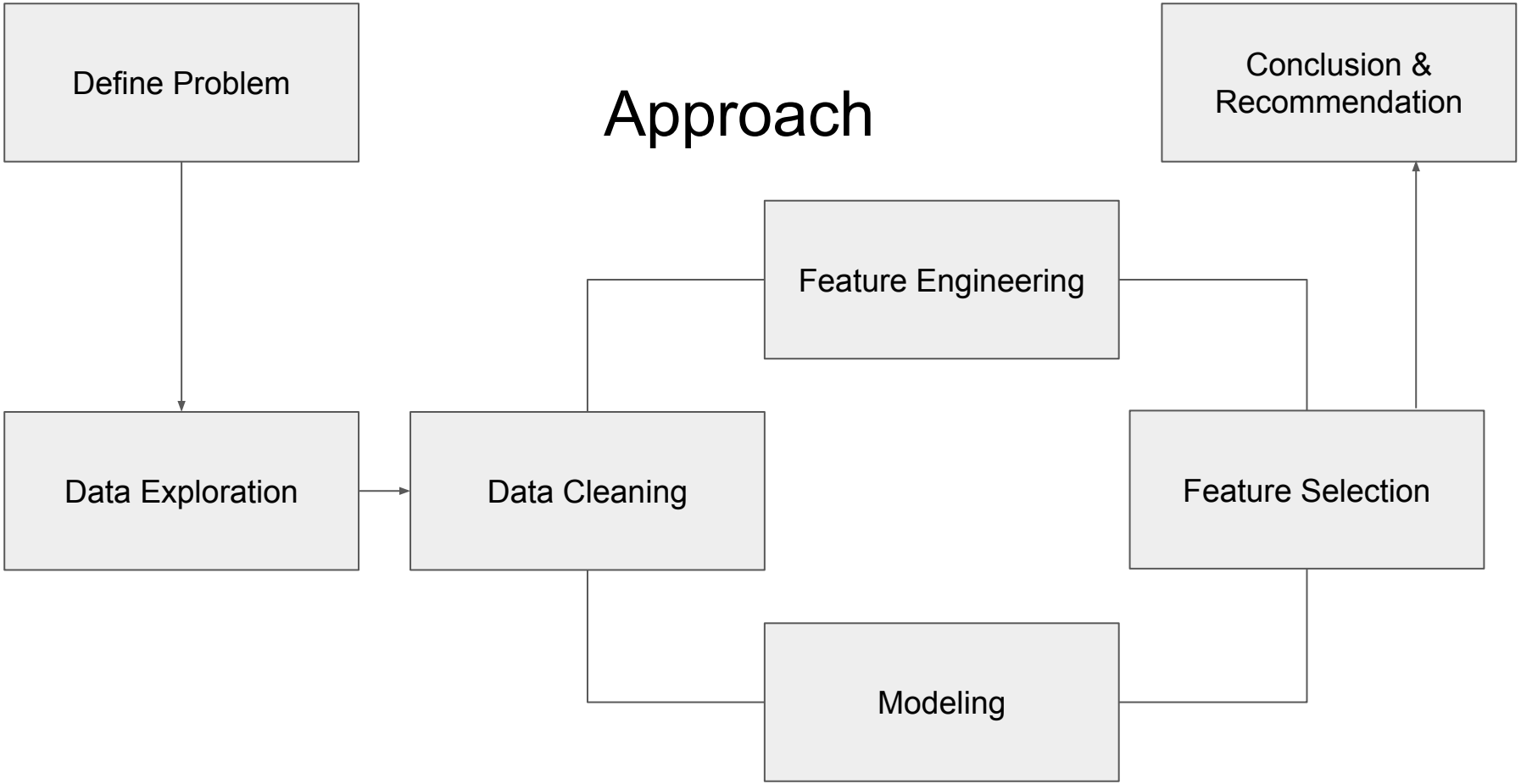
Prediction of Sale Price of Housing in Ames, Iowa

Nicholas Lim

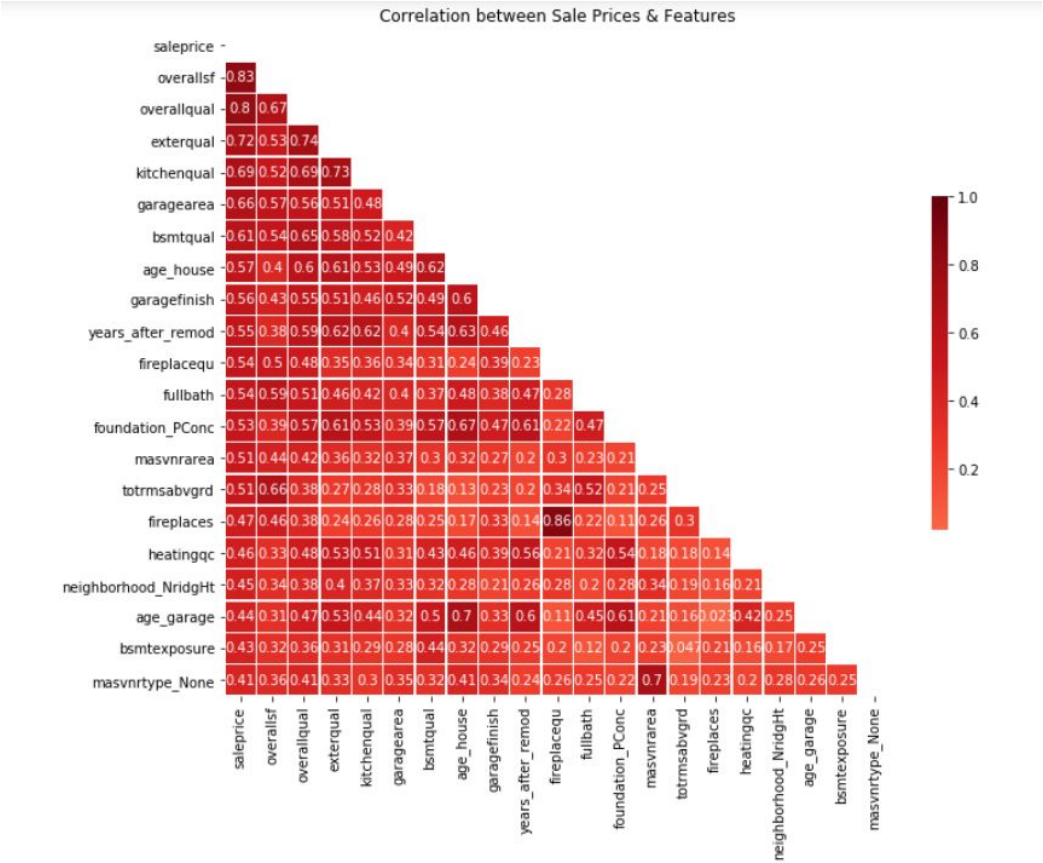
About the Ames Housing Dataset

- 2051 rows of test data in training set, sale price provided
- 879 rows of test data in test set
- 80 different features

Approach


Define Problem
Conclusion & Recommendation
Feature Engineering
Data Exploration
Data Cleaning
Feature Selection
Modeling

Model 1: Choosing the top 20 correlated features



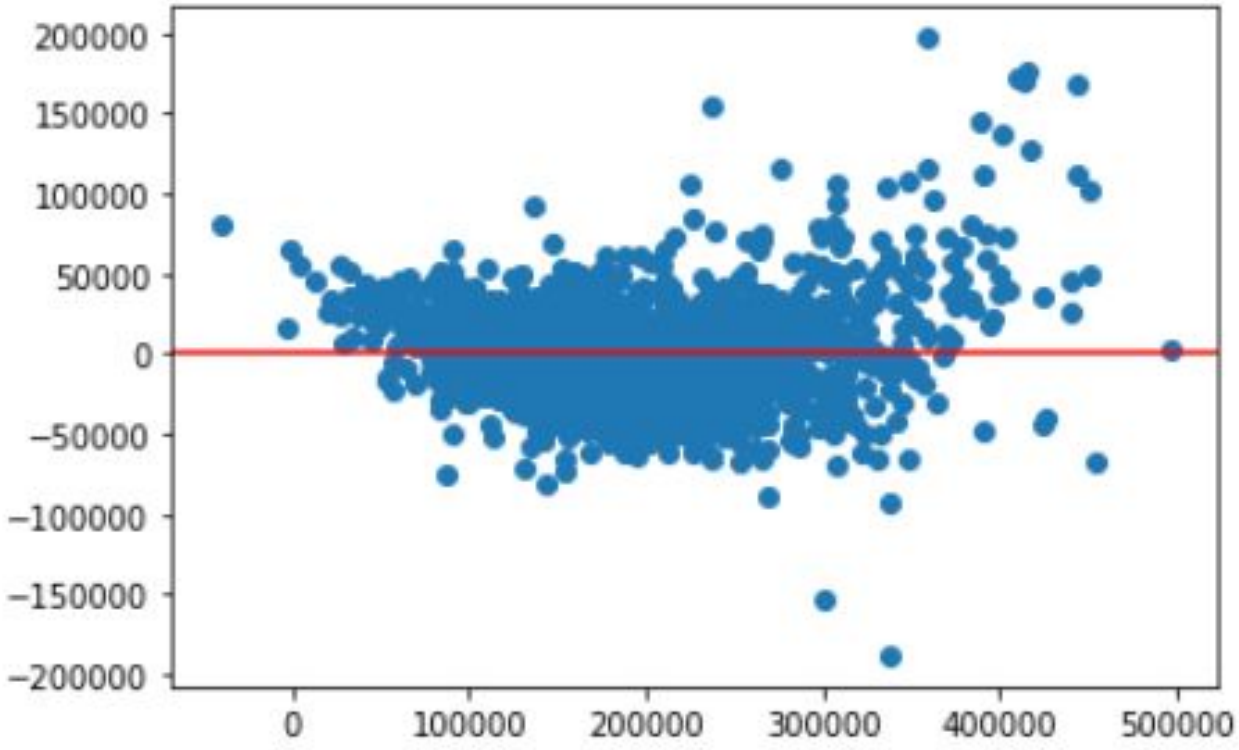
Observations:

- Correlation shown are based on absolute values
- Only 14 of the 20 features shown have correlation magnitude > 0.5

Model application:

- A cross-train split is done on the training set, before we measure the score of the train set
- Following regression techniques applied and the scores measured are:
  - Linear Regression: 0.8671427271526209
  - RIdge: 0.8672831470932177
  - Lasso: 0.8672310115497208
  - Elastic Net: 0.8602845817644343

A look at the residual plot...



Applying the test data on the Ridge model to get the predicted sale price and upload to Kaggle, we got the following RMSE:

- Private Score: 33345.50812
- Public Score: 34100.80685
- Aggregated Score: 33572.09774

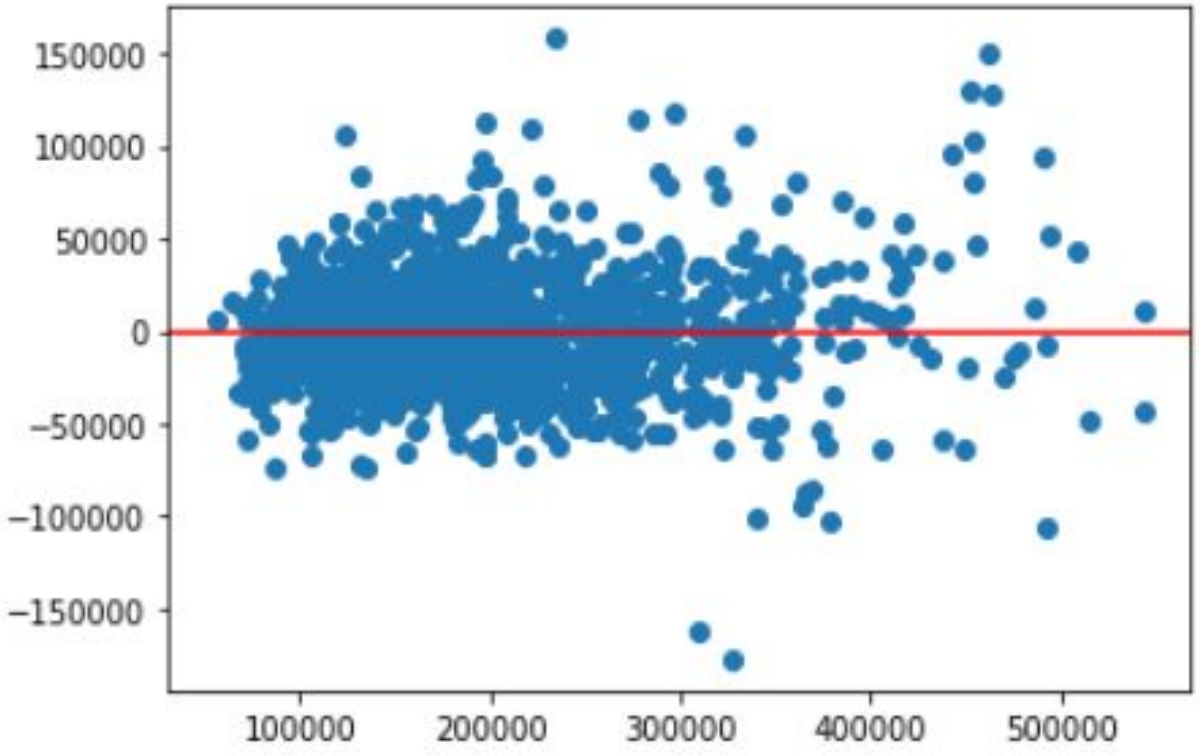
Model 2: Apply polynomial features to top 20 features

- In this model, we used the same 20 features but we apply polynomial features
- From here, once applied, we will take the top 20 interactions and fit into the models again to see if there’s an improvement in score

Model application:

- The scores are as follows for the various models:
  - Linear Regression: 0.8911682495617942
  - Lasso: 0.8916077898148467
  - Ridge: 0.8917525033060905
  - Elastic Net: 0.8774743099678481
- An improvement in the scores as compared to model 1, with ridge still leading as the best model among the rest

Looking at the residual plot again...



Overall, the residual points are closer to the zero mean, indicating a more accurate prediction

Applying the test data on the Ridge model to get the predicted sale price and upload to Kaggle, we got the following RMSE:

- Public Score: 27718.98259
- Private Score: 34613.57548
- Aggregated Score: 32545.19761

- Surprisingly, even though the public score improved, the private score seem to remain fairly constant.

Conclusion

- From the above 2 models, even though we observe a good improvement in score for the public score in model 2, the private score remains almost consistent across the 2 models, which sort of indicates a slight overfitting in the model itself. But model 2 still shows to be a better predictor as compared to model 1 since the RMSE for the private scores remain fairly consistent for the private scores but improves for the public ones.
- The way we clean our data may also impact our feature selection as well, so further experiments on the cleaning process can be done. For example, instead of replacing null values with default NAs or 0, we can try finding the mean/mode/median of the feature and replace those null values with it instead.