House Price Regression 2 Report

Introdcution:

The objective of this analysis is to predict the sale price of houses based on various features using statistical learning methods. The dataset used for this analysis includes information on different attributes of houses, such as their size, location, amenities, and condition. The aim is to build a model that accurately predicts the sale price of houses, which can be valuable for real estate agents, homeowners, and potential buyers.

Methods:

The initial step involved handling missing values in the dataset. The missing values in the variables "LotFrontage," "Heating," and "CentralAir" were imputed using appropriate strategies such as mean imputation and mode imputation. Feature engineering was performed to create new variables, such as calculating the age of each house based on the "YearBuilt" variable. Categorical variables were converted into a numerical representation using one-hot encoding.

Dataset:

The dataset used in this analysis contains information on various attributes of houses, including the sale price, lot frontage, heating type, central air conditioning, building type, year built, and other relevant variables. The dataset also includes an identifier column for each house. The data was initially loaded from the "Stat_380_train.csv" file and then preprocessed to handle missing values, perform feature engineering, and convert categorical variables.

Analyze and Conclusion:

The Lasso regression model was trained using the "cv.glmnet" function from the "glmnet" package. Cross-validation was performed to select the best lambda value, which represents the level of regularization. The model was then fitted using the selected lambda value, and predictions were made on the training set. The root mean squared error (RMSE) was calculated to evaluate the performance of the model.

Similar to the Lasso model, the Ridge regression model was trained using the "cv.glmnet" function and cross-validation to select the optimal lambda value. The model was fitted using the selected lambda, and predictions were made on the training set. The RMSE was computed to assess the model's performance.

Then the apply_model was used to build features in the test set and then applied train model to it and created a submission file. The analysis aimed to predict the sale price of houses based on various features. Lasso and Ridge regression models were trained using the preprocessed training data. The models were evaluated using the root mean squared error (RMSE) metric and Ridge had the lower one so I chose Ridge for submission file.

Lasso RMSE: 25232.58, Ridge RMSE: 25181.48

This was similar to my 25309.73285 on the leaderboard.