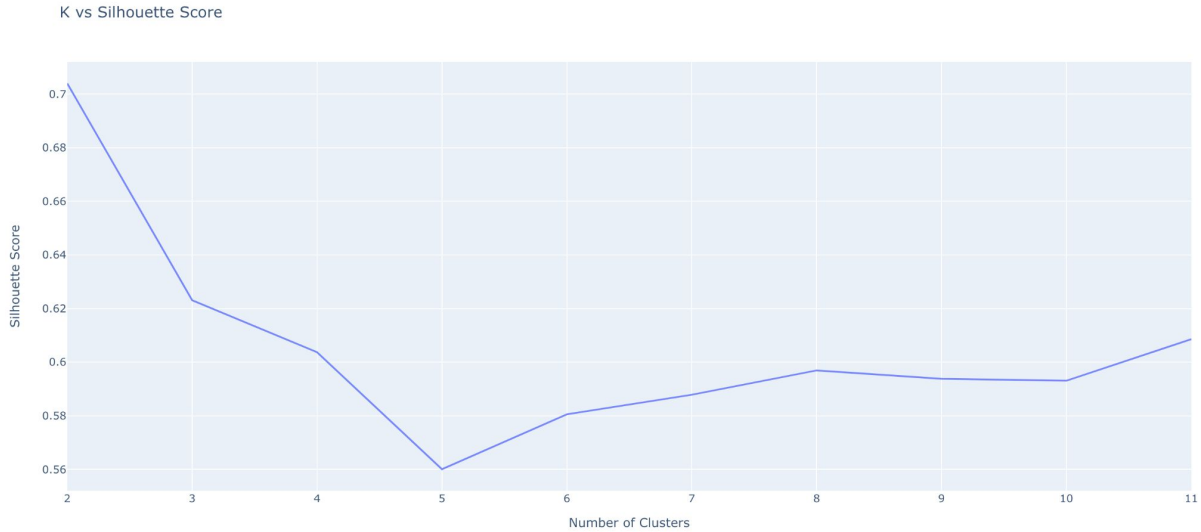
DSO 510: Business Analytics

Elmer, Issac, Wayne

Topic:
Project 2 Analysis

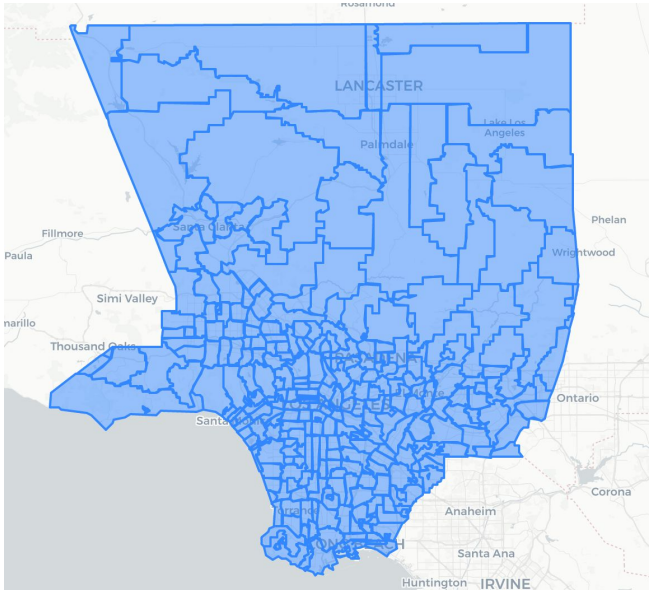
1. Executive Summary (WIP)
2. Goal & Variable Definition
3. Data Organization & Visualization
4. Analysis of Dependent Variable
5. Interpretation, Action, Adoption, Automation (WIP)

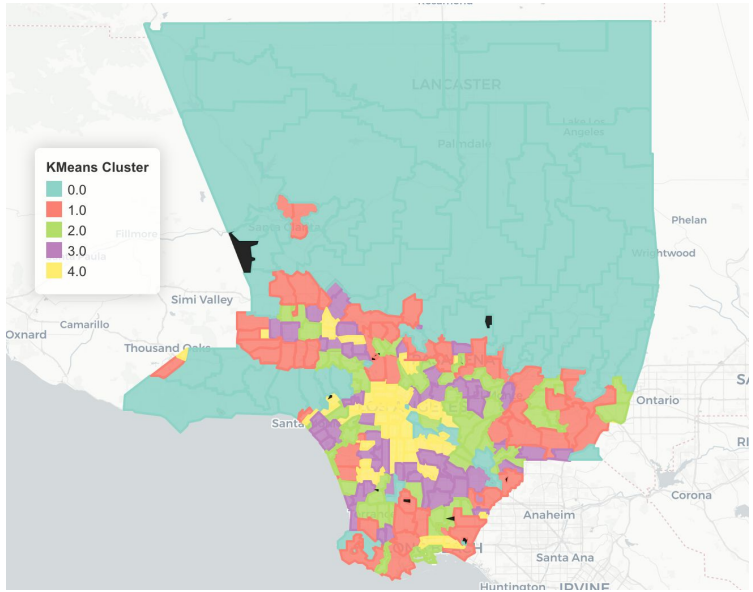
Kmeans Clustering On Zipcode



Based on guidance from previous class, We grouped 311 zipcodes using a Kmeans model (k=5) based off median age, median income, and population density from the 2022 US Census (most recently available).

Kmeans Clustering On Zipcode

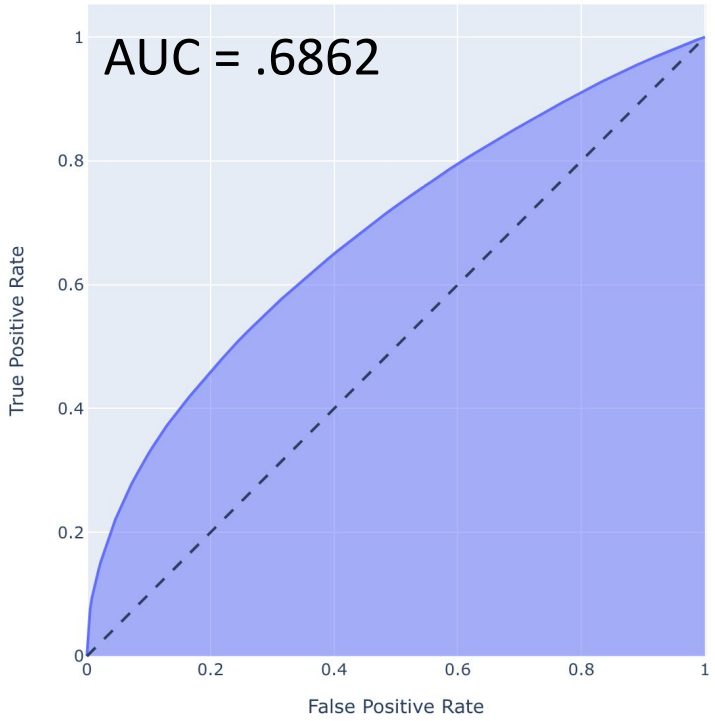




Logistic Regression

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.2395	0.3222	14.7949	0.0001
kmeans_cluster	0	1	0.1023	0.2008	0.2593	0.6106
kmeans_cluster	1	1	-0.5709	0.2010	8.0683	0.0045
kmeans_cluster	2	1	-0.1264	0.2008	0.3963	0.5290
kmeans_cluster	3	1	-0.2868	0.2078	1.9051	0.1675
kmeans_cluster	4	1	-0.1998	0.2008	0.9901	0.3197
kmeans_cluster	5	0	0	.	.	.
sq_feet		1	0.00105	5.996E-6	30406.3290	<.0001
bedrooms		1	-0.2310	0.00317	5301.6229	<.0001
bathrooms		1	0.2980	0.00446	4458.8651	<.0001
effective_year		1	-0.00137	0.000130	111.2542	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
kmeans_cluster 0 vs 5	1.108	0.747	1.642
kmeans_cluster 1 vs 5	0.565	0.381	0.838
kmeans_cluster 2 vs 5	0.881	0.595	1.306
kmeans_cluster 3 vs 5	0.751	0.500	1.128
kmeans_cluster 4 vs 5	0.819	0.552	1.214
sq_feet	1.001	1.001	1.001
bedrooms	0.794	0.789	0.799
bathrooms	1.347	1.335	1.359
effective_year	0.999	0.998	0.999

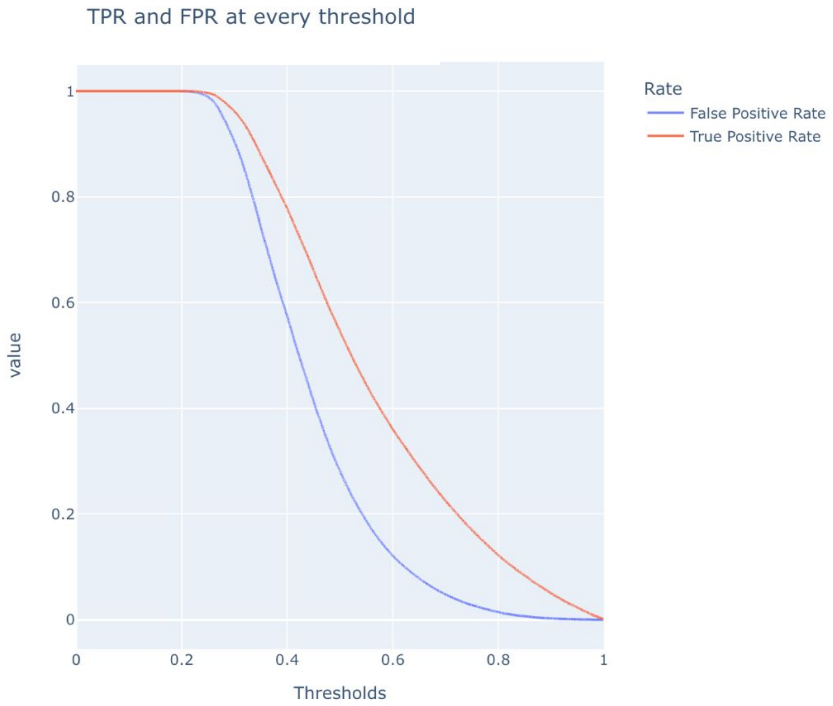


Logistic Regression

Model Information	
Data Set	MYLIB.PARCEL_FINAL
Response Variable	is_greater_than_lac_median
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1123121
Number of Observations Used	1123121

Response Profile		
Ordered Value	**is_greater_than_lac_median**	**Total Frequency**
1	False	561561
2	True	561560

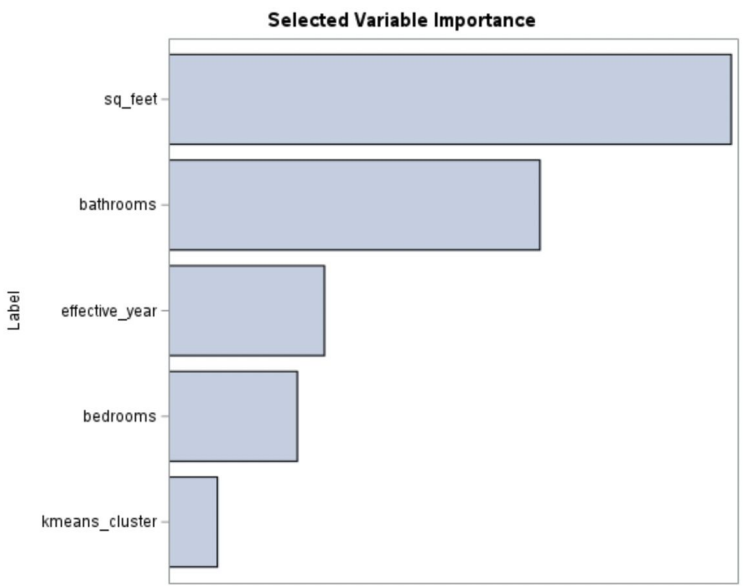


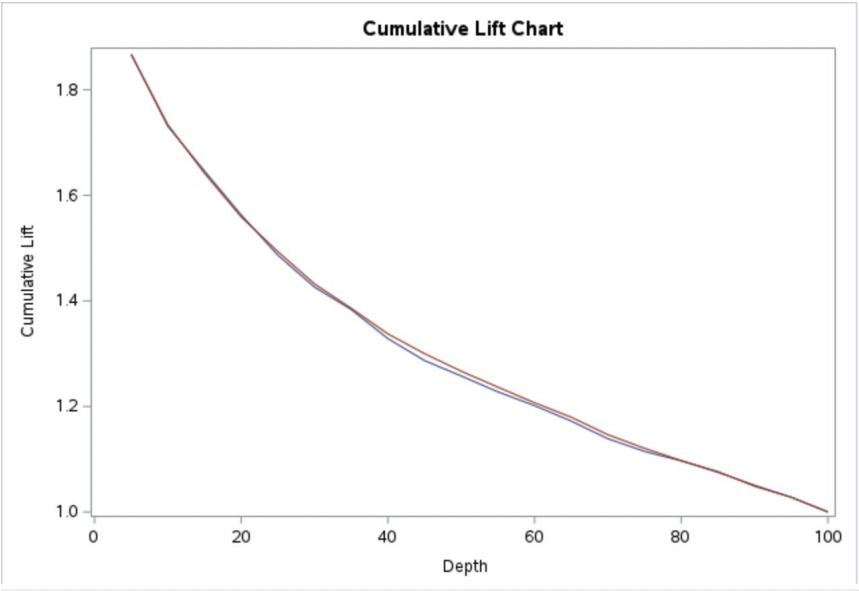
Rapid Predictive Modeler

Statistic	Train	Validation
Akaike's Information Criterion	60450.7535	.
Average Squared Error	0.2209	0.2204
Average Error Function	0.6293	0.6284
Maximum Absolute Error	0.9913	0.9794
Mean Squared Error	0.2210	0.2204
Sum of Frequencies	47999.0000	32001.0000
Root Average Square Error	0.4700	0.4695
Root Mean Squared Error	0.4701	0.4695
Schwarz's Bayesian Criterion	60643.8901	.
Sum of Square Errors	21208.8847	14109.2010
Misclassification Rate	0.3680	0.3641
Roc Index	0.6860	0.6900
Gini Coefficient	0.3720	0.3790
Kolmogorov-Smirnov Statistic	0.2690	0.2720
Kolmogorov-Smirnov Probability Cutoff	0.5220	.
Lift at 10%	1.5940	1.6003
Cumulative Lift at 10%	1.7309	1.7337
Captured Response at 10%	7.9704	8.0015
Cumulative % Captured Response at 10%	17.3089	17.3415

		Scorecard Points
bathrooms	1: LOW - 1.5	0.00
	2: 1.5 - 2.5	86.00
	3: 2.5 - HIGH	209.00
bedrooms	1: LOW - 1.5	187.00
	2: 1.5 - 2.5	199.00
	3: 2.5 - 3.5	137.00
	4: 3.5 - 4.5	65.00
	5: 4.5 - HIGH	0.00
effective_year	1: LOW - 1926.5	71.00
	2: 1926.5 - 1955.5	68.00
	3: 1955.5 - 1957.5	66.00
	4: 1957.5 - 1962.5	65.00
	5: 1962.5 - 1987.5	31.00
	6: 1987.5 - 1994.5	0.00
	7: 1994.5 - HIGH	99.00
kmeans_cluster	0	197.00
	1	0.00
	2	127.00
	3	96.00
	4	116.00
	5	133.00
sq_feet	1: LOW - 1555.5	0.00
	2: 1555.5 - 1935.5	159.00
	3: 1935.5 - HIGH	296.00

Rapid Predictive Modeler





Analytical Interpretation

- RPM and Logistic Regression performed similarly
- All variables besides Kmeans clusters were significant
  - Though mixed models may be explored since previous linear regression results showed significance in geography variables
- SqFt and Bathrooms increase odds of a home being over the LA County median home value
- Scorecards were not ideal, but there are indications that more bathrooms and more sq footage have more consistent results
  - Interesting finding that in more bedrooms does not mean more consistent positive cases

Decision Points

- Picking a threshold
  - Dependent on costs of assigning resources, a higher threshold can provide less false positives and save money

- Deciding on if a mixed model would be useful to look into if we are better able to summarize information about the geography a house lands in

- Should we continue with a global or localized model

Q&A (QUESTIONS & ANSWERS)

Questions?

Thank you!