Project Goal
Predict home values based on Bexar County data

Summary
I was tasked with predicting five home values based on Bexar County Appraisal District data. The 29 variables provided were defined, cleaned, and explored. Multiple types of regression models were run in order to predict home values, including: OLS (Ordinary least squares), GLM (generalized linear model), random forest, GAM (generalized additive model), SVM (support vector machine), and MARS (multivariate adaptive regression splines, using earth).

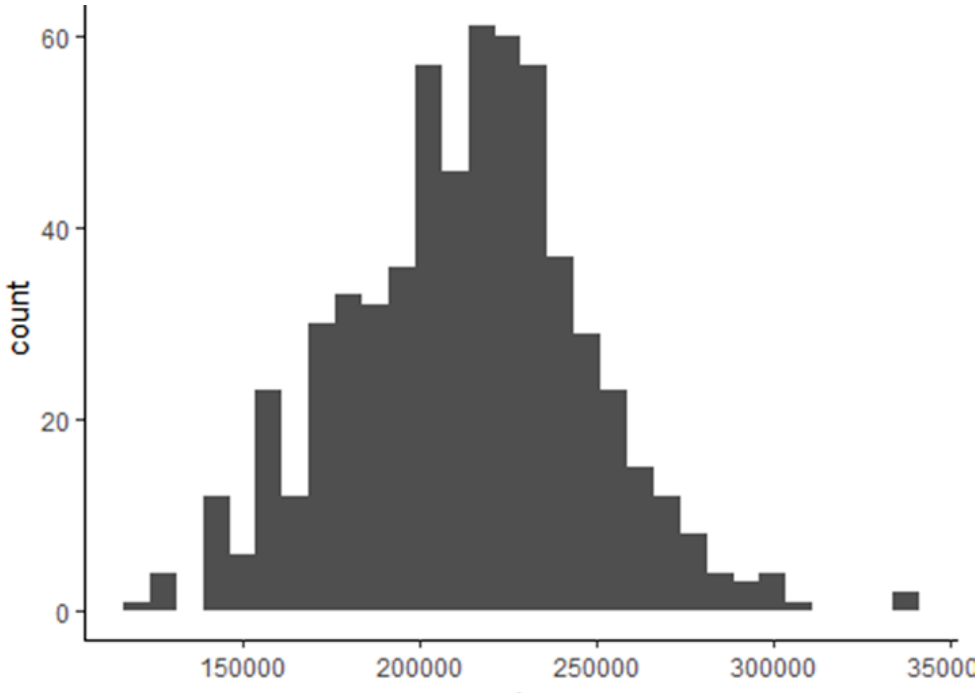
First, all models were run with all available predictors. The models were then run with 10, 11, and 12 of the original predictors. The reduction in predictors was due to both degenerate variables and high correlation between variables. 10-fold cross validation was used for model validation.

Included
R model and results

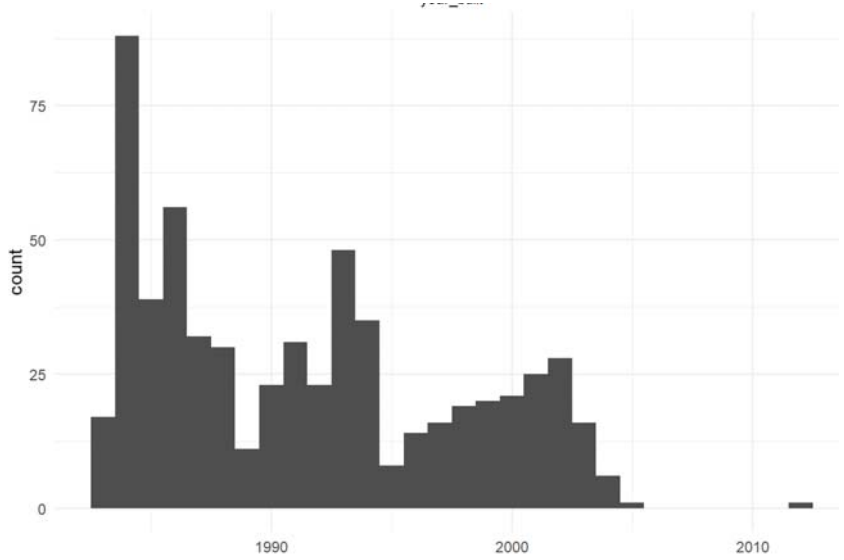
Predictors

 # Define variables
 # Source: BCAD, Bexar County Appraisal District
 # http://www.bcad.org/mapSearch/

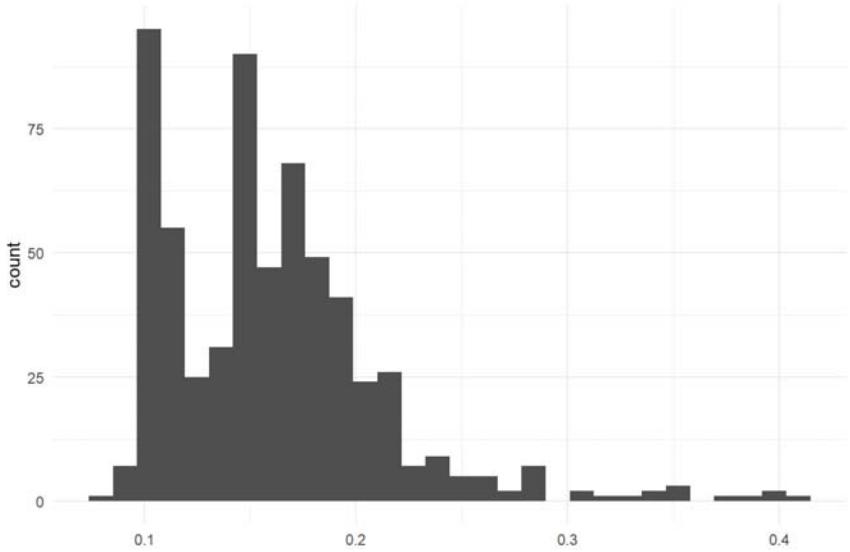
```
# AG = attached garage
# CP = attached carport
# DCK = attached wood deck
# DCK2 = attached wood deck second level
# DCKC = deck with cover
# ENC = enclosure
# GAR = detached garage
# LA = living area
# LA1 = additional living area
# LA2 = living area second level
# OP = attached open porch
# OP2 = attached open porch second level
# OPP = detached open porch
# PA = terrace(patio slab)
# PAC = patio covered, terrace with cover
# PTO = detached patio
# RMS = residential misc shed
# RSH = shed
# RSW = swimming pool
# SPA = spa/hot tub
# UTL = attached utility
# UTL2 = second story utility
# WD = attached wood deck
# WDD = detached wood deck
# eff_front = effective frontage (avg frontage and rear lot line)
# eff_depth = effective depth (avg the sidelines)
# year_built, acres, value, property_id
```



Home value distribution



Home year built distribution

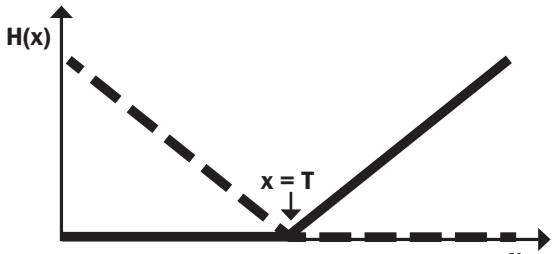


Home acres distribution

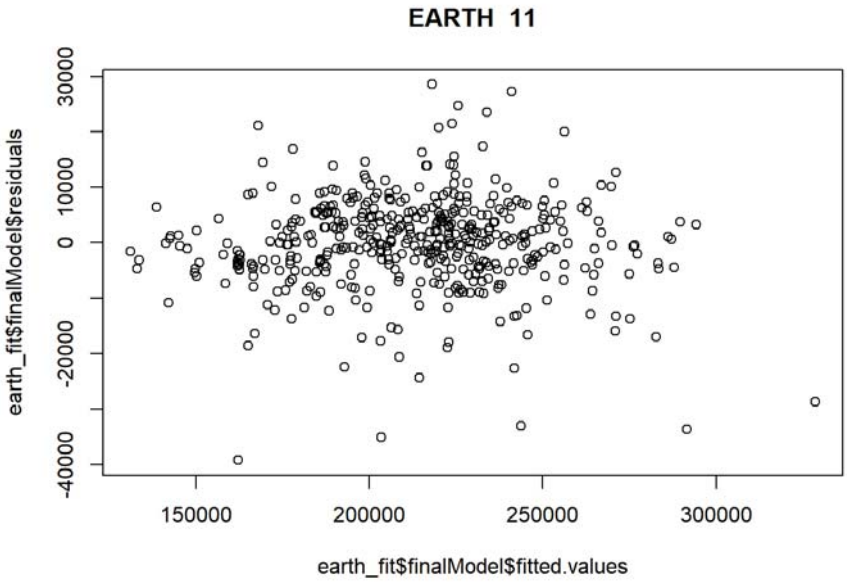
	model	num_pred	mse_train	rmse_train	mae_train	r2_train	mse_test	rmse_test	mae_test	r2_test
1	EARTH	29	71501592.00	8455.86	5897.81	0.94	62408710.00	7899.92	5755.12	0.94
2	EARTH	28	72728599.00	8528.11	6028.48	0.93	63745720.00	7984.09	5690.44	0.94
3	GLM	29	65745010.00	8108.33	6184.20	0.94	70575694.00	8400.93	6422.92	0.93
4	GAM	29	65745010.00	8108.33	6184.20	0.94	70575694.00	8400.93	6422.92	0.93
5	GLM	28	66607672.00	8161.35	6280.35	0.94	71071786.00	8430.41	6563.11	0.93
6	GAM	28	66607672.00	8161.35	6280.35	0.94	71071786.00	8430.41	6563.11	0.93
7	EARTH	11	77884966.00	8825.25	6280.82	0.93	71778914.00	8472.24	6098.31	0.93
8	LM	29	79772584.00	8931.55	6872.06	0.93	72580306.00	8519.41	6572.86	0.93
9	LM	28	79450724.00	8913.51	6937.24	0.93	73154189.00	8553.02	6712.26	0.93
10	EARTH	12	77294453.00	8791.73	6383.04	0.93	77481790.00	8802.37	6050.68	0.92
11	EARTH	10	78941148.00	8884.88	6484.52	0.93	79358279.00	8908.33	6327.88	0.92
12	GLM	12	91789457.00	9580.68	7336.08	0.91	83979337.00	9164.02	7080.65	0.92
13	GAM	12	91789457.00	9580.68	7336.08	0.91	83979337.00	9164.02	7080.65	0.92
14	RF	28	102537915.00	10126.10	6981.23	0.91	102103007.00	10104.60	6779.88	0.90
15	LM	12	101122839.00	10055.99	7613.06	0.91	103000838.00	10148.93	7538.70	0.90
16	RF	10	101471499.00	10073.31	7037.62	0.91	104251066.00	10210.34	6770.37	0.90
17	SVM	12	65010378.00	8062.90	5418.44	0.94	106428222.00	10316.41	7181.38	0.90
18	RF	12	102579114.00	10128.14	7058.47	0.91	107179864.00	10352.77	6916.07	0.90
19	RF	29	105469988.00	10269.86	7134.05	0.91	107184804.00	10353.01	7220.27	0.90
20	GLM	11	104083724.00	10202.14	7901.12	0.90	108423545.00	10412.66	8022.53	0.89
21	GAM	11	104083724.00	10202.14	7901.12	0.90	108423545.00	10412.66	8022.53	0.89
22	GLM	10	104968227.00	10245.40	7944.39	0.90	108972226.00	10438.98	8122.01	0.89
23	GAM	10	104968227.00	10245.40	7944.39	0.90	108972226.00	10438.98	8122.01	0.89
24	RF	11	104791101.00	10236.75	7187.70	0.91	109569002.00	10467.52	7330.40	0.89
25	SVM	10	74160528.00	8611.65	5883.65	0.93	115039261.00	10725.64	7315.36	0.89
26	SVM	11	70851282.00	8417.32	5638.93	0.93	118244638.00	10874.04	7264.27	0.88
27	LM	11	113750319.00	10665.38	8141.93	0.90	126284926.00	11237.66	8437.16	0.88
28	LM	10	114247508.00	10688.66	8149.79	0.90	126680677.00	11255.25	8536.54	0.88
29	SVM	28	1085130600.00	32941.32	26123.82	-0.01	1028092896.00	32063.89	25408.76	0.00
30	SVM	29	1085197111.00	32942.33	26125.02	-0.01	1028120039.00	32064.31	25409.23	0.00

Basic Hinge Function used in Mars/Earth models

Interactions between predictors are modeled with products of hinge functions.



h(x) = max(0, x - t)
or max(0, t - x)



Results

Ultimately, the earth model produced the best predictions with 11 predictor variables and test RMSE of $8,472.24, approximately 4% of the median home value of the 605 observations. The actual error, based on the actual home values provided after model completion, was off by an average of $5,334.20.

RMSE was used as a performance metric because there were unexpected values, potential outliers, that needed to be considered. An additional benefit of using RMSE was maintaining the units of the original data, in this case, dollars.