# Report

## Question 1:

- **Report the mean squared error on the test samples using all the features.**

  I have used L2 regularisation (for hyperparameters such as learning rate etc. and parameters list please refer **README**), and the mean squared error on the test samples is given below:

  **MSE = 1351.4953760361** ...(source: **kaggle RMSE: 32.76269)**

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

  For the gradient descent algorithm to halt I have incorporated two stopping criteria, anyone of which when satisfied the learning algorithm will halt.

  - **Validation ( held-out set ):**

    I have taken a validation set of size approximately **20%** of the given data set (for exact indices please refer: **README**) .

    If the cost (**MSE**) obtained on validation set on $i^{th}$ iteration is greater than $(i-1)^{th}$ iteration then the learning algorithm will halt. And then weight for $(i-1)^{th}$ iteration is returned.

  - **Error rate:**

    **Error rate = abs(cost of train-set at $i^{th}$ iteration - cost of train-set at $(i-1)^{th}$ iteration)**

    If error rate reduces < **(less than)** $10^{-18}$ then it will halt.

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

  Model got too much overfitted, hence MSE had increased. (For hyperparameters such as learning rate etc. and parameters list please refer **README**)

  **MSE = 1590.355781476** ...(source: **kaggle RMSE: 39.87926)**

**Question 2:**

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

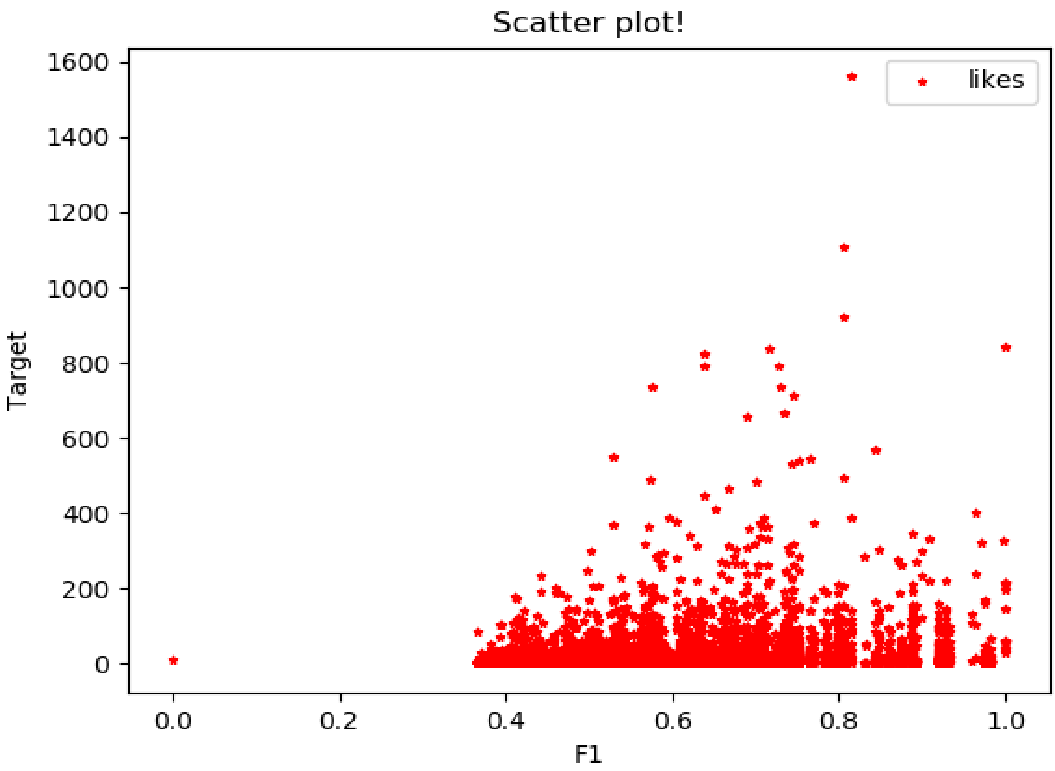
For determining the top three features, I have used following concepts (For model settings such as feature scaling, hyperparameters etc. Please refer: **README**):

- Since we are using a linear model in our incorporation, linear relationship between **feature set** and the **target** value can be exploited for extracting the top influential features
- So for the all the features whose influence is more on determining the curve must have a significant large weight associated to it. (Since data taken is of normal distribution so with increase in the value of these features will increase the target values but till certain points as after that there would be no more increase in the value **see: the curves**)
- Also we should also incorporate the concept of negative learning, where if the feature gives us the decreasing plot, we should also decrease the value of target prediction

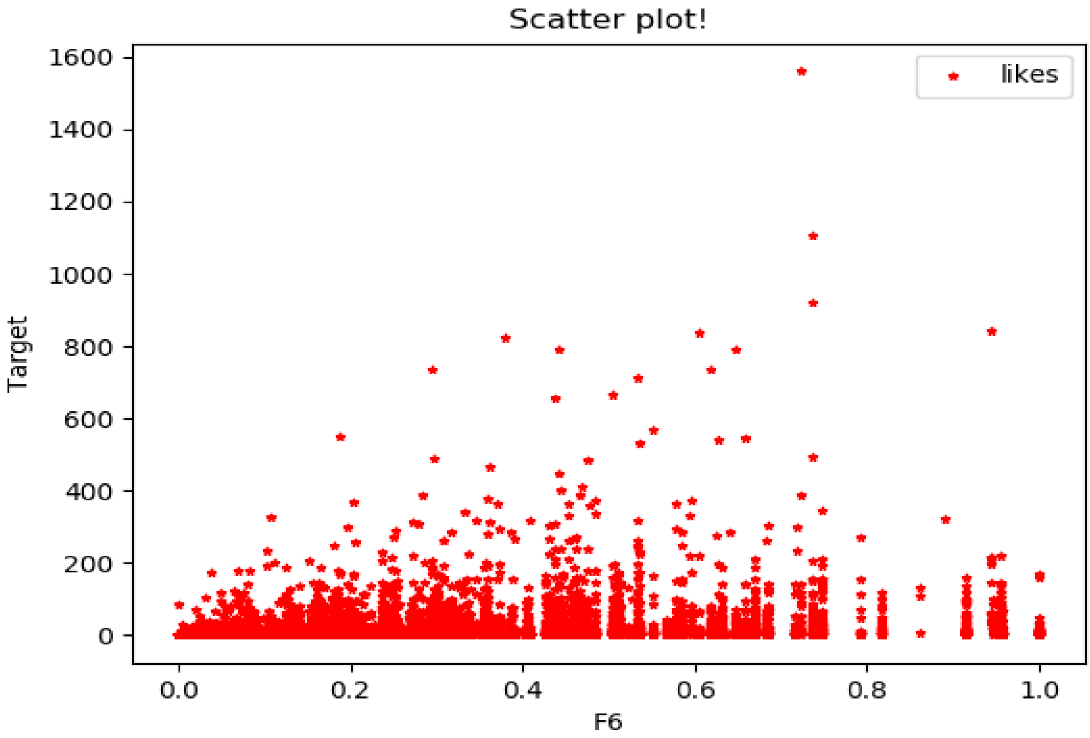
**Using these concepts the top three features along with their plot is shown below:**

**NOTE: Feature set is normalized using min-max normalization and L2 regularization is used.**

- **F1** ( features is aggregated by page, by calculating min, max, average, median and standard deviation of essential features.)

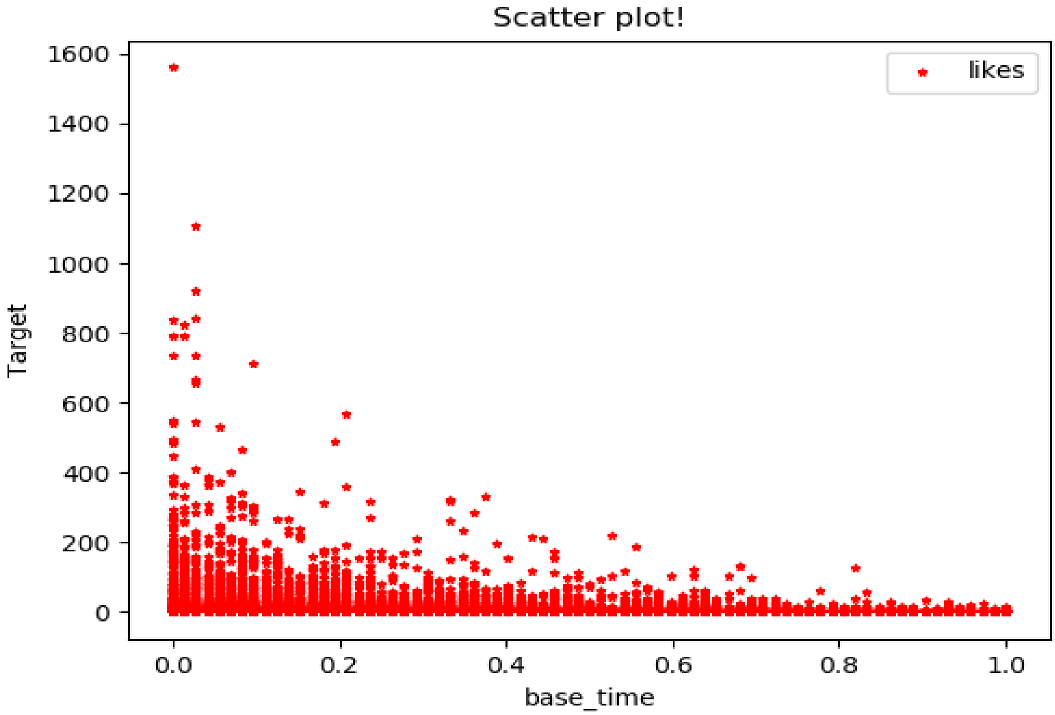


- **F6** ( features is aggregated by page, by calculating min, max, average, median and standard deviation of essential features.)



- **base_time -** Selected time in order to simulate the scenario.

  ( Target prediction decreases linearly with increase in base_time)



**Question 3:**

- **Modify your implementation to allow for a $p$-norm regularizer where $p$ can take any one of the three values, $\{2,4,6\}$. Your regularization term now becomes $\lambda||\mathbf{w}||p$ where p-norm of a $d$-dimensional $\mathbf{w}$ is defined as $||\mathbf{w}||p=(\sum di=1|wi|p)1p$. Report mean squared errors on the test samples when $p=4$ and $p=6$.**

(For model settings  such as feature scaling, hyperparameters etc. Please refer: **README**)

➢ P = 4

I have taken lambda **λ = 109000**, which had given me a better performance on test samples than **L2** regularised model.

**MSE = 1071.5723686144** ...(source: **kaggle RMSE: 32.73488)**

➢ P = 6

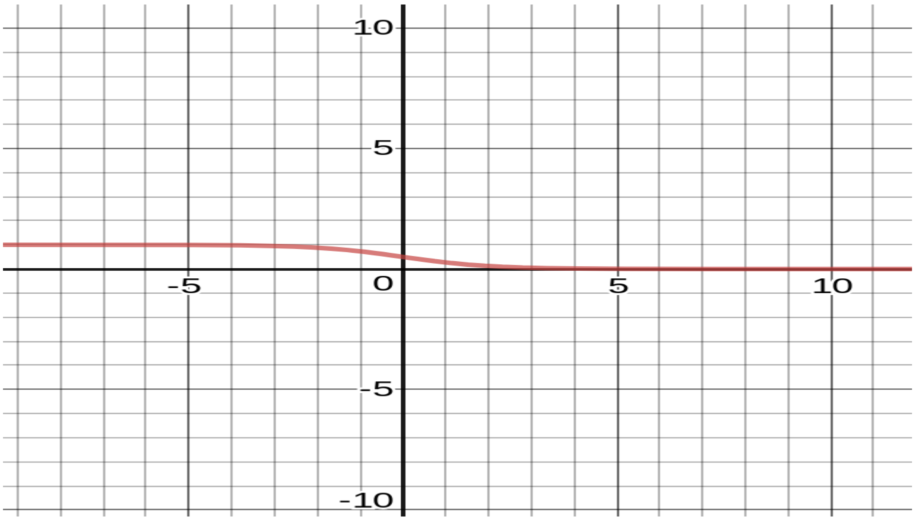
I have taken lambda **λ = 119000**, which had given me a better performance on test samples than **L2** regularised model also than p = 4 norm.

**MSE = 1070.1429860416** ...(source: **kaggle RMSE: 32.71304)**

**Question 4:**

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

I have kept in mind the distribution of model (min-max normalization), before applying the basis function, I want something like decreasing (**shown below**):



Source : https://www.desmos.com/calculator (Inverse-sigmoid)

This curve will put more emphasis on the features having less value so that their essence should not be ignored in the curve also, less emphasis on the outliers that can drive the curve drastically.

- **Inverse Sigmoidal** $\frac{1}{(1+e^{x})}$

  **Note: see the graph shown above**

  Choice of this basis function helps me alot while driving my target prediction. As model performs better than all previous norms. (Regularizer)**λ = 740**

  **MSE = 1062.9693023041** ...(source: **kaggle RMSE: 32.60321)**

- **Gaussian Function** $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  Gaussian also behaves similar to the curve we need, hence I used it. Performance is not better than sigmoidal because of its behaviour. (Regularizer)**λ =13000**

  **MSE = 1067.0296637764** ...(source: **kaggle RMSE: 32.66542)**

**Question 5:**

- **You can implement any enhancements to the regularized linear regression model that you think your model could benefit from.**

  For this I have incorporated the knowledge of both p=6 norm and sigmoidal basis function, fand therefore I have got the best result till now. (Regularizer)**λ = 706**

  **MSE = 1062.6745897161** ...(source: **kaggle RMSE: 32.59869)**