Capstone Project

Home Sales prices

By: Ryan Mark

Contents

I. 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.

II. Data preparation, exploration, visualization

The initial Exploratory Data Analysis (EDA) was focused on understanding the target variable 'SalePrice'. In cell [10] a distribution of 'SalePrice' was performed to visualize the distribution of home prices. From this information it can be determined that the volume of prices was between 100k and 200k. Next, in cell [11] a correlation Data-Frame was created to identify the columns that had the highest correlation with 'SalePrice'. This output indicated that Overall Quality had a strong positive correlation with Sale Price. The top 6 positive correlations were then graphed in cell [12] in order to visualize the relationships between these variables. Next, the removal of missing values was implemented in the training and test datasets. In the training set any column with missing values above a count of 81 was removed. Likewise, any column in the test set that had a missing value above the threshold of 78 was removed.

```
#saleprice correlation matrix
k = 15 #number of variables for heatmap
plt.figure(figsize=(16,8))
corrmat = train_df.corr()

# picking the top 15 correlated features
cols = corrmat.nlargest(k, 'SalePrice')['SalePrice'].index
cm = np.corrcoef(train_df[cols].values.T)
sns.set(font_scale=1.25)
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, fmt='.2f', annot_kws={'size': 10}, yticklabels=co
ls.values, xticklabels=cols.values)
plt.show()
```

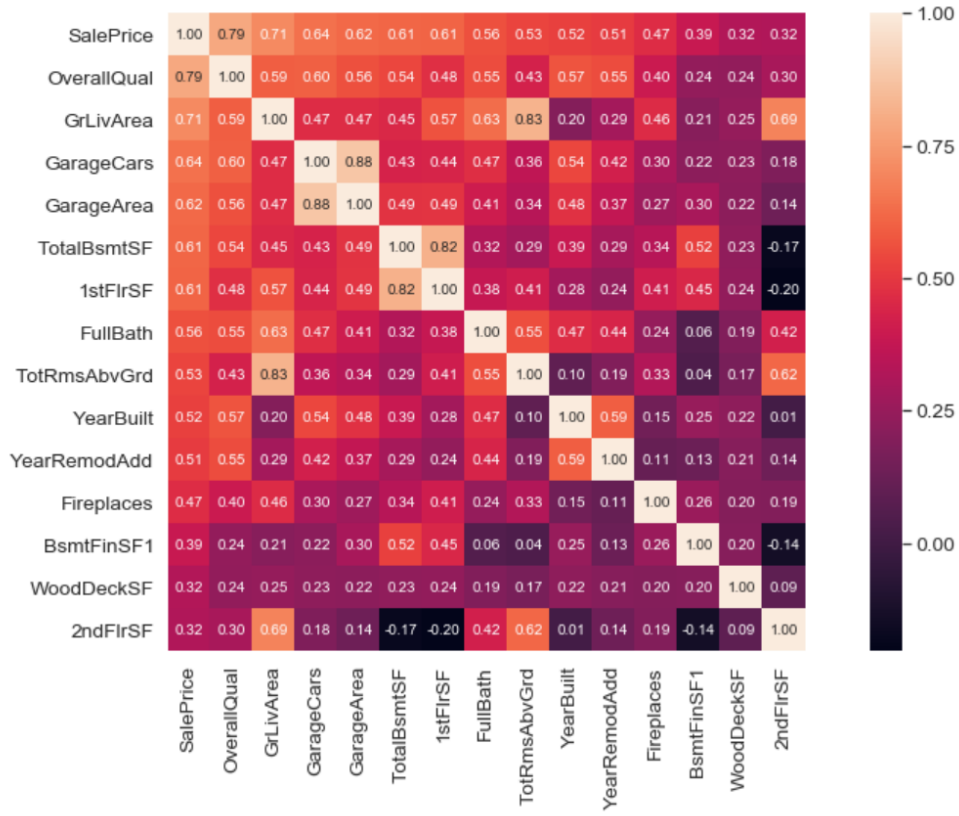

Figure 1 – Correlation Heat Map

```
# Top 6 correlated columns to SalePrice graphed
fig, axes = plt.subplots(2, 3, figsize=(25, 12), sharex=False)

sns.scatterplot(x='OverallQual', y='SalePrice' ,data=train_df, color='green',ax=axes[0, 0])
sns.scatterplot(x='GrLivArea', y='SalePrice' ,data=train_df, color='red',ax=axes[1, 0])
sns.scatterplot(x='GarageCars', y='SalePrice' ,data=train_df, color='purple',ax=axes[0, 1])
sns.scatterplot(x='GarageArea', y='SalePrice' ,data=train_df, color='orange',ax=axes[1, 1])
sns.scatterplot(x='TotalBsmtSF', y='SalePrice' ,data=train_df, color='blue',ax=axes[0, 2])
sns.scatterplot(x='1stFlrSF', y='SalePrice' ,data=train_df, color='maroon',ax=axes[1, 2])

plt.show()
```

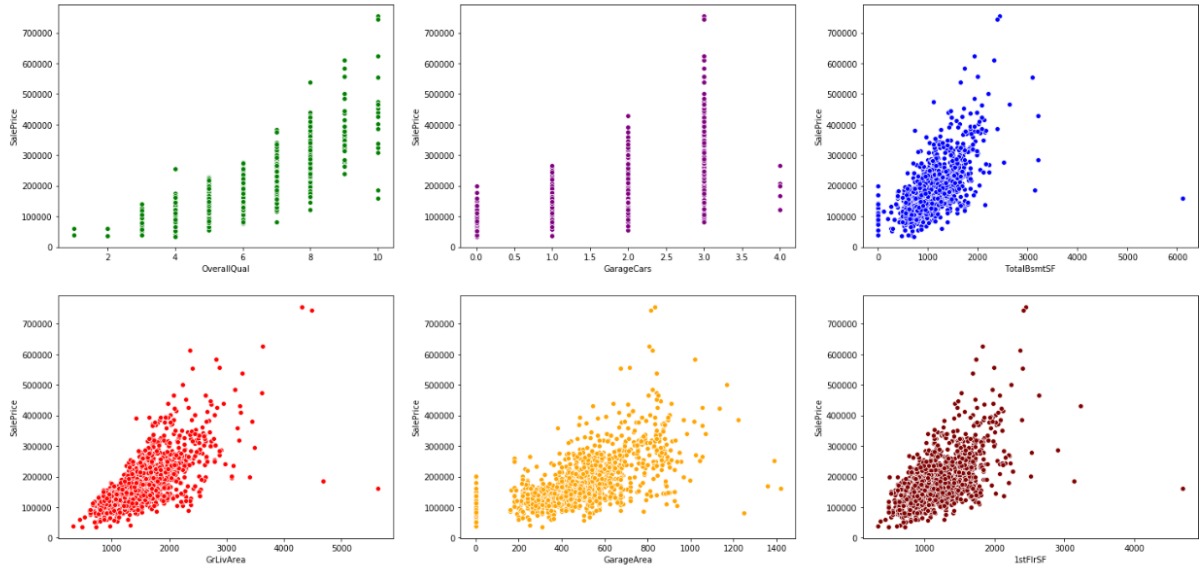

Figure 2 – Top Correlated Predictors of Sales Price

III. 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.

IV. Review results, evaluate models

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.

V. Implementation and programming

RMSE scores aside, the Kaggle prediction scores from unscaled data yielded the best scores. This was evident with both the Random Forest and Gradient Boosting models. An opportunity in this area would be to determine why the Kaggle performance scores were lower for our top models when the data was scaled compared to when the data was not scaled.

VI. 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.