HIV/AIDS

Predictive Analytics for Viral Suppression

DAV-Data Science

Group-2:
Gagan Preet Singh
Karina Thapa
Vidhika Jain

TABLE OF CONTENTS

01 **ABSTRACT & INTRODUCTION**

02 **PROJECT PROPOSAL**

- Scope & Aim
- Research Questions

03 **DATA PROFILING**

- Pre-Profiling
- Data Collection
- Data Cleaning
- Post-Profiling

04 **MODEL DEVELOPMENT**

- Logistic Regression
- Random Forest
- XGBoost
- Neural Networks

05 **MODEL EVALUATION & SELECTION**

06 **ADDRESSING RESULTS**

- Research Questions

07 **CONCLUSIONS RECOMMENDATION & LEARNING**

01

ABSTRACT INTRODUCTION

Objective
&
Background

Objective:

Analyze New York City's detailed HIV/AIDS data to identify factors influencing viral suppression, a key indicator of successful treatment.

Background:

New York City, known for its diversity and population density, provides a rich dataset through its annual HIV/AIDS reports. This project focuses on understanding how demographics such as age, gender, borough, and race impact viral suppression among residents diagnosed with HIV.

02

PROJECT PROPOSAL

Scope Aim
&
Research Questions

Scope & Aim:

- Scope: Utilize New York City's comprehensive HIV/AIDS data to conduct a deep analysis of factors affecting viral suppression among the city's diverse population.
- Aim: To enhance understanding of demographic impacts on viral suppression rates and provide recommendations to improve public health strategies.

Research Question:

- Demographic Influence: Which demographics have the highest rate of viral suppression in NYC, and what factors contribute to these outcomes?
- Temporal Trends: Are there significant changes over time in viral suppression rates among different boroughs?
- Geographical Impact: Does the United Hospital Fund neighborhood correlate with the likelihood of achieving viral suppression?
- Age and Gender Effects: How do age and gender impact the rates of viral suppression among those diagnosed with HIV?
- Predictive Modeling: Can machine learning models effectively predict the proportion of individuals likely to achieve viral suppression?

03

Data PROFILING

- Pre-Profiling
- Data Cleaning
- Post-Profiling

Data Summary

Data Source: NYC-OpenData https://opendata.cityofnewyork.us/

Total Columns (Features): There are 18 features in the dataset.

Numerical Features:

- HIV diagnoses: Number of HIV diagnoses among individuals aged 13 or older.
- HIV diagnosis rate: Rate of HIV diagnoses per 100,000 population.
- Concurrent diagnoses: Number of HIV diagnoses with a concurrent AIDS diagnosis within 31 days.
- % linked to care within 3 months: Percentage of new HIV diagnoses with an HIV viral load or CD4 test within 3 months of diagnosis.
- `AIDS diagnoses: Number of AIDS diagnoses among individuals aged 13 or older.
- AIDS diagnosis rate: Rate of AIDS diagnoses per 100,000 population.
- PLWDHI prevalence: Estimated number of people living with diagnosed HIV infection per 100 population.
- % viral suppression: Percentage of people living with diagnosed HIV infection aged 13 or older with viral load ≤200 copies/mL.
- Deaths: Number of deaths from any cause among people with HIV/AIDS aged 13 or older.
- Death rate: Deaths per 1,000 mid-year people living with HIV/AIDS.
- HIV-related death rate: Death rate for those assigned an HIV-related cause of death.
- Non-HIV-related death rate: Death rate for those assigned a non-HIV-related cause of death.

Categorical Features:

- Year: Calendar year of the report.
- Borough: Borough of residence at different stages of HIV/AIDS (diagnosis, treatment, death).
- UHF: United Hospital Fund neighborhood at different stages of HIV/AIDS.
- Gender: Gender of the individual, including a category for transgender.
- Age: Age at different stages of HIV/AIDS.
- Race: Race/ethnicity of the individual.

Pre-EDA and Challenges

Key Numerical Insights

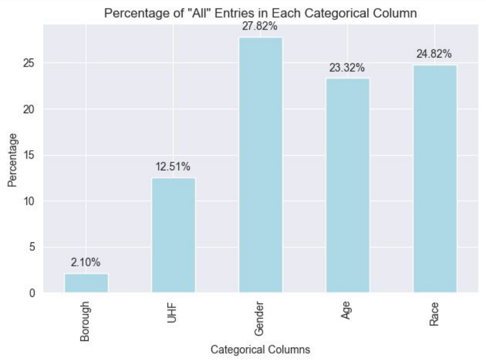
- Variability in Diagnoses: Extensive range in HIV and AIDS diagnosis rates, with some outliers potentially indicating data entry errors or demographic-specific outbreaks.
- Care and Viral Suppression:
  - 81.78% average linkage to care within three months post-diagnosis.
  - 68.20% average for achieving viral suppression, underscoring potential areas for healthcare improvement.

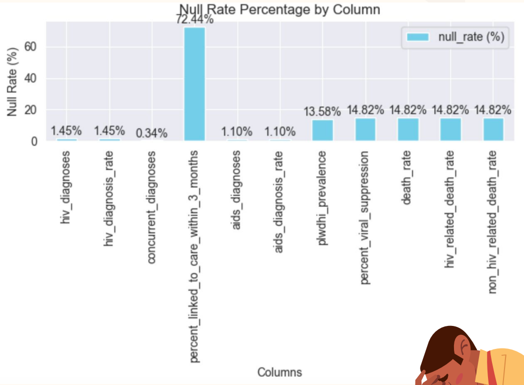
Categorical Data Distribution

- High Frequency of 'ALL' Entries:
  - Common across boroughs, age, and race categories, aggregating data that may hide specific subgroup trends.
- Geographic and Demographic Details:
  - All five NYC boroughs included, with Brooklyn showing the highest data entries.
  - Data spans diverse UHF neighborhoods and demographic groups, enabling detailed subgroup analysis.

Data Integrity and Issues:

- Missing Data: Key variables such as HIV diagnosis rates and viral suppression percentages have notable missing values, comprising about 4.5% of the dataset, which could skew analysis results.
- Significant Zeros: Presence of zeros in crucial categories like HIV diagnoses and deaths suggests underreporting or non-reporting which might affect the accuracy of statistical analysis.
- Highly Correlated Variables: Identifying highly correlated variables like AIDS diagnoses and AIDS diagnosis rate, which might lead to multicollinearity if used together in predictive models.
- Data Skewness: Data distribution for many key variables is heavily skewed, necessitating transformations for more accurate modeling and analysis.



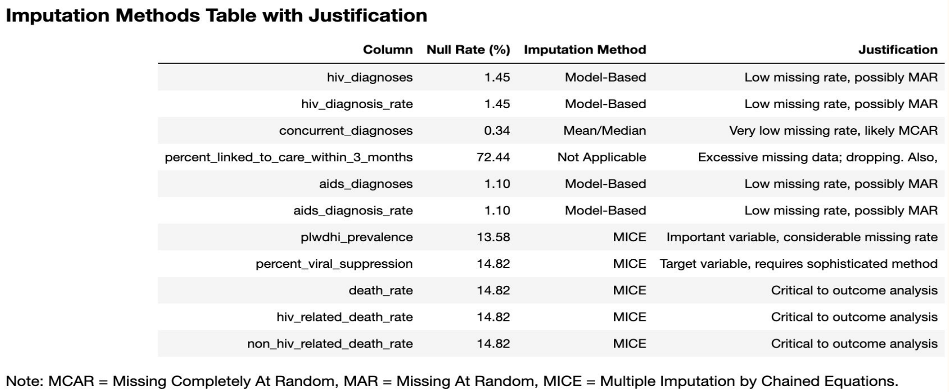


- PreEDA_Data_Profiling_Report - PDF
- PreEDA_Data_Profiling_Report - HTML

Issues Identification

- PreEDA_Data_Profiling_Report - PDF
- PreEDA_Data_Profiling_Report - HTML

- **Missing Data and Outliers:** Significant missing data in key metrics; outliers in rates suggest potential inaccuracies.
- **Skewness and Correlation:** Data distributions are skewed; high correlations between certain metrics could lead to multicollinearity in modeling.
- **Addressing Data Challenges:** Refinement of 'ALL' Data: Exclude 'ALL' during detailed analyses to avoid diluting subgroup-specific trends.
- **Normalization and Imputation:** Apply transformations to normalize distributions and impute missing data to maintain analysis integrity.
- **Data Integrity Checks:** Implement rigorous checks to ensure data cleaning and transformations preserve accurate information.



Imputation Methods Table with Justification

Column	Null Rate (%)	Imputation Method	Justification
hiv_diagnoses	1.45	Model-Based	Low missing rate, possibly MAR
hiv_diagnosis_rate	1.45	Model-Based	Low missing rate, possibly MAR
concurrent_diagnoses	0.34	Mean/Median	Very low missing rate, likely MCAR
percent_linked_to_care_within_3_months	72.44	Not Applicable	Excessive missing data; dropping. Also,
aids_diagnoses	1.10	Model-Based	Low missing rate, possibly MAR
aids_diagnosis_rate	1.10	Model-Based	Low missing rate, possibly MAR
plwdhi_prevalence	13.58	MICE	Important variable, considerable missing rate
percent_viral_suppression	14.82	MICE	Target variable, requires sophisticated method
death_rate	14.82	MICE	Critical to outcome analysis
hiv_related_death_rate	14.82	MICE	Critical to outcome analysis
non_hiv_related_death_rate	14.82	MICE	Critical to outcome analysis

Note: MCAR = Missing Completely At Random, MAR = Missing At Random, MICE = Multiple Imputation by Chained Equations.

- Post-EDA_Data_Profiling_Report - PDF
- Post-EDA_Data_Profiling_Report - HTML

Post-EDA

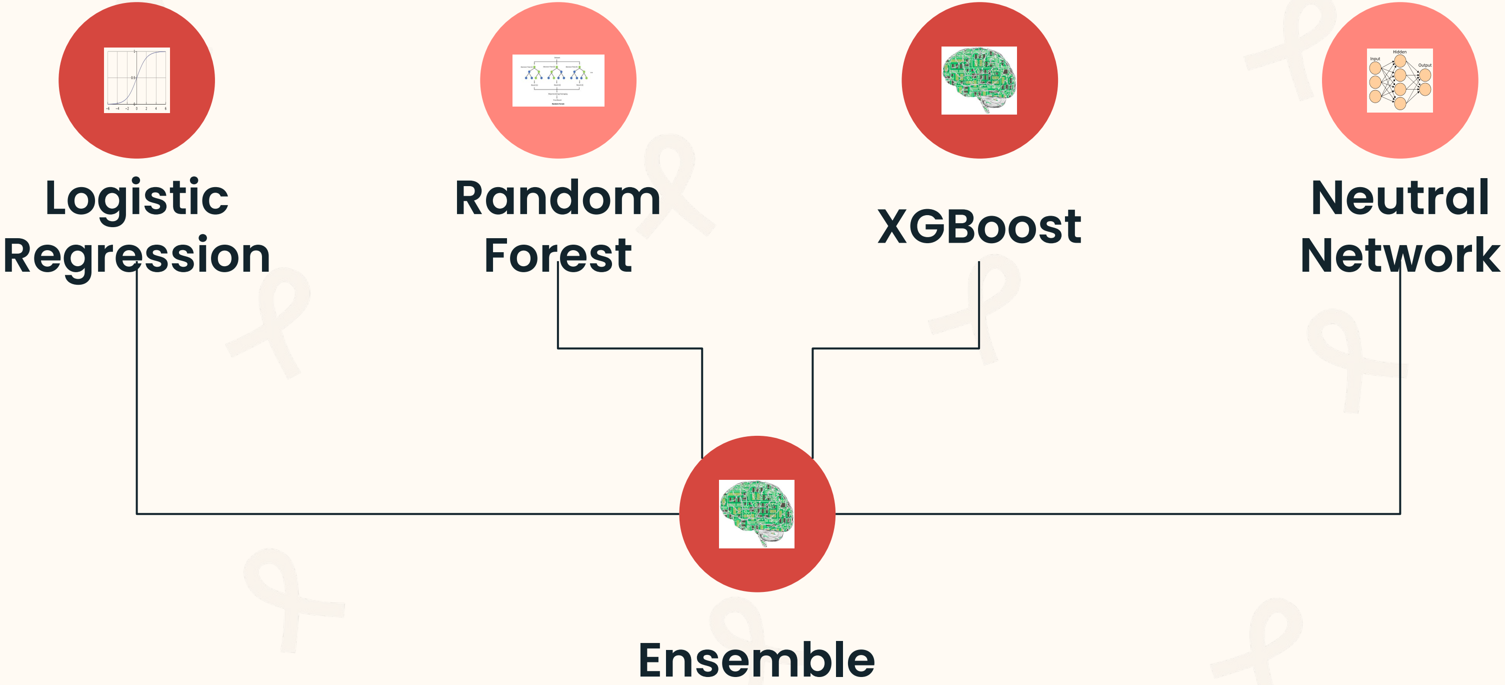
Data Overview:

- **Completeness & Efficiency**: The dataset now features complete data with no missing values across 10,500 observations and 17 variables, stored efficiently at 1.4 MiB.
- **Key Variable Insights:**
  - **Numerical Variables:**
    - **Year Distribution**: Data spans evenly across five years, offering a solid basis for trend analysis.
    - **Health Metrics Correlations:** Notable high correlations are observed among health metrics like 'HIV diagnoses' and 'AIDS diagnoses', which could influence predictive modeling due to multicollinearity.
    - **Zero Values**: A significant presence of zero values in 'HIV diagnoses' and 'deaths' suggests potential underreporting or true no-case scenarios, which are crucial for accurate model interpretation.
  - **Categorical Variables:**
    - **Borough and UHF Correlation:** High correlation between borough and UHF areas suggests redundancy, indicating a potential for dimensionality reduction.
    - **Demographic Representation**: Even distribution across gender, age, and race categories ensures a representative dataset for inclusive analysis.
- **Target Variable:**
  - **Percent Viral Suppression:** Shows a well-distributed range post-imputation, ideal for unbiased modeling.
- **Data Quality Concerns:**
  - **Skewness**: High skewness in death rates may require data transformation to improve model fit and predictions.
  - **High Zero Counts**: Prevalent zeros in key metrics might affect certain modeling techniques and need to be handled carefully.
- **Modeling Recommendations:**
  - **Reduce Multicollinearity**: Employ dimensionality reduction or feature selection techniques to manage high correlations effectively.
  - **Transform Skewed Variables**: Consider logarithmic or similar transformations to normalize data distributions, enhancing model accuracy and robustness.
- **Improving Modeling with 'ALL' Category Handling:**
  - **Refining Granularity**: By excluding or appropriately handling the 'ALL' category in categorical variables, models can achieve greater specificity and avoid biases towards general averages. This allows for more precise predictions and insights at more granular subgroup levels.

04

ML MODELING

MODELING



Logistic Regression

Random Forest

XGBoost

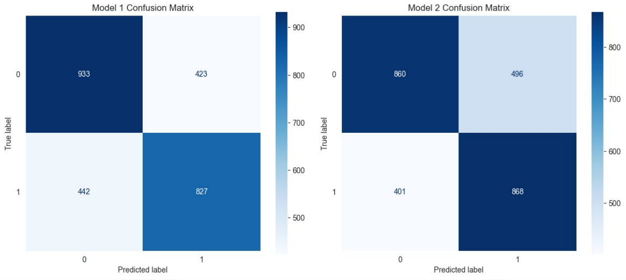
Neutral Network

Ensemble

Logistic Regression

Purpose:

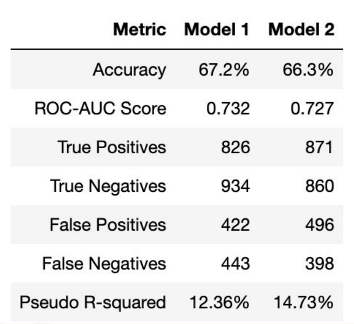
Predict viral suppression among HIV patients using demographic, clinical, and geographic data.



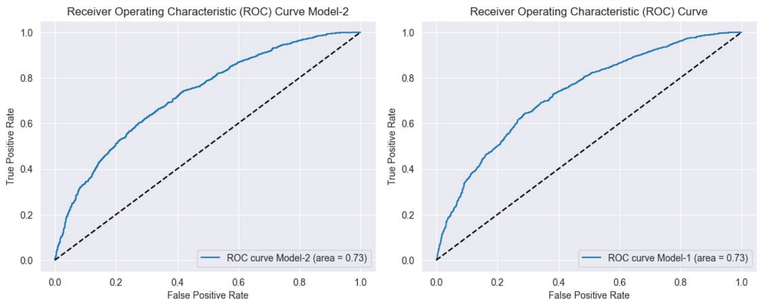
Model Setup:

- **Features:** Age, gender, race, HIV/AIDS diagnoses, borough, and UHF neighborhoods.
- **Target:** Binarized 'percent_viral_suppression' for binary classification.

Performance Matrix:



Metric	Model 1	Model 2
Accuracy	67.2%	66.3%
ROC-AUC Score	0.732	0.727
True Positives	826	871
True Negatives	934	860
False Positives	422	496
False Negatives	443	398
Pseudo R-squared	12.36%	14.73%



Insights & Recommendations:

- **Significant Predictors:** Older age and female gender are more likely to achieve suppression. High diagnosis rates correlate with lower suppression.
- **Geographic Insights:** Model 2 provides detailed local insights, useful for targeted interventions.
- **Public Health Strategy:** Use model insights for targeted interventions and resource allocation, continuously updating models with new data.

Conclusion:

Model 2, with detailed geographic insights, is slightly **more effective** for targeted public health strategies despite lower overall accuracy compared to Model 1. Both models provide valuable insights into factors affecting HIV suppression and can guide public health planning.

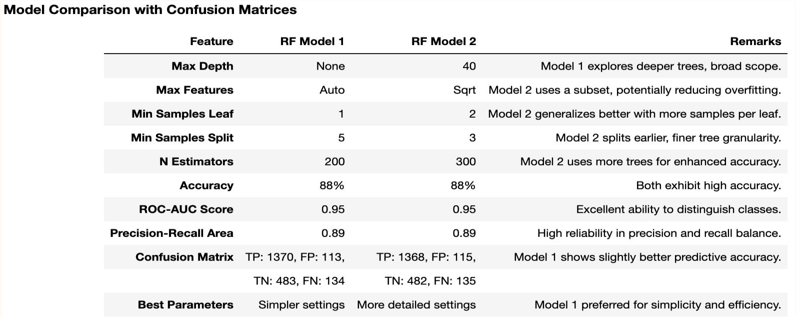
Random Forest

Purpose: Predict the likelihood of viral suppression among HIV patients using a comprehensive set of demographic, clinical, and geographical data.

Model Setup:

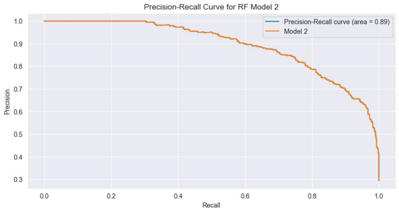
- **Features Used:** Demographic data (age, gender, race), clinical data (HIV/AIDS diagnoses), geographical data (borough, UHF neighborhoods).
- **Target Variable:** Binarized 'percent_viral_suppression' to facilitate binary classification.

Performance Matrix: Training: Models were trained using GridSearchCV, which included cross-validation to ensure generalization and robustness.

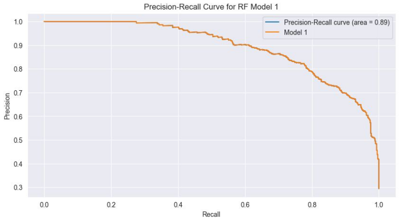


Model Comparison with Confusion Matrices

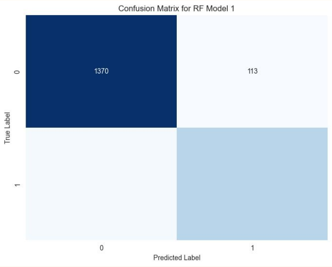
Feature	RF Model 1	RF Model 2	Remarks
Max Depth	None	40	Model 1 explores deeper trees, broad scope.
Max Features	Auto	Sqrt	Model 2 uses a subset, potentially reducing overfitting.
Min Samples Leaf	1	2	Model 2 generalizes better with more samples per leaf.
Min Samples Split	5	3	Model 2 splits earlier, finer tree granularity.
N Estimators	200	300	Model 2 uses more trees for enhanced accuracy.
Accuracy	88%	88%	Both exhibit high accuracy.
ROC-AUC Score	0.95	0.95	Excellent ability to distinguish classes.
Precision-Recall Area	0.89	0.89	High reliability in precision and recall balance.
Confusion Matrix	TP: 1370, FP: 113, TN: 483, FN: 134	TP: 1368, FP: 115, TN: 482, FN: 135	Model 1 shows slightly better predictive accuracy.
Best Parameters	Simpler settings	More detailed settings	Model 1 preferred for simplicity and efficiency.



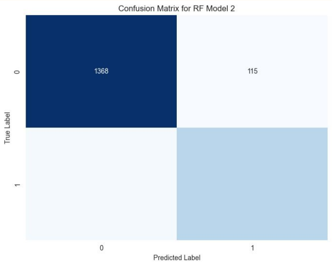
Precision-Recall Curve for RF Model 2



Precision-Recall Curve for RF Model 1



Confusion Matrix for RF Model 1



Confusion Matrix for RF Model 2

Insights & Recommendations:

- **Significant Predictors:** Integration of diverse features allows for exploring the impact of de l suppression.
- **Geographic Insights:** RF Model 2 offers more detailed local insights but with a slight trade-off in certain metrics.
- **Public Health Strategy:** Insights from these models should guide targeted interventions and efficient resource allocation. Continuous updates with new data are recommended to adapt strategies to changing health dynamics.

Conclusion:

- **Best Model:** RF **Model 1** is favored for its balanced precision-recall, simpler configuration, and slightly more accurate confusion matrix results.
- **Application:** The insights derived from Model 1 are vital for precise public health strategies and interventions focused on enhancing HIV viral suppression outcomes.

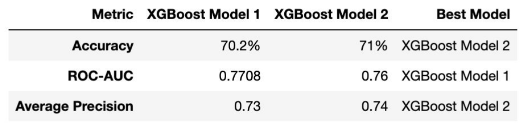
XGBoost

Purpose: Predict viral suppression in HIV patients using a combination of demographic and clinical data with XGBoost, an advanced machine learning technique.

Model Setup:

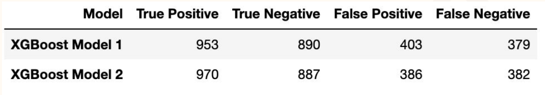
- **Features Used:** Demographic data (age, gender, race), clinical data (HIV/AIDS diagnoses), geographical data (borough, UHF neighborhoods).
- **Target Variable:** Binarized 'percent_viral_suppression' to facilitate binary classification.

Performance Matrix: Training: Models were trained using GridSearchCV, which included cross-validation to ensure generalization and robustness.



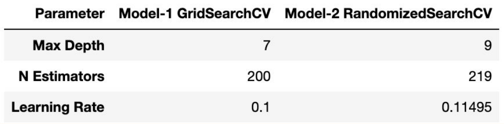
Metric	XGBoost Model 1	XGBoost Model 2	Best Model
Accuracy	70.2%	71%	XGBoost Model 2
ROC-AUC	0.7708	0.76	XGBoost Model 1
Average Precision	0.73	0.74	XGBoost Model 2

Confusion Matrix &

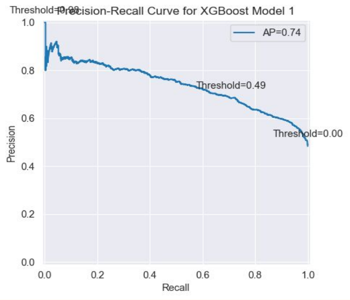


Model	True Positive	True Negative	False Positive	False Negative
XGBoost Model 1	953	890	403	379
XGBoost Model 2	970	887	386	382

Hyperparameter Details:



Parameter	Model-1 GridSearchCV	Model-2 RandomizedSearchCV
Max Depth	7	9
N Estimators	200	219
Learning Rate	0.1	0.11495

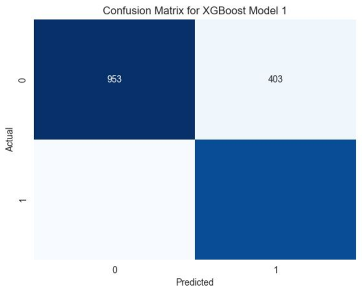


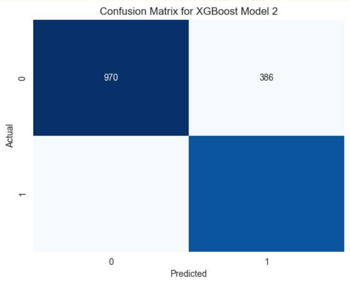
Insights & Recommendations:

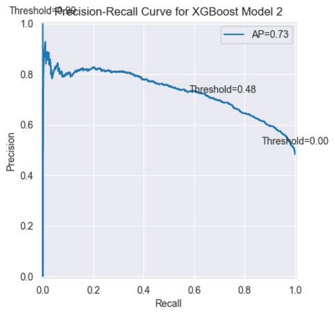
- **Model 2's Advantages:** Slight improvement in minimizing false negatives, crucial for reducing under-treatment scenarios.
- **Precision-Recall Balance:** Model 2 maintains a steadier balance between precision and recall, beneficial for clinical application.
- **Public Health Strategy:** Utilize insights from these models to enhance HIV treatment strategies and resource allocation.

Conclusion:

- **Best Model: Model 2**, with better precision-recall balance and slightly higher accuracy, is recommended for deploying in public health strategies. Both models effectively utilize advanced machine learning to provide deep insights into factors influencing viral suppression, aiding in more informed public health decision-making.







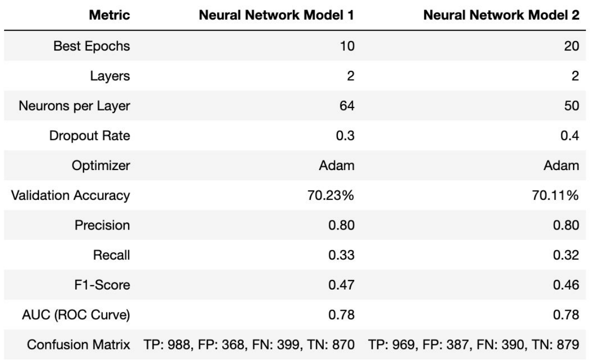
Neural Network

Purpose: To utilize neural network architectures to predict viral suppression in individuals diagnosed with HIV, aiding effective disease management.

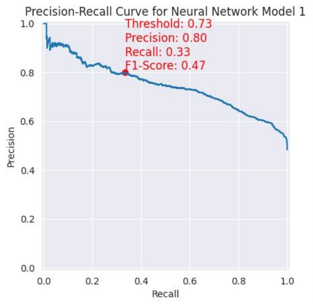
Model Setup:

- **Features Used:** Demographic (borough, gender, age, race) and clinical data (HIV and AIDS diagnosis rates).
- **Target Variable:** Binarized 'percent_viral_suppression' to facilitate binary classification.

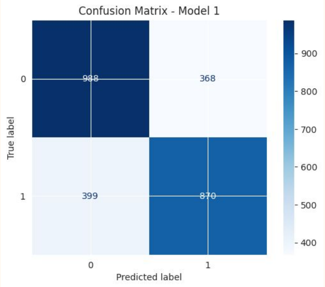
Performance Matrix: Training: Models were trained using GridSearchCV, which included cross-validation to ensure generalization and robustness.



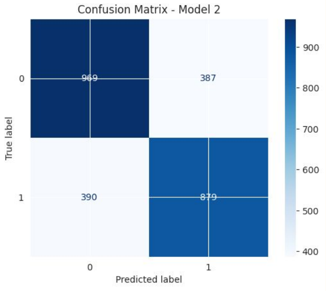
Metric	Neural Network Model 1	Neural Network Model 2
Best Epochs	10	20
Layers	2	2
Neurons per Layer	64	50
Dropout Rate	0.3	0.4
Optimizer	Adam	Adam
Validation Accuracy	70.23%	70.11%
Precision	0.80	0.80
Recall	0.33	0.32
F1-Score	0.47	0.46
AUC (ROC Curve)	0.78	0.78
Confusion Matrix	TP: 988, FP: 368, FN: 399, TN: 870	TP: 969, FP: 387, FN: 390, TN: 879



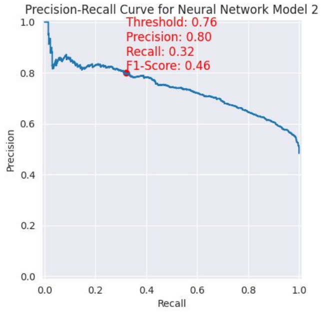
Precision-Recall Curve for Neural Network Model 1



Confusion Matrix - Model 1



Confusion Matrix - Model 2



Precision-Recall Curve for Neural Network Model 2

Insights & Recommendations:

- **Significant Predictors:** Demographic and clinical factors are integral to predicting viral suppression, with borough and age showing particular significance.
- **Model Recommendation:** Model 1 is recommended based on its superior recall and validation accuracy, crucial for medical applications where missing a case of viral suppression can have significant implications.

Conclusion:

- **Best Model:** Neural Network Model 1 is favored for its balanced performance and higher sensitivity in identifying patients likely to achieve viral suppression. This model is instrumental in guiding public health strategies and enhancing treatment outcomes for HIV patients across NYC.

Ensemble Model

Purpose: To enhance predictive accuracy for HIV viral suppression using a combination of machine learning models.

Target Variable: Percent_viral_suppression, binary classified.

Model Composition:

Model Type	Identifier	Description	Hyperparameters	Special Notes
Logistic Regression	LR1	Binary classification with demographic and clinical data.	Solver: 'liblinear', max_iter: 1000	Pipeline with preprocessing.
Logistic Regression	LR2	Similar to LR1, includes SMOTE for balancing.	Solver: 'liblinear', max_iter: 1000	SMOTE for class balance.
Random Forest	RF1	Optimized with GridSearchCV.	Trees: 200, Depth: 20, Features: 'sqrt'	Extensive hyperparameter tuning.
Random Forest	RF2	Variant with different hyperparameters.	Trees: 400, Depth: 40, Features: 'sqrt'	Variant from GridSearchCV.
XGBoost	XGB1	Optimized with GridSearchCV including demographic and clinical features.	Depth: 7, Estimators: 200, Rate: 0.1	Focused on detailed parameter tuning.
XGBoost	XGB2	Similar setup to XGB1, with parameters optimized via RandomizedSearchCV.	Depth: 9, Estimators: 131, Rate: ~0.09	More exploratory parameter search.
Neural Network	NN1	Basic model with one to two layers.	Layers: 1-2, Neurons: 32-64, Dropout: 0.2	Simple but effective architecture.
Neural Network	NN2	More complex with additional layers and neurons.	Layers: 2-3, Neurons: 50-100, Dropout: 0.3-0.4	Extended training duration.

Performance Matrix Table:

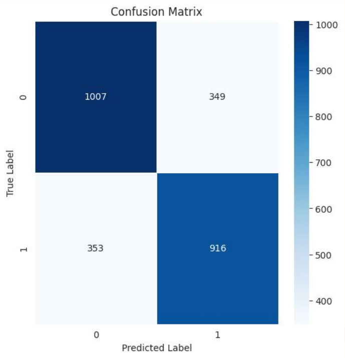
Metric	Value
Accuracy	73.26%
ROC AUC	0.80
Precision	74% (Non-suppressed), 72% (Suppressed)
Recall	74% (Non-suppressed), 72% (Suppressed)
F1-Score	74% (Non-suppressed), 72% (Suppressed)
Confusion Matrix	True Positives: 916, False Positives: 349, False Negatives: 353, True Negatives: 1007

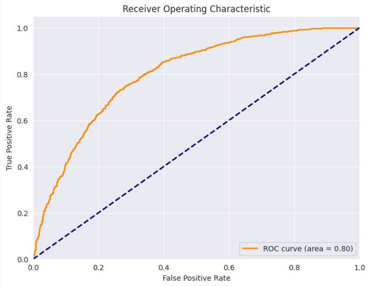
Key Insights:

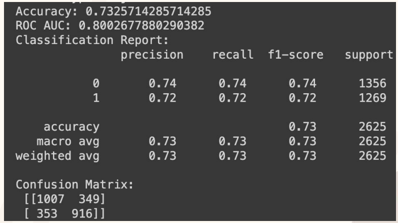
- The Ensemble Model demonstrates strong capabilities in distinguishing between outcomes, crucial for tailoring HIV treatment plans. Balances precision and recall well, minimizing the risk of false negatives, which is vital for patient care in viral suppression scenarios.

Conclusion:

- The Ensemble Model, provides robust predictions essential for strategic health interventions focused on HIV viral suppression, thereby supporting more informed public health decisions.







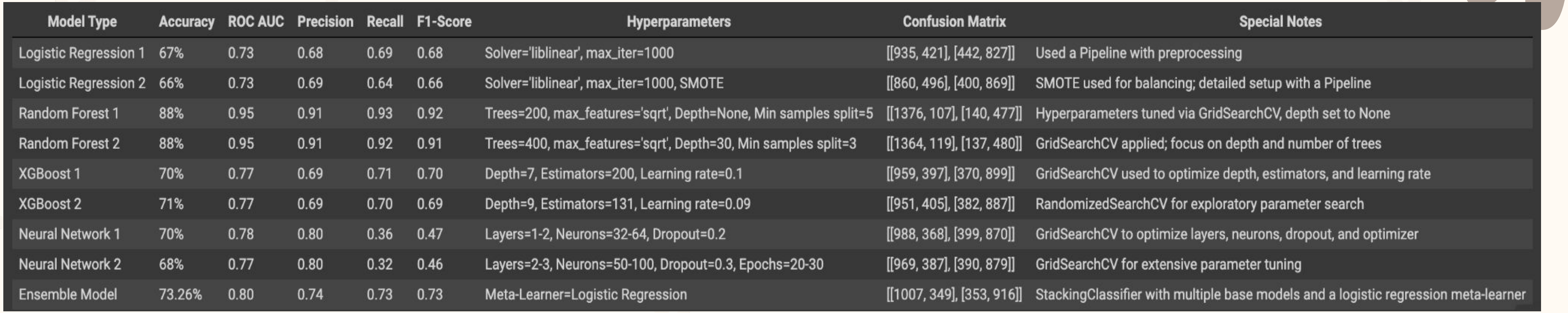
```
Accuracy: 0.7325714285714285
ROC AUC: 0.8002677880290382
Classification Report:
               precision    recall  f1-score   support

           0       0.74      0.74      0.74      1356
           1       0.72      0.72      0.72      1269

    accuracy                           0.73      2625
   macro avg       0.73      0.73      0.73      2625
weighted avg       0.73      0.73      0.73      2625

Confusion Matrix:
 [[1007  349]
 [ 353  916]]
```

Selecting Best Models



Model Type	Accuracy	ROC AUC	Precision	Recall	F1-Score	Hyperparameters	Confusion Matrix	Special Notes
Logistic Regression 1	67%	0.73	0.68	0.69	0.68	Solver='liblinear', max_iter=1000	[[935, 421], [442, 827]]	Used a Pipeline with preprocessing
Logistic Regression 2	66%	0.73	0.69	0.64	0.66	Solver='liblinear', max_iter=1000, SMOTE	[[860, 496], [400, 869]]	SMOTE used for balancing; detailed setup with a Pipeline
Random Forest 1	88%	0.95	0.91	0.93	0.92	Trees=200, max_features='sqrt', Depth=None, Min samples split=5	[[1376, 107], [140, 477]]	Hyperparameters tuned via GridSearchCV, depth set to None
Random Forest 2	88%	0.95	0.91	0.92	0.91	Trees=400, max_features='sqrt', Depth=30, Min samples split=3	[[1364, 119], [137, 480]]	GridSearchCV applied; focus on depth and number of trees
XGBoost 1	70%	0.77	0.69	0.71	0.70	Depth=7, Estimators=200, Learning rate=0.1	[[959, 397], [370, 899]]	GridSearchCV used to optimize depth, estimators, and learning rate
XGBoost 2	71%	0.77	0.69	0.70	0.69	Depth=9, Estimators=131, Learning rate=0.09	[[951, 405], [382, 887]]	RandomizedSearchCV for exploratory parameter search
Neural Network 1	70%	0.78	0.80	0.36	0.47	Layers=1-2, Neurons=32-64, Dropout=0.2	[[988, 368], [399, 870]]	GridSearchCV to optimize layers, neurons, dropout, and optimizer
Neural Network 2	68%	0.77	0.80	0.32	0.46	Layers=2-3, Neurons=50-100, Dropout=0.3, Epochs=20-30	[[969, 387], [390, 879]]	GridSearchCV for extensive parameter tuning
Ensemble Model	73.26%	0.80	0.74	0.73	0.73	Meta-Learner=Logistic Regression	[[1007, 349], [353, 916]]	StackingClassifier with multiple base models and a logistic regression meta-learner

Goal: our goal of accurately predicting viral suppression in HIV patients, the Ensemble Model stands out as the best choice due to several compelling reasons:

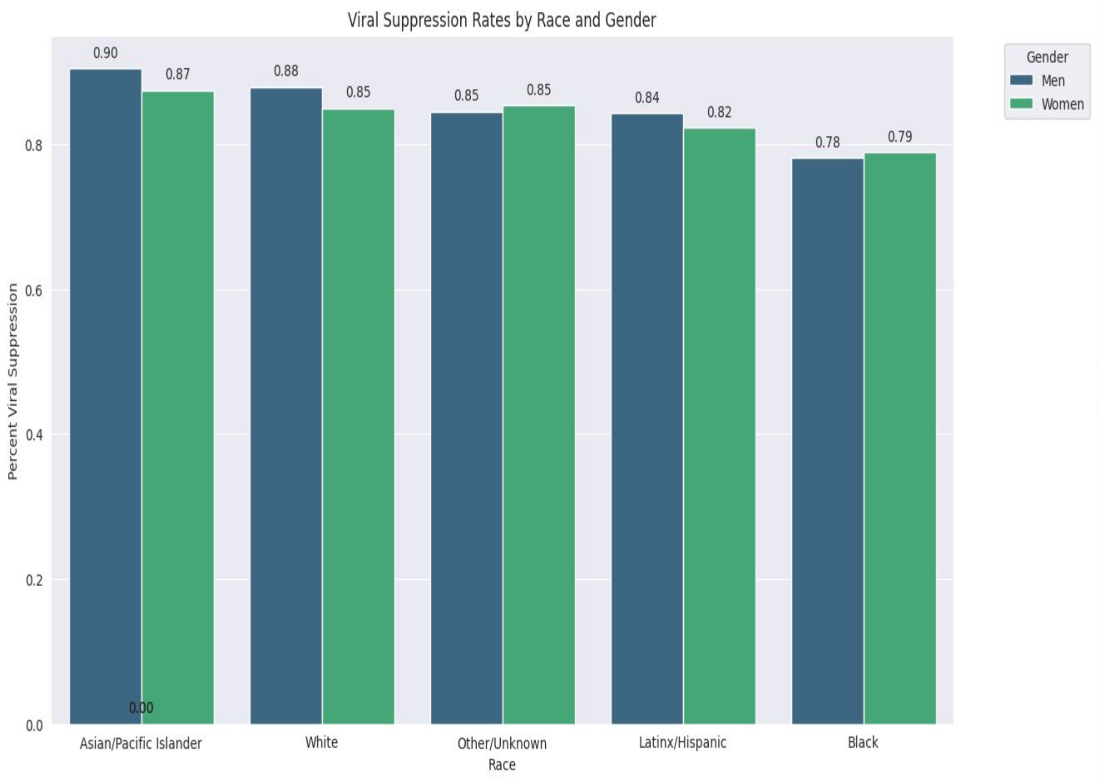
- **Accuracy:** At ***73.26%***, it achieves the highest accuracy among all models, making it highly reliable for making predictions in a healthcare setting.
- **ROC AUC:** With a ***score of 0.80***, it demonstrates superior ability to differentiate between those who will achieve viral suppression and those who will not, which is vital for targeted medical interventions.
- **Balanced Metrics:** It offers a well-rounded performance with a ***precision of 74% and recall of 73%***, ensuring it effectively identifies true cases of suppression without a high rate of false positives or negatives.
- **Comprehensive Approach:** By leveraging the strengths of various base models, the Ensemble Model harnesses a broader spectrum of data patterns, enhancing its predictive accuracy and robustness.
- **Confusion Matrix: *TP: 1007, TN: 916*** Demonstrates the model’s ability to classify both positive and negative cases accurately.

These characteristics make the ***Ensemble Model*** particularly suited to our needs, ensuring that our predictions are not only accurate but also actionable, supporting optimized treatment plans and resource allocation.

Research Question

- Demographic Influence: Which demographics have the highest rate of viral suppression in NYC, and what factors contribute to these outcomes?
- Temporal Trends: Are there significant changes over time in viral suppression rates among different boroughs?
- Geographical Impact: Does the United Hospital Fund neighborhood correlate with the likelihood of achieving viral suppression?
- Age and Gender Effects: How do age and gender impact the rates of viral suppression among those diagnosed with HIV?
- Predictive Modeling: Can machine learning models effectively predict the proportion of individuals likely to achieve viral suppression?

- Demographic Influence: Which demographics have the highest rate of viral suppression in NYC, and what factors contribute to these outcomes?



Summary:

The analysis shows disparities in HIV viral suppression rates among races and genders. Asian/Pacific Islanders and Whites have the highest rates, over 85%, indicating better healthcare access or management. Men usually have higher rates than women, suggesting differences in treatment adherence or access. The lowest rates are observed in Black populations, pointing to underlying socio-economic and healthcare inequalities.

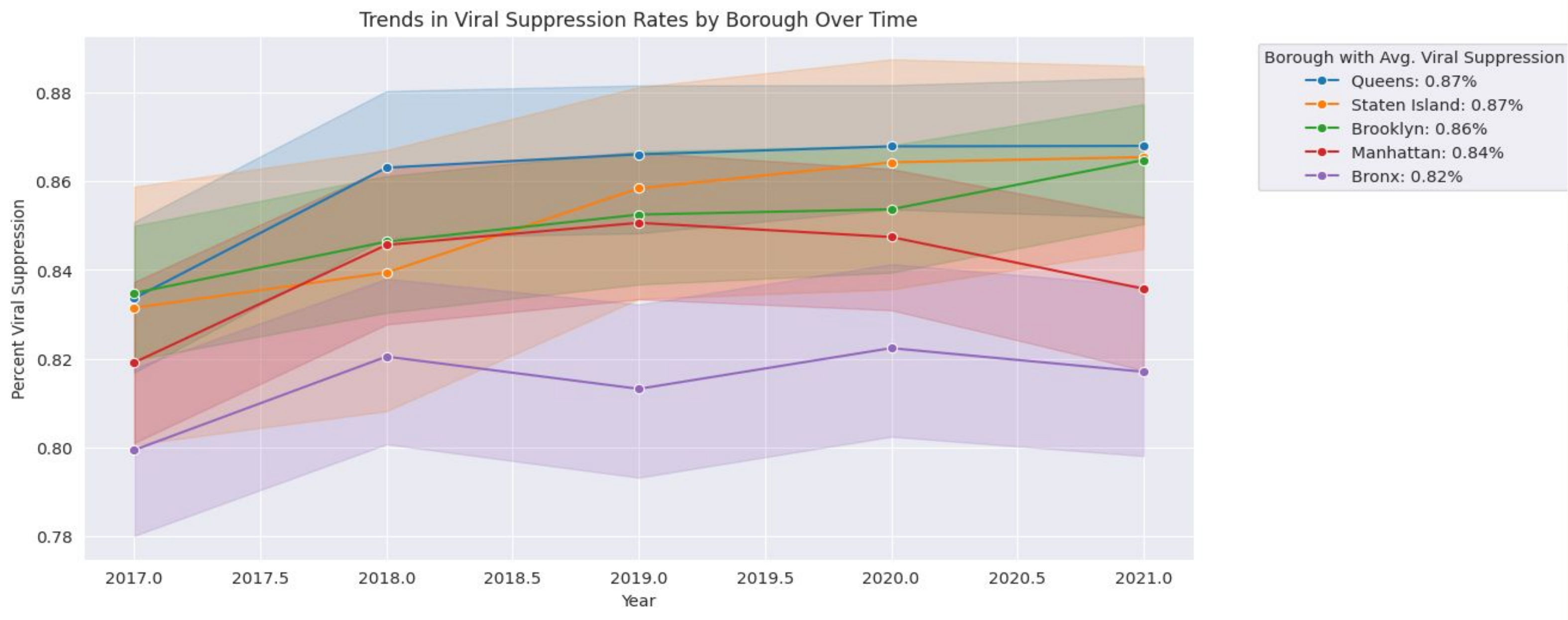
Key Observations:

- Highest rates: Asian/Pacific Islanders and Whites.
- Lowest rates: Black individuals.
- Men typically outperform women in suppression rates.

Insights:

- **Racial Disparities:** There's a clear impact of socio-economic and systemic factors on health outcomes.
- **Gender Differences:** Men's higher rates indicate potential gender-specific healthcare barriers.

- **Temporal Trends:** Are there significant changes over time in viral suppression rates among different boroughs?



Summary:

A time series analysis shows increasing HIV viral suppression rates across NYC boroughs, with Queens and Staten Island leading at over 87%.

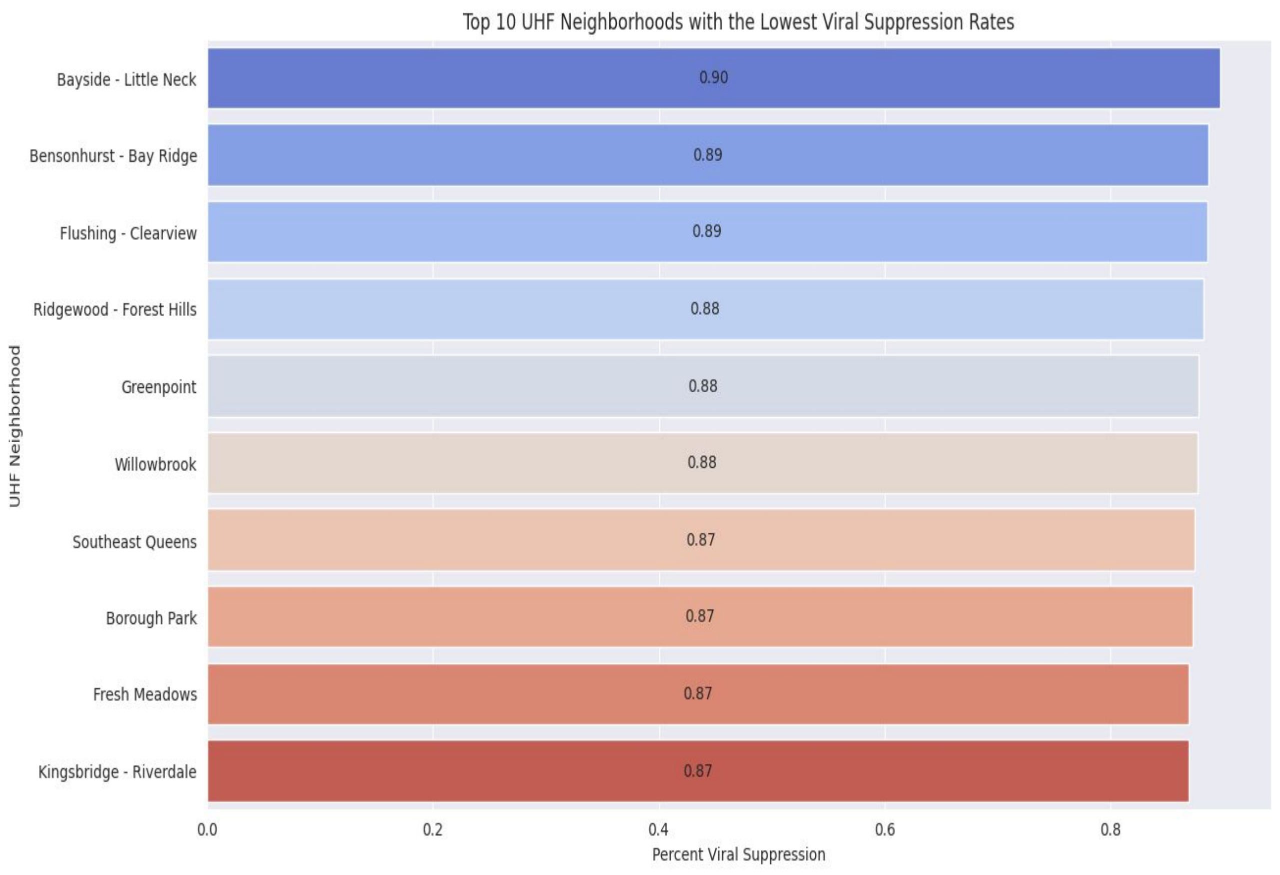
Key Observations:

- All boroughs are improving.
- Highest rates in Queens and Staten Island; Bronx has the lowest.

Insights:

- **Borough Variation:** Rates vary significantly, indicating the need for customized health initiatives.
- **Progress Over Time:** The trend is positive, emphasizing the importance of sustained health efforts.

- Geographical Impact: Does the United Hospital Fund neighborhood correlate with the likelihood of achieving viral suppression?



Summary:

The plot highlights neighborhoods with the lowest viral suppression rates, pinpointing areas like Bayside - Little Neck and Bensonhurst - Bay Ridge for targeted health interventions due to their relatively better performance within this group. This analysis directs where health resources are most needed.

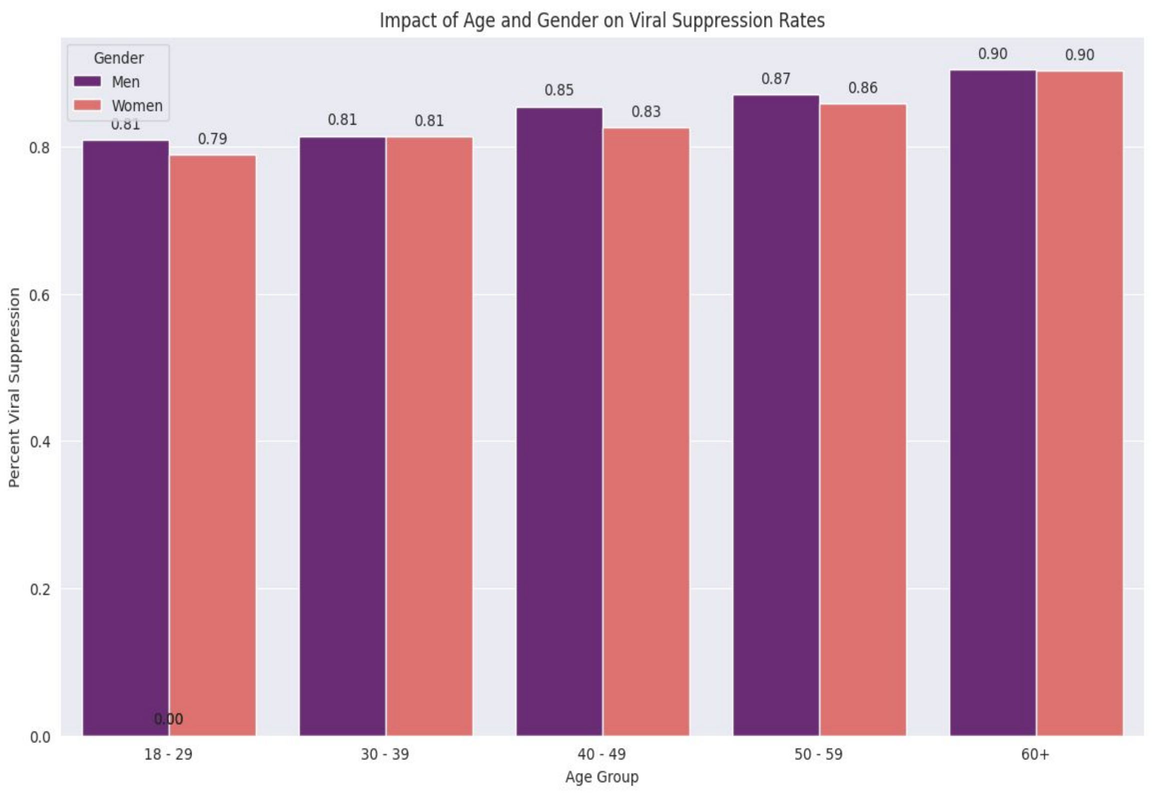
Key Observations:

- Bayside - Little Neck and Bensonhurst - Bay Ridge show better performance among the neighborhoods with the lowest rates.

Insights:

- **Critical Areas for Intervention:** Focusing on neighborhoods with the lowest rates aids in targeted interventions and more efficient resource allocation.

- **Age and Gender Effects:** How do age and gender impact the rates of viral suppression among those diagnosed with HIV?



Summary:
Data shows that viral suppression rates increase with age, peaking near 90% in individuals aged 60 and above, possibly due to better treatment adherence or stable healthcare. Across all age groups, women consistently show slightly lower suppression rates than men, highlighting gender-specific healthcare challenges.

Key Observations:

- Suppression rates rise with age, highest in those 60+.
- Women generally have lower rates than men across all ages.

Insights:

- **Age-Related Compliance:** Older age groups likely have better compliance and more stable healthcare.
- **Gender-Specific Healthcare Needs:** Persistent lower rates among women suggest the need to address gender-specific barriers in healthcare access and adherence.

- **Predictive Modeling:** Can machine learning models effectively predict the proportion of individuals likely to achieve viral suppression?

- **Key Performance Metrics of the Ensemble Model:**
  - **Accuracy: 73.26%** - Indicates high reliability in predictions.
  - **ROC AUC: 0.80** - Demonstrates strong discriminative ability between patient outcomes.
  - **Precision: 74%** - High accuracy in positive predictions.
  - **Recall: 73%** - Effective in identifying actual cases of viral suppression.
  - **Confusion Matrix:**
    - **True Positives**: 1007
    - **False Positives**: 349
    - **False Negatives**: 353
    - **True Negatives**: 916
- **Conclusion:**
  - The Ensemble Model proves highly effective in predicting viral suppression, supported by its accuracy and balanced performance metrics.
  - This model is instrumental for healthcare planning and interventions, helping tailor treatment plans and optimize resource allocation effectively.

CONCLUSION & RECOMMENDATIONS LEARNING

CONCLUSION

- **Model Efficacy:** The Ensemble Model excels in accuracy and ROC AUC, affirming its utility in predicting viral suppression.
- **Predictive Accuracy:** Balances precision and recall well, ensuring dependable forecasts for healthcare use.
- **Data Integration:** Leverages varied data to enhance prediction across different patient groups.
- **Reliable Outcomes:** Offers reliable predictions crucial for effective treatment strategies.

RECOMMENDATIONS

- **Regular Updates:** Continually update the model with fresh data to maintain accuracy.
- **Clinical Deployment:** Implement the model in clinical settings to support treatment decisions.
- **Enhance Capabilities**: Integrate newer algorithms and data for refined predictions.
- **Policy Development**: Use insights to craft policies that improve viral suppression rates.

LEARNING

- **Technique Integration:** Utilizing multiple machine learning techniques boosts predictive performance.
- **Algorithm Strengths:** The ensemble approach harnesses the strengths of various models for robust predictions.
- **Insight Depth:** Provides deep insights into factors affecting viral suppression, enhancing healthcare strategies.
- **Actionable Predictions**: Supports precise health interventions, improving patient outcomes.

THANKS!

Do you have any questions?