Boston House Price Prediction

Linear Regression

‘Femi Bolarinwa

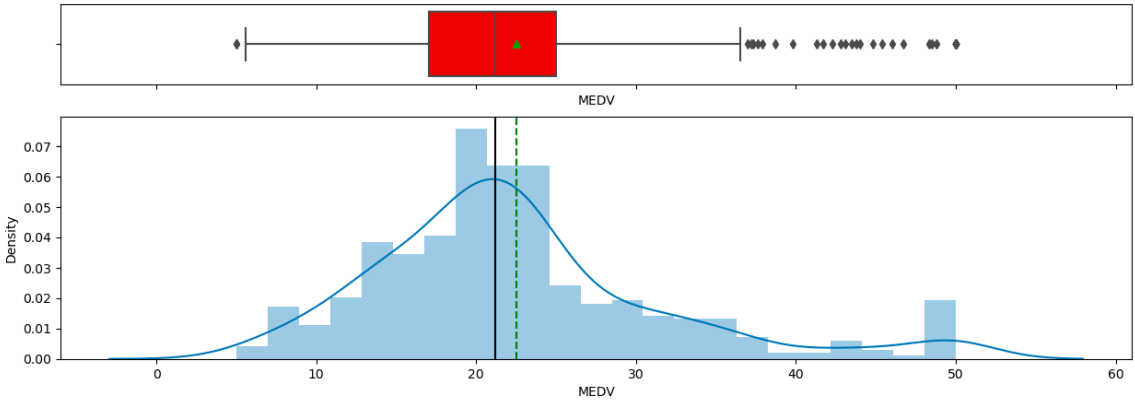
Data Snapshot

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
501	0.06263	0.0	11.93	0	0.573	6.593	69.1	2.4786	1	273	21.0	9.67	22.4
502	0.04527	0.0	11.93	0	0.573	6.120	76.7	2.2875	1	273	21.0	9.08	20.6
503	0.06076	0.0	11.93	0	0.573	6.976	91.0	2.1675	1	273	21.0	5.64	23.9
504	0.10959	0.0	11.93	0	0.573	6.794	89.3	2.3889	1	273	21.0	6.48	22.0
505	0.04741	0.0	11.93	0	0.573	6.030	80.8	2.5050	1	273	21.0	7.88	11.9

Target Variable

MEDV (Price)

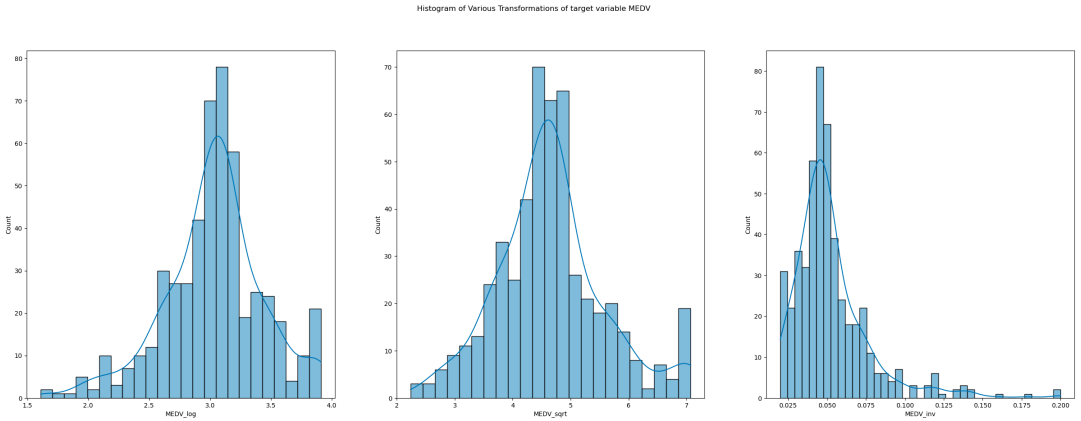
- Target variable is skewed
- Significant outliers
- Distribution non-normal. Transformation will be required for good modelling.



Target Variable

Possible Transformations

- Log and square-root transformations appear to be closer to normal distribution.
- Log transformation will be used for building my prediction model



Multi-collinearity of Variables

VIF

- One assumption of linear regression is that no multi-collinearity among predictor variables.
- Variance Inflation Factor (VIF) help to indicate multicollinearity in the data.
- Feature (TAX) having a VIF score > 10 has been dropped.

feature	VIF
const	532.025529
CRIM	1.923159
ZN	2.483399
INDUS	3.270983
CHAS	1.050708
NOX	4.361847
RM	1.857918
AGE	3.149005
DIS	4.333734
RAD	2.942862
PTRATIO	1.909750
LSTAT	2.860251

Model 1

Statistical Summary

- Three coefficients (ZN, INDUS, and AGE) have p-values greater than 5% (level of significance) which mean we fail to reject the null hypothesis that those coefficients are zero. They are not statistically significant enough or have enough predictive power to predict the target variable. I'll drop them and rebuild model.

- Decent R-squared and adjusted R-squared scores of 0.769 and 0.761 respectively. Meaning model captures ~77% of variance in target variable Log(Price)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.769
Model:	OLS	Adj. R-squared:	0.761
Method:	Least Squares	F-statistic:	103.3
Date:	Sun, 14 May 2023	Prob (F-statistic):	1.40e-101
Time:	14:39:58	Log-Likelihood:	76.596
No. Observations:	354	AIC:	-129.2
Df Residuals:	342	BIC:	-82.76
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6324	0.243	19.057	0.000	4.154	5.111
CRIM	-0.0128	0.002	-7.445	0.000	-0.016	-0.009
ZN	0.0010	0.001	1.425	0.155	-0.000	0.002
INDUS	-0.0004	0.003	-0.148	0.883	-0.006	0.005
CHAS	0.1196	0.039	3.082	0.002	0.043	0.196
NOX	-1.0598	0.187	-5.675	0.000	-1.427	-0.692
RM	0.0532	0.021	2.560	0.011	0.012	0.094
AGE	0.0003	0.001	0.461	0.645	-0.001	0.002
DIS	-0.0503	0.010	-4.894	0.000	-0.071	-0.030
RAD	0.0076	0.002	3.699	0.000	0.004	0.012
PTRATIO	-0.0452	0.007	-6.659	0.000	-0.059	-0.032
LSTAT	-0.0298	0.002	-12.134	0.000	-0.035	-0.025

Omnibus:	30.699	Durbin-Watson:	1.923
Prob(Omnibus):	0.000	Jarque-Bera (JB):	83.718
Skew:	0.372	Prob(JB):	6.62e-19
Kurtosis:	5.263	Cond. No.	2.09e+03

Model 2

Statistical Summary

- Adjusted R-squared score remains unaffected. Validates the previous assertion that the removed variables were not statistically significant for prediction.
- All remaining coefficients have p-values of less than 5%. This means they are statistically significant or have enough predictive power to predict the target variable.
- The true value of the coefficients of the remaining independent variables lie within the confidence interval shown with a 95% likelihood.
- Each ‘coef’ represents the change in the output Log(MEDV) due to a change of one unit in the variable (everything else held constant)
- ‘std err’ reflects the level of accuracy of the coefficients. The lower it is, the more accurate the coefficients are.

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.762
Method:	Least Squares	F-statistic:	142.1
Date:	Sun, 14 May 2023	Prob (F-statistic):	2.61e-104
Time:	14:41:03	Log-Likelihood:	75.486
No. Observations:	354	AIC:	-133.0
Df Residuals:	345	BIC:	-98.15
Df Model:	8		
Covariance Type:	nonrobust		

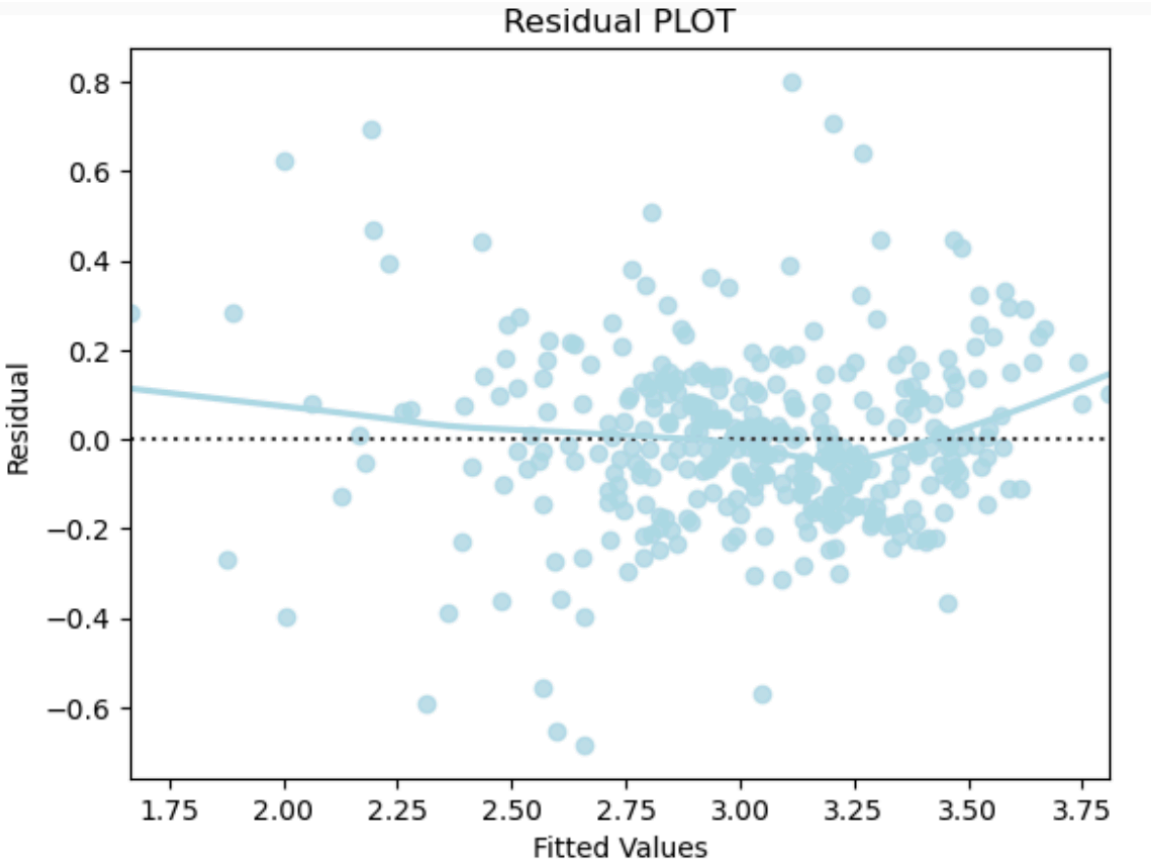
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6494	0.242	19.242	0.000	4.174	5.125
CRIM	-0.0125	0.002	-7.349	0.000	-0.016	-0.009
CHAS	0.1198	0.039	3.093	0.002	0.044	0.196
NOX	-1.0562	0.168	-6.296	0.000	-1.386	-0.726
RM	0.0589	0.020	2.928	0.004	0.019	0.098
DIS	-0.0441	0.008	-5.561	0.000	-0.060	-0.028
RAD	0.0078	0.002	3.890	0.000	0.004	0.012
PTRATIO	-0.0485	0.006	-7.832	0.000	-0.061	-0.036
LSTAT	-0.0293	0.002	-12.949	0.000	-0.034	-0.025

Omnibus:	32.514	Durbin-Watson:	1.925
Prob(Omnibus):	0.000	Jarque-Bera (JB):	87.354
Skew:	0.408	Prob(JB):	1.07e-19
Kurtosis:	5.293	Cond. No.	690.

Checking Linear Regression Assumptions

Linearity of Variables

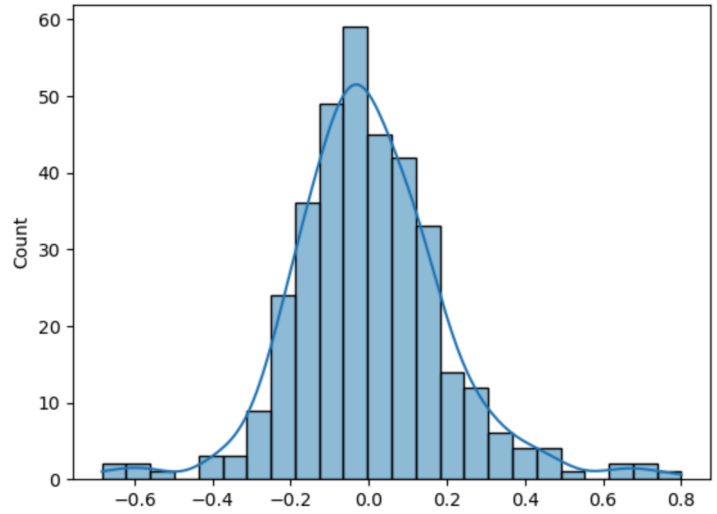
- Predictor variables must have a linear relation with the target variable.
- Plot of residual vs the fitted/ predicted shown.
- Residuals do not form a strong pattern. They are randomly and uniformly scattered on the x-axis.
- Assumption satisfied.

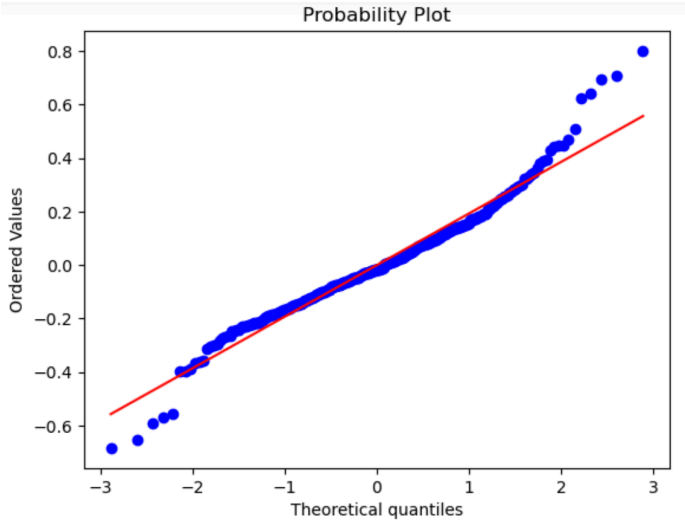


Checking Linear Regression Assumptions

Normality of Residuals

- Error terms/residuals should be normally distributed.
- If not, confidence intervals may become unreliable.
- Histogram and QQ-plot of residual shown. Residual appears normal.
- Assumption satisfied.





Checking Linear Regression Assumptions

Homoscedasticity

- Residuals must be symmetrically distributed across the regression line.
- Goldfeldquandt Test:
  - Null Hypothesis: Residuals are homoscedastic
  - Alternate Hypothesis: Residuals are heteroscedastic
  - Level of significance (alpha) = 5%
- Test Result: p-value = 30%
- Since p-value > alpha, I cannot reject the Null Hypothesis that the residuals are homoscedastic.
- Corresponding assumption is satisfied

```
In [249]: from statsmodels.stats.diagnostic import het_white
          from statsmodels.compat import lzip
          import statsmodels.stats.api as smsname = ["F statistic", "p-value"]
          test = sms.het_goldfeldquandt(y_train, X_train)
          lzip(name, test)

Out[249]: [('F statistic', 1.0835082923425283), ('p-value', 0.3019012006766869)]
```

Checking Linear Regression Assumptions

Zero Residual Mean

- Mean of residuals/error values must be zero.
- Practically zero mean of residuals obtained
- Assumption satisfied

```
In [247]: residuals = model2.resid
          np.mean(residuals)

Out[247]: -5.583731844753188 5e-15
```

Cross-validation

K-Fold

- Cross validating model2 across 10-fold of training dataset
- Result:
  - r2-score = 0.73
- Similar r2 score as previously obtained (0.77)
- Model appears to be good fit

In [254]:
```
# Import the required function

from sklearn.model_selection import cross_val_score
# Build the regression model and cross-validate
linearregression = LinearRegression()
cv_Score11 = cross_val_score(linearregression, X_train, y_train, cv = 10)
cv_Score12 = cross_val_score(linearregression, X_train, y_train, cv = 10, scoring = 'neg_mean_squared_error')
print("RSquared: %0.3f (+/- %0.3f)" % (cv_Score11.mean(), cv_Score11.std() * 2))
print("Mean Squared Error: %0.3f (+/- %0.3f)" % (-1*cv_Score12.mean(), cv_Score12.std() * 2))
```

```
RSquared: 0.729 (+/- 0.232)
Mean Squared Error: 0.041 (+/- 0.023)
```

Model Performance

Testing on unseen data

- Original data randomly split into training and test datasets at 70:30 ratio
- r2-score identical for both training and unseen test data (~0.77)
- Model correctly predicts 77% of the target variable (log(MEDV)) with mean absolute error of ~0.14, mean absolute percentage error of ~5% and root mean square error of ~0.2
- Model not overfitted.
- Model giving a generalized performance.

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.195504	0.143686	0.767174	0.761082	4.981813

Business Insight and Recommendation

- NOX has the most impact on house prices. A unit increase in Nitric Oxide concentration in a neighbourhood will cause about a unit decrease in the log of house price. Thats ten fold reduction in the house value. Developers, Home owners and city council need to work together to keep NO at lowest levels possible.
- Charles River (CHAS) has the second most impact on prices, albeit positive. The river needs to be preserved and protected from pollution.
- Intuitively, number of rooms (RM) and assess to highways (RAD) affect price positively.
- Houses closer to top five Boston employment centers (DIS) appear to cost more.
- Parents seem to prefer neighbourhood with better Pupil-to-teacher ratio. (PTRATIO).
- Crime rate (CRIM) affects prices negatively. Home-owners associations should work with the police to keep crime rate down