Анализ данных с использованием языка программирования R

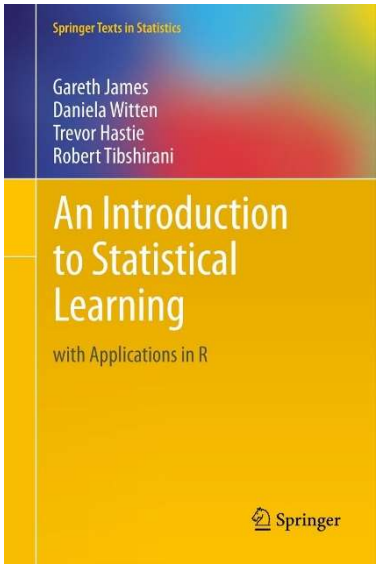
Тема 4
Подготовка данных для моделирования. Модели классификации данных

Минюкович Екатерина Александровна
к.э.н., доцент

miniukovich@bsu.by

Reference

An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, http://www-bcf.usc.edu/~gareth/ISL/ (available online for free)



Reference

Introduction to Machine Learning with R
by Dr. Dimitrios Gouliermis

http://www.mpia.de/homes/dgoulier/MLClasses/Course%20-%20Introduction%20to%20Machine%20Learning%20for%20Scientists%20with%20R.html

Reference

H2O documentation

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html

Supervised vs. Unsupervised Learning

Supervised	Unsupervised
Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation; 3) response Y measured on same n observations **Y** Continuous: Regression Discrete: Classification	Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation Clustering…

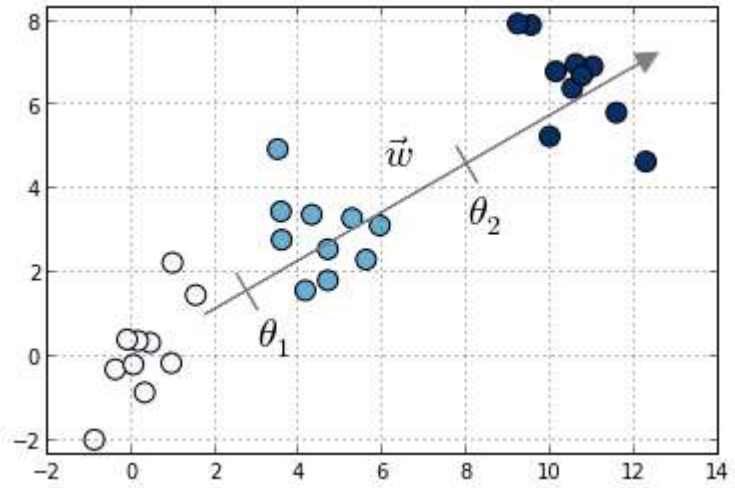
Classification

Binary	**Multiclass or multinomial**
2 classes	more than 2 classes

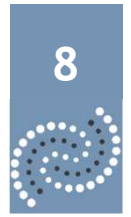
Ordinal classification (regression) or ranking learning

Ordinal classification (regression) or ranking learning
https://en.wikipedia.org/wiki/Ordinal_regression

H2O GLM (family = ordinal)



Regression / Classification / Clustering Problem



Steps to solve

- ***Working with data***
- ***Modeling***

Working with data

- Tidy data
- Types of variables and actions
- Missing data and imputation
- Feature engineering

Working with data Tidy Data

- Tidy data is a standard way of mapping the meaning of a dataset to its structure. This is Codd's 3rd normal form and the focus put on a single dataset rather than the many connected datasets common in relational databases.
- In tidy data:
  1. **Each variable forms a column.**
  2. **Each observation forms a row.**
  3. **Each type of observational unit forms a table.**

Which table below is tidy?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

More about tidy data:

ftp://cran.r-project.org/pub/R/web/packages/tidyr/vignettes/tidy-data.html

Working with data

Types of variables and actions

Types of variables	Actions
Categorical	Convert to factor *(automatically will be converted to n binary vars (n - number of labels) when building a model)*
Text	*Options:* • *Scrap a pattern and convert it to factor* • *Convert text to numbers (Word2Vec)* • *Drop text variable*
Numerical 	*Read if algorithm require standardization of numerical variables (often such algorithms do it by default).* *Standardization = mean removal + variance scaling*

Working with data

Missing data and imputation

- **Missing data:** NaN
- **Imputation**
  - Mean, median or mode
  - Prediction

Examples:
https://www.kaggle.com/kernels search on “Missing data imputation”

Working with data

Feature Engineering

- Based on variables meaning
- Technical approaches

Examples:

https://www.kaggle.com/kernels search on "Feature engineering"

Working with data

Example

- dataset: Titanic
  https://www.kaggle.com/c/titanic

- classification_titanic_part1.R
- classification_titanic_part2.R

Modeling

- Choose a class of model
- Fit the model to data
- Validate the model and optimize hyperparameters
- Predict for unknown data

Some models for Classification in h2o

- Generalized Linear Model (GLM)

  (*family is binomial or multinomial*)

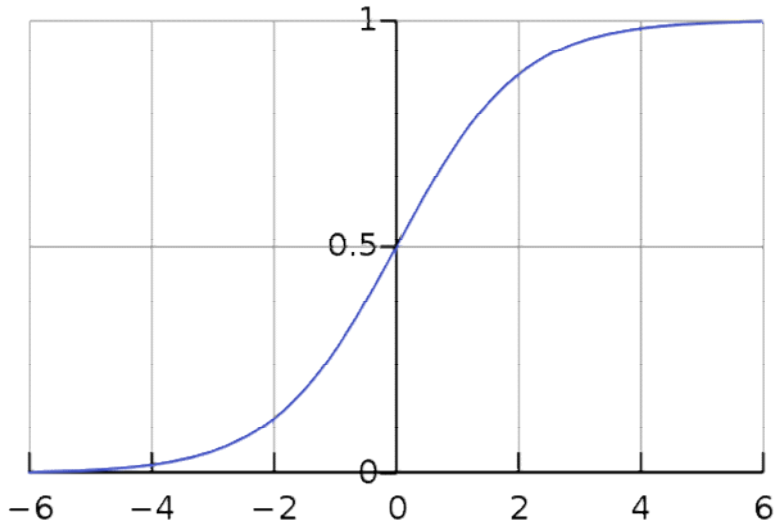
- Ensemble methods
  - Distributed Random Forest (DRF)
  - Gradient Boosting Machine (GBM)
  - Stacked Ensembles

Classification

Logistic Regression

To model p(X) = Pr(Y = 1|X) we need function that gives outputs between 0 and 1 for all values of X

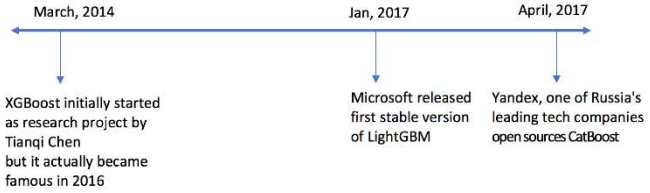
$$\hat{y} = p(X) = \frac{e^{\theta_0+\theta_1 x_1+\cdots+\theta_m x_m}}{1+e^{\theta_0+\theta_1 x_1+\cdots+\theta_m x_m}} = \frac{e^{X\theta}}{1+e^{X\theta}}$$



$$f(x) = \frac{e^x}{1+e^x}$$

Classification & Regression

Decision trees
Bagging or Bootstrap aggregation
Random Forest
Gradient Boosting (XGBoost, Light GBM, Catboost)



Useful links

GBM http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html (demo) https://habr.com/ru/company/ods/blog/327250/

XGBoost https://xgboost.readthedocs.io/en/latest/tutorials/model.html, https://mlexplained.com/2018/01/05/lightgbm-and-xgboost-explained/

LGBM https://lightgbm.readthedocs.io/en/latest/Features.html, https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

Catboost https://catboost.ai/docs/concepts/educational-materials-videos.html (videos by Yandex), https://www.youtube.com/watch?v=V5158Oug4W8&list=PLVlY_7IJCMJeRfZ68eVfEcu-UcN9BbwiX&index=20&t=1290s (video from mlcourse.ai, Catboost starts from 9th minute), https://github.com/catboost/tutorials/blob/master/python_tutorial.ipynb (Catboost tutorial on Titanic)

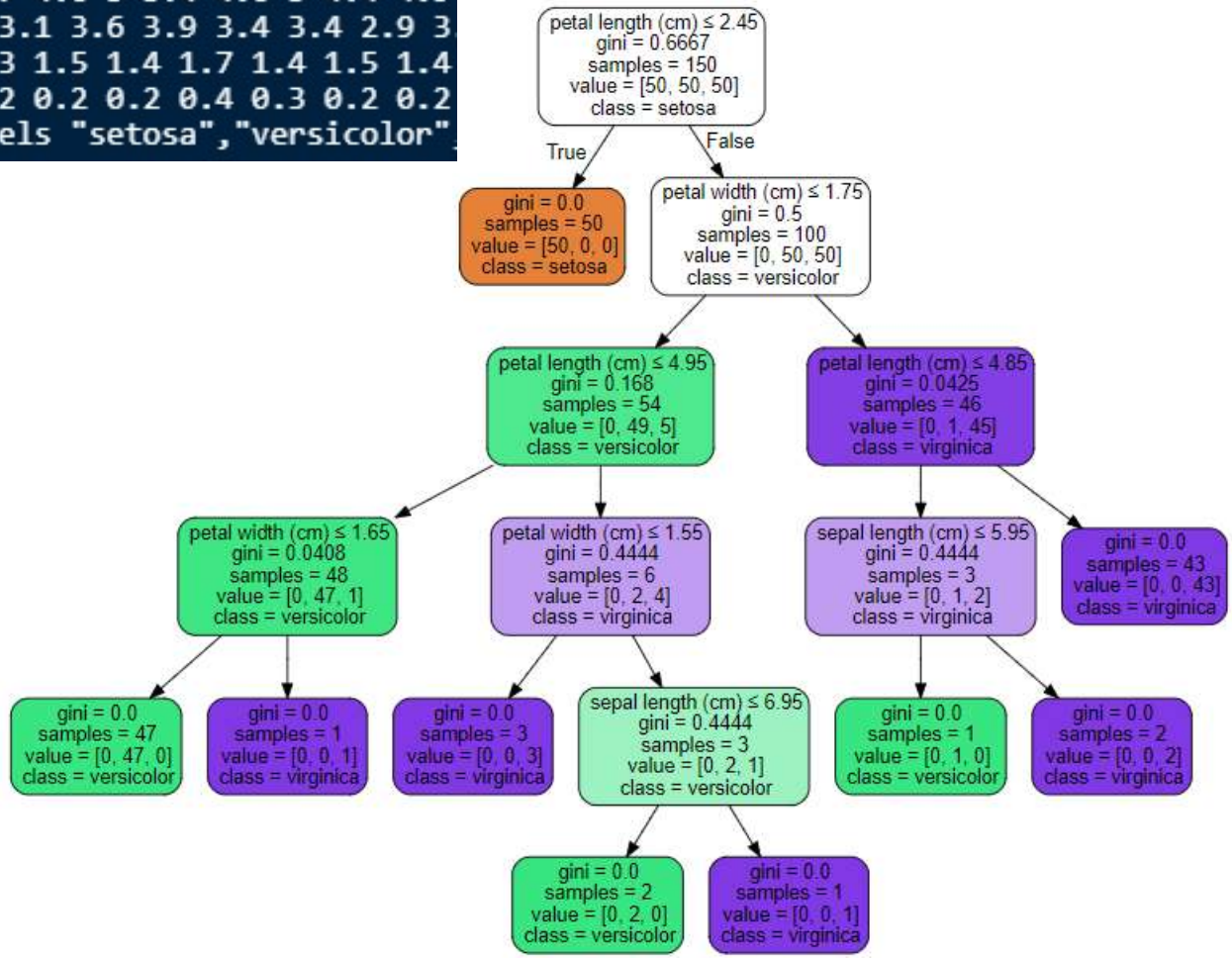
Classification tree

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2
 $ Species     : Factor w/ 3 levels "setosa","versicolor"
```

Iris Setosa

Iris Versicolor

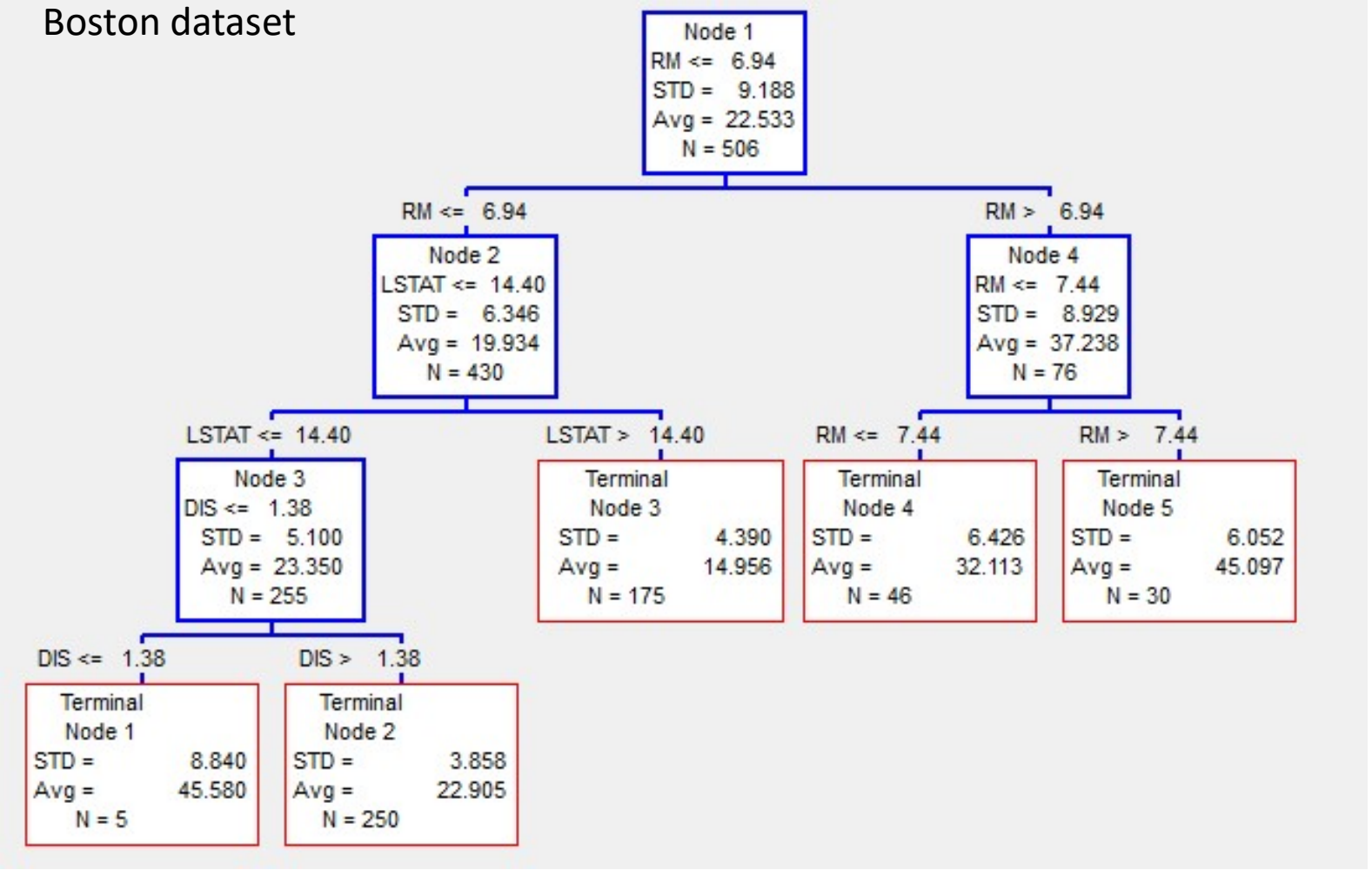
Iris Virginica



https://habr.com/ru/company/ods/blog/322534/

Regression tree

Boston dataset



https://habr.com/ru/company/ods/blog/322534/

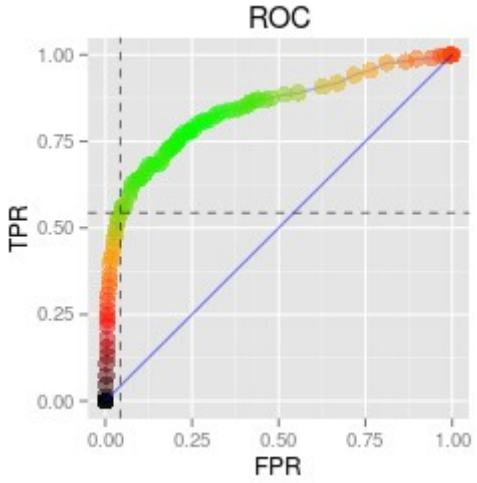
Classification metrics

Confusion matrix

Survived (S) -1; Not Survived (NS) - 0

Actual/Predicted	*0*	*1*	*Error*
0 (N)	**TN** (NS as NS)	**FP** (NS as S)	**FPR=FP/N** (False Positive Rate)
1 (P)	**FN** (S as NS)	**TP** (S as S)	**FNR=FN/P**(False Negative Rate)

Receiver operating characteristic curve



Accuracy = (TP+TN)/(P+N)

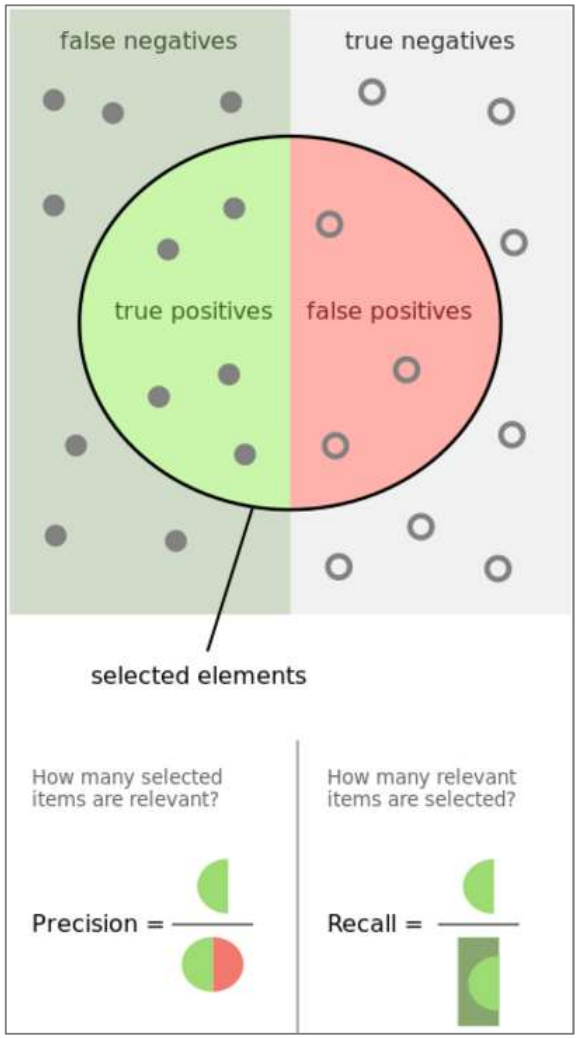
Precision = TP/(TP+FP) **Recall** = **TPR** = TP/P

F1 = 2*Precision*Recall/(Precision+Recall) - *harmonic mean Precision and Recall*

AUC - Area Under ROC Curve (**the closer to 1, the better a model is**)

More: https://en.wikipedia.org/wiki/Precision_and_recall

Classification metrics



Confusion matrix

Survived (S) -1; Not Survived (NS) - 0

Actual/Predicted	***0***	***1***	***Error***
0 (N=438)	TN=365	FP=?	FPR=FP/N = ?
1 (P=274)	FN=?	TP=212	FNR=FN/P = ?
Total			(FN+FP)/(N+P) = ?

FN - ошибка первого рода; FP - ошибка второго рода

Accuracy = (TP+TN)/(P+N) - ?

Precision = TP/(TP+FP) -?

Recall = **TPR** = TP/P-?

http://scikit-learn.org/stable/modules/classes.html#classification-metrics

Multiclass Classification

Some classification algorithms naturally permit the use of more than two classes

- **GLM**
- **Random Forest, Gradient Boosting**

example in mclass.R

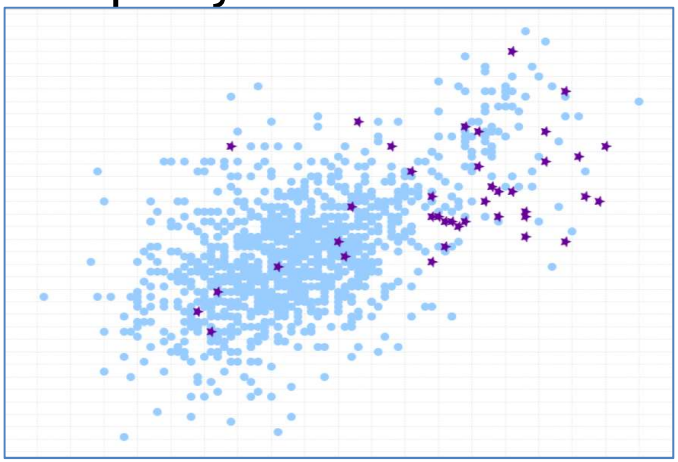
Techniques of transformation to binary

- One vs. All
- One vs. One

Read more:
https://en.wikipedia.org/wiki/Multiclass_classification
http://scikit-learn.org/stable/modules/multiclass.html

Classification: Unbalanced classes

Unbalanced classes - classes are not represented equally



Accuracy Paradox

Tactics to Combat Unbalanced Classes

1) Collect more data

2) Resample Your Dataset

3) Generate Synthetic Samples

Imbalanced-learn

https://imbalanced-learn.readthedocs.io/en/stable/user_guide.html

4) Change Your Performance Metric

(e.g. Absolute MCC (Matthews Correlation Coefficient), AUCPR (Area Under the Precision-Recall Curve in h2o)

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/performance-and-prediction.html#metric-best-practices-classification*)*

5) Use special hyperparameters

(*e.g. balance classes in h2o*)

Read more: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Modeling

Hyperparameters optimization

- Parameters to optimize
- Good range of values

More about parameters to optimize and good range of values
https://www.linkedin.com/pulse/approaching-almost-any-machine-learning-problem-abhishek-thakur?trk=hp-feed-article-title-like

Practise

classification_titanic_part1.R
classification_titanic_part2.R
mclass.R

Managed Independent Work

pr_classification.R