SUPERVISED MACHINE LEARNING: REGRESSION

COURSE PORJECT REPORT

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

ABOUT THE DATA

We will be working with a data set based on [housing prices in Ames, Iowa](https://www.kaggle.com/c/house-prices-advanced-regression-techniques). It was compiled for educational use to be a modernized and expanded alternative to the well-known Boston Housing dataset. This version of the data set has had some missing values filled for convenience.

Predictor

- SalePrice: The property's sale price in dollars.

Features

- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale
- MSSubClass: The building class
- MSZoning: The general zoning classification

OBJECTIVES FROM THE ANALYSIS

1. Train test split

2. Simple EDA

- Descriptive statistics and data cleaning
- Numerical features
- Categorical features

3. Model variations

- Apply One-hot encoding
- Apply nplog transformation
- Apply Standard scaling
- Add Polynomial features

4. Cross-validation and Regularization

- Linear Regression
- Lasso Regression (L1)
- Ridge Regression (L2)
- Elastic Net Regression (L1 + L2)
- Compare the metrics

5. Conclusion

METHODS USED

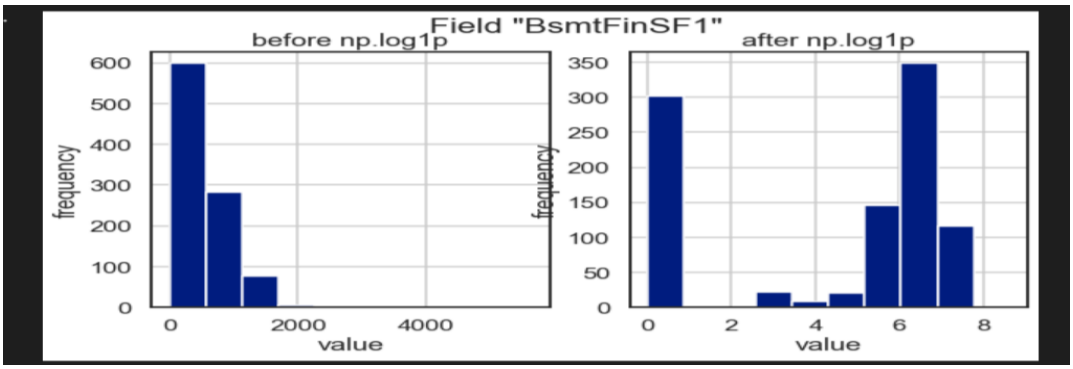
1. So I first started with importing all the libraries and data into the jupyter notebook.
2. Then we applied the train test split on it.

3. Then I focused on removing the duplicated data and focused on finding the information using data.describe()
4. Then we separated the categorical and numerical values of the dataset and proceeded and found the skewness in the numerical data.

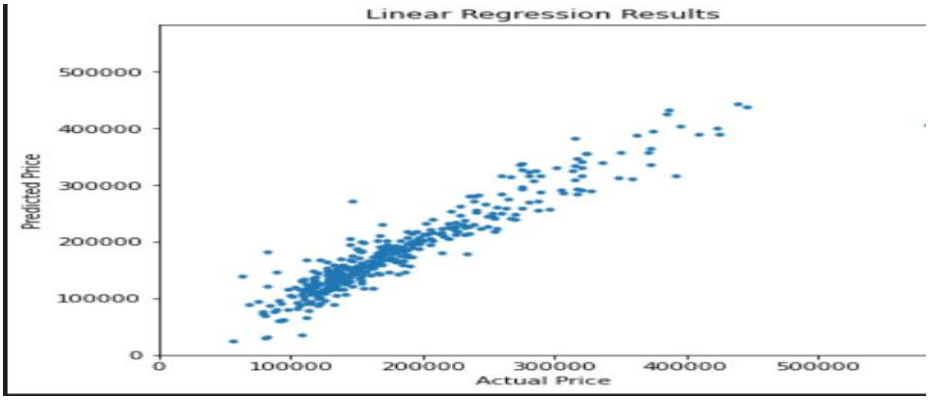
5. Then we printed the histogram for various of the numerical values
6. For the saleprice we plotted a pairplot with other values.

```
# Pairplot of transformed features and the target
sns.pairplot(data_sqrt.join(train['SalePrice']), plot_kws=dict(alpha=.2, edgecolor='none'));
```

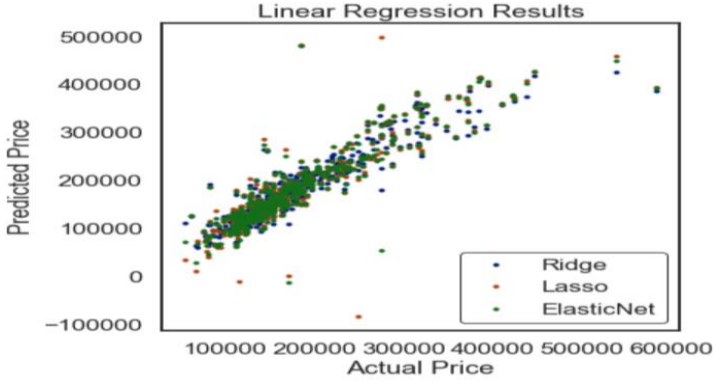
7. Then we encode the data using one hot encoder.
8. We also applied np/log as a feature.



9. Then we separated features from predictor.
10. We found the mean_squared_error and ran the linear regression



11. We do the same for lasso regression, ridge regression and elastic regression and including scaling as well by using MinMaxScaler
12. Then we compare the values of all the three regressions and make a plot of it.



13. Comparing the metrics

```python
rmse_vals = [linearRegression_rmse, ridgeCV_rmse, lassoCV_rmse, elasticNetCV_rmse]

labels = ['Linear', 'Ridge', 'Lasso', 'ElasticNet']

rmse_df = pd.Series(rmse_vals, index=labels).to_frame()
rmse_df.rename(columns={0: 'RMSE'}, inplace=1)
rmse_df
```

Python

	RMSE
Linear	306369.683423
Ridge	32169.176206
Lasso	39257.393991
ElasticNet	35001.234296

CONCLUSION

This analysis shows that feature engineering can have a large effect on the model performance, and if the data are sufficiently large, cross-validation should be preferred over train-test-split to construct model evaluation. In my case, the coefficients of predictors were not shrunk by the Lasso model, and it is shown that regularization does not always make big improvement on a given model. In the end, the Lasso regression has the highest when predicting on the test set, and categories of housing model appear to be the most important features to predict a house price. Also, Lasso did shrink some of the features that are not so important in terms of prediction.

While researching further analysis, I found a suggestion of using grouped Lasso when a model have categorical features, which is worth trying in this case.