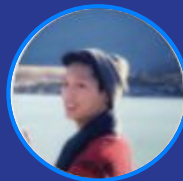





Machine Learning for Housing Price Prediction

Michelle Lai, Prashant Shukla, Dominic Woolridge, and Mark Wu



Agenda

- Dataset Overview
 - Project Ideation & Problem Statement
 - Exploratory Data Analysis: Data Pre-processing
 - EDA: Univariate, Bivariate, Visualization & Correlation Analysis
 - Model Exploration & Results
 - Multiple Linear Regression
 - Lasso Regression
 - Deep Learning Neural Networks
 - Model Comparisons & Technical Analysis
 - Conclusions & Business Application
 - Key Takeaways
 - Team Github Information
- 

Project Ideation and Problem Statement

Weigh different factors effectively when predicting housing prices and draw insights on how they interact with each other

Challenge 1	Challenge 2	AI/ML Fit	AI/ML Fit
<p>Housing Market is fluid and always evolving</p> <p>New factors influence the property prices. How can new drivers be calibrated in new model?</p>	<p>Create insights from immense amount of data</p> <p>Immense amount of data already exists. However, it is challenging to detect shifts in market that are just starting to happen</p>	<p>Great tool for High-Dimensional Property datasets</p> <p>ML algorithms don't just analyze data, they also train on the provided data to improve the predictions. ML can work on structured and unstructured data which can furnish further insights into hidden patterns</p>	<p>Big Data and available computing power</p> <p>Volumes of data and vastly improved computing has driven enterprise focus, academic research and widespread adoption of the AI/ML domain</p>

Dataset Overview

Source

[Kaggle](#) (Scraped from Zillow using their API)

Dependent Variable

Latestprice which is most recent available price

Austin, Texas House Listings

The Austin housing market is one of the hottest real estate markets in the country. The dataset contains sold houses in and around Austin

Data Size

15,171 Listings
Jan-2018 - Jan-2021
47 Columns

Independent Variable

45 Features
1 Key Column

+

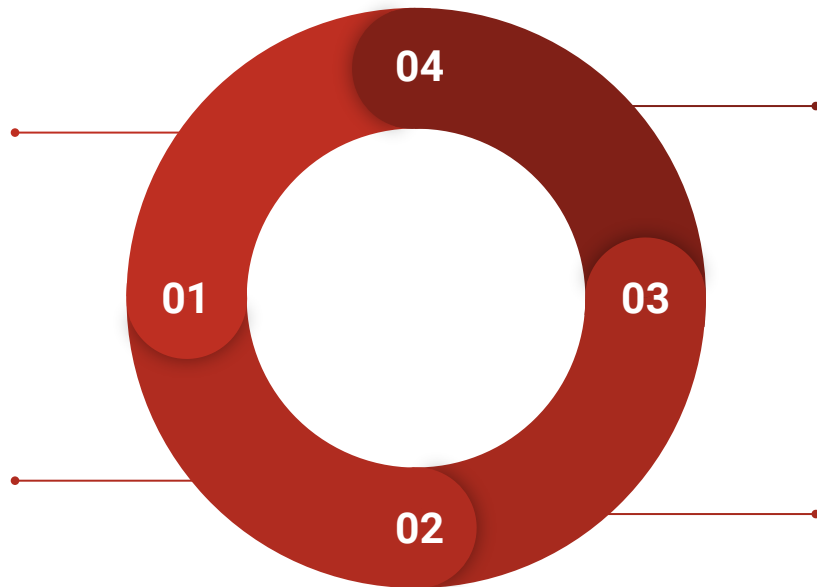
EDA: Dataset Pre-Processing

Treat Suspect Data

Correct (remove) data which looks suspect and represents very small fraction of entire population.
Check for missing values

Outlier Correction

Understand the skew of the distribution and correct for outliers. Price outliers were corrected, plot sales were removed



Data Transformation

Convert booleans to numeric and transform categorical data using one hot encoding

Treat multicollinearity

Assess columns exhibiting multicollinearity and treat

Preprocess Home Desc using NLP

Home descriptions can play an influential role in gaining public interest in a property, which in turn effects the final sale price. To capture this data, the following was performed:

- Text cleaning
- Encoding text into numeric vectors
- Splitting vectors into new DataFrame
- Dimensionality reduction via PCA
- Concat with original DataFrame

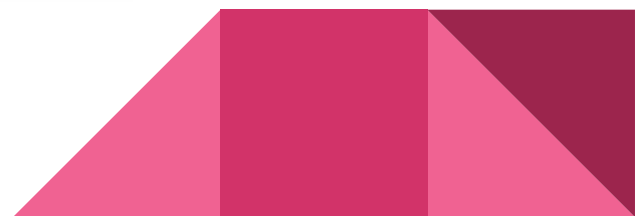
description
Great starter home on a corner lot. This home has 3 beds, 2 baths and a spacious storage shed in the back. Great location, about 12 miles to downtown Austin.

Figure 1

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents

Figure 2

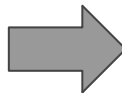


Home Desc contd: PCA

```
✓ [96] print(df_nlp.shape)
0s      print(df_nlp.head())

(15171, 100)
   0  1  2  3  ...  96  97  98  99
0  0.000000  0.0  0.000000  0.000000  ...  0.0  0.0  0.0  0.472661
1  0.000000  0.0  0.156604  0.110749  ...  0.0  0.0  0.0  0.000000
2  0.000000  0.0  0.255752  0.000000  ...  0.0  0.0  0.0  0.000000
3  0.306982  0.0  0.000000  0.000000  ...  0.0  0.0  0.0  0.000000
4  0.000000  0.0  0.000000  0.000000  ...  0.0  0.0  0.0  0.000000

[5 rows x 100 columns]
```



```
✓ [25] print(df_pca.shape)
0s      print(df_pca.head())

(15171, 2)
   0  1
0  7.126972 -0.766971
1 -1.045979  3.891180
2 -2.246494 -3.316062
3 -0.602707  3.010375
4 -2.240411 -1.590358
```

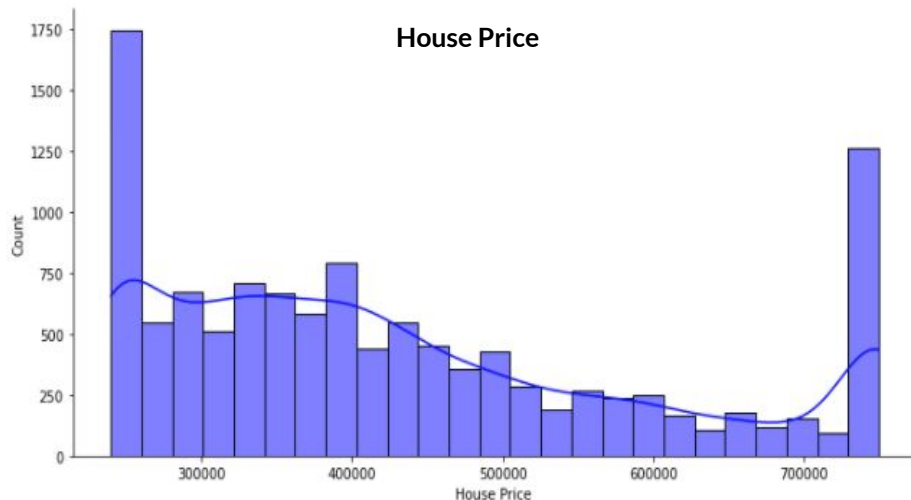
- Utilized SVD decomposition to reduce dimensionality from 100 to 2
- First two components have highest variance. Remaining components carry minimal information.

EDA: Univariate Analysis

Main Variable of Interest: House Price

	House Price in \$s
Count	11,749
Mean	434,213.70
Std	161,109.05
Min	239,900.00
25%	300,000.00
50%	398,000.00
75%	535,000.00
Max	749,801.00

Summary Statistics

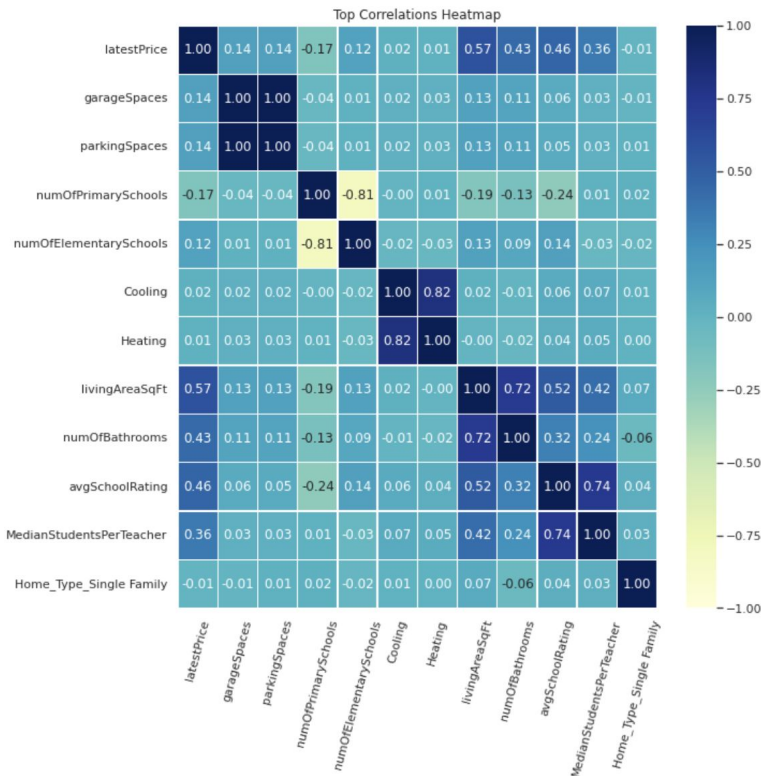


Observations:

- Closest to a Bimodal Distribution
- Peaks at the end and prices fairly distributed across the dataset

EDA: Visualization & Correlation Analysis

Bivariate Analysis to explore relationship between pairs of key features



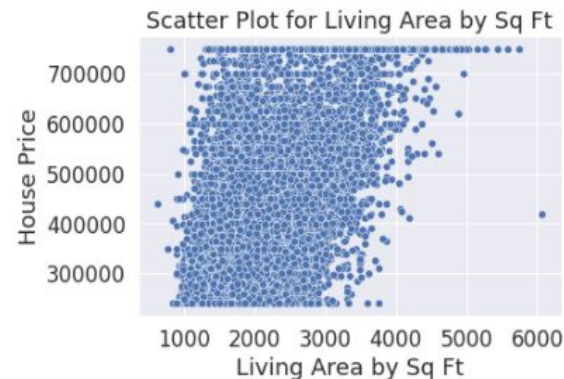
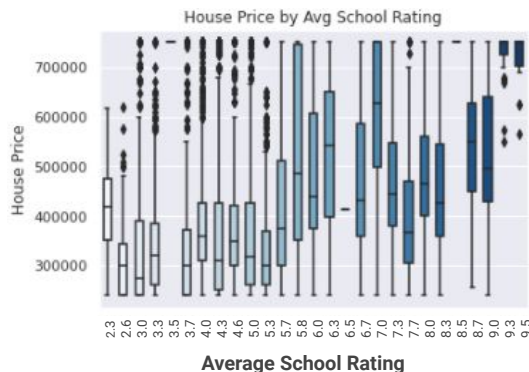
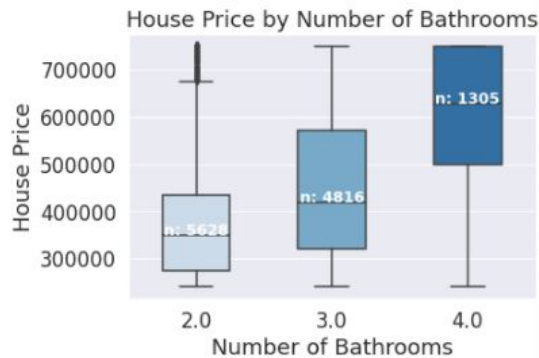
- Pairplot for all variables
- Correlation & Heatmap
 - Strongest correlated pairs:
 - Garage Spaces & Parking Spaces
 - No. of Primary Schools & No. of Elementary Schools
 - Cooling & Heating
 - Living Area Sq. Ft. & No. of Bathrooms
 - Average School Rating & Median Students per Teacher

EDA: Visualization & Correlation Analysis

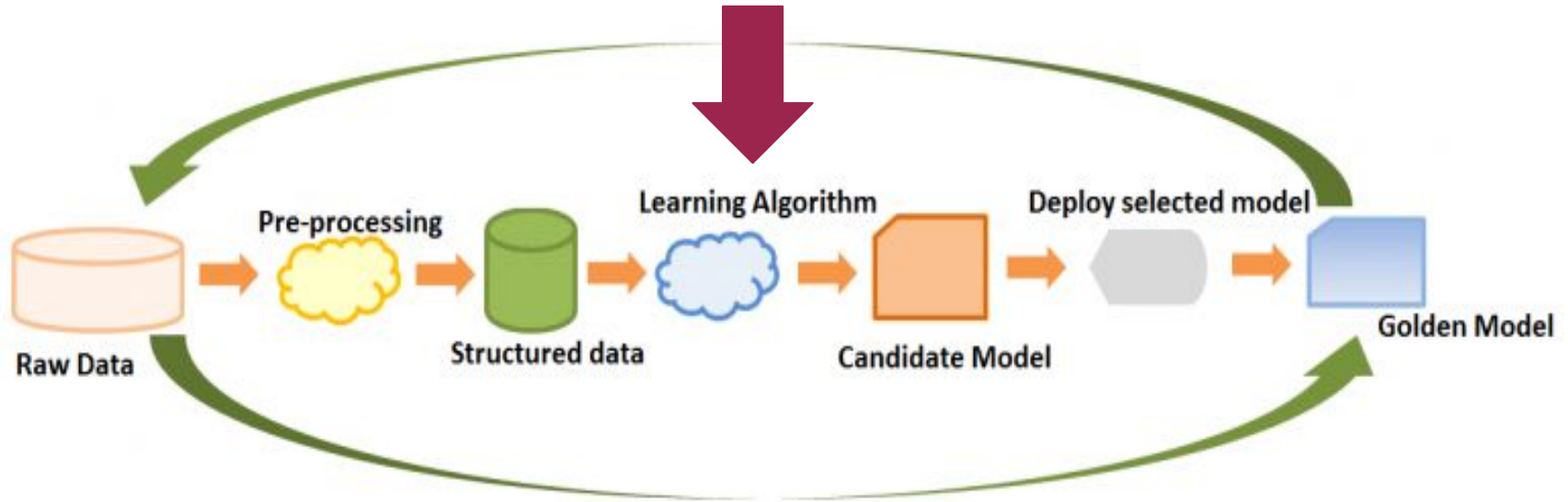
Bivariate Analysis to explore relationship between the Target Variable (House Price) and Features

Features with the Highest Correlations to House Price:

- Living Area by Square Foot: 0.57
- Average School Rating: 0.46
- Number of Bathrooms: 0.43
- Median Students per Teacher: 0.36



Model Building to select our ML Algorithm



Multiple Linear Regression

- Dependent Variable (Y): **House Price**
- Independent Variables (X): **Key Features**

Features (X)	Coefficient	Standard Error	T-stat	P> t	[0.025	0.975]
Number of Accessibility Features	7678.5166	5750.734	1.335	0.182	-3593.883	1.90E+04
View	6850.6445	2321.696	2.951	0.003	2299.733	1.14E+04
Number of Patio & Porch Features	4098.8827	1251.549	3.275	0.001	1645.637	6552.128
Number of Window Features	2961.0822	2085.638	1.42	0.156	-1127.117	7049.281
Median Students Per Teacher	2959.4221	1050.937	2.816	0.005	899.409	5019.435
Number of Security Features	2476.4604	1356.415	1.826	0.068	-182.34	5135.261
Home Type: Single Family	1594.8584	3.99E+04	0.04	0.968	-7.66E+04	7.98E+04
Number of Appliances	1166.3447	519.145	2.247	0.025	148.734	2183.955
Living Area by Sq. Foot	120.346	2.401	50.132	0	115.64	125.052
Number of Bathrooms	3.02E+04	2364.827	12.786	0	2.56E+04	3.49E+04
Year Built	-969.6139	65.108	-14.892	0	-1097.237	-841.991
Zipcode	-2461.7804	67.883	-36.265	0	-2594.842	-2328.719
Number of Price Changes	-7786.8087	380.047	-20.489	0	-8531.764	-7041.854
Average School Distance	-9124.4283	994.446	-9.175	0	-1.11E+04	-7175.149
Number of Bedrooms	-2.14E+04	2276.088	-9.405	0	-2.59E+04	-1.69E+04

Feature Significance

- Highest Coefficients drive largest changes in House Price
- T-statistic > 2 & < -2
- P-value $< .05$

Multiple Linear Regression Results & Performance Analysis

Most Significant Features based on MLR:

- Living Area by Square Foot
- Whether the house has a View
- Number of Patio & Porch Features

Limitations & Issues:

- Possible multicollinearity amongst variables
- Linear model does not take into account the interaction between the multiple variables
- Neural Networks as a better method to capture interactions between features and possibly yield stronger performance

Linear Regression Metrics on Test Set

Mean Absolute Error	76,221.67
Root Mean Squared Error	100,468.60
Adjusted R2	63.40%

Lasso Regression Model & Results

Feature	Coefficient
zipcode	-45734.37
latitude	-4136.62
longitude	14262.02
propertyTaxRate	-34300.57
garageSpaces	-8847.55
parkingSpaces	3540.83
yearBuilt	-19196.44
numPriceChanges	-20316.89
latest_salemonth	3498.84
latest_saleyear	13453.63
numOfPhotos	3923.54
numOfAccessibilityFeatures	423.66
numOfAppliances	1832.19
numOfParkingFeatures	7713.51
numOfPatioAndPorchFeatures	4136.04
numOfSecurityFeatures	1416.69
numOfWaterfrontFeatures	1657.92
numOfWindowFeatures	2060.15
numOfCommunityFeatures	-3454.02
lotSizeSqFt	-605.45
livingAreaSqFt	88365.53
numOfPrimarySchools	13876.45
numOfElementarySchools	17205.05
numOfMiddleSchools	-2878.73
numOfHighSchools	-11741.91
avgSchoolDistance	-8055.06

Feature	Coefficient
avgSchoolRating	59660.11
avgSchoolSize	-21850.57
MedianStudentsPerTeacher	4225.29
numOfBathrooms	18780.97
numOfBedrooms	-13060.54
numOfStories	-9177.80
Association	-33793.21
Cooling	1445.82
Heating	-1748.73
spa	4599.90
view	2923.82
Home_Type_Condo	13493.11
Home_Type_Mobile	-755.96
Home_Type_MultiFamily	-2475.63
Home_Type_Multiple Occupancy	1248.09
Home_Type_Other	3118.22
Home_Type_Residential	3519.51
Home_Type_Single Family	9006.48
Home_Type_Townhouse	8022.74
Home_Type_Vacant Land	-814.28
City_austin	32479.73
City_driftwood	1454.48
City_dripping springs	0.00
City_manchaca	2048.36
City_manor	4460.72
City_pflugerville	3348.32

Purpose: Feature selection/reduction

Lasso Regression Score

MSE	10,094,225,918
-----	----------------

Results

- Only one feature was equalized to zero
- Extremely High MSE
- Better to move forward with another ML method

Neural Networks

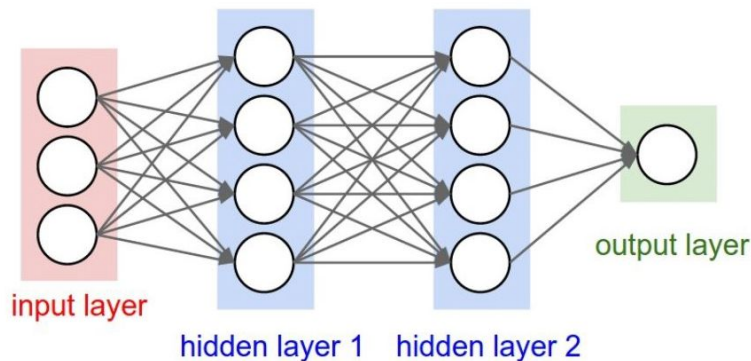


Figure 1

DL NN Algorithm Steps Summarized

- Forward pass from Input to Output layer and measure of the loss function.
- Backpropagation to measure loss contribution from each connection.
- Gradient descent to minimize cost function.

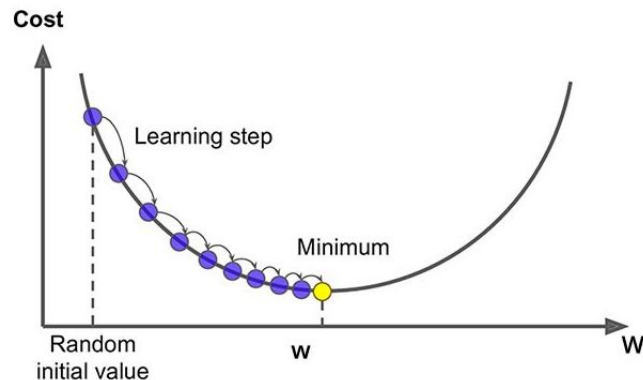


Figure 2

Neural Networks - Keras Setup

```
✓ [75] model.add(Dense(64, activation='relu', input_dim = 55))  
0s model.add(Dense(64, activation='relu'))  
model.add(Dense(64, activation='relu'))  
model.add(Dense(1)) #Returns number, not sigmoid 0,1 function  
  
model.compile(optimizer='adam',  
              loss='mse',  
              metrics=['mse', 'mae'])
```

```
✓ [76] early_stopping = EarlyStopping(  
0s     monitor='val_loss',  
     mode='min',  
     verbose=2,  
     patience=25)
```

```
[77] model.fit(X_train, y_train, epochs=2000,  
              batch_size=64, validation_data=(X_test, y_test),  
              callbacks=[early_stopping])
```

64 Neurons per layer

Two hidden layers

Output layer = Linear Activation

Optimizer: ADAM algorithm

EPOCHS: 2000 (Early Stop enabled)

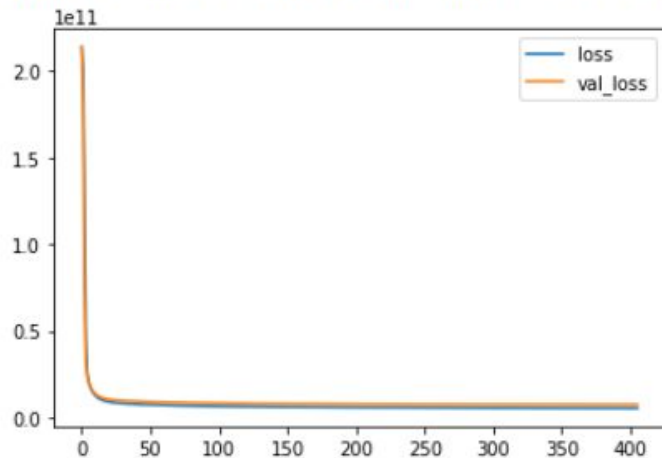
Metrics (MSE & MAE)



Neural Networks Results & Analysis

```
✓ [78] pd.DataFrame(model.history.history)[['loss', 'val_loss']].plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7b10dcd750>
```



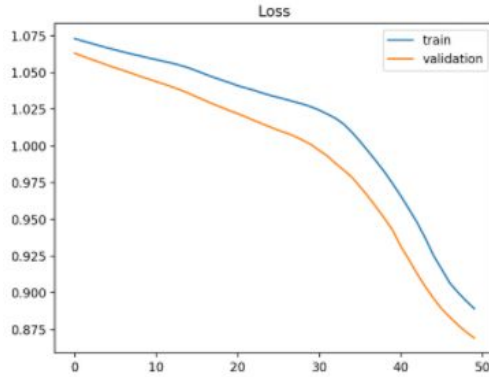
MSE (Loss): 5.3B - RMSE: 73K

Mae: 53K

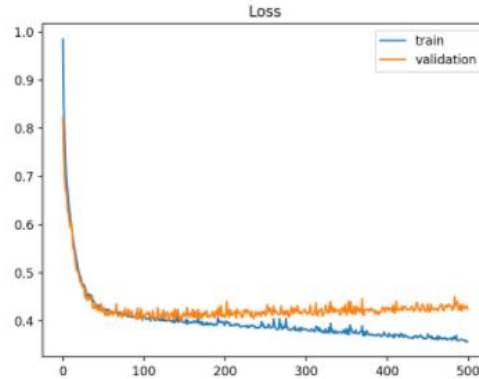
Validation Loss: 7.5B - RMSE: 86K

Validation MAE: 61K

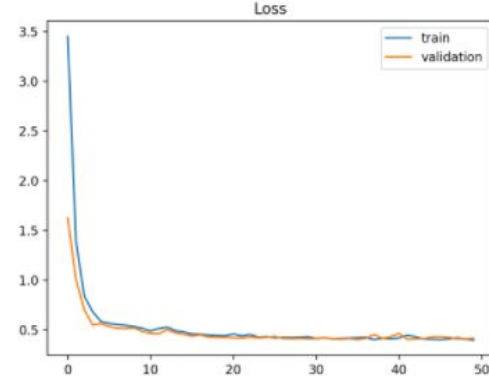
Neural Networks: Sample Graphs



Underfit



Overfit



Good fit

- Neural Network loss graph mimics Good fit sample graph
 - Achieved by using early stopping

ML Models Comparison & Technical Analysis

Validation Metrics			
Model	RMSE	MSE	MAE
Linear Regression	100K	10B	76K
Neural Network	86K	7.5B	59K

Neural Network outperforms linear regression, with ~20% lower validation loss.

What does this mean?

ML Techniques predicted top features associated with housing prices in Austin, Texas

1st

Living Area Square
Footage

2nd

House View

3rd

of Porch & Patios



How can this be used?



Real Estate Investors

Real Estate Agents



Key Takeaways and What's next?

Data, Data, Data!

**DATA IS
EVERYTHING**

What else could we have done?

- **Narrower Focus**
 - **Additional Models**
- 

Useful Packages



TextHero

Gensim

TensorFlow (DL)

SKlearn PCA



Team GitHub Accounts



Michelle Lai: <https://github.com/michylai>



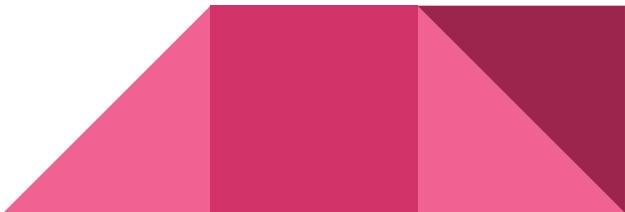
Prashant Shukla: <https://github.com/prashantslink>



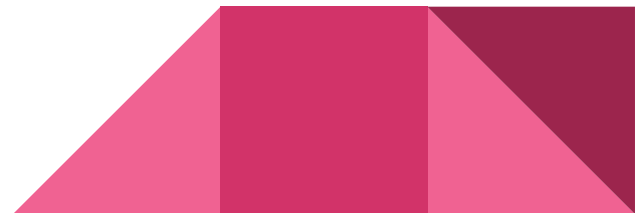
Dominic Woolridge: <https://github.com/dowoolridge>



Mark Wu: <https://github.com/markyo8>



THANK YOU!



Links to Pictures

<https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>

https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fquick-guide-to-gradient-descent-and-its-variants-97a7afb33add&psig=AOvVaw2VfJDsfMCS-4KulF0Qw68A&ust=1638327482646000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCPjQgc-Lv_QCFQAAAAAdAAAAABAD

<https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>

