HOUSE PRICE PREDICTION USING CLASSIFICATION TECHNIQUES

Classification is a technique for classifying data into a set of categories. The principal purpose of a classification problem is to determine which category or class new data will belong to. Logistic regression is a classification algorithm that is used to predict a binary outcome (either the event occurs or does not occur) in accordance with a set of independent variables.

The dataset given for the analysis consists of 1460 observations and 51 variables which describe the overall condition of a house depending on various factors such as lot frontage, year built, sale price, number of rooms etc. The problem statement requires to classify the overall condition of a house as Poor, Good, Average based on the rating from 1 to 10. There are three outcomes to the dependent variables, therefore multinomial logistic regression can be applied on the data to classify the overall condition of a house. Multinomial logistic regression is an extension of binary logistic regression which enables the prediction of more than two classes of the target variable.

The multinomial logistic regressor is fit on the given dataset as follows:

- The dataset is loaded into R using read.csv () function. The libraries like dplyr, mice which are useful in building a model are loaded into R.
- The data set is checked for missing values. Missing values cause certain problems on the decisions taken on analyzing the data. However, excluding missing data from the evaluation causes loss of data and results in inaccurate conclusions. The ideal way to deal with missing values is to impute them with certain values which complete the dataset without disturbing the original structure.
- In this analysis, MICE (Multivariate Imputation by Chained Equations) imputation technique is used to replace the missing data. This feature automatically detects the columns with missing values and imputes them accordingly using methods such as predictive mean matching, logistic regression, depending on type of the variable.

- The dataset is divided into training and test sets on the overall condition variable, the model is trained on the training data and predictions are made on the test data.
- Multinom () function from the nnet package is used to build the model on training data. Using the relevel () function, one of the levels of the target variable(Average) is set as a baseline.
- The summary of the model describes about the coefficients for Poor and Good conditions in comparison to the baseline Average.
- The predictions of the model are calculated on the training data to check the efficiency of the model. In this analysis sensitivity of the confusion matrix is considered as a metric of evaluating the model's performance. Sensitivity is the percentage of positives found that are expected to be positive. The sensitivity of the model for all the three levels(Poor,Average,Good) is found to be 1 which implies the model predicted with 100% accuracy on the training data.
- The same model is applied to the test data and the results observed are as follows. The sensitivity of the class average given by the model is 96.9% which indicates that the model predicted the Average value accurately around 97% of the time. Like wise the model predicted the class good correctly 98% of the time.
- However the model did not classify the class poor accurately most of the times. The sensitivity value is 0.33 which means the model predicted the class poor precisely only 33% of the time.