Plot it to Understand it Better:
Creating Visualizations in R to Support Students in Interpreting Results of Machine Learning Algorithms

Overview

Objective: Making interpretation of machine learning algorithm results easier for university students

- Students cannot get familiar with the methods before the lecture is over ✘
- Parametrization is far from trivial for the first time ✘
- Results are hard to interpret without prior experience ✘

Suggested approach: tailored visualizations in R for classification methods

Easy-to-interpret results, more enthusiastic students ☺

Methodology

Step 1: Characteristics of individual classes

- **Tableplots** hide the details and are efficient at emphasizing the common characteristics of record groups
- Variables are ordered based on their importance in the model
- Records are ordered by the target variable

Step 2: Detailed view of misclassified cases

- **Parallel coordinate plots** are efficient in highlighting groups if they exist
- Variables are ordered based on their importance in the model
- One plot for each misclassified subset
- Overlapping is solved with line width

Step 3: In-depth analysis of the model itself

- Besides the general plots, special plots are created for Random Forest (RF) and Naïve Bayes (NB) models
- Random forest results are represented with **heatmaps**
- Naïve Bayes are represented with **conditional density plots and mosaic plots**

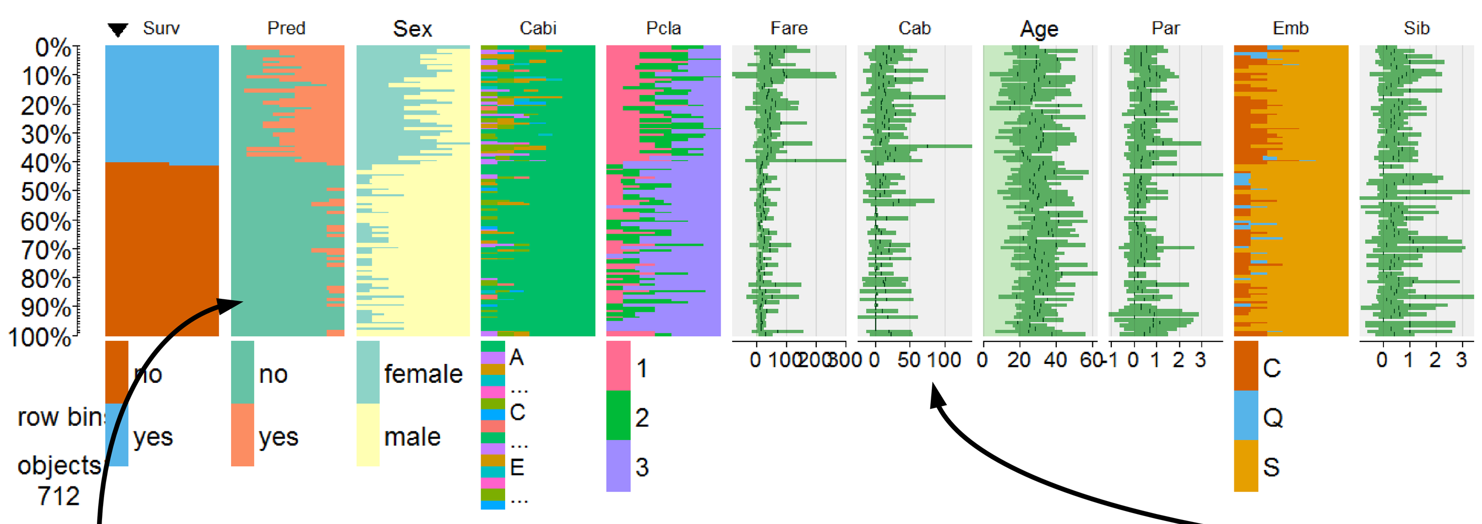
Use Case Data Set: Survival on the Titanic

The (cleaned) Titanic data set contains nine features of individuals (passengers and crew) who were on board at the tragic voyage. The classification exercise is predicting of their survival.

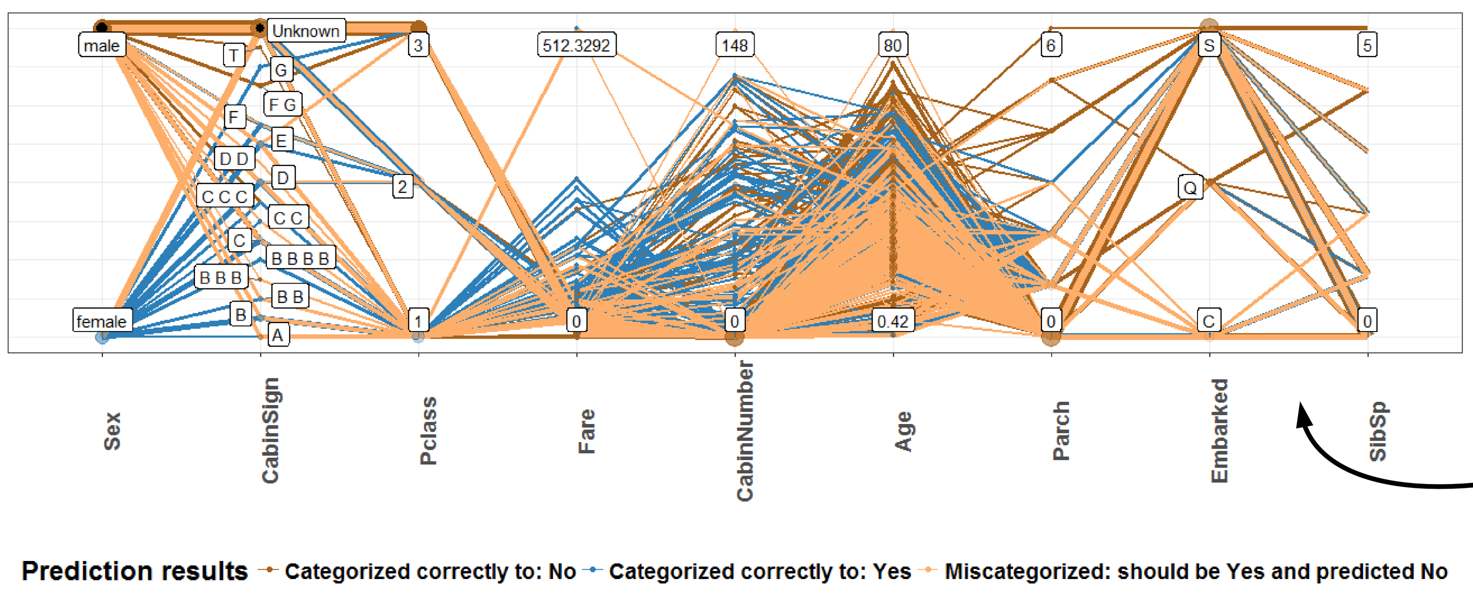
Features:

- **Survived –** Target variable indicating whether the passenger survived the tragedy
- **Pclass** – Indicator whether the passenger travelled on first, second or third class
- **CabinNumber** and **CabinSign –** cabin information of the passengers, highly sparse
- **Embarked** – Port of embarkation (Cherbourg, Queenstown, Southampton)
- **Fare –** Passenger fare
- **Parch –** Number of parents and children on-board
- **Age –** Age of the passenger in years
- **SibSp –** Number of siblings and spouses on-board

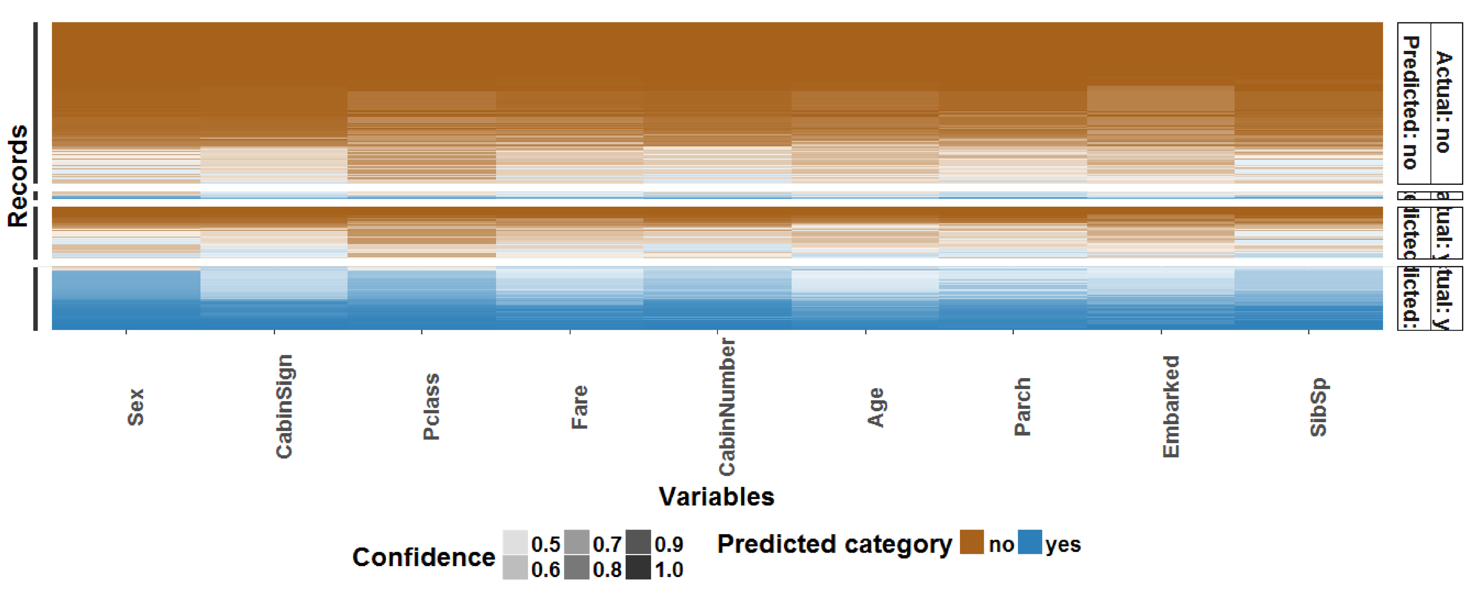
Random Forest



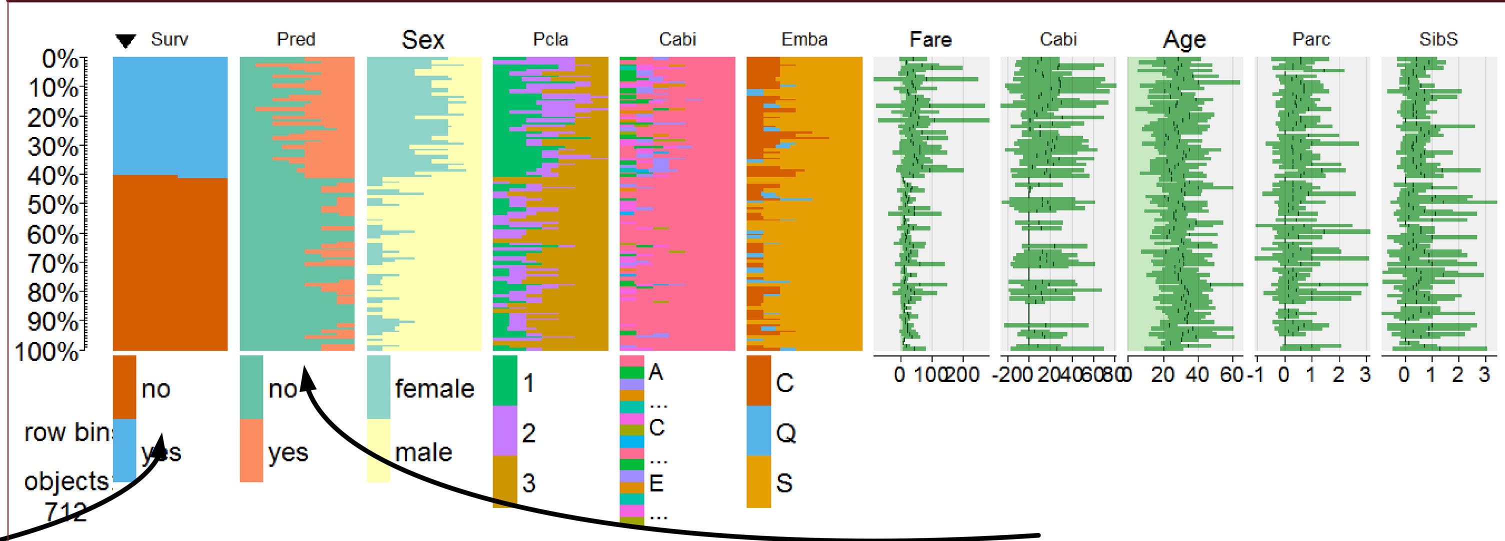
Misclassified cases are almost always false negatives



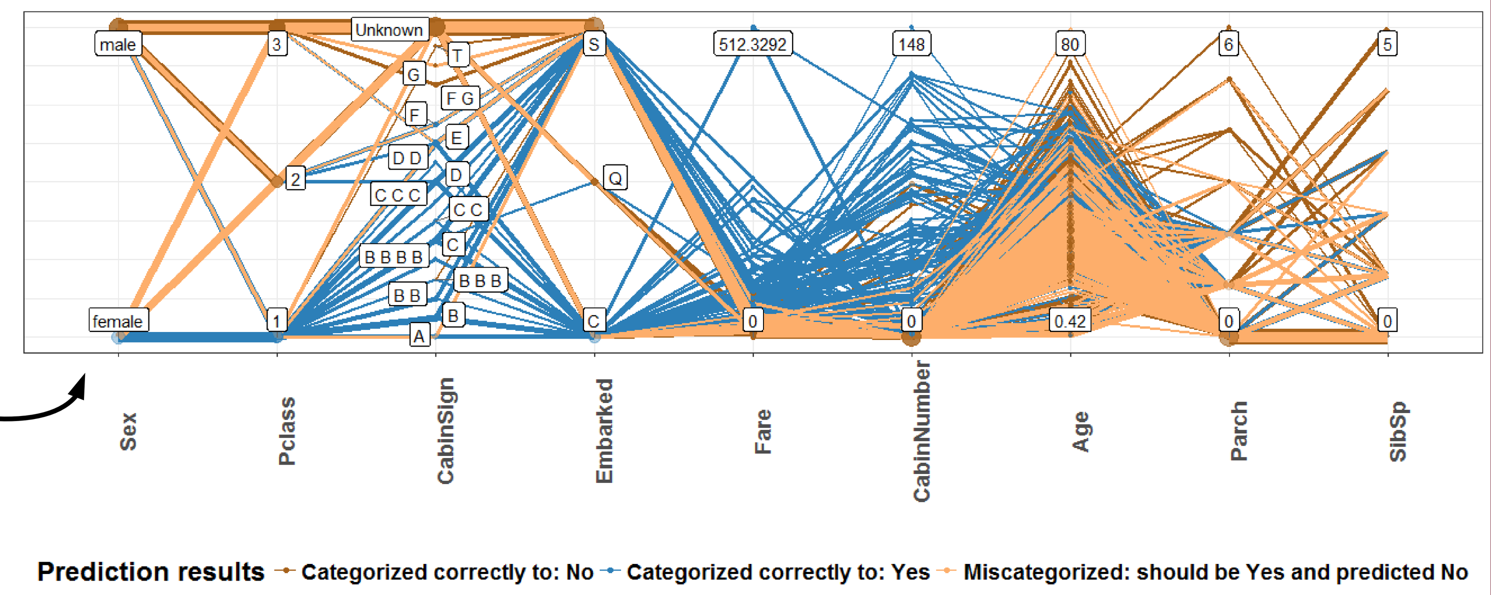
Causal relationship with cabin details?



Naïve Bayes



Misclassified cases are much more balanced here than they in case of the RF



The NB model seems more conservative than the RF: the independency condition is obviously violated

Survivors are typically:

- women or first-class passengers,
- with higher fare,
- travelling alone or with smaller company
- closer to 30 than 40 in age

Most of misclassified cases are men with a low fare. Number of relatives on-board are irrelevant.

RF: cells are colored based on the predictive confidence of the trees the variable is part of.
NB: conditional density plots help to understand standard deviations as well, not only the means

Conclusions

- The two general plots can help even in the exploratory phase
- Painfully missing interactivity, e.g. linked highlighting
- Detailed view is only informative until ~10,000 rows

References

The poster idea is originated from the thesis work of Adam Bereczki: *"Visualization of Machine Learning Algorithms"* (Advisors: Ágnes Salánki & Gábor Szárnyas)

Cool related visualization projects: [1] Welling, Soeren H., et al. "*Forest Floor Visualizations of Random Forests.*" arXiv preprint arXiv:1605.09196 (2016). & [2] ML Demos: http://mldemos.b4silio.com/download.html

Contact Information

For further information, contact Ágnes Salánki
Email: salanki.agnes@gmail.com

BRUSSELS useR! 2017

R Ladies Budapest