ISEN 613

HONEWORK 5

Q1)

I performed Linear Regression, KNN Regression and Decision Tress on the Snails data set. All the models predict *log (Rings)* with different predictors for each model. The Validation set, LOOCV and K-fold cross validation R^2 values are higher for Linear Regression model (results are summarized in the table below). Also, the RMSE and MAE values are low for linear regression model. Therefore, the best model that I would recommend is the Linear Regression model (with appropriate transformations on the input parameters).

Q2)

I performed Linear Regression, KNN Regression and Decision Tress on the Snails data set. After building the models, all the cross-validation methods are performed on each model to check how the model performs on the training data (new data). The results are summarized in the below table. R^2 is the metric used. Seed is set to 829 (set.seed(829)) for all the models. All the three validation methods give the same results: Linear Regression> KNN Regression> Decision Trees (By comparing cross validated R^2 values). Therefore, we can conclude that Linear Regression predicts best compared to other two models as the K-Fold cross validated R^2 is high. I would recommend Linear Regression model (with appropriate transformations on the input parameters).

R^2 Results:

Sl.No	Model	Validation Set	LOOCV	K-Fold
1	Linear Regression	0.6393605	0.6413814	0.6434824
2	KNN Regression	0.6231237	0.6362332	0.6355098
3	Trees	0.5064469		0.5177638

Q3)

In the data set given, a predictor is categorical variable (Type: M, F, I). One-Hot encoding is used to convert this predictor to numerical one.

Linear Regression (Recommended Model):

First, I performed Linear Regression using all the predictors given in the data set. It is seen that few predictors are not significant in determining the Rings (by observing p-values). To find the significant predictors, subset selection method is performed. All the metrics BIC, Cp, adjusted R2, or AIC indicate the best model contain 7 predictors and gave same set of predictors. After obtaining these 7 predictors, for finding the relationship between Rings and predictors, graphs are plotted. Based on the plots, appropriate transformations are performed on predictors and Rings are transformed to *log (Rings)*. Linear regression model is fitted and the obtained R^2 value is 0.6461. This model is validated using the cross-validation methods (Validation set, LOOCV, K-fold). The results are shown in table above. The R^2 value obtained after performing K-Fold cross validation is 0.6434824. The *log (Rings)* increases with increase in *sqrt (Height), log (WholeWeight)* and *log (ShellWeight),* decrease with increase in *TypeI, exp (log (Diameter)), log (ShuckedWeight)* and *log((Viscera Weight))*. All the predictors p-values are less than 0.05, hence all are

very much useful in predicting the Age of the snails. The F-statistic is very high and residual standard error is also low. Around 64% of variability of the data is explained by this model.

KNN Regression Model:

Using the predictors obtained by subset selection (except type) and transformations, KNN model is built. Inclusion of Type predictor doesn't have much significance on R^2 value. KNN model also predicts *log (Rings)*. To find the best *k* value, three cross validation methods are performed. Validation set approach gives k=14 (using seed 829). Both K-fold and LOOCV gives k=19. Validation set approach gives different k values for different seeds, there is randomness in the process. Hence, k=19 is selected based on results obtained from LOOCV and K-fold cross validation methods. Final KNN model: *exp (log (Diameter)), (sqrt(Height)), log(WholeWeight), log(ShuckedWeight), log((VisceraWeight)), log(ShellWeight)* as predictors with k=19. The R^2 value for this model is 0.6712178. To validate the model all the three cross validation methods are performed. The results are shown in table above. To know how the model performs on unseen data or new data, cross validation is performed. The R^2 value obtained after performing K-Fold cross validation is 0.6355098. We can say that this model explains 63% of variability.

Decision Trees:

Using the predictors obtained by subset selection, a tree model is built. This tree model uses only two predictors *"ShellWeight" and "ShuckedWeight"*. Applying log to these predictors and *log (Rings)* is predicted using trees. R2 value is computed and found to be 0.554279. Cross validation is used to find the most complex tree under (Pruning). The R2 value for this one is 0.5435613. Hence the normal R^2 value is high, we consider the normal model (unpruned tree model). Validation set approach and K-fold cross validation methods are used to validate the model built. The R^2 values obtained by these methods are shown in table above. We use the results of K-fold cross validation (Validation set approach has randomness in the process). The R^2 value obtained after performing K-Fold cross validation is 0.5177638. We can say that this model explains 51% of variability using the two predictors.