Regularized Regression

When training a machine learning model, the model can be easily overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit the model to our test set. Regularization techniques help reduce the possibility of overfitting and help us obtain an optimal model.

When it comes to training models, there are two major problems one can encounter: overfitting and underfitting.

- Overfitting happens when the model performs well on the training set but not so well on unseen (test) data.
- Underfitting happens when it neither performs well on the train set nor on the test set.

Particularly, regularization is implemented to avoid overfitting of the data, especially when there is a large variance between train and test set performances.
There are different ways of reducing model complexity and preventing overfitting in linear models. This includes ridge and lasso regression models.

Lasso Regression:

This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression. Lasso regression is also referred to as **L1 Regularization**.
In this shrinkage technique, the coefficients determined in the linear model from equation 1.1. above are shrunk towards the central point as the mean by introducing a penalization factor called the alpha α (or sometimes lamda) values.

$$L_1 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 + \lambda \sum_{i=1}^{p} |B_i|$$

Alpha (α) is the penalty term that denotes the amount of shrinkage (or constraint) that will be implemented in the equation. With alpha set to zero, you will find that this is the equivalent of the linear regression model. Larger value penalizes the optimization function. Therefore, lasso regression shrinks the coefficients and helps to reduce the model complexity and multi-collinearity.
Alpha (α) can be any real-valued number between zero and infinity; the larger the value, the more aggressive the penalization is.

- Due to the fact that coefficients will be shrunk towards a mean of zero, less important features in a dataset are eliminated when penalized.
- The shrinkage of these coefficients based on the alpha value provided leads to some form of automatic feature selection, as input variables are removed in an effective approach.
- λ is the regularization parameter that controls the amount of regularization applied.
- $\beta_1, \beta_2, ..., \beta_p$ are the coefficients.

1. **Linear regression model**: LASSO regression starts with the standard linear regression model, which assumes a linear relationship between the independent variables (features) and the dependent variable (target). The linear regression equation can be represented as
   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$ Where:
   - y is the dependent variable (target).
   - $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients (parameters) to be estimated.
   - $x_1, x_2, ..., x_p$ are the independent variables (features).
   - ε represents the error term.
2. **L1 regularization**: LASSO regression introduces an additional penalty term based on the absolute values of the coefficients. The L1 regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter λ
   $L_1 = \lambda * (|\beta_1| + |\beta_2| + ... + |\beta_p|)$ Where:
   - λ is the regularization parameter that controls the amount of regularization applied.
   - $\beta_1, \beta_2, ..., \beta_p$ are the coefficients.

3. **Objective function**: The objective of LASSO regression is to find the values of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values, while also minimizing the L1 regularization term:makefileCopy codeMinimize: RSS + L_1 Where:
    - RSS is the residual sum of squares, which measures the error between the predicted values and the actual values.

4. **Shrinking coefficients**: By adding the L1 regularization term, LASSO regression can shrink the coefficients towards zero. When λ is sufficiently large, some coefficients are driven to exactly zero. This property of LASSO makes it useful for feature selection, as the variables with zero coefficients are effectively removed from the model.
5. **Tuning parameter λ**: The choice of the regularization parameter λ is crucial in LASSO regression. A larger λ value increases the amount of regularization, leading to more coefficients being pushed towards zero. Conversely, a smaller λ value reduces the regularization effect, allowing more variables to have non-zero coefficients.
6. λ denotes the amount of shrinkage.
7. $\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model
8. $\lambda = \infty$ implies no feature is considered i.e, as λ closes to infinity it eliminates more and more features

Ridge Regression:

Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square. Ridge regression is also referred to as **L2 Regularization**.

- It reduces the model complexity by coefficient shrinkage.
- It shrinks the parameters, therefore it is mostly used to prevent multicollinearity.

$$L_2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 + \lambda \sum_{i=1}^{p} B_i^2$$

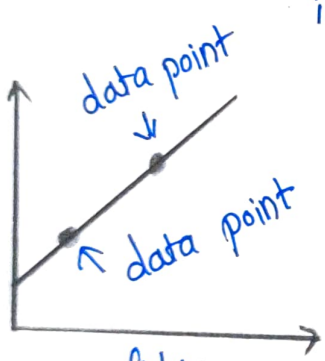
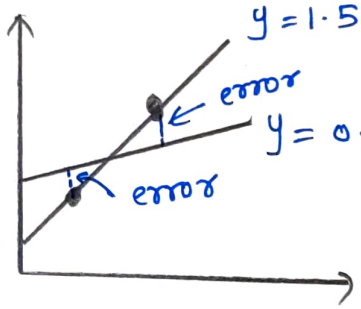


fig. overfitting



loss for line L

$\lambda = 1$

$0 + (1.5)^2$

$= 2.25$

loss for line 2 :

$\lambda = 1$

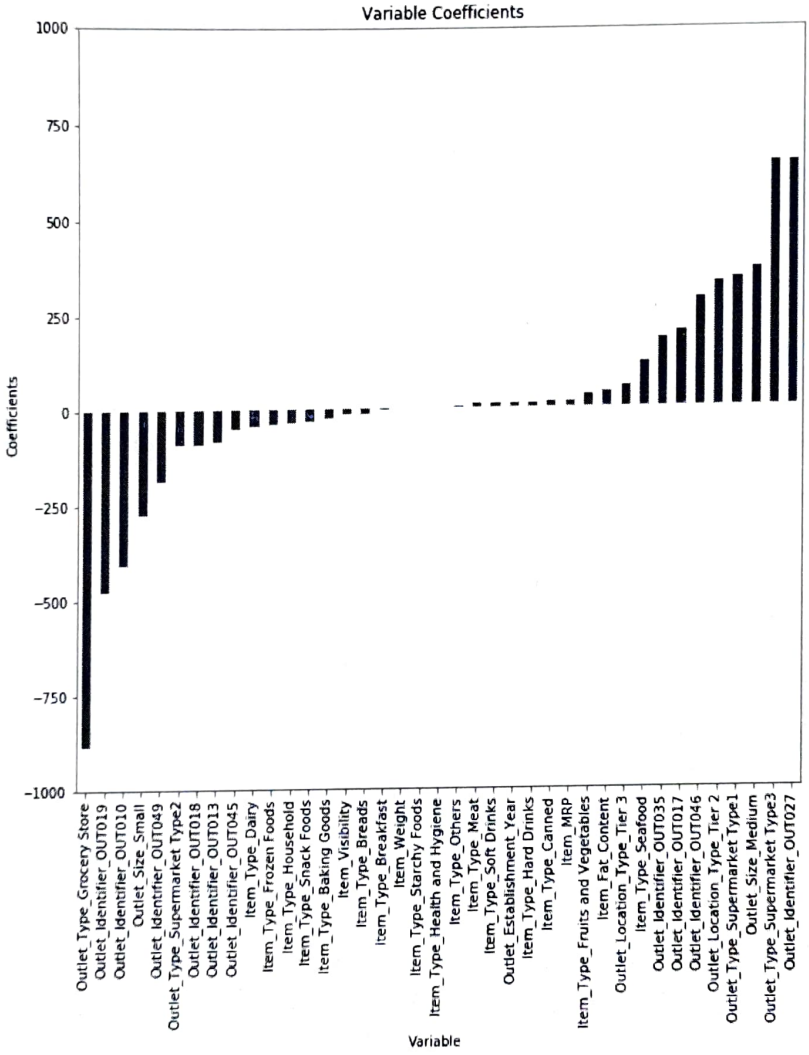
$(2.3 - 0.9 - 1.5)^2 +$

$(5.3 - 2.7 - 1.5)^2 + (0.9)^2$

$= 2.03$

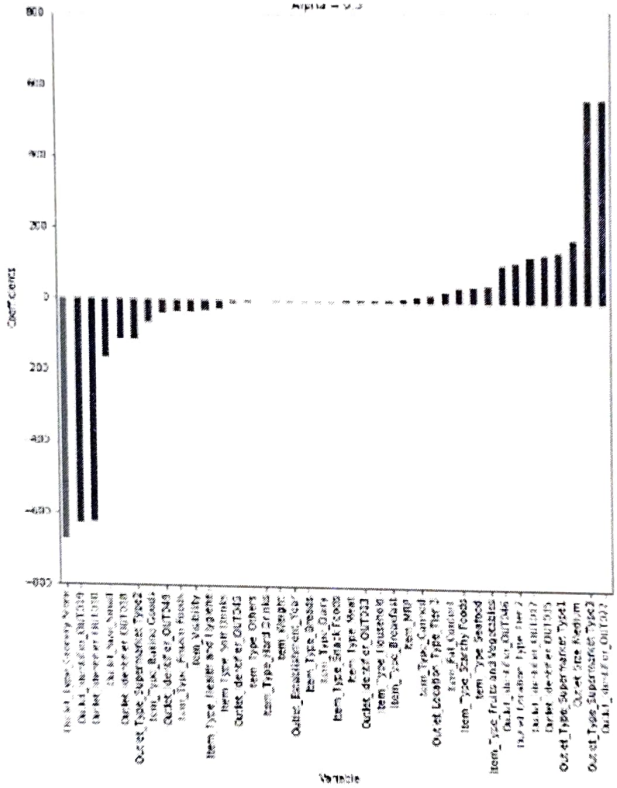
As you can see ML will choose line No. 2 as best fit line, though it is giving error in training. Because due to Rige regression λ is added and cost is reduced.

Look at the coefficients of feature in regression model.



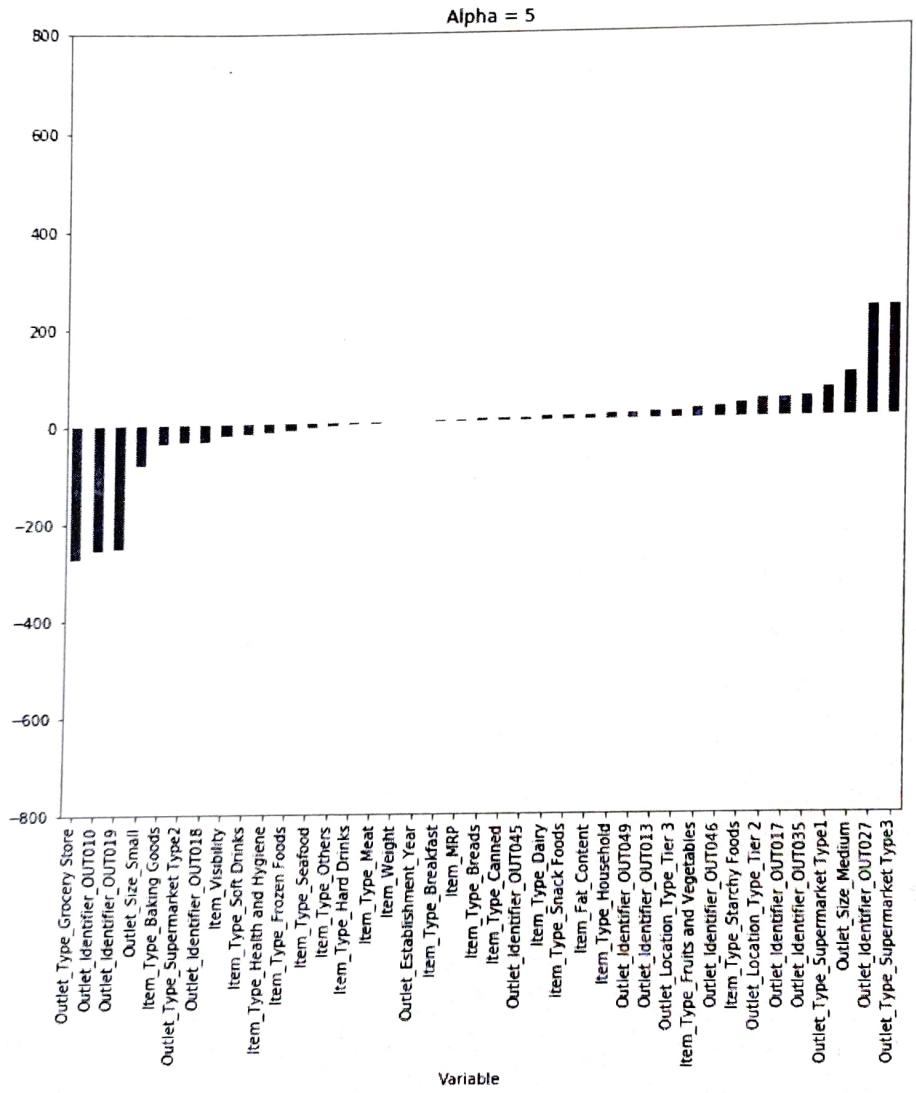
We can see that coefficients of Outlet_Identifier_OUT027 and Outlet_Type_Supermarket_Type3(last 2) is much higher as compared to rest of the coefficients. Therefore the total sales of an item would be more driven by these two features.

consider alpha = 0.5



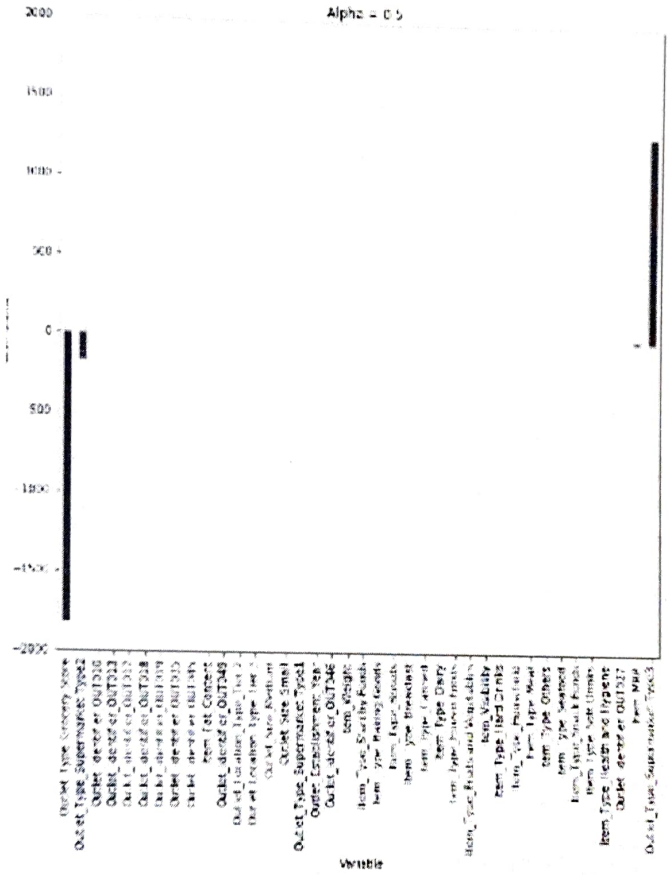
Consider alpha = 5

You can see that, as we increase the value of alpha, the magnitude of the coefficients decreases, where the values reaches to zero but not absolute zero.



For Lasso regression:

For Alpha=0.5



So, we can see that even at small values of alpha, the magnitude of coefficients have reduced a lot. Even at smaller alpha's, our coefficients are reducing to absolute zeroes. Therefore, lasso selects the only some feature while reduces the coefficients of others to zero. This property is known as feature selection and which is absent in case of ridge.