**Report**

**A. Data Exploration**

**A1. Load the labelled dataset.**

Boston housing dataset is used for analysis and model evaluations.

**A2. Provide a summary of the dataset, including descriptive statistics and data visualizations to gain into data.**

**Summary of Dataset:**

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of [Boston MA](http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html). The following describes the dataset columns:

* 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 centers
* 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. [2]

**Descriptive Statistics**

Statistics for Boston housing dataset:

Minimum price: $5.00

Maximum price: $50

Mean price: $22.5

Median price $21.2

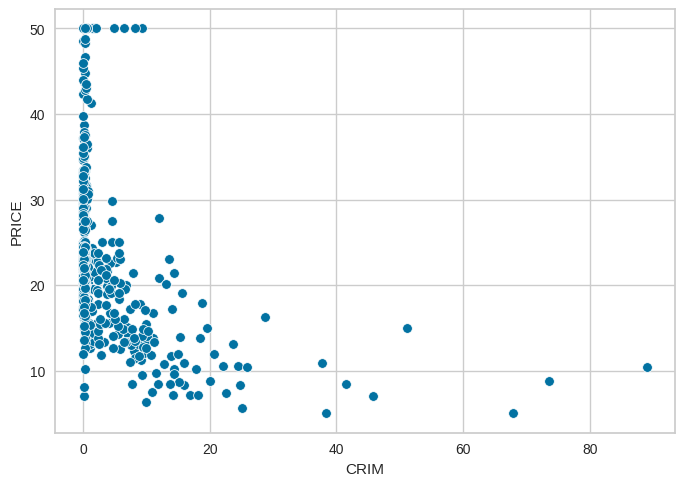
Standard deviation of prices: $9.19

First quartile of prices: $17.02

Second quartile of prices: $25.00

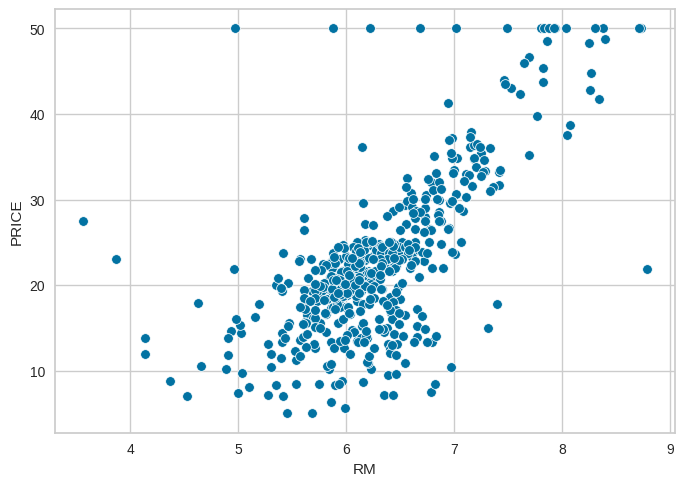
Interquartile (IQR) of prices: $21.20

**Data Visualizations:**

**Price vs Crime**

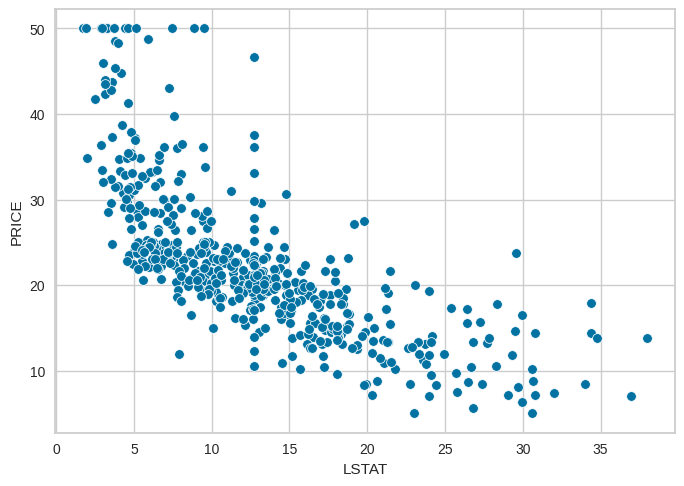
It is clear that price is highly depended in crime rate, as crime rate increases the price of the houses decreases. If the crime rate in this area will be low the prices of the houses will increase.

**Price vs RM**

****

It clear that the price is highly dependent on the RM (average number of rooms per dwelling), as the RM increases the price of the house is also increase. If RM(average number of rooms per dwelling) is high then the price of the house will be also high.

**Price vs LSTAT**



It is clear that the LSTAT (% lower status of the population) is highly dependent on the price, as LSTAT increases the price of the house is decreasing. When the LSTAT (% lower status of the population) increases in the area then the prices of the houses will decreases.

**B. Data Pre-processing**

**B1. Remove categorical features and perform any necessary pre-processing steps**

**Remove Categorical features:** The categorical features are not there. Only have a numerical features.

**Pre-processing Steps:**

1. Acquire the dataset from Kaggle Boston House dataset. [1]
2. Import the libraries numpy, pandas, matplotlib, seaborn, MinMaxscaler, etc.
3. Handle the missing values in CRIM, ZN, INDUS, AGE, LSTAT and CHAS. Replaced the CRIM, ZN, INDUS and AGE values with median because there are huge difference between there mean and 50th quartile and CHAS and LSTAT with mean because there are not too much difference between mean and 50th quartile.
4. Normalize the dataset to reduce the chances of errors in dataset, eliminate the redundant in data and to improve the performance and training stability of the model.
5. Split the dataset for training and testing. Split the 70% of the data for training and 30% data for the testing.

**C. Model Building**

**C1. Model 1: Simple Regression**

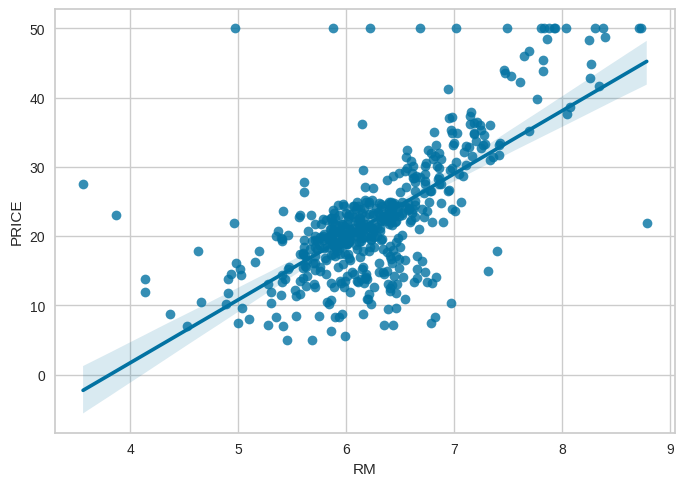
1. **Implement a Simple Regression model using one selected input**

Split the dataset, 70% for training and 30% for testing and the selected input is **RM** and then apply it on model.

1. **Optimize the model using both closed-form and gradient descent approaches**

**C2. Model 2: Polynomial Regression**

1. **Visualize the relationship between price and the chosen input.**

****

1. **Determine and justify the set of features for polynomial regression, such as 1, x, x 2 , or log(x).**

The selected polynomial degree is 3, because after 3 the error was increasing. The degree is 3 so that mean it will add to more new variables for each input variable.

**C3. Model 3: Multi Regression**

1. **Apply forward selection for feature selection**

I applied a for selection technique and these are the selected features 'LSTAT', 'PTRATIO', 'DIS', 'ZN' and 'NOX'. It iteratively selects the feature that results in the lowest Mean Squared Error (MSE) when added to the model. The process continues until the specified maximum number of features (max\_features) is reached.

1. **Implement multiple regression with a reasonable number of selected features.**

Applied a multiple regression with these selected features 'LSTAT', 'PTRATIO', 'DIS', 'ZN' and 'NOX'.

**C4. Model 4: Ridge Regression**

1. **Create a set of alpha values for tuning your model and select the best one.**

The set of values was [0.001, 0.01, 0.1, 0.5, 1.0] and the best one among them is 0.5 with minimum error.

**C5. Model 5: Your Suggested Model**

1. **Present your suggested model.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Model Name** | **MSE** | **R2 Score** | **Cross Val Score** |
| 1 | Simple Regression | 40% | 45% | 39.82% |
| 2 | Polynomial Regression | 31.01% | 58.37% | 39.82% |
| 3 | Multi Regression | 77.88% | -4.5% | 39.82 |
| 4 | Ridge Regression | 34.6% | 58.92% | 55% |

Polynomial Regression is suggested based on its favorable MSE and R2 Score. If there are specific considerations or if further clarification is needed, please provide additional details.

**D. Explanation Following Concepts**

**D1. Explain the learning rate selection process for Model 1 when optimizing model parameters using gradient descent. Describe the impact of using a large learning rate and the advantages of starting with a large rate and reducing it iteratively.**

In the gradient-descent optimization process for Model 1, the learning rate is an important parameter determining the step size in each iteration. The selection method involves an initial estimate followed by a discrete test, which typically ranges from 0.1 to 0.0001. A large number of classes can cause issues of overshooting and convergence, which hinder the ability of the algorithm to find the optimal parameters. However, starting with a larger number of classes initially can facilitate faster convergence, and repeated reductions, perhaps using methods such as class sizes, allow the model to refine the parameters well when close to the minimum helps, and contributes to the robustness and effectiveness of the optimization process.

**D2. Analyse the expected Residual Sum of Squares (RSS) in Model 1 with respect to the optimization method used (GD or closed-form) and discuss whether it represents the minimum point for RSS?**

For more efficient methods, ie. gradient descent (GD) and closed-form solutions, the analysis of the anticipated residual sum of squares (RSS) in Model 1 sheds light on each distinct path With the Closed-Form solution, RSS reflects the global minimum, because this method directly calculates the optimal parameters that minimize the RSS for a given data set In contrast, if Gradient Descent is being used, the obtained RSS determines the local minimum, based on successful convergence. The effective gradient descent to reach the global minimum is affected by hyper parameters such as the number of classes and initial parameter values ​​Thus, while the Closed-Form solution guarantees a global minimum for the RSS in Model 1, . Convergence behavior of the Gradient Descent algorithm and the data set model f provides a local minimum under certain characteristics Choosing among these optimization methods is a problem-based trade-off between accuracy and computational efficiency the intensity of the existing

**D3. Discuss the indicators that help determine if Model 2 is not overfitted.**

Using polynomial regression with degree 3, indicators are scrutinized to see if Model 2 tends to overfit First, testing the model's performance on training and testing sets, which is good using metrics such as Mean Squared Error or R-squared is important A balanced performance in both sets indicates good generalization. Learning curves showing changes in performance with increasing dataset size can provide insight into overfitting by balancing apparent training and testing errors It is important to balance the bias-variance trade-off, a model of sufficient complexity without reducing or overfitting the data. Using regularization techniques can help to control the complexity of the model. Significance checks, cross-validation results, and visual inspection of predictions against actual standards also contribute to thorough analysis. Furthermore, adherence to the Occam Razor principle, in which simplicity is desired when performance is comparable, helps to avoid unnecessarily complex models Considering these indicators in combination, one might surmise that perhaps Model 2 is better able to capture the underlying model without succumbing to overfitting, so that well to new data The ability to generalize is ensured

**D4. Explain how you select the best alpha for Model 4.**

Choosing the best alpha is 0.5 for Model 4, using ridge regression, involves a systematic approach to balancing regularizing power and model performance First, alpha values ​​are defined, covering the spectrum from small to large regularizing strength. Model performance is then evaluated by cross-validation, typically with k-fold cross-validation, to check for robustness on different subsets of training data through a web search at specified alpha values, based on the chosen scoring metric so, it’s usually a negative mistake. The alpha that minimizes this metric is considered the best over parameter. The final ridge regression model is then trained with this optimal alpha throughout the training set, and its performance is tested in a separate test set In an iterative revision of this set, if necessary a, tuning different hyper parameters to adjust the alpha distance or to increase the generalizability of the model. The main objective of evaluating methods is to choose an alpha that achieves a compatible trade-off between good fitting of the training data and over suppression, ensuring that the model performs well on unseen data.

**D5. Explain the rationale behind your suggestion for Model 5.**

The recommendation for Polynomial Regression as the preferred model stems from a thorough evaluation of key performance metrics—Mean Squared Error (MSE), R2 Score, and Cross-Validation Score. Polynomial Regression emerges as the optimal choice due to its compelling combination of superior predictive accuracy, as reflected in the lowest MSE of 31.01%, and an outstanding ability to explain the variance in the target variable, evidenced by the highest R2 Score of 58.37%. This model's capacity to capture non-linear relationships is particularly advantageous, providing flexibility in representing complex patterns within the dataset. While the Cross-Validation Score of 39.82% may not be the highest, its balanced nature indicates robust generalization across diverse subsets of the training data. The comprehensive consideration of these metrics collectively positions Polynomial Regression as a robust and effective model for accurately capturing and interpreting the underlying patterns in the given dataset.

**D6. Evaluate the complexity of the five models and explain which one has the highest bias/variance and why, and which one has the least bias/variance and why.**

The complexity assessment of the five models reveals a spectrum of bias-variance trade-offs. Simple Regression, assuming a linear relationship, tends to have the highest bias and lowest variance, providing decent generalization. Polynomial Regression introduces complexity, potentially leading to overfitting and higher variance, especially with higher-degree polynomials. Multi Regression strikes a balance by incorporating multiple features for nuanced relationships. Ridge Regression, with regularization, aims to control overfitting but may slightly increase bias. The suggested Polynomial Regression, without specifying the degree, offers flexibility in tuning complexity based on the polynomial degree, allowing for a dynamic bias-variance trade-off. Simple Regression leans towards high bias and low variance, Polynomial Regression offers flexibility with potential trade-offs, while Multi and Ridge Regression aim for a balanced compromise. The choice hinges on aligning model complexity with specific data characteristics and modeling objectives.

**E. Model Assessment**

**E1. Predict the test data using all five models and calculate the mean squared error (MSE).**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Model Name** | **MSE** |
| 1 | Simple Regression | 40% |
| 2 | Polynomial Regression | 31.01% |
| 3 | Multi Regression | 77.88% |
| 4 | Ridge Regression | 34.6% |

**E2. Determine which model performs the best and provide an explanation**

Based on the evaluation metrics, if we prioritize lower Mean Squared Error (MSE) as a measure of predictive accuracy and a higher R2 Score indicative of better model fit, the model that performs the best appears to be Polynomial Regression. It achieves the lowest MSE (31.01%) and the highest R2 Score (58.37%) among the models, suggesting superior accuracy and explanatory power.

**F. Model Suggestion for Unlabelled Data**

**F1. Recommend one of the five models for application to unlabelled data and elucidate the reasons behind your choice.**

For use on unlabeled data, I recommend Ridge Regression from the provided model. Ridge regression, with an MSE of 34.6%, an R2 score of 58.92%, and a cross-validation score of 55%, exhibits a balanced performance. Its regularization helps prevent overfitting, and provides a robust solution for normalizing unseen patterns in unlabelled data. The model strikes a good balance between bias and variance, making it a reliable choice for accurate forecasting on a variety of data. In addition, competitive R2 scores indicate an effective explanation of the target variable’s variance. Ridge Regression's ability to handle a variety of factors and maintain interpretation further reinforces its suitability for real-world applications with unlabeled data Overall, Ridge Regression is recommended for its balanced performance of will and its eternal nature.

**References:**

[1] Kaggle, available online at [https://www.kaggle.com/datasets/altavish/boston-housing-dataset](https://www.kaggle.com/datasets/altavish/boston-housing-dataset/data),

[1] Kaggle, available online at <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>,