Subjective Questions

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:
i) The optimal value of Alpha for ridge and lasso regression is given below:

	Ridge	**Lasso**
Alpha	5	0.0001

ii) If the value for alpha is doubled then our model metrics will change and in general become worse.
Changes in metrics for Ridge and Lasso regression when optimal value for alpha is changed are shown below.

Ridge Regression:

Metric	Optimal Alpha	Double Optimal Alpha
Alpha	**5**	**10**
R2 Score (Train)	0.946	0.936
R2 Score (Test)	0.911	0.911
RSS (Train)	1.022	1.216
RSS (Test)	0.671	0.674
RMSE (Train)	0.034	0.037
RMSE (Test)	0.042	0.042

Observations:
When alpha is doubled for Ridge regression we see following changes:

- R-squared for training set is slightly reduced
- Residual sum of squares for the training set has increased by about 18% so the model fit has become worse.
- Root mean squared error has increased for both test and training set which mean the models ability to predict has worsened.
- Model performance has become slightly worse in both training and test sets.

Lasso Regression:

Metric	Optimal Alpha	Double Optimal Alpha
Alpha	**0.0001**	**0.0002**
R2 Score (Train)	0.944	0.932
R2 Score (Test)	0.915	0.917
RSS (Train)	1.052	1.291
RSS (Test)	0.643	0.623
RMSE (Train)	0.034	0.038
RMSE (Test)	0.041	0.040

Observations:
When the value for alpha is doubled for Lasso regression we see following changes:

- R-squared for the training set is slightly reduced and for the test set it is slightly increased.
- Residual sum of squares for the training set has increased by about 23% so the model fit has become worse but decreased slightly for the test set.
- Root mean squared error has increased slightly for the test but decreased slightly for the training set.
- In summary, the model now fits slightly better to test and slightly worse to the training dataset.

iii) After this change the most important predictors variables are shown below:

Ridge Regression (alpha = 10)

Feature	Coefficient
OverallQual	0.071480
TotalArea	0.066400
OverallCond	0.062895
GrLivArea	0.062853
1stFlrSF	0.059179
BsmtFinSF1	0.041259
TotalBsmtSF	0.040604
GarageArea	0.039016
LotArea	0.032335
Neighborhood_StoneBr	0.031810

Lasso Regression (alpha = 0.0002)

Feature	Coefficient
TotalArea	0.397626
OverallQual	0.172024
OverallCond	0.131049
PropertyAge	-0.087930
SaleCondition_Partial	0.046233
TotalBath	0.044495
LotArea	0.037514
Neighborhood_Crawfor	0.037507
MSZoning_FV	0.036177
BsmtFinSF1	0.031743

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:
The optimal value of lambda for ridge and lasso regression is found to be 5 and 0.0001 respectively by using cross validation techniques.

One of the way to decide which model to use is by looking at and comparing the model metrics for ridge and regression:

Metric	Ridge Regression	Lasso Regression
R2 Score(Train)	0.946	0.944
R2 Score(Test)	0.911	0.915
RSS(Train)	1.022	1.052
RSS(Test)	0.671	0.643
MSE(Train)	0.001	0.001
MSE(Test)	0.002	0.002
RMSE(Train)	0.034	0.034
RMSE(Test)	0.042	0.041

Based on above metrics, it is evident that the Lasso regression performs slightly better than the Ridge regression.

Another important thing to consider when choosing between the models is the number of features used in the model. In the case of Lasso, features whose coefficients are reduced to zero are effectively removed from the model. In our model, **331** features have been removed by Lasso regression. This makes it easier to identify important features that predict the target variable from a business perspective.

Due to better metrics, and for being a simpler model with less number of features I would choose the **Lasso regression** for this task.

Question 3

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:
If the five most important predictor variables in the lasso model were not available in the incoming data out model would change as shown below:

<table>
<tr><th></th><th>Not-missing</th><th>Missing</th></tr>
<tr><td>Top 5 important features</td><td>
<table>
<tr><th>Feature</th><th>Coefficient</th></tr>
<tr><td>TotalArea</td><td>0.347536</td></tr>
<tr><td>OverallQual</td><td>0.153322</td></tr>
<tr><td>OverallCond</td><td>0.137154</td></tr>
<tr><td>PropertyAge</td><td>-0.102042</td></tr>
<tr><td>MSZoning_FV</td><td>0.066882</td></tr>
</table>
</td><td>
<table>
<tr><th>Feature</th><th>Coefficient</th></tr>
<tr><td>GrLivArea</td><td>0.288891</td></tr>
<tr><td>TotalBsmtSF</td><td>0.110742</td></tr>
<tr><td>GarageArea</td><td>0.067936</td></tr>
<tr><td>ExterQual_Fa</td><td>-0.065093</td></tr>
<tr><td>Functional_Maj2</td><td>-0.058814</td></tr>
</table>
</td></tr>
<tr><td>R-Squared</td><td>R2 Score(Train): 0.944
R2 Score(Test): 0.915</td><td>R2 Score(Train): 0.934
R2 Score(Test): 0.887</td></tr>
</table>

If the five most important predictor variables in the lasso model were not available in the incoming data “Above grade (ground) living area”, ”Total square feet of basement area”, ”Garage Area”, “Fair quality of material on exterior” and “Major Deductions 2 functionality” become the top 5 most important predictor features.

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

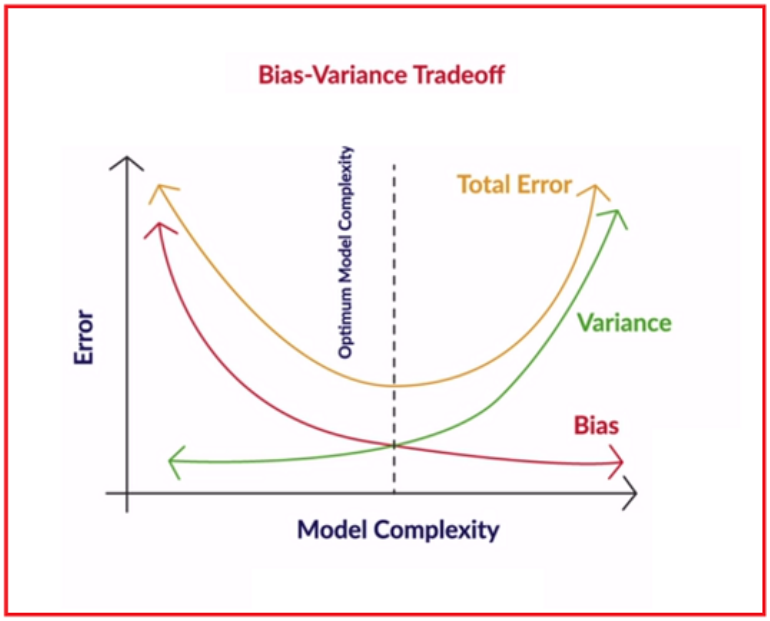
Answer:
It is essential to ensure that a model is robust and generalisable in order to be effective in practical use. Robust model is less sensitive to noise and small changes in the data and is able to capture the underlying patterns and relationships within the data without overfitting to the training data. A robust model is generalisable and stable hence its performance extends beyond the training data and it performs well also on new or unseen data.

Ultimately the robustness and generalisability of the model has to do with optimizing the complexity of the model. Some of the important concepts around managing complexity of the model include the following:

Occam's Razor
It is a philosophical and scientific principle that in the context of model building suggests that among two models with similar predictive performance a simpler model with fewer features or parameters should be favored. As the model becomes more complex, it might fit the training data very well but may not generalize effectively to unseen data. Therefore complexity in model building should be penalized.

Overfitting
Overfitting happens when a model learns the training data too well. It generally performs very well on the training data but is highly sensitive to fluctuations in the data and doesn't perform effectively on data not seen by the model during the training stage. Overfitted models are usually too complex and have low bias and high variance.

Bias-variance tradeoff
Bias and variance are two types of errors that affect the robustness and generalisability of the model. A model with very high bias is simpler but under fitted and is unable to capture important underlying patterns and relationships within the data. It leads to poor fit of the training data and hence is unable to perform well on real world scenarios.
A model with very high variance is too complex and has to memorize the training dataset. It is overfitted to the training data therefore it is highly sensitive to small fluctuations or noise in the data. It usually performs well on training sets but performs poorly on unseen or new data.
The goal is to find an equilibrium between bias and variance where the model complexity of the model is just right to capture important underlying patterns in the data while not fitting too close to the noise in the data.



Source: Upgrad Learning platform (learn.upgrad.com)

Model complexity

Complexity of the model increases as the number of features, size of parameters, degree of polynomial required to represent the model increases. Highly complex models are prone to overfitting hence complexity of the model must be penalized in order to maintain good balance of model bias and variance.

Regularization

Regularization is a technique to prevent overfitting of the model by introducing a penalty term that prevents the model from becoming overly complex. The two common types of regularization are Lasso and Ridge. By using regularization techniques we can ensure the model doesn't overfit with the training data. This is one of the important techniques used to build a robust and generalisable model.

Robust and generalisable models generally show a satisfactory accuracy in both training data as well as test data or unseen new dataset.

A model with high bias generally has low accuracy because it oversimplifies the data while a model with high variance has high training accuracy but low test accuracy because of its tendency to overfit.

A good model that makes use of optimisation techniques like regularization, feature selection and feature engineering to strike a balance between bias and variance tend to show reasonably good accuracy against both training and test datasets.