Project 2: Ames Housing Data & Kaggle Challenge

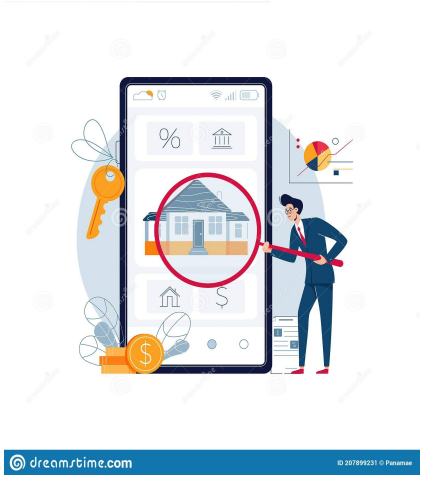
Malcolm Lau, DSI26

Scope

- Background & Business Problem
- Broad Methodology
- Observations from EDA
- Modelling Results
- Conclusion & Future Work

Background & Business Problem

Property owners who wish to put their property up for sale tend to want to obtain a preliminary valuation for consideration.



ROLE. As a Data Scientist for a Property-Tech Company in the USA, we are tasked to build the algorithm behind this web app to predict the property sale price and recommend the optimal number of features to provide a balance between accuracy of prediction and user experience (because nobody likes to fill in lengthy forms!).

CONSTRAINT. To ensure that User Experience is optimized, the user is not going to input more than 10 fields in our web app as it would increase the cost of engagement of the user and may result in lower utilisation of our web app.

Broad Methodology (1/2)

1. Address Missing Values in the Dataset

2. Select list of features to build regression model
    a. Remove outliers
    b. Remove features that do not provide enough information about the dataset as a whole (eg. Pool Area, almost none of the properties have a pool)
    c. Remove features that provide similar information (eg. MS Zoning vs Bldg Type & Housestyle)
    d. Create features that aggregate information (eg. has basement, has fireplace)

Broad Methodology (2/2)

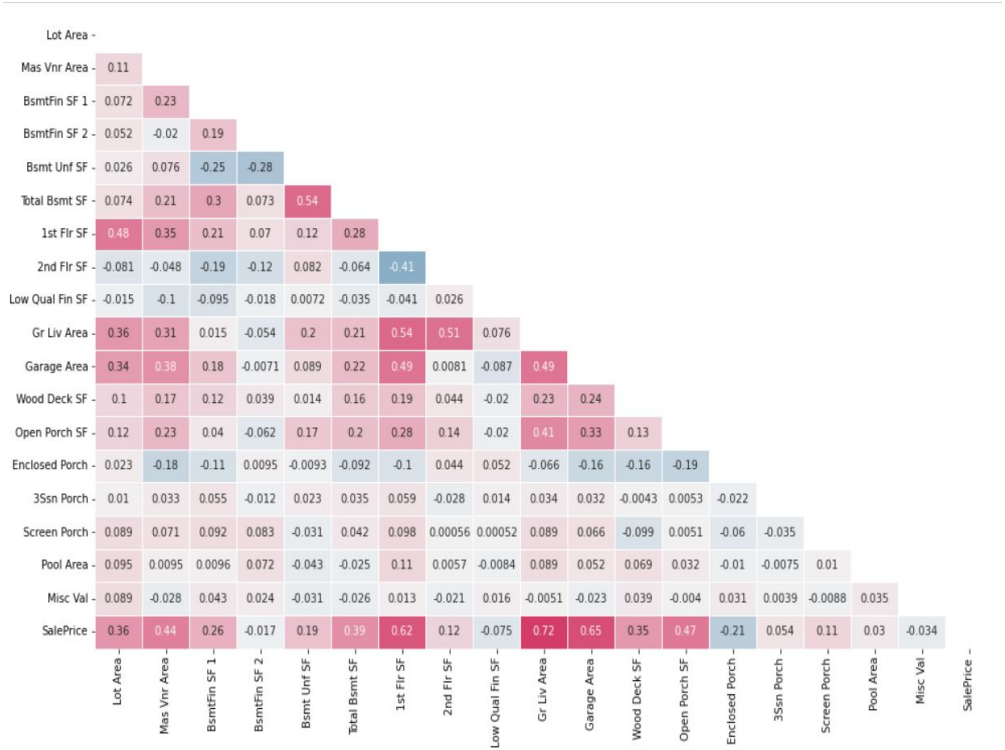
1. 1st Run of OLS, LASSO and Ridge to extract the top 5/10/15/20/25/30 features of each model (by ranking the coefficients)

2. Subsequent Run of OLS, LASSO and Ridge on the top 5/10/15/20/25/30 and select best performing model of each run for submission to Kaggle to score on the unseen data (scoring by RMSE)

3. Use the validation set scores during training as a baseline to evaluate and select the optimum number of features for the web app

Observations From EDA: Correlation w Target

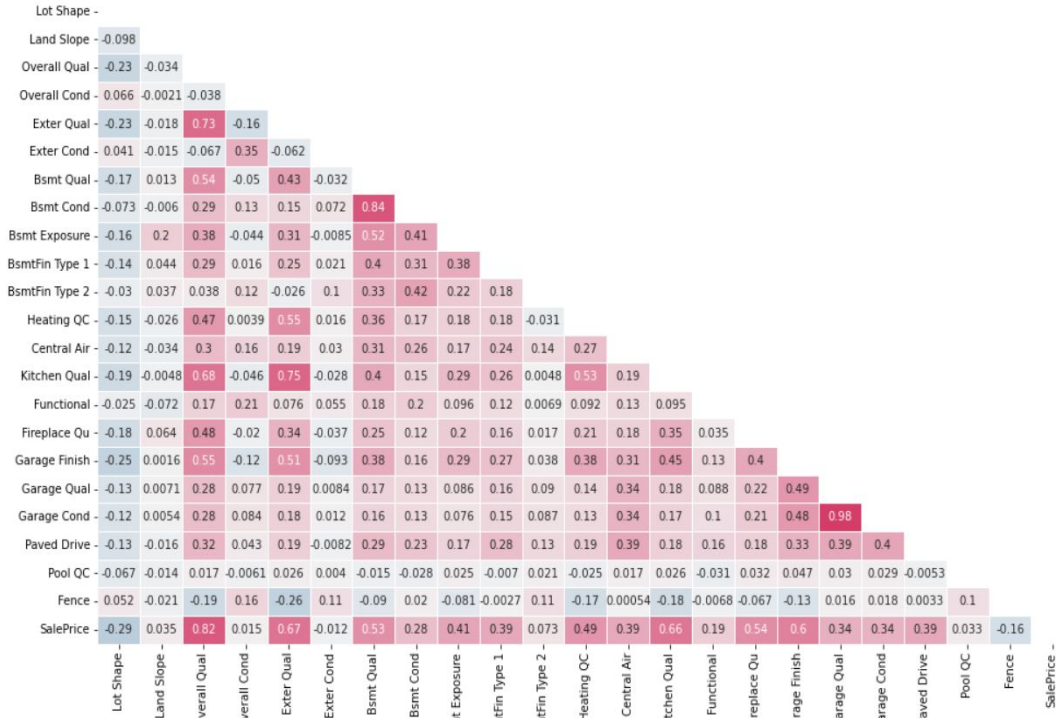
- Gr Liv Area and 1st Flr SF exhibits a moderate to strong correlation with SalePrice
  - Reasonable as the larger a property, the more expensive it tends to be
  - Both features are proxies of property size



Observations From EDA: Correlation w Target

- Overall Qual is strongly correlated to SalePrice
  - Reasonable as properties made with higher quality materials tend to be more expensive

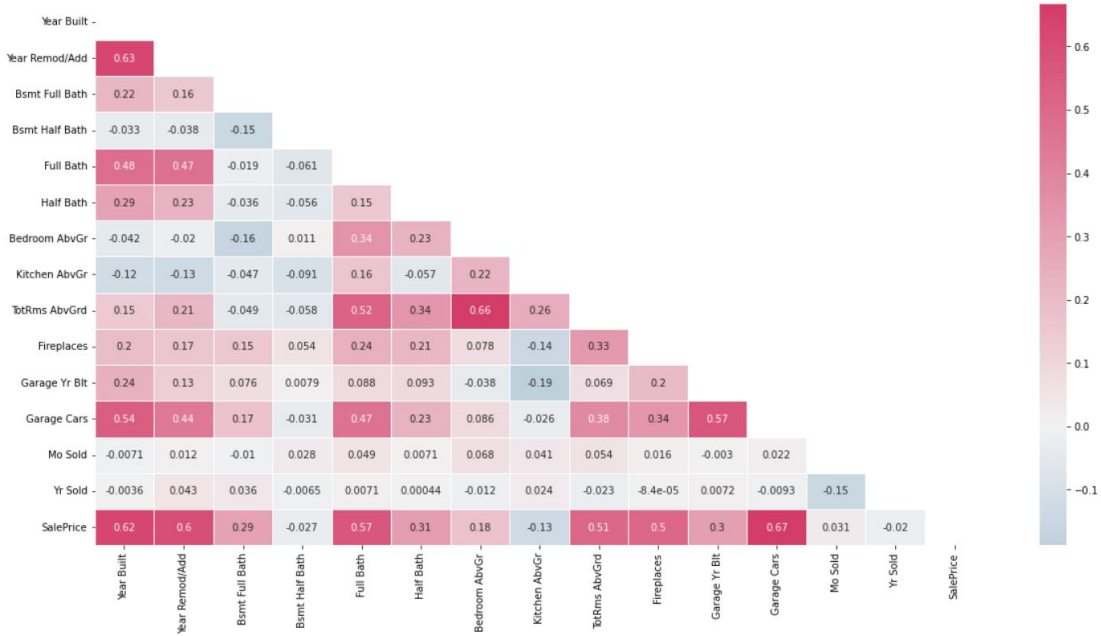
- Garage Qual is strongly correlated with Garage Cond
  - Suggests that Garages of higher quality tend to be in better condition
  - Good quality Garages may retain their condition better (?)



Other observations from Correlation with Target

- The later the year a property is built/remodded is strongly correlated to the SalePrice. This is reasonable as newer properties tend to be more expensive due to inflation and are likely of better condition.

- There is almost zero correlation between `Yr Sold` and `Mo Sold` with SalePrice, this suggests that timing the market to maximise property value does not yield higher prices.



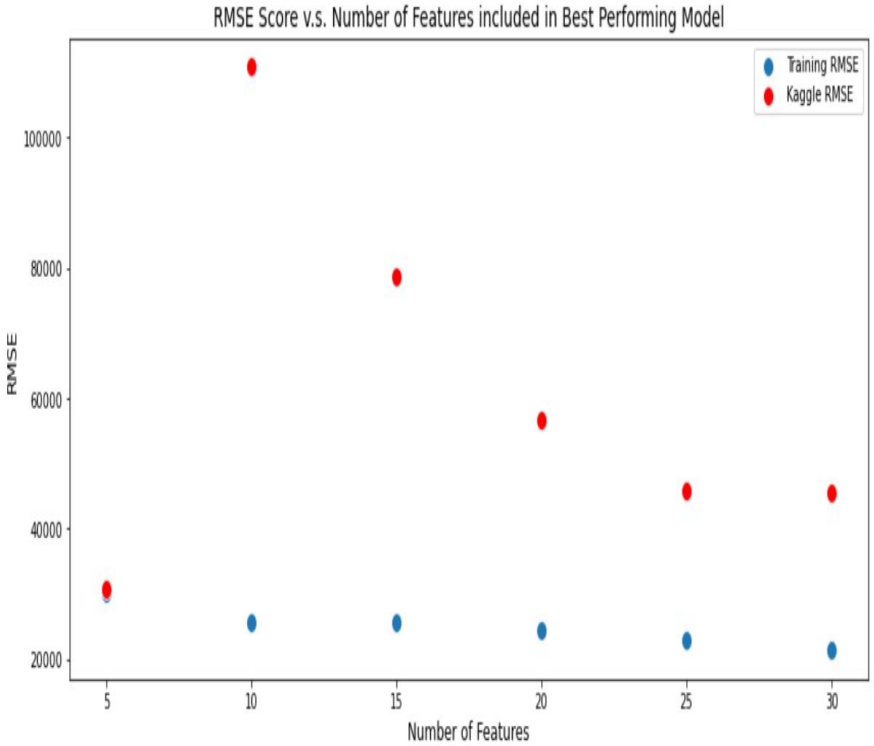
Modelling Results

Best Performing Models

Number of Features	Model Chosen	Validation Set RMSE
5	Ridge	30,273
10	LASSO	25,563
15	LASSO	25,613
20	LASSO	24,522
25	LASSO	22,861
30	LASSO	21,400
90 (All Features)	Ridge	20,290

Modelling Results

- Models submitted to Kaggle scored worse except for the Ridge model using 5 coefficients
- RMSE went as high as > 100K for 10 features and narrowed as more features are added
- Possible Explanantion for Poorer Performance on Unseen Data
    - The model likely overfitted when we take 10 features and above, as shown from the higher RMSE score on the unseen data when compared to when we used 5 features, even though the RMSE score improved after the inclusion of additional features.
    - There may be outliers in the unseen data, which was not picked up during data cleaning due to the format of a Kaggle competition, regression models are sensitive to outliers as the coefficients would be weighted to take the outliers into account.



Conclusion & Future Work

- Selecting 5 features led to a RMSE (taking the kaggle test set) that is approximately 18% (or ~$30K) of the mean SalePrice of the training set
    - Acceptable given that the valuation obtained via our web app is supposed to be a preliminary valuation.

- Therefore, we will select **features = 5** as the number of features in our web app that requires user input. This is in agreement with existing User Experience (UX) literature to keep fields in a form to below 10 fields for optimal user engagement and conversion.

Conclusion & Future Work

The 5 features selected are as follows (ranked in order of magnitude)

Feature	Coefficient
Overall Cond	0.308
has_bsmt	-0.27
Overall Qual	0.211
Gr Liv Area	0.140
Lot Area	0.043

Conclusion & Future Work

The accuracy of the model can be further improved by:

- Collecting more data from our US cities as certain preferences may be unique to Ames, Iowa. Collecting more samples will allow the model to generalize better to unseen data.

- There are many hidden features inside the `Neighborhood` feature as many other possible other features characterize whether a property in a particular neighborhood is of value, such as crime rate, proximity to good schools, proximity to employment centers or near rivers/lakes. Such features should be collected instead of stating the neighborhood of the property.

End