HOUSE PRICE PREDICTION

Assignment

upGrad

Problem Statement

i) How price of the house is dependent on various features
ii) The top features that influences the price of the house.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy to enter the market.

2:17 PM Sat 19 Dec 64%

Assignment

upGrad

Steps to proceed with the assignment.

1. Perform EDA to understand the data.
2. Check missing values.
3. For continuous variables, try to impute the missing values with mean or median. Perform EDA to find out which one fits best.
4. For categorical columns.
   a. Check all categories carefully. → For those categorical columns, where you have meaningful missings, you should not use mode approach over here.

Cx
A 90%
B 5%
C 3%
D 2%
} Skewed categorical columns

Basement quality
1
2
3
→ NA: "No Basement"

→ Categorical columns that are highly skewed

Activate Windows
Go to PC settings to activate Windows.

10

Try to drop skewed categorical models.

How to solve..Continue.

upGrad

What to do?

5. Check if the target variable is normally distributed or not?
6. Create dummies for categorical data.
   - You can create groups of the the categories to reduce the number of categories and then create dummies.
   - This is an optional method.
7. Handling year columns.
   - There are 4 columns that contain year. What to do with them?
   - How to convert them?

Activate Windows
Go to PC settings to activate Windows.

11

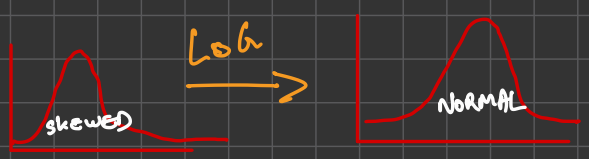
Linear Regression cannot extrapolate
- if $100 < y < 1000$
  then it will always predict between $100 < P < 1000$
- Should not have outliers
  → capping
  → Drop outliers
  ] not acceptable for assignment

→ Model wont work for range of data that was considered as outliers

Not removing outliers will influence the best fit line.

How to handle outliers?

1- Do Log transformation for skewice



Categorical Variable

- Group categories with lesser % of data into others
- combine these categories to 'Rare' category

YEAR COLUMNS

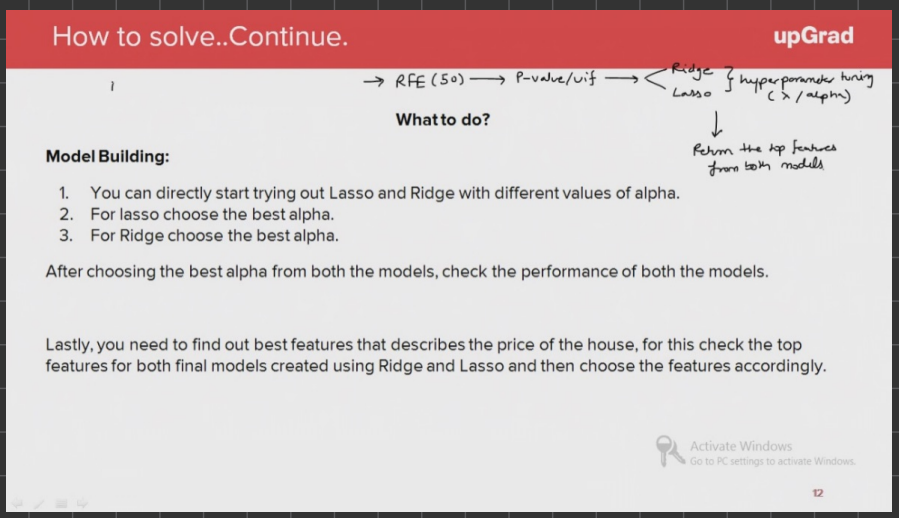
1- calculate age from Yearbuilt column

2- use as continous variable

3- drop other Year features

MODELLING

1- Start with RFE(50)
2- P-value/VIF
3- Ridge/Lasso] Tune hyperparameter alpha
4- Return top features



Lasso → Top 5 features
↓
Remove from X_TRAIN
↓
RETRAIN MODEL
↓
MEASURE PERFORMANCE

SUBJECTIVE QUESTIONS

- use code to answer in Jupyter

JUPYTER NOTEBOOK
PDF (SUBJECTIVE ANSWERS)

2 if

3:05 PM Sat 19 Dec

Quiz Time — upGrad

Poll Questions

Question-1: Suppose we have a regularized linear regression model: argmin$||Y-\beta x||^2+\lambda||\beta||$. What is the effect of increasing λ on bias and variance?

- A. Increases bias, increases variance
- B. Increases bias, decreases variance
- C. Decreases bias, increases variance
- D. Decreases bias, decreases variance
- E. Not enough information to tell

Question-2: After applying a regularization penalty in linear regression you find that some of the coefficients are zeroed out. Which of the following penalties might have been used?

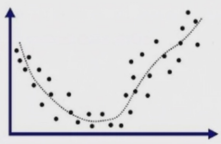
- A. L0 norm
- B. L1 norm
- C. L2 norm
- D. either (A) or (B)
- E. any of the above

3:09 PM Sat 19 Dec

Quiz Time — upGrad

Poll Questions

Question-3: Suppose you used a degree-3 polynomial to fit the data and it look like as shown in the image, what will happen if we use a degree-4 polynomial?



- A. There are high chances that degree 4 polynomial will over fit the data
- B. There are high chances that degree 4 polynomial will under fit the data
- C. Can't say
- D. None of these

Question-4: Which of the following is true for the given fit(refer to the image)?

- A. Bias will be high, variance will be high
- B. Bias will be low, variance will be high
- C. Bias will be high, variance will be low
- D. Bias will be low, variance will be low

3:13 PM Sat 19 Dec

Quiz Time — upGrad

Poll Questions

Question-5: Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

- A. 1 and 2
- B. 2 and 3
- C. 1 and 3
- D. 1, 2 and 3

USED FOR AVOIDING OVER-FITTING

Question-6: If the fitted model is underfitting then which of following regularization algorithm would you prefer?

- A. L1
- B. L2
- C. Any
- D. None of these

Dont need to - SC

3:24 PM Sat 19 Dec

Quiz Time — upGrad

Poll Questions

Train Test Train Test

Question-7: Let's say, a "Linear regression" model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?

- A. You will always have test error zero
- B. You can not have test error zero
- C. Can't say

$\rightarrow \sum_{i=1}^{n} (y_i - \hat{y})^2 + \lambda\beta_1 + \lambda\beta_2 + \lambda\beta_3 + \cdots \lambda\beta_{100}$

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{100} x_{100}$

Small value 1L 0.1L

Question-8: Suppose we fit "Lasso Regression" to a data set, which has 100 features (X1,X2...X100). Now, we rescale one of these feature by multiplying with 10 (say that feature is X1), and then refit Lasso regression with the same regularization parameter. Now, which of the following option will be correct?

- A. It is more likely for X1 to be excluded from the model
- B. It is more likely for X1 to be included in the model
- C. Can't say
- D. None of these

scale of β_1 depends on scale of X

X increases
β_1 decreases
hence model will not penalize more

Quiz Time

upGrad

Poll Questions

Question-1: Suppose we have a regularized linear regression model: $\text{argmin}||Y-\beta x||^2+\lambda||\beta||$. What is the effect of increasing λ on bias and variance?

- A Increases bias, increases variance
- B Increases bias, decreases variance
- C Decreases bias, increases variance
- D Decreases bias, decreases variance
- E Not enough information to tell

Question-2: After applying a regularization penalty in linear regression you find that some of the coefficients are zeroed out. Which of the following penalties might have been used?

- A L0 norm
- B L1 norm
- C L2 norm
- D either (A) or (B)
- E any of the above