8 Tipps, die Sie gerne wissen möchten, bevor Sie Ihren ersten SAS Predictive Modeling Hackathon für Ihre Studierenden veranstalten

Gerhard Svolba
Analytic Solutions Architect
SAS Austria

Credits for Input to:
Tamara Fischer

Twitter: https://twitter.com/gsvolba
https://github.com/gerhard1050
https://www.linkedin.com/in/gerhardsvolba/

sas
THE POWER TO KNOW.

Lecturer for Data Science Topics

FH OBERÖSTERREICH UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Marketing Information Systems for „Global Sales and Marketing“ (Master)

FH Burgenland

Visual Analytics and Business Intelligence for Business and Process Engineering (Master)

universität wien

Data Science Case Studies with SAS (Bacc+Master)

MEDIZINISCHE UNIVERSITÄT WIEN

Methodological Seminar (MD)

sas

Why running a machine learning hackthon with students?

- Modern contemporary way to involve people in programming and data science topics. Found in several areas (kaggle, climathon …)
- Increases the students‘ motivation to deal with the business subject and methodological subject
- Not more effort compared to a classical lecture assignment and its evaluation
- Possibility to extend the hackathon task in the lecture with presentation and interpretation tasks

General setup of my hackathons

- Hackathon based on marketing data.
- Build a predictive models with SAS (modeling procedures or SAS Visual Analytics) to predict who might be a responder (based on data from a historic campaign).
- Students received 90% of the data (100,223 records)
- I withhold 10% (10,892 obs) for the evaluation.
- Out of these test data I used the top 20 % scored to check whether they have a response or not and calculate the hitrate per student.
- Dataset contains 11 variables + 1 target variable.

Lecture: Data science case studies with SAS

Statistics Students

- Students worked with SAS programming only (Datastep, SAS procedures)
- Students could choose between
  - SAS University Edition (local on their laptop, Virtual machine)
  - SAS On Demand for Academics (cloud version).
- The students submit their model logic (SAS scorecode or bin file).
  - I performed the scoring.
  - An alternative would have been to give them also the unlabeled test data and to return the scores to me
- 8 lectures (4 hours each): Hackathon task in the 7th and 8th lecture
- 32 students, 8 groups

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA
OBERÖSTERREICH

Marketing Information Systems for „Global Sales and Marketing“

- Students only worked with SAS Visual Analytics (no coding)
  - 2 lectures (5 hours each) to explain idea and methods of customer analytics and usage of SAS Visual Analytics
  - 1 lecture to run presentations
- Used a tenant in the SAS Viya cloud environment
- 3 parallel lectures (full time, part time, triple degree), 57 students, 15 groups
- Simulated the situation: Retail company sends out an RFI for marketing analytics.
  - Students had to perform presentations, presenting what they have found with their models and what marketing actions they are suggesting to run.

Summary

- Was highly impressed how well they did and how far they came when
  - defining customer segmentation,
  - describing the features of the customer base and
  - identify the high affinity customer groups and describing their features.
- And they did extremely well:
  - I saw interactive software demos in the presentations.
  - Segments defined by a decision tree in SAS Visual Analytics directly interpreted for marketing actions.
  - Many findings were illustrated with persona roles to relate business actions to these segments


Recommendations


ClickShare
L'OSTERIA
GRANDE AMORE
GRANDE PIZZA, GRANDE GUSTO!

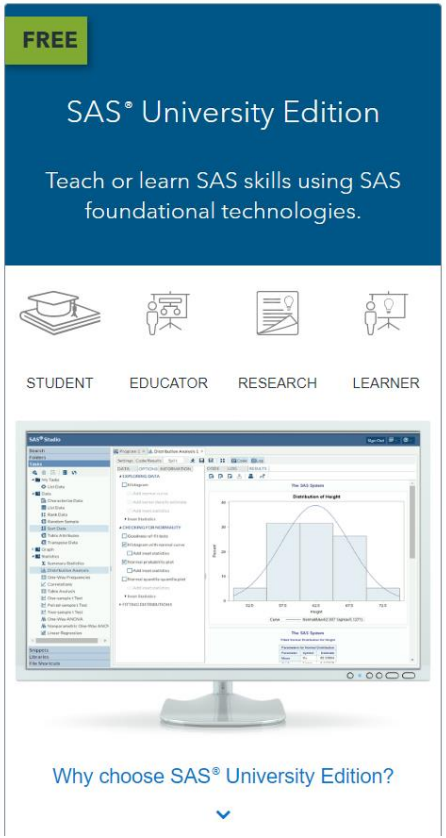
My experiences:

#1

Ensure that all students have access to the same set of modeling procedures

SAS® Academic Software

https://www.sas.com/de_de/learn/academic-programs/software.html



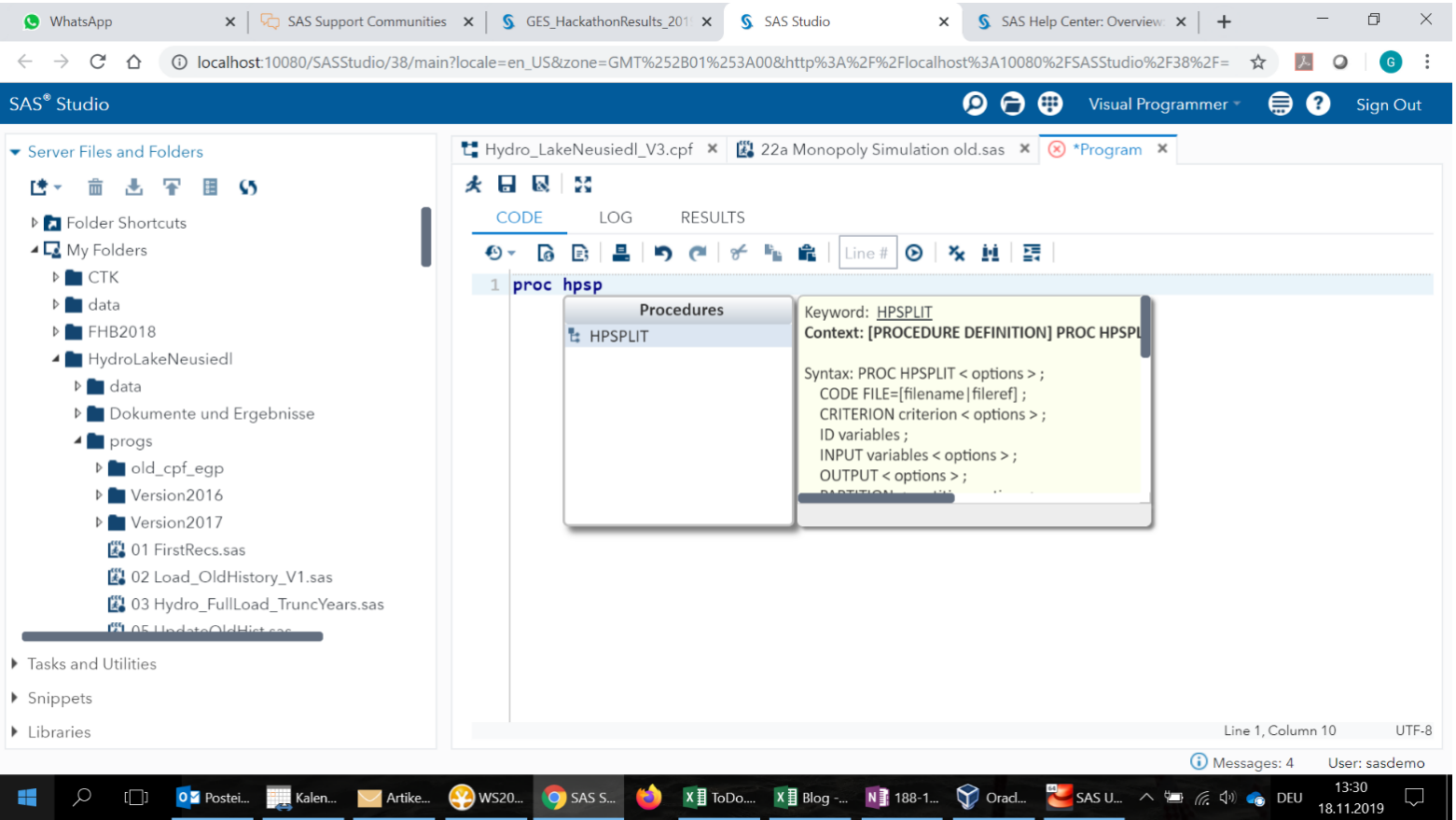
SAS University Edition or SAS On Demand for Academics?

- **SAS University Edition** only contains the procedures of classical SAS STAT.
- Limits the students using the SAS UnivEdition to models like LogisticRegression (PROC LOGISTIC), discriminant analysis (PROC DISCRIM)

- **SAS On Demand for Academics** also contains the High Performance Analytics procedures.
- Students using SAS On demand for Academics also have access to procedures like and decision Trees (PROC HPSPLIT) and HPFOREST for Random Forests or HPNEURAL for Neural Networks.

- In order to compensate for the different procedure sets, I defined a subgroup ranking and also honored the students who just built regression models.
- Note that it is also misleading because the syntax suggestion in SAS Studio suggests FOREST or TREESPLIT, even if this procedure is not available in the package.

Syntax Suggestion in SAS Studio



#2
Be very rigid when defining the submission procedure

Example for the need to individualize the evaluation

```
/** TD */
%HackEval(TD1);
%HackEval(TD2);
%HackEval(TD3);
%HackEval(TD4);
%HackEval(TD5);
/**/

/*** FT  **/
%HackEval(FT1,probvar=P__va_c_TargetBuy_Catpurchase);
%HackEval(FT2);
%HackEval(FT3);
%HackEval(FT4,probvar=P__va_c_BUYYes);
%HackEval(FT5);
/**/

/* PT */
%HackEval(PT1,probvar=P__va_c_Purchase_Desipurchase);
%HackEval(PT2);
%HackEval(PT3);
%HackEval(PT4);
%HackEval(PT5);
%HackEval(PT6,probvar=P__VA_C_TARGETBUY1RESPONDERS);
```

#3

Prepare an automatic evaluation program that stores and scores the result

Prepare your evaluation program in a highly automated way

- You might have to apply it more than once
- SAS macro to loop over all student contributions
  - group-id variable as a unique key
  - automatically consumes the respective score logic and the name that datasets that contain their scores.

Datastep Scorecode for LOGISITC/TREE

```
*** 3. Create Test Data with Student ID;
data work.tst_&id.;
 set work.bigorganics_tst;
run;

*** 4. Perform Score Code;
data work.tst_&id.;
 set work.tst_&id.;
 %include "&path./&id._ScoreCode.sas";
 *P_targetbuy1 = P_targetamt1;
run;

*** 5. Select top 2178 Records;
proc sort data=work.tst_&id.;
 by descending P_targetbuy1 ;
run;

data work.Top2178_&id.;
 set work.tst_&id.(obs=2178);
run;
```

```
*** 6. Evaluate Hits;
proc freq data=work.Top2178_&id.;
 title Treffer in Baseline für Gruppe &id.;
 table TargetBuy;
run;
title;
```

The HPFOREST procedures stores the score logic in a binary file (SAS Viya: ASTORE).

```
proc hpforest data=bigorganics_trn maxtrees=2 ;
    input demaffl demage PromSpend PromTime /level= interval;
    input DemCluster demgender DemTVReg PromClass /level=nominal;
    target targetbuy /level= binary;
    ods output FitStatistics=fitstats;
    partition fraction (validate=0.3);
    save file="&path./GES_ScoreCode.bin";
    score out=scoreRF;
    run;
```

```
%let id = GES;
```

```
data tst_&id;
  set work.bigorganics_tst;
run;
```

```
proc hp4score data=work.tst_&id.;
score file= "&path./GES_ScoreCode.bin"
      out=work.tst_&id.;
run;
```

#4

When defining the grades:

Consider that the hackathon hit rate and the students effort not necessarily has a full positive correlation

#5

Use a simple datastep to split of the test data

Withholding test data for the evaluation

```
data em.bigorganics_TRN
     em.bigorganics_TST;
 call streaminit (13);
 set em.bigorganics;
 if rand('Uniform') < 0.9 then output em.bigorganics_TRN;
 else                          output em.bigorganics_TST;
run;
```

#6

When working with „Big Data“,
make sure that your data is a true big data set
from individual analysis subjects

Just replicating your data rewards overfitting

- If your base dataset contains only ~10,000 observations and you want to provide a „big“ dataset with ~200,000 observations
- Don‘t just replicate your data 20 times into the final dataset.

```
DATA Bigdata;
 SET SMALL_DATA SMALL_DATA … SMALL_DATA
RUN;
```

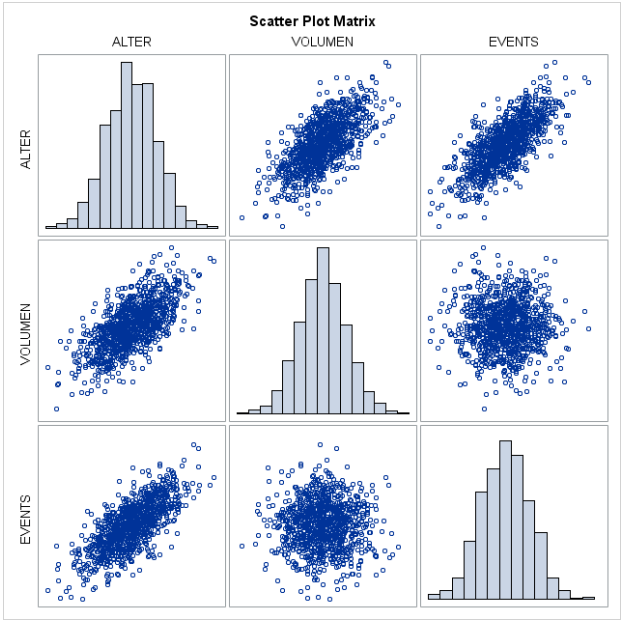
- As you have identical observations in you data, this data will reward overfitting a predictive model.
  - → Because the same obervations might end in the training AND the test data.

- For internal variable you can create syntethica data on basis of the co-variances.
- See also my presentation from KSFE 2016, Greifswald
  „ **Simulationen und Mathematische Programmierung mit SAS“**

Tipp #5

SO SIMULIEREN SIE DATEN AUS EINER MULTIVARIATEN VERTEILUNG

```
proc iml;

Mean = {42, 5200, 280}; /* population means */
Cov =
{12 48 25, /* population covariances */
48 420 0,
25 0 100};

N = 1000; /* sample size */
call randseed(123);
X = RandNormal(N, Mean, Cov); /* x is a 1000 x 3 matrix */

SampleMean = mean(X);
SampleCov = cov(X);
varNames = {Alter Volumen Events};

print SampleMean[colname=varNames],
SampleCov[colname=varNames rowname=VarNames];

/* write sample to SAS data set for plotting */
create MVN from X[colname=varNames]; append from X; close MVN;
quit;
```



Proc CORR and PROC SIMNORMAL

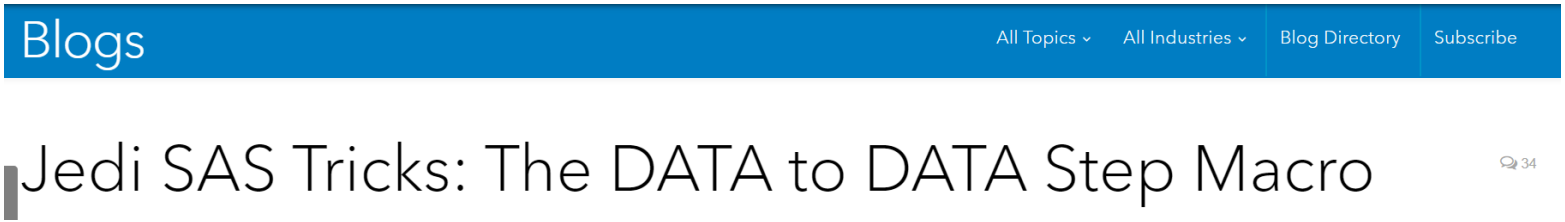
- Note that this can also be performed with the SIMNORMAL procedure (part of SAS/STAT).
- Covariances can be easily created with the CORR procedures.

```
proc corr data=em.hmeq out=hmeq_cov cov noprint nocorr;
 var BAD LOAN MORTDUE VALUE YOJ DEROG DELINQ CLAGE NINQ
CLNO DEBTINC;
run;

proc simnormal data=hmeq_cov(type=cov)
               out =hmeq_10000 numreal= 10000 seed= 123456;
 var BAD LOAN MORTDUE VALUE YOJ DEROG DELINQ CLAGE NINQ
CLNO DEBTINC;
run;
```

#7

Use the %data2datastep macro to simply transfer data to other SAS environments

- %***data2datastep***(bigorganics_trn,em,work,c:\tmp\bigorganics_trn.sas);

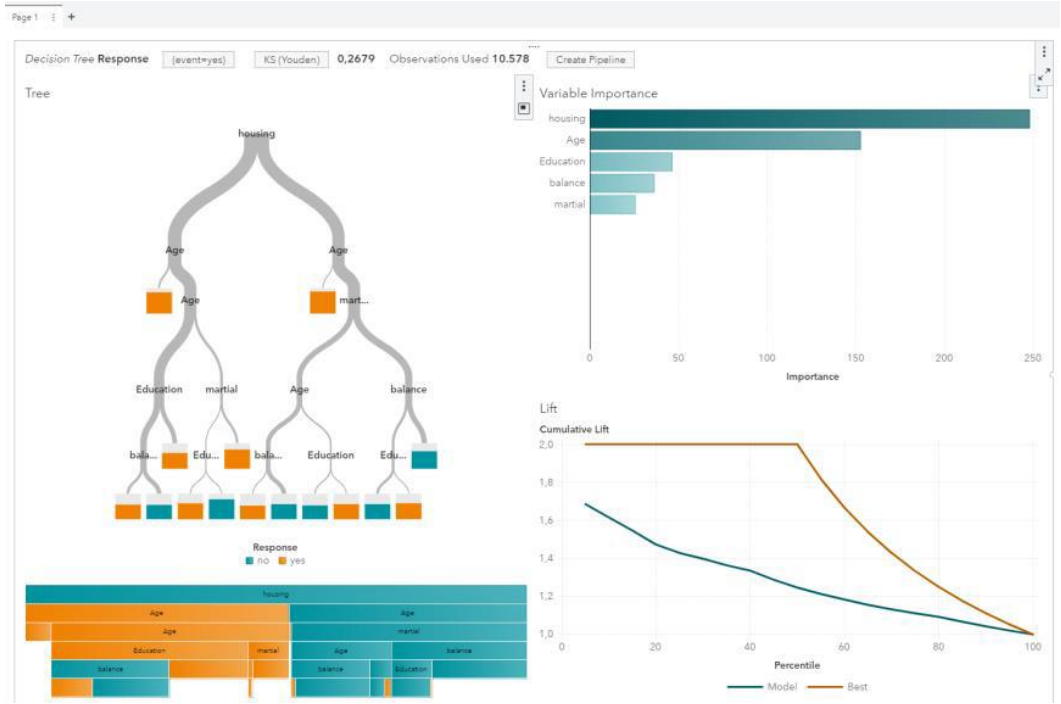


https://blogs.sas.com/content/sastraining/2016/03/11/jedi-sas-tricks-data-to-data-step-macro/

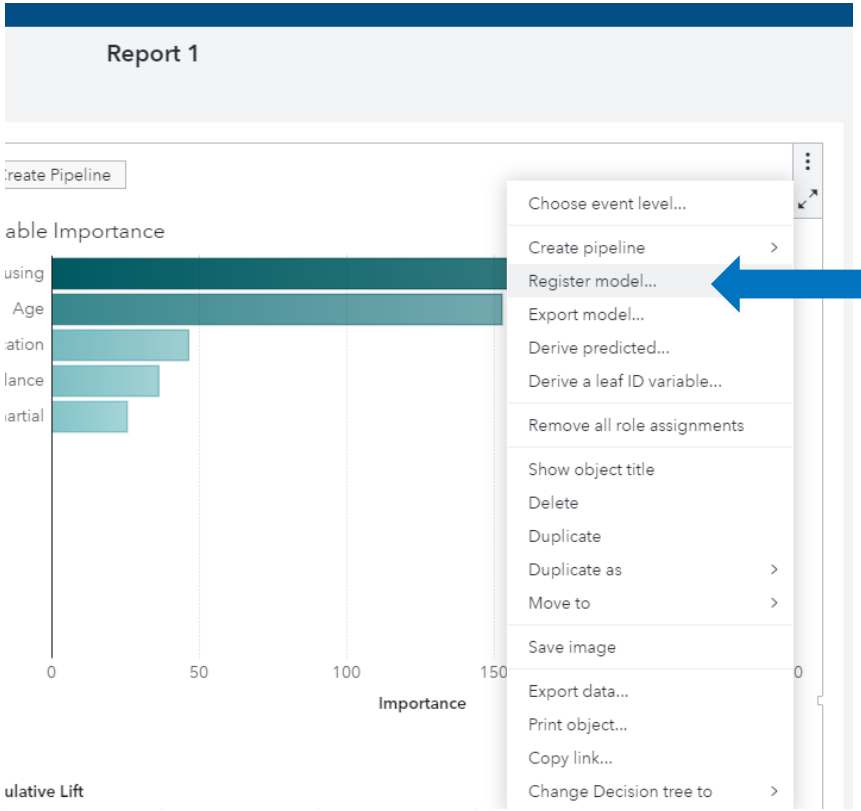
#8

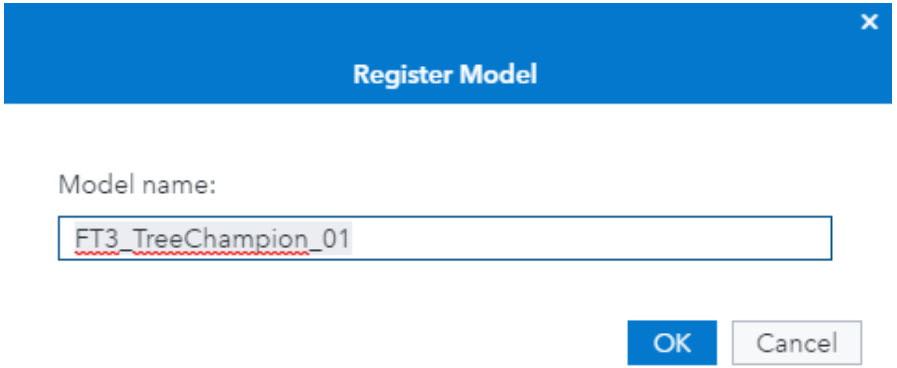
How SAS Model Manager became my best friend – Consuming, Documenting and Evaluating Predictive Models in a SAS Predictive Modeling Hackathon in Academics

Building a model in SAS Visual Analytics

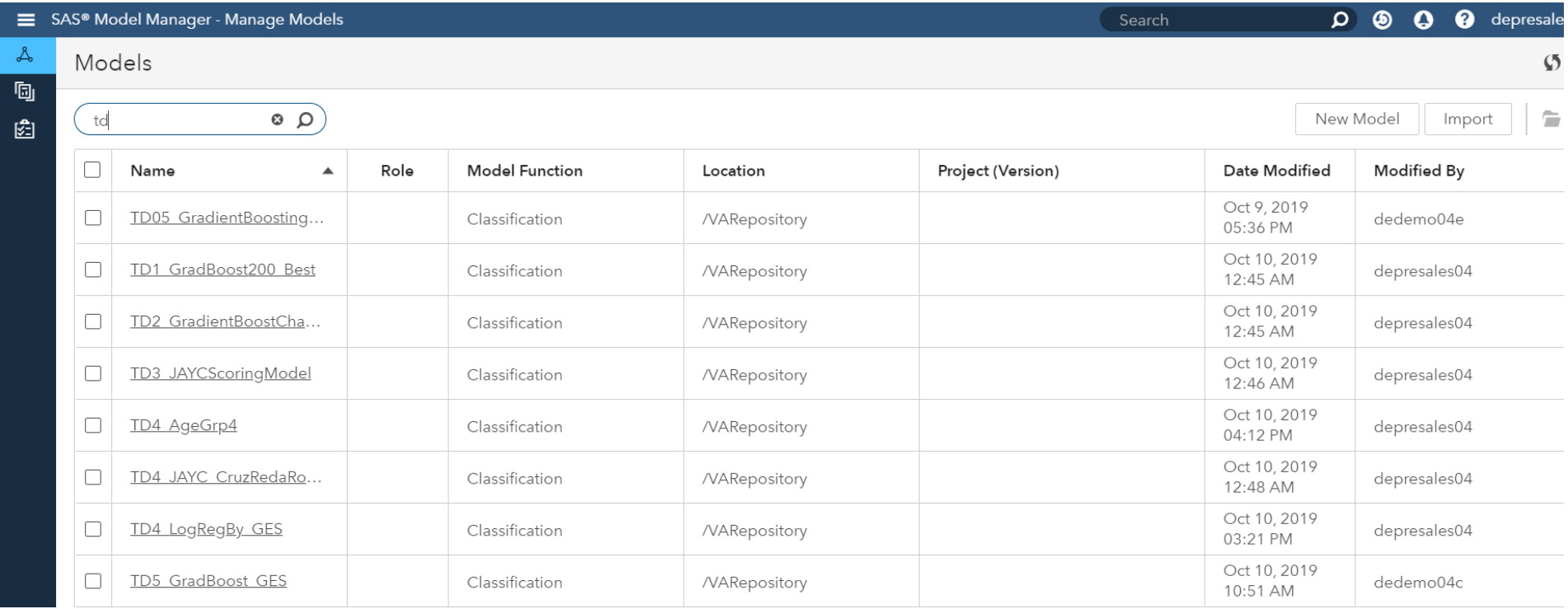


Registering and naming the model





Viewing the model in SAS Model Manager



Name	Role	Model Function	Location	Project (Version)	Date Modified	Modified By
TD05_GradientBoosting…		Classification	/VARepository		Oct 9, 2019 05:36 PM	dedemo04e
TD1_GradBoost200_Best		Classification	/VARepository		Oct 10, 2019 12:45 AM	depresales04
TD2_GradientBoostCha…		Classification	/VARepository		Oct 10, 2019 12:45 AM	depresales04
TD3_JAYCScoringModel		Classification	/VARepository		Oct 10, 2019 12:46 AM	depresales04
TD4_AgeGrp4		Classification	/VARepository		Oct 10, 2019 04:12 PM	depresales04
TD4_JAYC_CruzRedaRo…		Classification	/VARepository		Oct 10, 2019 12:48 AM	depresales04
TD4_LogRegBy_GES		Classification	/VARepository		Oct 10, 2019 03:21 PM	depresales04
TD5_GradBoost_GES		Classification	/VARepository		Oct 10, 2019 10:51 AM	dedemo04c

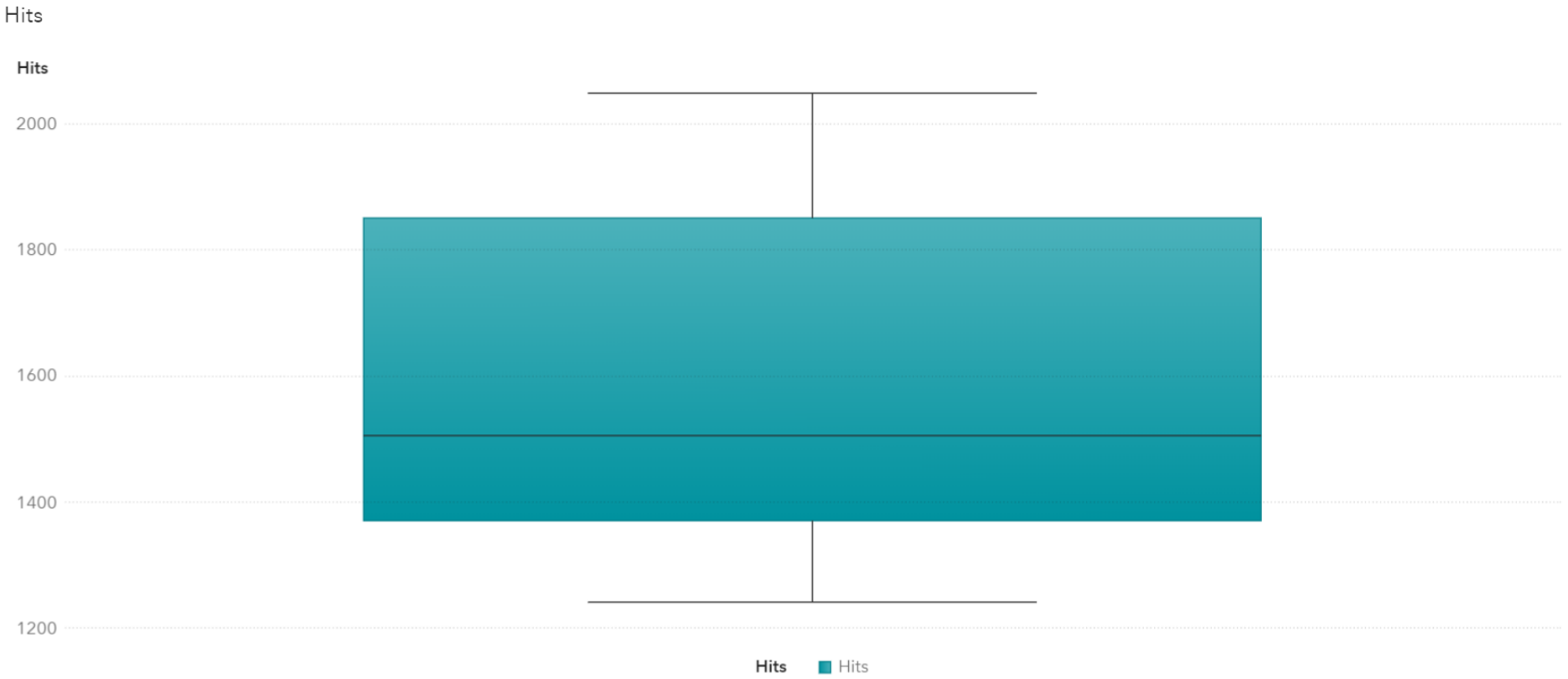
Automatically scoring each model with SAS Code
(SAS Viya)

- Students did not have to submit SAS Score, binary files or SAS ASTORES
- They just registered their model
- I could simply access this models in SAS Model Manager
- Could use the following DS2-Code to score and evaluation the models

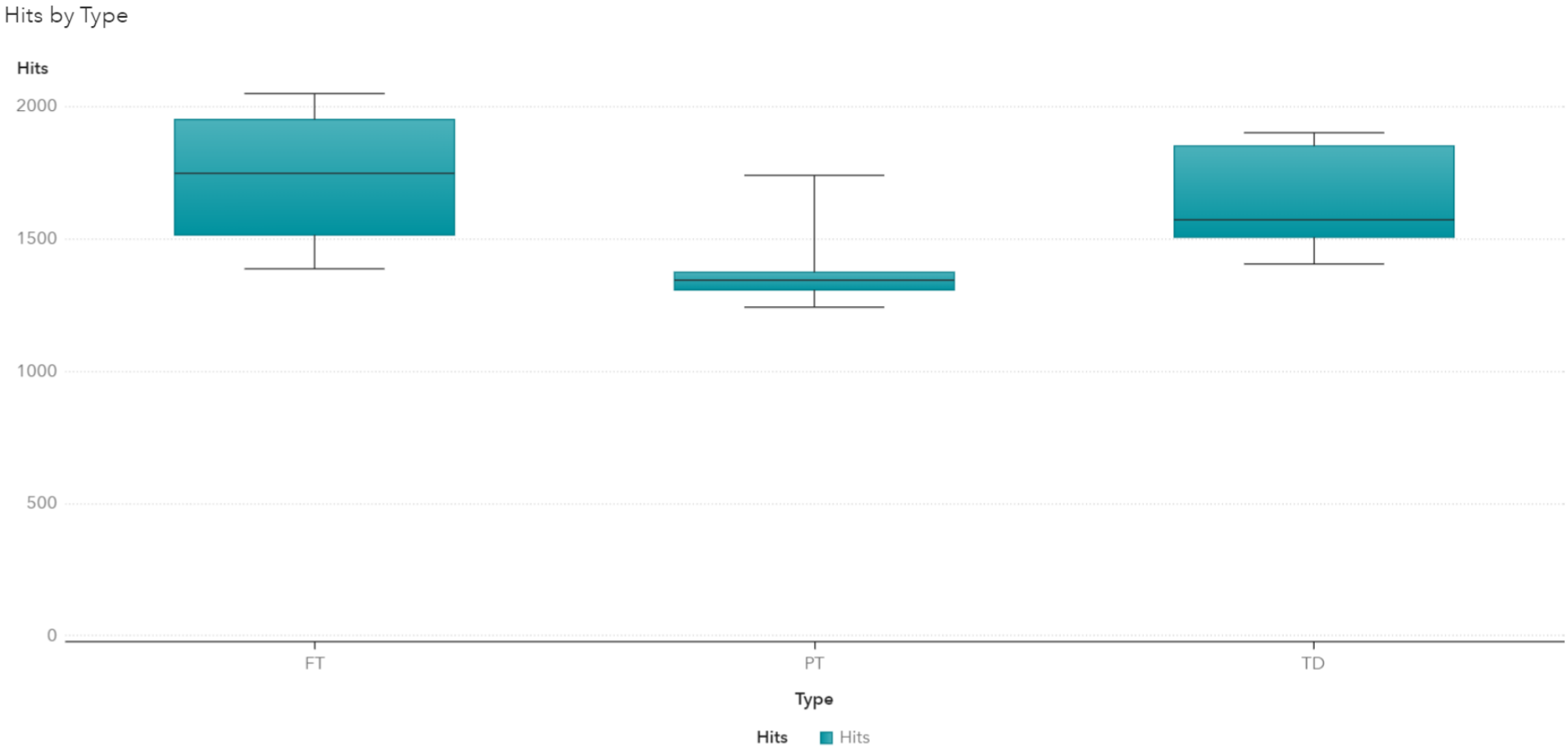
```
proc cas;
loadactionset "ds2";
action runModel submit /
    modelTable={name="SAS_MODEL_TABLE", caslib="PUBLIC"}
    modelName="HMEQ_GB"
    table={name="HMEQ", caslib="PUBLIC"}
    casOut={name="HMEQ_SCORED", caslib="CASUSER"}
strictLevel="IGNORE";
run;
quit;
```

#9
Displaying the Hackathon Results step-by-step (with SAS Visual Analytics)

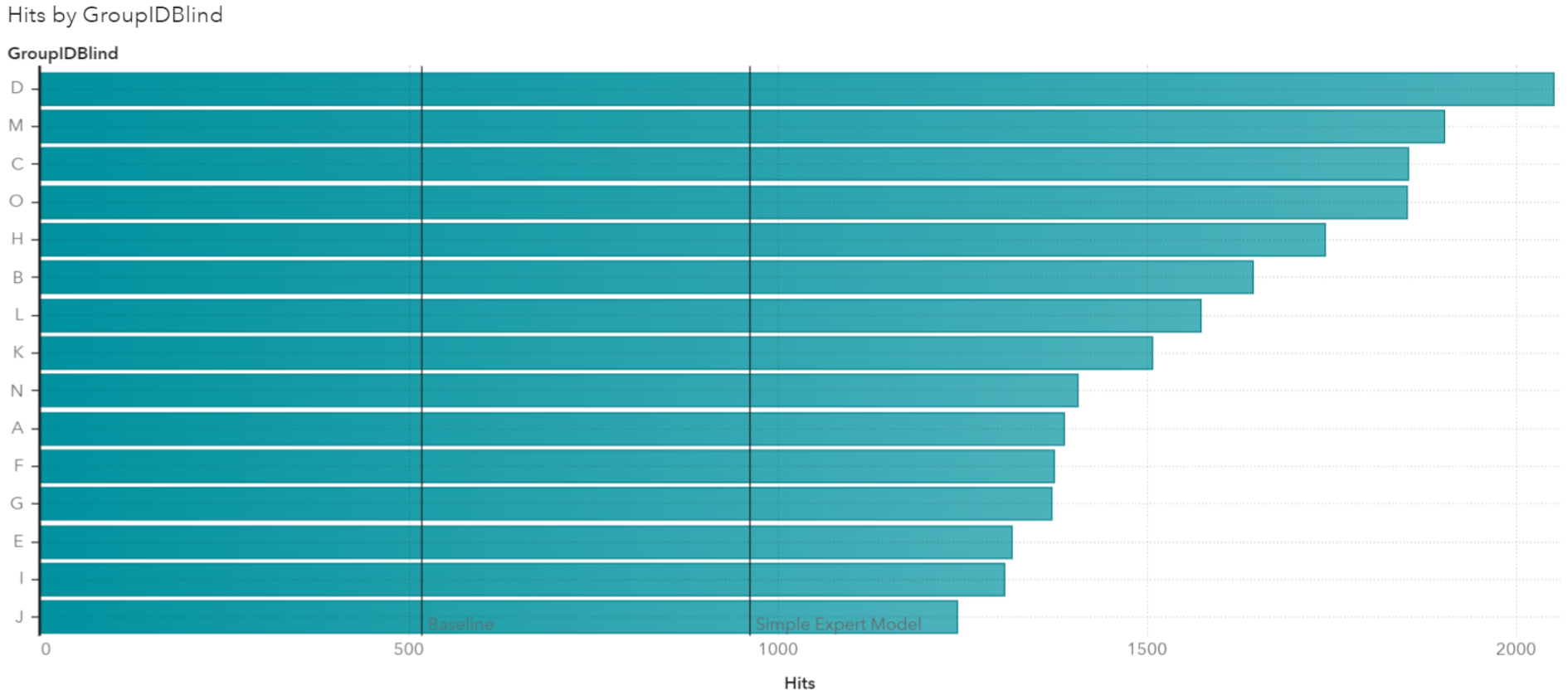
Hackathon: Hit Range



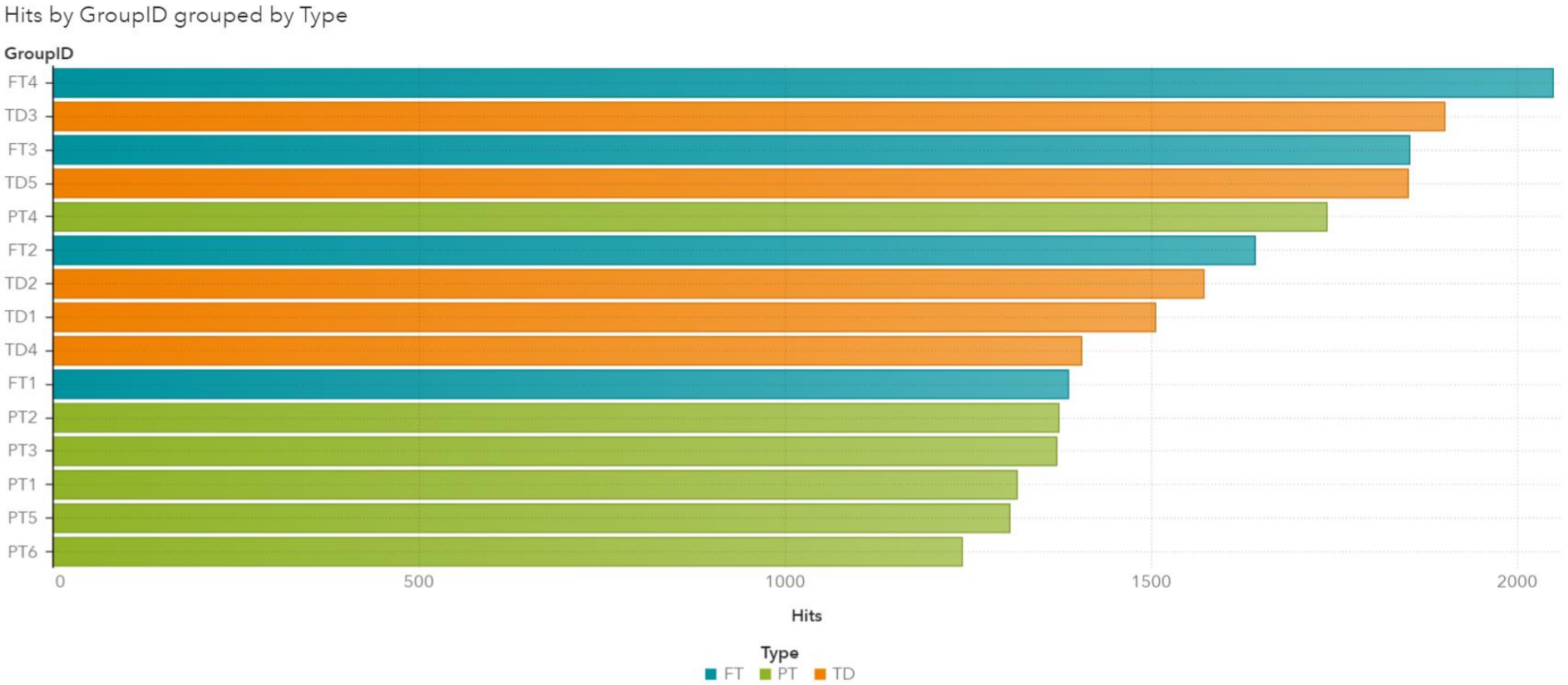
Hackathon: Hit Range by Group



Hackathon: Hit Rate Blinded



Hackathon: Hitrate unblinded for Teams



Students:
Two bonus tips for students

Students #1
First prepare a minimum framework that delivers the required result set, only then move to fine-tune your model

Students #2
Minor mistakes can ruin your performance

```
0 <= x <  5
5 <  x <= 8
```

Conclusion

- I can recommend running a hackathon in your lecture
- SAS offers free software for students (non-students)
- Make your life as easy as possible (SAS Macros, Requirements, …)