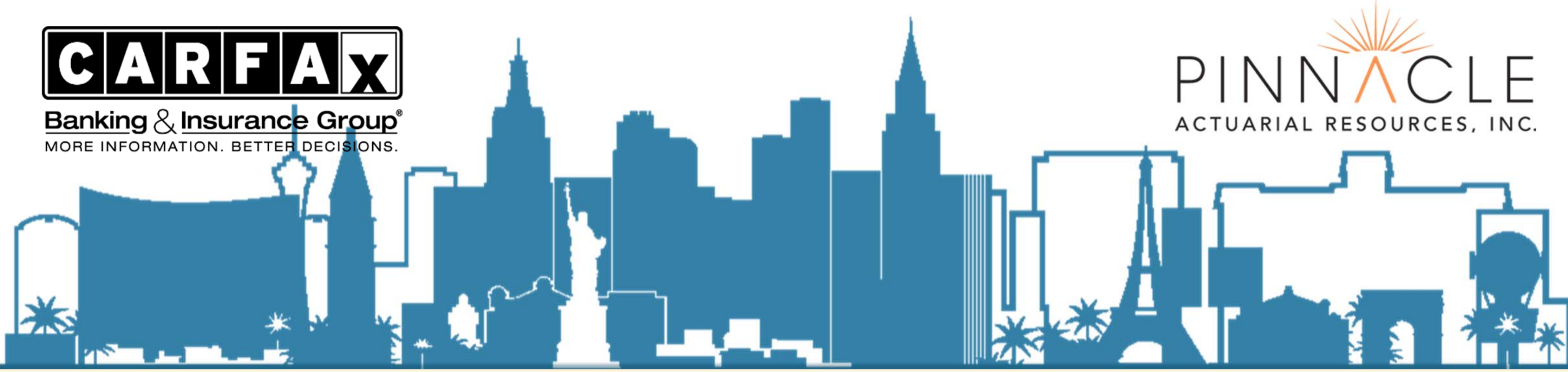


GLMS, MACHINE LEARNING, & MORE, OH MY!

Michael Chen, Pinnacle Actuarial Resources, Inc.,
Gary Wang, Pinnacle Actuarial Resources, Inc.,
Don Hendriks, CARFAX

2019 CAS ANNUAL MEETING – HONOLULU, HI

About the Presenters

- **Donald Hendriks**, ACAS, MAAA
- CARFAX
- National Business Consultant
- Greater Detroit Area, MI

- **Michael Chen**, FCAS, MAAA, CSPA
- Pinnacle Actuarial Resources, Inc.
- Consulting Actuary
- Des Moines, IA

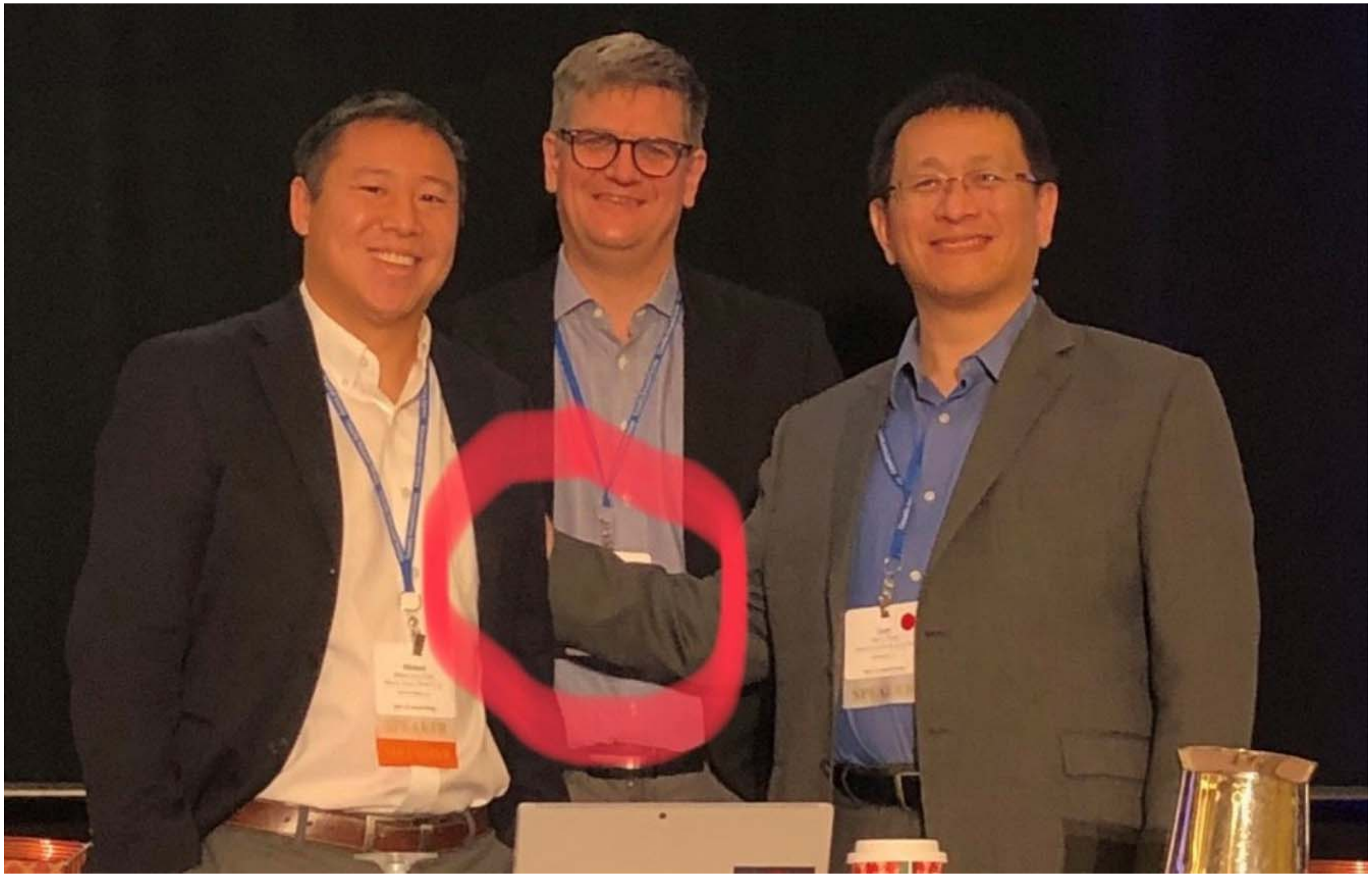
- **Gary Wang**, FCAS, MAAA, CSPA
- Pinnacle Actuarial Resources, Inc.
- Senior Consulting Actuary
- Bloomington, IL

CARFAX

PINNACLE ACTUARIAL RESOURCES, INC.

Introduction

Introduction

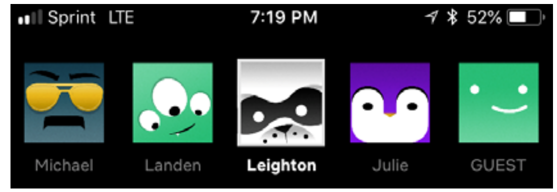


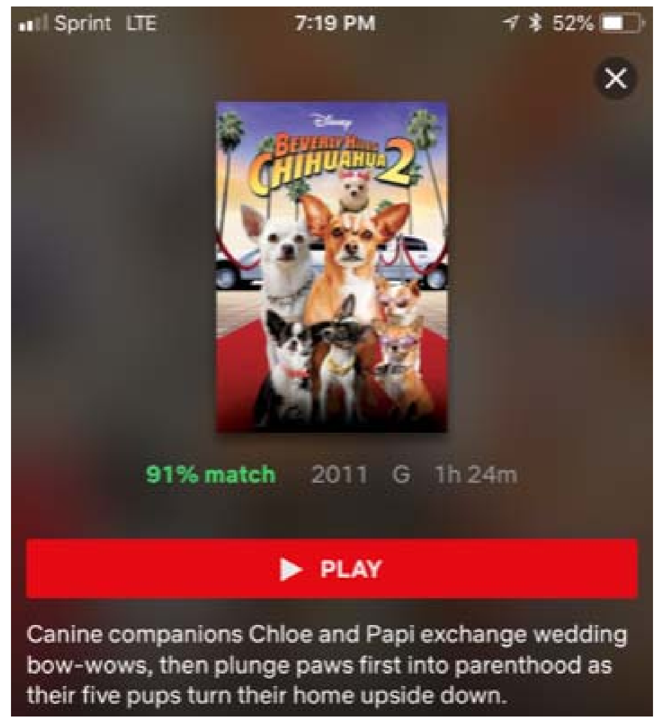
Agenda

- Introduction
- Residual Analysis – Feature Engineering
- Leveraging Competing Models in Exploration
- Machine Learning

Introduction

Netflix Prize



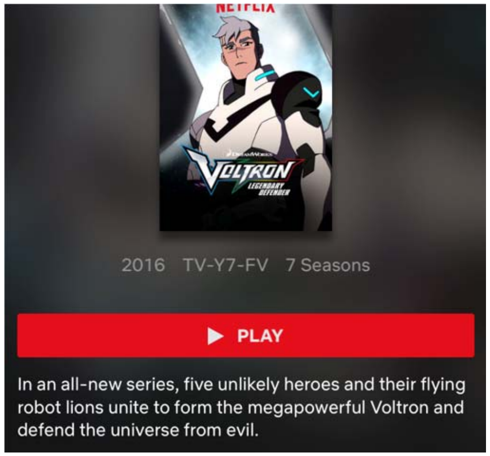


Netflix Prize

- Competition began 10/2/2006 for Netflix which at the time was mainly a DVD-rental service (video streaming was just beginning)
- Prize for best collaborative filtering algorithm to predict user ratings for films
- $1,000,000 prize to be awarded for making the company’s recommendation engine 10% more accurate.
- The Data Set (nothing like it at the time, before Kaggle was founded)
  - Over 100 Million ratings
  - 17,770 movies
  - 480,189 customers

Netflix Prize

1. By 2007 and 2008 while many teams had improved on the algorithm none had gained the 10% improvement
   - The solution? Combine forces!
   - Teams generally do better as their members become familiar with one another.



Netflix Prize

2. $1,000,000 awarded in 9/21/2009 to BellKor's Pragmatic Chaos which bested Netflix's own algorithm for predicting ratings by 10.06%
   - Netflix never implemented the winning algorithm
   - Netflix implemented earlier simpler algorithm with 8.43% improvement. Additional accuracy determined to cost too much of an engineering effort for the result.

Netflix Prize

3. Since deemed a success, why no Netflix Prize 2?
    - Netflix involved in a multi-million dollar lawsuit claiming the data could be de-anonymized using background knowledge from the Internet Movie Database.
    - Environment had changed
        - Netflix moving toward streaming service
            - More data
            - What people were actually watching, not just rating

Computers vs People

- In 1997 Deep Blue beat World Champion, Garry Kasparov in Chess
- In 2011 Watson beat former champions Ken Jennings and Brad Rutter at Jeopardy!

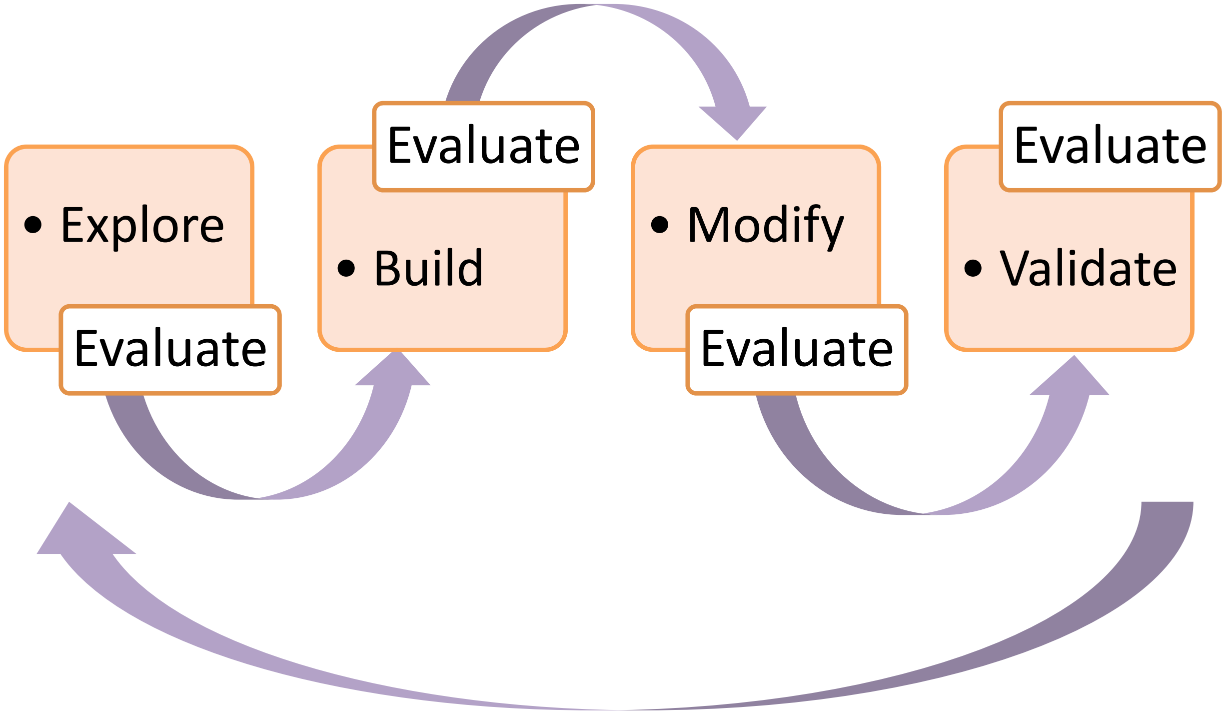
Freestyle Chess: Human-Computer Teams

- Freestyle Chess: Human-Computer Teams
  - Freestyle Chess is a variant form where teams are also allowed and, within the established time limits, every possible form of consultation.
  - Two amateur players with a computer able to be other teams comprised of grandmasters that also had computers.

Freestyle Chess: Human-Computer Teams

1. Human-computer teams can out perform computers or experts
2. The people working the smart machine doesn't necessarily have to be an expert in the task at hand.
3. People should be cognizant of their own limits

Modeling Process Overview



Feature Engineering

- Feature engineering, refers to the process of creating new input features.
  - Feature engineering is an effective method of improving predictive models.

Feature Engineering, creating useful features

- Calculate statistics like the minimums, maximums, averages, medians and ranges.
  - Investigating the extremes (or the lack) may help define interesting behaviors.

- Create flags and count occurrences of events, highlighting statistically interesting habitual behaviors.
  - NSF notice on renewals for retention analysis
  - Examples: Younger drivers may separate themselves more clearly in a Non-Standard book than older drivers

- Create ratios seeking to add predictive value to already meaningful variables.
  - density, population/land area
  - vehicle to driver ratio (often used in a Matrix)

Feature Engineering, creating useful features

- Develop quintiles across variables of interest seeking to create expressive segments of the population while also dealing with extreme values.
  - Creating bins to transform variates to categorical variables

- Apply dimensionality reduction techniques, ranks, clustering etc. expecting that grouping those with similar behaviors will be statistically beneficial.
  - Principal Components
  - Clustering

Feature Engineering, creating useful features

- Consider the element of time as an important interaction with any feature that has been developed.
  - Recognizing newer policies are biased toward no claim history, opportunity to tap into external data

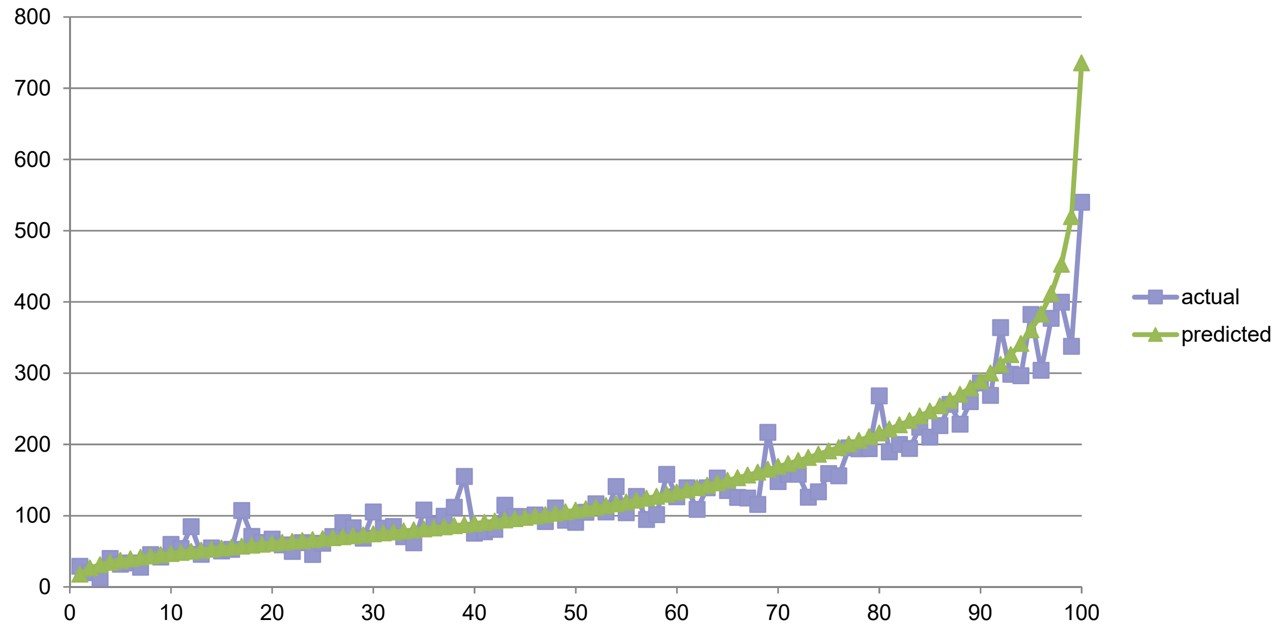
- Use regression to identify trends in continuous variables thinking that moves up or down (whether fast or slow) will be interesting.
  - Looking at univariate and doing a fit

Feature Engineering, using non-linear techniques

- Have GLM
- Create residuals from your Actual and Predicted Values
- Model on the residuals using non-linear techniques
- Practical considerations concerning algorithm from the models
- Score your original GLM dataset
- Model with new variable either as a variate or could group as a categorical
  - Modeling techniques like decision trees naturally produce bins (categorical), bins have values so could also be treated as a variate
- Alternatively could do similar exercise on the actual for example vehicle symboling

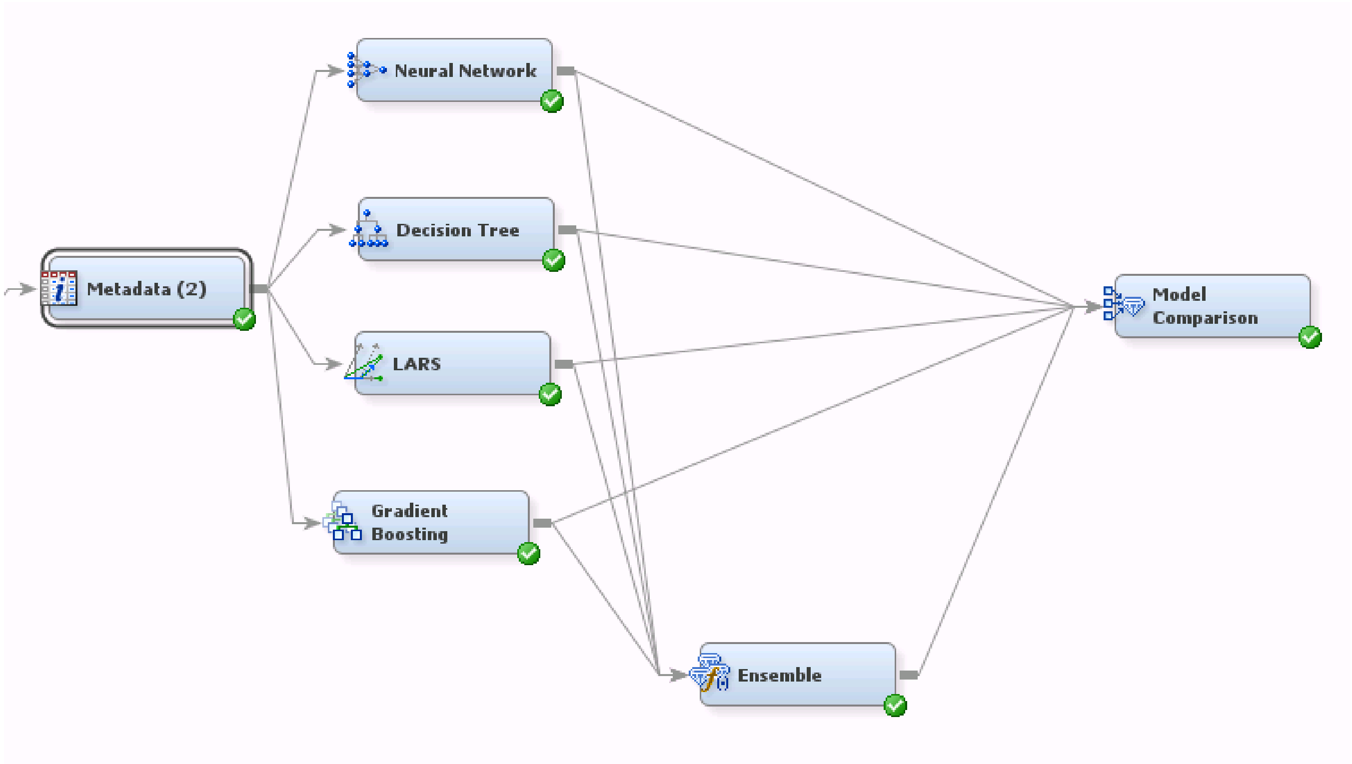
Feature Engineering, using non-linear techniques

- Initial GLM has been completed
  - Validates fairly well but may have room for improvement



Feature Engineering, using non-linear techniques

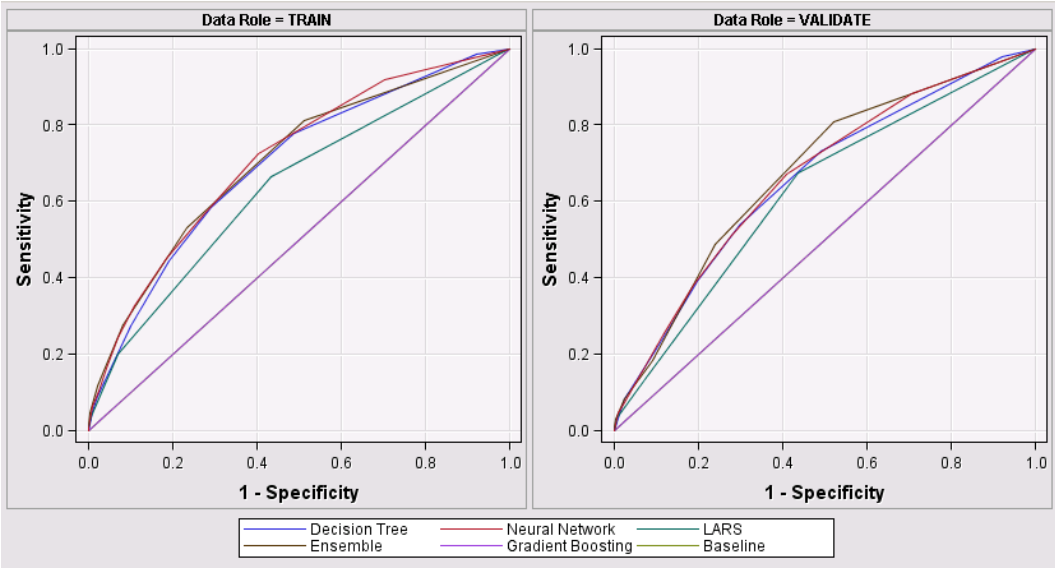
- Model on the residuals using non-linear techniques



Feature Engineering, using non-linear techniques

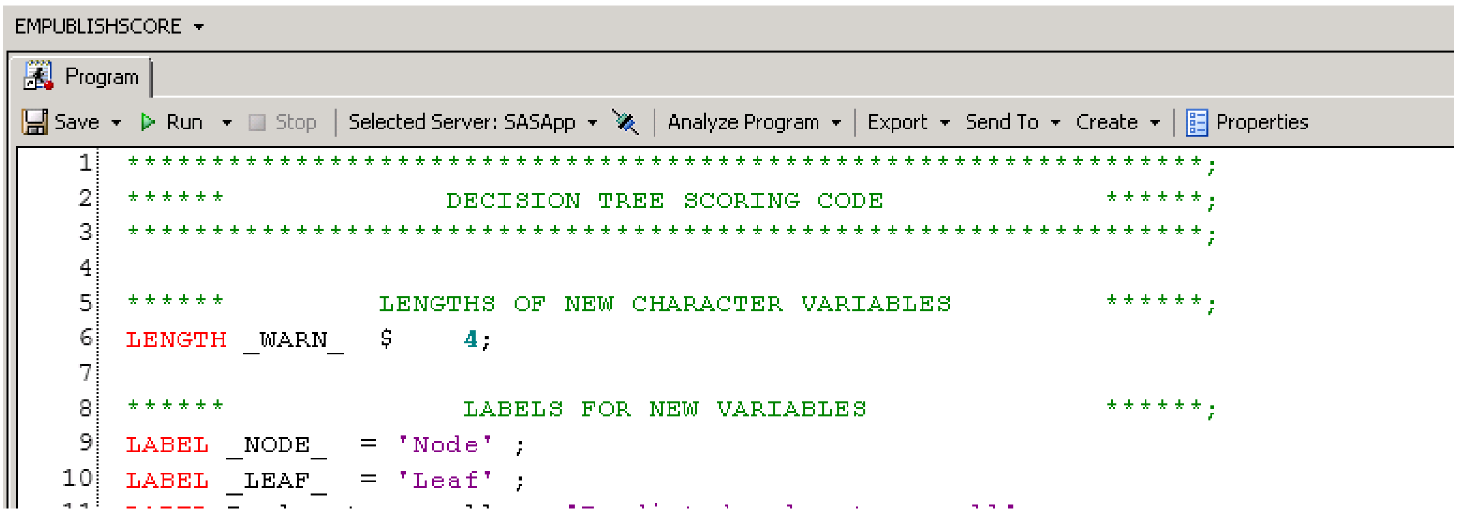
- Practical considerations concerning algorithm from the models

Selected Model	Model Description
Y	Ensemble
	Neural Network
	Decision Tree
	Gradient Boosting
	LARS



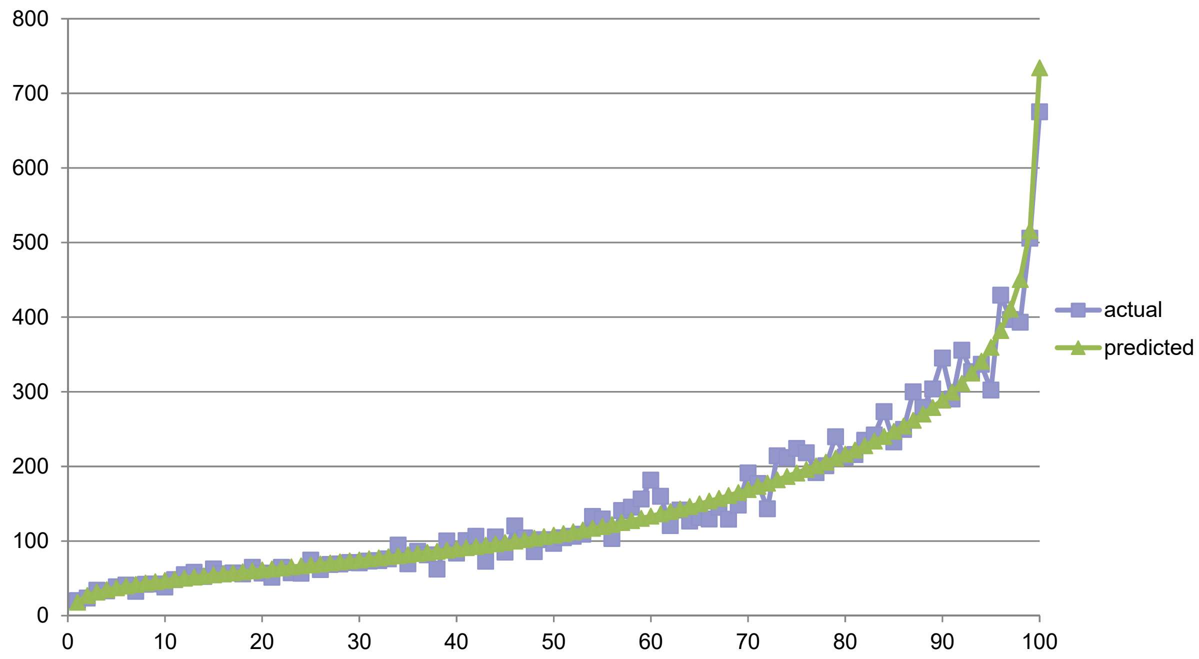
Feature Engineering, using non-linear techniques

- Use developed non-linear model to score your original GLM dataset



Feature Engineering, using non-linear techniques

- Model with new variable either as a variate or could group as a categorical

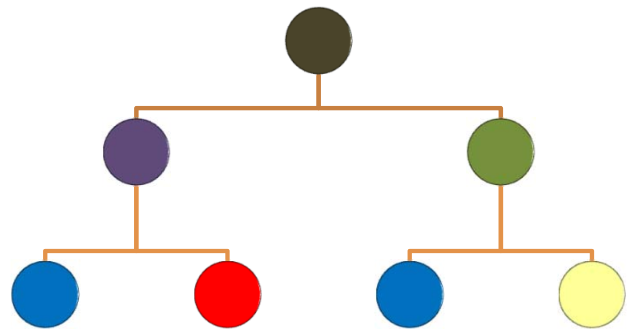


Decision Trees - Methodology

Split data according to measures of similarity

If the Target Variable is: **Categorical → Classification Tree**
Continuous → Regression Tree

Two Competing Objectives: **Purity → Measure of Variation**
Parsimony → Desire for Simple



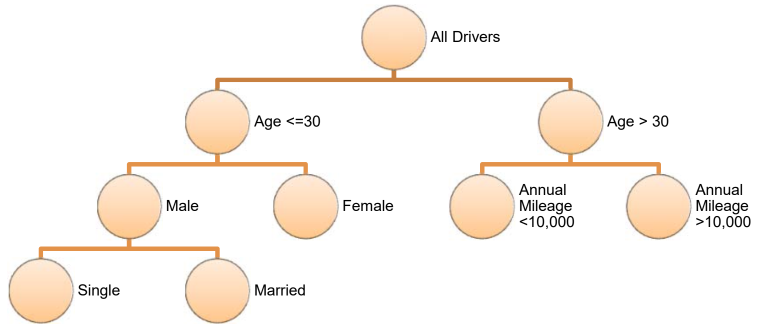
Applications of Decision Trees

- Enhancing GLMs
  - Screening predictor variables
  - Analyzing residuals
  - Identifying transformations and/or interactions
- Portfolio diagnostics
- Checking or quality control

Decision Trees – Finding Interactions

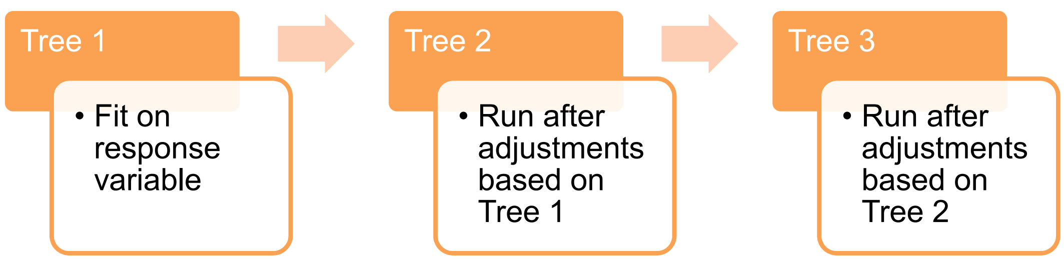
Decision Tree automatically captures interactions

- Two explanatory variables ***interact*** if they combine non-additively to affect the target
- Traditional regression requires an explicit interaction term identified upfront
- A Decision Tree of depth D can capture interactions of order up to D



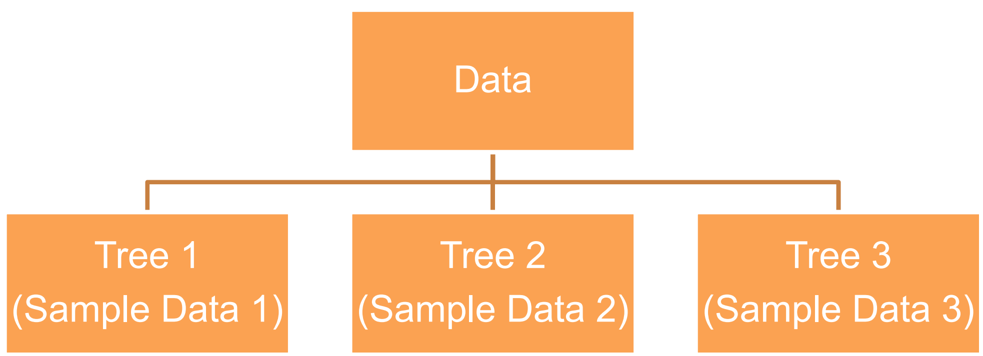
Boosting

- **Gradient Boosting**
  - Models built sequentially
  - New model built on the residual
- **AdaBoost**
  - Adaptive boosting
  - Iteratively changes weights of training observations based on errors of previous prediction



Bagging

- **Bootstrap Aggregation**
  - Bootstrapped samples of the data
  - Models built in parallel
  - Result based on average of the model predictions
- **Random Forest**
  - Bootstrap Aggregation
  - Sampling of Features during tree creation



Exploratory Data Set

Target: Collision Claim during Policy Year		
Policy Year State Policy Term	Tenure Multi-Policy Vehicle Count Driver Count	Deductible Age Gender Marital Status Model Year

Benchmarking with Random Forest

	TRAIN		VALIDATION	
Model Type	**Area Under ROC**	**GINI Coefficient**	**Area Under ROC**	**GINI Coefficient**
HP Tree B2D10	0.644	0.287	0.597	0.193
HP Forest	0.609	0.218	0.590	0.180
HP Tree B3D5	0.609	0.219	0.583	0.165
HP GLM Step	0.582	0.163	0.575	0.149
HP GLM Full	0.582	0.164	0.575	0.149

Exploratory Model Gini Coefficients

	TRAIN		VALIDATION	
Model Type	Area Under ROC	GINI Coefficient	Area Under ROC	GINI Coefficient
HP GLM Full	0.582	0.164	0.575	0.149
HP GLM Step	0.582	0.163	0.575	0.149
HP GLM N Step	0.582	0.164	0.575	0.150
HP GLM N I Step	0.582	0.164	0.575	0.150
HP GLM N P3 Step	0.601	0.202	0.585	0.171
HP GLM N P3 I Step	0.601	0.202	0.585	0.171
HP GLM C Step	0.613	0.226	0.588	0.176
HP Forest	0.609	0.218	0.590	0.180

Decision Tree (DT B2D10) Variable Importance

Variable Name	Number of Splitting Rules	Sum of Square Errors	Importance	Validation Sum of Square Errors	Validation Importance
raw_tenure	24	6.235	1.000	2.895	0.935
raw_age2	61	5.706	0.915	2.928	0.946
raw_modelyear	44	4.958	0.795	3.096	1.000
raw_veh_count2	14	4.391	0.704	2.434	0.786
raw_pol_eff_year	33	3.461	0.555	1.822	0.588
raw_state	13	3.203	0.514	1.446	0.467
raw_drv_count2	10	2.198	0.352	1.230	0.397
raw_female	11	1.951	0.313	0.795	0.257
raw_married	4	1.709	0.274	0.638	0.206
raw_ded_coll	7	1.596	0.256	0.662	0.214
raw_homeauto	1	0.857	0.138	0.271	0.087
raw_term_annual	2	0.538	0.086	0.000	0.000

Random Forest Variable Importance

Variable Name	Number of Splitting Rules	Gini Reduction	Margin Reduction	OOB Gini Reduction	OOB Margin Reduction
raw_modelyear	112	3.88E-05	7.77E-05	1.45E-05	9.42E-06
raw_veh_count2	106	4.40E-05	8.81E-05	2.58E-05	3.64E-05
raw_tenure	94	7.18E-05	1.44E-04	4.05E-05	1.07E-04
raw_age2	70	2.78E-05	5.55E-05	7.85E-06	6.24E-05
raw_married	65	3.87E-05	7.74E-05	2.14E-05	8.87E-05
raw_pol_eff_year	60	8.23E-06	1.65E-05	5.19E-07	1.47E-05
raw_female	51	5.44E-06	1.09E-05	1.22E-06	-3.84E-05
raw_state	46	1.31E-05	2.62E-05	-6.59E-08	-1.15E-05
raw_drv_count2	44	7.15E-06	1.43E-05	3.12E-06	1.25E-05
raw_term_annual	*42*	*6.56E-06*	*1.31E-05*	*1.61E-06*	*1.77E-05*
raw_ded_coll	*41*	*4.13E-06*	*8.27E-06*	*-9.74E-07*	*-7.76E-06*
raw_homeauto	*17*	*1.84E-06*	*3.67E-06*	*-4.47E-07*	*-1.40E-05*

PINNACLE ACTUARIAL RESOURCES, INC.

Type 3 Results from GLM Model

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
raw_age2	1	35.37	<.0001
raw_ded_coll	*1*	*36.09*	*<.0001*
raw_drv_count2	1	57.16	<.0001
raw_female	1	6.96	0.0083
raw_homeauto	1	0.09	0.767
raw_married	1	80.69	<.0001
raw_modelyear	1	98.29	<.0001
raw_pol_eff_year	1	23.7	<.0001
raw_state	17	82.91	<.0001
raw_tenure	1	32.81	<.0001
raw_term_annual	1	1.22	0.2699
raw_veh_count2	1	98.18	<.0001

Type 3 Results for Coll Ded (as Char Var)

Analysis Of Maximum Likelihood Parameter Estimates									
Parameter		DF	ß	SD Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq	Exp (ß)
raw_ded_coll	100	1	-0.0087	0.0957	-0.1963	0.1788	0.01	0.9273	0.991
raw_ded_coll	200	1	0.1690	0.0626	0.0463	0.2917	7.29	0.0069	1.184
raw_ded_coll	250	1	0.1189	0.0333	0.0536	0.1842	12.75	0.0004	1.126
raw_ded_coll	500	0	0	0	0	0	.	.	1.000
raw_ded_coll	1000	1	-0.2131	0.0459	-0.3031	-0.1232	21.58	<.0001	0.808

Leveraging the Strengths of Random Forests

- Easy to set up and run
- Strong results with minimal adjustments
- Resistance to overfitting
- Invariant to monotonic transformations

LAZY LEARNING

Regression using nearest neighbors

Donald F.J. Hendriks, ACAS, ASA, FCA, MAAA
CARFAX Banking & Insurance Group

2019 Casualty Actuarial Society Annual Meeting

Lazy Learning

<u>Lazy Learning Algorithm</u> A machine learning algorithm in which no abstraction occurs

- Nonparametric
  - Purely deterministic in nature
  - Target is based on only the data put into the learning system
- Allows regression or classification of new observations based on existing classification system
- Learns as it goes
  - Fast to train
  - Slow to predict

Nearest Neighbors Regression Algorithm

Nearest Neighbors Regression Algorithm

1. New observation
2. Compare features to existing observations
3. Select the k most similar observations
4. Take the average of the k observations

- Uses *similarity* to determine appropriate comparisons
- Assumes the new observation's outcome will be similar to other observations with similar features

Measuring Similarity with Distance

DISTANCE FUNCTIONS

- **Euclidian Distance**

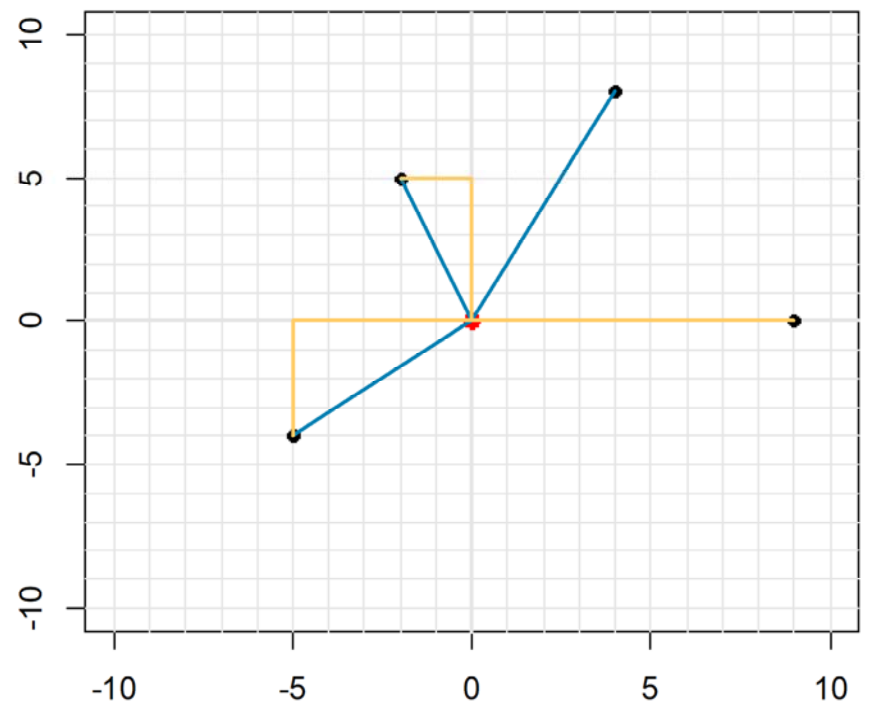
$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

  - Most common default

- **Manhattan Distance**

$$d = \sum_{i=1}^{n} |p_i - q_i|$$

  - May be preferable on high-dimensional data



The Nearest Neighbor Algorithm

STRENGTHS

- Simple and effective
- Requires no assumptions regarding underlying data
- Fast to train
- Can be applied to regression or classification
- Well suited for multi-class problems

WEAKNESSES

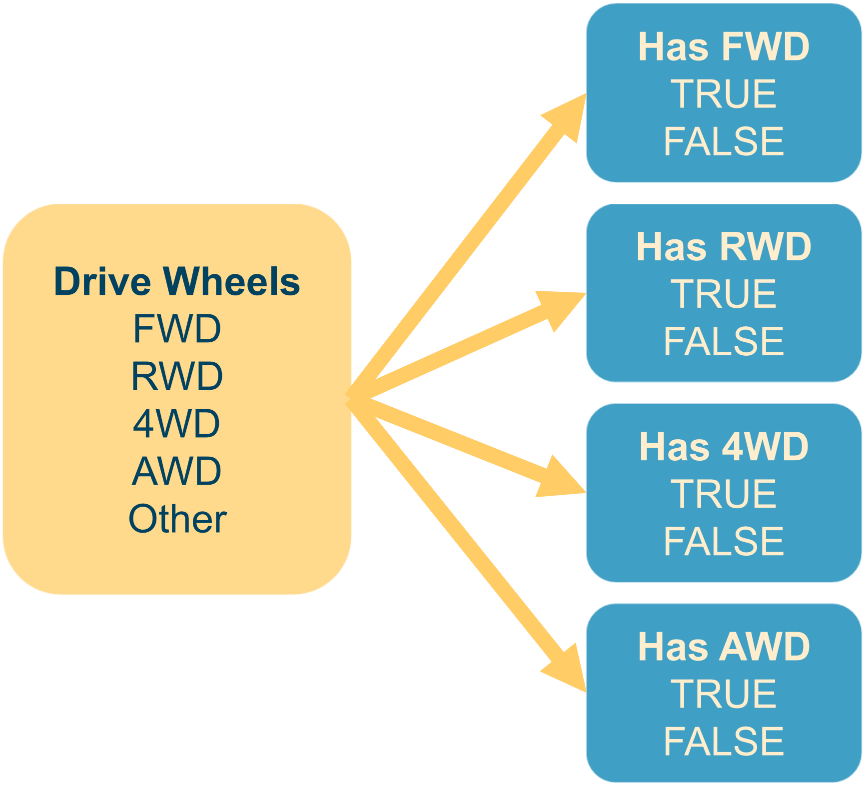
- Does not produce a model
- Requires selection of appropriate k value
- Subject to scale biases
- Slow to predict, memory intensive algorithm
- Ill-suited for rare-event target data

Nearest Neighbors Example: Estimating Mileage

- Target variable: Annual miles driven per year, `AvgAnnMiles`
- Database of 469,469 vehicles and vehicle characteristics (26 variables)
  - Year, make, model, body style, engine
  - Number of owners, length of ownership, and registration type
  - Vehicle maintenance and damage history
  - Most recent mileage reading
- Demographic data for garaging ZIP Code (33 variables)
  - Population density, household types, places of employment
  - Drive times, percent employed, type of employment
  - School enrollment and educational attainment

Dealing with categorical variables

- One-hot encoding
  - Converts categorical variables into ‘dummy’ variables.
  - Greatly increases dimensionality.
- Simple to perform with most modeling software.
- Always leave a category out to prevent multicollinearity.
- Good time to consider related categories.



CARFAX

Banking & Insurance Group®
MORE INFORMATION. BETTER DECISIONS.

AVERAGE ANNUAL MILES

An example in R

1. Exploring vehicle data

Categorical Data

- `ZIP`: Registration ZIP code
- `Make`: Vehicle make
- `EngDisp`: Engine displacement (L)
- `EngConfig`: Engine configuration
- `BodyStyle`: Vehicle body style
- `Trim`: Vehicle trim description
- `Color`: Vehicle color
- `DriveWheels`: Drive system
- `CurrOwnershipType`: Current Registration Type
- `HistOwnershipType`: Current Registration Type

Numerical Data

- `Date`: Evaluation date of data
- `Cylinders`: Number of cylinders
- `LastOdometer`: Most recent odometer reading for the current owner
- `LengthOwnership`: Days since most recent title event
- `Retail`: Retail value of vehicle

1. Exploring vehicle data

Logical (Boolean) Vehicle History Data

- SevereProblem
- BrandedTitle
- StructuralDamage
- NonsevereAccident
- Damage
- FailedEmissions
- FailedSafety
- OdometerProblem
- Rollback
- Export
- Lien
- Repossessed
- Stolen
- PriorCPO
- ServicedFar
- ServiceHist
- RegOilChg
- OpenRecall

1. Exploring demographic data

Population and Commute (ZIP Code Level)

SQMILES	ZIP Code Area (sq. mi.)	**TRANCAR1**	PoP driving to work alone
DENSITY	Population density	**TRANCARP**	PoP carpooling to work
POP17	Population (1/1/2017)	**TRANPUBLIC**	PoP taking public transit to work
POPGROW17	Population growth (2017/2010)	**TRANWALKBIKE**	PoP walking or biking to work
POPFORE22	Population forecast (2022/2017)	**TRAVHOME**	PoP working from home
AVGHHSIZE	Average household size	**TRAVL15**	PoP commute < 15 minutes
URBAN.PCT	PoP in urban area	**TRAV15.29**	PoP commute 15-29 minutes
MEDAGE	Median population age	**TRAV30.59**	PoP commute 30-59 minutes
POP.65P.PCT	PoP aged 65 years or older	**TRAV60.89**	PoP commute 60-89 minutes

1. Exploring demographic data

Education, Employment and Other (ZIP Code Level)

SE.K.12	PoP enrolled in K - 12 school	**EMP.LABFRC**	PoP in labor force
SE.COLL	PoP enrolled in college or university	**EMP.UNEMP**	PoLF unemployed
ED.LHS	PoP without high school education	**WHCOLROCC**	PoLF in white collar jobs
ED.HS	PoP with high school education	**EMP.SELF2**	PoLF self-employed
ED.COLL.DEG	PoP with bachelor's degree or higher	**EMP.GOVT**	PoLF in government jobs
MEDAGHHER	Median Householder Age	**POV.TOTAL**	PoP in poverty
MEDVEHICLE	Median Vehciles per Household	**VET.TOTAL**	PoP veterans
VEH.0	Percent of Household with no vehicle		

1. Exploring Data

```
> summary(CFX.data)
     ZIP                 Make              CurrOwnershipType    LengthOwnership
         : 30738   CHEVROLET: 96113  Personal       :366406   Min.   :    1
02816    :  4549   FORD     : 81631                 : 64227   1st Qu.:  552
02914    :  2991   TOYOTA   : 35866  Personal lease: 16926    Median : 1292
02861    :  2869   DODGE    : 33557  Commercial Use:  9134    Mean   : 1761
02895    :  2627   HONDA    : 25973  Corporate      :  5734   3rd Qu.: 2506
02893    :  2618   BUICK    : 19549  Rental         :  3316   Max.   :12388
(Other):423077   (Other)  :176780  (Other)        :  3726   NA's   :23319
 ServiceHist       RegOilChg         AvgAnnMiles
Mode :logical   Mode :logical   Min.   :   200
FALSE:397697    FALSE:469452    1st Qu.:  7247
TRUE :71772     TRUE :17        Median : 11626
                                Mean   : 12505
                                3rd Qu.: 16678
                                Max.   : 59956
                                NA's   :119350
```

1. Exploring Data

```
> summary(CFX.data)
     ZIP                Make               CurrOwnershipType   LengthOwnership
         : 30738   CHEVROLET: 96113   Personal        :366406   Min.   :     1
02816    :  4549   FORD     : 81631                   : 64227   1st Qu.:   552
02914    :  2991   TOYOTA   : 35866   Personal lease: 16926   Median : 1292
02861    :  2869   DODGE    : 33557   Commercial Use:   9134   Mean   : 1761
02895    :  2627   HONDA    : 25973   Corporate     :   5734   3rd Qu.: 2506
02893    :  2618   BUICK    : 19549   Rental        :   3316   Max.   :12388
(Other):423077   (Other)  :176780   (Other)       :   3726   NA's   :23319
 ServiceHist      RegOilChg        AvgAnnMiles
Mode :logical   Mode :logical   Min.   :    200
FALSE:397697    FALSE:469452    1st Qu.:   7247
TRUE :71772     TRUE :17        Median : 11626
                                Mean   : 12505
                                3rd Qu.: 16678
                                Max.   : 59956
                                NA's   :119350
```

1. Exploring Data

```
> summary(CFX.data)
      ZIP                Make              CurrOwnershipType   LengthOwnership
          : 30738   CHEVROLET: 96113   Personal        :366406   Min.   :    1
02816     :  4549   FORD     : 81631                   : 64227   1st Qu.:  552
02914     :  2991   TOYOTA   : 35866   Personal lease: 16926   Median : 1292
02861     :  2869   DODGE    : 33557   Commercial Use:   9134   Mean   : 1761
02895     :  2627   HONDA    : 25973   Corporate       :   5734   3rd Qu.: 2506
02893     :  2618   BUICK    : 19549   Rental          :   3316   Max.   :12388
(Other):423077   (Other)  :176780   (Other)         :   3726   NA's   :23319
 ServiceHist      RegOilChg        AvgAnnMiles
Mode :logical   Mode :logical   Min.   :   200
FALSE:397697    FALSE:469452    1st Qu.:  7247
TRUE :71772     TRUE :17        Median : 11626
                                Mean   : 12505
                                3rd Qu.: 16678
                                Max.   : 59956
                                NA's   :119350
```

CARFAX Banking & Insurance Group® MORE INFORMATION. BETTER DECISIONS.

PINNACLE ACTUARIAL RESOURCES, INC.

1. Exploring Data

```
> summary(CFX.data)
     ZIP                 Make             CurrOwnershipType    LengthOwnership
         : 30738   CHEVROLET: 96113   Personal      :366406   Min.   :    1
02816    :  4549   FORD     : 81631                 : 64227   1st Qu.:  552
02914    :  2991   TOYOTA   : 35866   Personal lease: 16926   Median : 1292
02861    :  2869   DODGE    : 33557   Commercial Use:  9134   Mean   : 1761
02895    :  2627   HONDA    : 25973   Corporate     :  5734   3rd Qu.: 2506
02893    :  2618   BUICK    : 19549   Rental        :  3316   Max.   :12388
(Other):423077   (Other)  :176780   (Other)       :  3726   NA's   :23319
 ServiceHist      RegOilChg        AvgAnnMiles
Mode :logical   Mode :logical   Min.   :   200
FALSE:397697    FALSE:469452    1st Qu.:  7247
TRUE :71772     TRUE :17        Median : 11626
                                Mean   : 12505
                                3rd Qu.: 16678
                                Max.   : 59956
                                NA's   :119350
```

1. Exploring Data

```
> summary(CFX.data)
      ZIP                 Make              CurrOwnershipType    LengthOwnership
          : 30738   CHEVROLET: 96113   Personal        :366406    Min.   :     1
02816     :  4549   FORD     : 81631                   : 64227    1st Qu.:   552
02914     :  2991   TOYOTA   : 35866   Personal lease: 16926    Median : 1292
02861     :  2869   DODGE    : 33557   Commercial Use:  9134    Mean   : 1761
02895     :  2627   HONDA    : 25973   Corporate       :  5734    3rd Qu.: 2506
02893     :  2618   BUICK    : 19549   Rental          :  3316    Max.   :12388
(Other):423077   (Other)  :176780   (Other)         :  3726    NA's   :23319
  ServiceHist        RegOilChg          AvgAnnMiles
Mode  :logical    Mode  :logical    Min.   :    200
FALSE:397697     FALSE:469452     1st Qu.:   7247
TRUE  :71772      TRUE  :17          Median :  11626
                                     Mean   :  12505
                                     3rd Qu.:  16678
                                     Max.   :  59956
                                     NA's   :119350
```

1. Exploring Data

```
> summary(MDDB.data)
     ZIP              SQMILES            DENSITY               POP17           POPGROW17
 Length:11176       Min.   :   0.00   Min.   :     0.0   Min.   :     0   Min.   :-37.920
 Class :character   1st Qu.:  11.83   1st Qu.:    33.0   1st Qu.:  1421   1st Qu.: -1.780
 Mode  :character   Median :  35.36   Median :   130.7   Median :  5508   Median :  0.710
                    Mean   :  65.20   Mean   :  1469.2   Mean   : 12854   Mean   :  1.881
                    3rd Qu.:  77.23   3rd Qu.:  1101.5   3rd Qu.: 20288   3rd Qu.:  4.143
                    Max.   :3277.18   Max.   :131914.8   Max.   :115933   Max.   :103.570

    TRANCAR1          TRANCARP           TRAVL15          MEDHHSIZE        SE.K.12
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.3961   1st Qu.:0.03979   1st Qu.:0.1077   1st Qu.:2.500   1st Qu.:0.2171
 Median :0.4525   Median :0.05163   Median :0.1462   Median :2.600   Median :0.2442
 Mean   :0.4421   Mean   :0.05450   Mean   :0.1605   Mean   :2.648   Mean   :0.2457
 3rd Qu.:0.5003   3rd Qu.:0.06586   3rd Qu.:0.1980   3rd Qu.:2.800   3rd Qu.:0.2707
 Max.   :1.0000   Max.   :0.29144   Max.   :0.6514   Max.   :5.000   Max.   :1.0000
 NA's   :36       NA's   :36        NA's   :36       NA's   :36      NA's   :36
```

1. Exploring Data

```
> summary(MDDB.data)
    ZIP                SQMILES            DENSITY               POP17          POPGROW17
 Length:11176       Min.   :   0.00   Min.   :     0.0   Min.   :     0   Min.   :-37.920
 Class :character   1st Qu.:  11.83   1st Qu.:    33.0   1st Qu.:  1421   1st Qu.: -1.780
 Mode  :character   Median :  35.36   Median :   130.7   Median :  5508   Median :  0.710
                    Mean   :  65.20   Mean   :  1469.2   Mean   : 12854   Mean   :  1.881
                    3rd Qu.:  77.23   3rd Qu.:  1101.5   3rd Qu.: 20288   3rd Qu.:  4.143
                    Max.   :3277.18   Max.   :131914.8   Max.   :115933   Max.   :103.570

   TRANCAR1          TRANCARP           TRAVL15          MEDHHSIZE         SE.K.12
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.3961   1st Qu.:0.03979   1st Qu.:0.1077   1st Qu.:2.500   1st Qu.:0.2171
 Median :0.4525   Median :0.05163   Median :0.1462   Median :2.600   Median :0.2442
 Mean   :0.4421   Mean   :0.05450   Mean   :0.1605   Mean   :2.648   Mean   :0.2457
 3rd Qu.:0.5003   3rd Qu.:0.06586   3rd Qu.:0.1980   3rd Qu.:2.800   3rd Qu.:0.2707
 Max.   :1.0000   Max.   :0.29144   Max.   :0.6514   Max.   :5.000   Max.   :1.0000
 NA's   :36       NA's   :36        NA's   :36       NA's   :36      NA's   :36
```

2. Preparing Data

- One-hot encode categorical data
- Newest Neighbors is sensitive to scale
- Some data should be normalized
  - Percentages – already defined on [0, 1]
  - Factor (dummy) variables – already defined on [0, 1]
  - Other numerical variables – Convert to Z-score:

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

- Normalized data can be weighted if needed for further refinement
- Output must be well-defined for all observations in training set

2. Preparing Data

CARFAX Data

- Remove data with no target variable.
- One-hot encode categorical data.
  `ade4: acm.disjonctif`
- Remove sparse data.

Demographic Data

- Remove 39 weird ZIPs.

```
> # Move data with no target to a validation set
> CFX.NoTarget <- CFX.data[
+                     is.na(CFX.data$AvgAnnMiles), ]
> CFX.data <- CFX.data[
+                     !is.na(CFX.data$AvgAnnMiles), ]
>
> # One-hot encode categorical data
> library (ade4)
> CFX.data <- cbind (CFX.data,
+              acm.disjonctif (CFX.data[, .(Color)]))
...
> # Remove sparse CARFAX data
> CFX.data$Color.Orange <- NULL
> CFX.data$COT.Police <- NULL
  ...
> # Remove ZIPs with no information
> MDDB.data <- MDDB.data[
+                   !is.na (MDDB.dta$TRANCAR1), ]
```

CARFAX
Banking & Insurance Group®
MORE INFORMATION. BETTER DECISIONS.

PINNACLE
ACTUARIAL RESOURCES, INC.

2. Preparing Data

CARFAX Data

- Convert all numeric variables to Z-scores.
- Convert logical values to numerical values.
- ‘Unconvert’ target variable.

```
> # Convert all CFX numeric variables to Z-scores
> num.field <- unlist (lapply (CFX.data, is.numeric))
> CFX.scaled <- cbind (CFX.data[, !num.field],
+                      scale (CFX.data[, num.field]))
>
> # Convert all CFX logical variables to integers (1 or 0)
>   log.field <- unlist (lapply (CFX.data, is.logical))
>   CFX.scaled[log.field] <- apply (CFX.scaled[log.field],
+                                   2,
+                                   function (x)
+                                      as.integer (x))
>
> # Replace target variable with de-scaled version
> CFX.scaled$AvgAnnMiles <- CFX.data$AvgAnnMiles
>
```

2. Preparing Data

- Merge your tables.
  - Leaves us with 246,735 obserations.
- Split into training and validation datasets.
- Segregate the target from the training set.

```
> # Merge feature data into a single dataset
> all.data <- merge (x = CFX.data, by.x = 'ZIP',
+                    y = MDDB.data, by.y = 'ZIP')
>
> # After merging data, ZIP is not needed
> all.data$ZIP <- NULL
>
> # Split into training and validation datasets
> train <- sample (1:nrow(all.data),
+                  size = floor (nrow(all.data) * .75))
> train.data <- all.data[train, ]
> validate.data <- all.data[-train, ]
>
> # Remove target from training set and store
> train.target <- train.data$AvgAnnMiles
> train.data$AvgAnnMiles <- NULL
```

3. Run the model

- Run the model for several values of k.
- Once the data is fully prepared, building the model is easy.

Generate results by feeding the model data through the `test` parameter.

```
> library (FNN)   # watch the CAPS!
>
> NN.05 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 5)
> NN.10 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 10)
> NN.20 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 20)
> NN.30 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 30)
> NN.40 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 40)
> NN.50 <- knn.reg (train = train.data, test = test.data,
+                   y = train.target, k = 50)
```

3. Choosing the Best k Value

- Measure of fit – Root mean squared error

$$RMSE = \sqrt{\frac{\sum(x_i - \widehat{x_i})^2}{N}}$$

```
rmse <- function (predicted, actual) {
   sqrt(mean((predicted - actual)^2)) }
```

- Run additional models near your best.
- Other measures of fit may be preferable.

Model	k	RMSE
NN.05	5	6496.912
NN.10	10	6345.539
NN.11	11	6321.834
NN.12	12	6320.361
NN.13	13	6343.337
NN.15	15	6481.418
NN.20	20	6726.696
NN.30	30	6886.385
NN.40	40	6950.291
NN.50	50	6980.672

3. Other refinements

- Look at removing data that may be non-predictive.
- Consider scaling data in different ways.
  - Convert all values to Z-scores.
  - Convert all variables to [0, w].
  - Weight individual variables.
- Use dimensionality reduction techniques.
- Add sophistication at the neighborhood level.
  - Run a simple generalized linear on the k nearest neighbors.