# Predictive Modeling for Gym Injury Prevention

## Identifying High-Risk Individuals for Targeted Intervention

Sujith Roy S

28 January 2025

# Overview

Goal:

ACC's Initiative aims to reduce gym-related injuries by offering a free personal training session to its clients.

Challenges:

- Accurately identifying high-risk individuals.
- Effectiveness of a single session.
- Class imbalance.
- Ensuring fairness and equity in the selection process.

# Assessment

1. Timeframe: The data spans from 2006 to 2016.

2. Size: Includes approximately 80,000 rows, representing individuals eligible for the initiative.

3. Key Features:

- Demographics: Includes information like age and socioeconomic status.

- Injury history: Contains details about prior injuries, their nature, and recurrence timelines.

- Lifestyle factors: Includes working conditions and habits related to physical activity.

- Geographic information: Features like Areaunit_score, indicating geographic socioeconomic classifications.

- Missing data: Some fields, such as area unit_score and numeric injury counts (e.g., arm_sprain_all), have missing values that require imputation.

4. Outcome variable: A binary column (y) indicates whether an individual made a gym-related injury claim within 12 months.

# Methodology

-Predictive Modeling: Estimate the probability of injury using factors like injury history, age, socioeconomic status, and lifestyle habits. Individuals are ranked by likelihood of injury.

- Model 1 (Baseline): Random Forest classifier trained on the original dataset without balancing to serve as a performance benchmark.
- Model 2 (Balanced): Random Forest classifier trained on SMOTE-balanced data to improve prediction accuracy for the minority class (injuries).

- Stratified Sampling: Results from Model 2 are subject to sampling where we divide the population into subgroups (e.g., by age, ethnicity) to ensure fair representation and accurate targeting.

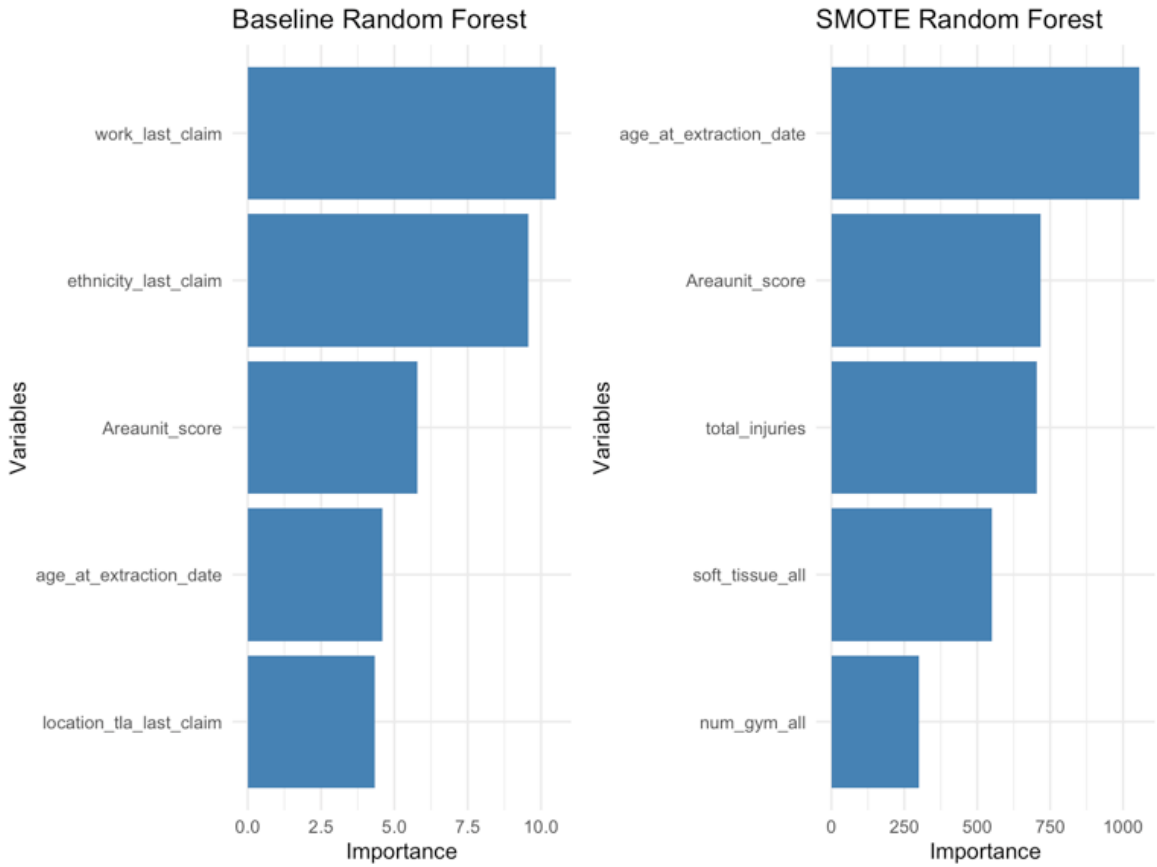
This multi-step approach improves prediction accuracy while ensuring fairness and equity in the targeted intervention.

# Model Summary & Feature Importance

| Metric | Model 1 (Baseline) | Model 2 (SMOTE Applied) |
|---|---|---|
| **AUC (Area Under Curve)** | 0.96 | 0.979 |
| **Out-of-Bag (OOB) Error Rate** | 4.88% | 6.65% |
| **True Negatives (N)** | 75,751 | 79,868 |
| **False Positives (N predicted as Y)** | 247 | 130 |
| **False Negatives (Y predicted as N)** | 3,655 | 7,837 |
| **True Positives (Y)** | 345 | 163 |
| **Class Error for 'N'** | 0.00325 | 0.0019 |
| **Class Error for 'Y'** | 0.91375 | 0.9796 |
| **Key Features** | Weight training injury, Work nature, Ethnicity, Gym injuries, Age, Total injury, Area unit score, Location | Age, Total injuries, Area unit score, Soft tissue injuries, Gym injuries, Knee injuries |
| **Model Performance** | Struggles to predict minority class ('Y') | Better performance for predicting 'Y' after SMOTE, but still biased |

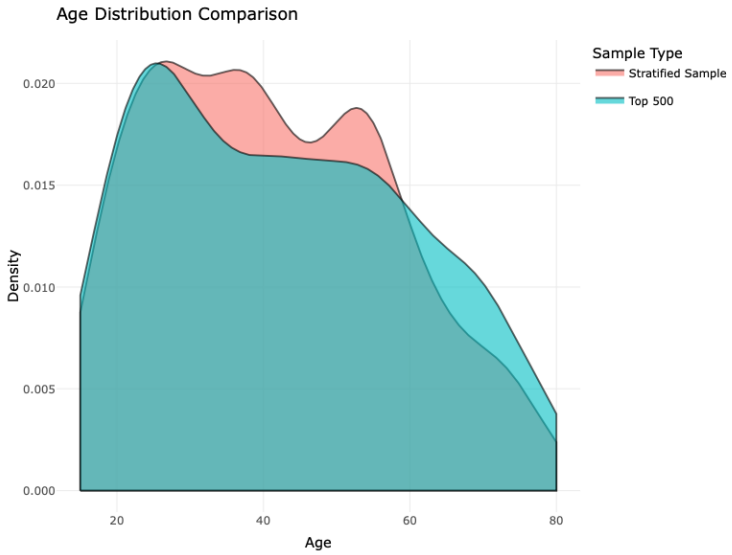
## Key Insights

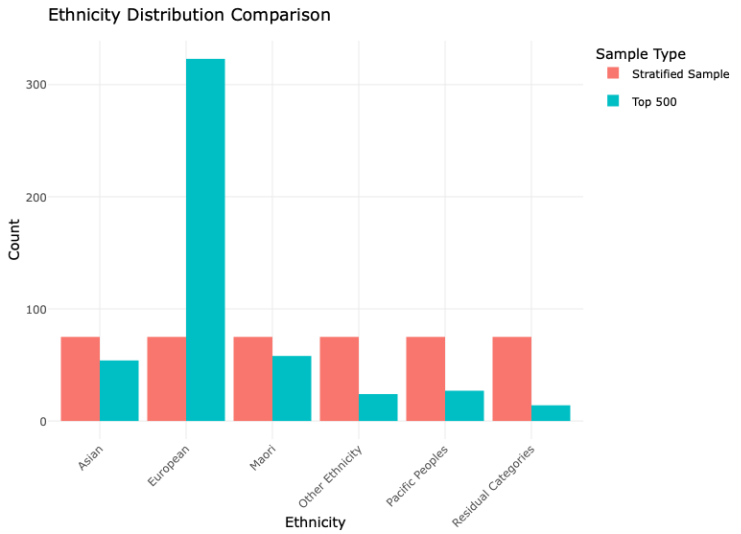
- AUC: Model 2 (SMOTE applied) shows a slight improvement in AUC, indicating better overall discriminatory ability between the two classes.
- Class Errors: Both models perform well for the "No Injury" class (N), but the class error for "Injury" (Y) is high in Model 1, though it improves with SMOTE in Model 2.
- OOB Error Rate: Model 1 has a lower OOB error rate, suggesting it might be more stable, but Model 2 performs better on AUC, indicating that balancing the classes with SMOTE helps improve overall predictive performance.

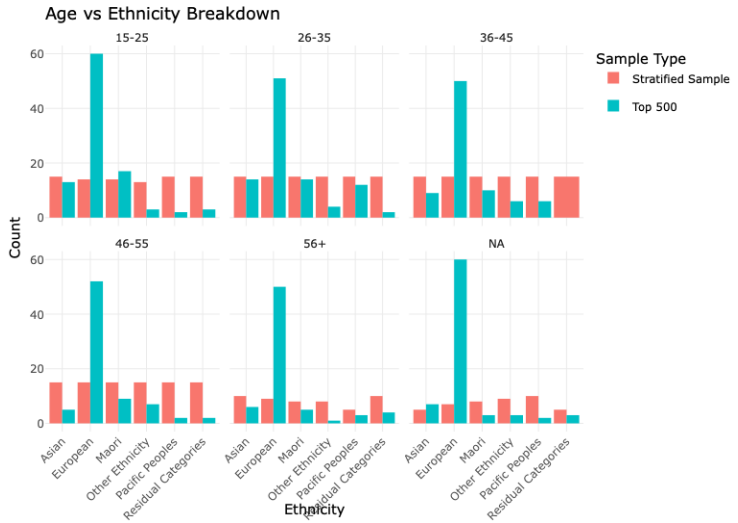


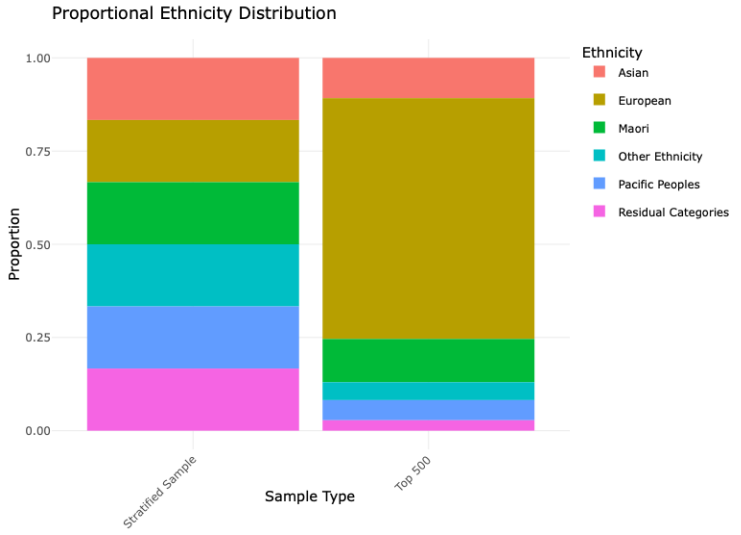

Given that predicting the minority class is a priority for the injury prevention initiative, Model 2 is more suitable despite the slightly higher error rate, as it focuses on improving predictions where it matters most. Further fine-tuning and optimization will likely enhance its ability to accurately predict injuries, making it the more effective model for this task.

# Stratified sampling



# Conclusions

1. Proof of Concept: The current model is a proof of concept and has not yet undergone peer review. While the results are promising, further validation and peer-reviewing is necessary before drawing definitive conclusions.

2. Optimization: The model has yet to be fully optimized for best performance. Further hyperparameter tuning and refinement are necessary to improve accuracy and reduce potential biases. The model relies on Random Forests and SMOTE balancing techniques. Alternative modeling approaches should be explored to ensure the best possible predictive performance.

3. Class Imbalance: Despite addressing the class imbalance using SMOTE, the model's performance for predicting the minority class ('Y' - injury) is still a concern. Further analysis and potentially different balancing techniques should be considered.

4. Monitoring the Initiative: We recommend implementing an automated system to track injury reports among the selected individuals over the next 12 months. This will allow us to measure whether the program is leading to a reduction in injuries and to make adjustments to the intervention if needed. This ongoing evaluation will be essential for refining the initiative and ensuring its long-term success.

5. Additional data: The data provided was only available up until 2016. With a larger and more recent dataset, such as data up to 2024, we will have a much bigger sample size to work with. This expanded data would allow us to split it into training and testing sets, enabling us to evaluate and refine the model more effectively. Additional data such as physical attributes, access to trainers, activity levels, injury causes and injury timeline will add further predictive value.

6. It is also recommended to evaluate the effectiveness of a single session to reduce gym related injuries.