<u>Question 1:</u>

- Report the mean squared error on the test samples using all the features with L2 Norm:

Mean Squared Error: **1073.589785862**

- Describe what stopping criterion you used to determine when the gradient descent algorithm should halt.

The gradient descent has been cross-validated using a 80% (training) 20% (validation) dataset division.
The training will continue until the validation MSE decreases.

- Report the mean squared error on the test samples when the predictor is unregularized i.e. . $\lambda = 0$.

Mean Squared Error: **1590.099363602**

Question 2:

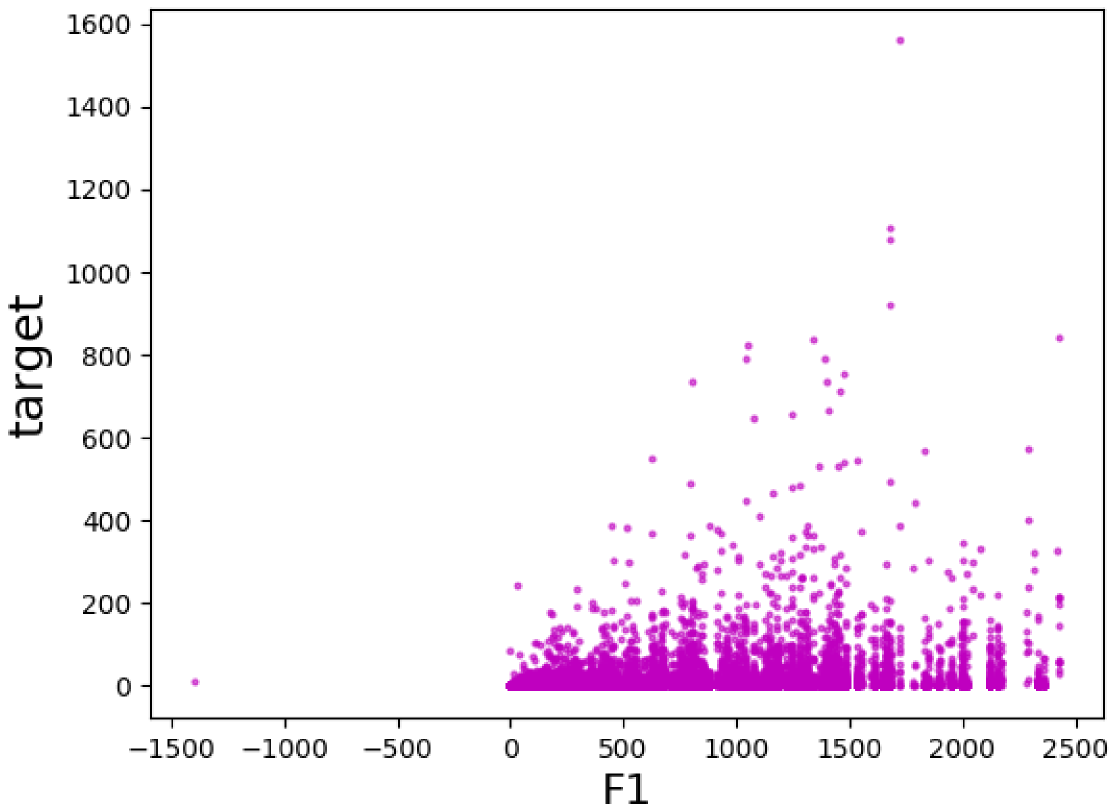
What do you consider to be the three most important features that contributed to your predictor's performance?

After training the Regression model with L2 normalisation the weight vector is as following:

[1.01724951 0.78462184 0.86763448 1.00545727 0.16841069 **2.36252291**
  1.46447325 1.26536943 2.07997951 0.8097293   **2.8847474**   1.04668431
  0.9732983   1.61757686 1.57861202 0.89703365 1.65938838 1.03812405
 -2.4727472   0.75157846 0.82018822 0.7436405   1.07270332 0.61462759
  0.7856113   0.82135161 0.73265588 0.83860968 0.71740518 0.97178461
  0.61169759 0.6987022   0.88578762 0.84625084 0.66462522 0.7708094
  1.00417298]

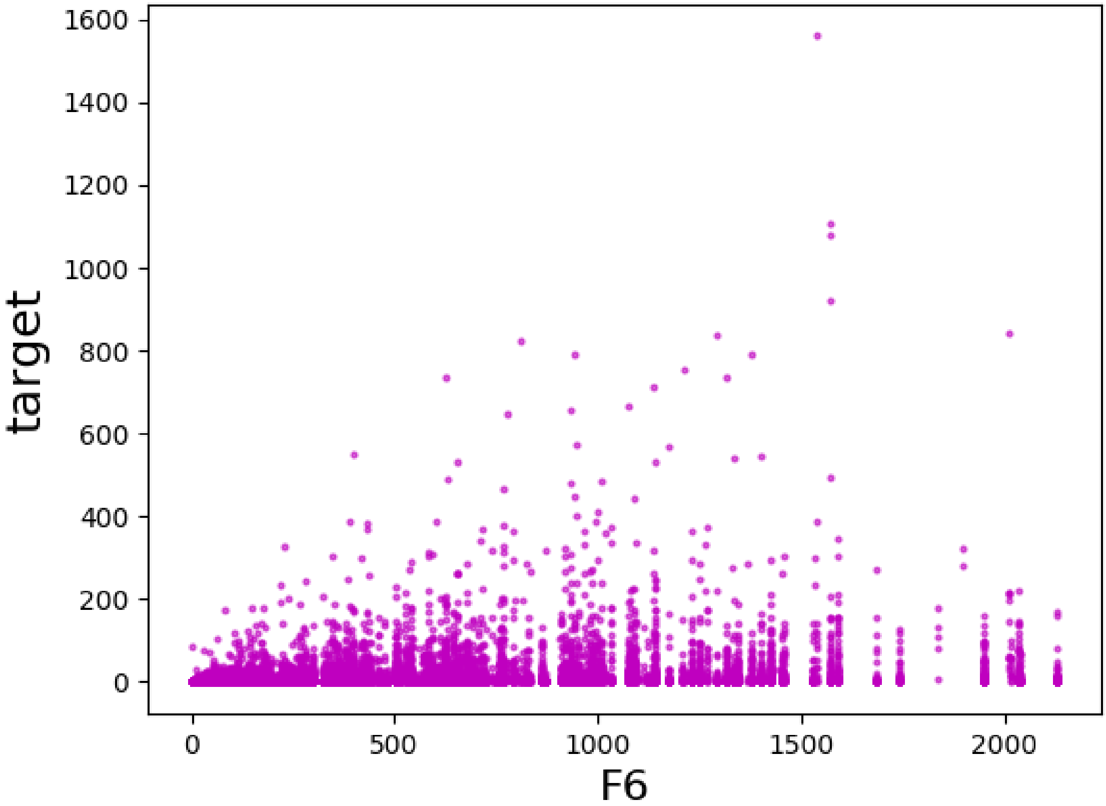
Taking the absolute values of the weights from the above vector, the top 3 features are (F1, F6 and base_time):

F1: 2.36252291



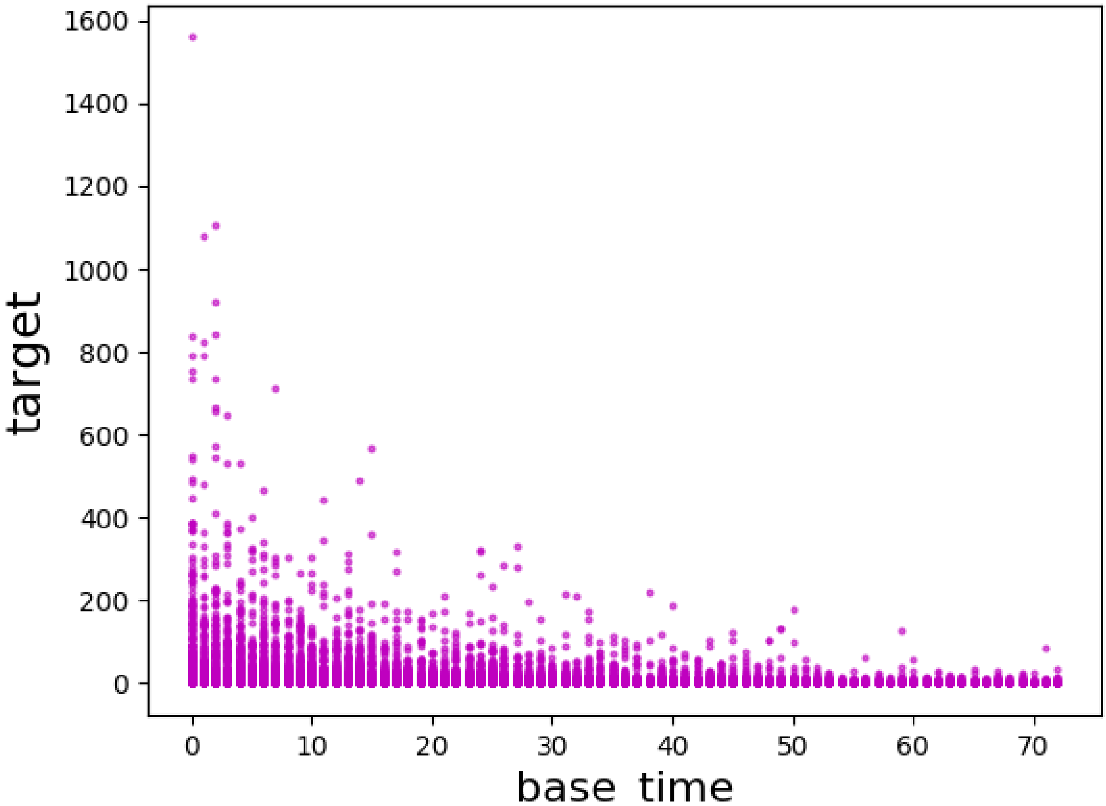
The above plot shows an almost normal distribution, which positively influences the model. However, there are some outlier values which can lead to some false predictions.

F6: 2.8847474



The F6 parameter is also an almost normal distribution, However, compared to F1, it is better distributed with lower outliers, and hence, the model rightly gives more emphasis to F6 compared to F1.

base_time: -2.4727472



As can be clearly observed from the graph, there is a strong negative dependence between the target and the parameter base_time which decreases as base_time increases.
Hence, the model correctly puts a strong negative weight to the parameter.

Question 3:

Modify your implementation to allow for a p-norm regularizer where p can take any one of the three values, [2, 4, 6]. Report mean squared errors on the test samples when p=4 and p=6.

P = 4
Mean Squared Error: **1075.075239956**

P = 6
Mean Squared Error: **1072.544160078**

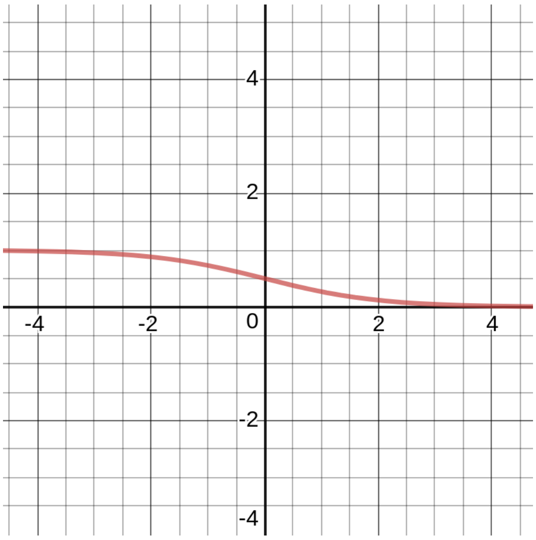
Question 4:

Implement two different basis functions that will be applied to your input features with the L2- regularized model.

Basis Function 1:

$$\Phi(x)=\frac{1}{1+e^{x}}$$

The function behaves as below:



Graph generated using https://graphsketch.com/

The Feature scaling used in the model is Min-Max. So all data points are within the interval of 0 and 1.

As evident from the plot above, the function scales up values in the lower ranges closer to 0 while downscales values which are closer to 1 or extremely high value datapoints. This ensures equal representation of all datapoints disabling high value datapoints (more likely to be outliers) from over overtly influencing the model.

Mean Squared Error: **1062.96212961**

Basis Function 2:

$$\Phi(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(Normal Distribution with mean ω and standard deviation σ)

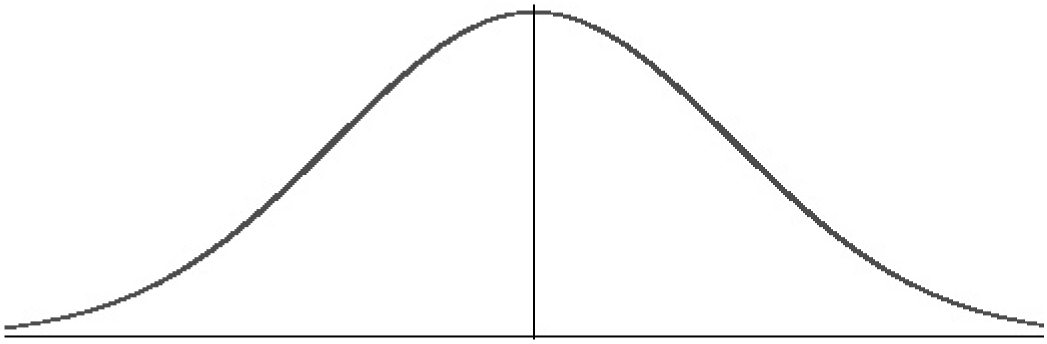


Image courtesy: http://www.statisticshowto.com/probability-and-statistics/normal-distributions/

As seen from the above curve the function penalises both sides of the spectrum extremeties extremely low and extremely low values, favouring values more around the mean thus penalising ourliers.

Mean Squared Error: **1067.306031122**

Question 5:

For this part, you have free reign. You can implement any enhancements to the regularized linear regression model that you think your model could benefit from. Please describe your innovations within the file.

For this phase Basis Function 1, i.e $\Phi(x)=\frac{1}{1+e^{x}}$ was used along with p-norm of order 6.
Mean Squared Error: **1062.749568026**