STATS 506 - Final Project - Report

Shenyi Tang

December 2024

1 GitHub Repository

Shenyi Tang's STATS 506 GitHub Repo: https://github.com/shenyi-tang/stats506-computing-methods-and-tools.git

2 Data Introduction

The Project use two datasets of Year 2020, "Medicare Physicians and Other Practitioners-by Providers and Services" and "Individual Income Tax Statistics by Zip Code Level".

The medicare dataset covers healthcare provider and service information. It includes provider demographics, locations, service types, beneficiary counts, and payment information. The data captures both individual and organizational providers, medical services rendered, and various payment metrics including submitted charges, allowed amounts, and standardized Medicare payments that account for geographic variations.

The SOI tax dataset contains comprehensive tax return information by zip code and AGI (Adjusted Gross Income) level in Tax Year 2020, focusing on income sources, deductions, credits, and tax payments. The data includes taxpayer demographics, filing statuses, income brackets, and details about COVID-19 related payments.

3 Research Question

The project focuses on the factors driving high medicare payments amount for pediatric medicine in different income level (categorized by low, median and high income level by AGI value). Factors may come from the medicare information itself or from the local tax related information.

PS: In the proposal, I propose to choose the cardiology-related medical services. However, due to the high volume of data in cardiac medical services increase the calculating time sharply, I use pediatric medicine part instead.

4 Research Approaches

During the data cleaning and merging steps, all the amount-related variables started with "A" in SOI tax data are aggregated to `STATEFIPS - Zip Code` level using the variable `N2`, number of individuals, as weight. For the variable `AGI_STUB`, which refers to the size of adjusted gross income, it is aggregated to the `STATEFIPS - Zip Code` level according to the times appeared in a zip code based on `N2`, that is, taking the mode.

For the variables selected for the feature importance model, they covers almost all the variables in the medicare dataset except addresses and standardized payments which is highly correlated with the medicare payment. In the selection of variables in the tax dataset, variables related with children, education and medicine are specifically chosen besides some general tax and income variables.

The analytical framework XGBoost to do the feature selection. A random search strategy is used to explore 70 different combinations of parameters across key model parameters including tree depth (3-10), learning rate (0.01-0.3), and sampling ratios (0.6-1.0).he model's performance was validated using 5-fold cross-validation with early stopping mechanisms to prevent overfitting. This approach ensured robust model performance while maintaining computational efficiency.

Post-model development, the analysis deep dived into the the relationship between various factors and high payment probability. High payments were defined using the 75th percentile threshold of Medicare payment amounts. The prediction ability of each feature was evaluated using logistic regression models and quantified through AUC (Area Under the Curve) metrics. For continuous variables, adaptive binning techniques were implemented using quintile-based breaks or equal-width intervals when appropriate. This resulted in standardized categories of "Very Low," "Low," "Medium," "High," and "Very High" for all continuous variables. For each feature category, the probability of high payments was calculated along with its standard error, accounting for sample size variations. These relationships were visualized through bar plots with error bars, where the height of each bar represents the probability of high payments for that category

To investigate how these relationships varied across income levels, the data was stratified into three income groups (low, medium, and high) based on `AGI_STUB` values. Same approached mentioned above are utilized in each income group. This stratified analysis provided insights into how the importance and impact of various features shifted across different economic contexts.

5 Results

The top 20 importance of features related to the medicare payment amount of pediatric medicine is illustrated as Figure. 1

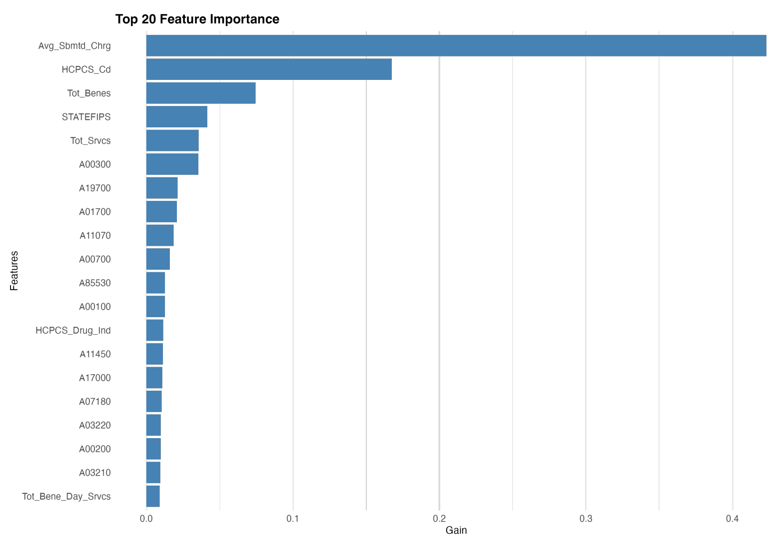


Figure 1: Feature Importance

The analysis reveals complex patterns in the factors influencing Medicare payments for pediatric medicine across different income levels. Submitted charges (Avg_Sbmtd_Chrg) emerge as the dominant predictor, explaining 42.3% of the variation in Medicare payments, with remarkably high predictive power across all income groups (AUC 0.89). This suggests that provider billing practices are the primary driver of payment variations, regardless of local income levels.

For tax related variables, Charitable Contributions (A19700) demonstrate meaningful predictive power (Gain = 0.021), particularly in higher-income areas. This might reflect the relationship between community wealth and healthcare infrastructure quality. Similarly, Taxable Tensions and Tannuities (A01700) show influence (Gain = 0.0209) to the medicare important, suggesting that areas with higher retirement income may have distinct patterns in pediatric Medicare payments.

Table 1: Feature Predicting Ability by Income Level (AUC Values)

Feature	High	Low
Avg_Sbmtd_Chrg	0.906	0.896
HCPCS_Cd	0.702	0.640
Tot_Srvcs	0.555	0.565
Tot_Bene_Day_Srvcs	0.542	0.550
A11450	0.531	0.549
A85530	0.546	0.545
A11070	0.561	0.542
A17000	0.542	0.530
A07180	0.550	0.529
A00200	0.544	0.525
A19700	0.538	0.524
A01700	0.533	0.522
A00300	0.563	0.521
A00100	0.544	0.520
A03210	0.555	0.516
A03220	0.561	0.515
STATEFIPS	0.548	0.511

For output of stratified income groups, due to my processing to the `AGI_STUB` at the beginning, there is no `Median` income group for the regions having pediatric medicine services. Among these variables, Qualified sick and family leave credit amount(A11450), Additional Child Tax Credit Amount(A11070) and Educator Expenses Amount(A03200) shows AUC above or around 0.55 both in high income and low income groups, which indicating that taxes amounts related to family, children and education indeed contribute to the Medicare payments.

These findings suggest that tax-related variables serve as important indicators of community characteristics that influence Medicare payments, which could be valuable for policymakers in designing more equitable healthcare payment systems that account for local economic conditions.

6 Attribution of Sources

* Using LLM. Model Claude to go through the data dictionaries of two datasets and to understand US Medicare system.

7 Appendix

1. The table of feature Import Analysis Results

Table 2: Feature Important Analysis Results

Feature	Gain	Cover	Frequency
Avg_Sbmtd_Chrg	0.4231	0.2375	0.1300
HCPCS_Cd	0.1674	0.1394	0.1314
Tot_Benes	0.0744	0.0347	0.0888
STATEFIPS	0.0415	0.0328	0.0394
Tot_Srvcs	0.0356	0.0369	0.0805
A00300	0.0355	0.0332	0.0321
A19700	0.0212	0.0321	0.0289
A01700	0.0209	0.0255	0.0290
A11070	0.0186	0.0399	0.0355
A00700	0.0160	0.0332	0.0343
A85530	0.0126	0.0276	0.0251
A00100	0.0126	0.0207	0.0318
HCPCS_Drug_Ind	0.0115	0.0071	0.0049
A11450	0.0112	0.0317	0.0286
A17000	0.0109	0.0255	0.0244
A07180	0.0104	0.0418	0.0274
A03220	0.0097	0.0287	0.0262
A00200	0.0097	0.0217	0.0249
A03210	0.0095	0.0330	0.0312
Tot_Bene_Day_Srvcs	0.0090	0.0314	0.0435
A02500	0.0086	0.0260	0.0280
A07225	0.0084	0.0320	0.0246
agi_stub	0.0075	0.0039	0.0097
Rndrng_Prvdr_Crdntls	0.0072	0.0142	0.0178
Place_Of_Srvc	0.0049	0.0067	0.0098
Rndrng_Prvdr_Gndr	0.0021	0.0030	0.0125

2. The table of features predicting abilities without stratification (AUC values)

Table 3: Feature Predicting Ability (Top 20)

Rank	Feature	AUC	Feature Type
1	Avg_Sbmtd_Chrg	0.897	binned
2	HCPCS_Cd	0.667	original
3	Tot_Srvcs	0.559	binned
4	A01700	0.547	binned
5	Tot_Bene_Day_Srvcs	0.543	binned
6	A00300	0.533	binned
7	A11070	0.532	binned
8	A03210	0.530	binned

Continued on next page

Table 3 continued

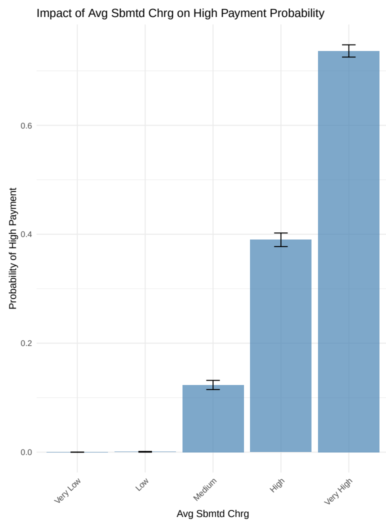
Rank	Feature	AUC	Feature Type
9	Tot_Benes	0.528	binned
10	A11450	0.526	binned
11	A17000	0.526	binned
12	STATEFIPS	0.525	original
13	A85530	0.524	binned
14	A07180	0.524	binned
15	A00200	0.523	binned
16	A03220	0.523	binned
17	A19700	0.520	binned
18	A00700	0.518	binned
19	HCPCS_Drug_Ind	0.517	original
20	A00100	0.514	binned

3. The full table of features predicting abilities by income level (AUC values) (Top 20)

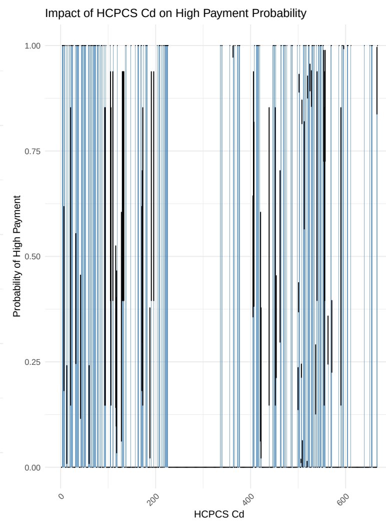
Table 4: Feature Predicting Ability by Income Level (Top 20)

No.	Feature	High	Low
1	Avg_Sbmtd_Chrg	0.906	0.896
2	HCPCS_Cd	0.702	0.640
3	Tot_Srvcs	0.555	0.565
4	Tot_Bene_Day_Srvcs	0.542	0.550
5	A11450	0.531	0.549
6	A85530	0.546	0.545
7	A11070	0.561	0.542
8	Tot_Benes	0.523	0.537
9	A17000	0.542	0.530
10	A07180	0.550	0.529
11	A00200	0.544	0.525
12	A19700	0.538	0.524
13	A01700	0.533	0.522
14	HCPCS_Drug_Ind	0.511	0.521
15	A00300	0.563	0.521
16	A00100	0.544	0.520
17	A03210	0.555	0.516
18	A03220	0.561	0.515
19	A00700	0.515	0.513
20	STATEFIPS	0.548	0.511

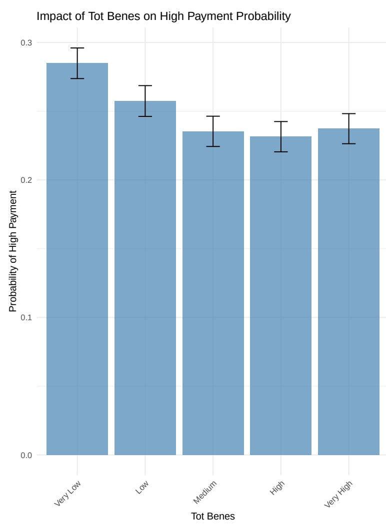
4. The impact of variables on high payment probability



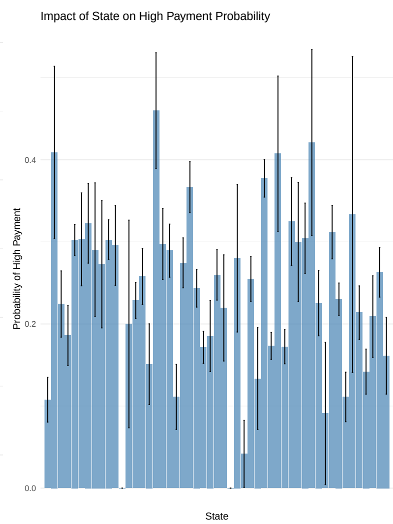
Impact of Avg Sbmtd Chrg on High Payment Probability
Probability of High Payment
0.0
0.2
0.4
0.6
Very Low
Low
Medium
High
Very High
Avg Sbmtd Chrg



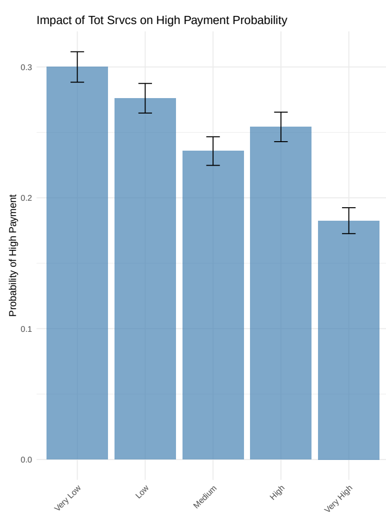
Impact of HCPCS Cd on High Payment Probability
Probability of High Payment
0.00
0.25
0.50
0.75
1.00
0
200
400
600
HCPCS Cd



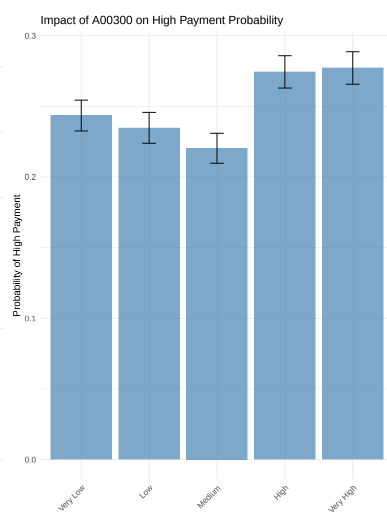
Impact of Tot Benes on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Benes



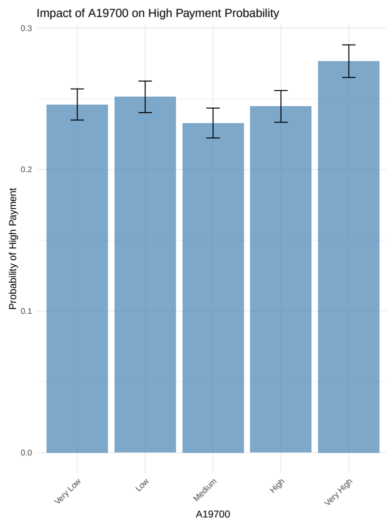
Impact of State on High Payment Probability
Probability of High Payment
0.0
0.2
0.4
State

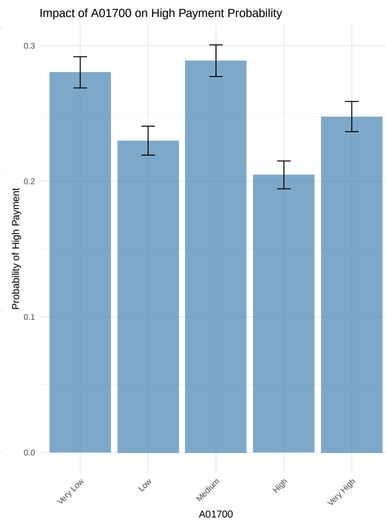


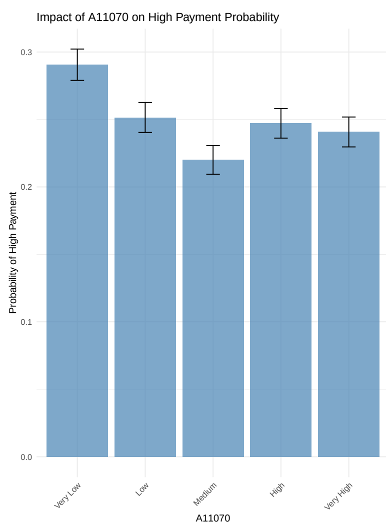
Impact of Tot Srvcs on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Srvcs

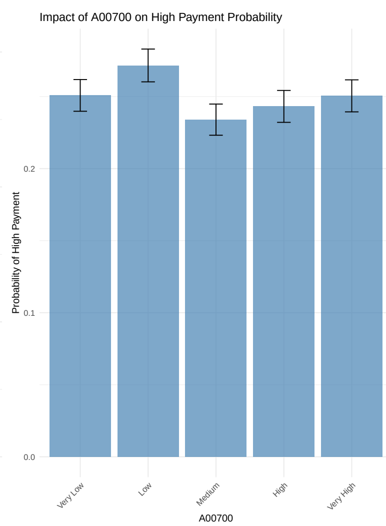


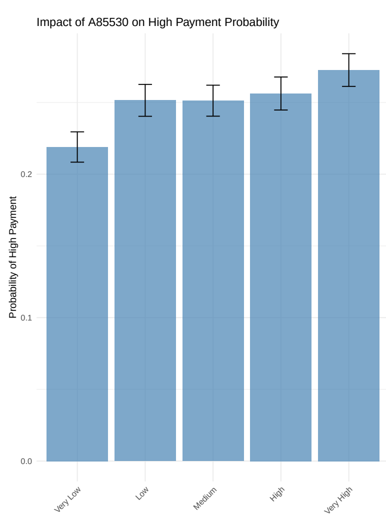
Impact of A00300 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A00300

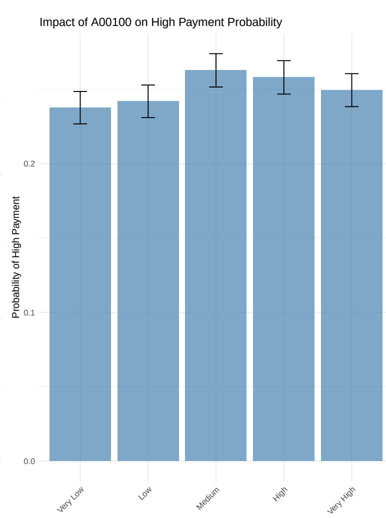

Impact of A19700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A19700

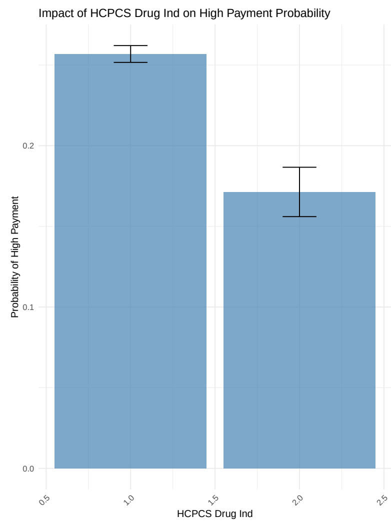

Impact of A01700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A01700

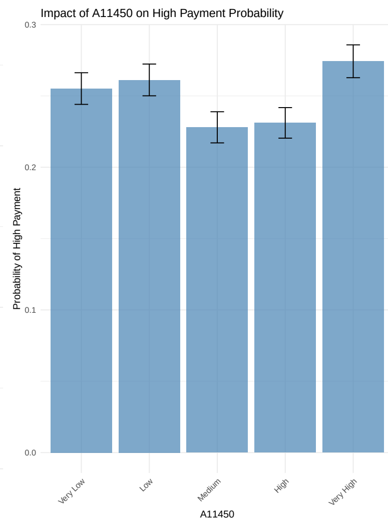

Impact of A11070 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A11070

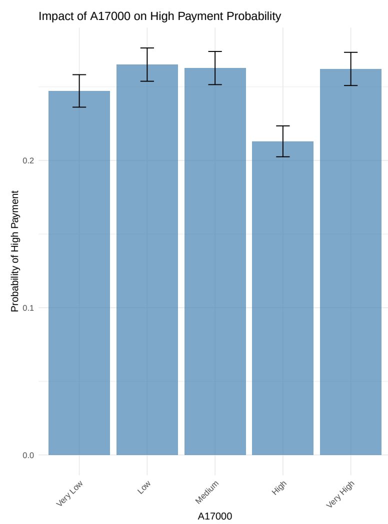

Impact of A00700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A00700

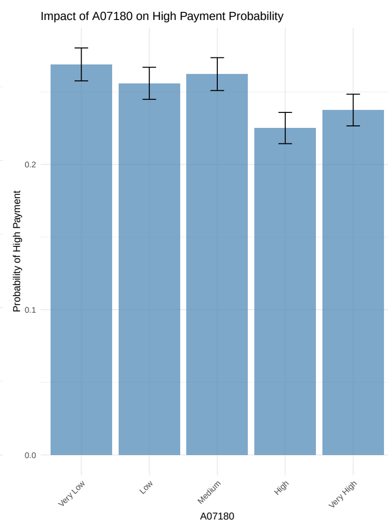

Impact of A85530 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A85530

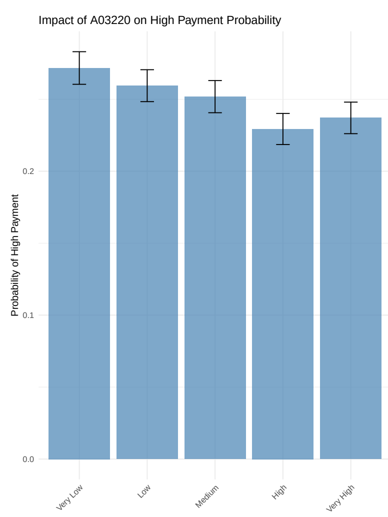

Impact of A00100 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A00100

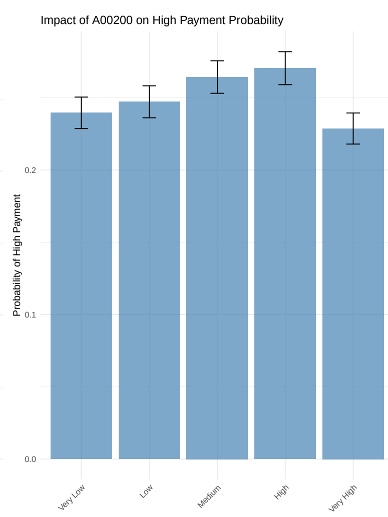

Impact of HCPCS Drug Ind on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.5
1.0
1.5
2.0
2.5
HCPCS Drug Ind

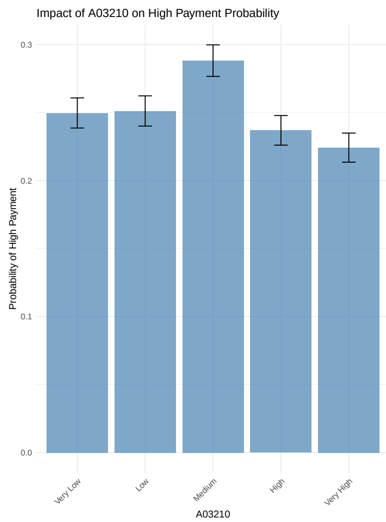

Impact of A11450 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A11450

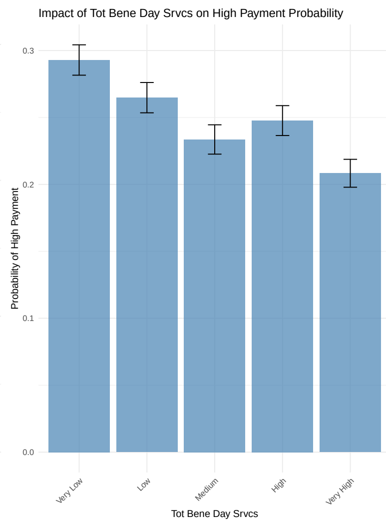

Impact of A17000 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A17000

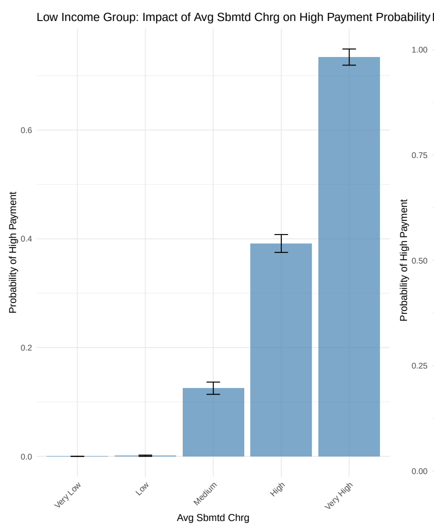

Impact of A07180 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A07180

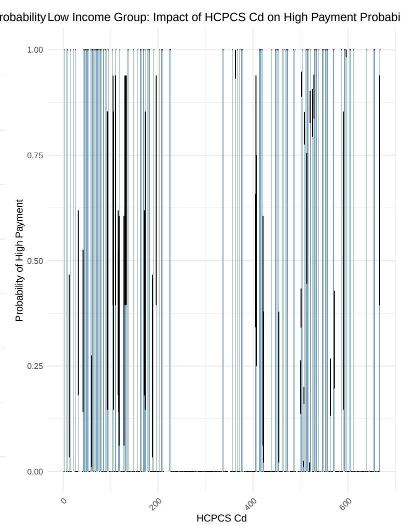

Impact of A03220 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A03220

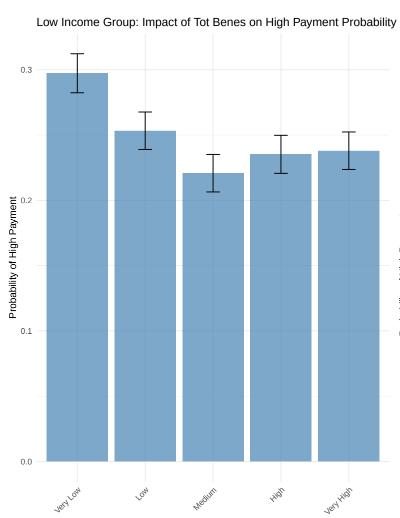

Impact of A00200 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A00200

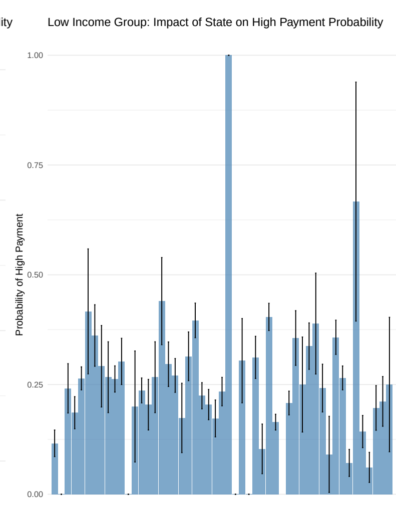

Impact of A03210 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A03210

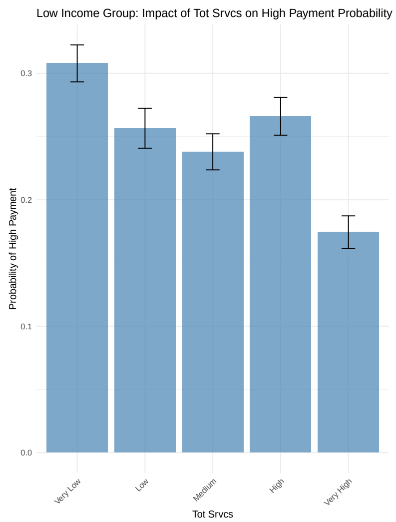

Impact of Tot Bene Day Srvcs on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Bene Day Srvcs

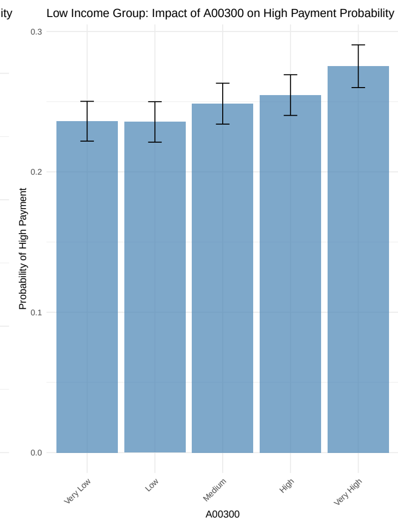

Low Income Group: Impact of Avg Sbmtd Chrg on High Payment Probability
Probability of High Payment
0.0
0.2
0.4
0.6
Very Low
Low
Medium
High
Very High
Avg Sbmtd Chrg

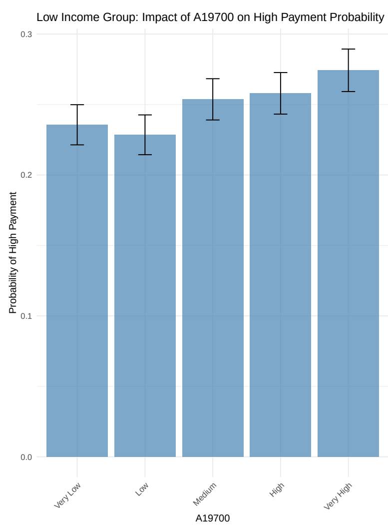

Low Income Group: Impact of HCPCS Cd on High Payment Probabi
Probability of High Payment
0.00
0.25
0.50
0.75
1.00
0
200
400
600
HCPCS Cd

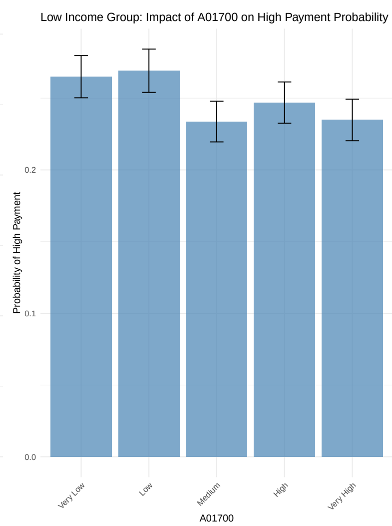

Low Income Group: Impact of Tot Benes on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Benes

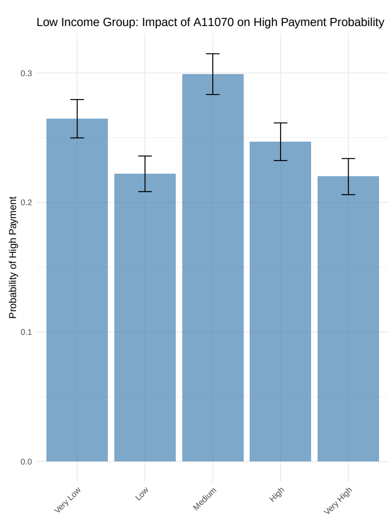

Low Income Group: Impact of State on High Payment Probability
Probability of High Payment
0.00
0.25
0.50
0.75
1.00
State

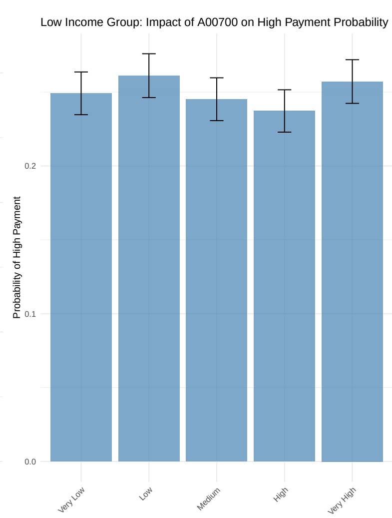

Low Income Group: Impact of Tot Srvcs on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Srvcs

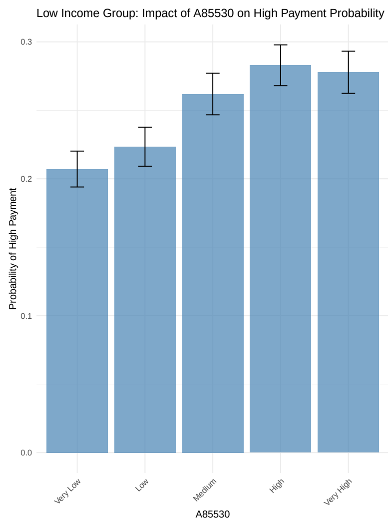

Low Income Group: Impact of A00300 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A00300

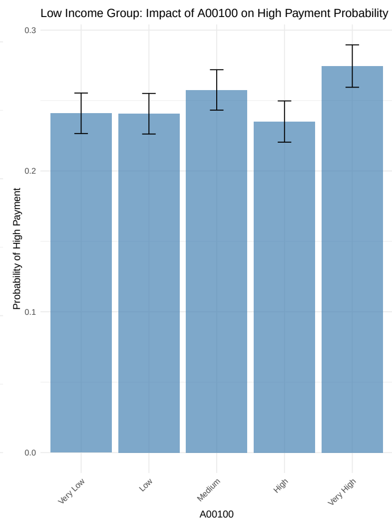

Low Income Group: Impact of A19700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A19700

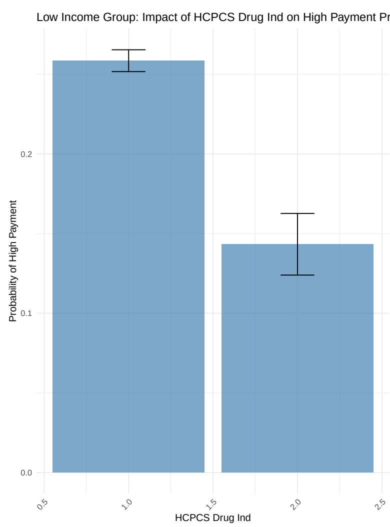

Low Income Group: Impact of A01700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A01700

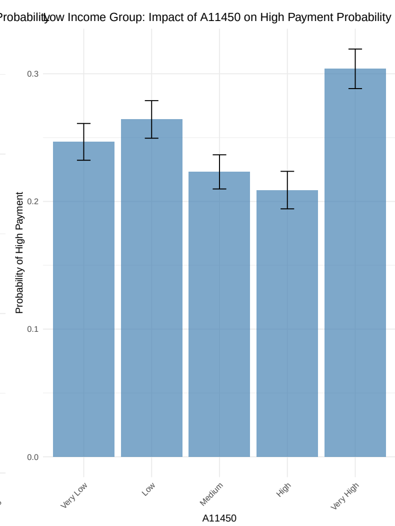

Low Income Group: Impact of A11070 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A11070

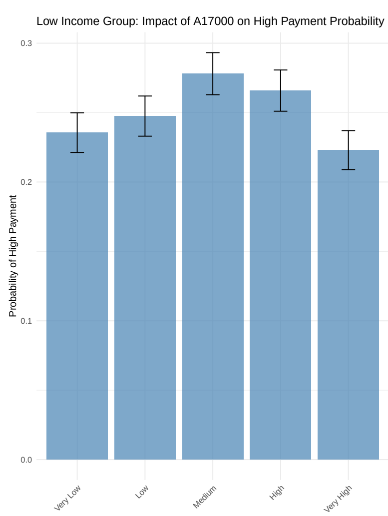

Low Income Group: Impact of A00700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A00700

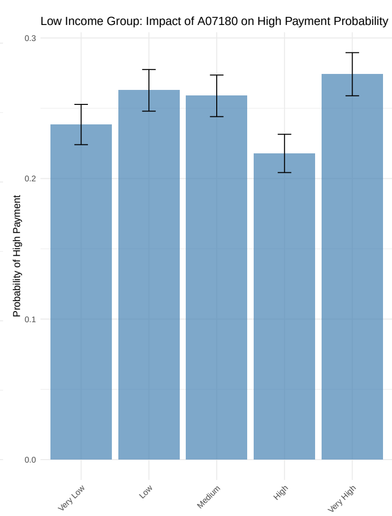

Low Income Group: Impact of A85530 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A85530

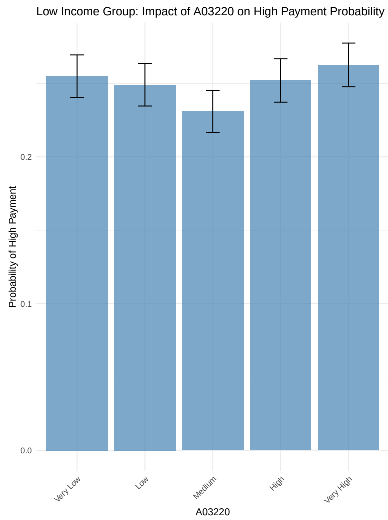

Low Income Group: Impact of A00100 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A00100

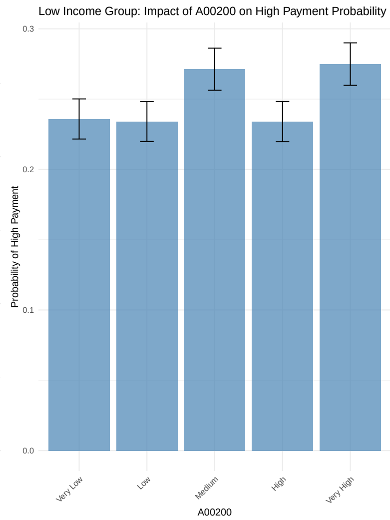

Low Income Group: Impact of HCPCS Drug Ind on High Payment Probability
Probability of High Payment
0.2
0.1
0.0
0.5
1.0
1.5
2.0
2.5
HCPCS Drug Ind

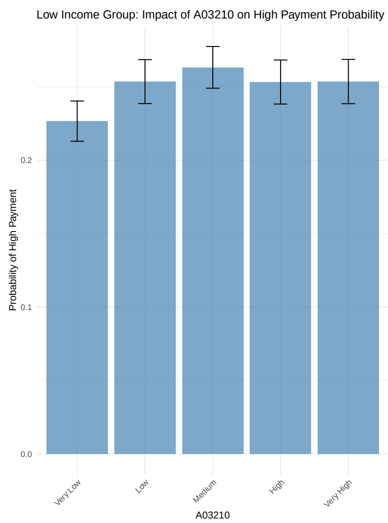

Low Income Group: Impact of A11450 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A11450

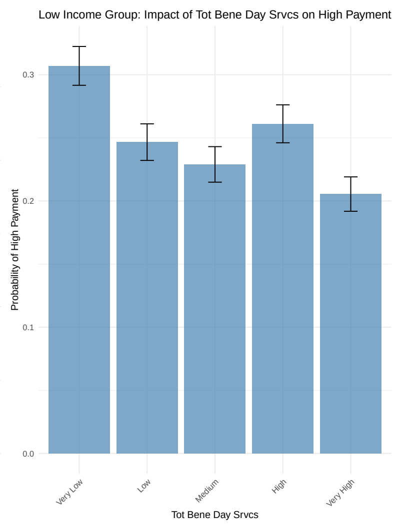

Low Income Group: Impact of A17000 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A17000

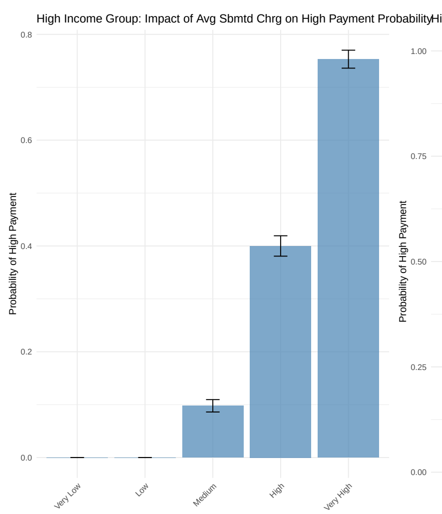

Low Income Group: Impact of A07180 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A07180

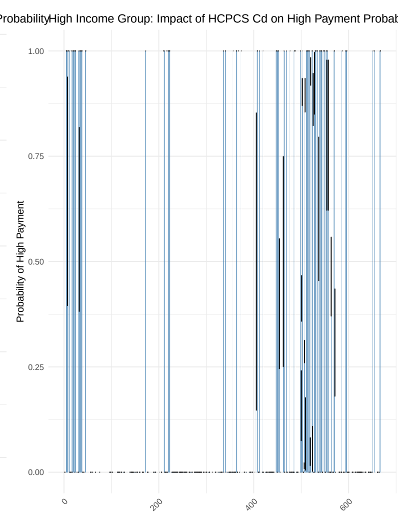

Low Income Group: Impact of A03220 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A03220

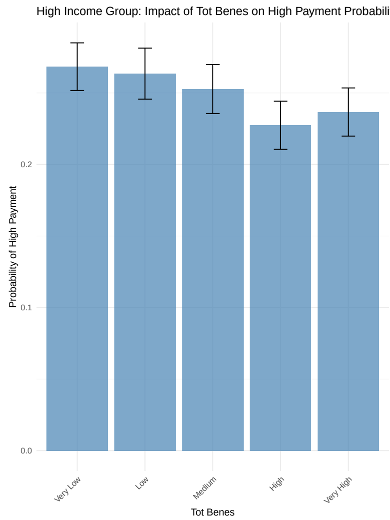

Low Income Group: Impact of A00200 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A00200

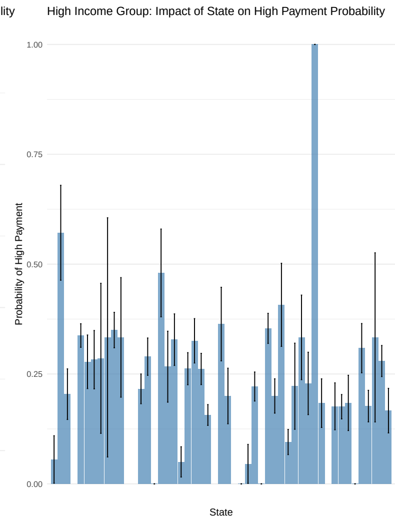

Low Income Group: Impact of A03210 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A03210

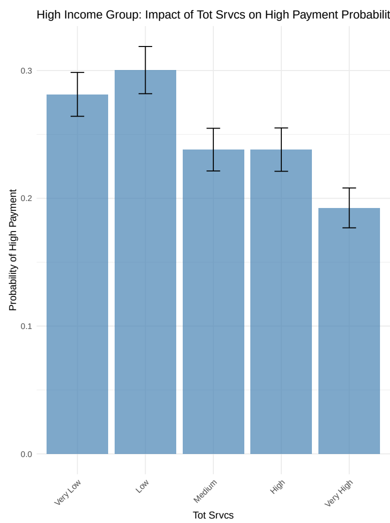

Low Income Group: Impact of Tot Bene Day Srvcs on High Payment I
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Bene Day Srvcs

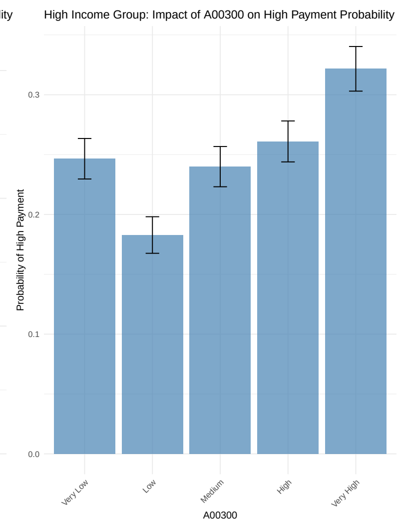

High Income Group: Impact of Avg Sbmtd Chrg on High Payment Probability
Probability of High Payment
0.0
0.2
0.4
0.6
0.8
Very Low
Low
Medium
High
Very High
Avg Sbmtd Chrg

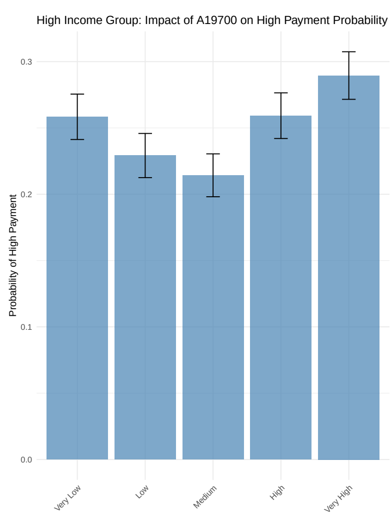

High Income Group: Impact of HCPCS Cd on High Payment Probab
Probability of High Payment
0.00
0.25
0.50
0.75
1.00
0
200
400
600
HCPCS Cd

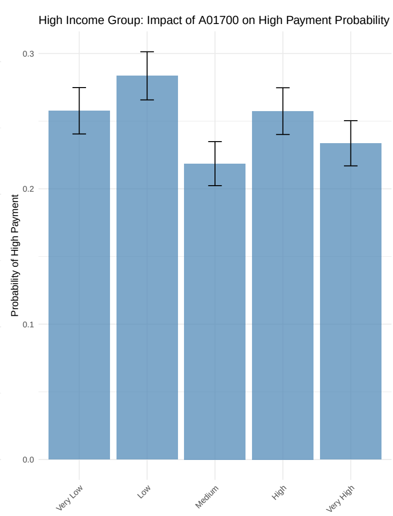

High Income Group: Impact of Tot Benes on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
Tot Benes

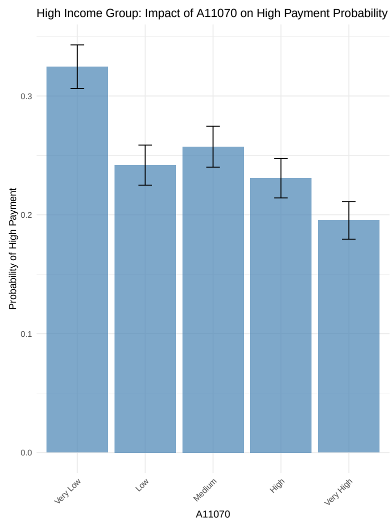

High Income Group: Impact of State on High Payment Probability
Probability of High Payment
0.00
0.25
0.50
0.75
1.00
State

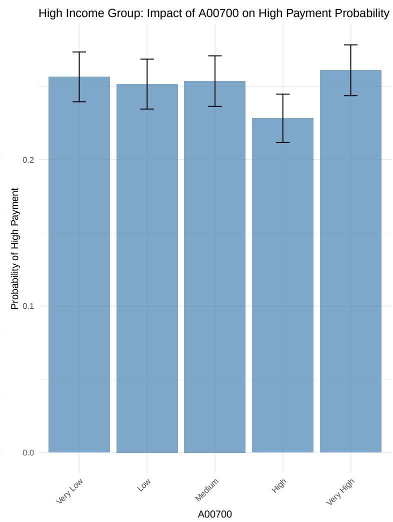

High Income Group: Impact of Tot Srvcs on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
Tot Srvcs

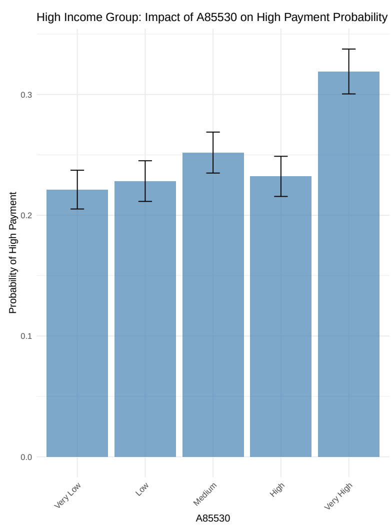

High Income Group: Impact of A00300 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A00300

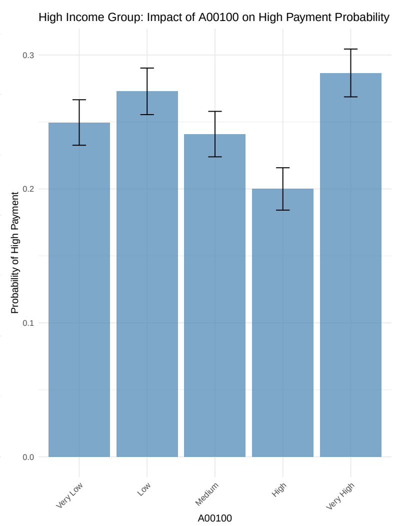

High Income Group: Impact of A19700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A19700

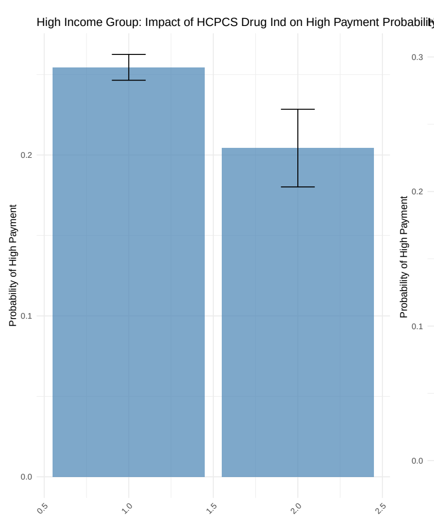

High Income Group: Impact of A01700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A01700

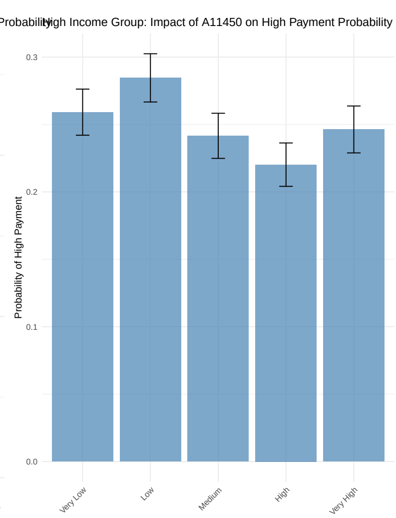

High Income Group: Impact of A11070 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A11070

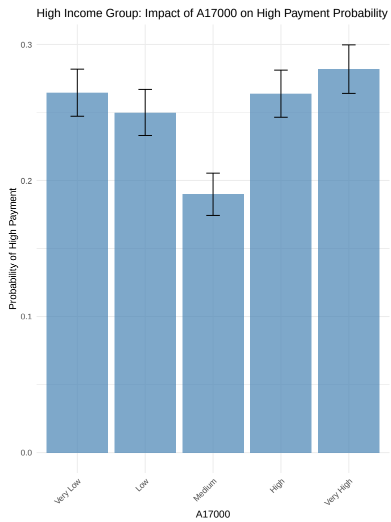

High Income Group: Impact of A00700 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
Very Low
Low
Medium
High
Very High
A00700

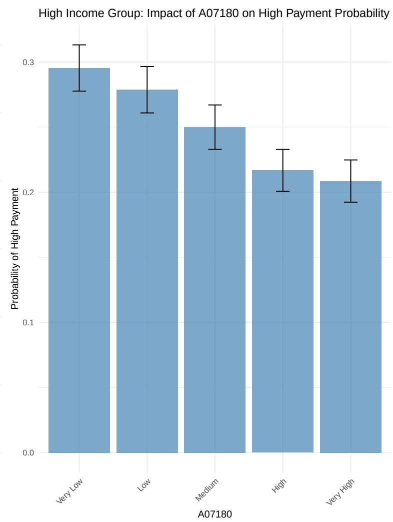

High Income Group: Impact of A85530 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A85530

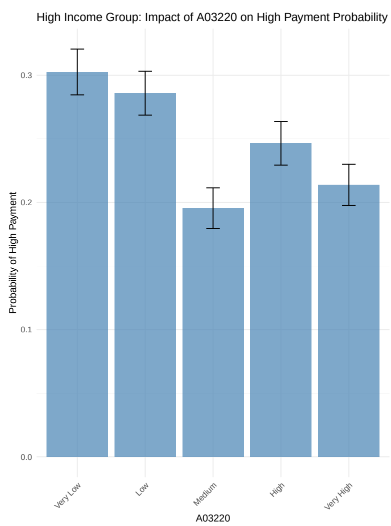

High Income Group: Impact of A00100 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A00100

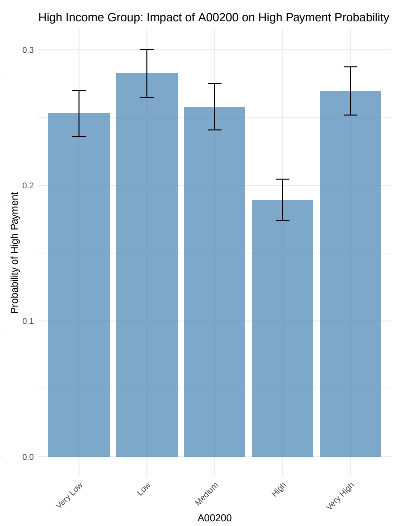

High Income Group: Impact of HCPCS Drug Ind on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.5
1.0
1.5
2.0
2.5
HCPCS Drug Ind

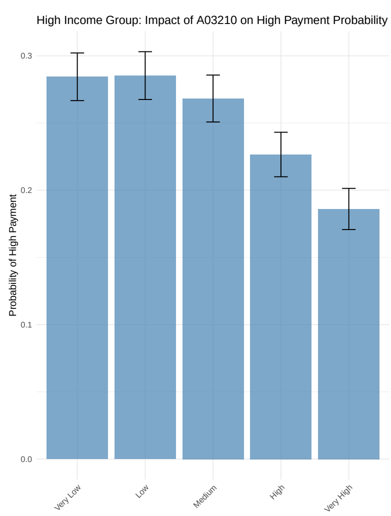

High Income Group: Impact of A11450 on High Payment Probability
Probability of High Payment
0.0
0.1
0.2
0.3
Very Low
Low
Medium
High
Very High
A11450

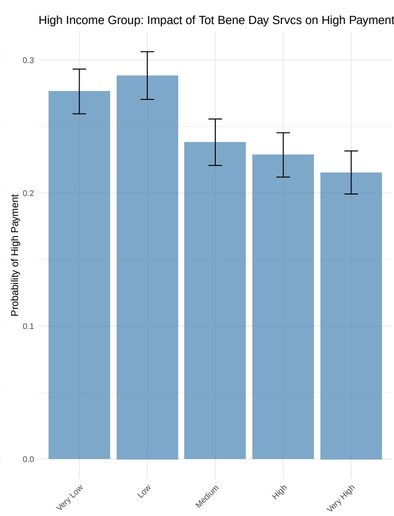

High Income Group: Impact of A17000 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A17000


High Income Group: Impact of A07180 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A07180


High Income Group: Impact of A03220 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A03220


High Income Group: Impact of A00200 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A00200


High Income Group: Impact of A03210 on High Payment Probability
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
A03210


High Income Group: Impact of Tot Bene Day Srvcs on High Payment
Probability of High Payment
0.3
0.2
0.1
0.0
Very Low
Low
Medium
High
Very High
Tot Bene Day Srvcs