Machine Learning

Time Series Module

SUPERVISED

CLASSIFICATION REGRESSION

TARGET:

Discrete
Categorical

Continuous
numeric, ordered

SUB-METHODOLOGIES:

time series
Anomaly detection
NLP

time series
anomaly detection

ALGORITHMS:

Logistic Regression
SVM-classifier
Decision Tree
Random Forest
KNN
Neural Network
Naive Bayes

OLS (linear regression)
GLM (tweedie regressor)
Polynomial Regression
Support Vector Regressor
Decision Tree Regressor
Neural Networks
LASSO, LARS
Holt-Winters
Auto Regression

REINFORCEMENT

SUB-METHODOLOGIES

Anomaly Detection
NLP
Recommender Systems

ALGORITHMS:

Collaborative Filtering
Content Filtering
Neural Networks

UNSUPERVISED

CLUSTERING DIMENSIONALITY REDUCTION

SUB-METHODOLOGIES:

Anomaly Detection
NLP

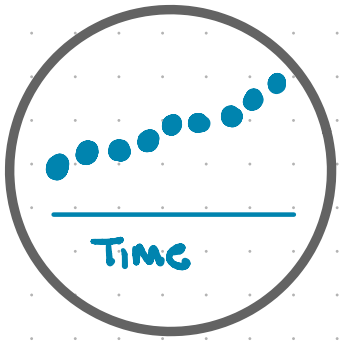
ALGORITHMS:

K-Means
DBSCAN
Hearchical

PCA
t-SNE
factor Analysis

TIME SERIES ANALYSIS

"Time is not a line but a series of 'now' points"
Taisen Deshimaru



TSA is finding patterns in temporal data & making predictions.

TSA is a sub-methodology to Regression & Classification. ∴ it is a form of supervised machine learning

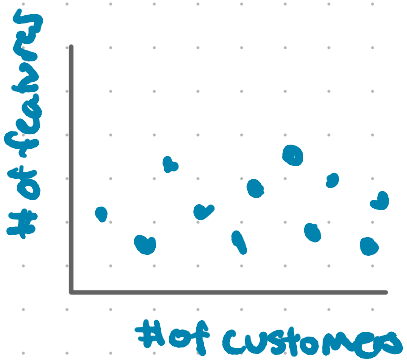
TSA - a special case of regression - WHY?

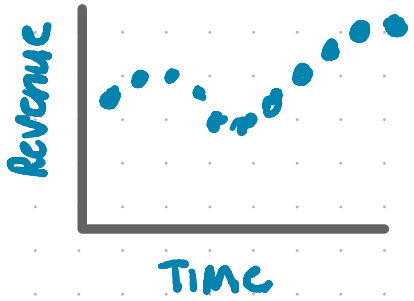
Because... **ASSUMPTIONS**

Time Series data does NOT meet required assumptions for linear regression.

Specifically... **COLLINEARITY**

The features in TSA are date/times. By their nature, they are dependent on each other.





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	← today
last week						yest.		

X = dependent

ASSUMPTION: features are independent of each other!

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

TO TSA or not to TSA ...THAT is the question

SCENARIO: Predict number of new customers we will have (customer growth)

OPTIONS

TIME SERIES REGRESSION — METHODOLOGY — ORDINARY REGRESSION

Use previous OUTCOMES as predictors to predict future OUTCOMES

	Customer growth		
	TRAIN	VALIDATE	TEST
Jan 20	100	100	100
Feb 20	90	90	90
Mar 20	95	95	95
Apr 20	80	80	80
May 20	105	105	105
Jun 20	110	110	110
Jul 20	110	110	110
Aug 20		90	90
Sep 20			70
Oct 20			

TRAIN
LABELED OUTCOME
predictors

VALIDATE
LABELED OUTCOME
predictors

TEST
LABELED OUTCOME
predictors

Use Attributes to predict OUTCOME

	Customer growth	Sales staff	Mktg $	# campaigns	# new products
Jan 20	100				
Feb 20	90				
Mar 20	95				
Apr 20	80				
May 20	105				
Jun 20	110				
Jul 20	110				
Aug 20	90				
Sept 20	70				
Oct 20					

Attributes = predictors = features = independent vars.
outcome = target = dependent variable

Time Series Vocabulary

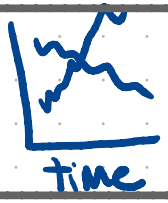
RESAMPLING (dates) Changing frequency of data points.

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

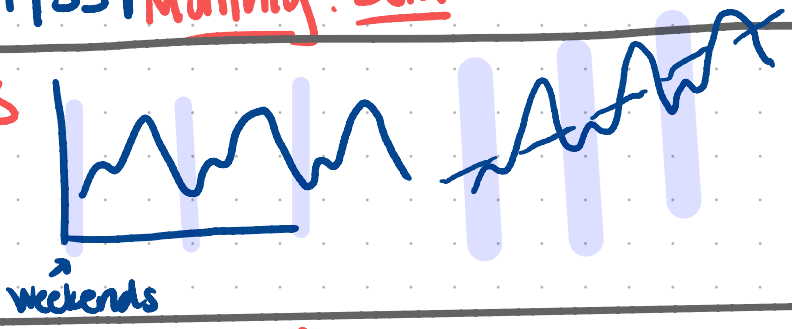
Resample Monthly & SUM

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

TREND: Long term progression (increasing or decreasing)

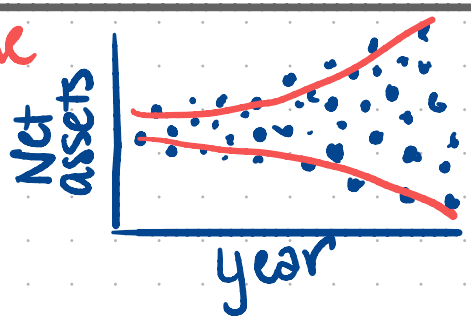


SEASONALITY: Series is influenced by seasonal factors
e.g. Month of year, day of week.
Always a fixed, known period. ⇒
Seasonal series == PERIODIC series



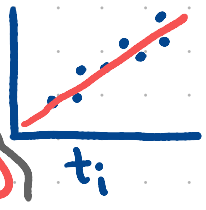
CYCLIC: fluctuations that are NOT of a fixed period. Duration of fluctuations -> 2yrs
e.g. housing market.

HETEROSKEDASTICITY: Changes in variance over time



COLLINEARITY when independent variables are highly correlated

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)

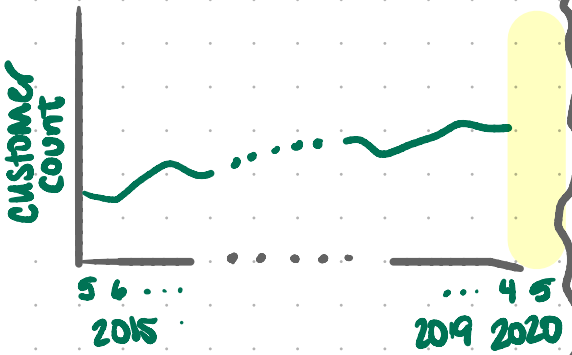
Time Series

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	

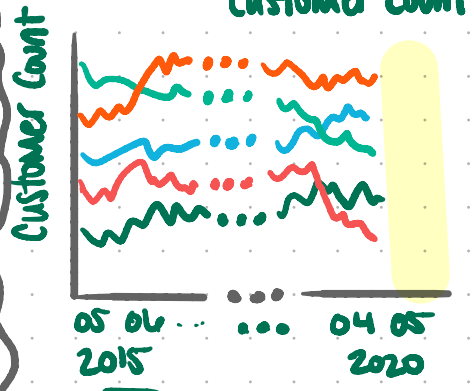


1 Model

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				

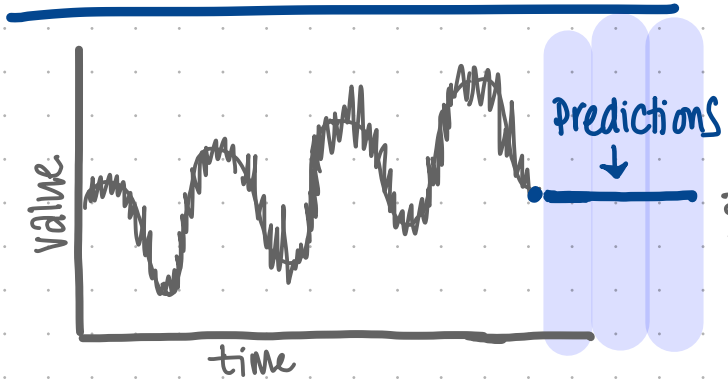
Model patterns from previous months to be able to give a 30d. forecast of customer count.

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

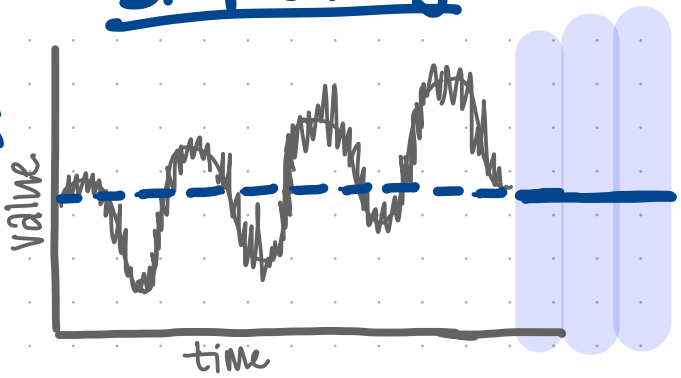
- Savings acct
- Checking acct
- Home insurance
- Auto insurance
- Renters insurance

Forecasting, Predicting, Modeling Time Series Data

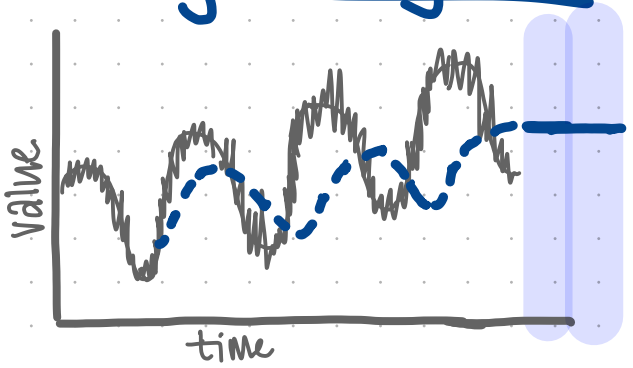
Last observed value



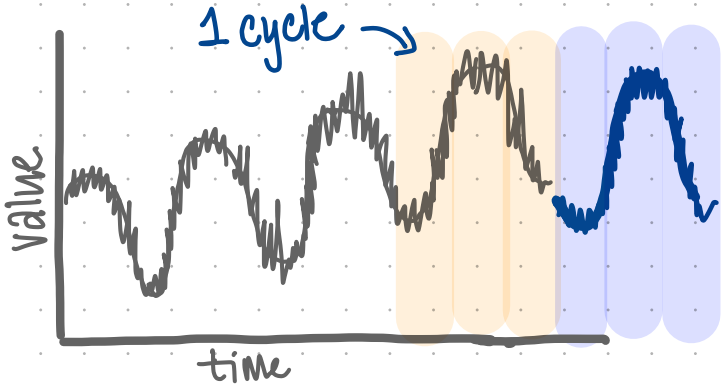
Simple Avg



Moving/Rolling Avg.

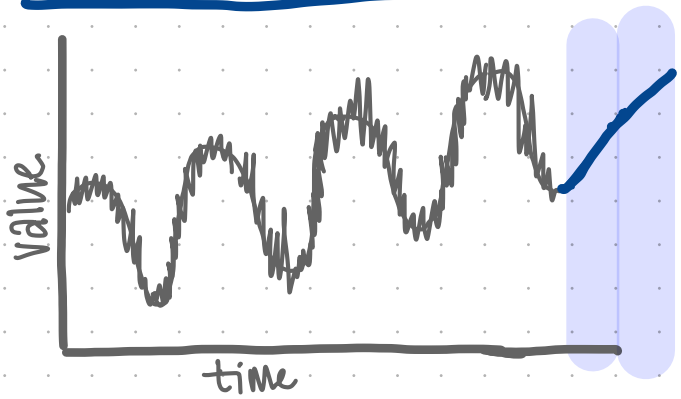


Previous Cycle



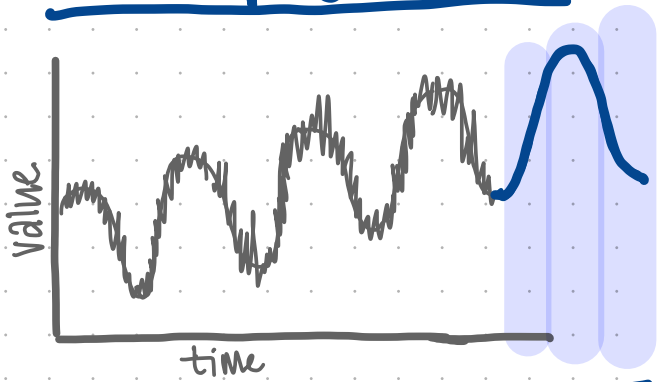
[define a cycle, predict the next cycle to be the values of the previous cycle]

Holt's Linear Trend



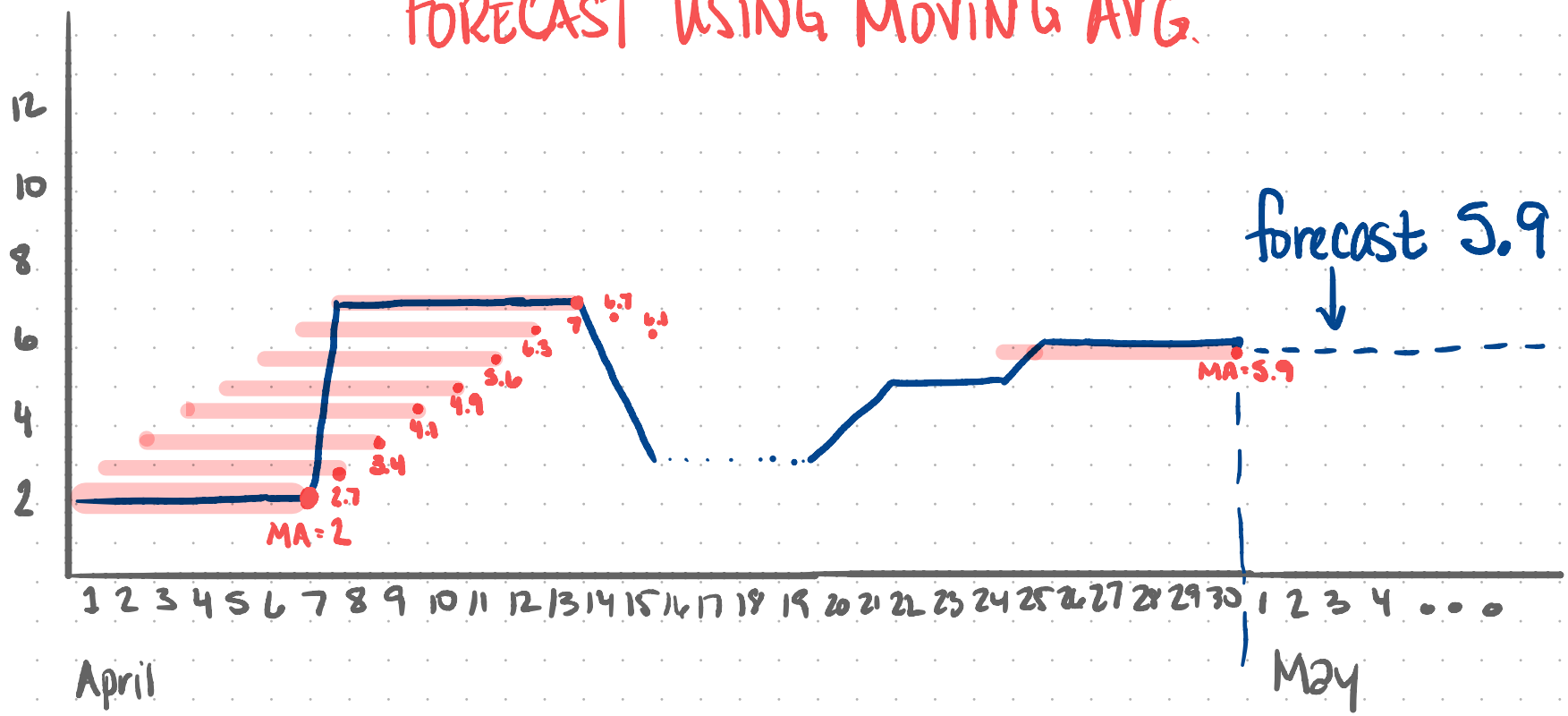
[Exponential Smoothing applied to both the average & the trend (slope)]

FB Prophet Model



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

FORECAST USING MOVING AVG.



7-day Moving/Rolling Average

Time Series Cross Validation - Prophet