Scenario Questions

Ans 1

Let’s first define what Multi-col-linearity is? And how does it affect our model?

Multi-col-linearity is a phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. That means, the predictor variables will have a high correlation. Thus, it would it becomes difficult to determine the individual effects of each independent variable(predictor variable) on the dependent variable(response variable) accurately.

When multi-col-linearity is present, the estimated regression Coefficients become ***large and unpredictable***, leading to unreliable inferences about the effects of the predictor variables on the response variable, That means we might have a ***high variance***.

In our case there is a high correlation between the size of the house and its age. So, to eliminate this dependency between this features we use ***two methods.***

1. ***Regularization***
   - Using ridge, lasso or Elastic net regression we can reduces model variance by adding a penalty term to counteract the high variability caused by correlated predictors.
2. ***Feature engineering***
   - In this way, we change the size_of_house to other derived feature like number_of_rooms and see how the new derived feature would be correlated with age. If it doesn’t, the problem is solved.

Ans 2

let’s first look at the equation of each performance measures

$$\text{R- squared} = 1 - \frac{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2}{\sum_{i=1}^{n} \left(y^{(i)} - \bar{y}^{(i)} \right)^2}$$

R-squared tells us that how much less variation is captured around this line than the mean. R-squared is .85 means that we have 85% less variation around the line than the mean would have. Which is good on its own but we need further comprehensive understanding of model performance to quantify how much the average variation the model delivers.

$$MSE(x,h) = \frac{1}{n_i} \sum_{i=1}^{n} \left(h(x^{(i)}) - y^{(i)} \right)^2$$

MSE measures the average of the squares of the errors. MSE gives ***higher weight to large errors*** due to squaring, which can be useful if you want to penalize large errors more, But on the flip side, since MSE is squared, its units are not the same as the original target variable, making ***it harder to interpret directly.***

$$RMSE(x,h) = \sqrt{\frac{1}{n_i} \sum_{i=1}^{n} \left(h(x^{(i)}) - y^{(i)} \right)^2}$$

RMSE is the square root of the mean squared error, providing a measure of the magnitude of the error.

RMSE is in the same units as the original target variable(Label), which makes it easier to interpret the magnitude of the error but it treats all errors equally, so it may not penalize large errors as much as MSE.

Having defining what each performance measures are capable of, during choosing of the performance measure, ***we should consider the following points:***

- R-squared is not enough to measure the performance. It only tells us that how much less the variation gets relative to the mean. It doesn’t tell us the exact variation.
- For interpret-ability RMSE is better because it has the same unit with the label.
- If we want to be sensitive to outliers, We ought to use MSE

After considering those points, we select one or combination of two to the table.

*** Read and summarize Lesson 3

In lesson 3, which is Classification, the following points are covered.

- Classification is a supervised learning algorithm.
- Types of classification
    - Binary classifier
    - Multi-class(or multi-nomial) classifier
- Classification process
- Data processing for classification
- Logistic regression
- Gradient Descent
- Model evaluation matrix(example: MSE and Cross-entropy loss)
- Accuracy, Precision, Recall and F1 score
- Confusion Matrix