Time Series

2020-05-11
Mags Giust

MACHINE LEARNING

Unsupervised Learning

DIMENSIONALITY REDUCTION

- PCA
- t-SNE
- Factor Analysis

CLUSTERING

AD, NLP may be solved with clustering.

Algorithms

- KMeans
- DBScan
- Hierarchical

Reinforcement Learning

AD, NLP, RS may be solved with reinforcement learning

- Collaborative Filtering
- Content Filtering
- Neural Networks (DL)

Supervised Learning

REGRESSION

- Continuous Target

TSA, AD may be regression probs.

- Linear Regression
- GLM
- Polynomial Reg.
- SVR
- DecisionTree Regr.
- Neural Networks

CLASSIFICATION

- Discrete, Categorical target

TSA, NLP, AD may be classification problems

- Logistic Regression
- Decision Tree
- Random Forest
- SVC
- KNN
- Neural Networks

SUB-METHODOLOGIES

- **TSA**: time series analysis
- **AD**: anomaly detection
- **NLP**: natural language Processing
- **RS**: recommendation System

TIME SERIES analysis

"Time is not a line but a series of Now Points" ~Taisen Deshimaru

Time Series Forecasting in Minutes: https://www.youtube.com/watch?v=wGUV_XqchbE

- finding patterns in temporal data, and making predictions
- a sub-methodology to Regression / classification - (predominantly regression)

Why do we treat it differently than "normal" regression?

general Linear Regression

X = features									target
x1	x2	x3	x4	x5	x6	x7	x8	x9	y

Time Series

X = features = time!							target
t_{-7}	t_{-6}	t_{-5}	t_{-4}	t_{-3}	t_{-2}	t_{-1}	t_0

ASSUMPTION: features are independent of each other!

BUT: these features (consecutive date/time stamps) are, by their nature, dependent on each other!

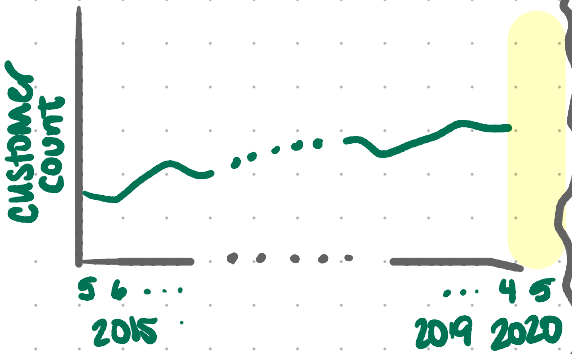
Time Series

forecast/predict # of new customers next month using historical monthly new customer count

- features = each historical month
- target = next month

1 dimension

Date	Customer Count
2015-05	
2015-04	
⋮	⋮
2020-04	
2020-05	

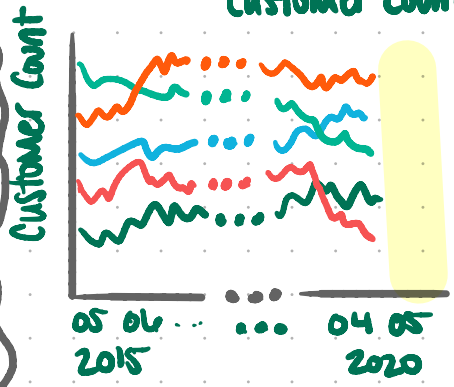


1 Model

5 dimensions

Product	5/15	6/15	...	4/20	5/20
savings acct					
checking acct					
home insurance					
auto insurance					
renters ins.					

Customer Count



Multiple Models

Not Time Series

Predict # of new customers each month using sales/marketing activity.

- observation = Month/Year
- features (e.g.) = prev. month campaign count, 30d. Δ in sales staff, # new prod. features
- target = customer count

Features | y

Month-Year	# of campaigns	Δ in Sales Staff	...	Customer Count
5-15				
6-15				
⋮				
3-20				
4-20				

Model patterns from previous months to be able to give a 30d. forecast of customer count.

To predict by product, build multiple models, 1 per product

- Savings acct
- Checking acct
- Home insurance
- Auto insurance
- Renters insurance

Time Series Vocabulary

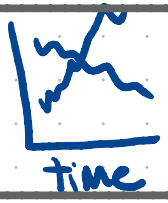
RESAMPLING (dates) Changing frequency of data points.

date	New customers
2020-04-01	50
2020-04-02	125
⋮	⋮
2020-04-30	75
2020-05-01	53

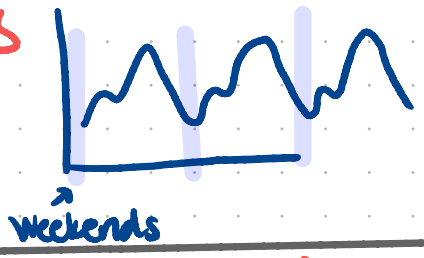
Resample Monthly & SUM

date	New customers
2020-04	50+125+...+75
2020-05	53+...

TREND: Long term progression (increasing or decreasing)

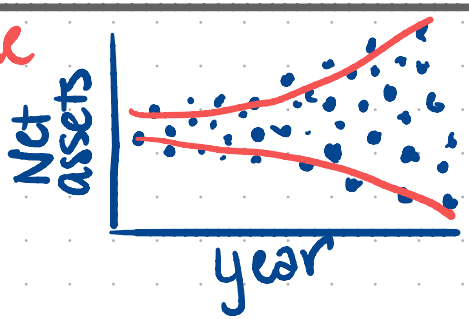


SEASONALITY: Series is influenced by seasonal factors
e.g. Month of year, day of week.
Always a fixed, known period. ⇒
Seasonal series == PERIODIC series

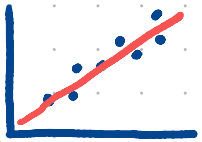


CYCLIC: fluctuations that are NOT of a fixed period. Duration of fluctuations > 2yrs
e.g. housing market.

HETEROSKEDASTICITY: Changes in variance over time



AUTOCORRELATION: "Regression of Self"
Used to detect non-randomness in data.
It is a correlation coefficient, but instead of between 2 different variables, it is between the values of the same variable at 2 different times



"tomorrow's" value is dependent on "today's" value.

LAG VARIABLE: Previous time step
yesterday (t_{-1}) is a lag var to today (t_0)

New Skills

Acquisition: gather data using a RestAPI

Prep: Working with dates - resample()
asfreq()

filling missing values - fillna() ffill, pad, bfill

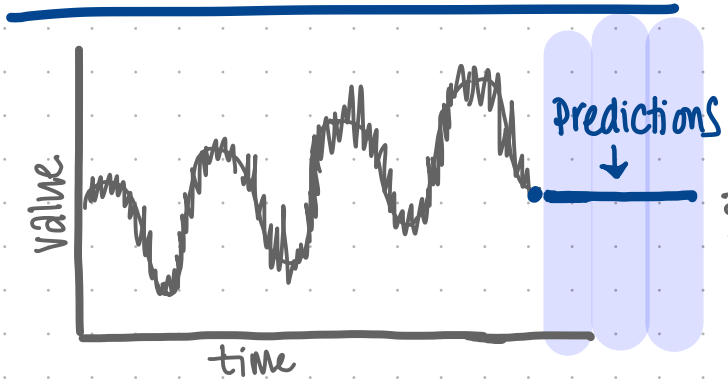
Explore: Splitting TS data into train/test
- sklearn.model_selection.TimeSeriesSplit
- using date cutoff

Viz time series data: plot simple aggregate (resample().mean().plot())
Rolling aggregate (rolling().mean().plot())
Pd.over Pd.diff. (resample().mean().diff().plot())
Customize datetime axis Matplotlib.dates
Explore 'Seasonality'
Merge Multiple frequencies into same plot
Time Series Decomposition
Analyzing Lag

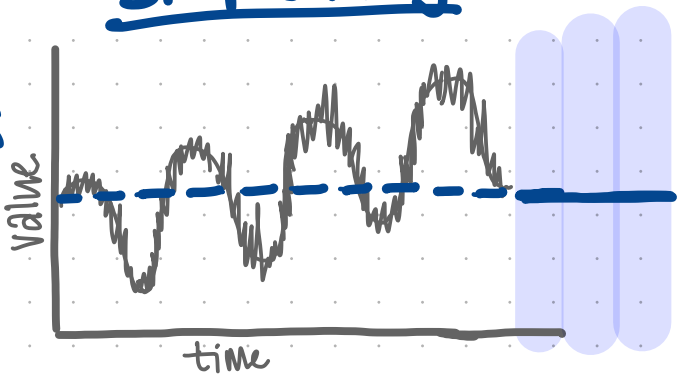
Model: Forecast/Predict Methods include
- Last observed value
- Simple avg.
- Moving/rolling avg
- Holt's linear trend
- previous cycle
- FB Prophet Module.

Forecasting, Predicting, Modeling Time Series Data

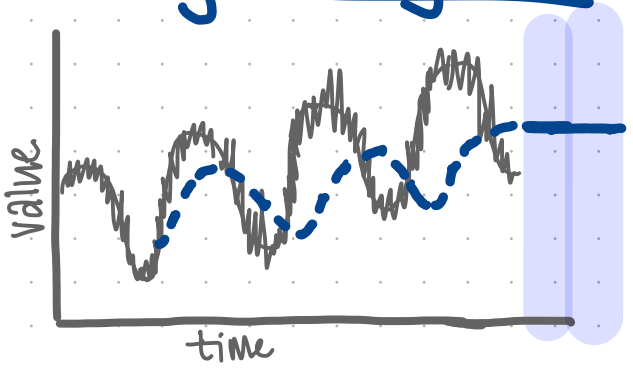
Last observed value



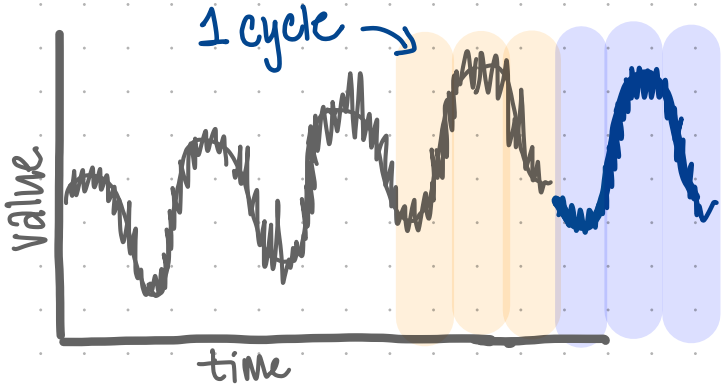
Simple Avg



Moving/Rolling Avg.

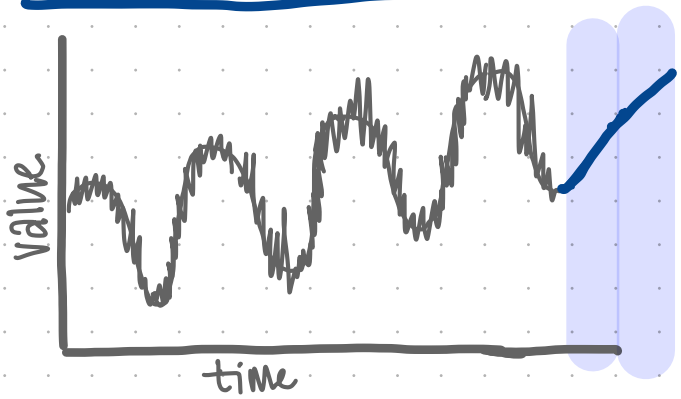


Previous Cycle



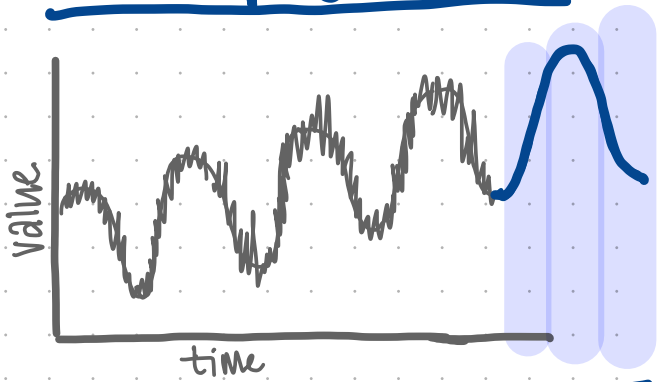
[define a cycle, predict the next cycle to be the values of the previous cycle]

Holt's Linear Trend



[Exponential Smoothing applied to both the average & the trend (slope)]

FB Prophet Model



[Non-linear trends fit with yearly, weekly, daily seasonality + holiday effects]