**Explanation of the Chi-Square Test and Its Application to Disenrollment by Plan**

**What is a Chi-Square Test?**

* A **Chi-Square test** is a statistical method used to determine if there is a significant association between two categorical variables. It assesses whether the observed frequencies (counts) in each category match the expected frequencies if there were no association between the variables.

**P-Value in Chi-Square Test**:

* A p-value is the probability that the observed association occurred by random chance. In the context of the Chi-Square test:
  + **Low p-value (e.g., <0.05)**: Indicates a significant association, meaning the variable (e.g., health plan) is likely related to disenrollment.
  + **High p-value**: Suggests no significant relationship between the two variables.

**Application in This Analysis**:

* In our case, we used the Chi-Square test to determine if **disenrollment rates** varied significantly across different **health plans**. Specifically:
  1. We created a contingency table showing the count of active vs. disenrolled members for each health plan.
  2. The Chi-Square test compared the observed disenrollment rates across plans to the rates we would expect if there were no difference between plans.
  3. A low p-value (typically <0.05) would indicate a significant difference, suggesting that certain plans have different disenrollment rates compared to others.

**How the Data Was Analyzed**

To provide a clear view of the analysis steps, this section outlines each step from data preprocessing through advanced statistical analysis and modeling.

**Step 1: Data Cleaning and Preprocessing**

* **Missing Values**: We began by identifying and addressing missing values to ensure data quality.
* **Encoding Categorical Variables**: Categorical variables, such as healthPlan, were converted into numerical formats for compatibility with the logistic regression model.
* **Scaling Numerical Variables**: Key numerical features, like memberMonthsCount, were standardized to ensure consistency in the analysis.

**Step 2: Exploratory Data Analysis (EDA)**

* **Distribution Analysis**: We analyzed distributions of key variables such as age, memberMonthsCount, and activeFlag to understand baseline characteristics.
* **Descriptive Statistics**: Basic summary statistics were calculated for each variable, helping us understand the range, mean, median, and variability of the data.

**Step 3: Chi-Square Test for Categorical Variables**

* The Chi-Square test was applied to determine if categorical variables (such as healthPlan) were significantly associated with disenrollment. A low p-value indicated significant associations, highlighting variables that may influence member retention.

**Step 4: Correlation Analysis**

* We calculated correlations between continuous variables (e.g., memberMonthsCount, age) and activeFlag to identify potential relationships. This step helped in selecting variables that could be relevant in predicting disenrollment.

**Step 5: Logistic Regression Modeling**

* **Single Variable Models**: We initially tested individual variables in logistic regression models to assess their individual impact on disenrollment.
* **Multivariate Model**: We then combined selected variables (e.g., memberMonthsCount, healthPlan, hospice, esrd, age) into a multivariate logistic regression model. This approach allowed us to evaluate the combined influence of multiple factors.

**Step 6: Model Evaluation with ROC AUC**

* To measure the effectiveness of the multivariate logistic regression model, we calculated the ROC AUC score. The ROC AUC score, around 0.78, indicates the model's ability to distinguish between active and disenrolled members.

**Step 7: Visualization and Interpretation**

* **Feature Impact**: A bar chart of logistic regression coefficients visualizes each variable’s influence on disenrollment likelihood.
* **ROC Curve**: The ROC curve illustrates the trade-off between sensitivity and specificity, showing the model’s classification accuracy.

1. **Feature Selection for Multivariate Logistic Regression**
   * **First Model**: Included memberMonthsCount, healthPlan, noPcpChange, and onePriorAuthDenial.
     + These variables were selected based on initial exploratory data analysis, where they showed potential associations with disenrollment.
     + The rationale was to observe the impact of member tenure, insurance plan type, frequency of PCP changes, and prior authorization denials on disenrollment.
   * **Second Model**: Focused on memberMonthsCount, healthPlan, age, hospice, and esrd.
     + In response to your question about including disease- and insurance-related variables, this model incorporated health conditions (e.g., hospice and esrd) and member age, which are factors closely related to healthcare usage and cost.
2. **Evaluation Metrics and Interpretation**
   * Both models were evaluated using metrics such as accuracy, ROC AUC, and logistic regression coefficients.
   * The ROC AUC score provides a measure of the model's ability to distinguish between active and disenrolled members.

**Model 1: Analysis with memberMonthsCount, healthPlan, noPcpChange, and onePriorAuthDenial**

* **Results**:
  + **Accuracy**: 71%
  + **ROC AUC**: 0.756
  + **Coefficient Analysis**:
    - memberMonthsCount (0.86): Positive impact; longer membership duration is associated with a higher likelihood of staying active.
    - healthPlan (0.06): Low impact; minimal influence from plan type alone.
    - noPcpChange (-0.81): Negative impact; members with no PCP changes are slightly more likely to disenroll.
    - onePriorAuthDenial (0.43): Positive impact; prior authorization denials are associated with a higher likelihood of disenrollment.

**A graph showing a graph with several squares

Description automatically generated with medium confidence**

**Interpretation**:

* This model provides a modest accuracy of 71% and an ROC AUC of 0.756, indicating that these variables capture some level of predictive value, though additional factors may improve the model's discriminative ability.
* The coefficient values show that membership duration is a strong predictor of staying active, while prior authorizations are a concern for disenrollment, albeit with lower significance than health conditions.

**Model 2: Analysis with memberMonthsCount, healthPlan, age, hospice, and esrd**

* **Results**:
  + **Accuracy**: 74%
  + **ROC AUC**: 0.778
  + **Coefficient Analysis**:
    - memberMonthsCount (0.94): Strong positive impact; longer membership duration correlates with active status.
    - hospice (-3.74): Strong negative impact; members receiving hospice care are much more likely to disenroll.
    - esrd (-0.99): Negative impact; ESRD patients show a higher likelihood of disenrollment.
    - healthPlan (0.05) and age (-0.018): Minimal impact; limited influence from plan type and age alone.

A graph with blue squares

Description automatically generated with medium confidence

**Interpretation**:

* The second model achieved a slightly higher accuracy and ROC AUC score, with an accuracy of 74% and an ROC AUC of 0.778. This suggests that including disease-related factors like hospice and ESRD improves the model’s ability to predict disenrollment.
* The coefficients highlight the significance of health-related variables. hospice and esrd have strong negative associations with staying active, indicating that members with severe or terminal health conditions are more likely to disenroll.

**Insights and Recommendations**

**1. Target Members in Specific Health Plans**:

* Since certain health plans are associated with higher disenrollment rates, these plans may benefit from tailored member engagement strategies.

**2. Prioritize Health-Related Support**:

* Members with conditions like ESRD or in hospice care have higher disenrollment rates. Enhanced support for these groups could improve retention.

**3. Engage New Members Early**:

* Members with shorter membership durations are more likely to disenroll, indicating that early-stage engagement strategies may help increase long-term retention.

**4. Address Authorization Denials**:

* Members facing authorization denials are more likely to disenroll, suggesting a need for streamlining authorization processes to improve the member experience.