Model Evaluation Metrics in Machine Learning. Understand the problem and a practical approach

Hans de Wit, Senior Data Scientist, Telenor Norway

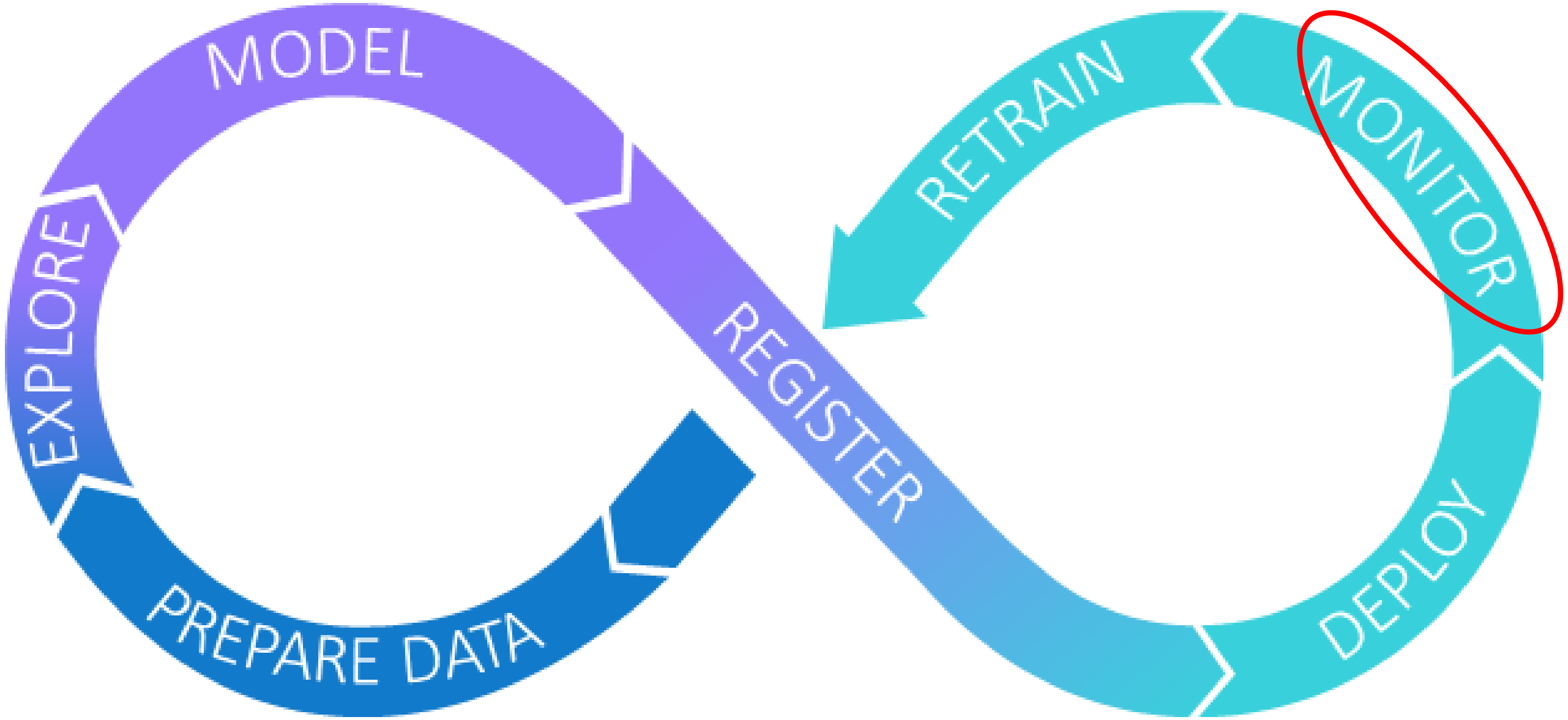
Hans de Wit

- **My passion:**
- **Making the unreal happen**

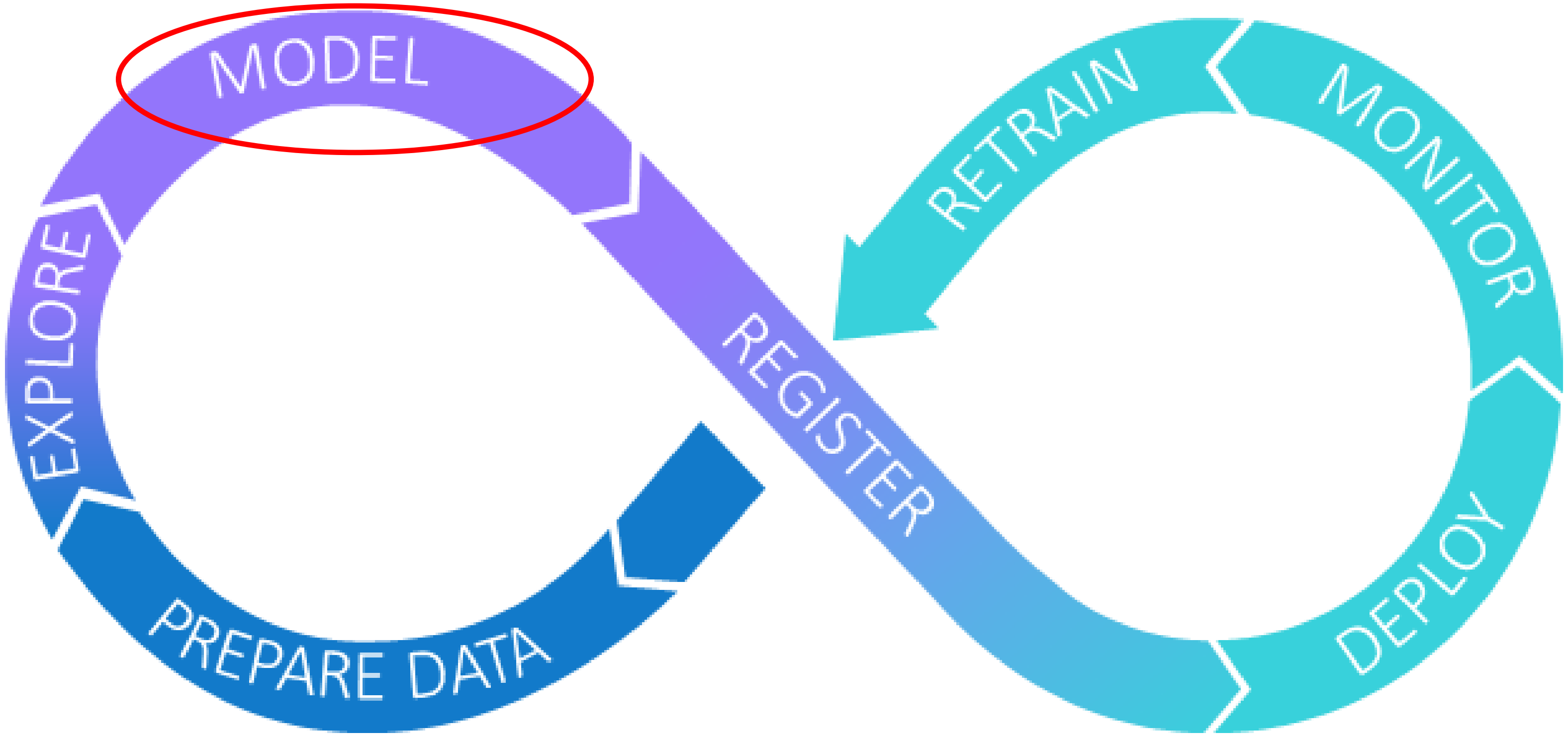
- **Telenor Mobile/IT Norway (since 2013)**
  - Advanced Analytics & Data Science Manager
- **ING Bank, The Netherlands**
  - Senior member 'Model'/Innovation-team ING Retail Customer Intelligence
  - Member analytical campaign management ING Bank Customer Intelligence department, 1997-2005
- **ING Card, 2005-2008**
  - Direct Marketing, Credit Risk, Fraud

– Master of Marketing (SRM) and bachelor of Commercial economics and Direct Marketing.

During this presentation we will focus on Model performance (Monitor).



Let’s start with the Model building/develop phase, with the focus on model comparison.



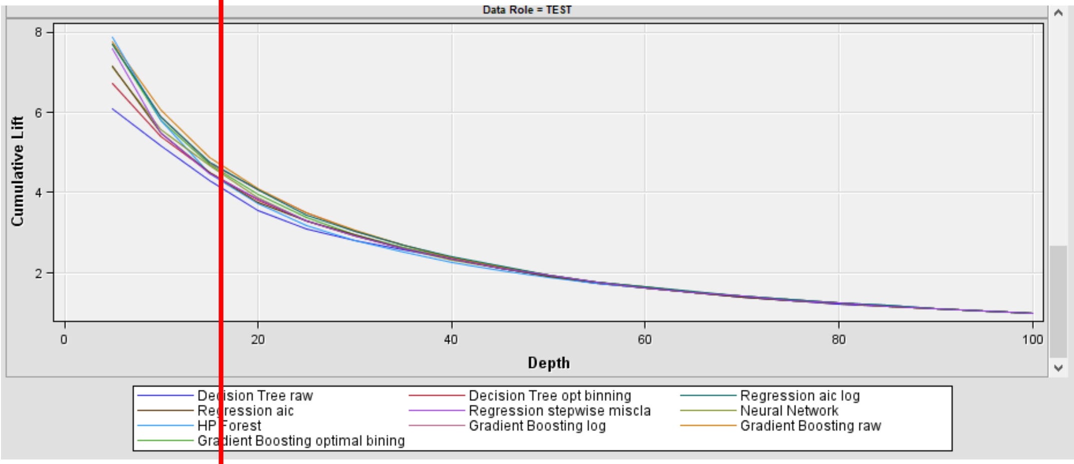
Assesment measures to choose the best model to solve your problem. Overall measurement.

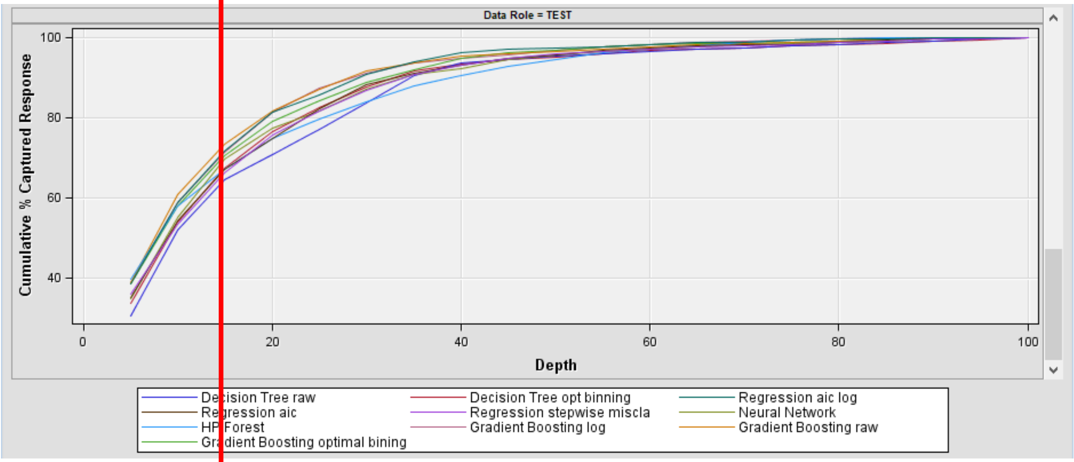
- Binary target
  - Decision
    - Accuracy
    - Misclassification
    - Profit/Loss
  - Ranking
    - C statistics (AUC)
    - Gini Coefficient
  - Estimate
    - Average square error
    - AIC
    - Root mean square error

- Interval target
  - Average squared error (ASE)
  - Root average squared error (RASE)
  - Root mean absolute error (RMAE)
  - Root mean squared logarithmic error (RMSLE)

In marketing campaigns, it is quite common to use lift or cumulative lift of the top xx% to choose the best model. Like top20%. **Measures at Pre-Specified Cutoff Points**

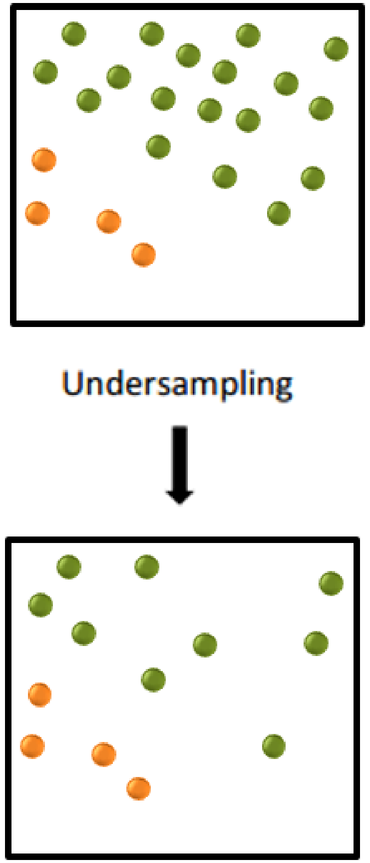
- Most (1:1) campaigns require a decision on which customer would be eligible for a specific campaign based on a model.
  - Cut off could be cumulative above 2.
  - They want the best model in the top 20 percent of the population.
- Business is not interested in how the model performs after the cut-off.





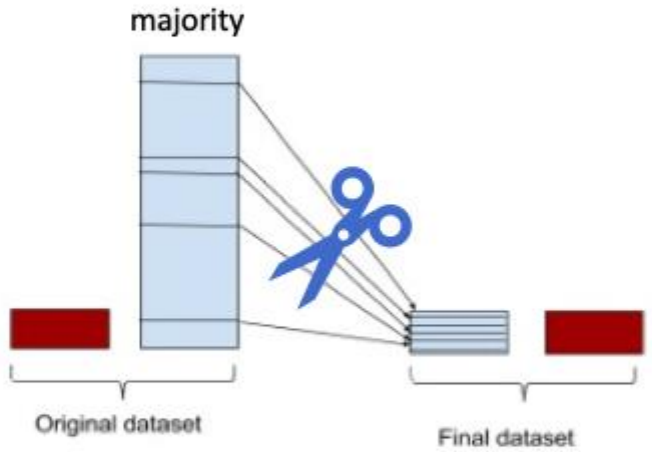
In many branches, like the Finance industry or Telco, it is quite common to have to deal with rare events. The event rate is really low <5%.

- Examples
  - Marketing
  - Fraud
  - Churn
- Algorithms have a hard time discriminating between events and non events.
- (under) Sampling is needed.



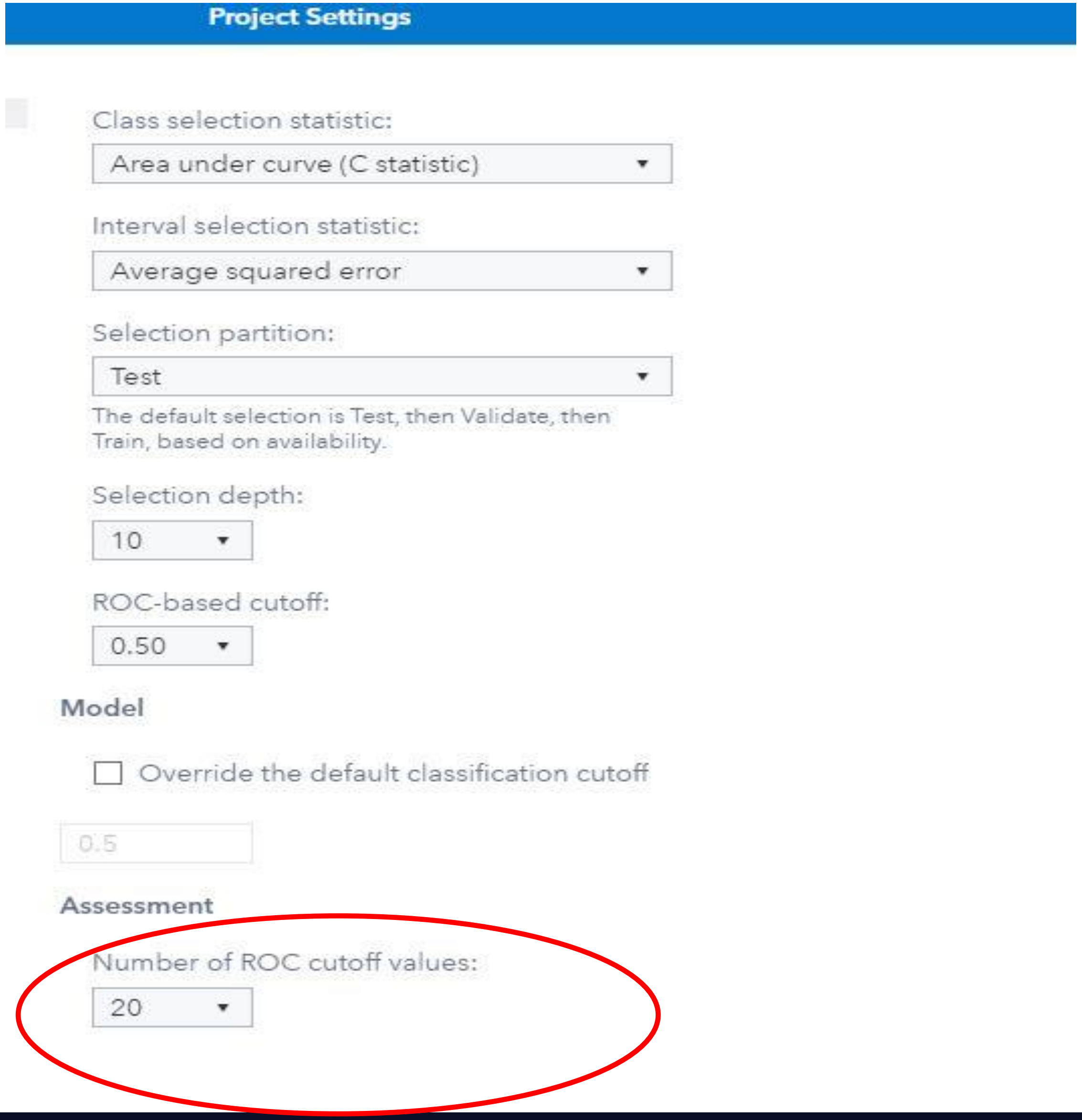
Why you should use sampling in SAS Model Studio or SAS Enterprise Miner and not before?

- Sas will recalculate the result to the real event rate, when you are using rare event sampling.
- If you do the sampling before SAS Model Studio, some fit statistics will not be correct. (precision, recall, f1 score, misclassification rate).
  - What is the max lift in case of 50/50 (balanced dataset?
  - AUC will be correct, because AUC is independent of the event rate.
- Once you deploy a model, you are using the correct event rate (probabilities) especially when you have to compare different models. (NBA)
- You can implement the prior in Sas Enterprise Miner.

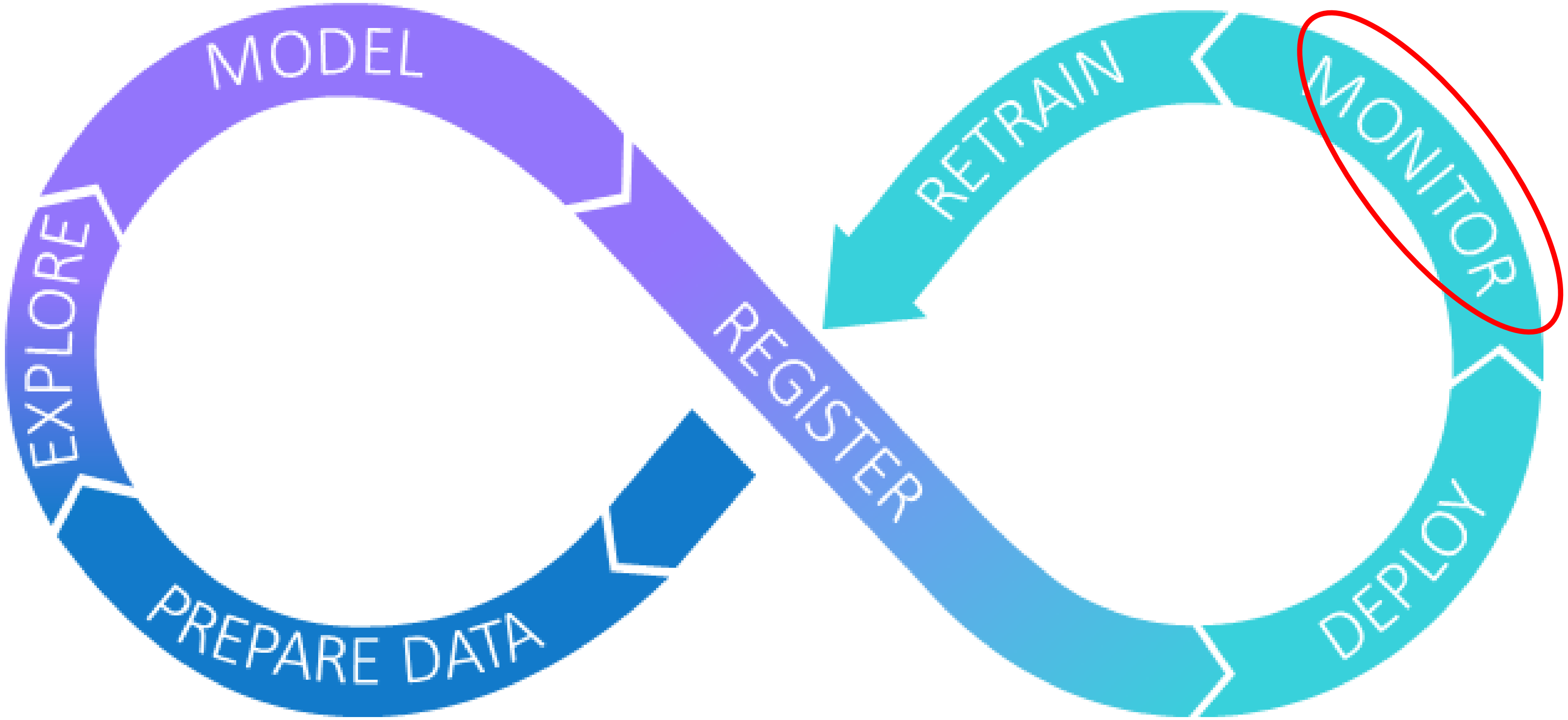


Rare events and SAS Model studio

- Read the abt, without sampling in Sas Model Studio.
- Click advanced and choose event-based-sampling.
  - Choose like 20/80 or 50/50 or your choice.
- For very rare event (<1%) you have to go to Project Settings and change the assessment settings ‘number of ROC cutoff values’
  - From 20(default) to 500.
  - Otherwise the AUC and other fit statistics will be have a different outcome.



Monitoring model in Sas Model manager



Why do we monitor models?

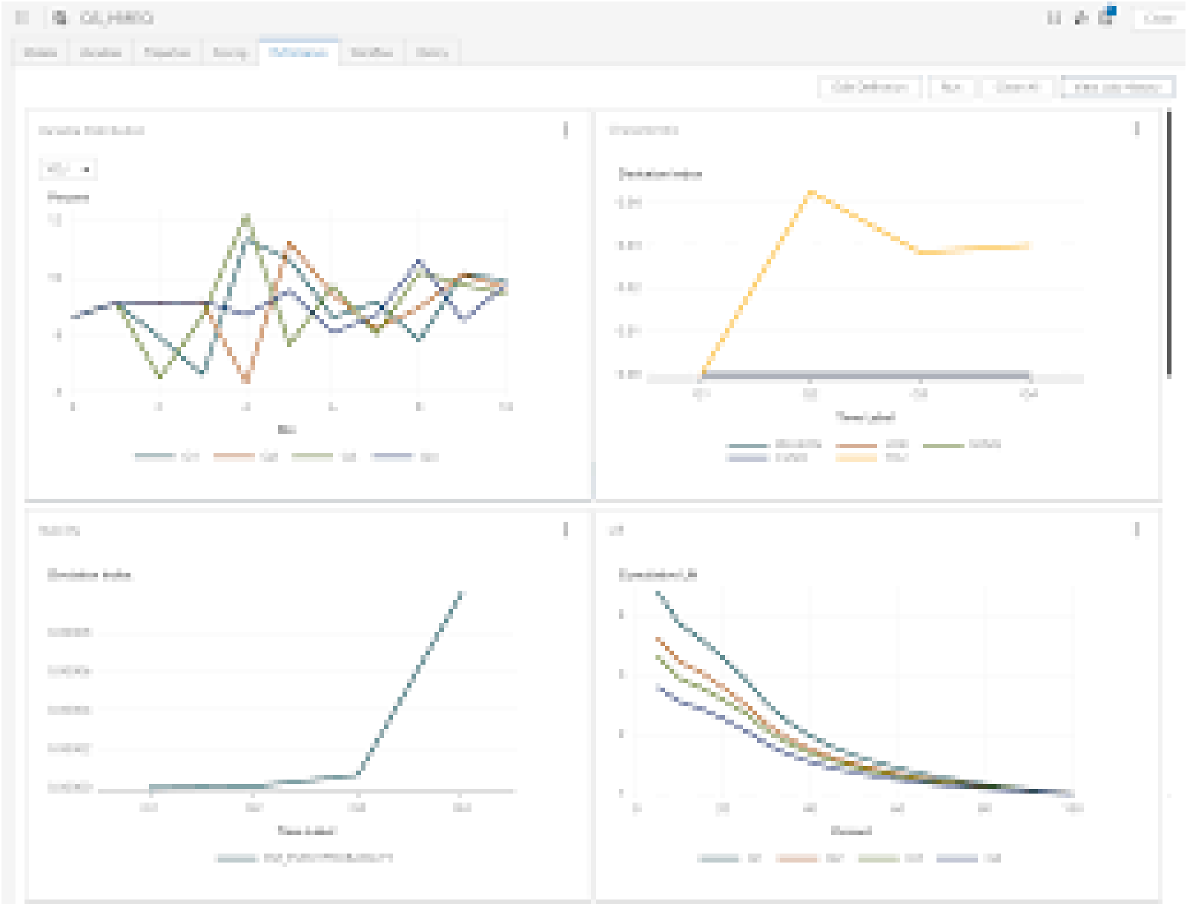
- Every day we make decisions based on models. You would like to make the correct decision.
- Following data quality/definitions. Have the distribution of the features changed?

2 reasons why a model is not performing anymore

- The distribution of the features of the models have changed over time
- The profile of the target has changed.
  - Example: There are more and more older people buying an Iphone, compared to when the model was initially build

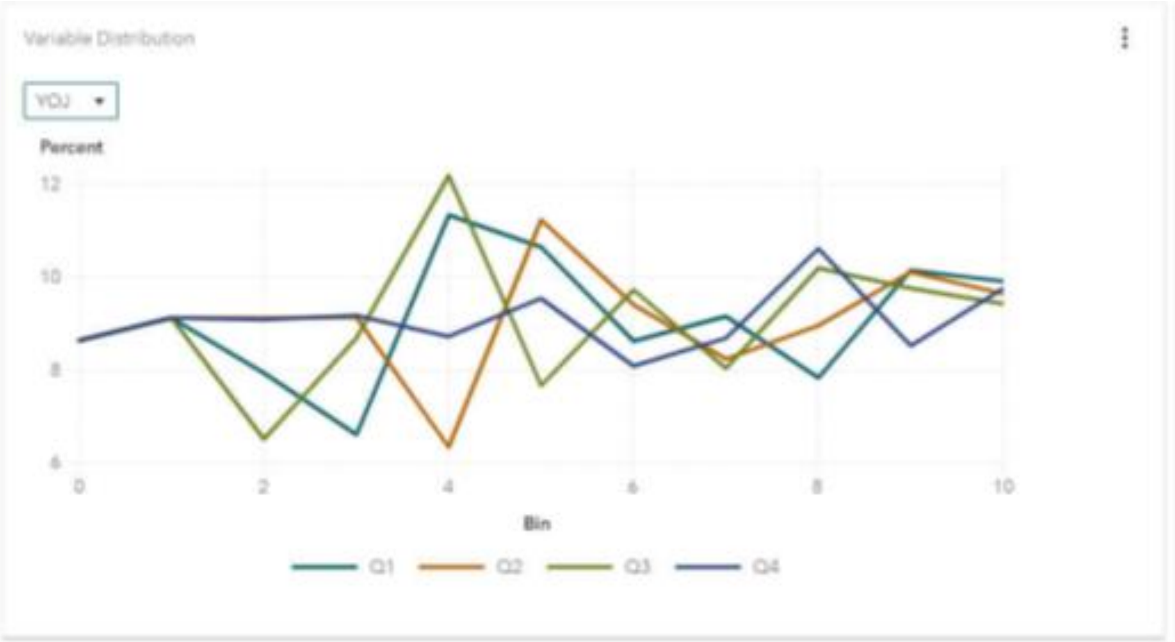
Monitor models

- Variable distribution
- Model stability
  - Deviation index
  - Lift deviation
- Model accuracy
  - AUC /Gini index
  - KS (Kolmogorov Smirnov statistics)



Variable distribution, did the distribution of the features change over time

- Some times a data manager changed the definition of feature. From minute to second.
- New categories are added or old categories are dropped.



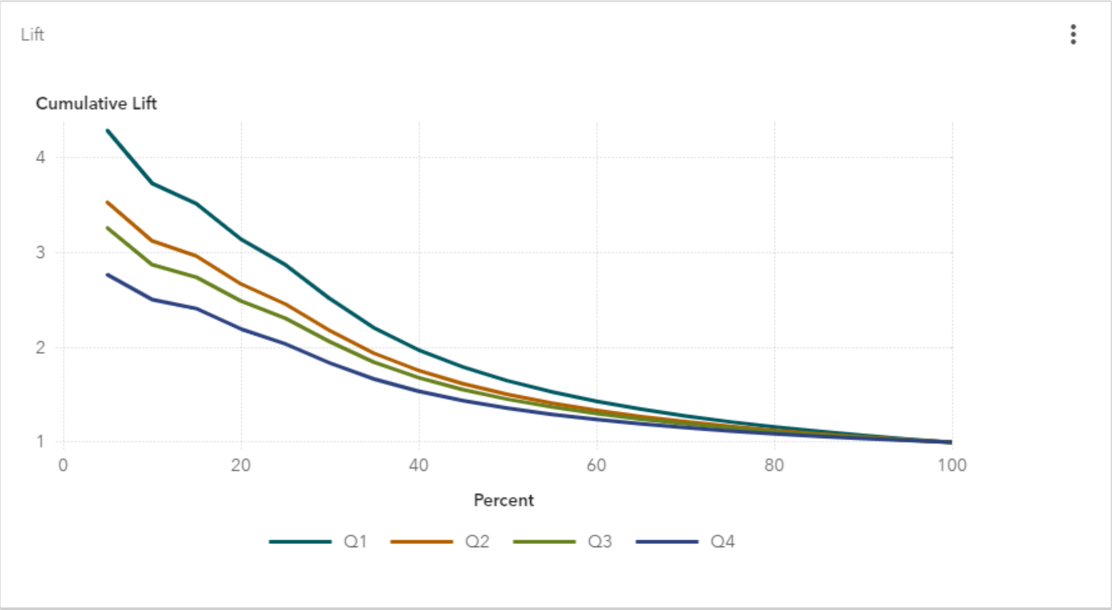
Deviation Index

- a) Stability of the output variables, i.e. the estimated propensity probability
- b) Stability of the input variables, i.e. the model input variables
- c) Stability of the macro-economic environment, i.e. the population event rate

- **KPI**:
- > No alert/warning: A deviation of <10% is considered as non-significant.
- > Warning Condition: A deviation of [10%, 25%] is considered as relatively significant and should be examined along with the input variables and macro-economic environment deviation to see which one caused the problem and recalibrate it.
- > Alert Condition: A deviation of >25% is considered as very significant and a possible redevelopment should be examined.

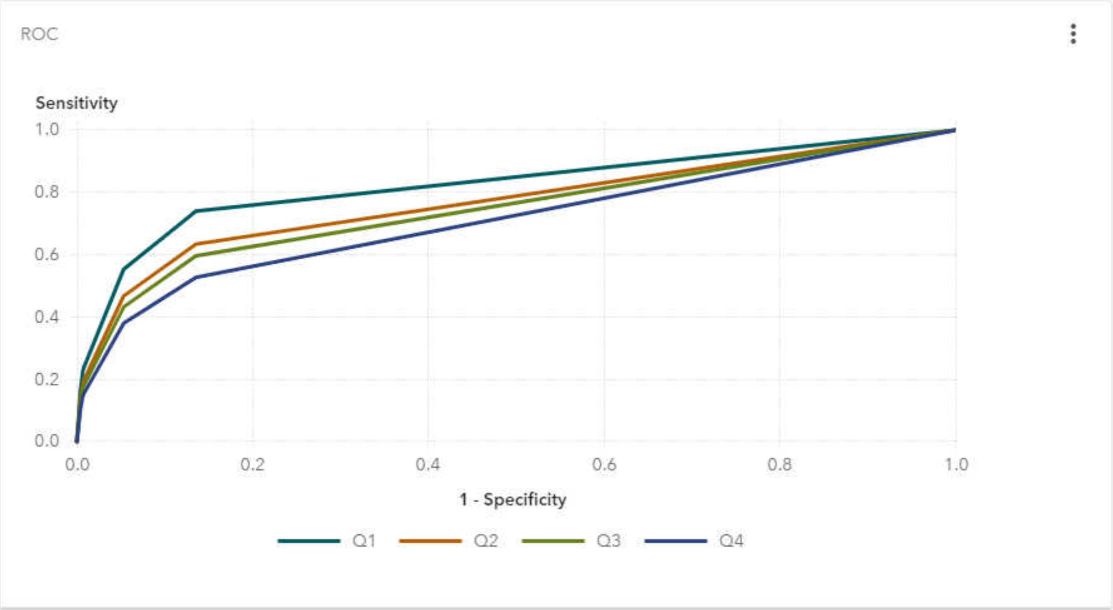
Lift

- **Decision**
- No alert/warning: A deviation of <10% in the top 20% of the population.
- Warning Condition: A deviation of [10%, 25%] in the top 20% of the population and no >25% deviation in any of these two groups.
- Alert Condition: A deviation of >25% in the top 5% or 10% of the population



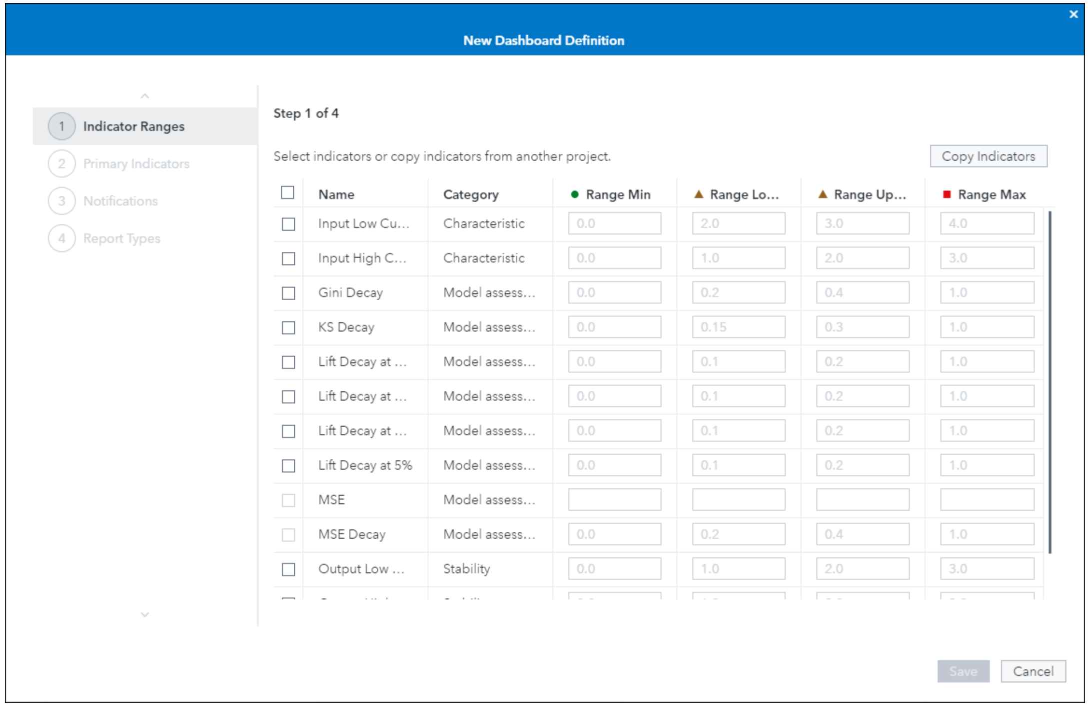
Area under Curve (AUC)

- AUC (Area Under the Curve): Measures the models ability/probability as a correct classifier of events.
- **KPI:**
  - No alert/warning: A deviation of <10% of AUC decay is considered as non-significant.
  - Warning Condition: A deviation of [10%, 25%] of AUC decay should be an indication of model investigation and possible recalibrating.
  - Alert Condition: A deviation of >25% of the AUC decay should commence the model retirement process.



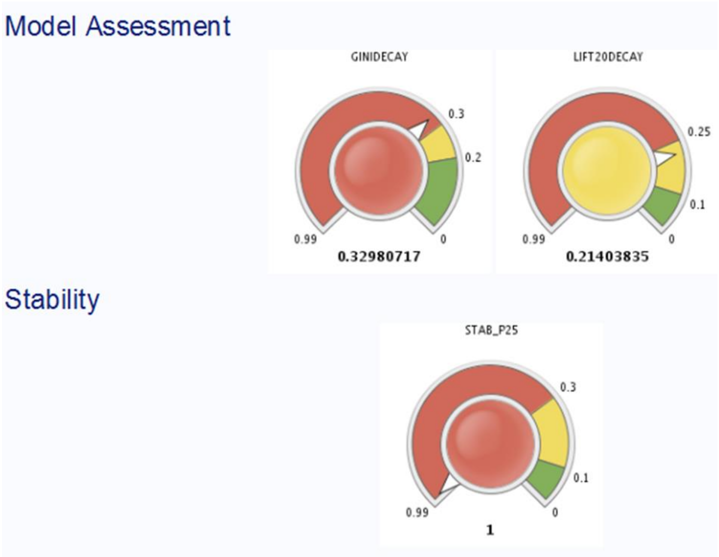
In Sas Model Manager 14.3 (sas 9.4) you can create alerts and a dashboard

- **Create alerts**



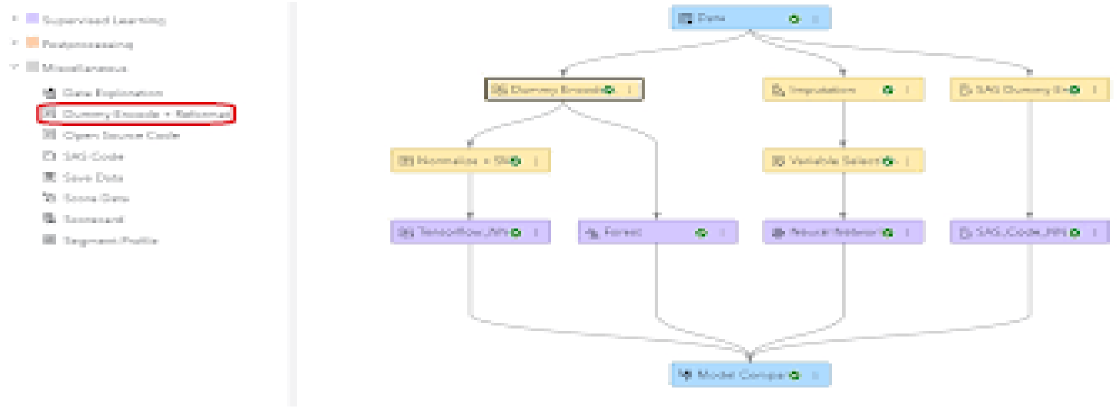
Name	Category	Range Min	Range Lo...	Range Up...	Range Max
Input Low Cu...	Characteristic	0.0	2.0	3.0	4.0
Input High C...	Characteristic	0.0	1.0	2.0	3.0
Gini Decay	Model assess...	0.0	0.2	0.4	1.0
KS Decay	Model assess...	0.0	0.15	0.3	1.0
Lift Decay at ...	Model assess...	0.0	0.1	0.2	1.0
Lift Decay at ...	Model assess...	0.0	0.1	0.2	1.0
Lift Decay at ...	Model assess...	0.0	0.1	0.2	1.0
Lift Decay at 5%	Model assess...	0.0	0.1	0.2	1.0
MSE	Model assess...				
MSE Decay	Model assess...	0.0	0.2	0.4	1.0
Output Low ...	Stability	0.0	1.0	2.0	3.0

- **Dashboard**
- Overview of the models



What to do when the model performance analysis creates an alert

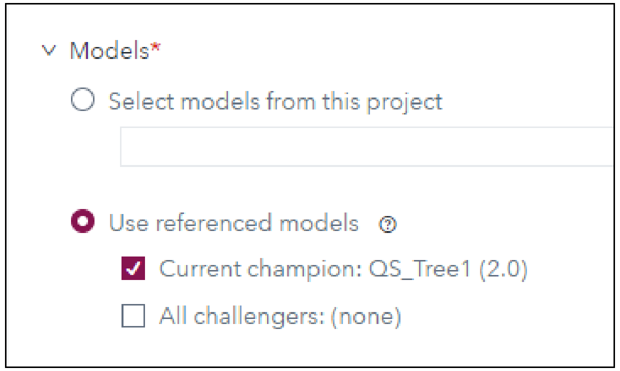
- **Retrain Model**
- It will use the same pipeline of SAS Model Studio or Sas Enterprise Miner.
  - No new features will be used
  - Fast creation of a new model (one button does it all)



- **Rebuild**
- You have to develop a new pipeline to create a new model.
- You have to spend more time.

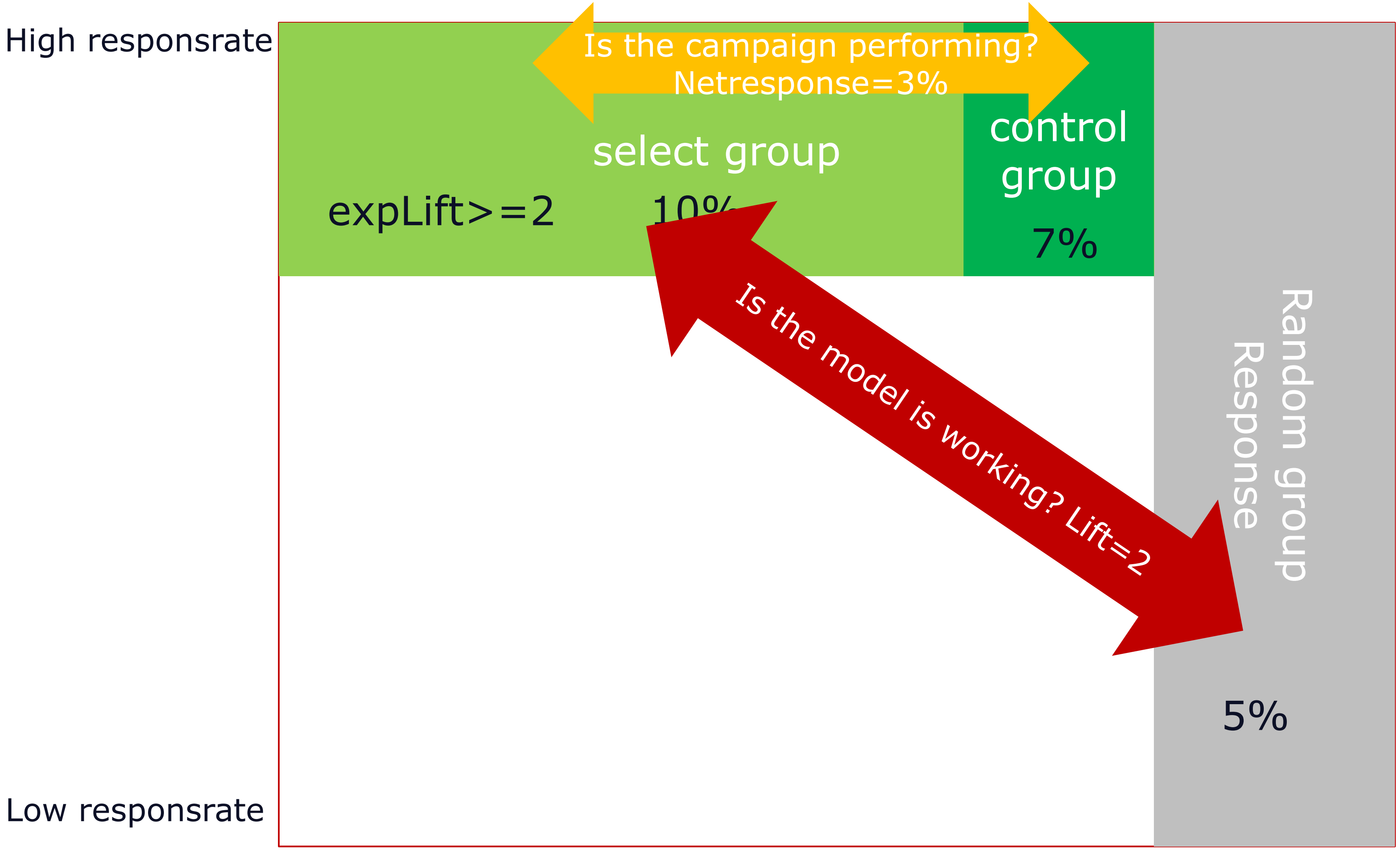
Lessons learned from monitoring models on Telenor data

- Most of the time we see that the more advanced model, like a Gradient Boosting, Random Forrest, performed better on the train, validation and test set, then Logistic Regression which uses tree based binning. The Logistic Regression performes better after some months.
- Import not only the champion model, but also the challenger(easier) model into Model Manager. So you can monitor both models over time.
- Gradient Boosting and Random Forrest works most of the time with more features and the models are more sensitive for small changes over time.



Model monitoring in action

A way of using and measuring an analytical model in an outbound campaign (sms, Email, mail).



Thank you

Hans de Wit, Telenor Mobile Norway, +47 48 29 1399