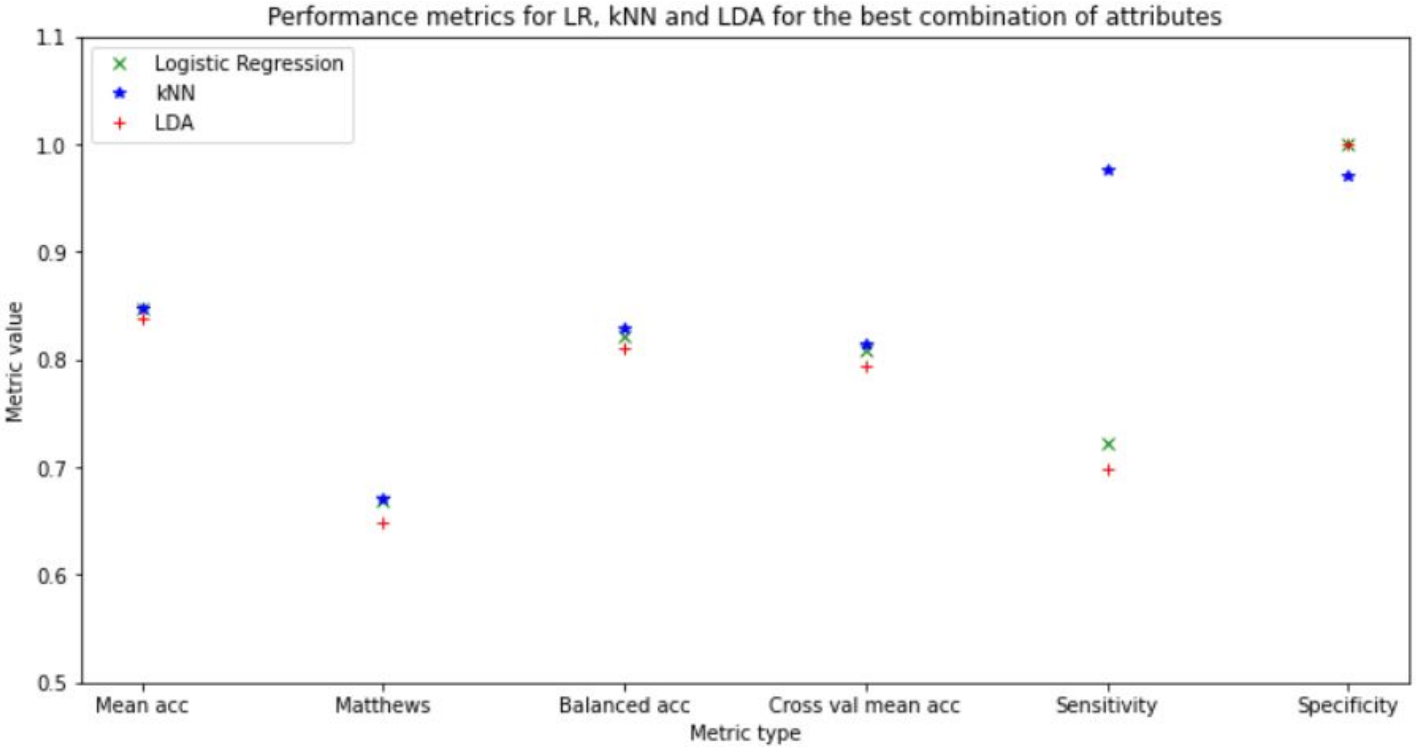
We took all combinations of 1, 2, … 7 attributes ->

And trained LR, kNN (k=5) and LDA models (with the default threshold)->

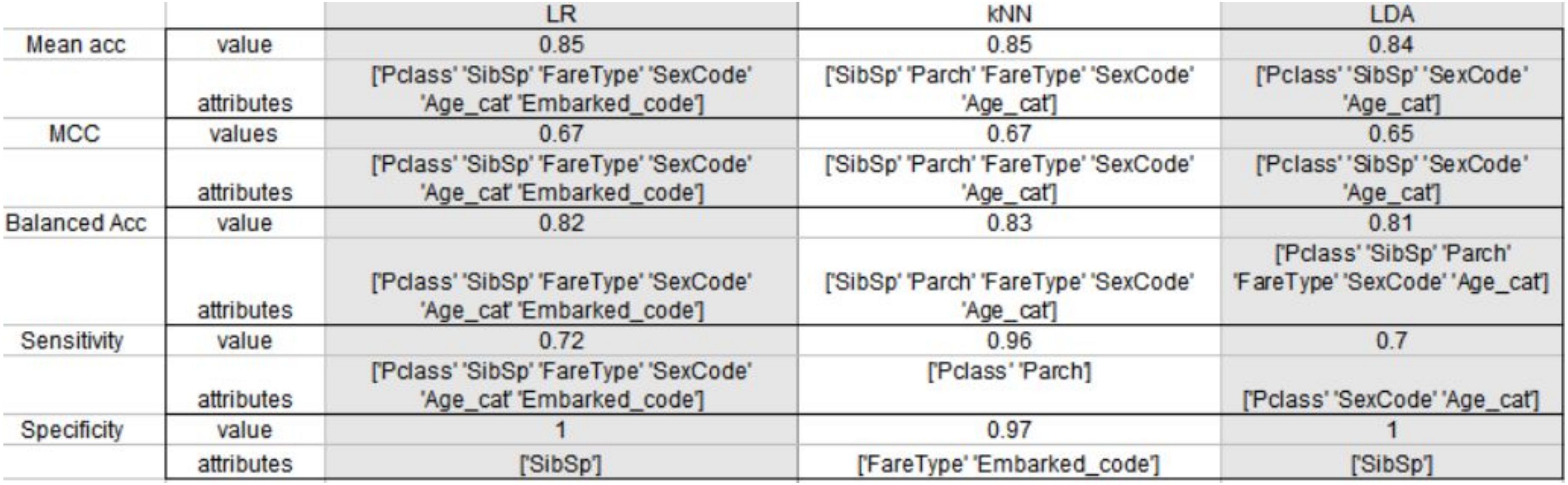
For which we looked at the metrics on the next slide ->

Performance metrics for Regression, kNN, LDA

- Mean accuracy
- Matthews correlation coefficient
- Balanced accuracy
- Cross validation mean accuracy
- Sensitivity
- Specificity



The effect of the choice of attributes



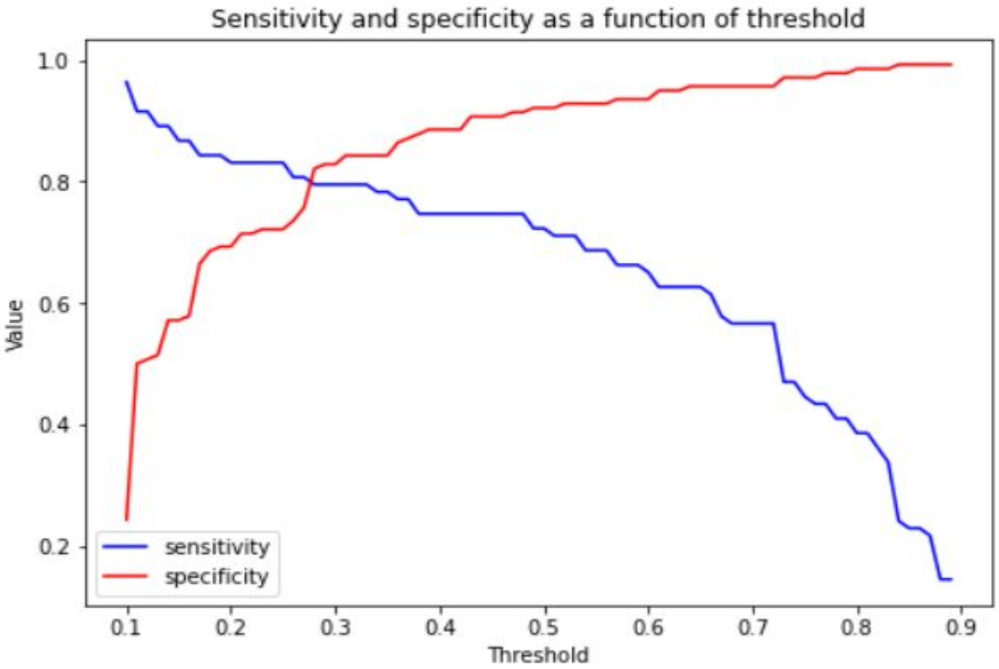
		LR	kNN	LDA
Mean acc	value	0.85	0.85	0.84
	attributes	['Pclass' 'SibSp' 'FareType' 'SexCode' 'Age_cat' 'Embarked_code']	['SibSp' 'Parch' 'FareType' 'SexCode' 'Age_cat']	['Pclass' 'SibSp' 'SexCode' 'Age_cat']
MCC	values	0.67	0.67	0.65
	attributes	['Pclass' 'SibSp' 'FareType' 'SexCode' 'Age_cat' 'Embarked_code']	['SibSp' 'Parch' 'FareType' 'SexCode' 'Age_cat']	['Pclass' 'SibSp' 'SexCode' 'Age_cat']
Balanced Acc	value	0.82	0.83	0.81
	attributes	['Pclass' 'SibSp' 'FareType' 'SexCode' 'Age_cat' 'Embarked_code']	['SibSp' 'Parch' 'FareType' 'SexCode' 'Age_cat']	['Pclass' 'SibSp' 'Parch' 'FareType' 'SexCode' 'Age_cat']
Sensitivity	value	0.72	0.96	0.7
	attributes	['Pclass' 'SibSp' 'FareType' 'SexCode' 'Age_cat' 'Embarked_code']	['Pclass' 'Parch']	['Pclass' 'SexCode' 'Age_cat']
Specificity	value	1	0.97	1
	attributes	['SibSp']	['FareType' 'Embarked_code']	['SibSp']

Different algorithms perform best with different selections of attributes

Bonus: Effect of threshold

Specificity and Sensitivity (function of threshold)

- We selected the attributes for which kNN had the best accuracy
- We trained a kNN model on these attributes (k=5)
- We saved the predicted probabilities
- And we varied the threshold that divides the 2 classes
- Sensitivity and Specificity obtained ->

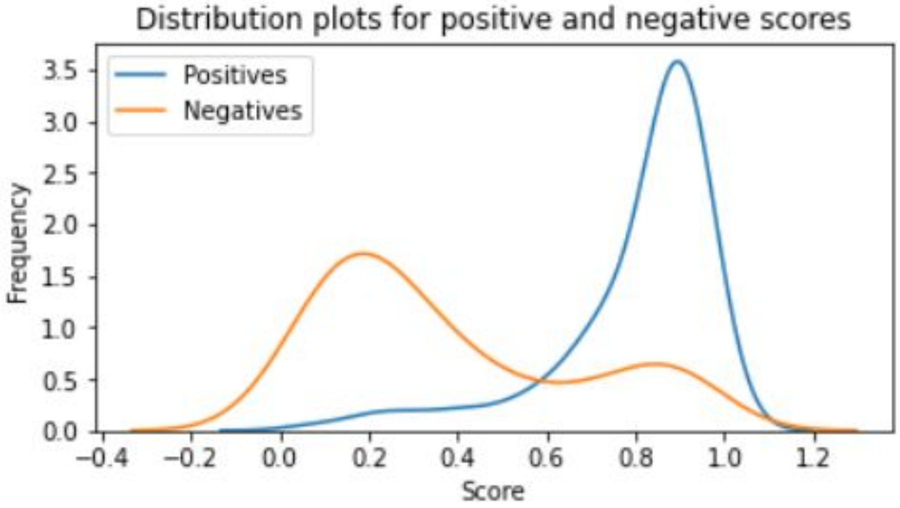


Sensitivity / True Positive (Recognition) rate $\alpha(s) = {}^{TP}/_{RP}$

Specificity / True Negative (Recognition) rate $\beta(s) = {}^{TN}/_{RN}$

Distribution plots for the two classes

- We used the same model from the previous slide
- And we plotted the frequency of scores for the two classes
- This would help us visualize the best threshold value


Distribution plots for positive and negative scores
Positives
Negatives
Frequency
Score