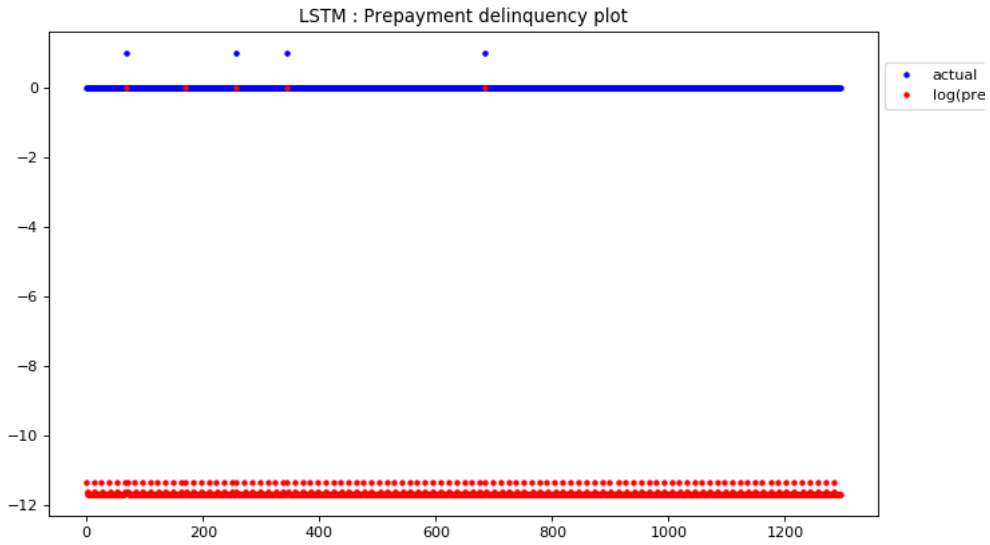
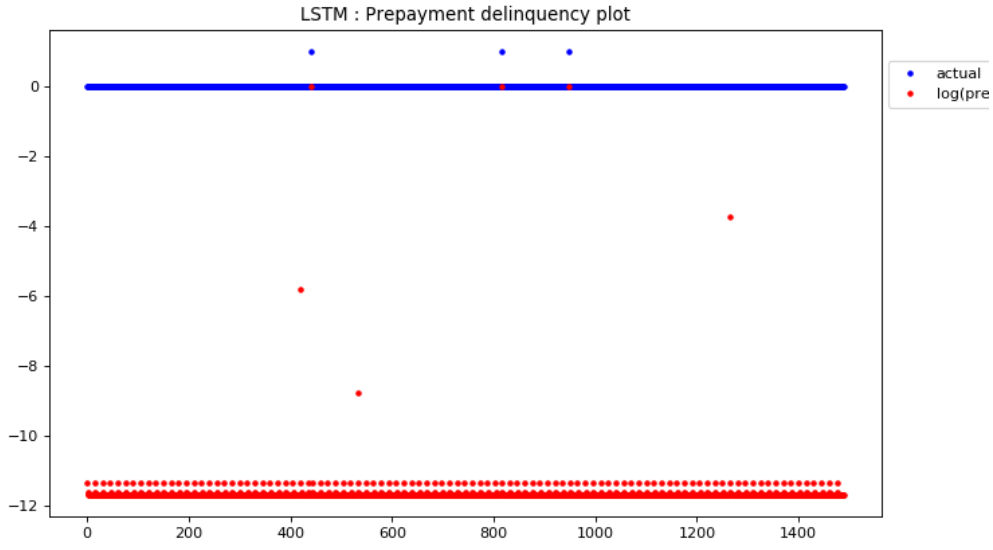
# Data Analysis of the Predicted Probabilities of Prepayment

The following plots are for **out-of sample loans originating between 2016-2017**. X-axis shows the indices. **Blue Dots** are the values of True values Prepayment (0 means no prepayment, 1 means prepayment). **Red Dots** are log(Predicted Probability) in case of Feed forward and LSTM network, and in case of logistic regression they mean Predicted Probabilities.

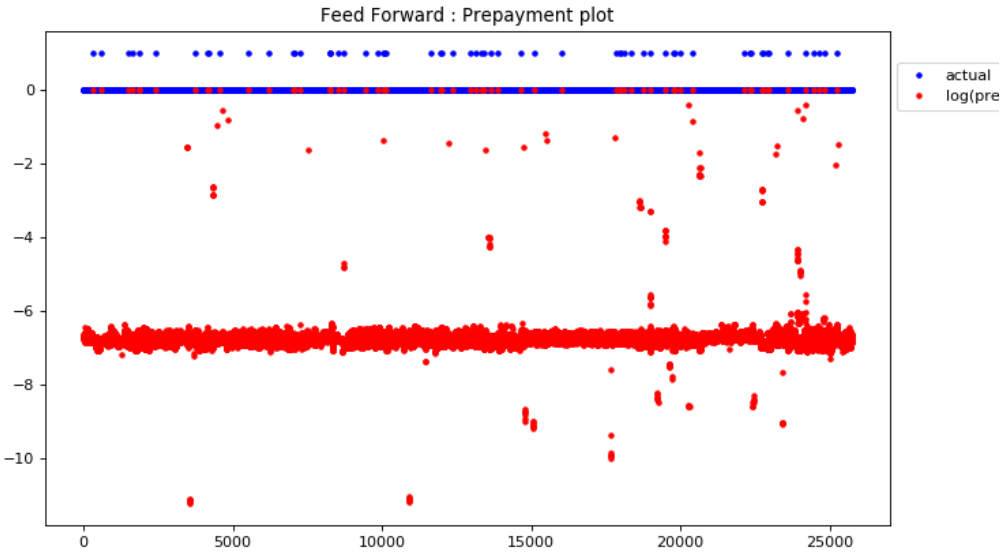
**For LSTM:** Note that for every positive prepayment the predicted probability is very high compared to other case.

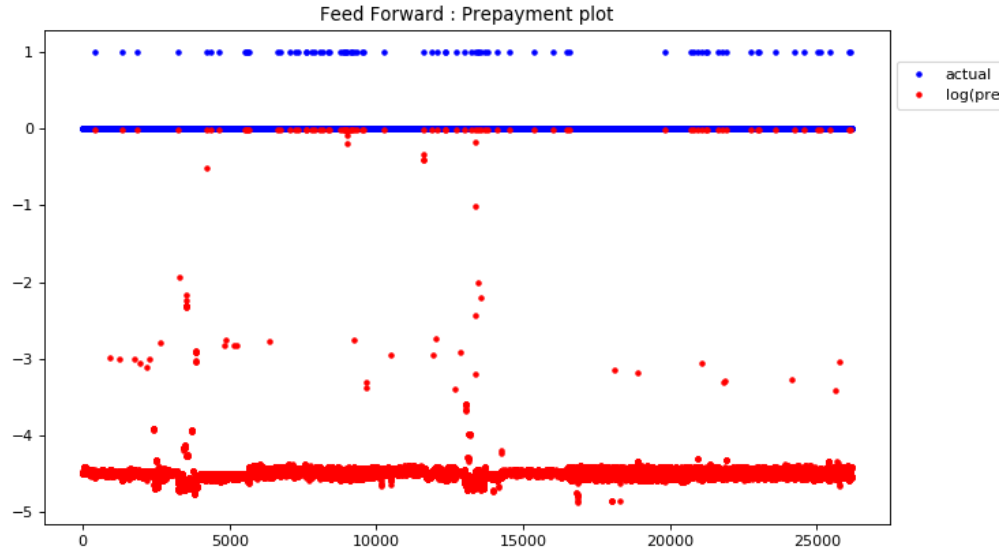



**LSTM 128 hidden units, dropout 0.5**
training AUC: [0.9986, 0.9994, 0.9995, 0.9995]
testing AUC: [0.9995, 0.9995, 0.9995, 0.9991]
test loss: `0.08477056358854673`

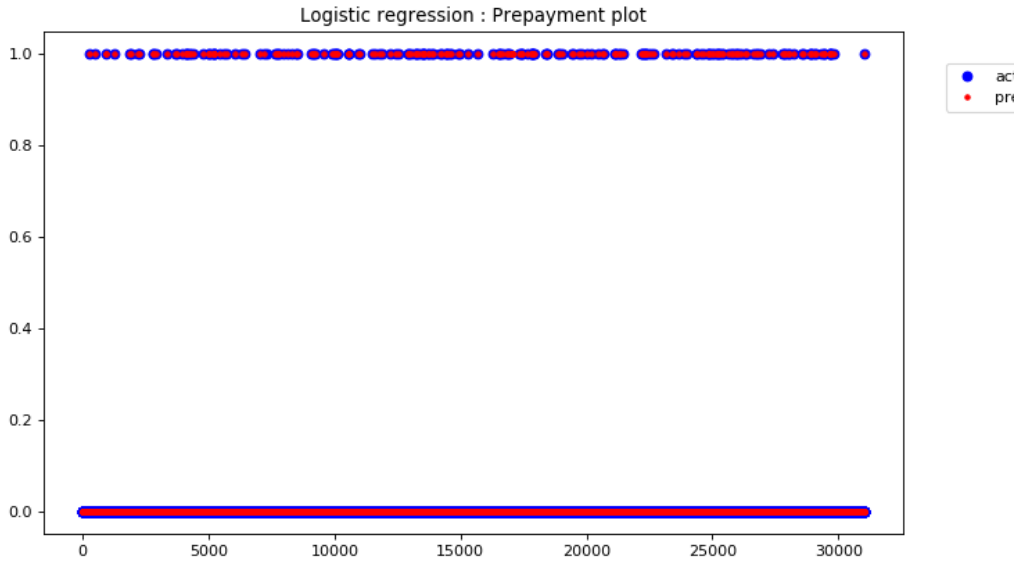
**For Feed Forward:** Note that for every positive prepayment the predicted probability is very high compared to other case. The other cases are much more noisier than the LSTM.

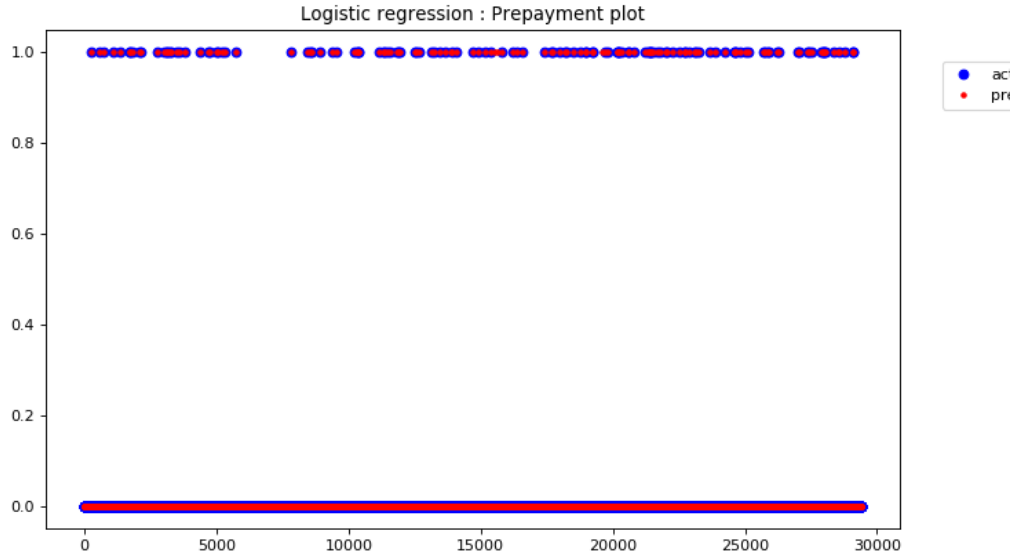



**5 layer FF : Number of hidden units in each hidden layer = [90, 70, 60, 50]**
training AUC: [0.9908, 0.9989, 0.9997]
testing AUC: [0.9988, 0.9999, 0.9999]
test loss: `0.6411`

**For Logistic Regression:** For Logistic Regression the predicted probabilities look very accurate but are even more noisy. Log(Predicted Probability) is in the range of -2000 to -4000 so I had to plot the Predicted Probabilities instead of the log values. **This is the reason the loss value for Logistic Regression is so high.**



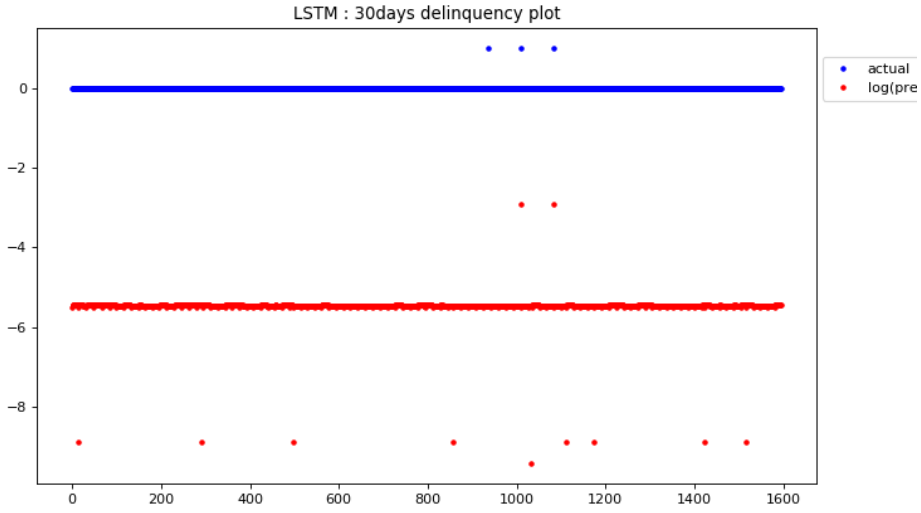



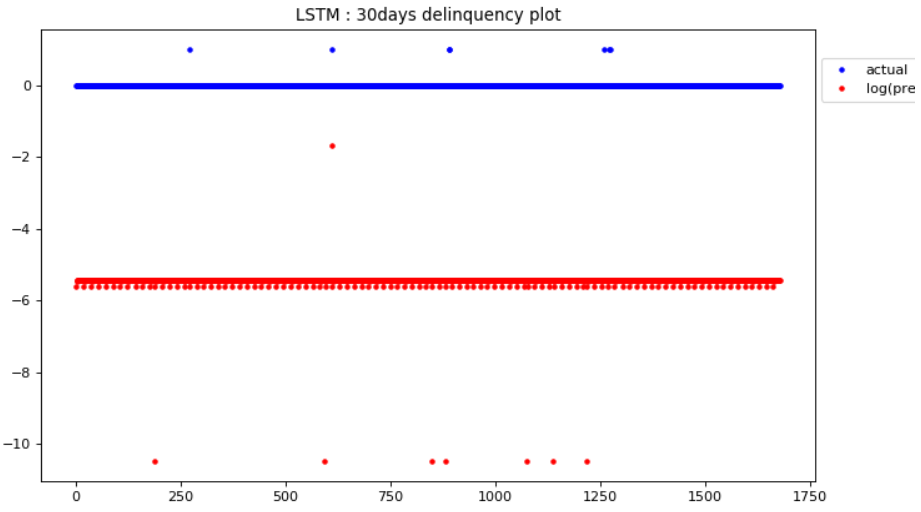

**Logistic Regression**
training AUC: [0.9689, 0.9982, 0.9982, 0.9983, 0.9983]
testing AUC: [0.9999, 0.9999, 0.9999, 0.9999, 0.9999]
test loss: 6.3246

# Data Analysis of the Predicted Probabilities of 30days delinquency

The following plots are for **out-of sample loans originating between 2016-2017**. X-axis shows the indices. **Blue Dots** are the values of True values 30 days delinquency (0 means no delinquency, 1 means 30 days delinquent). **Red Dots** are log(Predicted Probability) in case of Feed forward and LSTM network, and in case of logistic regression they mean Predicted Probabilities.

**For LSTM:** For 30 Days delinquency the AUC is not that good as we can see the predicted probabilities for positive cases of delinquency are not high. **LSTM performs better than Feed Forward network in predicting 30 days delinquency.**

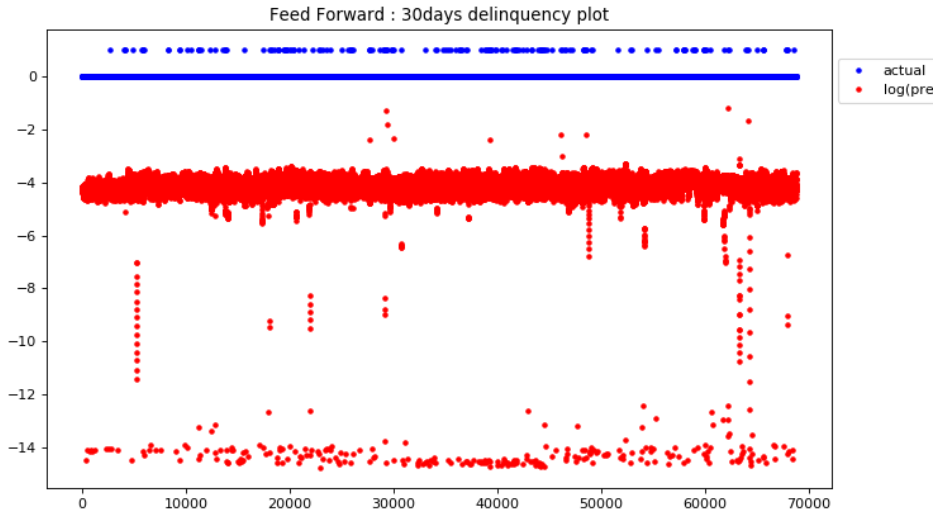


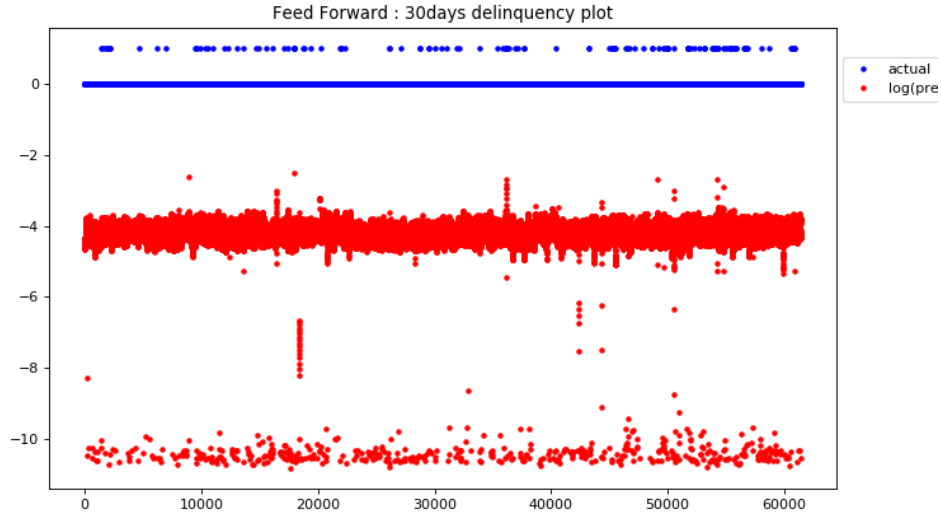




**LSTM 128 hidden units, dropout 0.5**
testing AUC: [0.6619, 0.7712, 0.7018, 0.7005]

test loss: `0.08477056358854673`

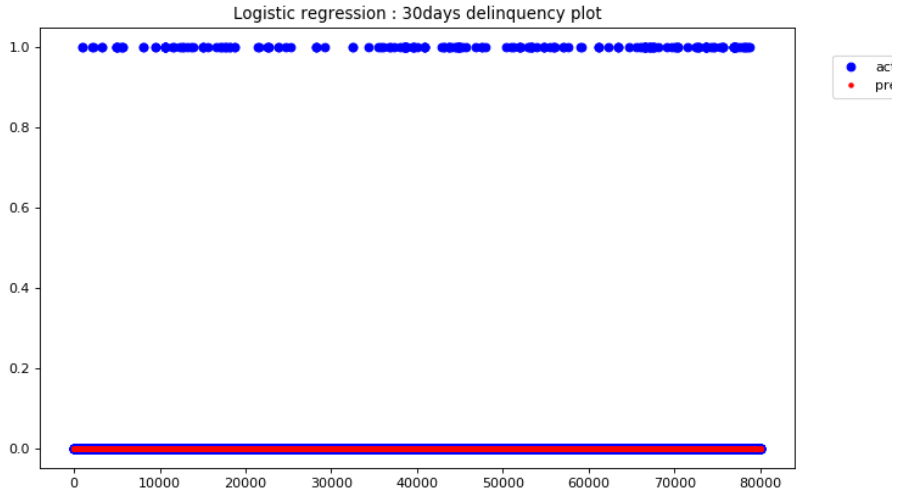
**For Feed Forward:** Even in Feed forward the predicted probabilities for 30 days delinquency status has a lot of noise and not very accurate. This seems to be a harder problem.

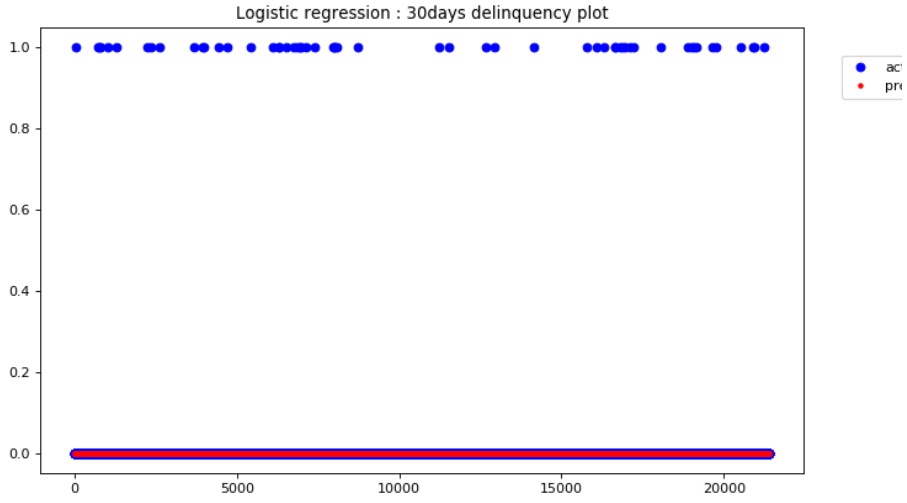



**5 layer FF : Number of hidden units in each hidden layer = [90, 70, 60, 50]**
testing AUC: [0.6798, 0.6833, 0.6234]
test loss: `0.6411`

**For Logistic Regression:** The worst performance in this problem is by logistic regression since almost it predicts all the examples of 30 days delinquency wrongly. It is as good as random guessing. Log(Predicted Probability) is in the range of -2000 to -4000 so I had to plot the Predicted Probabilities instead of the log values.




**Logistic Regression**
testing AUC: [0.4694, 0.4611, 0.4655, 0.4737, 0.4850]
test loss: 6.3246