Mini Project 3

House Price Classification Modelling Report

Max Kim

Objectives

- Multiple different factors influencing house price and it is difficult to estimate house price with just few features

- Indicate house buyers to know the price range of their dream house

- Provide property investors to understand the trend of housing prices

- Help developers to determine the selling price range of a house

Dataset

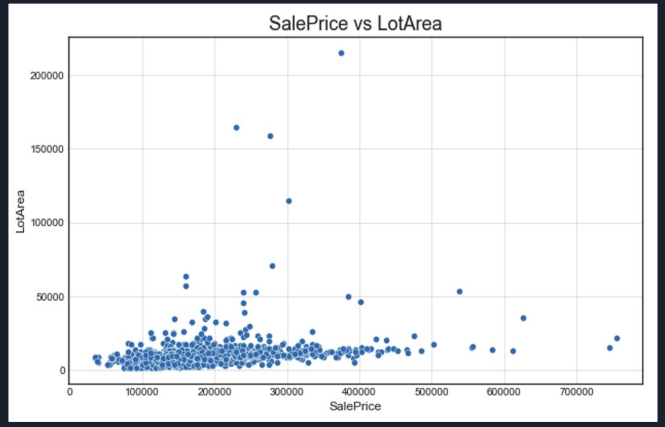
- Data source from Kaggle
  - https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview
  - Containing various features with house prices stated
- Size & Volume
  - The dataset contains 1460 records of house prices with 79 features
- Data Dictionaries examples
  - GrLivArea: Above grade (ground) living area square feet
  - LotArea: Lot size in square feet
  - SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
  - OverallQual: Overall material and finish quality
  - GarageArea: Size of garage in square feet
  - Fence: Fence quality
  - Kitchen: Number of kitchens

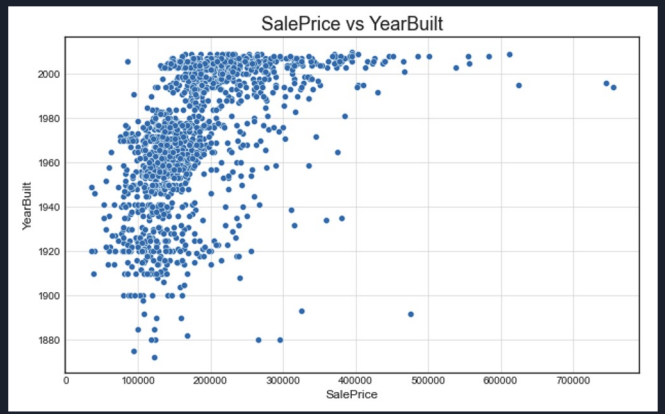
Target Variable

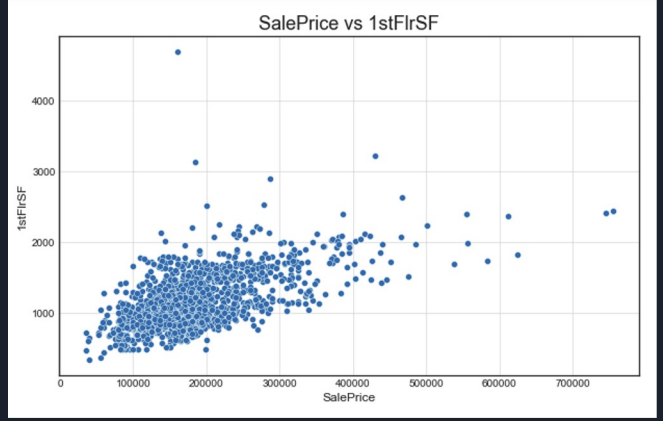
- Target Variable: Sale Price
  - Our target variable is stated as exact amount of the sales price. Divide these prices into 4 different categories (based on quantile) to predict range of prices that a house will fall into.

Category	Range
0	<=110K
1	Between 110K - 163K
2	Between 163K - 214K
3	>214K

Relationship







Base Modelling

Model	Accuracy	ROC AUC	Precision	Recall	F1-Score
Logistic Regression	0.67	0.88	0.67	0.67	0.67
KNN	0.66	0.85	0.67	0.66	0.67
SVC	0.53	0.81	0.39	0.53	0.42
Decision Tree	0.75	0.89	0.75	0.75	0.75
Random Forest	**0.78**	**0.94**	**0.78**	**0.78**	**0.78**

- **Logistic Regression**: High area under the curve but poor score in class 0 & 2 with overall not a good model score
- **KNN**: Similar to Logistic Regression model with poor score in class 2
- **SVC**: Low overall accuracy with 0 precision and recall for class 0
- **Decision Tree:** High accuracy and area under the curve. Overall balanced values across all classes for precision and recall
- **Random Forest:** Highest accuracy with area under curve. Overall better score for precision and recall for all classes than decision tree

Bagging & Boosting

- Attempt bagging and boosting technique to improve on our models.

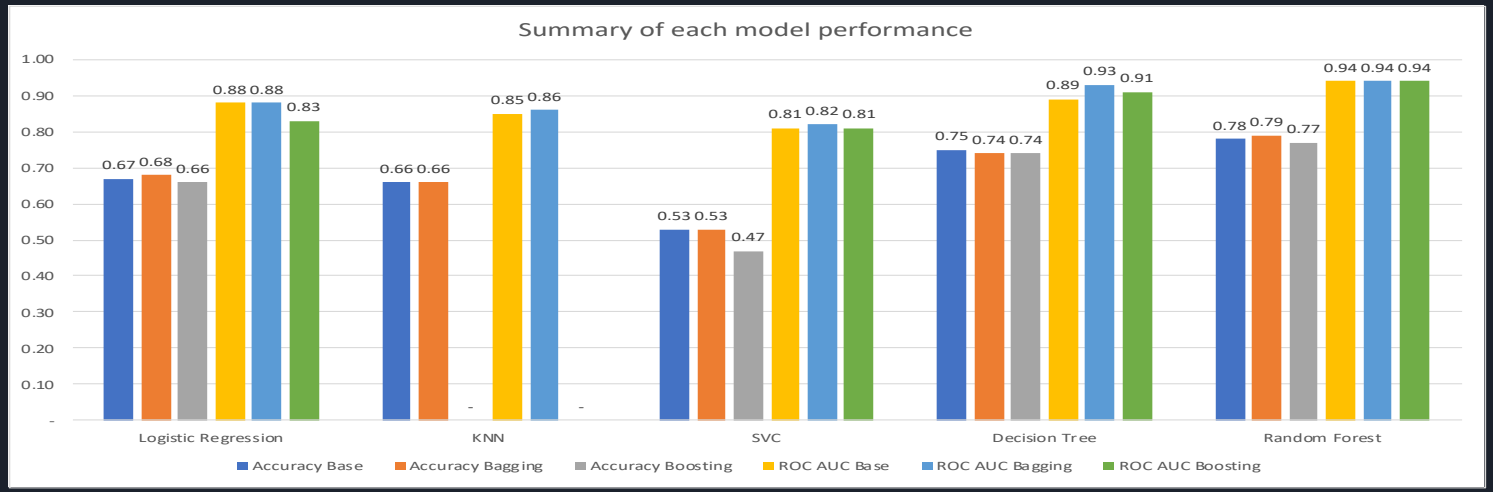
Bagging Model	Accuracy	ROC AUC	Precision	Recall	F1-Score
Logistic Regression	0.68	0.88	0.68	0.68	0.68
vs Base	+0.01	-	+0.01	+0.01	+0.01
KNN	0.66	0.86	0.67	0.66	0.66
vs Base	-	+0.01	-	-	-0.02
SVC	0.53	0.82	0.52	0.52	0.52
vs Base	-	+0.01	+0.13	-	+0.10
Decision Tree	0.74	0.93	0.74	0.74	0.74
vs Base	-	+0.06	-	-	+0.01
Random Forest	0.79	0.94	0.79	0.79	0.79
vs Base	+0.01	-	-	+0.01	+0.01

Bagging & Boosting

AdaBoosting Model	Accuracy	ROC AUC	Precision	Recall	F1-Score
Logistic Regression	0.66	0.83	0.65	0.66	0.65
vs Base	-0.01	-0.05	-0.02	-0.01	-0.02
SVC	0.47	0.81	0.49	0.47	0.39
vs Base	-0.06	-	+0.10	-0.06	-0.03
Decision Tree	0.74	0.91	0.76	0.74	0.74
vs Base	-	+0.04	+0.02	-	+0.01
Random Forest	0.77	0.94	0.77	0.77	0.77
vs Base	-0.01	-	-0.02	-0.01	-0.01

No boosting for KNN since it did not have the attributes for supporting sample weighting

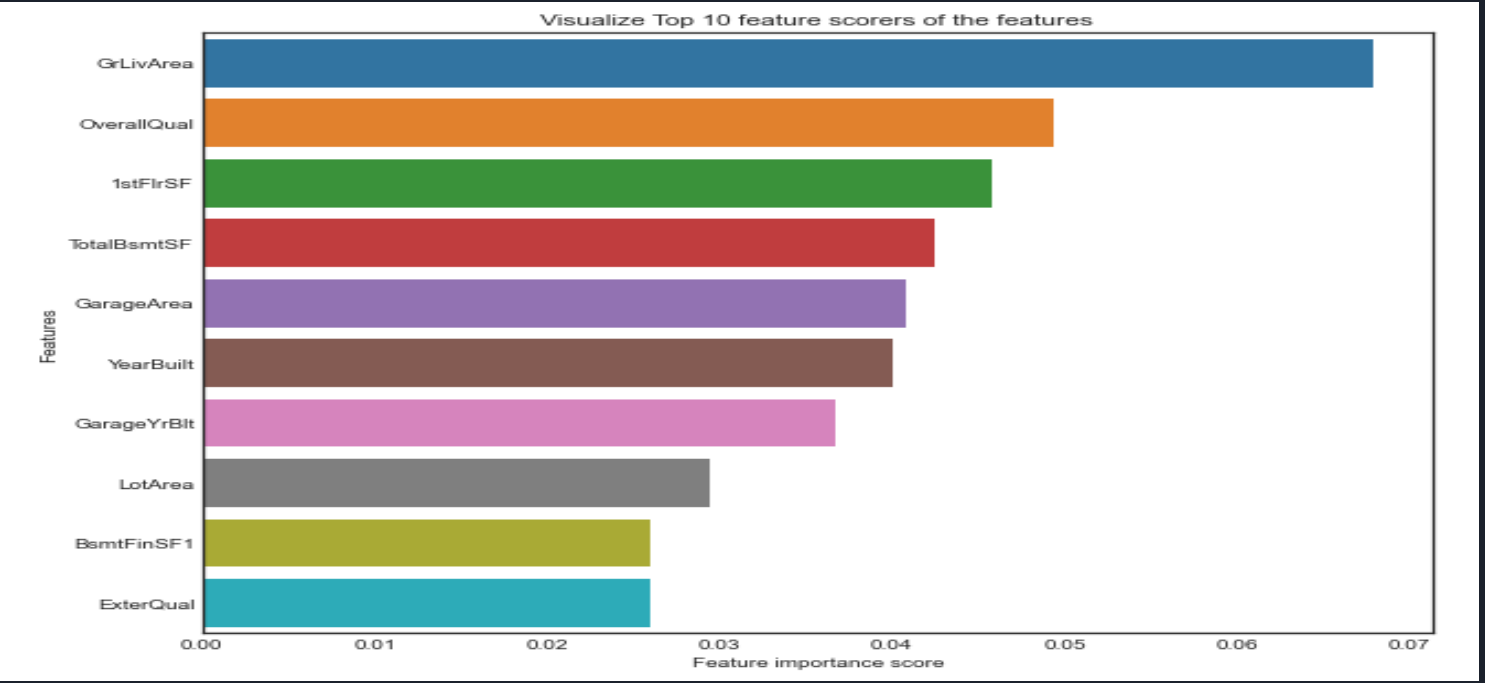
Modelling Summary



- Overall, random forest had highest performance out of these models. Boosting method mostly decreased the performance
- From the summary, we can observe random forest with bagging achieved the highest results but has very small improvement on accuracy comparing to the base model (+0.1%)

Feature Importance

- Using Random Forest to see what are the important features to consider when deciding the price range of the house



Conclusion

- Train multiple classification models with number of different features of houses
- Observed that the feature that strongly correlated with house price is the living area size and overall material quality
- Best base model performance was from random forest. There were slight improvement on the accuracy for this model from bagging, however, there was no significant improvement.
- Potentially further optimized through parameter adjustment as there are still room for improvement since highest performance at the moment is 0.79 accuracy rate
- As time pass, the trend and feature importance can be change. In order to be a practical use, we need to input latest sales price and features.

Reference

- https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview
- https://www.kaggle.com/code/pmarcelino/comprehensive-data-exploration-with-python