Business Analytics

Introduction to the DMC

Decision Sciences & Systems (DSS)

Department of Informatics

TU München

Tutorial Business Analytics

Outline

Today's topics:

- Dates & Grading for Data Mining Cup
- Rules of Data Mining Cup
- Steps of Data Mining Cup
- Example dataset + Script
- Presentation of dataset

Tutorial Business Analytics

Dates & Grading for Data Mining Cup

Date

- 26.05. 11:30 am – 15.06. 11:55 pm

Grading for the DMC

- Best 25%: +8 points
- Next 25%: +6 points
- Next 25%: +4 points
- Minimum 2 points if you perform better than 0-R

Note: Only **“serious”** submissions are taken into account for the ranking.

Tutorial Business Analytics

Rules of Data Mining Cup

Teams

- Team size: 1 – 4 members.
- **Teams must be built before the first submission (teams will be fixed after first submission!).**
- Each student can only be member of one team within one Data Mining Cup.

Submissions

- Maximum number of valid submissions for each DMC: 10.
- Best ranked submission, **only**, will be taken into account for the ranking.
- For reasons of traceability you must use a fixed seed of 42 (`set.seed(42)`).

Disqualification reasons:

- **Non-reproducible** submissions (submitted predictions **must be reproducible** using the submitted R script)
- **Hard-coded** classifications (even if the best ranked submission is not hard-coded!)
- **Copies** from other groups (disqualification of both teams)

Tutorial Business Analytics

Steps of Data Mining Cup

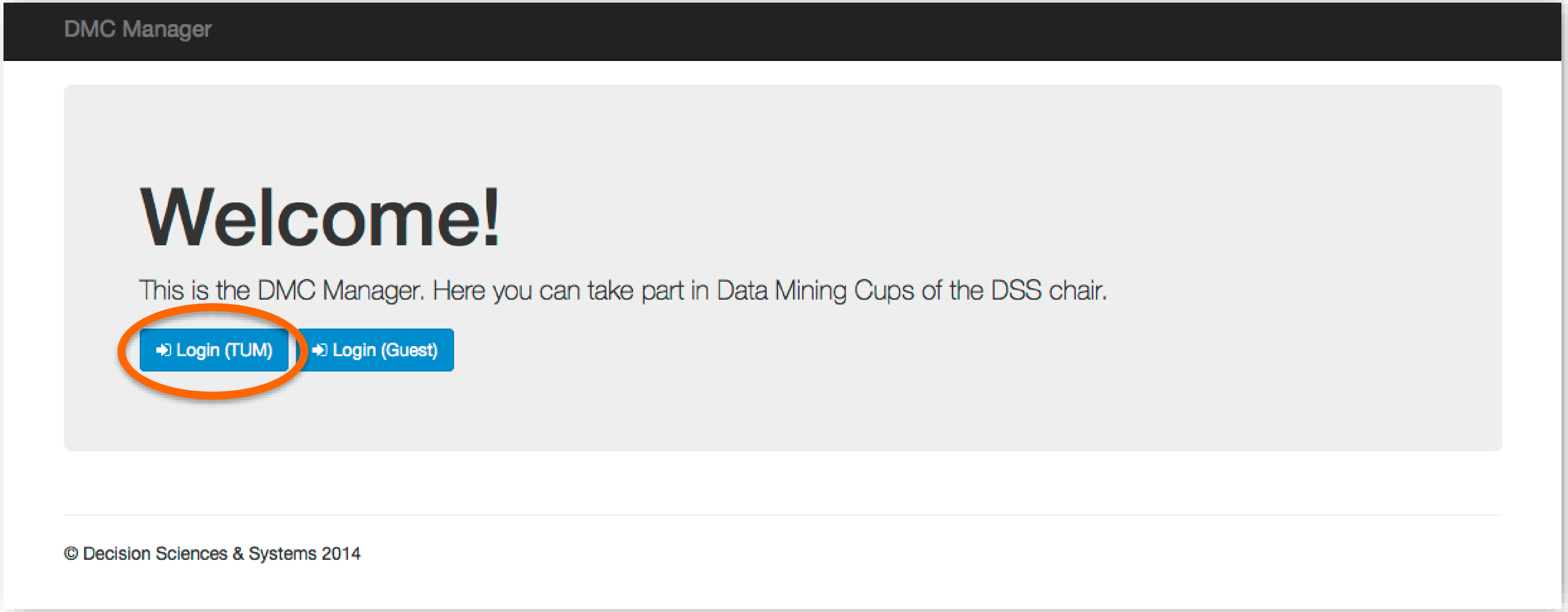
1. Build a Team in the DMC Manager
2. Load & Explore the Data Set
   - Summary statistics
   - Plotting
3. Data Preparation
   - Feature Selection
   - Discretization
4. Training & Evaluation
   - Classification Methods
   - Metrics
   - Resampling Methods
5. Predict Classes in Test Data
6. Export the Predictions
7. Upload the Predictions and the Corresponding R Script on DMC Manager

Source: http://topepo.github.io/caret/

1. Build Team in DMC Manager

Login with your TUM login data ("TUM Kennung")

https://dmc.dss.in.tum.de/dmc/



DMC Manager

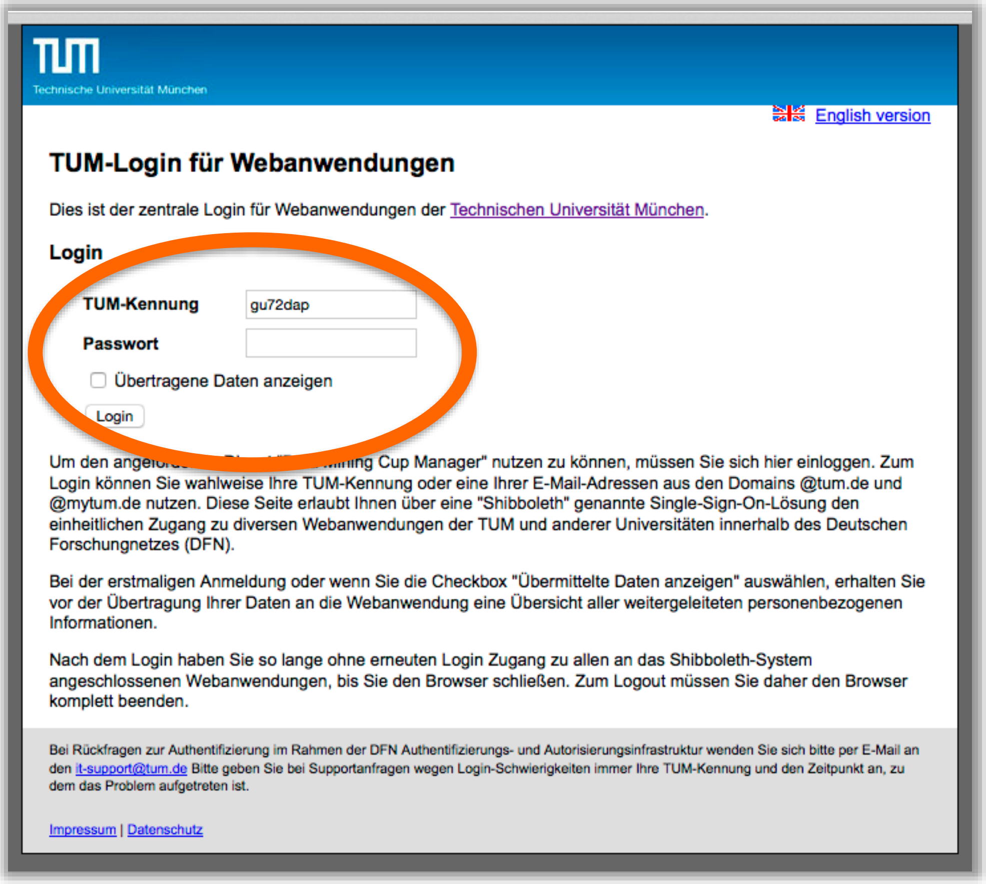
Welcome!

This is the DMC Manager. Here you can take part in Data Mining Cups of the DSS chair.

© Decision Sciences & Systems 2014

1. Build Team in DMC Manager

Login via “Shibboleth” with your TUM login data (“TUM Kennung”)



TUM-Login für Webanwendungen

Dies ist der zentrale Login für Webanwendungen der Technischen Universität München.

Login

TUM-Kennung gu72dap

Passwort

□ Übertragene Daten anzeigen

Login

Um den angeforderten Dienst "Data Mining Cup Manager" nutzen zu können, müssen Sie sich hier einloggen. Zum Login können Sie wahlweise Ihre TUM-Kennung oder eine Ihrer E-Mail-Adressen aus den Domains @tum.de und @mytum.de nutzen. Diese Seite erlaubt Ihnen über eine "Shibboleth" genannte Single-Sign-On-Lösung den einheitlichen Zugang zu diversen Webanwendungen der TUM und anderer Universitäten innerhalb des Deutschen Forschungnetzes (DFN).

Bei der erstmaligen Anmeldung oder wenn Sie die Checkbox "Übermittelte Daten anzeigen" auswählen, erhalten Sie vor der Übertragung Ihrer Daten an die Webanwendung eine Übersicht aller weitergeleiteten personenbezogenen Informationen.

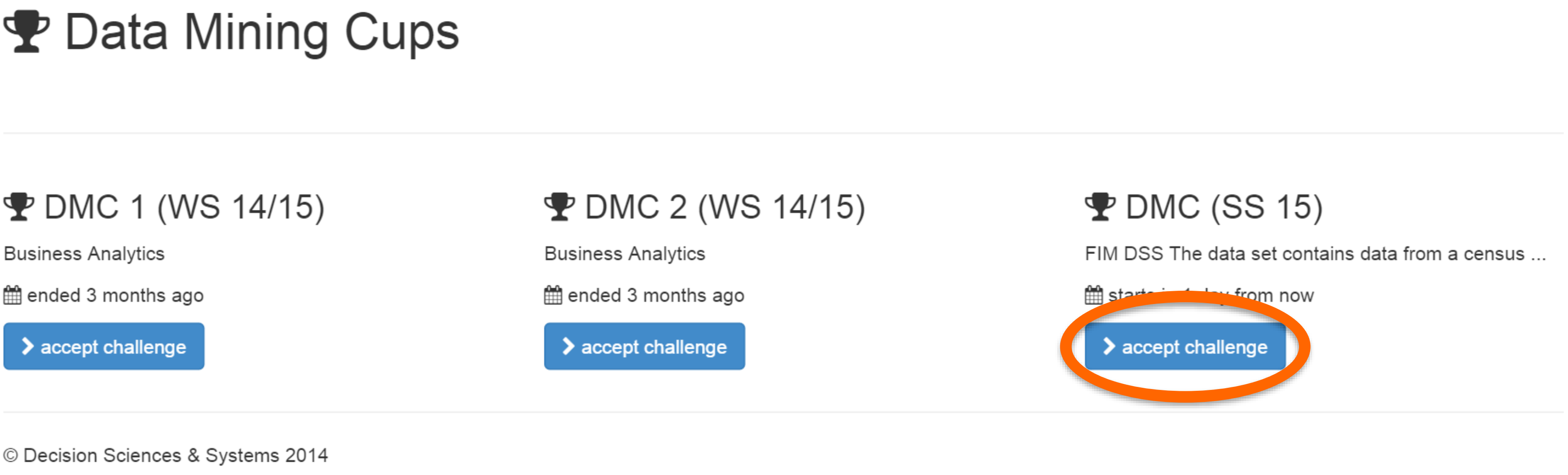
Nach dem Login haben Sie so lange ohne erneuten Login Zugang zu allen an das Shibboleth-System angeschlossenen Webanwendungen, bis Sie den Browser schließen. Zum Logout müssen Sie daher den Browser komplett beenden.

Bei Rückfragen zur Authentifizierung im Rahmen der DFN Authentifizierungs- und Autorisierungsinfrastruktur wenden Sie sich bitte per E-Mail an den it-support@tum.de Bitte geben Sie bei Supportanfragen wegen Login-Schwierigkeiten immer Ihre TUM-Kennung und den Zeitpunkt an, zu dem das Problem aufgetreten ist.

Impressum | Datenschutz

1. Build Team in DMC Manager

Choose the DMC instance in the DMC Manager



🏆 Data Mining Cups

🏆 DMC 1 (WS 14/15)

Business Analytics

📅 ended 3 months ago

🏆 DMC 2 (WS 14/15)

Business Analytics

📅 ended 3 months ago

🏆 DMC (SS 15)

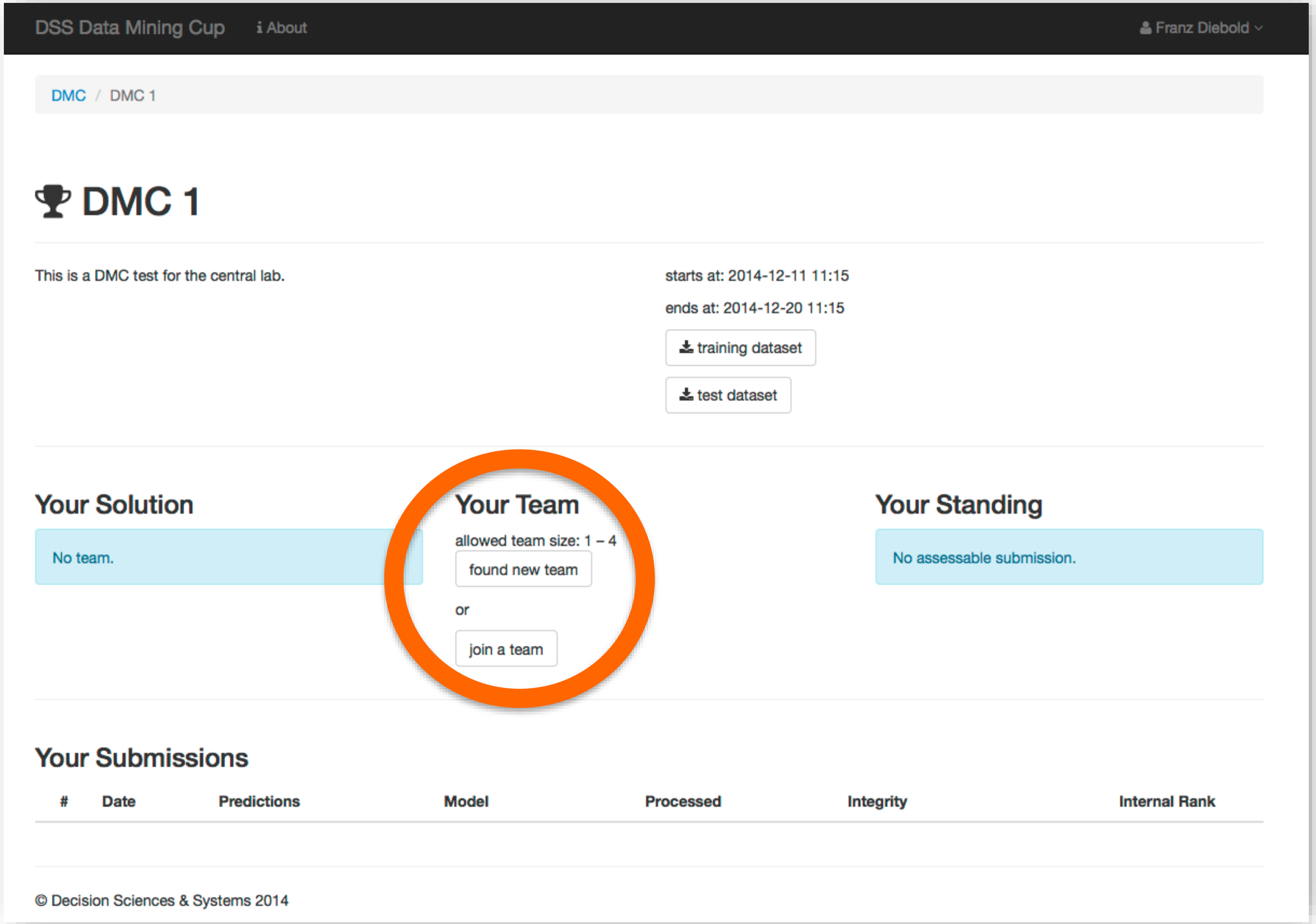
FIM DSS The data set contains data from a census ...

📅 starts in 1 day from now

© Decision Sciences & Systems 2014

1. Build Team in DMC Manager

Found new team or join an existing team



DSS Data Mining Cup   ℹ About   Franz Diebold

DMC / DMC 1

DMC 1

This is a DMC test for the central lab.

starts at: 2014-12-11 11:15

ends at: 2014-12-20 11:15

training dataset

test dataset

Your Solution

No team.

Your Team

allowed team size: 1 – 4

found new team

or

join a team

Your Standing

No assessable submission.

Your Submissions

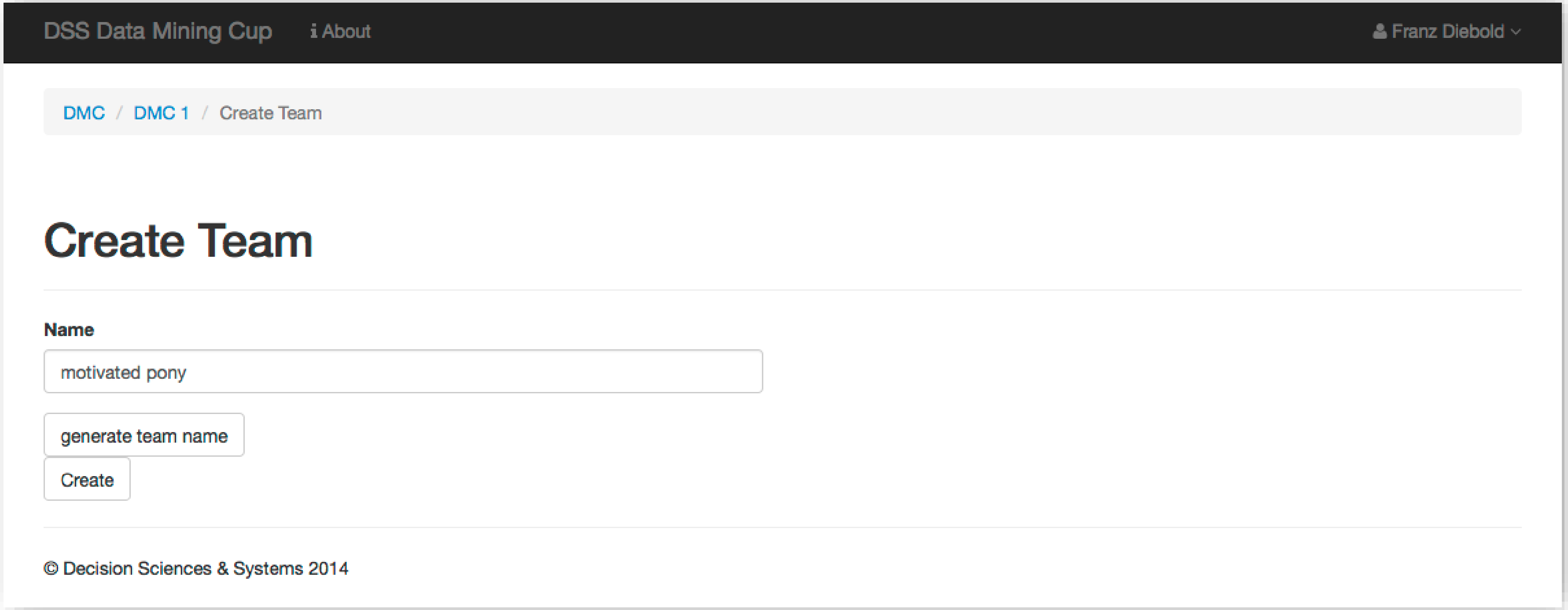
#	Date	Predictions	Model	Processed	Integrity	Internal Rank

© Decision Sciences & Systems 2014

1. Build Team in DMC Manager

Creating a new team

- Team size: 1-4 members



Create Team

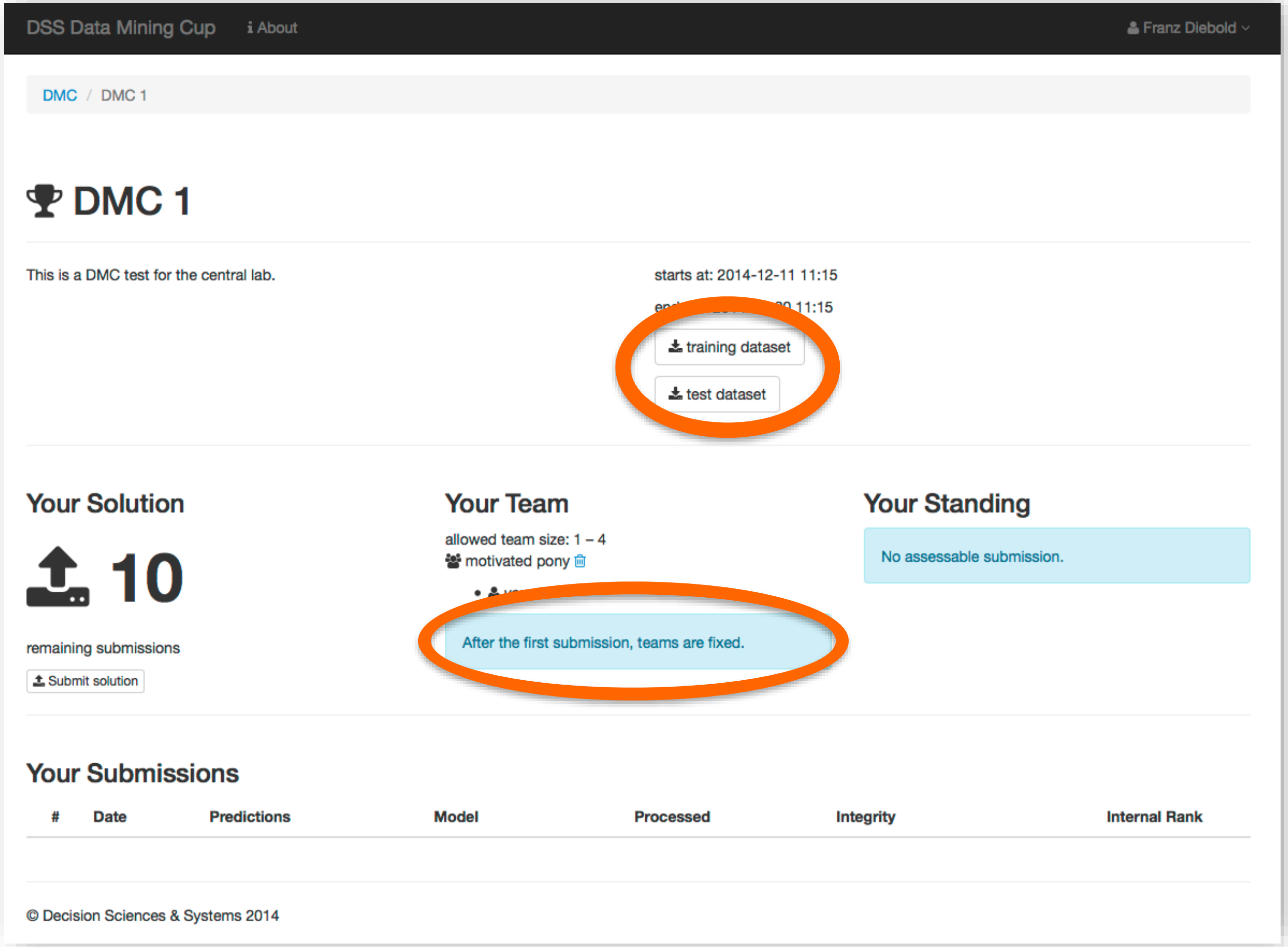
Name

motivated pony

© Decision Sciences & Systems 2014

2. Load & Explore the Data Set

Download the training and test datasets from the DMC Manager



DSS Data Mining Cup ℹ About 👤 Franz Diebold

DMC / DMC 1

🏆 DMC 1

This is a DMC test for the central lab.

starts at: 2014-12-11 11:15

en 20 11:15

Your Solution

10

remaining submissions

Your Team

allowed team size: 1 – 4

motivated pony

After the first submission, teams are fixed.

Your Standing

No assessable submission.

Your Submissions

#	Date	Predictions	Model	Processed	Integrity	Internal Rank

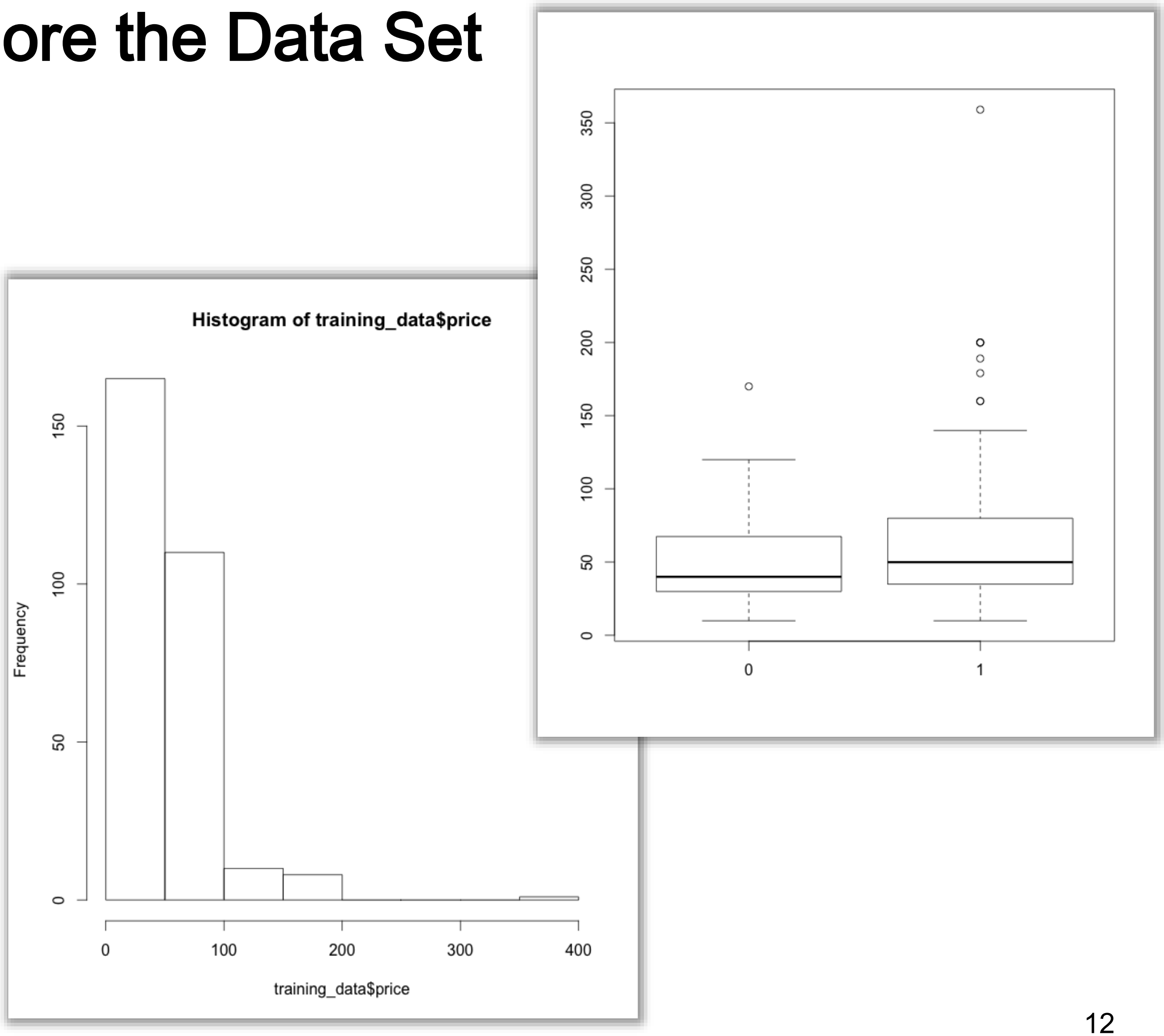
© Decision Sciences & Systems 2014

2. Load & Explore the Data Set

Load & Explore in R

- Load data sets into R

- Explore the Data Set
  - Get an overview
  - Statistics
  - Plotting



3. Data Preparation

- Possible Data Preparation steps
  - Nominal attributes
  - Ordinal attributes
  - Unified date format
  - Missing values
  - Fix errors and outliers
  - Zero variance and correlation
  - Discretization/Binning
  - Feature Selection

- ALL changes in both training & test dataset!
- Do NOT DELETE any instances in the test data!

4. Training & Evaluation

Classification Methods

Name	***method* Argument in *train* Function**	**Tuning Parameters**
OneRule	OneR	-
Naïve Bayes	nb	fL, usekernel
Decision Trees	J48	C (pruning factor), M
k-Nearest Neighbors	kknn	kmax, distance, kernel
Ensemble Methods	ada, LogitBoost, logicBag	iter; maxdepth; nu, nIter, nleaves, ntrees

```
> model = train(Class~., data=training, method=”J48")
```

. = all attributes

More classifiers: http://topepo.github.io/caret/modelList.html

Source: http://topepo.github.io/caret/

4. Training & Evaluation

Classification Methods – Tuning Parameters

- `tuneLength`: number of tuning parameter values
- `tuneGrid`: for specific tuning parameter values
  - data frame, where each row is a tuning parameter setting and each column is a tuning parameter

```
> model = train(Class~., data=training, method=”J48”,
                 tuneGrid=data.frame(C=c(0.1, 0.2, 0.3),M=c(2,2,2))
```

nothing given -> algorithm chooses them on its own
here: every combination is build -> 9 DTs and go with majority

Where to find parameters?

http://topepo.github.io/caret/train-models-by-tag.html

Or in R:

```
> getModelInfo()$J48$parameters
```

4. Training & Evaluation

Metrics

Name	***metric* in *train* Function**	**Description**
Accuracy	Accuracy	*=(tp + tn) / (tp + fp + tn + fn)*
Kappa	Kappa	see below
ROC Curve	ROC	area under the ROC curve

relevant for submission

```
> model = train(Class~., data=training, method=”J48”,
        metric=“Kappa”)
```

Kappa

- Ratio, which compares a classification method with a random classifier
  - < 0: worse than random classifier
  - > 0: better than random classifier

Source: http://topepo.github.io/caret/

4. Training & Evaluation

Resampling Methods

Name	*method* Argument in *trainControl* Function
Bootstrapping (Holdout method, default)	boot
Repeated K-fold Cross Validation	repeatedcv 10fold is used often
Leave-one-out	LOOCV

```
> # 2 x repeated 3-fold cross validation
> fitCtrl = trainControl(method=”repeatedcv”, number=3, repeats=2)

> model = train(Class~., data=training, method=”J48”,
                 trControl=fitCtrl)
```

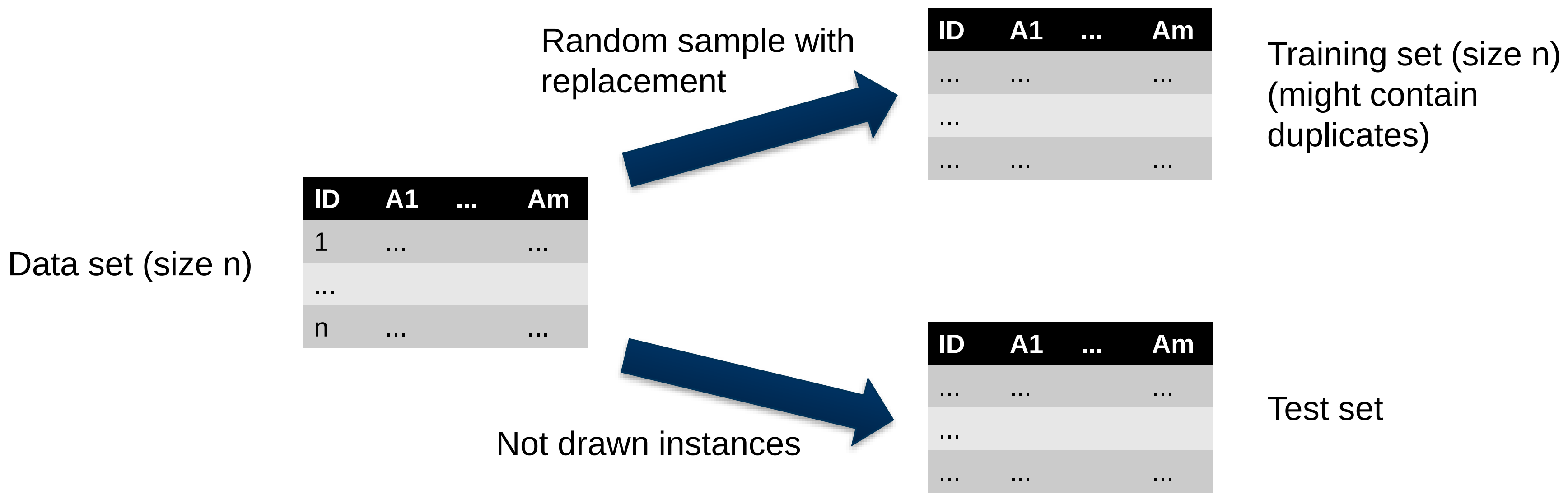
Source: http://topepo.github.io/caret/

4. Training & Evaluation

Bootstrapping

Bootstrapping

- Resampling method



Data set (size n)

ID	A1	...	Am
1	...		...
...			
n	...		...

Random sample with replacement

ID	A1	...	Am
...	...		...
...			
...	...		...

Training set (size n) (might contain duplicates)

Not drawn instances

ID	A1	...	Am
...	...		...
...			
...	...		...

Test set

4. Training & Evaluation

Balanced Samples using the “ROSE” package

not too relevant for this cup
as our measure is accuracy
relevant, when few instances are true

- „ROSE“ package: http://cran.r-project.org/web/packages/ROSE/index.html
- Balanced samples by over-/under-sampling the minority/majority instances

```
> library(ROSE)
> training_data = ovun.sample(class ~ ., data=training_data,
            method="over", N=10000, na.action="na.pass")$data
```

method	Description
over	over-sampling of minority instances
under	under-sampling of majority instances
both	combination of over- and under-sampling

create new instances that are true

randomly remove
negative instances

4. Training & Evaluation

Comparing the models

- Can compare several trained models
- The models should be using the same resampling

```
> res = resamples(list(dt = model_dt, nb = model_nb))
> summary(res)

…
Accuracy
     Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
dt 0.4457  0.4810 0.4946 0.4910  0.5041 0.5275    0
nb 0.5000  0.5163 0.5246 0.5192  0.5275 0.5275    0
```

even worst case solution of NB was better than DT

5. Predict Classes in Test Data

- Use the trained model to predict the classes in the test dataset.

```
> prediction_classes = predict.train(object=model,
       newdata=test_data, na.action=na.pass)
> predictions = data.frame(id=test_data$id,
       prediction=prediction_classes)
```

ignore not availables

6. Export the Predictions

- Export predictions into csv-file
  - Format: id, prediction
  - Must contain all instances of the original test dataset

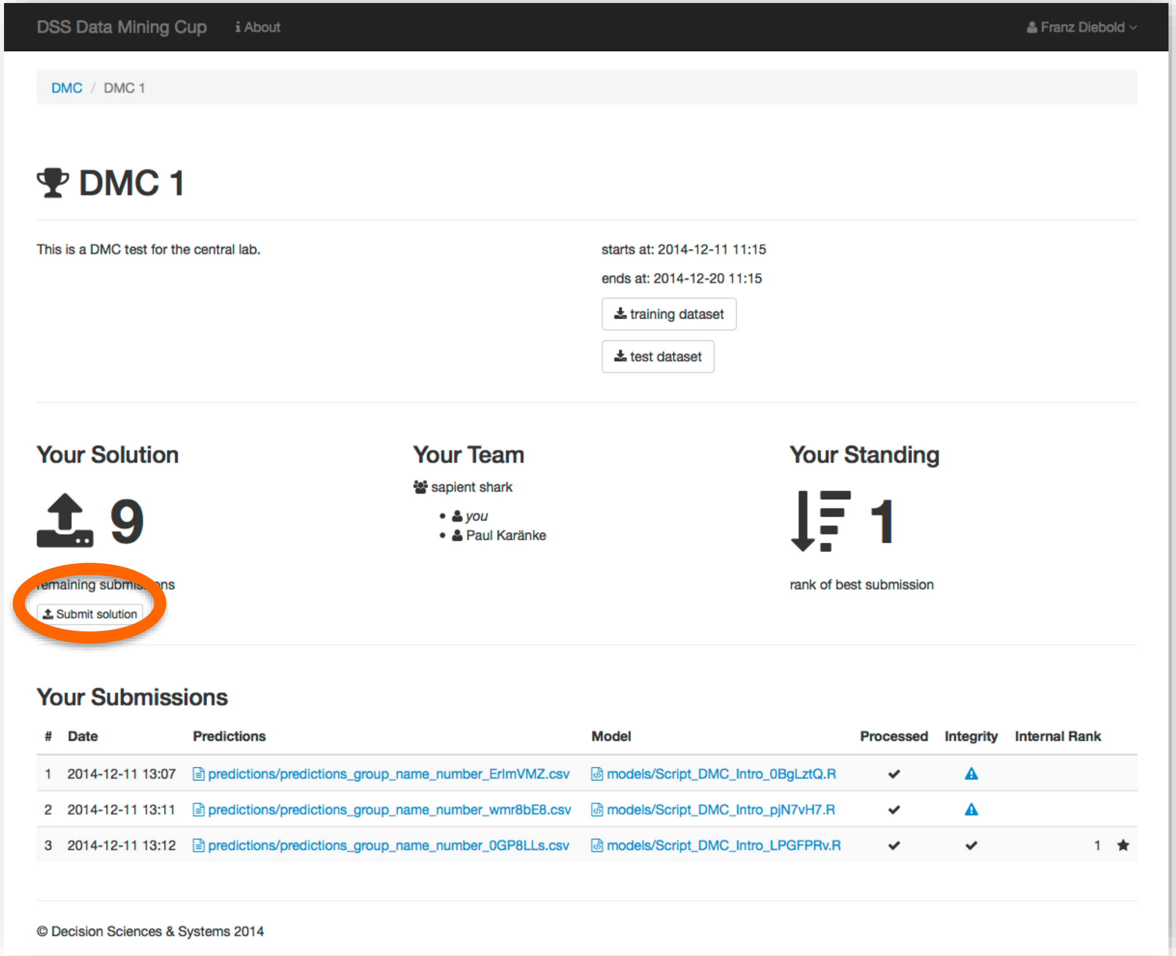
```
> write.csv(predictions, file="predictions_group_name_number.csv",
        row.names=FALSE)
```

predictions_group_name_number.csv

```
"id","prediction"
130200,"1"
394720,"0"
87847,"1"
228637,"1"
189299,"0”
262991,"1”
…
```

check this first

7. Upload the Predictions and the Corresponding R Script on DMC Manager



DMC / DMC 1

DMC 1

This is a DMC test for the central lab.

starts at: 2014-12-11 11:15

ends at: 2014-12-20 11:15

Your Solution

9

remaining submissions

Your Team

sapient shark

- *you*
- Paul Karänke

Your Standing

1

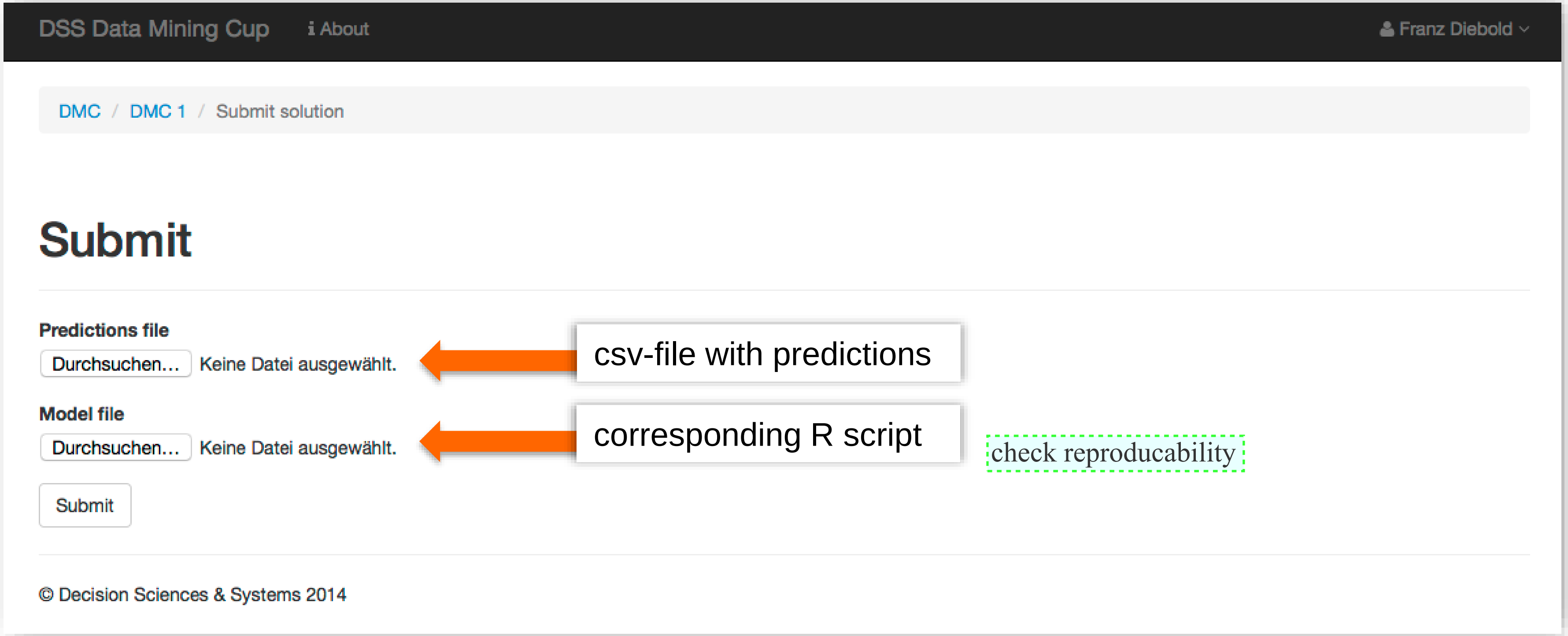
rank of best submission

Your Submissions

#	Date	Predictions	Model	Processed	Integrity	Internal Rank
1	2014-12-11 13:07	predictions/predictions_group_name_number_ErlmVMZ.csv	models/Script_DMC_Intro_0BgLztQ.R	✓	⚠	
2	2014-12-11 13:11	predictions/predictions_group_name_number_wmr8bE8.csv	models/Script_DMC_Intro_pjN7vH7.R	✓	⚠	
3	2014-12-11 13:12	predictions/predictions_group_name_number_0GP8LLs.csv	models/Script_DMC_Intro_LPGFPRv.R	✓	✓	1 ★

© Decision Sciences & Systems 2014

7. Upload the Predictions and the Corresponding R Script on DMC Manager



Submit

Predictions file

Keine Datei ausgewählt.

csv-file with predictions

Model file

Keine Datei ausgewählt.

corresponding R script

check reproducability

© Decision Sciences & Systems 2014

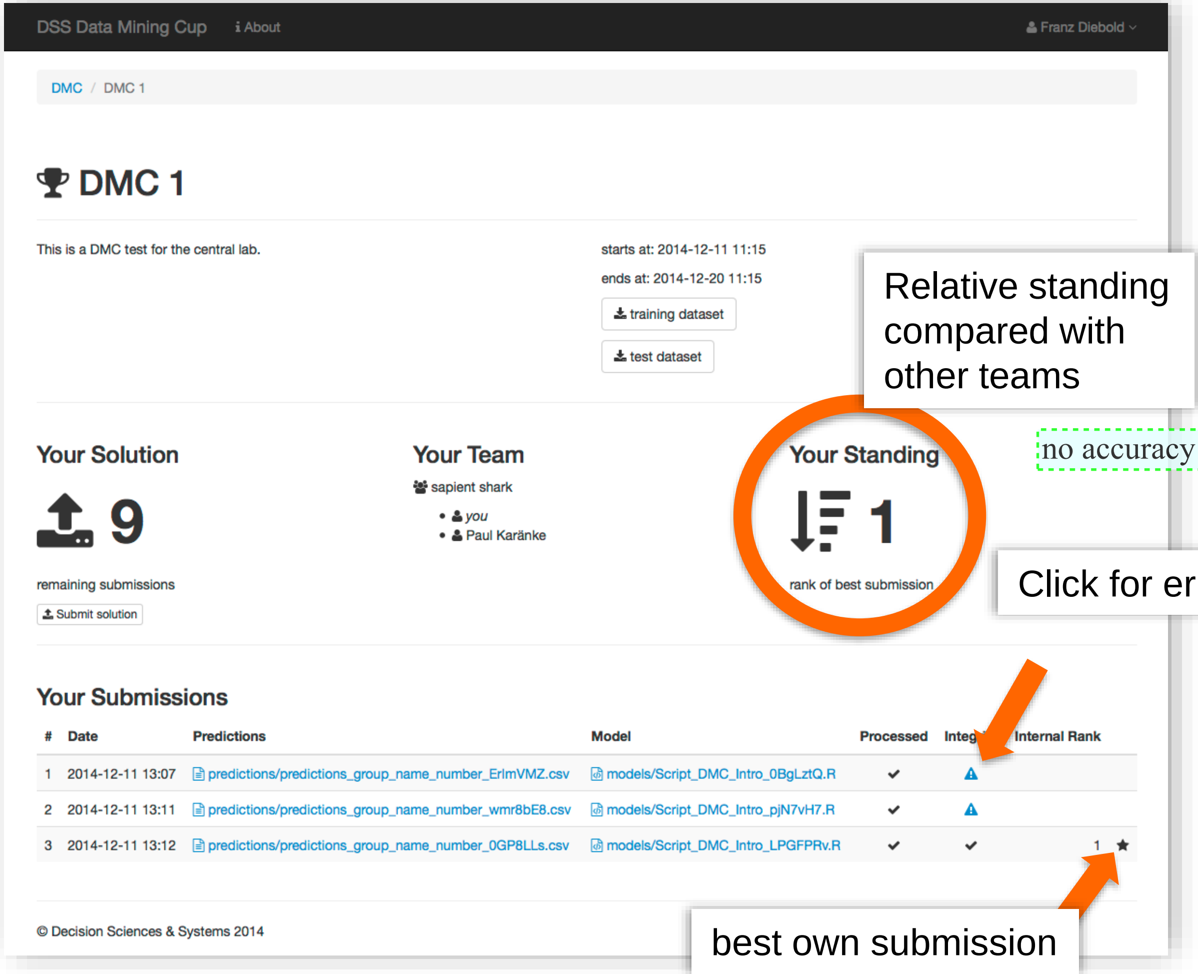
7. Upload the Predictions and the Corresponding R Script on DMC Manager

Submissions & Possible Errors

- Maximum number of submission: 10 (valid submissions)
  - Best submission counts

- Possible errors
  - Wrong column names
  - Unknown IDs (if not in Test Data)
  - Missing IDs (if in Test Data but not in Predictions)
  - Wrong file format
  - …

7. Upload the Predictions and the Corresponding R Script on DMC Manager



DSS Data Mining Cup   ℹ About   Franz Diebold

DMC / DMC 1

🏆 DMC 1

This is a DMC test for the central lab.

starts at: 2014-12-11 11:15

ends at: 2014-12-20 11:15

training dataset

test dataset

Your Solution

9

remaining submissions

Submit solution

Your Team

sapient shark

- *you*
- Paul Karänke

Your Standing

1

rank of best submission

Your Submissions

#	Date	Predictions	Model	Processed	Integ	Internal Rank
1	2014-12-11 13:07	predictions/predictions_group_name_number_ErlmVMZ.csv	models/Script_DMC_Intro_0BgLztQ.R	✓	⚠	
2	2014-12-11 13:11	predictions/predictions_group_name_number_wmr8bE8.csv	models/Script_DMC_Intro_pjN7vH7.R	✓	⚠	
3	2014-12-11 13:12	predictions/predictions_group_name_number_0GP8LLs.csv	models/Script_DMC_Intro_LPGFPRv.R	✓	✓	1 ★

© Decision Sciences & Systems 2014

Relative standing compared with other teams

no accuracy given

Click for error description

best own submission

Comparing Classifiers

- Classifiers are hard to compare [1]
  - Different datasets
  - Limited collection of publically available datasets
  - Different data preparation
  - Tuning
  - Statistically significant claims
  - Etc.

- No best classifier
  - under certain assumptions, no classifier is better than another one [2]

[1] Salzberg S., On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach

[2] Wolpert D., On the Connection between In-sample Testing and Generalization Error

Comparing Classifiers

Many studies make mistakes when comparing Classifiers [3]

- Not using statistical tests at all
- Apply unsuitable tests or ignore assumptions
- [3] addresses these problem for…

Comparison of Two Classifiers:

- T-test: checks whether average difference in performance is significant from 0
  - Often inappropriate due to calculating using the averages
  - E.g.: Outliers can have unwanted strong effect on data and increases the variance which decreases the test power
  - Assumes the difference between random variables to be normal distributed (N<30; both often not given)
- Wilcoxon Signed-Ranks Test: non-parametric, ranks the differences in performance and compares them
  - Does not assume normal distribution and is less affected by outliers

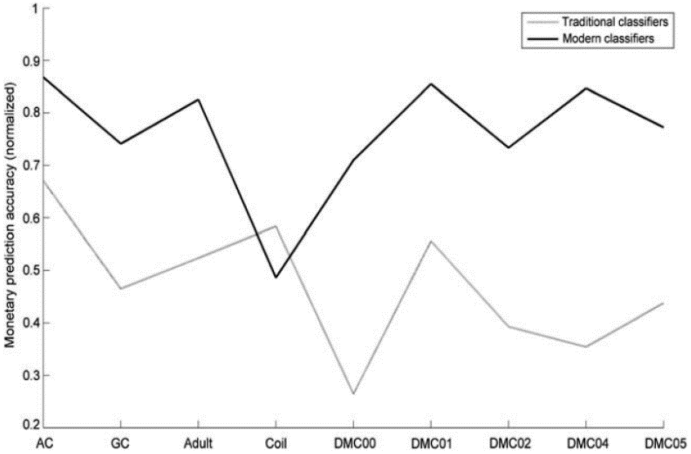
Comparison of Multiple Classifiers

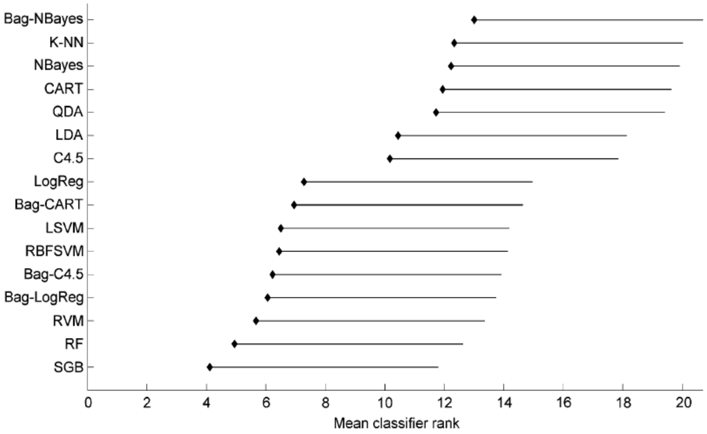
[3] Demsar J., Statistical Comparisons of Classifiers over Multiple Data Sets

Comparing Classifiers

However, there is a number of studies, which can provide useful guidelines on classifier selection

- Modern vs Traditional Classifiers [4]





[4] Lessmann S., Voß S., A Benchmarking Study of Novel Versus Established Classification Models

Questions?

Information about the „caret package“

http://topepo.github.io/caret/

Example dataset raw_data_large

Data

- History of purchase of an online shop
- Both information about good and customer

Task

- Predict if there would be a return

Column name	Description	Range of values	Missing values
ID	Order id	Natural number	No
od	Order date	Date	No
dd	Delivery date	Date	Yes
size	Item size	String	No
price	Price of item	Positive real number	No
tax	Tax	Positive real number	No
a6	Salutation	String	No
a7	Date of birth	Date	Yes
a8	State	String	No
a9	Return shipment	{0,1}	No