The Features App

Utilizing Data Science to aid rapid expansion into new markets by REMIX real-estate company.

by Okechukwu Ofili

February 21st, 2021

80

The Number of Features We Analyzed in The Ames, Iowa DataSet

The Problem Statement

The REMIX, real-estate company is looking to expand rapidly across America, starting with **Ames, Iowa**. This presentation aims to **identify key features** that drive home sales prices and **utilize prediction performance** to recommend data processing models for the **REMIX features app**.

Our Approach

STEP 1
Data Analysis and Feature Engineering

STEP 2
Feature Selection

STEP 3
Model analysis and hypertuning

STEP 4
Hyper Tuning

STEP 5
Conclusion and recommendatio ns

About Our Data

The **Ames Housing dataset**, was compiled by Professor Dean De Cock. The dataset contains a total of 2927 observations split across different explanatory variables:
23 nominal
23 ordinal
14 discrete
20 continuous

Step 1

Feature Engineering

Feature Engineering Overview

1. MISSING DATA

4 Features missing more than 80% of values where highlighted

2. DOMINANT DATA

31 Features where Single Variable had more than 78% Dominance. e.g Street Pavement

3. CATEGORICAL

These were binned in some instances (e.g Pool or No Pool). And the rest were dummified.

4. CONTINUOUS

Continuous data was used as-is in some cases. And in other cases were combined (e.g. Outdoor areas, all combined into one)

5. DISCRETE

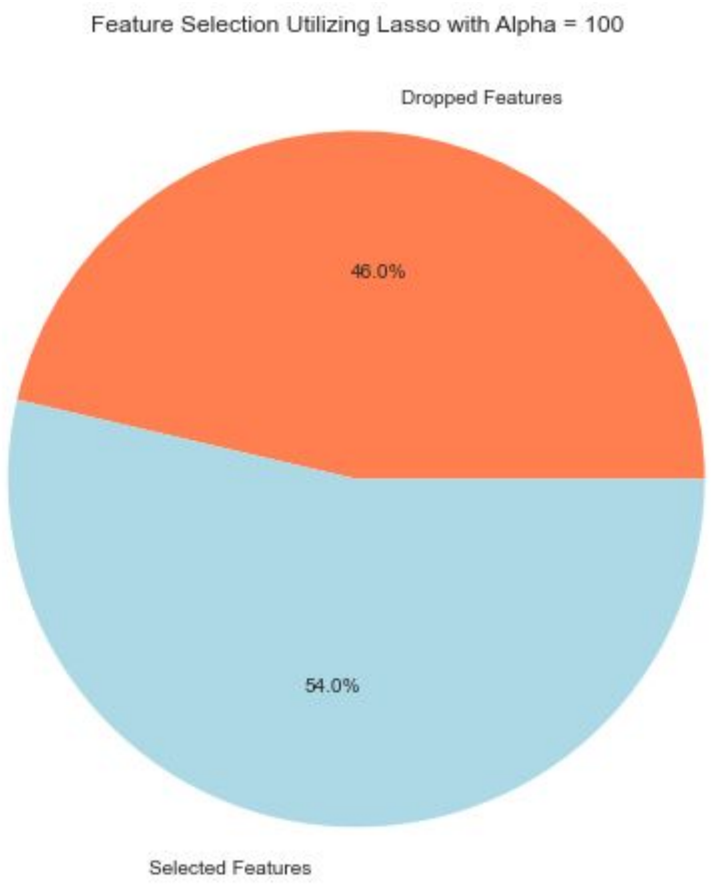
Some of the discrete data like Months Sold had to be converted to Strings (Categorical), so that **October is** not **ten times better than January**

6. ORDINAL

These were Categorical data that had logical trend to them. We converted these to numbers. So an excellent Kitchen was 5 times better than a poor kitchen.

Step 2

Feature Selection



Reduce Noise

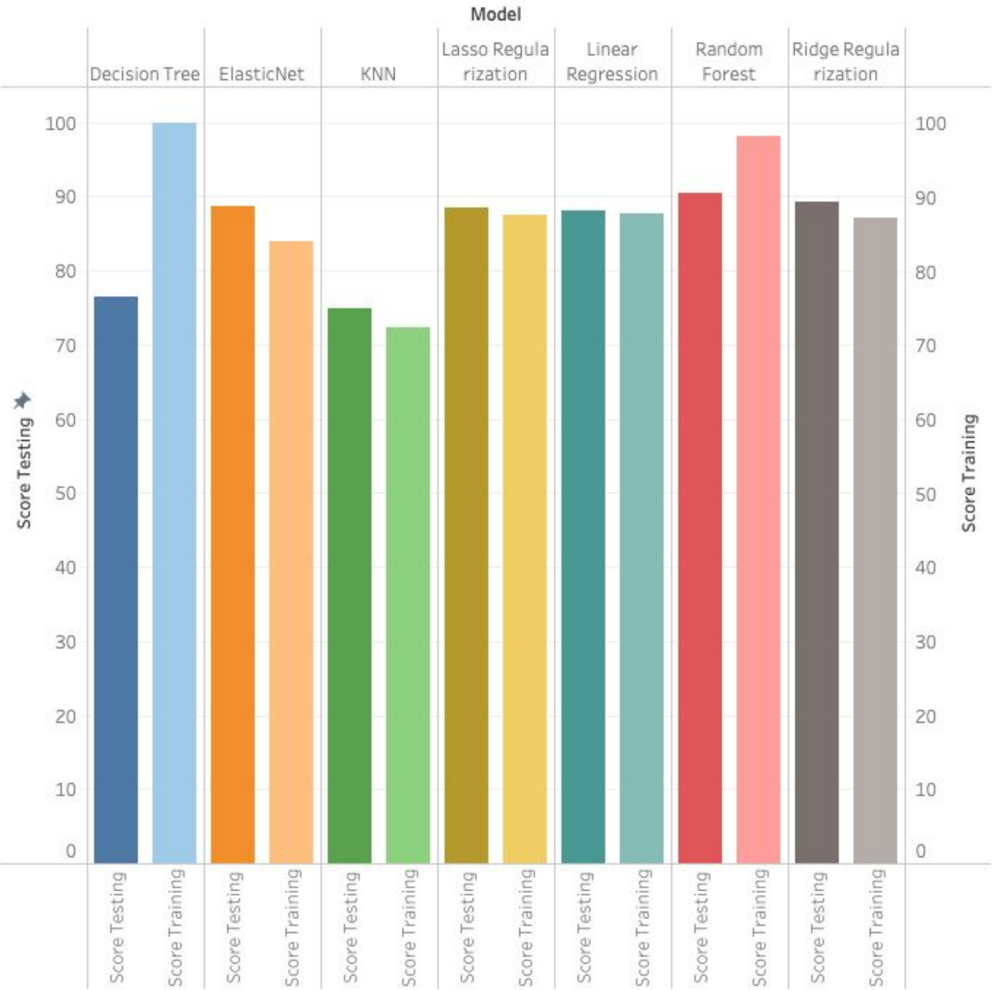
After feature engineering, we applied a Lasso Regularization on our Data, and found that with an Alpha of 100, over 50% of our features were dropped.

Step 3

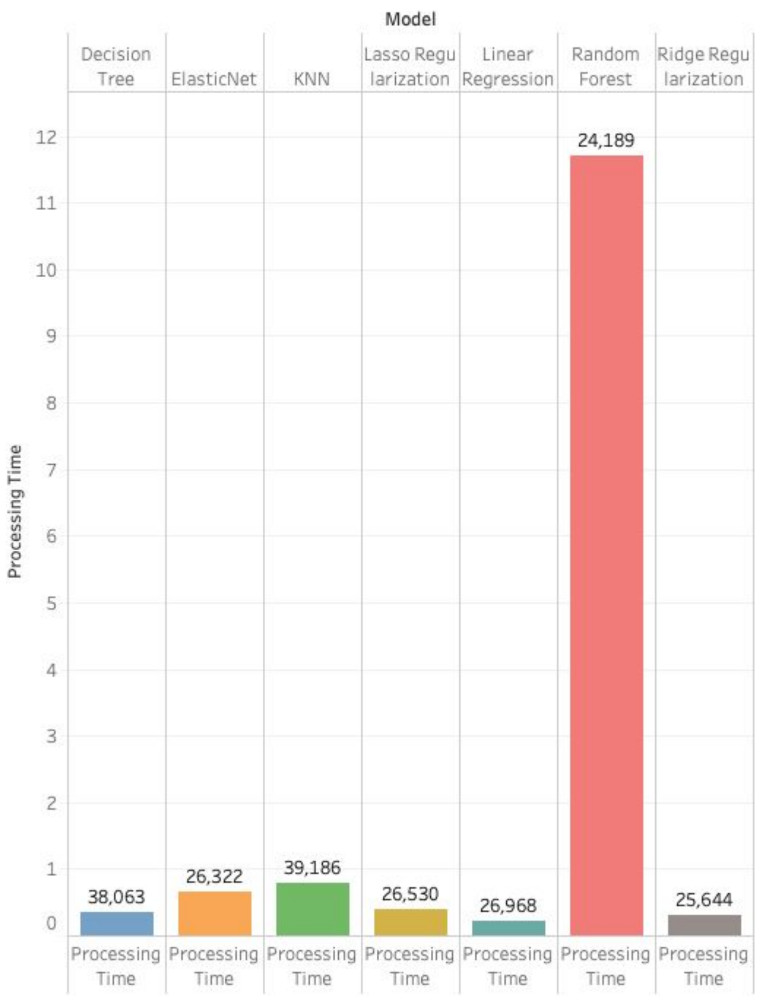
Model Analysis and Selection

Algorithm Performances

	Score Training	Score Testing	Score Testing Scaled	Score Testing Scaled	Process Time (sec)	RMSE Score	Kaggle RMSE
Random Forest	98	**90**	98	**91**	**11.7**	**24189**	29079
Ridge	87	89	88	88	0.3	25643	33026
Elastic Net	84	89	87	88	0.7	26322	-
Lasso	88	89	88	89	0.4	26530	-
Linear Regression	88	88	88	88	**0.2**	26968	**33017***
Decision Tree	**100**	76	**100**	75	0.4	38062	-
KNN	73	75	82	83	0.8	39185	-



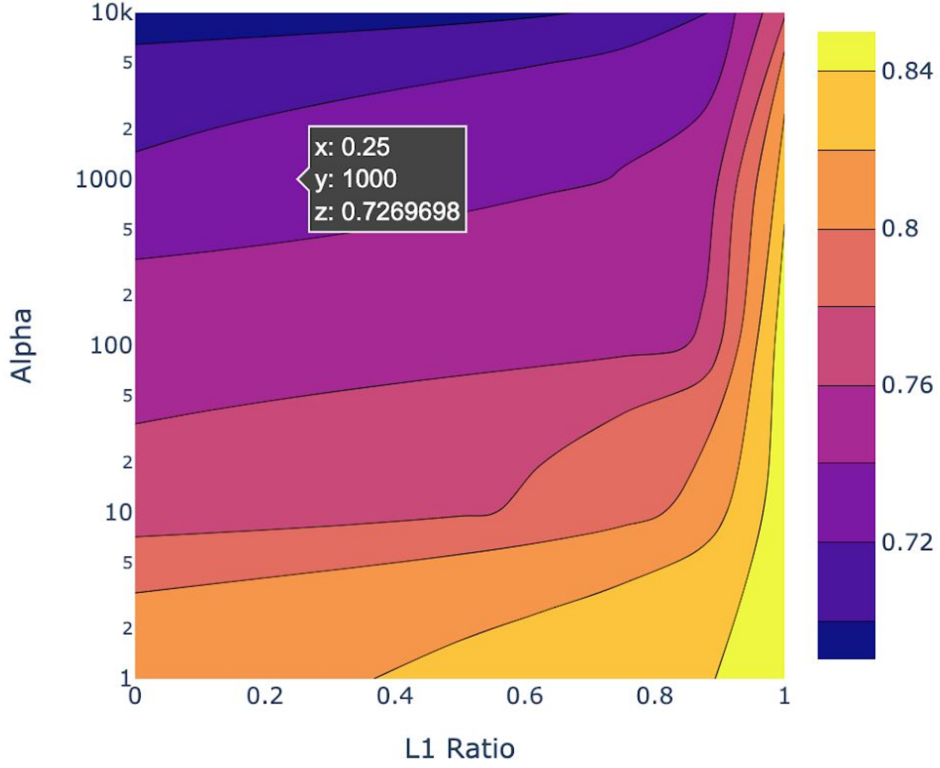
From here we visually see observations from our table. The overfitted decision tree is dominant and the stable performance of Linear Regression is clear.



Although our **Random Forest** had the best RMSE score, it took the longest time to process. Approximately 4 times the time for all the other Algorithms combined!

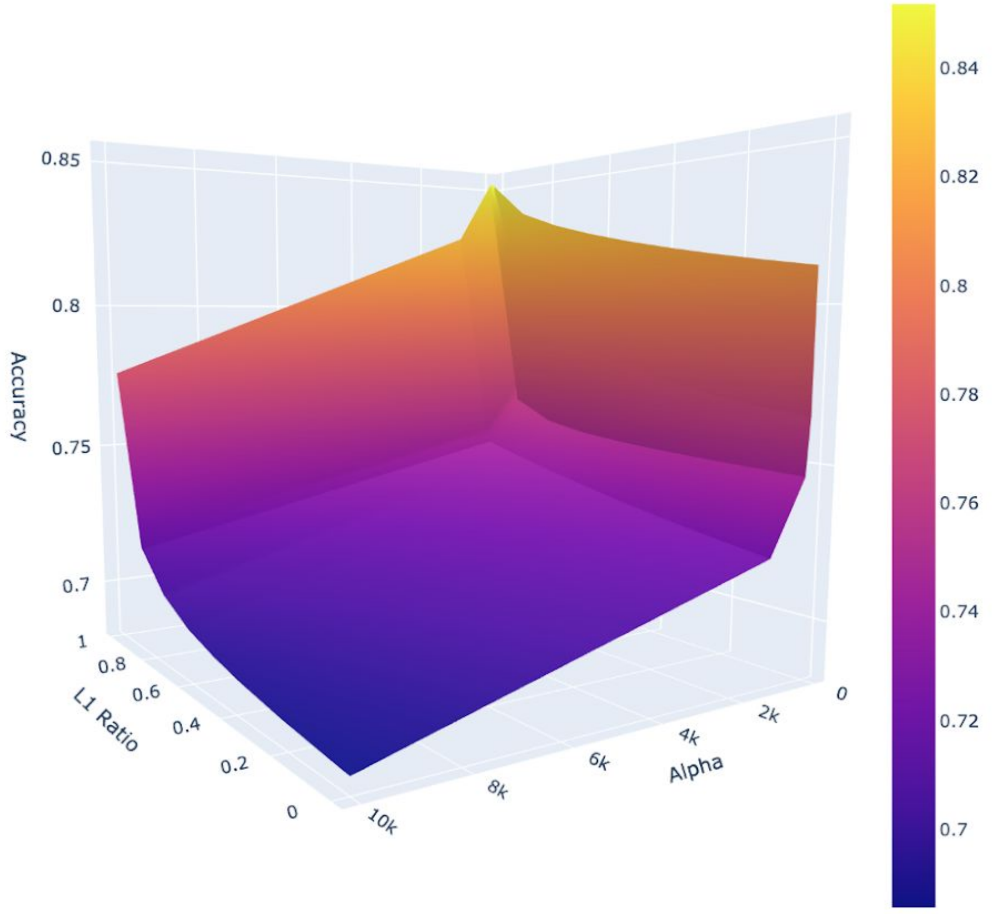
Step 4

Hyper Tuning



Hypertuning Contour Plot

I picked the Elastic Net algorithm because it has properties of both Lasso Regularization (L1) and Ridge Regularization (L2).

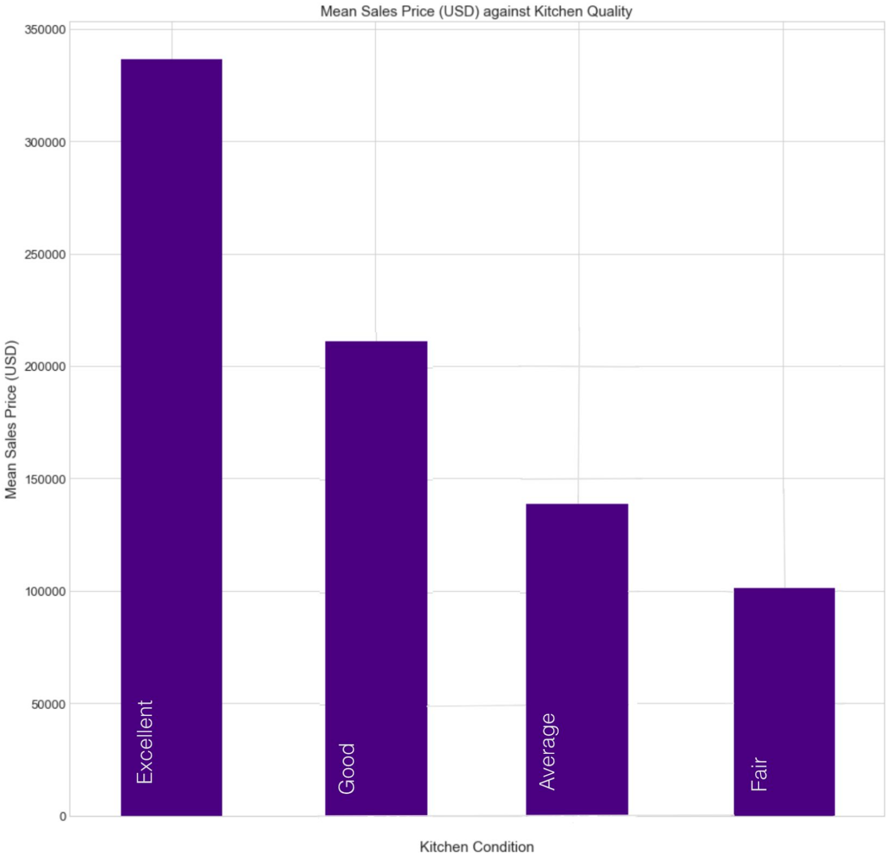


3-D Hyper Tuning Contour Plot

This rendering gives us a depth feel for how our model performs. At an alpha greater than 2,000 and L1 ratio greater than 0.8 we see a drop off.

Step 5

Conclusion and Recommendations

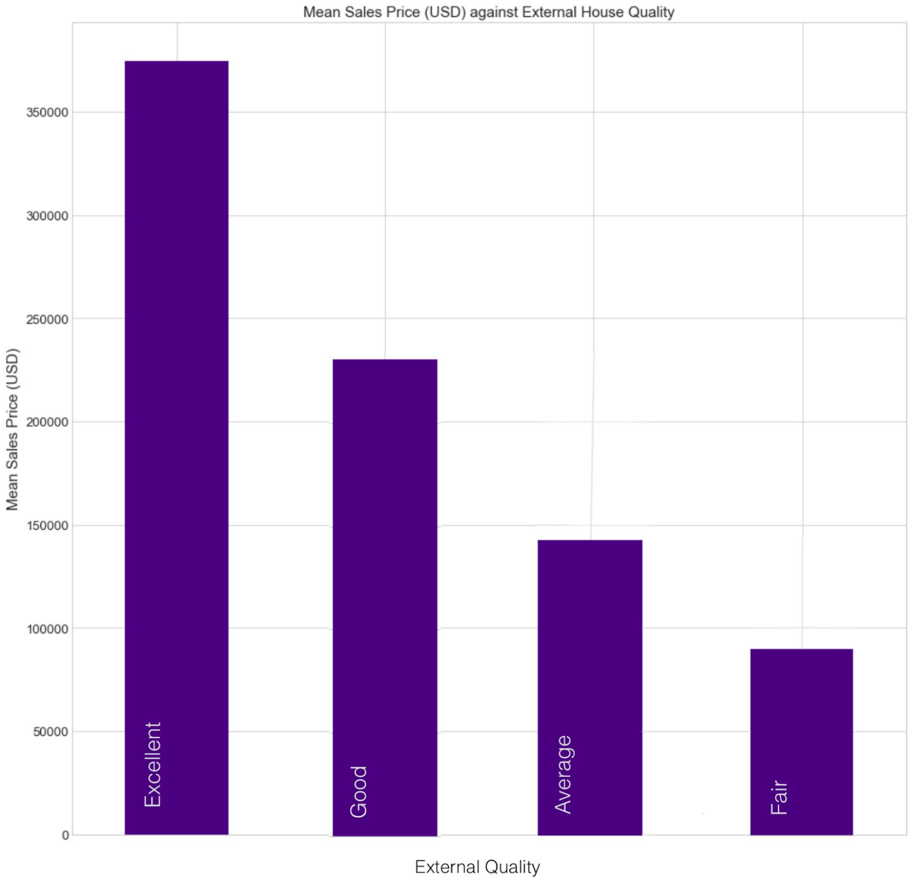


Mutable Features

These are features that can be modified, they are a profit mine for our client. The top mutable features we would recommend for our client to focus on in Ames, Iowa is:

Kitchen

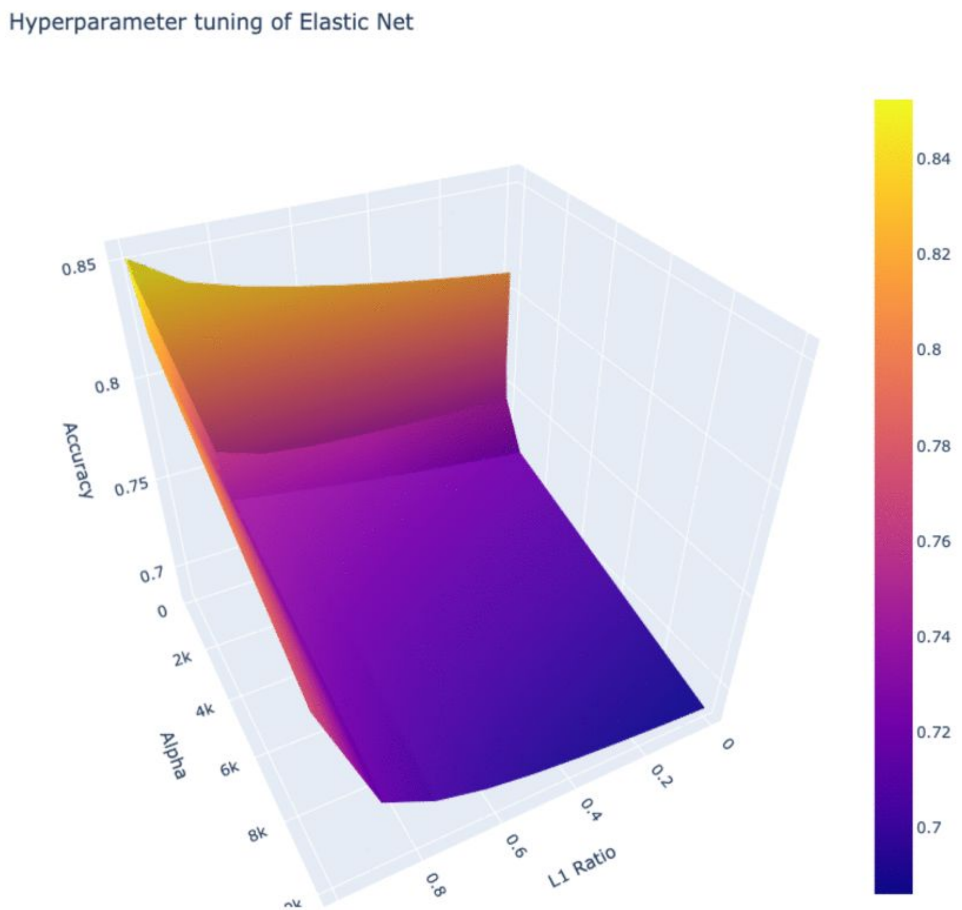
As we can see in the bar plot, having just a Good Kitchen on Average sold for more than **200,000 USD**, while a Fair kitchen barely attracted **100,000 USD** on average.



Immutable Features
These are features that our company should access before buying properties. They are immutable because they are very hard to change.

Ground Living Area and Basement Size: This is valid across most cities, the larger the Living area the more costly the house is.

External Quality of the House: This can be visualized here, the houses with an [excellent quality grade] sold on average for more than **350,000 USD** and those with fair quality sold on average for less than **100,000 USD.** So think brick versus Hardiplank.



App Model

With all the time spent hypertuning our parameters, the best performance of our models in predicting RMSE scores of unseen data was a simple **Linear Regression** that has had a **Lasso Regression** applied to it to suppress features.

This **Linear Regression Model** also works best when a **log** is applied to the skewed data.

Thanks!

Does anyone have any questions?

oaofili@gmail.com
+1 832 685 4145
ofilispeaks.com