MIDS W207

Final Project

Kaggle - House Prices Advanced Regression

Fall 2021

Parham Motameni (pmotameni@berkeley.edu)
Radia Wahab (radiawahab@berkeley.edu)
Jun Qian (junqian@berkeley.edu)

Instructor: Uri Schonfeld

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction

Description of Data and Data Source

The Ames Housing dataset was compiled by **Dean De Cock**, with 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

This data set (https://www.kaggle.com/c/house-prices-advanced-regression-techniques) is being used for this project to demonstrate the use of various Machine Learning techniques, to be able to have the algorithm perform the difficult task of deciding what the house price should be.

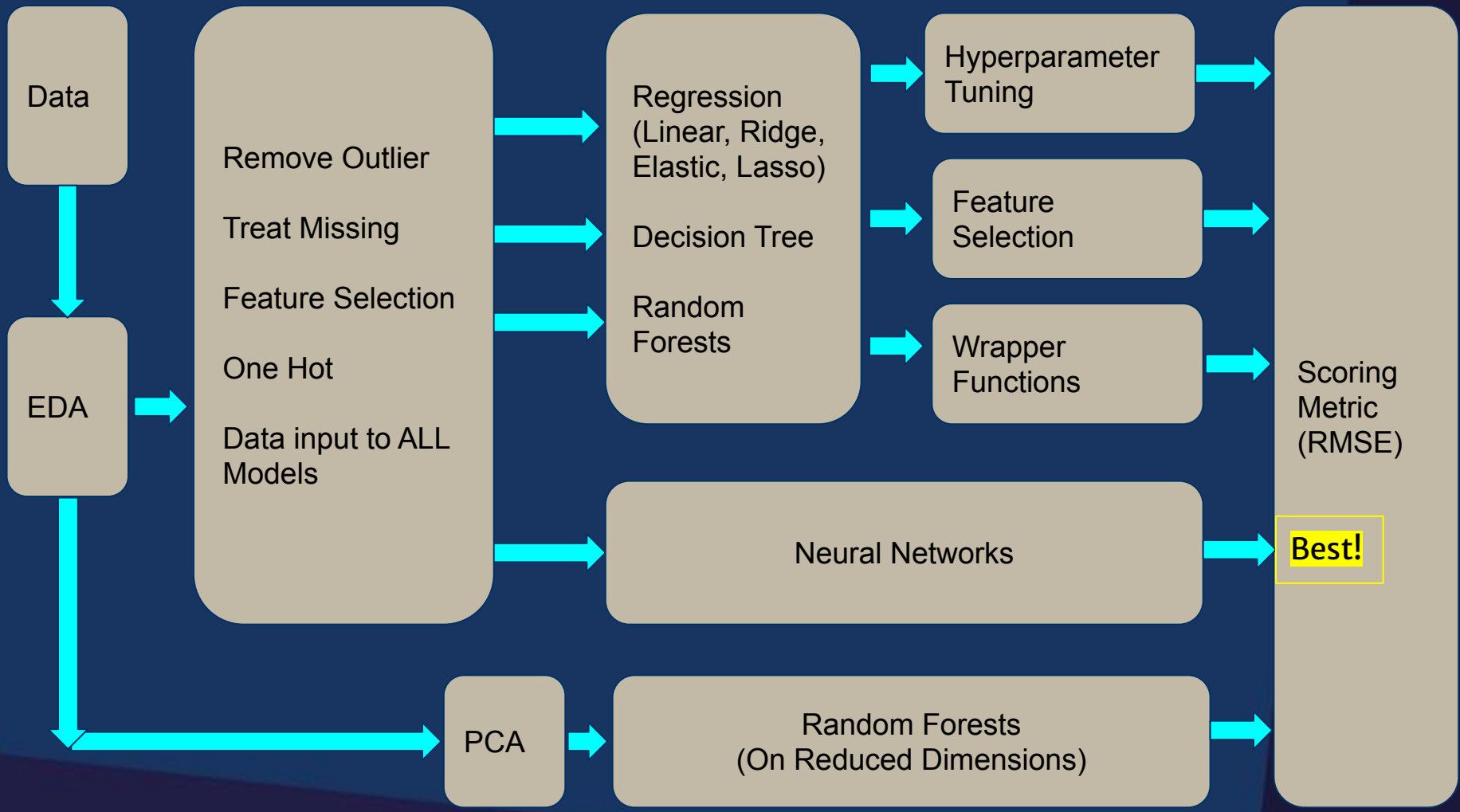
Berkeley
UNIVERSITY OF CALIFORNIA

Model Exploration

Full Workflow

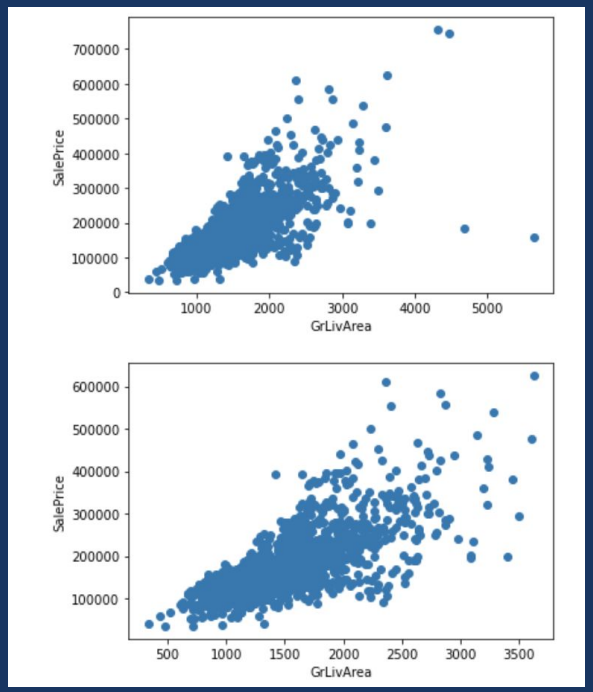
Berkeley
UNIVERSITY OF CALIFORNIA

Model Exploration



Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis



1. Data_Description.txt
2. The paper where the original data was published
3. Reviewing House Price Websites for Subject Matter

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock
Truman State University

Journal of Statistics Education Volume 19, Number 3(2011),
www.amstat.org/publications/jse/v19n3/decock.pdf

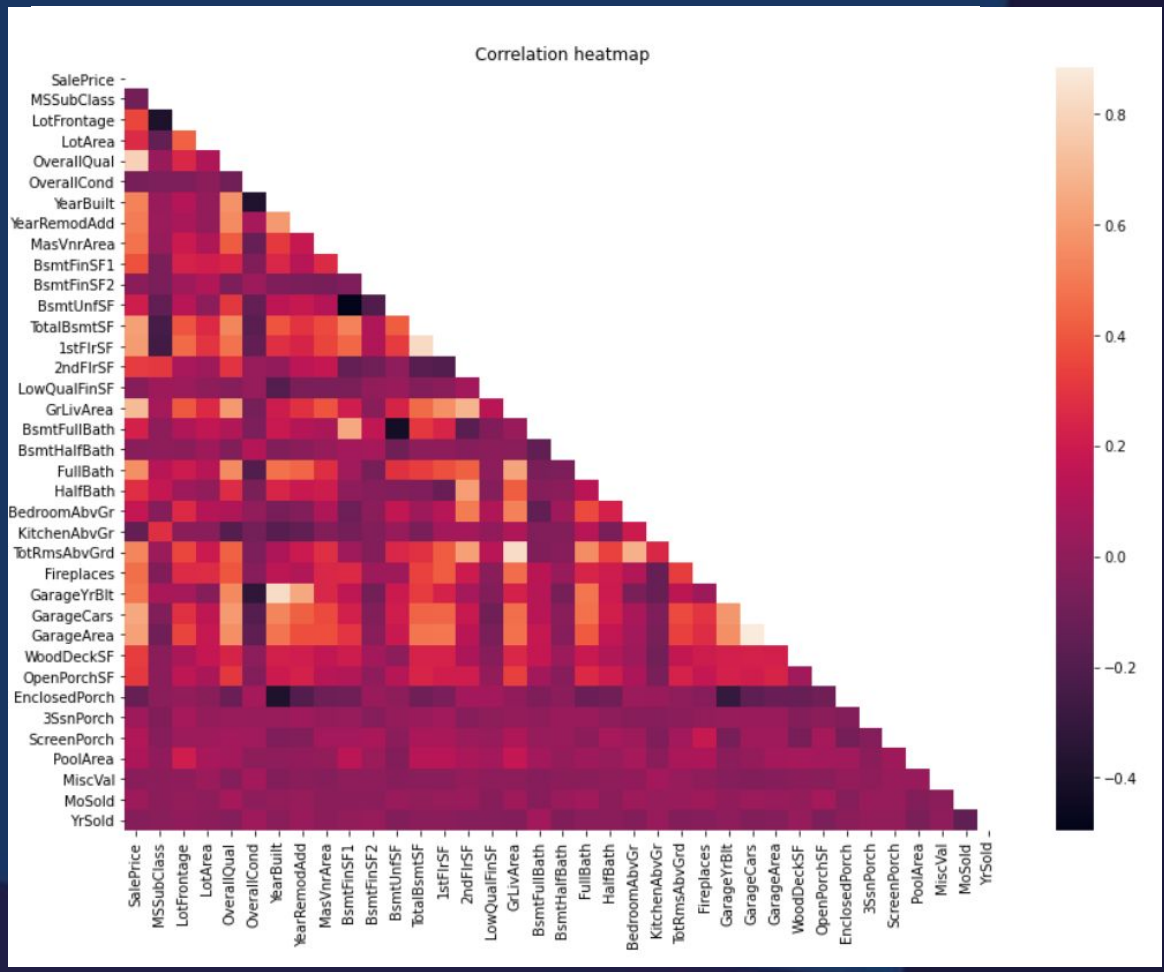
Potential Pitfalls (Outliers): Although all known errors were corrected in the data, no observations have been removed due to unusual values and all final residential sales from the initial data set are included in the data presented with this article. There are five observations that an instructor may wish to remove from the data set before giving it to students (a plot of SALE PRICE versus GR LIV AREA will quickly indicate these points). Three of them are true outliers (Partial Sales that likely don't represent actual market values) and two of them are simply unusual sales (very large houses priced relatively appropriately). I would recommend removing any houses with more than 4000 square feet from the data set (which eliminates these five unusual observations) before assigning it to students.

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis

Numerical Variables

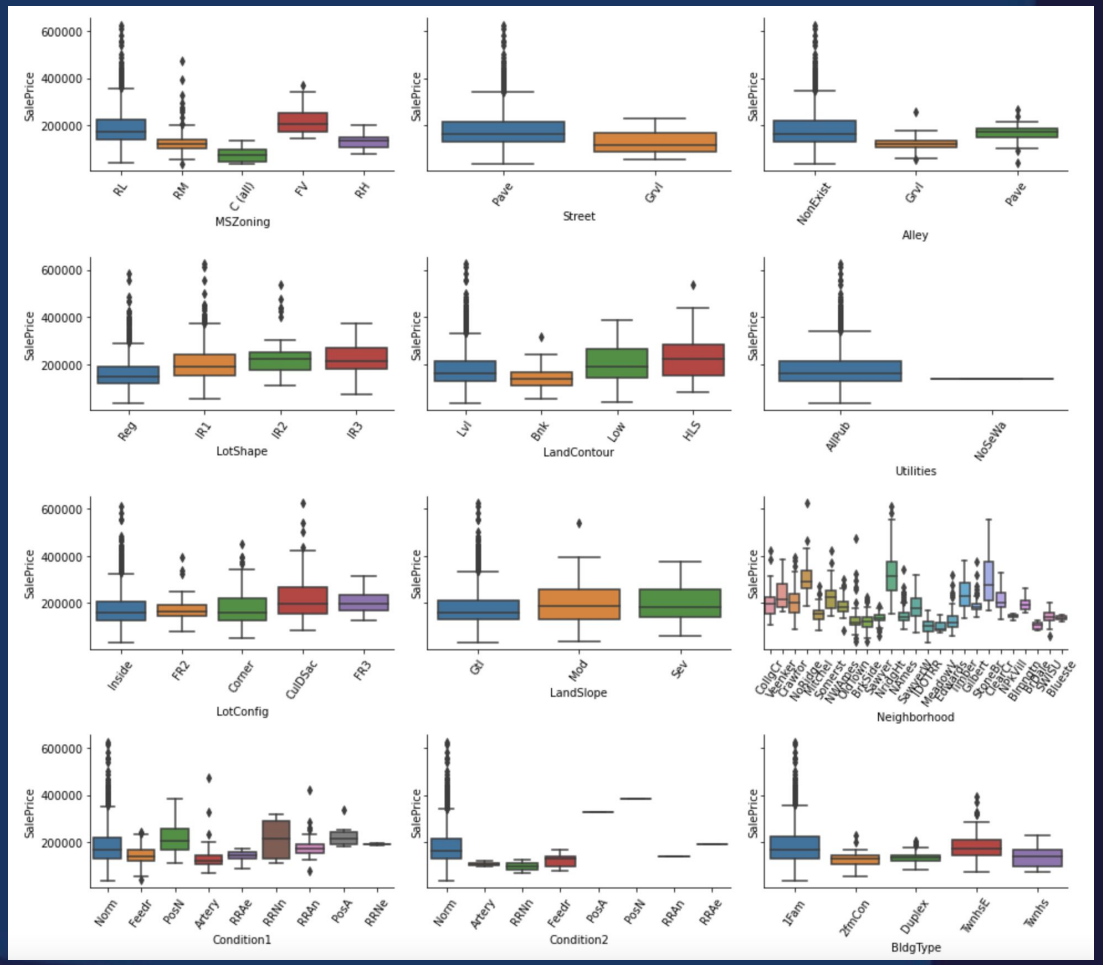
- Distribution Analysis
- Correlation Analysis



Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis

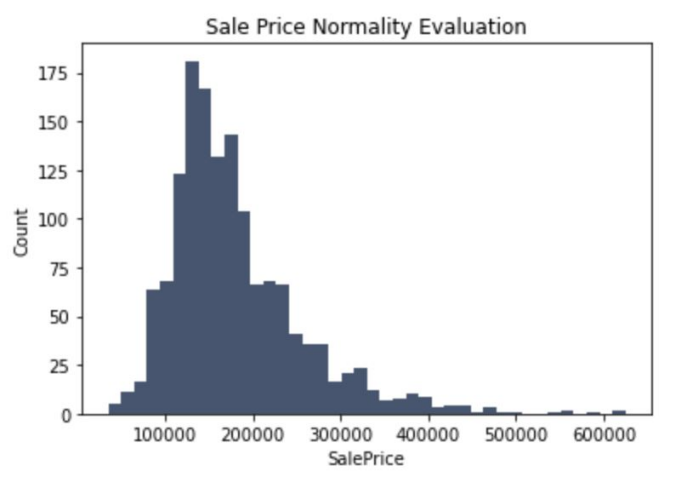
Categorical Variables



Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis

1. Normality of Target Variable
2. Check for Missing Values
3. All our EDA fed into the next steps of Data preparation



	Missing Values	Percentage Missing
PoolQC	1451	99.656593
MiscFeature	1402	96.291209
Alley	1365	93.750000
Fence	1176	80.769231
FireplaceQu	690	47.390110
LotFrontage	259	17.788462
GarageYrBlt	81	5.563187
GarageQual	81	5.563187
GarageFinish	81	5.563187
GarageType	81	5.563187
GarageCond	81	5.563187
BsmtExposure	38	2.609890
BsmtFinType2	38	2.609890
BsmtFinType1	37	2.541209
BsmtCond	37	2.541209
BsmtQual	37	2.541209
MasVnrArea	8	0.549451
MasVnrType	8	0.549451
Electrical	1	0.068681

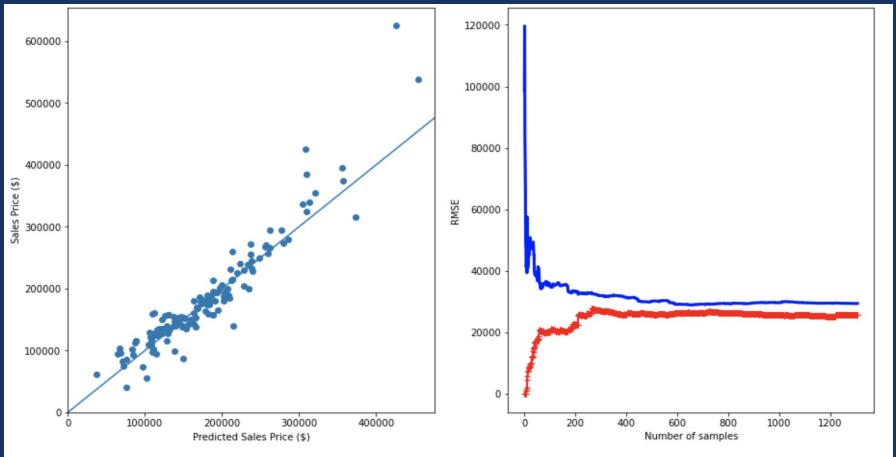
Berkeley
UNIVERSITY OF CALIFORNIA

Regression and Hyperparameter Tuning

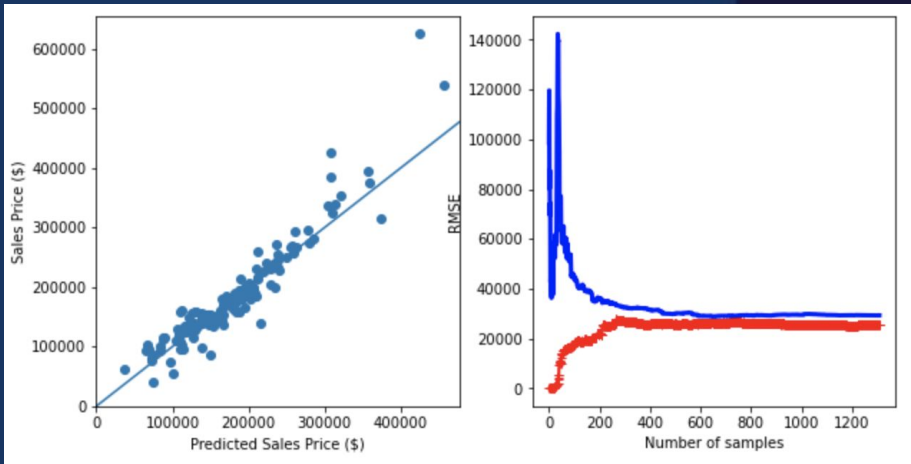
Berkeley
UNIVERSITY OF CALIFORNIA

Hyperparameter Tuning

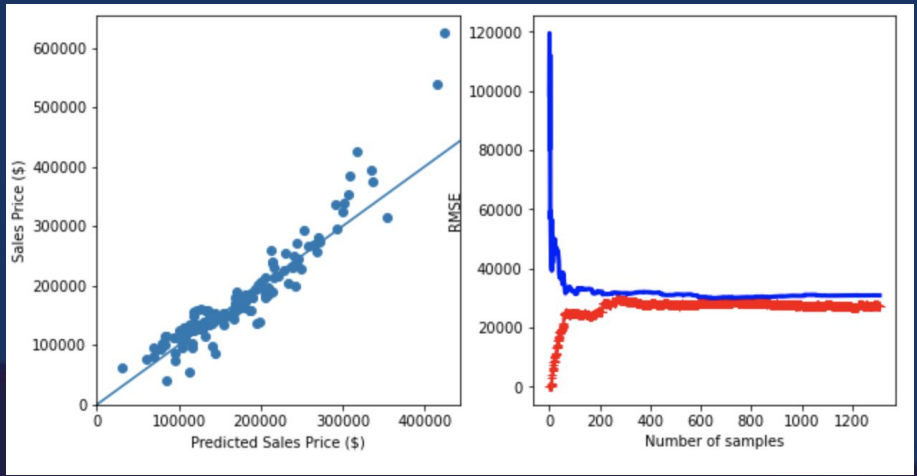
Ridge Regression



Lasso Regression



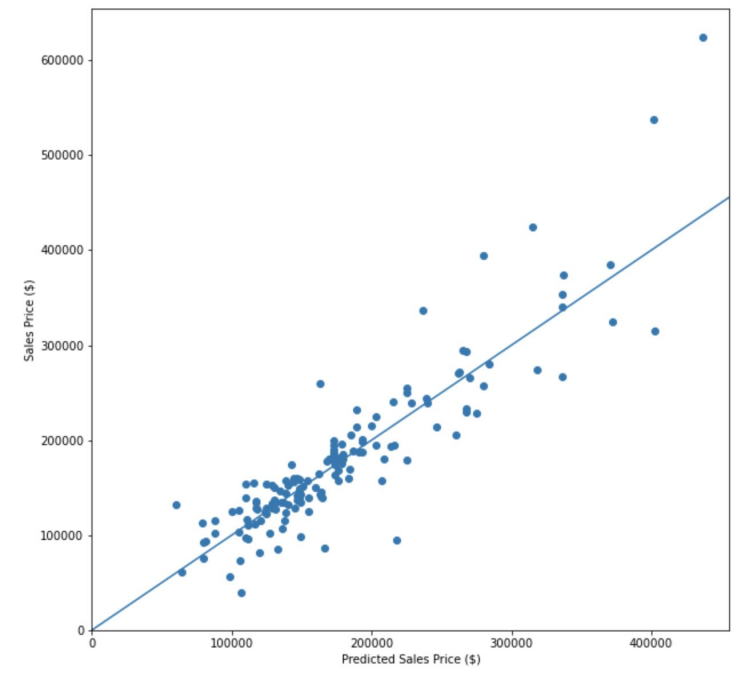
Elastic Net Regression



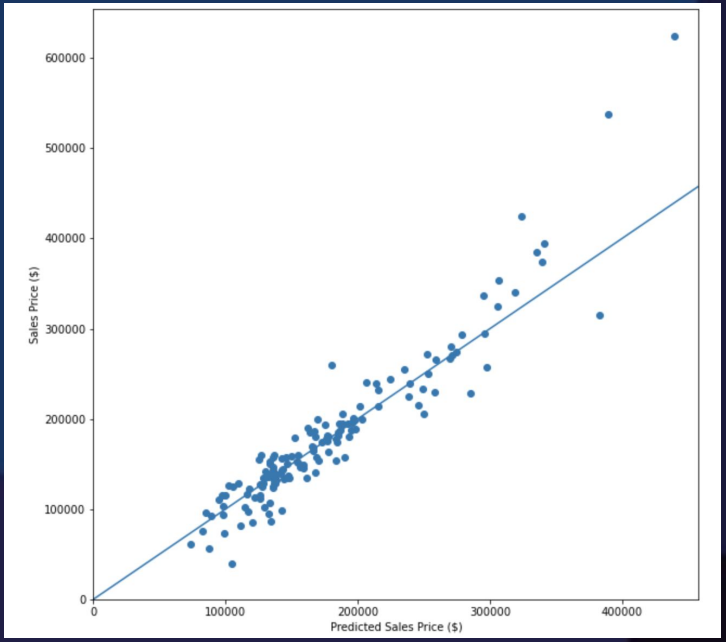
Berkeley
UNIVERSITY OF CALIFORNIA

Hyperparameter Tuning

Decision Tree

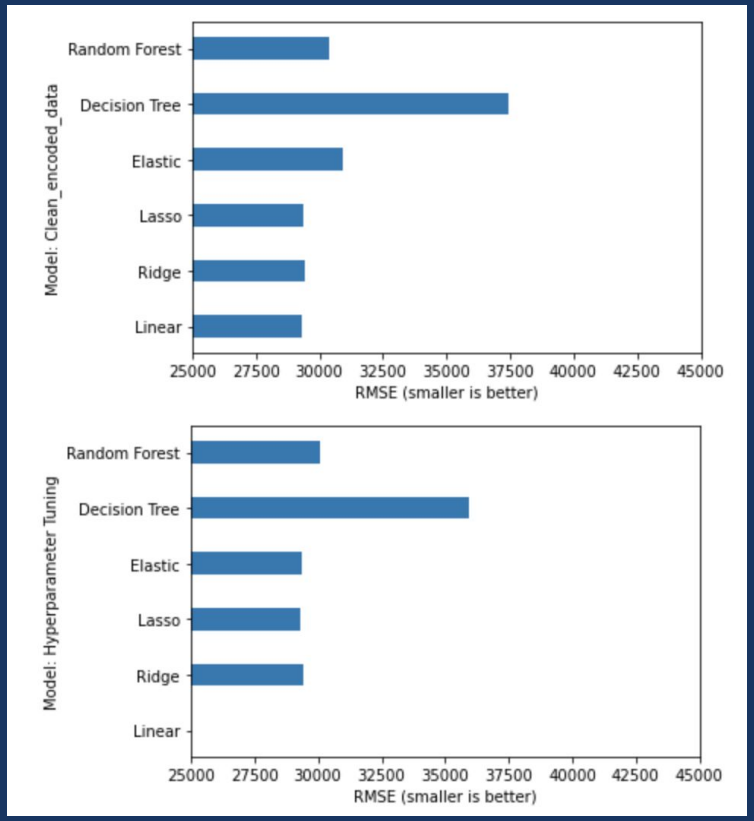


Random Forest Regression



Berkeley
UNIVERSITY OF CALIFORNIA

Overall Score



	Regressor	Baseline	Hyperparameter_tuning
0	Linear	29325.596202	0.000000
1	Ridge	29406.413053	29409.794982
2	Lasso	29340.796981	29325.822247
3	Elastic	30925.944372	29340.796981
4	Decision Tree	37407.315384	35942.875202
5	Random Forest	30394.642472	30210.036331

Berkeley
UNIVERSITY OF CALIFORNIA

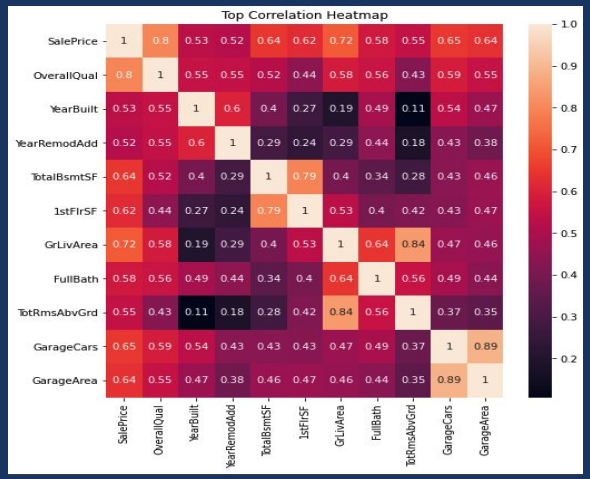
Feature Selection

Berkeley
UNIVERSITY OF CALIFORNIA

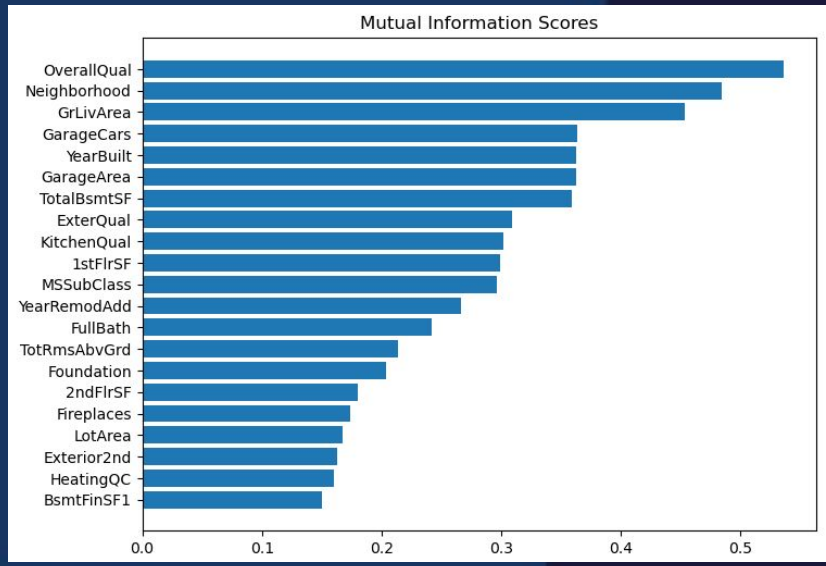
Feature Selection|

1. Filtered Method

- Correlation (corr>0.5)
- Mutual Information (mi score > 0.15)




```
21 features are selected

'GarageCars', 'FullBath', 'OverallQual',
'Exterior2nd', 'MSSubClass', 'LotArea',
'Neighborhood', 'ExterQual', 'HeatingQC',
'TotRmsAbvGrd', 'KitchenQual', 'Fireplaces',
'YearRemodAdd', '1stFlrSF', 'TotalBsmtSF',
'YearBuilt', 'GrLivArea', 'Foundation',
'2ndFlrSF', 'GarageArea', 'BsmtFinSF1'
```

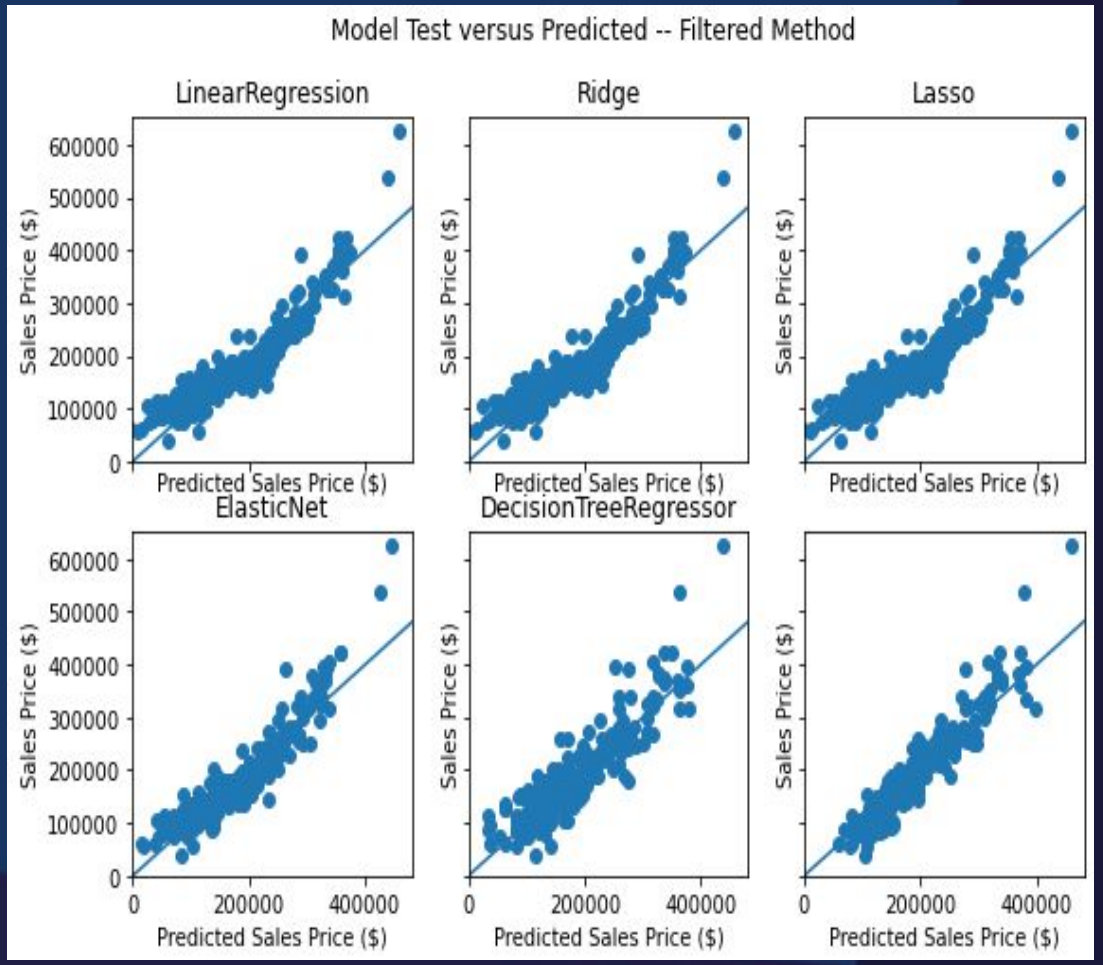


Berkeley
UNIVERSITY OF CALIFORNIA

Feature Selection|

Filtered Method - Results

	RMSE
LinearRegression	31542.3456
Ridge	31516.065
Lasso	31538.9221
ElasticNet	29546.5483
DecisionTreeRegressor	36009.2444
RandomForestRegressor	27720.2856



Berkeley
UNIVERSITY OF CALIFORNIA

Feature Selection|

Wrapper Method

- A recursive feature elimination process using sklearn RFE function, threshold=35

- Some of the top features selected.
- Liner, Ridge, Lasso share relatively similar features
- Decision Tree based have difference selections

OriginalFeatures	Linear	Ridge	Lasso	ElasticNet
OverallCon	1	1	1	1
ExterQual	1	1	1	1
LandSlope	1	1	1	5
SaleType	1	1	1	6
SaleCondit	1	1	1	8
Condition1	1	1	1	12
BedroomAt	1	1	1	14
Street	1	1	1	
Utilities	1	1	1	
CentralAir	1	1	1	
PavedDrive	1	1	1	
BsmtHalfBa	1	1	2	
HeatingQC	1	1	4	1
RoofMatl	1	4	1	1
KitchenAbv	1	7	7	1
HalfBath	1	12	16	1
MasVnrTyp	1	16	13	1
GarageCar	1	17	27	1
BldgType	1		1	1
HouseStyle	1			1
ExterCond	1			
BsmtFullBa	3	21	22	1
OverallQua	4	1	6	1
MSZoning	5	3	10	1
Heating	6	5	5	1
RoofStyle	7	1	3	1
Foundation	8	1	1	1
LotConfig	10	8	1	24
FullBath	14	1	1	13
Fireplaces	18	29	20	16
KitchenQua	20	1	1	21
Electrical	21	1	1	

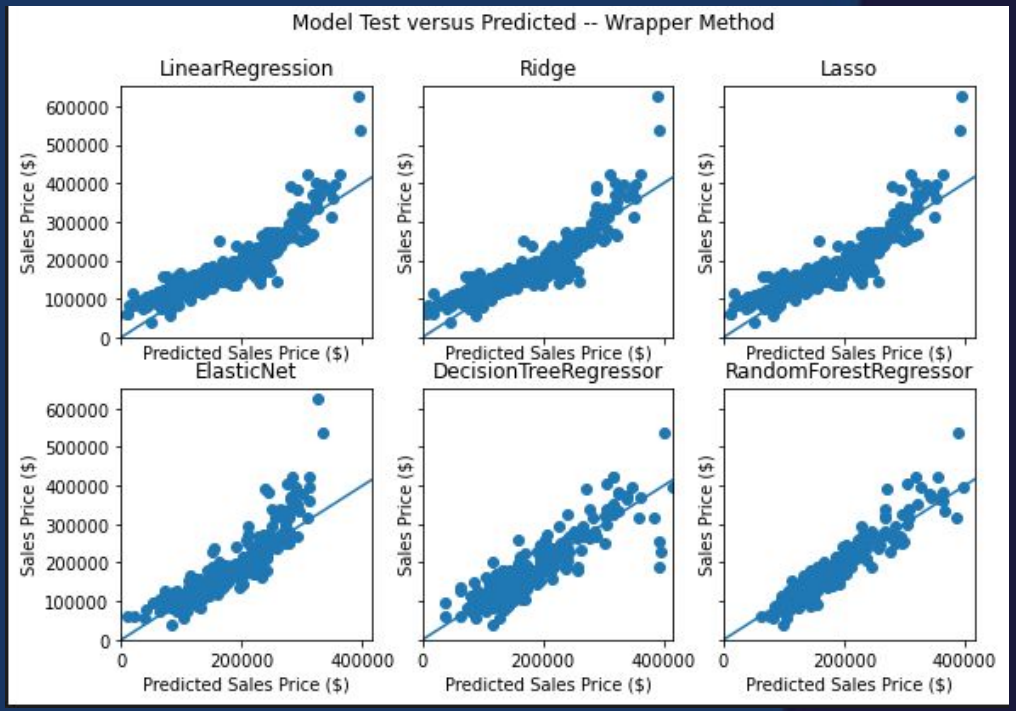
Original Features	Decision Tree	Random Forest
SaleType	1	1
BedroomAt	1	1
CentralAir	1	1
HeatingQC	1	
MasVnrTyp	1	
GarageCar	1	1
Fireplaces	1	12
KitchenQua	1	1
Functional	1	
YrSold	1	9
Neighborhc	1	1
LotArea	1	1
BsmtUnfSF	1	1
TotalBsmtS	1	1
1stFlrSF	1	21
GrLivArea	1	1
WoodDeck	1	1
OpenPorch	1	1
SaleCondit	2	20
HalfBath	3	
RoofStyle	4	1
BsmtFinSF	6	1
OverallCon	7	28
MSZoning	8	1
TotRmsAbv	12	17
YearBuilt	13	5
Exterior1st	15	15
2ndFlrSF	16	1
ExterCond	17	
YearRemoc	18	10
OverallQua	19	14
GarageArea	21	13

Berkeley
UNIVERSITY OF CALIFORNIA

Feature Selection|

Wrapper Method - Results

	RMSE
LinearRegression	38633.08
Ridge	39027.536
Lasso	38558.299
ElasticNet	41114.857
DecisionTreeRegressor	38439.478
RandomForestRegressor	28009.02



Neural Network

Berkeley
UNIVERSITY OF CALIFORNIA

Neural Net | Architecture and Implementation

Considering the size of data and nature of the problem, a tiny tower architecture designed for this regression problem.

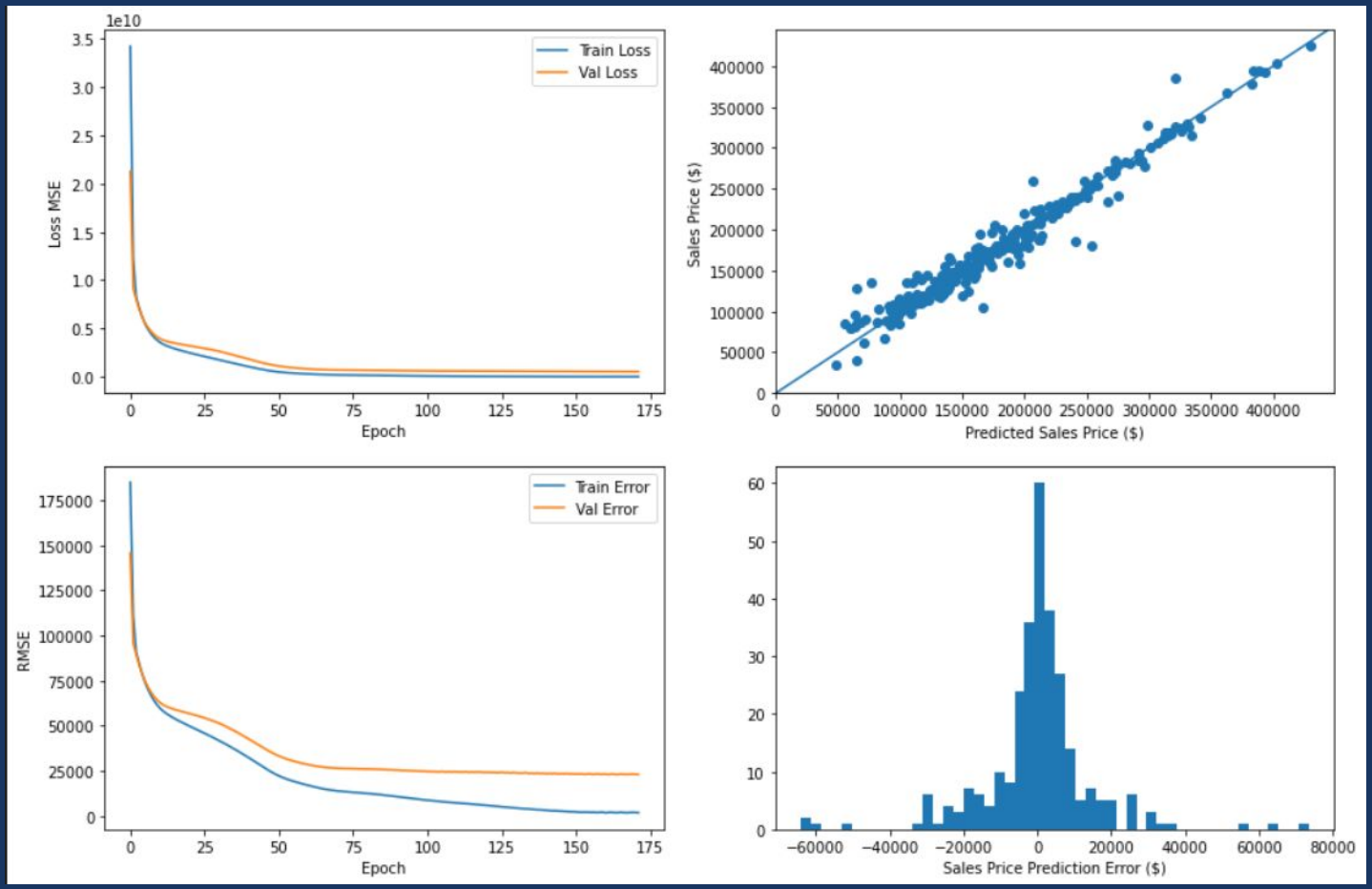
- Sequential
- 2 Hidden Layers
- Relu Activation Functions

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               14592

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 22,913
Trainable params: 22,913
Non-trainable params: 0
_________________________________________________________________
```

- TensorFlow
- Keras
- Optimizer: Adam
- Loss: MSE
- Metrics: RMSE
- 1000 Epochs
- Call backs
  - Save
  - Early Stopping
- Save, resume

Berkeley
UNIVERSITY OF CALIFORNIA

Neural Net | Result



RMSE for test: 16309.35

Berkeley
UNIVERSITY OF CALIFORNIA

Modules

- Regressors
  - Base Regressor
  - Decision Tree
  - Elastic Net
  - Lasso
  - Linear
  - Neural Net
  - Random Forest
  - Ridge
- Helpers
  - DataLoader
  - BaseRegressorPlot
- Configurations
  - num epochs
  - resume_nn
  - enable cache
  - log level
  - saved model file
  - data path
  - sections to skip

Berkeley
UNIVERSITY OF CALIFORNIA

Notebook Organization

Notebook 1:

- Project Overview
- EDA
- Missing Values
- Distribution of Nominal and categorical variables

Notebook 2:

- Regression and Hyperparameter Tuning

Notebook 3: Neural Networks

Notebook 4: Features and Wrapper Functions

Notebook 5: PCA and Random Forest on Reduced Dimensions

Notebook 6: Conclusion and Further-work

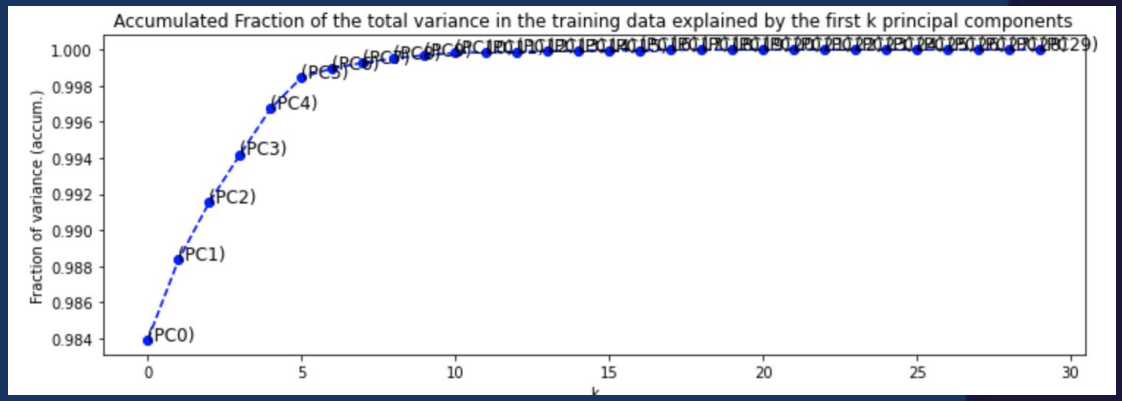
Berkeley
UNIVERSITY OF CALIFORNIA

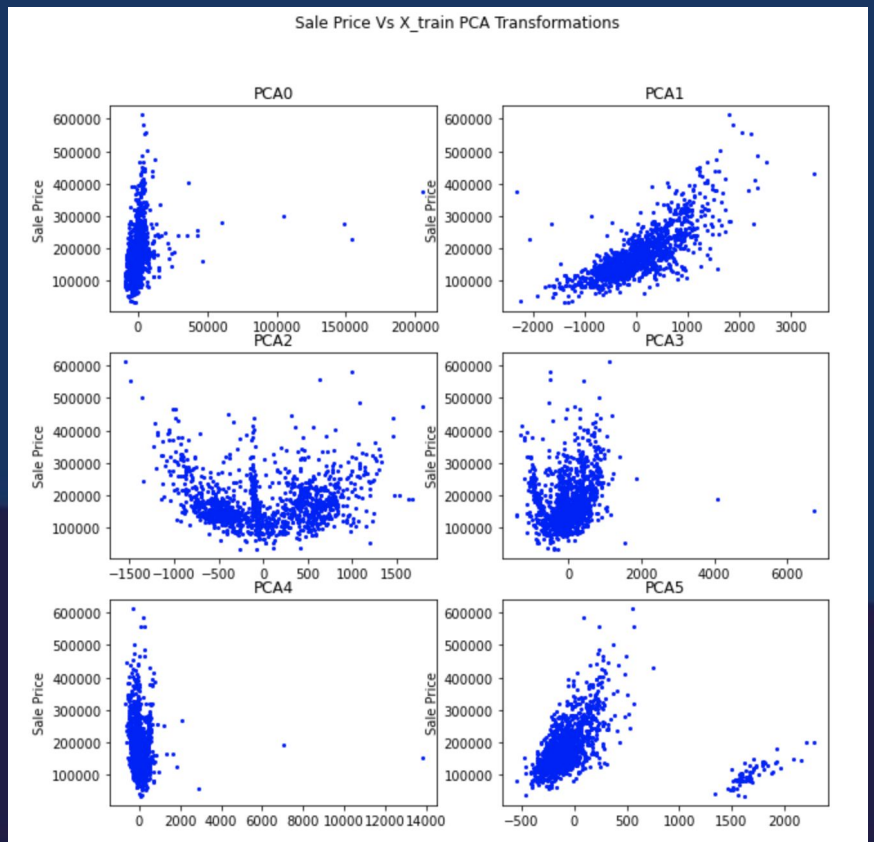
PCA (Failed Attempt)

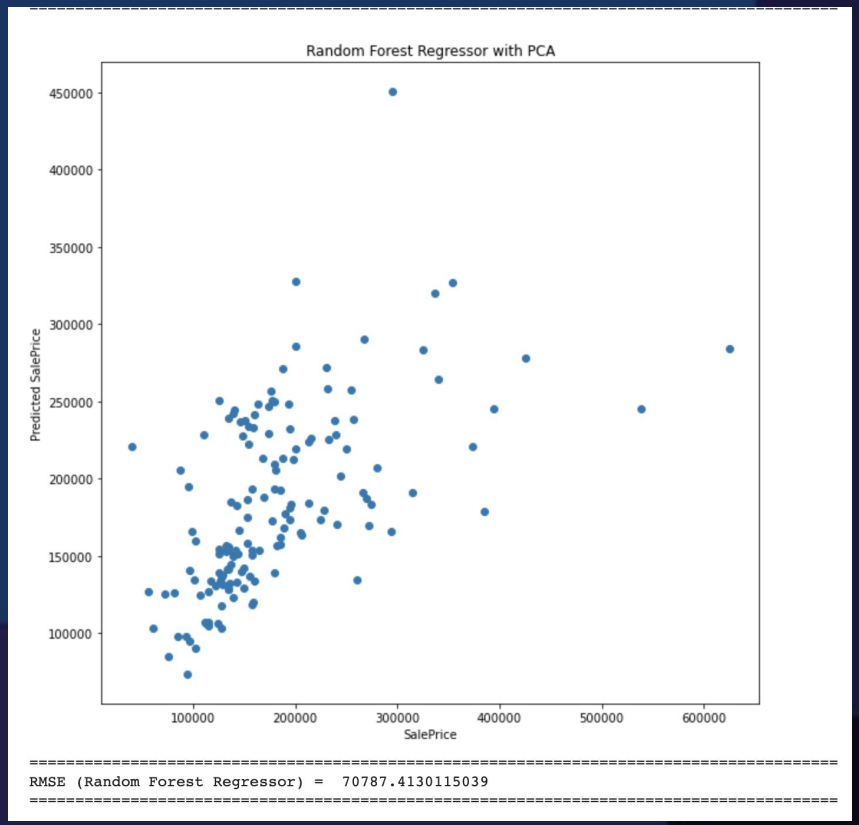
Berkeley
UNIVERSITY OF CALIFORNIA

PCA

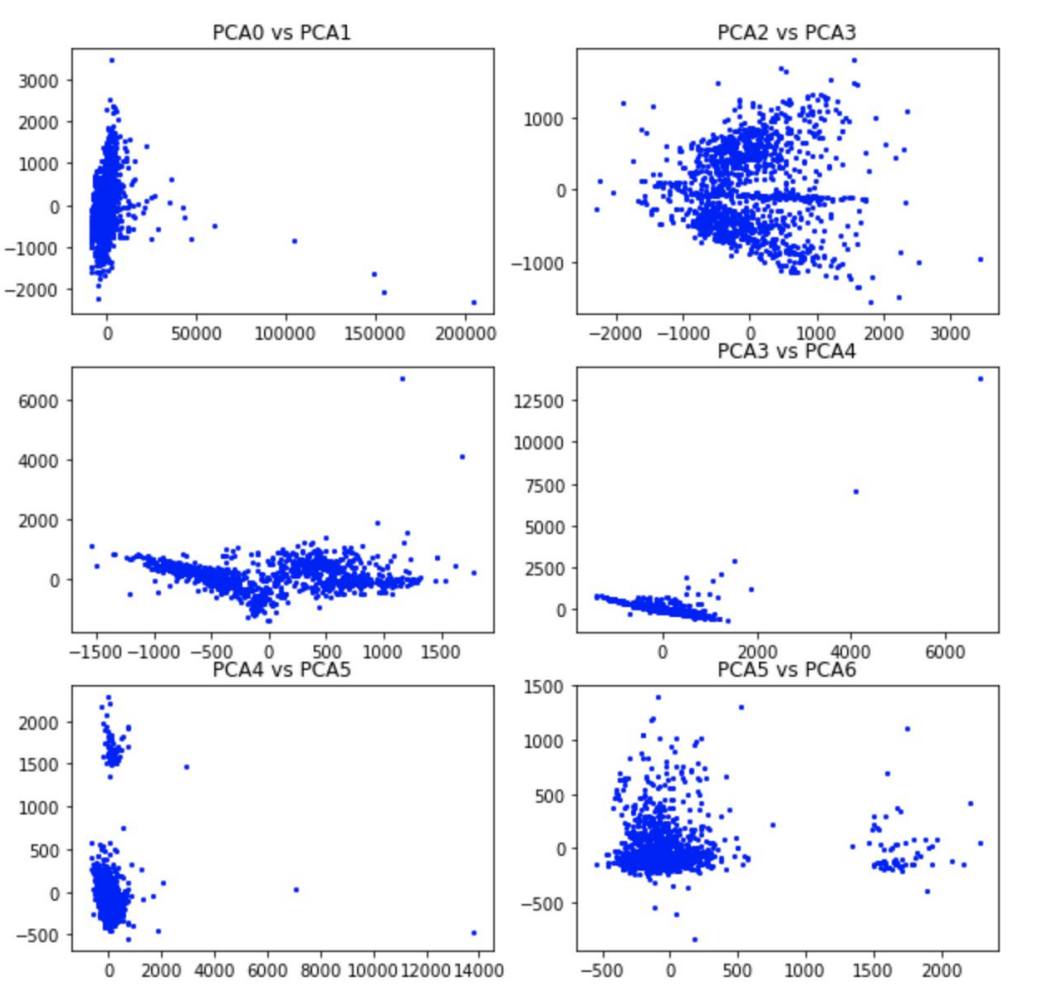
Numerical Variables

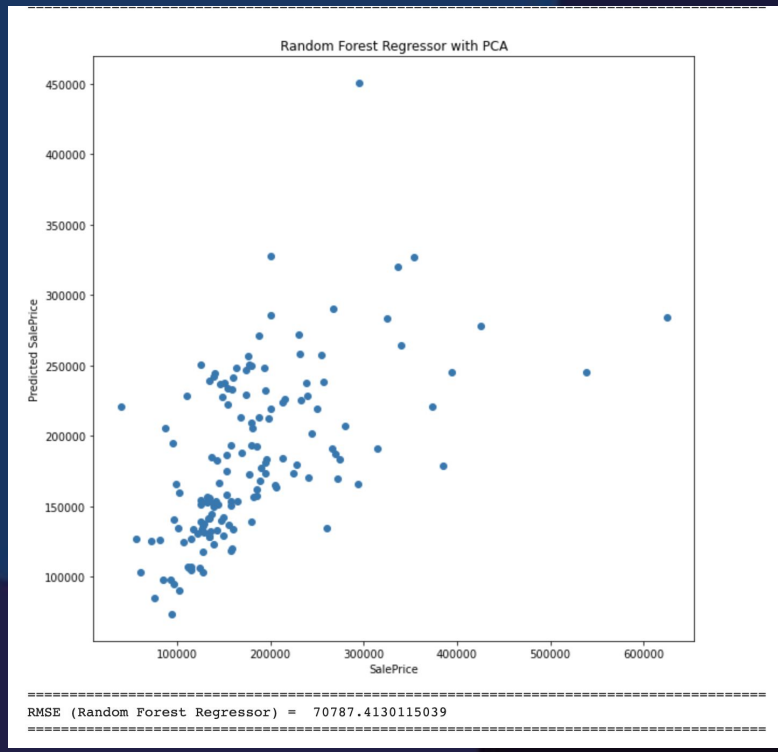






PCA





Berkeley
UNIVERSITY OF CALIFORNIA

Thanks!

Berkeley
UNIVERSITY OF CALIFORNIA

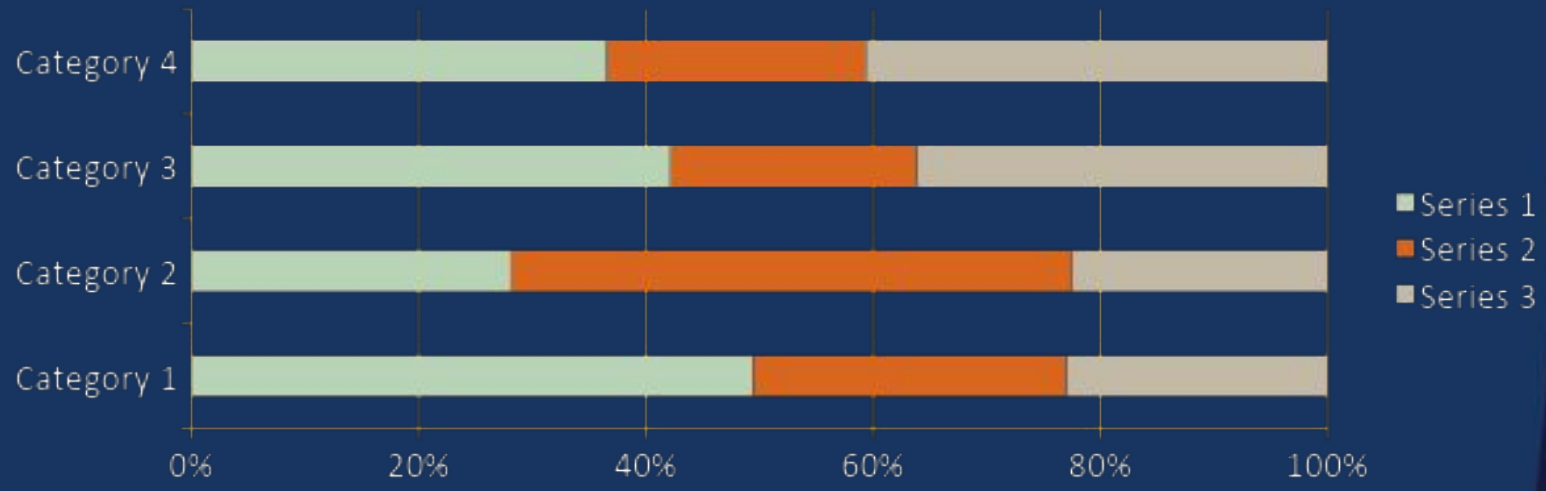
Header Georgia 22 pt

Lucida Grande 14 pt

Nemo enim ipsam volptatem quia voluptas sit aspernatur aur adit amet eius modi tempora incidunt ut labore et dolore magnam

Berkeley
UNIVERSITY OF CALIFORNIA

Header Georgia 42 pt



Berkeley
UNIVERSITY OF CALIFORNIA

Header Georgia 22 pt

Lucida Grande 14 pt
Nemo enim ipsam volptatem quia voluptas sit aspernatur aur adit amet eius modi tempora incidunt ut labore et dolore magnam

Lorem	Dolor	It	Enim	Color
43	60000	5600	1700	34000
35	55000	89056	4359	0349
435	3245	23243	25567	123
56	7	45	09867	456
345	768	0980	43138	2389
Fuga	Modi	Tempora	Nihil	Incidunt
Cor	Et vel	Wquis	Autem	Nulla

Header Georgia 22 pt

Lucida Grande 18 pt

Berkeley
UNIVERSITY OF CALIFORNIA

Header Georgia 42 pt

- Body copy, Lucida Grande 18 pts
  - Add your secondary bullet point here
    - Add your tertiary bullet point here

Berkeley
UNIVERSITY OF CALIFORNIA