Regression Methods to Predict the Compressive Strength of Concrete

Term Project for
REGRESSION ANALYSIS AND TIME SERIES MODELS (MA60280)

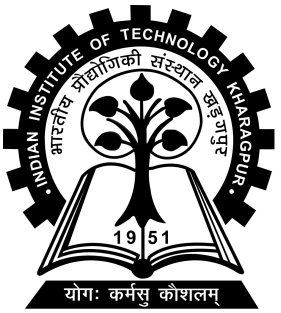
Indian Institute of Technology Kharagpur

by

Arup Baral (20MA20010)
Atharv Bajaj (20MA20014)
Kattunga Lakshmana Sai Kumar (20MA20026)
Padmanabhuni Jitendra Chandra Prabhakar (20MA20039)
Rangoju Bhuvan (20MA20048)
Shatansh Patnaik (20MA20067)

Under the supervision of

Dr. Buddhananda Banerjee



Department of Mathematics

Indian Institute of Technology Kharagpur

Spring Semester, 2024-25

April 15, 2024

Contents

Chapter 1

Regression Analysis

1.1 Introduction

The primary goal of the project is to perform multiple regression and fit a model to the given data. In this case since the number of features in the dataset is high, so as to tackle this problem we shall use various techniques such as Principal Component Regression, Ridge Regression and Fitting of Orthogonal Polynomials.

1.1.1 Description of the Dataset

The dataset (Kaggle Link: `http://surl.li/spmnl`) consists of concrete mixtures with various components, and the target variable is the concrete compressive strength measured in MPa. The dataset comprises **1030 instances** with **9 attributes**, including **8 quantitative input variables** and **1 quantitative output variable** representing the concrete compressive strength measured in MPa. There are **no missing attribute values**. Given below is a table which contains all information related to the components (input and output) in the dataset.

Name	**Measurement**	**Description**
Cement	kg in a m^3 mixture	Input Variable
Blast Furnace Slag	kg in a m^3 mixture	Input Variable
Fly Ash	kg in a m^3 mixture	Input Variable
Water	kg in a m^3 mixture	Input Variable
Superplasticizer	kg in a m^3 mixture	Input Variable
Coarse Aggregate	kg in a m^3 mixture	Input Variable
Fine Aggregate	kg in a m^3 mixture	Input Variable
Age	Day (1 to 365)	Input Variable
Concrete Compressive Strength	MPa	Output Variable

Table 1.1: Variable Information

1.1.2 Using Exploratory Data Analysis to derive Valuable Information

Measures of Central Tendency:

We can observe the values of Mean, Count Values, Quartiles, Maximum Values and Standard Deviation Values for each of the features in the table given below:

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.166408	73.894854	54.187379	181.564854	6.203204	972.918932	773.579515	45.662136	35.817961
std	104.507710	86.279340	63.995962	21.355663	5.973035	77.753954	80.175801	63.169912	16.705742
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.330000
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.710000
50%	272.900000	22.000000	0.000000	185.000000	6.300000	968.000000	779.500000	28.000000	34.445000
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.135000
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.600000

FIGURE 1.1: Measures of Central Tendency

The features mostly exhibit a significantly large standard deviation, suggesting that the dataset encompasses a wide range of input data.

Correlation Plot:

A correlation plot visually represents the relationships between variables in a dataset using colors to indicate the strength and direction of correlation. The plot is given below:

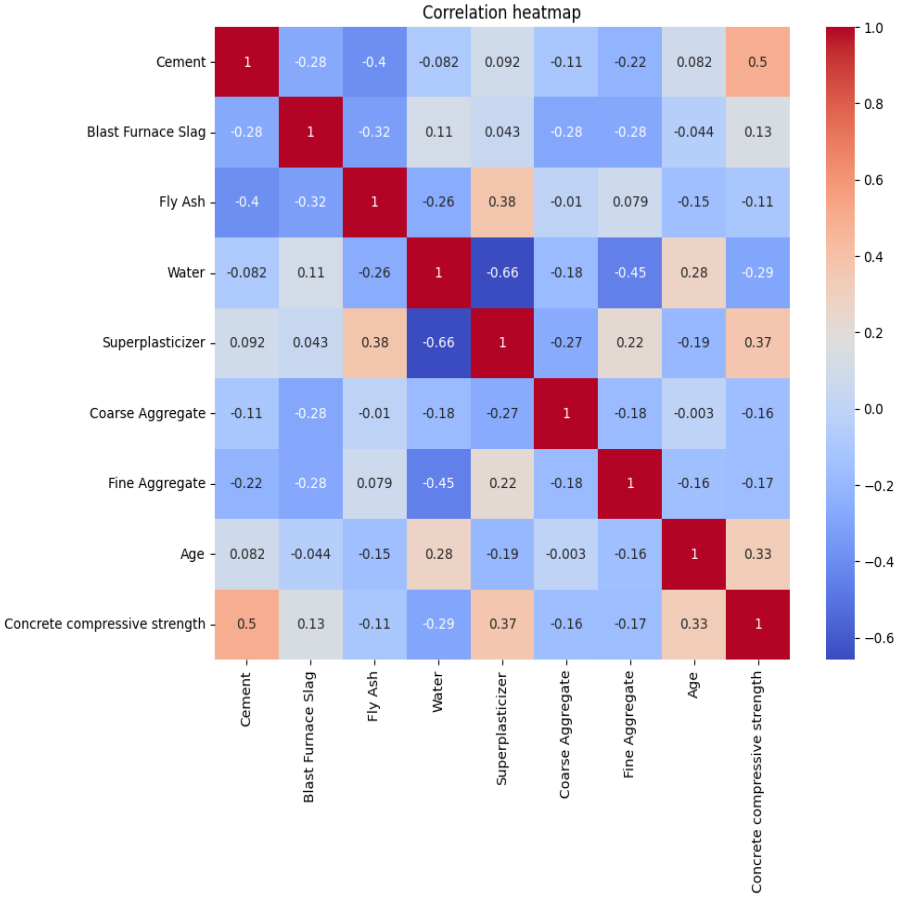


FIGURE 1.2: Correlation Plot

We can observe that the correlation between Water and Superplasticier is -0.66 and the correlation between flyash and cement is -0.4. This indicates a fairly potential multicollinearity in regression analysis. Multicollinearity occurs when independent variables are highly correlated, leading to unstable coefficient estimates and difficulty in interpreting the model. Addressing multicollinearity is essential for obtaining reliable regression results.

Box Plot:

A box plot provides summary statistics including median, quartiles, and outliers, displaying the distribution of a dataset in a compact visual format. For the given dataset, the box plot of all the fetaures can be depicted as follows:

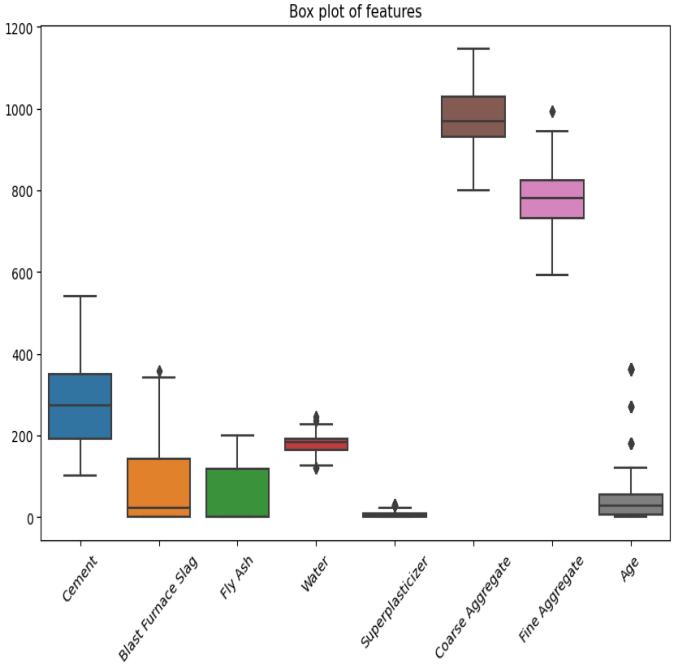


FIGURE 1.3: Box Plot

We can infer from the box plot that features like Cement, Blast Furnace Slag, Flyash have longer lengths and hence, these have higher variability and a potential presence of outliers, meanwhile features like Superplasticizer and Water have shorter lengths, which suggest that the datat points have low variability and noise.

Violin Plot:

A violin plot depicts the summary statistics such as median and quartiles, offering insights into the shape, spread, and skewness of the dataset's distribution. For the given dataset, the violin plot of all the fetaures can be depicted as follows: We can observe that some features Cement, Blast Furnace Slag, Flyas have very thin violin plots, which

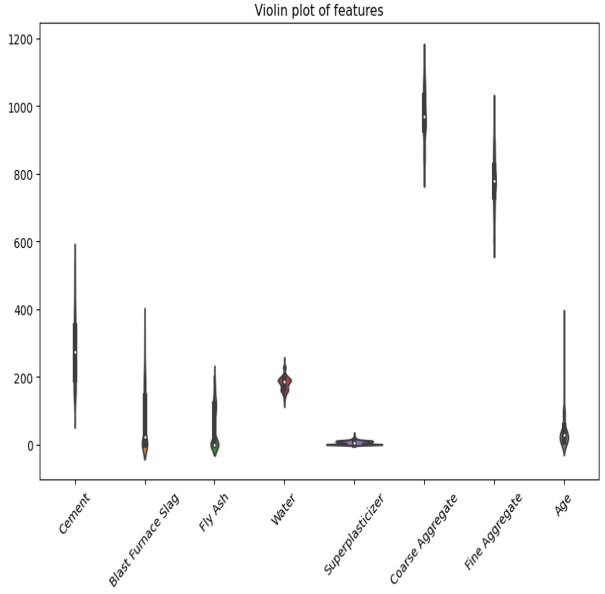


FIGURE 1.4: Violin Plot

indicates that these features have large variability and large skewness in the data meanwhile as observed Water and Superplasticizer have flat violin plots, which implies that these features have low skewness and variability.

The dataset is scaled accordingly before proceeding to regression to avoid overflow errors. Also, the train-test split ratio is 4:1 for all the models applied further.

1.2 Multiple Linear Regression

1.2.1 Estimation of Parameters of the Model

The regression coefficients $\boldsymbol{\beta}$ are estimated using the method of Ordinary Least Squares (OLS), which minimizes the sum of the squared residuals. The estimated regression coefficients are given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{1.1}$$

where:

- $\hat{\boldsymbol{\beta}}$ is the $(p+1) \times 1$ vector of estimated regression coefficients
- $\mathbf{X}$ is the $n \times (p+1)$ design matrix, with the first column being a vector of 1's and the remaining columns containing the values of the independent variables
- $\mathbf{y}$ is the $n \times 1$ vector of observed values of the dependent variable

For the given dataset, we can obtain the estimates of the parameters as follows: Calculating the Confidence Intervals for all parameters of the model, we obtain:

```
The value of beta0 is -0.00021576303923781515
The value of beta1 is 0.7442830808489095
The value of beta2 is 0.5604572694493724
The value of beta3 is 0.31477298177979063
The value of beta4 is -0.17348621824147636
The value of beta5 is 0.1103940838895006
The value of beta6 is 0.08286745779972085
The value of beta7 is 0.11547595433661348
The value of beta8 is 0.44031656445551937
```

1.2.2 Estimation of Confidence Intervals

In multiple linear regression, the estimated coefficients $\hat{\boldsymbol{\beta}}$ are unbiased and have a multivariate normal distribution under assumptions of linearity, independence, homoscedasticity, normality, and low multicollinearity.

1.2.2.1 Confidence Intervals for Parameters

The confidence interval for a coefficient β_i is given by:

$$\hat{\beta}_i \pm t_{\alpha/2, n-p-1} \times SE(\hat{\beta}_i)$$

where $t_{\alpha/2,n-p-1}$ is the critical value of the t-distribution, and $SE(\hat{\beta}_i)$ is the standard error of the coefficient.

The standard error $SE(\hat{\beta}_i)$ is calculated as:

$$SE(\hat{\beta}_i) = \sqrt{\text{Var}(\hat{\beta}_i)}$$

The variance of the coefficient is:

$$\text{Var}(\hat{\beta}_i) = \sigma^2 \left((X^T X)^{-1} \right)_{ii}$$

1.2.2.2 Confidence Interval for Variance (σ^2)

The confidence interval for the variance of the error term σ^2 is given by:

$$\left(\frac{(n-p-1)s^2}{\chi^2_{\alpha/2,n-p-1}}, \frac{(n-p-1)s^2}{\chi^2_{1-\alpha/2,n-p-1}} \right)$$

where s^2 is the estimated variance of the error term, and $\chi^2_{\alpha/2,n-p-1}$ and $\chi^2_{1-\alpha/2,n-p-1}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the chi-square distribution with $n-p-1$ degrees of freedom, respectively.

Calculating the Confidence Intervals for all parameters of the model, we obtain:

```
95% confidence interval for beta 0 : -0.04357194196999485 , 0.04314041589151922
95% confidence interval for beta 1 : 0.6290678321777943 , 0.8594983295200247
95% confidence interval for beta 2 : 0.44571402464663595 , 0.6752005142521087
95% confidence interval for beta 3 : 0.20841096555894353 , 0.4211349980006377
95% confidence interval for beta 4 : -0.28420438219430366 , -0.06276805428864905
95% confidence interval for beta 5 : 0.037174279568743265 , 0.18361388821025793
95% confidence interval for beta 6 : -0.012118036218921518 , 0.17785295181836322
95% confidence interval for beta 7 : 0.003852235131426754 , 0.22709967354180022
95% confidence interval for beta 8 : 0.3926173998091377 , 0.48801572910190105
95% confidence interval for sigma square : 0.3650118521278395 , 0.4432931461749914
```

1.2.3 Significance Testing of Parameters of the Model

Significance testing for parameters in multiple linear regression involves evaluating the null hypothesis that a predictor's coefficient is zero against the alternative that it is not. This is typically done using t-tests or F-tests, with p-values indicating the probability of observing the data if the null hypothesis were true.

For the above model we get the following output on applying Significance Testing for the coefficients involved in the model:

```
beta 0  is not significant
beta 1  is significant
beta 2  is significant
beta 3  is significant
beta 4  is significant
beta 5  is significant
beta 6  is not significant
beta 7  is significant
beta 8  is significant
```

1.2.4 ANOVA

In the context of multiple linear regression, the Analysis of Variance (ANOVA) technique is used to evaluate the overall significance of the regression model and to partition the total variation in the dependent variable into different sources of variation.

Mathematically, the squared error components essential to the model are defined as follows:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{1.2}$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1.3}$$

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1.4}$$

where:

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ is the predicted value of the dependent variable
- $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the mean of the dependent variable

It can be shown that the Total Sum of Squares (SST) is the sum of the Regression Sum of Squares (SSR) and the Residual Sum of Squares (SSE):

$$\text{SST} = \text{SSR} + \text{SSE} \tag{1.5}$$

The ANOVA table for the multiple linear regression model is as follows:

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-statistic
Regression	p	SSR	$\frac{\text{SSR}}{p}$	$\frac{\text{MSR}}{\text{MSE}}$
Residual	$n-p-1$	SSE	$\frac{\text{SSE}}{n-p-1}$	
Total	$n-1$	SST		

The null hypothesis for the ANOVA F-test is that all the regression coefficients are zero, except for the intercept:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{1.6}$$

The alternative hypothesis is that at least one of the regression coefficients is non-zero:

$$H_1 : \text{At least one } \beta_j \neq 0, \text{ for } j = 1, 2, \ldots, p \tag{1.7}$$

The F-statistic is computed as:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} \tag{1.8}$$

If the null hypothesis is true, the F-statistic follows an F-distribution with $(p, n-p-1)$ degrees of freedom. The p-value of the F-test is the probability of observing an F-statistic greater than or equal to the calculated value, under the assumption that the null hypothesis is true, else we reject the hypothesis.

Applying the Analysis of Variances to the given model we obtain:

```
The Sum of Sqaured Errors for the model ( SSE ) : 327.053201997045
The Mean Squared error for the model ( MSE ) : 0.3973914969587424
The Sum of Squares Regression for the model ( SSR ) : 512.5336404616726
The Total error for the model ( SST ) :  839.5868424587176
The Model is  significant
```

1.2.5 Calculation of Coefficient of Determination

The coefficient of determination, denoted as R^2, is a measure of the proportion of the total variation in the dependent variable y that is explained by the multiple linear regression model.

The coefficient of determination is defined as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{1.9}$$

For the given dataset we have the following Result:

```
Multiple Regression Model Summary:
----------------------------------------
R-squared            : 0.6105
Adjusted R-squared   : 0.6066
Number of significant terms   : 7
Number of non-significant terms: 2
Model significance   : significant
SSE on test set: 70.91222451511904
MSE on test set: 7.879136057235448
----------------------------------------
```

The R^2 score obtained is not optimal and we might need to fit a higher degree polynomial for further improvement.

1.2.6 Studentized residuals

Studentized residuals are standardized measures used in regression analysis to assess model fit and identify outliers. They're obtained by dividing the residual (the difference between observed and predicted values) by its estimated standard deviation, adjusted for the model's degrees of freedom. These residuals should ideally follow a normal distribution. The studentized residual plot, displaying residuals against predicted values, helps detect patterns or outliers that may indicate model shortcomings. A random scatter around zero in the plot suggests a well-fitted model.

For the multiple linear regression model, the residual plot is obtained as follows

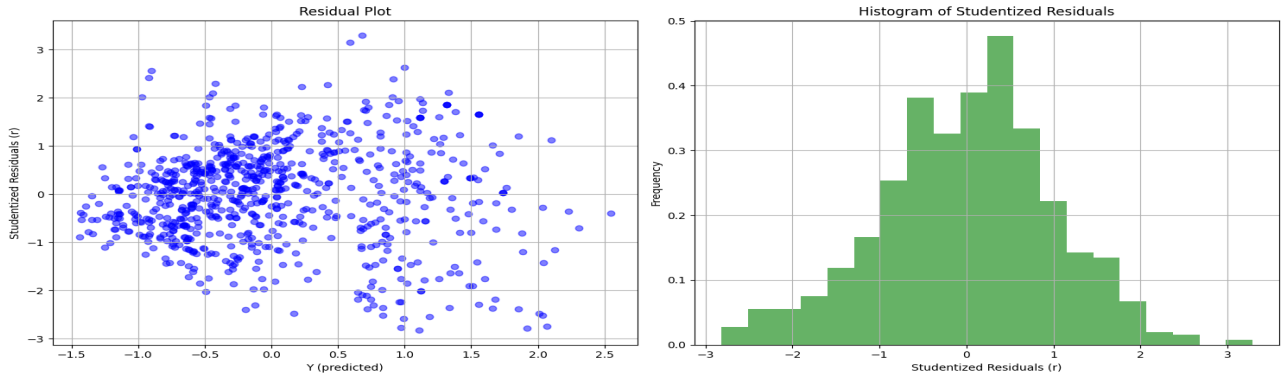


FIGURE 1.5: Residual Plot and Histogram of Studentized Residuals

We can see that the residuals are not concentrated around 0 which is sub-optimal which also explains the average R^2 score obtained.

1.3 Polynomial Regression

The relationship between parameters involved in engineering problems is usually non-linear. Polynomial Regressions are therefore preferred compared to a purely linear regression. However, using a polynomial regression comes with its own pitfalls, and we need to keep a few considerations in mind while building our model.

Using a low-order model in a transformed variable is typically more desirable than employing a high-order model in the original metric. Indiscriminately fitting high-order polynomials constitutes a significant misuse of regression analysis. It is crucial to uphold a sense of parsimony, favoring the simplest model that adequately reflects the data and aligns with our understanding of the problem context.

As the order of the polynomial increases, the X^TX matrix becomes ill-conditioned. This means that matrix inversion calculation becomes inaccurate, and a considerable variance is introduced into the model parameters, thereby giving us a bad model.

1.3.1 Forward Selection

The forward selection method is an approach that fits a polynomial regression while keeping the above problems in mind. Here, we start with a low-order regression while successively increasing the regression order and checking the significance of the t-statistic of the highest-order terms. We stop increasing the regression order when the highest-order terms are no longer significant.

The possible ill-conditioning in this model is resolved by checking the correlations of the features involved and removing one of the terms if its correlation with another term is higher than a threshold value.

An alternative approach that could be used is backward elimination, where we start with a high-order regression and delete terms until the highest-order term has a significant t-statistic. We have opted for forward elimination since the computation time for fitting a polynomial regression with a large degree on 8 parameters is fairly significant.

The results obtained upon applying forward elimination are shown below

TABLE 1.2: Regression Model Summary

Model	SSRegression	SSError	R^2	R^2_{adjusted}	Model Significance	SSE on Test Set
Degree 1	512.53	327.05	0.6105	0.6066	Significant	70.91
Degree 2	682.68	156.91	0.8131	0.8026	Significant	41.07
Degree 3	761.44	78.14	0.9069	0.8843	Significant	41.83
Degree 4	809.62	29.97	0.9643	0.9172	Significant	1520.94

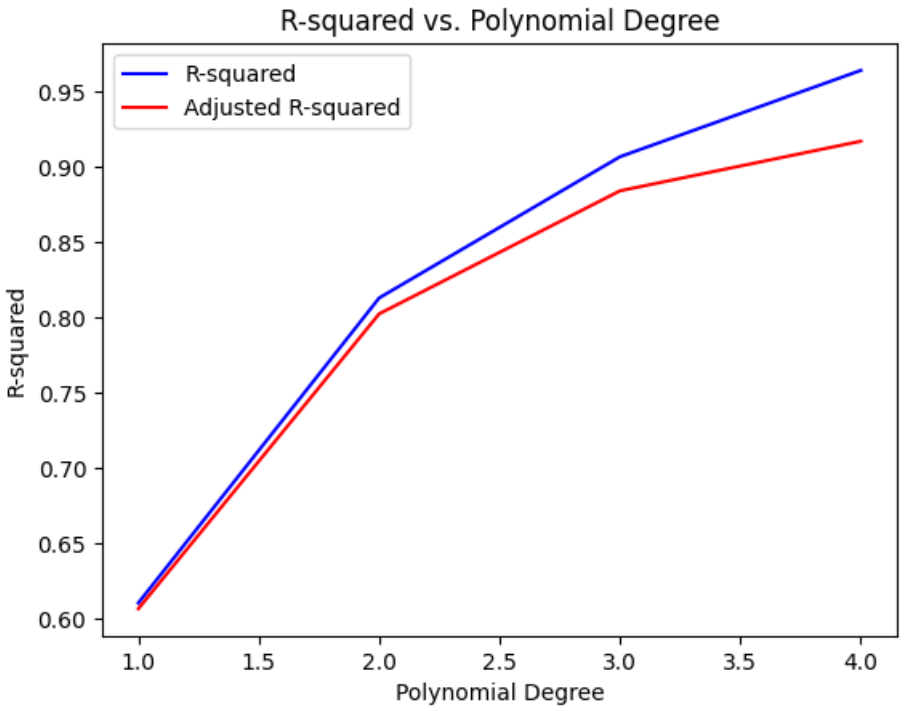


FIGURE 1.6: Variation of R^2 and $R^2_{adjusted}$ with the Degree while continuing forward substitution

Applying forward substitution on the dataset gives us the optimal polynomial regression to be of order 4 with $R^2 = 0.9643$ and $R^2_{adjusted} = 0.9172$, which is a good departure from our original multi-linear regression model with $R^2 = 0.6105$ and $R^2_{adjusted} = 0.6066$

The residual plots for models of increasing degrees are obtained as follows:

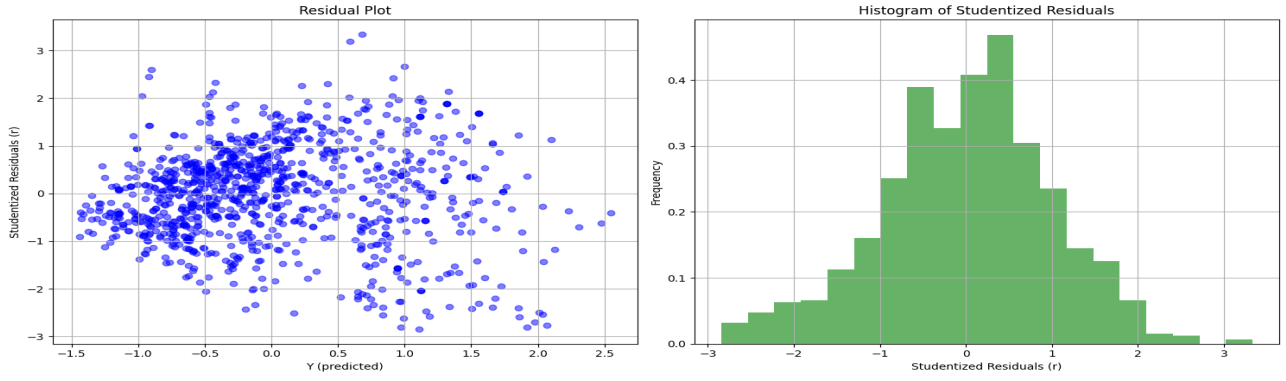


FIGURE 1.7: Residual Plot and Histogram of Studentized Residuals (Degree 1)

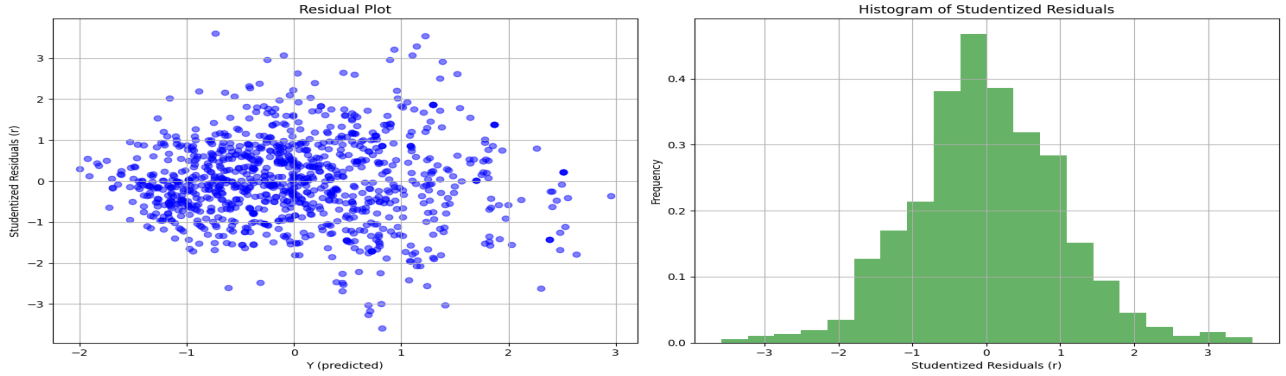


FIGURE 1.8: Residual Plot and Histogram of Studentized Residuals (Degree 2)

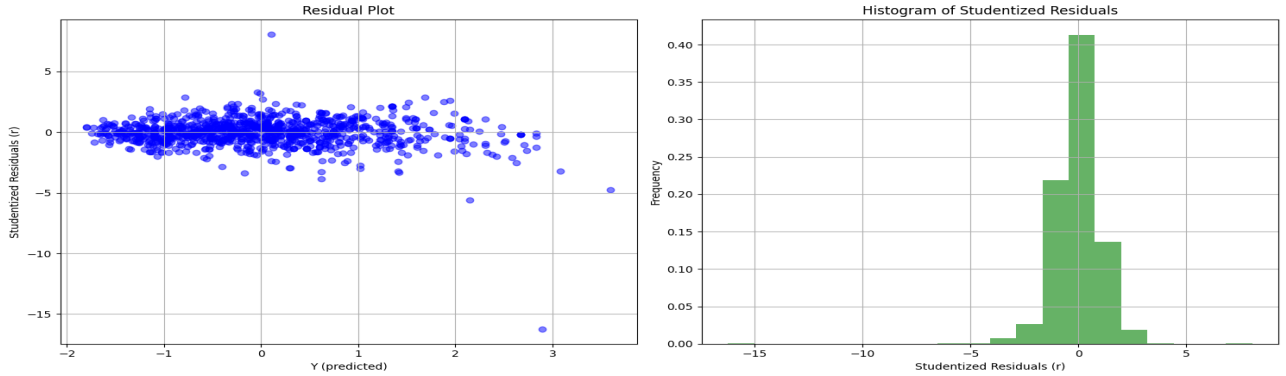


FIGURE 1.9: Residual Plot and Histogram of Studentized Residuals (Degree 3)

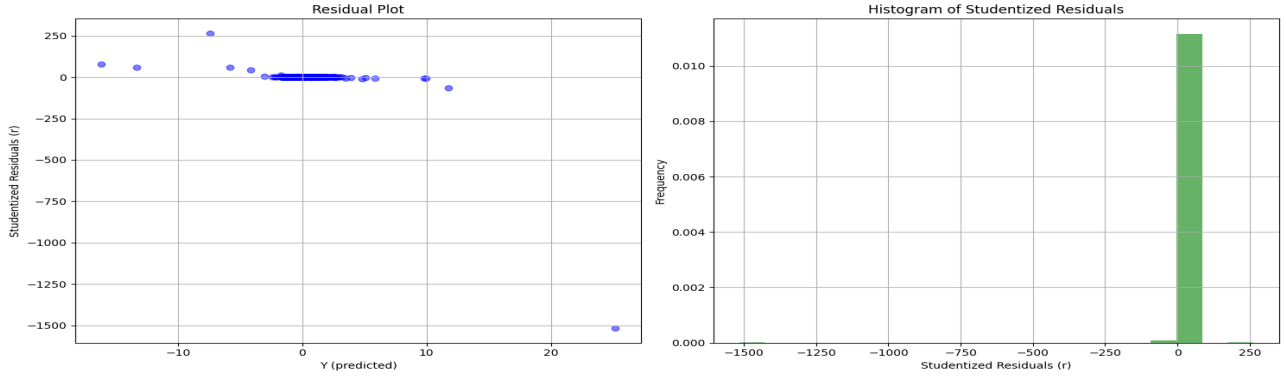


FIGURE 1.10: Residual Plot and Histogram of Studentized Residuals (Degree 4)

However, one can see that despite the order 4 regression having the lowest $SSError$ on the training set, it has the highest $SSError$ on the Test set. This is a classic case of overfitting, which occurs when the values on the test set lie outside that of the training set. We may choose the 3rd-order regression to mitigate this since we get the best of both worlds here. This can also be observed from the residual plots which mostly converge to zero at degree 4 while being enougly scattered at degree 3.

1.3.2 Orthogonal Polynomial Regression

Orthogonal Polynomial Regression is a technique used to fit a polynomial regression model where the basis functions (polynomials) are orthogonal to each other. This property simplifies the computation of regression coefficients and can improve numerical stability. The Gram-Schmidt orthogonalization process is typically used to construct these orthogonal polynomials.

1.3.2.1 Orthogonal Process using Gram-Schmidt Orthogonalization

The Gram-Schmidt orthogonalization process is used to construct a set of orthogonal polynomials from a set of basis polynomials. Given a set of basis polynomials $\{1, x, x^2, x^3, \ldots\}$, the process proceeds as follows:

1. **Initialization:** Set $q_0(x) = 1$ (normalized).
2. **Iteration:** For $i > 0$, calculate $q_i(x)$ as:

$$q_i(x) = x^i - \sum_{j=0}^{i-1} \frac{\langle x^i, q_j \rangle}{\langle q_j, q_j \rangle} q_j(x)$$

   where $\langle f, g \rangle$ is the inner product defined as:

$$\langle f, g \rangle = \int_a^b f(x)g(x)w(x)\,dx$$

   Here, $w(x)$ is the weight function and $[a, b]$ is the interval of orthogonality.
3. **Normalization:** Normalize each $q_i(x)$ such that $\langle q_i, q_i \rangle = 1$.
4. **Orthogonal Polynomials:** The resulting polynomials $\{q_0(x), q_1(x), q_2(x), \ldots\}$ are orthogonal polynomials.

1.3.2.2 Mathematical Formulation

In orthogonal polynomial regression, the model can be written as:

$$Y = \beta_0 + \beta_1 q_1(X) + \beta_2 q_2(X) + \ldots + \beta_p q_p(X) + \epsilon$$

where:

- Y is the dependent variable,
- $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients,
- X is the independent variable,
- $q_1(X), q_2(X), \ldots, q_p(X)$ are orthogonal polynomials of degrees 1 to p,
- ϵ is the error term.

The regression coefficients can be estimated using least squares estimation:

$$\beta = (Q^T Q)^{-1} Q^T Y$$

where Q is the matrix of orthogonal polynomials evaluated at the data points. This formulation ensures that the resulting polynomials are orthogonal, which can lead to more stable and interpretable regression results.

Orthogonal regression is easier to apply when there is a single independent variable,yet to test the model for our data, we opted to change it a bit and add polynomials of the features independently (i.e., there are no cross-over terms in the polynomial). We ensured that rest of the algorithm remains the same.

Now applying Orthogonal Regression to our data set we have the following Results:

TABLE 1.3: Orthogonal Polynomial Regression Summary

Model	SSRegression	SSError	R^2	R^2_{adjusted}
Degree 1	633.93	396.07	0.6155	0.6125
Degree 2	797.43	232.57	0.7742	0.7706
Degree 3	865.28	164.72	0.8401	0.8363
Degree 4	897.90	132.10	0.8717	0.8676
Degree 5	904.07	125.93	0.8777	0.8728

Plot between R^2 and Polynomial Degree:

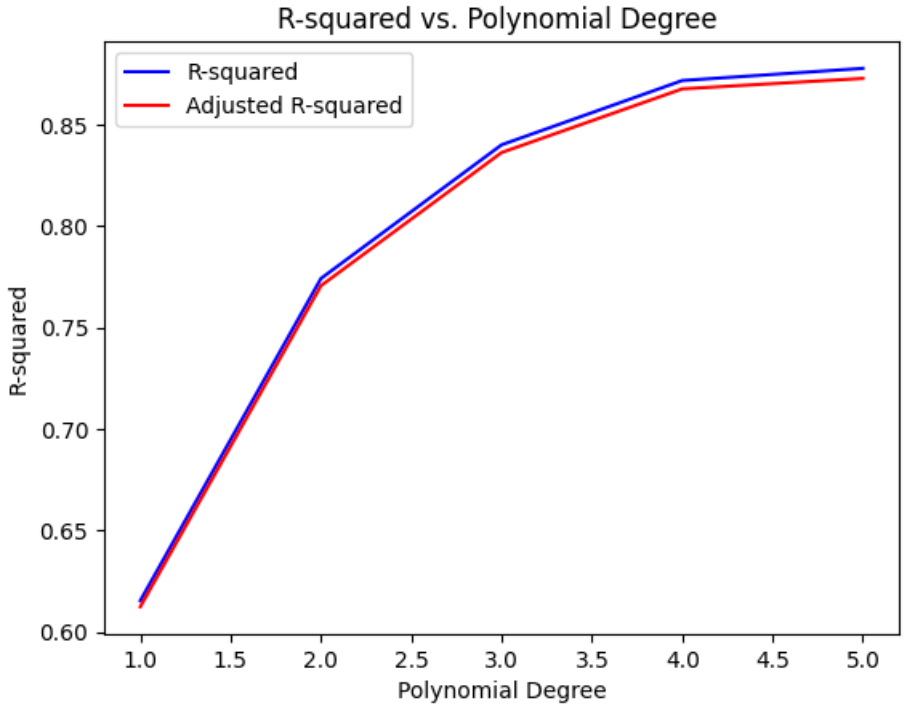


FIGURE 1.11: R^2 vs Polynomial Degree

We can observe that the R^2 score is increasing with an increase in the degree of the polynomials added which suggests that there is a better fitting higher degree polynomial than the linear model and to investigate that, we will use PCR to tackle multicollinearity.

1.3.3 Principal Component Regression

Principal Component Regression (PCR) is a technique used in regression analysis to handle multicollinearity and high dimensionality in predictor variables. It combines the concepts of Principal Component Analysis (PCA) and linear regression.

1.3.3.1 Principal Component Analysis (PCA)

PCA is a method to reduce the dimensionality of a dataset while retaining most of the variability present in the data. Given a dataset X of n observations and p predictors, PCA finds a set of $k < p$ orthogonal vectors $\{v_1, v_2, \ldots, v_k\}$, called principal components, that capture the maximum variance in the data.

The first principal component, Z_1, is the linear combination of the predictors that explains the most variance. Subsequent components, $Z_2, Z_3, \ldots, Z_k$, explain the remaining variance in decreasing order.

1.3.3.2 Mathematical Formulation

1. **Principal Components:** The principal components are calculated as:

$$Z_i = X \cdot v_i$$

where X is the standardized predictor matrix, v_i is the ith eigenvector of the covariance matrix of X, and $\cdot$ denotes matrix multiplication.

2. **Eigenvalue Decomposition:** The covariance matrix Σ of X is decomposed as:

$$\Sigma = V \Lambda V^T$$

where V is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues.

3. **Regression Model:** PCR fits a regression model using the first k principal components:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k + \epsilon$$

4. **Regression Coefficients:** The regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are estimated using least squares regression:

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y$$

PCR is effective in handling multicollinearity because it projects the predictors onto a lower-dimensional subspace spanned by the principal components, which are orthogonal and uncorrelated. This reduces the impact of multicollinearity on the regression estimates. **Performing PCR on the dataset we get the following Output:**

TABLE 1.4: Principal Component Regression Summary

Model	SSRegression	SSError	R^2	R^2_{adjusted}	Model Significance	SSE on Test Set
Degree 1 PCR	633.93	327.05	0.6105	0.6066	Significant	70.91
Degree 2 PCR	834.93	156.91	0.8131	0.8026	Significant	41.07
Degree 3 PCR	927.58	57.86	0.9311	0.9139	Significant	29.75
Degree 4 PCR	986.64	58.54	0.9303	0.8256	Significant	37.17

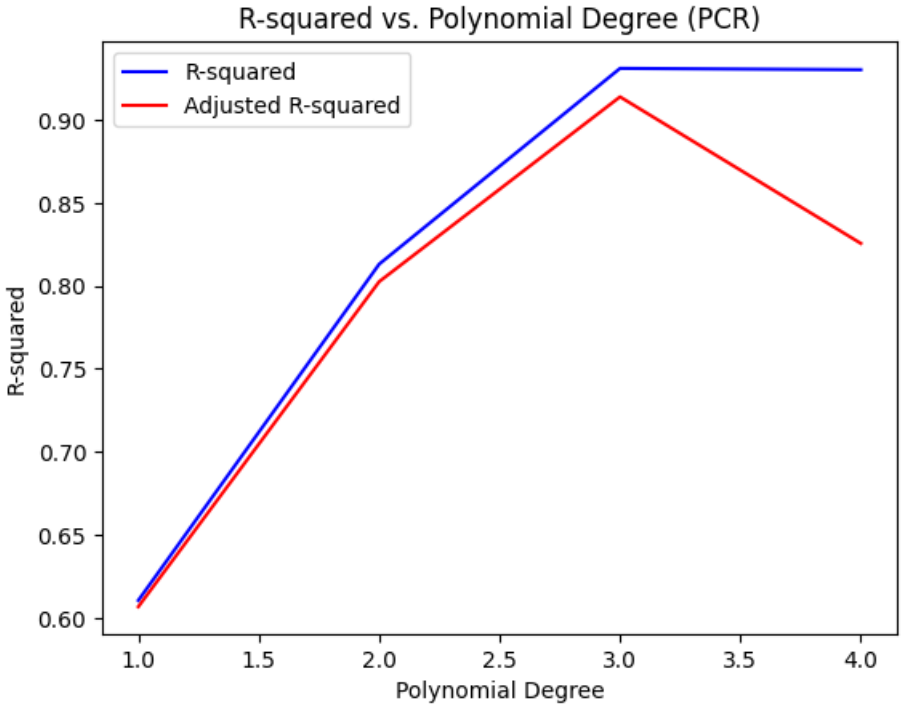


FIGURE 1.12: Variation of R^2 and $R^2_{adjusted}$ with the Degree while continuing Principal Component Regression

Similar to what we have seen in Table 1.2, the above table also suggests that the degree 3 polynomial is the best-fitting polynomial because the degree 4 polynomial is overfitting the train dataset. Instead of avoiding multicollinearity, we can directly minimize the variance of fitted parameters obtained which is implemented in Ridge regression model that follows.

1.3.4 Ridge Regression

Ridge regression serves as a pivotal strategy to tackle multicollinearity and enhance the stability of coefficient estimates, thereby bolstering the model's robustness. By introducing a penalty term into the regression objective function, ridge regression effectively restrains the magnitudes of coefficients while still allowing them to contribute meaningfully to predictive accuracy. This approach strikes a fine balance between minimizing the discrepancy between observed and predicted values and regulating the variability of coefficient estimates. Consequently, ridge regression not only addresses multicollinearity but also mitigates the risk of overfitting, promoting more reliable and generalizable model outcomes.

The mathematical formulation of the ridge regression objective function is as follows:

$$\text{Minimize} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

Where:

- n is the number of observations
- p is the number of predictor variables
- y_i is the observed target value for the ith observation
- $\hat{y}_i$ is the predicted value for the ith observation
- β_j is the coefficient of the jth predictor variable
- λ is the regularization parameter (also known as the shrinkage parameter) that controls the strength of the penalty term.

Applying Ridge Regression on the Dataset we get the following results:

Table 1.5: Ridge Regression Summary

Model	SSRegression	SSError	R^2	R^2_{adjusted}	SSE on Test Set
Degree 1	494.77	344.82	0.5893	0.5853	75.14
Degree 2	623.90	215.69	0.7431	0.7286	51.83
Degree 3	707.58	132.00	0.8428	0.8036	44.41
Degree 4	758.73	80.86	0.9037	0.7591	34.84

Plot between R2 and Polynomial Degree:

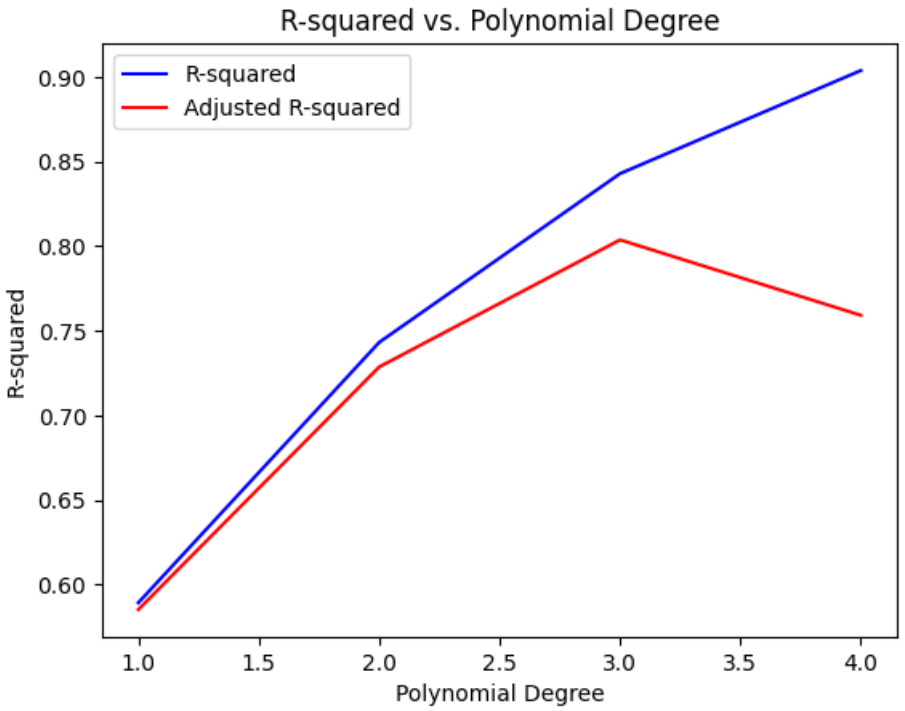


FIGURE 1.13: Variation of R^2 and $R^2_{adjusted}$ with the Degree while continuing Ridge Regression

We can observe the $R^2_{adjusted}$ on the training data declined at degree 4 simultaneously overfitting train data and not being able to generalize on the test data. So, we can infer that degree 3 is the best-fitting polynomial for the given data and the coefficients are also found a=out accordingly for the same.

1.4 Conclusion

Based on the primary goal of the project, which is to perform multiple regression and fit a model to the given data, several techniques have been employed to tackle the high number of features in the dataset. These techniques include Principal Component Regression, Ridge Regression, and Fitting of Orthogonal Polynomials.

The use of Principal Component Regression helps in reducing the dimensionality of the dataset by transforming the original features into a set of principal components, thus simplifying the regression model. Ridge Regression, on the other hand, introduces a regularization term to the regression model, which helps in reducing the impact of multicollinearity and overfitting, especially useful in datasets with high dimensionality.

Furthermore, the fitting of orthogonal polynomials allows for the modeling of complex relationships between the predictors and the response variable, providing a more flexible model that can capture non-linear patterns in the data.

Overall, these techniques enable the creation of more robust and interpretable regression models, improving the accuracy and reliability of the predictions made on the given dataset. Finally, we were able to get an R^2 score of 0.9 for a degree 3 polynomial which also doesn't overfit the training data indicating that the models applied work well on the chosen dataset.