Predicting Member Engagement in Preventive Care in LPPO Plans: A Data-Driven Approach to Improving Health and Stars Ratings

Humana-Mays Healthcare Analytics
2024 Case Competition

Humana

TEXAS A&M UNIVERSITY
Mays Business School

Table of Contents

1. Executive Summary

Humana's Local Preferred Provider Organization (LPPO) plans face a significant challenge with members underutilizing preventive care services compared to Health Maintenance Organization (HMO) plans. This report presents a comprehensive analysis of this issue and proposes data-driven strategies to improve preventive care utilization among LPPO members.

Key Findings:

1. Complex Factors: Preventive care utilization is influenced by a combination of demographic, geographic, health status, and plan-related factors.
2. High-Risk Clusters: Three distinct clusters were identified within the high-risk population, each requiring tailored intervention strategies.
3. Digital Engagement: Members who regularly interact with online services show higher preventive care utilization.
4. Veteran-Specific Challenges: A significant portion of high-risk members are veterans with unique healthcare utilization patterns.
5. Cost Considerations: Both high and low-cost members are at risk of missing preventive visits.

Methodology: We employed advanced machine learning techniques, including CatBoost modeling and SHAP (SHapley Additive exPlanations) analysis, to identify key factors influencing preventive care gaps and to develop targeted strategies.

Proposed Strategies:

1. Preventive Care Rebate Program: Targeting high-cost, non-veteran members with financial incentives to encourage preventive care visits.
   - Estimated ROI: 255.47%
   - Potential Impact: 5,070 additional members using preventive care
2. Veteran-Centric Care Coordination Program: A peer-led outreach program to improve preventive healthcare engagement among veterans.
   - Estimated ROI: 405.3%
   - Potential Impact: 856 additional veterans using preventive care

Additional Recommendations:

- Implement a digital engagement strategy to increase online interaction and awareness of health services.
- Provide at-home healthcare preventive checks for members with disabilities.

- Explore network effects on healthcare behaviors by collecting and analyzing peer influence data.

Expected Outcomes: Implementation of these strategies is expected to lead to improved health outcomes for LPPO members, enhanced CMS Star Ratings, more accurate risk adjustment, and a strengthened competitive position in the growing Medicare Advantage LPPO market.

This data-driven approach not only addresses the immediate challenge of preventive care utilization but also establishes a framework for ongoing refinement of member engagement strategies in Humana's LPPO plans.

2. Case Background

2.1 The Importance of Preventive Care in Medicare Advantage Plans

Preventive care, encompassing regular check-ups, screenings, and early interventions, plays a crucial role in maintaining public health and controlling healthcare costs. This is particularly critical for individuals who often face increased health risks due to age and chronic conditions. According to the Centers for Disease Control and Prevention (CDC), ninety percent of the nation's $4.5 trillion in annual health care expenditures are for people with chronic and mental health conditions[1]. Interventions to prevent and manage these diseases can have significant health and economic benefits.

The benefits of preventive care extend to multiple stakeholders in the healthcare system. For patients, preventive care offers early detection of diseases, better health outcomes, lower out-of-pocket costs in the long term, and improved quality of life. Patients who engage in regular preventive care are more likely to detect health issues early, leading to more effective treatments and better long-term outcomes. This proactive approach not only improves their quality of life but also reduces the financial burden of treating advanced diseases.

Healthcare providers also benefit significantly from regular preventive care visits. These appointments provide opportunities for early intervention, better management of chronic conditions, and reduced hospitalization rates. Moreover, they foster improved patient-provider relationships, enabling more personalized and comprehensive care. Regular preventive visits allow healthcare professionals to develop a more holistic understanding of their patients' health, leading to more effective and tailored treatment plans.

From a payer's perspective, such as Humana, preventive care is a key factor in controlling healthcare costs and improving plan performance. It leads to lower overall healthcare costs, improved Centers for Medicare & Medicaid Services (CMS) Star Ratings, more accurate risk adjustment and appropriate funding, and enhanced member satisfaction and retention. These factors are crucial for Medicare Advantage plans, as they directly impact the plan's competitiveness, financial performance, and ability to provide high-quality care to members.

Despite these clear benefits, Humana faces a unique challenge with preventive care engagement in its growing Local Preferred Provider Organization (LPPO) plans. LPPO plans have a higher percentage of unengaged members compared to Health Maintenance Organizations (HMOs). While LPPO plans offer members greater flexibility in choosing healthcare providers, this flexibility comes with the drawback of potentially reduced engagement in preventive care services. This disengagement poses a critical challenge for Humana in ensuring optimal health outcomes for its members and maintaining competitive performance in the Medicare Advantage market.

[1] https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html, accessed Oct 20, 2024

Addressing this challenge is crucial for Humana's success. By developing innovative solutions to enhance preventive care utilization in LPPO plans while maintaining the flexibility members value, Humana can strengthen its position in the Medicare Advantage market, improve member health outcomes, and optimize its operational efficiency. The key lies in finding effective strategies to engage LPPO members in preventive care, thereby realizing the full potential of these plans for both members and the company. This balance between flexibility and engagement will be essential in navigating the evolving landscape of Medicare Advantage plans and ensuring the best possible outcomes for all stakeholders involved.

2.2 Business Problem

Humana's Local Preferred Provider Organization (LPPO) plans face a significant challenge: members are underutilizing preventive care services compared to those in Health Maintenance Organization (HMO) plans. This disparity is particularly concerning as Humana's LPPO market share expands, potentially compromising member health outcomes and Humana's business performance.

Our analysis addresses this issue through a three-step approach:

1. Develop a predictive model to accurately identify LPPO members likely to miss preventive care appointments.
2. Analyze the model to determine key factors contributing to low preventive care engagement among LPPO members.
3. Formulate data-driven, actionable strategies to boost preventive care utilization in LPPO plans based on these insights.

By implementing this approach, we aim to:

- Enhance health outcomes for LPPO members
- Improve Humana's performance in the CMS Stars program
- Facilitate more accurate risk documentation
- Bolster Humana's competitive position in the growing Medicare Advantage LPPO market

2.3 Key Performance Indicators for Step 1

To assess the effectiveness of our predictive model, we will use the following metrics:

A. AUC Score (Area Under the Curve):

The AUC score measures our model's ability to distinguish between members who will and won't attend preventive care visits. It ranges from 0 to 1, where 1 indicates perfect prediction and 0.5 represents random guessing. A higher AUC score suggests better model performance.

B. Classification Report:

This report will provide a comprehensive view of our model's performance, including:

1. Accuracy: The overall proportion of correct predictions made by our model.
2. Class-specific metrics:
   - Precision: The proportion of correct positive predictions for each class.
   - Recall: The proportion of actual positive cases correctly identified for each class.
   - F1-Score: A balanced measure between precision and recall for each class.
   - Support: The number of instances for each class in our dataset.
3. Averaged metrics:
   - Macro Average: Unweighted mean of the metrics, treating all classes equally.
   - Weighted Average: Mean of the metrics weighted by the number of instances in each class.

These metrics will help us understand how effectively our model distinguishes between attendees and non-attendees of preventive care visits. Additionally, they will reveal whether our model performs consistently across both classes and shows if there is any bias towards one class. Ultimately, they will give us a comprehensive view of the overall reliability of our predictions, ensuring that we have a solid foundation for our subsequent analysis and strategy development.

3. Comprehensive Data Processing

We have designed a robust data pipeline for identifying the missing preventive healthcare individuals in the LPPO plan. There were multiple large datasets, ranging from 1.5 million to 33 million rows of data from patient level to claim level. Our approach ensures data integrity while preserving crucial information necessary for analysis.

3.1 Data Loading and Initial Exploration

We started the process by importing 11 different member-level datasets. Each of these datasets contained 1,527,904 rows. The data ranges from target members to additional features, control points, cost and utilization, demographics, pharmacy utilization, sales channels, social determinants of health, web activity, and member details. The number of columns varies within these datasets; all were included to provide a full view of each member profile and healthcare interaction.

3.2 Member Level Data Merging and Initial Cleaning

3.2.1 Data Merging

To create a unified dataset, we merged the 11 member-level datasets based on the common identifier *id*. We used an inner join to ensure data consistency across all datasets, resulting in a base dataset of 1,527,904 rows and 248 columns. This approach guaranteed that all rows were utilized in the merged dataset, providing a complete picture of each member. We called this merged dataset base member data.

3.2.2 Data Type Conversions

To optimize our data for analysis, we performed several data type conversions. Boolean variables such as *disabled_ind*, *dual_eligible_ind*, *lis_ind*, and *veteran_ind* were converted from 'Y'/'N' to True/False. Categorical variables like *sex_cd*, *tenure_band*, *pbp_segment_id plan_benefit_package_id*, and *mco_contract_nbr* were converted to the Categorical type.

3.2.3 Handling Missing Values

We took a nuanced approach to handling missing values. For categorical columns, we filled missing values with 'Unknown' to retain these records in our analysis. For numeric columns, we left the missing values as-is, allowing our machine learning model to handle them appropriately during the modeling phase.

3.3 Feature Engineering and Data Integration for Other Datasets

3.3.1 Member Conditions Data Processing

The Member Conditions dataset, containing over 4 million rows, required special attention due to its structure of multiple rows per member, each representing a different condition. We created unique codes (CD1 to CD129) for 129 condition descriptions to standardize the data. We then performed one-hot encoding for both condition keys and condition description codes.

To aggregate this data to the member level, we created binary indicators for each condition key (COND_KEY) and condition descriptions codes (cond_desc_code)[2], as well as for CMS model versions (V24 and V28). For members without records in this dataset, we assigned 'Unknown' to the Chronicity field and 0 to the condition columns, assuming these members don't have known chronic disease.

3.3.2 Member Visit Claims Processing

The Member Visit Claims dataset, with 19 million rows, provided detailed information on member healthcare interactions. We aggregated these claims at the member level and created a pivot table with unique member ids as rows and 552 columns representing 23 visit types across 24 months (January 2021 to December 2022). Each cell in this pivot table contained the sum of visits for that member, visit type, and month, allowing us to maintain granularity while making the data manageable for analysis.

3.3.3 Quality Data Processing

The Quality Data, our largest dataset with 33 million rows, required extensive processing. We created a mapping dictionary to convert measure descriptions to abbreviated names and aggregated the data by measure and year for each member. We created columns for each measure-year combination (2020, 2021, 2022) and summed the *compliant_cnt* for each member across measures and years[3]. This approach allowed us to capture the quality measures effectively while reducing the dataset's complexity.

3.4. Data Integration and Final Preparation

The final step in our data processing pipeline involved merging all the processed datasets: the base member data, aggregated Member Conditions data, aggregated Visit Claims data, and aggregated Quality data. This integration resulted in a comprehensive final dataset with 1,527,904 rows (one per member) and 1073 columns, including the *id* and dependent variable *preventive_visit_gap_ind*.

To ensure data completeness, we filled empty values in the newly generated one-hot columns from the last three tables (Member Conditions, Member Claims, and Quality Data), with 0 and used 'Unknown' for missing categorical variables columns. This approach allowed us to maintain the

[2] We include both condition key and condition description, because some condition key is matched with more than one condition description, and some condition description is matched with more than one condition key.

[3] We've tried to aggregate each measurement by more granular dates like Quarter, but the prediction made by our model thereafter using this level of granularity is not better compared to aggregate them with respect to different measure_year. Thus we aggregate the number of complaints for each measure, each member on each measure year.

maximum amount of information while preparing a clean, analysis-ready dataset for our predictive modeling efforts.

The following diagram illustrates the workflow for our data processing pipeline, detailing each step from data loading and merging to feature engineering, cleaning, and model training.

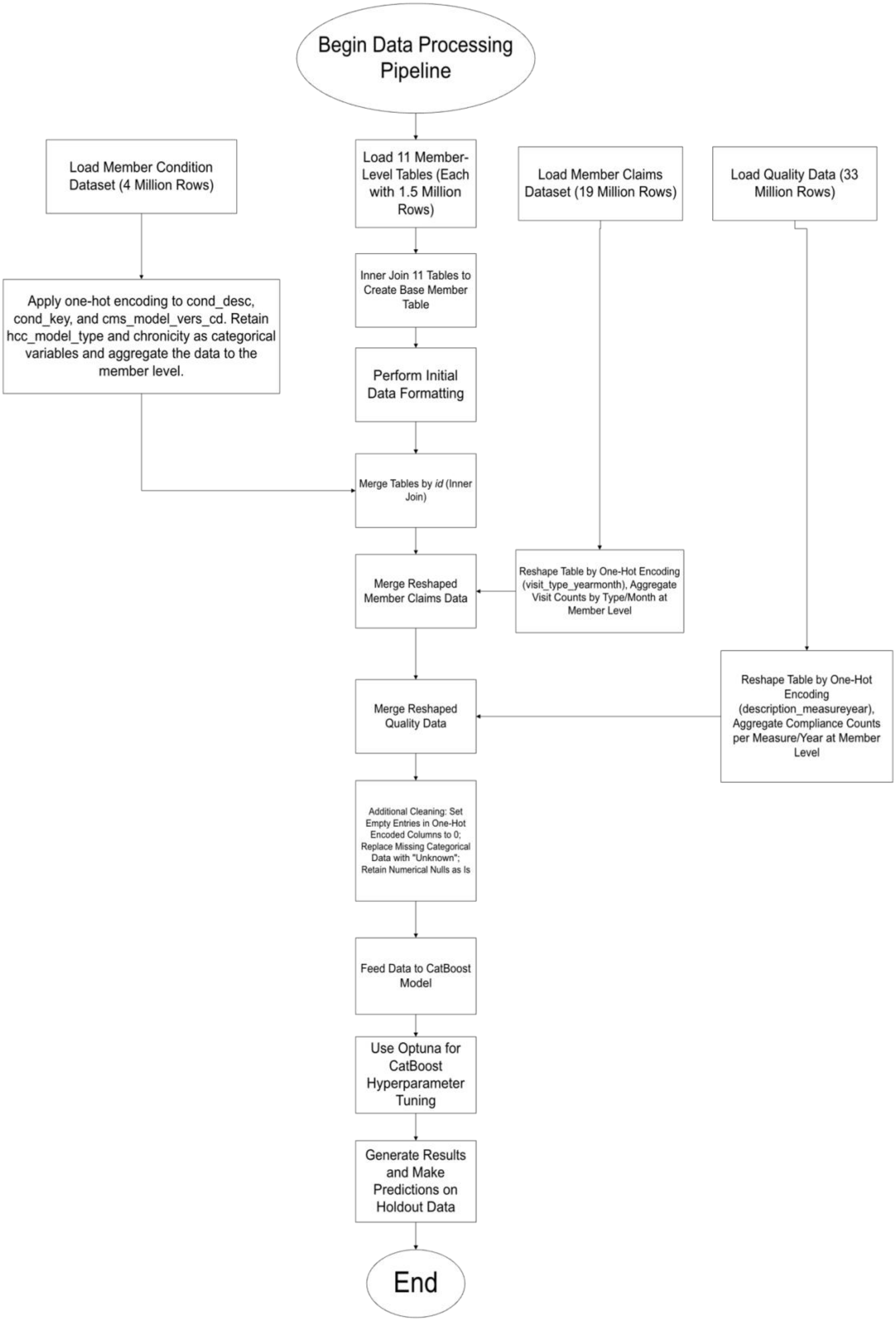

Begin Data Processing Pipeline
Load Member Condition Dataset (4 Million Rows)
Load 11 Member-Level Tables (Each with 1.5 Million Rows)
Load Member Claims Dataset (19 Million Rows)
Load Quality Data (33 Million Rows)
Apply one-hot encoding to cond_desc, cond_key, and cms_model_vers_cd. Retain hcc_model_type and chronicity as categorical variables and aggregate the data to the member level.
Inner Join 11 Tables to Create Base Member Table
Perform Initial Data Formatting
Merge Tables by *id* (Inner Join)
Merge Reshaped Member Claims Data
Reshape Table by One-Hot Encoding (visit_type_yearmonth), Aggregate Visit Counts by Type/Month at Member Level
Merge Reshaped Quality Data
Reshape Table by One-Hot Encoding (description_measureyear), Aggregate Compliance Counts per Measure/Year at Member Level
Additional Cleaning: Set Empty Entries in One-Hot Encoded Columns to 0; Replace Missing Categorical Data with "Unknown"; Retain Numerical Nulls as Is
Feed Data to CatBoost Model
Use Optuna for CatBoost Hyperparameter Tuning
Generate Results and Make Predictions on Holdout Data
End

Figure 3.1 Data Process Workflow

4. Exploratory Data Analysis, modeling and interpretation:

4.1 Exploratory Data Analysis

Here we are going to take a closer look at some of the features that may have an impact on the preventive visit gap among the members within the LPPO plan. To explore the relationships between various features and the preventive visit gap, we utilized Tableau to create a series of visualizations. These visualizations helped to uncover insights and trends in the data that would be useful for further analysis.

4.1.1 General Data Overview:

First, let us examine the target variable (preventive visit gap). As illustrated in figure 4.1., the pie chart visualizes the distribution of the preventive visit gap among LPPO members. The data shows a near-equal split between members who missed preventive visits and those who completed them. There are approximately 55.03% of LPPO members (840.84K members) who do not have a preventive visit gap, which means they are up to date with their preventive health visits. However, 44.97% of members (687.06K members) have a preventive visit gap, which means they have missed preventive care. The results suggest that a substantial proportion of LPPO members are not consistently participating in preventive healthcare visits, which can have broader implications for health outcomes.

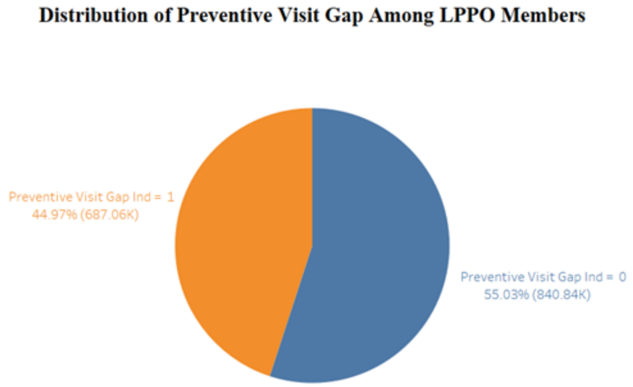


Figure 4.1 Distribution of Preventive visit gap Among LPPO members

Figure 4.2 shows three pie charts that show the gender distribution of the overall population, members who have a preventive visit gap, and members who are not having a preventive visit gap, from which it is possible to infer the following:

- The overall distribution of the dataset is 46% male and 54% female. Based on this, the two other groups (with and without a preventive visit gap) can be compared.

- The gender distribution among members who did not have a preventive visit gap is 41% male and 59% female. This indicates that the proportion of females is greater than the males who have completed their preventive visits.
- For members with a preventive visit gap, the distribution is 51% male and 49% female. This shows a higher proportion of males in this group than the overall dataset. This suggests that males are more likely to miss preventive visits than females.

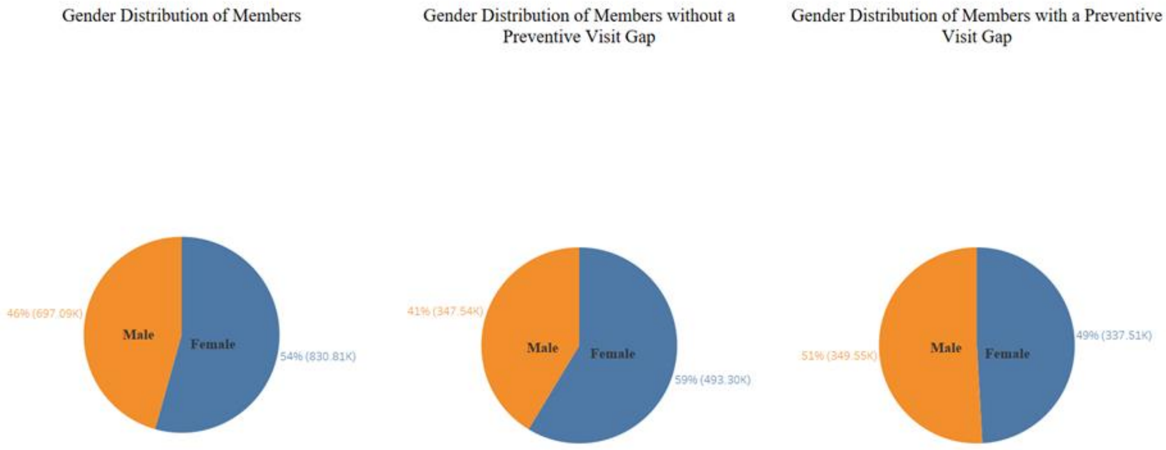


Fig 4.2 Gender distribution

The overall conclusion we can draw from these insights is that gender plays a role in preventive visit behavior, with males being more likely to have a preventive visit gap and females being more likely to complete their preventive visits.

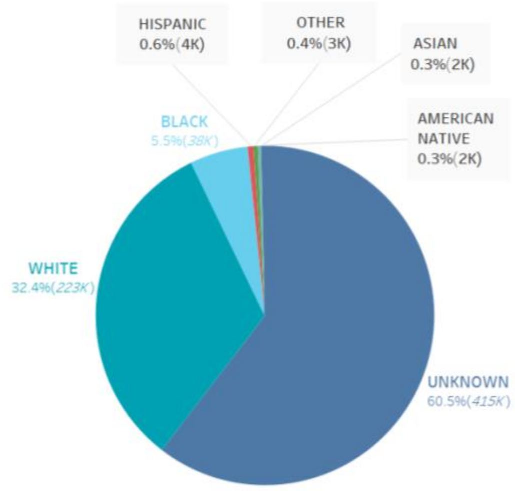


Fig 4.3 *Racial Distribution of Members with a Preventive Visit Gap*

Figure 4.3 shows the racial composition of LPPO members with preventive visit gaps. This visualization provides the following key insights:

- Most members with preventive visit gaps (60.5% or 415K) fall into the category of "Unknown". Based on this information, it is evident that a large portion of the race-related information is missing or not reported.
- The number of white members with a preventive visit gap is 32.4% (223K) of the total number of members in this category, making them the largest identified racial group in this category.
- Among those with preventive visit gaps, Black members make up 5.5% (38K), while other minority groups, such as Hispanic, Asian, and American Native members each represent a very small proportion of those with such gaps.
- There are less than 2% of members with a preventive visit gap among Hispanics, Asians, American Natives, and other groups combined. This may be an indication that these racial groups are underrepresented in the LPPO population, or it may be indicative that these groups are less likely to miss preventive visits than others. This assumption, however, is hard to verify given the large amount of "unknown" data available.

Overall, the overwhelming "Unknown" category suggests a significant gap in racial data for this population, which limits understanding of racial disparities in preventive visits. In the identified racial data, white members have the largest number of preventive visit gaps, indicating they are most likely to miss preventive visits. Black members have the second-largest preventive visit gap, though their proportion is lower. Hispanic, Asian, and American Native populations are very small in this analysis, making conclusions about their preventive visit behaviors difficult.

4.1.2 Preventive Visit Gap Based on Geographical Area

Let us look at how some of the features can help us uncover insights about the preventive visit gap. Figure 4.4 shows the top 5 counties in the US where the majority of LPPO members missed their preventive visits, with Wake County, North Carolina having the highest number (nearly 5000 members) having missed their preventive visits. Among the other counties in the top 5 are Cook, Illinois, Jefferson, Kentucky, Clark, Nevada and Palm Beach, Florida.

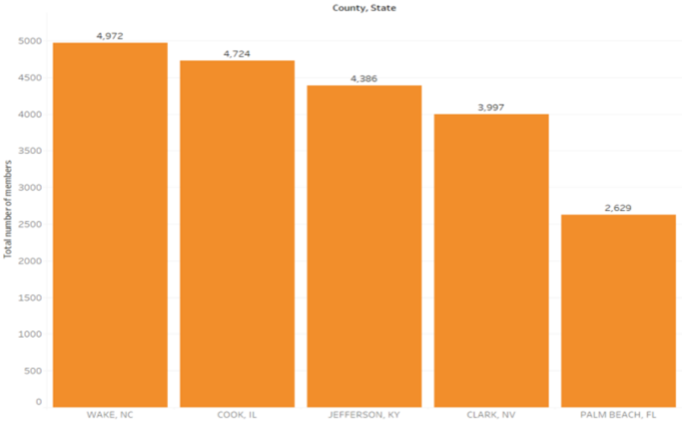


Fig 4.4 *Top 5 Counties with the Highest Number of LPPO Members Missing Preventive Visits*

Considering the states, we can observe from figure 4.5 that Texas, North Carolina, Georgia, Florida, and Kentucky are among the top five states with the highest number of LPPO members failing to attend preventive appointments. The number of missed visits for these states is significantly higher than for other states, with Texas having the largest number of missed visits with over 60,000 missed visits.

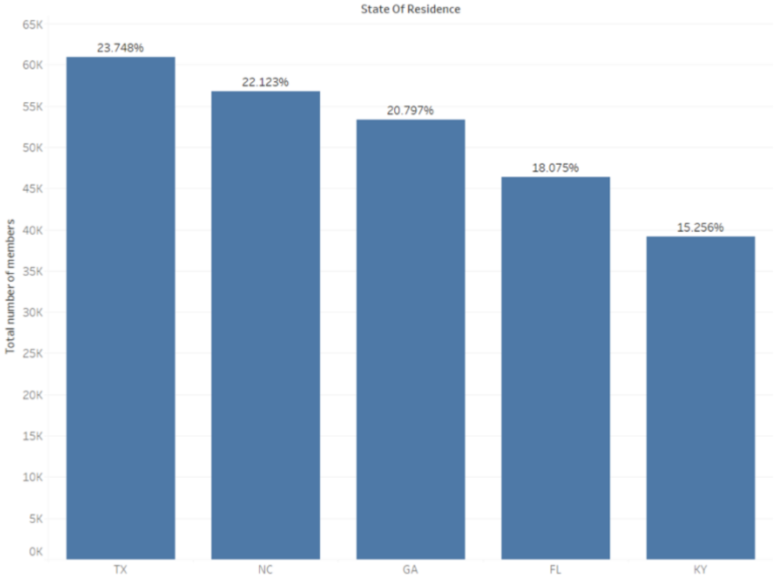


Fig 4.5 Top 5 States with the Highest Number of LPPO Members Missing Preventive Visits

This can be further confirmed by the shaded map in Fig 4.6 which shows the distribution of LPPO members missing preventive visits across the states of the United States. The states are color-coded, with darker shades of color indicating the number of members who have missed the preventive visits.

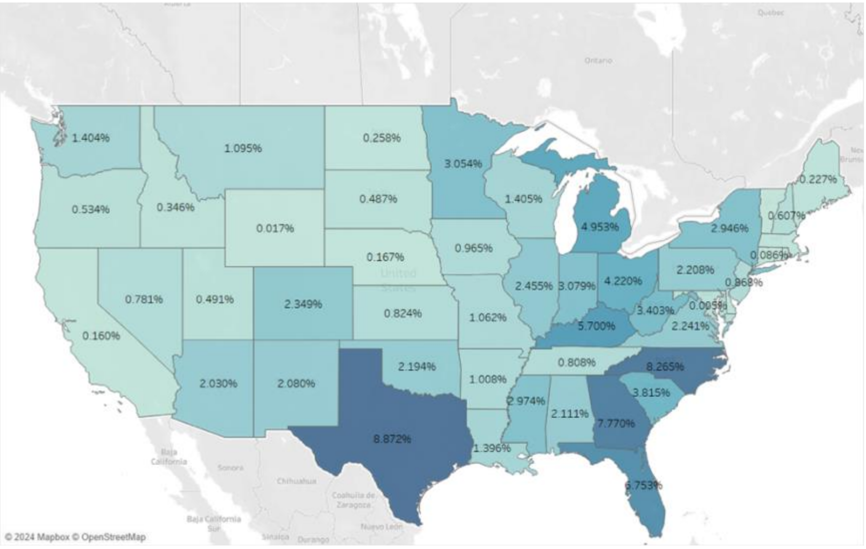


Fig 4.6 State wise Distribution of LPPO Members Missing Preventive Visits

In conclusion, there is a clear regional pattern, with southern states and southern eastern states (especially Texas, Georgia, and Florida) showing a higher concentration of missed preventive visits among LPPO members. In addition, we can also see from the map that these regions account for up to 30% of the LPPO members across the United States of America. Targeted health outreach programs may increase preventive care compliance and improve overall health outcomes for LPPO members in these regions.

4.1.3 Preventive Visit Gap for Veterans

In the below fig 4.7, we can see a bar chart that compares veterans who have and have not completed preventive care visit. It shows that 53% of veterans have not completed a preventive visit, while 47% have. This highlights a substantial gap in preventive care engagement among veterans.

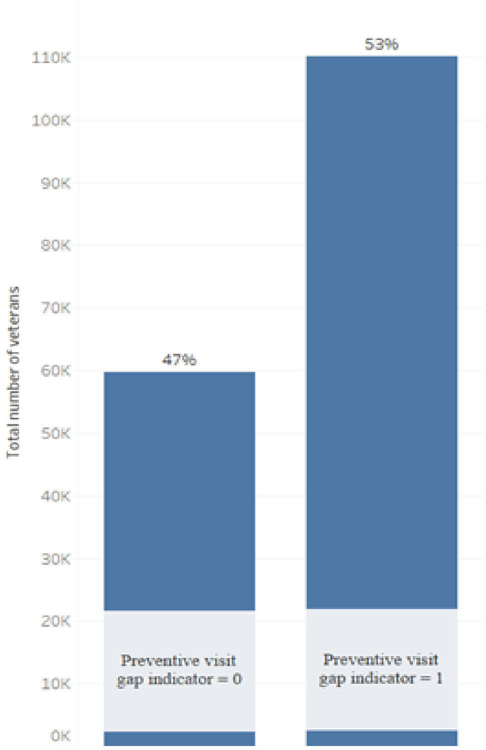


Fig 4.7 Distribution of Veterans over preventive visit gap

4.2 Model Selection

To predict the target variable, we evaluated several advanced machine learning algorithms suitable for our dataset of 1.5 million rows and 1,073 features, including 28 categorical variables. After careful consideration and experimentation, we selected CatBoost as our primary predictive model.

4.2.1 Introduction to CatBoost

CatBoost (Prokhorenkova et al., 2017), developed by Yandex, is a gradient boosting algorithm designed to handle categorical data efficiently without extensive preprocessing. The name "CatBoost" stands for "Categorical Boosting," highlighting its unique capability to process categorical variables directly. CatBoost incorporates innovative techniques to address common challenges such as handling categorical features, reducing overfitting, and improving prediction accuracy.

4.2.2 Mathematical Formulation of CatBoost

Like other gradient boosting algorithms, CatBoost builds an ensemble of trees sequentially to minimize a specified loss function. The prediction at the m-th iteration is given by:

$$E_m(X) = E_{m-1}(X) + \alpha_m h_m(X)$$

Where:

- $Em-1(X)$ is the prediction from the previous iteration
- αm is the learning rate
- hm is the decision tree (weak learner) built at iteration m

The model minimizes the loss function L over the training data:

$$\alpha_m, h_m = \arg\min_{\alpha,h} \sum_{i=1}^{N} L(Y_i, E_{m-1}(X_i) + \alpha h(X_i))$$

Where N is the number of samples, Y_i is the target variable, and X_i represents the feature vector for the i-th sample.

4.2.3 Key Innovations in CatBoost

1. Handling Categorical Variables:

CatBoost introduces Ordered Target Statistics to handle categorical variables. For a categorical feature c and a data point X_i, the Ordered Target Statistic $\widehat{c_\iota}$ is computed as:

$$\hat{c}_i = \frac{\sum_{j \in D_i} Y_j}{|D_i| + a}$$

Where D_i is the set of data points preceding X_i in a given permutation, Y_j is the target value for data point j, a is a prior hyperparameter, and $|D_i|$ is the cardinality of D_i.

2. Ordered Boosting:

CatBoost constructs trees using only information available prior to each data point, preventing target leakage. The gradient for each data point i at iteration m is computed using an ordered target:

$$g_{i,m} = \frac{\partial L(Y_i, E_{m-1}(X_i))}{\partial E_{m-1}(X_i)}$$

3. Symmetric Tree Structures:
CatBoost builds symmetric trees, where the structure is the same for all leaves at a given depth, reducing model variance and speeding up training and prediction.

4.2.4 Advantages of CatBoost

1. Efficient handling of categorical variables without manual preprocessing
2. Ease of use with minimal hyperparameter tuning required

4.2.5 Comparison with Other Algorithms

Compared to other methods of Boosting, CatBoost handles categorical variables, particularly those with a high cardinality, much better. It involves much less preprocessing and tuning of hyperparameters and is, therefore, much more efficient on real-world large datasets that contain incomplete data.

In our pilot test, we also compared CatBoost to neural network approaches, represented by TabNet, on tabular data. On the base member level dataset, TabNet has an AUC of about 0.75, which is low compared to CatBoost's 0.76. Meanwhile, TabNet has over 21 hours of training and tuning with 50 rounds, and even more time is spent preparing the missing values. Fine-turning with CatBoost for 50 rounds in this study will take just a few hours. With these criteria, we have chosen CatBoost with a balance between performance and efficiency.

4.3 Model Training and Hyperparameter Optimization

4.3.1 Data Preparation and Splitting

Before proceeding with model training and optimization, we split our data into training and testing sets using an 80-20 ratio. This approach allowed us to train our model on a substantial portion of the data while retaining a significant amount for testing and validation.
Moreover, to ensure reproducibility of our results, we set a fixed random state of 42 for all random operations, including the data split. This practice allows for consistent results across multiple runs and facilitates easier debugging and validation of our model.

4.3.2 Hyperparameter Tuning and Training

We defined an objective function that trains a CatBoost model with given hyperparameters and returns the inverse of the AUC score (1 - AUC) as the optimization target. This approach allows Optuna to minimize the objective, effectively maximizing the AUC score.
Key components of our optimization process:

- Used GPU acceleration for faster training
- Implemented early stopping to prevent overfitting

– Utilized cross-validation for robust performance estimation

4.3.3 Key Hyperparameters Tuned

– **iterations**: Number of boosting iterations (500 to 1500)
– **learning_rate (η)**: Step size for each iteration (0.01 to 0.15, log scale)
– **depth**: Maximum depth of the trees (6 to 12)
– **l2_leaf_reg**: L2 regularization term (1 to 10, log scale)
– **border_count**: Number of splits for numerical features (32 to 255)
– **bagging_temperature**: Controls randomness in bagging (0.0 to 1.0)
– **random_strength**: Amount of randomness for scoring splits (1e-5 to 10, log scale)

4.3.4 Optimization Results and Discussion

After running 50 trials, the best performing model achieved the following parameters:
iterations: 1349,
learning_rate: 0.10258099808532092,
depth: 9,
l2_leaf_reg: 1.7890664895548984,
border_count: 179,
bagging_temperature: 0.04182400529131734,
random_strength: 1.600911166731653e-05

The final model achieved these results:
- AUC: 0.7783620885664624
- Accuracy: 0.7091998520850445

Classification Report:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.71	0.8	0.75	168169
1	0.71	0.59	0.65	137412
Accuracy			0.71	305581
Macro Avg	0.71	0.7	0.7	305581
Weighted Avg	0.71	0.71	0.71	305581

Our model demonstrates moderate performance, with an overall accuracy of 71%. This indicates that the model correctly classifies nearly three-quarters of all cases, providing a useful but imperfect tool for identifying potential gaps in preventive care.

The model's performance varies between the two classes. For clients without a preventive visit gap (class 0), the model achieves a recall of 80%, indicating it successfully identifies 4 out of 5 cases where no gap exists. However, for clients with a preventive visit gap (class 1), the recall drops to 59%, suggesting the model fails to identify about 2 in 5 cases where a gap is present. This discrepancy is noteworthy, especially considering the relatively balanced distribution of the target variable (55.03% for class 0 and 44.97% for class 1).

Interestingly, the precision is consistent at 71% for both classes, meaning that when the model predicts either outcome, it is correct about 71% of the time. This balance in precision, coupled with the imbalance in recall, results in F1-scores of 0.75 for class 0 and 0.65 for class 1.

The performance metrics, particularly the F1-scores of 0.75 for class 0 (no preventive visit gap) and 0.65 for class 1 (preventive visit gap), highlight the model's varying effectiveness in predicting different outcomes. This discrepancy suggests that certain factors may be more influential in determining whether a member will have a preventive visit gap. To better understand these factors and their relative importance, we turn to model interpretation techniques. By examining the key features identified by our CatBoost model through the lens of Andersen's Behavioral Model, we can gain valuable insights into the determinants of preventive care utilization among LPPO members.

4.3.5 Potential Improvements for Our Model

While our CatBoost model has shown promising results, there are several avenues we could explore to potentially enhance our predictive capabilities:

1. **Advanced Tabular Neural Networks**: Explore models like FT-Transformer (Feature Tokenizer Transformer), which has shown strong performance on tabular data. FT-Transformer's ability to capture complex interactions between features could potentially improve our predictions.

2. **Ensemble Methods**: Implement stacking or blending techniques, combining CatBoost with other algorithms like LightGBM or XGBoost to leverage the strengths of multiple models.

3. **Incorporating External Data**: Integrate additional external datasets, such as more local socioeconomic indicators beyond the ones provided in this dataset, to provide more context for predictions and potentially uncover new patterns in preventive care utilization.

By exploring these avenues, we aim to not only improve the accuracy of our predictions but also gain deeper insights into the factors influencing preventive care utilization among LPPO members.

5. Model Interpretation

Our CatBoost model identified several key features that influence preventive care utilization among LPPO members. To provide a structured and theoretically grounded interpretation of these results, we use Andersen's Behavioral Model of Health Services Use as our primary framework.

5.1 Andersen's Behavioral Model and Our Analytical Approach

Andersen's Behavioral Model, first developed in the 1960s and refined over time, posits that health services use is determined by three main components:

1. **Predisposing Factors**: Characteristics that predispose individuals to use or not use services (e.g., demographics, social structure, health beliefs).
2. **Enabling Factors**: Conditions that facilitate or impede the use of services (e.g., income, health insurance, access to care).
3. **Need Factors**: Perceived or evaluated health status that necessitates health service use.

Our analysis extends this model to include additional categories that are particularly relevant to our LPPO member population:

4. **Healthcare Costs**: A complex factor that could be either enabling or disabling.
5. **Member Engagement and Utilization**: Factors reflecting members' interaction with the healthcare system.

We first identified the top 30 features by importance:

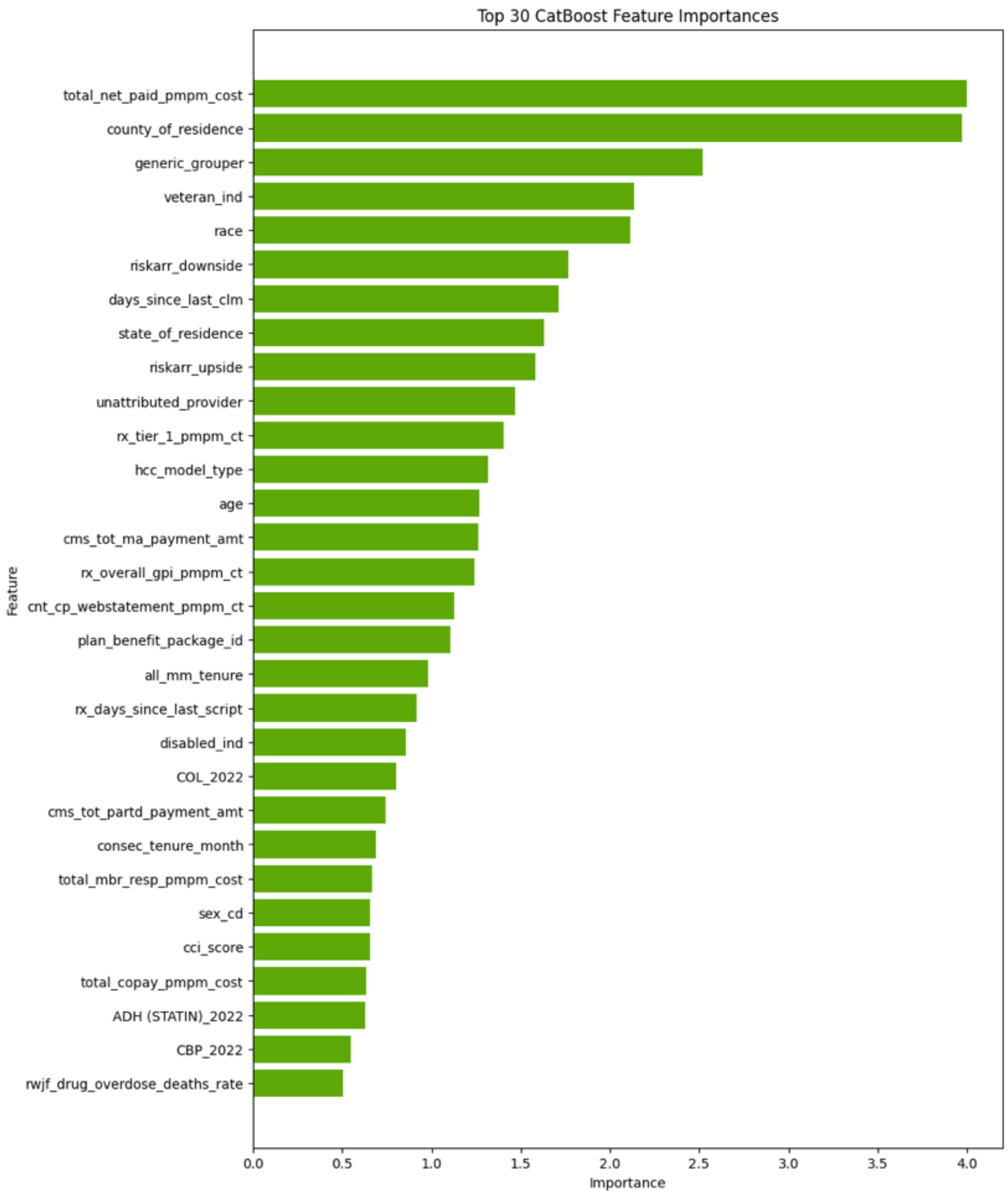


Figure 5.1 Top 30 Features by Importance

The following table categorizes these features according to our extended Andersen's Model framework:

Demographic and Geographic Factors (Predisposing Factor)	**Healthcare Costs (Could be Enabling or disabling Factor)**	**Plan Characteristics (Enabling Factor)**	**Health Status and Comorbidities (Need Factor)**	**Member Engagement and utilization**
county_of_residence (3.975)	total_net_paid_pmpm_cost (4.001)	generic_grouper (2.521)	cci_score (0.654)	days_since_last_clm (1.709)

veteran_ind (2.133)	cms_tot_ma_payment_amt (1.260)	riskarr_downside (1.768)	disabled_ind (0.858)	rx_tier_1_pmpm_ct (1.403)
race (2.116)	cms_tot_partd_payment_amt (0.741)	riskarr_upside (1.582)		rx_overall_gpi_pmpm_ct (1.242)
state_of_residence (1.633)	total_mbr_resp_pmpm_cost (0.665)	unattributed_provider (1.470)		cnt_cp_webstatement_pmpm_ct (1.127)
age (1.266)	total_copay_pmpm_cost (0.635)	hcc_model_type (1.315)		all_mm_tenure (0.981)
sex_cd (0.658)		plan_benefit_package_id (1.103)		rx_days_since_last_script (0.915)
rwjf_drug_overdose_deaths_rate (0.502)				COL_2022 (0.799)
				consec_tenure_month (0.685)
				ADH (STATIN)_2022 (0.625)
				CBP_2022 (0.545)

5.2 Detailed Analysis of Factor Categories

Building on the framework established in section 5.1, we now take a closer look at each factor category. Our analysis explores how specific features within these categories shape preventive care utilization, shedding light on the complex mix of factors that influence healthcare-seeking behavior among our LPPO members.

1. Demographic and Geographic Factors (Predisposing and Enabling Factors) Key features: *county_of_residence*, *state_of_residence*, *race*, *veteran_ind*, *age*, *sex_cd*; Geographic factors, particularly *county_of_residence*, are highly influential in predicting preventive care gaps. This suggests significant regional variations in preventive care utilization, possibly due to differences in healthcare access, local health initiatives, or socioeconomic factors. Demographic factors like race, veteran status, and age also play important roles. These factors reflects disparities in healthcare access or utilization patterns among different demographic groups.
2. Healthcare Costs (Complex Factor - Potentially Enabling or Disabling) Key features: *total_net_paid_pmpm_cost*, *cms_tot_ma_payment_amt*, *cms_tot_partd_payment_amt*, *total_mbr_resp_pmpm_cost*, *total_copay_pmpm_cost* The high importance of *total_net_paid_pmpm_cost* suggests that overall healthcare spending is strongly correlated

with preventive care utilization. By using logistic regression, we found that on average, members with higher healthcare costs may be more likely to attend preventive visits, possibly due to greater health needs or more frequent interactions with the healthcare system.

3. Plan-related features significantly influence preventive care utilization. The *generic_grouper*, risk arrangement flags (*riskarr_downside* and *riskarr_upside*), and *hcc_model_type* suggest that a plan's approach to medications, risk management, and health status assessment correlate with preventive care patterns. Additionally, *unattributed_provider* and *plan_benefit_package_id* indicate that specific plan designs and provider attribution impact utilization. Collectively, these factors highlight how plan structure, risk approaches, and benefit designs are associated with members' likelihood of using preventive care services, potentially reflecting differences in care management strategies and incentives across various plan characteristics.
4. Health Status and Comorbidities (Need Factors) Key features: *cci_score*, *disabled_ind*. These features reflect the impact of overall health status on preventive care utilization. For example, by simply using logistic regression, we found the *cci_score* (Charlson Comorbidity Index) suggests that members with more complex health needs have higher preventive care utilization. However, also using the logistic regression, *disabled_ind* shows an opposite trend: members with disabilities are less likely to use preventive care, possibly due to accessibility issues.
5. Member Engagement and Healthcare Utilization Key features: *days_since_last_clm*, *rx_tier_1_pmpm_ct*, *rx_overall_gpi_pmpm_ct*, *rx_days_since_last_script*, *cnt_cp_webstatement_pmpm_ct*, *all_mm_tenure*, *consec_tenure_month*, *COL_2022*, *ADH (STATIN)_2022*, *CBP_2022*. Recent healthcare utilization, represented by *days_since_last_clm*, is predictive of preventive care attendance. Members who have had recent claims are more likely to attend preventive visits, suggesting that engagement with the healthcare system promotes preventive care. Medication-related variables also show high importance. Their impact on *preventive_care_ind* varies, indicating that members who are actively managing their health through regular medication use are more likely to attend preventive visits. Indicated by *cnt_cp_webstatement_pmpm_ct*, member engagement with their health plan, particularly through digital means, is associated with higher preventive care use. A simple logistic regression also reveal that longer tenure with the plan also correlates with better preventive care adherence. Past preventive care behavior is predictive of future behavior. Logistic regression also reveal that members who have completed colorectal cancer screening, adhere to statin medication, or have controlled blood pressure in 2022 are more likely to continue engaging in preventive care.

5.3 Additional Analysis Using SHAP

While the CatBoost model provides overall feature importance, it doesn't capture how these factors might affect individual members differently. To address this, we used SHAP (SHapley Additive exPlanations) values to provide a more nuanced understanding of our model's predictions.

SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

We used SHAP for several reasons:

1. Individual-level insights: SHAP values show how each feature contributes to a prediction for each individual member, allowing us to understand decision-making patterns at a granular level.
2. Interaction effects: SHAP can reveal how features interact with each other to influence predictions.
3. Consistency with global importance: While providing local explanations, SHAP values are consistent with the overall feature importance, offering a bridge between local and global interpretability.
4. Handling of complex models: SHAP works well with complex models like CatBoost, providing interpretability without sacrificing model performance.

By using SHAP, we can understand how different factors influence preventive care decisions for different subgroups of members, providing a more comprehensive view than the overall feature importance alone. Here are the top 30 average SHAP values across our test dataset.

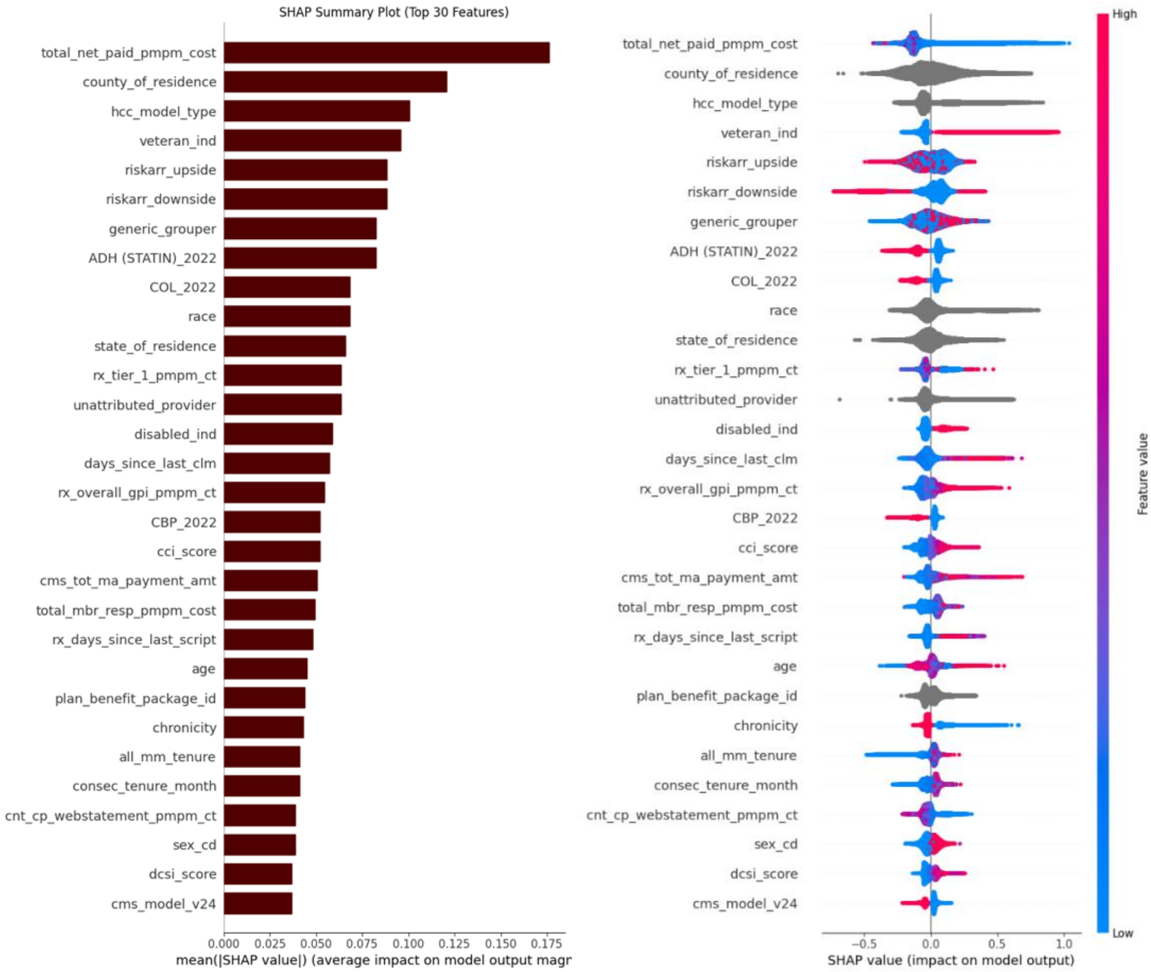


Figure 5.2 (Left) Top 30 features ranked by average SHAP values, showing the most impactful predictors on the model's output, with total_net_paid_pmpm_cost leading in influence.
(Right) SHAP summary plot visualizing the distribution of feature impacts on model predictions, with colors representing feature values (blue = low, red = high) for the top 30 predictors.

Here are insights we found from the two graph above:

1. **Feature Importance Confirmation**: The SHAP summary plot largely confirms the CatBoost feature importance ranking, with total_net_paid_pmpm_cost, county_of_residence, and hcc_model_type as top influencers.

2. **Non-linear Relationships**: Many features, including total_net_paid_pmpm_cost and days_since_last_clm, show non-linear impacts on preventive care utilization predictions.

3. **Geographic Variation**: county_of_residence and state_of_residence show highly variable impacts, suggesting local factors significantly influence preventive care utilization.

4. **Risk Factors**: Different hcc_model_types and risk arrangement factors (riskarr_upside and riskarr_downside) have distinct impacts on predictions, indicating the importance of risk profiling in preventive care utilization.

5. **Health Status Effects**: Features like cci_score and dcsi_score generally show that poorer health status decrease the predicted likelihood of preventive care utilization, which is contrast from our previous finding on simple logistic regression.

6. **Engagement Metrics**: More recent healthcare interactions (lower values in days_since_last_clm and rx_days_since_last_script) generally increase the predicted likelihood of preventive care utilization.

These insights reveal complex interactions between features and preventive care utilization, emphasizing the need for nuanced, personalized approaches in designing interventions to improve preventive care utilization across different member segments.

In the next step we plot the SHAP values, here is some interesting local pattern we've found:

1. **Age and Tenure Interaction:**
    - Under 75: Longer tenure associated with higher likelihood of missing preventive visits.
    - Over 75: Shorter tenure associated with missing preventive visits.

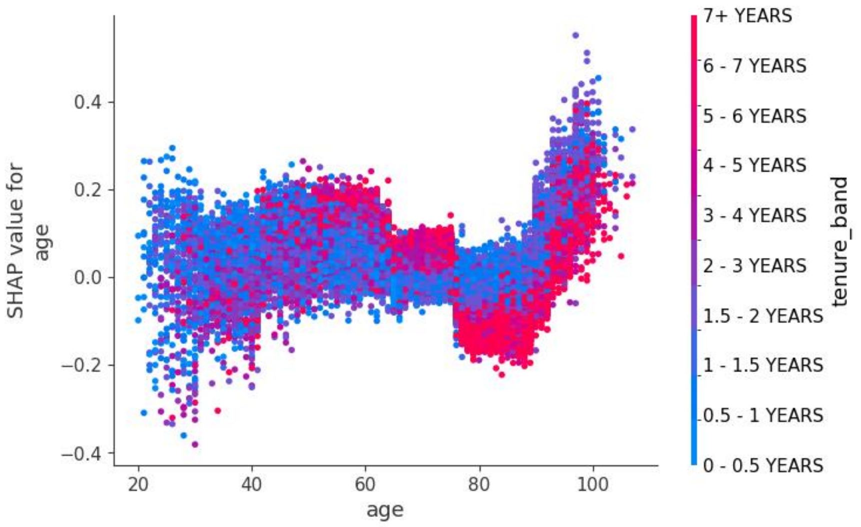


Figure 5.3 SHAP values for age across tenure bands, showing age's impact on model predictions, with tenure increasing from blue to red

2. **Gender and Tenure Interaction:**
    - Females: Longer tenure associated with higher likelihood of missing preventive visits.
    - Males: Shorter tenure associated with higher likelihood of missing preventive visits.

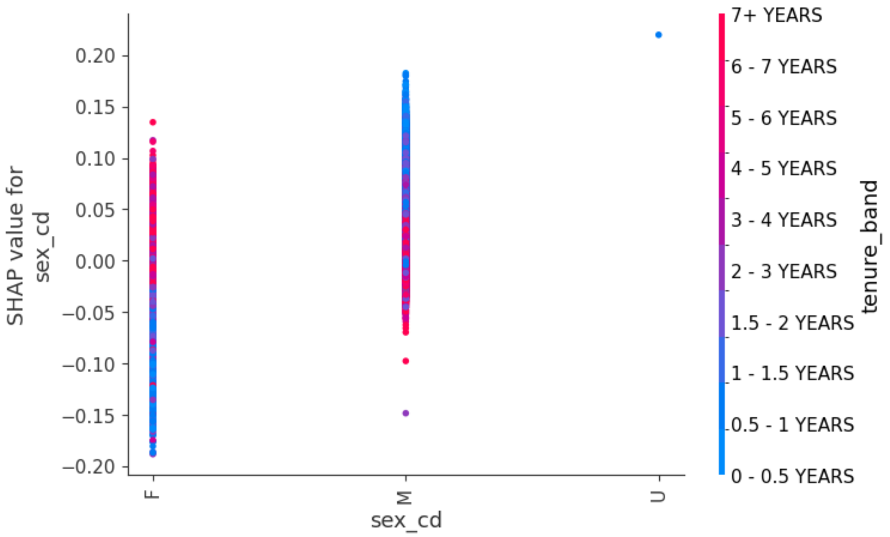


Figure 5.4 SHAP values for sex code (F, M, U), illustrating the impact of gender on model predictions, with color indicating tenure bands from 0 to 7+ years

3. **Veteran Status and Healthcare Utilization:**
    - Non-veterans: Lower utilization associated with lower likelihood of missing preventive care.
    - Veterans: Lower utilization associated with higher likelihood of missing preventive care.

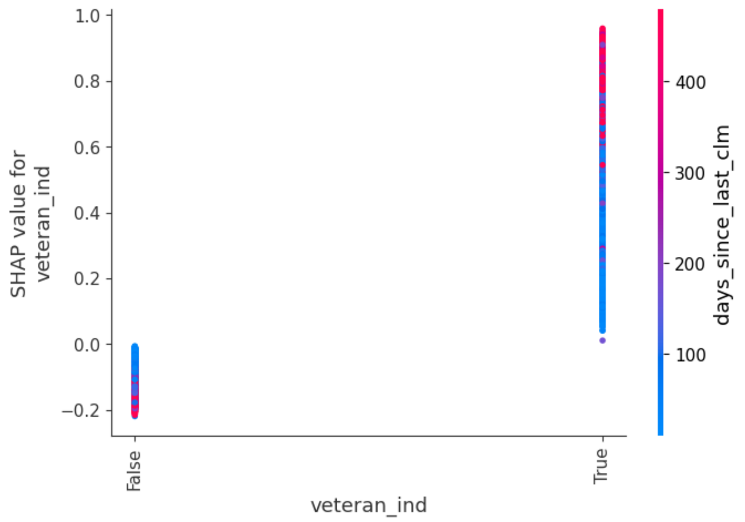


Figure 5.5 SHAP values for veteran status (True/False), showing the influence of veteran indication on model predictions, with colors representing days since last claim.

4. **Disability Status and Chronic Conditions:**
   - Without disabilities: No chronic conditions associated with lower likelihood of missing preventive care.
   - With disabilities: No chronic conditions associated with higher likelihood of missing preventive care.

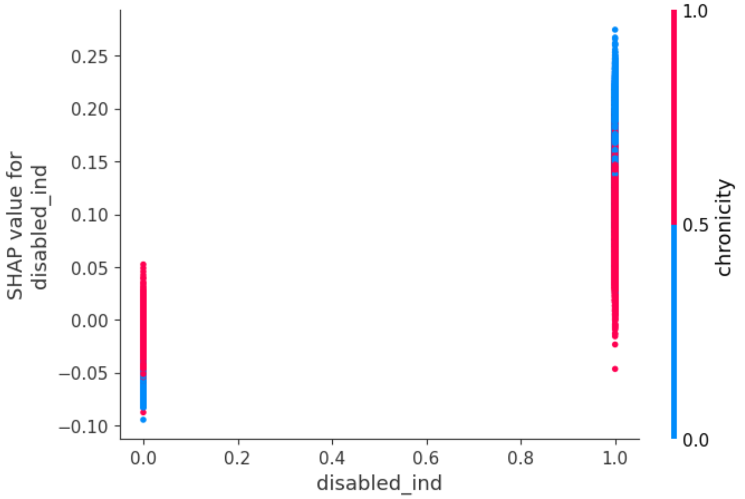


Figure 5.6 SHAP values for disability status (0 = False, 1 = True), illustrating the effect of disability indication on model predictions, with color representing the degree of chronicity.

The SHAP analysis reveals complex interactions between member characteristics and their likelihood of missing preventive care visits. These insights suggest that:

1. Tailored outreach strategies may be necessary for different age groups, genders, and veteran statuses.
2. The impact of tenure on preventive care utilization varies significantly across different demographic groups, highlighting the need for personalized retention and engagement strategies.

3. Veterans with low healthcare utilization may require special attention to ensure they receive preventive care.
4. The relationship between chronic conditions and preventive care adherence differs for people with disabilities, potentially indicating accessibility issues.

In the next step, we will build our business recommendation based on what we’ve learnt from the data.

6. Business Implications

6.1 Market landscape on Preventive Care incentive

Before presenting our suggestions, it's crucial to examine the strategies our industry peers are implementing in this field. Major health insurance competitors are introducing innovative approaches to enhance preventive care utilization, recognizing its vital role in improving health outcomes and reducing overall healthcare costs. These initiatives reflect a broader industry shift towards proactive, personalized approaches to healthcare delivery.

UnitedHealthcare has made significant strides in this area with the launch of its Surest plan[4]. This innovative offering eliminates deductibles and provides upfront pricing information, leading to remarkable improvements in preventive care utilization:

- 20% increase in physician visits
- 9% increase in preventive physical exams
- 15% improvement in preventive mammograms
- 34% higher rate of preventive colonoscopies

Building on this success, UnitedHealthcare has also introduced NavigateNOW[5], a virtual-first health plan that combines digital convenience with personalized care. This plan offers:

- 24/7 access to personalized care teams for primary, urgent, and behavioral health services
- $0 copays for virtual and in-person primary care and behavioral health visits
- An integrated approach that seamlessly combines virtual and in-person care

Cigna has taken a community-focused approach to improving preventive care access. The company leverages partnerships with local organizations to provide free preventive care screenings and wellness education events. Cigna also utilizes its proprietary Evernorth Social Determinants of Health Index (ESD)[6] to identify areas where additional preventive care resources are needed most. This data-driven approach allows Cigna to target its efforts effectively and address specific community needs.

Aetna has implemented the Aetna Community Care Program[7], which takes a holistic approach to preventive care. This program employs nurse care managers, social workers, and health educators to provide personalized care and address social determinants of health that may impede access to preventive services. Aetna has also embraced value-based care models[8] that emphasize preventive care and early detection. By leveraging data analytics and technology, Aetna proactively identifies care gaps and reaches out to patients who may be due for preventive services.

[4] https://www.uhc.com/news-articles/newsroom/surest-preventive-care

[5] https://www.unitedhealthgroup.com/newsroom/2021/2021-10-18-uhc-virtual-first-health-plan.html

[6] https://newsroom.thecignagroup.com/healthier-you-preventive-care-health-equity

[7] https://www.aetnabetterhealth.com/impact/community-care-program.html

[8] https://www.aetna.com/employers-organizations/resources/value-based-care.html

Across these major insurers, several common strategies emerge:

- Expanding telehealth and digital health solutions to improve access to preventive care
- Addressing social determinants of health to reduce barriers to preventive services
- Implementing value-based care models that incentivize preventive care and better health outcomes
- Leveraging data analytics and technology to identify at-risk patients and encourage preventive care utilization
- Forming partnerships with local healthcare providers and community organizations to expand preventive care offerings

These initiatives demonstrate the industry's commitment to reimagining healthcare delivery, with a strong focus on prevention and early intervention. By combining technological innovation, data-driven insights, and community engagement, these insurers are working to create a more proactive and accessible healthcare system that prioritizes preventive care. As these strategies continue to evolve, they have the potential to significantly improve health outcomes, enhance patient experiences, and ultimately reduce the overall cost of healthcare.

6.2 Data Driven Recommendation from our Team

While our competitors have implemented various strategies to improve preventive care utilization, our team has taken a unique, data-driven approach to address this challenge specifically for Humana's LPPO members. By leveraging advanced analytics and machine learning techniques, we've developed targeted recommendations that complement and potentially enhance the industry-wide strategies.

6.2.1 Target Clusters

Using our pre-trained CatBoost model, we identified high-risk patients (predicted probability >0.8 of missing preventive visits) in our test set of 305,581 members from the test dataset[9]. This high-risk group comprises 33,361 members (10.9% of the test dataset).

We then performed K-means clustering on these high-risk patients using SHAP values of the top 50 features, resulting in three distinct clusters[10]:

1. **Cluster 0: Diverse Population** (31% of high-risk members)
   - Lower healthcare costs, racially diverse, moderate veteran representation (39.7%)
   - Key factors: Healthcare costs, HCC model type, veteran status, time since last claim
   - Insight: Potential underutilization of healthcare services

2. **Cluster 1: Non-Veteran, High-Cost Group** (41.7% of high-risk members)
   - Highest healthcare costs, no veterans, lack of attributed providers
   - Key factors: Race, county of residence, unattributed provider, healthcare costs

[9] We use the test set to train the SHAP explainer because the full dataset is too large for the memory.
[10] We will explain why we split them into 3 groups in the Appendix of this Chapter.

- Insight: The low utilization of preventive care visits might be due to the financial burden of high healthcare costs.

3. **Cluster 2: Veteran-Exclusive Group with Low Healthcare Utilization** (27.3% of high-risk members)
   - All veterans, lowest healthcare costs, longest time since last claim
   - Key factors: Veteran status (extremely high influence), healthcare costs, time since last claim
   - Insight: Veteran-specific factors critically important

Note that Cluster 2 is align with our finding in the last section, such that veterans who don't utilize healthcare often are more likely to miss preventive care visit.

6.2.2 Business Recommendations based on our model

Our recommendations focus on Clusters 1 and 2, as Cluster 0 represents a mix between the other two clusters. Assuming we are using the over 1.5 million members from both the training and test dataset.

6.2.2.1 Cluster 1: Preventive Care Rebate Program

Aim: Encourage preventive care visits by offering a financial rebate (e.g. $50) to high-risk members, since their spending on healthcare cost is high compared to the average member in our test dataset.

Step	Calculation
Total Member Population	1,527,904
High-risk members	1,527,904 × 10.9% = 166,542
Target Group (Cluster 1)	166,542 × 41.7% ≈ 69,448
Current Utilization	69,448 × 11.37% ≈ 7,896
Projected utilization after program	11.37% + 7.3%[11] = 18.67%
New preventive care users	69,448 × 18.67% ≈ 12,966 members
Incremental increase in users	12,966 - 7,896 = 5,070 members
Total rebate cost	12,966 × $50 = $648,300
Administrative cost (10%)	$648,300 × 10% = $64,830
Estimated healthcare savings	5,070 × $500 = $2,535,000
Total program cost (Rebates + Administrative costs)	$648,300 + $64,830 = $713,130
Net benefit (Estimated savings - Total program cost)	$2,535,000 - $713,130 = $1,821,870
ROI ((Net benefit / Total program cost) × 100%)	($1,821,870 / $713,130) × 100% ≈ 255.47%

Key Outcomes:
• Total members impacted: 69,448
• Additional members using preventive care: 5,070

[11] An estimation of the 7.3% comes from this research(Green et al., 2019), it's in the middle of 7.1% and 7.7%

- Net financial benefit: $1,821,870
- ROI: 255.47%

6.2.2.2 Cluster 2: Veteran-Centric Care Coordination Program

Aim: To improve preventive healthcare engagement among veterans through a peer-led, veteran-specific care coordination program, while ensuring the highest standards of privacy, consent, and ethical practices.

Strategy: Our approach leverages the 4.7% of veterans in Cluster 2 who currently utilize preventive care to encourage their peers. Key elements include:

1. **Recruitment**: Target the 2,098 veterans already using preventive care as potential volunteers.

2. **Training**: Equip volunteers to share personal experiences with preventive care effectively.

3. **Messaging**: Focus on the tangible benefits volunteers have experienced from preventive care.

4. **Peer Matching**: When possible, connect volunteers with peers of similar age or background to enhance relatability.

5. **Consent-Based Outreach**: Assign volunteers to contact only those veterans who have explicitly agreed to participate in the program.

This strategy aims to increase preventive care utilization from 4.7% to 6.58% among the target veteran population, fostering a culture of proactive health management within the veteran community.

Step	Calculation
Total Member Population	1,527,904
High-risk members	1,527,904 × 10.9% = 166,542
Target Group (Cluster 2)	166,542 × 27.3% = 45,466
Current preventive care utilization	45,466 × 4.7% = 2,137 members
Potential volunteer pool	2,137 (members already using preventive care)
Volunteers (10% of group)	2,137 × 10% = 214 volunteers
Each volunteer reaches out to	20 peers over a year
Total outreach	214 × 20 = 4,280 veterans
Increase success rate to	20% (more experienced volunteers)
New preventive care users	4,280 veterans × 20% success rate = 856 new users
New utilization rate	(2,137 + 856) / 45,466 = 6.58% (up from 4.7%)
Gift card incentive	214 × $50 = $10,700

Administration (part-time coordinators)	$1,000/month × 12 months × 6 coordinators = $72,000
Communication materials	$2,000
Total program cost	$10,700 + $72,000 + $2,000 = $84,700
Healthcare savings	856 new users × $500 = $428,000
Net benefit	$428,000 - $84,700 = $343,300
Return on Investment (ROI)	($343,300 / $84,700) × 100 ≈ 405.3%

Key Outcomes:

- Volunteers: 214
- Additional veterans using preventive care: 856
- New utilization rate: 6.58% (40% increase)
- Net financial benefit: $343,000
- ROI: 405.3%

Note on Legal and Ethical Considerations for Cluster 1 and Cluster 2 Targeting

To ensure the legality and ethical implementation of both our Preventive Care Rebate Program (Cluster 1) and Veteran-Centric Care Coordination Program (Cluster 2), we will adhere to the following principles:

1. **Informed Consent**:
    - Cluster 1: Members will be fully informed about the rebate program, its terms, and conditions before participation.
    - Cluster 2: We will only contact veterans who have provided explicit, informed consent to participate in the peer outreach program.
2. **HIPAA Compliance**: All aspects of both programs, including data handling and communication, will strictly adhere to HIPAA regulations to protect members' health information.
3. **Privacy Protection**:
    - Cluster 1: Personal information used for the rebate program will be strictly protected.
    - Cluster 2: Volunteers will receive comprehensive training on maintaining confidentiality and protecting the privacy of their peers.

4. **Opt-out Mechanism**: Participants in both programs will have the right to opt-out at any time through an easily accessible process, with no negative consequences to their healthcare.
5. **Data Security**: All personal information will be stored securely, with access limited to authorized personnel only. Data will be used solely for the purposes of these programs.
6. **Ethical Targeting**:
   - Cluster 1: The rebate offer will be made available to all eligible high-risk members without discrimination.
   - Cluster 2: Peer matching will be based on relevant factors such as age, background, or shared experiences, always prioritizing participant consent and privacy.
7. **Regulatory Compliance**: We will ensure compliance with all relevant federal and state regulations governing healthcare incentives, outreach, and veteran services.
8. **Transparency**: Clear information about both programs' goals, processes, and data usage will be provided to all participants.
9. **Regular Audits**: We will conduct regular audits to ensure compliance with these principles and to identify any areas for improvement in both programs.
10. **Equal Access**:
    - Cluster 1: We will ensure that the rebate program doesn't inadvertently disadvantage members who may have financial constraints in accessing preventive care.
    - Cluster 2: We will provide multiple channels of communication for veteran outreach to ensure accessibility for all.
11. **Cultural Sensitivity**:
    - Cluster 1: Communication about the rebate program will be culturally appropriate and available in multiple languages as needed.
    - Cluster 2: Veteran volunteers will be trained in cultural competence to effectively engage with diverse veteran populations.
12. **Financial Integrity**:
    - Cluster 1: The rebate process will be transparent, with clear eligibility criteria and timely disbursement.
    - Cluster 2: Any financial incentives for volunteers will be clearly disclosed and compliant with relevant regulations.

By implementing these measures, we aim to create programs that respect the rights and privacy of all our members, including our veteran population, while effectively improving their access to preventive care. These considerations will help ensure that our initiatives are not only effective but also ethically sound and legally compliant.

6.3 More General Recommendations from our Team

Besides detailed targeted strategy, our analysis has also led to several general recommendations to improve preventive care utilization. These recommendations focus on digital engagement, accessibility for members with disabilities, and exploring network effects.

First, we suggest implementing a digital engagement strategy. We've noticed that members who regularly check their statements online are more likely to attend preventive care visits. Therefore, we recommend nudging people to check their billings more frequently. This could be another way to increase people's awareness of health, and thus lead to an increase in the utilization of preventive care visits.

Moreover, we advise considering the needs of members with disabilities. Providing at-home healthcare preventive checks for people with disabilities could significantly improve access to these crucial services. This approach ensures that mobility challenges or other disabilities do not become barriers to receiving preventive care.

Last but not the least, we recommend exploring the potential of network effects on healthcare behaviors. Some research(Latkin & Knowlton, 2015) has shown that people's health behaviors are influenced by their peers. Humana may consider including peer information in their dataset. For example, if both the member and their spouse are in the dataset, this relationship should be noted. Additionally, including a column to indicate if a member joined the program because of a referral from another member could be valuable. By collecting and analyzing this network data, we could have a better understanding of how members' behavior is influenced by their peers. This insight could potentially be leveraged to improve outreach strategies and increase preventive care utilization across the member base.

6.4 Appendix: Clustering Method and Cluster Characteristics

We began our clustering exploration by using the Elbow method in K-Means clustering. K-Means is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping subgroups (clusters) where each data point belongs to only one group.

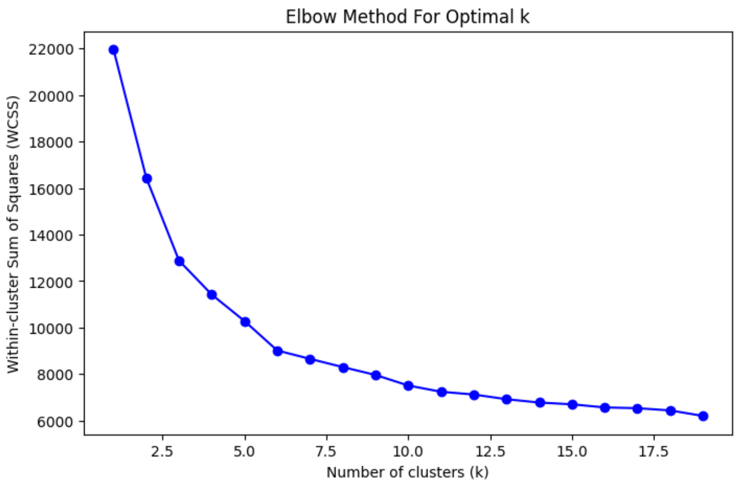
The Elbow method is a technique used to determine the optimal number of clusters (K) in K-Means clustering. It works by running K-Means clustering on the dataset for a range of K values (e.g., 1 to 10) and for each K, calculating the sum of squared distances from each point to its assigned center (inertia). As K increases, the inertia will decrease because points will be closer to their assigned centers.

The Elbow method gets its name from the way the graph looks: when you plot the inertia against the number of clusters, the line chart resembles an arm. The "elbow" of this arm is considered to be the optimal K value. This point represents where adding another cluster doesn't give much better modeling of the data, indicated by a sudden flattening of the average distance change within the groups.

In our analysis, although the graph suggested that six clusters could be optimal, when we grouped the high-risk members into six clusters, the characteristics were difficult to distinguish, and some groups seemed to share similar characteristics. This is a common occurrence in data analysis, where statistical measures may suggest one thing, but practical interpretability suggests another.

When we adjusted our approach to use three clusters, the characteristics of each group became more distinct and interpretable. This decision to use three clusters instead of six demonstrates the importance of balancing statistical measures with practical considerations and domain knowledge in data analysis.

This approach allowed us to identify more clearly defined and meaningful segments within our high-risk member population, which in turn enables more targeted and effective intervention strategies.



Here are the characteristic of the clusters:
Cluster 0:

Feature	SHAP Value	Description
total_net_paid_pmpm_cost	0.287849	Mean: $175.13. Lower than high-risk group and test set, indicating low healthcare utilization.
hcc_model_type	0.238158	72.99% Unknown, higher than test set (24.68%), suggesting potential gaps in risk assessment.
veteran_ind	0.152174	39.71% veterans, higher than test set (11.14%), indicating a significant veteran population.
days_since_last_clm	0.150878	Mean: 260.25 days. Higher than high-risk group and test set, suggesting infrequent healthcare visits.
rx_tier_1_pmpm_ct	0.136041	Mean: 0.111. Lower than high-risk group and test set, indicating lower use of tier 1 medications.

Cluster 1:

Feature	SHAP Value	Description
race	0.219375	More diverse: 49.09% White, 35.79% Unknown. Different distribution compared to other clusters.
county_of_residence	0.217104	indicates geographic importance.
unattributed_provider	0.184915	54.94% unattributed, higher than test set (16.21%), suggesting issues with provider continuity.

total_net_paid_pmpm_cost	0.155772	Mean: $741.83. Highest among clusters, indicating high healthcare utilization.
generic_grouper	0.149905	79.70% True, higher than test set (41.11%), suggesting higher generic medication use.

Cluster 2:

Feature	**SHAP Value**	**Description**
veteran_ind	0.709969	100% veterans, significantly higher than test set (11.14%), exclusive veteran population.
total_net_paid_pmpm_cost	0.570227	Mean: $7.90. Significantly lower than all groups, indicating extremely low healthcare utilization.
days_since_last_clm	0.260621	Mean: 339.82 days. Highest among all groups, suggesting very infrequent healthcare visits.
hcc_model_type	0.237655	71.80% Unknown, higher than test set (24.68%), indicating potential gaps in risk assessment.
rx_tier_1_pmpm_ct	0.180210	Mean: 0.029. Lowest among all groups, suggesting very low use of tier 1 medications.

High-Risk Group (no SHAP values, using statistical information):

Feature	**Description**
total_net_paid_pmpm_cost	Mean: $366.04. Lower than test set, but higher than Clusters 0 and 2.
veteran_ind	40.85% veterans, higher than test set (11.14%).
hcc_model_type	61.09% Unknown, higher than test set (24.68%).
days_since_last_clm	Mean: 236.82 days. Higher than test set, indicating less frequent healthcare visits.
rx_tier_1_pmpm_ct	Mean: 0.504. Lower than test set, suggesting lower use of tier 1 medications.

Full Test Dataset (no SHAP values, using statistical information):

Feature	Description
total_net_paid_pmpm_cost	Mean: $576.05. Higher than high-risk group and Clusters 0 and 2, lower than Cluster 1.
veteran_ind	11.14% veterans. Lowest among all groups.
hcc_model_type	24.68% Unknown, 74.54% Medical. More complete risk assessment than other groups.
days_since_last_clm	Mean: 74.71 days. Lowest among all groups, indicating more frequent healthcare visits.
rx_tier_1_pmpm_ct	Mean: 1.309. Highest among all groups, suggesting higher use of tier 1 medications.

These tables provide a comprehensive view of the most influential features for each cluster, along with comparisons to the high-risk group and full test dataset. This information can be used to justify and tailor intervention strategies for each group based on their unique characteristics and risk factors.

7. Conclusion

Our comprehensive analysis of preventive care utilization among Humana's LPPO members has revealed significant insights and opportunities for improving member engagement and health outcomes. Through advanced machine learning techniques, including CatBoost modeling and SHAP analysis, we've identified key factors influencing preventive care gaps and developed targeted strategies to address them.

Key findings include:

1. Complex interplay of factors: Preventive care utilization is influenced by a combination of demographic, geographic, health status, and plan-related factors, underscoring the need for multifaceted intervention strategies.
2. Distinct high-risk clusters: We identified three distinct clusters within the high-risk population, each requiring tailored approaches to improve preventive care engagement.
3. Importance of digital engagement: Members who regularly interact with online services show higher preventive care utilization, highlighting the potential of digital strategies.
4. Veteran-specific challenges: A significant portion of high-risk members are veterans with unique healthcare utilization patterns, necessitating targeted interventions.
5. Cost considerations: Healthcare costs play a complex role in preventive care utilization, with both high and low-cost members at risk of missing preventive visits.

Based on these insights, we've proposed targeted strategies, including a Preventive Care Rebate Program and a Veteran-Centric Care Coordination Program. These data-driven approaches promise significant improvements in preventive care utilization and substantial returns on investment.

Looking forward, implementing these strategies could lead to:

- Improved health outcomes for LPPO members through increased preventive care utilization
- Enhanced CMS Star Ratings for Humana's LPPO plans
- More accurate risk adjustment and appropriate funding
- Strengthened competitive position in the growing Medicare Advantage LPPO market

As the healthcare landscape continues to evolve, Humana's commitment to leveraging data-driven insights will be crucial in addressing the unique challenges of LPPO plans. By focusing on personalized, targeted interventions and embracing digital innovations, Humana can set new standards for preventive care engagement in the Medicare Advantage space.

8. References

Green, B. B., Anderson, M. L., Cook, A. J., Chubak, J., Fuller, S., Kimbel, K. J., Kullgren, J. T., Meenan, R. T., & Vernon, S. W. (2019). Financial Incentives to Increase Colorectal Cancer Screening Uptake and Decrease Disparities. *JAMA Network Open*, *2*(7), e196570. https://doi.org/10.1001/jamanetworkopen.2019.6570

Latkin, C. A., & Knowlton, A. R. (2015). Social Network Assessments and Interventions for Health Behavior Change: A Critical Review. *Behavioral Medicine*, *41*(3), 90–97. https://doi.org/10.1080/08964289.2015.1034645

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: unbiased boosting with categorical features*.