Predicting Property Sale Prices

Letting clients know what their house is worth

...accurately

Data Science @ Regression Realty, Inc

Akram Sadek, James Goudreault, Rishi Goutam, Srikar Pamidi

Internal Briefing, March 10, 2011

Today's Discussion

Predicting House Prices in Ames, IA

- Data
  - Cleaning
  - Feature Engineering
- Notable Findings
- Predictive Models
  - Accuracy Scores
  - Model Complexity
  - Statistical Validation
  - Hyperparameters
- Q&A

Data

Cleaning and Feature Engineering

The data - cleaning

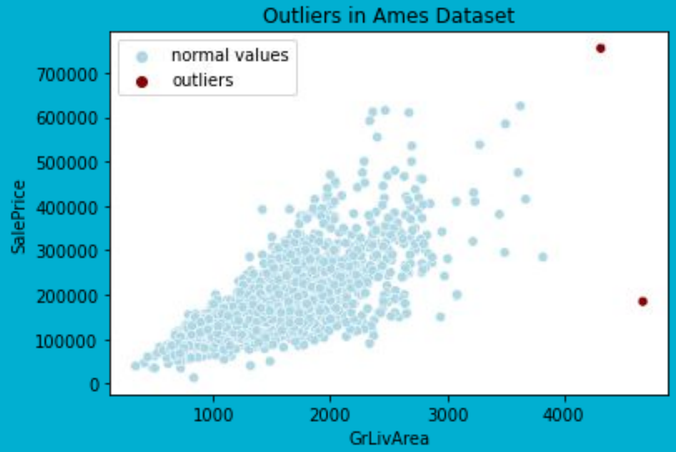
Dataset of **2580** properties sold in Ames, Iowa from 2006-2010

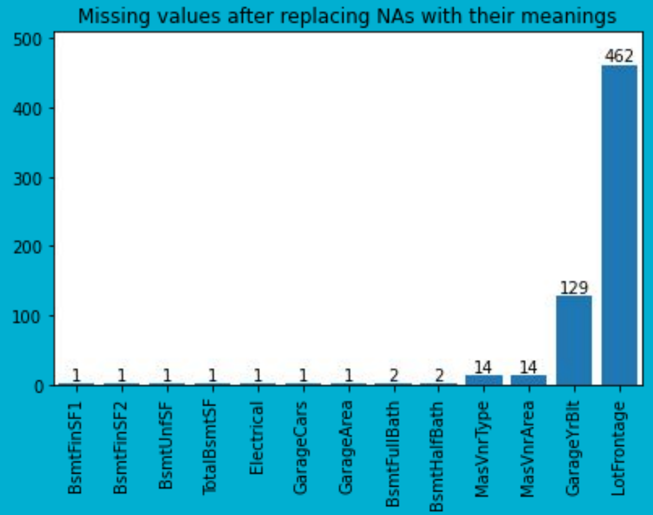
80+ features From Ames City Assessor's Office (not traditional MLS data sources)

Cleaning *Dropped 1 duplicate*

Removed 2 outliers

Imputed missing values by mean/median/mode of group or 0 where value might not exist





The data - feature engineering

New features

- The **school district** a property lies in (Edwards, Fellows, Mitchell, Meeker, Sawyer, *Unknown*)
- **Interest Rate** (TNX index) for the month the property was sold

Derived features

- **Ordinalize** some categorical features
  (*Qual/*Cond, Neighborhood, etc)
- **Combined** multiple features into single feature
  (StreetAlley, Total Outdoor SF, etc)
- **Collapse** features into smaller set of categories
  (MSSubClass, etc)
- *Others*
  (number of floors, property age, etc)

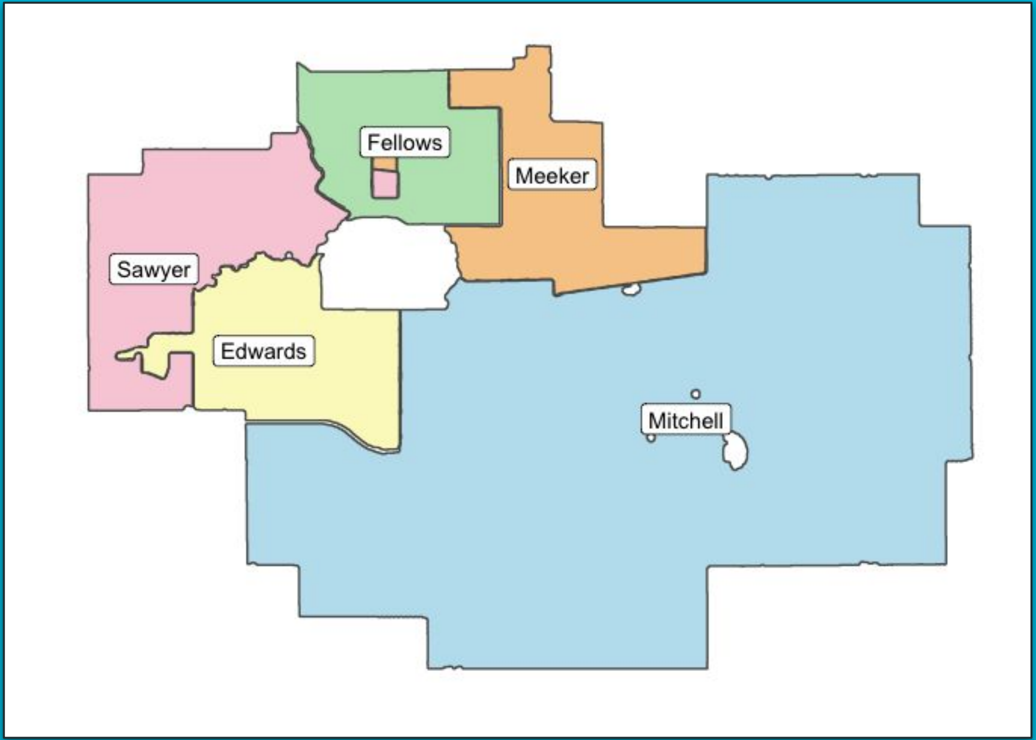
- Boolean Indicators
  Whether a property
  - is near an arterial road or near a railroad
  - is in a Planned Unit Development (PUD)
  - is near a **park**, green-belt, or other positive amenity
  - has been **renovated** or has a **pool**
  - *others*

New Feature: School Districts

Ames has 5 public school districts

Determine if a property's coordinates fall within school district region to get district

*We might expect house prices to vary based on the district**



* School districts found to be highly multicolinear with Neighborhood

Notable Findings

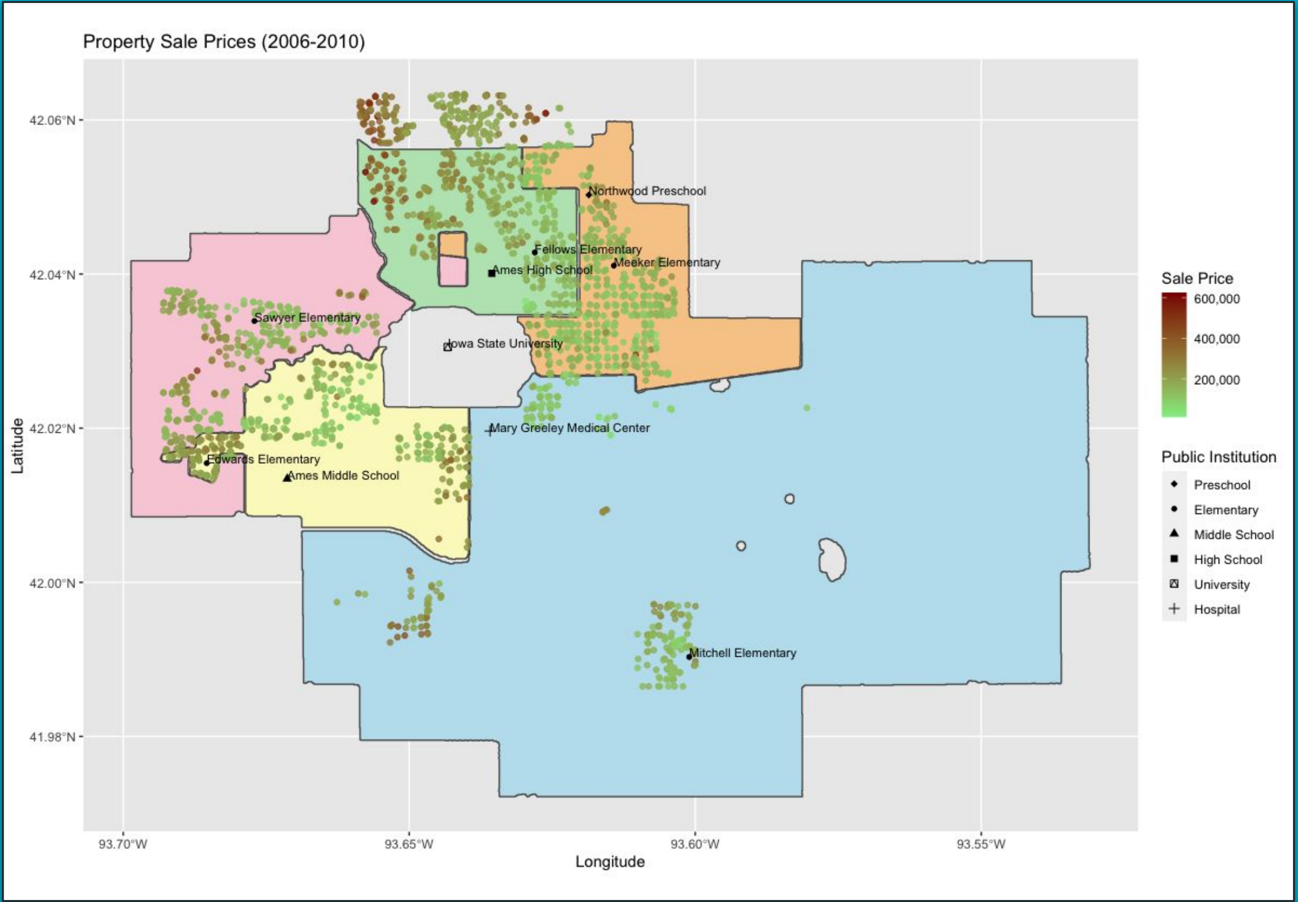
Exploratory data analysis informed feature selection and engineering

Sale Price

Houses at the outskirts of town cost more than those at the center

Perhaps cheaper housing caters to Iowa State University students?

Or, are there other reasons?



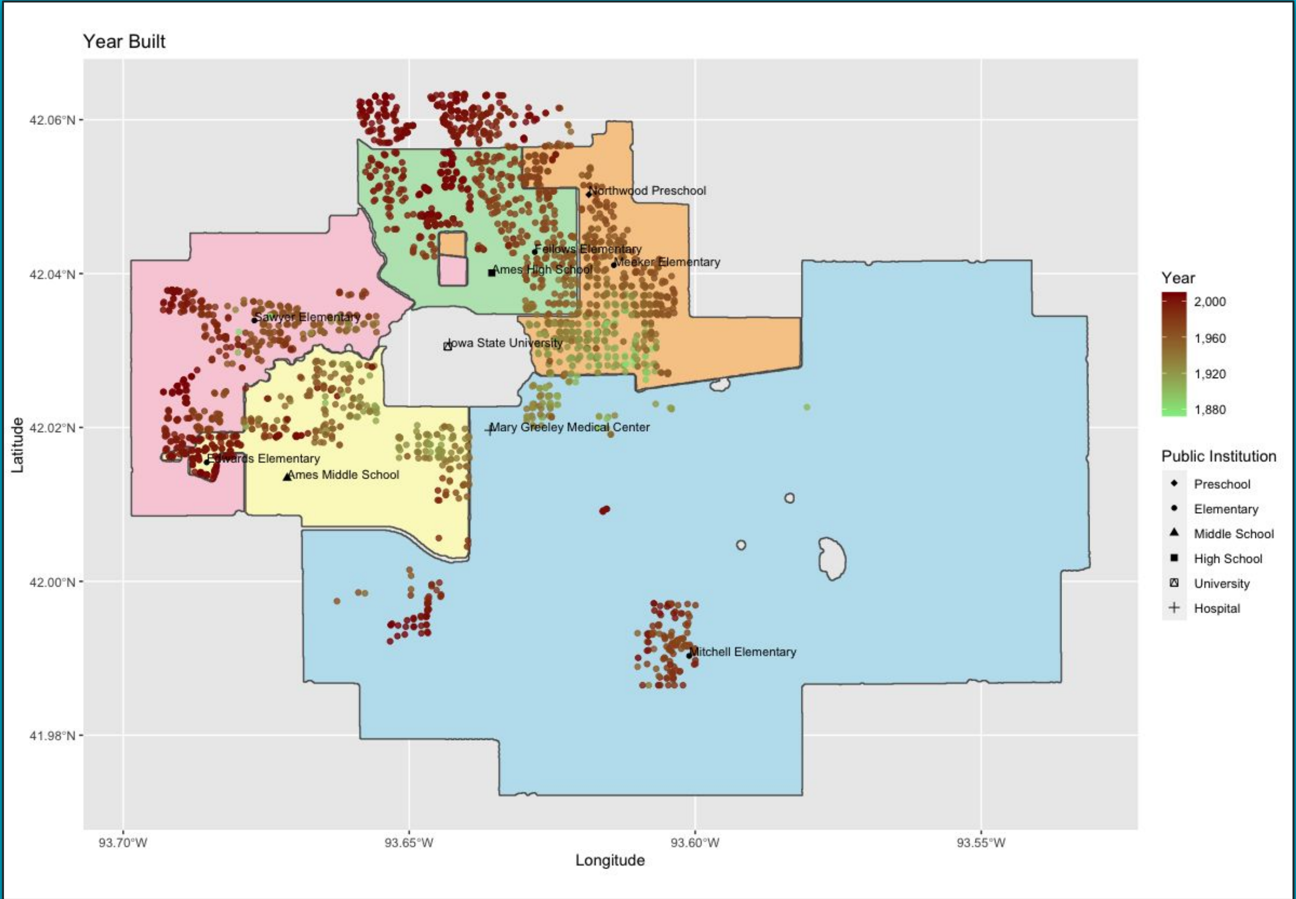
Year Built

Newer houses tend to be more expensive

They are also built on the outskirts, away from the older parts of the city

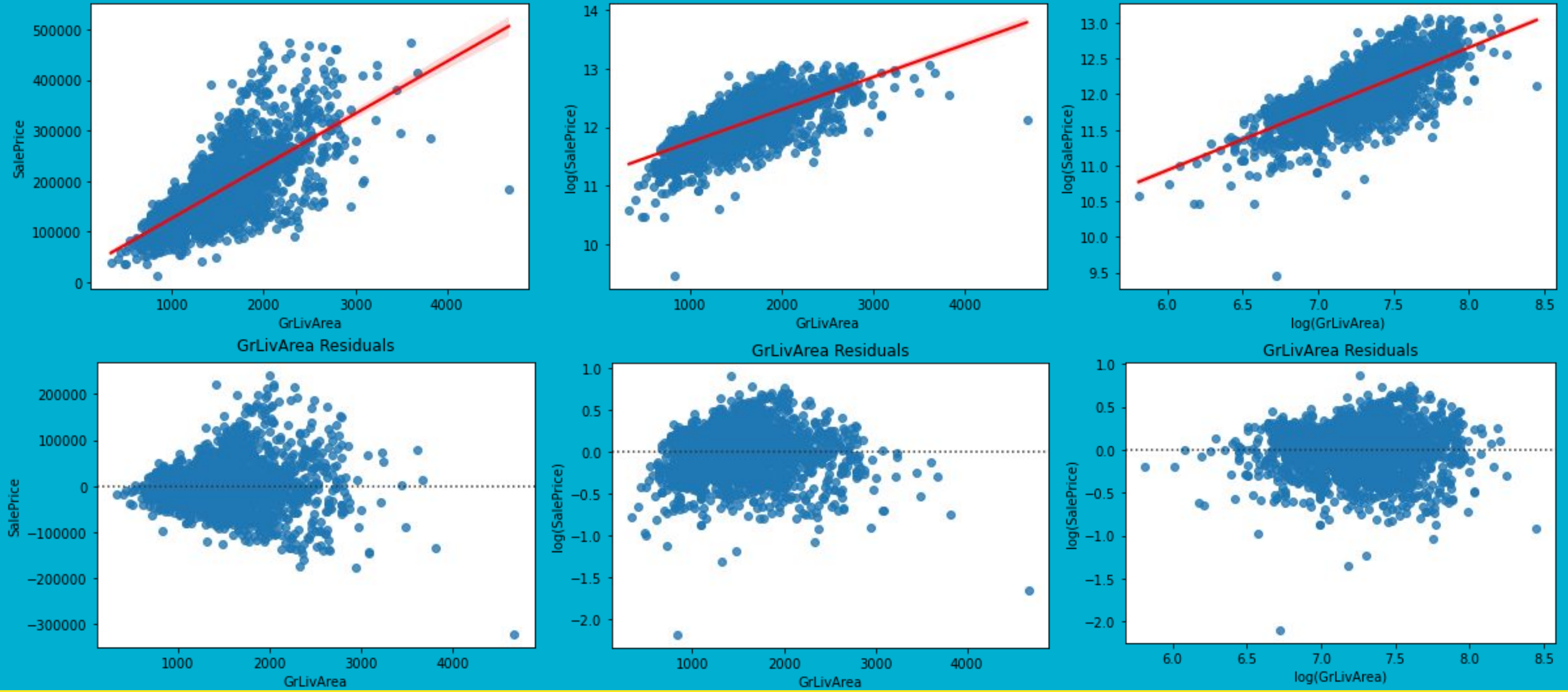
…and away from the university

…not much market incentive to provide students with new housing near ISU



Area features display increasing variance

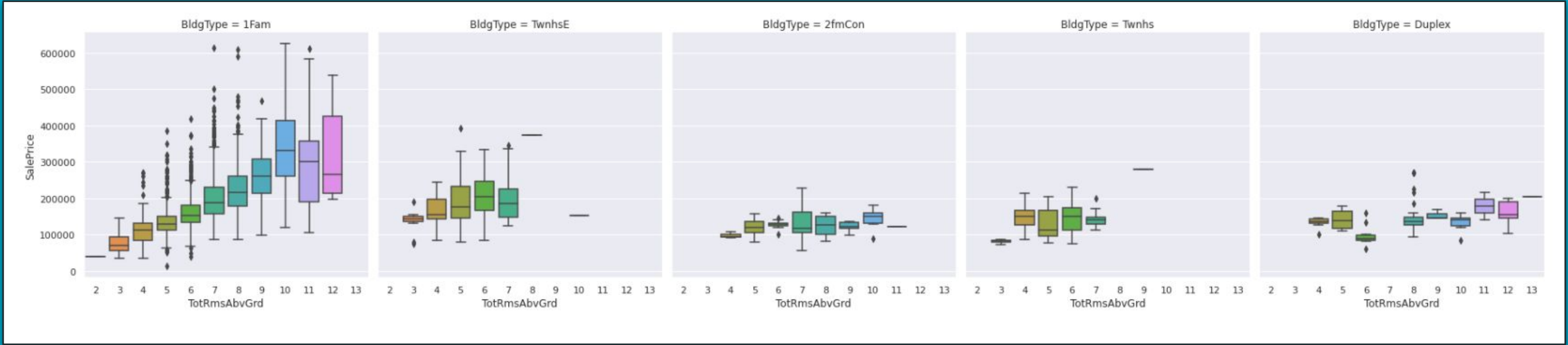
Linear-Linear → big improvement → Log-Linear → some improvement → Log-Log



Size matters...in some cases

The price of a **single family home** is strongly correlated with the **number of rooms** in the home.

Not so with other home types



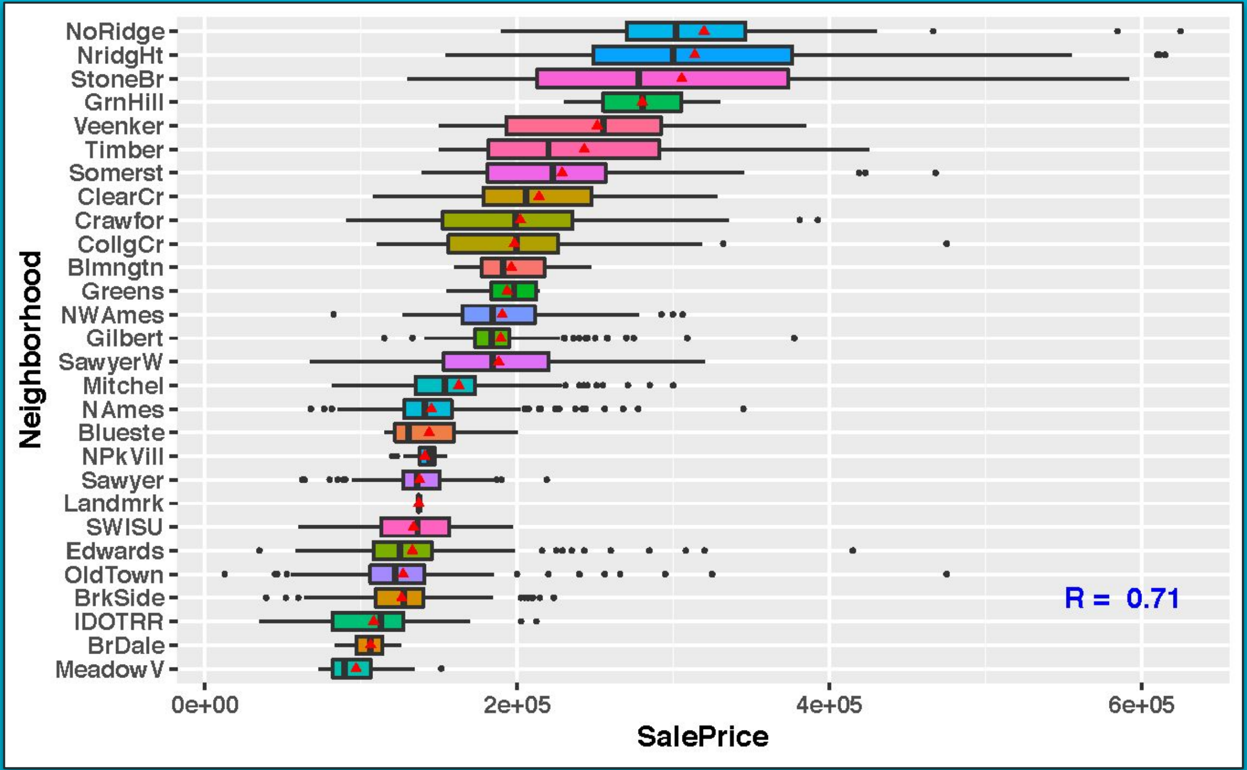
Neighborhood is Strongly Predictive

Neighborhood is highly correlated with home sale price

Neighborhoods were arranged in order of mean sale price to obtain a ranking

Ranking was used to create a new feature to describe each neighborhood as an ordinal integer

(0, 1, 2, … , 26, 27)



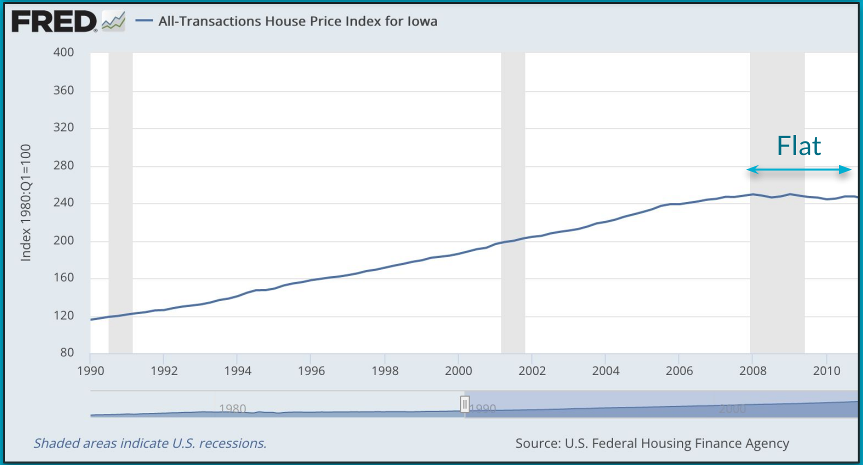
2006-2010: An anomaly?

Ames house prices have been flat in this period…

…instead of showing the steady increase seen across Iowa in the recent past.

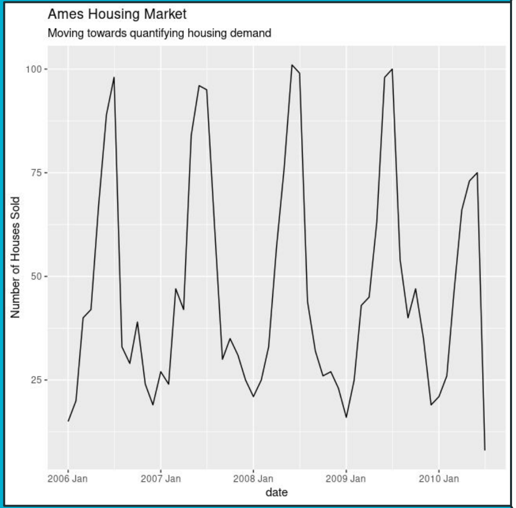
This is despite interest rates dropping after 2008

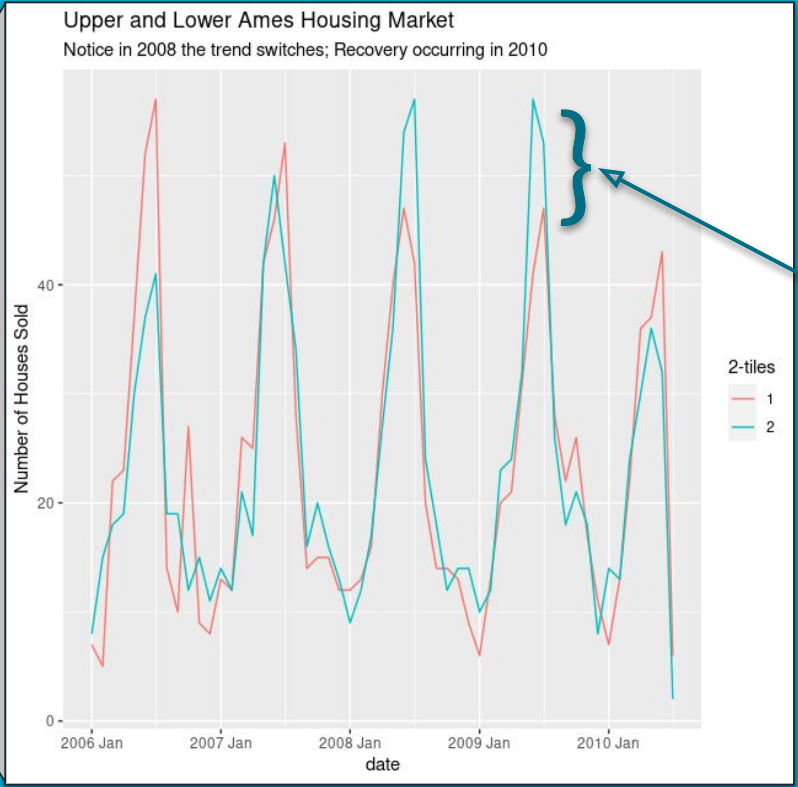
```
> #Ljung-Box Test---We want to reject NH (), looking for p< .05
> test1<- Box.test(pur.ts,type="Ljung-Box", lag= log(nrow(pur.ts))) # p = .99
> test2<- Box.test(o.ts,type="Ljung-Box", lag= log(nrow(o.ts))) # p =.14
> test3<- Box.test(t.ts,type="Ljung-Box", lag= log(nrow(t.ts))) # p = .34
> test4<- Box.test(th.ts,type="Ljung-Box", lag= log(nrow(th.ts))) # p = .94
> test5<- Box.test(fo.ts,type="Ljung-Box", lag= log(nrow(fo.ts))) # p = .33
> test6<- Box.test(fi.ts,type="Ljung-Box", lag= log(nrow(fi.ts))) # p = .31
```



Seasonality and Effect of 2008 Crash

Sales exhibit seasonality…





…with more **summer** sales than sales in **winter**

We see a shift towards more sales of **cheap** houses than **expensive** houses

Are affluent owners hoping to weather the storm?

Predictive Models

Linear Regression & Elastic-Net
Tree Models
SVR
Neural Network
Time Series

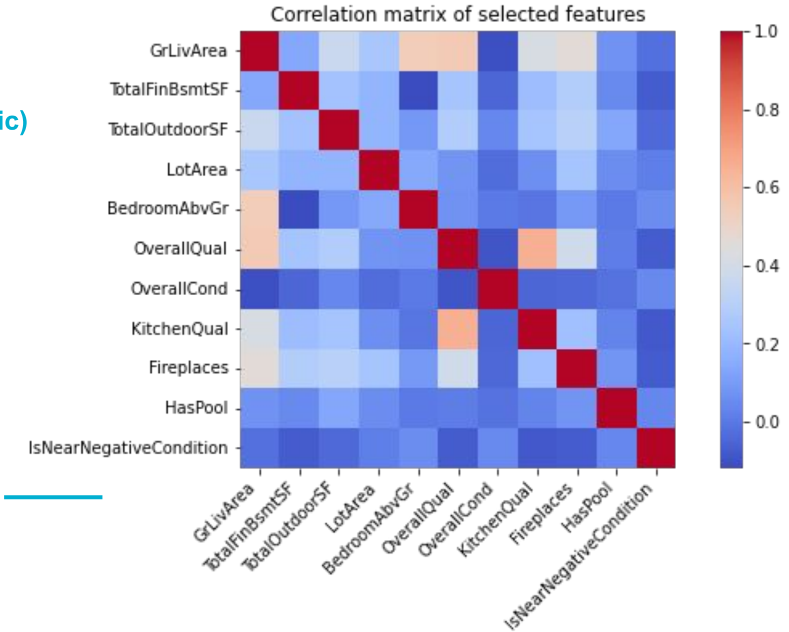
Linear Regression & Elastic-Net

Model Scores + Statistical Validation

Model Scores

	Basic	Final	Elastic-Net
R^2 Train	0.9184	0.9378	0.9331
R^2 Test	0.9020	0.9113	**0.9221**
RMSE	0.1232	0.0501	**0.0462**
AIC†	3489.04	3377.0	229.4
BIC†	13017.63	12652.6	850.7

ρ (Basic)

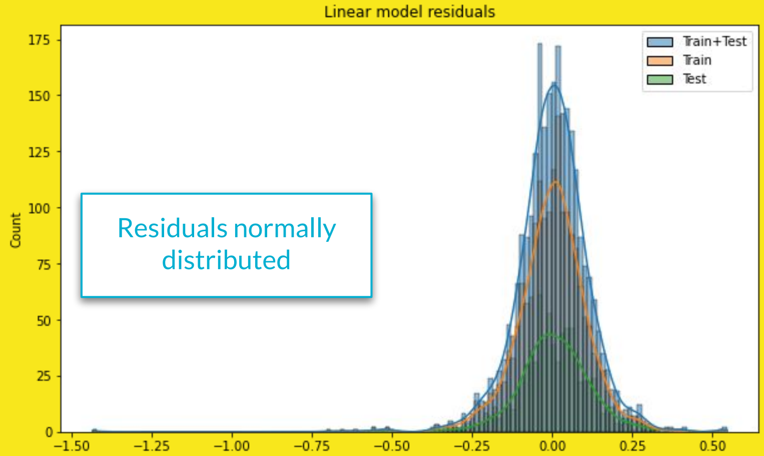


†Used custom AIC/BIC function to apply same function on Elastic-Net as on Linear Regression

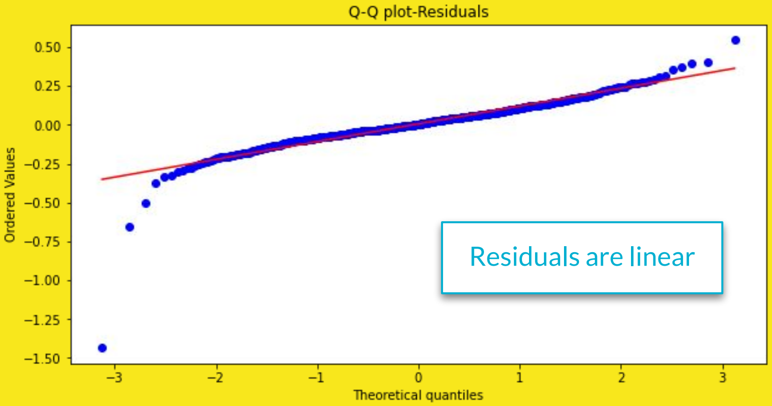
Checking Assumptions
(Basic Model)

vif(model)	GVIF	df	GVIF^(1/(2*Df))
GrLivArea	4.40	1	2.10
TotalFinBsmtSF	1.57	1	1.25
TotalOutdoorSF	1.39	1	1.18
LotArea	1.48	1	1.22
BedroomAbvGr	2.33	1	1.53
MSSubClass	98.42	14	1.18
MSZoning	37.02	6	1.35
OverallQual	3.42	1	1.85
OverallCond	1.48	1	1.22
Neighborhood	2261.42	27	1.15
KitchenQual	2.20	1	1.48
	1.51	5	1.04
	1.21	4	1.02
CentralAir	1.51	1	1.23
Fireplaces	1.72	1	1.31
HasPool	1.06	1	1.03
IsNearNegativeCondition	1.12	1	1.06
LandContour	1.87	3	1.11

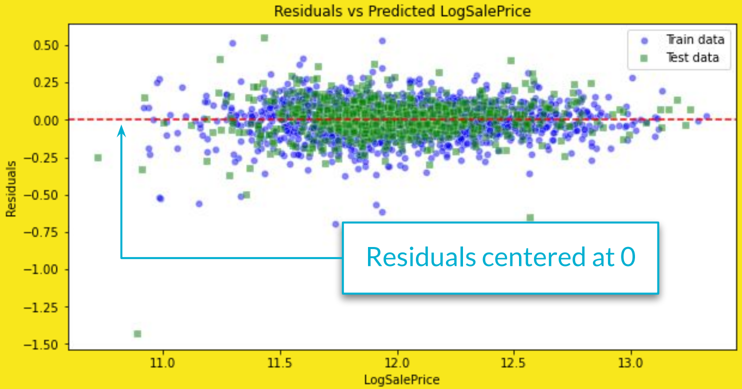
Low multicolinearity



Residuals normally distributed

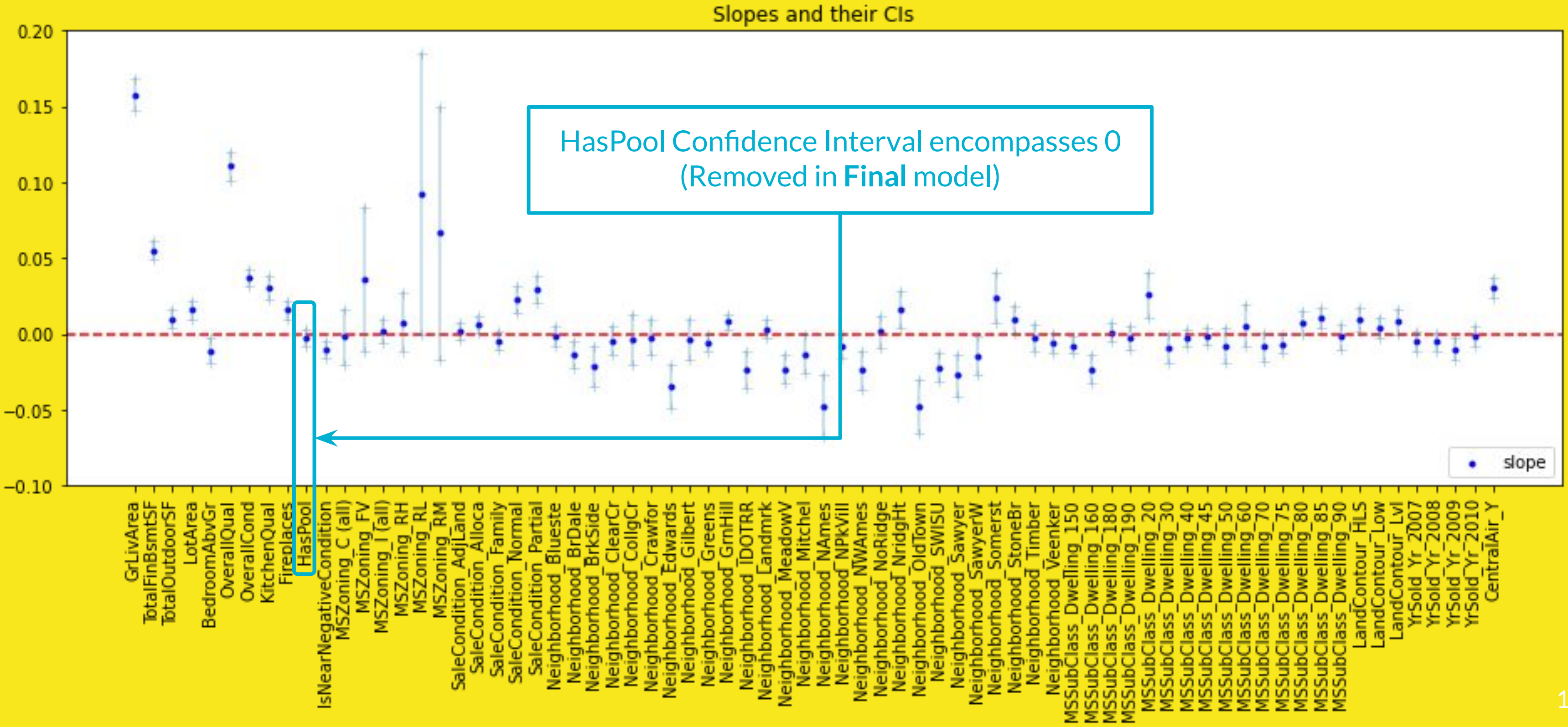


Residuals are linear



Residuals centered at 0

Coefficients
(Basic Model)



Effect of 1-unit change on mean Sale Price

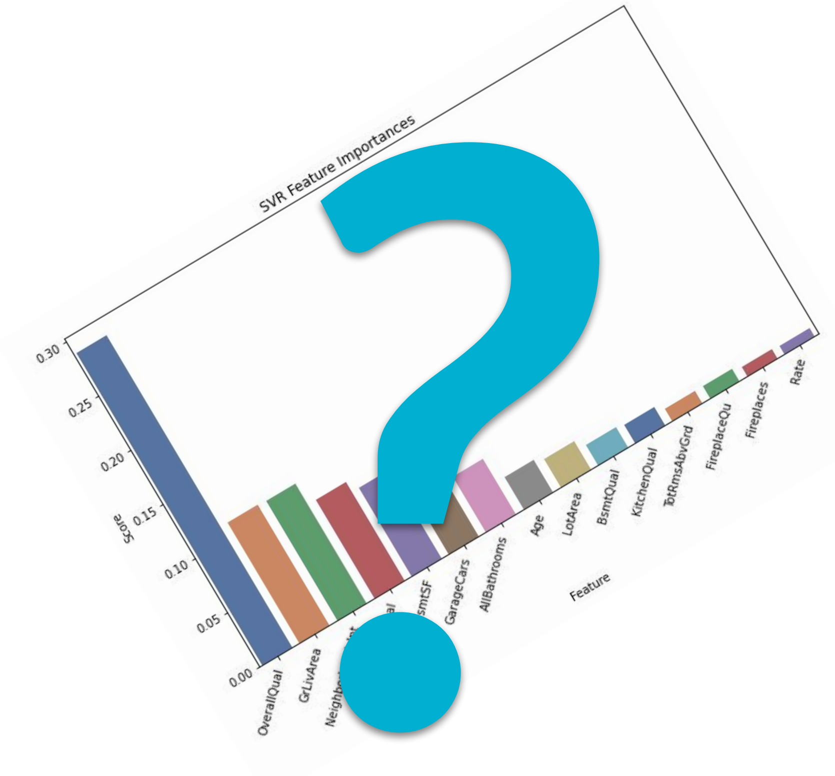
Linear regression can show the effect of a 1-unit change in a feature, *ceteris paribus*

- Increasing the living area of a home by 1 sq. ft. yields an additional $57.43 to the property's average sale price
- But the same for basement and outdoor areas are just $21.84 and $10.64 respectively
- Being near a high-traffic road or rail line decreases value by $7675!

What can an agent **recommend**?

- A fireplace can add $4380 to the sale price
  - However, might not be cost effective as cost of installation is between 2-5k. Agent to make determination based on fireplace type (gas, electric, etc)
- Adding a central air unit increases price by $23599
  - Given that cost of installation is between 3-15k, an agent can recommend installation

Support Vector Regression



Support Vector Regression

Model Scores

	Before Standardizing		After Standardizing	
	Default SalePrice	Default LogSalePrice	Default logSalePrice	Tuned logSalePrice
R^2 Train	-0.063	0.687	0.951	0.928
R^2 Test	-0.063	0.676	0.852	0.922
RMSE	x	x	0.385	0.279

Hyper-parameters ⁑

grid search performed to tune hyperparameters, best params:

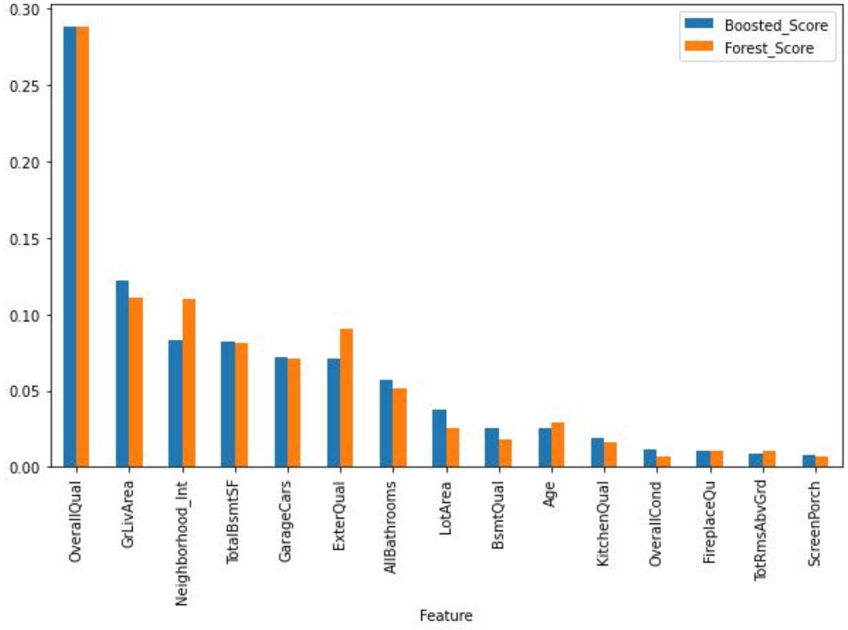
- C = 1e5
- gamma= 1e-6

* Reported models using logSalePrice for consistency. Models using SalePrice showed marginal improvement (~10-15% in R^2), but were more overfit

⁑ Attempted gridsearch to explore other parameters such as `kernel=linear/poly`but lacked time (>5 hours to run)

Tree Models

Random Forest
&
Gradient Boosting



Tree Models - Random Forest & Gradient Boosting

Model Scores

	Random Forest		Boosted	
	Default	Tuned	Default	Tuned
R^2 Train	0.987	0.986	0.985	0.994
R^2 Test	0.909	0.915	0.920	0.927
RMSE	0.050	0.047	0.045	0.043

Hyper-parameters

grid search performed to tune hyperparameters, best params:

- n_estimators = 300
- max_depth = 24
- min_samples_split = 2
- max_features = 21

grid search performed to tune hyperparameters, best params:

- n_estimators = 1000
- learning_rate = 0.0257
- max_depth = 5
- max_features = 25
- subsample = 0.5

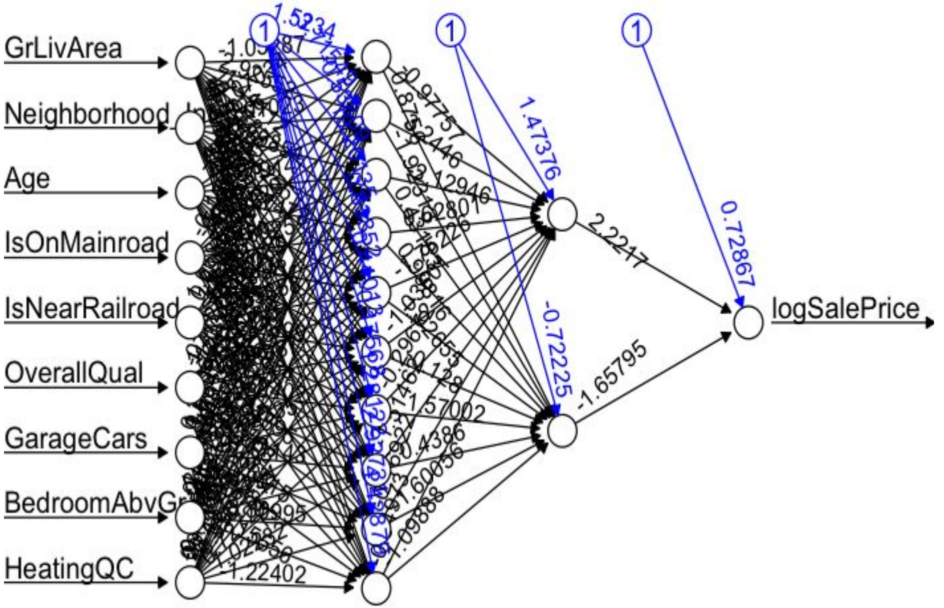
*reported models using logSalePrice for consistency, models using SalePrice were marginally better (~10-15% improvement in R^2)

*attempted gridsearch to explore other parameters such as loss=huber but lacked time (>2hrs to run)

Neural Network

Backpropagation

Sub-network of Complete Neural Net



Backpropagation Model

Perceptron - artificial 'neuron' that performs weighted sum on synaptic inputs. Output function dependant on weighted sum - could be 'analog' or 'digital' (fire to 1 if threshold exceeded)

$$h_j=\theta\left[\Sigma(h_iW_{ij}+b_j)\right]=\theta(s_j)$$

Backpropagation - a forward propagating neural network where backward propagating synapses have been added to the neurons to adjust the synaptic weights based on the final output error

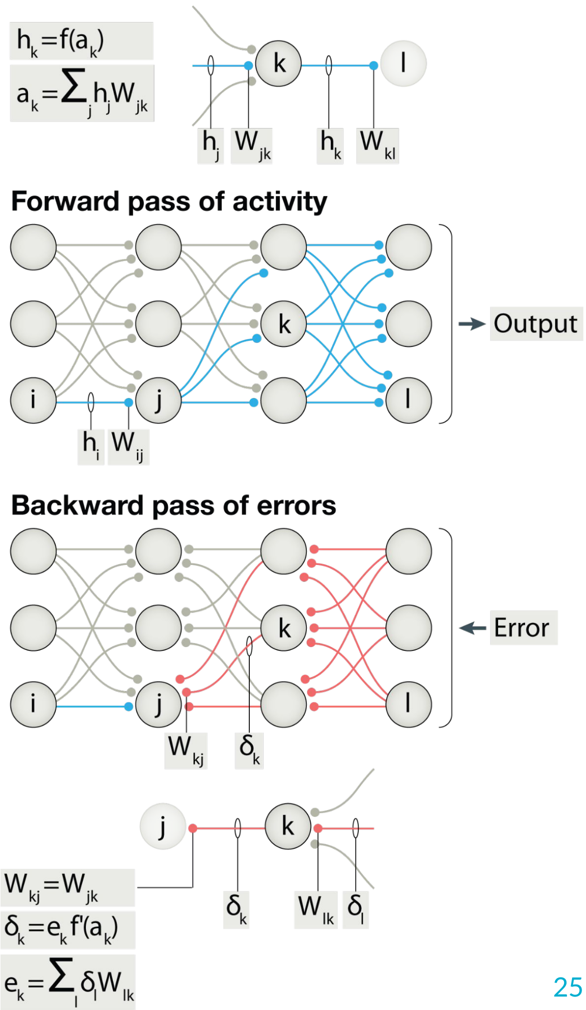
$$E=\frac{1}{2}\sum_l(h_l-t_l)^2$$

Backprop method - compute for every synaptic weight the gradient of the error ∂E/∂W with respect to the current weight. Use gradient to adjust weight proportionally. Key insight is error can be computed recursively via the chain rule

$$\Delta W_{ij}=-\eta\frac{\partial E}{\partial W_{ij}}=-\eta\frac{\partial E}{\partial s_j}\cdot\frac{\partial s_j}{\partial W_{ij}}=-\eta\cdot\frac{\partial E}{\partial s_j}\cdot h_i=-\eta\delta_j h_i$$

$$\delta_l=\frac{\partial E}{\partial s_l}=\frac{\partial E}{\partial h_l}\cdot\frac{\partial h_l}{\partial s_l}=(h_l-t_l)\cdot\theta'(s_l)=(h_l-t_l)\left[1-\theta^2(s_l)\right]$$

$$\delta_k=\frac{\partial E}{\partial s_k}=\sum_l\frac{\partial E}{\partial s_l}\cdot\frac{\partial s_l}{\partial h_k}\cdot\frac{\partial h_k}{\partial s_k}=\sum_l\delta_l\cdot W_{kl}\cdot\left[1-\theta^2(s_k)\right]$$



Backpropagation Results

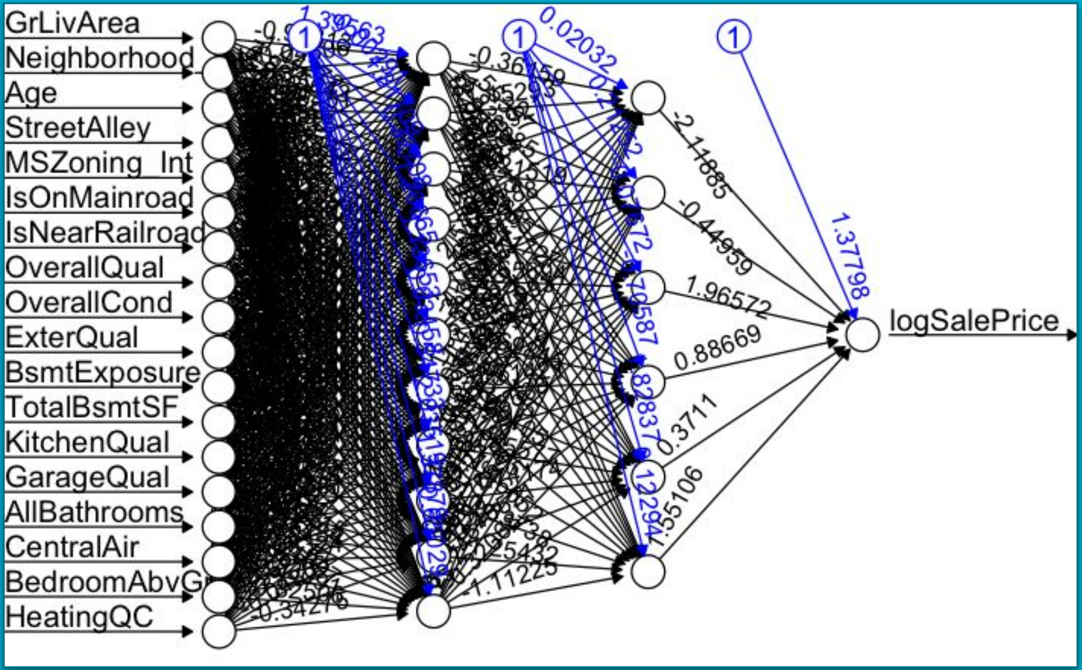
Backpropagation is accurate with far fewer features, making it ideal for our agents, who would need to enter less information for a good sale price estimate

18 inputs
(c.f. with 47 inputs using Elastic-Net)

Two hidden layers (11, 6)

Substantive improvement in model accuracy will become possible with further work.

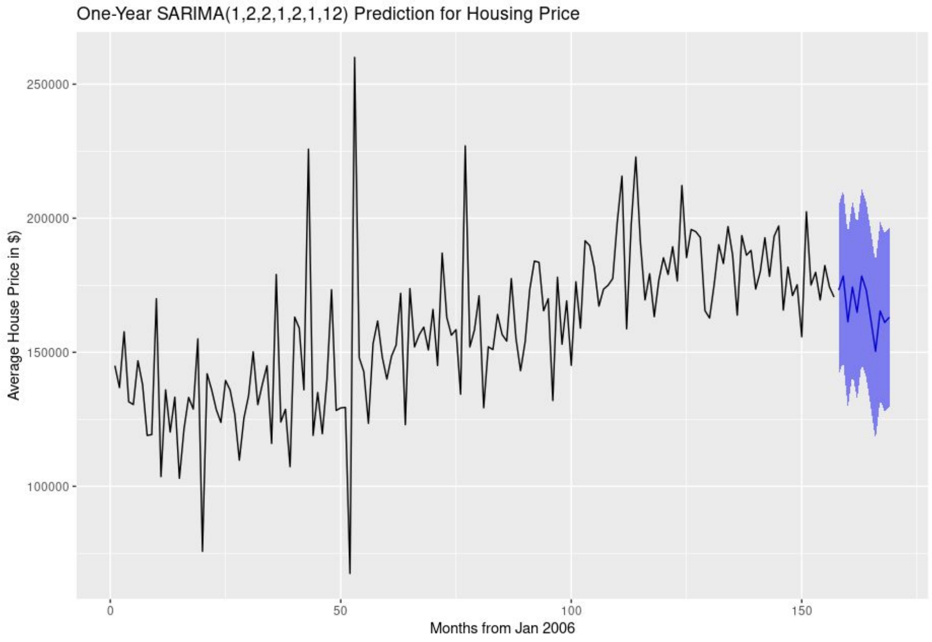
	Backprop
R^2 Train	0.9366
R^2 Test	0.8950
RMSE	0.0320
AIC	577.1
BIC	2160.3



Time Series

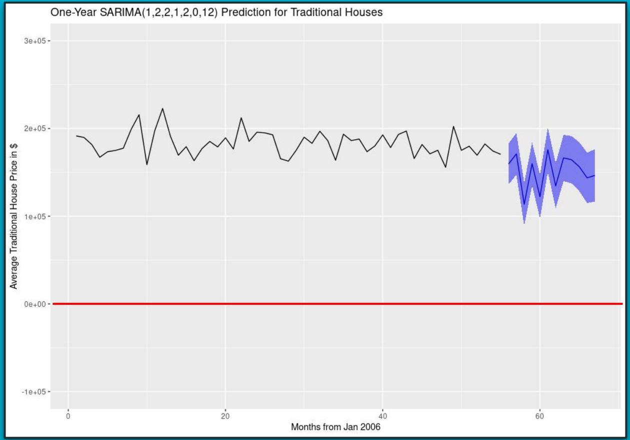
SARIMA

Prediction - All House Types

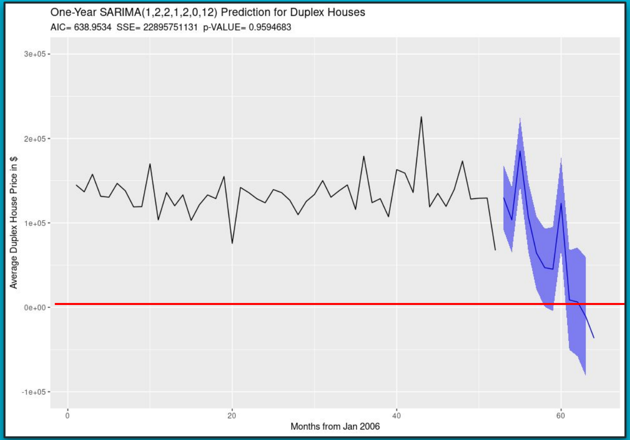


Price Time Series– House Type (Collapse_MSSubClass)

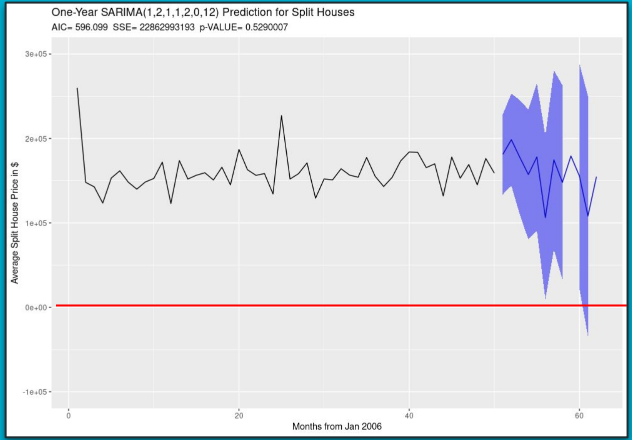
Traditional



Duplex

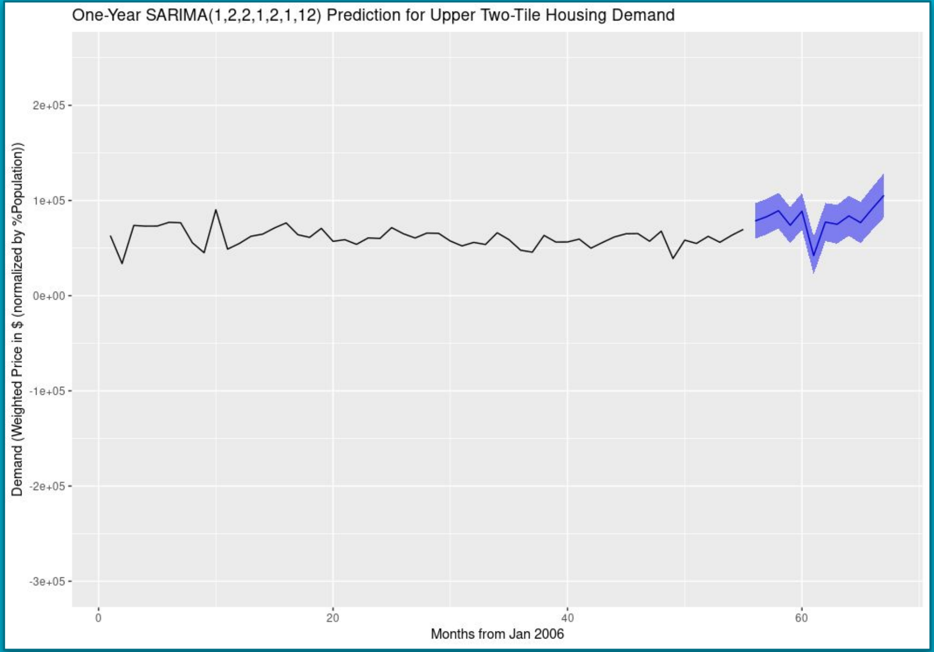


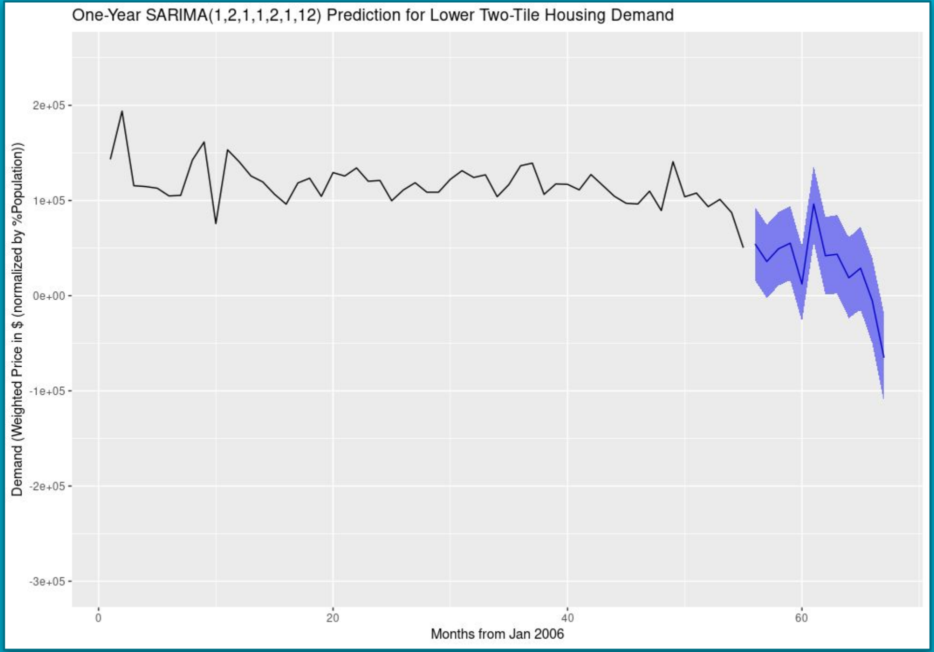
Split



“Demand” Time Series– Upper and Lower 2-Tile

*Reveals that “demand” of upper two-tile homes will recover in the next two years**





*Assuming our dataset contains all houses on Ames market at that month

Time Series - Average Price

Model Scores

- As SARIMA predictions are not based on linear Correlation, there will not be strong Coeff. Of Determination, nor will it have much significance.
- We checked improvement of our model, if the RMSE decreased between train and test.

	SARIMA	
	Test	Train
R^2	0.077	0.528
RMSE	\$18,655.99	\$9014.61

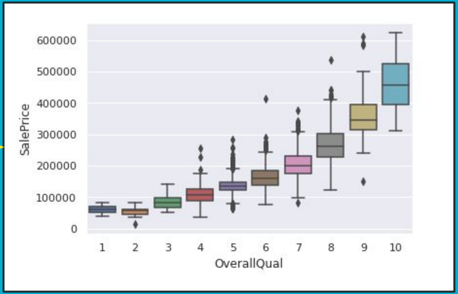
Model Parameters and Coefficients

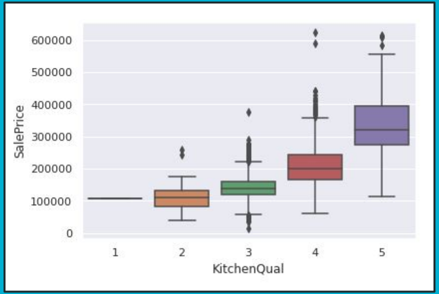
SARIMA(1,2,2,1,2,1,12)

- This is an **AR(1)+MA(2)**, **twice-differenced** for Weak Stationarity, with seasonal **sAR(1)+sMA(1)**, **twice-differenced** for Weak Stationarity, over a seasonal period of 12 months.
- This means that we have a strong *Moving Average* component to our model, and it is accounting for many *random shocks* present in the data. The small number of *Auto-regressive* components implies there is not as strong a dependence on previous value dictating future value
- The *Autoregressive* components show *historical values* are accounted for twice: once locally, and once in the seasonal autocorrelation. This explains why our sparse models were not resilient for long prediction timescales.

Future Development

- **Ensemble** multiple machine learning models to bring even more accurate predictions
- Use log-log for (sale price ~ area features) rather than log-linear
- Incorporate **MLS data** and **Household Income data** into our model
- Investigate impact of distance from Iowa State University
- Add **interaction** and **polynomial** features
  (E.g., some *Qual features appear to be quadratic)





Takeaway

Regression Realty Agents can now Accurately Predict Sale Price!

Just input the property's details into our app‡ and get estimated home value

Listing agents
Know the true value of a home instead of guessing. Ensure a home isn't mispriced, thus on the market for too long or short a time

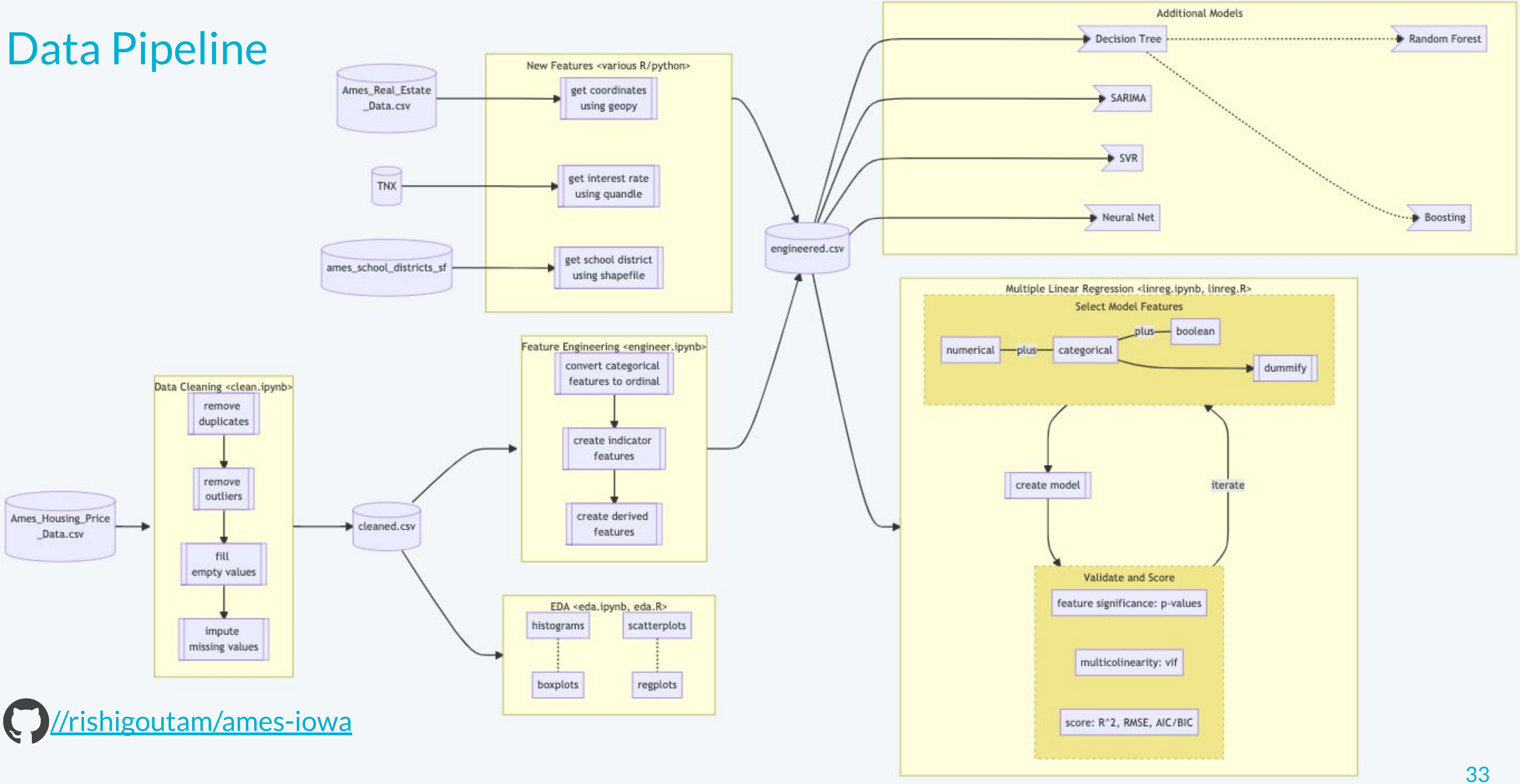
Selling agents
Know if a home is over- or under-priced and offer bidding advice to your client

‡ *Expected by Q3, 2011*

	R^2 Train	R^2 Test	RMSE
Elastic-Net	0.933	0.922	0.046
Random Forest	0.986	0.915	0.047
Gradient Boosting	0.994	0.927	0.043
SVR	0.926	0.922	0.279
Backprop	0.937	0.895	0.032
SARIMA†	0.528	0.077	$14,167.58

† SARIMA used SalePrice (not logSalePrice)
Time series R^2 cannot be compared to other models as the model is not linear

Data Pipeline



//rishigoutam/ames-iowa

References

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project
http://jse.amstat.org/v19n3/decock.pdf

Iowa House Prices over Time
https://fred.stlouisfed.org/series/IASTHPI

Treasury Yields (TNX)
https://finance.yahoo.com/quote/%5ETNX/

Ames Neighborhoods
https://www.cityofames.org/home/showpublisheddocument/1024/637356764775500000

Ames School Districts
https://github.com/topepo/AmesHousing/blob/master/data/ames_school_districts_sf.rda
https://www.ames.k12.ia.us/boundaries/

Kaggle Ames Housing Dataset
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Fireplace Installation Cost
https://homeguide.com/costs/fireplace-installation-cost

Central Air Installation Cost
https://www.remodelingcalculator.org/cost-install-central-air/

Backpropagation
Lillicrap, T. P. *et. al.*, "Backpropagation and the brain". *Nat. Rev. Neurosci.* **21**, 335-346 (2020)

Q&A

Data Science @ Regression Realty, Inc

Akram Sadek, James Goudreault, Rishi Goutam, Srikar Pamidi