House Price Prediction System

Yuen Yee Lo, PhD

Data Science Intensive Capstone Project, May 2022

Springboard

Contents

- Introduction
- Data
- EDA
- Modelling
- Results and Analysis
- Conclusion

Why important?

- Help people buy a house
- Know the price range in the future
- Plan their finance
- Beneficial for property investors
- Know the trend of housing prices in a certain location

Who May Care?

House buyer

Real estates

PORCHLIGHT REAL ESTATE GROUP
ART ALEXRUBINTEAM
QUEST REALTY
THE AGENCY REALTY GROUP INC.
ANDREW ESTATES REALTY · GROUP
RealistREALTY International LTD
Vizelli & PARTNERS LUXURY HOMES | FAIR SERVICE
DAVE PERRY-MILLER REAL ESTATE
ROSE ASSOCIATES

Data and Data Wrangling

Data

- 1461 entries
- 81 explanatory variables describing (almost) every aspect
    - 36 numerical data
    - 43 category data

- Residential homes in Ames, Iowa
- Collected : 2006-2010

Example of data:

- LotArea: Lot size in square feet
- Bedroom: Number of bedrooms above basement level
- YearBuilt: Original construction date
- MSZoning: The general zoning classification
- Neighborhood: Physical locations within Ames city limits
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- Street: Type of road access
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Fireplaces: Number of fireplaces
- GarageType: Garage location
- PoolArea: Pool area in square feet
- Fence: Fence quality
- SaleCondition: Condition of sale
-
- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

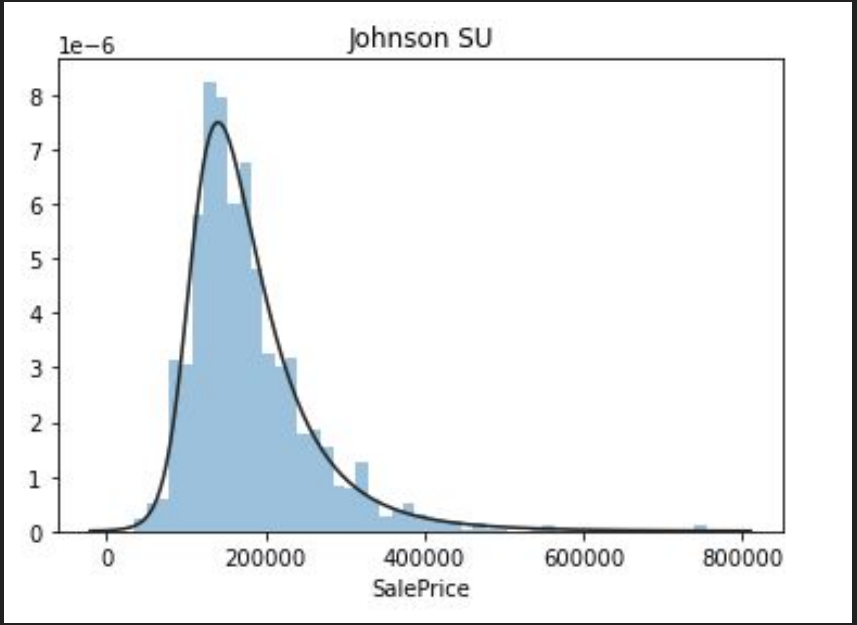
Handling missing data

- Some replaced 0 , e.g. number of pool
- Some replaced by mean, like lotsize
- For categories , replaced by “unknown”
- For some categories with lot of difference, grouped them into subgroups. E.g. neighborhood

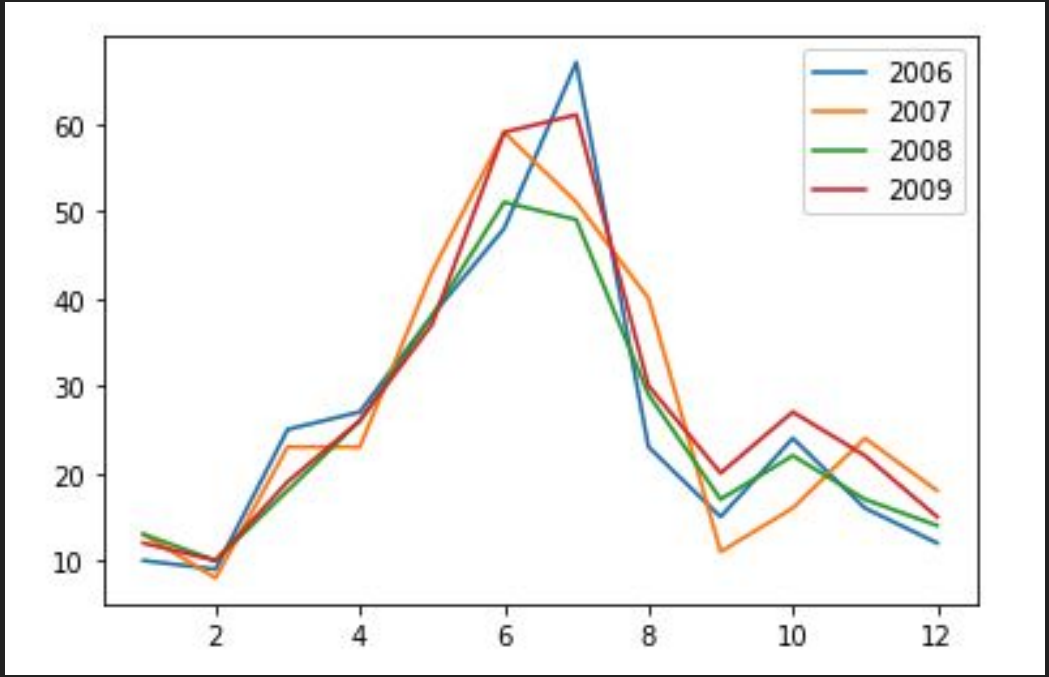
EDA

EDA

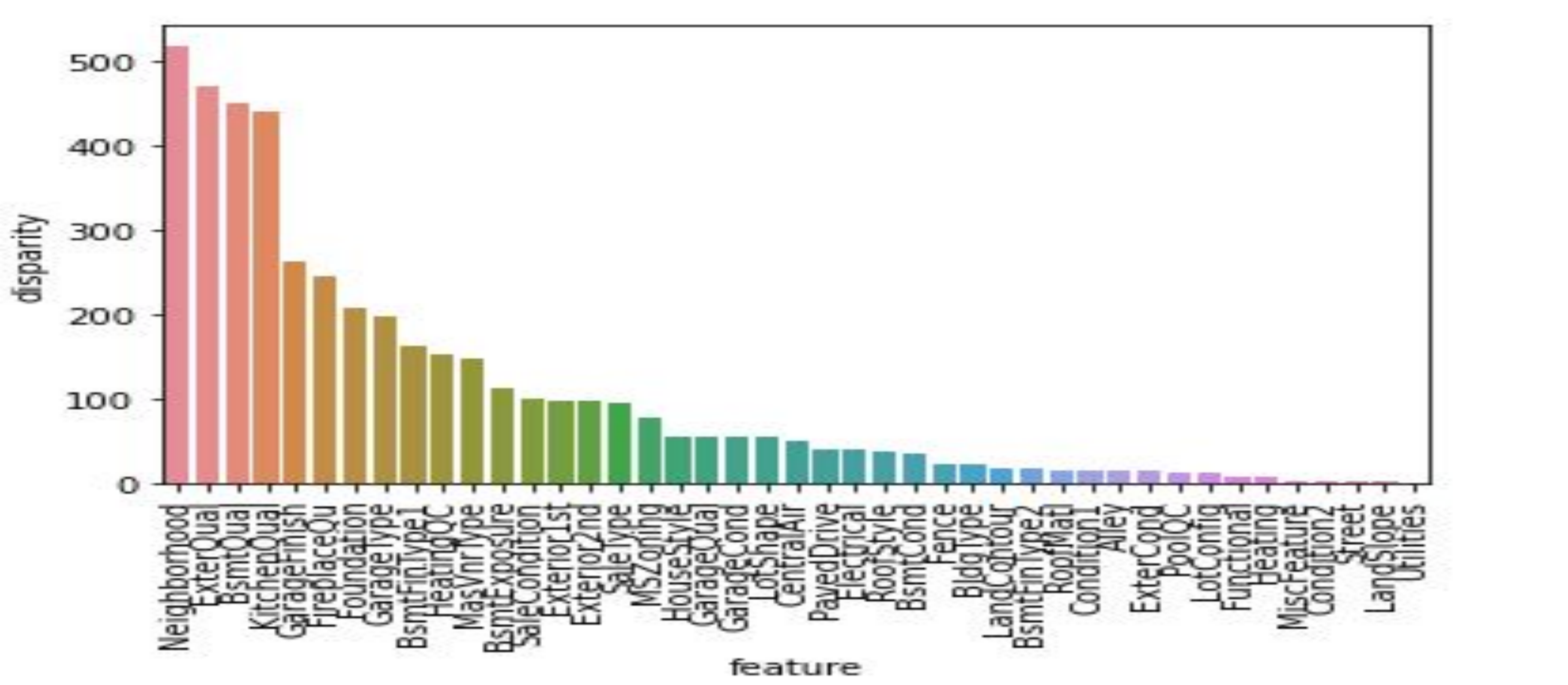
Sale Price distribution



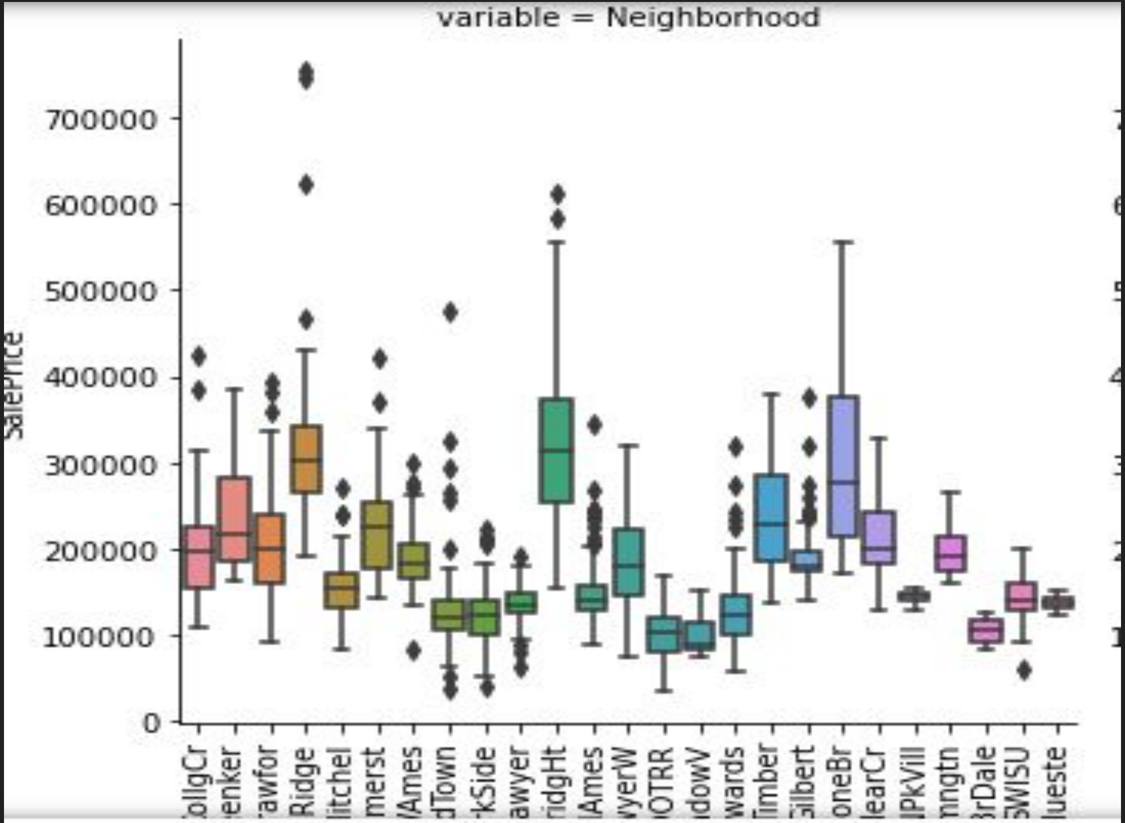
EDA: Number of house sold each month over year



EDA: Category data : Disparity



EDA: Neighborhood has big impact on house prices



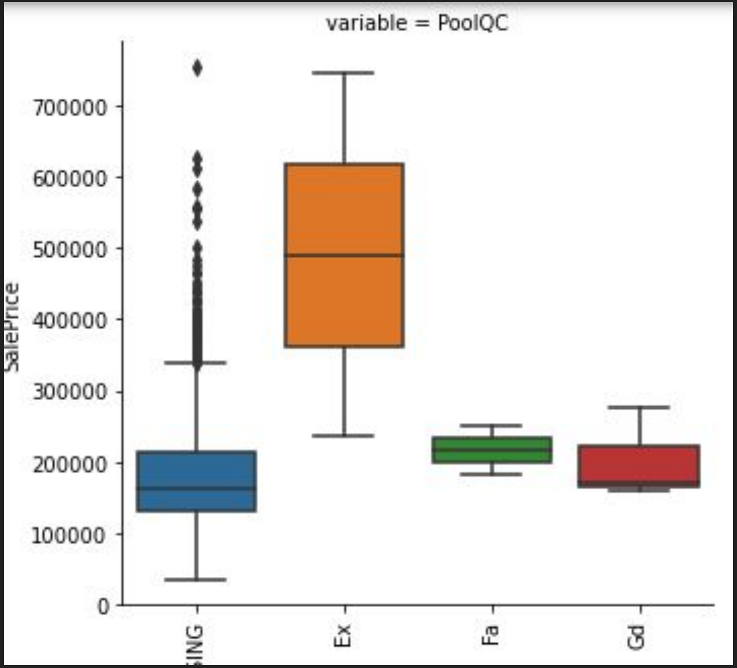
EDA: Having pool on property seems to improve price

Ex: excellent

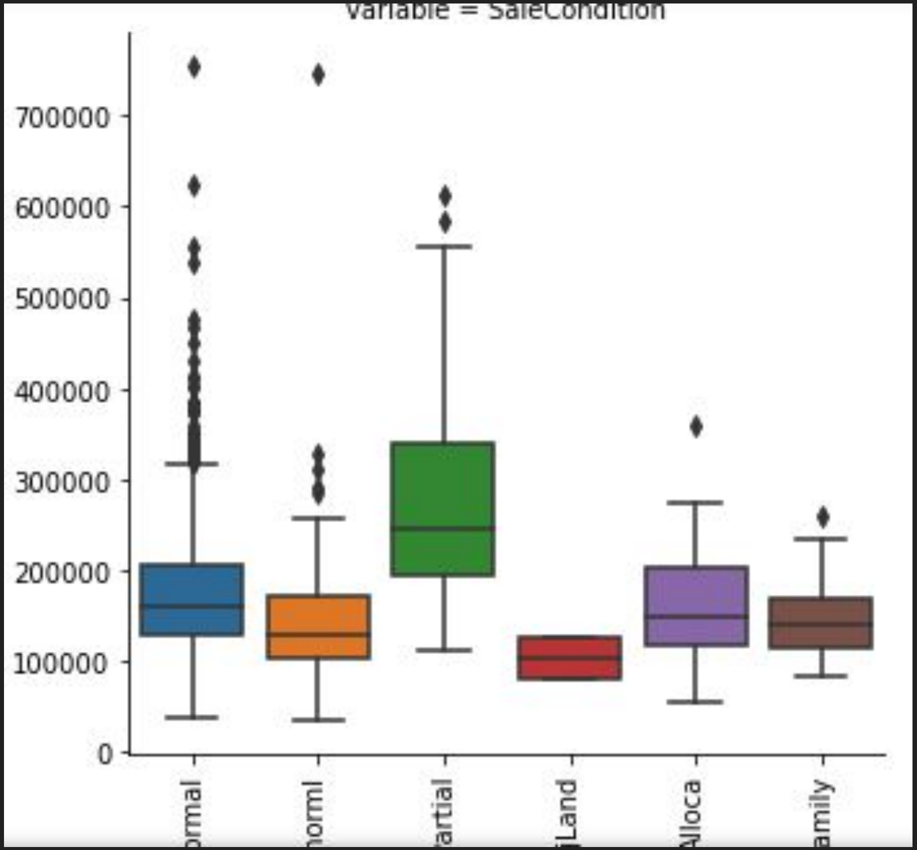
Gd: Good

Fa: Fair

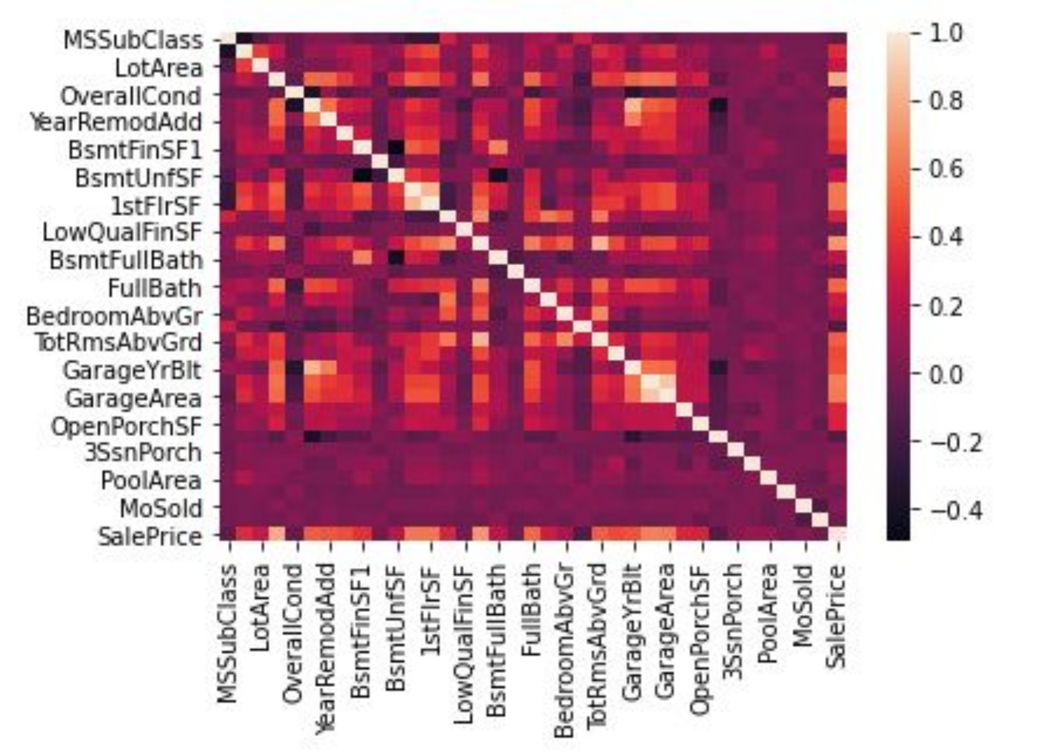
Missing



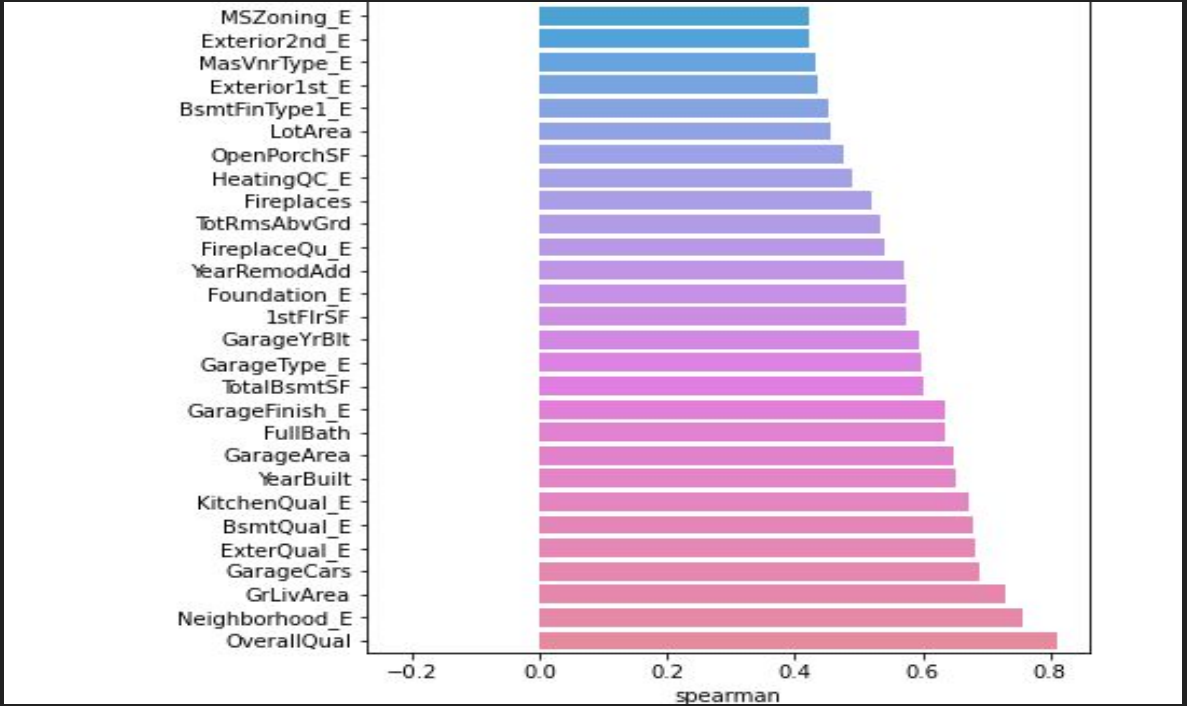
EDA: Partial Sale Condition has impact



Heatmap


MSSubClass
LotArea
OverallCond
YearRemodAdd
BsmtFinSF1
BsmtUnfSF
1stFlrSF
LowQualFinSF
BsmtFullBath
FullBath
BedroomAbvGr
TotRmsAbvGrd
GarageYrBlt
GarageArea
OpenPorchSF
3SsnPorch
PoolArea
MoSold
SalePrice
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Spearman's correlation measures the strength and direction of monotonic association between two variables.



Modelling

Training and testing data

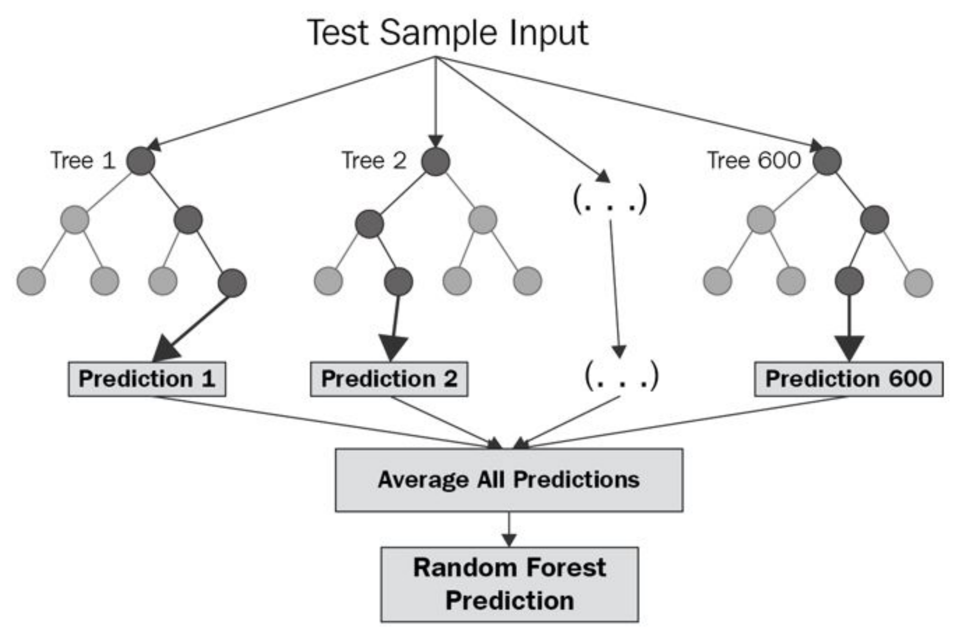
Split data for (80%) training and (20%) testing

Modelling: Baseline Linear Regression

- Linear approach for modelling the relationship between a scalar response and one or more explanatory variables
- Simple model
- Only numerical data

Modelling: Random Forest Regression

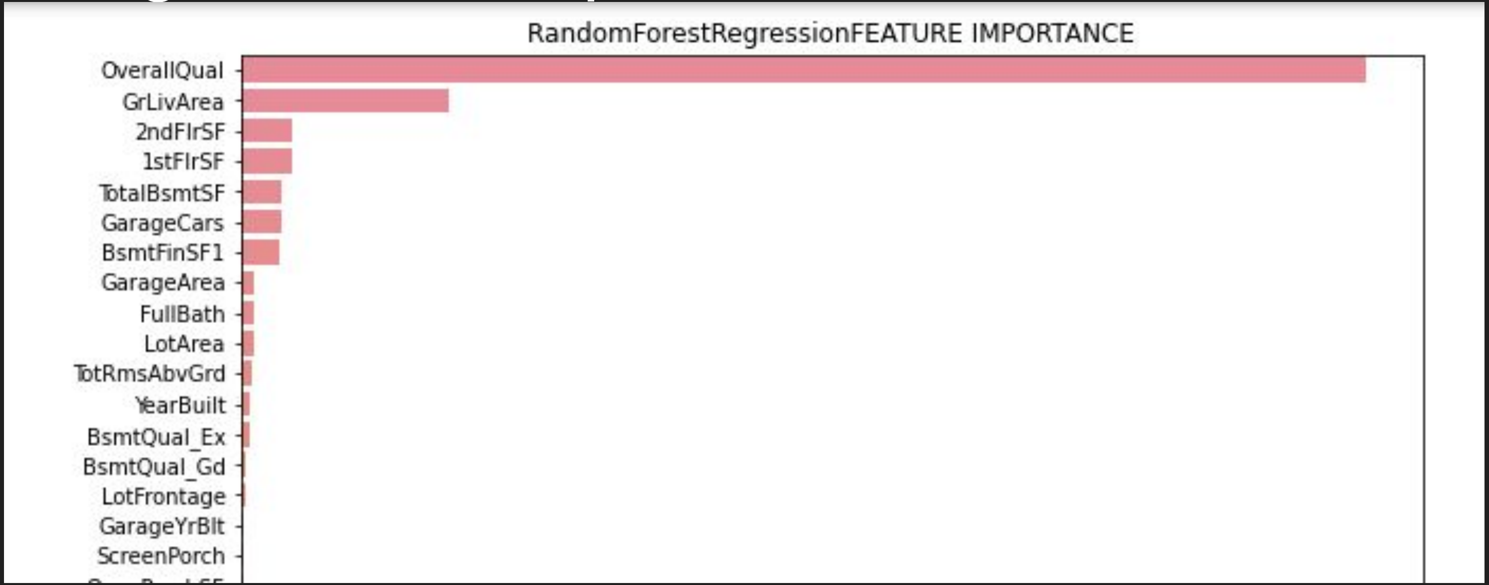
- Ensemble technique
- Regression and classification
- Bootstrap multiple DTs
- Reduce overfitting
- Good for interpretation
- Easy to spot outliner
- Provide features importance



Modelling: Feature Importance

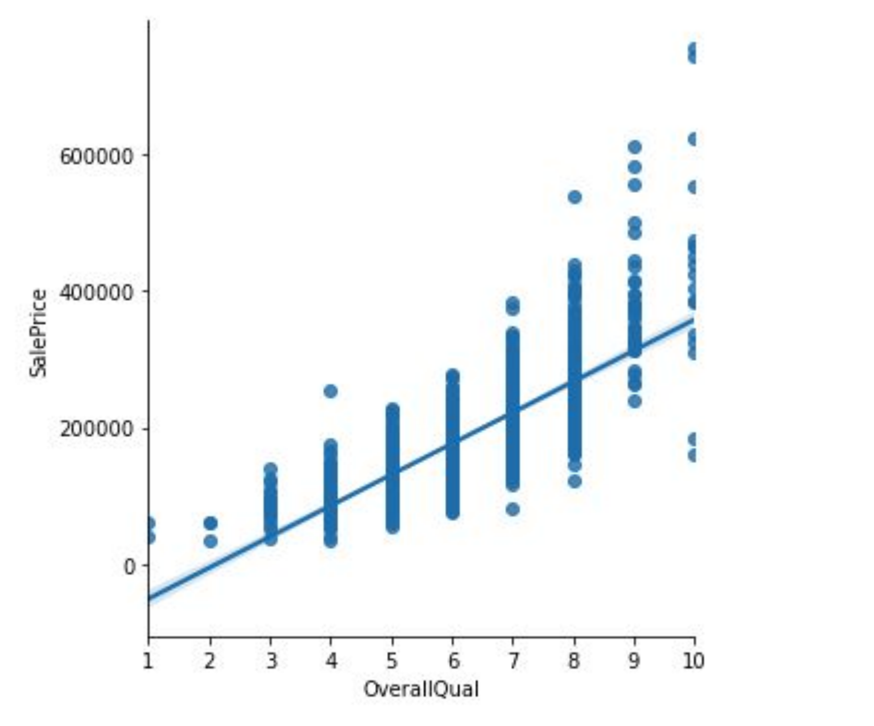
- Random forest can give us insight about the feature importance
- Reduce number of features
- Reduces the complexity of a model
- Easier to interpret
- Improves the accuracy if the right subset is chosen

Modelling: Feature Importance



- Feature “OverallQual” is the most important features
- Top10 important features are numerical data
- Category data are not significant

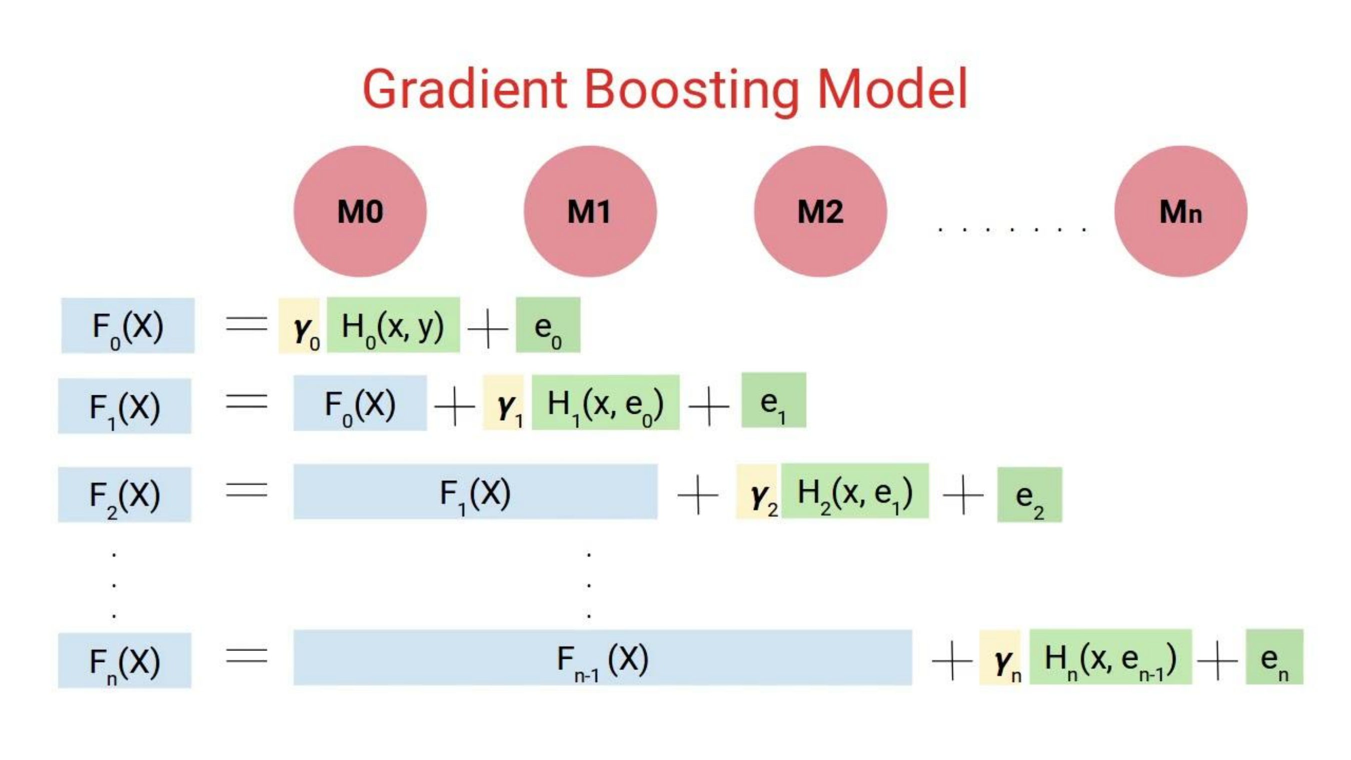
OverallQual vs. SalePrice



Modelling: **Gradient Boosting Regression**

- Trains many models in a gradual, additive and sequential manner
- The major difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners (eg. decision trees)
- Loss function would be based off the error between true and predicted house prices

Gradient Boosting



Gradient Boosting Model

$$F_0(X) = \gamma_0 \; H_0(x, y) + e_0$$

$$F_1(X) = F_0(X) + \gamma_1 \; H_1(x, e_0) + e_1$$

$$F_2(X) = F_1(X) + \gamma_2 \; H_2(x, e_1) + e_2$$

$$\vdots \qquad \vdots$$

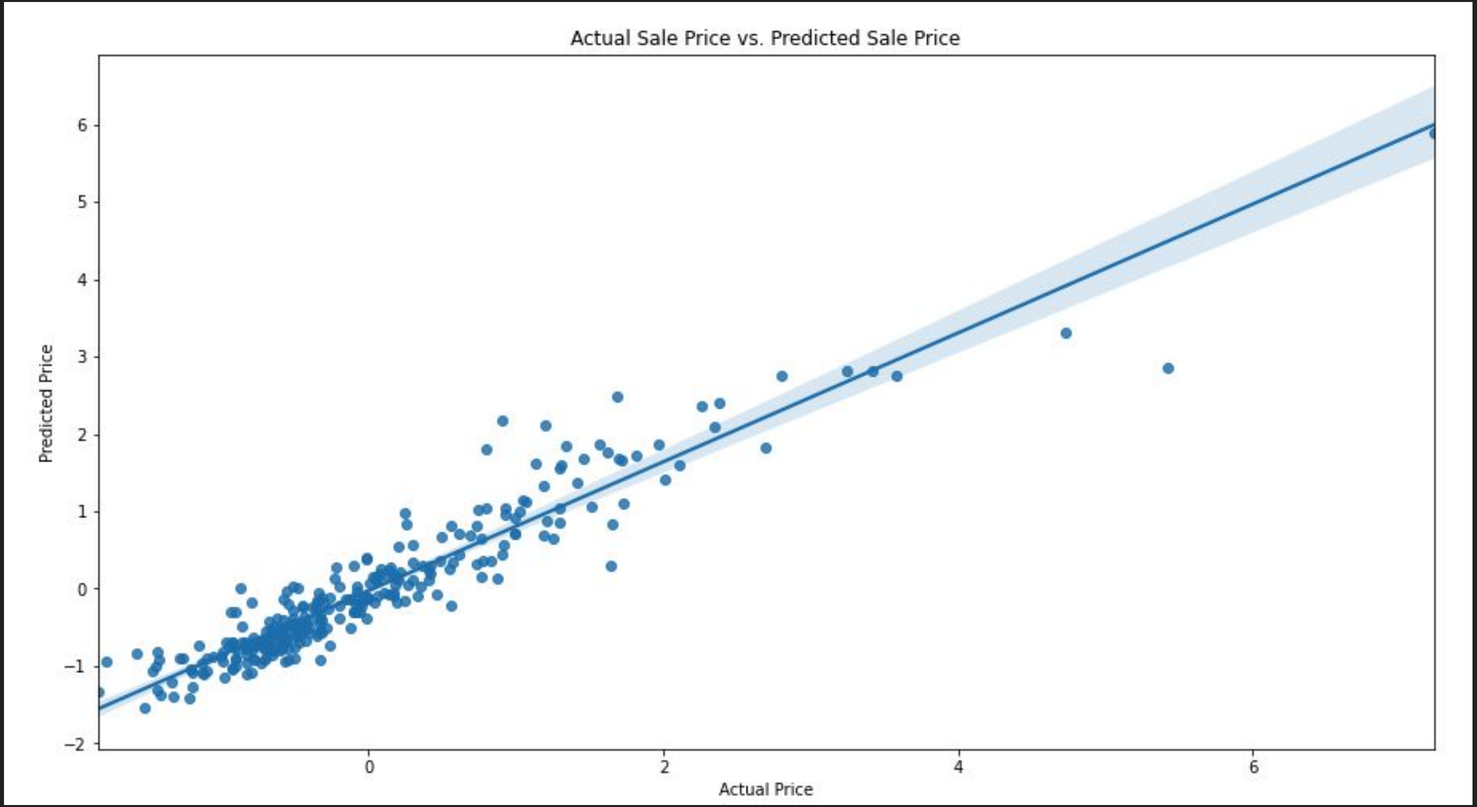
$$F_n(X) = F_{n-1}(X) + \gamma_n \; H_n(x, e_{n-1}) + e_n$$

Results:

- MAE, MSE, RMSE for result matric
- Gradient boosting performs the best
- Reduce features can improve the performance

	MAE	MSE	RMSE
Linear Regression	0.289	0.215	0.463
Random Forest	0.288	0.199	0.447
Random Forest Top10 features	0.288	0.187	0.433
Gradient Boosting	**0.241**	**0.133**	**0.365**

Predicted housing price (test data)



Conclusions

- Advanced regression techniques
- Built a baseline model using linear regression
- Compare to the random forest regression and gradient boosting regression
- Feature selection from the features importance of RF
- Categories data show no significant impact
- Gradient boosting regression performs the best
- Further improvement:
    - parameter tuning
    - deep learning (if we have more data)

Thank You

Question?

Contact

Yuen Yee Lo

Email: yylo7775@gmail.com

https://www.linkedin.com/in/yuen-yee-lo-4b74865/

https://github.com/yuenyeelo/

http://www.y2nlp.com

Project directory: https://github.com/yuenyeelo/springboard/tree/main/Capstone2