International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Comparative Assessment of Regression Models Based On Model Evaluation Metrics

Abhishek V Tatachar

Undergraduate Student, Department of Information Science and Engineering, Global Academy of Technology

---

Abstract **–** *Supervised Learning is a prominent task of machine learning which maps inputs to corresponding outputs. Regression is one such supervised learning technique that models a relationship between independent and dependent variables. Given that regression is a powerful machine learning technique used to make predictions, different regression models can be made use of – Linear Regression, Ridge Regression, Support Vector Regression, Lasso Regression, and Polynomial Regression being prominent ones. This paper aims to implement these models and evaluate the predictions based on certain metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MEA), R^2, and Adjusted R^2.*

Keywords – Regression, Supervised Learning, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MEA), R^2, and Adjusted R^2.

I. INTRODUCTION

Artificial Intelligence and Machine Learning is in action for a long time now. Sir Alan Turing in his paper [1] first asked the question "Can a machine think?". In [1] Sir Alan Turing has defined this question in the form of an Imitation Game. In the year 1959, a pioneer in the field of computer gaming and AI, Arthur Samuel, coined the term "Machine Learning". Artificial Intelligence allows systems to solve problems or make a decision like human beings. Artificial Intelligence provides the systems with human-like cognitive capacity. Machine learning is one of the branches of Artificial Intelligence that gives the machines to learn and improve from experience without having to be explicitly programmed. The main aim of machine learning is to reduce the number of human interventions required to perform menial tasks. Machine Learning is categorized into three types – Supervised Learning, Unsupervised Learning, and Reinforced Learning. Machine Learning has a broad scope and range of applications such as Image Recognition, Speech Recognition, Online fraud detection, self-driving cars, and so on.

Supervised Learning is one of the machine learning types in which the machines are trained using labeled data. That is, in supervised learning, the learning machines are provided with the input as well as the correct target output. The supervised learning algorithm is expected to generate a mapping function that maps the input variable(s) to the output variable [2]. How does this work? The models in the supervised learning learn about the data which is called training. Once training is completed, the model is tested against the testing data (this is the unseen data for which the output is not known) and the predictions for the test data are obtained. There are two types of supervised learning techniques – Regression and Classification.

Regression is one of the supervised learning techniques that is used for the prediction of continuous data. The regression technique is used to model relationships between dependent and independent variables and produce a line of best fit. That is regression is a statistical method that is used in many disciplines including finance and healthcare that is used to determine the strength and the character of the relationship between the dependent and independent variables. Unlike classification problems, in regression, the output or projected features are continuous [3].

The following diagram provides how each of these topics, i.e., Artificial Intelligence, Machine Learning, and Supervised learning fall into each other.

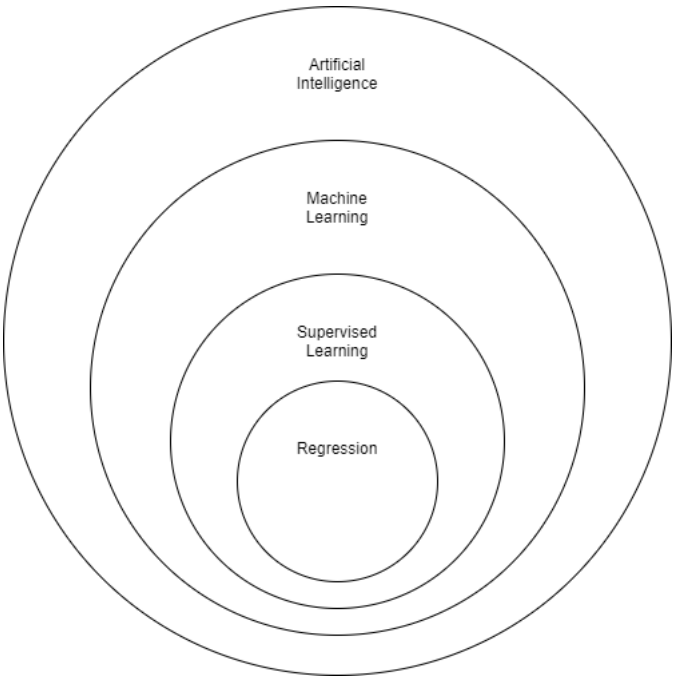

Figure 1: Concept Flow

Regression is of many types, the ones discussed in this paper include - Linear Regression, Lasso Regression, Support Vector Regression, Ridge Regression, Polynomial Regression.

In this paper, we have evaluated these models against certain metrics. These are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MEA), R^2, and Adjusted R^2.

Mean Squared Error (MSE) – Mean Squared Error which is also called the Mean Squared Deviation is the squared difference between the actual values and the predicted values. That is MSE tells us how close the line of best fit is to the set of points. MSE is always a positive value. The square is taken to eliminate negative signs. The closer the MSE is to 0, the more accurate is the prediction.

The formula for Mean Squared Error is given as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE – Mean Squared Error
n – Number of predictions
Y_i – Observed values
$\hat{Y}_i$ – Predicted values

Root Mean Squared Error (RMSE) – The Root Mean Squared error which is also called the Root Mean Squared Deviation is the square root of the mean of squares of all the errors [4]. In other words, the RMSE is simply the Standard deviation of the errors. RMSE again tells us how close the line of best fit is to the set of points.

The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

RMSE – Root Mean Squared Error
n – Number of predictions
Y_i – Observed values
$\hat{Y}_i$ – Predicted values

Mean Absolute Error (MAE) – The Mean Absolute error which is also called the Mean Absolute deviation provides us the average of the absolute difference between the observed value and the predicted values. The difference between the MAE and MSE is that MAE takes the absolute difference between the predicted values and the observed values whereas the MSE takes the squared difference.

The formula for MAE is given as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

MAE – Mean Absolute Error
n – Number of predictions
Y_i – Observed values
$\hat{Y}_i$ – Predicted values

R-Squared (R^2) – R^2 is called the coefficient of Determination. R-Squared determines the proportion of variance in the dependent variable that can be explained by the independent variables. R-squared provides us with the goodness of fit (the extent to which the observed values match the predicted values) for the regression model.

$$R^2 = 1 - \frac{SSR}{TSS}$$

SSR – Sum of Squares of residuals
TSS – Total Sum of Squares

Adjusted R-Squared – Adjusted R-Squared or Adjusted R^2 also provides the goodness of fit like R^2. The difference is that the Adjusted R-Squared is adjusted for the number of predictors for the model. The value of the Adjusted R-Squared will decrease with the increase in the number of irrelevant variables.

The formula for Adjusted R-Squared is given as:

$$R^2{}_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

R^2 – R-Squared
p – Number of predictors
N – Total number of samples

These metrics are available in the sklearn library which is a Machine Learning library in Python.

The dataset being used for this comparative study is the house sale prices for King County, which includes Seattle. The dataset includes the houses sold between May 2014 to May 2015. The dataset is a no copyright dataset that can be downloaded from Kaggle.

The dataset comprises 21 columns, which includes the date – which is the id, date of sale, number of bedrooms, number of bathrooms, and so on. Out of these, the variable for which we have to perform prediction is the price – the price of the house.

id	date	price	bedrooms	bathrooms
7129300520	20141013T000000	221900.0	3	1.00
6414100192	20141209T000000	538000.0	3	2.25
5631500400	20150225T000000	180000.0	2	1.00
2487200875	20141209T000000	604000.0	4	3.00
1954400510	20150218T000000	510000.0	3	2.00

Figure 2a: Dataset

sqft_living	sqft_lot	floors	waterfront	view	condition	grade
1180	5650	1.0	0	0	3	7
2570	7242	2.0	0	0	3	7
770	10000	1.0	0	0	3	6
1960	5000	1.0	0	0	5	7
1680	8080	1.0	0	0	3	8

Figure 2b: Dataset Continued

sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
1180	0	1955	0	98178
2170	400	1951	1991	98125
770	0	1933	0	98028
1050	910	1965	0	98136
1680	0	1987	0	98074

Figure 2c: Dataset Continued

lat	long	sqft_living15	sqft_lot15
47.5112	-122.257	1340	5650
47.7210	-122.319	1690	7639
47.7379	-122.233	2720	8062
47.5208	-122.393	1360	5000
47.6168	-122.045	1800	7503

Figure 2d: Dataset continued

To understand the correlation between the variables of the dataset, let us consider the following heatmap.

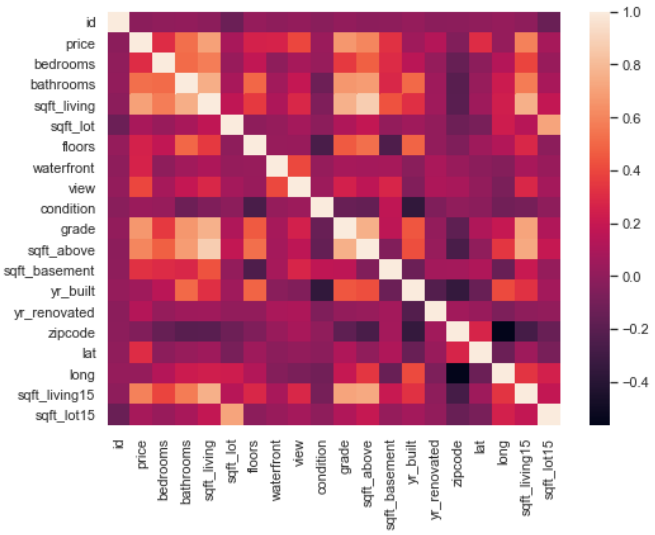

Figure 3: Heatmap showing correlation between variables

The paper presents a comparative assessment or evaluation of the different regression models based on the above metrics for this dataset.

II. LITERATURE SURVEY

In [5] Serkan ETİ et al have presented a study, comparing the regression models explaining the profitability base on financial data. They have made use of and evaluated multiple linear regression and logistic regression. From their study, they concluded that the multiple linear regression model returned an R^2 value of 0.912 and Logistic regression returned an R^2 value of 0.47 and multiple linear regression gave a better performance. They concluded that the optimal model must be selected based on the purpose of the analysis.

In [6] Mohan S Acharya et al have provided a comparative study of regression models to predict graduate admissions. They have compared different regression algorithms such as Linear Regression, SVR (Support Vector Regression), Decision Trees Regression, and Random Forest Regression, for the given profile of the student. To select the best model, they have computed the error functions for these models and compared their performance. They concluded that Linear regression performed best on their dataset with a low Mean Squared Error value and a high R^2 value.

In [7] J.García-Gutiérrez et al have proposed a comparison of machine learning regression algorithms for the LiDAR-derived estimation of forest variables. In this paper, they have compared the classic Multiple Linear Regression with regression techniques in machine learning (such as neural networks, SVMs, nearest neighbor, and ensemble methods such as random forests) with emphasis on regression trees. They concluded that Support Vector Regression with

kernels outperforms the other techniques statistically. Also, it was found that the machine learning techniques outperformed the classic MLR technique.

In [8] Gursev Pirge has presented a comparative study of regression analysis algorithms. The author has evaluated the algorithms based on the metrics such as MAE, MSE, RMSE, and R-squared methods. The author has compared 5 different regression algorithms such as the Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression, and XGBOOST Regression. From this study, the author concluded that the best results were obtained using the XGBoost algorithm, which was followed by the Random Forest method. He also concluded that the findings of the linear, Lasso, and Ridge regression methods were quite similar.

III. METHODOLOGY

Initially, we divide the dataset into testing and training data. Training data is used to allow the machines to learn and the testing data is used to test the model. We can do so using sklearn's train_test_spilt. Once the data is divided into testing and training data, we can fit these values to the different regression algorithms and perform predictions. As discussed in the earlier section, we evaluate the performance based on metrics such as MSE, RMSE, MAE, R^2, and Adjusted R^2.

1. Linear Regression

Linear regression is a linear procedure to model the relationship between a dependent variable and an independent variable. The relationship obtained will be linear. That is the equation obtained from a linear regression model will be of the form y = mx +c. The regression lines are of two types a positive regression line that is obtained when y increases with increase in x and a negative regression line that is obtained when y decreases with increase in x.

There are two types of linear regression depending on the number of independent variables being used.

- When the number of independent variables is one we call this a simple linear regression. In this case, the formula is of the form
  $y = \beta_0 + \beta_1 x_1$
  Where y is the dependent variable, x is the independent variable, β_0 is the y-intercept and β_1 is the regression coefficient representing the change in y with respect to the change in x also called the slope.
- When there is more than one independent variable this type of linear regression is called the multiple linear regression. In this case, the formula is given as
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$
  Where y is the dependent variable, *x1, x2… x_i* are the independent variables, β_0 is the y-intercept and β_i is the regression coefficient representing the change in y with respect to the change in x_i.

In [9] Khushbu Kumari et al have presented the basic concepts of linear regression and how linear regression can be implemented using SPSS and Excel. According to the authors linear regression is important for two main reasons – one, it helps in examining the strength of association or relationship between a dependent and independent variable, and two because it adjusts for the effects of the confounders or the covariates. In this paper the authors have also provided the assumptions for linear regression and examples to illustrate Linear Regression.

For our application, we make use of multiple linear regression. We predict the price attribute against the other independent attributes. Let us consider the plot of actual values and predicted values. This plot will tell us how close the two values were.

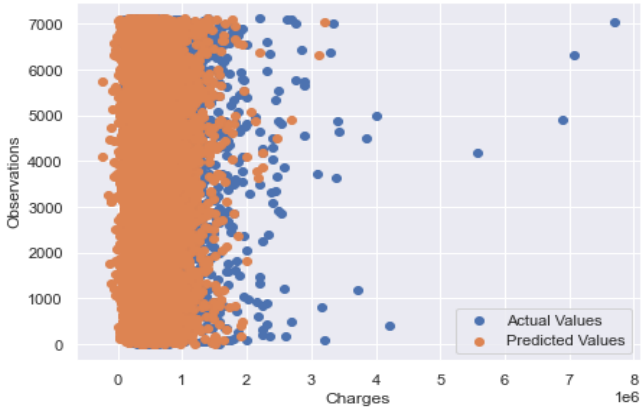

Figure 5: Plot of actual and predicted values for Linear Regression

The values for various metrics are charted down in the following table.

Metric	Value
MSE	45563379684.940384
RMSE	213455.80264996403
MAE	123832.69104834474
R^2	0.6846362760787215
Adjusted R^2	0.6838383358157776

Table 1: Metrics for Linear Regression

2. Ridge Regression

Ridge regression is used to overcome the problem of overfitting. Overfitting happens when the model has higher or better performance with the training dataset and poor performance with the testing dataset. In [10] Xue Ying has given an overview of overfitting from the perspective of their causes and methods to solve this. The author has explained 4 major strategies - early-stopping, network-reduction, data-expansion, and regularization. Ridge regression makes use of L2 regularization.

Ridge regression penalizes those features that have higher slopes, that is we add a penalty to the square of the magnitude of coefficients, to reduce the cost function.

Ridge regression is a model tuning technique using which L2 regularization can be performed.

The mathematical for ridge regression is given as:

$$\sum_{i=1}^{n}(y_i - \hat{y_i}) + \lambda m^2$$

Here, λ is the penalty and m denotes the magnitude of the coefficient or the slope. The value of λ should always be greater than 0. That is $\lambda > 0$.

Following results were obtained when we made use of Ridge Regression as the model for prediction for our dataset.

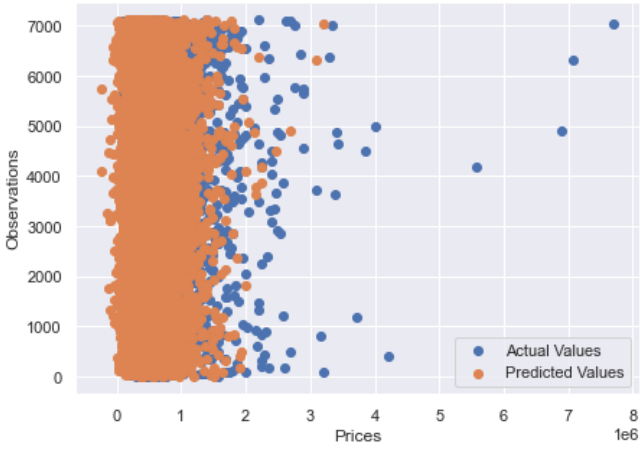

Figure 6: Plot of actual and predicted values for Ridge Regression

Metric	Value
MSE	45554694203.61722
RMSE	213435.45676296903
MAE	123807.43082224767
R^2	0.6846963920260691
Adjusted R^2	0.6838986038698236

Table 2: Metrics for Ridge Regression

3. Lasso Regression

The aim of using Lasso regression is somewhat similar to that of Ridge regression, that is to reduce overfitting. However, additionally, Lasso regression also solves the purpose of feature selection. Lasso regression makes use of L1 Regularization and is specially used when there is more number of features. L1 regularisation adds a penalty proportional to the absolute value of the coefficient's magnitude. The formula is given as:

$$\sum_{i=1}^{n}(y_i - \hat{y_i}) + \lambda|m|$$

Here again, λ is the penalty and m denotes the magnitude of the coefficient or the slope. The value of λ should always be greater than 0. That is $\lambda > 0$.

Following results were obtained when we made use of Lasso Regression as the model for prediction for our dataset.

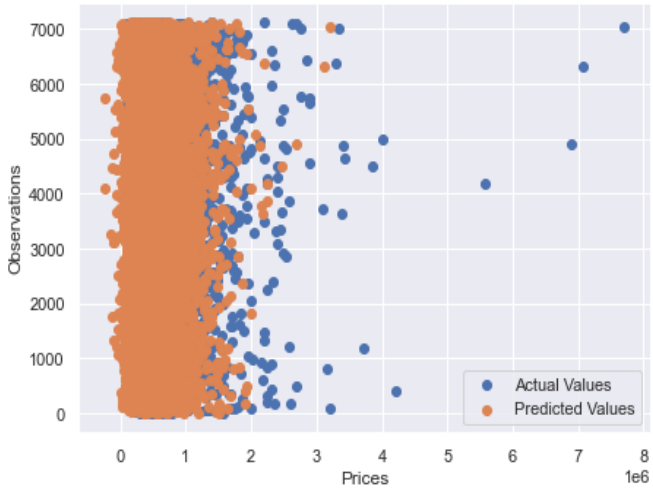

Figure 7: Plot of actual and predicted values for Lasso Regression

Metric	Value
MSE	45563226388.53266
RMSE	213455.44356734655
MAE	123831.67538767449

R^2	0.6846373371090124
Adjusted R^2	0.6838393995307108

Table 3: Metrics for Lasso Regression

4. Support Vector Regression

The Support Vector Regression is based on the concept of Support Vector Machines. The Support Vector Machines are supervised Learning algorithms that are used to perform classification. They make use of a hyperplane to fit or map the dependent and independent variables. The concept behind SVR is to compute a linear regression function in a high-dimensional feature space where the input data are mapped using a nonlinear function [11].

SVR or Support Vector Regression is based on the principle of SVMs and aims at finding the line of best fit for the given data. Here, the hyperplane serves as the line of best fit. A mathematical function, called the kernel is made use for transforming the data into the required from and the boundaries are draw at ε distance which defines the margin between the data points.

Following results were obtained when we made use of Support Vector Regression as the model for prediction for our dataset.

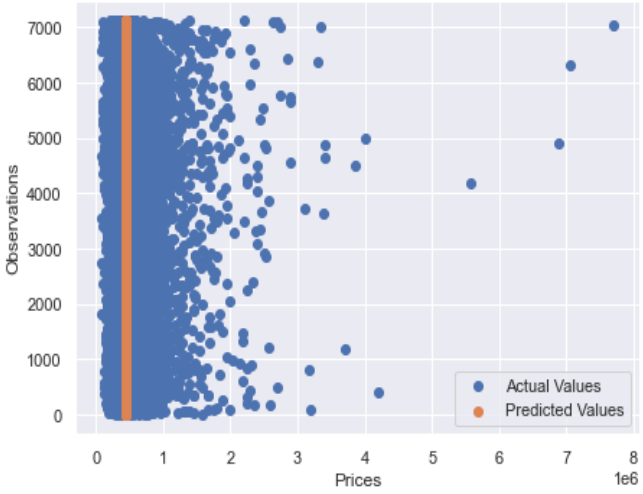

Figure 8: Plot of actual and predicted values for Support Vector Regression

Metric	Value
MSE	152589483778.5077
RMSE	390627.039231167
MAE	223235.2604451575
R^2	-0.05613736663877744
Adjusted R^2	-0.05880962874160267

Table 4: Metrics for Support Vector Regression

5. Polynomial Regression

Polynomial Regression can be considered as one of the cases of linear regression. In this method the relationship between the dependent and the independent variable is modelled as a nth degree polynomial. Even though the polynomial regression allows for non-linear relationship it is still considered under linear regression. In polynomial regression the original characteristics are transformed into Polynomial features of the desired degree (2,3,...,n) and then modelled using a linear model. The formula for polynomial regression is given as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \beta_1 x_n^n$$

Here, β_0 is the y intercept, $\beta_1, \beta_2 \ldots \beta_n$ are the magnitude of coefficients.

Following results were obtained when we made use of Polynomial Regression as the model for prediction for our dataset.

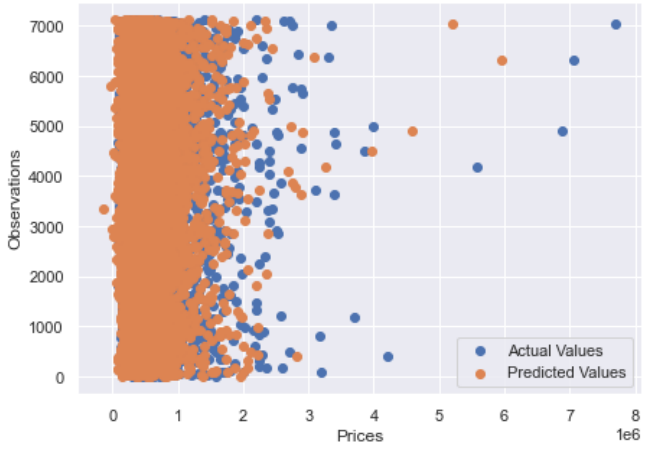

Figure 9: Plot of actual and predicted values for Polynomial Regression

Metric	Value
MSE	25935368489.23436
RMSE	161044.616455299
MAE	99958.6195001264
R^2	0.8204901733674757
Adjusted R^2	0.8200359736374524

Table 4: Metrics for Polynomial Regression

IV. COMPARISION

In this section let us compare the performance of each of the regression algorithms that we have used to conclude which algorithm performed well. Support Vector regression had the highest value hence the model did not perform well for the dataset.

Based on MSE – We know that lesser the MSE, better is the accuracy. The following graph shows the MSE values of each of the models.

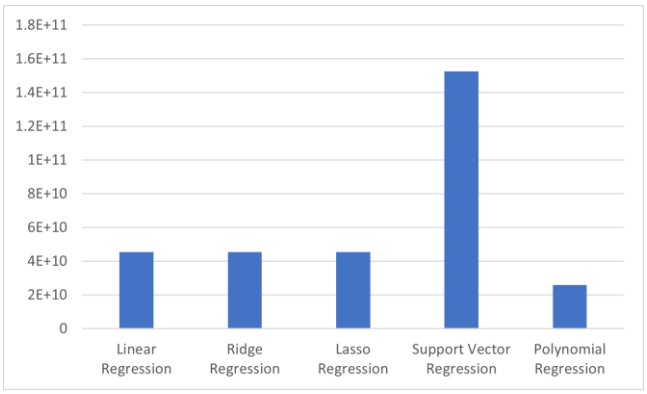

Figure 10: Comparison of MSE

From the graph we can conclude that Polynomial Regression has the least MSE Value, hence it performed well.

Based on RMSE – We know that lesser the RMSE, better is the accuracy. The following graph shows the RMSE values of each of the models.

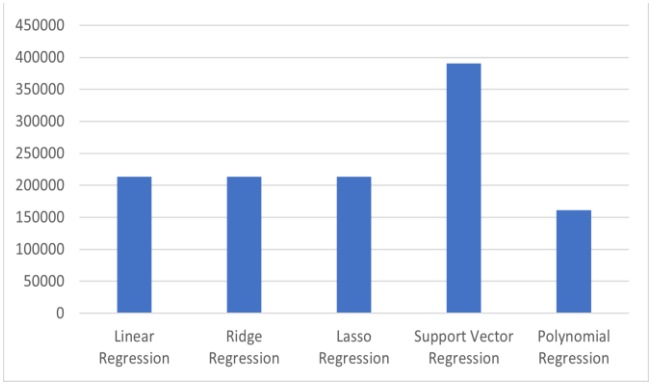

Figure 11: Comparison of RMSE

From the graph we can conclude that Polynomial Regression has the least RMSE Value, hence it performed well. Support Vector regression had the highest value hence the model did not perform well for the dataset.

Based on MAE – We know that lesser the MAE, better is the accuracy. The following graph shows the MAE values of each of the models.

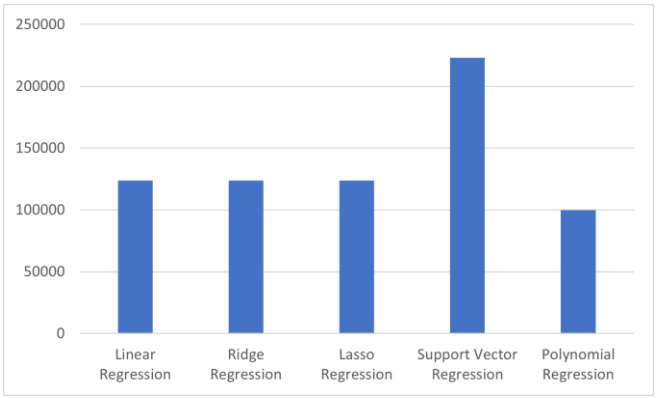

Figure 12: Comparison of MAE

From the graph we can conclude that Polynomial Regression has the least MAE Value, hence it performed well. Support Vector regression had the highest value hence the model did not perform well for the dataset.

Based on R^2 - We know that higher the R^2, better is the accuracy. The following graph shows the R^2 values of each of the models.

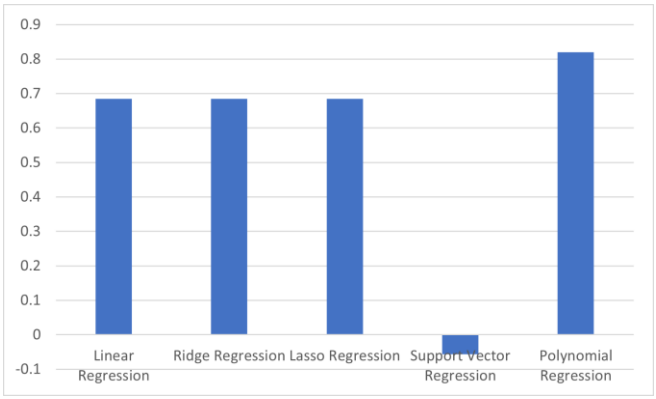

Figure 13: Comparison of R^2

From the graph we can conclude that Polynomial Regression has the highest R^2 Value, hence it performed well. Support Vector regression had the least value hence the model did not perform well for the dataset.

Based on Adjusted R^2 - We know that higher the adjusted R^2, better is the accuracy. The following graph shows the R^2 values of each of the models.

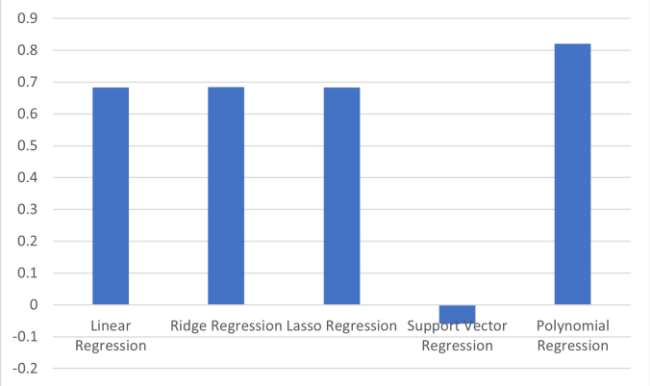

Figure 14: Comparison of Adjusted R^2

From the graph we can conclude that Polynomial Regression has the highest Adjusted R^2 Value, hence it performed well. Support Vector regression had the least value hence the model did not perform well for the dataset.

V. CONCLUSION

The main aim of this paper was to compare and assess the different regression models based on the model evaluation metrics. The dataset used was the house sale prices for King County, which includes Seattle. From the study we were able to explore the different regression algorithms such as the Linear regression (Multiple Linear Regression), Ridge Regression, Lasso Regression, Support Vector Regression and the Polynomial Regression, and applied these algorithms on the dataset. The study was performed in a python environment and the models were obtained from the scikit learn library.

From this study, on our dataset, polynomial regression performed the best with around 80% R^2 score. On the contrary Support Vector Regression performed the least.

However, the choice of regression algorithm should be based on the type of data being considered, the distribution of the data and the parameters under considerations.

ACKWNOWLEDGEMENTS

I convey my sincere thanks to Prof. Dheeraj D, Assistant Professor, Department of Information Science and Engineering, Global Academy of Technology, for guidance and mentorship and guiding through the process.

Further I also put forward my heartfelt thanks to @harlfoxem for making the copyright free dataset available on Kaggle.

REFERENCES

[1]. A. M. TURING, I.—COMPUTING MACHINERY AND INTELLIGENCE, Mind, Volume LIX, Issue 236, October 1950, Pages 433–460,

[2]. Nasteski, Vladimir. (2017). An overview of the supervised machine learning methods. HORIZONS.B. 4. 51-62. 10.20544/HORIZONS.B.04.1.17.P05.

[3]. Uysal, İlhan & Güvenir, Halil Altay. (1999). An overview of regression techniques for knowledge discovery. Knowledge Engineering Review. 14. 319-340. 10.1017/S026988899900404X.

[4]. Simon P. Neill, M. Reza Hashemi, Chapter 8 - Ocean Modelling for Resource Characterization, Editor(s): Simon P. Neill, M. Reza Hashemi, In E-Business Solutions, Fundamentals of Ocean Renewable Energy, Academic Press, 2018, Pages 193-235, ISBN 9780128104484, https://doi.org/10.1016/B978-0-12-810448-4.00008-2.

[5]. Eti, Serkan & İnel, Mehmet. (2016). A research on comparison of regression models explaining the profitability base on financial data. International Journal of Business and Management. 4. 470-475.

[6]. Acharya, Mohan & Armaan, Asfia & S Antony, Aneeta. (2019). A Comparison of Regression Models for Prediction of Graduate Admissions. 1-5. 10.1109/ICCIDS.2019.8862140.

[7]. J. García-Gutiérrez, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables, Neurocomputing, Volume 167, 2015, Pages 24-31, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2014.09.091.

[8]. Gursev Pirge, Comparison of Regression Analysis Algorithms, Dec 2020, https://gursev-pirge.medium.com/comparison-of-regression-analysis-algorithms-db710b6d7528

[9]. Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

[10]. Ying, Xue. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 1168. 022022. 10.1088/1742-6596/1168/2/022022.

[11]. Basak, Debasish & Pal, Srimanta & Patranabis, Dipak. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews. 11.