A project report on predicting which customers are happy customers?

Pallavi Karan | CPSC 6820 –Data Science | April 25, 2016

ABSTRACT

Nurturing relationships with your customers is a crucial part of growing a successful business. As we know that unhappy customers don't stick around so Santander Bank asked Kagglers to help them identify dissatisfied customers early in their relationship. This will help Santander Bank to improve the customers' happiness before it's too late and help them stick around. The problem is to predict if a customer is satisfied or dissatisfied with their banking experience based on various anonymized features.

MOTIVATION

This particular project caught my eyes because it has 371 anonymized predictors, hence no domain expertise will help and no biased feature selections.

DATASET

The given datasets are in CSV file hence no conversion of data required.
There are *76,022 rows* and *371 columns* (370 predictors and 1 response variable) present in the dataset. The dataset is fairly large in terms of predictor values containing a large number of numeric variables.
The TARGET column is the variable to predict. The response variable, TARGET, is a binary outcome, 0 for a satisfied customer and 1 for a non-satisfied customer.

DATA CLEANING

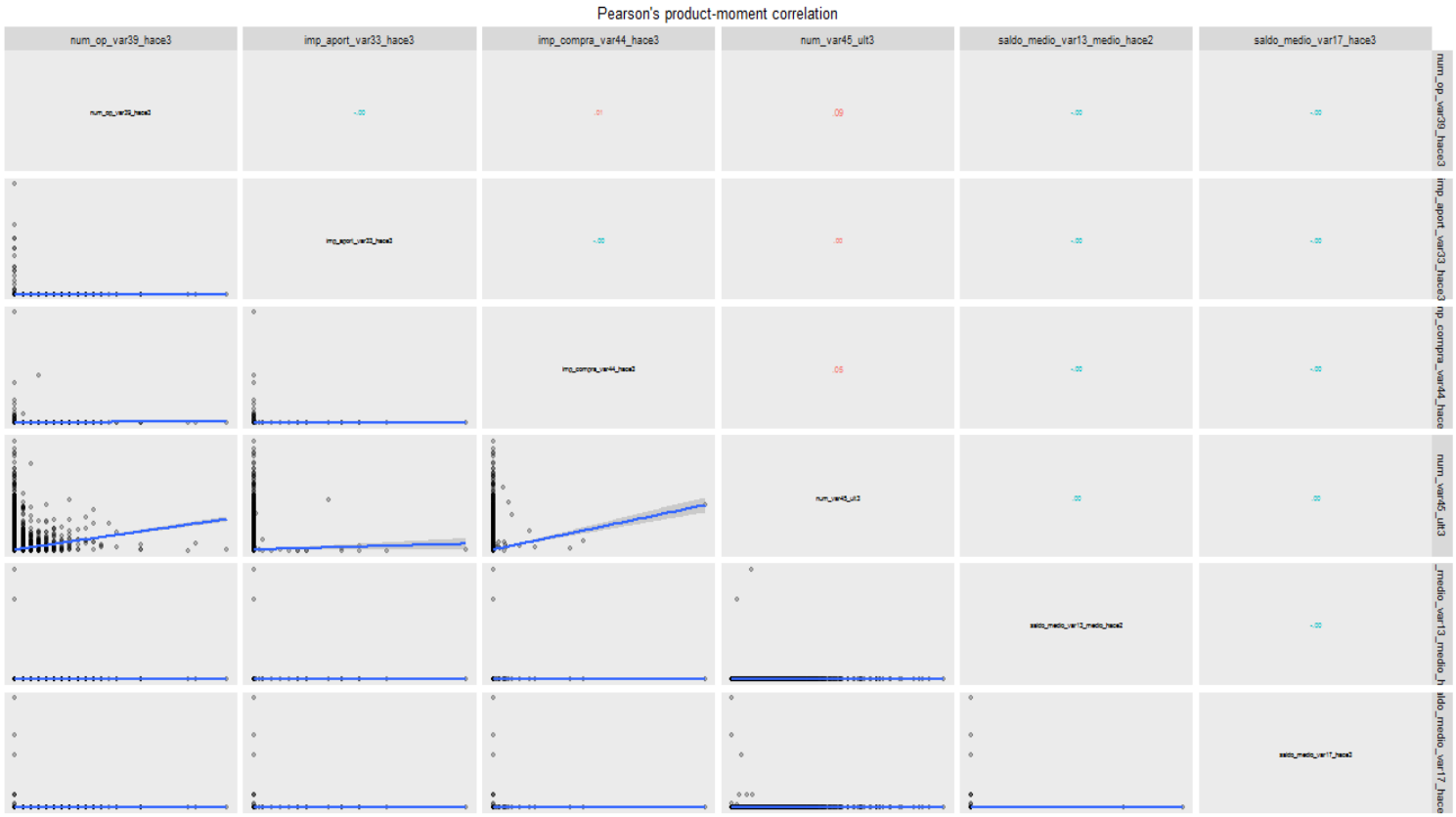
- The dataset is quantitative hence no text processing is required.
- The dataset has no missing values.
- Removing the columns that have maximum and minimum value=0 i.e the column has only one value=0. In total 34 columns were removed. Below are the listed columns.

ind_var2_0	ind_var2	ind_var27_0	ind_var28_0
ind_var28	ind_var27	ind_var41	ind_var46_0
ind_var46	num_var27_0	num_var28_0	num_var28
num_var27	num_var41	num_var46_0	num_var46
saldo_var28	saldo_var27	saldo_var41	saldo_var46
imp_amort_var18_hace3	imp_amort_var34_hace3	imp_reemb_var13_hace3	imp_reemb_var33_hace3
imp_trasp_var17_out_hace3	imp_trasp_var33_out_hace3	num_var2_0_ult1	num_var2_ult1
num_reemb_var13_hace3	num_reemb_var33_hace3	num_trasp_var17_out_hace3	num_trasp_var33_out_hace3
saldo_var2_ult1	saldo_medio_var13_medio_hace3		

- Removing Outliers: Removing outlier through Box Plot. Box Plot of all attributes are drawn. The values which are absurd and are not making sense to the dataset are removed.
- Summary of data set.

EXPLORATORY DATA ANALYSIS

- This problem is a classic supervised learning problem with binary response. Hence, the possible approaches to solve this problem are using basic models like Logistic regression, Support Vector machine and Decision Trees.
- Pearson Co-Relation



Pearson's co-relation can range from -1 to 1. We can see none of the graphs have negative co-relation. And the strong co-relation is between num_op_var39_hace3 & num_var45_ult3, and imp_compra_var44_hace3 & num_var45-ult3.

DATA MODELING

Since the response variable had 0 and 1 values and all the predictors were quantitative whole numbers hence logistic regression was an apt choice from the exploratory dataset.

Randomly split the dataset into train set and validation set using index
sample(Rindex,trunc(length(Rindex)*0.3))

LOGISTIC REGRESSION

Output of the training model:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-8.49	0.00	0.00	0.00	8.49

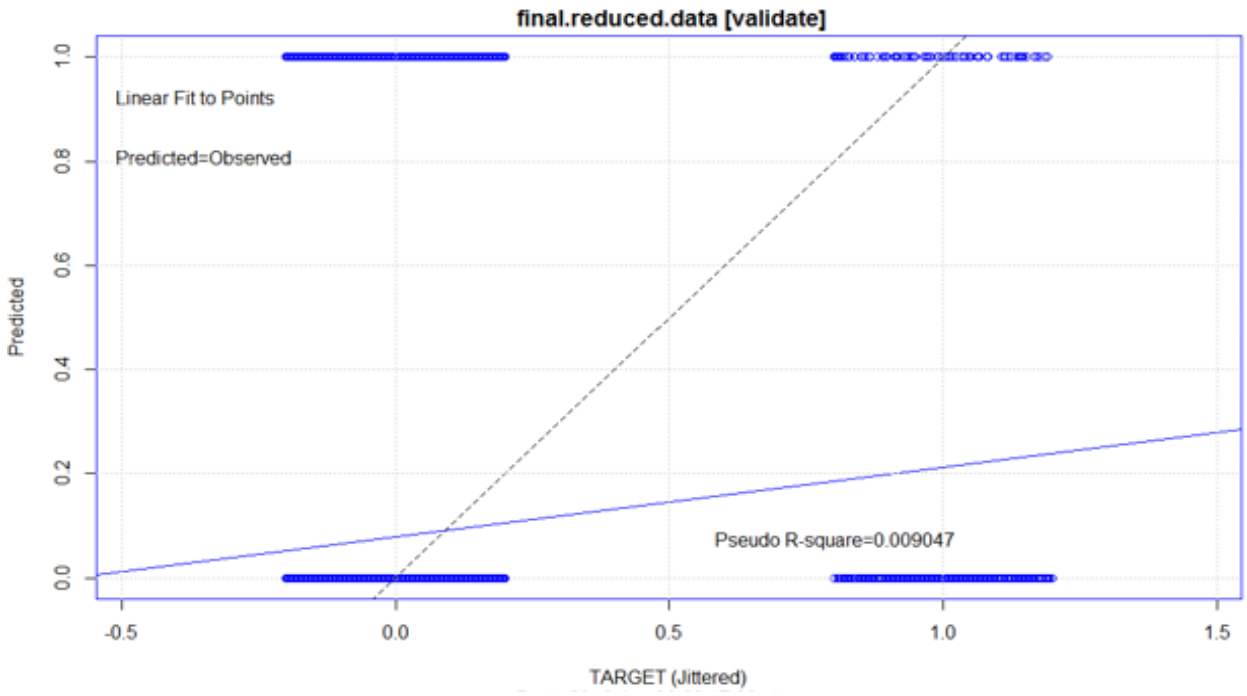
Number of Fisher Scoring iterations: 25
Log likelihood: -201267.761 (230 df)
Null/Residual deviance difference: -384849.877 (229 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.09847712
Error matrix for the Linear model on final.reduced.data [validate] (counts):

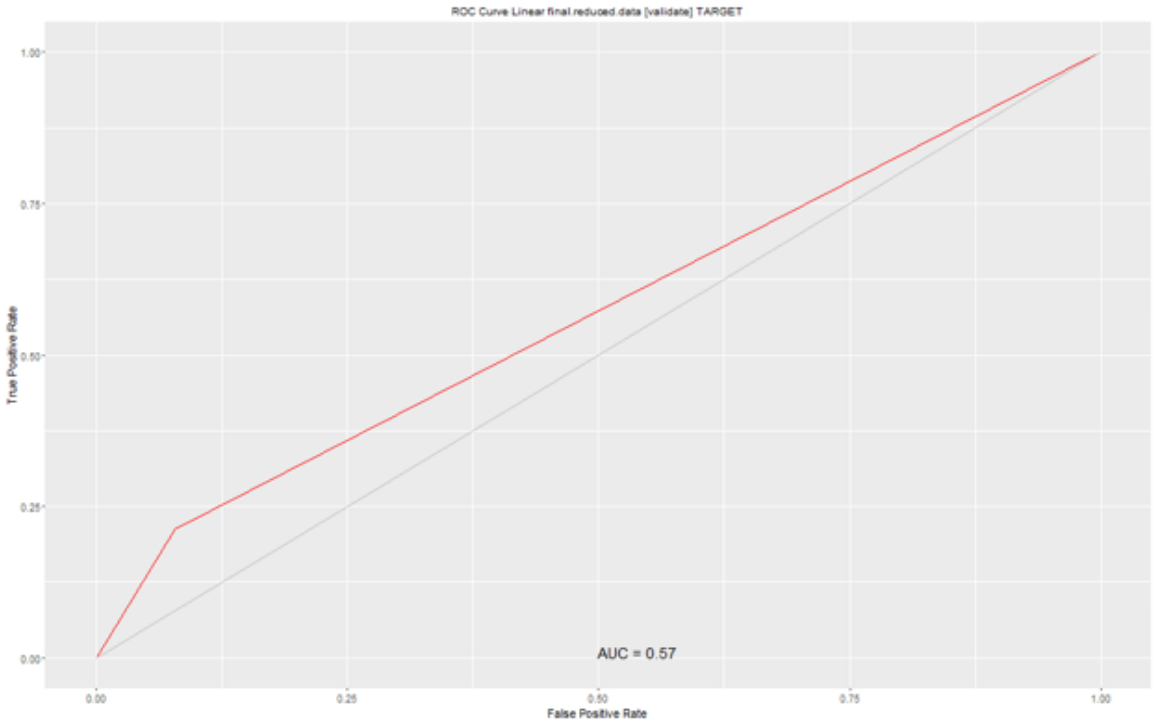
	Predicted	
Actual	0	1
0	10091	855
1	360	97

Error matrix for the Linear model on final.reduced.data [validate] (proportions):

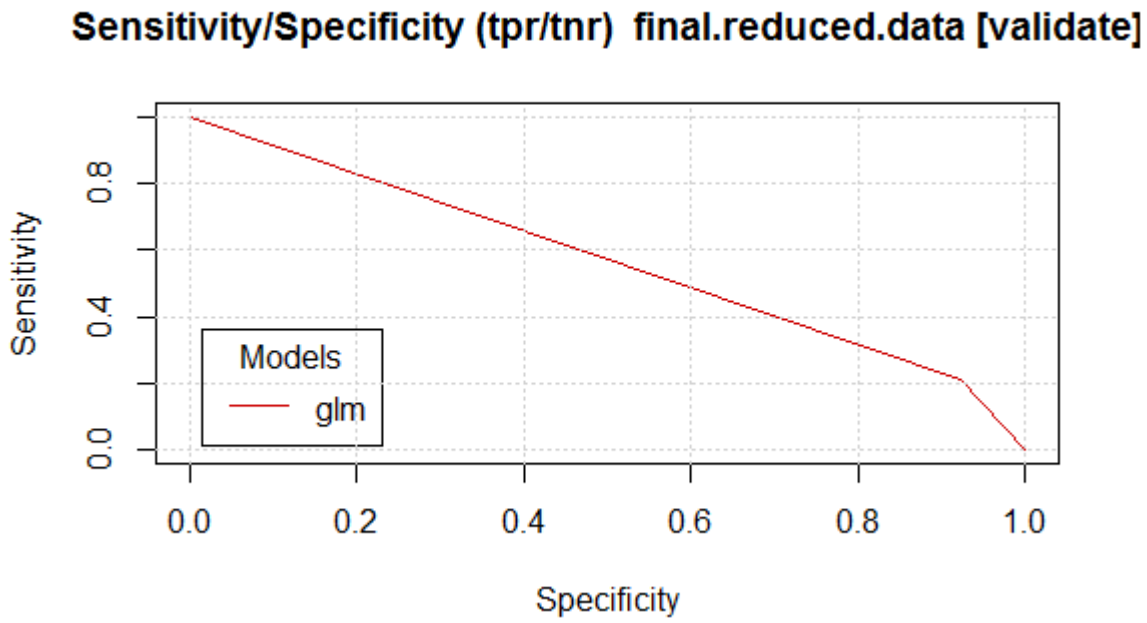
	Predicted		
Actual	0	1	Error
0	0.88	0.07	0.08
1	0.03	0.01	0.79

Overall error: 11%, Averaged class error: 44%
Accuracy= 89.3%
Sensitivity= 21%
Specificity=92.18%

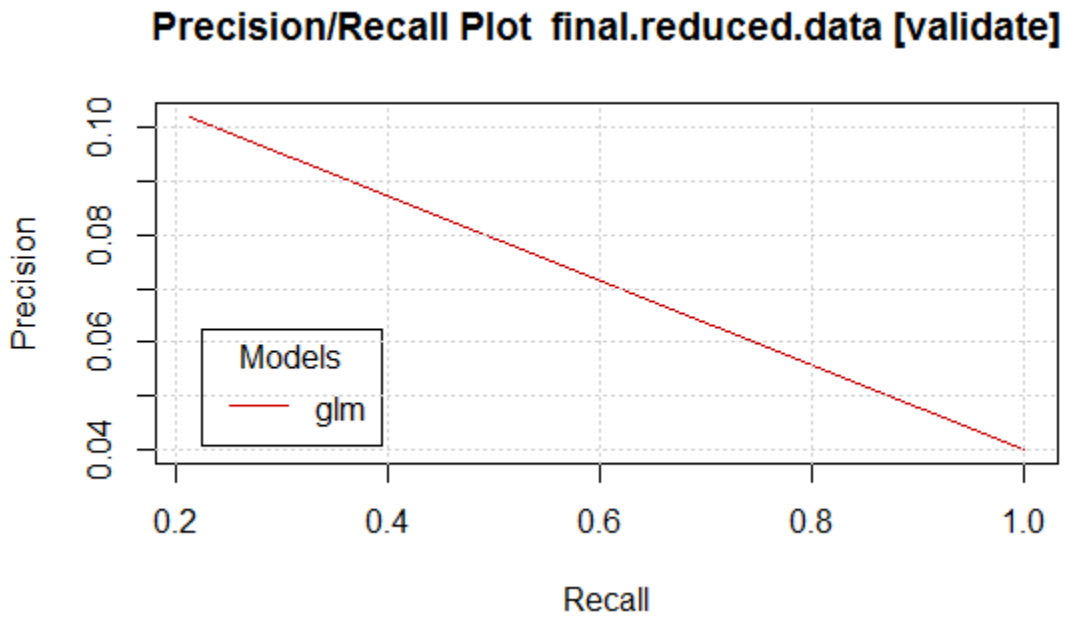




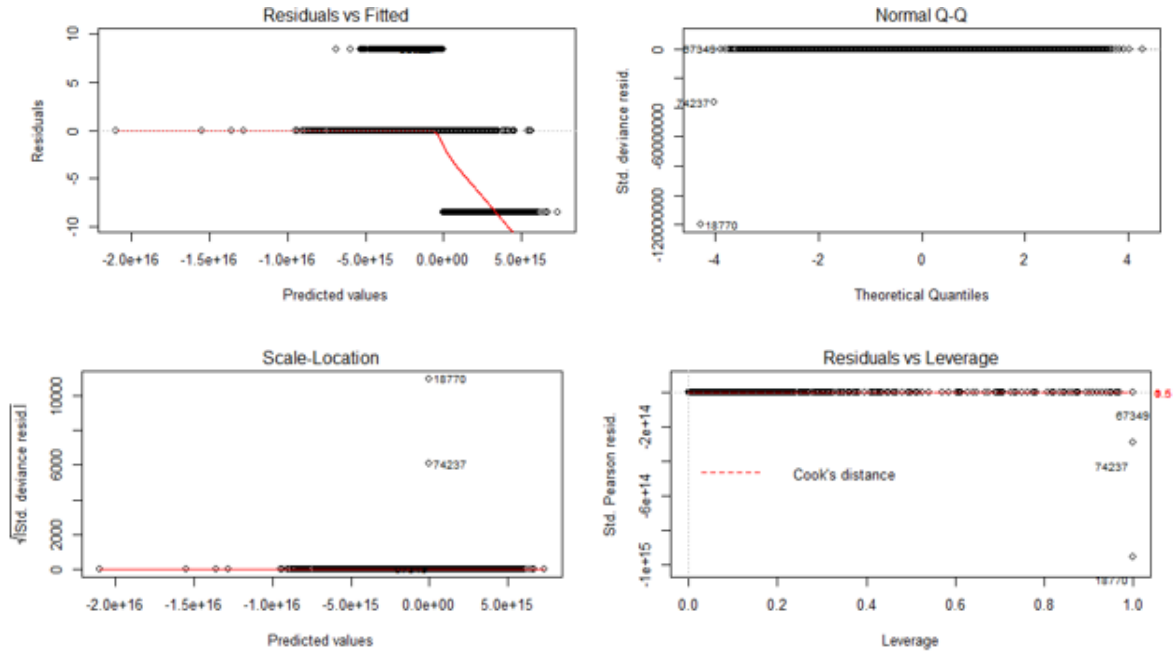
The area under curve is 0.57 and the ROC curve's hug is leaning towards True Positive Rate.



The true positive rate decreases as the true negative rate increases.



The precision decreases as the recall increases.



SUPPORT VECTOR MACHINE

Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)

parameter : cost C = 1

Gaussian Radial Basis kernel function.

Hyperparameter : sigma = 0.103848374897822

Number of Support Vectors : 16030

Objective Function Value : -3980.579

Training error : 0.034239

Probability model included.

	Predicted	
Actual	0	1
0	10946	0

1 456 1

Predicted

Actual 0 1 Error

0 0.96 0 0

1 0.04 0 1

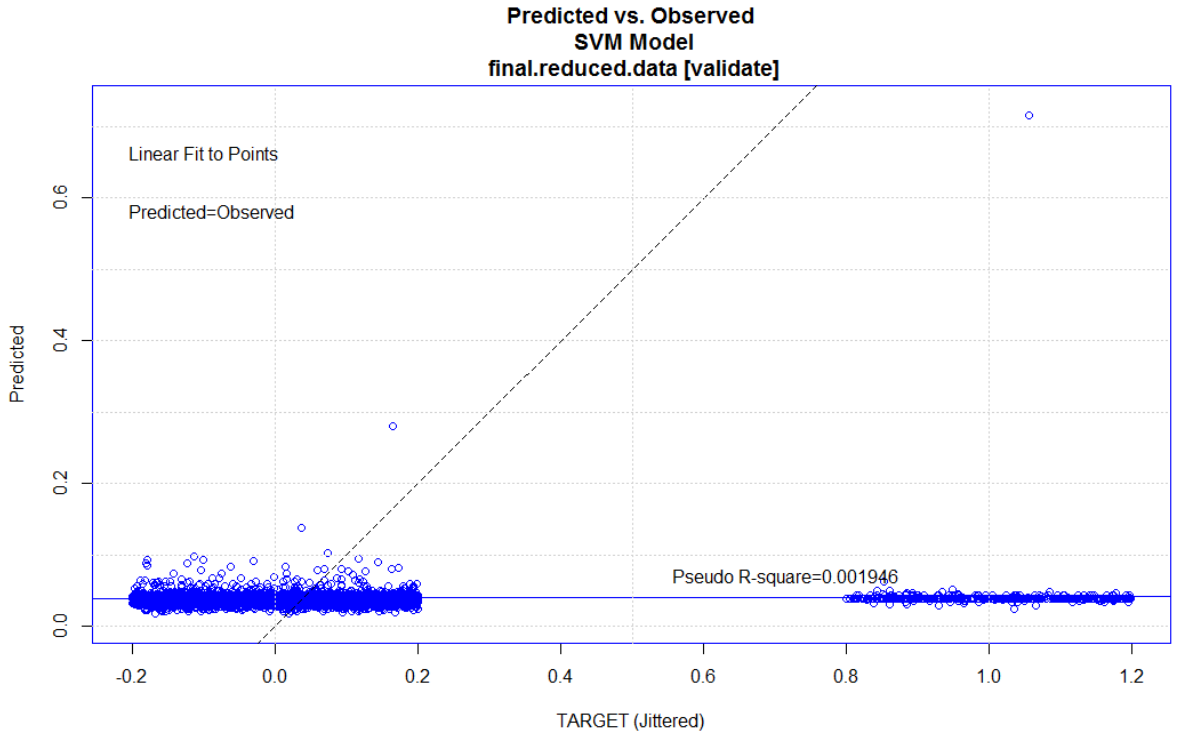
Overall error: 4%, Averaged class error: 50%

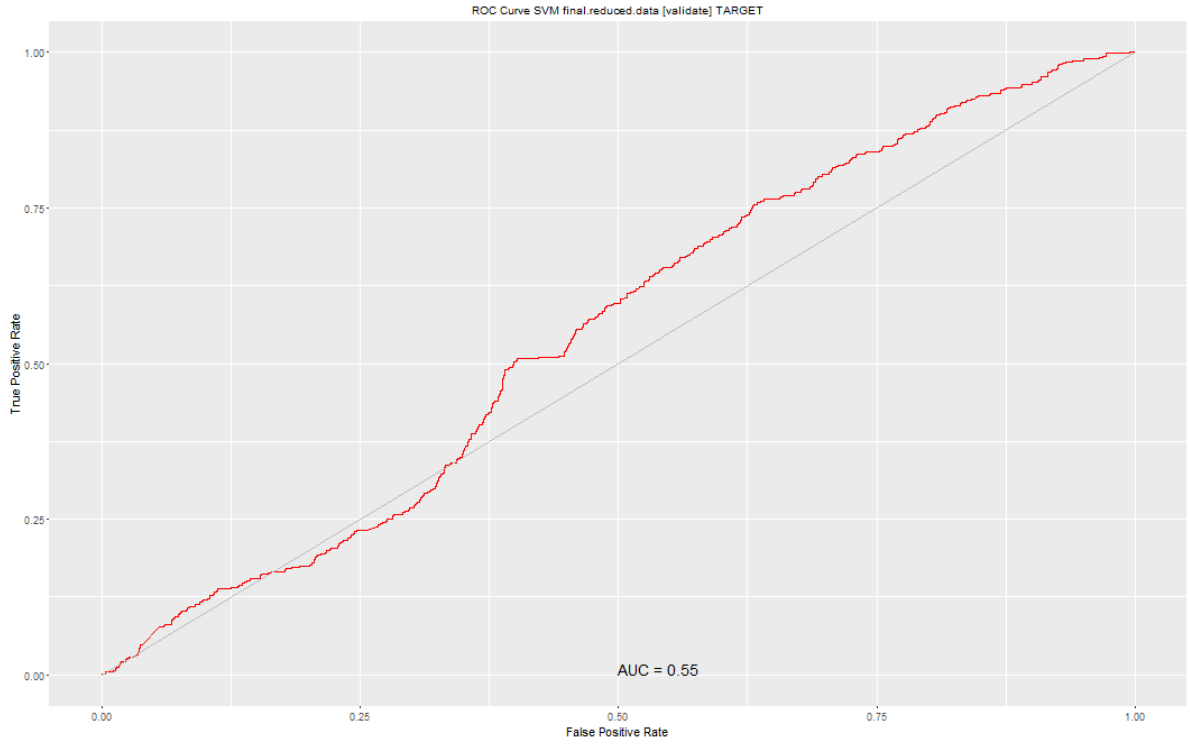
Accuracy=96.001%

Sensitivity= 0.21%

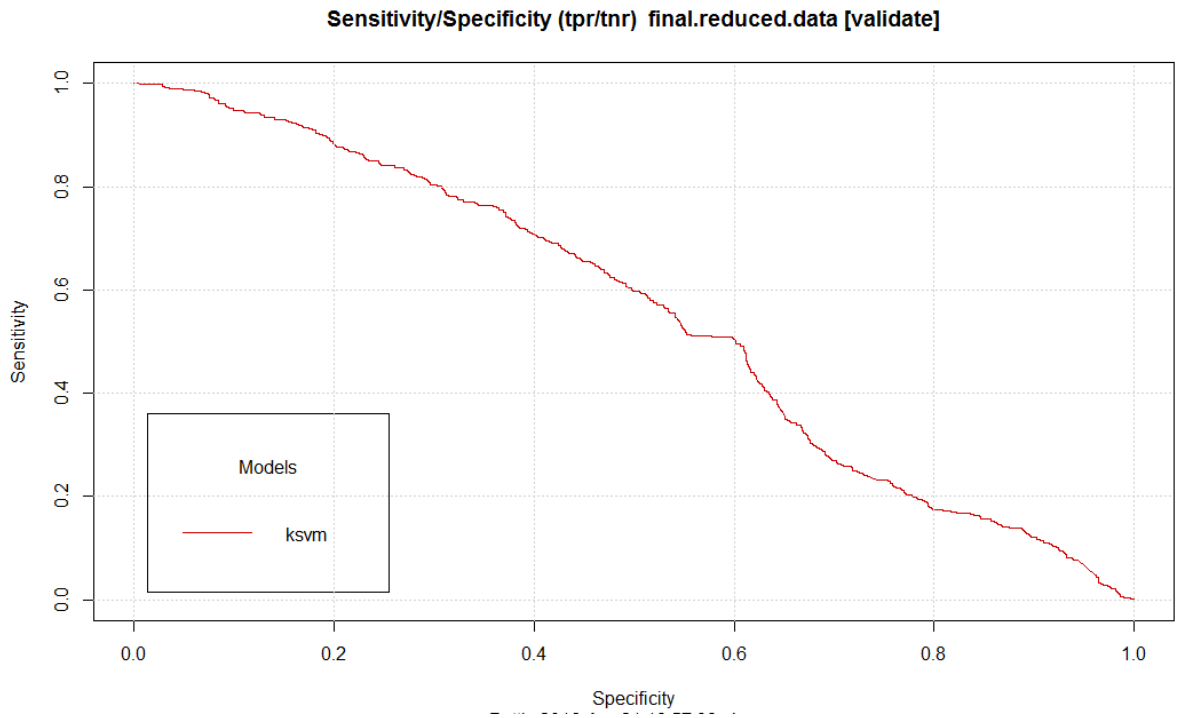
Specificity=100%

Area under the ROC curve for the ksvm model on final.reduced.data [validate] is 0.5520

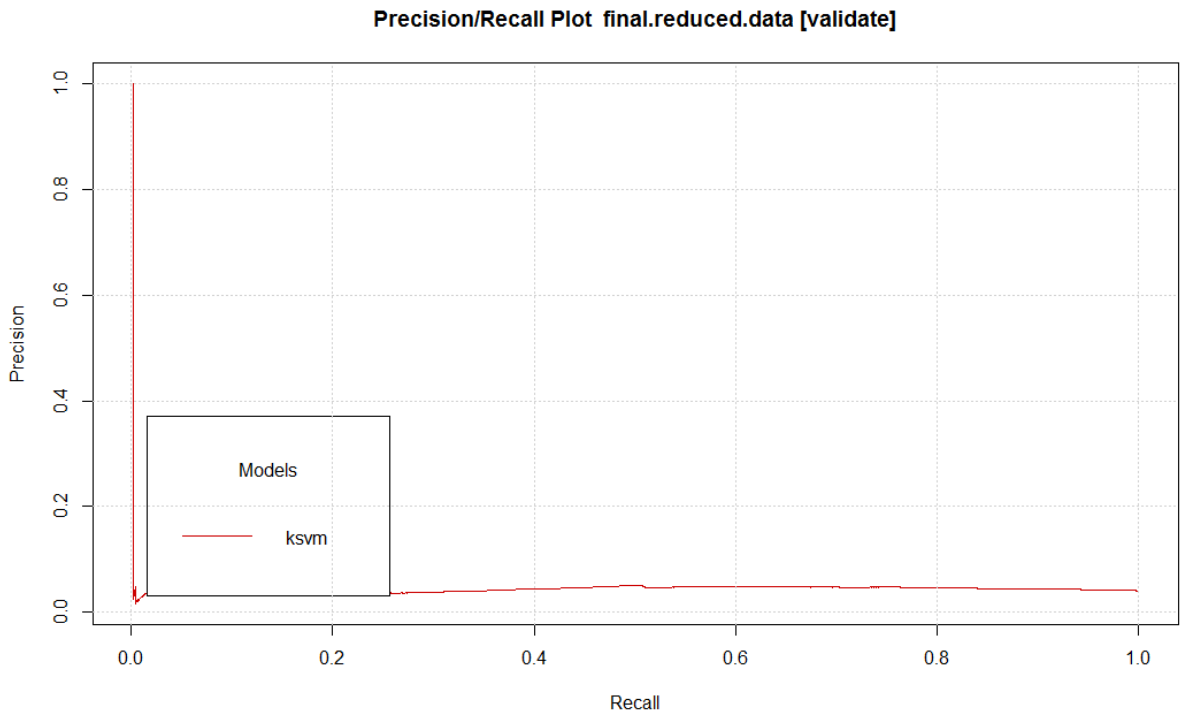




The ROC curve is not smooth and at around 39.5 at False Positive Rate the ROC curve leans towards the left True Positive Rate.



The overall true positive rate decreases as the true negative rate increases but the ksvm line is not smooth indicating that for certain true negative rate's increase was increasing for the increase in true positive rate.



Recall value at 0.0 has a very high precision value and as the Recall value increases the Precision value is almost constant between 0.0 to 0.01 Precision value.

NEURAL NETWORKS

Summary of the Neural Net model (built using nnet):

A 301-5-1 network with 1817 weights.

Output: as.factor(TARGET).

Sum of Squares Residuals: 3845.0000.

Error matrix for the Neural Net model on final.reduced.data [validate] (counts):

	Predicted	
Actual	0	1
0	10556	390
1	452	5

Error matrix for the Neural Net model on final.reduced.data [validate] (proportions):

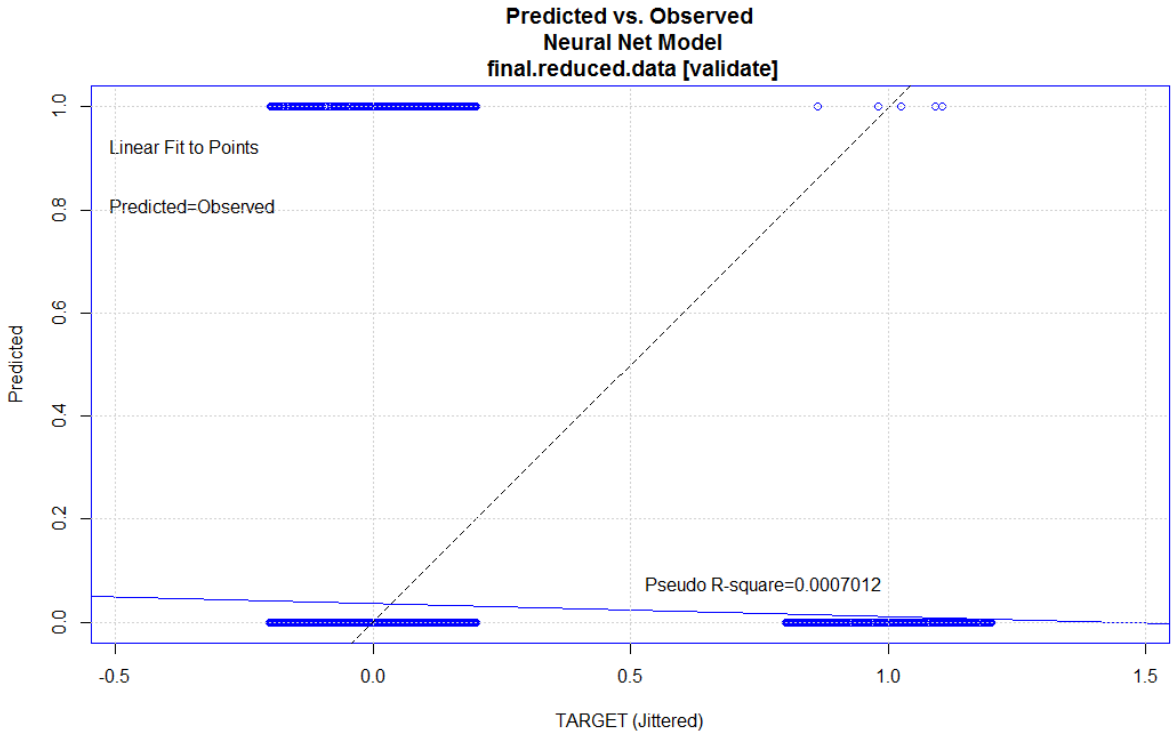
Actual	Predicted 0	1	Error
0	0.93	0.03	0.04
1	0.04	0.00	0.99

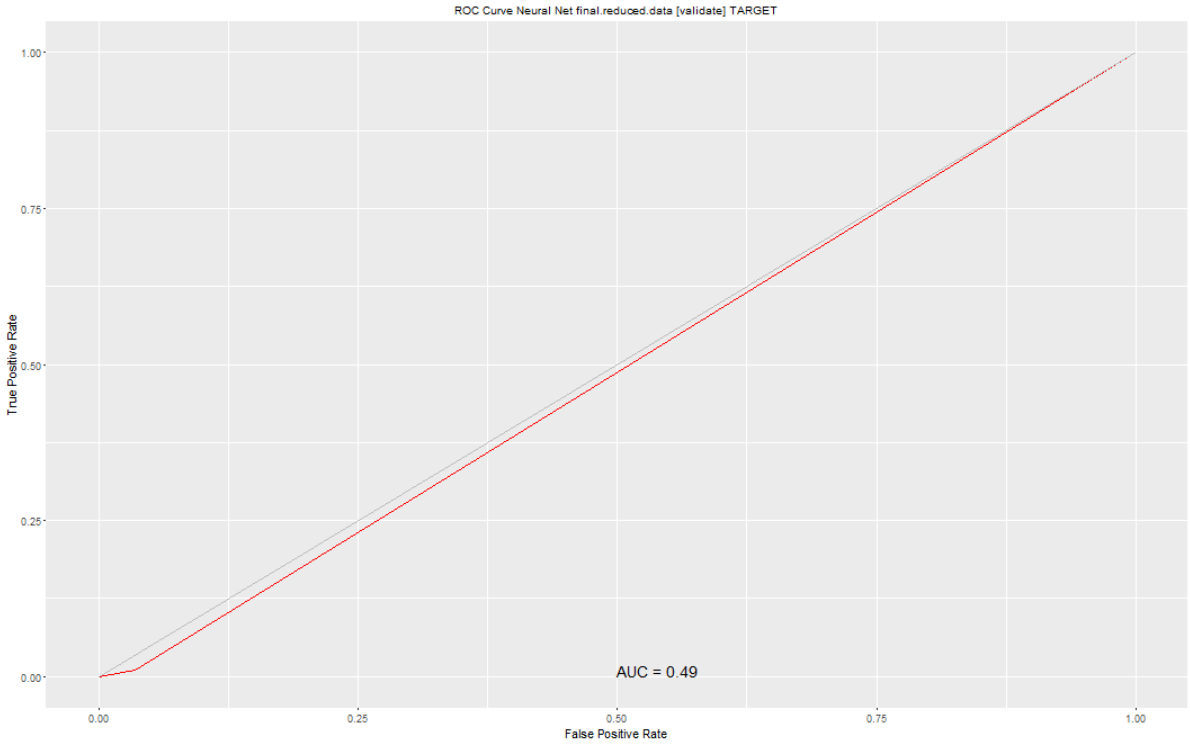
Overall error: 7%, Averaged class error: 52%

Accuracy=92.691%

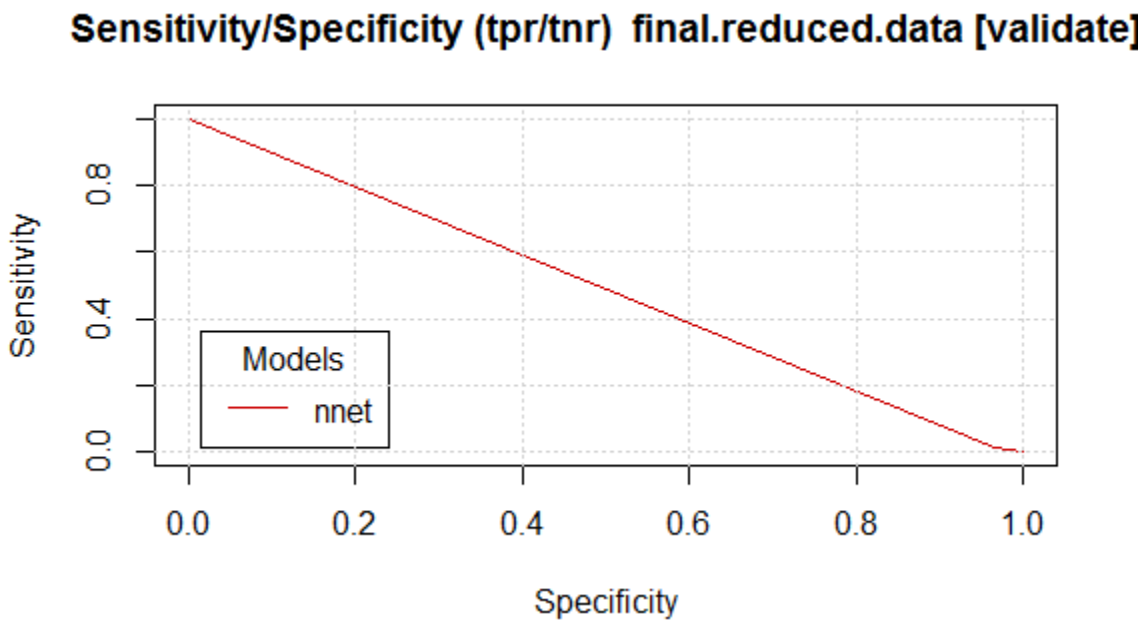
Sensitivity= 1.09%
Specificity=96.43%

Area under the ROC curve for the nnet model on final.reduced.data [validate] is 0.4877

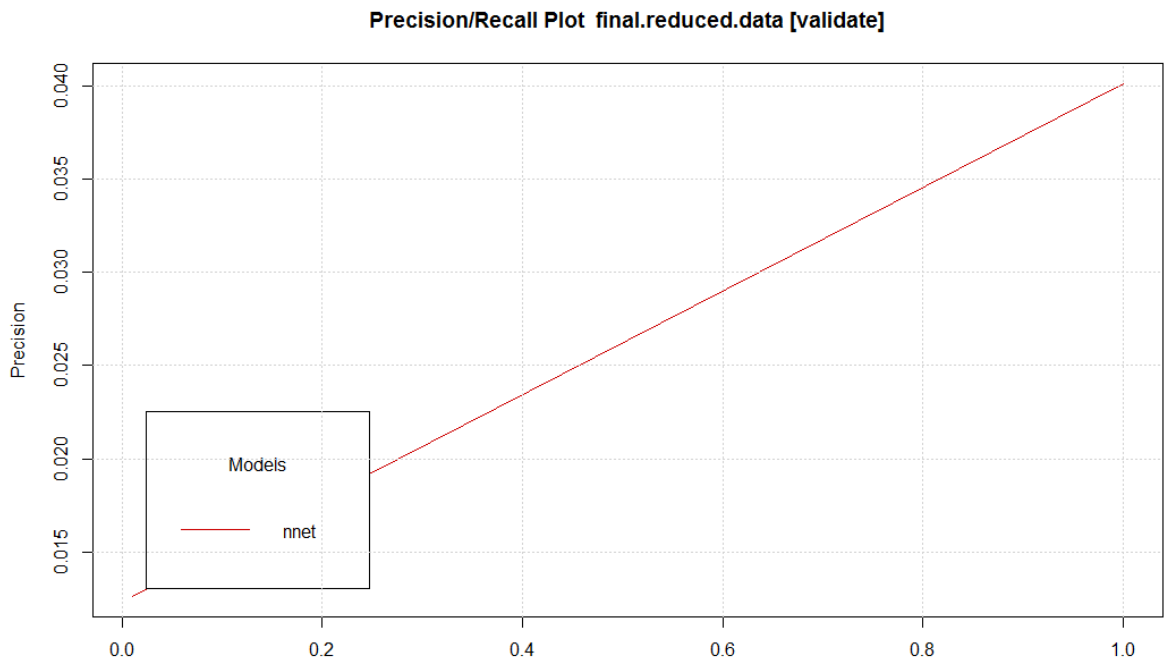




The ROC curve is negatively skewed leaning towards the False Positive Rate and hence is a very poor value of Area Under the Curve.



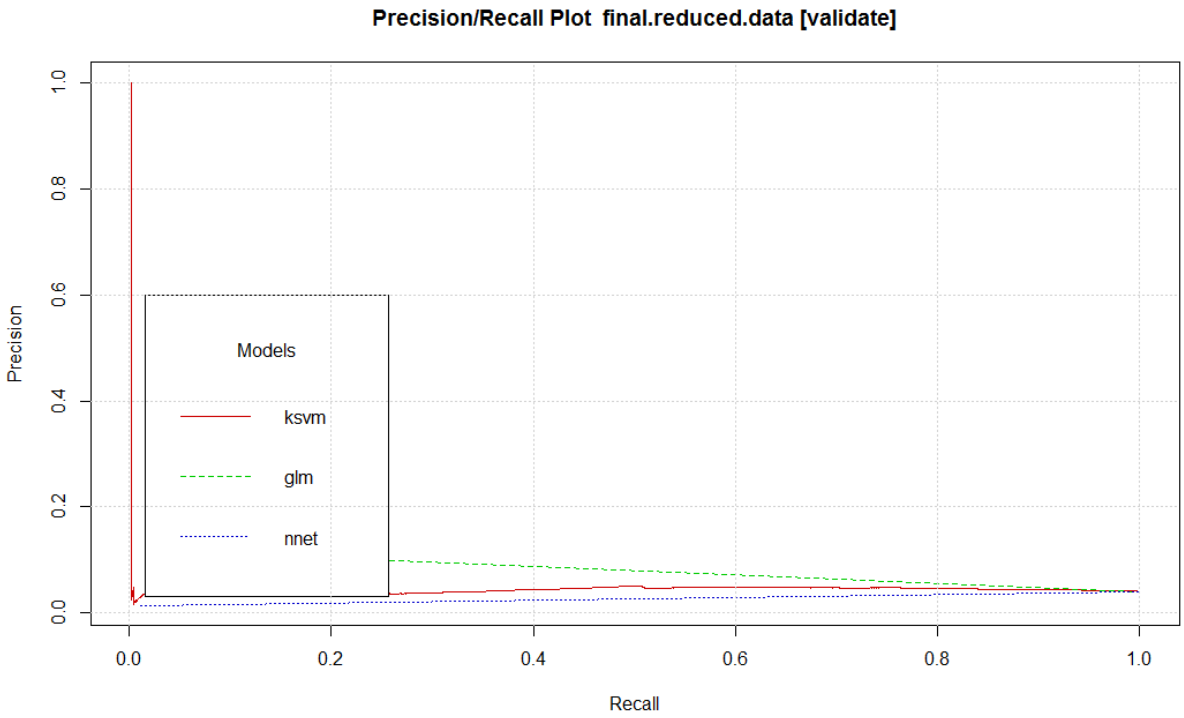
There is a smooth decline in the neural network model as True Negative Rate increases the Tre Positive Rate decreses.



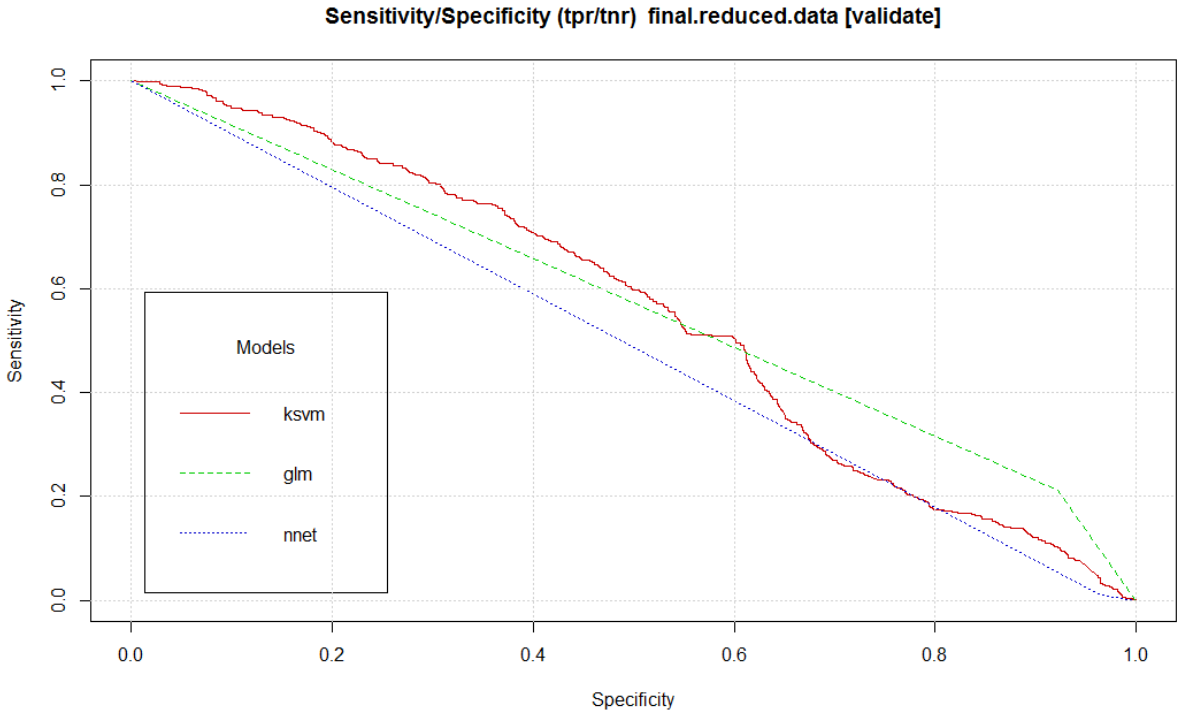
The precision and True positive rate increases at the same pace.

ASSESING MODEL ACCURACY

Model	Accuracy	Sensitivity	Specificity	Time	Area under ROC curve	Pseudo R-Square
LOGISTIC REGRESSION	89.3%	21%	92.18%	9.27 Hrs	0.57	0.00904
SUPPORT VECTOR MACHINE (Kernel= Radial)	96.001%	0.21%	100%	1.2 Hrs	0.5520	0.00194
NEURAL NETWORK (Hidden Layers=5)	92.691%	1.09%	69.43%	56 Mins	0.4877	0.00070



As there is an increase in True Positive Rate, Logistic Regression's precision value decreases at a much slower rate than SVM and Neural Networks. Although at 0.0 Recall rate SVM provides very high precision value.



As the True Negative Rate increases the True Positive Rate decreases for all three models but it is very smooth for Neural Network making it worse than Logistic Regression and

SVM. SVM has better Specificity value than Logistic regression which has better Sensitivity value.

Area Under Curve for ROC is best for Logistic Regression since the curve for ROC is hugging the top left of graph towards the True Positive Rate more than SVM or Neural Networks.

For choosing the best model below are the categories it should satisfy:

1. Accuracy should be high.
2. Sensitivity should be high.
3. Specificity should be high.
4. R squared value should be high.
5. Area under curve should be >0.5, 0.9 is excellent.

Since there is no straight answer that satisfies all of the above perfectly, ranking their preferences based on the dataset knowledge was done.

- Since the dataset's response variable is a binary outcome, hence ROC is the best parameter as it is a graphical plot that illustrates the performance of a binary classifier system. ROC's AUC indicates a high true positive rate and low false rate if the ROC curve is skewed to the left corner in the graph. Moreover Pseudo R squared can't classify binary variables since it is designed to predict probabilities which include values that are less than zero and greater than one too.
- Accuracy falls in the second option since the dataset is heavily unbalanced and heavily biased, classifying all the samples with columns having 95% of the same values over 5% different values.
- Sensitivity and specificity are both important since both classify true and false positive rate respectively.
- Hence preferences would be ROC's AUC > Accuracy > Sensitivity and Specificity > Pseudo R Squared> Time *for this dataset.*

  Logistic Regression => AUC =0.57 | Accuracy =89.3%
  SVM => AUC=0.55 | Accuracy =96%

CONCLUSION

Considering the factors listed above in the model assessing sequence for this dataset, Logistic Regression and SVM's performance is very similar and has only 0.002 difference in their AUC curve. But picking one would be *Logistic Regression, a better*

classifier than SVM based on ROC's AUC value although the accuracy is higher for SVM on this unbalanced & biased dataset.

To make the prediction better, in future, normalization or scaling up the predictors can be helpful. Algorithms like Random Forest, Decision Tree etc can be applied.

APPENDIX

Code Link & Box Plot Links:
https://github.com/pallavikaran/DataScienceSantanderKaggle

REFERNCES

- Kaggle Dataset Source: https://www.kaggle.com/c/santander-customer-satisfaction/data