

Bereket Eshete Kebede
UID - U00827234

M.S. Student
University of Memphis, Fall 2022
COMP 7745 Machine Learning

Implementing a Linear Regression Model
Implementing a Logistic Regression Model

Instructor: Prof. Zahangir Alom
Assignment #1 Report

Table of contents

Implementing a Linear Regression Model	3
Introduction	3
Methodology	3
Model descriptions	4
Experiment and Results	4
(a) Database	4
(b) Training and testing logs	6
(c) Discussion and comparison	6
Conclusion	7
References	7
Implementing a Logistic Regression Model	8
Introduction	8
Methodology & Model descriptions	8
Experiment and Results	9
(a) Database	9
(b) Training and testing logs	9
(c) Discussion and comparison	9
Conclusion	9
References	9

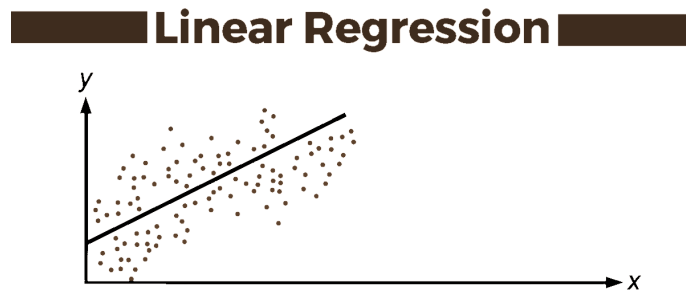
Implementing a Linear Regression Model

Introduction

In this report we will Implement a linear regression model to predict the house price for the provided datasets. Linear regression analysis is used to predict the value of a variable based on the value of another variable. If you ask a prospective homeowner to describe their ideal home, they usually won't start by talking about how high the basement ceiling is or how close the house is to an east-west railroad. However, the dataset for this report shows that factors other than the number of beds and a white picket fence have a significant impact on price negotiations. We will estimate the final price of each residential property in Ames, Iowa, given the 79 explanatory factors that describe (nearly) every feature of residential properties there [1].

Methodology

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables [2]. Linear Regression is the simplest algorithm in machine learning and it can be trained in different ways [3].



Many regression models rely on distance metrics to determine the convergence to the best result. Usually the metrics used are the Mean Average Error (MAE), the Mean Squared Error (MSE) or the Root Mean Squared Error (RMSE) [3,4].

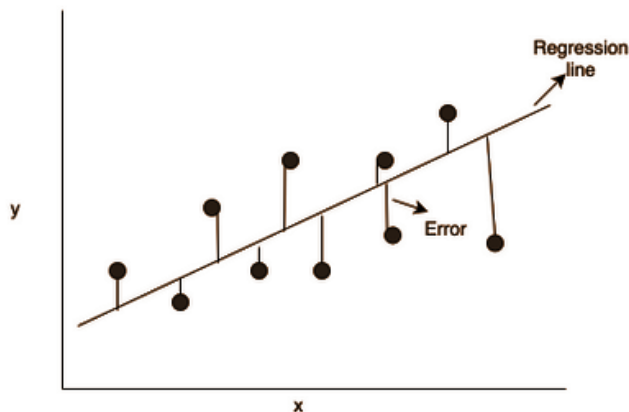
$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i^{real} - y_i^{pred}|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i^{real} - y_i^{pred})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{real} - y_i^{pred})^2}$$

Model descriptions

The regression model operates by minimizing the error between the point distribution and the linear prediction line as shown in the figure below [3]



Experiment and Results

(a) Database

train.csv - the training set

test.csv - the test set

data_description.txt - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here.

<p>SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.</p> <p>MSSubClass: The building class</p> <p>MSZoning: The general zoning classification</p> <p>LotFrontage: Linear feet of street connected to property</p> <p>LotArea: Lot size in square feet</p> <p>Street: Type of road access</p> <p>Alley: Type of alley access</p> <p>LotShape: General shape of property</p> <p>LandContour: Flatness of the property</p> <p>Utilities: Type of utilities available</p> <p>LotConfig: Lot configuration</p>	<p>BsmtUnfSF: Unfinished square feet of basement area</p> <p>TotalBsmtSF: Total square feet of basement area</p> <p>Heating: Type of heating</p> <p>HeatingQC: Heating quality and condition</p> <p>CentralAir: Central air conditioning</p> <p>Electrical: Electrical system</p> <p>1stFlrSF: First Floor square feet</p> <p>2ndFlrSF: Second floor square feet</p> <p>LowQualFinSF: Low quality finished square feet (all floors)</p> <p>GrLivArea: Above grade (ground) living area square feet</p> <p>BsmtFullBath: Basement full bathrooms</p>
---	--

<p>LandSlope: Slope of property</p> <p>Neighborhood: Physical locations within Ames city limits</p> <p>Condition1: Proximity to main road or railroad</p> <p>Condition2: Proximity to main road or railroad (if a second is present)</p> <p>BldgType: Type of dwelling</p> <p>HouseStyle: Style of dwelling</p> <p>OverallQual: Overall material and finish quality</p> <p>OverallCond: Overall condition rating</p> <p>YearBuilt: Original construction date</p> <p>YearRemodAdd: Remodel date</p> <p>RoofStyle: Type of roof</p> <p>RoofMatl: Roof material</p> <p>Exterior1st: Exterior covering on house</p> <p>Exterior2nd: Exterior covering on house (if more than one material)</p> <p>MasVnrType: Masonry veneer type</p> <p>MasVnrArea: Masonry veneer area in square feet</p> <p>ExterQual: Exterior material quality</p> <p>ExterCond: Present condition of the material on the exterior</p> <p>Foundation: Type of foundation</p> <p>BsmtQual: Height of the basement</p> <p>BsmtCond: General condition of the basement</p> <p>BsmtExposure: Walkout or garden level basement walls</p> <p>BsmtFinType1: Quality of basement finished area</p> <p>BsmtFinSF1: Type 1 finished square feet</p> <p>BsmtFinType2: Quality of second finished area (if present)</p> <p>BsmtFinSF2: Type 2 finished square feet</p>	<p>BsmtHalfBath: Basement half bathrooms</p> <p>FullBath: Full bathrooms above grade</p> <p>HalfBath: Half baths above grade</p> <p>Bedroom: Number of bedrooms above basement level</p> <p>Kitchen: Number of kitchens</p> <p>KitchenQual: Kitchen quality</p> <p>TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)</p> <p>Functional: Home functionality rating</p> <p>Fireplaces: Number of fireplaces</p> <p>FireplaceQu: Fireplace quality</p> <p>GarageType: Garage location</p> <p>GarageYrBlt: Year garage was built</p> <p>GarageFinish: Interior finish of the garage</p> <p>GarageCars: Size of garage in car capacity</p> <p>GarageArea: Size of garage in square feet</p> <p>GarageQual: Garage quality</p> <p>GarageCond: Garage condition</p> <p>PavedDrive: Paved driveway</p> <p>WoodDeckSF: Wood deck area in square feet</p> <p>OpenPorchSF: Open porch area in square feet</p> <p>EnclosedPorch: Enclosed porch area in square feet</p> <p>3SsnPorch: Three season porch area in square feet</p> <p>ScreenPorch: Screen porch area in square feet</p> <p>PoolArea: Pool area in square feet</p> <p>PoolQC: Pool quality</p> <p>Fence: Fence quality</p> <p>MiscFeature: Miscellaneous feature not covered in other categories</p> <p>MiscVal: \$Value of miscellaneous feature</p> <p>MoSold: Month Sold</p> <p>YrSold: Year Sold</p> <p>SaleType: Type of sale</p> <p>SaleCondition: Condition of sale</p>
--	--

Table 1. Dataset features and the descriptions

Data preprocessing:

We need to fill in the gaps where there is no data available in the columns. Hence, we create dummy variables for the categorical features and replace the numeric missing values (NaN's) with the mean of their respective columns.

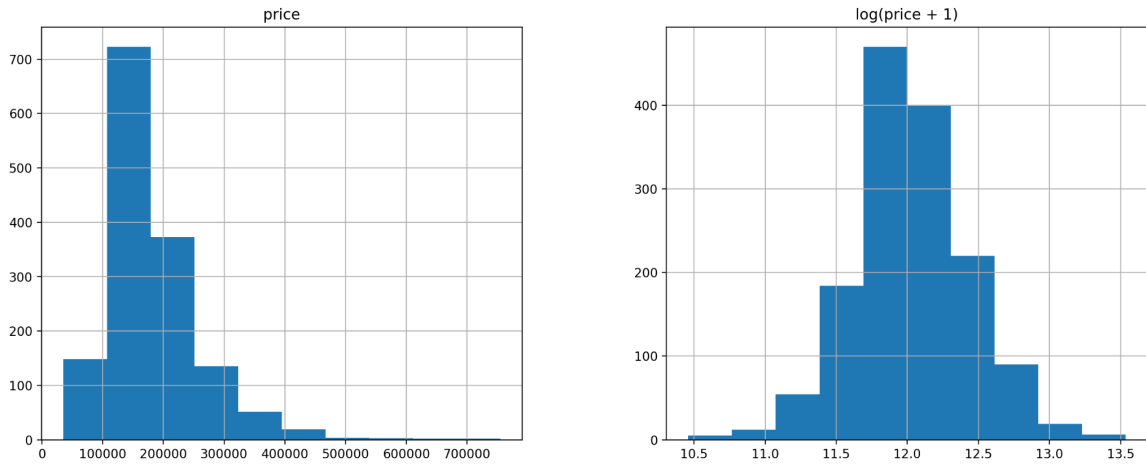


Figure 1. Data histogram plot

(b) Training and testing logs

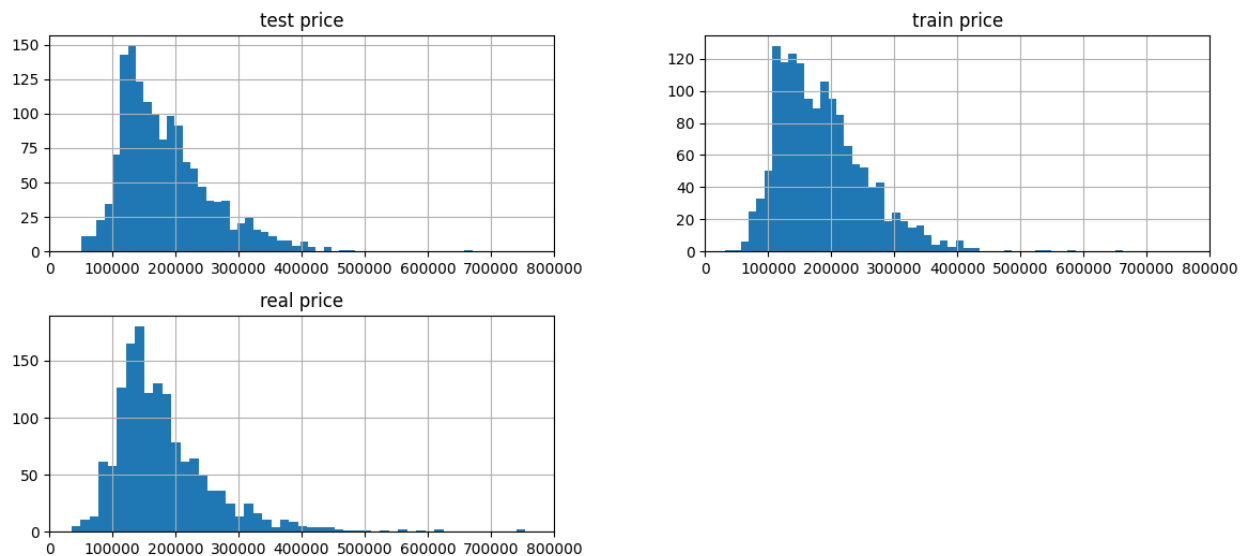


Figure 2. Training and testing plot

(c) Discussion and comparison

Table 2. Analysis of the results with Means Squared Error (MSE), absolute errors, and Root MSE (RMSE).

MAE:	79489.39
MSE:	11379036010.16
RMSE:	106672.56

Conclusion

Linear regression can be used to predict the price of houses with a good accuracy as shown in figure 2.

References

- [1] <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
- [2] https://en.wikipedia.org/wiki/Linear_regression
- [3] <https://community.cloudera.com/t5/Community-Articles/Understanding-Linear-Regression/ta-p/281391>
- [4] <https://towardsdatascience.com/comparing-robustness-of-mae-mse-and-rmse-6d69da870828>

Implementing a Logistic Regression Model

Introduction

Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems. Linear regression provides a continuous output but Logistic regression provides a discrete output.

- Elastic Net - In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods.
- Lasso Net - is a method for feature selection in neural networks, to enhance interpretability of the final network.
- Ridge Net - Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated.

Methodology & Model descriptions

In elastic Net Regularization we added the both terms of L_1 and L_2 to get the final loss function. This leads us to reduce the following loss function. Where alpha is between 0 and 1. when alpha = 1, It reduces the penalty term to L_1 penalty and if $\alpha = 0$, it reduces that term to L_2 penalty.

$$L_{\text{elastic-Net}}(\hat{\beta}) = \left(\sum (y - x_i^T \hat{\beta})^2 \right) / 2n + \lambda \left((1 - \alpha) / 2 * \sum_{j=1}^m \hat{\beta}_j^2 + \alpha * \sum_{j=1}^m \|\hat{\beta}_j\| \right)$$

Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares, $\alpha \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α , the greater the amount of shrinkage and thus the coefficients become more robust to collinearity. Ridge regression is an L_2 penalized model. Add the squared sum of the weights to the least-squares cost function.

$$\min_w \|Xw - y\|_2^2 + \alpha \|w\|_2^2$$

Lasso regression, is a linear model that estimates sparse coefficients. Mathematically, it consists of a linear model trained with ℓ_1 prior as regularizer. The objective function to minimize is:

$$\min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha \|w\|_1$$

Experiment and Results

(a) Database

Similar to Task (A), train.csv and test.csv

(b) Training and testing logs

Table 3. Metric values for the three regression models

	RMSE	MSE	MAE
Ridge	0.131	0.018	0.085
Lasso	0.123	0.015	0.080
Elastic	0.204	0.042	0.149

(c) Discussion and comparison

In this report we covered the common linear regression models (Ridge, Lasso and ElasticNet). We saw the representation used by the model. We discussed rules of thumb to consider when preparing data for use with linear regression and finally we saw how to evaluate a linear regression model.

Conclusion

We implemented three logistic regression models and evaluated their performance metrics.

References

- [1] <https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/>
- [2] Elastic Net, https://en.wikipedia.org/wiki/Elastic_net_regularization
- [3] Lasso Net, <https://lassonet.ml/>

[3] Ridge Net,

<https://www.kaggle.com/code/faressayah/practical-introduction-to-10-regression-algorithm/notebook#%E2%9C%94%E2%8F%A5-Ridge-Regression>