MET CS 677 Term Project:

House Sales predict in King County, USA

Hao Wu

11/29/2021

**1.Description**

*1.1 Research Scenario Description*

The price of a house is often an important part of human society and economy, and the price of a house is usually affected by the size of the house, the number of rooms, and various parameters of the house. In this case, we often want to buy a house at a lower price, but different room types will have different prices. If we can summarize the main factors affecting house prices and create models to predict house prices, we can Get a home at a reasonable price.

*1.2 Data Set Description*

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

id – The unique identification of house

date – The date of the house sold

price – The price of the house sold

bedrooms – The number of the house’s bedrooms

bathrooms – The number of the house’s bathrooms

sqft\_living – The square-feet of the house’s living room

sqft\_lot – The square-feet of the house’s lot

floors – How many floors are the house

view – The grade of the house’s environment

condition – The house is good condition or bad

grade – The house’s grade

sqft\_above - The square-feet of the houses above floor

sqft\_basement The square-feet of the house’s basement

yr\_build – Year of the house build

yr\_renovated – Year of the house renovated

zipcode – Where is the house

**2.Visualization**

*2.1 Correlation*

图表, 树状图

描述已自动生成

Figure 1.1 Correlation Heatmap of Housing Sales

We need to focus on the second line-price correlation, because this research is to select the decisive features that affect housing prices.

We rank the correlation from top to bottom as sqft\_living, grade, sqft\_above, bathrooms, view, sqft\_basement, bedrooms. This is the feature we selected through **Figure 1.1**.

*2.2 Boxplot*

*图表, 箱线图

描述已自动生成*

Figure 1.2 Boxplot between grade and price

图表, 箱线图

描述已自动生成

Figure 1.3 Boxplot between view and price

图表, 箱线图

描述已自动生成

Figure 1.4 Boxplot between bedrooms and price

From Figure 1.2, Grade and price are positively correlated. As the grade increasing, the price keeps increasing; therefore, grade should be an important feature when we are modeling.

From Figure 1.3, as view increasing, the price of the house still increasing although the slope small.

Figure 1.4 is a classic left skew distribution which was important in regression as a feature.

*2.3 Scatterplot*

图表, 散点图

描述已自动生成

Figure 1.5 Scatterplot between price and sqft\_living

图表, 散点图

描述已自动生成

Figure 1.6 Scatterplot between price and sqft\_basement

Through Figure 1.5 and 1.6, sqft\_basement has many outlier, it will significant impact the regression result.

**3. Model Evaluation**

*Linear Regression. Decision Tree. Random Forest*

文本

描述已自动生成 *文本

描述已自动生成* *文本

描述已自动生成*

Through the figure, Decision Tree model has a good perform in training set; however, when it applies on testing set, the score is decreasing. It is meaning the model overfitting.

图表, 条形图

描述已自动生成

Compare with other model’s r2\_score, we can recognize Random Forest is the best model in those three models.

图表, 条形图

描述已自动生成

After we omitted each feature and do regression, we find out the grade feature was the most import feature when we do regression. If we omitted grade feature, we will get the lowest r2\_score.

**4.Conclusion**

The housing price may change with other factors, but if the real estate developer has rated the property, then this rating is credible, because after model selection, when we use the optimal model to filter the features, we find that Grade is always the most important factor affecting house prices.