Machine Learning w/ R and/or Azure

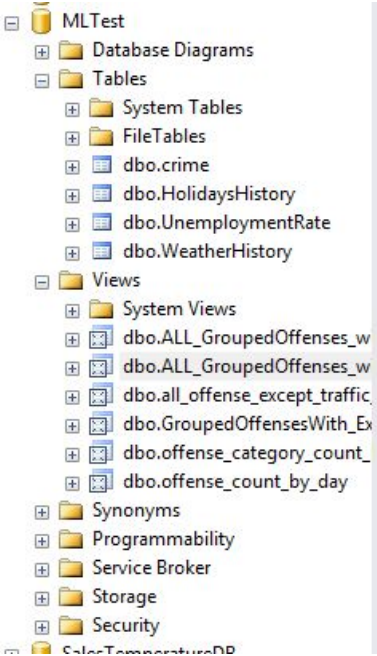
Machine Learning Steps

1. Ask the Right Question
2. Prepare the Data
3. Select the Algorithm
4. Training the Model
5. Test the model.
6. Repeat over and over.

Data Prep

Usually 50-80% of the Time/Effort.

Everyone underestimates the complexity.




```
SELECT *
  FROM [MLTest].[dbo].[ALL_GroupedOffenses_without_traffic_With_External_Less_Holiday]
  order by crime_date
```

100 %

Results | Messages

	offense_count	crime_date	Year	Week	WeekDay	DayofYear	Holiday	Before	After	HolidayN...	Unemplo...	minTemp	maxTemp	precip
142	183	2014-05-23	2014	21	6	143	0	0	0	0	5.1	45.30	71.50	0.40
143	180	2014-05-24	2014	21	7	144	0	0	0	0	5.1	50.70	69.60	0.21
144	159	2014-05-25	2014	22	1	145	0	1	0	6	5.1	46.60	67.20	0.07
145	166	2014-05-26	2014	22	2	146	1	0	0	6	5.1	48.00	76.00	0.00
146	199	2014-05-27	2014	22	3	147	0	0	1	6	5.1	48.20	83.50	0.00
147	208	2014-05-28	2014	22	4	148	0	0	0	0	5.1	52.50	88.50	0.00
148	195	2014-05-29	2014	22	5	149	0	0	0	0	5.1	57.60	86.40	0.01
149	206	2014-05-30	2014	22	6	150	0	0	0	0	5.1	56.80	71.70	0.15
150	186	2014-05-31	2014	22	7	151	0	0	0	0	5.1	49.10	80.60	0.00
151	166	2014-06-01	2014	23	1	152	0	0	0	0	4.9	53.80	84.00	0.00
152	225	2014-06-02	2014	23	2	153	0	0	0	0	4.9	52.20	85.20	0.00
153	182	2014-06-03	2014	23	3	154	0	0	0	0	4.9	54.40	91.90	0.00

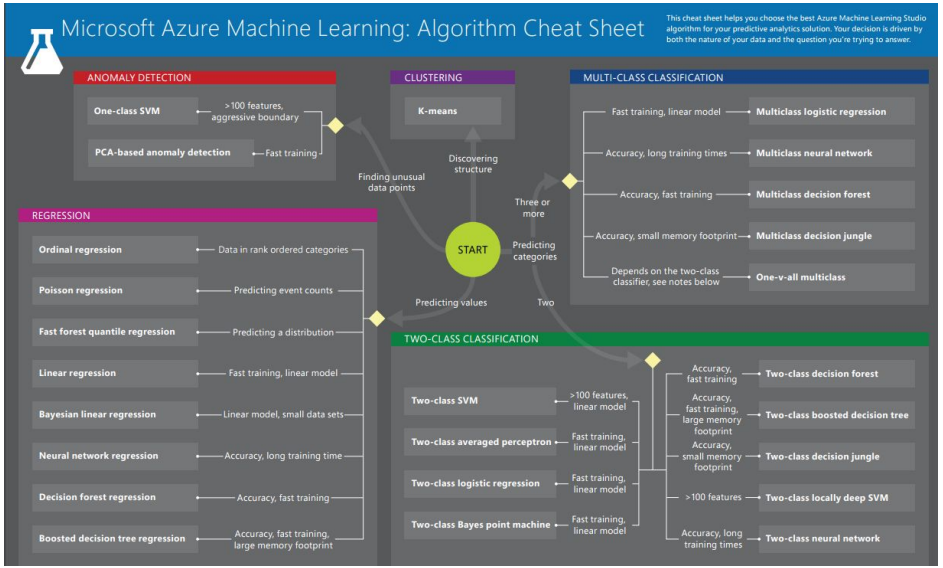
Data Split / Select Algorithm / Train

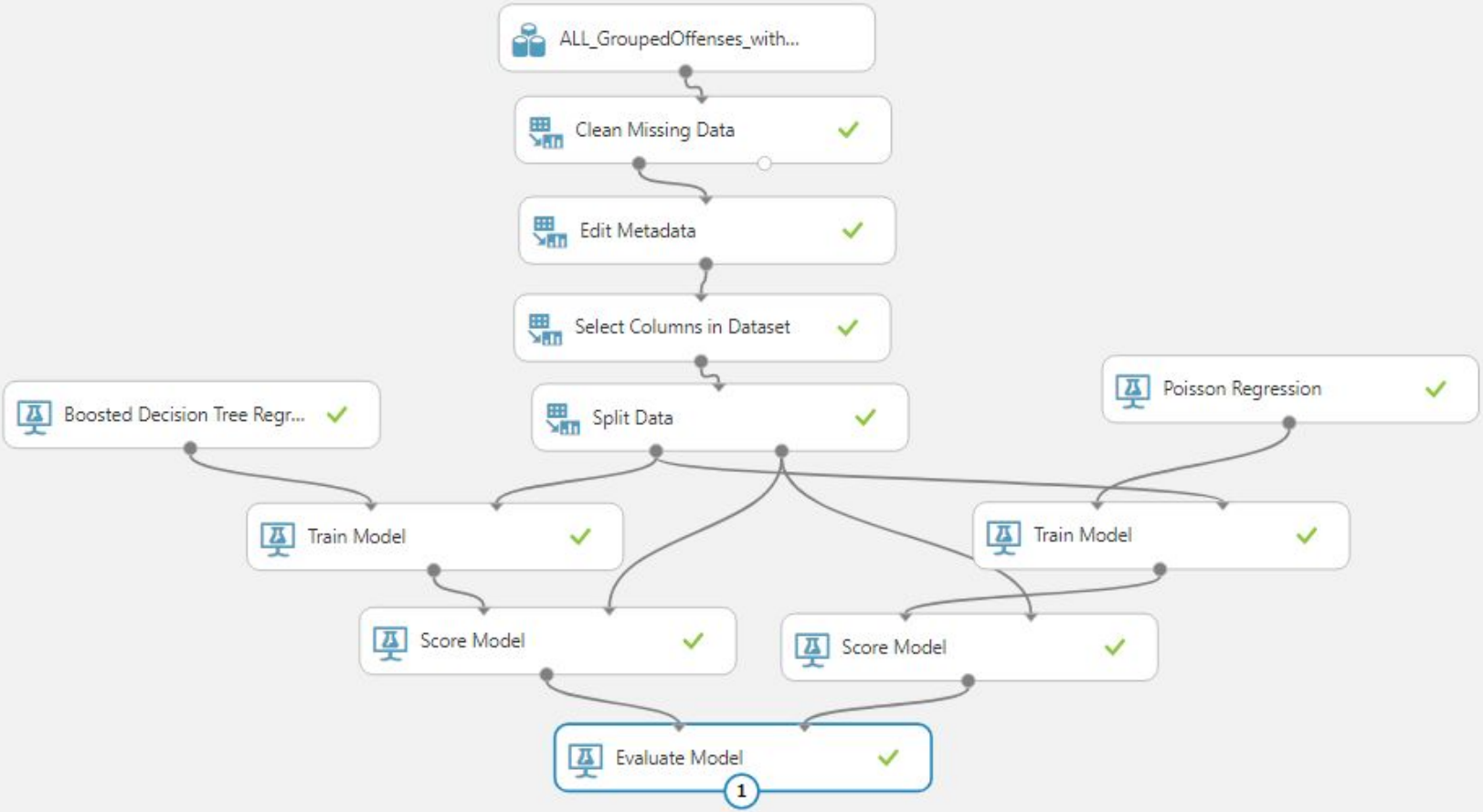
Normal is around 70% Training Data / 30% Test Data.

You must split your data, or you will by definition overstate your accuracy if you use your Test data for training.

<u>4 Primary Algorithm Types</u>
Anomaly Detection
Classification (2 Class / MultiClass)
Clustering
Regression

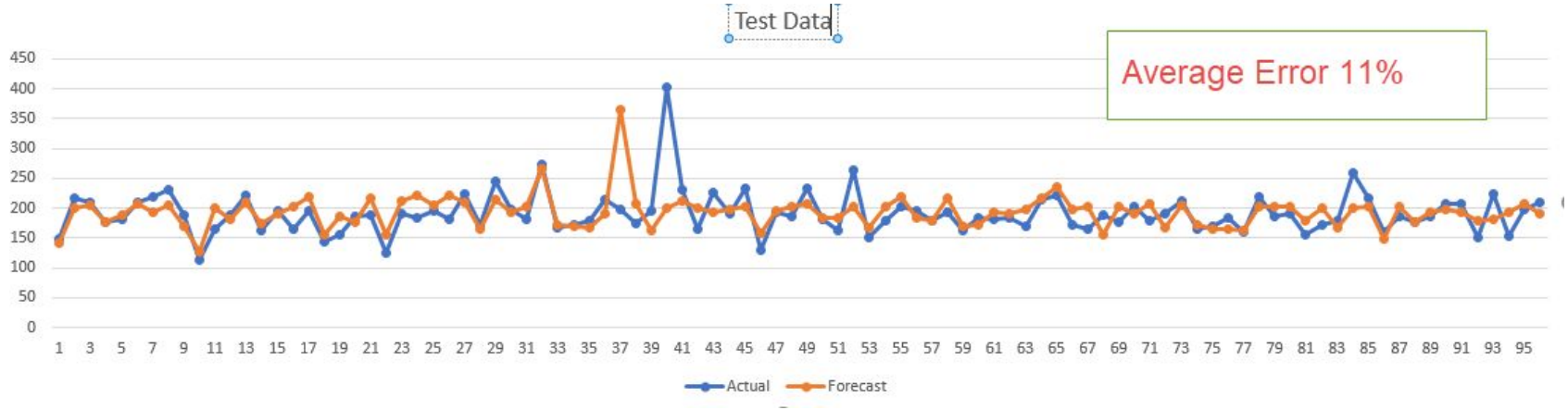
Link to Algorithm
Cheat Sheet on last slide->

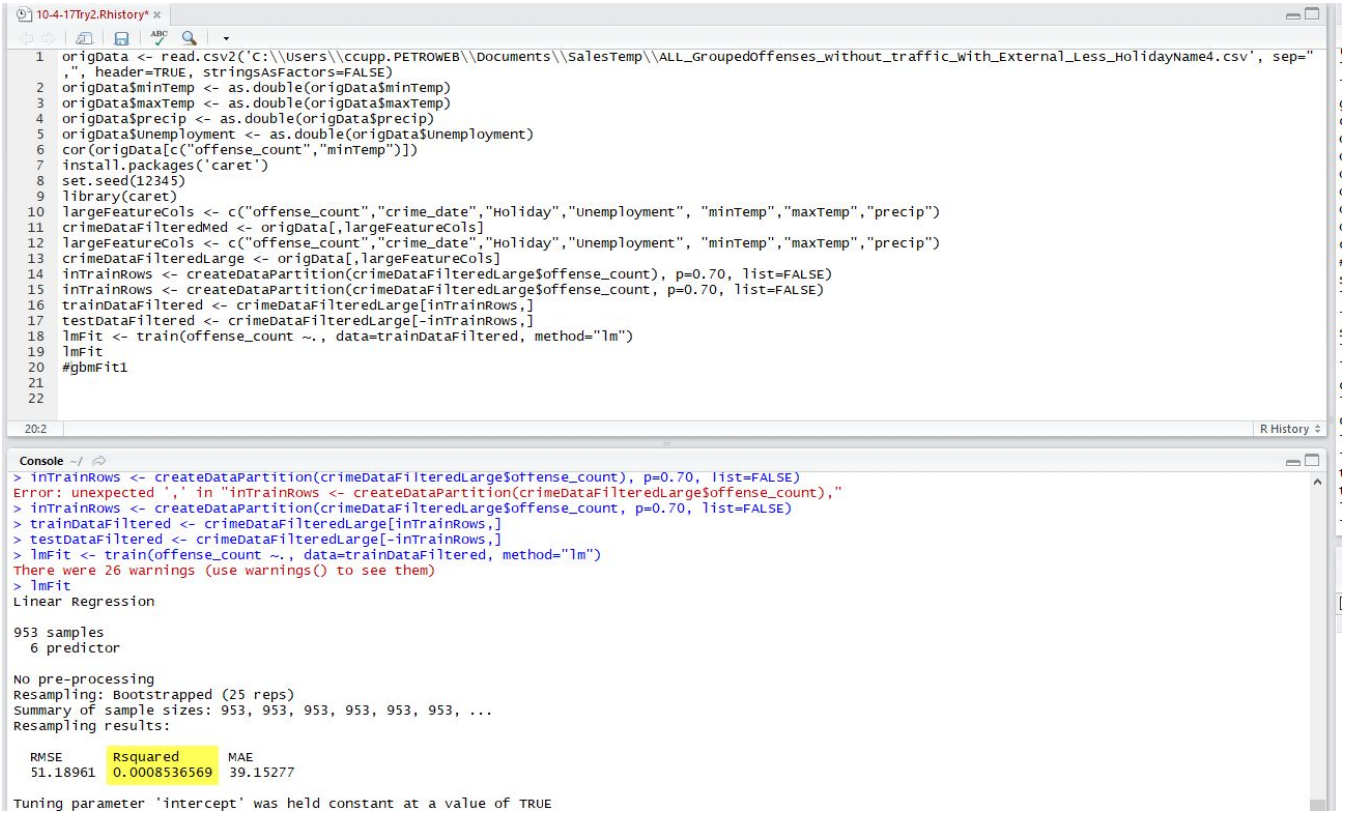



ALL_GroupedOffenses_with...
Clean Missing Data
Edit Metadata
Select Columns in Dataset
Split Data
Boosted Decision Tree Regr...
Poisson Regression
Train Model
Train Model
Score Model
Score Model
Evaluate Model
1

Evaluation of Model

Mean Absolute Error	20.379047
Root Mean Squared Error	29.311785
Relative Absolute Error	0.771106
Relative Squared Error	0.582272
Coefficient of Determination	0.417728




10-4-17Try2.Rhistory*
1 origData <- read.csv2('C:\\Users\\ccupp.PETROWEB\\Documents\\SalesTemp\\ALL_GroupedOffenses_without_traffic_With_External_Less_HolidayName4.csv', sep=","
, header=TRUE, stringsAsFactors=FALSE)
2 origData$minTemp <- as.double(origData$minTemp)
3 origData$maxTemp <- as.double(origData$maxTemp)
4 origData$precip <- as.double(origData$precip)
5 origData$Unemployment <- as.double(origData$Unemployment)
6 cor(origData[c("offense_count","minTemp")])
7 install.packages('caret')
8 set.seed(12345)
9 library(caret)
10 largeFeatureCols <- c("offense_count","crime_date","Holiday","Unemployment", "minTemp","maxTemp","precip")
11 crimeDataFilteredMed <- origData[,largeFeatureCols]
12 largeFeatureCols <- c("offense_count","crime_date","Holiday","Unemployment", "minTemp","maxTemp","precip")
13 crimeDataFilteredLarge <- origData[,largeFeatureCols]
14 inTrainRows <- createDataPartition(crimeDataFilteredLarge$offense_count), p=0.70, list=FALSE)
15 inTrainRows <- createDataPartition(crimeDataFilteredLarge$offense_count, p=0.70, list=FALSE)
16 trainDataFiltered <- crimeDataFilteredLarge[inTrainRows,]
17 testDataFiltered <- crimeDataFilteredLarge[-inTrainRows,]
18 lmFit <- train(offense_count ~., data=trainDataFiltered, method="lm")
19 lmFit
20 #gbmFit1
21
22
20:2
R History
Console ~/
> inTrainRows <- createDataPartition(crimeDataFilteredLarge$offense_count), p=0.70, list=FALSE)
Error: unexpected ',' in "inTrainRows <- createDataPartition(crimeDataFilteredLarge$offense_count),"
> inTrainRows <- createDataPartition(crimeDataFilteredLarge$offense_count, p=0.70, list=FALSE)
> trainDataFiltered <- crimeDataFilteredLarge[inTrainRows,]
> testDataFiltered <- crimeDataFilteredLarge[-inTrainRows,]
> lmFit <- train(offense_count ~., data=trainDataFiltered, method="lm")
There were 26 warnings (use warnings() to see them)
> lmFit
Linear Regression
953 samples
6 predictor
No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 953, 953, 953, 953, 953, 953, ...
Resampling results:
RMSE Rsquared MAE
51.18961 0.0008536569 39.15277
Tuning parameter 'intercept' was held constant at a value of TRUE

R

Now what.

More Data.... Always more Data.

- More History
- Other Sources
  - Student Schedule
  - Events (Broncos, Rockies Games)

- Realistic Scenario.. This would be a far superior/usable model if it was by precinct or neighborhood... and you would need GIS Skills.

Links

https://github.com/gcaseycupp/MLwithRandAzure
gcaseycupp@gmail.com

SalesTemperature: http://www.salestemperature.com/ **ForecastER:** http://www.patientforecaster.com/

Pluralsight - Azure ML: https://app.pluralsight.com/library/courses/azure-machine-learning-getting-started

Pluralsight - R: https://app.pluralsight.com/library/courses/r-understanding-machine-learning

Azure Algorithm Cheat Sheet: https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet

Denver MileHigh MapTime Meetup : R Intro: https://github.com/rsteve388/Maptime-Introduction-to-R-/

Getting Data Science with R and ArcGIS: https://community.esri.com/videos/3269

R Bridge for ArcGIS:
https://learn.arcgis.com/en/projects/analyze-crime-using-statistics-and-the-r-arcgis-bridge/lessons/install-the-r-arcgis-bridge-and-start-statistical-analysis.htm https://community.esri.com/videos/3343

Denver Crime Stats Home :
https://www.denvergov.org/content/denvergov/en/police-department/crime-information/crime-statistics-maps/2015-crime-statistics-maps.html