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
The optimal value of alpha or the regularization parameter for ridge and lasso regression depends on the particular data set and problem. Alpha is a compromise between fitting the training data well and avoiding overfitting.

In ridge regression, increasing alpha causes the coefficient estimates to shrink toward zero. Higher values of alpha lead to more regularization, resulting in a simpler model with smaller coefficient sizes. This helps to mitigate the influence of less important predictors and prevent overfitting.

Lasso regression, on the other hand, not only shrinks the coefficient estimates, but also performs variable selection by forcing some coefficients to be exactly zero. Increasing alpha in Lasso regression reinforces this aspect of variable selection and leads to sparser models with less important predictors.

If the value of alpha is doubled for both ridge and lasso regression, the following changes can be expected:

Ridge regression:

Increased regularization: higher alpha values lead to further regularization of the model, resulting in even smaller coefficient sizes and less dependence on individual predictor variables.
Smaller coefficient sizes: With increased regularization, the model assigns less weight to individual predictors, resulting in smaller coefficient sizes. The model emphasizes a combination of predictors rather than relying heavily on a particular predictor.

Lasso Regression:

Increased parsimony: doubling alpha in Lasso Regression increases the parsimony effect, forcing more coefficients to zero. This increases the number of excluded predictor variables and promotes model parsimony.
Reduced number of important predictors: the more coefficients are forced to zero, the simpler the model becomes and the fewer predictors are needed for prediction. The remaining predictors with non-zero coefficients are considered the most important in the updated model.

In ridge regression, after doubling the value of alpha:
The most important predictor variables are likely to be those that have a relatively stronger relationship with the outcome variable and are more resistant to increased regularization. These

important predictors generally have larger coefficient sizes than others, even after regularization. However, the differences in coefficient sizes can be reduced by increased regularization.

In lasso regression, after doubling alpha:
The most important predictor variables are those that are not excluded and have nonzero coefficients in the updated model. These predictors have strong relationships with the outcome variable and can be considered the most influential in making predictions. It is noteworthy that the number of important predictor variables decreases as more coefficients are forced to zero. The predictors remaining after regularization are the most important in the updated model.

As in our case the alpha increase for ridge and lasso has below effect

Optimal value of alpha for ridge: 50
Optimal value of alpha for lasso: 100
On doubling above mentioned behaviour was seen

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

Ridge regression:
Ridge regression is appropriate when there are a large number of predictors or when there is the possibility of multicollinearity (high correlation) between predictors.
It helps shrink the coefficient estimates toward zero without completely eliminating any predictor, so that all predictors contribute to the model to some degree. Larger lambda values in ridge regression increase the amount of regularization applied to the model. This leads to greater shrinkage of the coefficient estimates toward zero. As lambda increases, the size of the coefficient estimates decreases, resulting in a simpler model with smaller coefficients. The model is less sensitive to individual predictor variables, reducing the risk of overfitting. However, extremely large lambda values can cause the coefficients to become too small, potentially leading to an underfitting model that performs poorly on training and test data.

Lasso regression:
Lasso regression is beneficial when you suspect that only a subset of the predictors really contribute to the outcome and want to perform feature selection. It tends to set the coefficient estimates of irrelevant or less important predictors exactly to zero, effectively excluding them from the model. Lasso regression is appropriate when you have a large number of predictors and want to identify the most influential predictors for prediction or interpretability.

In lasso regression, the effect of lambda is more pronounced. Increasing the lambda value leads to stronger regularization and variable selection. As the lambda value increases, more

coefficients are set to exactly zero, resulting in sparser models with fewer predictors considered important. This facilitates feature selection because the less important predictors are effectively excluded from the model. Larger lambda values in Lasso regression prioritize simplicity and sparsity of the model over an accurate fit to the training data. However, if lambda is too large, too many predictors may be excluded, resulting in an inadequate fit to the training data

Lasso tends to do well if there are a small number of significant parameters and the others are close to zero (when only a few predictors actually influence the response). Ridge works well if there are many large parameters of about the same value (when most predictors impact the response).

We will choose Lasso as its giving feature selection option. It has features which are not required from model without affecting the model accuracy making model generalized simple and accurate.

Question 3

After building the model, you realised 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

After removing the variables identified as the top five predictor variables in the original Lasso model from data set create a new model. Use the updated dataset without the five excluded variables and create a new model. Assess the significance of the predictor variables. Examine the coefficients or feature importance measures of the new model to determine the importance of the predictor variables in the updated model. The variables with larger coefficients or higher feature importance values may be considered the most important in the new model.

Order the predictor variables based on their coefficient sizes or trait importance values. The five variables with the highest values can be considered the five most important predictor variables in the new model.

Other options would be to collect or impute missing data by including the missing predictors back into the model, we can recover their effect and improve the performance of the model.
Refine the feature selection by Re-evaluating the remaining predictors in the model and possibly re-running feature selection. Identifying the next set of important predictors among the available variables and building a new model based on this updated selection.
Considering alternative models if the absence of key predictors severely affects the performance of the model, we can explore alternative modeling techniques that can handle missing data or better adapt to the new circumstances. For example, we might consider random forest, gradient boosting, or other methods that are more robust to missing predictors.

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

To ensure that a model is robust and generalizable, we can use the following best practices:

Sufficient and representative data: make sure data set is large enough and representative of the population to which we want to generalize. A small or biased data set can lead to overly optimistic or limited model performance. Collecting diverse and comprehensive data helps capture the underlying patterns and variability in the problem domain.

Train-Test Split: Split data into separate training and testing sets. The training data set is used to train the model, while the testing data set is reserved for evaluating the model's performance on unseen data. A common split is 70-30 or 80-20, but the specific ratio may vary depending on the size of the data set.

Cross-validation: apply cross-validation techniques, such as k-fold cross-validation, to obtain more reliable performance estimates. Cross-validation involves dividing the data into multiple subsets, training the model on a combination of these subsets, and evaluating its performance on the remaining subset. This allows the ability of the model to generalize to different subsets of data to be assessed and reduces the impact of random variation in the data.

Feature selection: Perform feature selection to identify the most important predictors for model. Informative feature selection helps reduce noise, improve interpretability, and focus the model on the most important variables. Various techniques such as correlation analysis, forward/backward selection, or regularization can help in selecting robust features.

Regularization: apply regularization techniques, such as ridge regression or lasso regression, to prevent overfitting and improve model generalization. Regularization adds a penalty term to the model's objective function that prevents complex or large coefficient estimates. This helps reduce the model's sensitivity to noisy or irrelevant predictors, leading to improved generalization performance.

Hyperparameter tuning: Optimize the model's hyperparameters using techniques such as grid search, random search, or Bayesian optimization. Hyperparameters control the behavior of the model, and finding the optimal values can significantly affect its robustness and generalizability. Experiment with different hyperparameter settings and select the ones that perform best on cross-validation or validation data.

Avoid data leaks: ensure that there are no data leaks during model training or preprocessing. Data leaks occur when information from the test set unintentionally affects the model during

training, resulting in overly optimistic performance estimates. Strictly separate training and test data throughout the modeling process.

Evaluation metrics: Use appropriate evaluation metrics based on the problem at hand. Accuracy alone may not be sufficient, especially for imbalanced data sets or in specific domains. Consider metrics such as precision, recall, F1 score, area under the ROC curve (AUC-ROC), or mean square error (MSE) depending on the nature of the problem and desired model performance.

External validation: if possible, validate the model against external or independent data sets to assess its generalizability to different data sources. This will help assess whether the model can perform well beyond the specific data set used for training and testing.

Model monitoring and maintenance: continuously monitor and evaluate the performance of the model over time. As new data become available, periodically re-train and re-evaluate the model to ensure that it remains robust and generalizable. Models can degrade or lose relevance as new trends or patterns emerge, so regular monitoring and updating is essential.

Ensuring that a model is robust and generalizable has important implications for its accuracy and overall performance. Here are the most important implications:

Improved accuracy on unseen data: A robust and generalizable model is designed to perform well on new, unseen data. By incorporating techniques such as training-test splitting and cross-validation, the model is better able to handle data it has not encountered during training. This reduces the risk of overfitting and allows the model to make accurate predictions for real cases.
Reduced overfitting: overfitting occurs when a model learns the specific details and noise in the training data to the point that it performs poorly on new data. By using regularization techniques such as ridge regression or lasso regression, a robust model reduces the complexity and variance of the model, mitigating the risk of overfitting. This improves the accuracy of the model by preventing it from learning incorrect patterns or noise in the training data.
Better generalization: a robust and generalizable model is able to capture the underlying patterns and relationships in the data that apply to different data sets. It learns to extract meaningful features and generalize them to unknown instances. This ability to generalize leads to more accurate predictions for new data, even when the distribution or features of the data differ from the training set.
Stability of performance: a model that is robust and generalizable tends to show stable performance across different data sets or subsets of the data. It is less sensitive to small variations in the data, resulting in consistent and reliable predictions. This stability increases the accuracy of the model by reducing the impact of random variations or characteristics specific to a given data set.
Increased reliability and trustworthiness: a robust and generalizable model inspires confidence in its predictions. It shows consistent performance across different scenarios and datasets, making it more reliable for decision making. Stakeholders can rely on the accuracy of the model's predictions, leading to greater adoption and usage.

Finally we are trying for model accuracy between 70-75%, P-value of all the features is < 0.05 and VIF of all the features are < 5