**Random Forest for Feature Importance**

Random Forest is an ensemble learning method used for classification and regression, which operates by constructing a multitude of decision trees at training time. For classification tasks like predicting member disenrollment, it outputs the class that is the mode of the classes (classification) of the individual trees. A key strength of Random Forest is its ability to rank the importance of variables in a regression or classification problem in a natural way.

**Methodology and Interpretation:**

1. **Feature Importance Calculation:** Random Forest measures the importance of a feature by looking at how much the tree nodes that use that feature reduce impurity on average across all trees in the forest. It computes this score automatically for each feature during training, providing a straightforward indicator of feature relevance.
2. **Importance of Features in Disenrollment Prediction:**
   * The initial analysis revealed maxPeriod as the most influential feature, which represents the maximum period of membership before disenrollment or censoring. It indicates that the length of time a member stays with the healthcare provider has a significant predictive power on whether they will leave.
   * Other significant features include memberMonthsCount, which directly measures the total number of months a member has been enrolled. This aligns with our understanding that longer membership periods could indicate higher satisfaction or more barriers to leaving.
   * Features related to healthcare service interaction like disenrollDate\_exists and various eligibility periods also emerged as important, suggesting that specific interactions and member status durations play key roles in disenrollment.

**Visual Interpretation from Detailed Feature Importances in Disenrollment Prediction:**

A graph with lines and text

Description automatically generated with medium confidence

* The graph shows that while maxPeriod and memberMonthsCount dominate, understanding nuanced interactions captured by eligibility periods and service interactions (like PCP changes) can provide deeper insights into member retention strategies.

**Updated Selected Feature Importances:**

A graph with purple bars

Description automatically generated

* After refining the focus to actionable variables, memberMonthsCount stands out significantly, suggesting targeted interventions to engage members more effectively during their membership lifecycle could reduce disenrollment rates.

**Discussion on Feature Selection**

While maxPeriod initially appeared as a key feature, on deeper analysis, it represents a similar aspect of member engagement as memberMonthsCount, and thus, we focus on the latter for actionable insights. We exclude direct health status markers like deceased and conditions leading to death (hospice, ESRD) from actionable interventions, as these disenrollments are not typically controllable through member engagement strategies.

**Recommendations Based on Random Forest Insights**

* **Member Engagement:** Enhance engagement strategies focusing on the duration of membership identified by memberMonthsCount. Tailoring services according to the membership lifecycle can potentially enhance member satisfaction and retention.
* **Service Interaction Improvements:** Address issues highlighted by changes in primary care providers and eligibility period transitions, which were marked as significant but less impactful than membership duration. Simplifying processes or improving service during these transitions could mitigate disenrollment risks.

This enhanced understanding directs our client to focus not only on broad metrics like membership length but also on specific service interactions that could be pivotal in improving member retention. By targeting these areas, the healthcare provider can develop more effective strategies to maintain and even grow their member base.

**Understanding ROC AUC for Individual Features**

The Receiver Operating Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers, which is a critical part of understanding feature performance in logistic regression. The Area Under the Curve (AUC) represents the classifier's ability to discriminate between those who will disenroll and those who will not, with a value of 1.0 representing a perfect classifier and 0.5 representing a useless one.

**Interpretation of ROC AUC Results:**

* The ROC AUC values from the analysis show a range of discriminative abilities across various features.
* Features such as memberMonthsCount, which deal with the duration of membership, demonstrate better predictive abilities (moderate AUC scores), suggesting that the length of time a member stays within the healthcare plan is indicative of their likelihood to remain enrolled or disenroll.
* The chart illustrates variability in predictive power, with some features performing close to a random classification (AUC ~0.5) and others showing more promise (AUC >0.5 but less than 0.75), indicating moderate effectiveness.

A graph with blue lines and white text

Description automatically generated

**Impact of Logistic Regression Coefficients**

The coefficients in a logistic regression model indicate the strength and direction of the relationship between each feature and the likelihood of the outcome (in this case, staying enrolled or disenrolling).

**Coefficient Insights:**

* **Positive Coefficients:** Indicate an increase in the likelihood of staying enrolled as the value of the feature increases. For example, features like age showing small positive coefficients suggest that as members age, they might be slightly more inclined to stay.
* **Negative Coefficients:** Suggest that an increase in the feature value is associated with higher disenrollment rates. For instance, institutional having a negative coefficient implies that members who are in institutional settings are more likely to disenroll, contrary to what might be expected. This could be due to various factors such as dissatisfaction with services or better alternatives outside the institutional setting.

**Significant Variables from Logistic Regression Coefficients:**

* The analysis highlights that not all features have a meaningful impact on disenrollment predictions. For example, while disenrolledDueToDeath has a very negative coefficient, indicating a strong relationship with disenrollment, it is a non-actionable insight as these disenrollments are due to mortality rather than modifiable factors within the healthcare provider's control.
* The variable memberMonthsCount shows a high importance in both the ROC AUC and coefficient analysis, reinforcing its critical role in predicting member behavior regarding disenrollment.

**A graph with green lines and white text

Description automatically generated with medium confidence**

**Recommendations Based on Logistic Regression Insights**

* **Enhance Member Retention Strategies:** Focus on length of membership and the specific characteristics such as member engagement and satisfaction levels during their tenure. Initiatives could include improving member services, offering loyalty benefits, or personalizing healthcare plans to increase retention rates.
* **Address Service Gaps in Institutional Settings:** Investigate the underlying causes behind the negative coefficients associated with institutional settings and develop targeted interventions to address these gaps, potentially improving satisfaction and reducing disenrollment rates in these environments.

**1. ROC Curve for Multivariate Logistic Regression**

* **What It Shows**: The ROC curve presented in the graph illustrates the model's ability to distinguish between active and disenrolled members. The area under the curve (AUC) is 0.80, indicating that our logistic regression model has good discriminative power.
* **Interpretation**: An AUC of 0.80 means that our model has an 80% chance of correctly distinguishing between a randomly chosen active member and a randomly chosen disenrolled member. The curve being far from the diagonal (which represents random guessing) shows that our model has predictive value.
* **Key Takeaway**: The ROC curve's shape and AUC score indicate the model's strength in differentiating between outcomes, which is crucial for assessing the model's overall performance.

A graph with a line drawn on it

Description automatically generated

**2. Feature Impact on Disenrollment Prediction (Multivariate Logistic Regression)**

* **What It Shows**: This chart displays the coefficients of various features in our logistic regression model, indicating their impact on the probability of members being active or disenrolled.

A graph with green squares

Description automatically generated

* **Interpretation**:
  + **Positive Coefficients**: Features like memberMonthsCount and eligible\_3\_12\_Months have positive coefficients, suggesting that they increase the likelihood of a member remaining active.
  + **Negative Coefficients**: Features like eligible\_37\_Or\_More\_Months and noPcpChange have negative coefficients, indicating that they increase the probability of disenrollment. Specifically, eligible\_37\_Or\_More\_Months has a strong negative impact, suggesting that members in this category are significantly more likely to disenroll.
* **Why Coefficients Differ in Single vs. Multivariate Analysis**:
  + **Single-Variable Analysis**: When we analyze variables one by one, we only observe the direct impact of each feature on the outcome, without accounting for other features. This can sometimes show a positive relationship if the variable alone appears to increase the active status.
  + **Multivariate Analysis**: Here, we consider the combined effect of all features simultaneously. This can lead to changes in the coefficients' signs because some variables may capture effects already explained by others. For example, memberMonthsCount might appear negative in a single analysis but positive in multivariate analysis when adjusted for related features, such as eligibility duration.

**3. Why Some Coefficients Are Negative or Positive**

* **Example**: eligible\_37\_Or\_More\_Months had a positive coefficient in single-variable analysis but a negative coefficient in multivariate analysis. This is because, when considered alone, the feature might seem to indicate active membership. However, when accounting for other variables like memberMonthsCount, the adjusted relationship reveals that longer eligibility (without interaction with healthcare services) correlates with higher disenrollment risk.
* **Key Insight**: The multivariate model better captures the complexity of real-world behavior by considering interdependencies between features. As a result, some variables' impacts may flip when other related variables are taken into account.

**Model Performance Summary:**

* **Precision, Recall, F1-Score**: These metrics from the classification report show that the model is more accurate in predicting active members (1) compared to predicting disenrolled members (0).
* **Accuracy**: The overall accuracy of 74% indicates that the model correctly predicts the status of members in 74% of cases.
* **Feature Selection**: The features with significant impact on disenrollment include:
  + memberMonthsCount (positive impact on remaining active)
  + eligible\_3\_12\_Months and eligible\_13\_24\_Months (positive impact)
  + eligible\_37\_Or\_More\_Months (negative impact, increasing risk of disenrollment)
  + noPcpChange (negative impact)
  + Changes in PcpChange and prior authorization denials have smaller but notable effects.

**Analysis of Disenrolled Patients**

In the analysis of disenrolled patients, we aimed to identify features that influence the likelihood of members leaving the service. By focusing on features that have a meaningful impact, we sought to derive actionable insights that could help in designing strategies to improve member retention.

**Correlation Analysis for Numerical Features**

The **Correlation Values for Numerical Features** chart helps us understand the strength and direction of relationships between each numerical variable and disenrollment. Key observations include:

* **memberMonthsCount**: A strong positive correlation (~1.0) indicates that as the memberMonthsCount increases, the likelihood of staying active also increases, which makes sense since members who have stayed longer are less likely to disenroll.
* **eligible\_3\_12\_Months**: A negative correlation suggests that members in this eligibility range are more likely to disenroll compared to those with longer eligibility periods.
* **eligible\_37\_Or\_More\_Months**: A positive correlation with active status implies that members eligible for 37 or more months are less likely to disenroll.
* **Age**: A moderate positive correlation indicates older members are slightly more inclined to stay active, potentially due to increased reliance on health benefits.

These correlations guide us in understanding which numerical features to focus on when modeling disenrollment risk.

**Chi-Square Analysis for Categorical Features**

The **Chi-Square Statistics for Categorical Features** chart illustrates the strength of associations between categorical features and disenrollment. Key insights include:

* **noPcpChange and onePcpChange**: High chi-square values suggest these features significantly impact disenrollment. For example, not having a primary care physician (PCP) change is associated with higher stability in membership.
* **establishedCare\_No\_Visit**: A high chi-square value implies that members without recent established care visits are more likely to disenroll.
* **Institutional and Hospice**: These features have the highest chi-square values. However, since they are related to critical conditions or end-of-life scenarios, they may not provide actionable insights for general retention strategies.
* **HealthPlan**: While the chi-square value is lower, variations in health plans could still provide insights into retention strategies.

By focusing on features with high statistical significance, we can better understand the drivers of disenrollment.

A graph with blue squares

Description automatically generatedA graph of a bar

Description automatically generated with medium confidence

**Conclusion from the Analysis**

The comprehensive analysis provided insights into the variables significantly impacting member disenrollment. Using Random Forest, Logistic Regression (both single and multivariate analyses), and additional correlation and chi-square tests, we identified the most critical features driving disenrollment. Here are the key takeaways and actionable insights:

1. **Significant Features Impacting Disenrollment:**
   * **memberMonthsCount**: This feature emerged as a consistent and highly significant predictor, with a strong positive influence on the likelihood of members staying active. The longer a member remains enrolled, the less likely they are to disenroll.
   * **eligible\_3\_12\_Months & eligible\_13\_24\_Months**: These eligibility periods showed a positive impact on remaining active, suggesting that members in these brackets are more engaged or satisfied with their plans.
   * **eligible\_37\_Or\_More\_Months**: This feature had a negative coefficient in the multivariate analysis, indicating a higher risk of disenrollment for members in this eligibility group. It highlights that, over time, certain members may become disengaged or may not find sufficient value in remaining enrolled.
   * **noPcpChange**: This variable negatively influenced active status, indicating that a lack of change in primary care provider (PCP) is associated with a higher risk of disenrollment. This could reflect members not receiving adequate care or engagement.
   * **Changes in PcpChange**: Variables representing PCP changes (e.g., onePcpChange, twoPcpChange, threeOrMorePcpChange) had smaller but notable effects, suggesting that service interaction and continuity of care are important in retaining members.
   * **Prior Authorization Denials**: Features like noPriorAuthDenial and other denial categories showed some impact on disenrollment, hinting that administrative hurdles may influence members' decisions to leave.
2. **Key Insights and Recommendations:**
   * **Focus on Membership Duration**: Since memberMonthsCount has the strongest impact, developing strategies to enhance member satisfaction over time is crucial. Consider loyalty programs, personalized engagement, or incentives to encourage long-term membership.
   * **Address Eligibility Concerns**: Members in the eligible\_37\_Or\_More\_Months category require targeted interventions. This may include personalized communication, better benefits tailored to their needs, or strategies to re-engage and provide additional value.
   * **Improve PCP Engagement**: Analyze and optimize PCP interactions, as stability or dissatisfaction in PCP relationships can drive disenrollment. Implement measures to improve the quality of care and member experience with healthcare providers.
   * **Review Prior Authorization Policies**: Simplifying or making prior authorization processes more member-friendly could help reduce dissatisfaction and potential disenrollment.
3. **Why Coefficient Signs Change**:
   * The differences between single and multivariate analyses, such as the change from a positive to a negative coefficient for eligible\_37\_Or\_More\_Months, highlight the complexities of interactions between features. Multivariate analysis accounts for the relationships between variables, revealing a more nuanced understanding of the factors affecting disenrollment.