ISA 444: Business Forecasting

27 - Advanced Topics

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Spring 2021

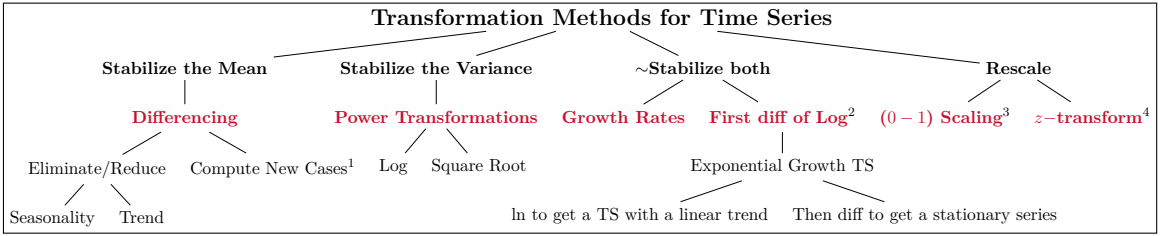
Outline

Recap of What we Have Covered This Semester

Main Learning Outcomes

- **Explain the purpose of forecasting in a business setting.**
- **Use the basic tools of forecasting including plots, summary measures, transformations, measures of forecast accuracy, and prediction intervals.**
- **Forecast a nonseasonal time series using simple exponential smoothing.**
- **Forecast a nonseasonal time series using linear exponential smoothing.**
- **Use decomposition methods and Holt-Winters smoothing methods to forecast a seasonal time series.**
- **Use ARIMA models to forecast a time series.**
- **Use simple and multiple linear regression models to forecast a time series.**

Recap: Guidelines for Transforming Time-Series Data



A classification of common transformation approaches for time series data.[5]

[1]The COVID19 package returns cumulative cases, i.e. a first difference $\longrightarrow$ new confirmed cases.

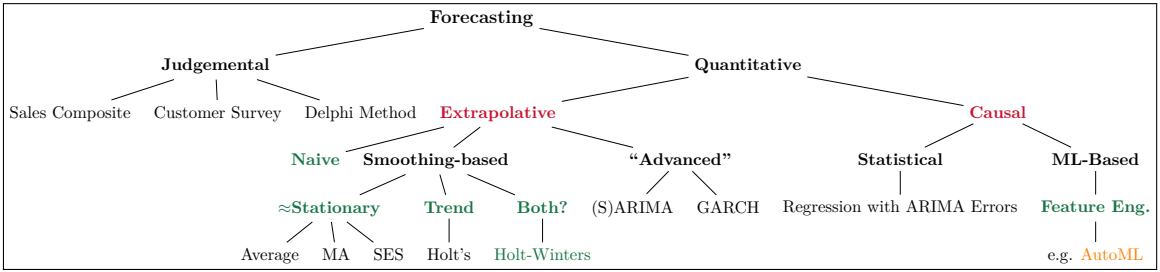
[2]First difference of LOG $\approxeq$ percentage change. This is almost exact if the percentage change is small, but for larger percentage changes, it may differ greatly (see here for more details).

[3]Rescaling of the data from the original range so that all values are within the range of 0 and 1. Mathematically, speaking this can be achieved by calculating $y_t = \frac{x_t - \min}{\max - \min}$.

[4]One can normalize a time-series by $z_t = \frac{x_t - \mu}{\sigma}$.

[5]My (incomplete) attempt to provide you with a taxonomy for time series data transformations.

Recap: A 10,000 Foot View of Forecasting Methods



A 10,000 foot view of forecasting techniques[6]

[6] An (incomplete) classification of forecasting techniques. Note that these focus on univariate time-series. Hence, they exclude popular approaches used in multivariate time series forecasting.

Learning Outcomes for Today's Class

Main Learning Outcomes

- **Examine the Use of autoML for TS Prediction.**

Outline

Definition

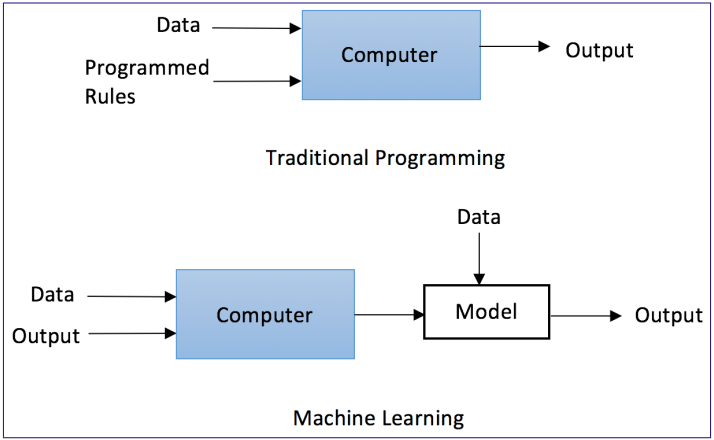
Mitchell (2006) has elgontely defined the scientific field of machine learning to be centered around answering the following question:

> *"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

In his view, machine learning is the study of algorithms that:

- **improve its performance** P
- **at task** T
- **following experience** E

A Paradigm Shift in Programming

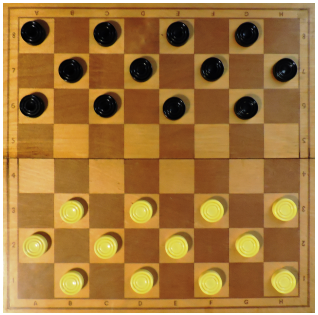


Source: Image is from Yuxi Lui (2019). Python Machine Learning by Example. Packt Publishers (click on image for more details).

Defining the Learning Task [1]

Improve on task T, with respect to performance metric P, based on experience E

- **T: Playing checkers**
- **P: Percentage of games won against an arbitrary opponent**
- **E: Playing practice games against itself**



Note: This idea in Samuel (1959) led to the popularization of machine learning.

Defining the Learning Task [2]

Improve on task T, with respect to performance metric P, based on experience E

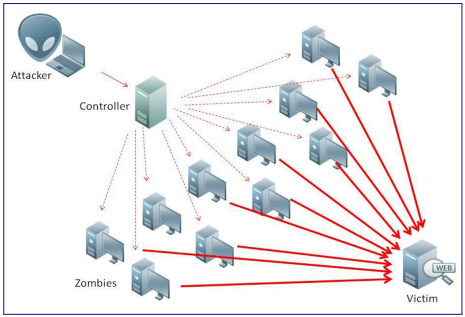
- **T: Autonomous driving using LADAR sensing**
- **P: Average distance traveled before human-judged error**
- **E: A sequence of images and steering commands recorded while observing a human driver**

Sources: Image by Dllu - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=64517567 and text from https://www.seas.upenn.edu/~cis519/fall2017/lectures/01_introduction.pdf

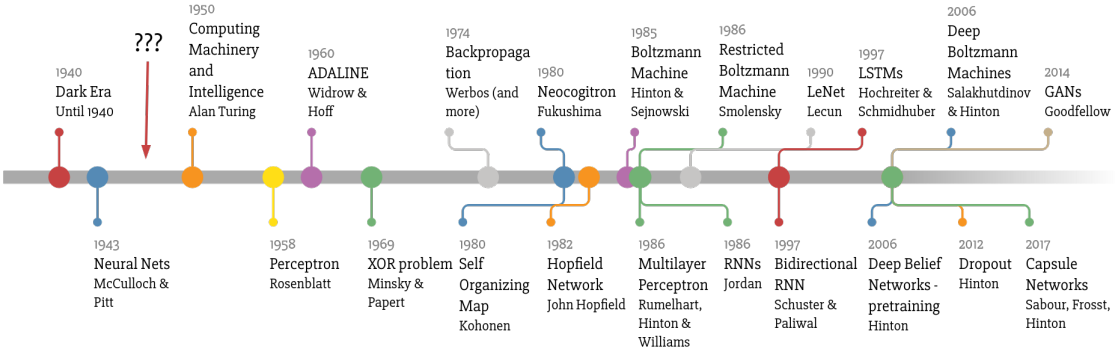
Defining the Learning Task [3]

Improve on task T, with respect to performance metric P, based on experience E

- **T: Categorizing network traffic as begnin or portmap (or another DDoS attack)**
- **P: Percentage of correctly categorized observations in each group**
- **E: Database of network traffic, with human given labels**

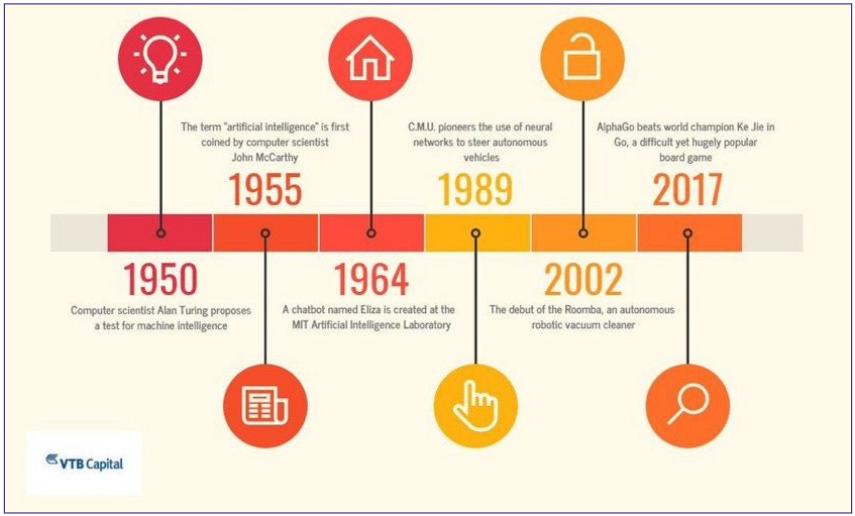


History of Machine Learning (click for source) [1]

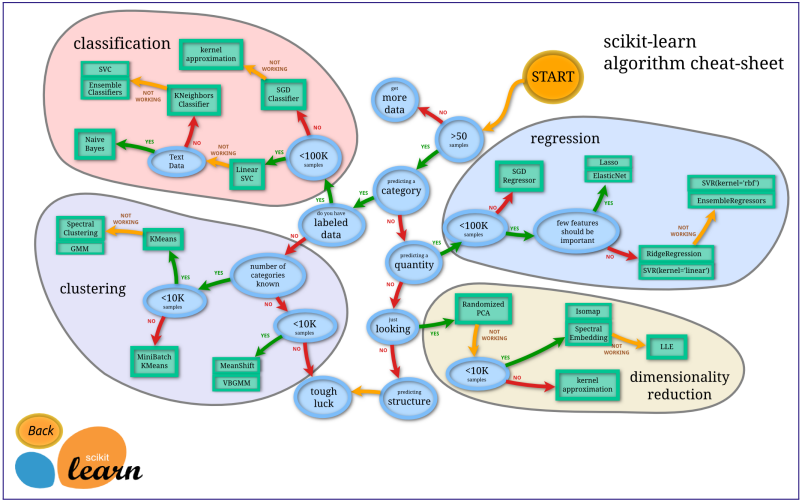


Made by Favio Vázquez

History of Machine Learning (click for source) [2]



From Task (T) to Model Type: Types of Learning



Source: SciKit-Learn "Choosing the Right Estimator" (2020).

Outline

autoML Installation Guidelines

Based on http://h2o-release.s3.amazonaws.com/h2o/rel-zipf/2/index.html:

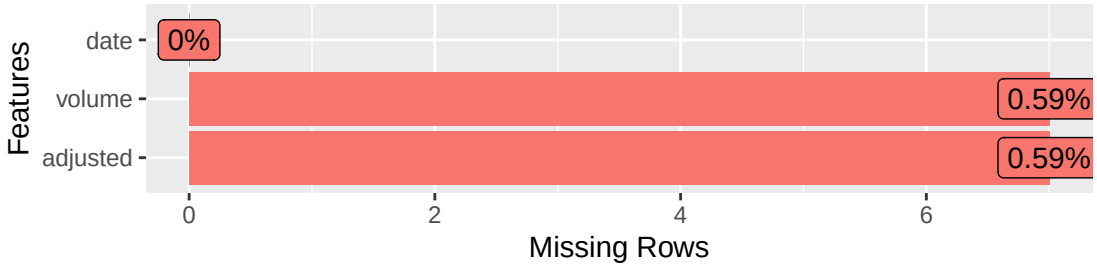
```
# The following two commands remove any previously installed H2O packages for
if ("package:h2o" %in% search()) { detach("package:h2o", unload=TRUE) }
if ("h2o" %in% rownames(installed.packages())) { remove.packages("h2o") }

# Next, we download packages that H2O depends on.
pkgs <- c("RCurl","jsonlite")
for (pkg in pkgs) {
  if (! (pkg %in% rownames(installed.packages()))) { install.packages(pkg) }
}

# Now we download, install and initialize the H2O package for R.
install.packages("h2o", type="source",
                 repos="http://h2o-release.s3.amazonaws.com/h2o/rel-zipf/2/R")
```

Let us Predict the Price of $ZIL: Missing Values

```
pacman::p_load(tidyverse, tidyquant, DataExplorer, lubridate, h2o, imputeTS)
zil = tidyquant::tq_get('ZIL-USD', from = '2018-01-01', to = Sys.Date() - 1) %
  select(date, adjusted, volume)
plot_missing(zil, theme_config = list(legend.position = c("none")) )
```



Let us Predict the Price of $ZIL: Feature Engineering [1]

```
cryptoFeatures = zil %>%
  mutate(adjusted = na_interpolation(adjusted),
         year = lubridate::year(date),
         quarter = quarter(date) %>% as.factor(),
         day = wday(date, label = T) %>% as.character() %>% as.factor,
         mday = mday(date),
         qday = qday(date),
         lagPrice = lag(adjusted),
         ma7 = rollmeanr(adjusted, k = 7, na.pad = T) %>% lag,
         ma30 = rollmeanr(adjusted, k = 30, na.pad = T) %>% lag,
         lagVolume = lag(volume)) %>%
  select(-volume)
```

Let us Predict the Price of $ZIL: Feature Engineering [2]

```
btc = tq_get('BTC-USD', from = '2018-01-01', to = Sys.Date() - 1) %>%
  select(date, adjusted, volume) %>%
  mutate(adjusted = na_interpolation(adjusted),
         volume = na_interpolation(volume),
         lagAdjBTC = lag(adjusted),
         lagVolBTC = lag(volume)) %>%
  select(date, lagAdjBTC, lagVolBTC)

cryptoFeatures = left_join(cryptoFeatures, btc, by = 'date') %>%
  na.omit()
```

Let us Predict the Price of $ZIL: Train, Validation & Test

```
trainData = cryptoFeatures %>% filter(year < 2020)
validData = cryptoFeatures %>% filter(year == 2020)
testData  = cryptoFeatures %>% filter(year == 2021)

# Set names for h2o
y = "adjusted"
x = setdiff(names(trainData), c(y, 'date') )
```

Let us Predict the Price of $ZIL: Fitting the h2o Model [1]

```
h2o.init() # Fire up h2o
h2o.no_progress() # Turn off progress bars

# Convert to H2OFrame objects
train_h2o = as.h2o(trainData)
valid_h2o = as.h2o(validData)
test_h2o  = as.h2o(testData)
```

Let us Predict the Price of $ZIL: Fitting the h2o Model [2]

```
automl_models_h2o = h2o.automl(
    x = x,
    y = y,
    training_frame = train_h2o,
    validation_frame = valid_h2o,
    leaderboard_frame = test_h2o,
    max_runtime_secs = 60,
    stopping_metric = "RMSE")
```

Let us Predict the Price of $ZIL: Best Model [1]

```
automl_leader = automl_models_h2o@leader
pred_h2o = h2o.predict(automl_leader, newdata = test_h2o)
h2o.performance(automl_leader, newdata = test_h2o)
```

```
## H2ORegressionMetrics: glm
##
## MSE:  0.000199702
## RMSE:  0.0141316
## MAE:  0.009613009
## RMSLE:  0.0121932
## Mean Residual Deviance :  0.000199702
## R^2 :  0.9260148
## Null Deviance :1.554524
## Null D.o.F. :121
## Residual Deviance :0.02436365
```

Let us Predict the Price of $ZIL: Best Model [2]

```
h2o.init() # Fire up h2o
# Investigate test error
error_tbl <- cryptoFeatures %>%
    filter(year == 2021) %>%
    add_column(pred = pred_h2o %>% as_tibble() %>% pull(predict)) %>%
    rename(actual = adjusted) %>%
    mutate(
        error     = actual - pred,
        error_pct = error / actual
        )
```

Let us Predict the Price of $ZIL: Best Model [3]

```
error_tbl
```

```
## # A tibble: 122 x 16
##    date       actual  year quarter day    mday  qday lagPrice    ma7    ma30
##    <date>      <dbl> <dbl> <fct>   <fct> <int> <dbl>    <dbl>  <dbl>  <dbl>
##  1 2021-01-01 0.0791  2021 1       Fri       1     1   0.0827 0.0799 0.0484
##  2 2021-01-02 0.0727  2021 1       Sat       2     2   0.0791 0.0822 0.0500
##  3 2021-01-03 0.0680  2021 1       Sun       3     3   0.0727 0.0803 0.0513
##  4 2021-01-04 0.0704  2021 1       Mon       4     4   0.0680 0.0791 0.0526
##  5 2021-01-05 0.0705  2021 1       Tue       5     5   0.0704 0.0763 0.0538
##  6 2021-01-06 0.0782  2021 1       Wed       6     6   0.0705 0.0744 0.0549
##  7 2021-01-07 0.0761  2021 1       Thu       7     7   0.0782 0.0745 0.0564
##  8 2021-01-08 0.0738  2021 1       Fri       8     8   0.0761 0.0736 0.0579
##  9 2021-01-09 0.0764  2021 1       Sat       9     9   0.0738 0.0728 0.0593
## 10 2021-01-10 0.0737  2021 1       Sun      10    10   0.0764 0.0733 0.0607
```

Let us Predict the Price of $ZIL: Best Model [4]

```
error_tbl %>% summarise(me = mean(error),
                        rmse = mean(error^2)^0.5,
                        mae  = mean(abs(error)),
                        mape = 100*mean(abs(error_pct)) ) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 4
## $ me   <dbl> 0.006849469
## $ rmse <dbl> 0.0141316
## $ mae  <dbl> 0.009613009
## $ mape <dbl> 7.059708
```

Let us Predict the Price of $ZIL: Comparison with Naive Forecast

```
library(fpp2)
naive = data.frame(year = cryptoFeatures$year,
                   adjusted = cryptoFeatures$adjusted,
                   naiveFC = cryptoFeatures$adjusted %>% lag)
naiveResults = naive %>% filter(year == 2021)
forecast::accuracy(object = naiveResults$naiveFC, x = naiveResults$adjusted)
```

```
##                      ME       RMSE         MAE      MPE     MAPE
## Test set 0.0009974754 0.01268581 0.008879902 0.388961 6.322648
```

Outline

Today's Learning Objectives

Main Learning Outcomes

- **Examine the Use of autoML for TS Prediction.**

References [1]

"Choosing the Right Estimator." 2020. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

Mitchell, Tom Michael. 2006. "The Discipline of Machine Learning." Machine Learning Department. School of Computer Science, Carnegie Mellon University. http://ra.adm.cs.cmu.edu/anon/usr0/ftp/anon/ml/CMU-ML-06-108.pdf.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–29. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560.

ISA 444: Business Forecasting

27 - Advanced Topics

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Spring 2021