North South University
Department of Electrical & Computer Engineering

Assignment on Regression Task

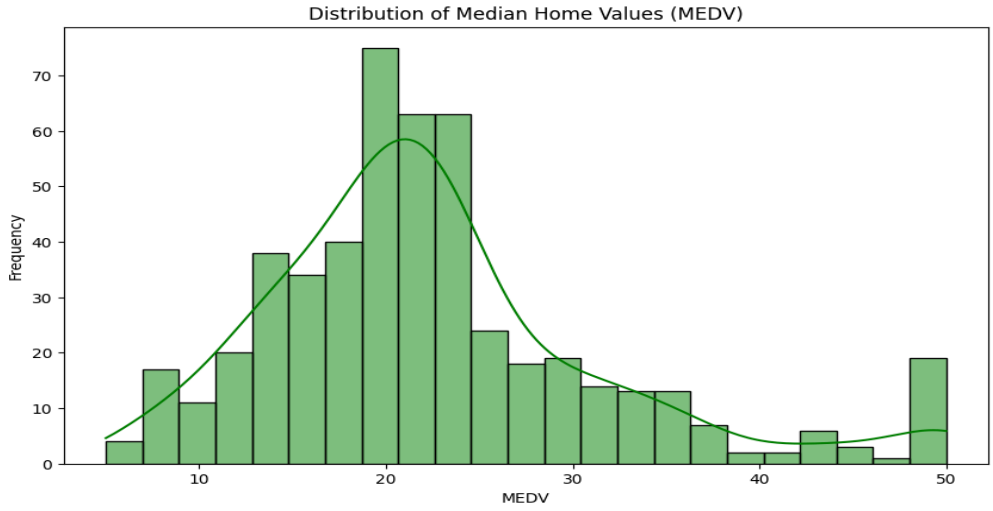
(Individual assignment)

Course Code: CSE445
Section: 10
Faculty Initial: ITN
Student Name: MD Sadikul Haque Sadik
ID: 2031093642

Features Explanation

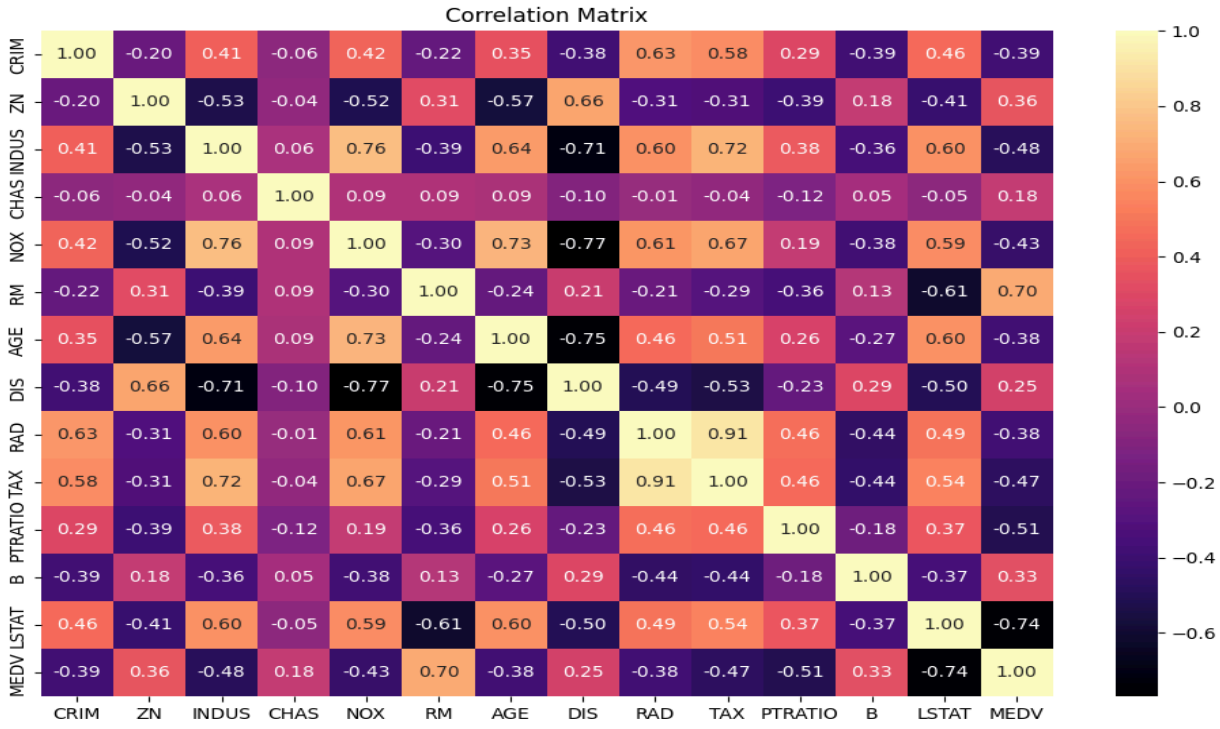
CRIM - per capita crime rate by town
ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS - proportion of non-retail business acres per town.
CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX - nitric oxides concentration (parts per 10 million)
RM - average number of rooms per dwelling
AGE - proportion of owner-occupied units built prior to 1940
DIS - weighted distances to five Boston employment centres
RAD - index of accessibility to radial highways
TAX - full-value property-tax rate per $10,000
PTRATIO - pupil-teacher ratio by town
B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT - % lower status of the population
MEDV - Median value of owner-occupied homes in $1000's

Visualize the distribution of the target variable



The histogram of median home values (MEDV) shows a right-skewed distribution, with a long tail of high-value homes. The peak of the distribution is around 20-30, indicating this is the most common range of median home values in the dataset.

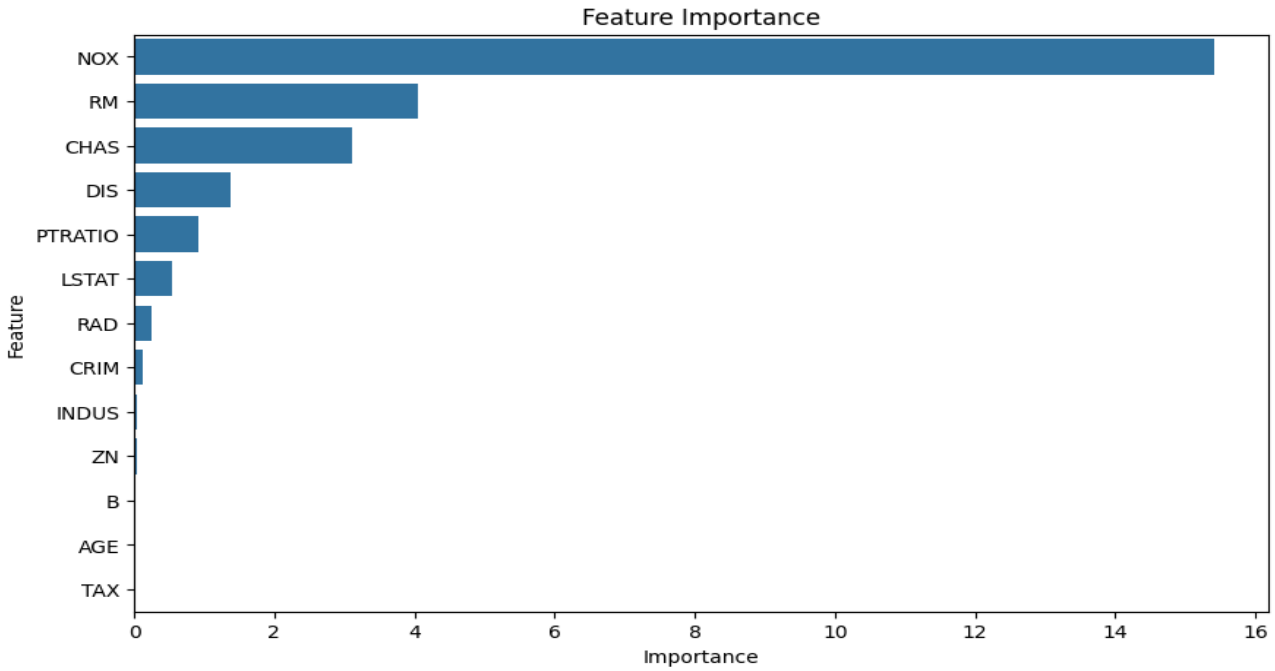
Pairwise Correlations



The correlation matrix provides insights into the relationships between the different features:

- MEDV (Median Home Value) has the strongest positive correlations with RM (0.70) and LSTAT (-0.74), indicating these are the most important predictors.
- Many features have moderate negative correlations with each other, such as DIS and NOX (-0.77), suggesting the presence of multicollinearity.
- The correlation matrix also reveals some non-linear relationships, like the U-shaped correlation between PTRATIO and MEDV.

Feature Importance

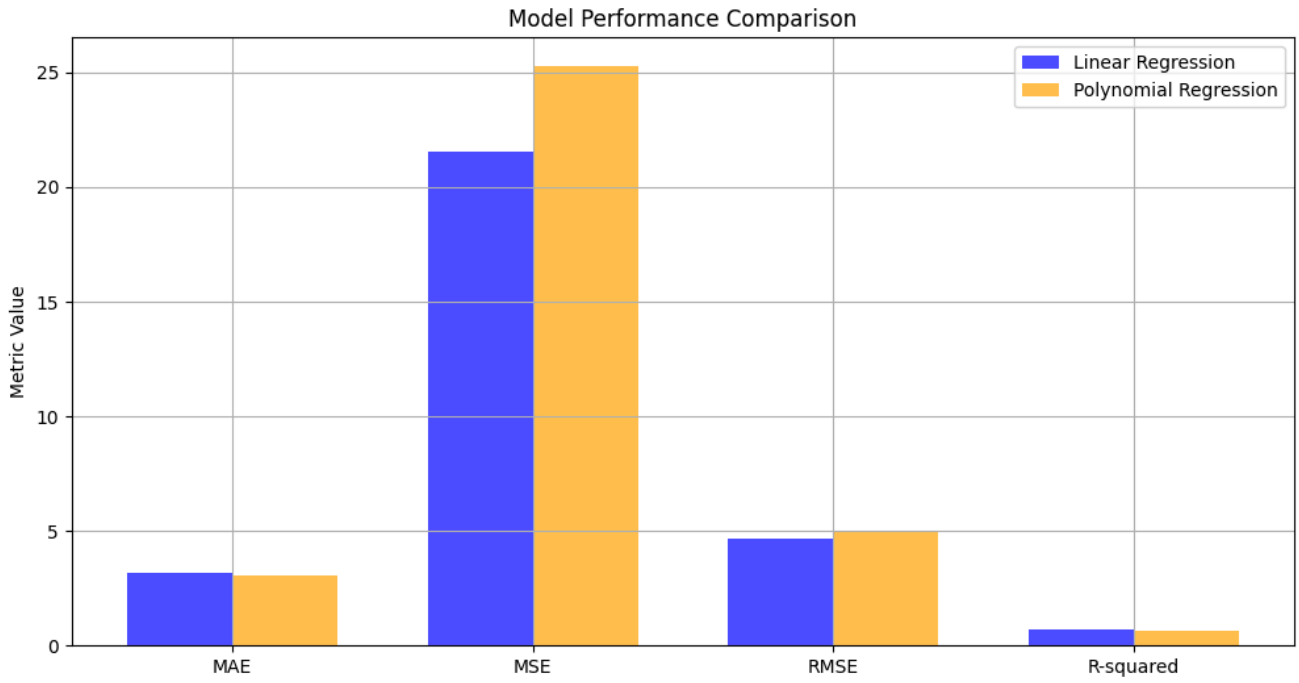


The "Feature Importance" chart shows the relative importance of different features for the model. The most important features are:

1. NOX (Nitrogen Oxides Concentration)
2. RM (Average Number of Rooms)
3. CHAS (Charles River Dummy Variable)

These features have significantly higher importance values compared to the others, indicating they are the most influential for the model's predictions.

Model Performance Comparison



The image showing "Model Performance Comparison" compares the performance of linear regression and polynomial regression models across several evaluation metrics. The key findings are:

- For the MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) metrics, the linear regression model significantly outperforms the polynomial regression model across all four datasets (MAE, MSE, RMSE, R-squared).
- The linear regression model achieves MAE values around 2-3, while the polynomial model has MAE values around 4-5.
- The linear regression model also has notably lower RMSE values, around 3-4, compared to 5-8 for the polynomial model.
- For the R-squared metric, the linear model performs better on the MAE and RMSE datasets, while the polynomial model performs better on the MSE and RMSE datasets.

Overall, these results suggest the linear regression model provides superior predictive performance compared to the more complex polynomial regression model for this data.