1.1

a.

$$NLL(D,w) = -\sum_{i=1}^{N}[(1-y_i)\ \log\left(1-\sigma(w^T x_i)\right) + y_i \log \sigma\,(w^T x_i)]$$

$$= -\sum_{i=1}^{N}[\frac{\partial}{\partial w_j}(1-y_i)\ \log\left(1-\sigma(w^T x_i)\right) + \frac{\partial}{\partial wj} y_i \log \sigma\,(w^T x_i)]$$

$$= -\sum_{i=1}^{N}[-\frac{(1-y_i)}{1-\sigma(w^T x_i)} + \frac{y_i}{\sigma(w^T x_i)}]\frac{\partial}{\partial w_j}\sigma(w^T x_i)$$

$$= -\sum_{i=1}^{N}[-\frac{(1-y_i)}{1-\sigma(w^T x_i)} + \frac{y_i}{\sigma(w^T x_i)}]\sigma(w^T x_i)(1-\sigma(w^T x_i))x_i$$

$$= -\sum_{i=1}^{N}[\frac{y_i - \sigma(w^T x_i)}{\sigma(w^T x_i)[1-\sigma(w^T x_i)]}]\sigma(w^T x_i)(1-\sigma(w^T x_i))x_i$$

$$\frac{\partial NLL(w)}{\partial(w)} = -\sum_{i=1}^{N} x_i\,[-\sigma(w^T x_i) + \ y_i]$$

1.2

a.

$$l(w) = (1-y_t)\ \log\left(1-\sigma(w^T x_t)\right) + y_t\, log\ \sigma(w^T x_t)$$

b.

$$w_t = \ \ n(x_t(-\sigma(w_{t-1}^T x_t) + y_t)\,) + w_{t-1}$$

c.

If x_t is very sparse, then the time complexity will be O(n) (or close to O(1)), where n are (x_t, y_t) pairs.

d.

With a large n we can cover more ground with fewer iterations. But a large n has a greater probability of missing the convergence/lowest point since the slope of the hill is changing constantly. With a very small n, we can move in the direction of the negative gradient since we are recalculating it repeatedly. A low n is more precise but calculating the gradient and to also to get to the lowest point will be time consuming.

e.

$$\frac{\partial(l-\mu\|\mathbf{w}\|_2^2)}{\partial(w)} = -2\mu w + x_t(-\sigma(w^T x_t) + y_t)$$

$$w_t = n(x_t(-\sigma(w_{t-1}^T x_t) + y_t)) + w_{t-1} - 2n\mu w_{t-1}$$

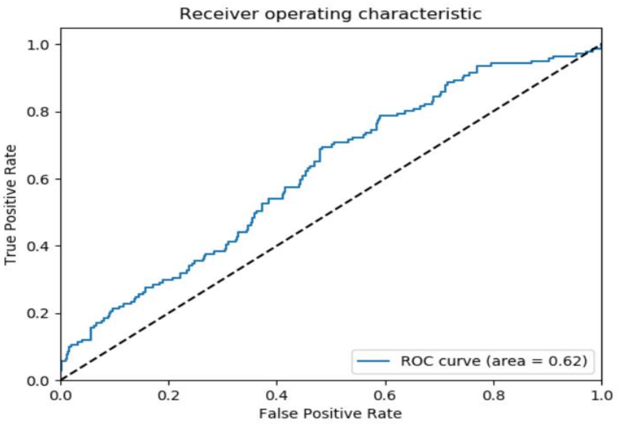
The time complexity of w_t with n dimensions will be O(n).

2.1

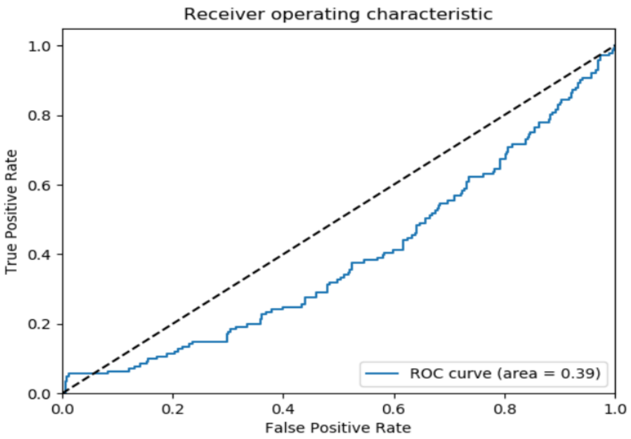
Metric	Deceased patients	Alive patients
Event Count		
1. Average Event Count	1027.7385229540919	683.1552587646077
2. Max Event Count	16829	12627
3. Min Event Count	2	1
Encounter Count		
1. Average Encounter Count	24.839321357285428	18.695492487479132
2. Max Encounter Count	375	391
3. Min Encounter Count	1	1
Record Length		
1. Average Record Length	157.04191616766468	194.70283806343906
2. Median Record Length	25.0	16.0
3. Max Record Length	5364	3103
4. Min Record Length	0	0
Common Diagnosis	DIAG320128 DIAG319835 DIAG313217 DIAG197320 DIAG132797	DIAG320128 DIAG319835 DIAG317576 DIAG42872402 DIAG313217
Common Laboratory Test	LAB3009542 LAB3023103 LAB3000963 LAB3018572 LAB3016723	LAB3009542 LAB3000963 LAB3023103 LAB3018572 LAB3007461
Common Medication	DRUG19095164 DRUG43012825 DRUG19049105 DRUG956874 DRUG19122121	DRUG19095164 DRUG43012825 DRUG19049105 DRUG19122121 DRUG956874

2.3 The below ROC curves indicates that when regularization parameter gets larger then the Area Under the Curve (AUC) gets smaller and as the learning rate increases the ROC becomes smoother. From the below models, the highest AUC 0.65, which could be considered a better model when compared to the rest.

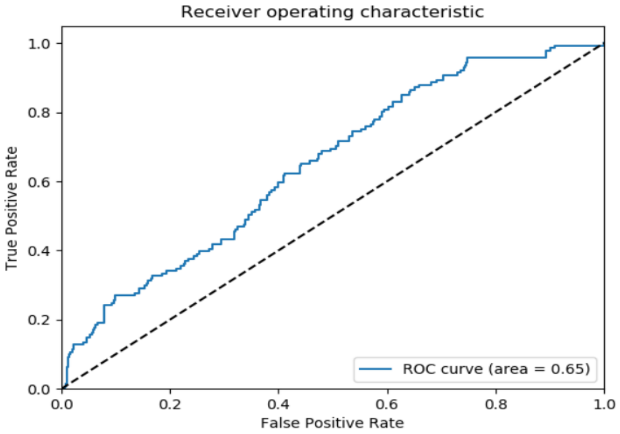
For default of n = 0.01 and mu = 0.0



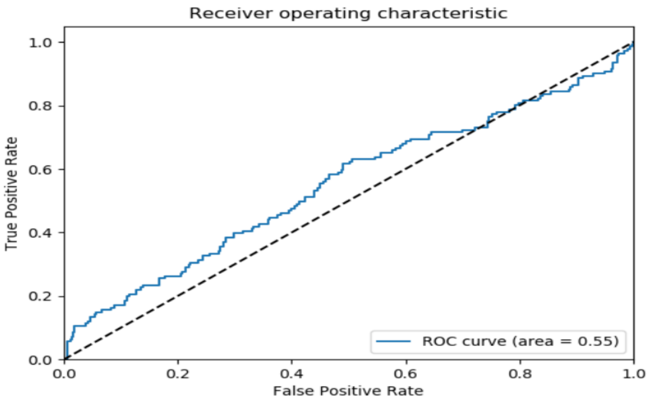
For n= 0.01 and mu = 0.5



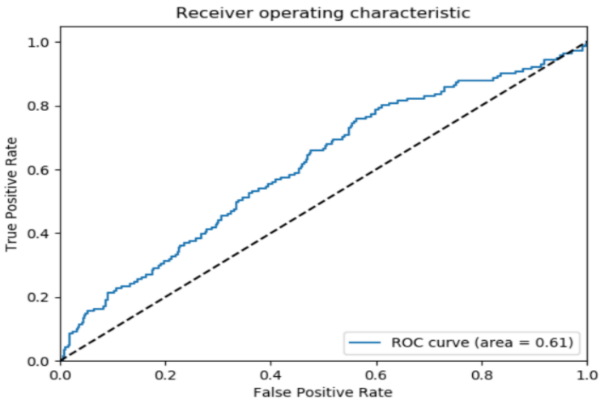
For n= 0.1 and mu = 0.0



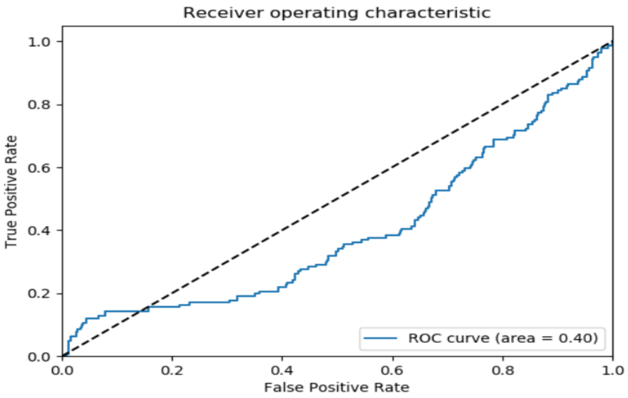
For n= 0.1 and mu = 0.01



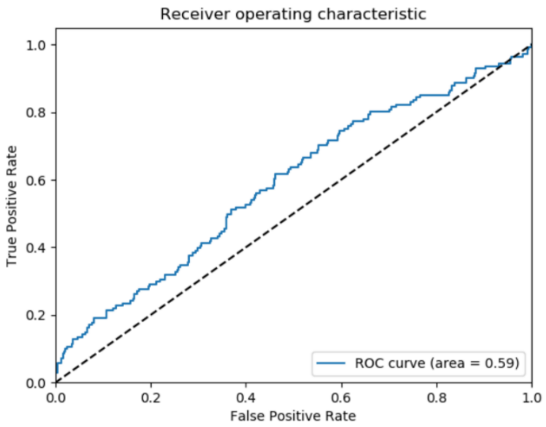
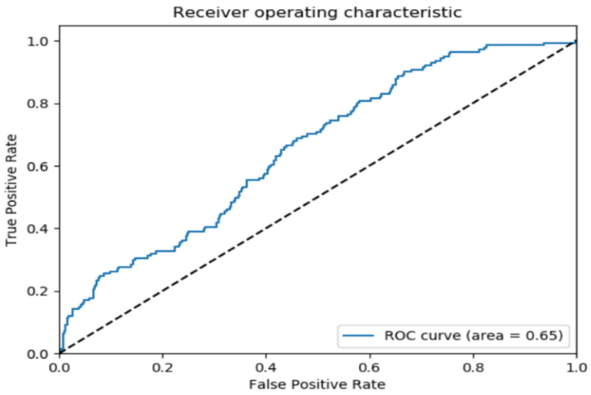
For n = 0.5 and mu = 0.0



For n = 0.5 and mu = 0.01



2.4 c. Ensemble methods combine multiple models thus yielding a better predictive performance with reduced variance and bias. It is surprising to see that with the default values the performance reduced when compared to the previous model. This could be because ensemble techniques typically involve using several models (not just a few) and here, we just have 5 which might not be enough. But with the number of models increased to 60, the performance increased significantly compared to the previous model.

Default n = 5 and r = 0.4  	n = 60 and r = 0.4  Receiver operating characteristic True Positive Rate False Positive Rate 0.0 0.2 0.4 0.6 0.8 1.0 ROC curve (area = 0.65)