**IA0**

Yongsung Cho

Computer Science, Oregon State University

AI 534: Machine Learning

Prof. Fern

Due Oct 1, 2021

**Part 1**

Please check readme.txt file.

**Part 2**

Before removing id feature

Table

Description automatically generated

After removing id feature

Table

Description automatically generated

I’ve used pd.drop() function. It removed id feature from dataframe.

**Question: Is it a good idea to use this feature in predicting the price of the house? why?**

I think using the id feature in predicting the price of the house is essential. So, I shouldn’t remove the id feature from this table (data frame). Because if I don’t know the id, I can’t figure out which one is the correct one. In addition, the duplicated row will be made because the id feature works as an identifier. If we are going to use database, the id feature is going to be key attribute. Therefore, the id feature is essential.

1. table[['month','day','year']] = table['date'].str.split("/",expand=True)

table = table[['id','month','day','year','bedrooms','bathrooms','sqft\_living','sqft\_lot','floors','waterfront','view','condition','grade','sqft\_above','sqft\_basement','yr\_built','yr\_renovated','zipcode','lat','long','sqft\_living15','sqft\_lot15','price']]

First of all, I split date feature to month, day, and year. When I run first line of code, new features were on the last part of the table. So, I changed the order of data frame.

Table

Description automatically generated

This is a result data frame. Finally, I can compare the date of contract easily.

**Question: date feature is useful for this problem? Can you think of better ways of using this date feature than splitting them into three numerical features?**

I think date feature is useful for this problem. However, previous data feature was inconvenient for compare and calculate. So, splitting date feature is going to be good way to predicting the price of the house.

In the three numerical features, I think that the day feature is useless. Because the most important thing for sale is when it is sold. On the other hand, year and month is more important than day. I think that the price of the house isn’t changed by the day of the date.

1. Because of using matplotlib, I can see the plot at a glance and easily understand graphs.

Box plot grouped by bedrooms

Chart, box and whisker chart

Description automatically generated

This plot shows that the house with many bedrooms is usually expensive.

Box plot grouped by bathrooms

Chart, bar chart

Description automatically generated

Most of the house have 1 to 4 bathrooms and the house with many bathrooms is usually expensive, like box plot grouped by bedrooms plot.

Box plot grouped by floors

Chart, box and whisker chart

Description automatically generated

The mean of the house with 2.5 floors is highest, but there are not many houses were sold. Most of the house that was sold have 1 or 2 floors.

1. Scatter plot sqft\_living against sqft\_living15 / Scatter plot sqft\_lot against sqft\_lot15

Shape

Description automatically generatedChart

Description automatically generated

I think this scatter plot doesn’t have any meaning. Because co-variance matrix of sqft\_living against sqft\_living15 has only 4 numbers and it was too big or too small number for showing plot. Sqft\_lot against sqft\_lot15 co-variance matrix also has same result. Therefore, these features are redundant.