Capstone Project

Home Sales Prices

By: Ryan Mark

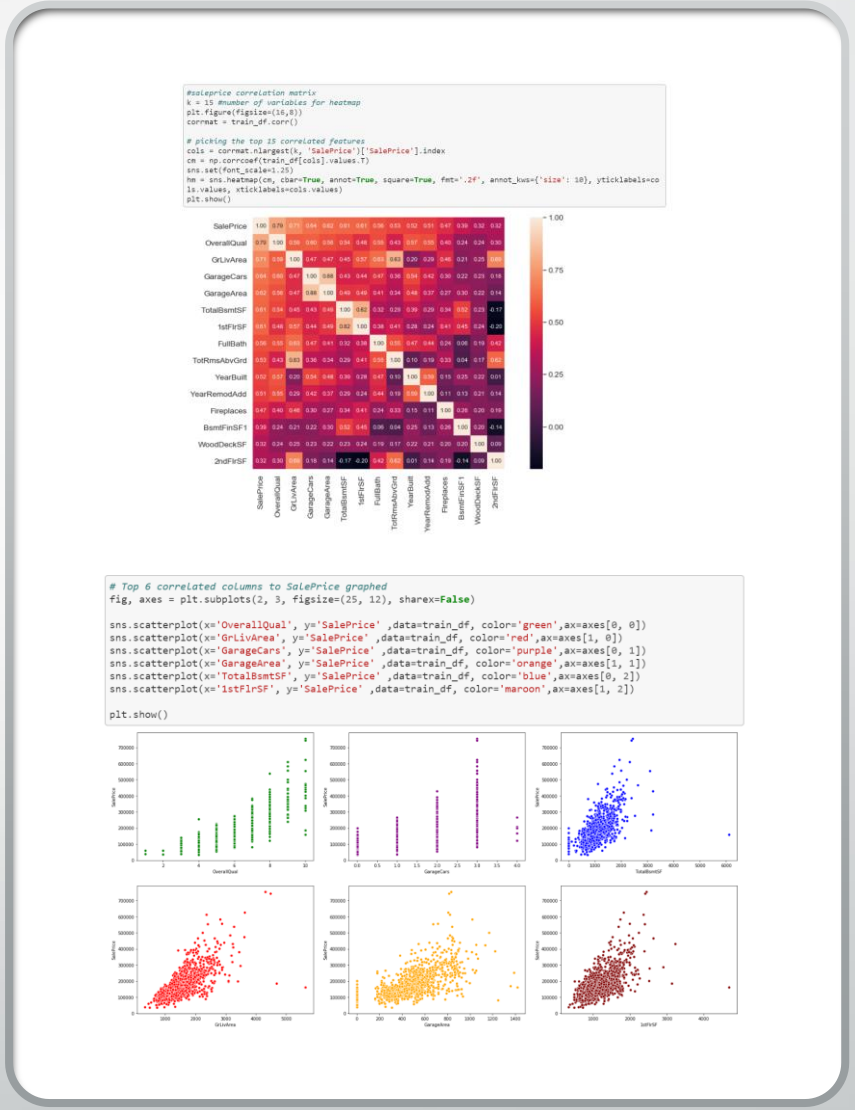
Contents

Executive Summary

- The Ames, Iowa housing dataset captures various details about residential properties in that area from 2006 to 2010. The objective is to facilitate the use of machine learnings models to predict the value of home sales prices.

Exploratory Data Analysis (EDA)

- The initial Exploratory Data Analysis (EDA) was focused on understanding the target variable ‘SalePrice’.
- From this information it can be determined that the volume of prices was between 100k and 200k.
- This output indicated that Overall Quality had a strong positive correlation with Sale Price.



Review research design and modeling methods

- The modeling methods used in our analysis consisted of Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting. The intent was to understand the performance of Random Forest and Gradient Boosting in respect to the previously used models of Linear / Ridge regression. Through Random Forest Regressor a mean prediction can be obtained from the multiple decision trees within that model. Through Gradient Boosting, it can strengthen a model with weak predictions and make it better.

Results

- Robust Scaling was used on all models to scale the data based on the interquartile range. This was done to address the robust approach of this scaling method to outliers. This scaling approach was used on the Random Forest and Gradient Boosting models initially as a test. It should be known that in principle tree models do not require scaling it is unaffected by monotonic transformations.
- Our results indicated that the Gradient Boosting Model performed the best in terms of RMSE. This was further evidenced by the Kaggle score submission which was the highest score achieved (score = .164680). This indicates that the predictor values in our top model was the best in terms of having the lowest prediction error. Furthermore, Both the Random Forest and Gradient Boosting performed distinctively better than the Linear Regression and Ridge Regression models.

Exposition, problem description, and management recommendations

- The comparison of Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting indicated a significant performance when utilizing multiple tree methods. In particular it was observed that Random Forest and Gradient Boosting superior predictive results compared to our base regression models. Based on the model comparisons, Gradient Boosting is the best model for predicting home prices.

```
data = {'Model':        ['Linear Regression','Ridge Regression', 'Ra
ndom Forest_2', 'Gradient Boosting_2'],
        'RMSE Score':   [.454, .454, .378, .354, 31341.180, 29687.79
        'Kaggle Score': [4.70255, 4.70256, 10.76463, 11.21304, .1660
        'Scaling?':      ['Scaled','Scaled','Scaled','Scaled','Not S
       }

results = pd.DataFrame(data, columns = ['Model', 'RMSE Score','Kaggl
results.style.highlight_max(color='red').highlight_min(color='green'
```

	Model	RMSE Score	Kaggle Score	Scaling?
0	Linear Regression	0.454000	4.702550	Scaled
1	Ridge Regression	0.454000	4.702560	Scaled
2	Random Forest	0.378000	10.764630	Scaled
3	Gradient Boosting	0.354000	11.213040	Scaled
4	Random Forest_2	31341.180000	0.166010	Not Scaled
5	Gradient Boosting_2	29687.793000	0.164680	Not Scaled