Experiment

- Used loans originating in 2006-2015 as training set and loans originating in 2016-2017 as test set.
- Each epoch was trained with a subset of the training set and not the entire training set.
- After each epoch testing was done on the entire testset.
- Ran 40 Epochs to reduce cross entropy loss.
- AUC was calculated for transition to Prepayment status and transition to 30 days delinquency status.

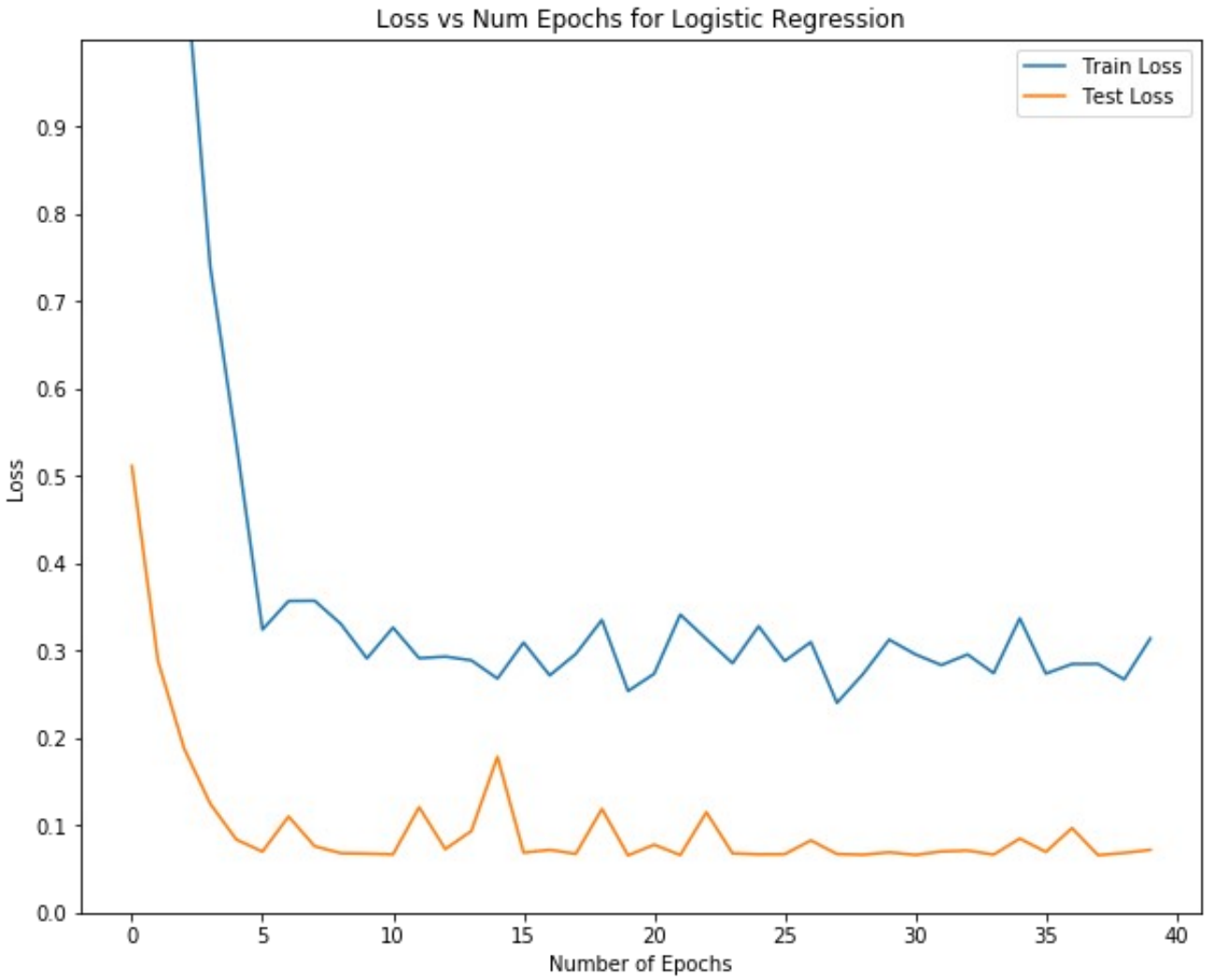
Following are the best features for the respective models wrt AUC.
- Feed forward network with 5 layers, Logistic regression with features:
1) ORIGINAL COMBINED LOAN-TO-VALUE
3) ORIGINAL DEBT-TO-INCOME (DTI) RATIO
4) ORIGINAL LOAN-TO-VALUE (LTV)
8) CREDIT SCORE
11) CURRENT INTEREST RATE
14) FIRST TIME HOMEBUYER FLAG - one hot encoded
16) LOAN AGE
26) NUMBER OF BORROWERS
27) NUMBER OF UNITS
28) OCCUPANCY STATUS - one hot encoded
29) ORIGINAL INTEREST RATE
30) PRODUCT TYPE - one hot encoded
31) LOAN PURPOSE - one hot encoded
32) STATE - one hot encoded

- LSTM were the models explored with all the following features:
1) ORIGINAL COMBINED LOAN-TO-VALUE
3) ORIGINAL DEBT-TO-INCOME (DTI) RATIO
4) ORIGINAL LOAN-TO-VALUE (LTV)
5) ORIGINAL UPB
6) ACTUAL LOSS CALCULATION
7) CHANNEL - one hot encoded
8) CREDIT SCORE
10) CURRENT DEFERRED UPB
11) CURRENT INTEREST RATE
12) DEFERRED PAYMENT MODIFICATION - one hot encoded
13) EXPENSES
14) FIRST TIME HOMEBUYER FLAG - one hot encoded
15) LEGAL COSTS
16) LOAN AGE
17) MAINTENANCE AND PRESERVATION COSTS
18) MI RECOVERIES
19) MISCELLANEOUS EXPENSES
20) MODIFICATION COST
21) MODIFICATION FLAG - one hot encoded
22) MONTHLY REPORTING PERIOD
23) MORTGAGE INSURANCE PERCENTAGE (MI %)
24) METROPOLITAN STATISTICAL AREA (MSA)
25) NON MI RECOVERIES
26) NUMBER OF BORROWERS
27) NUMBER OF UNITS

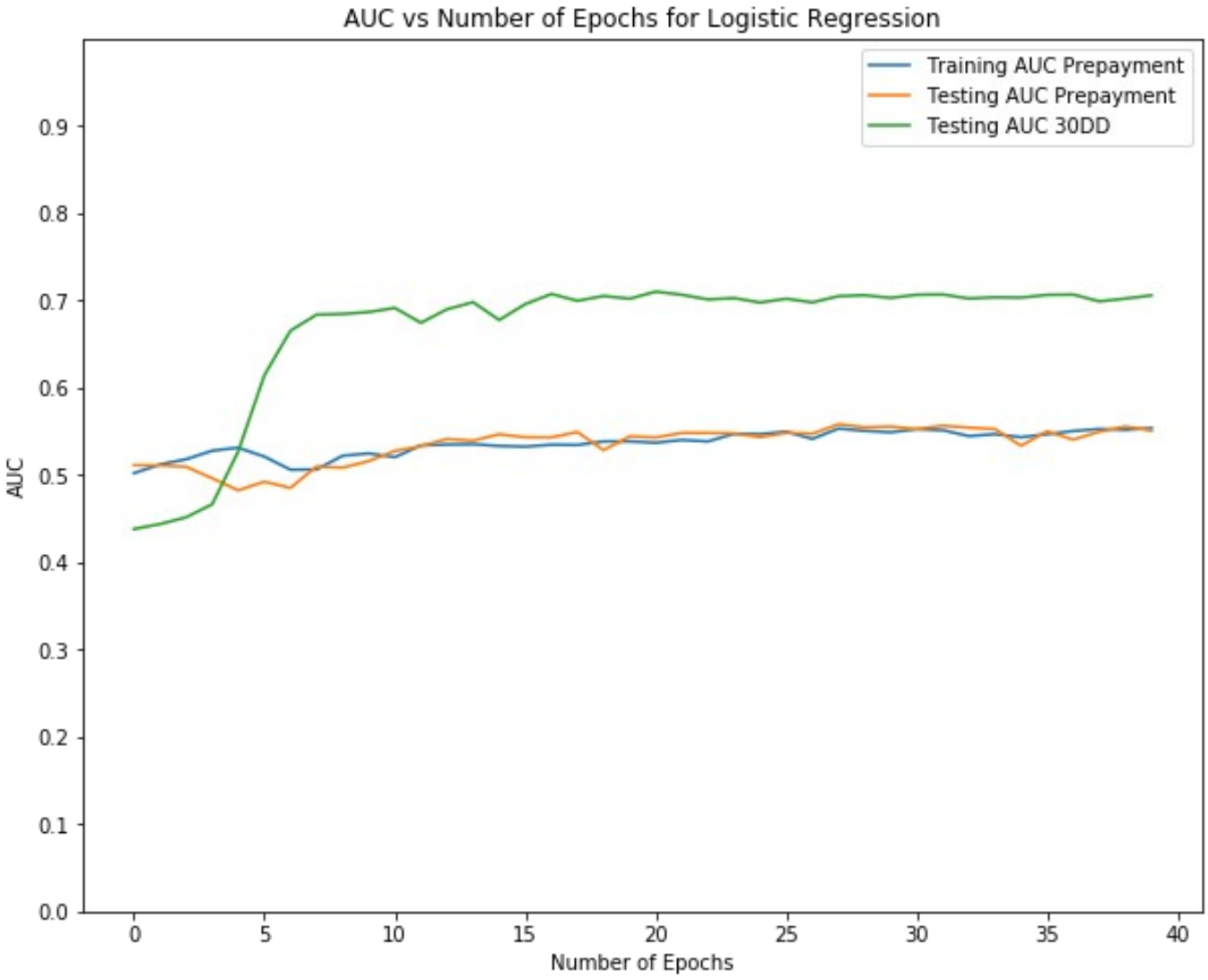
28) OCCUPANCY STATUS - one hot encoded
29) ORIGINAL INTEREST RATE
30) PRODUCT TYPE - one hot encoded
31) LOAN PURPOSE - one hot encoded
32) STATE - one hot encoded
33) STEP MODIFICATION FLAG - one hot encoded
34) SUPER CONFORMING FLAG - one hot encoded
35) TAXES AND INSURANCE

Logistic Regression

Plot of training and testing cross entropy loss against the number of epochs.



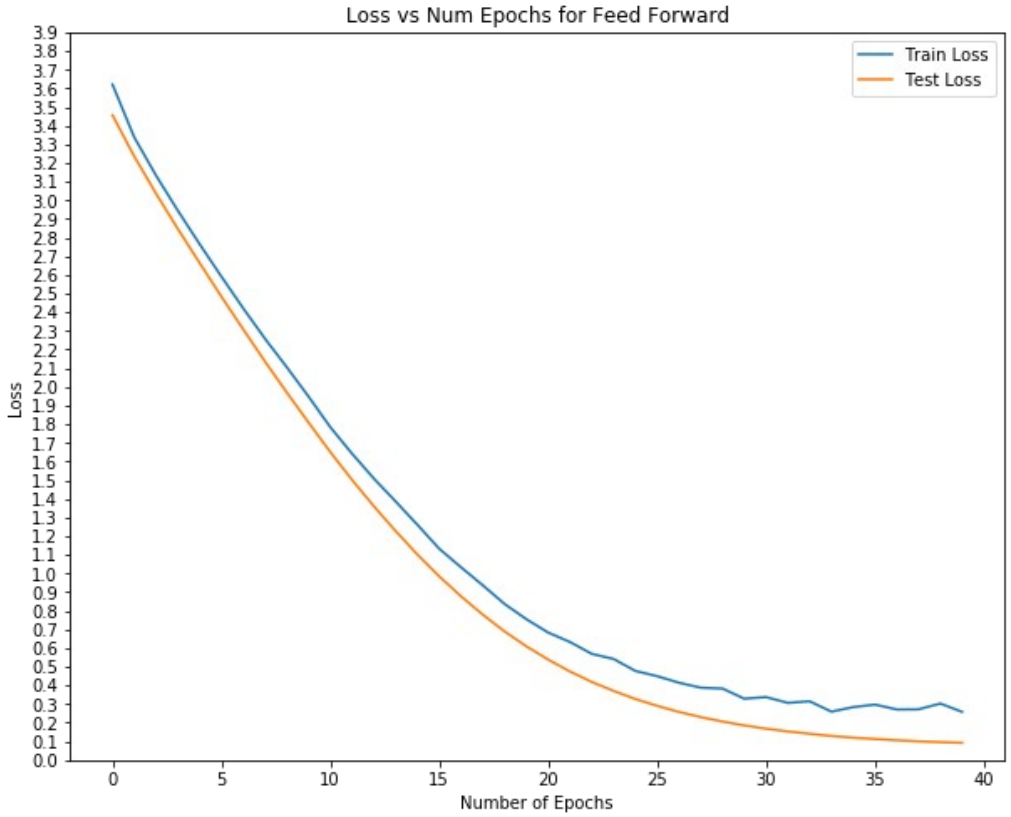
Plot of AUC for two types of transition:
- Orange line is test AUC for transition to Prepayment status
- Green line is test AUC for transition to 30 Days Delinquency status

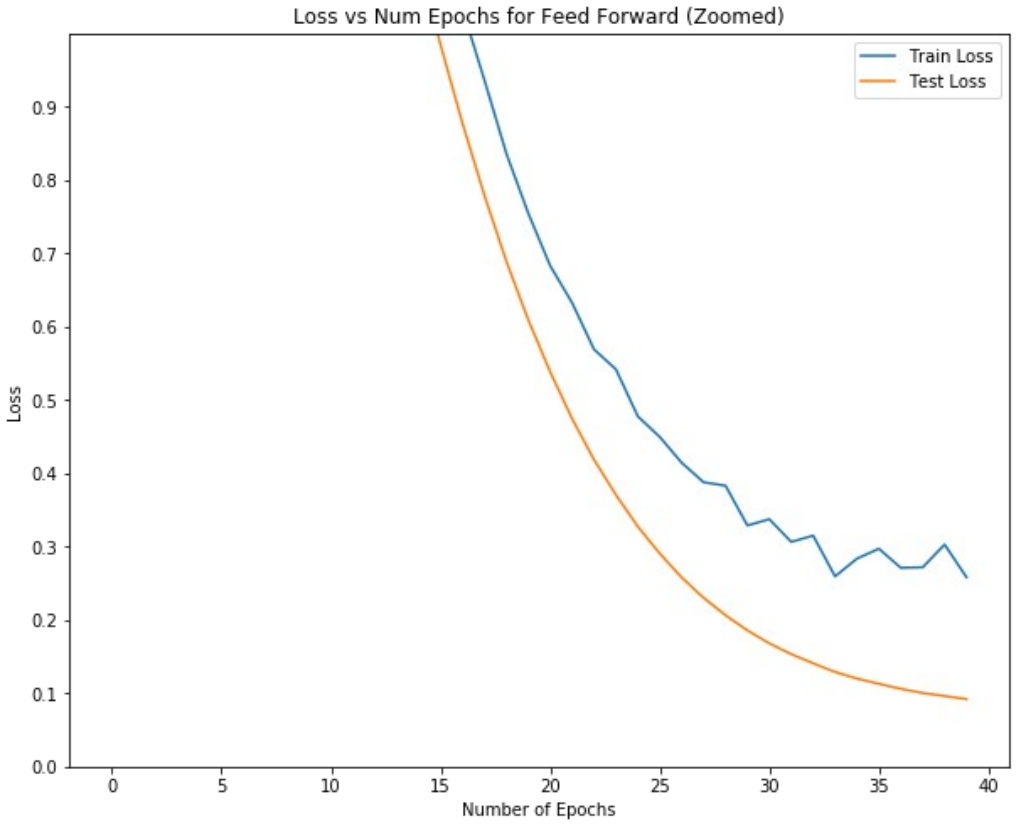


AUC for transition to Prepayment is around 0.5 but AUC for 30DD is around 0.7.
(Look at comparative study of the 3 models for more details)

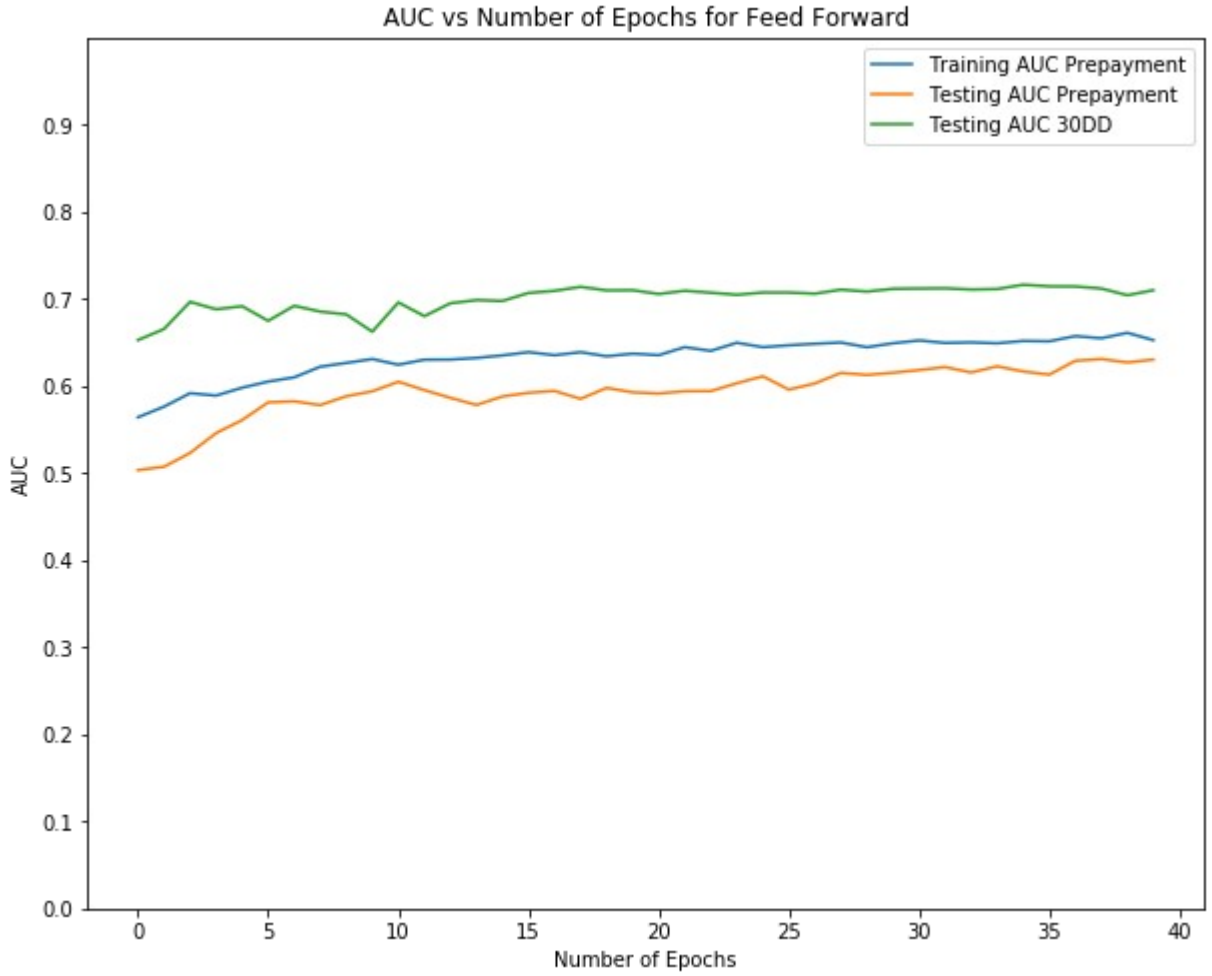
Feed Forward network

Plot of training and testing cross entropy loss against the number of epochs.





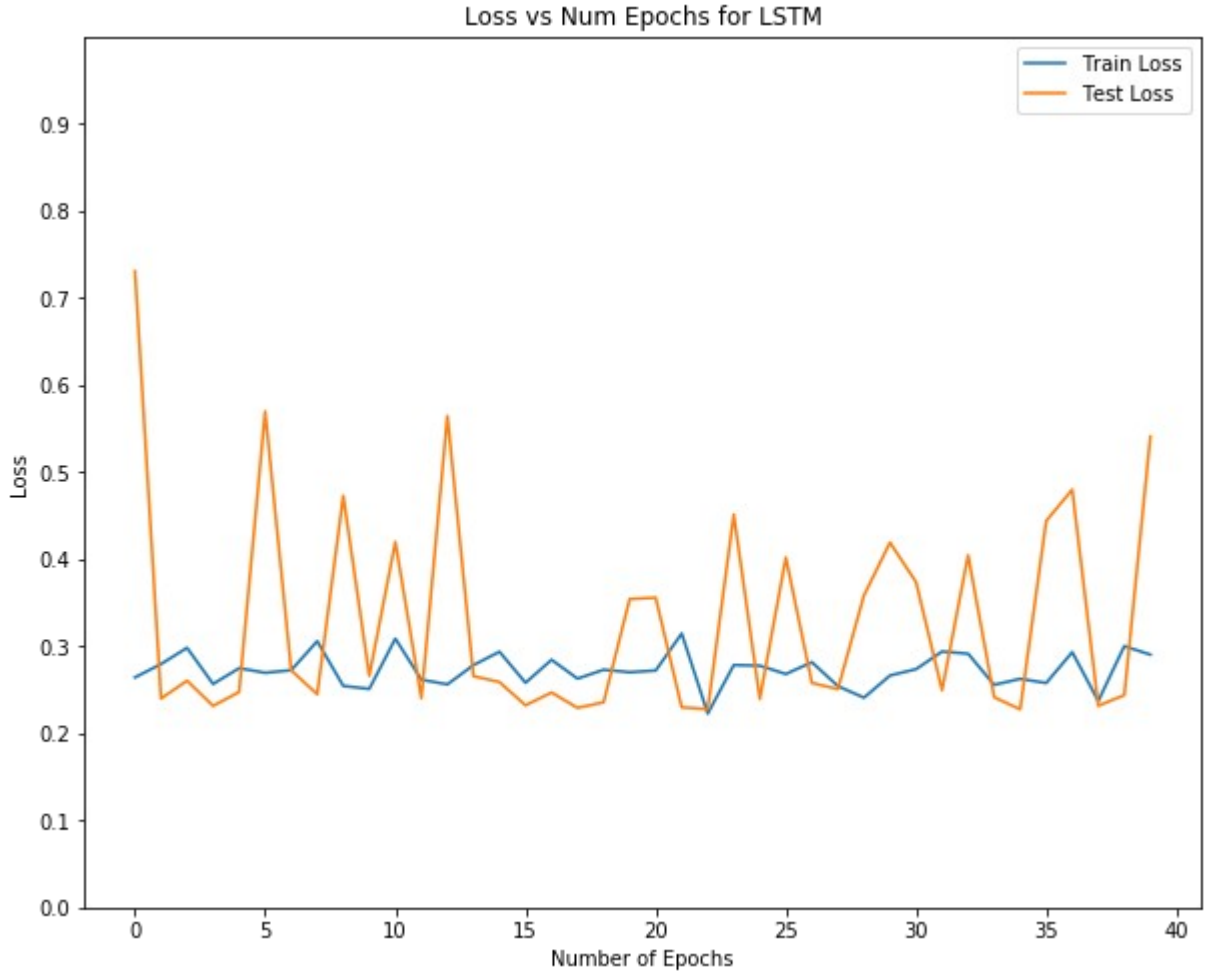
Plot of AUC for two types of transition:
- Orange line is test AUC for transition to Prepayment status
- Green line is test AUC for transition to 30 Days Delinquency status



AUC for transition to Prepayment is around 0.6 which is an improvement on the logistic regression results but AUC for 30DD is around 0.7 here as well. (Look at comparative study of the 3 models for more details)

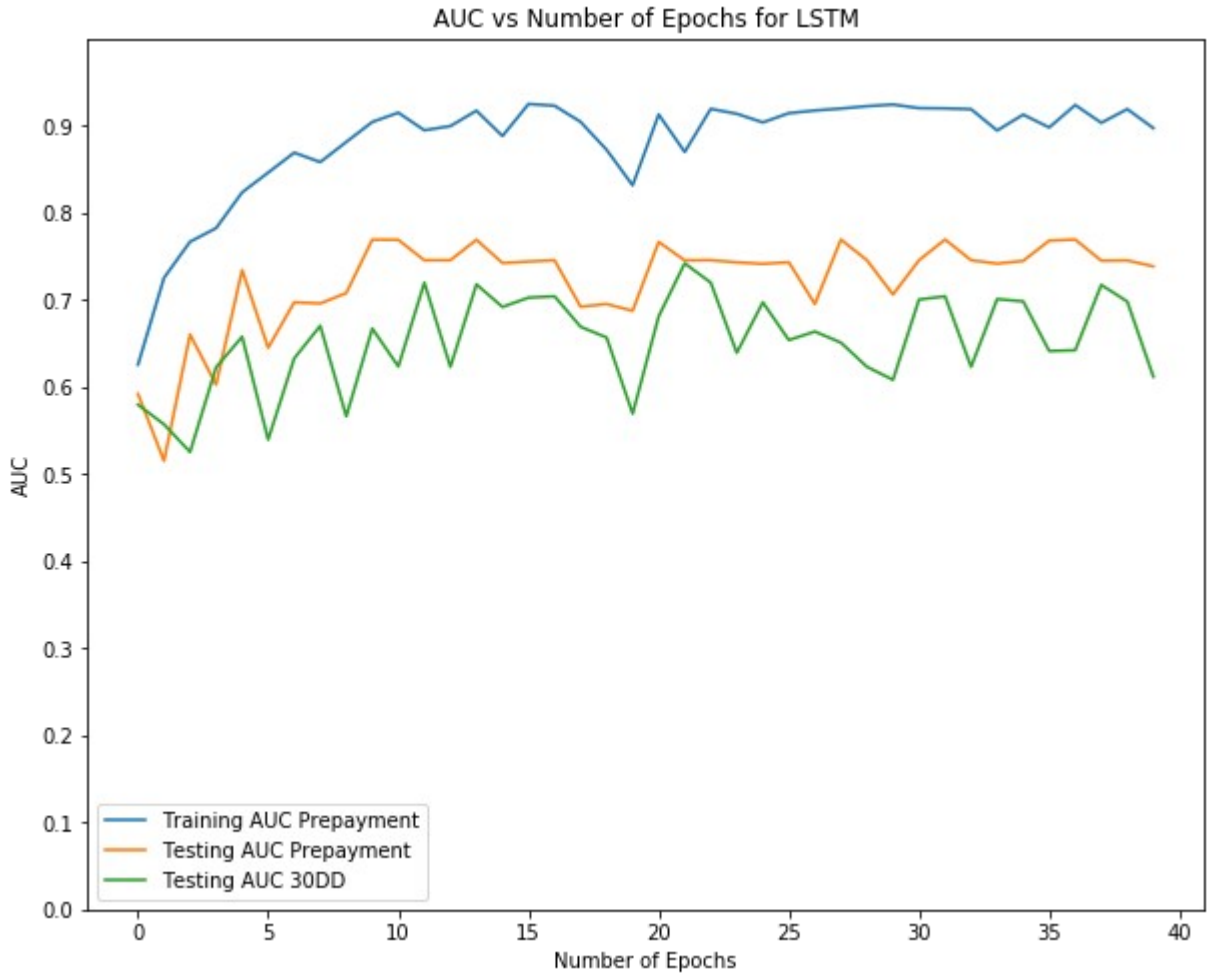
LSTM

Plot of training and testing cross entropy loss against the number of epochs.



The training loss (that is the loss that our models are trying to reduce) is lowest in LSTM compared to Logistic regression or Feed forward network. But test loss for LSTM is the worst, this is interesting. The Loss is calculated over all possible transitions and not just Prepayment, 30DD.

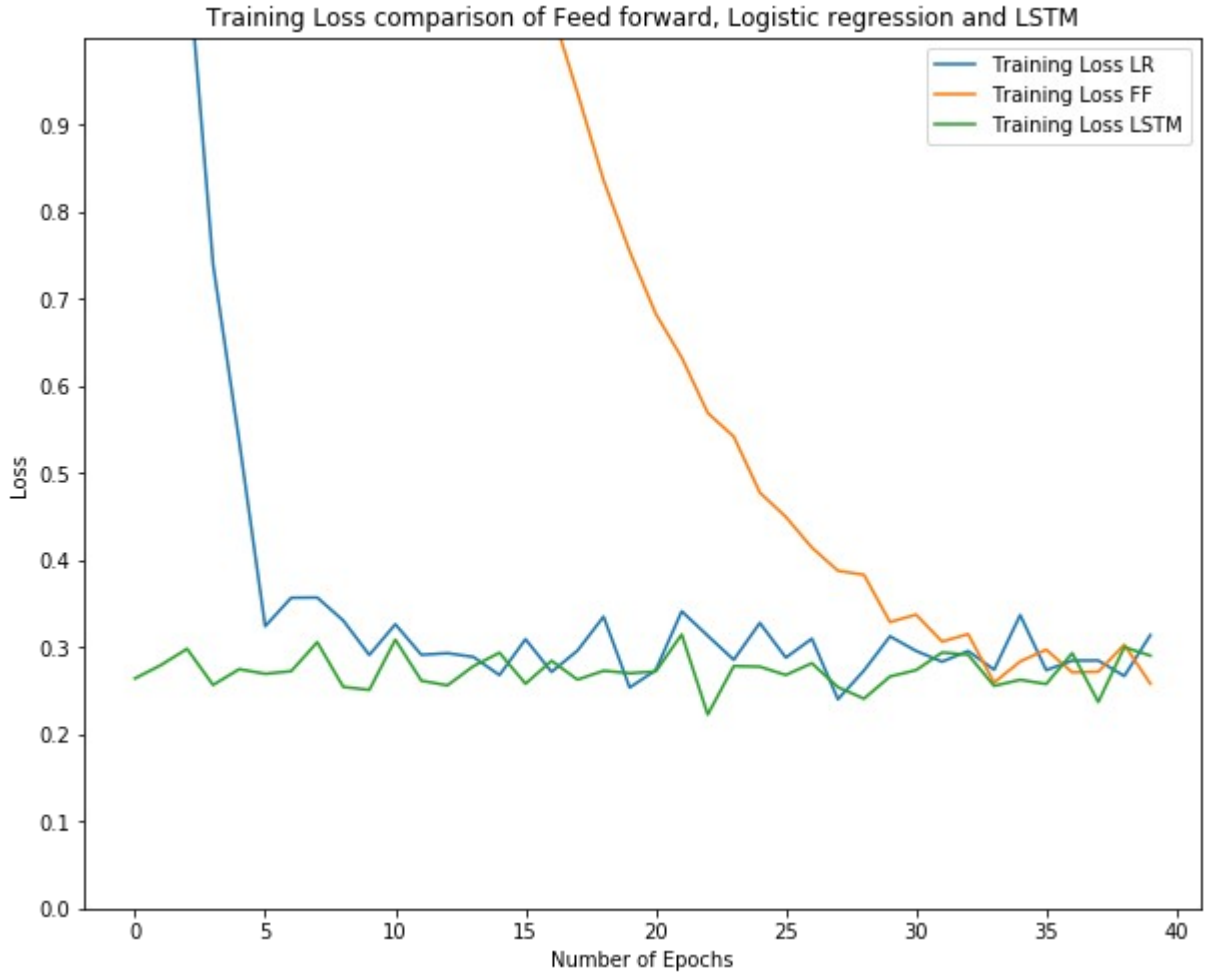
Plot of AUC for two types of transition:
- Orange line is test AUC for transition to Prepayment status
- Green line is test AUC for transition to 30 Days Delinquency status



Here also we see that AUC for prepayment is a little above 0.7 and AUC for 30DD is around 0.7. This is best out of the three models.
Now lets see the comparative study of the three models.

Comparison of Logistic regression, Feed forward network and LSTM

Plot of training cross entropy loss against the number of epochs.



- Feed forward takes time to converge.
- LSTM training cross entropy loss is the lowest (LSTM trains better)
- LSTM loss is quite smooth and does not fluctuate much, so it might be easier for the LSTM to converge.

Plot of AUC of transition to Prepayment on test set.

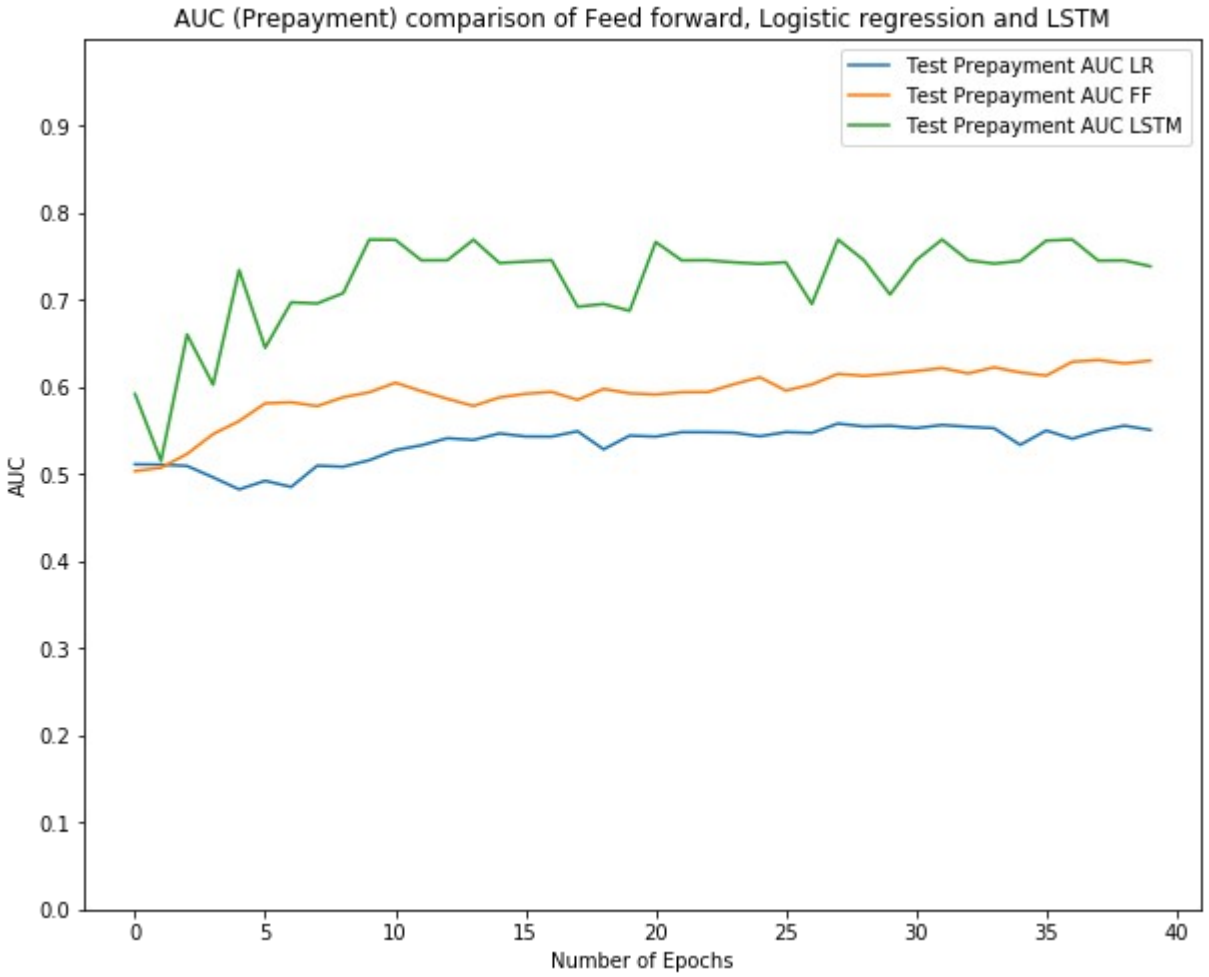


Table to show Max AUC and Mean AUC for Prepayment delinquency.

	Max AUC	**Mean of AUC**
Logistic Regression	0.558171	0.535332
Feed forward network	0.631124	0.593673
LSTM	0.769554	0.721963

Logistic Regression performs much worse than Feed forward network or LSTM network. LSTM as expected performs way better than Feed forward and logistic regression.

Plot of AUC of transition to 30DD test set.

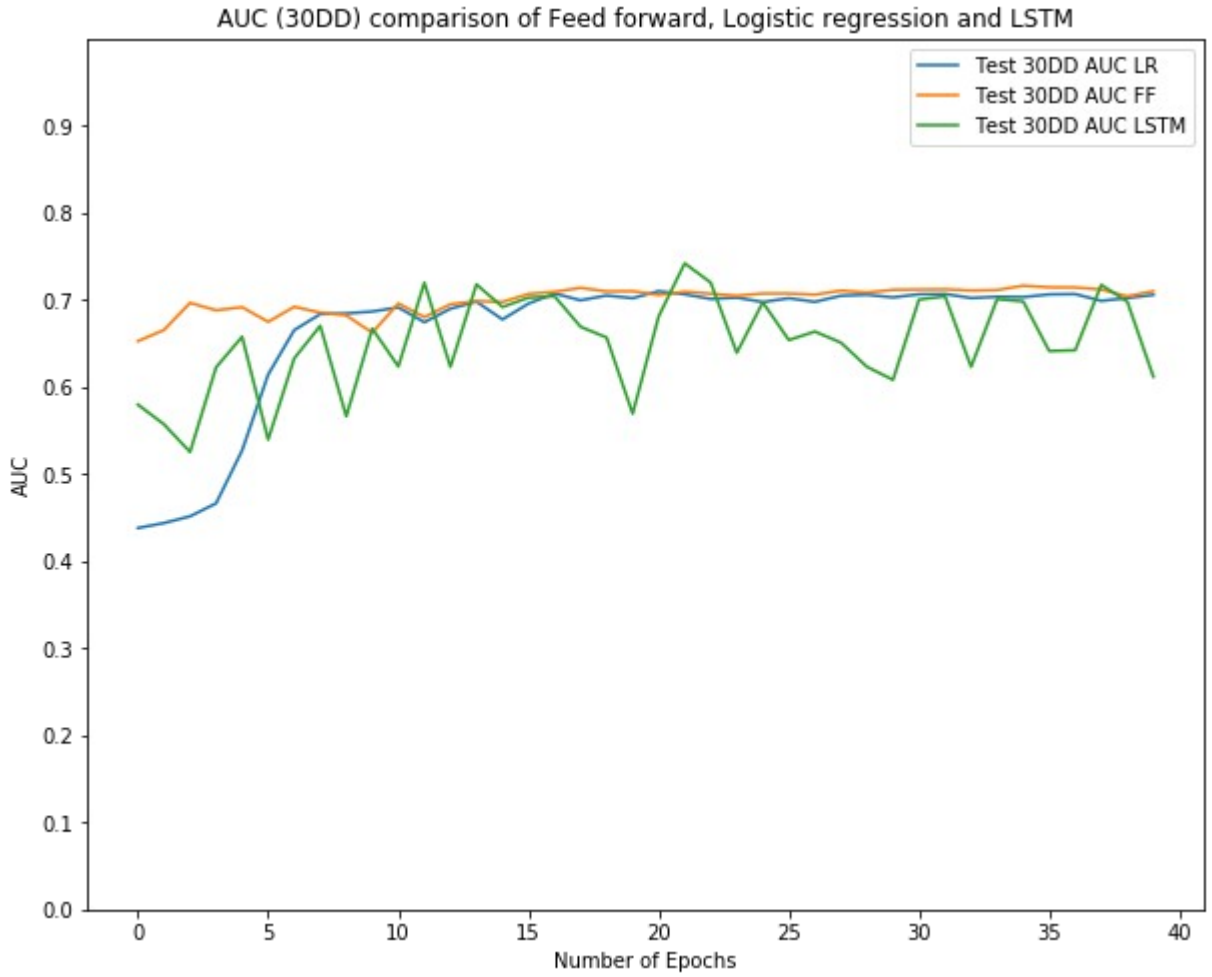


Table to show Max AUC and Mean AUC for 30 days delinquency.

	Max AUC	**Mean of AUC**
Logistic Regression	0.710236	0.667079
Feed forward network	0.716513	0.700304
LSTM	0.742221	0.653029

All the three models work similar in terms of performance.