Regularization

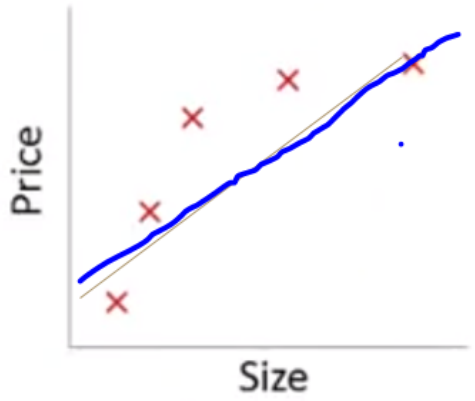
~Abhishek Kumar

Scope

- Problem of Overfitting
- What is regularization?
- Why do we need regularization?
- Types of regularization
- When to use them?
- Model Evaluation

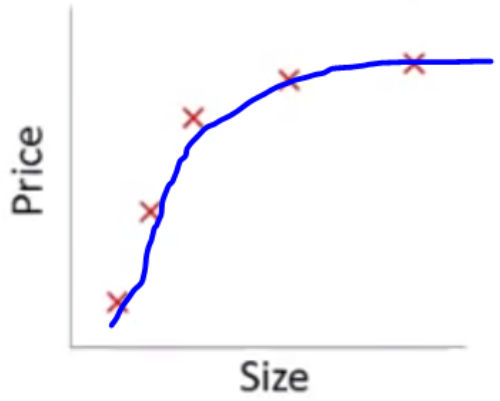
Overfitting

Example: Linear regression (housing prices)



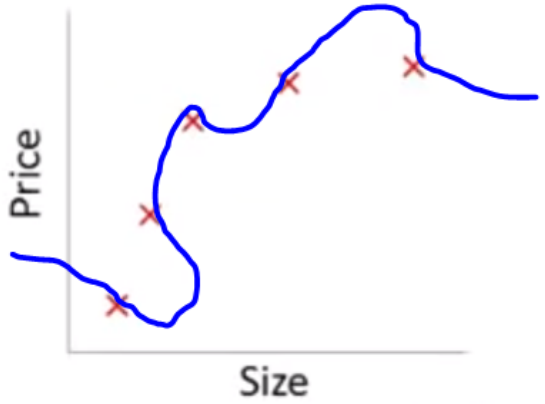
$mx + c$

"Underfitting"
High Bias



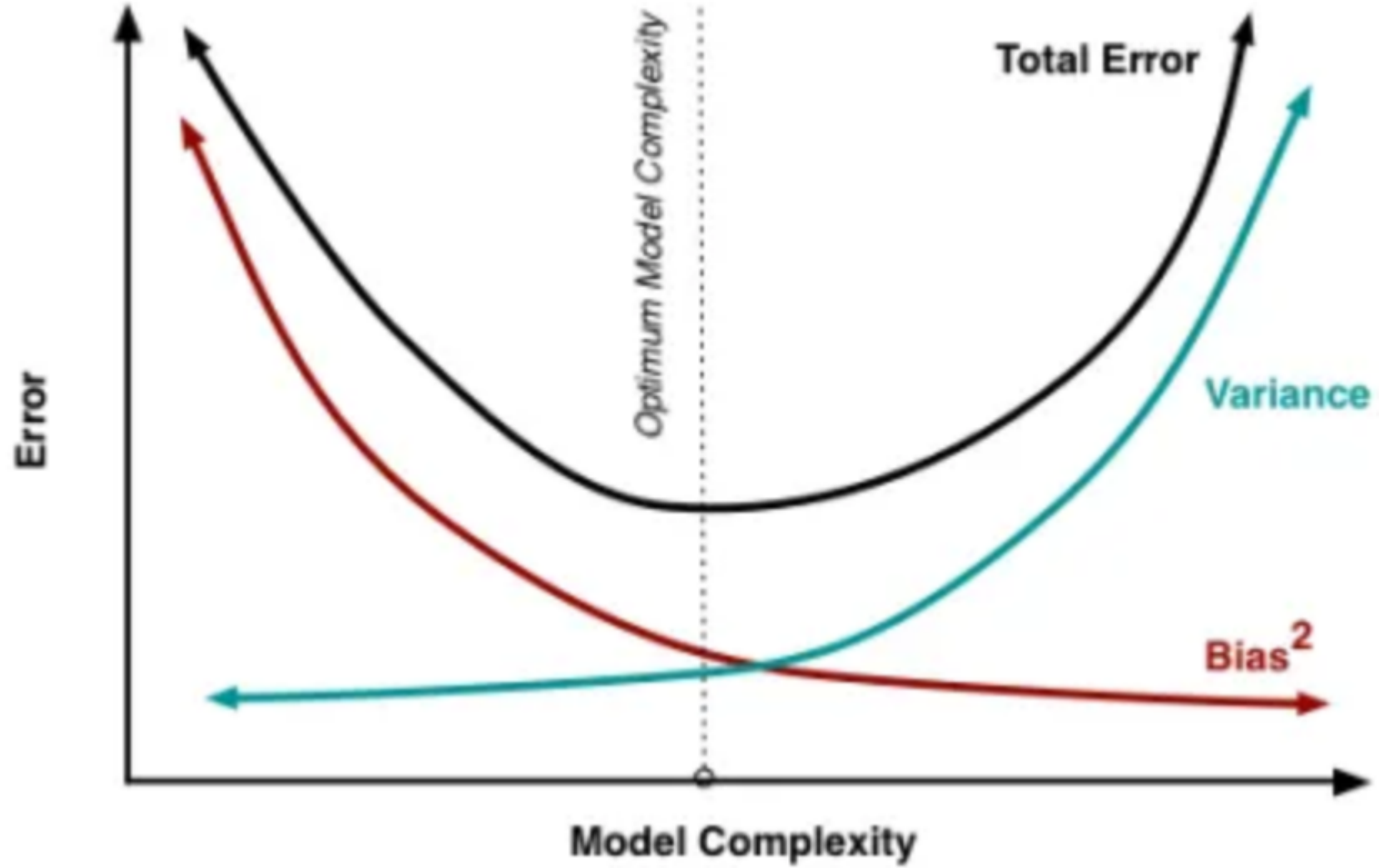
$m_1x + m_2x^2 + c$

"Just right"



$m_1x + m_2x^2 + m_3x^3 + m_4x^4 + c$

"Overfitting"
High Variance


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

Bias & Variance

- High Bias: It means the algorithm has a strong preconception / bias that housing prices are going to vary linearly with their size despite the evidence based on previous data.

- High Variance: If we fit a high order polynomial then hypothesis can fit to any function. So in this case our possible hypothesis is too large or too variable.

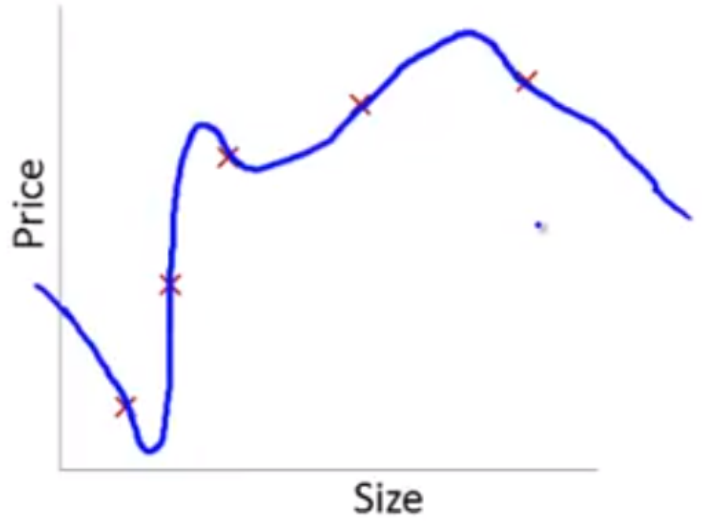
Regularization

- If we have too many features, the learned hypothesis may fit the training set very well(cost function almost zero) , but fail to generalize new examples.

- When we have a high dimensional data set, it would be highly inefficient to use all the variables since some of them might be imparting redundant information.

Addressing overfitting:

$x_1 =$ size of house
$x_2 =$ no. of bedrooms
$x_3 =$ no. of floors
$x_4 =$ age of house
$x_5 =$ average income in neighborhood
$x_6 =$ kitchen size
$\vdots$
x_{100}



Addressing Overfitting

- Reduce number of features
  - Manually select which features to keep
    1. Business Understanding & Domain knowledge
    2. Forward Selection
    3. Backward Elimination
- Regularization

Regularization

- Keep all /some of the features, but reduce magnitude/value of parameter m_j.
- Works well when we have lot of features, each of which contributes a bit to predicting y.

Cost Function

$m_1x + m_2x^2 + c$

$m_1x + m_2x^2 + m_3x^3 + m_4x^4 + c$

We penalize and make m_3 and m_4 really small

"*Everything should be made simple as possible, but not simpler – Albert Einstein*"

- Small values of parameters $m_1, m_2, \ldots\ldots\ldots, m_k$
- "Simple" hypothesis
- Less prone to overfitting

$$J(m) = \frac{1}{2n}\sum_{i=1}^{n}\left(h\ (x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{k} m_j^p$$

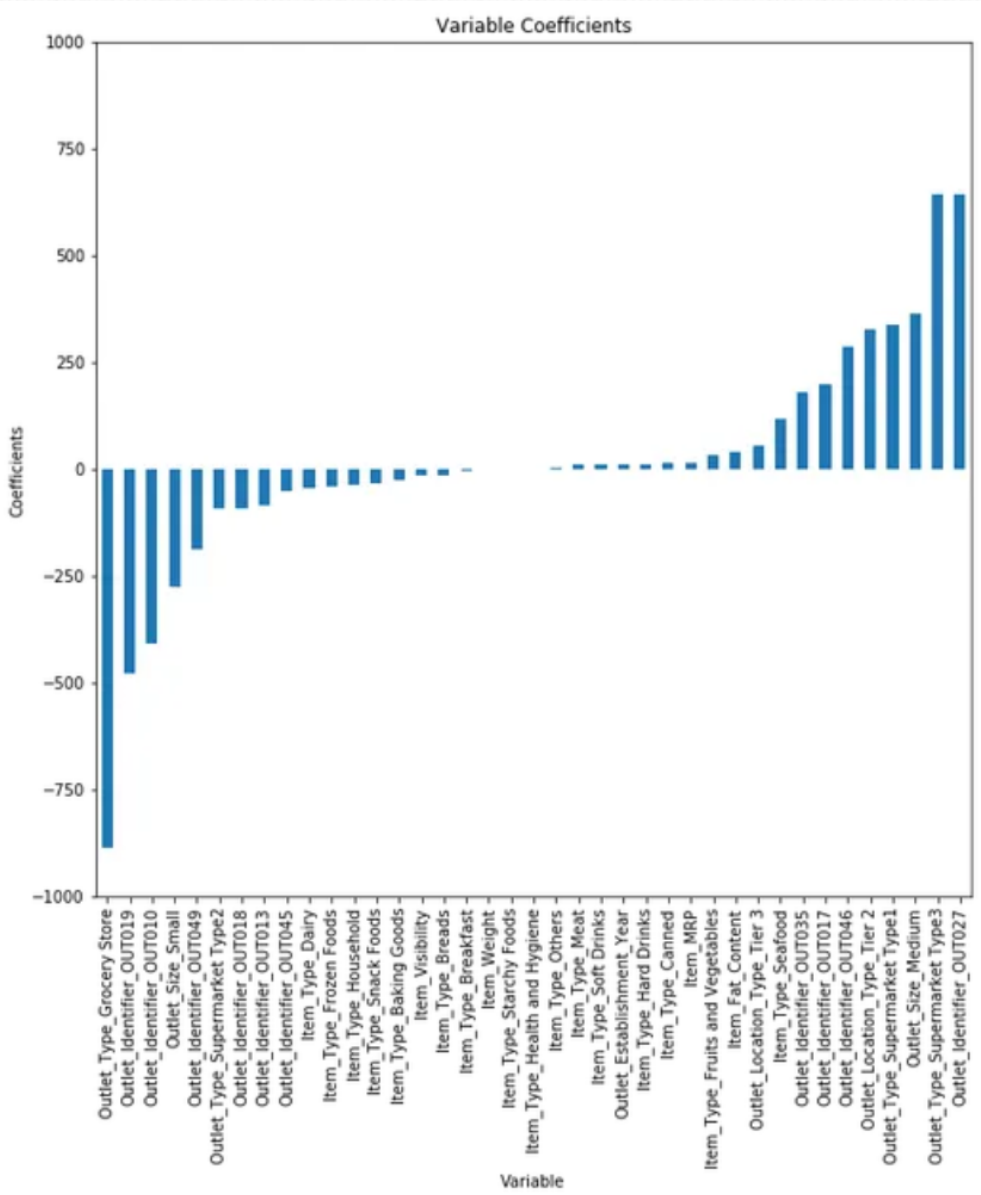
Regularization parameter

Regularization term

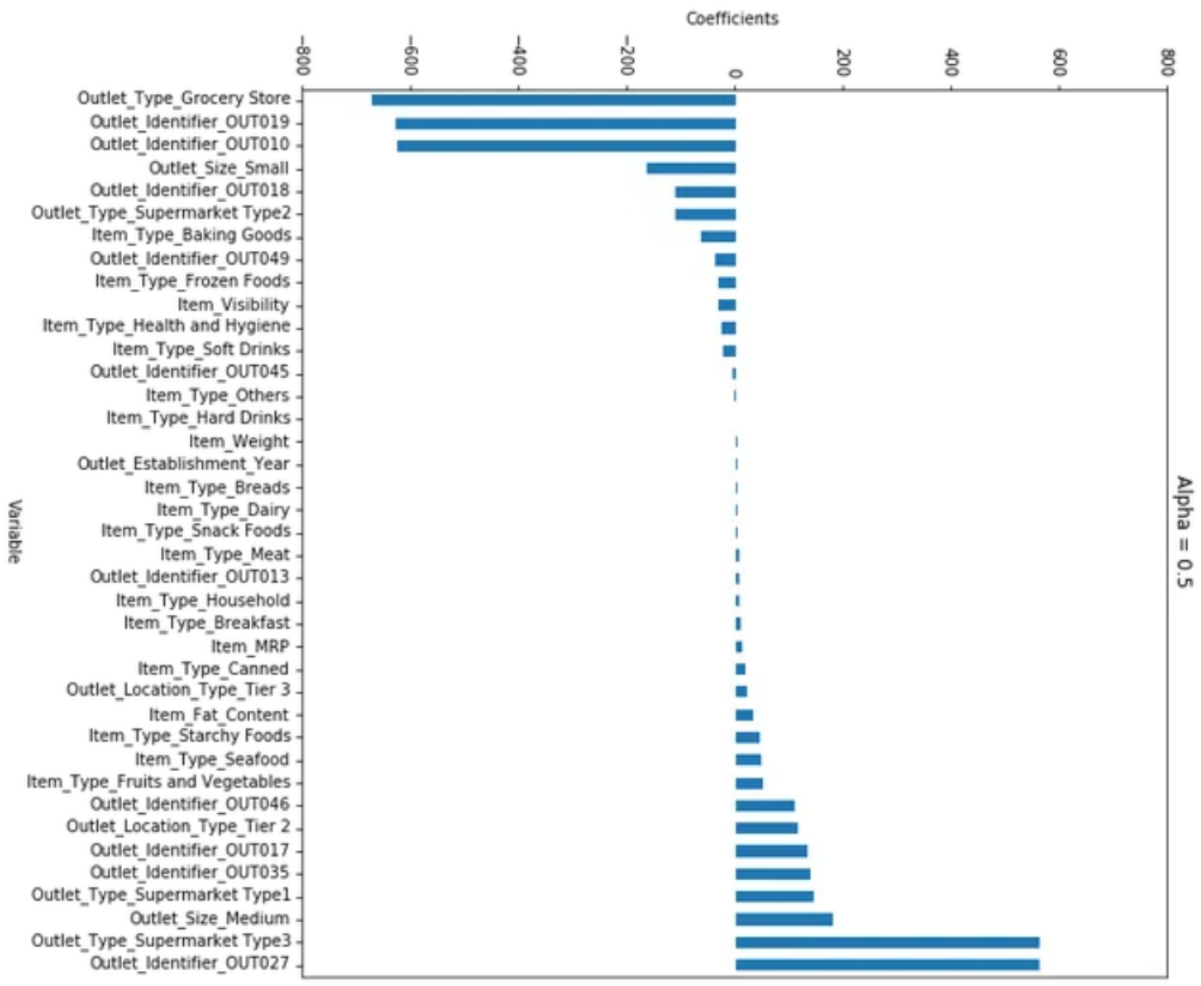
- min J(m)
- Lambda controls the trade off between fitting the training set well and setting the value of parameters m_j small and therefore keeping the hypothesis simple to aviod overfitting.

What happens if we set lamda very large?

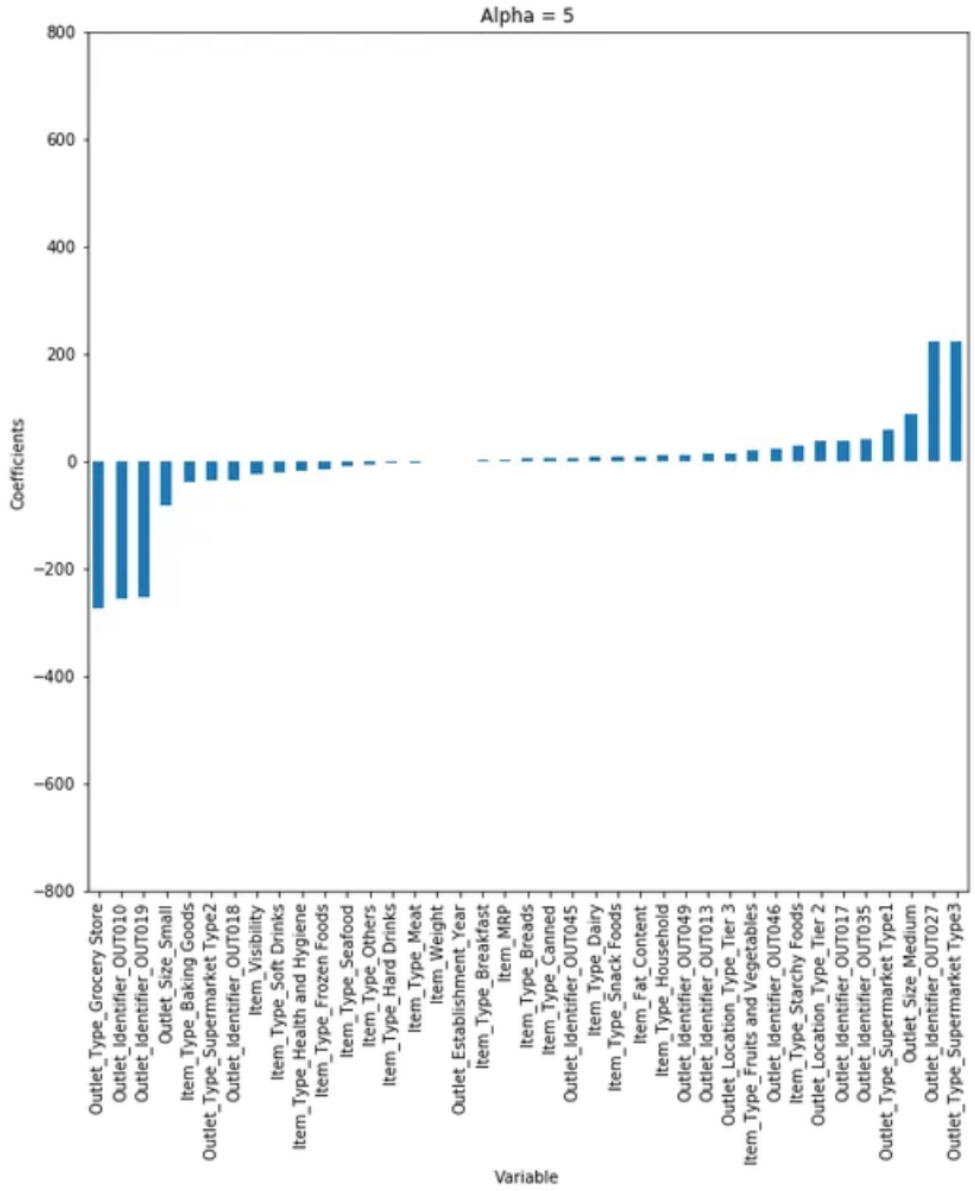
- It means we end up penalizing parameters m_j very heavily.
- Horizontal flat line a result, “Underfittng”
- So lamda should be choose very carefully.

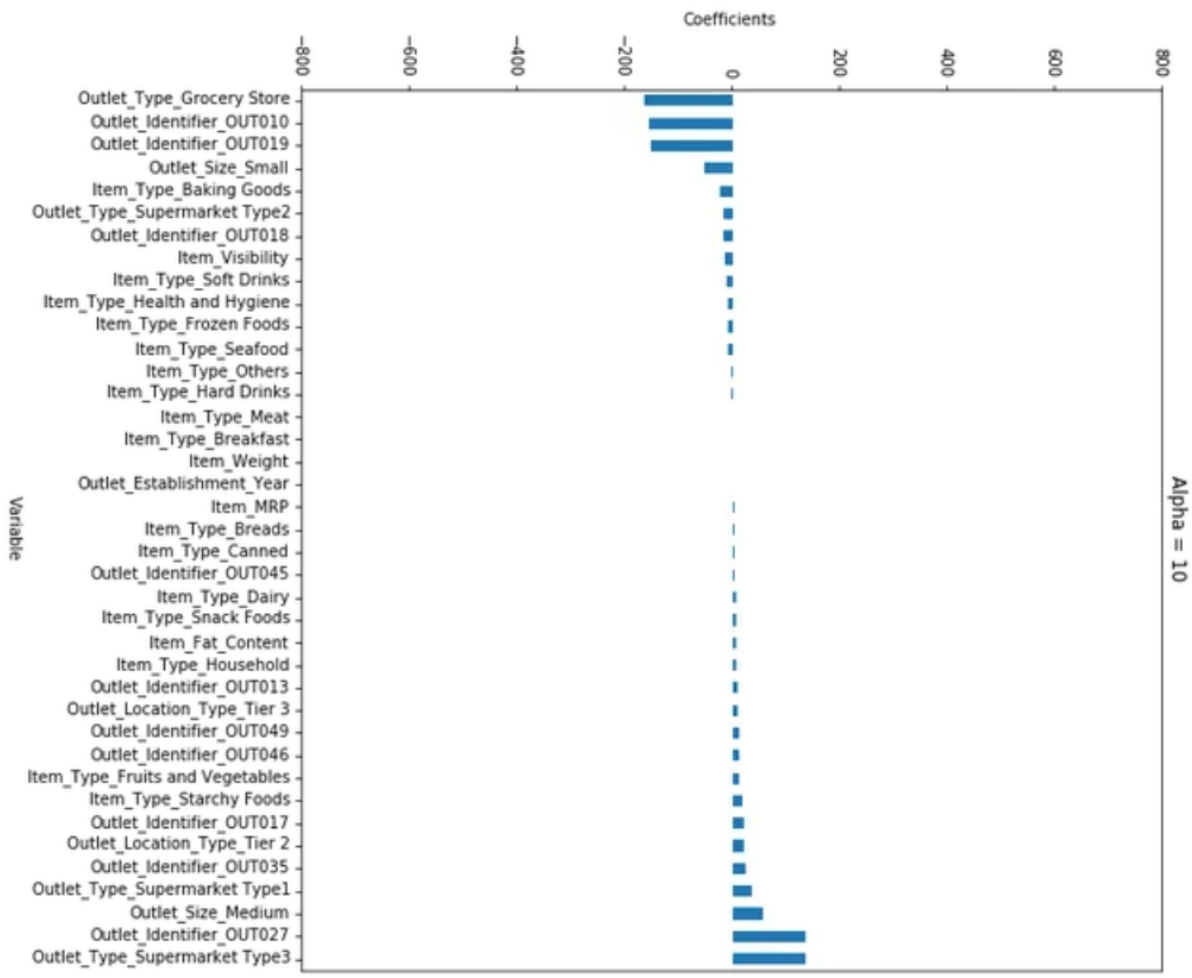

Variable Coefficients
Coefficients
1000
750
500
250
0
−250
−500
−750
−1000
Outlet_Type_Grocery Store
Outlet_Identifier_OUT019
Outlet_Identifier_OUT010
Outlet_Size_Small
Outlet_Identifier_OUT049
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT018
Outlet_Identifier_OUT013
Outlet_Identifier_OUT045
Item_Type_Dairy
Item_Type_Frozen Foods
Item_Type_Household
Item_Type_Snack Foods
Item_Type_Baking Goods
Item_Visibility
Item_Type_Breads
Item_Type_Breakfast
Item_Weight
Item_Type_Starchy Foods
Item_Type_Health and Hygiene
Item_Type_Others
Item_Type_Meat
Item_Type_Soft Drinks
Outlet_Establishment_Year
Item_Type_Hard Drinks
Item_Type_Canned
Item_MRP
Item_Type_Fruits and Vegetables
Item_Fat_Content
Outlet_Location_Type_Tier 3
Item_Type_Seafood
Outlet_Identifier_OUT035
Outlet_Identifier_OUT017
Outlet_Identifier_OUT046
Outlet_Location_Type_Tier 2
Outlet_Type_Supermarket Type1
Outlet_Size_Medium
Outlet_Type_Supermarket Type3
Outlet_Identifier_OUT027
Variable

Ridge Regression


Coefficients
-800
-600
-400
-200
0
200
400
600
800
Outlet_Type_Grocery Store
Outlet_Identifier_OUT019
Outlet_Identifier_OUT010
Outlet_Size_Small
Outlet_Identifier_OUT018
Outlet_Type_Supermarket Type2
Item_Type_Baking Goods
Outlet_Identifier_OUT049
Item_Type_Frozen Foods
Item_Visibility
Item_Type_Health and Hygiene
Item_Type_Soft Drinks
Outlet_Identifier_OUT045
Item_Type_Others
Item_Type_Hard Drinks
Item_Weight
Outlet_Establishment_Year
Item_Type_Breads
Item_Type_Dairy
Item_Type_Snack Foods
Item_Type_Meat
Outlet_Identifier_OUT013
Item_Type_Household
Item_Type_Breakfast
Item_MRP
Item_Type_Canned
Outlet_Location_Type_Tier 3
Item_Fat_Content
Item_Type_Starchy Foods
Item_Type_Seafood
Item_Type_Fruits and Vegetables
Outlet_Identifier_OUT046
Outlet_Location_Type_Tier 2
Outlet_Identifier_OUT017
Outlet_Identifier_OUT035
Outlet_Type_Supermarket Type1
Outlet_Size_Medium
Outlet_Type_Supermarket Type3
Outlet_Identifier_OUT027
Variable
Alpha = 0.5

Alpha = 0.05
800
600
400
200
0
−200
−400
−600
−800
Coefficients
Outlet_Type_Grocery Store
Outlet_Identifier_OUT019
Outlet_Identifier_OUT010
Outlet_Size_Small
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT018
Outlet_Identifier_OUT049
Item_Type_Frozen Foods
Outlet_Identifier_OUT045
Item_Type_Dairy
Item_Type_Baking Goods
Item_Type_Household
Item_Visibility
Item_Type_Snack Foods
Outlet_Identifier_OUT013
Item_Type_Breads
Item_Weight
Outlet_Establishment_Year
Item_Type_Health and Hygiene
Item_Type_Breakfast
Item_Type_Others
Item_Type_Soft Drinks
Item_MRP
Item_Type_Meat
Item_Type_Starchy Foods
Item_Type_Hard Drinks
Outlet_Location_Type_Tier 3
Item_Type_Canned
Item_Fat_Content
Item_Type_Fruits and Vegetables
Item_Type_Seafood
Outlet_Location_Type_Tier 2
Outlet_Identifier_OUT046
Outlet_Type_Supermarket Type1
Outlet_Identifier_OUT017
Outlet_Identifier_OUT035
Outlet_Size_Medium
Outlet_Type_Supermarket Type3
Outlet_Identifier_OUT027
Variable


Alpha = 5
800
600
400
200
0
−200
−400
−600
−800
Coefficients
Outlet_Type_Grocery Store
Outlet_Identifier_OUT010
Outlet_Identifier_OUT019
Outlet_Size_Small
Item_Type_Baking Goods
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT018
Item_Visibility
Item_Type_Soft Drinks
Item_Type_Health and Hygiene
Item_Type_Frozen Foods
Item_Type_Seafood
Item_Type_Others
Item_Type_Hard Drinks
Item_Type_Meat
Item_Weight
Outlet_Establishment_Year
Item_Type_Breakfast
Item_MRP
Item_Type_Breads
Item_Type_Canned
Outlet_Identifier_OUT045
Item_Type_Dairy
Item_Type_Snack Foods
Item_Fat_Content
Item_Type_Household
Outlet_Identifier_OUT049
Outlet_Identifier_OUT013
Outlet_Location_Type_Tier 3
Item_Type_Fruits and Vegetables
Outlet_Identifier_OUT046
Item_Type_Starchy Foods
Outlet_Location_Type_Tier 2
Outlet_Identifier_OUT017
Outlet_Identifier_OUT035
Outlet_Type_Supermarket Type1
Outlet_Size_Medium
Outlet_Identifier_OUT027
Outlet_Type_Supermarket Type3
Variable


Coefficients
−800
−600
−400
−200
0
200
400
600
800
Outlet_Type_Grocery Store
Outlet_Identifier_OUT010
Outlet_Identifier_OUT019
Outlet_Size_Small
Item_Type_Baking Goods
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT018
Item_Visibility
Item_Type_Soft Drinks
Item_Type_Health and Hygiene
Item_Type_Frozen Foods
Item_Type_Seafood
Item_Type_Others
Item_Type_Hard Drinks
Item_Type_Meat
Item_Type_Breakfast
Item_Weight
Outlet_Establishment_Year
Item_MRP
Item_Type_Breads
Item_Type_Canned
Outlet_Identifier_OUT045
Item_Type_Dairy
Item_Type_Snack Foods
Item_Fat_Content
Item_Type_Household
Outlet_Identifier_OUT013
Outlet_Location_Type_Tier 3
Outlet_Identifier_OUT049
Outlet_Identifier_OUT046
Item_Type_Fruits and Vegetables
Item_Type_Starchy Foods
Outlet_Identifier_OUT017
Outlet_Location_Type_Tier 2
Outlet_Identifier_OUT035
Outlet_Type_Supermarket Type1
Outlet_Size_Medium
Outlet_Identifier_OUT027
Outlet_Type_Supermarket Type3
Variable
Alpha = 10

Ridge Regression

- 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.
- But if you calculate R-square for each alpha, we will see that the value of R-square will be maximum at alpha=0.05.
- So we have to choose it wisely by iterating it through a range of values and using the one which gives us lowest error.

Ridge Regression

```
from sklearn.linear_model import
Ridge
## training the model
ridgeReg = Ridge(alpha=0.05,
normalize=True)
ridgeReg.fit(x_train,y_train)
pred = ridgeReg.predict(x_cv)
```

- Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients are reduced.

Ridge Regression

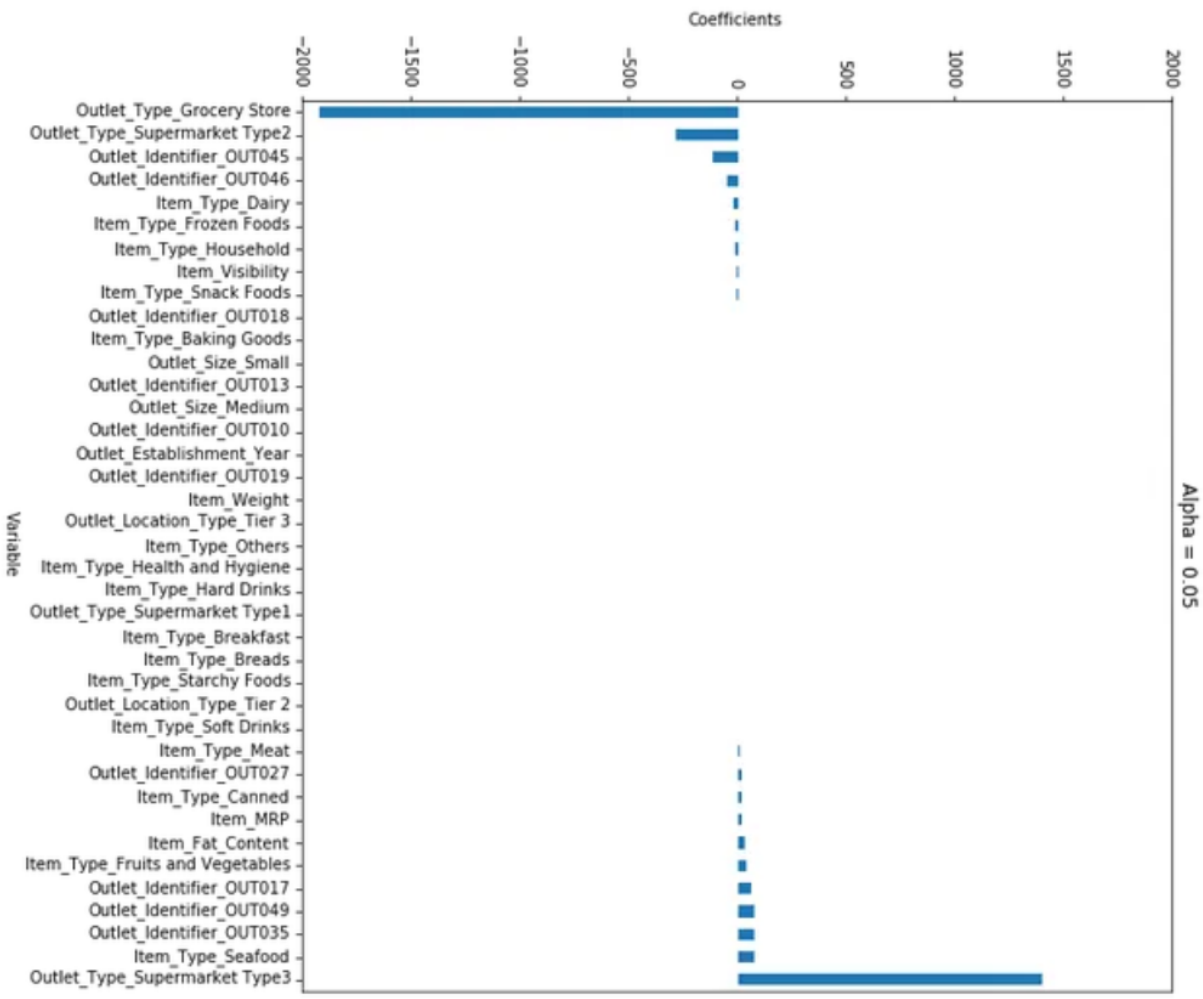
$$J(m) = \frac{1}{2n}\sum_{i=1}^{n}\left(h\ (x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{k} m_j^p$$

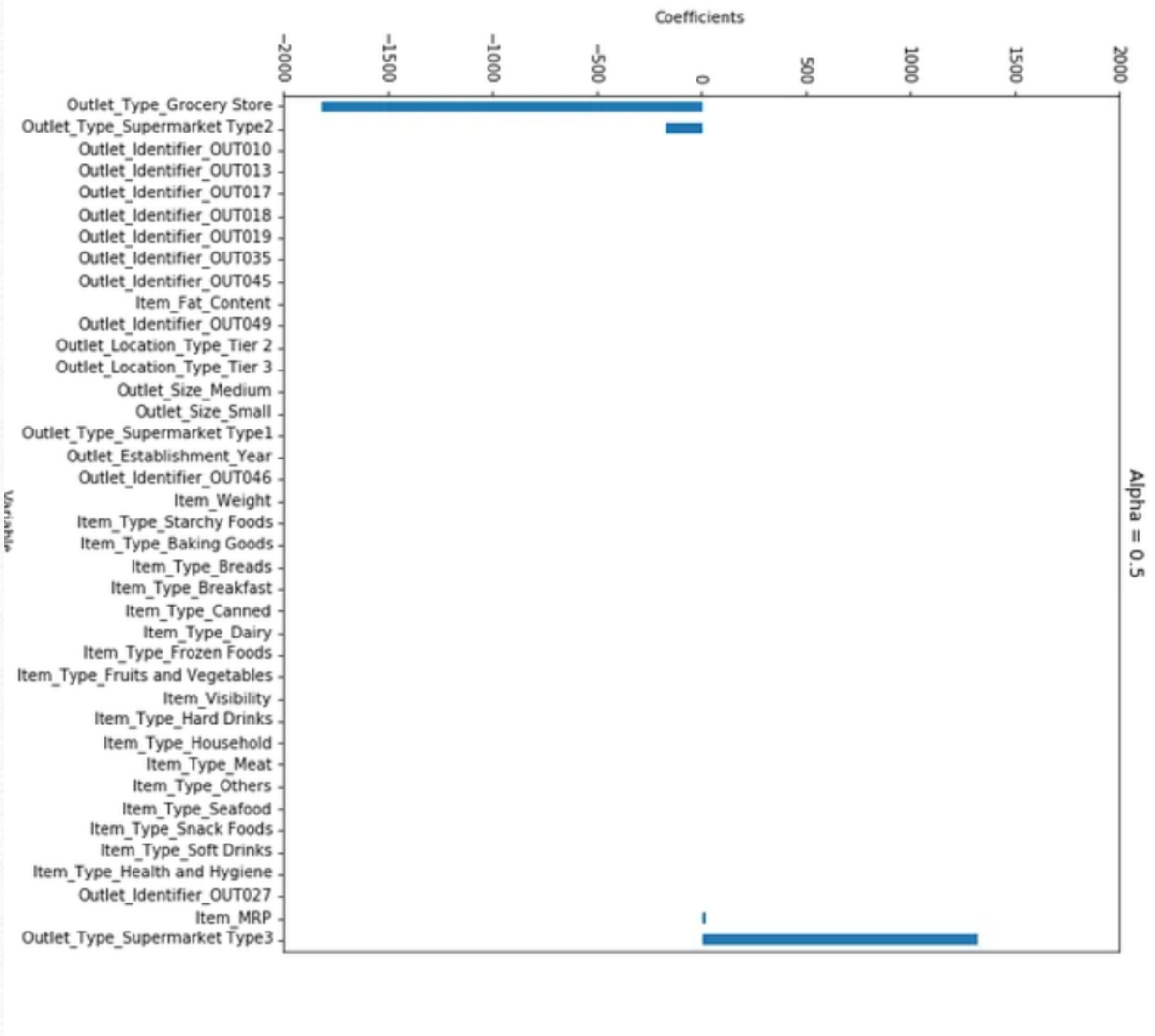
p = 2

It uses **L-2 regularization** technique.

Lasso regression

- LASSO (Least Absolute Shrinkage Selector Operator)


Coefficients
−2000
−1500
−1000
−500
0
500
1000
1500
2000
Outlet_Type_Grocery Store
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT045
Outlet_Identifier_OUT046
Item_Type_Dairy
Item_Type_Frozen Foods
Item_Type_Household
Item_Visibility
Item_Type_Snack Foods
Outlet_Identifier_OUT018
Item_Type_Baking Goods
Outlet_Size_Small
Outlet_Identifier_OUT013
Outlet_Size_Medium
Outlet_Identifier_OUT010
Outlet_Establishment_Year
Outlet_Identifier_OUT019
Item_Weight
Outlet_Location_Type_Tier 3
Item_Type_Others
Item_Type_Health and Hygiene
Item_Type_Hard Drinks
Outlet_Type_Supermarket Type1
Item_Type_Breakfast
Item_Type_Breads
Item_Type_Starchy Foods
Outlet_Location_Type_Tier 2
Item_Type_Soft Drinks
Item_Type_Meat
Outlet_Identifier_OUT027
Item_Type_Canned
Item_MRP
Item_Fat_Content
Item_Type_Fruits and Vegetables
Outlet_Identifier_OUT017
Outlet_Identifier_OUT049
Outlet_Identifier_OUT035
Item_Type_Seafood
Outlet_Type_Supermarket Type3
Variable
Alpha = 0.05


Coefficients
−2000
−1500
−1000
−500
0
500
1000
1500
2000
Outlet_Type_Grocery Store
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT010
Outlet_Identifier_OUT013
Outlet_Identifier_OUT017
Outlet_Identifier_OUT018
Outlet_Identifier_OUT019
Outlet_Identifier_OUT035
Outlet_Identifier_OUT045
Item_Fat_Content
Outlet_Identifier_OUT049
Outlet_Location_Type_Tier 2
Outlet_Location_Type_Tier 3
Outlet_Size_Medium
Outlet_Size_Small
Outlet_Type_Supermarket Type1
Outlet_Establishment_Year
Outlet_Identifier_OUT046
Item_Weight
Item_Type_Starchy Foods
Item_Type_Baking Goods
Item_Type_Breads
Item_Type_Breakfast
Item_Type_Canned
Item_Type_Dairy
Item_Type_Frozen Foods
Item_Type_Fruits and Vegetables
Item_Visibility
Item_Type_Hard Drinks
Item_Type_Household
Item_Type_Meat
Item_Type_Others
Item_Type_Seafood
Item_Type_Snack Foods
Item_Type_Soft Drinks
Item_Type_Health and Hygiene
Outlet_Identifier_OUT027
Item_MRP
Outlet_Type_Supermarket Type3
Variable
Alpha = 0.5

Any differences between ridge and lasso?

Lasso regression

- We can see that as we increased the value of alpha, coefficients were approaching towards zero.
- But if you see in case of lasso, 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.

Lasso regression

- It uses L1 regularization technique p = 1

$$J(m) = \frac{1}{2n}\sum_{i=1}^{n}\left(h\ (x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{k} m_j^p$$

- It is generally used when we have more number of features, because it automatically does feature selection.

Code

```
from sklearn.linear_model import Lasso
lassoReg = Lasso(alpha=0.3, normalize=True)
lassoReg.fit(x_train,y_train)
pred = lassoReg.predict(x_cv)
```

Which one to use?

- An example where we have a large dataset, lets say it has 10,000 features.
  - Ridge : model complex
  - Lasso: loose some information

Elastic Net Regression

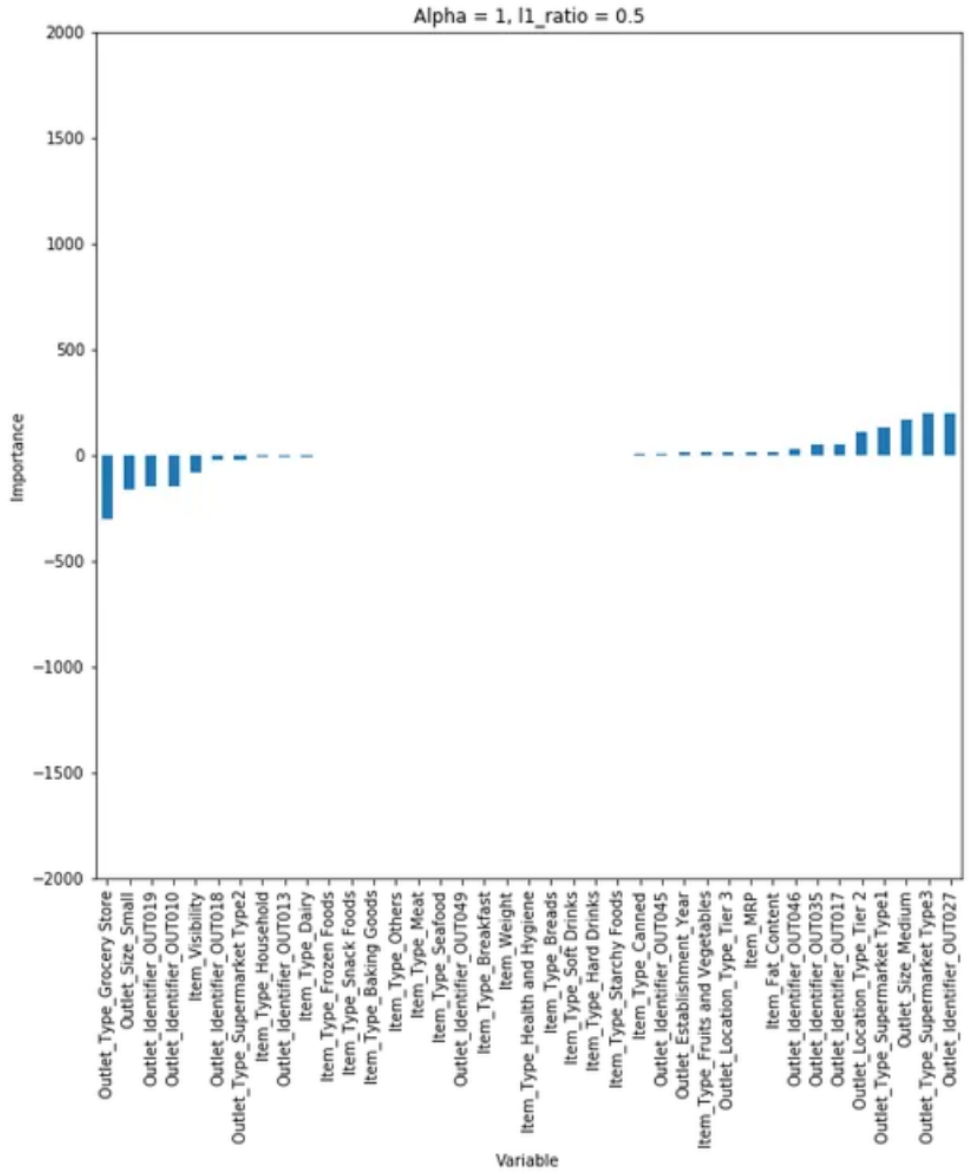
```
from sklearn.linear_model import
ElasticNet
ENreg = ElasticNet(alpha=1,
l1_ratio=0.5, normalize=False)
ENreg.fit(x_train,y_train)
pred_cv = ENreg.predict(x_cv)
```

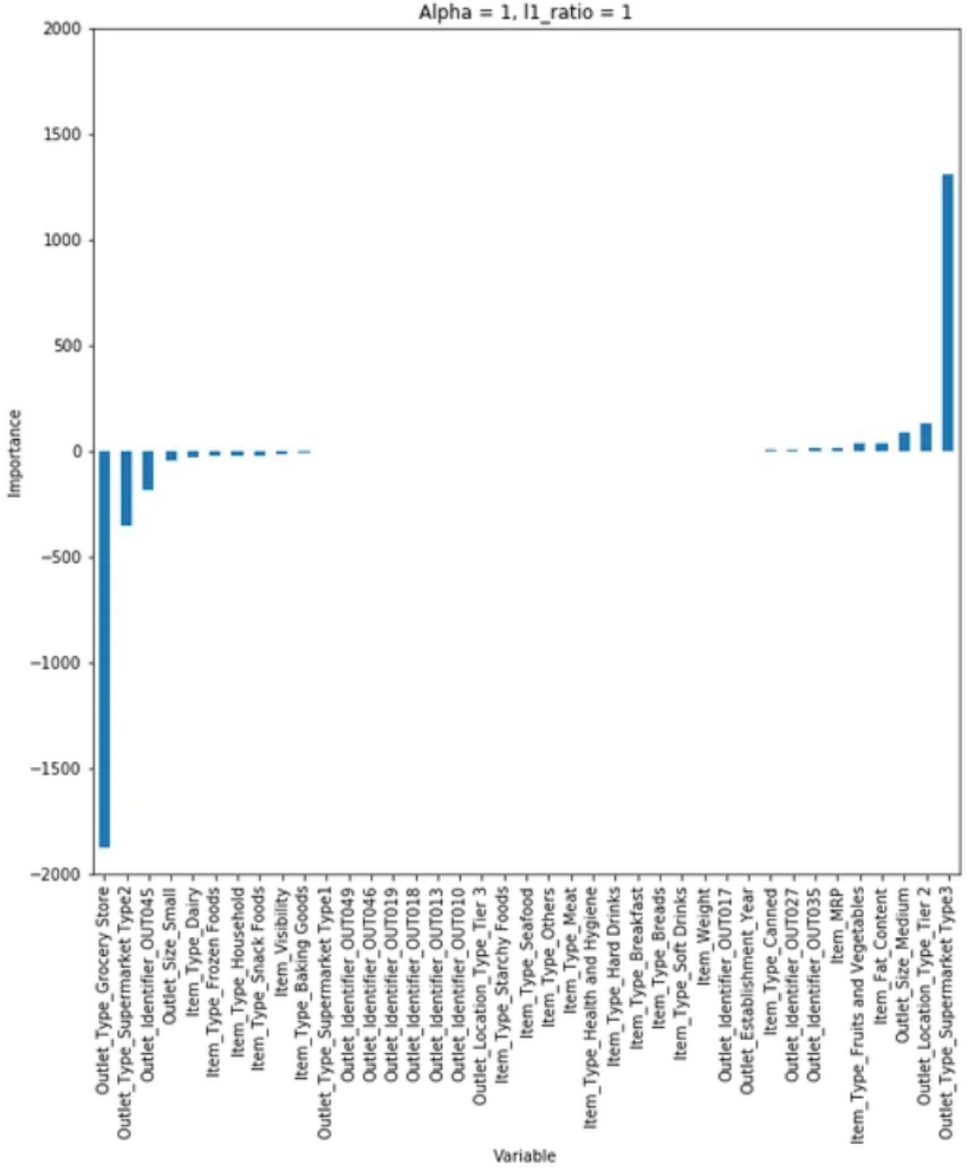
- It is basically a combination of both L1 and L2 regularization (you can implement both Ridge and Lasso by tuning the parameters)
- We need to define alpha and l1_ratio

Elastic Net Regression

- Alpha = a + b and l1_ratio = a / (a+b)

- Let alpha (or a+b) = 1, and now consider the following cases:
  1. If l1_ratio =1, therefore if we look at the formula of l1_ratio, we can see that l1_ratio can only be equal to 1 if a=1, which implies b=0. Therefore, it will be a lasso penalty.
  2. Similarly if l1_ratio = 0, implies a=0. Then the penalty will be a ridge penalty.
  3. For l1_ratio between 0 and 1, the penalty is the combination of ridge and lasso.


Alpha = 1, l1_ratio = 0.5
Importance
2000
1500
1000
500
0
−500
−1000
−1500
−2000
Outlet_Type_Grocery Store
Outlet_Size_Small
Outlet_Identifier_OUT019
Outlet_Identifier_OUT010
Item_Visibility
Outlet_Identifier_OUT018
Outlet_Type_Supermarket Type2
Item_Type_Household
Outlet_Identifier_OUT013
Item_Type_Dairy
Item_Type_Frozen Foods
Item_Type_Snack Foods
Item_Type_Baking Goods
Item_Type_Others
Item_Type_Meat
Item_Type_Seafood
Outlet_Identifier_OUT049
Item_Type_Breakfast
Item_Weight
Item_Type_Health and Hygiene
Item_Type_Breads
Item_Type_Soft Drinks
Item_Type_Hard Drinks
Item_Type_Starchy Foods
Item_Type_Canned
Outlet_Identifier_OUT045
Outlet_Establishment_Year
Item_Type_Fruits and Vegetables
Outlet_Location_Type_Tier 3
Item_MRP
Item_Fat_Content
Outlet_Identifier_OUT046
Outlet_Identifier_OUT035
Outlet_Identifier_OUT017
Outlet_Location_Type_Tier 2
Outlet_Type_Supermarket Type1
Outlet_Size_Medium
Outlet_Type_Supermarket Type3
Outlet_Identifier_OUT027
Variable


Alpha = 1, l1_ratio = 1
Importance
2000
1500
1000
500
0
−500
−1000
−1500
−2000
Outlet_Type_Grocery Store
Outlet_Type_Supermarket Type2
Outlet_Identifier_OUT045
Outlet_Size_Small
Item_Type_Dairy
Item_Type_Frozen Foods
Item_Type_Household
Item_Type_Snack Foods
Item_Visibility
Item_Type_Baking Goods
Outlet_Type_Supermarket Type1
Outlet_Identifier_OUT049
Outlet_Identifier_OUT046
Outlet_Identifier_OUT019
Outlet_Identifier_OUT018
Outlet_Identifier_OUT013
Outlet_Identifier_OUT010
Outlet_Location_Type_Tier 3
Item_Type_Starchy Foods
Item_Type_Seafood
Item_Type_Others
Item_Type_Meat
Item_Type_Health and Hygiene
Item_Type_Hard Drinks
Item_Type_Breakfast
Item_Type_Breads
Item_Type_Soft Drinks
Item_Weight
Outlet_Identifier_OUT017
Outlet_Establishment_Year
Item_Type_Canned
Outlet_Identifier_OUT027
Outlet_Identifier_OUT035
Item_MRP
Item_Type_Fruits and Vegetables
Item_Fat_Content
Outlet_Size_Medium
Outlet_Location_Type_Tier 2
Outlet_Type_Supermarket Type3
Variable

Evaluate the model

- How accurate do you think the model is?
- R square and adjusted R- square

$$R - Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})\text{^}2}{\sum(Y_{actual} - Y_{mean})\text{^}2}$$

R square

- It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable).
- The value of R-square is always between 0 and 1,
- where 0 means that the model does not model explain any variability in the target variable (Y)
- and 1 meaning it explains full variability in the target variable.

Adjusted R-square

- The only drawback of R^2 is that if new predictors (X) are added to our model, R^2 only increases or remains constant but it never decreases.
- We can not judge that by increasing complexity of our model, are we making it more accurate?
- The adjusted R-Square only increases if the new term improves the model accuracy.

"Knowledge is the treasure and practice is the key to it"

Discussion

Thank you!!!!!!