MCMC-GAN Report

# Dataset:

Raw Data Set: 88,647 samples

Legitimate (0): 58,000 samples, 65.43%

Phishing (1): 30,647 samples, 34.57%

In order to manipulate the dataset to give a more realistic unbalanced dataset, **unsampling method** was used to ensure integrity of the dataset so that it won’t sway the model’s accuracy.

Created 3 datasets:

* Dataset 1: 80% legitimate, 20% Phishing
* Dataset 2: 90% legitimate, 10% Phishing
* Dataset 3: 95% legitimate, 5% Phishing

**SelectKBest** is a feature selection method from scikit-learn that selects the top k features based on a scoring function. Each dataset was processed by:

* Loads the undersampled dataset.
* Applies SelectKBest using mutual\_info\_classif to get:
* Top 30 features
* Top 20 features
* Top 10 features
* Saves each of those as separate CSV files (e.g., top30.csv, top20.csv, top10.csv).
* Generated total of 9 datasets named:
* dataset80\_20\_top10.csv
* dataset80\_20\_top20.csv
* dataset80\_20\_top30.csv
* dataset90\_10\_top10.csv
* dataset90\_10\_top20.csv
* dataset90\_10\_top30.csv
* dataset95\_5\_top10.csv
* dataset95\_5\_top20.csv
* dataset95\_5\_top30.csv

# Scripts

2 scripts were created:

### **Script #1**

* "Start"
* "Read dataset"
* "Train GAN → Generate synthetic phishing"
* "Augment + Split dataset"
* "Train + Evaluate Models"
* "Save outputs: models, plots, logs"
* "End"

### Purpose:

* Loads dataset
* Augments minority class with GAN (phishing)
* Trains 5 models
* Evaluates models with full metrics
* Saves models, confusion matrices, logs, performance plots

### **Script #2**

* "Start"
* "Loop over dataset files"
* "Call GAN-MCMC.py for each dataset"
* "Print results/errors"
* "End"

### Purpose:

* Runs the Script #1and pipeline for **each dataset file** in a batch.

# Data Augmentation

## Run #1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Before Augmentation** | | | | **After Augmentation** | | | |
| **Legitimate** | **%** | **Phishing** | **%** | **Legitimate** | **%** | **Syn. Phishing** | **%** |
| 80\_20\_top10 | 40600 | 80 | 10150 | 20 | 40600 | 87.12 | 6000 | 12.88 |
| 80\_20\_top20 | 40600 | 80 | 10150 | 20 | 40600 | 87.12 | 6000 | 12.88 |
| 80\_20\_top30 | 40600 | 80 | 10150 | 20 | 40600 | 87.12 | 6000 | 12.88 |
| 90\_10\_top10 | 40597 | 90 | 4511 | 10 | 40597 | 87.12 | 6000 | 12.88 |
| 90\_10\_top20 | 40597 | 90 | 4511 | 10 | 40597 | 87.12 | 6000 | 12.88 |
| 90\_10\_top30 | 40597 | 90 | 4511 | 10 | 40597 | 87.12 | 6000 | 12.88 |
| 95\_5\_top10 | 40591 | 95 | 2136 | 5 | 40591 | 87.12 | 6000 | 12.88 |
| 95\_5\_top20 | 40591 | 95 | 2136 | 5 | 40591 | 87.12 | 6000 | 12.88 |
| 95\_5\_top30 | 40591 | 95 | 2136 | 5 | 40591 | 87.12 | 6000 | 12.88 |

## Run#2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Before Augmentation** | | | | **After Augmentation** | | | |
| **Legitimate** | **%** | **Phishing** | **%** | **Legitimate** | **%** | **Syn.Phishing** | **%** |
| 80\_20\_top10 | 40600 | 80 | 10150 | 20 | 40600 | 50 | 40600 | 50 |
| 80\_20\_top20 | 40600 | 80 | 10150 | 20 | 40600 | 50 | 40600 | 50 |
| 80\_20\_top30 | 40600 | 80 | 10150 | 20 | 40600 | 50 | 40600 | 50 |
| 90\_10\_top10 | 40597 | 90 | 4511 | 10 | 40597 | 50 | 40597 | 50 |
| 90\_10\_top20 | 40597 | 90 | 4511 | 10 | 40597 | 50 | 40597 | 50 |
| 90\_10\_top30 | 40597 | 90 | 4511 | 10 | 40597 | 50 | 40597 | 50 |
| 95\_5\_top10 | 40591 | 95 | 2136 | 5 | 40591 | 50 | 40591 | 50 |
| 95\_5\_top20 | 40591 | 95 | 2136 | 5 | 40591 | 50 | 40591 | 50 |
| 95\_5\_top30 | 40591 | 95 | 2136 | 5 | 40591 | 50 | 40591 | 50 |

# Model Training Accuracy

## Run #1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.915573 | 0.91492949 | 1.00 | 0.916700 | 0.916700 | 0.916904 | 0.9344 | 0.914929 | 0.916904 |
| Random\_Forest | 1.000000 | 0.99984672 | 1.00 | 1.000000 | 1.000000 | 1.000000 | 0.9327 | 0.999847 | 1.000000 |
| XGBoost | 0.999969 | 0.99963213 | 1.00 | 0.999448 | 0.999448 | 0.999663 | 0.9325 | 0.999632 | 0.999663 |
| CatBoost | 1.000000 | 0.99984672 | 1.00 | 1.000000 | 1.000000 | 1.000000 | 0.9331 | 0.999847 | 1.000000 |
| Stacking\_Ensemble | 0.999969 | 0.99981606 | 1.00 | 0.999939 | 0.999939 | 0.999939 | 0.9253 | 0.999816 | 0.999939 |

## Run #2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.90197044 | 0.90344828 | 0.90478536 | 0.90166271 | 0.90373889 | 0.9053928 | 0.9039189 | 0.90351417 | 0.9782673 |
| Random\_Forest | 0.95814567 | 0.96203378 | 0.96291344 | 0.9729744 | 0.97591273 | 0.98031143 | 0.9864677 | 0.98687244 | 0.9865028 |
| XGBoost | 0.95144265 | 0.95524279 | 0.95582336 | 0.97075746 | 0.97271048 | 0.97563121 | 0.9827370 | 0.98365214 | 0.9833705 |
| CatBoost | 0.95221675 | 0.95594652 | 0.95606967 | 0.97065189 | 0.97255212 | 0.97436439 | 0.9834409 | 0.98321221 | 0.9831242 |
| Stacking\_Ensemble | 0.95348346 | 0.95788177 | 0.95953554 | 0.97125011 | 0.97383654 | 0.97835841 | 0.9847255 | 0.98484875 | 0.9835113 |

# Model Test Accuracy

## Run #1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.915573 | 0.91981402 | 1.00 | 0.915379 | 0.916309 | 1.000000 | 0.918658 | 0.919814 | 0.91722707 |
| Random\_Forest | 1.000000 | 0.99985694 | 1.00 | 0.999928 | 1.000000 | 1.000000 | 1.000000 | 0.999857 | 1 |
| XGBoost | 0.999969 | 0.99957082 | 1.00 | 0.999857 | 0.999356 | 1.000000 | 0.999499 | 0.999571 | 1 |
| CatBoost | 1.000000 | 0.99985694 | 1.00 | 0.999928 | 1.000000 | 1.000000 | 1.000000 | 0.999857 | 1 |
| Stacking\_Ensemble | 0.999969 | 0.99985694 | 1.00 | 0.999928 | 0.999928 | 1.000000 | 1.000000 | 0.999857 | 1 |

## Run #2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.9020936 | 0.90439245 | 0.90402299 | 0.902418 | 0.9056201 | 0.90537378 | 0.90498871 | 0.90712379 | 0.9782673 |
| Random\_Forest | 0.95 | 0.94938424 | 0.95192939 | 0.96551583 | 0.96945687 | 0.97155056 | 0.98078423 | 0.98160542 | 0.9865028 |
| XGBoost | 0.95012315 | 0.94819376 | 0.95192939 | 0.96563898 | 0.97089371 | 0.97126319 | 0.98016834 | 0.98189284 | 0.9833705 |
| CatBoost | 0.94995895 | 0.94922003 | 0.95102627 | 0.96609056 | 0.97085266 | 0.97237161 | 0.98119483 | 0.98222131 | 0.9831242 |
| Stacking\_Ensemble | 0.9500821 | 0.9499179 | 0.95184729 | 0.96551583 | 0.97052424 | 0.97233056 | 0.98197495 | 0.98226237 | 0.9835113 |

# Model Performance Summary

## Model Runtime

## Run #1: Total Runtime (64.78s) 1 minute, 4 seconds

Time in seconds for each run on different datasets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.064932 | 0.04133987 | 0.070000 | 0.031000 | 0.054985 | 0.051403 | 0.038674 | 0.041340 | 0.05064917 |
| Random\_Forest | 0.375537 | 0.57116985 | 0.430000 | 0.343205 | 0.373059 | 0.416190 | 0.36863 | 0.571170 | 0.43906379 |
| XGBoost | 0.110879 | 0.13680887 | 0.130000 | 0.097424 | 0.097696 | 0.146045 | 0.101069 | 0.136809 | 0.14215589 |
| CatBoost | 3.399250 | 3.81893301 | 4.440000 | 3.270048 | 4.051742 | 4.648555 | 3.508223 | 3.818933 | 4.23389912 |
| Stacking\_Ensemble | 2.460165 | 3.35549188 | 2.790000 | 2.143260 | 2.416204 | 2.639271 | 2.349248 | 3.355492 | 2.65028095 |

## Run #2: Total Runtime (194.25s) 3 minutes, 14 seconds

Time in seconds for each run on different datasets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.4524941 | 0.6486461 | 1.3210160 | 0.29852605 | 0.43748617 | 0.4434852 | 0.39855099 | 0.30516696 | 0.64577103 |
| Random\_Forest | 1.8515760 | 2.5587520 | 3.0445470 | 1.57795811 | 2.00496626 | 2.3060731 | 1.42719221 | 1.77323818 | 1.93439889 |
| XGBoost | 0.2794430 | 0.2944991 | 0.3189139 | 0.27831006 | 0.28667283 | 0.3102090 | 0.27777004 | 0.29276419 | 0.31539297 |
| CatBoost | 5.2230460 | 6.0892400 | 6.7655007 | 5.57427096 | 5.69559288 | 6.7576439 | 4.89800906 | 5.67196298 | 6.39673185 |
| Stacking\_Ensemble | 11.458267 | 14.902006 | 19.640628 | 9.94195199 | 11.4385309 | 15.745398 | 8.9191699 | 11.0642629 | 11.9862711 |

## Evaluation- Run #1

## Precision

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.611829 | 0.61622732 | 1.000000 | 0.603419 | 0.606061 | 1.000000 | 0.61287 | 0.616227 | 0.60872506 |
| Random\_Forest | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| XGBoost | 0.999445 | 1.000000 | 1.000000 | 1.000000 | 0.999442 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| CatBoost | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Stacking\_Ensemble | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

## Recall

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 1.000000 | 1.000000 | 1.000000 | 0.031000 | 1.000000 | 1.000000 | 0.038674 | 1.000000 | 1.000000 |
| Random\_Forest | 1.000000 | 0.99888889 | 1.000000 | 0.343205 | 1.000000 | 1.000000 | 0.36863 | 0.998889 | 1.000000 |
| XGBoost | 1.000000 | 0.99666667 | 1.000000 | 0.097424 | 0.995556 | 1.000000 | 0.101069 | 0.996667 | 1.000000 |
| CatBoost | 1.000000 | 0.99888889 | 1.000000 | 3.270048 | 1.000000 | 1.000000 | 3.508223 | 0.998889 | 1.000000 |
| Stacking\_Ensemble | 1.000000 | 0.99888889 | 1.000000 | 2.143260 | 0.999444 | 1.000000 | 2.349248 | 0.998889 | 1.000000 |

## F1- Score

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.759173 | 0.76255031 | 1.000000 | 0.031000 | 0.754717 | 1.000000 | 0.038674 | 0.762550 | 0.75677948 |
| Random\_Forest | 1.000000 | 0.99944414 | 1.000000 | 0.343205 | 1.000000 | 1.000000 | 0.36863 | 0.999444 | 1.000000 |
| XGBoost | 0.999722 | 0.99833055 | 1.000000 | 0.097424 | 0.997495 | 1.000000 | 0.101069 | 0.998331 | 1.000000 |
| CatBoost | 1.000000 | 0.99944414 | 1.000000 | 3.270048 | 1.000000 | 1.000000 | 3.508223 | 0.999444 | 1.000000 |
| Stacking\_Ensemble | 1.000000 | 0.99944414 | 1.000000 | 2.143260 | 0.999722 | 1.000000 | 2.349248 | 0.999444 | 1.000000 |

## Evaluation- Run #2

## Precision

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.8421 | 0.8444 | 0.8446 | 0.8453 | 0.8478 | 0.8484 | 0.8511 | 0.8547 | 0.9915 |
| Random\_Forest | 0.9633 | 0.9652 | 0.9616 | 0.9789 | 0.9802 | 0.9834 | 0.9888 | 0.9905 | 0.9906 |
| XGBoost | 0.9559 | 0.9592 | 0.9592 | 0.9773 | 0.9817 | 0.9827 | 0.9882 | 0.9915 | 0.9892 |
| CatBoost | 0.9607 | 0.9616 | 0.9554 | 0.9788 | 0.9820 | 0.9824 | 0.9886 | 0.9913 | 0.9908 |
| Stacking\_Ensemble | 0.9458 | 0.9523 | 0.9512 | 0.9745 | 0.9774 | 0.9797 | 0.9870 | 0.9899 | 0.9885 |

## Recall

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.9898 | 0.9915 | 0.9902 | 0.9851 | 0.9888 | 0.9872 | 0.9817 | 0.9810 | 0.9676 |
| Random\_Forest | 0.9356 | 0.9324 | 0.9415 | 0.9516 | 0.9583 | 0.9593 | 0.9726 | 0.9726 | 0.9731 |
| XGBoost | 0.9438 | 0.9362 | 0.9440 | 0.9534 | 0.9597 | 0.9594 | 0.9719 | 0.9721 | 0.9740 |
| CatBoost | 0.9383 | 0.9358 | 0.9462 | 0.9528 | 0.9593 | 0.9620 | 0.9736 | 0.9730 | 0.9733 |
| Stacking\_Ensemble | 0.9548 | 0.9473 | 0.9525 | 0.9561 | 0.9633 | 0.9647 | 0.9768 | 0.9745 | 0.9752 |

## ROC-AUC

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.9021 | 0.9044 | 0.9040 | 0.9024 | 0.9056 | 0.9054 | 0.9050 | 0.9071 | 0.9797 |
| Random\_Forest | 0.9500 | 0.9494 | 0.9519 | 0.9655 | 0.9695 | 0.9716 | 0.9808 | 0.9816 | 0.9820 |
| XGBoost | 0.9501 | 0.9482 | 0.9519 | 0.9656 | 0.9709 | 0.9713 | 0.9802 | 0.9819 | 0.9816 |
| CatBoost | 0.9500 | 0.9492 | 0.9510 | 0.9661 | 0.9709 | 0.9724 | 0.9812 | 0.9822 | 0.9821 |
| Stacking\_Ensemble | 0.9501 | 0.9499 | 0.9518 | 0.9655 | 0.9705 | 0.9723 | 0.9820 | 0.9823 | 0.9819 |

## F1-Score

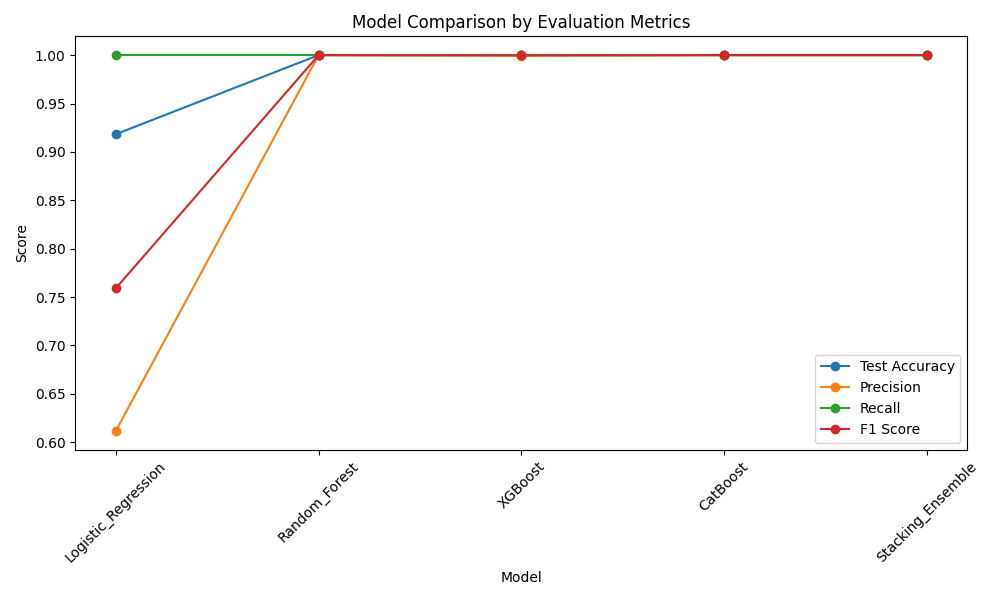
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **80:20** | | | **90:10** | | | **95:5** | | |
|  | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** | **10-F** | **20-F** | **30-F** |
| Logistic\_Regression | 0.9100 | 0.9121 | 0.9116 | 0.9099 | 0.9129 | 0.9125 | 0.9118 | 0.9135 | 0.9794 |
| Random\_Forest | 0.9493 | 0.9485 | 0.9514 | 0.9650 | 0.9691 | 0.9712 | 0.9806 | 0.9814 | 0.9818 |
| XGBoost | 0.9498 | 0.9476 | 0.9515 | 0.9652 | 0.9706 | 0.9709 | 0.9800 | 0.9817 | 0.9815 |
| CatBoost | 0.9494 | 0.9485 | 0.9508 | 0.9656 | 0.9705 | 0.9721 | 0.9811 | 0.9821 | 0.9820 |
| Stacking\_Ensemble | 0.9503 | 0.9498 | 0.9519 | 0.9652 | 0.9703 | 0.9721 | 0.9819 | 0.9821 | 0.9818 |

# Model Plot Comparison

## Run #1

### **Dataset 80%,20%**

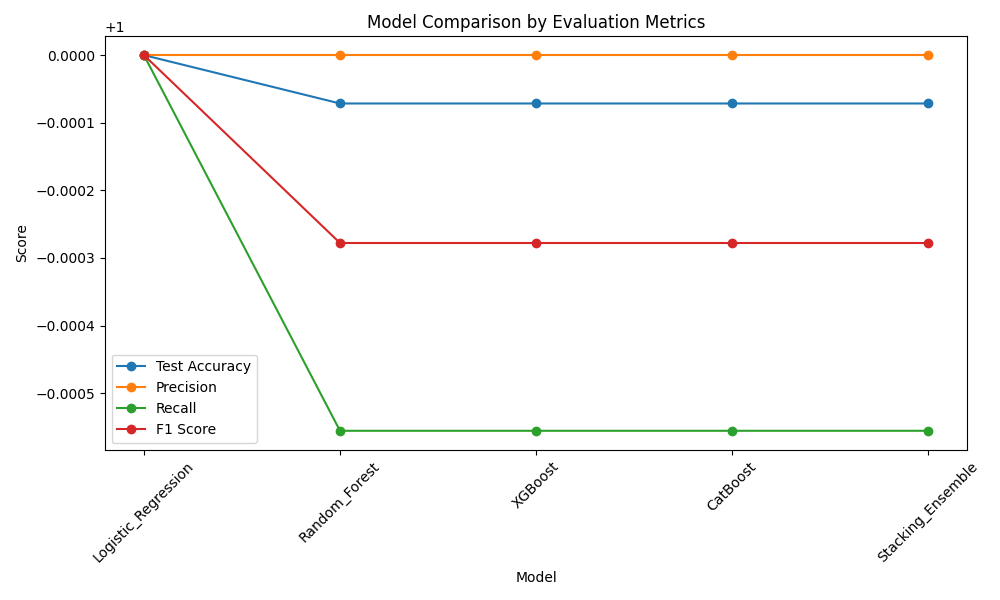
### Top 10 Features Model Comparison



### Top 20 Features Model Comparison

## 

### Top 30 Features Model Comparison

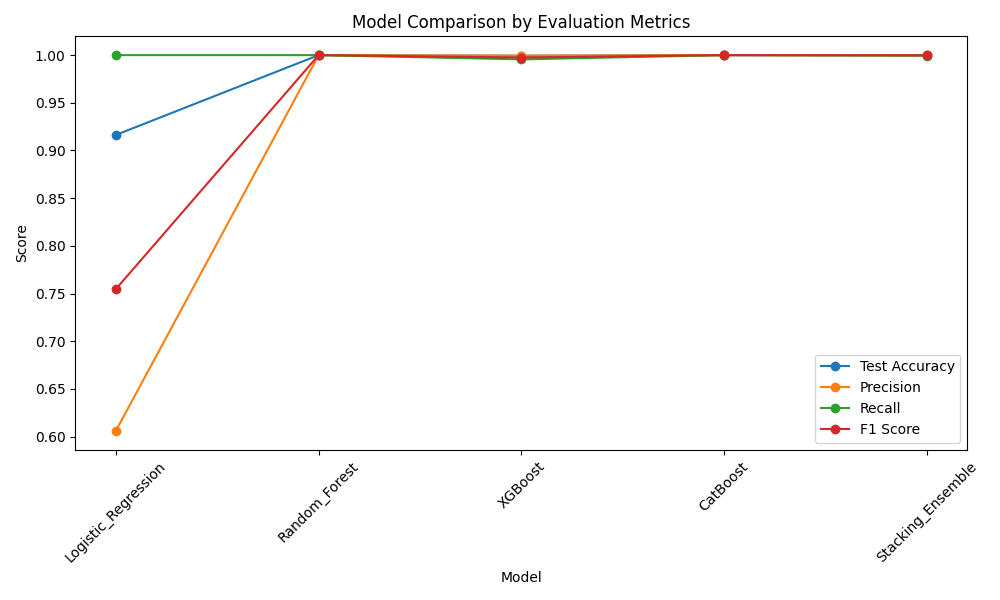


### **Dataset 90%,10%**

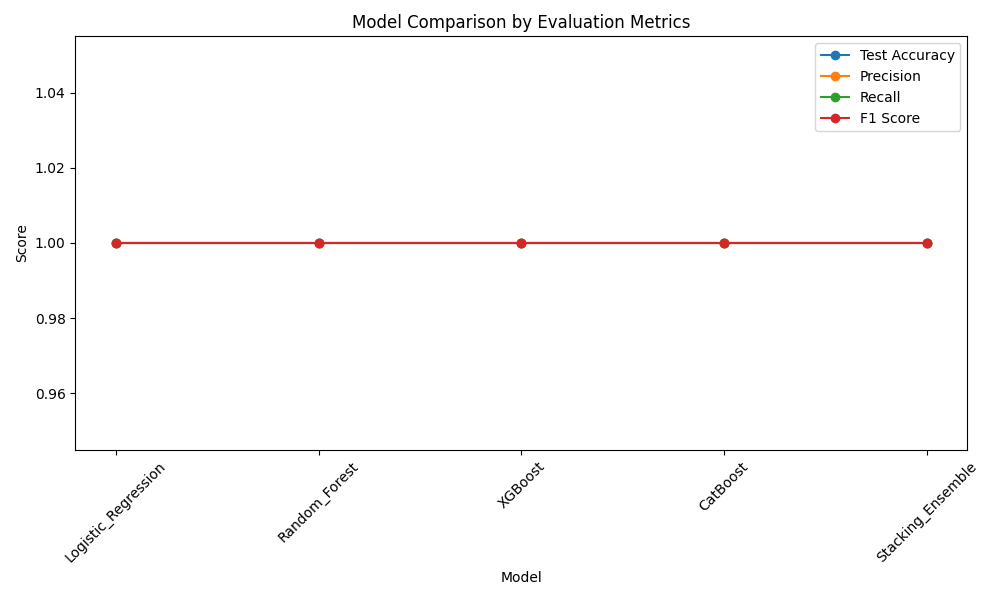
### Top 10 Features Model Comparison

## 

### Top 20 Features Model Comparison

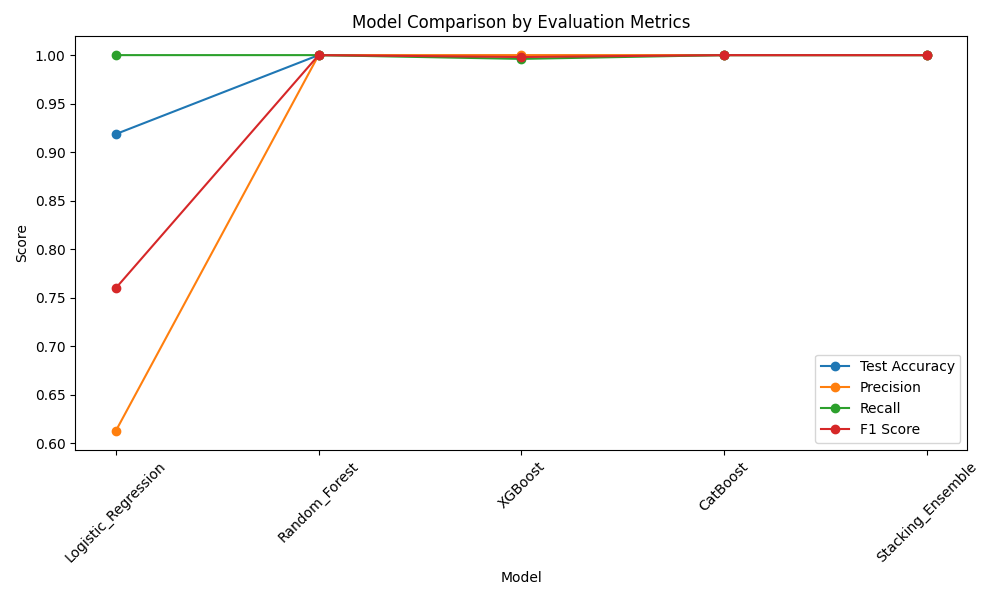


### Top 30 Features Model Comparison

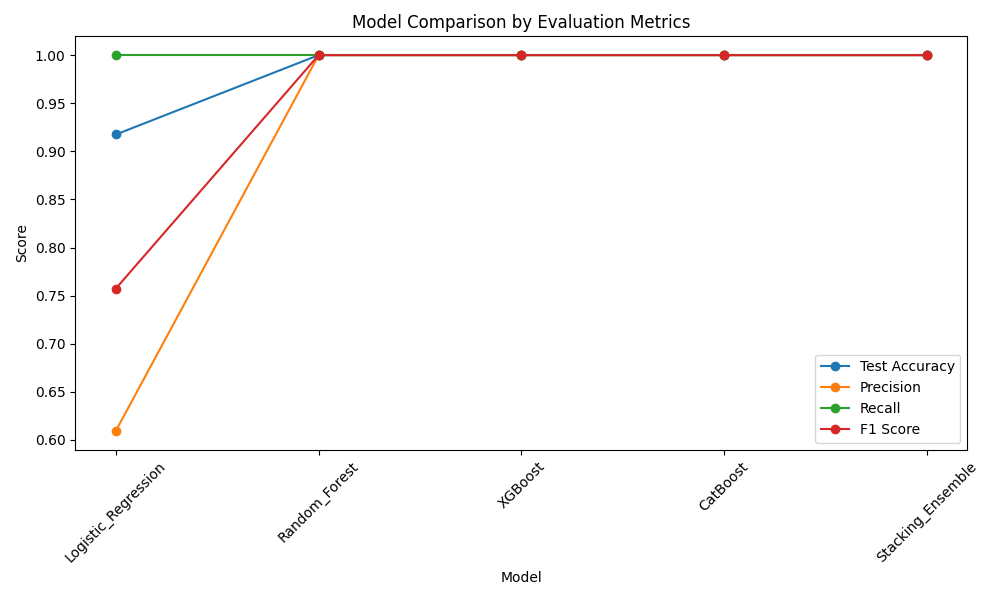


### **Dataset 95%,5%**

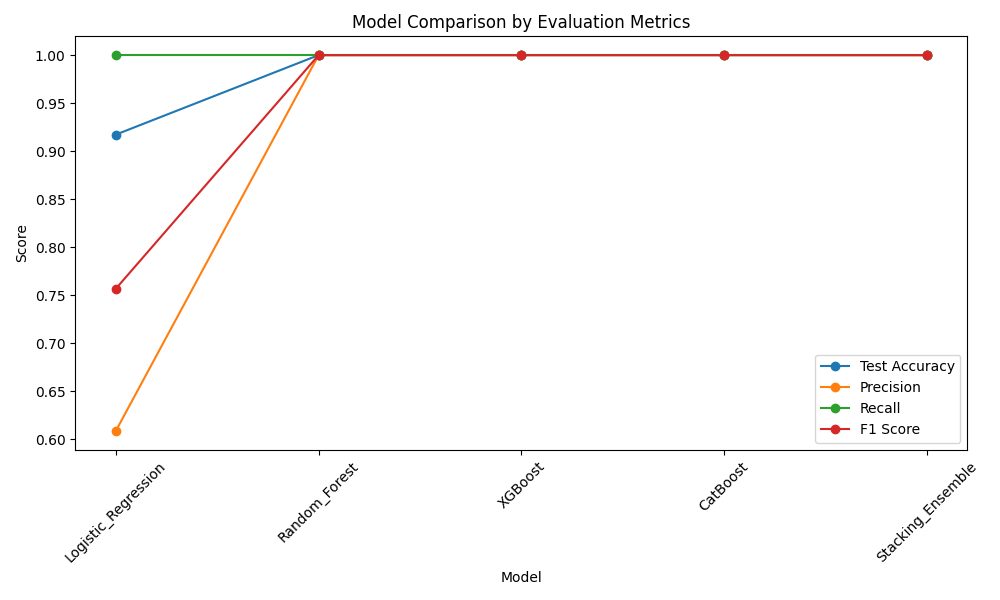
### Top 10 Features Model Comparison



### Top 20 Features Model Comparison



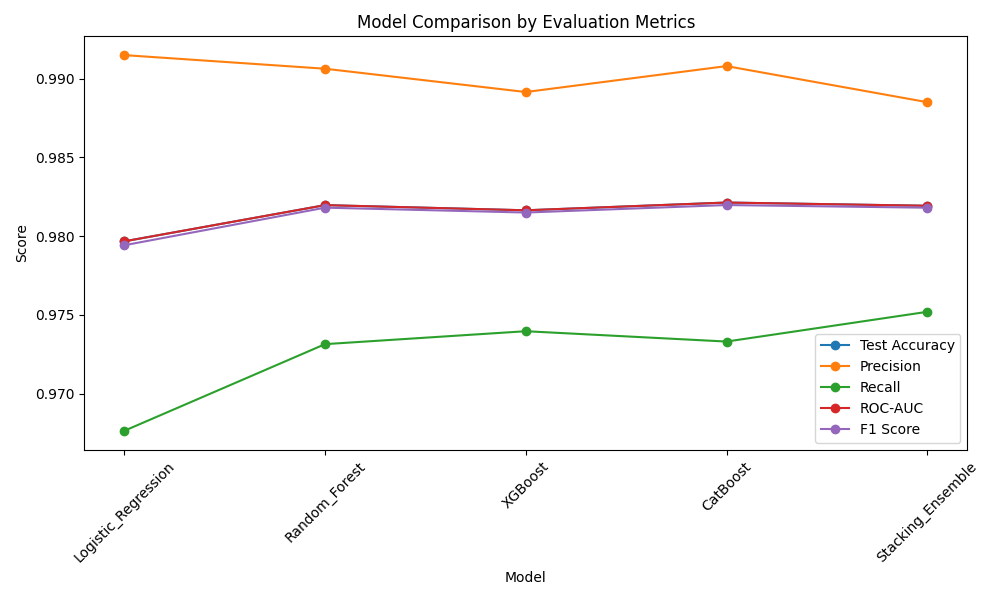
### Top 30 Features Model Comparison



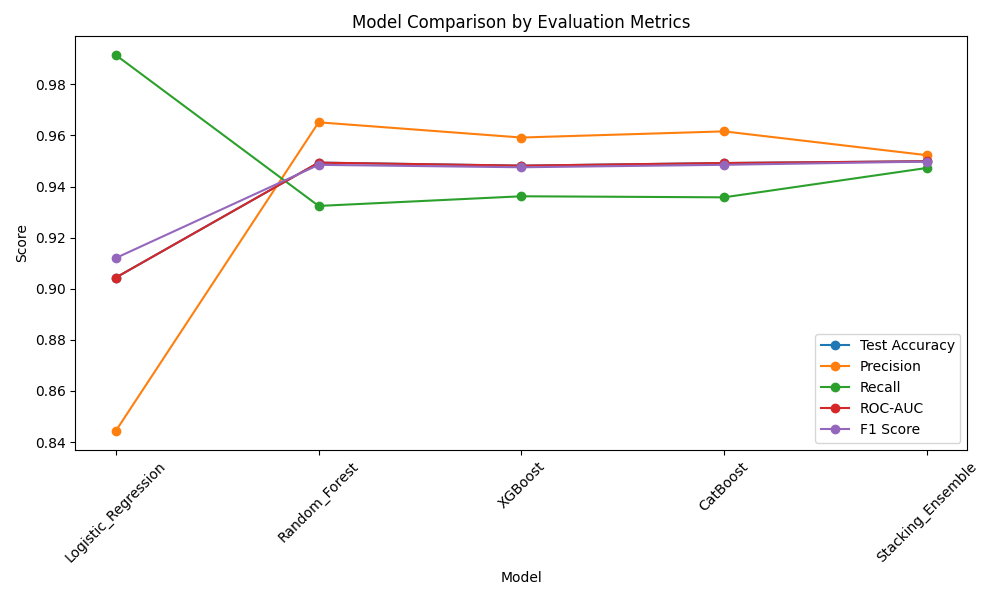
## Run #2

### **Dataset 80%,20%**

### Top 10 Features Model Comparison



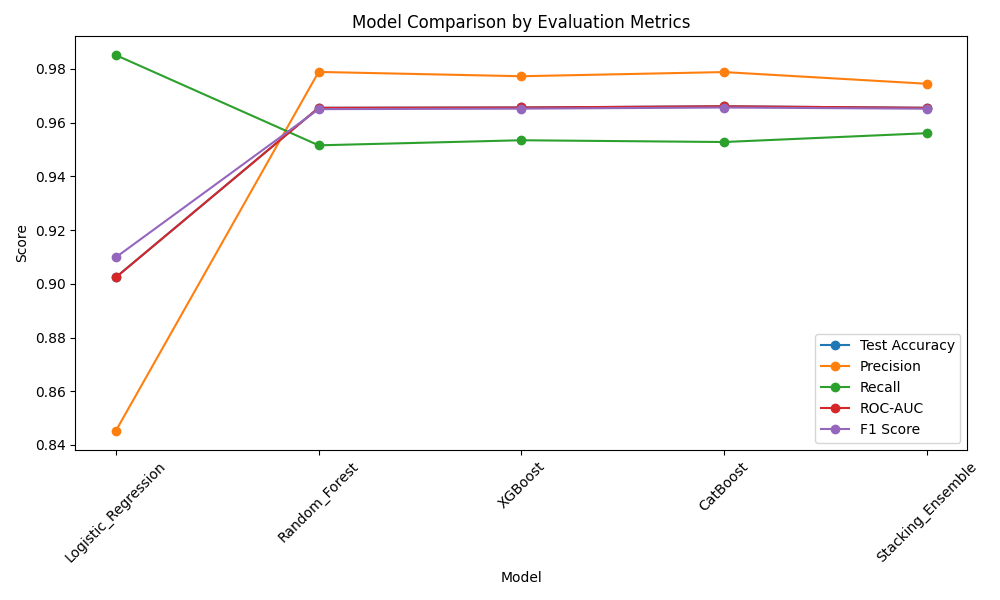
### Top 20 Features Model Comparison



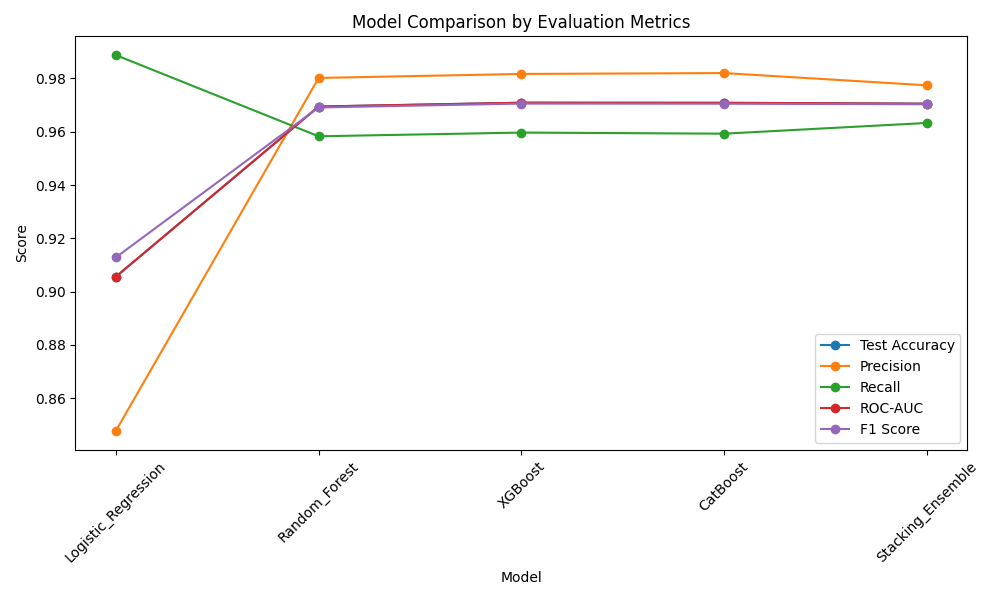
### Top 30 Features Model Comparison

### **Dataset 90%,10%**

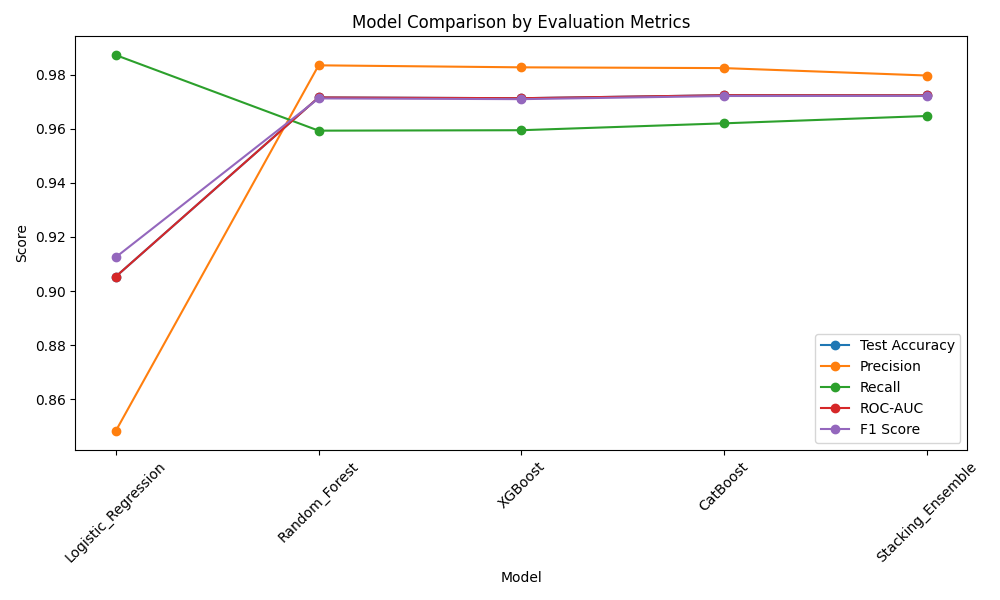
### Top 10 Features Model Comparison



### Top 20 Features Model Comparison

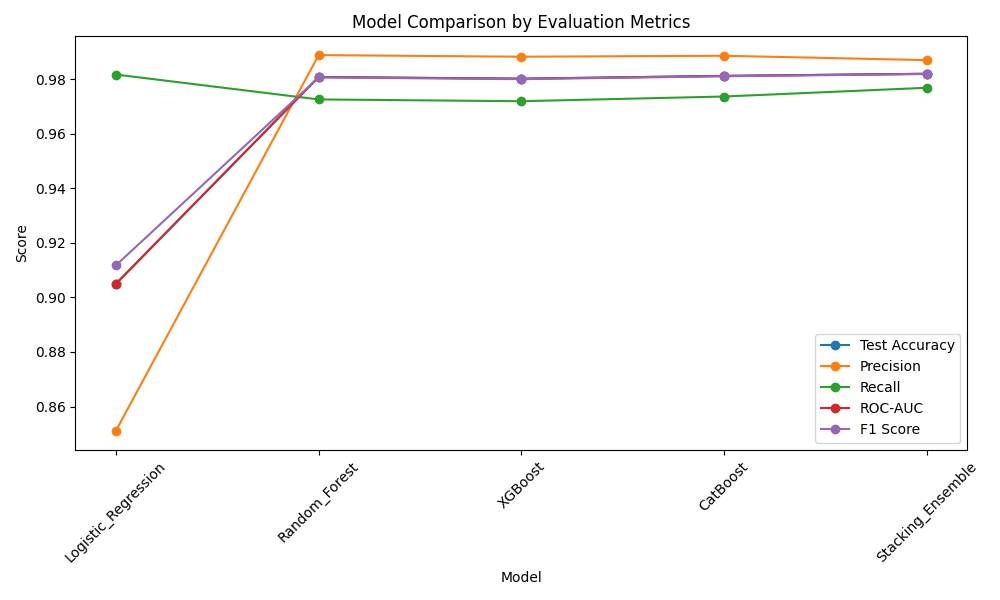


### Top 30 Features Model Comparison

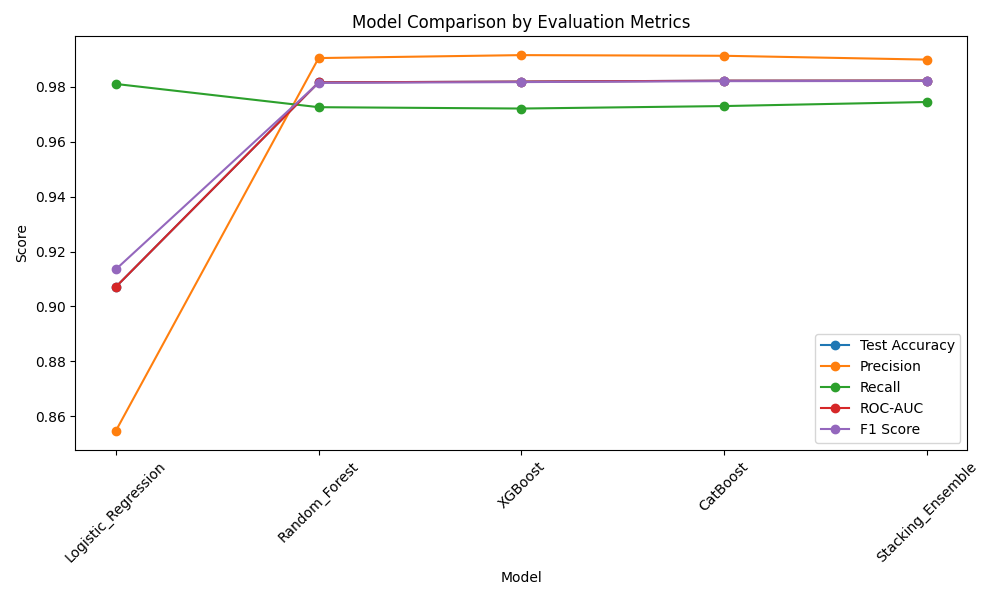


### **Dataset 95%,5%**

### Top 10 Features Model Comparison



### Top 20 Features Model Comparison



### Top 30 Features Model Comparisons

### 