Dealing with Common Issues in Analytics

Nicolette L. Ige

Master of science, Analytics Candidate

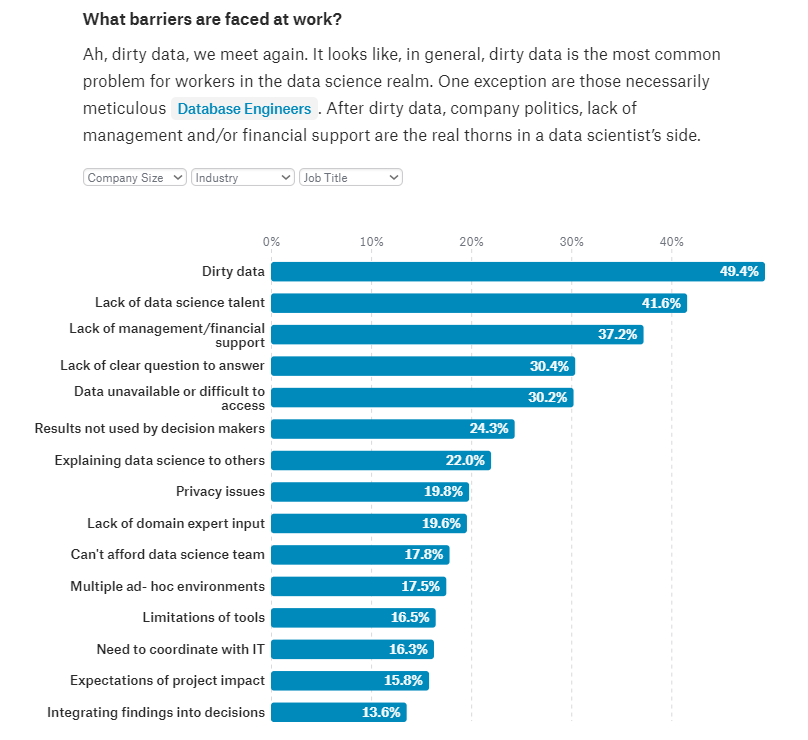
Louisiana state university

Version [0.0]

February 28, 2018

# Purpose

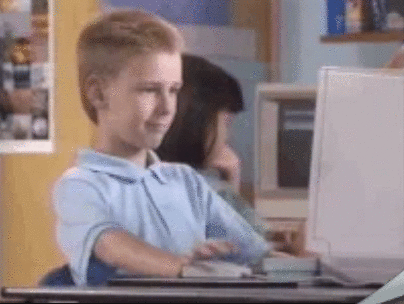
This lecture presents some of the situations and issues that MSA graduates are likely to encounter once they graduate. Most topics included are presented for the purpose of bringing awareness of things to be mindful of, promoting discussion of best practices and giving *(hopefully)* helpful tips.



<https://www.kaggle.com/surveys/2017>

# Data Quality





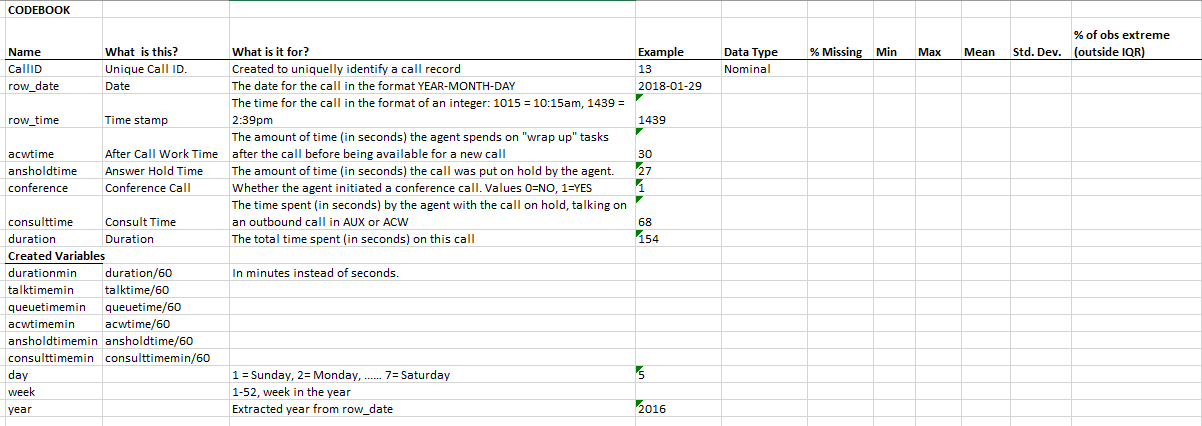
## Dirty Data

PRO TIP: Don’t assume the data is clean, and don’t assume you or someone else fully did it correctly, data quality and steps to correct quality should be reviewed by more than 1 person if possible.

**Read this** -> <https://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf>

## Create a codebook and data analysis plan

Example Codebook: Not just data dictionary but also % missing, min, max, mean, % of obs extreme outside of IQR, all possible values, if anything needs to be fixed.



Analysis plan should include the steps you want to take in data evaluation and analysis.

Evaluation questions and variables used for each analysis. Specific analyses/tests/graphics for each evaluation question.

## Comment your code and use a project diary

Write clean code- and document everything- you will not be able to remember all the steps you went through and everything you did and you will need to update boss/coworkers/client. Becoming good at this will be worth it. You will thank yourself and others will thank you as well. Practice makes perfect.

<https://www.slideshare.net/jamorrow/brief-introduction-to-the-12-steps-of-evaluagio>

## Do not fail to conduct frequency analysis of all variables

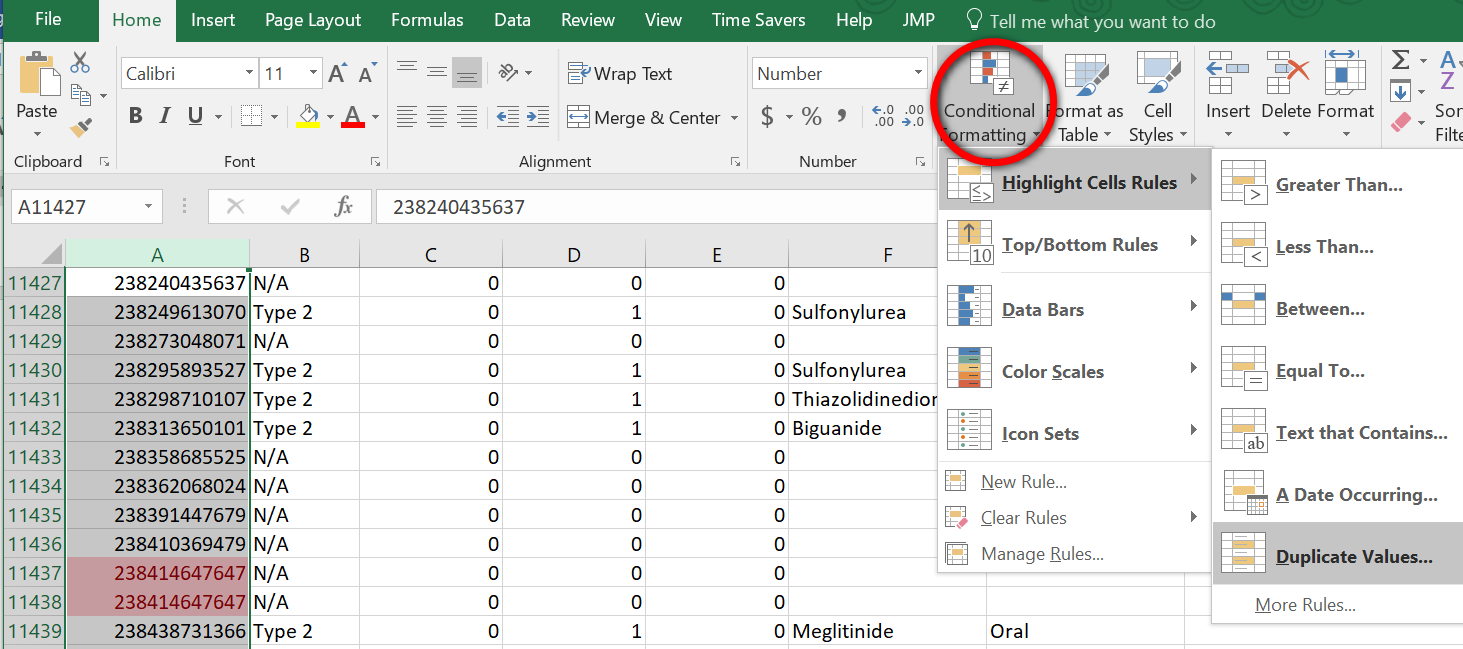
Can identify rare groups, or distributions that you might need to modify/transform.

Do this before and after removing or imputing outliers or modifying variables.

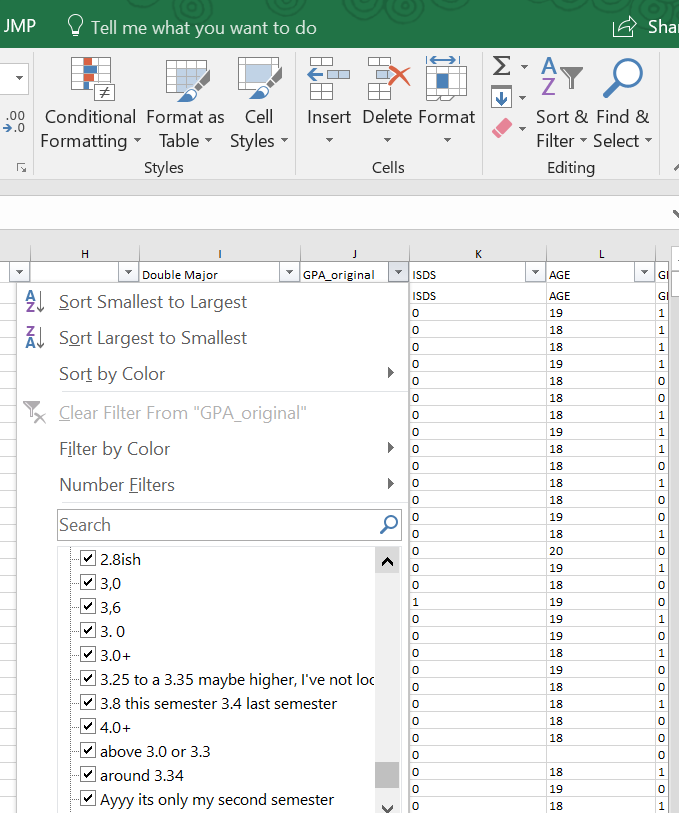
## compare/check values for inconsistencies or implausible values

Checking for duplicate records: Can do in Excel with Conditional Formatting. Or you can query, count how many distinct ID’s you have (is it less than the number of records you have?)

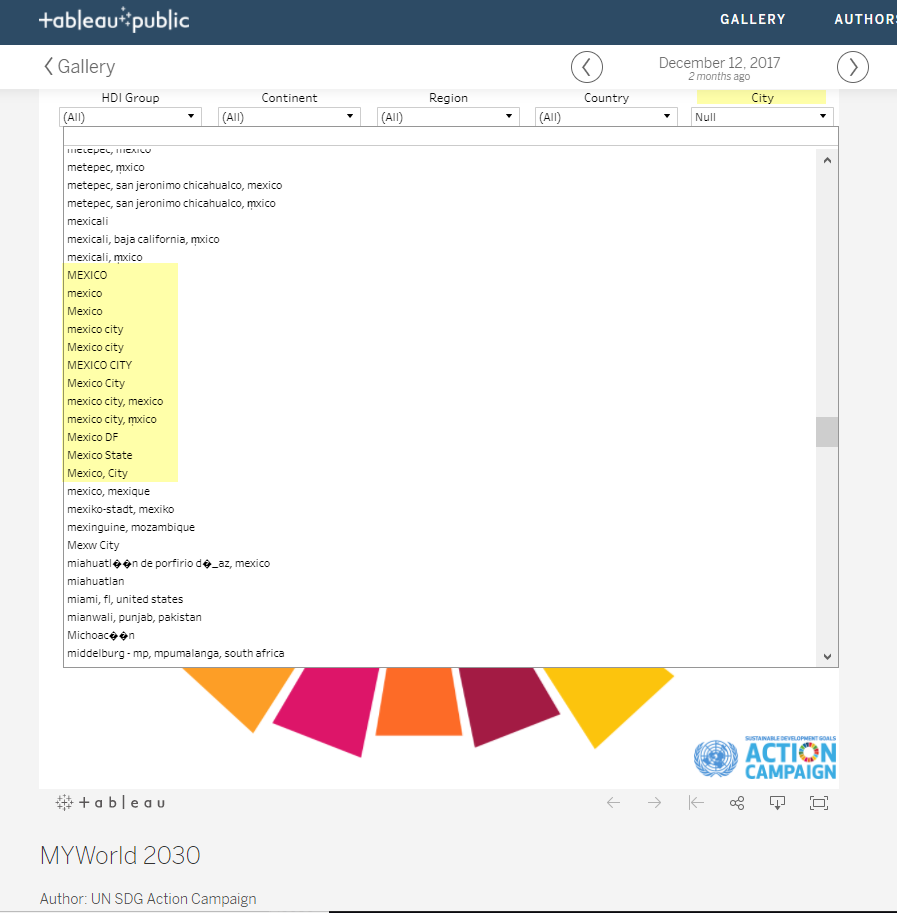
Some cases- Are you supposed to have duplicates and you don’t? Did the business mask the ID’s correctly? Example on call log. A transferred call should have same call ID, since it is the same customer and call.



Dealing with fields where values are typed in rather than selected.



Dealing with text fields: make all upper case or fix case- use regexp. Don’t be that person to not clean the data. This option of choosing the city is pointless because of not cleaning.



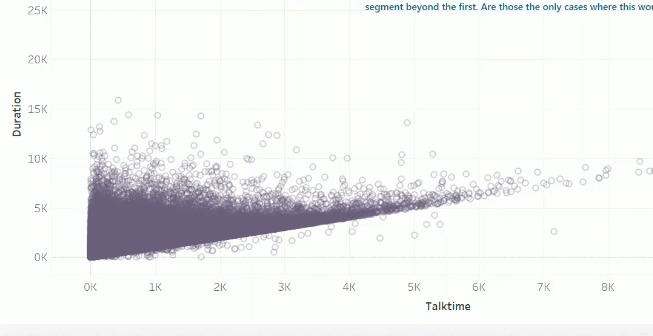
Inconsistencies: if you have column indicating if someone is married or not, filter to only those married, then can view ages to ensure there are no records that say someone is under a certain age. (ps. remember context of data when cleaning, in other cultures people can and do get married sometimes at very young ages)

## Using Data Visualization

PRO TIP: be careful to make sure visualization is showing what you think it is. Always use low opacity aka: high transparency on scatter plots if software has the option.

Key things to remember when using Tableau: Understanding Measure vs Dimension, and Discrete vs Continuous is EVERYTHING- most of your problems are likely due to those differences.

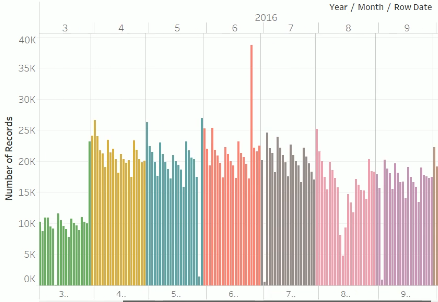
Sometimes spending 10 minutes exploring visually will help you come up with questions to help you understand the data better.



Example: We thought duration of a call would always be > or = to talk time,

but a few observations have it shorter?

This is data logged automatically in the system, how can it be?

Sometimes you will be asked to identify patterns or explain what is different about 2 groups.

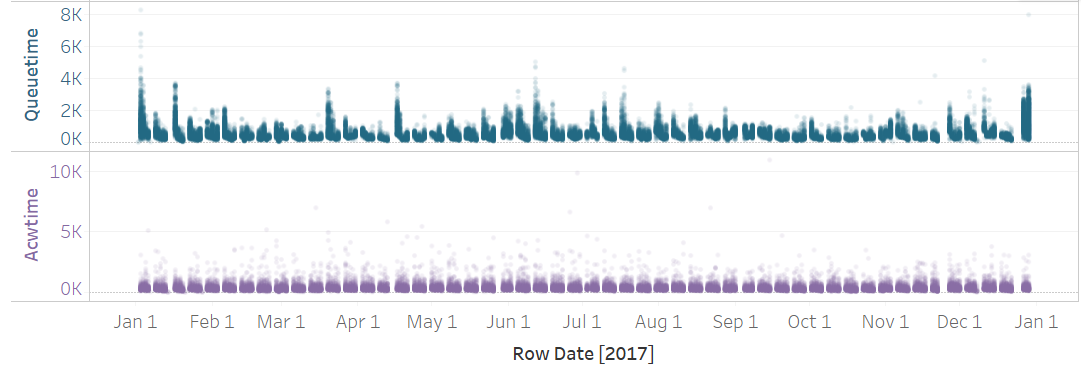
No predictive model is needed to identify *some* patterns.

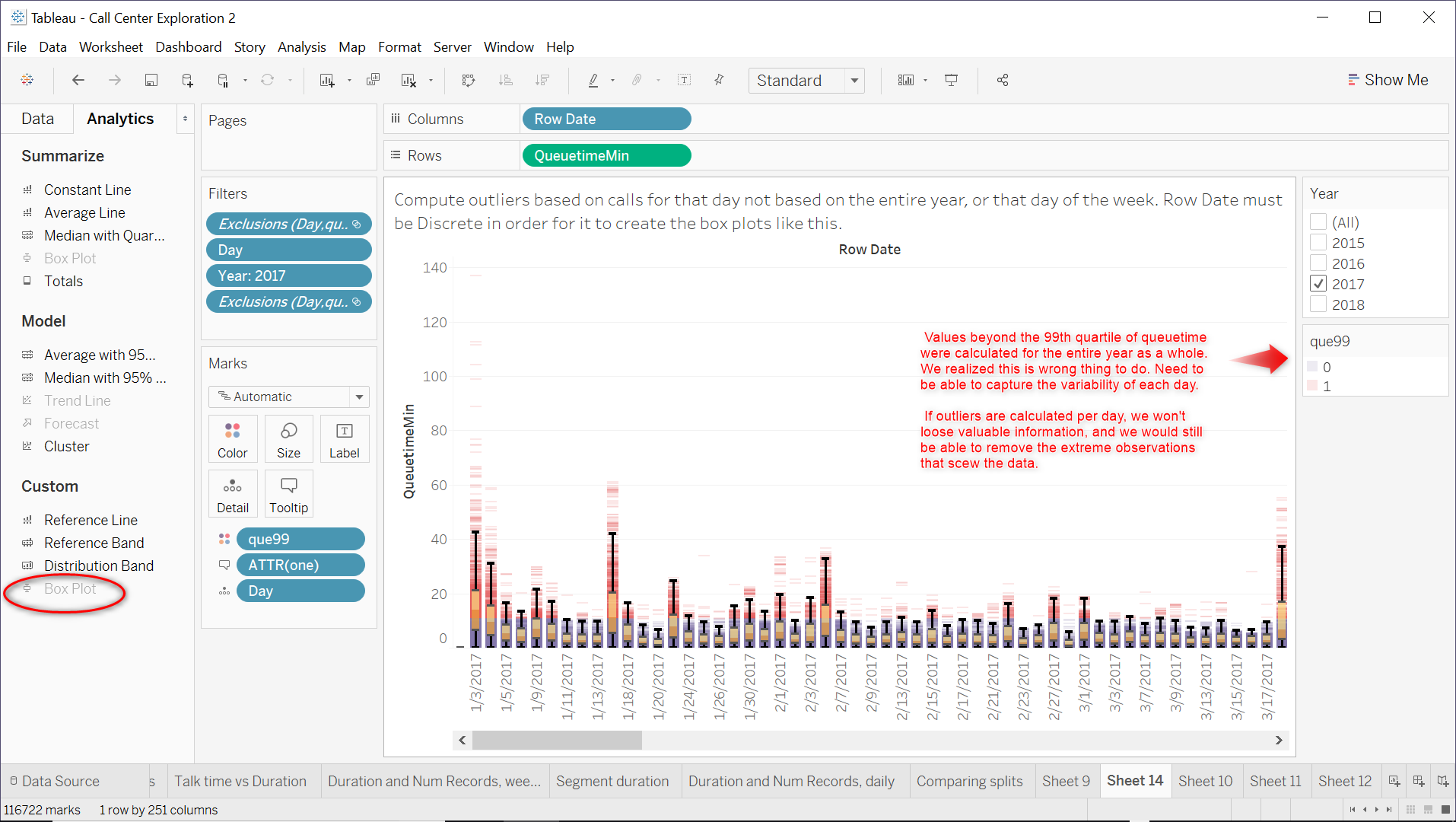
Some things can easily be shown by pie charts, bar graphs and scatterplots.

Don’t delete outliers before investigating, valuable information or valuable discoveries can be lost if you are not careful.

We are looking at call data throughout the year, business is not interested in patterns throughout the day but more so trends or patterns in year, and outside factors like extreme weather and how that effects the call center. All continuous variables have distributions that are extremely skewed.

For all variables except queue time, they seem to be random. We knew the outliers were mostly on Mondays and Tuesdays, but even calculating a cut off based on day of week was losing the valuable information of which specific days had the highest queue times. Maybe its not just its being a Monday and Tuesday, maybe it is the Monday/Tuesday after a holiday?





## Being an “interviewer”

As analytics people- we may be the first person to get data and analyze it, especially data from an organization or company that less sophisticated. You may be the person that needs to play interviewer/investigator/data auditor to make sure you are analyzing quality data for the purpose you are using it for. You may have to ask for the database schema to understand the data, you may even find yourself in a position wanting to discuss with the DBA to make changes in the schema to improve ability to analyze the data.

Also, being a good interviewer will help you if you run into “not having clear questions to answers”.

# predictive modelling

## variable Modification

Should we use a variable as is? If in logistic or linear regression model, think about why it matters so much. An increase in age by 1 unit is totally different than having split into groups and treated as nominal. Decision tree, don’t worry about it so much. It will find the but splitting point.

**Interval Variables:**

Binning transformations:

***Quantile: Data divided into groups with approx. the same frequency in groups.***

Bucket: Divided into even spaced intervals. (ex. 15-30, 31-46, ..) (the spacing of the intervals are based on the difference bw the max and min values)

**Optimal Binning: Option in SAS to bin the data in whatever way that maximizes the relationship with the target**

Best: SAS has this option, it’s GREAT, other tools may have something similar.

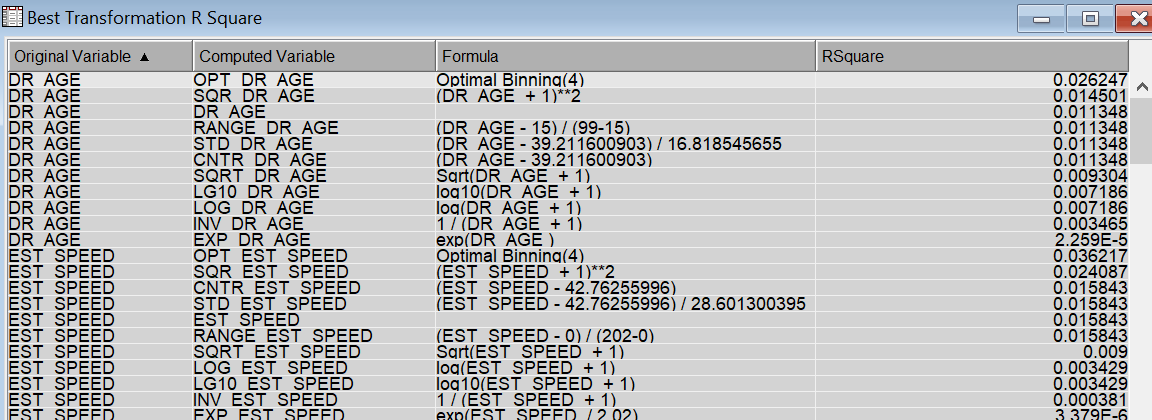
**Categorical Variables:**

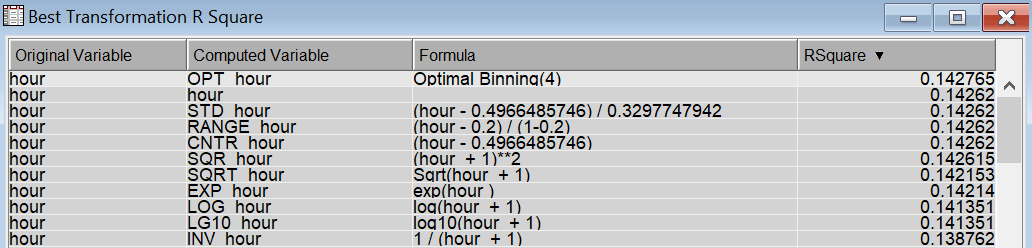
Group rare levels together

Create dummy variables

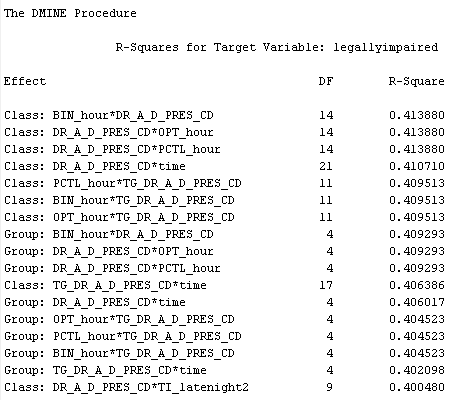
PRO-TIP: transform variables in different ways, merge to have a data set with multiple version of the variables and then do preliminary variable selection.

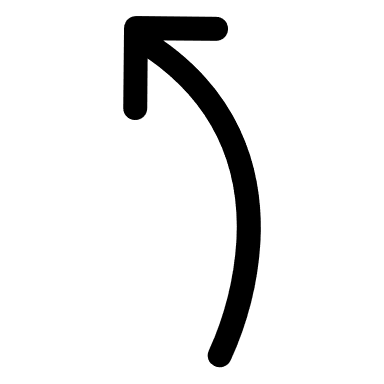
Or just use BEST and it tries them all and will show you results of all.





Variable Selection Node results:





Don’t be afraid of two-factor interaction terms!

Compared optimally binned hour variable and my binary variable (latenight).

R-squares is no diff and no difference in comparison in logistic regression model...

## Variables disguised as inputs

Ask yourself: would I know this attribute for a completely new record we would need to score. Beware of variables that are actually “after the fact”. Example, if a model is to predict how long a phone call is going to be at the beginning of the phone call; you are not going to know at the beginning of the call how many times the call was transferred. Number of times transferred might come up as the best predictor in variable selection, but it cannot be used as a predictor because of the nature of how the model is used and the business case.

Another example: predicting a college drop-out- question about full time status was asked after some had already dropped out of college, so all the records that had missing answer for that questions were all drop outs (also would catch if assessing randomness of outliers correctly and thoroughly).

PRO-TIP: If you do variable selection and variables you know are good predictors don’t show up, check for this mistake.

## Question completeness, avoid selection bias in your model

Not just missing values in records- question if you have all the real-world information, or data to represent every meaningful state, Wang 1996, “The extent to which data are of sufficient breadth, depth, and scope for the task at hand.”

Examples:

Dark skin not recognized by hand drying or soap machines.

Image recognition models predicting if a picture has bed in it, every picture they gave the model to train also had curtains in it. When the model was presented with a picture of a room with no bed and the model thought there was a bed, they realized they actually taught the model to recognize curtains. (heard on Ted Talk)

## Is your sample representative of your population? Do you need to use observation weights?

For most modeling projects, you often have to make do with whatever data is available, and most of the time what is available is a sample, not the entire target population, and many times it is not a “scientific sample”. You must verify that the **sample is a good representation of the target population** in which the model is created for.

How do you check for this? Compare distributions of some key variables in the sample and target population. Maybe age and income are key characteristics in your model, use those 2.

Are they different? If so, can calculate and use observation weights to **correct for any bias**:

Example with age and income: Split age and income into groups. Age: 3 groups and Income 4 groups. The target population has Nij people in the ith age group and the jth income group. Similarly, the sample has nij people for each combination of the groups. So, for someone in the ith age group and jth income group in the sample data- the appropriate observation weight is (Nij/N)/(nij/n). **Very easy**: Say you have a marketing campaign for bathing suit company: lets say target population has 30,000/70,000= 42.8% of their target population who are bw 25-30 and make 50-80k a year. But the sample only has 30 % of their observations falling into that IJ group combo. So a person in that combo would get obs weight of 42.8/30=1.426.

In SAS Enterprise Miner, assign this variable the **role of Frequency** and it considers the weights in estimating the model.

*Think about other cases: Ames Housing, how many houses in your sample are really small and old, how many homes in Ames Iowa are currently really small and old without any renovation, if not that many, you would want those obs to influence model less than those observations that are more likely to arise in population when applying model.*

## Are you trying to model a rare event? Should you Oversample?

Is there only a small portion (*subjective)* of your data that has the target event? You can help the “machine learn” by basically using less observations that are non-events. All records with an event are kept (they are over-represented) and only a randomly selected fraction of the non-events are kept. Other names for this is case-control sampling or biased sampling.

This introduces bias, but it is corrected by adjusting the predicted probabilities. In SAS Enterprise Miner- it is taken care of via the decision profile.

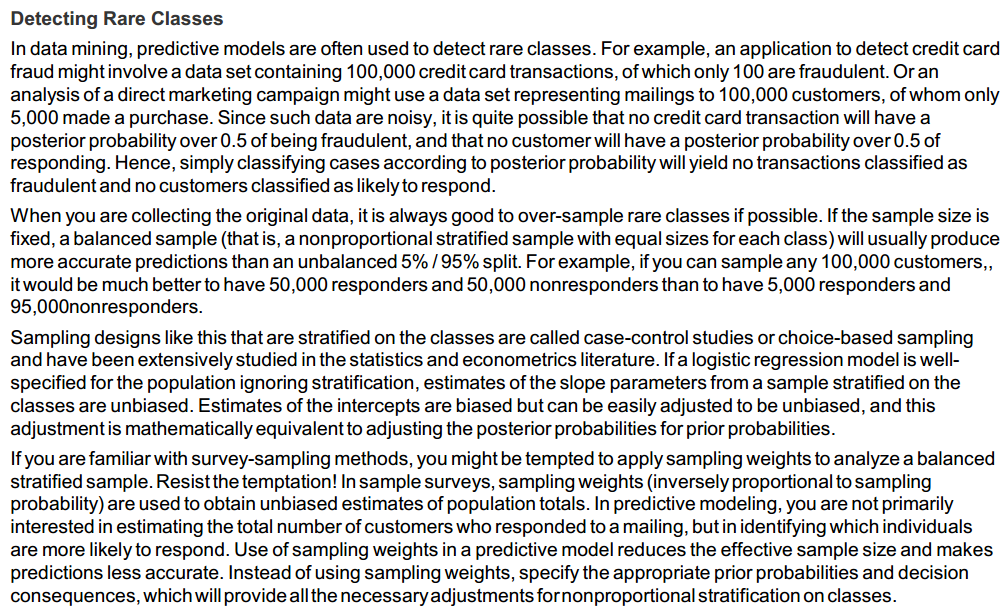
See Pages 13-18 of SAS Enterprise Miner Predictive Modeling Documentation for details regarding Prior Probabilities. Also explains how to retroactively adjust predicted probabilities of already scored data. Few main points below:

**Prior Probabilities:**

|  |  |  |
| --- | --- | --- |
| Do affect | Do not affect | In |
| model selection | estimating parameters | Regression node |
| early stopping | learning weights | Neural Network node |
| pruning | growing trees | Decision Tree node |
| computing total and avg profit or loss |  | A model with profit/loss matrix |

PS: Even if you don’t explicitly over sample the data you have been given or have access to, think about what is to be seen in reality if/when model is applied- SAS Enterprise Miner assumes that the validation and test sets are representative of operational data, so your test and validation stats and may not provide valid estimates of generalization if you don’t specify priors that are to be expected when applied.

What SAS Enterprise Miner Predictive Modeling Documentation says about **survey sampling weights** in the detecting rare events section on page 28.

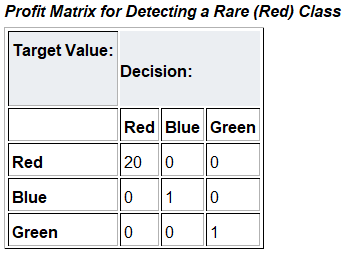


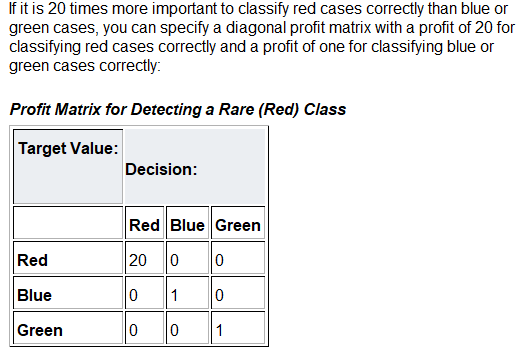
See page 31 for Neural Network with predicting 3 classes, 1 being a rare event, with Neural Nets, very cool.

## decision weights and profit/loss charts

Not only are decision weights or profit/loss matrices useful for calculating total profit or loss on your model results and can be used to select the best model; they can also be used to make tentative decisions. Profit/loss charts are much more useful than you realize can help you produce a model that is altered specifically for the business application.

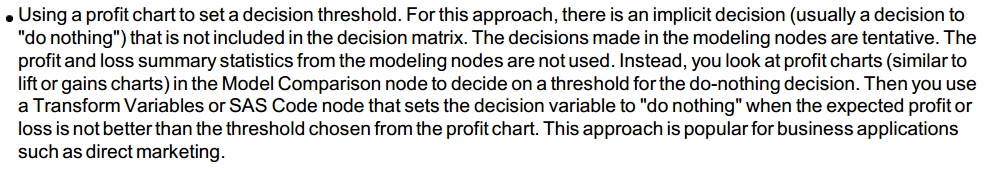
PRO-TIP: you don’t actually need to have a monetary profit or loss to use it, any target specific consequences that have practical interpretation do the same thing. You could translate it to *this* decision is 2 times worse *than that decision*.

Example from SAS:



Some business cases it may be preferred slightly low specificity(0’s) even if it means having high sensitivity(1’s). Misclassifying a truly fraudulent case as not fraudulent is worse than flagging a non-fraudulent case as being possibly fraudulent. If you have practical interpretation and not actual cost, then values computed regarding return on investment aren’t useful but that’s okay.

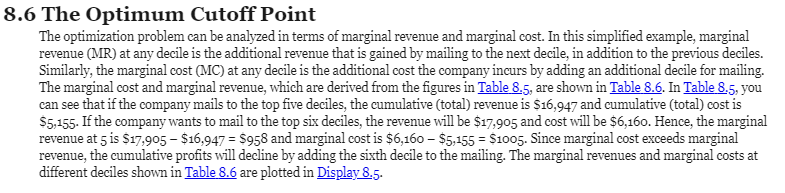
Decision Threshold:

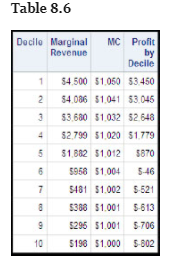
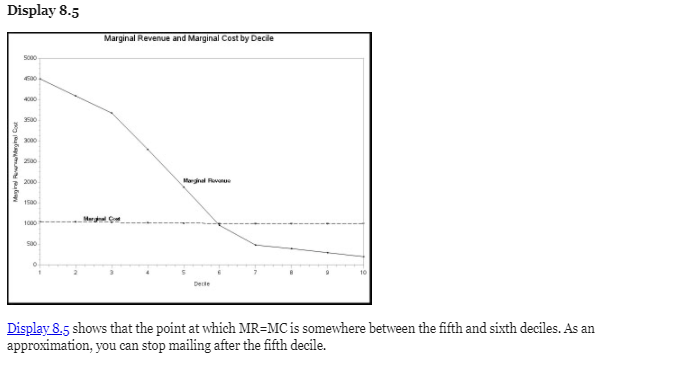


Use case: marketing campaign: Sometimes making no decision at all is best, maybe you should wait to decide on a record: Does doing something for an observation produce a profit above a certain threshold, no, then don’t do anything, if yes, then do it.

## Choosing an optimal cutoff point

First: Know cumulative lift, gain, % captured response





*Predictive Modeling with SAS Enterprise Miner: Practical solutions for Business Applications, Third Edition, Kattamuri S. Sarma*

## worried your model wont be used because decision makers won’t trust it? IS IT GENERALIZABLE?

Fully vet your model. Assess it as if someone else did the work. Be honest with the business decision makers about what makes you less or more confident in the model.

Terminology:

Underfitting/Undersmoothing- model is not sufficiently complex can fail to detect fully the signal in complex data.

Overfitting/Oversmoothing - model that is too complex might fit the noise, not just the signal

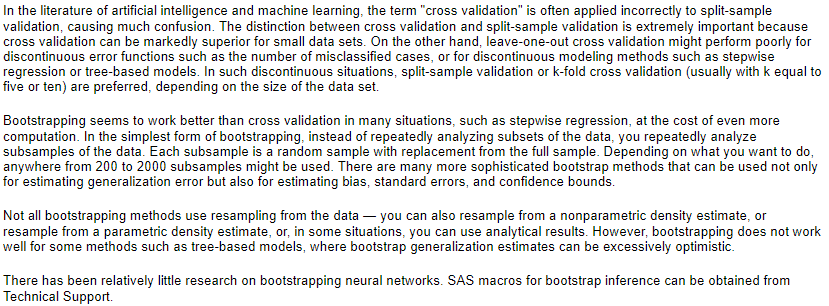
Good model strikes balance between simplicity and complexity. Bias- variance trade off. Too simple- introduce bias in predictions. Too complex- overfit and higher variance.

SAS documentation suggests:

For small data sets- cross validation.

For large data sets- Use a third/test sample.

What does it say about bootstrap, cross validation, and k-fold?



http://support.sas.com/documentation/cdl/en/emxndg/65358/HTML/default/viewer.htm#n002icfvzhfd57n1c2pmqx1ygovl.htm