Anastasia Simpson,
Data Analyst Candidate.
Take-Home Submission Report.
May 8, 2019

Analysis of Poor Performance of the Model in Henderson County.

Problem: *Property valuation model is performing significantly lower* ***0.49 Abs Error*** *in Henderson County.*

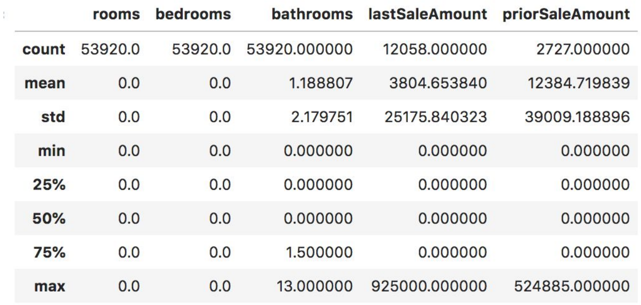
Based on high level overview of the data sets provided I conclude that Henderson County Model is an underfit model that has low variance and high bias due to the sparsity of the training data in addition to the presence of extreme outliers and multicollinearity of independent variables.
Below I list data issues I discovered during my short time analysis along with the recommendations for how to handle these issues.

1. Data Accuracy

1.1 Problem: Incorrect data.

Zeros are entered instead of NaN for the following features:

- *rooms,*
- *bedrooms,*
- *bathrooms ,*
- *lastSaleAmount,*
- *priorSaleAmount*



	rooms	bedrooms	bathrooms	lastSaleAmount	priorSaleAmount
count	53920.0	53920.0	53920.000000	12058.000000	2727.000000
mean	0.0	0.0	1.188807	3804.653840	12384.719839
std	0.0	0.0	2.179751	25175.840323	39009.188896
min	0.0	0.0	0.000000	0.000000	0.000000
25%	0.0	0.0	0.000000	0.000000	0.000000
50%	0.0	0.0	0.000000	0.000000	0.000000
75%	0.0	0.0	1.500000	0.000000	0.000000
max	0.0	0.0	13.000000	925000.000000	524885.000000

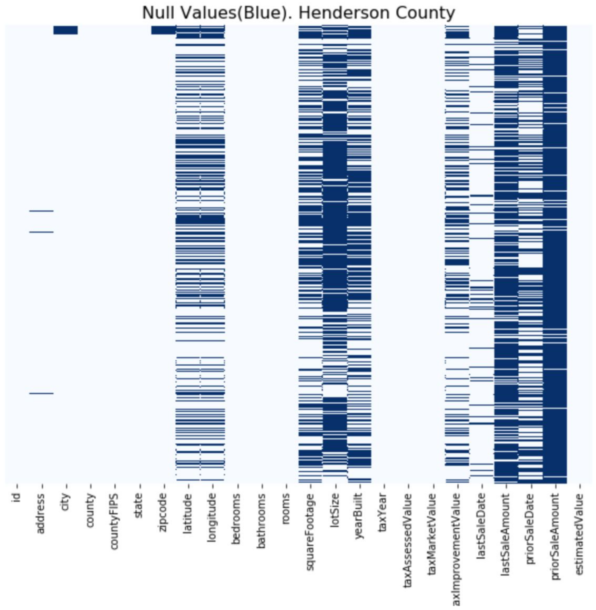
1.2. Recommendation:

- Replace zero values with NaN. Deploy the model w/ adjustment. Remove features with incorrect values.
- Collect data on rooms, bathrooms, bedrooms as they should be a solid predictor. Normally a property with more rooms/bathrooms/bedrooms has stronger capability to increase the price

2. Data Completeness

2.1. Problem: Sparse data in general.
Only **13.74%** of the all data was used training the model for Henderson County due to many null values.
Missing Values:

- taxImprovementValue is: 35.83%
- Square Footage is: 47.64%
- Lot Size is: 72.92%
- Year Built is: 56.04%

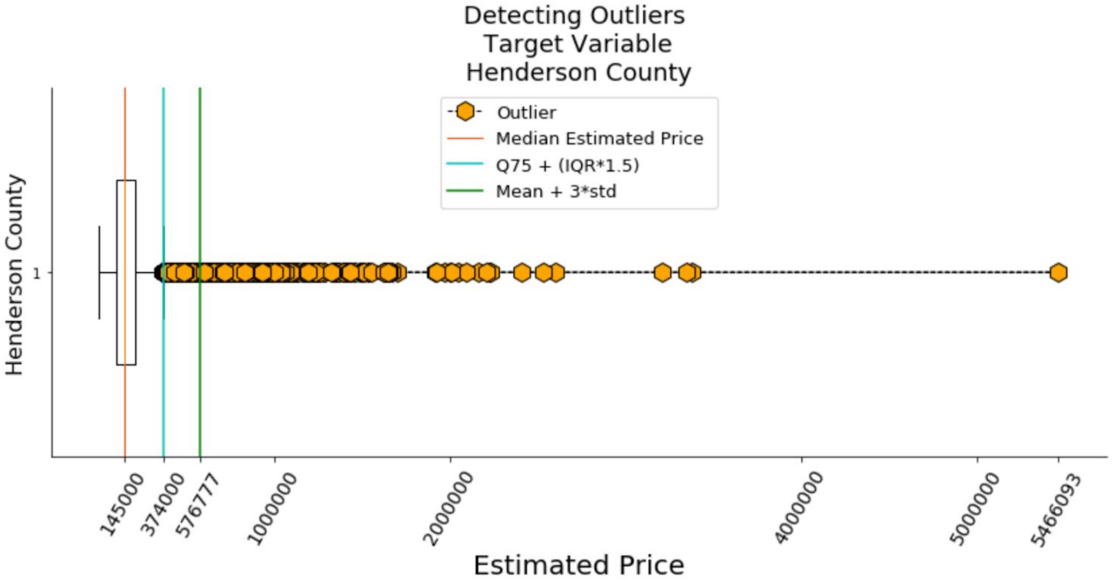


2.2 Recommendation:

- Collect more data for better predictive power of the model.
- Add additional features (see below)

3. Outliers

3.1. Problem: Observed extreme outliers in Henderson county.



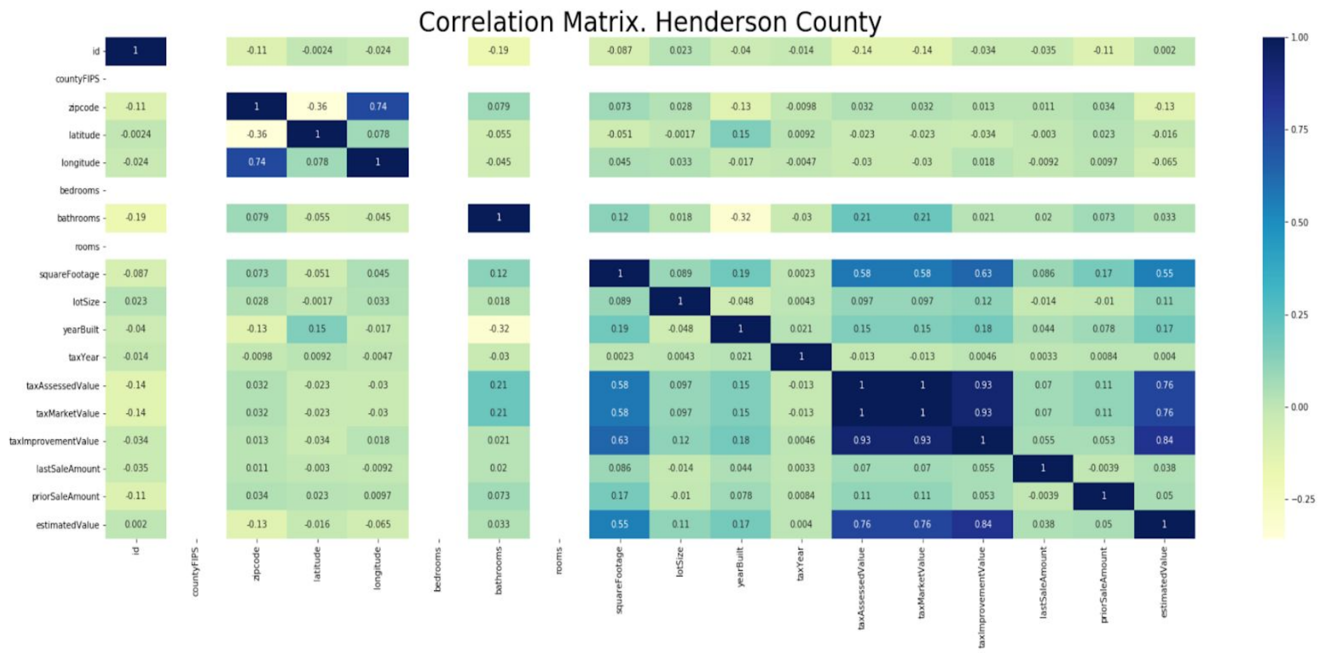
3.2. Recommendation:

- Treat outliers (remove, log transform, cap data, assign new values)
- Try models that are robust to outliers (i.e. regression tree model))
- Remove the most Extreme Outliers (i.e. Huge $ 5.466.093 outlier for Henderson County) and rerun the model to see if the score improves.
- Deploy an additional model for outlier houses (identify if the houses are luxury/ not as an additional feature)

4. Multicollinearity of Predictors

4.1. Problem: Observed multicollinearity among the following features: 'taxAssessedValue','taxMarketValue', 'taxImprovementValue' (>0.93 - 1 between each other), as well as the squareFootage (0.58 - 0.63).

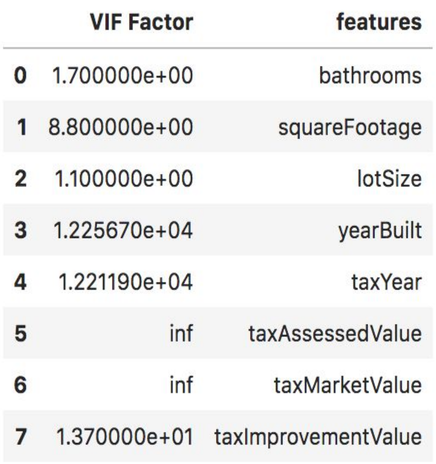
- Multicollinearity can yield to predictions that are wildly varying and possibly numerically unstable.



- **Variance Inflation Factor**
  8.80 for squareFootage
  Infinity for taxAssessedValue and taxMarketValue

4.2. Recommendation:

- Perform feature engineering for the tax related features due to multicollinearity. For example, we can use the AVG between these 3 features aggregating all three features into one or perhaps select the feature that we have the most data for.



	VIF Factor	features
0	1.700000e+00	bathrooms
1	8.800000e+00	squareFootage
2	1.100000e+00	lotSize
3	1.225670e+04	yearBuilt
4	1.221190e+04	taxYear
5	inf	taxAssessedValue
6	inf	taxMarketValue
7	1.370000e+01	taxImprovementValue

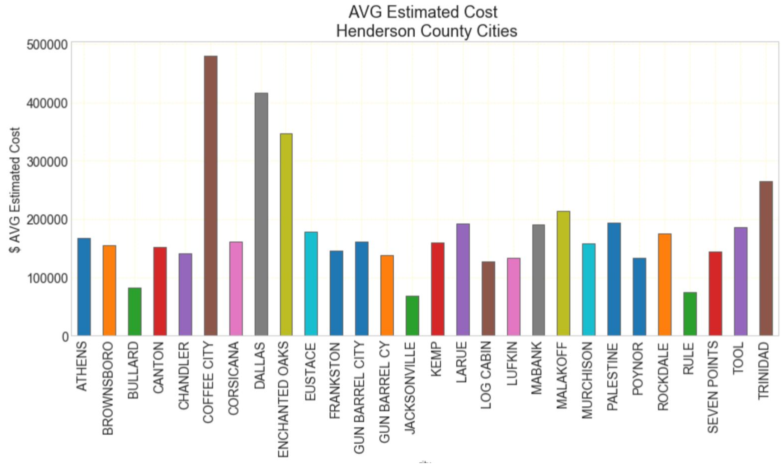
5. What additional data would you like to collect?

Some useful features may include:

- Luxury (binary yes/no), additional features that are suitable to predict luxury houses.
- Location-specific outcomes: number of permits issued to build swimming pools (restaurants, coffee shops), change in number of coffee shops within a one-mile;
- Building energy consumption relative to other structures in the same zip code
- renewal energy service
- Number of Yelp reviews for nearby businesses

6. What questions do you still have that you weren't able to answer?

- Look into more details on why other county perform better on the sparse data and determine if there are any other underlying factors.
- Is it possible that different models were deployed for different counties so that outliers are treated differently for other counties/or they were removed.
- What was the outlier treatment for the model?
- How was the problem of multicollinearity handled in building the models?
- Was there a practice to enter zero values for nan during data entry?
- Look into which cities the outliers are most common.


AVG Estimated Cost
Henderson County Cities
$ AVG Estimated Cost
ATHENS
BROWNSBORO
BULLARD
CANTON
CHANDLER
COFFEE CITY
CORSICANA
DALLAS
ENCHANTED OAKS
EUSTACE
FRANKSTON
GUN BARREL CITY
GUN BARREL CY
JACKSONVILLE
KEMP
LARUE
LOG CABIN
LUFKIN
MABANK
MALAKOFF
MURCHISON
PALESTINE
POYNOR
ROCKDALE
RULE
SEVEN POINTS
TOOL
TRINIDAD