Predicting home prices for Ames, Iowa

-Gouri.K

GOAL - Way to predict a house price ~ actuals

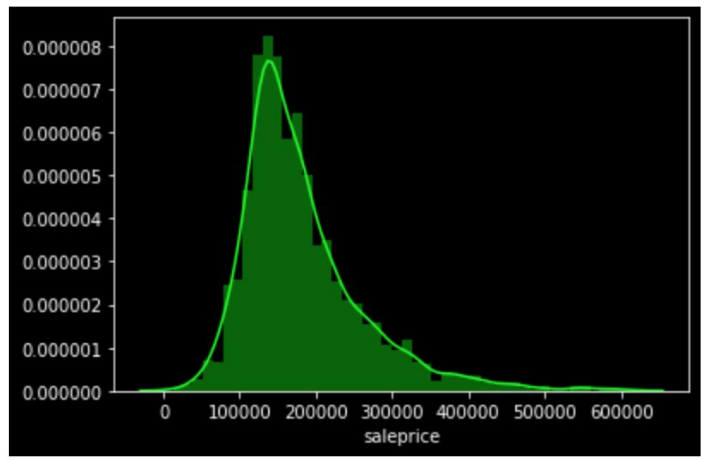
- Choose attributes to formulate **algorithms** using past data
- To **test** the theory , use the algorithm found, to predict the price of known houses and check how it fared by **measuring errors**.
- **Tweak** the algorithm to reduce the errors as much as possible
- Keep testing to the extent possible to find the best fit model (**repeat**)
- **Actual predictions** using the model(s) within a threshold of **+/- $20,000**
- Publish and present **the findings**

How to begin?

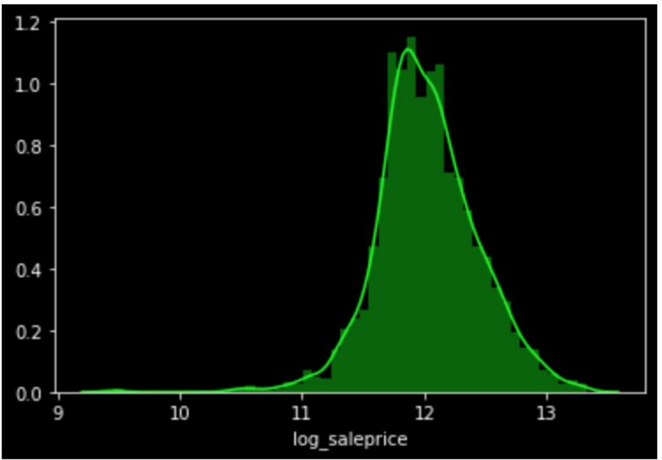
- What data do we have?
    - Metadata of 2051 observations with 81 features each .
    - Using that we have to predict Sale Price for 878 houses.
- How much data is missing?
    - Categorize each data into nominal, discrete, continuous and categorical/ordinal data
    - Lot of the missing data easily corrected. NA/None uploaded as missing
    - 330 empty lot frontage (huge actual missing chunk)
- Are there any invalid/incorrect data?
    - Garage year built was 2200 in row (1699). Changed it to 2000 .
- What steps to use to prepare the data to start exploring and predicting? - Correlation, Linear regression

Predicting Sale Price

SALE PRICE DISTRIBUTION

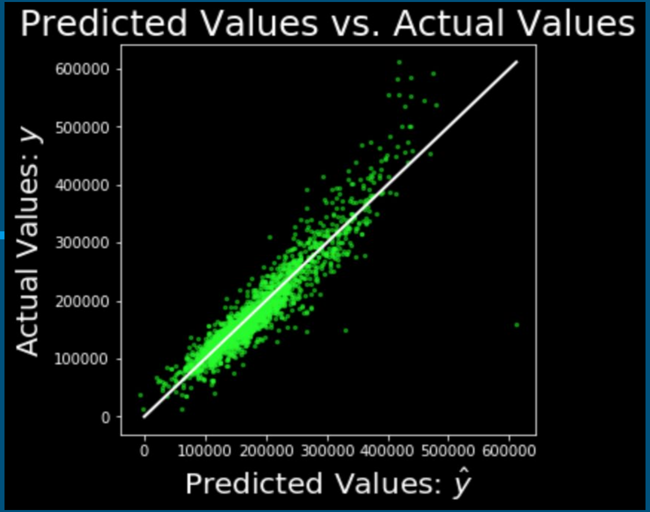


LOG SALE PRICE DISTRIBUTION



For best linear regression model predictions , it helps if the the target value distribution is ~ normal

SALE PRICE

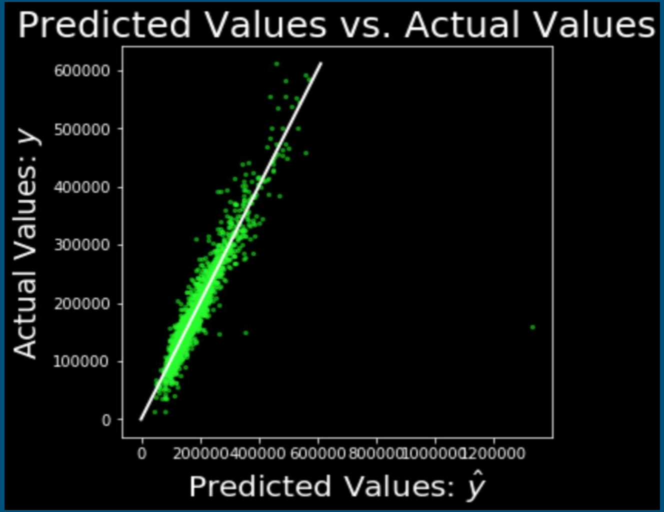


RMSE : 29390
Residual mean : 18705

Overfit Model

KAGGLE SCORE : 33747.88408

LOG SALE PRICE

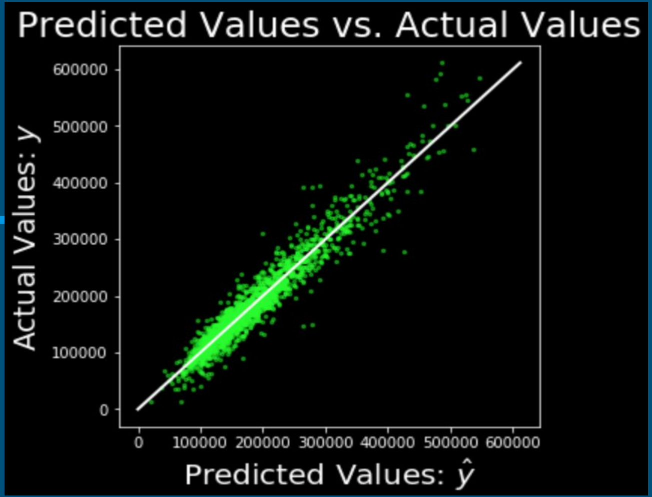


RMSE : 22611
Residual mean : 16256

Balanced model. Leaning towards Bias.

KAGGLE SCORE : 24183.55336

POLYNOMIAL FEATURES



KAGGLE SCORE : 23686.62794

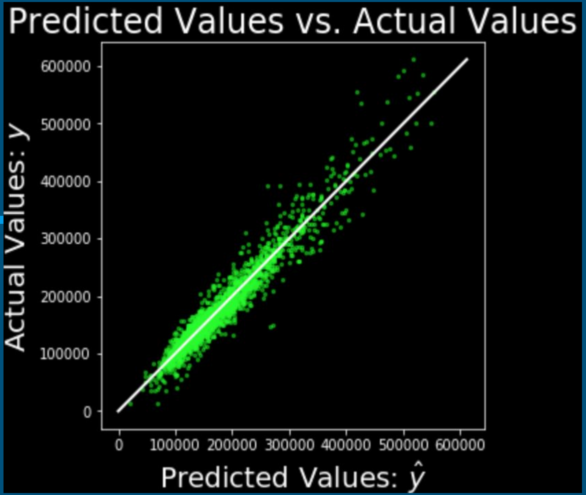
No of main features selected: 29
Created Poly features : 71
Total features used : 100

RMSE : 21582
Residual mean : 15138

FEATURES SELECTED USING CORRELATION AND P VALUE > 0.05

Leaning towards Bias. Overall balanced model

POLYNOMIAL + INTERACTION



KAGGLE SCORE : 21072.07275

No of main features selected: 29
Created Poly features : 50
Additional Interaction : 8
Total features used : 87

RMSE : 20630
Residual mean : 14455

Leaning towards Bias. Overall balanced model

Removed outliers. Dropped few columns from lot frontage and lot area

Business recommendation

- Overall quality , living area , age of the house, lot area, have a wood deck , open porch , ratio of bedroom to bathroom , garage area - add value

- Pool related data did not impact the housing prices.

- Overall quality of the house, remodelling helped house prices.

- Stone Brook, North Ames, North Ridge seem to be good investments

Conclusion

1. The goal was to present a **way to approach** a given data set to predict sale prices within a **threshold of +/- $ 20,000** . Came very close.
2. The algorithm itself cannot be re-used but the **idea** behind it can be
3. Given resources , we can try n number of combinations to come up with the best results with least errors but it **should work for the population** and not specific samples.

Additional data that can help

1. Taxation
2. Policies that can affect buying decisions
3. No of buyers looking for buying a house (Supply v/s demand)
4. Past data for the house (all the past transactions)

Further analysis for getting a better prediction

Time and resource permitting:

a. Check for more field interactions
b. Predict next year data with the models to see how they fare in the real world
c. To use other predictive models other than linear regression