Healthcare - Persistency of a Drug

Model Evaluation and Conclusions

Jonathan Allen | Matthew Iversen

Synthetic Minority Over-Sampling Technique (SMOTE)

In our analysis, SMOTE was utilized to address the significant class imbalance present in the dataset. Here's how it was applied:

- **Applied to Predictor Variables:** While SMOTE is typically used to balance the target variable in classification problems, in our unique case, we used SMOTE to address imbalances in key predictor variables (like race, gender, and age group) due to their skewed distribution.

- **Enhancing Model Training:** By creating a more balanced representation of different demographic groups, SMOTE allowed our models to learn from a dataset that more closely resembles a general population.

- **Improving Model Generalization:** The use of SMOTE aimed to improve the robustness and generalization of our models, ensuring they perform well not just for the majority group but across all segments of the population.

Baseline Model

```
Baseline Model Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       258
           1       0.62      1.00      0.77       427

    accuracy                           0.62       685
   macro avg       0.31      0.50      0.38       685
weighted avg       0.39      0.62      0.48       685
```

Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

Logistic Regression Model

Tuned Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.76	0.64	0.70	258
1	0.80	0.88	0.84	427
accuracy			0.79	685
macro avg	0.78	0.76	0.77	685
weighted avg	0.79	0.79	0.79	685

Resampled Tuned Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.69	0.73	0.71	258
1	0.83	0.81	0.82	427
accuracy			0.78	685
macro avg	0.76	0.77	0.76	685
weighted avg	0.78	0.78	0.78	685

Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

Random Forest Regression Model

Tuned Random Forest Classification Report

	precision	recall	f1-score	support
0	0.75	0.66	0.70	258
1	0.81	0.87	0.84	427
accuracy			0.79	685
macro avg	0.78	0.76	0.77	685
weighted avg	0.79	0.79	0.79	685

Resampled Tuned Random Forest Classification Report

	precision	recall	f1-score	support
0	0.66	0.72	0.69	258
1	0.82	0.78	0.80	427
accuracy			0.75	685
macro avg	0.74	0.75	0.74	685
weighted avg	0.76	0.75	0.76	685

Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

XGBoost Model

Tuned XGBoost Classification Report

	precision	recall	f1-score	support
0	0.77	0.67	0.71	258
1	0.81	0.88	0.84	427
accuracy			0.80	685
macro avg	0.79	0.77	0.78	685
weighted avg	0.80	0.80	0.80	685

Resampled Tuned XGBoost Classification Report

	precision	recall	f1-score	support
0	0.69	0.71	0.70	258
1	0.82	0.81	0.82	427
accuracy			0.77	685
macro avg	0.76	0.76	0.76	685
weighted avg	0.77	0.77	0.77	685

Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

Stacking Model

Stacking Model Classification Report

	precision	recall	f1-score	support
0	0.76	0.65	0.70	258
1	0.81	0.88	0.84	427
accuracy			0.79	685
macro avg	0.79	0.76	0.77	685
weighted avg	0.79	0.79	0.79	685

Resampled Stacking Model Classification Report

	precision	recall	f1-score	support
0	0.76	0.65	0.70	258
1	0.81	0.88	0.84	427
accuracy			0.79	685
macro avg	0.79	0.76	0.77	685
weighted avg	0.79	0.79	0.79	685

Pulling from the csv's data descriptions directly, Dexa_Freq_During_Rx is defined as "Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)". Numbers over 100 are highly suspicious as getting this many scans within a year period seems unusual. This would call for a talk with a stakeholder with more experience in the area as it appears to have outliers.

Final Recommendations

After testing different metrics with different data, XGBoost is the best choice for further development.

- **Highest Accuracy:** The Tuned XGBoost model consistently showed the highest accuracy (80%) compared to the other models. This indicates its superior ability to correctly classify instances overall.

- **Balanced Performance:** XGBoost demonstrated a good balance between precision and recall, especially for Class 1 (Persistency_Flag = True), which is crucial in medical predictions. High recall is particularly important to ensure that the model doesn't miss out on true cases.

- **Robust to Imbalance:** XGBoost inherently manages class imbalance well, especially with proper tuning of its parameters. It applies weighting schemes to the classes, making it more sensitive to the minority class.

Conclusion

The Tuned **XGBoost** model stands out as the most suitable choice given its current performance and potential for improvement. However, addressing the underlying issue of data **imbalance** and **diversity** is key to achieving higher accuracy and building a model that is truly reflective and fair for the entire population. Additionally, continual refinement in feature engineering, model tuning, and validation will play a crucial role in enhancing model accuracy and reliability.