"Homebuying Powered by Data Science"

Alice Zhang

in/xuyuan-zhang

github.com/AliceXuyuan

Joe Lu

in/joezlu

github.com/jzl4

Jordan Runge

in/jordanrunge

github.com/jordanrunge

Yunmei Zhang

in/yunmeizhang

github.com/zhangym1256


AUTOMATE
THE BORING STUFF
WITH PYTHON
PRACTICAL PROGRAMMING
FOR TOTAL BEGINNERS
AL SWEIGART
no starch press


THE ANNOTATED TEXT
THE POEMS
OF
T.S.
ELIOT
VOLUME I
COLLECTED & UNCOLLECTED POEMS

Data Science

Art of Realty

TSS Eliot

Data Abundance

Computing Power and Tools

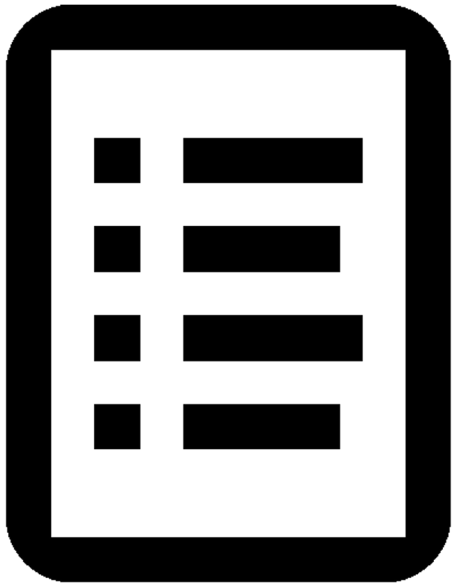
$35 Trillion Market


AMES
many windows
one community

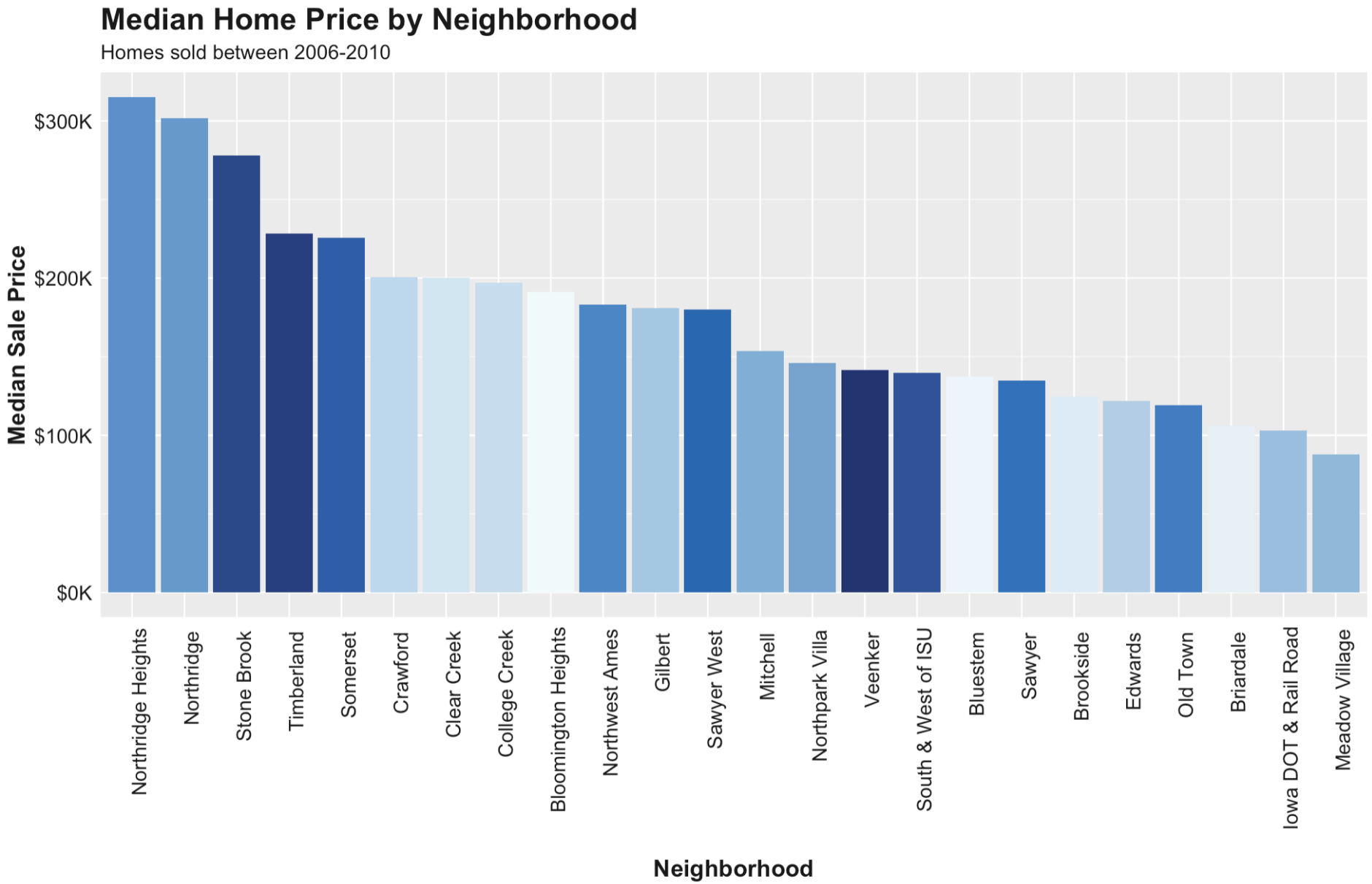
TSS Eliot

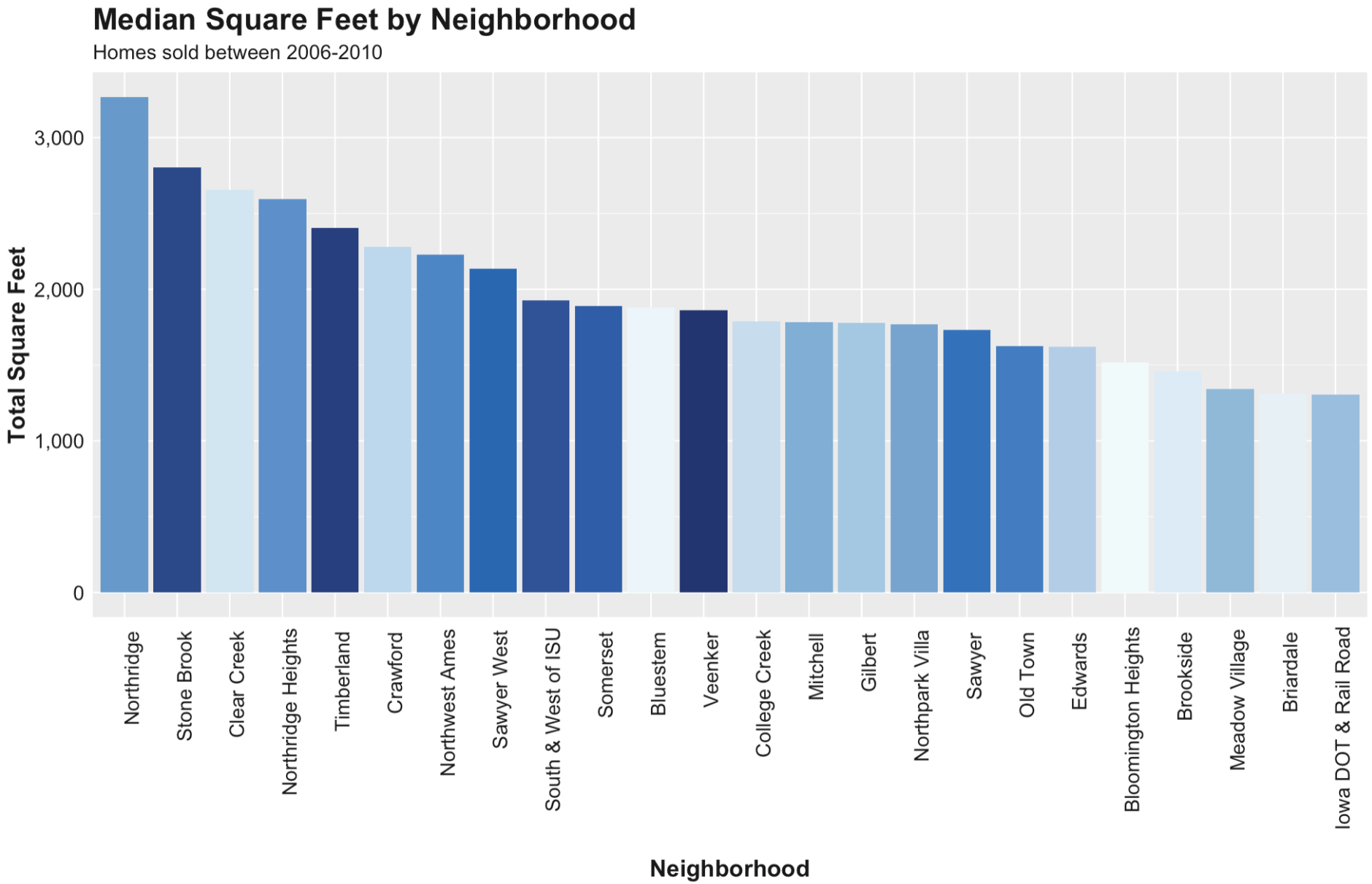
1460 Homes

80 Features



Cool. But how can we make sense of all this data?


Median Home Price by Neighborhood
Homes sold between 2006-2010
Median Sale Price
$300K
$200K
$100K
$0K
Northridge Heights
Northridge
Stone Brook
Timberland
Somerset
Crawford
Clear Creek
College Creek
Bloomington Heights
Northwest Ames
Gilbert
Sawyer West
Mitchell
Northpark Villa
Veenker
South & West of ISU
Bluestem
Sawyer
Brookside
Edwards
Old Town
Briardale
Iowa DOT & Rail Road
Meadow Village
Neighborhood


Median Square Feet by Neighborhood
Homes sold between 2006-2010
Total Square Feet
3,000
2,000
1,000
0
Northridge
Stone Brook
Clear Creek
Northridge Heights
Timberland
Crawford
Northwest Ames
Sawyer West
South & West of ISU
Somerset
Bluestem
Veenker
College Creek
Mitchell
Gilbert
Northpark Villa
Sawyer
Old Town
Edwards
Bloomington Heights
Brookside
Meadow Village
Briardale
Iowa DOT & Rail Road
Neighborhood

TSS Eliot

Interior

Features	*How 'big'?*	*How 'good'?*
Kitchen		1
Bathrooms	1	
Bedrooms		
Basement	1	2
Living Areas	1	
Utilities		1
Misc Inside	1	1
	4	*5*

Exterior

Features	*How 'big'?*	*How 'good'?*
Exterior		2
Garage	1	1
Lot	2	
Roof		
Outdoor Areas	1	
Misc Outside		
	4	*3*

Other

Features	*How 'good'?*
Location	2
House 'Meta'	2
Sale Context	1
	5

21 Most Important Features


Error of
0.124

How do we know we’re right?

Let’s get (a bit) technical.

TSS Eliot

1 Clean Data

2 Select Data

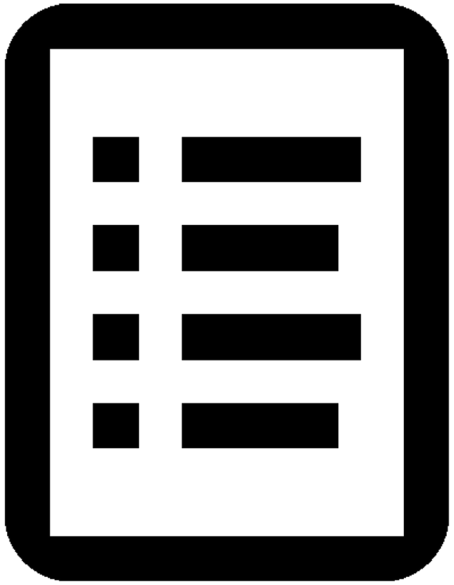
3 Create Models

- Linear
- Tree-based

4 Assist Homebuyers

TSS Eliot

80 Features

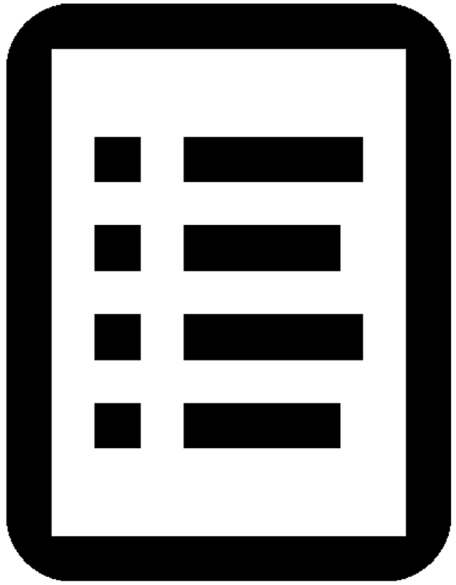


6% Missing

Variable	Method
Quality of Basement	Missing values are meaningful because there is no basement
Zoning	Look relevant feature (neighborhood), impute **most frequent value**
Area of Mason Veneer	Look at feature relevant feature (exterior type), **impute median value**

TSS Eliot

80 Features



- 33 Continuous
- 14 Ordinal
- 33 Categorical

TSS Eliot

Adjusted Basement Ft^2

- Finished 1
- Finished 2
- Unfinished

House Age

- Year Built
- Year Sold
- Year Remodeled

Adjusted Outdoor Ft^2

- 3 Season Porch
- Screen Porch
- Open Porch
- Enclosed Porch
- Wooden Deck

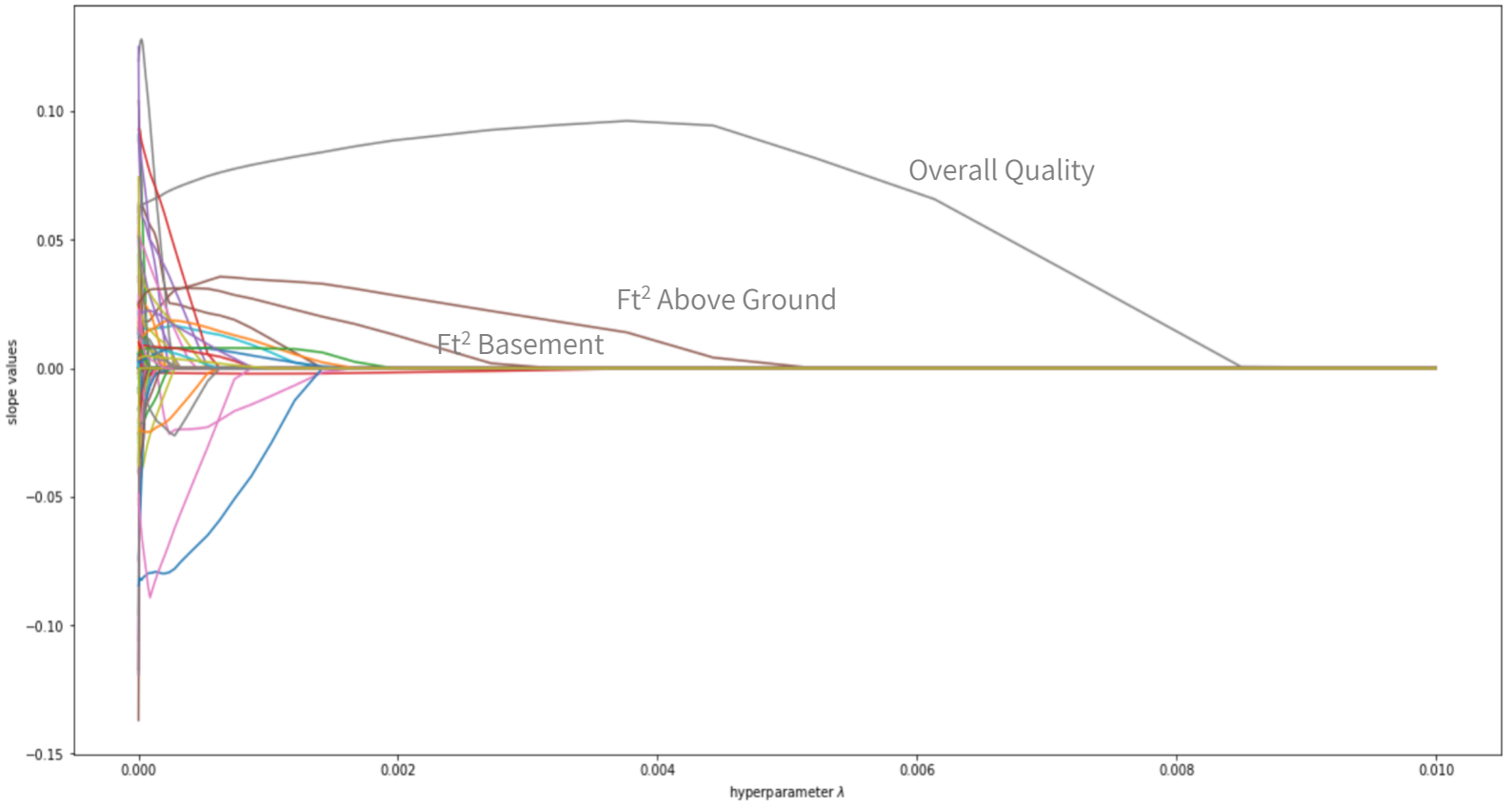
Total Baths

- Basement Full
- Basement Half
- Upstairs Full
- Upstairs Half


1
Data
VIF
17 Continuous Features
Eliminate 2 Features

TSS Eliot

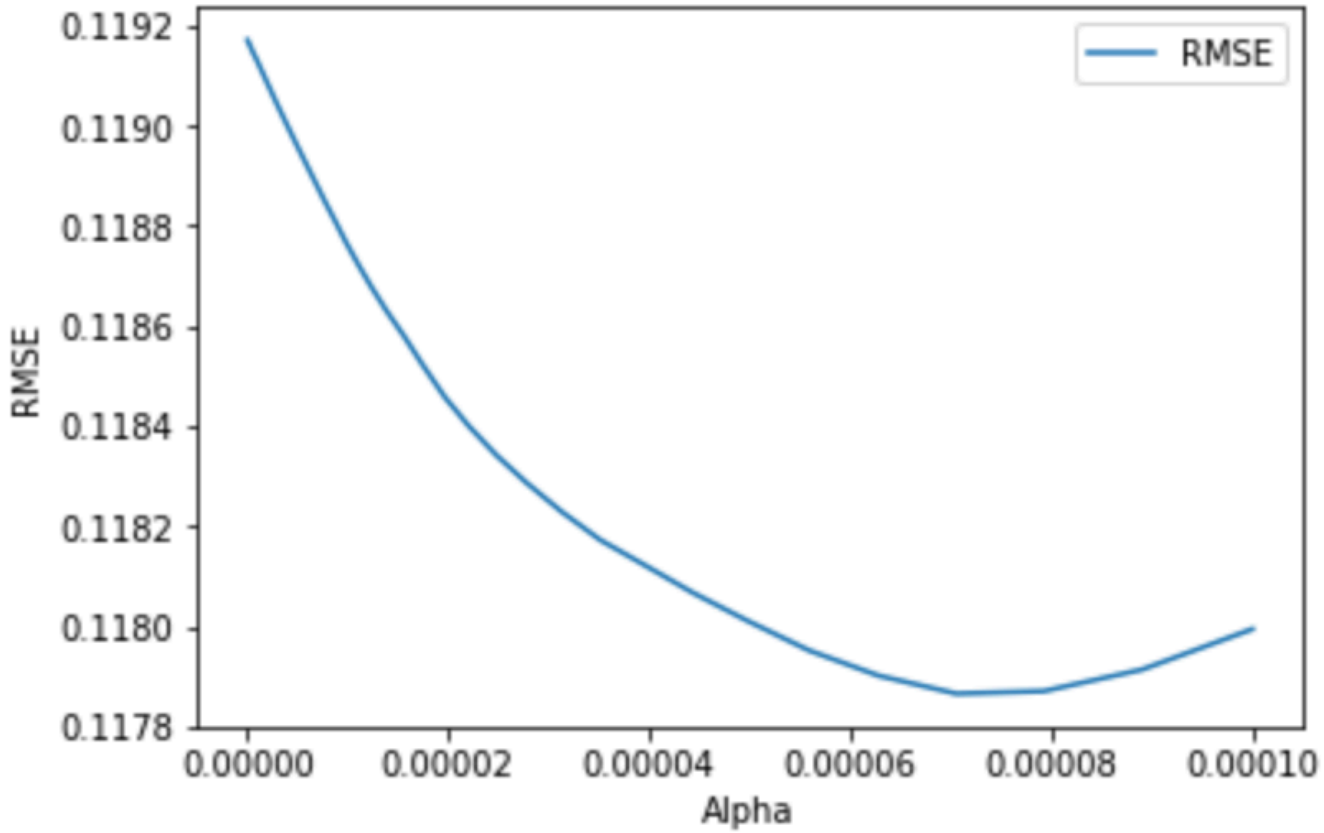
TSS Eliot



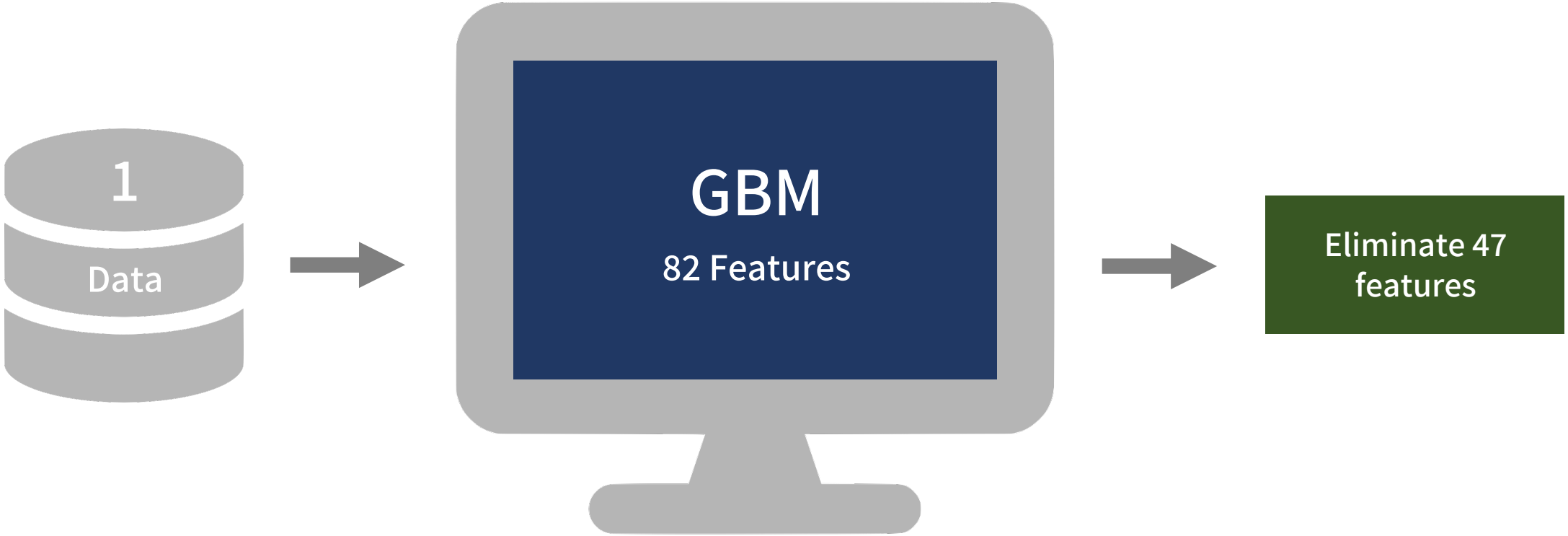
1. Overall Quality
2. Ft^2 Above Ground
3. Ft^2 Basement
4. Total Bathrooms
5. Ft^2 Garage
6. Age of House
7. Kitchen Quality
8. Fireplace Quality
9. Ft^2 Lot
10. Height of Basement
11. Exterior Quality
12. Number of Fireplaces

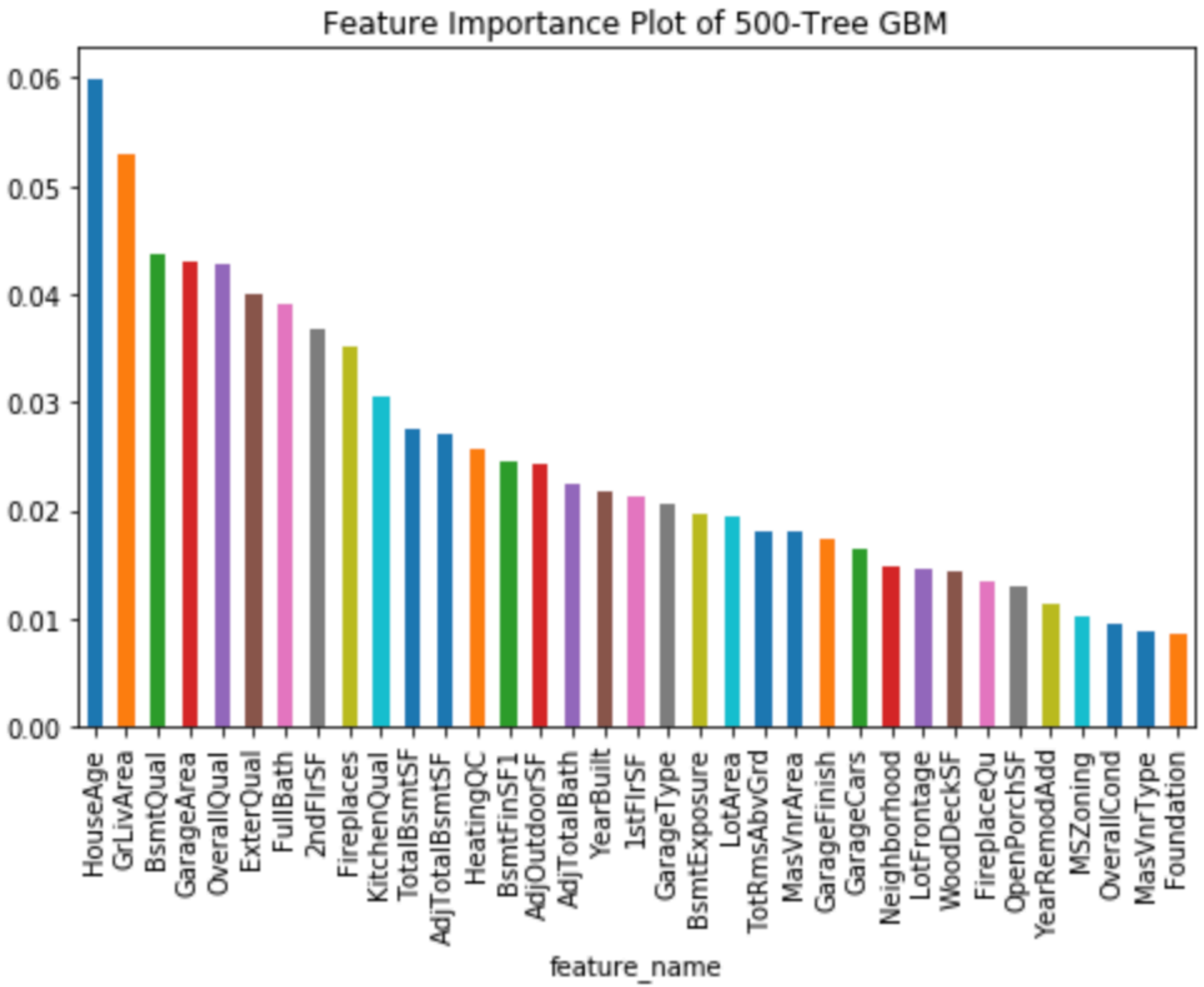
TSS Eliot


3
Data
Lasso
60 Features
Hyperparameter Tuning
Determine alpha
Drop outliers
Calculate RMSE


RMSE
0.1192
0.1190
0.1188
0.1186
0.1184
0.1182
0.1180
0.1178
RMSE
0.00000
0.00002
0.00004
0.00006
0.00008
0.00010
Alpha

Model	Error (RMSE)
Lasso	0.13306


1
Data
GBM
82 Features
Eliminate 47 features


Feature Importance Plot of 500-Tree GBM
0.06
0.05
0.04
0.03
0.02
0.01
0.00
HouseAge
GrLivArea
BsmtQual
GarageArea
OverallQual
ExterQual
FullBath
2ndFlrSF
Fireplaces
KitchenQual
TotalBsmtSF
AdjTotalBsmtSF
HeatingQC
BsmtFinSF1
AdjOutdoorSF
AdjTotalBath
YearBuilt
1stFlrSF
GarageType
BsmtExposure
LotArea
TotRmsAbvGrd
MasVnrArea
GarageFinish
GarageCars
Neighborhood
LotFrontage
WoodDeckSF
FireplaceQu
OpenPorchSF
YearRemodAdd
MSZoning
OverallCond
MasVnrType
Foundation
feature_name

TSS Eliot


2
Data
GBM
35 Features
Hyperparameter Tuning
Determine 3 hyperparameters
Calculate RMSE

Model	Error (RMSE)
Lasso	0.13306
GBM	0.13501

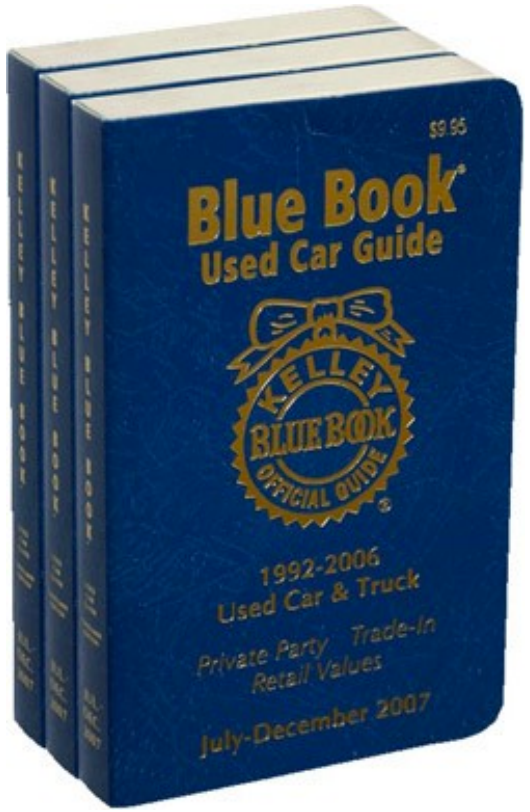

1
Data
Bayesian Ridge
80 Features
Hyperparameter Tuning
Determine 4 hyperparameters
Calculate RMSE

Model	Error (RMSE)
Lasso	0.13306
GBM	0.13501
Bayesian Ridge	0.13101

And then we combined models.

Model	Error (RMSE)
Lasso	0.13306
GBM	0.13501
Bayesian Ridge	0.13101
Stacked 25% Lasso 40% GBM 35% Bayesian Ridge	0.12412

Amazing. How will this help homebuyers?


$9.95
Blue Book
Used Car Guide
KELLEY BLUE BOOK OFFICIAL GUIDE
1992-2006
Used Car & Truck
Private Party Trade-In
Retail Values
July-December 2007

TSS Eliot

Recommendation Engines

Market Forecasting Tools

"Homebuying Powered by Data Science"

Alice Zhang

in/xuyuan-zhang

github.com/AliceXuyuan

Joe Lu

in/joezlu

github.com/jzl4

Jordan Runge

in/jordanrunge

github.com/jordanrunge

Yunmei Zhang

in/yunmeizhang

github.com/zhangym1256