Boston University

Final Report:

Aaron Taylor & Vineet Kowti

CS 699 Data Mining

Dr. Lee

4/2/2025

In this team project, we focused on the preprocessing and cleaning of a dataset to prepare it for machine learning analysis. The first step involved loading the necessary libraries and reading in the dataset (project_data.csv). Once the data was loaded, we examined its structure and dimensions to understand the variables and their types. Missing data was handled using different imputation strategies based on the type of variable. For example, blank spaces were converted into NA, and columns with a large number of missing values were removed entirely. Missing numerical values were imputed with the median, while categorical variables were filled using the mode. Additionally, some columns, like the 'Only Child' (OC) column, were reclassified to replace NA values with default values.

Feature selection played a critical role in the preprocessing pipeline. We removed columns with near-zero variance, as they would not contribute meaningful information to the model. Additionally, we identified and removed columns with high correlations (greater than 0.8), as these could introduce multicollinearity issues in the model. After cleaning the data, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, helping to reveal the underlying structure of the data. The explained variance of each principal component was visualized to assess how much information each component captured.

Outlier detection and handling were also important steps in our preprocessing. Using the interquartile range (IQR), we identified outliers and capped extreme values by replacing them with the 5th and 95th percentile values, a technique known as Winsorization. This helped to prevent outliers from distorting the model's performance. After completing these preprocessing steps, the cleaned dataset was saved to a new CSV file, ready for further modeling or analysis.

Our R code focuses on data preparation, feature selection, and modeling for a classification task, using various techniques such as stratified data splitting, class balancing, feature importance analysis, and machine learning models. Initially, the code splits the dataset df_clean into training and testing sets with an 80/20 split using stratified sampling, ensuring that the distribution of the target variable Class is preserved. The dataset is then visualized to show the class distribution before applying class balancing techniques using the ROSE package, including oversampling, undersampling, and a combination of both methods to address class imbalances. The training set is further processed by removing constant features using a custom function, which checks for columns with identical values across all rows, followed by subsetting the dataset to remove these uninformative columns.

Next, the code performs feature selection using Random Forest to determine which features are most important for predicting the target variable Class. Features with an importance score above a specified threshold are retained for further modeling, and the unimportant features are excluded. This feature selection process is done separately for the oversampled and combined (over- and undersampled) datasets. The selected features are then used to train Naive Bayes models, where predictions are made on the test set, and various performance metrics such as

True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and F1-score are computed to evaluate model performance. Additionally, the code includes ROC curve analysis to assess the model's ability to distinguish between classes, and the Area Under the Curve (AUC) is calculated for further performance evaluation.

Furthermore, weighted Random Forest models are trained on both the oversampled and combined datasets, where the sample sizes for each class are adjusted to ensure balance. The models are evaluated using similar metrics as the Naive Bayes models, with additional focus on metrics like Precision, F1-score, and Matthew's Correlation Coefficient (MCC). Lastly, the code includes a commented-out section where Boruta, a feature selection algorithm, is used to identify important features based on their relevance to the classification task. The various steps in the code highlight a comprehensive approach to tackling class imbalance and optimizing model performance through feature selection and evaluation.

This R code integrates various feature selection methods, including Random Forest, Near Zero Variance (NZV) and Correlation-based methods, to enhance the effectiveness of machine learning models such as Naive Bayes, Random Forest, and Neural Networks. The Near Zero Variance method is applied to filter out features that exhibit very little variation across the dataset, which often do not provide much predictive power. These features, typically constant or with minimal variance, are removed to streamline the modeling process. This feature selection technique ensures that only the most informative and varied attributes are considered, improving model efficiency and reducing the risk of overfitting. The removal of these near-zero variance features is especially crucial in the context of high-dimensional datasets, as it reduces noise and focuses on the more relevant patterns in the data.

Additionally, correlation-based feature selection is employed to remove highly correlated features, which could lead to multicollinearity and negatively impact model performance. In cases where two or more features are highly correlated, only one of them is retained for modeling, ensuring that redundant information does not distort the model's ability to learn from distinct features. By eliminating correlated features, the model is more likely to focus on the true relationships within the data, ultimately leading to better generalization on unseen data. This approach is applied in parallel with the Naive Bayes, Random Forest, and Neural Network models, where only the most important and non-redundant features are used to train the models. For each of these models, the selection of relevant features enhances the quality of the predictions by reducing noise and overfitting risks.

For Naive Bayes, applying Near Zero Variance and Correlation-based feature selection helps simplify the model by keeping only the most important features, improving interpretability while ensuring that the assumptions of feature independence are better met. In the case of Random Forest, these feature selection techniques contribute to the model's ability to identify important variables while reducing unnecessary complexity. By focusing on a smaller set of highly relevant features, the Random Forest algorithm can build more robust decision trees, leading to improved predictive accuracy. In the Neural Network model, the removal of near-zero variance and correlated features not only improves training efficiency but also helps in avoiding the curse of dimensionality. This ensures that the neural network focuses on the most significant features, resulting in more accurate and efficient learning, particularly when dealing with large and complex datasets.

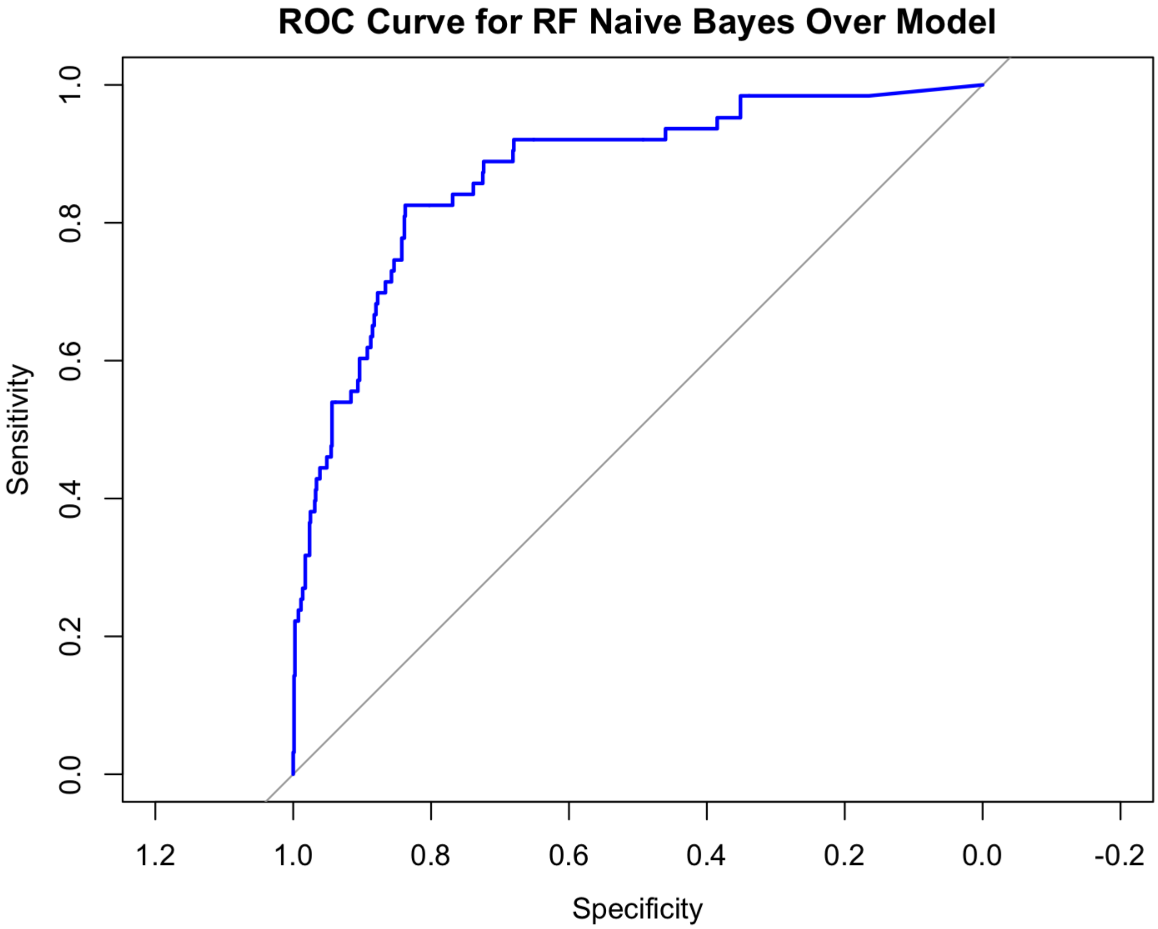
Through the use of these feature selection methods—Near Zero Variance and Correlation-based—across different models, the code aims to enhance the performance of Naive Bayes, Random Forest, and Neural Network classifiers. By reducing redundant or irrelevant information, each model is better equipped to make accurate predictions while maintaining simplicity and avoiding overfitting. The integration of these techniques with class balancing methods further strengthens the overall classification pipeline, enabling the models to perform optimally in real-world scenarios where class imbalances and high-dimensionality are common challenges.

After the intermediate report, we came together and both agreed to continue working together. We started meeting more frequently and tried to stick to deadlines. This project was very taxing so things kept getting pushed back but in the end we were able to come together to produce all 36 models. If we had more time we would pay more attention to deal when it comes to the output of the tables below. We would also work to make the script more robust and engineer more functions to do the heavy lifting. One more thing would be to create more tuning grids to further experiment with the dataset and strive for the extra-credit benchmark.

Model Results:
Best Model: (Near Zero Variance Neural Net trained on Both dataset)

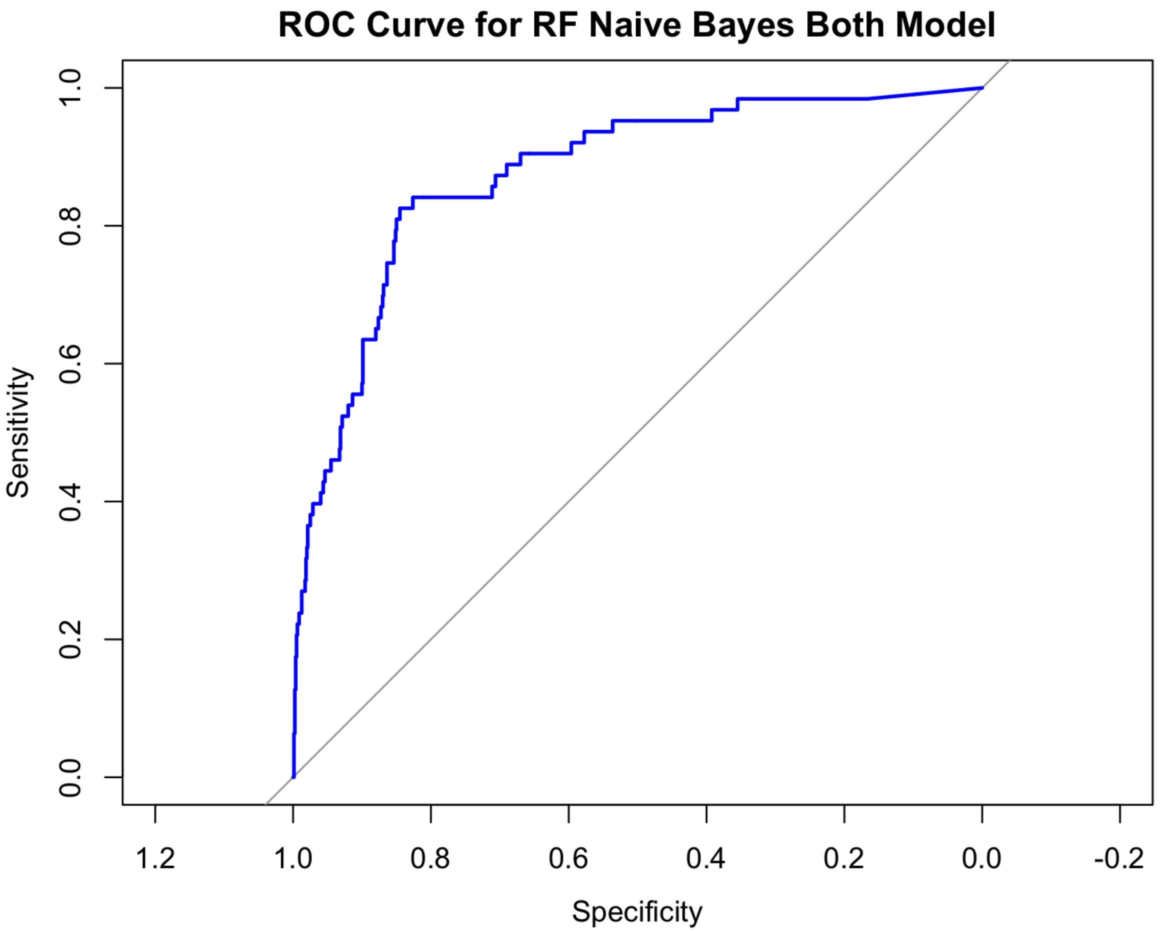
RF Naive Bayes Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.75	0.159	0.9836	0.75	0.907	0.8729	0.337	0.2475
Yes	0.841	0.25	0.2095	0.841	0.031	0.8729	-0.337	0.2475
Wt. Avg	.7955	.2045	.597	.796	.469	.8729	0	.2475



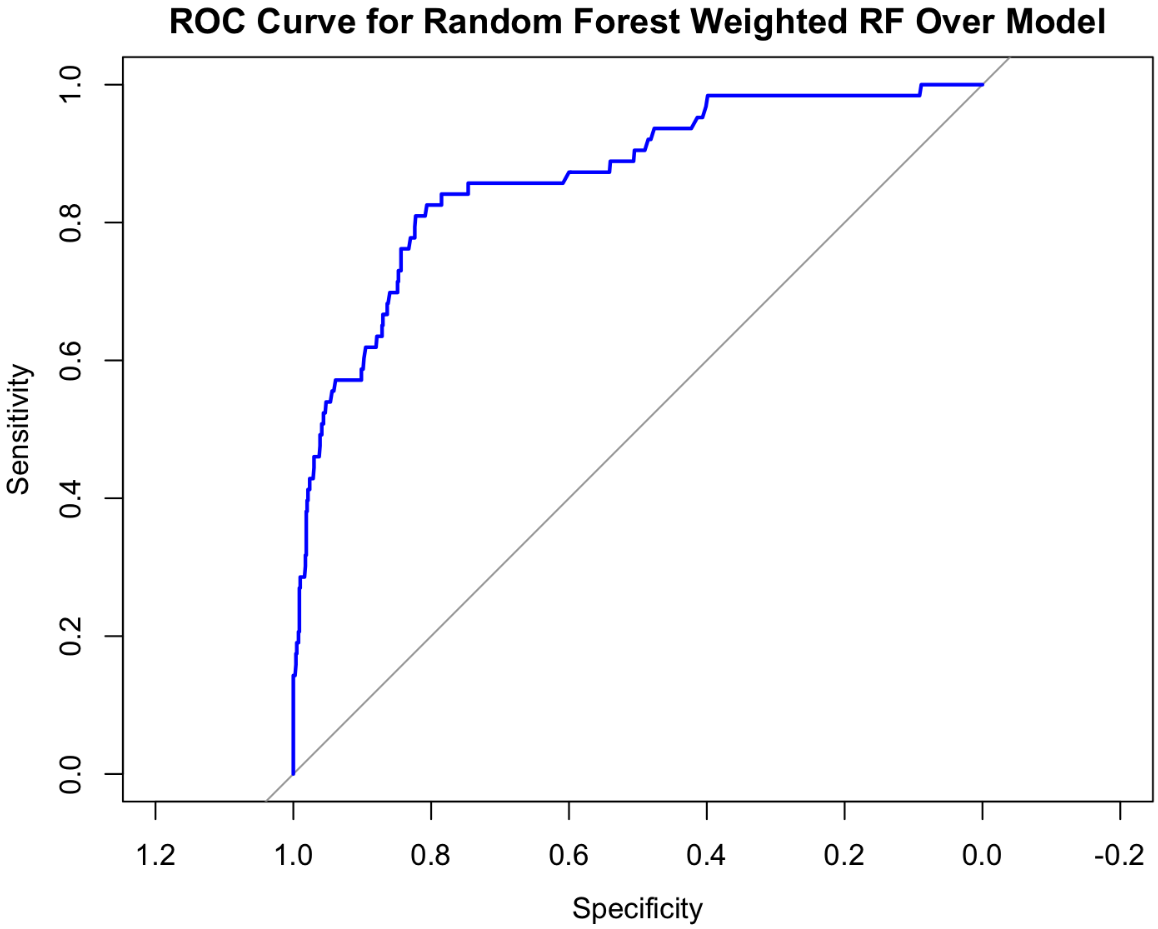
RF Naive Bayes Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7312	0.159	0.9836	0.7312	0.907	0.8748	0.321	0.2291
Yes	0.8413	0.27	0.017	0.8413	0.032	0.8748	-0.321	0.2291
Wt. Avg	.786	.2145	.5003	.786	.4695	0.8748	0	0.2291



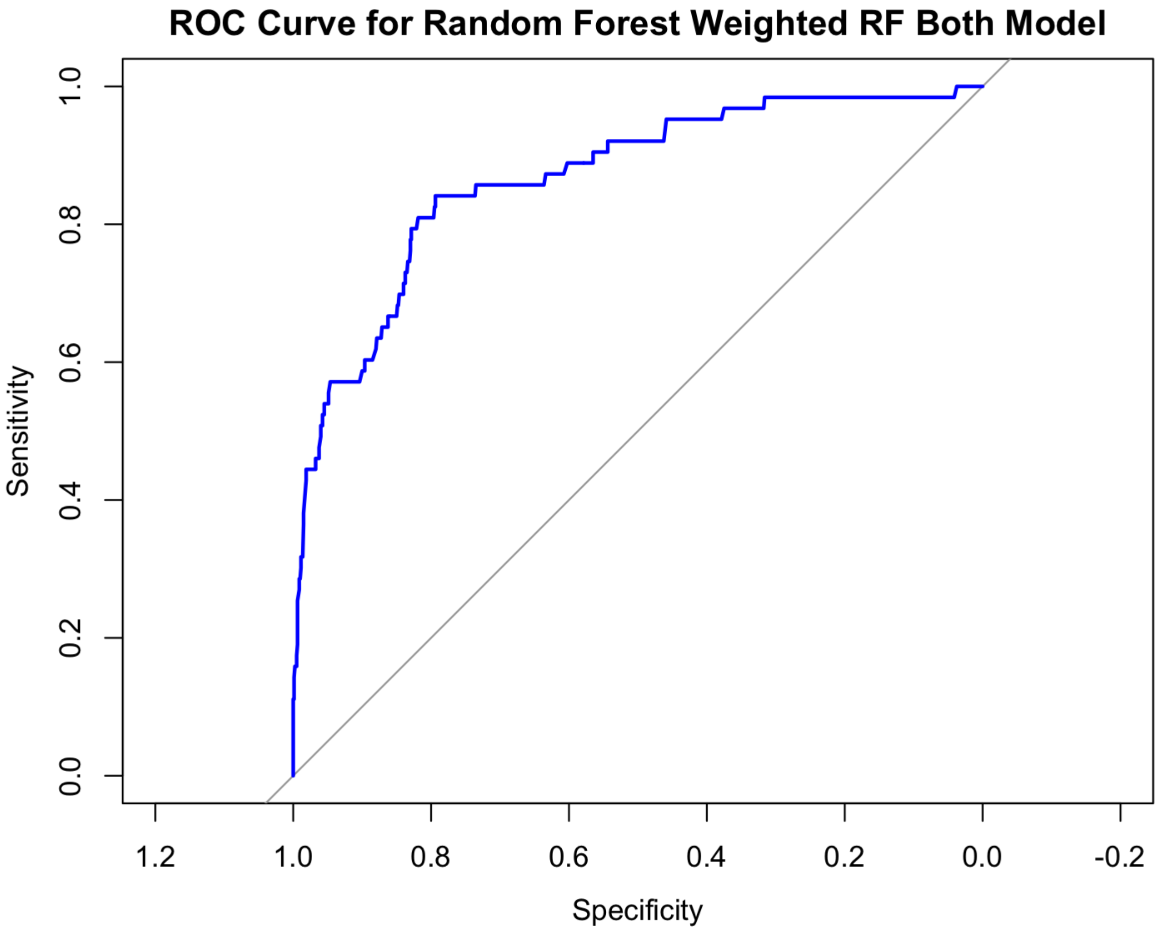
RF WRF Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.8150	0.190	0.9836	0.8150	0.887	0.8659	0.385	0.3131
Yes	0.8095	0.185	0.018	0.8095	0.0329	0.8659	-0.385	0.3131
Wt. Avg	.8	.188	.5003	.786	.4695	0.8659	0	0.3131



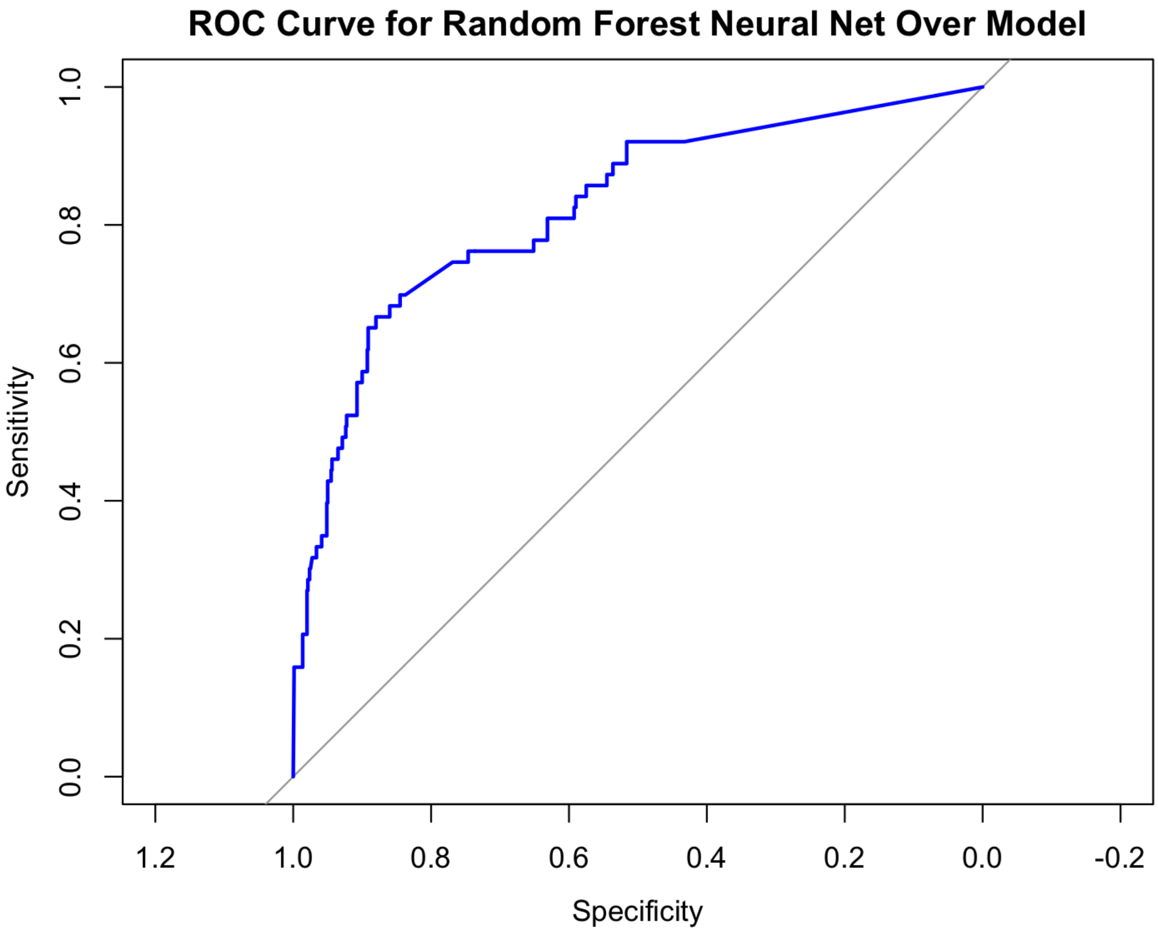
RF WRF Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7937	0.190	0.9830	0.7937	0.897	0.8666	0.385	0.2912
Yes	0.8254	0.185	0.2396	0.8254	0.0314	0.8666	-0.385	0.2912
Wt. Avg	.801	.1875	.6113	.801	.4642	0.8666	0	0.2912



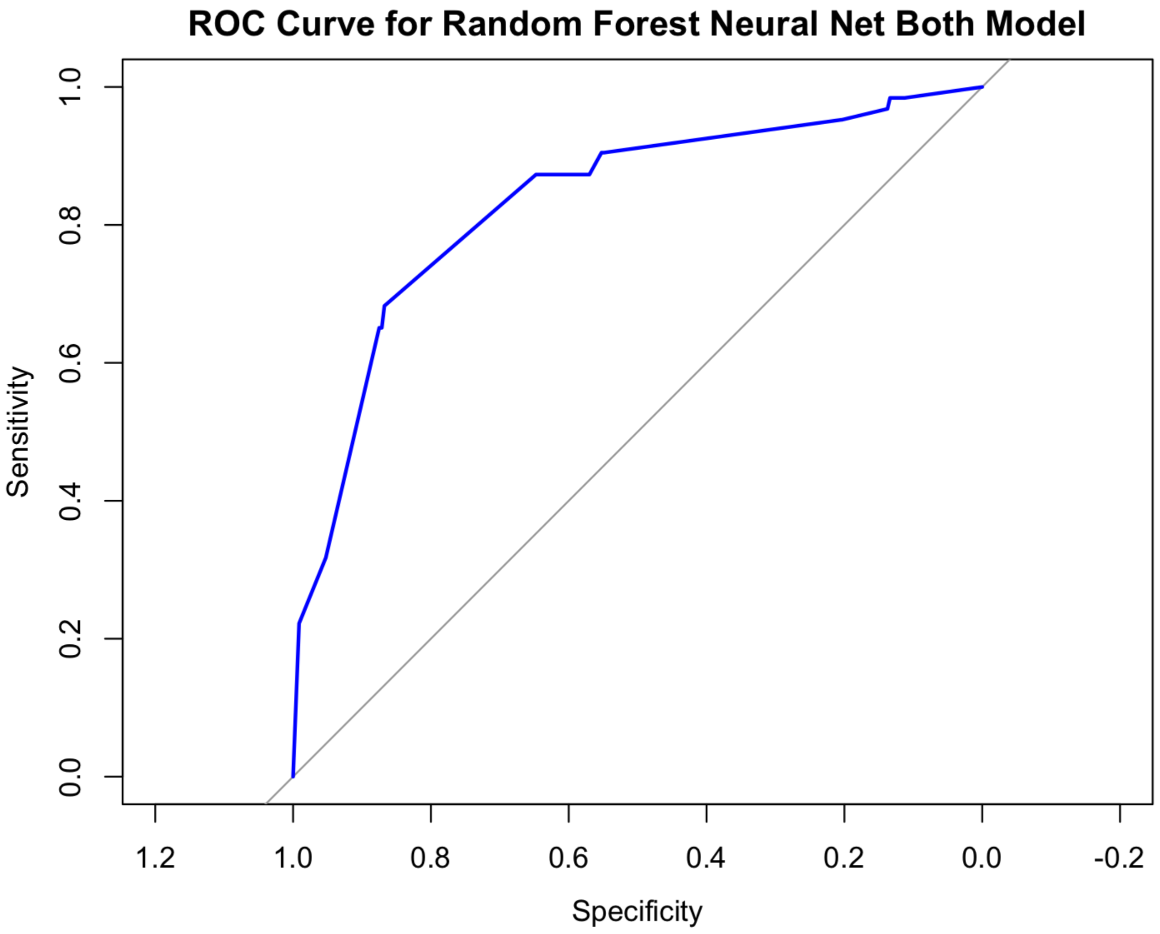
RF NN Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.9012	0.190	0.9639	0.9012	0.897	0.8234	0.3618	0.3425
Yes	0.5714	0.185	0.3130	0.5714	0.0314	0.8234	-0.3618	0.3425
Wt. Avg	0.9012	0.190	0.9639	0.9012	0.897	0.8234	0	0.3425



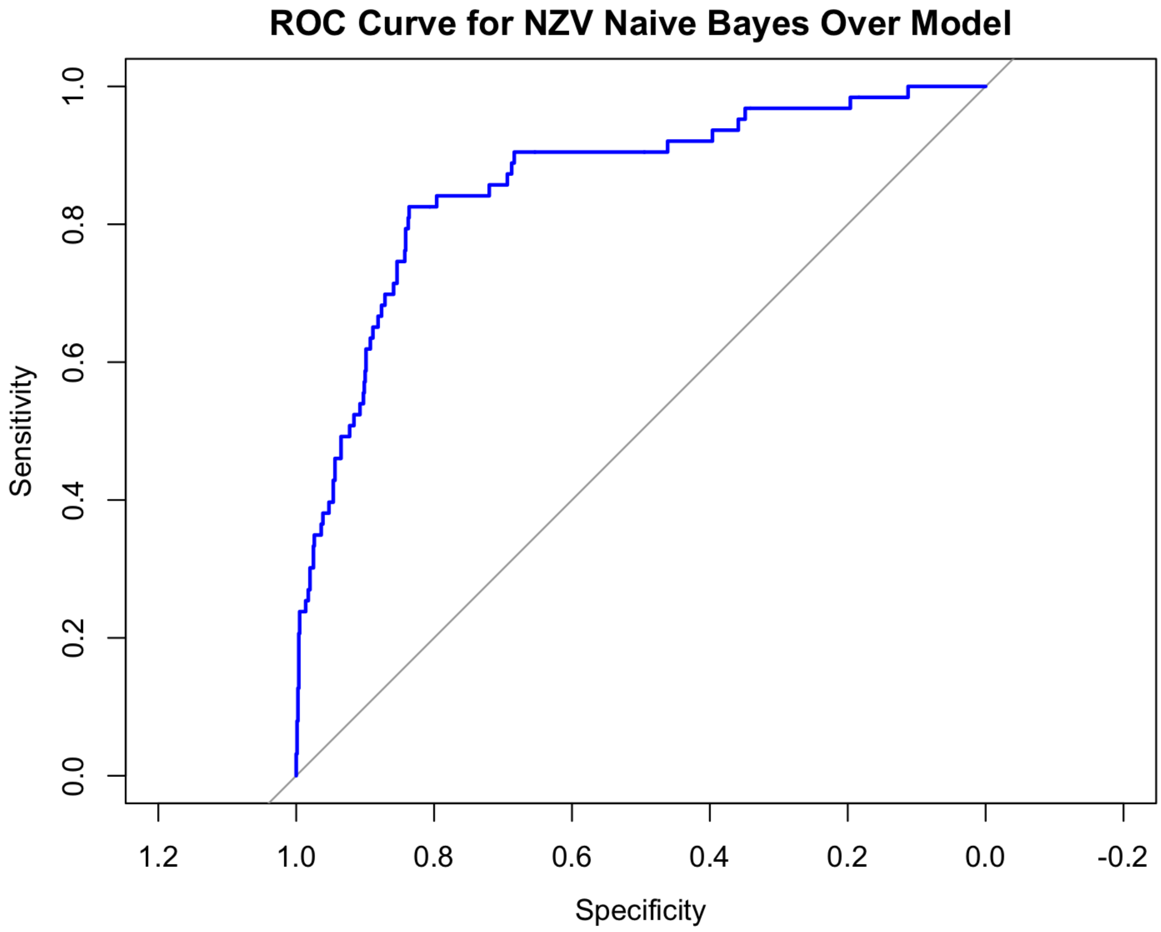
RF NN Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.8675	0.190	0.9720	0.8675	0.8019	0.8318	0.3618	0.3377
Yes	0.6825	0.185	0.2886	0.6825	0.0462	0.8318	-0.3618	0.3377
Wt. Avg	0.8675	0.190	0.9720	0.8675	0.8019	0.8318	0	0.3425



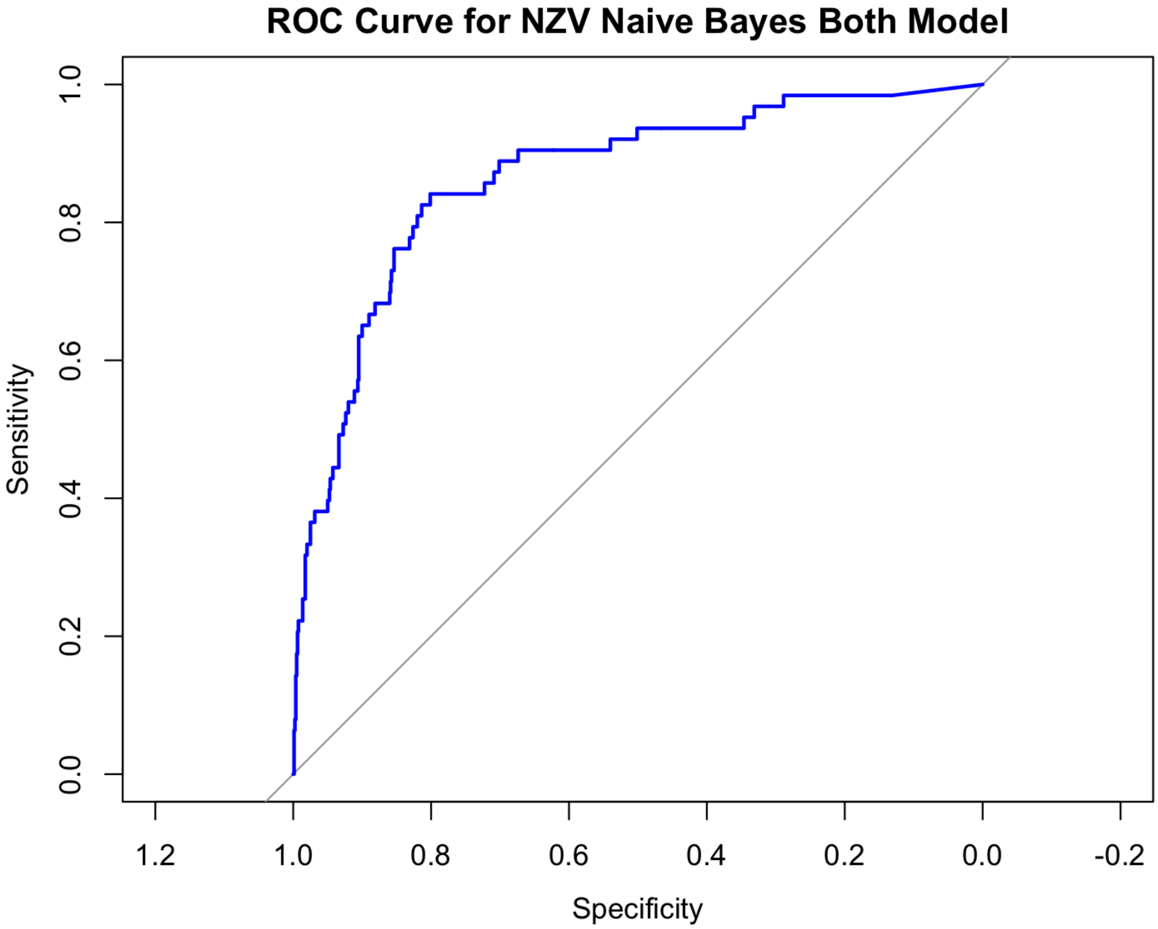
NZV Naive Bayes Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7788	0.190	0.9842	0.7788	0.907	0.8606	0.3648	0.2792
Yes	0.8413	0.185	0.2304	0.8413	0.029	0.8606	-0.3648	0.2792
Wt. Avg	0.7788	0.190	0.9842	0.7788	0.907	0.8606	0	0.2792



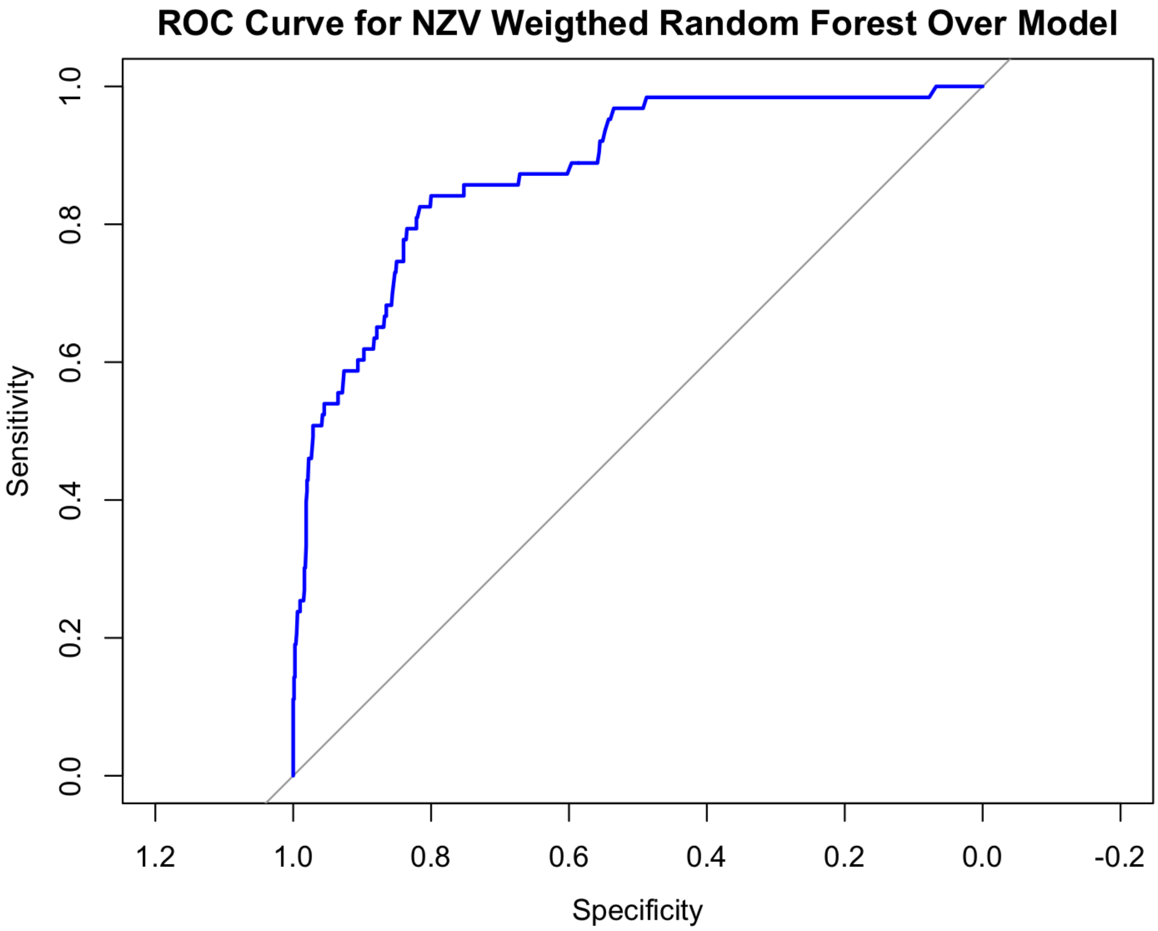
NZV Naive Bayes Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7588	0.190	0.9838	0.7588	0.907	0.8646	0.3457	0.2566
Yes	0.8413	0.185	0.2304	0.2154	0.03	0.8646	-0.3457	0.2566
Wt. Avg	0.7588	0.190	0.9838	0.7588	0.907	0.8646	0	0.2566



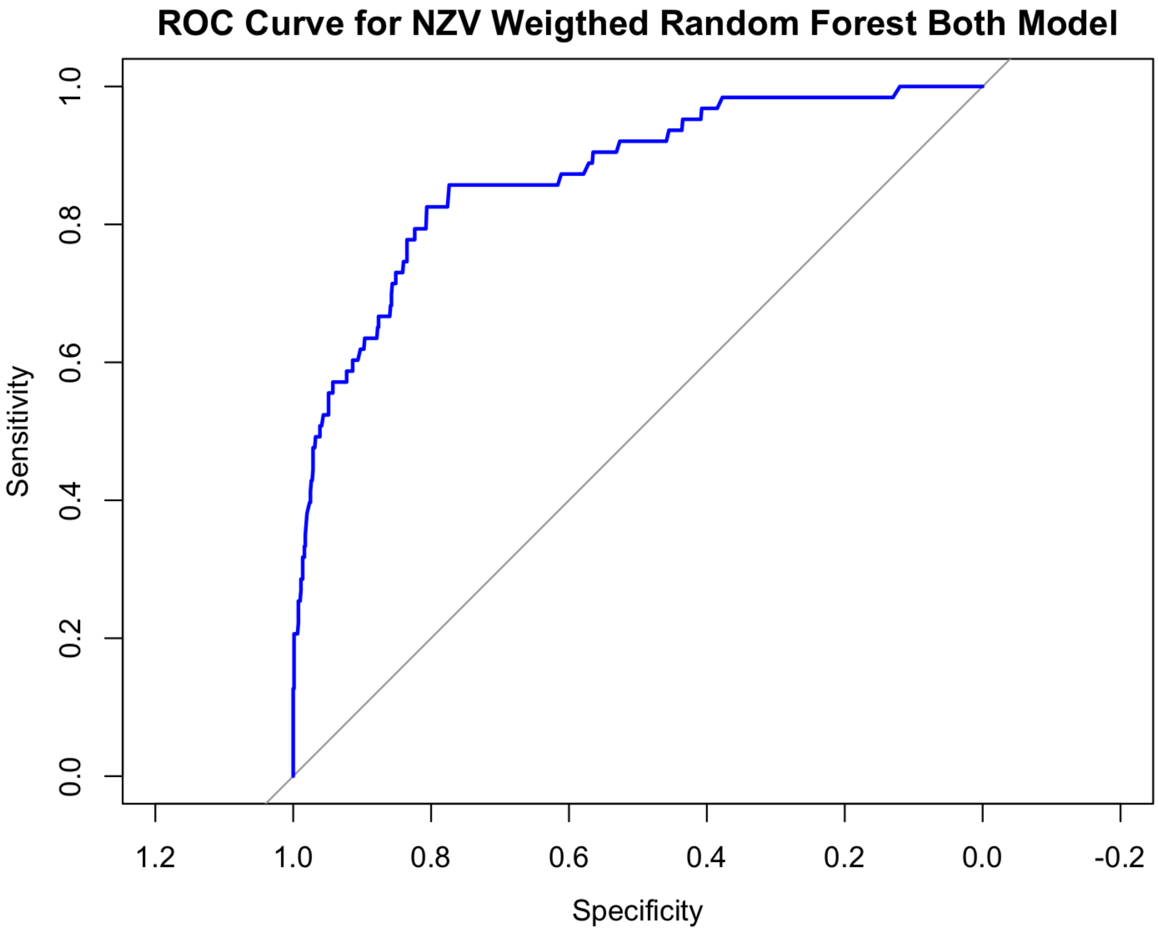
NZV Random Forest Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.8063	0.174	0.9832	0.8063	0.8974	0.8785	0.3848	0.3077
Yes	0.8254	0.193	0.2512	0.8254	0.0308	0.8785	0.3848	0.3077
Wt. Avg	0.8063	0.174	0.9832	0.8063	0.8974	0.8785	0.3848	0.3077



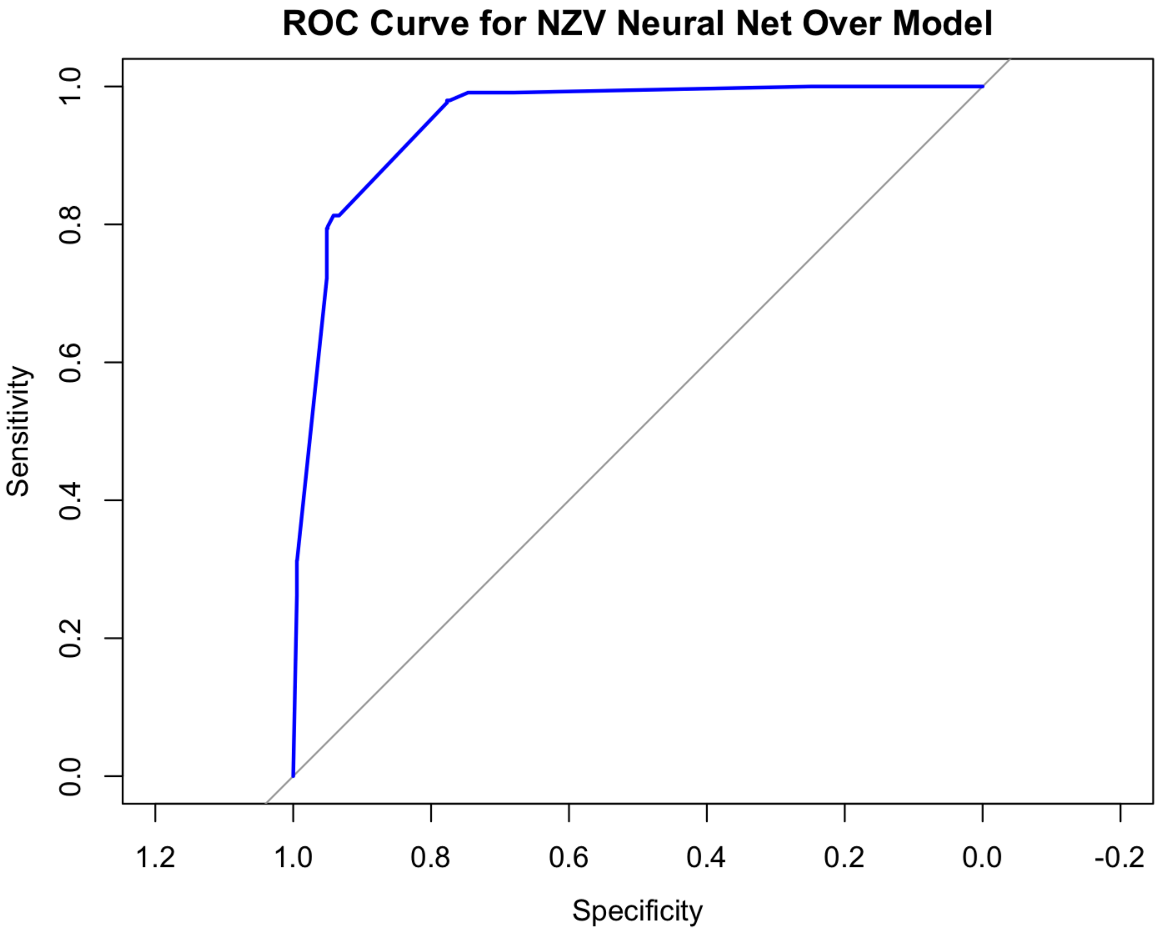
NZV Random Forest Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7887	0.175	0.9829	0.7887	0.897	0.8702	0.366	0.285
Yes	0.8254	0.211	0.2353	0.8254	0.0316	0.8702	0.366	0.285
Wt. Avg	0.7887	0.175	0.9829	0.7887	0.897	0.8702	0.366	0.285



NZV Neural Net Over Model:

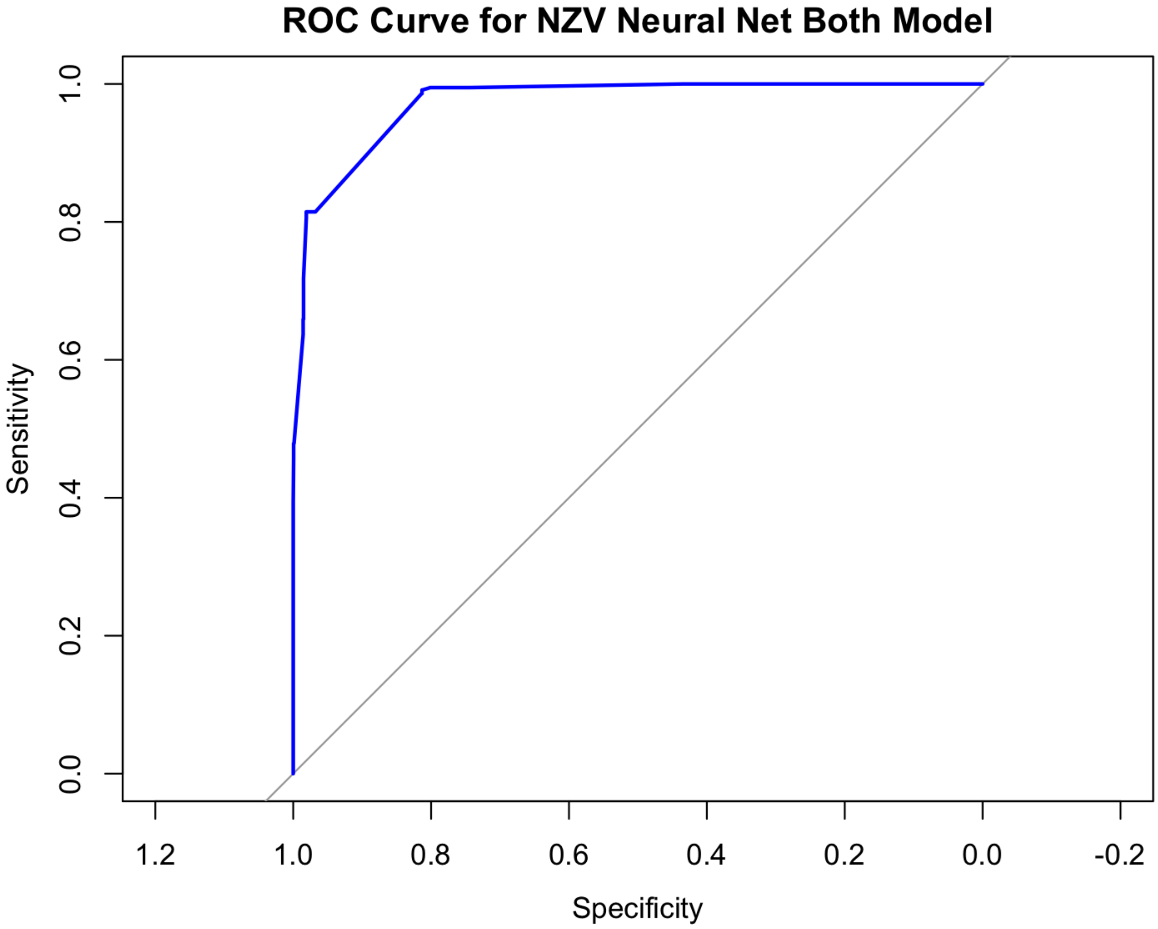
	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.9397	0.187	0.9829	0.7887	0.8241	0.9515	0.7591	0.3077
Yes	0.812	0.060	0.2353	0.8254	0.088	0.9515	0.7591	0.3077
Wt. Avg	0.9397	0.187	0.9829	0.7887	0.8241	0.9515	0.7591	0.3077

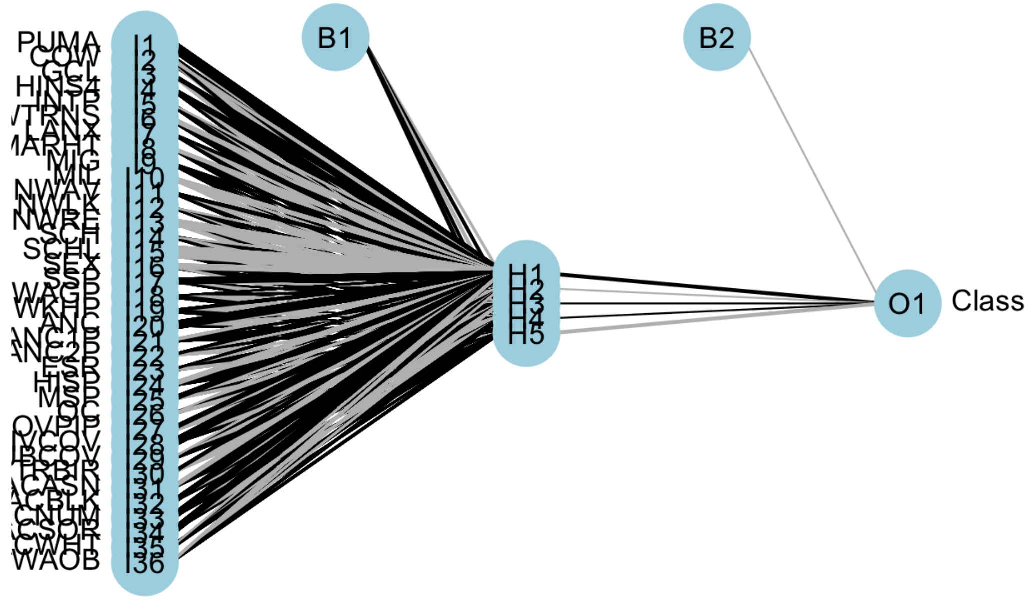


NZV Neural Net Both Model:

Best Model:

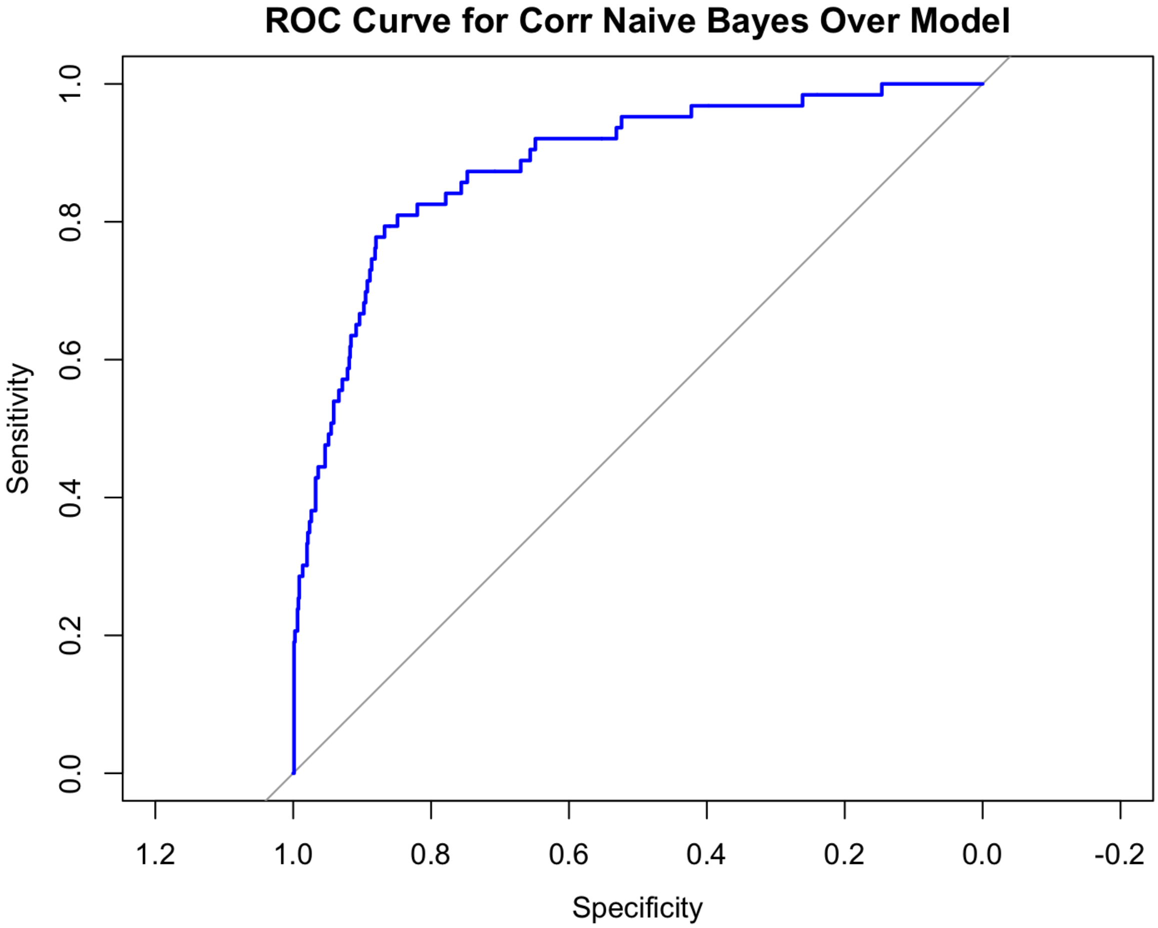
	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.8132	0.008	0.9895	0.8132	0.990	0.9734	0.8166	0.8034
Yes	0.9913	0.186	0.8395	0.9913	0.019	0.9734	0.8166	0.8034
Wt. Avg	.90225	.097	.9145	.90225	.5045	0.9734	0.8166	0.8034




PUMA
COW
GCL
HINS4
INTP
WTRNS
LANX
MARHT
MIG
MIL
NWAV
NWLK
NWRE
SCH
SCHL
SEX
SSP
WAGP
WKHP
ANC
ANC1P
ANC2P
ESR
HISP
MSP
OC
OVPIP
IVCOV
BCOV
TRBIR
ACASN
ACBLK
CNUM
CSOR
CWHT
WAOB
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
B1
B2
H1
H2
H3
H4
H5
O1
Class

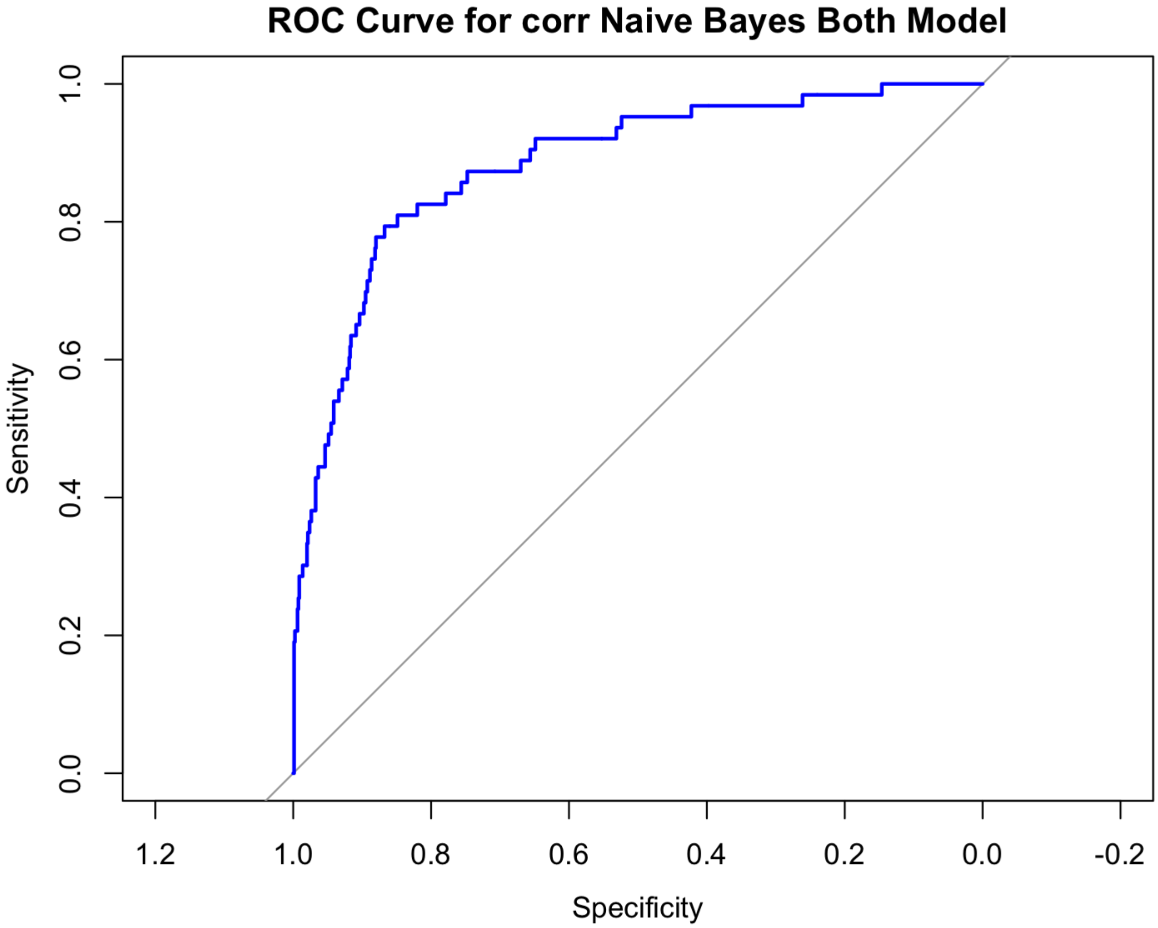
Correlation Naive Bayes Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.563	0.079	0.9890	0.563	0.9536	0.8828	0.2524	0.1378
Yes	0.9206	0.436	0.1425	0.9206	0.021	0.8828	0.2524	0.1378
Wt. Avg						0.8828	0.2524	0.1378



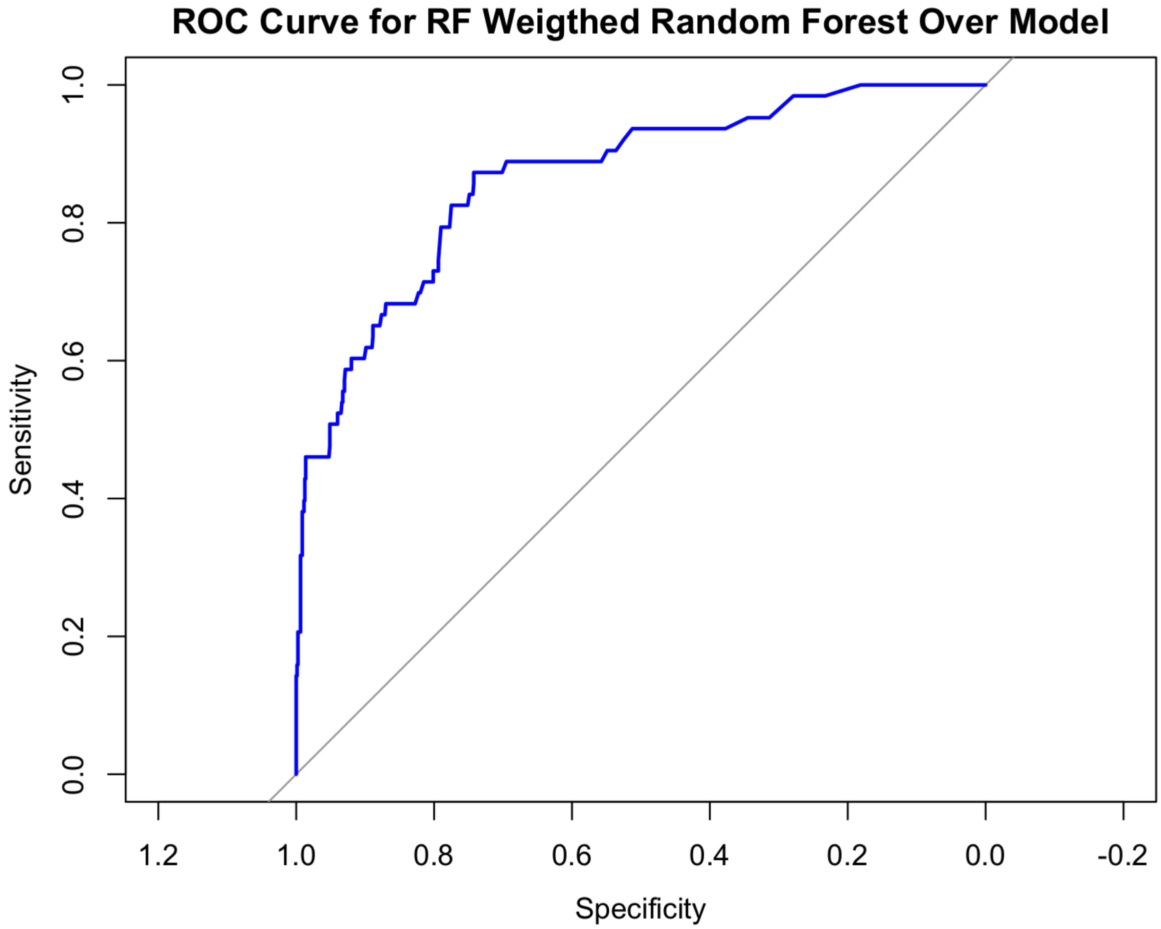
Correlation Naive Bayes Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.563	0.079	0.9890	0.563	0.9536	0.8828	0.2524	0.1378
Yes	0.9206	0.436	0.1425	0.9206	0.021	0.8828	0.2524	0.1378
Wt. Avg						0.8828	0.2524	0.1378



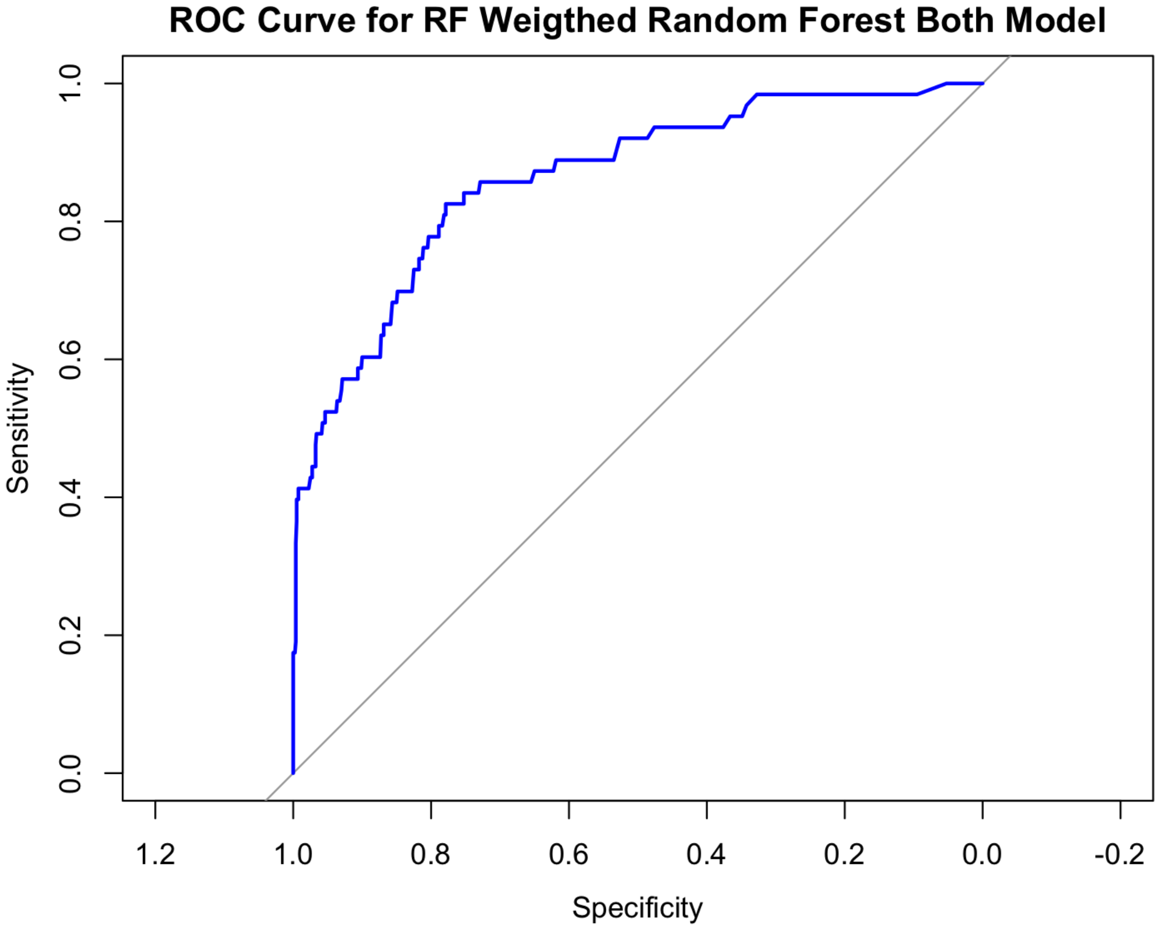
Correlation Random Forest Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.9775	0.539	0.9583	0.9775	0.622	0.8663	0.5018	0.4958
Yes	0.4603	0.02	0.6170	0.4603	0.029	0.8663	0.5018	0.4958
Wt. Avg	.75	.2495	.78	.78	.76	0.8663	0.5018	0.4958



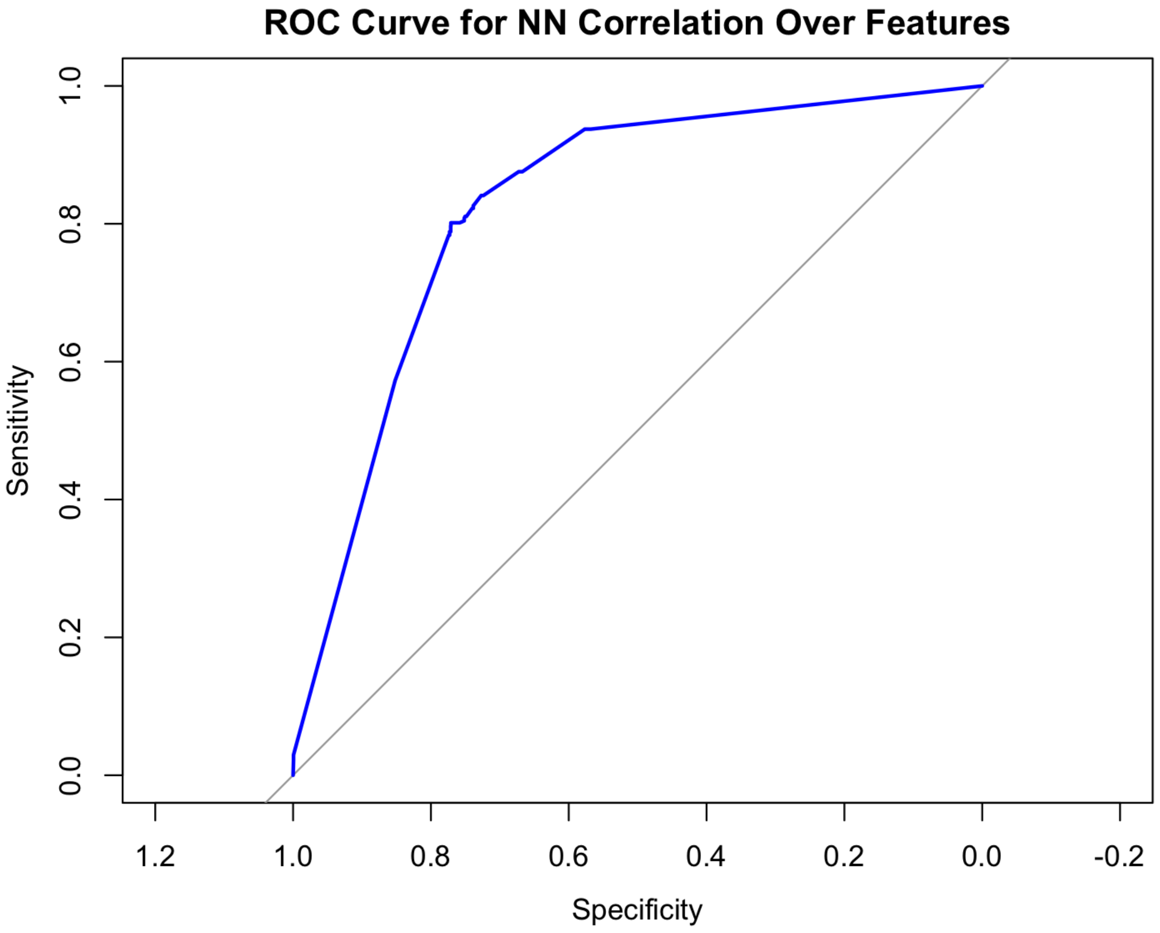
Correlation Random Forest Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.9287	0.428	0.9649	0.9287	0.7177	0.8626	0.4196	0.4102
Yes	0.5714	0.071	0.3871	0.5714	0.0469	0.8626	0.4196	0.4102
Wt. Avg	.75	.2495	.78	.78	.76	0.8626	0.4196	0.4102



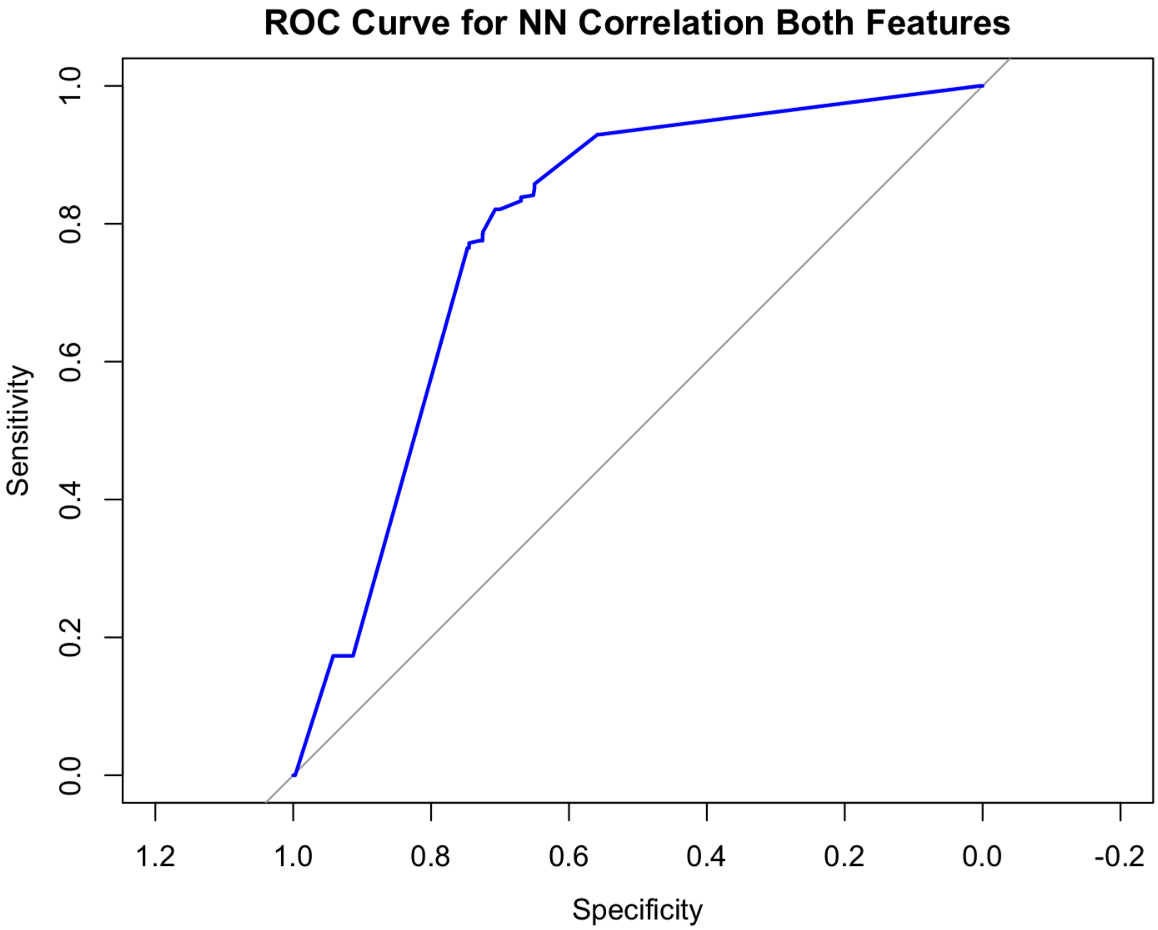
Correlation Neural Net Over Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.7238	0.428	0.8219	0.7238	0.7177	0.8272	0.5685	0.5644
Yes	0.8410	0.071	0.7503	0.8410	0.0469	0.8272	0.5685	0.5644
Wt. Avg	.76	.2495	.78	.78	.76	0.8272	0.5685	0.5644

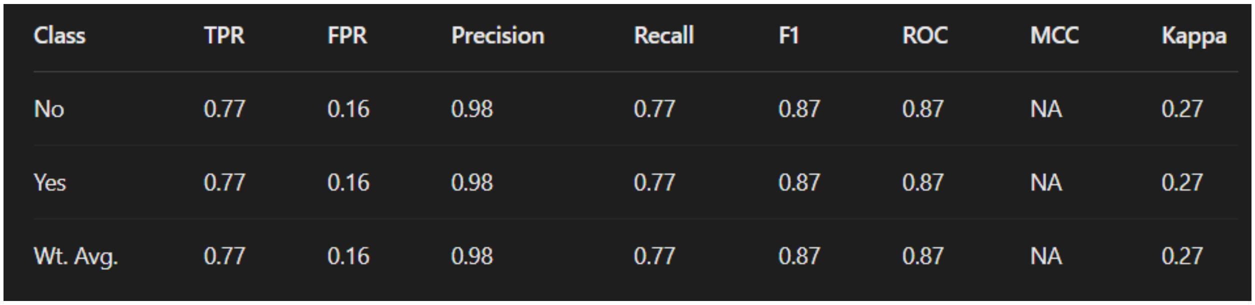


Correlation Neural Net Both Model:

	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
No	0.6695	0.166	0.8029	0.6695	0.7177	0.7872	0.5685	0.5022
Yes	0.8332	0.330	0.7131	0.8332	0.0469	0.7872	0.5685	0.5022
Wt. Avg	.77	.245	.756	.751	.3823	0.7872	0.5685	0.5022

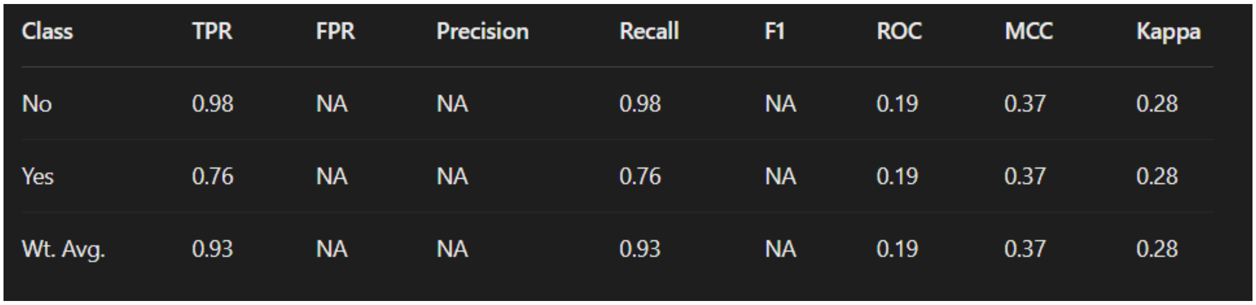


Logistic Regression NZV Over Sampling



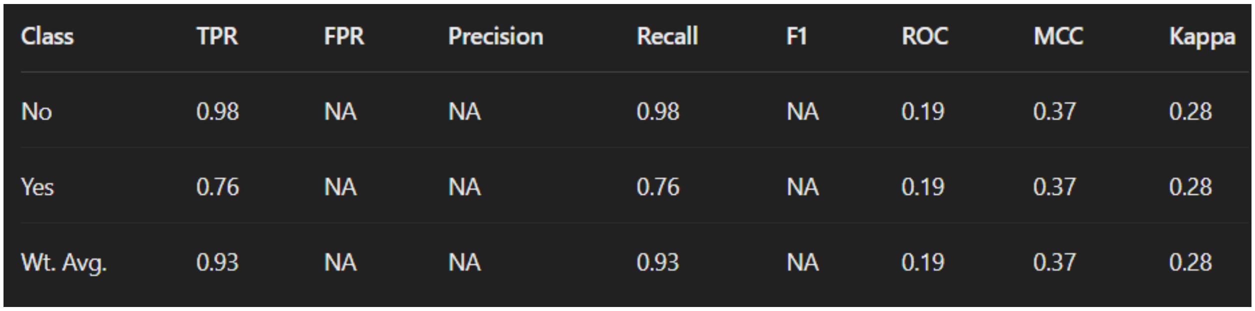
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27
Yes	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27
Wt. Avg.	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27

Logistic Regression NZV Both Sampling



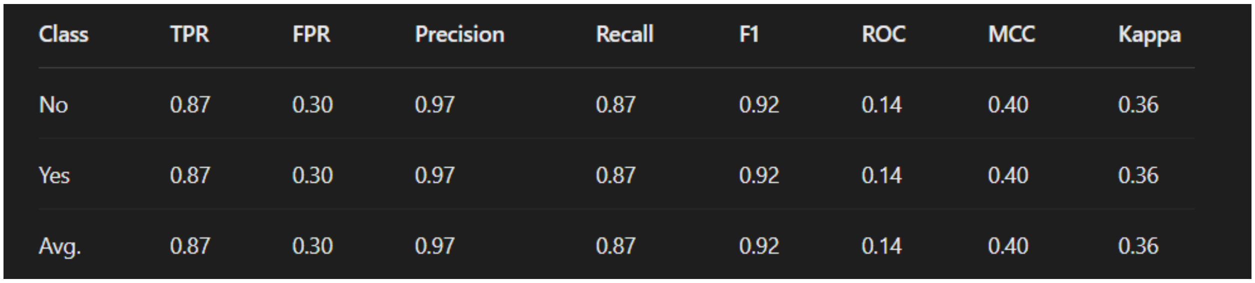
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.37	0.28
Yes	0.76	NA	NA	0.76	NA	0.19	0.37	0.28
Wt. Avg.	0.93	NA	NA	0.93	NA	0.19	0.37	0.28

Logistic Regression Correlation with ClassOver Sampling



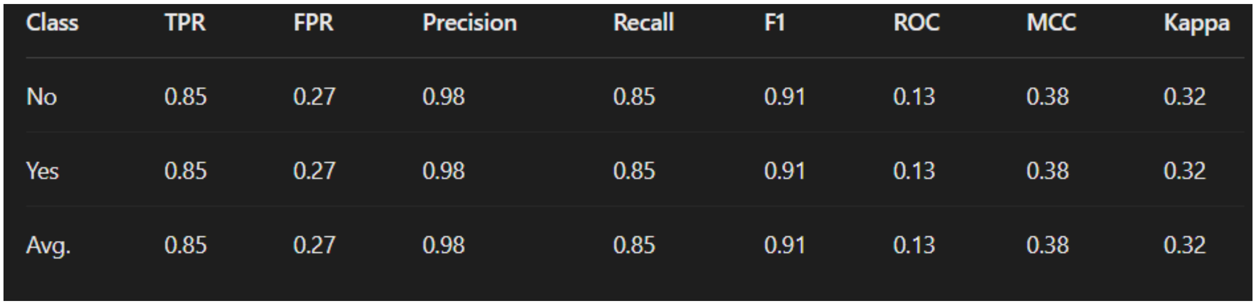
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.37	0.28
Yes	0.76	NA	NA	0.76	NA	0.19	0.37	0.28
Wt. Avg.	0.93	NA	NA	0.93	NA	0.19	0.37	0.28

Logistic Regression Correlation with ClassBoth Sampling



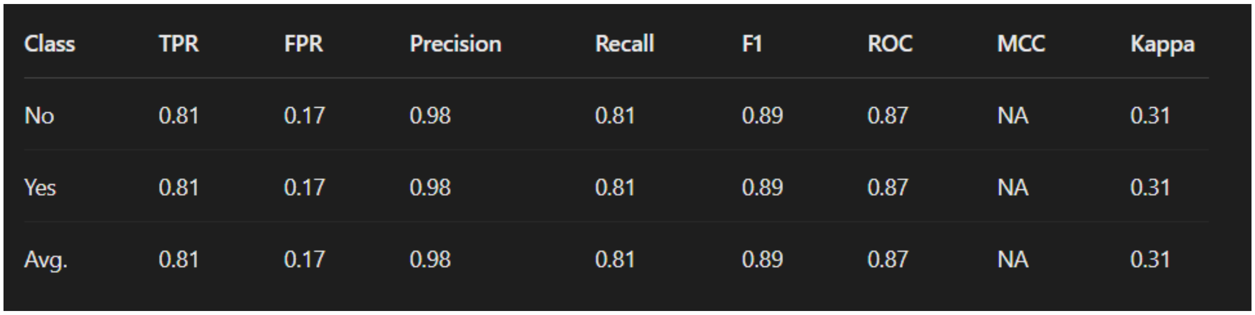
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.87	0.30	0.97	0.87	0.92	0.14	0.40	0.36
Yes	0.87	0.30	0.97	0.87	0.92	0.14	0.40	0.36
Avg.	0.87	0.30	0.97	0.87	0.92	0.14	0.40	0.36

Logistic Regression Random Forest Over Sampling



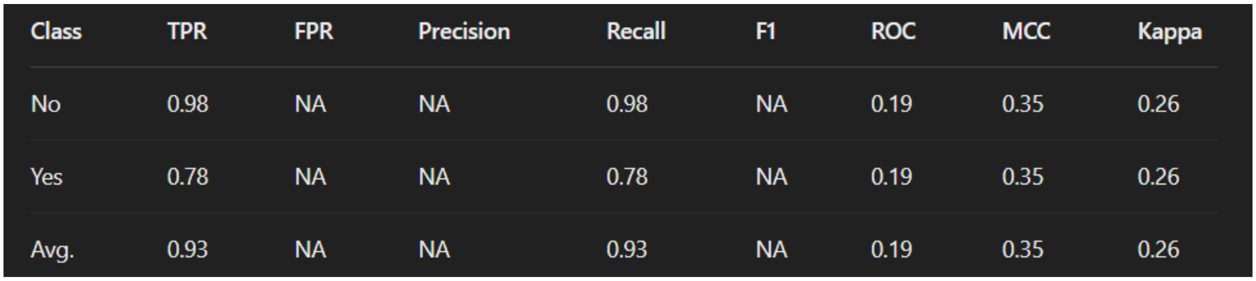
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32
Yes	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32
Avg.	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32

Logistic Regression Random Forest Both Sampling



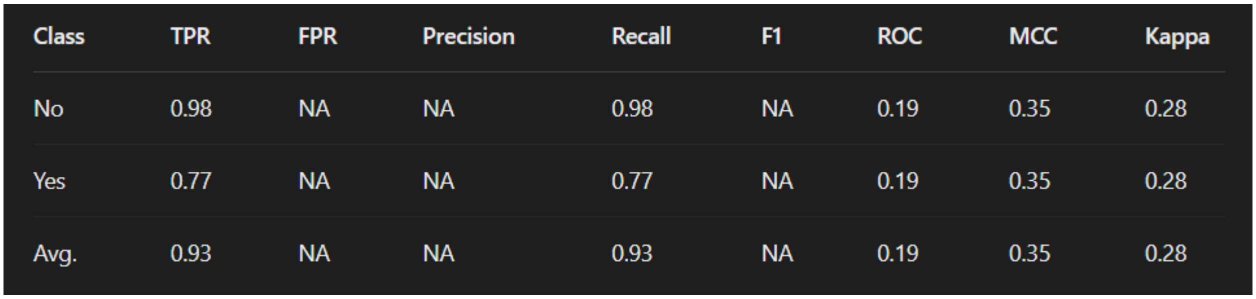
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.81	0.17	0.98	0.81	0.89	0.87	NA	0.31
Yes	0.81	0.17	0.98	0.81	0.89	0.87	NA	0.31
Avg.	0.81	0.17	0.98	0.81	0.89	0.87	NA	0.31

Decision Tree NZV Over Sampling



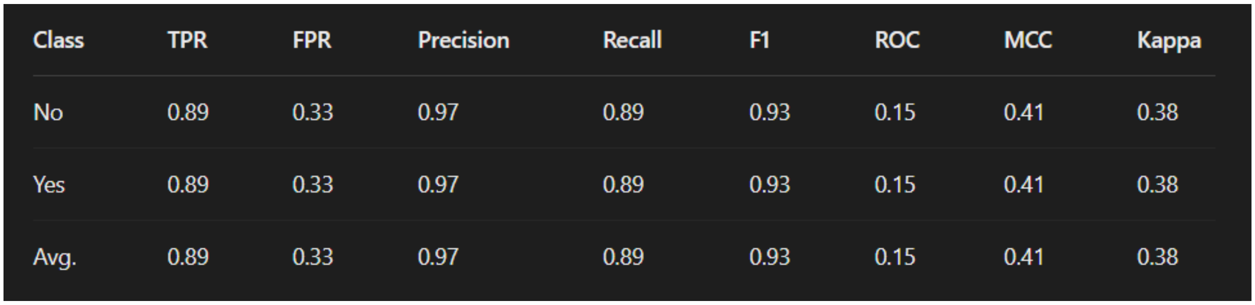
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.35	0.26
Yes	0.78	NA	NA	0.78	NA	0.19	0.35	0.26
Avg.	0.93	NA	NA	0.93	NA	0.19	0.35	0.26

Decision Tree NZV Both Sampling



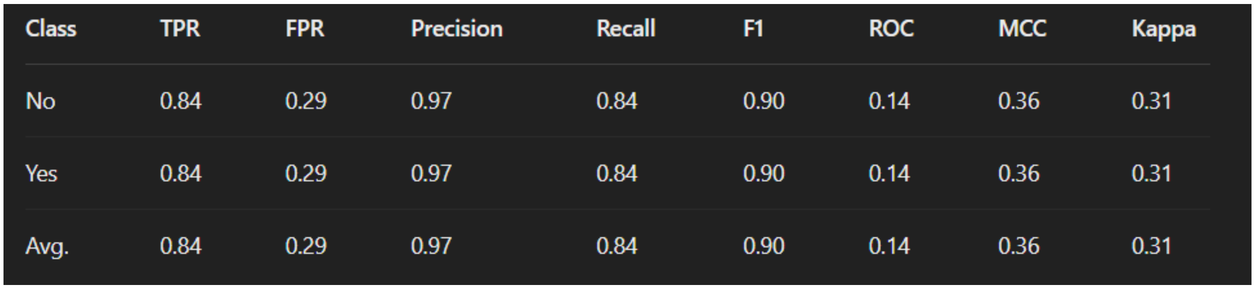
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.35	0.28
Yes	0.77	NA	NA	0.77	NA	0.19	0.35	0.28
Avg.	0.93	NA	NA	0.93	NA	0.19	0.35	0.28

Decision Tree Correlation with ClassOver Sampling



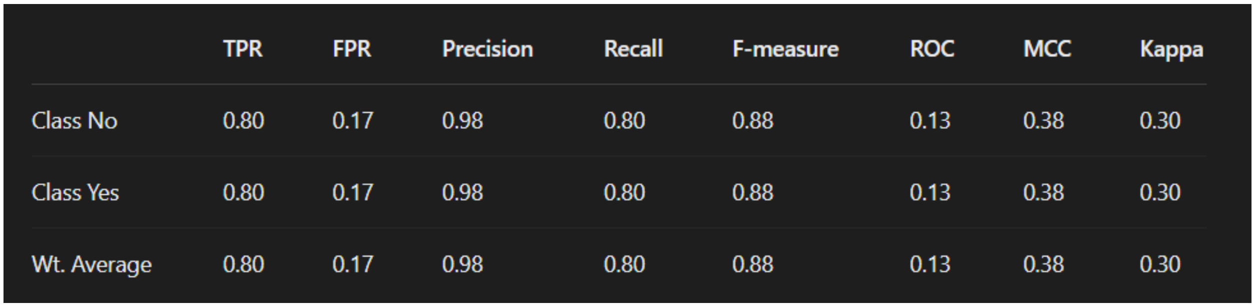
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.89	0.33	0.97	0.89	0.93	0.15	0.41	0.38
Yes	0.89	0.33	0.97	0.89	0.93	0.15	0.41	0.38
Avg.	0.89	0.33	0.97	0.89	0.93	0.15	0.41	0.38

Decision Tree Correlation with ClassBoth Sampling



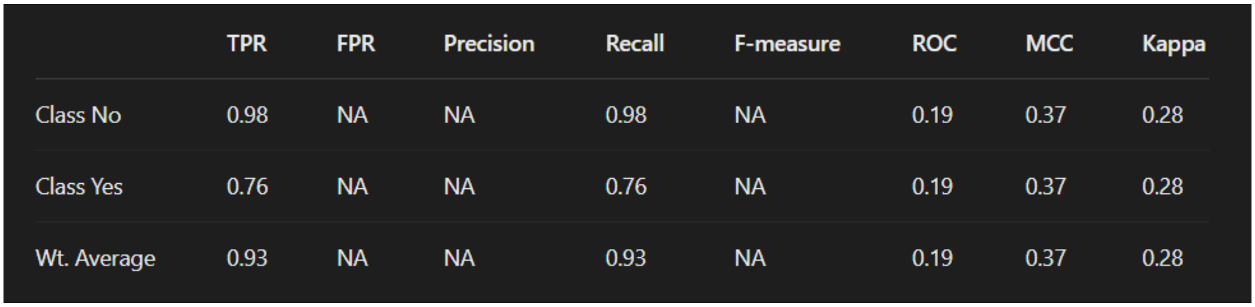
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31
Yes	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31
Avg.	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31

Decision Tree Random Forest Over Sampling



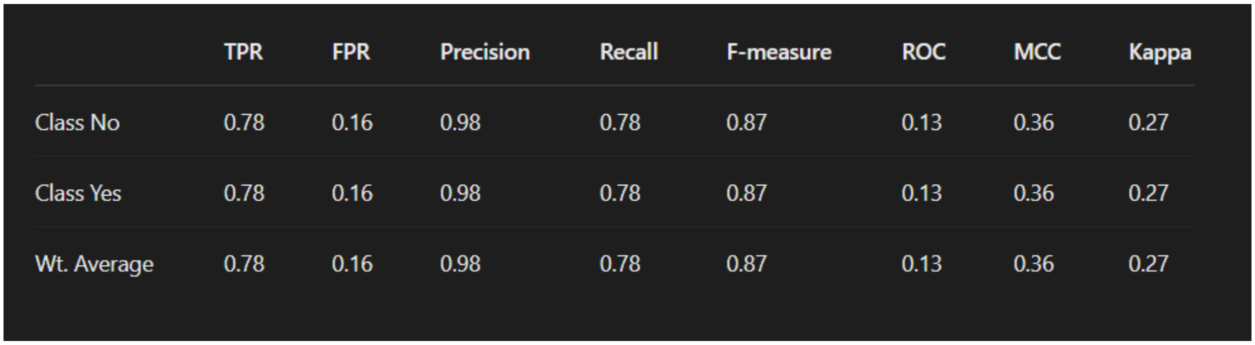
	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
Class No	0.80	0.17	0.98	0.80	0.88	0.13	0.38	0.30
Class Yes	0.80	0.17	0.98	0.80	0.88	0.13	0.38	0.30
Wt. Average	0.80	0.17	0.98	0.80	0.88	0.13	0.38	0.30

Decision Tree Random Forest Both Sampling



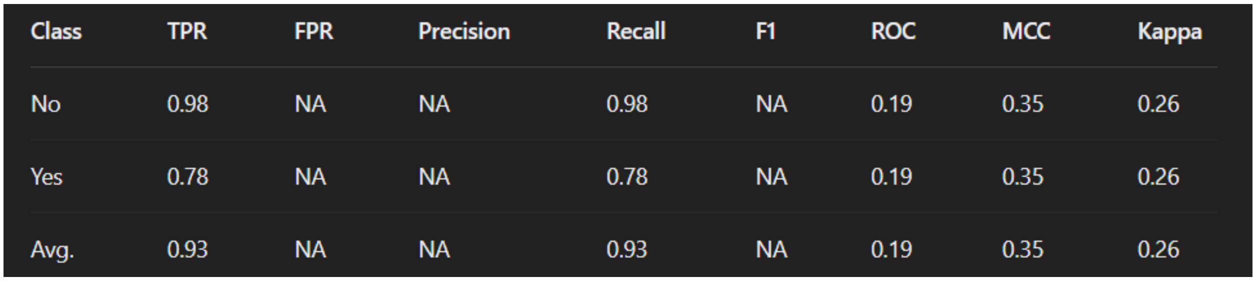
	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
Class No	0.98	NA	NA	0.98	NA	0.19	0.37	0.28
Class Yes	0.76	NA	NA	0.76	NA	0.19	0.37	0.28
Wt. Average	0.93	NA	NA	0.93	NA	0.19	0.37	0.28

SVM NZV Over Sampling



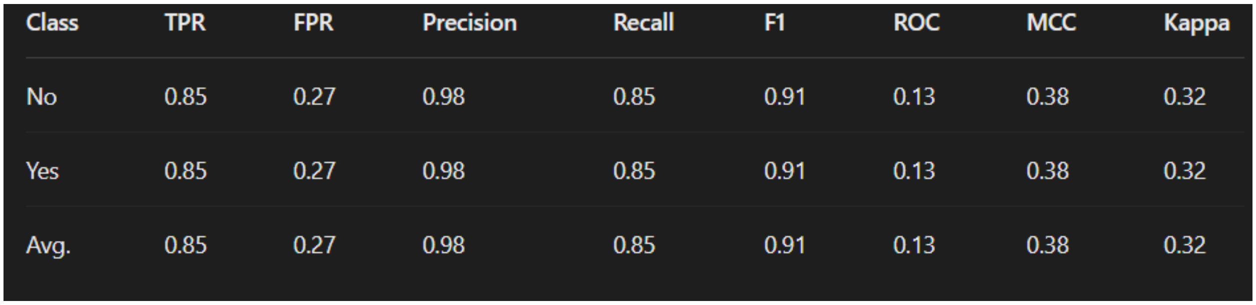
	TPR	FPR	Precision	Recall	F-measure	ROC	MCC	Kappa
Class No	0.78	0.16	0.98	0.78	0.87	0.13	0.36	0.27
Class Yes	0.78	0.16	0.98	0.78	0.87	0.13	0.36	0.27
Wt. Average	0.78	0.16	0.98	0.78	0.87	0.13	0.36	0.27

SVM NZV Both Sampling



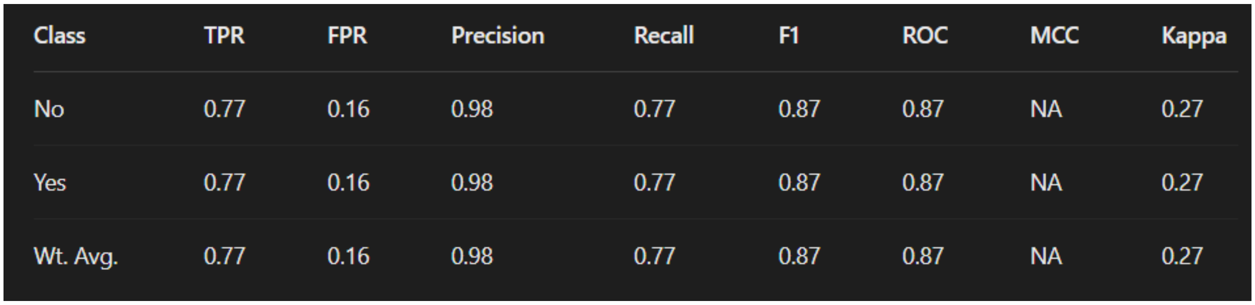
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.35	0.26
Yes	0.78	NA	NA	0.78	NA	0.19	0.35	0.26
Avg.	0.93	NA	NA	0.93	NA	0.19	0.35	0.26

SVM Correlation with ClassOver Sampling



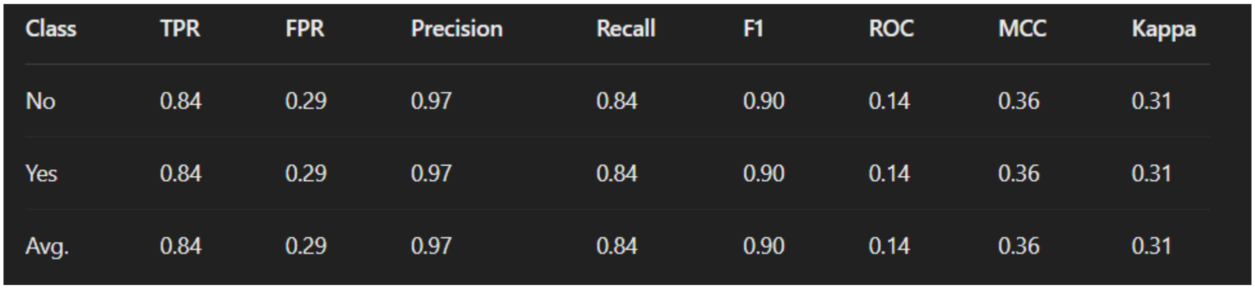
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32
Yes	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32
Avg.	0.85	0.27	0.98	0.85	0.91	0.13	0.38	0.32

SVM Correlation with ClassBoth Sampling



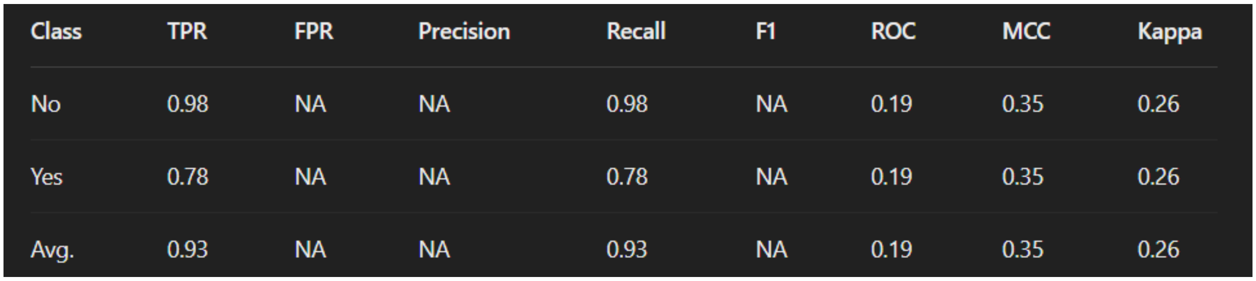
Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27
Yes	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27
Wt. Avg.	0.77	0.16	0.98	0.77	0.87	0.87	NA	0.27

SVM Random Forest Over Sampling



Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31
Yes	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31
Avg.	0.84	0.29	0.97	0.84	0.90	0.14	0.36	0.31

SVM Random Forest Both Sampling



Class	TPR	FPR	Precision	Recall	F1	ROC	MCC	Kappa
No	0.98	NA	NA	0.98	NA	0.19	0.35	0.26
Yes	0.78	NA	NA	0.78	NA	0.19	0.35	0.26
Avg.	0.93	NA	NA	0.93	NA	0.19	0.35	0.26