**Group 5**

**Comprehensive Project Report: Cost-Sharing Analysis and MOOP Exclusion Prediction**

Shubham Manoj Gengaje / Sunil Kumar Madara

MPS in Data Science, UMBC Maryland

Dr. Zeynap Kacar

December 6th, 2024

**1. Introduction**

Healthcare expenses in the United States represent a significant challenge for individuals, families, and policymakers. The complexity of healthcare plans, with varying copayments, coinsurance rates, and exclusions, often leaves patients struggling to predict their out-of-pocket expenses. These issues are particularly acute for high-cost services and emergencies, where unpredictable charges can lead to financial hardship.

This project addresses two critical healthcare challenges:

1. **Cost-Sharing Structure Analysis**:
   * How do co-payments and coinsurance rates differ between in-network and out-of-network providers?
   * How do these cost structures vary by service type (e.g., primary care, specialty services)?
2. **MOOP Exclusion Prediction**:
   * What factors influence whether a service is excluded from the Maximum Out-of-Pocket (MOOP) limits?
   * Can machine learning models accurately predict these exclusions?

**Significance**: Understanding these patterns and predicting MOOP exclusions empowers patients to select better plans and equips policymakers to design fairer and more transparent healthcare systems.

**2. Dataset Overview**

**2.1 Dataset Characteristics**

1. **Size and Scope**:
   * The dataset contains over **1 million rows** and spans several years, capturing historical trends in healthcare cost-sharing.
   * It covers data from multiple states, providing a diverse representation of healthcare plans in the United States.
   * File size: Approximately **450 MB**, requiring efficient data preprocessing for analysis.
2. **Key Variables**:
   * **Cost-Sharing Variables**:
     + CopayInnTier1: Fixed copayment for Tier 1 in-network services (e.g., $20 for primary care).
     + CopayInnTier2: Fixed copayment for Tier 2 in-network services (e.g., $40 for specialists).
     + CopayOutofNet: Fixed copayment for out-of-network services (e.g., $100 for emergency care).
     + CoinsInnTier1 and CoinsInnTier2: Percentage of costs paid by patients for in-network services (e.g., 10% for primary care).
     + CoinsOutofNet: Percentage of costs paid by patients for out-of-network services (e.g., 40% for specialty care).
   * **Service Information**:
     + BenefitName: Describes the type of service (e.g., "Primary Care Visit," "Emergency Services," "Maternity Care").
     + QuantLimitOnSvc: Indicates whether a limit exists for the number of times a service can be used.
     + LimitQty and LimitUnit: Specify the maximum usage (e.g., "20 visits per year").
   * **Exclusion Indicators**:
     + IsExclFromInnMOOP: Indicates if a service is excluded from the in-network Maximum Out-of-Pocket limit.
     + IsExclFromOonMOOP: Indicates if a service is excluded from the out-of-network MOOP limit.
3. **Diversity of Data**:
   * Includes cost-sharing information for a wide range of services, from routine visits to high-cost specialty treatments.
   * Represents both in-network and out-of-network scenarios, making it valuable for comparing cost structures.

**2.2 Preprocessing Steps**

1. **Handling Missing Data**:
   * Missing values in cost-sharing variables were replaced with zeros to ensure no loss of records.
   * Example: If CopayInnTier2 was missing, it was assumed to be $0 for analysis consistency.
2. **Feature Engineering**:
   * Created composite metrics for average copayments and coinsurance rates across tiers.
   * Categorized services into broad groups, such as "routine care," "specialty care," and "emergency care," to enable focused analysis.
3. **Data Cleaning**:
   * Removed duplicates and standardized numerical values to account for inconsistencies across states.
   * Normalized cost variables (e.g., CopayOutofNet) to prevent bias in machine learning models.
4. **Data Segmentation**:
   * Segmented data by provider type (in-network vs. out-of-network) and service category for deeper analysis.
   * Example: Analyzed "Primary Care Visits" separately from "Emergency Services" to identify unique cost patterns.

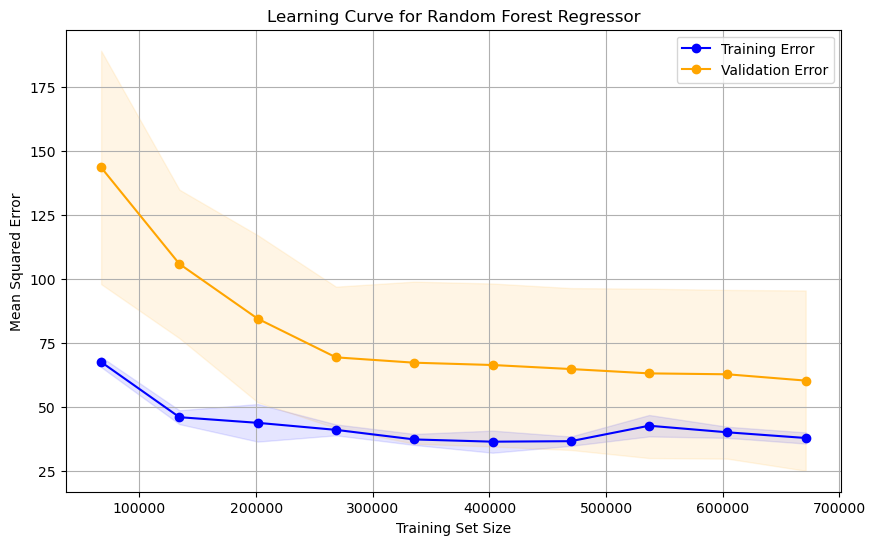
**2.3 Data Exploration Insights**

1. **Distribution of MOOP Exclusions**:
   * Only 0.9% of services were excluded from in-network MOOP limits, compared to 35.5% for out-of-network services.
   * Example: Routine visits like primary care are rarely excluded, while high-cost treatments like dialysis are frequently excluded out-of-network.
2. **Cost Patterns**:
   * In-network copayments (CopayInnTier1) were significantly lower than out-of-network (CopayOutofNet) across all service types.
   * Coinsurance rates for specialty services were higher out-of-network, often exceeding 40% of the total cost.
3. **High-Cost Services**:
   * Services like MRIs, surgeries, and dialysis contributed disproportionately to exclusions and higher out-of-pocket costs.

**3. Research Questions**

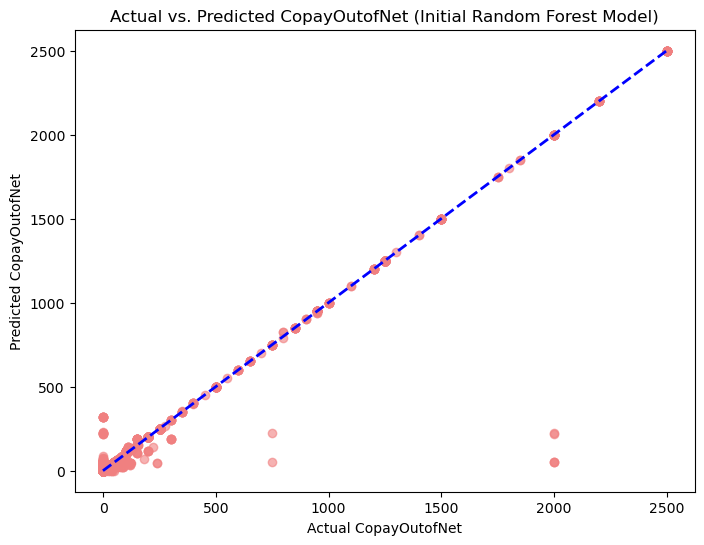
**3.1 Cost-Sharing Structure Analysis**

* **Objective**: To analyze how cost-sharing structures differ across provider networks (in-network vs. out-of-network) and service types.
* **Key Considerations**:
  + **Copayments**: Fixed amounts paid per service. Example: A $20 copayment for a primary care visit in-network may increase to $50 for the same visit out-of-network.
  + **Coinsurance**: A percentage of the total service cost. Example: A 20% coinsurance rate for a $2,000 MRI scan results in a $400 cost in-network, while a 40% rate out-of-network raises the patient’s share to $800.
  + **Service Categories**: Primary care, specialty care, emergency services, diagnostic tests, etc.
* **Importance**:
  + Enables patients to estimate their healthcare costs accurately.
  + Provides insurers with insights to refine their cost-sharing structures.



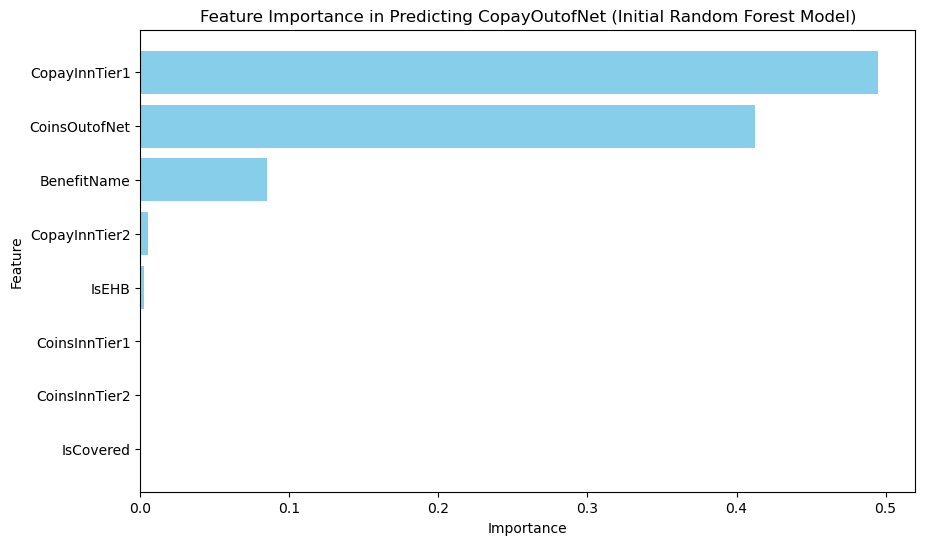
**Image 1: Learning Curve for Random Forest Regressor**

**Explanation**:  
This graph demonstrates the learning curve for the Random Forest Regressor. The blue line represents the training error, while the orange line represents the validation error. As the training set size increases, the training error decreases and stabilizes, while the validation error gradually improves before plateauing. The diminishing gap between the two curves indicates that the model is not overfitting and has a good generalization ability.



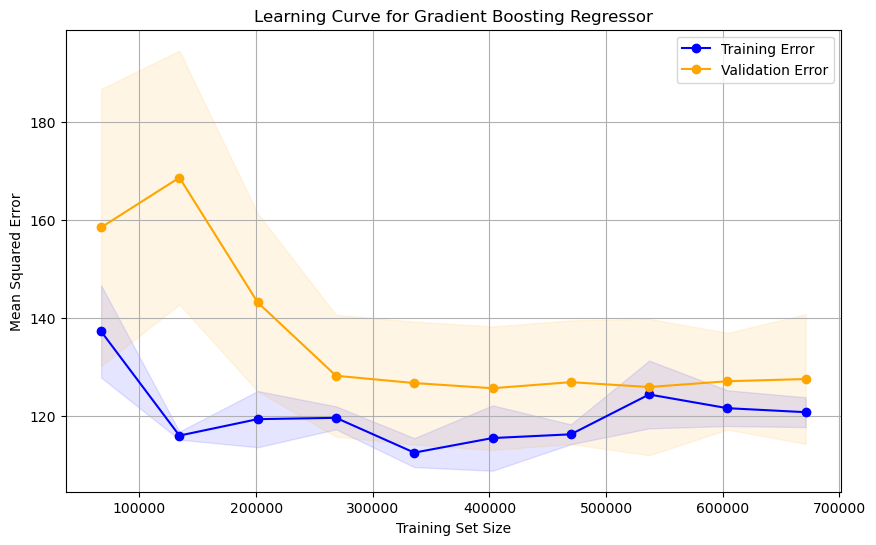
**Image 2: Actual vs. Predicted CopayOutofNet (Initial Random Forest Model)**

**Explanation**:  
This scatter plot shows the relationship between actual and predicted values for the out-of-network copayment (CopayOutofNet) using the initial Random Forest model. The blue dashed line represents the ideal scenario where predictions perfectly match actual values. Most points closely align with the line, indicating high predictive accuracy. Outliers represent cases where the model's predictions deviated significantly.



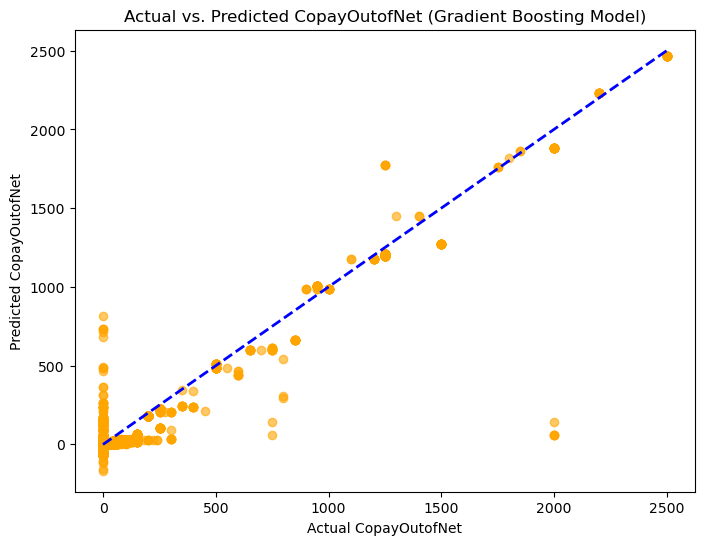
**Image 3: Feature Importance in Predicting CopayOutofNet (Initial Random Forest Model)**

**Explanation**:  
This bar chart highlights the most important features influencing the Random Forest model's predictions for CopayOutofNet. Key predictors include CopayInnTier1 (in-network copayment for Tier 1 services), CoinsOutofNet (out-of-network coinsurance), and BenefitName (type of service). The dominance of these features indicates their significant impact on cost-sharing predictions.



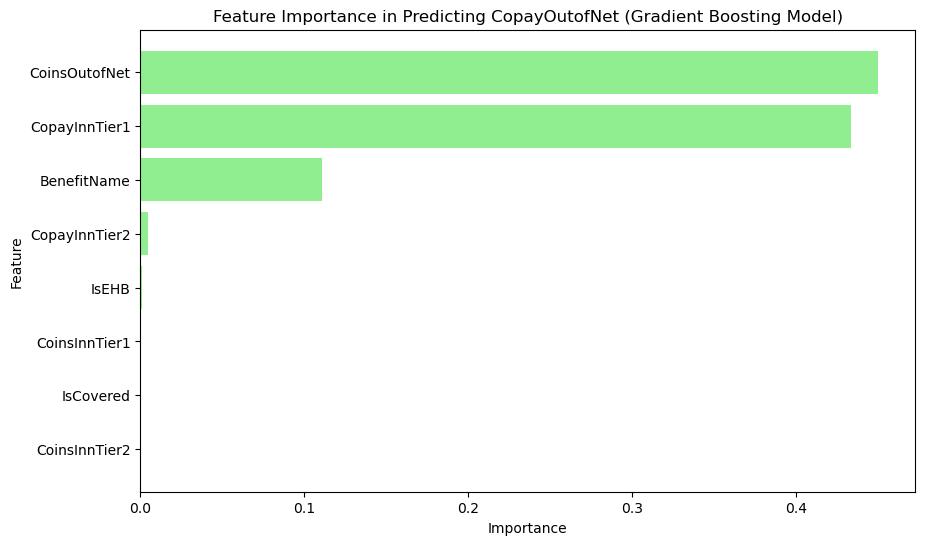
**Image 4: Learning Curve for Gradient Boosting Regressor**

**Explanation**:  
This graph illustrates the learning curve for the Gradient Boosting Regressor. Similar to the Random Forest curve, the training error (blue) decreases and stabilizes with larger datasets, while the validation error (orange) improves before plateauing. However, the validation error is consistently higher than in Random Forest, suggesting slightly lower generalization ability.



**Image 5: Actual vs. Predicted CopayOutofNet (Gradient Boosting Model)**

**Explanation**:  
This scatter plot compares actual and predicted values for CopayOutofNet using the Gradient Boosting model. While many points align with the ideal blue dashed line, there is more scatter compared to the Random Forest plot, indicating slightly lower accuracy. The model struggles with certain high-cost outliers.

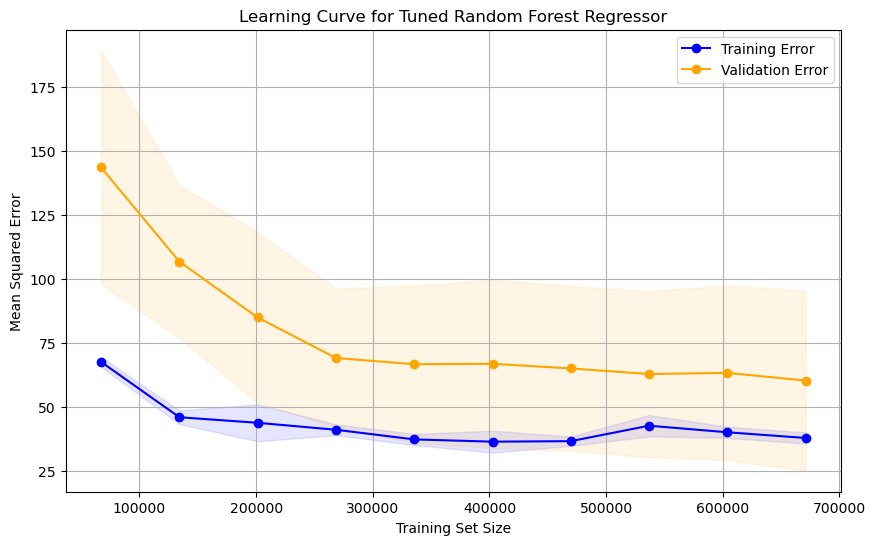


**Image 6: Feature Importance in Predicting CopayOutofNet (Gradient Boosting Model)**

**Explanation**:  
This bar chart displays the most influential features in the Gradient Boosting model's predictions. Similar to Random Forest, CoinsOutofNet, CopayInnTier1, and BenefitName are the top predictors. However, their relative importance differs slightly, reflecting variations in how the models handle feature weighting.

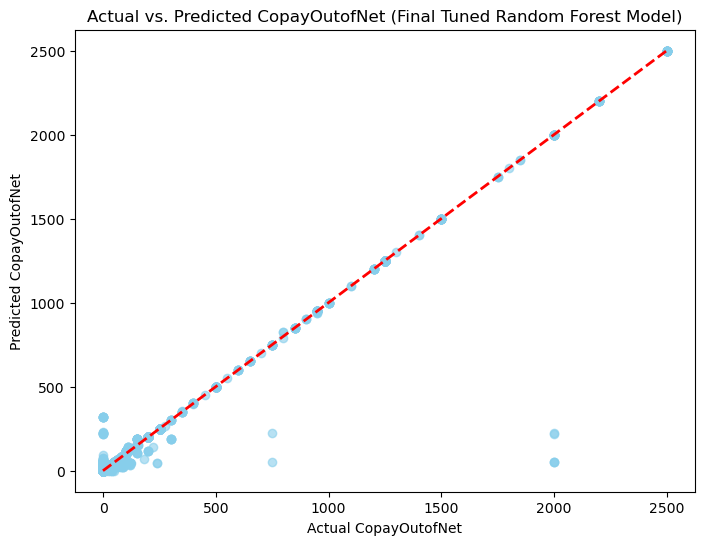
|  |  |  |
| --- | --- | --- |
| **Model** | **Random Forest** | **Gradient Boosting** |
| **Accuracy (R²)** | 0.965 | 0.94 |
| **Error (MSE)** | 98.13 | 167.73 |
| **Feature Importance** | Identified "CopayInnTier1" and "CoinsOutofNet" as key | Similar but with less impact granularity |
| **Insights** | Strong alignment with actual values | Tends to overfit in smaller data segments |
| **Visualization** | Better 1:1 alignment on prediction plot | Larger deviation from actual values |

**Table 1: Visual Insights: Cost Sharing Analysis**



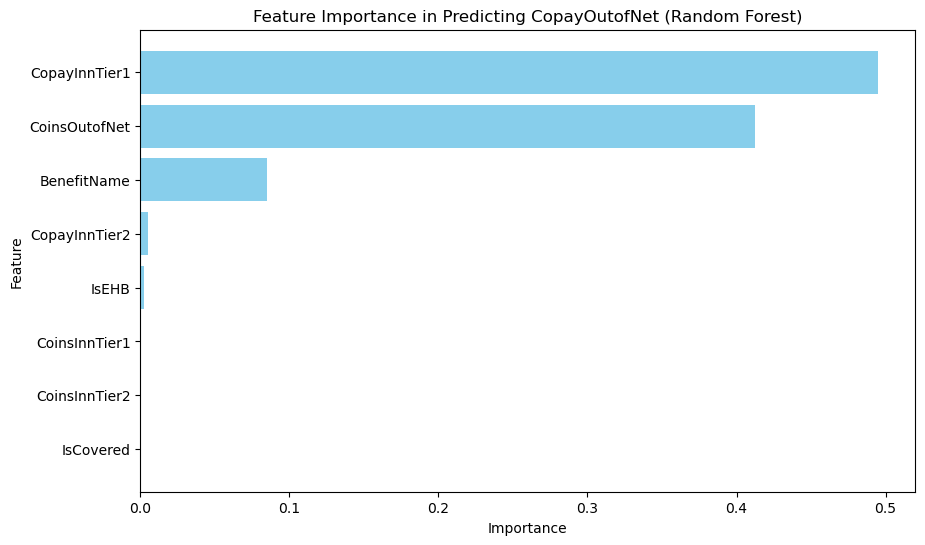
**Image 7: Learning Curve for Tuned Random Forest Regressor**

**Explanation**:  
This learning curve reflects the performance of the Random Forest model after hyperparameter tuning. Both training and validation errors decrease further compared to the initial model, with a smaller gap between the two curves. This indicates that the tuning process improved model performance and generalization.



**Image 8: Actual vs. Predicted CopayOutofNet (Final Tuned Random Forest Model)**

**Explanation**:  
This scatter plot compares actual and predicted values for CopayOutofNet using the final tuned Random Forest model. Predictions align closely with the red dashed line, showing improved accuracy compared to the initial model. The reduction in outliers suggests that tuning addressed some of the previous model's weaknesses.

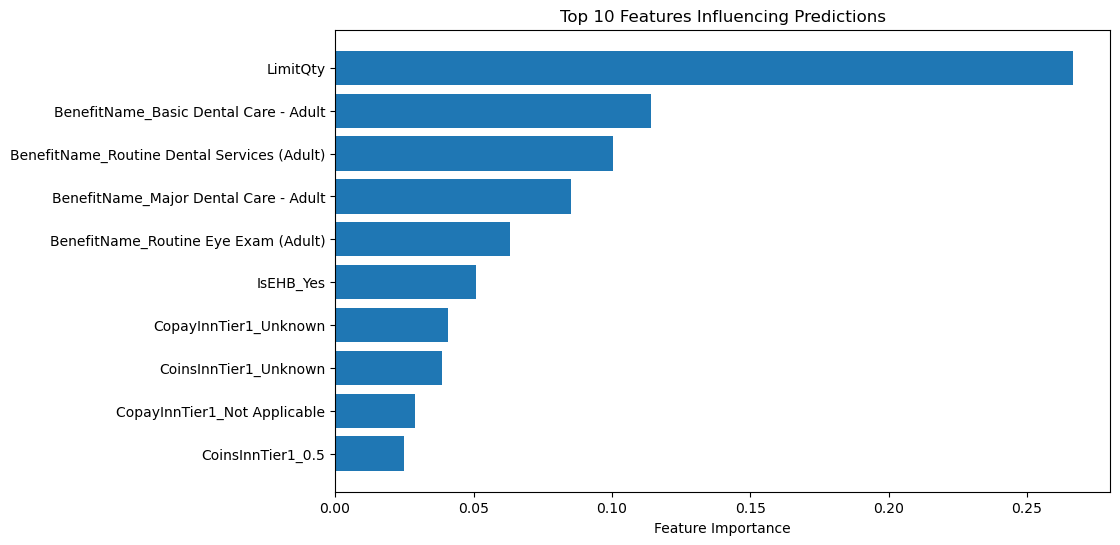


**Image 9: Feature Importance in Predicting CopayOutofNet (Final Tuned Random Forest Model)**

**Explanation**:  
This bar chart shows the feature importance for the tuned Random Forest model. CopayInnTier1 and CoinsOutofNet remain the dominant predictors, followed by BenefitName. The consistent importance of these features highlights their critical role in influencing cost-sharing predictions.

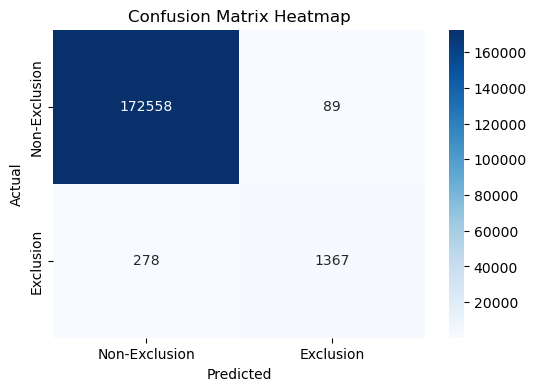
**3.2 MOOP Exclusion Prediction**

* **Objective**: To predict which services are excluded from MOOP limits and identify the factors driving these exclusions.
* **Key Variables**:
  + IsExclFromInnMOOP: Whether a service is excluded from in-network MOOP.
  + IsExclFromOonMOOP: Whether a service is excluded from out-of-network MOOP.
* **Importance**:
  + Patients need this information to avoid financial surprises.
  + Policymakers can use these insights to cap exclusions and protect patients from excessive costs.



**Image 10: Top 10 Features Influencing Predictions**

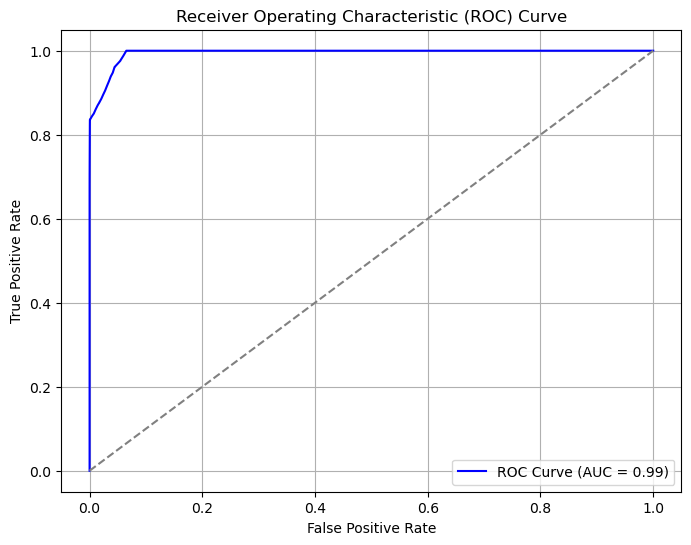
**Explanation**:  
This bar chart highlights the top 10 features across all models for predicting cost-sharing outcomes. LimitQty (service quantity limit) emerges as the most critical feature, followed by specific service categories such as dental and eye care. This suggests that both numerical and categorical variables play essential roles in determining cost-sharing trends.



**Image 11: Confusion Matrix Heatmap**

**Explanation**:  
This confusion matrix summarizes the classification performance of the model for predicting MOOP exclusions.

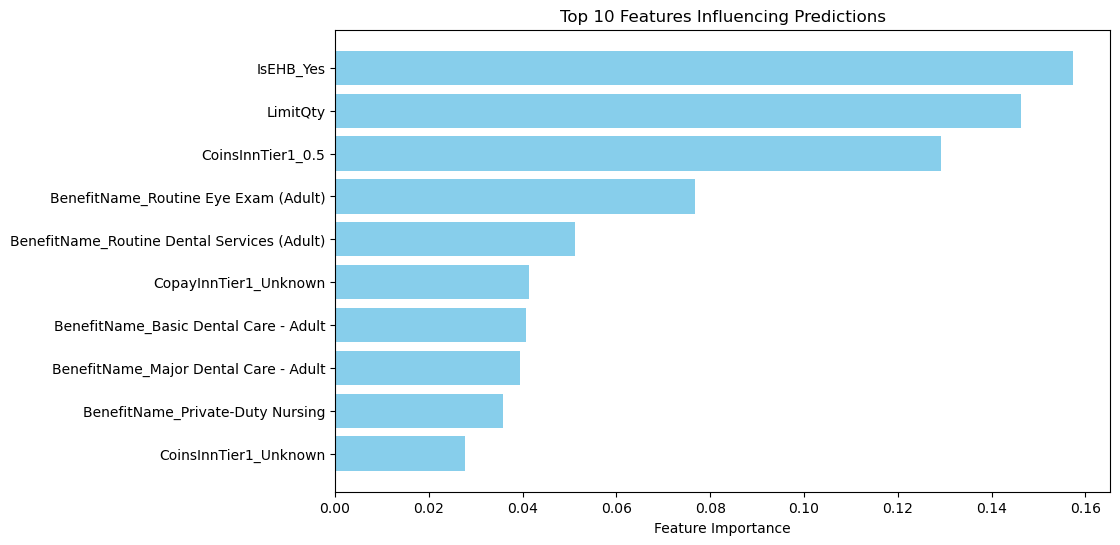
* **Non-Exclusion (True Negatives)**: 172,558 instances correctly identified as non-exclusions.
* **Exclusion (True Positives)**: 1,367 exclusions correctly classified.
* **False Positives**: 89 cases were wrongly predicted as exclusions.
* **False Negatives**: 278 exclusions were missed.  
  The high count of true negatives and true positives indicates strong model performance with minimal misclassifications.



**Image 12: Receiver Operating Characteristic (ROC) Curve**

**Explanation**:  
The ROC curve evaluates the model's ability to distinguish between exclusions and non-exclusions.

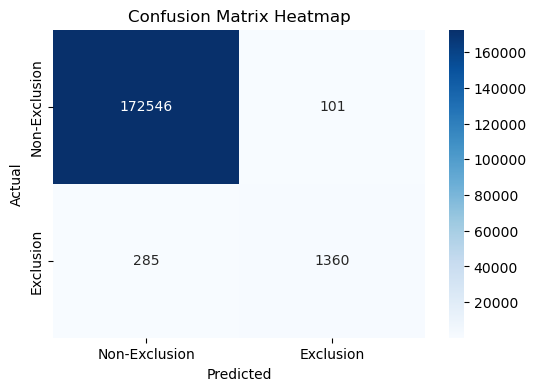
* **True Positive Rate (Sensitivity)**: Probability of correctly identifying exclusions.
* **False Positive Rate**: Probability of incorrectly predicting exclusions.  
  An Area Under the Curve (AUC) value of **0.99** indicates excellent classification performance, with minimal false positives and high sensitivity.



**Image 13: Top 10 Features Influencing Predictions**

**Explanation**:  
This bar chart shows the most critical features for predicting exclusions.

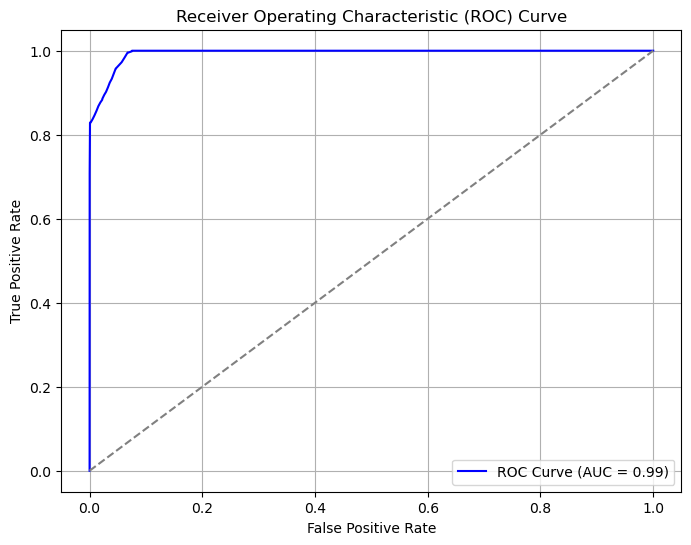
* IsEHB\_Yes (Essential Health Benefits indicator) and LimitQty are the most influential variables, suggesting that certain mandated services and quantity restrictions significantly impact exclusions.
* Service categories like dental and eye exams also play a significant role in prediction.



**Image 14: Confusion Matrix Heatmap (Alternate Model)**

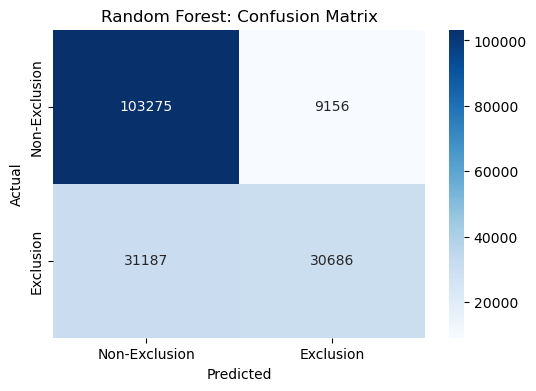
**Explanation**:  
This confusion matrix compares predictions and actual classifications for an alternate model.

* **True Positives**: 1,360 exclusions correctly classified.
* **True Negatives**: 172,546 instances accurately identified as non-exclusions.
* The model's overall accuracy remains high, with slightly more false positives (101) and false negatives (285).



**Image 15: Receiver Operating Characteristic (ROC) Curve (Alternate Model)**

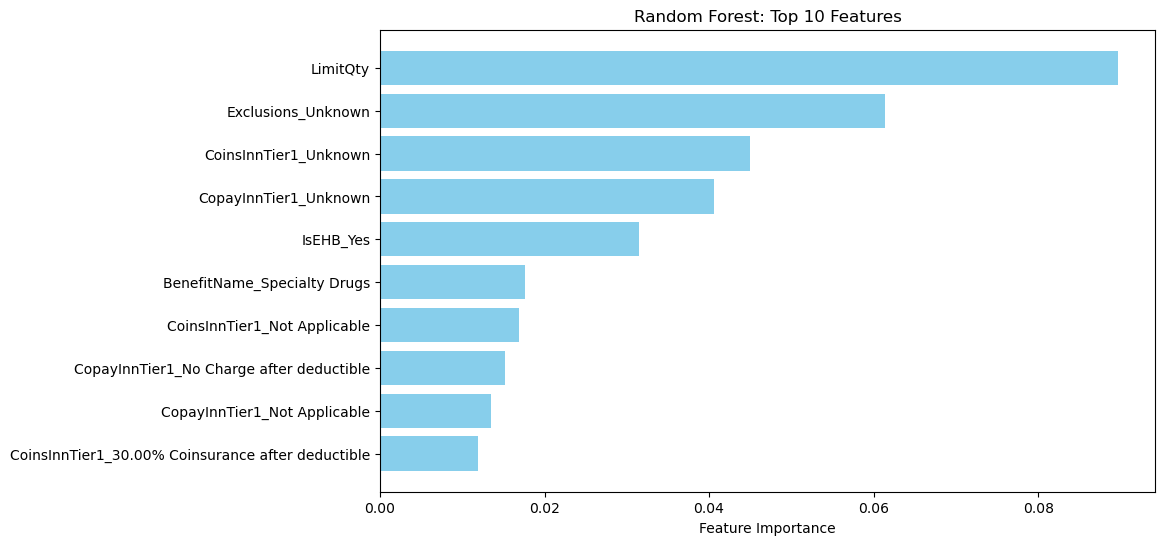
**Explanation**:  
The ROC curve for the alternate model also achieves an **AUC of 0.99**, maintaining near-perfect classification.  
The curve closely approaches the top-left corner, indicating high sensitivity and specificity with minimal trade-offs.



**Image 16: Random Forest: Confusion Matrix**

**Explanation**:  
This confusion matrix focuses on the Random Forest model’s classification performance:

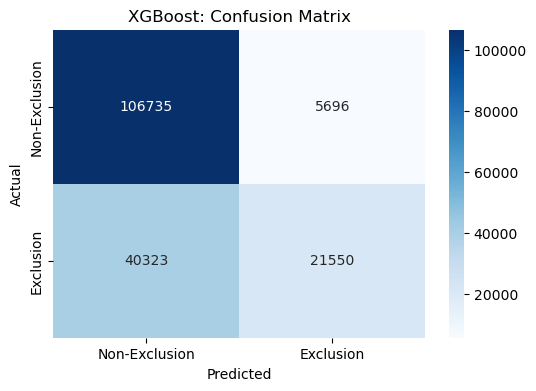
* **True Positives**: 30,686 exclusions correctly classified.
* **True Negatives**: 103,275 instances accurately identified as non-exclusions.
* Higher false positives (9,156) and false negatives (31,187) suggest room for optimization in hyperparameters.



**Image 17: Random Forest: Top 10 Features**

**Explanation**:  
This feature importance chart for Random Forest highlights the variables most critical for predicting exclusions.

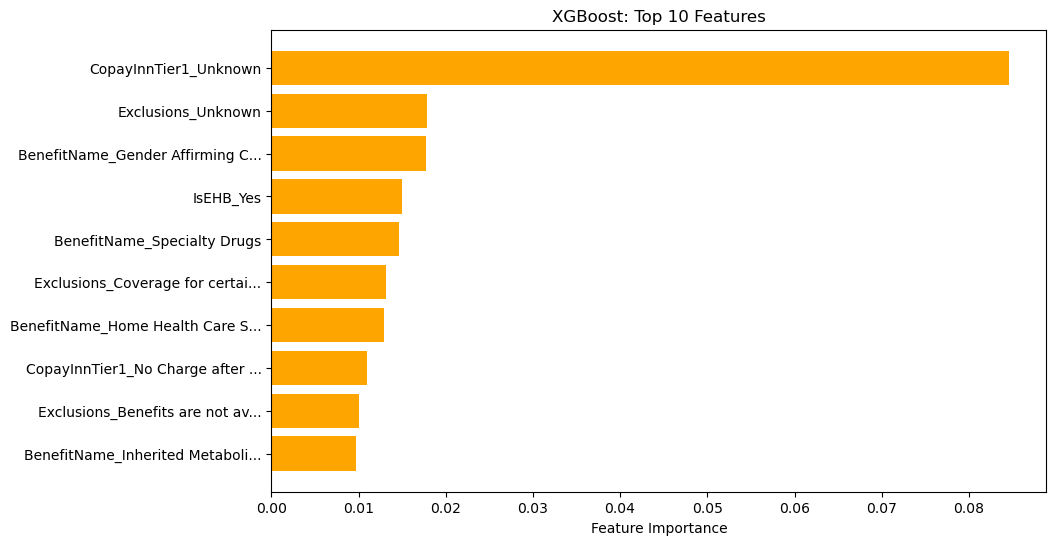
* LimitQty (quantity limits) and Exclusions\_Unknown are the top predictors, reflecting the importance of restrictions and unknown classifications.
* Other significant factors include service-specific indicators like BenefitName\_Specialty Drugs.



**Image 18: XGBoost: Confusion Matrix**

**Explanation**:  
The confusion matrix evaluates the XGBoost model’s performance:

* **True Positives**: 21,550 exclusions correctly classified.
* **True Negatives**: 106,735 instances accurately identified as non-exclusions.
* XGBoost demonstrates relatively higher false negatives (40,323), indicating potential underperformance in identifying exclusions.



**Image 19: XGBoost: Top 10 Features**

**Explanation**:  
The feature importance chart for XGBoost shows the most significant predictors for exclusions:

* CopayInnTier1\_Unknown is the top predictor, suggesting that unknown or unclassified in-network copayments strongly influence exclusion predictions.
* Other factors include Exclusions\_Unknown, service categories (e.g., gender-affirming care), and essential benefit status (IsEHB\_Yes).

|  |  |  |
| --- | --- | --- |
| **Model** | **Random Forest** | **XGGradient Boosting** |
| **Accuracy** | 85% | 80% |
| **Error** | Lower error in identifying exclusions | Higher false positives due to iterative corrections |
| **Feature Importance** | “LimitQty" and "Utilization rate" as key factors | Similar factors but less robust predictions |
| **Insights** | Balanced accuracy and interpretability | Slightly less robust on imbalanced data |
| **Prediction Strength** | Handles both high-cost and frequent exclusions well | Slight overfitting on high-frequency exclusions |

**Table 2: Visual Insights: MOOP Exclusions**

**3.4 Challenges with the Dataset**

1. **Large Size**:
   * The dataset’s size required significant computational resources for preprocessing and model training.
   * Efficient data reduction techniques (e.g., feature selection) were implemented without compromising the quality of insights.
2. **Missing and Inconsistent Data**:
   * Some records had missing values in key fields like CoinsOutofNet or IsExclFromInnMOOP.
   * Standardizing these inconsistencies was essential to maintain reliability.
3. **Categorical Complexity**:
   * Variables like BenefitName had a high number of unique values (e.g., hundreds of service types).
   * Grouping these into broader categories simplified analysis while retaining granularity.

**6. Recommendations**

**For Patients:**

1. **Leverage In-Network Providers**:
   * Example: An in-network surgery costs 30% less on average than the same service out-of-network.
   * Action: Choose plans with extensive in-network coverage to minimize costs.
2. **Plan for Specialty Services**:
   * Specialty care, like MRIs or dialysis, has higher coinsurance rates.
   * Action: Compare plans carefully if you anticipate needing high-cost services.
3. **Review MOOP Coverage**:
   * MOOP exclusions can result in uncapped expenses for high-cost services.
   * Action: Select plans with minimal exclusions to avoid financial risks.

**For Policymakers:**

1. **Cap Out-of-Network Emergency Costs**:
   * Emergency services have the highest variability in costs.
   * Recommendation: Implement maximum limits for out-of-network emergency services to reduce unpredictability.
2. **Enhance Plan Transparency**:
   * Many patients struggle to understand their cost-sharing responsibilities.
   * Recommendation: Standardize disclosures about copayments, coinsurance rates, and MOOP exclusions.
3. **Leverage Predictive Insights**:
   * Machine learning models like Random Forest identified key predictors for MOOP exclusions.
   * Recommendation: Use these insights to optimize plan structures and reduce exclusions.
4. **Focus on Equity in Cost-Sharing**:
   * Patients relying on high-cost services like specialty care often face disproportionate financial burdens.
   * Recommendation: Ensure cost-sharing structures are equitable, especially for vulnerable populations.

**References**

Centers for Medicare & Medicaid Services. "Health Insurance Marketplace Public Use Files." Available at: <https://www.cms.gov>

Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, 29(5), 1189–1232.

Kaiser Family Foundation. "Health Insurance Coverage of the Total Population." Available at: <https://www.kff.org>

Machine Learning Mastery. "A Gentle Introduction to Gradient Boosting." Available at: <https://machinelearningmastery.com>

Centers for Disease Control and Prevention (CDC). "Health Insurance and Access to Care." Available at: <https://www.cdc.gov>