**Statistics 101**

**Outlier Detection**

* You don’t get rid of all outliers, you must see which ones are influential
  + If there is reason to believe that these cases arise from a process different from that of the rest of the data, delete.
  + Show one case has influential points and one doesn’t, report to emphasize the impact of those points
* Use Weisberg test (w/ studentized residuals) will detect y outliers
* Hat elements or Mahalanobis distances will detect outliers in the x values
* Visualization of outliers
  + Boxplots
  + Scatterplots
  + Histograms
  + Normality plots (QQ Plots)
* Univariate way to detect outliers
  + Z-scores>2.5 outlier
  + In Excel can use quartiles and IQR
    - Quartile 1 = quartile(A$2:A$16,1)
    - Quartile 3= same as above except instead of 1 put 3
    - IQR= Q3-Q1
    - Upperbound= Q3+(1.5\*IQR) 1.5 for weak outlier and use 3 for strong outliers
    - Lowerbound= Q1-(1.5\*IQR)
* Other ways to detect outliers
  + Large residuals
* To determine if outliers are influential you can use Cook’s D, influential if Cook’s D>1 or 4/n

**Linear Regression**

* Assumptions that must be met
  + The sample is random ( x can be nonrandom provided that y’s are independent with identical conditional distributions)
  + The regression of y on x is linear
  + Homoscedasticity of the errors (constant variance)
  + Normality of the error distribution
    - Y variable doesn’t have to be normally distributed
      * Errors around prediction y must be normal which one can check with a normal plot of residuals
      * Errors must have constant variance which can check with predicted by residual plot(should have no pattern)
* How to determine if transformation is needed
  + Look at scatterplots of each regressors vs y
  + Weighted least squares for ex, 1/ave, 1/sqrt(ave) etc. good for small data sets, the ability to handle regression situations in which the data pts are of varying quality, however not so good with outliers
* Diagnostics other than ones described in outliers section
  + Once multiple linear regression is done check for influential observations
  + Multicollinearity issue
    - Use VIF for detecting multicollinearity issue (1/1-R^2)
    - Eigenvalues that are small represent a high but not critical source of multicollinearity
  + Make a ratio between the plain and the adjusted R-squared measure and check if their difference exceeds 20%. If it does, it means that we have introduced some redundant variables inside our model specification.
  + When a condition number is over the score of 30, there's a clear signal that unstable results are rendering the result less reliable.
  + Variances are not homogenous if residuals aren’t spread all over, like if they are skewed
  + Want BIC small, AIC small, don’t want t-test<0.05, don’t want DFIT>1 or abs(DFITS)>2\*sqrt(k/n), don’t want Cook’s D>1 or 4/n, don’t want leverage>3\*k/n, don’t want VIF>10, don’t want abs(DFBETA)>2/sqrt(n), correlation>0.5 important
* Can use RFECV to show the optimal number of variables to use and show the given variable names

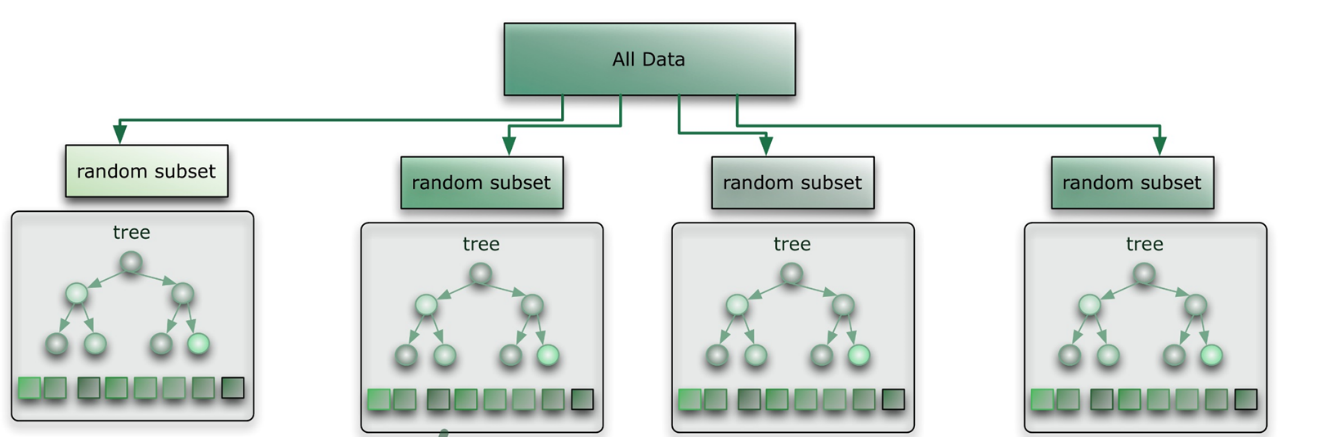
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **df** | **SS** | **MS** | **F** |
| Regression | K | SSr (want large) | MSr | MSr/MSres |
| Residual | n-p | SSres (want small) | Msres (want small) |  |
| Total | n-1 | SSt |  |  |

**Seasonality**

* **Autocorrelation** of a [random process](https://en.wikipedia.org/wiki/Random_process) describes the [correlation](https://en.wikipedia.org/wiki/Correlation) between values of the process at different times, as a function of the two times or of the time lag
* ARIMA (p,d,q)
  + p is the number of autoregressive terms
  + d is the number of nonseasonal differences needed for stationarity
  + q is the number of lagged forecast errors in the prediction equation (moving average term)
  + Rules on how to determine what the p,d,q terms should be: http://people.duke.edu/~rnau/arimrule.htm
  + Durbin Watson statistic: value near 2 indicates no first order serial correlation, positive correlation = DW below 2 and negative = DW above 3
  + If positive autocorrelation look at AR term not MA, if negative look at MA
  + Autocorrelation plot should be near 0 for randomness

**Machine Learning**

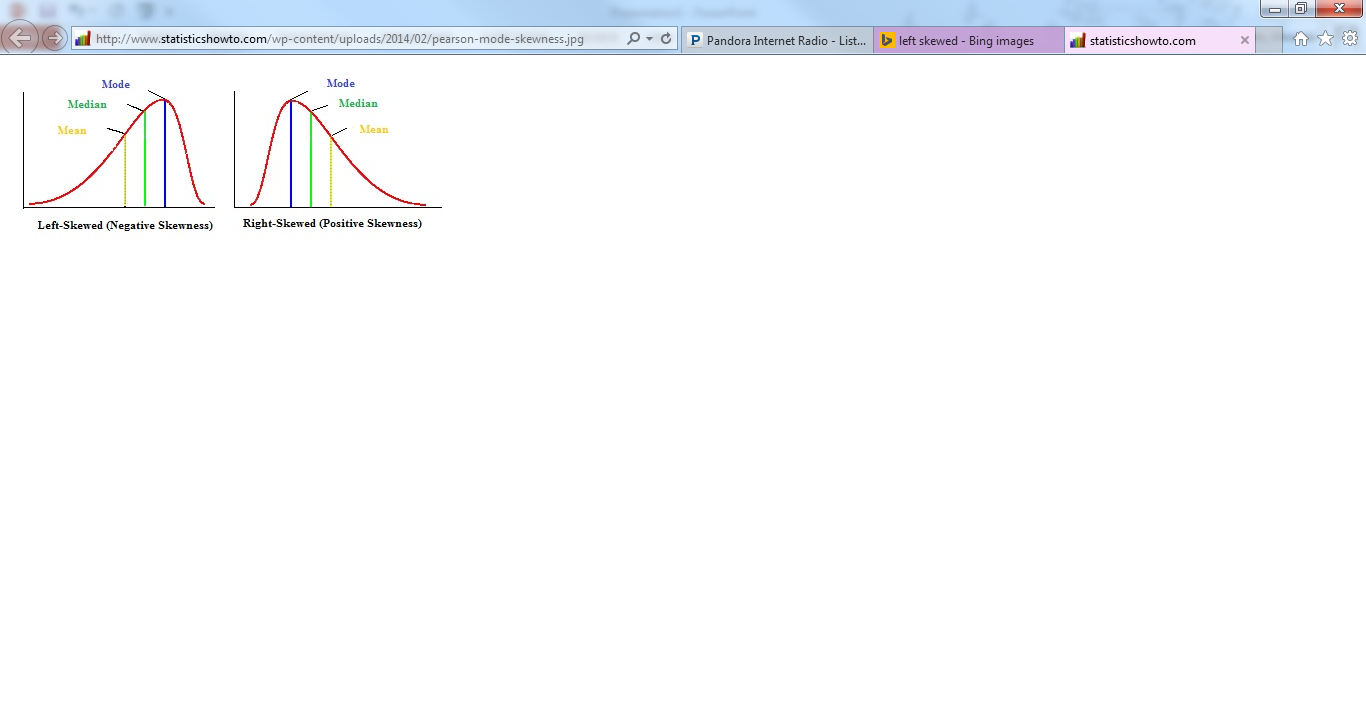
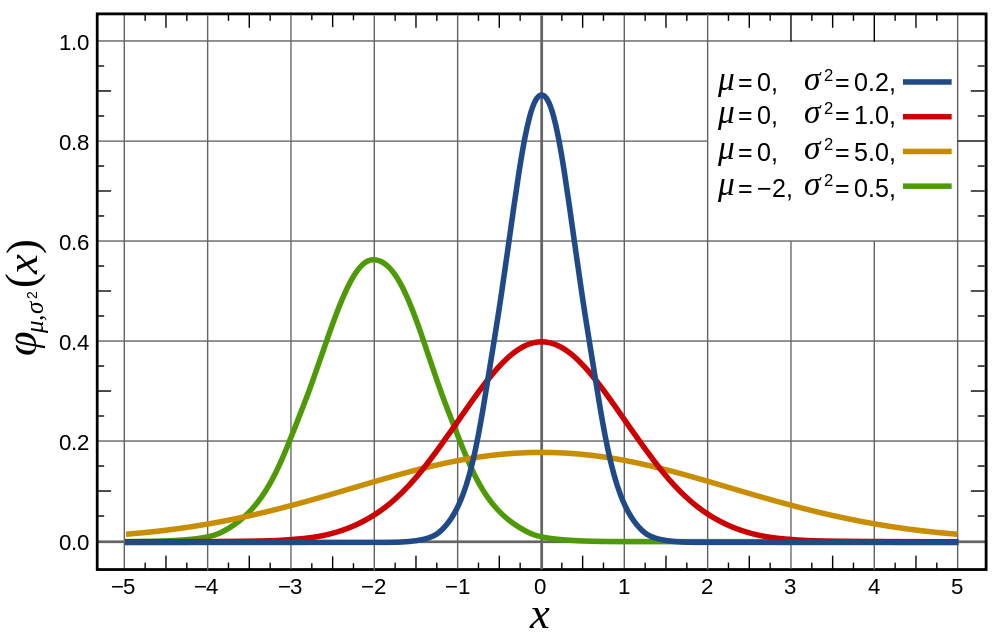
* Useful Vocab
  + **Bootstrapping** = random sampling with replacement
  + **Training set** is used to fit the models (usually 70%)
  + **Validation set** is used to estimate prediction error for model selection
  + **Test set** is used for assessment of the generalization error of the final chosen model. (about 30%)
* 5 main steps involved in training a machine learning algorithm
  + Selection of the features
  + Choosing a performance metric
  + Choosing a classifier and optimization algorithm
  + Evaluating the performance of the model
  + Tuning the algorithm
* Techniques/Models
  + Clustering great technique for structuring info and deriving meaningful relationships among data
  + Parametric vs Nonparametric Models
    - Using parametric models we estimate parameters from the training dataset to learn a function that can classify new data pts without requiring the original training set anymore
      * Perceptron, logistic regression, linear SVM
    - Nonparametric models can’t be characterized by a fixed set of parameters & the number of parameters grow with the training data
      * Decision tree, random forest, kernel SVM
* Random forest
  + Random forest regression is a machine learning technique, starts with “decision tree”
  + It’s an ensemble method so many weak learners (each individual tree) come together to become strong learner (averaging trees)
  + The random subset is about 70% of data, random sampling with replacement
  + Each split is a node and at each node small subset of m variables chosen at random
  + When the number of variables is large, but the fraction of relevant variables small, random forests are likely to perform poorly with small m.



<https://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/>

* Feature selection
  + If x variable had no variance wouldn’t want to include that feature
  + Using feature selection pick ones that have highest value, try to get x variables that aren’t as correlated to each other
* Diagnostics
  + Cross validation used see how well model will generalize its results on an independent dataset
  + ROC line is above benchmark line then the model is a better predictor than a random guess, want ROC closer to 1 better (can use in logistic regression for example)

**Normal Distribution Explanations**

<http://www.statisticshowto.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg> https://thecuriousastronomer.files.wordpress.com/2014/06/1000px-normal\_distribution\_pdf-svg.png

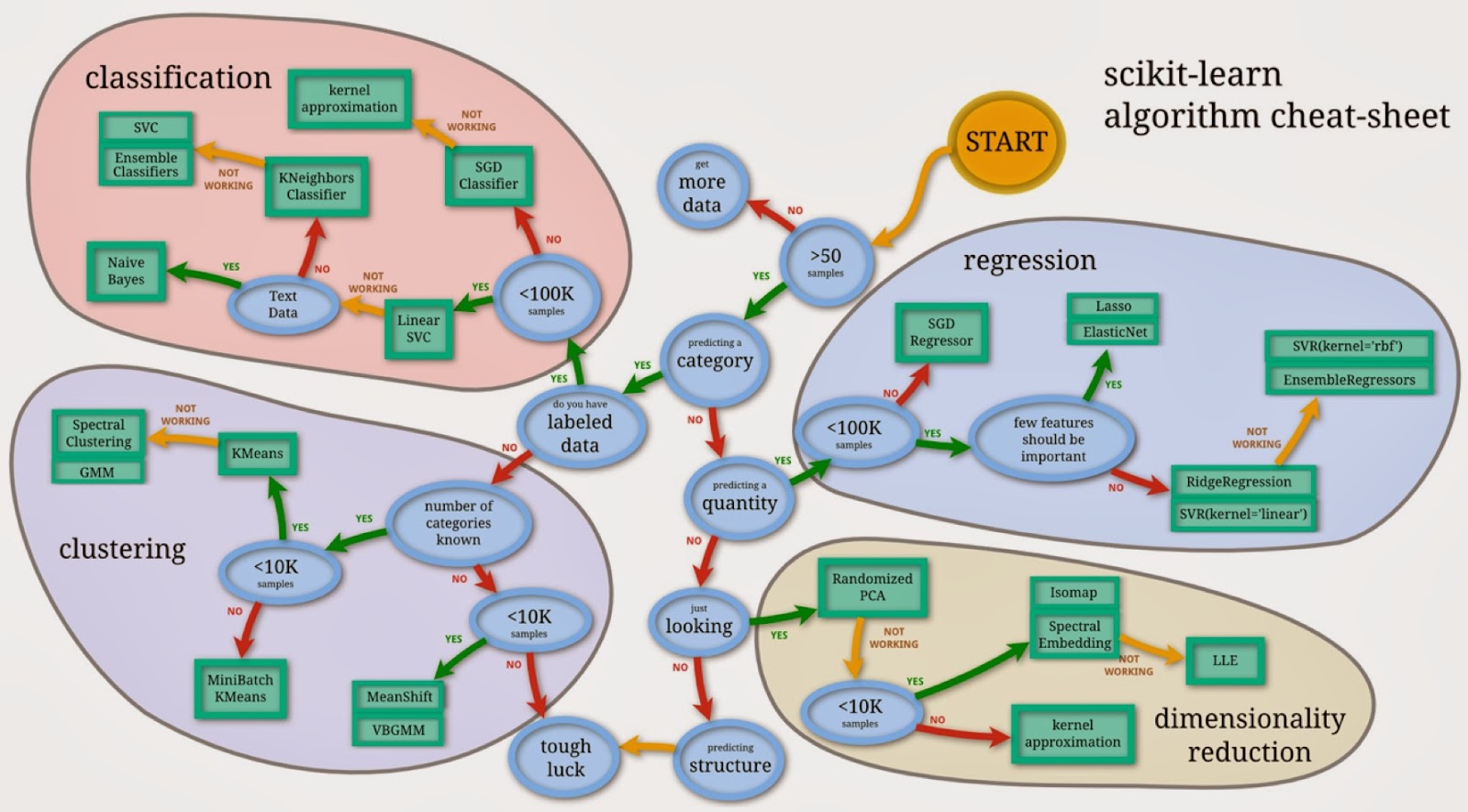
**Types of Errors**

* **Forecast error** = actual-projected
* **Percent Error** = abs[(projected-actual)/actual]\*100
* **Bias** : the difference between an estimator's expectations and the true value of the parameter being estimated (Wikipedia)
  + When finding lower confidence limit and upper confidence limit, if 0 is not included between then the bias is significant
* p-values > 0.05 indicate significance
* **Absolute value of error** = abs(error)
* **Squared error** = error^2
* **MSE** = average(all squared error)
* **TS** = sum(raw errors)/MAD
* **Mean absolute deviation (MAD**) = average(all absolute error)
* **Mean forecast error (Bias)** = the average error in the observations, large positive MFE means that the forecast is undershooting the actual observations
* **Mean absolute percentage error (MAPE)** = average (all absolute percentage error=actual-projected/actual)
  + MAPE would say 0.096 would say off by an average of 9.6%, do this for each of the regions and their products
  + SSE- square the errors and sum them
* **Exponential CAGR** = find ratio from 2nd prediction/1st prediction, so CAGR= ratio – 1

**Comparing Different Groups**

* Use Tukey’s HSD (the only down side is that in excel it only lets you compare 2 different groups at a time however Python you can do more)
  + Lakepepin excel used this analysis
* If assuming unequal sample variances use t-test
  + If t-stat < t critical two-tail test then groups not significantly different
* Chi-square = can be used to test whether the data observed differs significantly from the expected data

**Deciding on Which Model to Use**



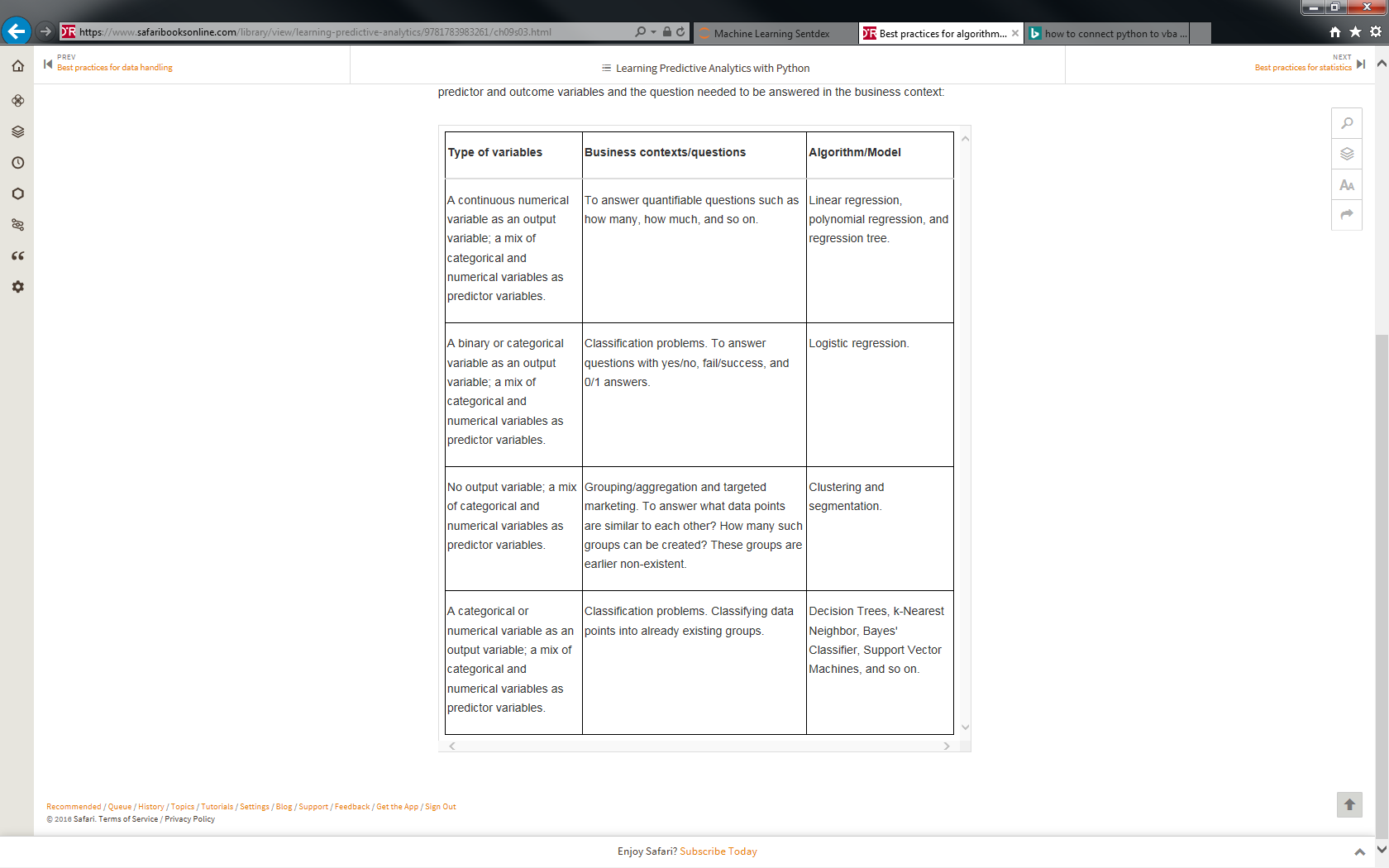


Table got from Learning Predictive Analytics with Python