House Price Prediction System

Yuen Yee Lo, PhD

Data Science Intensive Capstone Project, May 2022

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

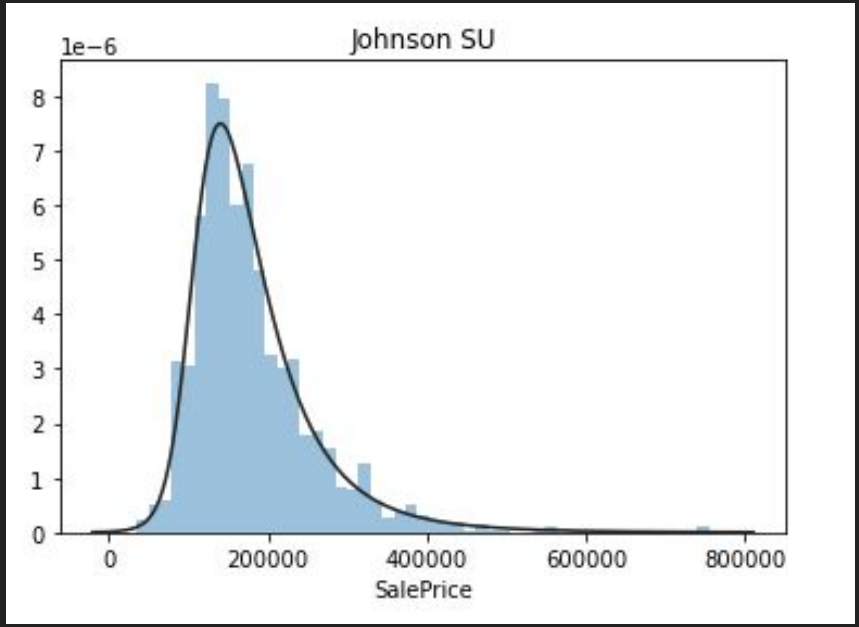
PorchLight Real Estate Group
ART AlexRubinTeam
Quest Realty
The Agency Realty Group Inc.
Andrew Estates Realty Group
RealListREALTY International LTD
Tirelli & Partners Luxury Homes | Fair Service
Dave Perry-Miller Real Estate
Rose Associates

Data

- 1461 entries
- 81 explanatory variables describing (almost) every aspect
    - 36 numerical data
    - 43 category data
- Residential homes in Ames, Iowa
- 2006-2010

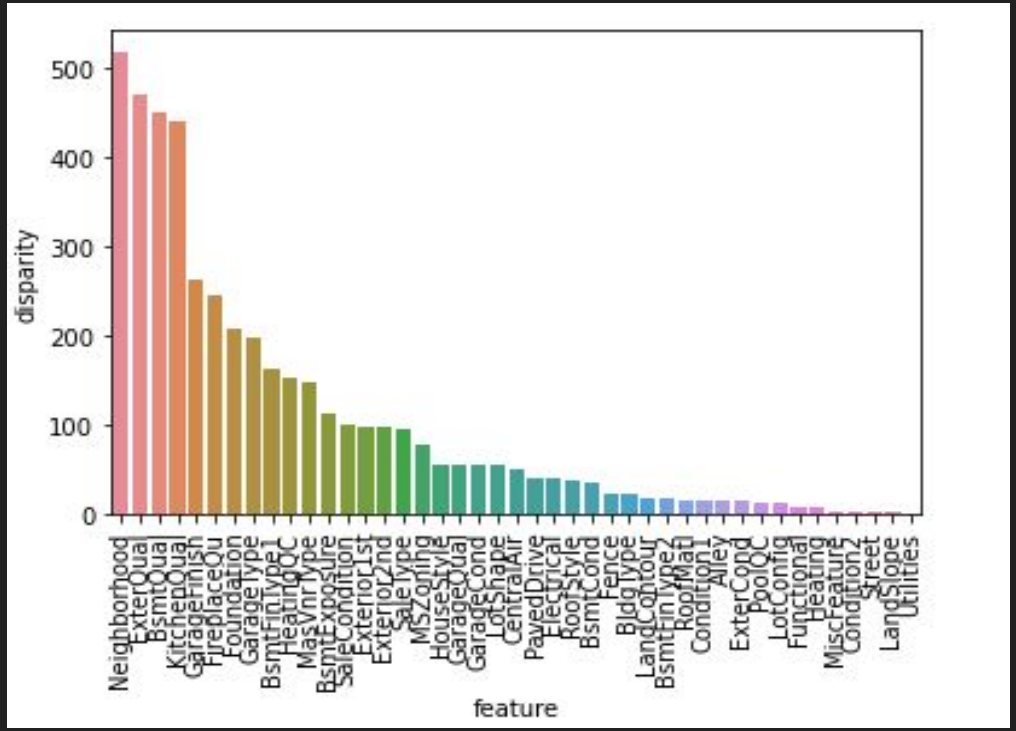
EDA

Sale Price distribution



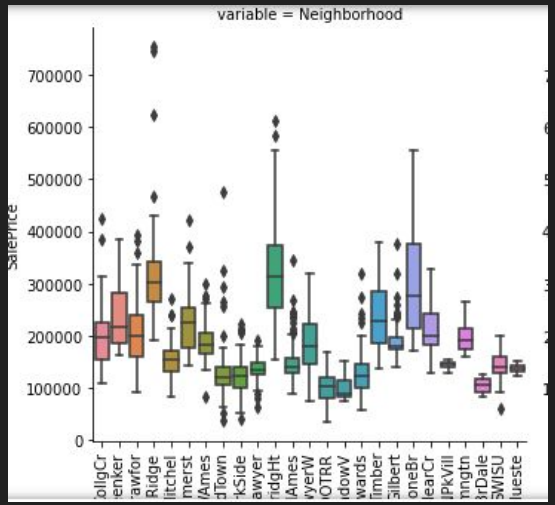
EDA: Category data : Disparity

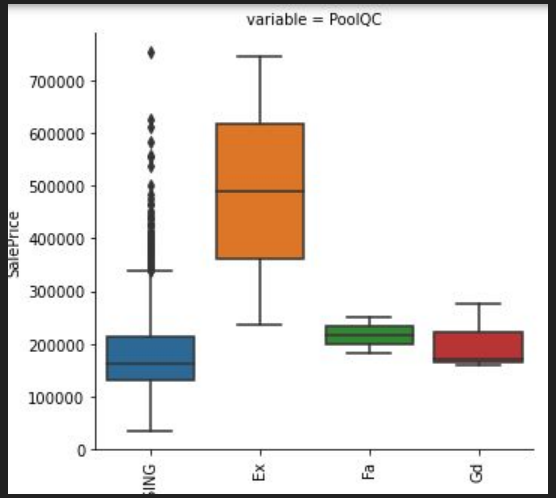
-

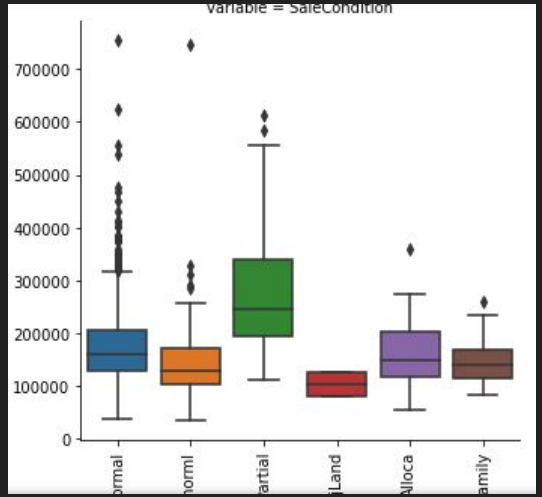


EDA:

- Some categories seem to more diverse
- Neighborhood has big impact on house prices
- Having pool on property seems to improve price
- Partial SaleCondition has impact







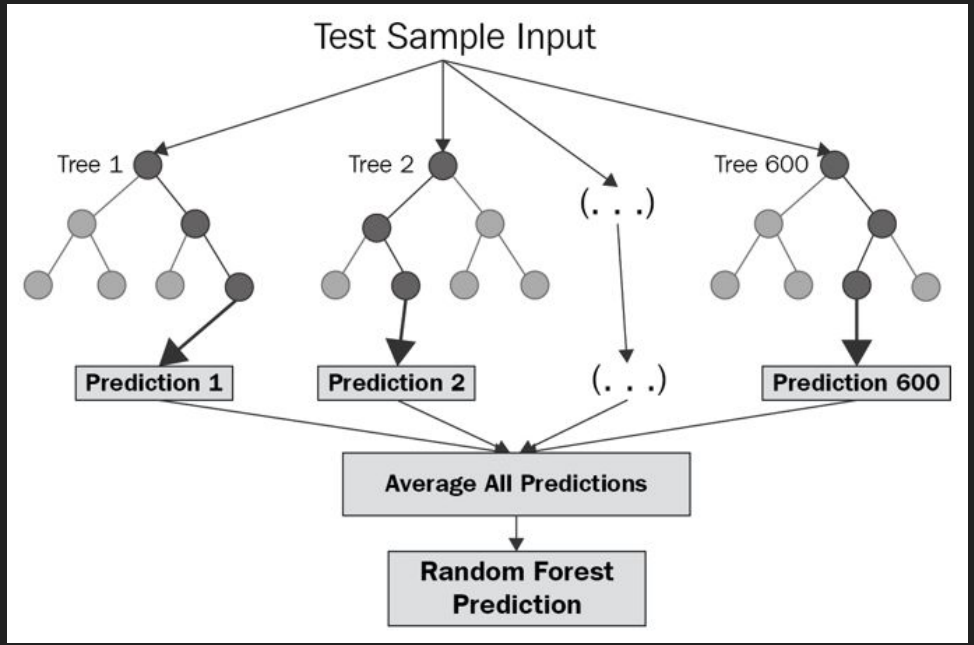
Modelling

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



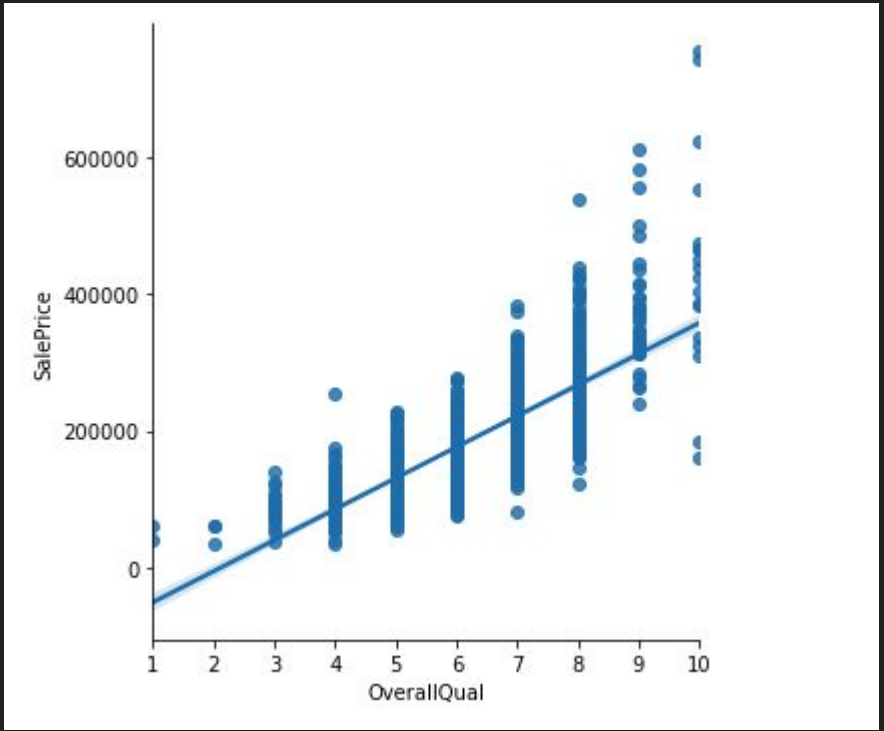
Modelling: Feature Importance

- 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

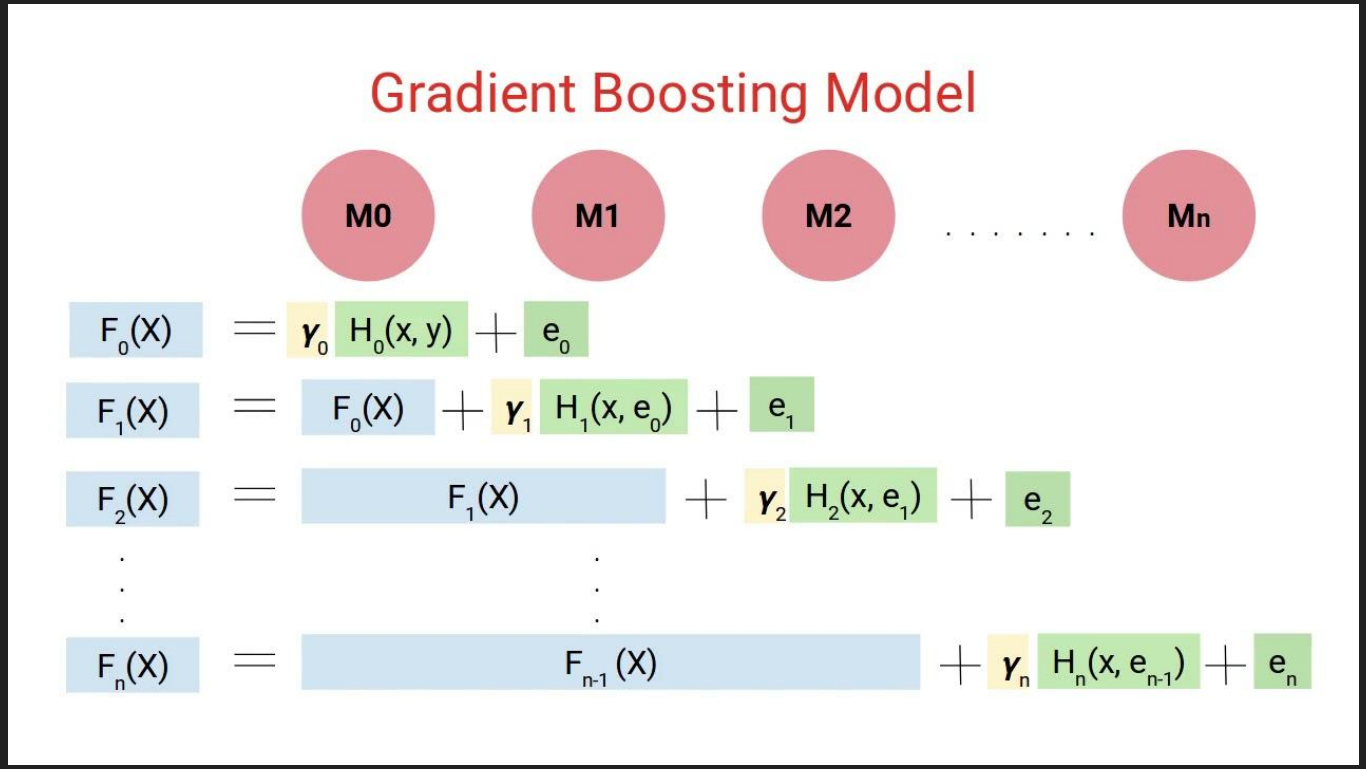
OverallQual vs. SalePrice



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

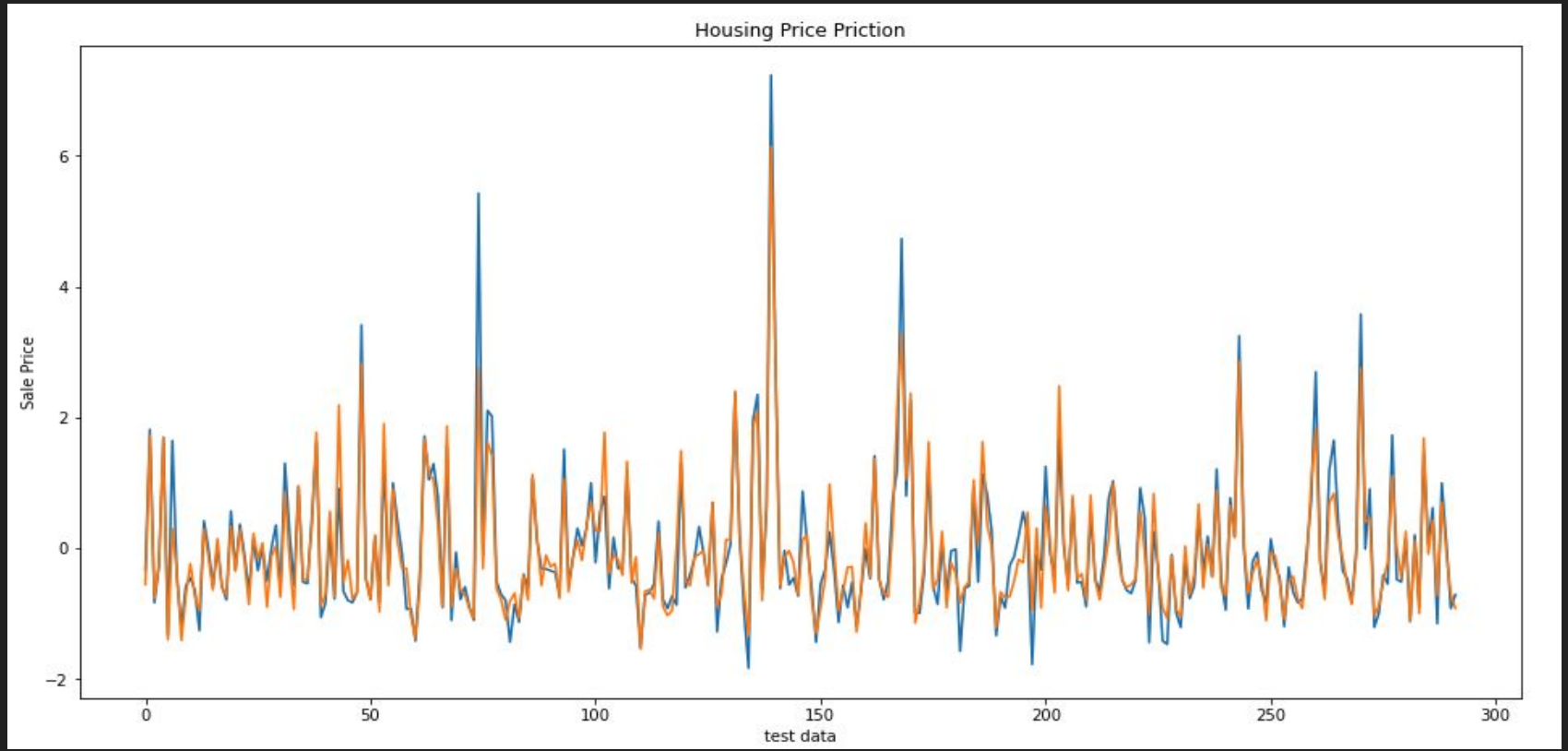


Results:

- 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

- 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)