**Introduction**

In an effort to understand the factors influencing member disenrollment, we embarked on a detailed data analysis project using various statistical techniques. Our goal was to identify key characteristics and behaviors that signal a higher likelihood of disenrollment from the healthcare services provided by our client. This report details the methodology, findings, and insights derived from our analysis, aimed at non-technical stakeholders requiring clarity on each step and its significance.

**Analysis Overview**

1. **Random Forest for Feature Importance**
   * **What is Random Forest?** Random Forest is a machine learning technique that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is widely used for classification and regression tasks.
   * **Purpose and Methodology:** We used Random Forest to determine the importance of each feature in predicting membership disenrollment. The algorithm provides a score that helps to identify which attributes are most significant in influencing the decision to leave the healthcare plan.
   * **Findings:** The analysis highlighted several key features such as memberMonthsCount, maxPeriod, and health status indicators (deceased, hospice). However, for further analysis, we focused on features under the client's control, excluding direct health status markers like deceased and hospice.
2. **Logistic Regression Analysis**
   * **Single Variable Analysis:** We employed logistic regression to understand the impact of individual features on the likelihood of disenrollment (activeFlag). This model estimates the probability that a given outcome is 1 for a binary dependent variable.
   * **Multivariate Analysis:** Expanding on the single variable approach, we applied logistic regression to multiple variables to observe their combined effect.
   * **Coefficients and ROC Analysis:** The coefficients from logistic regression indicate the strength and direction of the relationship between each feature and the likelihood of disenrollment. ROC (Receiver Operating Characteristic) curve, and its associated AUC (Area Under the Curve), measures the model's ability to discriminate between the classes effectively.
3. **Comparison of Random Forest and Logistic Regression**
   * **Feature Importance vs. Coefficients:** Random Forest identifies the overall importance of features without direction (positive or negative impact), while logistic regression coefficients provide a direction to the relationship, indicating whether a feature increases or decreases the likelihood of disenrollment.
   * **ROC vs. Accuracy:** ROC AUC is a performance measurement for classification problems at various threshold settings, showing the trade-off between sensitivity (true positive rate) and specificity (false positive rate). Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.

**Detailed Findings on Disenrolled Patients**

* **Analysis Approach:** We conducted separate analyses on the subset of disenrolled patients to identify patterns specific to this group.
* **Chi-Square and Pearson Correlation:** We used Chi-Square tests for categorical variables to understand the association between each feature and disenrollment. Pearson correlation was applied to numerical features to measure the strength of association.
* **Why Chi-Square and Correlation:** Logistic regression was not suitable here due to the uniform outcome (all were disenrolled). Instead, Chi-Square and correlation provided insights into how different features are associated with the disenrollment status.

**Key Metrics Explained**

* **Precision and Recall:** Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall (sensitivity) is the ratio of correctly predicted positive observations to all observations in actual class - yes.
* **F1 Score:** The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
* **Confusion Matrix:** This is a table that is often used to describe the performance of a classification model. It lays out the correct and incorrect predictions made by the model compared to the actual outcomes.

**Variables Impacting Disenrollment**

After refining our analysis and focusing on actionable insights, we identified key variables that significantly impact disenrollment:

* **Member Months Count (memberMonthsCount):** Longer durations of membership correlate with a higher likelihood of staying enrolled.
* **Eligibility Periods:** Various eligibility periods, such as eligible\_3\_12\_Months, eligible\_13\_24\_Months, also influence disenrollment, suggesting targeted retention strategies based on membership duration.
* **Changes in Primary Care Providers (noPcpChange, onePcpChange):** Frequent changes can signal member dissatisfaction.

**Conclusion and Recommendations**

The insights from this analysis should be leveraged to develop targeted interventions aimed at improving member retention. Our findings suggest focusing on enhancing member engagement and satisfaction through stable primary care assignments and monitoring members at critical durations of their membership to preempt potential disenrollment.