

House Prices - Advanced Regression

Robert III





Problem

Inspiration for Project

Why is it Important?

What I want to Learn

Here are some potential project ideas for your DTSA 5511 final project that have readily available datasets and are feasible within one to two months:

1. Predicting House Prices

- **Dataset:** Kaggle's *House Prices - Advanced Regression Techniques* ([link](#))
- **ML Approach:** Regression (Deep Learning-based Regression Model)
- **Why It Works:** Well-structured dataset, extensive features for EDA, and easy-to-train models.

2. Sentiment Analysis on IMDB Movie Reviews

- **Dataset:** IMDB Reviews dataset ([TensorFlow Dataset](#))
- **ML Approach:** NLP (LSTM, Transformers, or simple Deep Learning models)
- **Why It Works:** A straightforward binary classification problem with preprocessed data.

3. Handwritten Digit Recognition

- **Dataset:** MNIST dataset ([TensorFlow/Keras](#))
- **ML Approach:** Convolutional Neural Networks (CNNs)
- **Why It Works:** Classic deep learning project with a simple model training pipeline.

4. Fake News Detection

- **Dataset:** Kaggle's Fake News dataset ([link](#))
- **ML Approach:** NLP-based classification model (LSTMs, BERT, TF-IDF + Neural Networks)
- **Why It Works:** Useful real-world application with available preprocessed datasets.

5. Stock Price Prediction

- **Dataset:** Yahoo Finance or Kaggle stock market datasets ([link](#))
- **ML Approach:** Time-series forecasting using LSTMs/GRUs
- **Why It Works:** Readily available stock data and interesting patterns in historical trends.

6. Medical Image Classification (e.g., Pneumonia Detection)

- **Dataset:** Chest X-ray dataset ([Kaggle link](#))
- **ML Approach:** CNN-based classification
- **Why It Works:** A meaningful application with clear evaluation metrics.



Data

What is Featured in the Data

Dimensions of Data

Anna Montoya and DataCanary. House Prices - Advanced Regression Techniques.
<https://kaggle.com/competitions/house-prices-advanced-regression-techniques>, 2016. Kaggle.



Exploratory Data Analysis

Inspect:

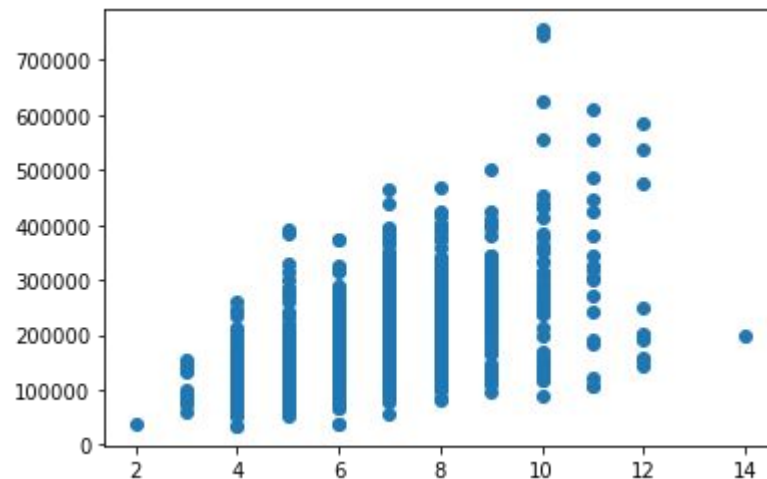
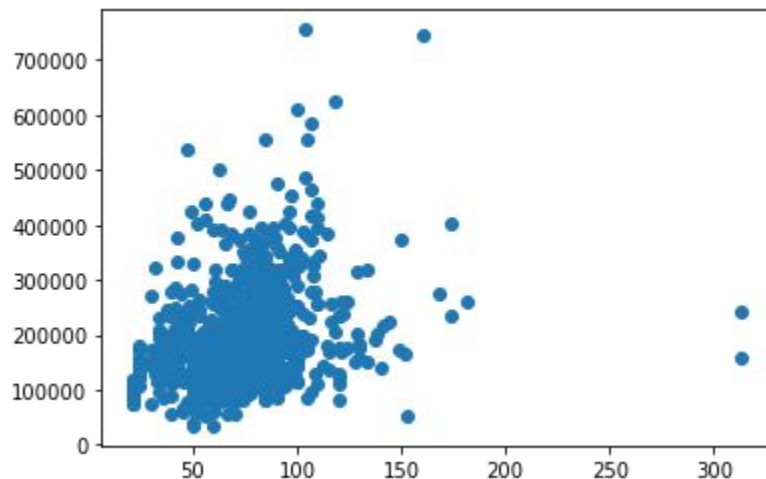
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoS
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeck
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	...	1460.000000
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	103.685262	443.639726	...	94.2442
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	181.066207	456.098091	...	125.3387
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	0.000000	...	0.000000
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	0.000000	0.000000	...	0.000000
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	0.000000	383.500000	...	0.000000
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	166.000000	712.250000	...	168.000000
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	857.000000



Exploratory Data Analysis

Visualize:





Exploratory Data Analysis

Clean:

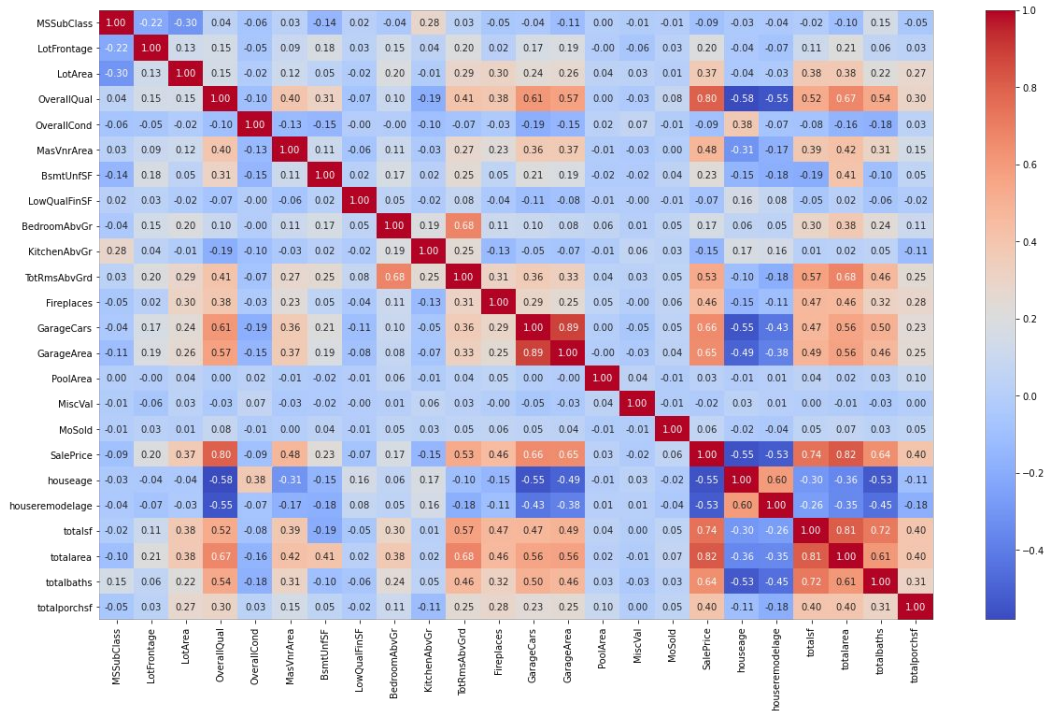
Outliers

Shown in Scatterplots

Null Values

Pool Quality

Misc Features





Models

Linear Regression -	9.510501616810296e+16
Random Forest Regression -	0.1339500658461279
XGB Regression -	0.11965682811517063
Ridge -	0.10894022173825084
Gradient Boosting Regression -	0.11333215220060736
LGBM Regression -	0.1275678077793979
Cat Boost Regression -	0.11313058072563338
Voting Regression -	0.11956123663283695



Results

Best-performing regression models (Gradient Boosting, XGBoost, CatBoost, LightGBM, Random Forest).

Stacking approach often leads to improved predictive performance compared to any of the individual base models

Final Result - 0.11966001781311514







Analysis

Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price.

House Prices - Advanced Regression Techniques

[Submit Prediction](#)[...](#)[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Submissions](#)

699	Satoshi Nagai		0.12542	10	2mo
700	Robert Comstock		0.12543	1	7h
 Your First Entry! Welcome to the leaderboard!					
701	Ilya Gredasov		0.12543	4	8d



Conclusion

Fun Project

Thanks for Viewing

Go Buffs