Regression Models for Predicting Medicaid Drug Reimbursement

DSCI 799: Graduate Capstone

Chelsea Heimiller

Prescription drugs are a major expense for many Americans and rising prescription drug prices are a concern for both the state and federal government as well as consumers. For the Medicaid program, net spending on prescription drugs increased 47% from 2017 to 2022 from $29.8 billion to $43.8 billion (Smith, 2023). This is potentially due to a variety of factors including but not limited to number of Medicaid beneficiaries, number of prescriptions, drug class, and economic factors. The goal of this project was to determine what impacts Medicaid drug price based on various factors obtained from multiple governmental sources. Understanding what drives Medicaid drug prices will allow policies and interventions to be implemented to reduce drug costs. It can also allow states to plan their Medicaid budgets based on how much they will be reimbursed by the federal government for drugs and the federal government can plan their budgets for Medicaid drug spending.

Background

Medicaid is a federally funded program that is managed at the state level. While eligibility is generally based on income and assets, states are responsible for determining eligibility and benefits. Therefore, what is covered by Medicaid may vary from state to state. Drug costs are an increasing part of the Medicaid budget with the amount of Medicaid dollars spent on prescription drugs increasing from 6% to 9% of total Medicaid spending between 2010-2015 (Dranove et al., 2021). The federal government is now able to negotiate Medicaid drug prices due to the Inflation Reduction Act which passed in 2022 which is expected to contribute to a decline in the amount a drug costs, however it is unclear if this will lead to lower overall spending as the number of Medicaid prescriptions increase.

The federal government reimburses the states a certain amount of money for the prescription drugs that are covered as a part of the Medicaid program. There are several factors that may contribute to the amount of dollars reimbursed by the federal government. These factors include number of prescriptions filled, units reimbursed by the federal government, number of Medicaid beneficiaries, and the drug prescribed among others. This project used data from a variety of sources to look at the trends in the Medicaid drug space to get an understanding of the factors surrounding federal Medicaid drug reimbursement. A regression model was built to identify what factors drive Medicaid drug reimbursement. Identifying what drives total Medicaid drug reimbursement will allow planning by both the states and federal government as well as can help to identify opportunities for reducing costs.

Data Sources

Data for this project comes from a variety of sources. The primary source is the Medicaid State Drug Utilization Data (SDUD) from 2018-2022 (Center for Medicaid and Medicare Services, 2023). The state drug utilization data (SDUD) files for 2018-2022 contain data on Medicaid drugs at the state level including units reimbursed, number of prescriptions filled, drug name, and dollars reimbursed (CMS.gov, 2023). The SDUD dataset contains the target variable which is total amount reimbursed for the drug.

The National Drug Code (NDC) directory data was obtained from the Federal Drug Administration (FDA) website and provides the drug’s proprietary name, classification, how the drug is taken, whether it is a brand or generic drug, and who makes the drug among many different variables as well as the 10-digit identification code for the specific drug (U.S. Food and

Drug Administration, 2023). The National Average Drug Acquisition Cost (NADAC) datasets are made available by CMS and show the average retail price of drugs based on what chain and independent retail pharmacies are charging for drugs (Center for Medicare and Medicaid Services, 2024). This data is based on surveys and is the basis for how Medicaid determines drug reimbursement rates. Drug price was in a weekly format and the weeks were averaged together to get the quarterly average cost of a drug per unit. The count of Medicaid beneficiaries by month was obtained from CMS and the records deemed unusable by CMS were dropped. The rest were averaged together by quarter to get the average number of Medicaid beneficiaries by state for a month in the quarter. This value was then multiplied by three to get the average member count per quarter for the state specified on the record.

Economic data was pulled from a few sources. Consumer price index (CPI) and state unemployment rate was obtained from the Bureau of Labor Statistics at the monthly level (U.S. Bureau of Labor Statistics, 2024). The monthly data was averaged together to the quarterly level to allow for a picture of the variable throughout the quarter. Gross domestic product (GDP) in billions of dollars was obtained from table T10105-Q on the Bureau of Economic Analysis' website at the quarterly level and represents the total dollars in goods and services produced by the United States (Bureau of Economic Analysis, 2024). Mortgage rates as percentages were available on the Federal Reserve Bank of St. Louis's website from Freddie Mac at the quarterly level (Freddie Mac, 2024). Finally, median household income as well as poverty data was obtained from the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program which provides these details at the state level for the whole year (United States Census Bureau, 2021).

Methodology

The dataset was first assembled using the SDUD as a base. All other datasets were appended to the SDUD based on defining characteristics of the record such as NDC package code, state, year, etc (Appendix A). Any records in the SDUD dataset that did not have a NDC package code associated with them in the NDC dataset were dropped as these products are not considered drug products by the FDA and are therefore not in the scope of this project. Monthly economic data was averaged to get a value for the entire quarter so that it matched with the quarterly format of the SDUD file and was a representative value of what happened during the specified time period.

Since data first needs to be reported to the state and then to the Center for Medicaid and Medicare Services (CMS), there is plenty of room for data to have been misreported or not reported at all. Missing data has an impact on the model building process and the quality of the resulting models. However, imputations run some risks. Namely they potentially introduce bias and may assign artificial importance to the imputed variable masking other relationships (Shadbahr et al., 2023). There is not one imputation method that is proven to be better than others, but leaving null values in the data will impact a model's performance and poorly imputed data may impact a model's generalizability. To prevent the missing values and potentially poor imputations from impacting the model building process and the assessment of what drives Medicaid drug dollars, any record with a missing value will be dropped resulting in a dataset that consists only of values that were already present in the dataset (Appendix B). The exception was DEA schedule since a null value in DEA schedule indicates that the drug is not a controlled substance. A total 1,167,935 of the records contained a null value and were dropped. This is approximately 23% of the records. The concern with creating imputations with only about 76%

of the data having complete records is that since only four years of data is included accurate imputations are not guaranteed to be feasible. The concern with dropping all records that have null values is that the loss of information from the dropped records may have an impact on the model's ability to generalize on a dataset it has not seen before. However, the impact of null values on the model is a bigger risk and therefore the tradeoff was accepted.

Data exploration indicated that the dataset had several outliers for multiple features. Outliers were identified with an outlier factor of 1.5 and extreme outliers were identified with an outlier factor of 3. The concern with outliers is their ability to influence a regression model, specifically linear regression models. Outliers may also offer valuable information to the model building process. Extreme outliers were dropped due to majority of the extreme outliers belonging to the same drug and the others appearing to be outliers when compared to the other records with the same drugs name (Appendix C). Outliers based on the outlier factor of 1.5 were kept due to their ability to capture important patterns on high-use/high-cost drugs and expected patterns with high population states and states with high poverty.

Variables that were lists with multiple items were split by tokenizing and using Multi-label Bianizer to encode the data so that each element in the list had its own flag column. This only impacted the pharmacological class column and dosage form columns. To reduce dimensionality of the data only the first five elements of the pharmacological class column were considered when splitting and data elements were cleaned to remove extraneous characters that would create duplicate columns for the same element. To reduce dimensionality further only classes that were present more than 70,000 times were kept in an effort to capture the most meaningful classes to the total amount reimbursed and prevent overfitting. Variables that are ordinal were encoded using label encoding with scikit learn. They were encoded in a logical order based on what they are portraying. For example: DEA Schedule can range anywhere from “Not Controlled” to “Class V” which is a controlled substance with no approved medical use that is highly addictive. The higher the schedule the more addictive the drug is considered making this an ordinal variable that label encoding is appropriate for. Variables that were not ordinal were for the most part encoded with one hot encoding.

The four features that had high cardinality were encoded using target encoding. Target encoding encodes features based on the mean of the target variable for each category being encoded to get a value directly related to the relationship of the category with the target variable. The category is replaced with the calculation instead of creating new features for each category resulting in one feature as opposed to multiple features. This reduces dimensionality which can prevent overfitting and allows a generalizable model. The downside is that if a category has a small sample size the encoding would be very similar to the value that is trying to be predicted which can lead to overfitting and a model that is not generalizable. To overcome this the data will be smoothed (Neural Ninja, 2023). The target encoder from category encoders will be used which defaults to leave-one-out encoding. Another concern is the inability of target encoding to handle a category that it has never seen before in data that is introduced outside of the training set. However, since target encoding provides context around the data in its relationship to the target without increasing dimensionality, target encoding will be used for the four features with high cardinality.

Target encoding is prone to data leakage if the data is not handled appropriately prior to encoding (Neural Ninja, 2023). Data was split into train and test sets and the encoding was fit to the training set and then the train and test set will be transformed so that the features are encoded. This prevents data leakage by ensuring that the training data is not exposed to data in the test set prematurely. If the data was not split the data in the training set would be encoded with an average that considers data in the test set which is data leakage. Data was split 80/20 with 80 percent of the data going to training and 20 percent of the data going to testing and k-fold cross-validation will be used for the validation set. Data was scaled using robust scaling to allow the most accurate models to be built by avoiding larger features overshadowing smaller ones.

Data was explored using a combination of the pandas library, matplotlib library, and tableau. Variables were investigated to identify if there were any trends among drug cost, number of prescriptions, or drug type among other variables before building a regression model to determine what drives Medicaid drug reimbursement costs.

Linear models and nonlinear models were explored due to the uncertainty about patterns in the data. Some variables appear to be linear whereas some appear to have no relationship. Linear models that were considered include linear regression (least squares) which is used as a baseline to compare other models to and ridge regression. While not robust to outliers, ridge regression is good for reducing dimensionality and handling multicollinearity. Both of these linear regression algorithms are not robust to outliers and the outliers in the dataset may be indicative of important patterns that influence the model in a way that negatively impacts its performance. Due to the influence that outliers can have on these algorithms two robust linear algorithms were explored: Huber regression and RANSAC regression. The Huber regression model was used because it is resistant to the influence of outliers while still taking their presence into account (Lewinson, 2023). Huber regression uses squared loss for data point close to the target and absolute loss for points that are considered outliers. RANSAC regression was also tested as it splits the data into two chunks, data points that are outliers and those that are inliers and then builds the model based on the inliers (Lewinson, 2023). The concern with RANSAC regression is that it may not lead to a generalizable model and fail to capture important patterns among high-use/high-cost drugs as well as states with large populations and high utilization due to only building the model on the inliers. Regardless RANSAC regression models were evaluated to determine if it was the best fit for this problem.

Since some of the variables are suspected to be nonlinear due to the lack of high correlation between the features and the target, nonlinear methods were also explored (Appendix D). Decision tree regression was used to build models as decision trees can capture nonlinear relationships and partitions features into regions based on values, so they are not as influenced by outliers as linear regression (AIML.com, 2023). The disadvantage is that decision trees are prone to overfitting. Ensemble methods were considered due to their ability to capture nonlinear relationships and ability to build models that are not prone to overfitting. Random forest regression was considered due to its resistance to overfitting through randomness introduced during training and its ability to lend itself to a generalizable model due to the averaging of predictions of all the trees created during the model building process (AIML.com, 2023). Random forest regression is also able to handle outliers without being influenced by them due to the partitioning of data that happens in the model building process. The concern with random forest regression is that it is not as easy to interpret as linear models or decision tree regression, but feature importance is still able to be assessed.

The benefit of using linear models is that they are easy to interpret and see how the features are interacting with the target which will assist in determining which factors are driving the cost of Medicaid drug prices (Satyavishnumolakala, 2020). Linear models are also computationally efficient which is an advantage due to the size of the dataset. The issue with the linear model is the initial evaluation of the linear models show that a few key features are greatly influencing the model and the remainder of the features have non-linear relationships with the target. Additionally, the two nonlinear regression methods are both less likely to be influenced by outliers due to the partitioning of data.

The models were assessed in a few different ways. All models were assessed for mean squared error and R-squared. Linear models had their coefficients assessed and nonlinear models were assessed for feature importance. Feature selection was done by first using recursive feature elimination to get an idea of which features were influencing the models. From there domain knowledge was used to refine features and select the features that were most relevant to the task and would allow the model to be explained.

Models were built using k-fold cross-validation with a factor of five first with default parameters to narrow down models and features. After features and models were selected grid search was used to tune hyperparameters to find the best fit for the problem at hand.

Features

Feature selection was done by first using recursive feature elimination to reduce the number of features and then modified using domain knowledge. A few different numbers were tried for how many variables should be selected using recursive feature elimination (RFE) but ultimately the best number of variables when combined with model building was about twenty (Figure 1). Recursive feature elimination was performed using linear regression as the estimator for computational efficiency. The drawback to linear regression is that it is possible that a variable that does not have a linear relationship is important to the determination of the target variable and may be left out. This is why domain knowledge will be used to add features that are relevant to the task at hand.

Units reimbursed, number of prescriptions, state name, quarter average NADAC per unit, NDC package code, proprietary name, pharmacological class allergens, pharmacological class anti-inflammatory agents, pharmacological class cell-mediated immunity, pharmacological class cyclooxygenase inhibitors, pharmacological class histamine H1 receptor agonists, pharmacological class non-steroidal, pharmacological class osmotic activity, pharmacological class osmotic laxative, dosage form injection, dosage form liquid, dosage form lozenge, dosage form patch, dosage form solution, dosage form spray.

Figure 1: Features Selected Using Recursive Feature Elimination for Twenty Variables

Initial investigation of the features selected using RFE indicated that the pharmacological class does not have a large impact on the total amount reimbursed and that majority of the dosage forms do not either. Economic factors were not selected using RFE likely because they are time based and four years' worth of data may not be enough time to see their true impact on the total amount reimbursed. However, it is expected that as the economy changes and legislation changes the economic variables will have an impact on the total amount reimbursed for Medicaid drugs.

The selected features were the same for all models (Figure 2). The state name is important because the states have control over certain aspects of the Medicaid program which can impact the drug spending and reimbursement in the state. The national average drug acquisition cost (NADAC) is the average price per unit of a drug based on retail pharmacy rates which are reported to CMS. This information is important because NADAC rates are the basis for Medicaid reimbursement rates making this feature relevant to the project. The NDC package code and proprietary name both lend identifying characteristics regarding the drug such as name, package size, and manufacturer. The quarter provides the time frame the record is from. The four dosage forms used in the model are representative of the form the drug is in. Any record that does not have one of these dosage forms as a flag is considered an 'other' dosage form. These four dosage forms make sense because tablet, powder, and capsule are likely all taken orally meaning they would be common drugs to be taken by the general population. The injection can capture drugs for diabetes, hormonal treatment, and blood thinners which may be used to treat chronic disease which may be prevalent in the Medicaid population.

Feature Name	Definition
Units Reimbursed	Units reimbursed for the drug by the federal government.
Number of Prescriptions	Number of prescriptions filled for the drug.
State Name	The name of the state the record belongs to.
Average NADAC Price Per Unit	National average price per unit of the drug from survey of retail pharmacies around the nation at the quarter level.
NDC Code	The unique identifier for the specific chemical, manufacture, and package size of the drug product. A single proprietary drug could have multiple NDC codes if multiple manufacturers produce it or if the drug has multiple package sizes.
Proprietary Drug Name	The brand name of the drug.
Quarter	The quarter the record is for in 'yyyyqq' format.
Dosage Form Tablet	Flagged if the drug is a tablet.
Dosage Form Powder	Flagged if the drug is a powder.
Dosage Form Injection	Flagged if the drug is given by injection.
Dosage Form Capsule	Flagged if the drug is a capsule.
GDP in Billions	Gross domestic product for the nation for the quarter.
Unemployment Rate	Unemployment percentage for the state the record belongs to for the quarter. This is the number of people not working over the number of people eligible in the state to work.
Poverty Percent, All Ages	Percentage of the state's population living below the poverty line for the quarter.
Median Household Income	The median income for the state by household.
CPI	Consumer price index for the nation for the quarter specified by the record.
Count Enrolled	The average number of Medicaid beneficiaries per month for the specified state and quarter multiplied by three to get average Medicaid beneficiaries by quarter.

Figure 2: Features Selected for Model Building

GDP in billions captures the gross domestic product of the country reflecting the economic health of the economy while consumer price index (CPI) captures the consumer confidence in the economy. The unemployment rate, poverty percent, and median household income are all state level variables that provide insight into the state's individual economy. The unemployment rate indicates how many people who are eligible to work are not working while poverty percent provides information on the percentage of the population of the state lives below the poverty line. Median household income provides information on the income level in the state. These economic factors are important because Medicaid eligibility is assessed based on income and assets so higher unemployment rates and poverty percentages will impact the number of Medicaid beneficiaries in the state which will contribute prescription drug use. The count of Medicaid beneficiary months is the sum of Medicaid members per month for the quarter. This is important since it provides information around trends in the Medicaid space. More Medicaid members likely indicates more prescriptions and higher reimbursement costs so therefore number of Medicaid members is an important feature for this project.

Of the features selected not all of them have a high correlation with the target variable, total amount reimbursed (Appendix D). The ones who do not have a strong correlation may still have a relationship with the variable it just may be nonlinear. Units reimbursed and number of prescriptions have the highest correlation followed by NDC package code, and state name.

Results

Over the course of 2018-2022 there was a general increase in total amount of dollars the federal government reimbursed the states for regarding prescription drugs (Figure 3). There is a larger increase between the second quarter of 2020 and the third quarter of 2020 which coincides with the COVID 19 pandemic. While there is a slight decrease in the fourth quarter of 2022, the time series indicates that the federal government has been reimbursing more money as the years go on for Medicaid drug products. Over time the number of Medicaid beneficiaries has been increasing, this is in line with the rise in all the previously explored time variables and indicates that the number of Medicaid beneficiaries may have an impact on total amount reimbursed (Figure 4).

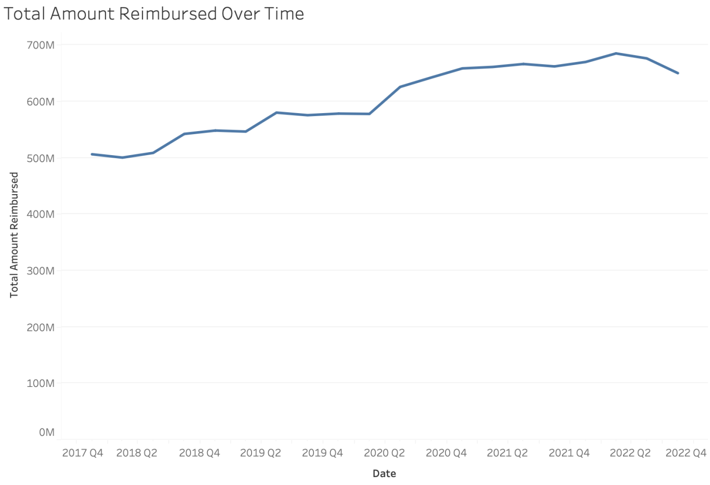


Figure 3: Total Amount of Medicaid Dollars Reimbursed to the States Over Time

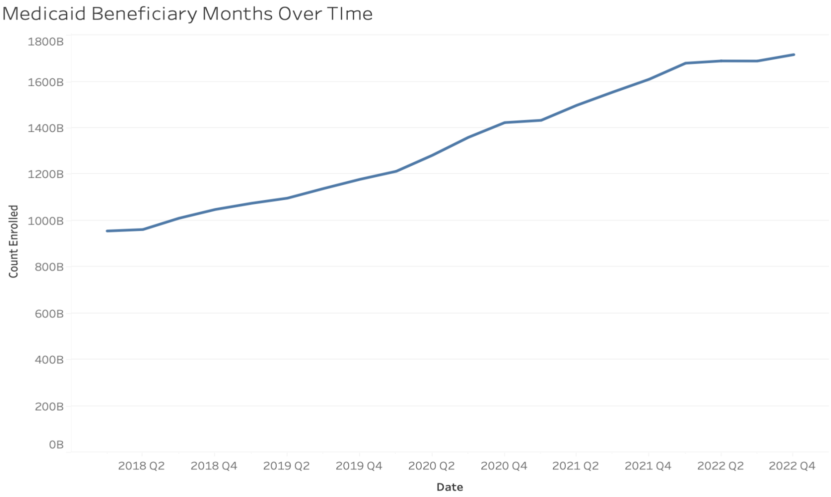


Figure 4: Total Count of Medicaid Beneficiaries Enrolled Over Time

Over time the total number of prescriptions filled, and the total units reimbursed by the federal government both increase at a steady rate (Figures 5 and 6). The upward trends with slight dips between Q2 2022 and Q3 2022 are similar to the total amount reimbursed over time trendline. This makes sense with total amount reimbursed because it would be logical for a higher number of prescriptions filled to lead to higher reimbursement dollars, and a higher number of units reimbursed to also lead to higher total dollars reimbursed. The catch regarding number of prescriptions filled is that just because a drug is filled does not necessarily mean that it will be reimbursed by the federal government. Additionally, units reimbursed for high-use or high-cost drugs may have an impact on the total amount reimbursed skewing the relationship. Due to this regression models were built to attempt to capture the relationships between other variables with units reimbursed and number of prescriptions filled to get the full picture of what is contributing to Medicaid drug prices.

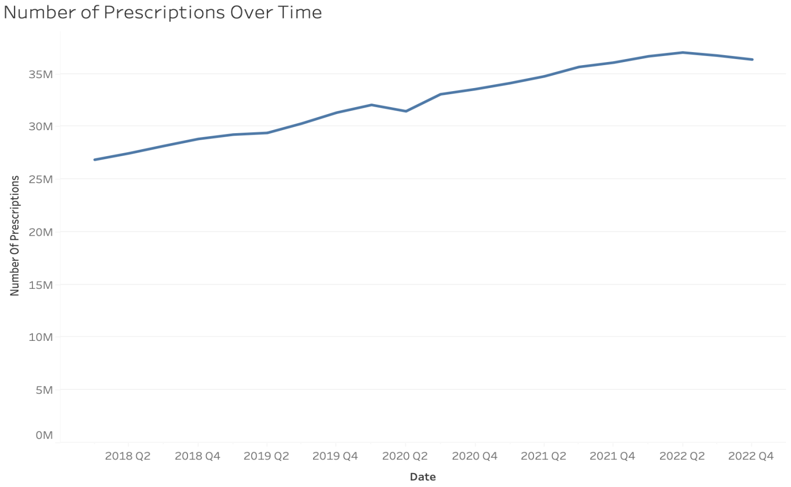


Figure 5: Number of Medicaid Prescriptions Filled Over Time

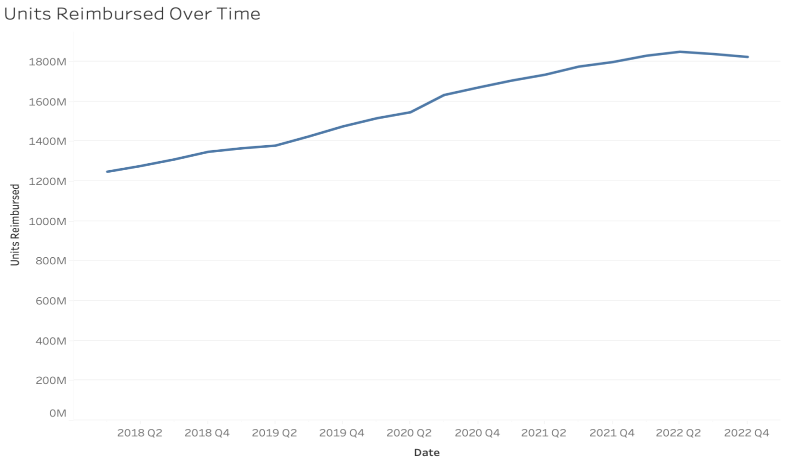


Figure 6: Number of Units Reimbursed for Medicaid Drugs Over Time

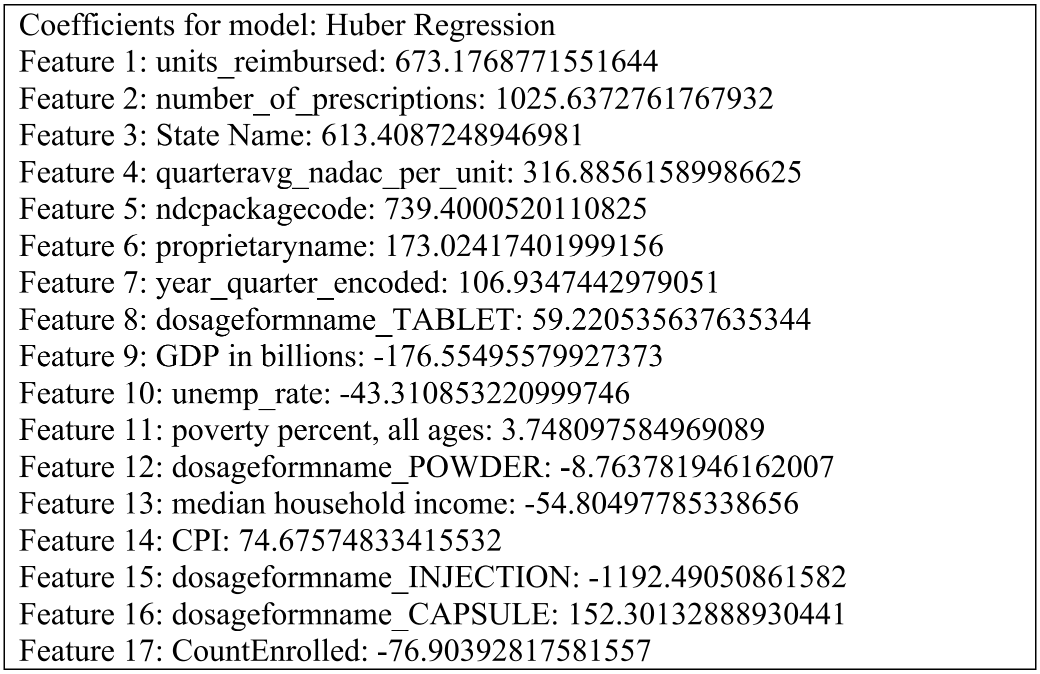
Results of Regression Models

Overall, the linear models performed poorly. None of the models exceeded an R-squared value of greater that 63.6% (Figure 7). The driving features of all linear models were primarily units reimbursed, NDC package code, and number of prescriptions (Appendices E through H). This is logical because as number of prescriptions filled for a drug goes up it is expected that the amount reimbursed to the state by the federal government would also increase. This would capture the relationship between drug use and drug cost meaning high-utilization drugs will have a higher amount reimbursed when compared to a low-use drug. NDC package code indicates which drug is being used and the size of the package the drug is in which also makes it a logical indicator for a linear relationship since a larger package size would be expected to have a larger reimbursement.

Model	Mean Squared Error	R-Squared
Linear Regression	3931403.97	63.8%
Ridge Regression	3931403.93	63.8%
RANSAC Regression	4176611.44	61.5%
Huber Regression	422677.11	61.0%
Decision Tree Regression	2490110.43	77.0%
Random Forest Regression	1272631.63	89.1%

Figure 7: Results of All Regression Models

The linear models that perform the best are the general linear regression and ridge regression. Huber regression and RANSAC regression both have lower R-squared values. However, these models are likely more indicative of what is happening in the dataset and are probably more generalizable due to their resistance to outliers. Of the linear models the Huber regression model is the preferred model due to its ability to capture the influence of outliers. The outliers in the Medicaid drug dataset are important because they capture the information surrounding high-utilization drugs and/or high-cost drugs that may be considered outliers in the dataset. The Huber model has an R-squared value of 61.0% meaning that the features explain 61.0% of the variability in the data and a mean-squared error of 422,677.11. Units reimbursed and number of prescriptions are expectedly the main drivers of this model due to their correlation with total amount reimbursed (Figures 8 and 9). State name also had a substantial positive impact likely due to the variations in population size and what Medicaid covers in each state. Surprisingly dosage form of injection has a negative impact indicating that records that are for drugs that are administered via injection have fewer dollars reimbursed when the linear model is referenced. Economic factors have some impact and are included to capture any variations. Due to the dataset being limited to four years of data there could very well be an important relationship and more years need to be included to determine the full impact of economic factors.



```
Coefficients for model: Huber Regression
Feature 1: units_reimbursed: 673.1768771551644
Feature 2: number_of_prescriptions: 1025.6372761767932
Feature 3: State Name: 613.4087248946981
Feature 4: quarteravg_nadac_per_unit: 316.88561589986625
Feature 5: ndcpackagecode: 739.4000520110825
Feature 6: proprietaryname: 173.02417401999156
Feature 7: year_quarter_encoded: 106.9347442979051
Feature 8: dosageformname_TABLET: 59.220535637635344
Feature 9: GDP in billions: -176.55495579927373
Feature 10: unemp_rate: -43.310853220999746
Feature 11: poverty percent, all ages: 3.748097584969089
Feature 12: dosageformname_POWDER: -8.763781946162007
Feature 13: median household income: -54.80497785338656
Feature 14: CPI: 74.67574833415532
Feature 15: dosageformname_INJECTION: -1192.49050861582
Feature 16: dosageformname_CAPSULE: 152.30132888930441
Feature 17: CountEnrolled: -76.90392817581557
```

Figure 8: Coefficients of the Variables for the Huber Regression Model

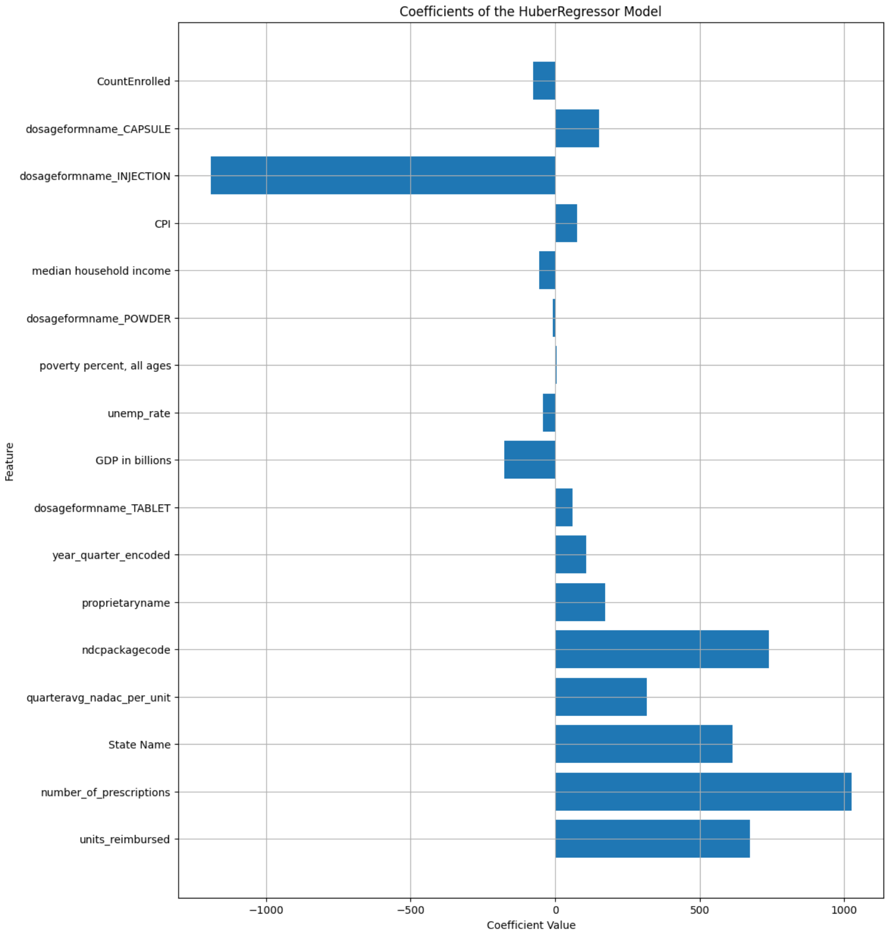


Figure 9: Visualization of Coefficients for the Huber Regression Model

The nonlinear models perform much better than the linear models when comparing mean squared error and R-squared on all feature subsets that were used (Figure 7). Random forest regression performed better than decision tree likely due to being an ensemble method and taking the best of the best when building multiple trees. The feature importance for the decision trees is nearly identical to the random forest regression models (Appendix I and Figure 10). This is likely because random forest is essentially a bunch of decision trees. The random forest regression model had an R-squared value of 89.1% meaning the model explains 89.1% of the variability in the total amount reimbursed and the mean squared error was 1272631.63 (Figure 7).

```
Feature Importance: Random Forest Regression
Feature 1: units_reimbursed: 0.1906
Feature 2: number_of_prescriptions: 0.3917
Feature 4: State Name: 0.0366
Feature 10: quarteravg_nadac_per_unit: 0.2387
Feature 11: ndcpackagecode: 0.0391
Feature 12: proprietaryname: 0.0216
Feature 75: year_quarter_encoded: 0.0048
Feature 71: dosageformname_TABLET: 0.0024
Feature 16: GDP in billions: 0.0044
Feature 5: unemp_rate: 0.0112
Feature 8: poverty percent, all ages: 0.0117
Feature 67: dosageformname_POWDER: 0.0017
Feature 9: median household income: 0.0108
Feature 15: CPI: 0.0046
Feature 60: dosageformname_INJECTION: 0.0002
Feature 55: dosageformname_CAPSULE: 0.0021
Feature 6: CountEnrolled: 0.0281
```

Figure 10: Feature Importance Random Forest Regression

The number of prescriptions is the most important feature for both models followed by quarter average NADAC per unit (Figure 11). The quarter average NADAC per unit is much more influential on the nonlinear models than in the linear models. This indicates that there is a nonlinear relationship between NADAC per unit that is important. This is because NADAC per unit has much smaller values that represent a single unit and therefore will not go up linearly when the total amount reimbursed goes up but will have an impact on total amount reimbursed when considered in conjunction with units reimbursed. The NADAC per unit generally is used as a baseline for Medicaid drug reimbursement as most states take NADAC into account for ingredient cost. The dosage forms are less important in the model however when they were removed there was a drop in accuracy. While increased accuracy does not necessarily mean the model is better the dosage forms were elected to be left in because they provide information about how common a drug may be. The oral forms such as tablet, capsule, and powder are easy to take and therefore highly prescribed and injections are used for common chronic diseases like diabetes. The economic factors are not as important as originally suspected but this could be due to only four years' worth of data being included in the model building process. It is possible as time goes on and more data is available that the economic factors and the year/quarter combination will show increased importance.

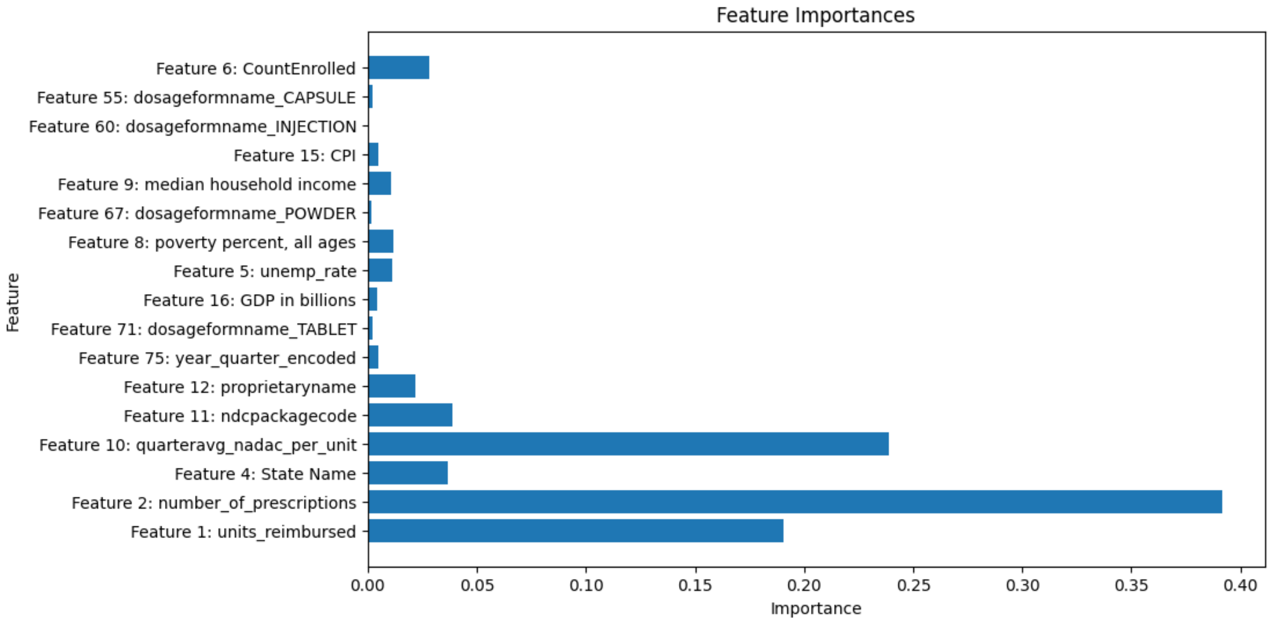


Figure 11: Random Forest Regression Feature Importance Visualization

Discussion

The model of choice for Medicaid drug price spending is the random forest regression model and the results of its feature importance can be used to determine the driving factors of Medicaid drug prices. This model is preferred for a couple of reasons. It has the best mean squared error and R-squared scoring. However, this alone is not a reason to pick this model. The random forest regression model has been chosen over decision tree model because the random forest regression is made up of multiple decision trees making random forest regression more resilient to overfitting. This is advantageous because the model is likely more generalizable to data it has not seen before. Random forest models also are not influenced by outliers due to the partitioning of the data so the important relationships for high-use and high-cost drugs can be captured without skewing the whole model as a linear model would. The random forest model is also able to capture relationships between variables that the linear models cannot. For example, random forest regression identified the relationship between NADAC per unit and total amount reimbursed which was not captured in any of the linear models. Due to this, the random forest regression model should be used to determine what variables are most important in driving Medicaid drug reimbursement.

The random forest regression model highlights a few variables that are the most important in driving the cost of Medicaid drug reimbursements. The most important factors are units reimbursed and number of prescriptions filled. While this may seem obvious it is important to note that just because a prescription is filled does not necessarily mean that it will be reimbursed, and the Medicaid beneficiary may have to pay out of pocket for some drugs. NADAC price per unit is also a driving factor and is an option for intervention. This was not a relationship that was captured in the linear models however it provides an important piece of information about what drives drug reimbursement dollars. Now that the federal government can negotiate drug costs, it is possible that price per unit will decline, and Medicaid drug reimbursement dollars will decrease. Negotiating the prices of high use and high-cost drugs is an area for savings that should be considered. The driving factors should be re-evaluated in a few years to determine if the NADAC is still relevant for this task due to this change in legislation. Looking into ways to incentivize providers to encourage lifestyle change could impact the number of prescriptions being filled by reducing chronic disease and the need for medication. Programs such as the Merit-Based Incentive Payment System (MIPS) could have measures developed to encourage providers to promote lifestyle change. While MIPS is a Medicare program, encouraging a MIPS measure will cascade to the entire population that the provider serves. The time and economic variables were not as impactful as originally thought but that is a pattern that may still be hidden and show itself as time goes on and more years can be added to the dataset.

There are several areas for potential improvement in the model. Namely, drug class was available but not able to be used as the classes were broad and not specific to a diagnosis. A single pharmacological class can be applied to drugs that treat multiple different diseases and drug class does not indicate what the drug is used to treat. Adding flags for common diseases or a flag that indicates that a drug is commonly used to treat a chronic disease can provide relevant information. This is knowledge that is outside the scope of this writer and is not easily available online. A consulting pharmacist will need to be brought on board to ensure the proper knowledge is available. Another piece of information that could be assessed is Medicaid beneficiaries with chronic diagnoses. This information was not a part of the scope of this project but investigating chronic disease presence in the Medicaid population and implementing interventions to reduce chronic disease and preventing the need for medication thus reducing the cost is an area for improvement. Since the most influential factor for total reimbursement is number of prescriptions filled finding ways to reduce utilization will have the highest impact on costs.

Since the dataset used to build this model had any record with a null value dropped there is a concern that valuable information may have been excluded from the model building process. However, since the records present in the dataset are guaranteed to be correct the regression model is likely more accurate than if imputations had been used. After 2023's data is available the model should be reassessed to determine if important patterns that were not available due to data incompleteness are now able to be added to the model to determine if the driving factors surrounding Medicaid drug reimbursement have shifted due to the new information.

There are a variety of factors that influence total Medicaid dollars reimbursed to the states by the federal government. The random forest regression model provides the most information about the relationship between the selected variables and total amount reimbursed by explaining 89.1% of the variability. The most influential is number of prescriptions and units reimbursed followed by the cost of the drug per unit which can be negotiated by the federal government directly with the pharmaceutical companies. It is important to highlight that this process was done without expert knowledge regarding prescription drugs and therefore important features may have been missed that a pharmacist would be able to identify. To work on improving the model a pharmacist should be consulted with to ensure the appropriate information about what a drug is used to treat can be considered to determine if there is a relationship between common/chronic diseases and what Medicaid drug reimbursement. The random forest regression model could benefit from further work, but it is a good starting point for determining what drives Medicaid drug spending and explaining much of what is happening with drug spending. However, due to about 11% of the variability in total drug reimbursement not being explained it is not recommended that this model be used to forecast Medicaid drug spending at this time. The previously made recommendations should be considered and more years of data should be added to determine if the time-based variables are more important than the current model would suggest. It is possible that with these changes that the model would be able to be used to forecast total amount reimbursed to the state by the federal government.

References

AIML.com. (2023, October 3). *What are the advantages and disadvantages of Decision Tree Model?* . AIML.com Machine Learning Resources. https://aiml.com/what-are-the-advantages-and-disadvantages-of-using-a-decision-tree/

AIML.com. (2023, October 3). *What are the advantages and disadvantages of Random Forest?*. AIML.com Machine Learning Resources. https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/

Anderson, L. A. (2023, November 21). *National Drug Codes explained: What you need to know*. National Drug Codes Explained. https://www.drugs.com/ndc.html

Bureau of Economic Analysis. (2024). *Section 1 Domestic Product and Income*. National Income and Product Accounts. https://apps.bea.gov/iTable/?isuri=1&reqid=19&step=4&categories=flatfiles&nipa_table_list=1

Centers for Medicare and Medicaid Services. (2024). NADAC Comparison. Medicaid. https://data.medicaid.gov/dataset/a217613c-12bc-5137-8b3a-ada0e4dad1ff

Centers for Medicare and Medicaid Services. (2024). Program Information for Medicaid and CHIP beneficiaries by Month. Medicaid. https://data.medicaid.gov/dataset/3da9f4e6-7976-43a8-8d1b-72f2c557a5ca

Centers for Medicare and Medicaid Services. (2023). State Drug and Utilization Database 2018. Medicaid. https://data.medicaid.gov/dataset/a1f3598e-fc71-51aa-8560-78e7e1a61b09/data

Centers for Medicare and Medicaid Services. (2023). State Drug and Utilization Database 2019. Medicaid. https://data.medicaid.gov/dataset/daba7980-e219-5996-9bec-90358fd156f1

Centers for Medicare and Medicaid Services. (2023). State Drug and Utilization Database 2020. Medicaid. https://data.medicaid.gov/dataset/cc318bfb-a9b2-55f3-a924-d47376b32ea3

Centers for Medicare and Medicaid Services. (2023). State Drug and Utilization Database 2021. Medicaid. https://data.medicaid.gov/dataset/eec7fbe6-c4c4-5915-b3d0-be5828ef4e9d

Centers for Medicare and Medicaid Services. (2023). State Drug and Utilization Database 2022. Medicaid. https://data.medicaid.gov/dataset/200c2cba-e58d-4a95-aa60-14b99736808d

Dranove, D., Ody, C., & Starc, A. (2021). A dose of managed care: Controlling drug spending in Medicaid. *American Economic Journal: Applied Economics*, *13*(1), 170–197. https://doi.org/10.1257/app.20190165

Freddie Mac (2024). 30-Year Fixed Rate Mortgage Average in the United States [MORTGAGE30US], retrieved from FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/MORTGAGE30US

Lewinson, E. (2023, June 12). *3 robust linear regression models to handle outliers*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/dealing-with-outliers-using-three-robust-linear-regression-models/

Ninja, N. (2023, October 11). *Target encoding: Categories guided by outcomes*. Let's Data Science. https://letsdatascience.com/target-encoding/

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Torné, R. V., Sala, E., Lió, P., Patel, M., Preller, J., Selby, I., Breger, A., Weir-McCall, J. R., Gkrania-Klotsas, E., Korhonen, A., Jefferson, E., Langs, G., … Schönlieb, C.-B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, *3*(1). https://doi.org/10.1038/s43856-023-00356-z

United States Census Bureau. (2021, October 8). *SAIPE datasets*. Census.gov. https://www.census.gov/programs-surveys/saipe/data/datasets.html

U.S. Bureau of Labor Statistics (2024). Consumer Price Index for All Urban Consumers (CPI-U) 2018-2022. BLS Beta Labs. https://beta.bls.gov/dataViewer/view/338ff7677f4e44a187efbe20dca439b2

U.S. Bureau of Labor Statistics (2024). Local Area Unemployment Rate Statewide Unemployment Rate 2018-2022 Seasonally Adjusted. BLS Beta Labs. https://beta.bls.gov/dataViewer/view/9764712c09b7444a836c489e28d3e999

U.S. Food and Drug Administration. (2023). National Drug Code Directory (NDC). openFDA. https://open.fda.gov/data/ndc/

Appendix A: Joining Variables for Datasets

- SDUD with unemployment rate by state, year, and quarter,
- SDUD with Medicaid beneficiaries by state, year, and quarter,
- SDUD with income and poverty estimates by state, year, and quarter,
- SDUD with NADAC on NDC, year, and quarter,
- SDUD with the FDA NDC dataset on NDC,
- SDUD with CPI on year and quarter,
- SDUD with GDP on year and quarter,
- SDUD with mortgage interest rates on year and quarter.

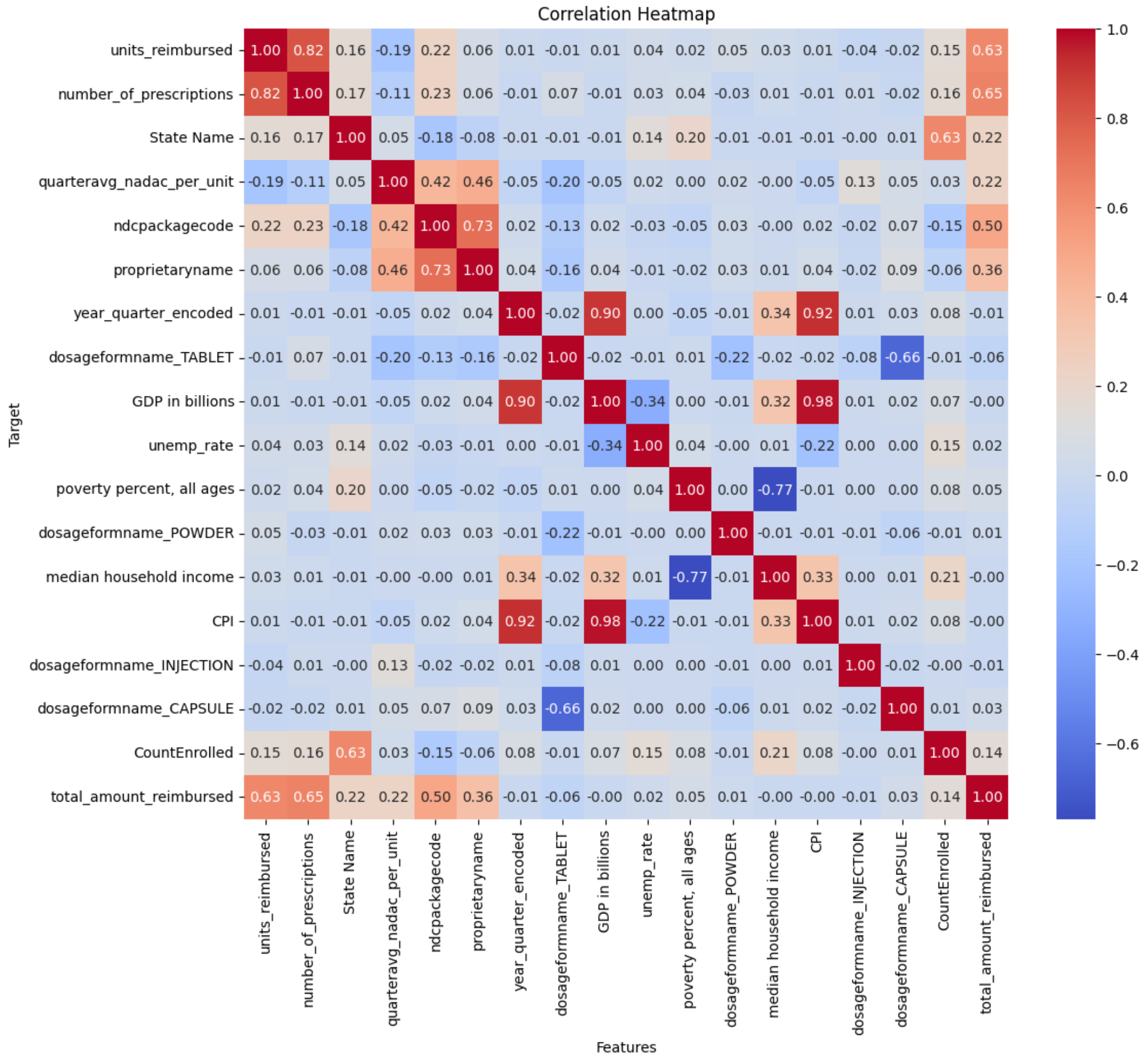
Appendix B: Null Value Counts by Variable

units_reimbursed: 0
number_of_prescriptions: 0
total_amount_reimbursed: 0
year: 0
product_name: 0
State Name: 0
Region: 0
year-quarter: 0
unemp_rate: 0
CountEnrolled: 120632
poverty estimate, all ages: 0
poverty percent, all ages: 0
median household income: 0
quarteravg_nadac_per_unit: 1970479
classification_for_rate_setting: 1970479
ndcpackagecode: 0
producttypename: 0
proprietaryname: 0
dosageformname: 0
marketingcategoryname : 0
labelername: 0
substancename: 140841
pharm_classes: 550892
deaschedule: 6544641
CPI: 0
GDP in billions: 0
MORTGAGE30US: 0

Appendix C: Extreme Outlier Counts

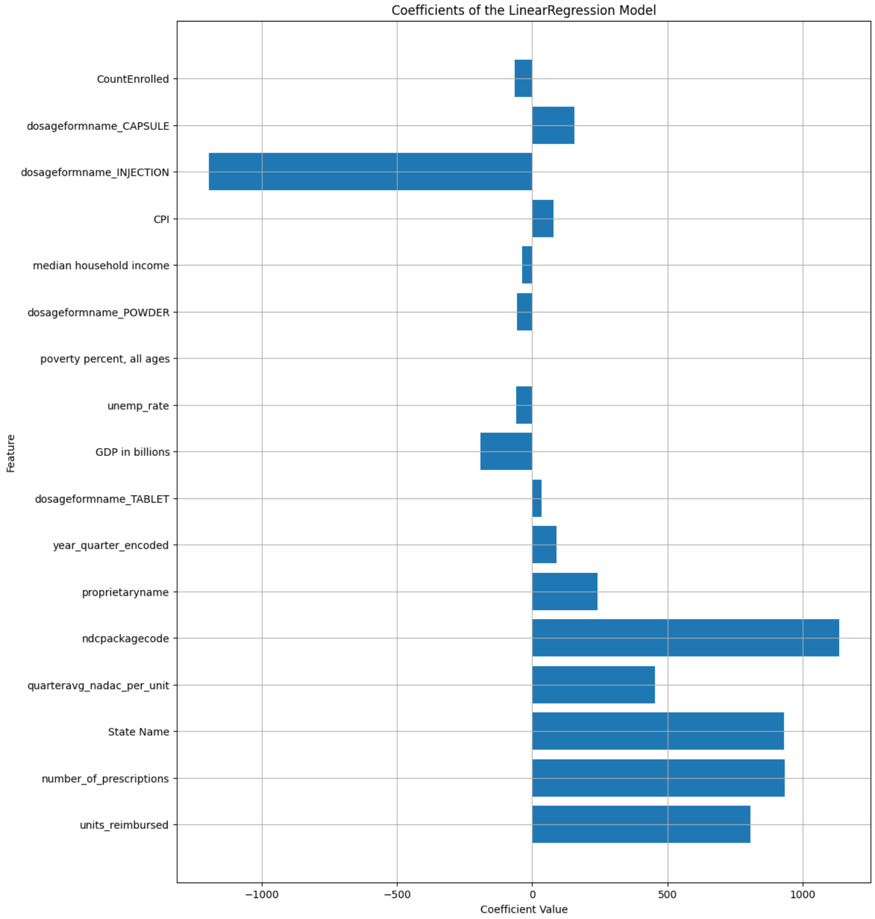
- Units reimbursed had 684,222 extreme outliers.
- Number of prescriptions had 653,332 extreme outliers.
- Total amount reimbursed had 724,513 extreme outliers.
- Count enrolled had 386267 extreme outliers.
- Poverty estimate, all ages had 264,340 extreme outliers.
- Average NADAC per unit had 1,003,380 extreme outliers.
- Unemployment rate had 299,985 extreme outliers

Appendix D: Correlation Heatmap



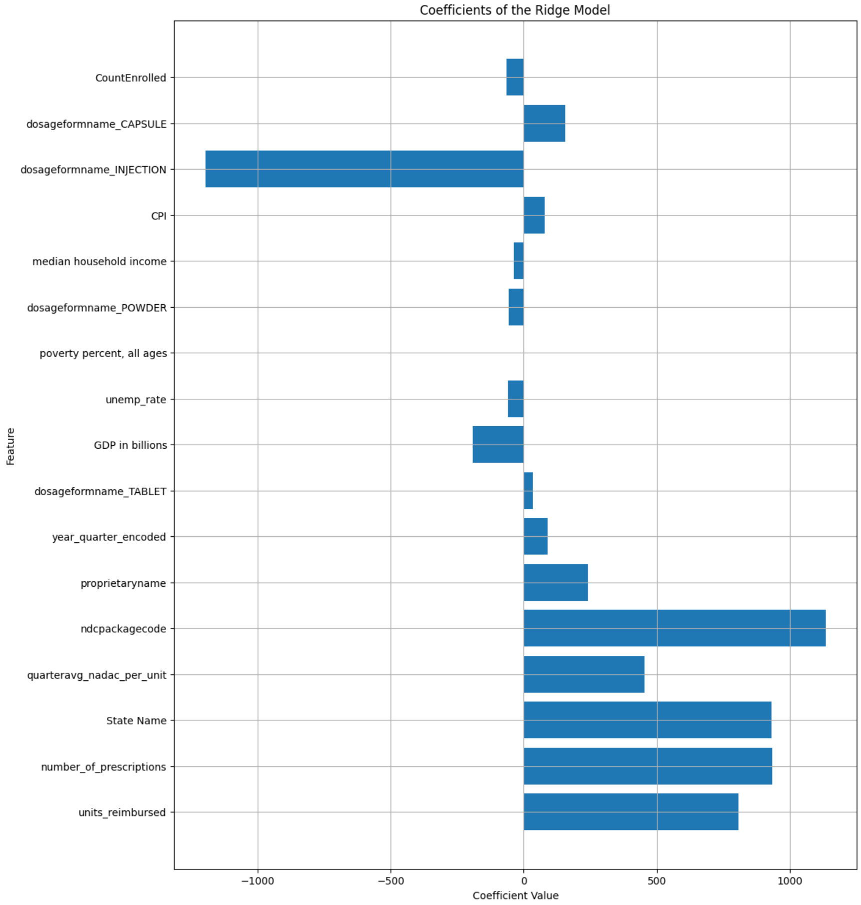
Appendix E: Coefficients for Linear Regression

Feature 1: units_reimbursed: 805.8478155284607
Feature 2: number_of_prescriptions: 932.7958398765895
Feature 3: State Name: 932.3349528298947
Feature 4: quarteravg_nadac_per_unit: 454.8684008496773
Feature 5: ndcpackagecode: 1135.391758038455
Feature 6: proprietaryname: 242.3077990013939
Feature 7: year_quarter_encoded: 89.44613299640392
Feature 8: dosageformname_TABLET: 34.97993416728937
Feature 9: GDP in billions: -190.54914213656212
Feature 10: unemp_rate: -58.820648757696006
Feature 11: poverty percent, all ages: -0.24574099155354467
Feature 12: dosageformname_POWDER: -55.964809918187015
Feature 13: median household income: -36.02967121074727
Feature 14: CPI: 79.90266377680516
Feature 15: dosageformname_INJECTION: -1197.0317150741007
Feature 16: dosageformname_CAPSULE: 155.45680017126196
Feature 17: CountEnrolled: -63.745673446674886



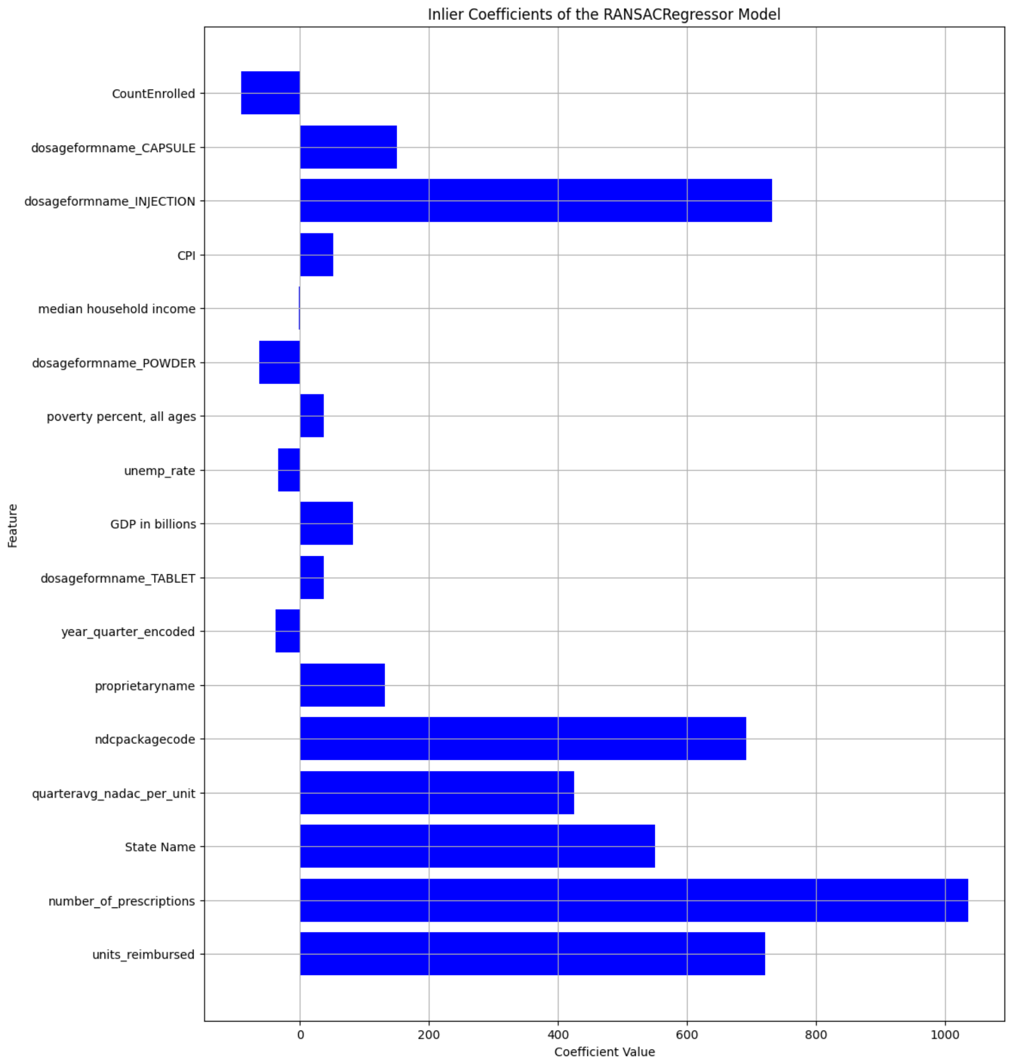
Appendix F: Coefficients for model: Ridge Regression

Feature 1: units_reimbursed: 805.848449730306
Feature 2: number_of_prescriptions: 932.7949024212112
Feature 3: State Name: 932.3339467047196
Feature 4: quarteravg_nadac_per_unit: 454.8673440545028
Feature 5: ndcpackagecode: 1135.3913948960185
Feature 6: proprietaryname: 242.30898629224043
Feature 7: year_quarter_encoded: 89.44091485177724
Feature 8: dosageformname_TABLET: 34.98250491785089
Feature 9: GDP in billions: -190.5338192321284
Feature 10: unemp_rate: -58.819241496662464
Feature 11: poverty percent, all ages: -0.24646591446453872
Feature 12: dosageformname_POWDER: -55.96163507320959
Feature 13: median household income: -36.03041123468374
Feature 14: CPI: 79.89467579256768
Feature 15: dosageformname_INJECTION: -1196.8737200129895
Feature 16: dosageformname_CAPSULE: 155.45908044559712
Feature 17: CountEnrolled: -63.74540322804764

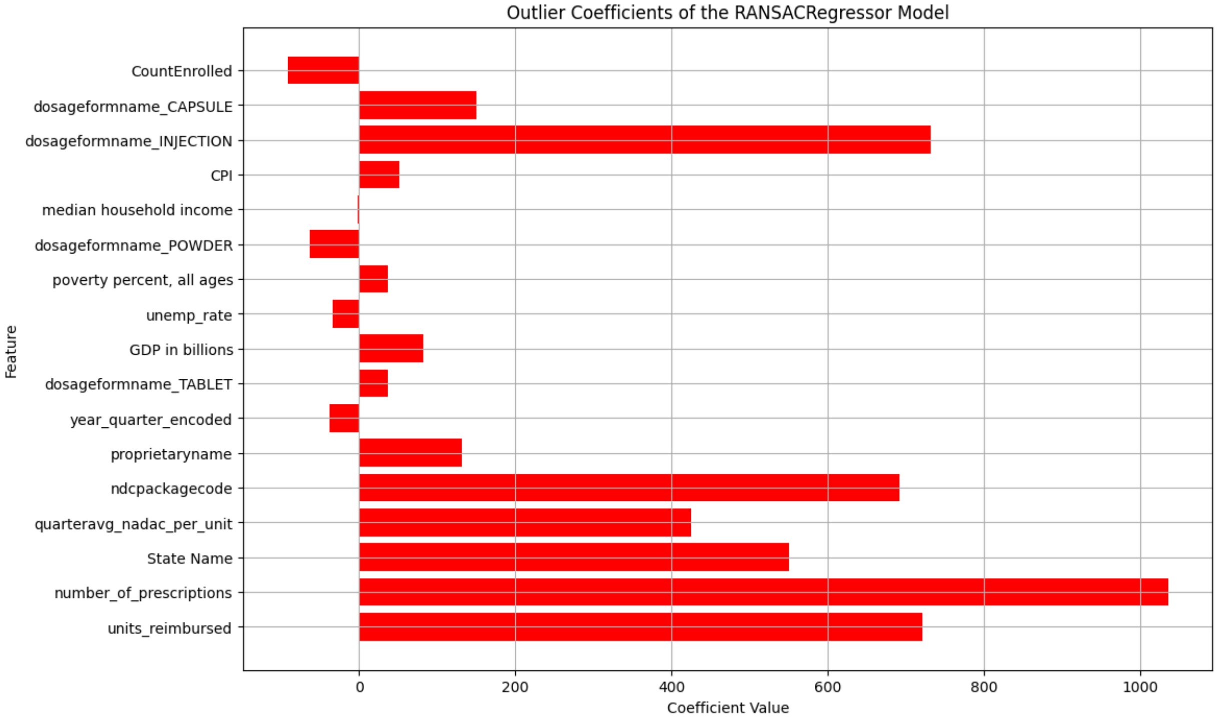


Appendix G: Coefficients for model: RANSAC Regression

Inlier Coefficients for model: RANSACRegressor
Feature 1: units_reimbursed: 721.7316831202443
Feature 2: number_of_prescriptions: 1036.0988959966419
Feature 6: proprietaryname: 132.38858604305346
Feature 8: dosageformname_TABLET: 37.67916300059159
Feature 9: GDP in billions: 82.98792740555349
Feature 11: poverty percent, all ages: 36.93635402642627
Feature 13: median household income: -0.9248505540393417
Feature 14: CPI: 51.96259068105955
Feature 15: dosageformname_INJECTION: 732.3840888344008
Feature 16: dosageformname_CAPSULE: 150.6985896857148
Feature 17: CountEnrolled: -91.54647119817338

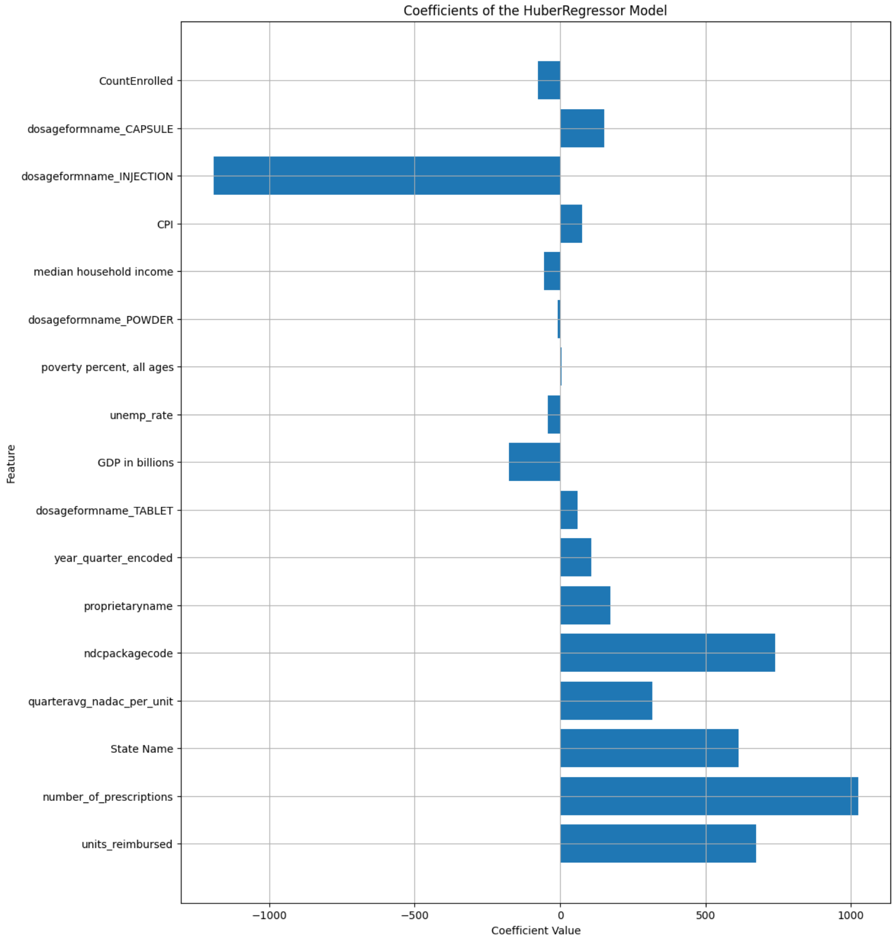


Outlier Coefficients for model: RANSAC Regression
Feature 3: State Name: 551.0401793079877
Feature 4: quarteravg_nadac_per_unit: 425.8881354379106
Feature 5: ndcpackagecode: 692.2350492612529
Feature 7: year_quarter_encoded: -37.63489910006101
Feature 10: unemp_rate: -34.18304792061245
Feature 12: dosageformname_POWDER: -63.25762200873235



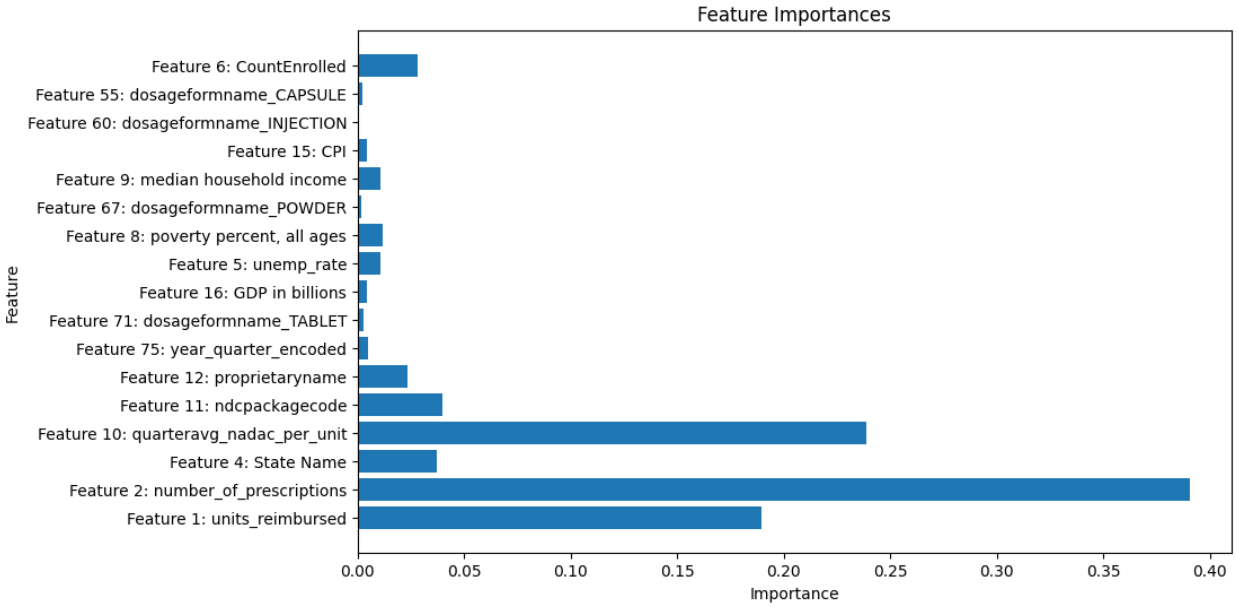
Appendix H: Coefficients for model: Huber Regression

Feature 1: units_reimbursed: 673.1768771551644
Feature 2: number_of_prescriptions: 1025.6372761767932
Feature 3: State Name: 613.4087248946981
Feature 4: quarteravg_nadac_per_unit: 316.88561589986625
Feature 5: ndcpackagecode: 739.4000520110825
Feature 6: proprietaryname: 173.02417401999156
Feature 7: year_quarter_encoded: 106.9347442979051
Feature 8: dosageformname_TABLET: 59.220535637635344
Feature 9: GDP in billions: -176.55495579927373
Feature 10: unemp_rate: -43.310853220999746
Feature 11: poverty percent, all ages: 3.748097584969089
Feature 12: dosageformname_POWDER: -8.763781946162007
Feature 13: median household income: -54.80497785338656
Feature 14: CPI: 74.67574833415532
Feature 15: dosageformname_INJECTION: -1192.49050861582
Feature 16: dosageformname_CAPSULE: 152.30132888930441
Feature 17: CountEnrolled: -76.90392817581557



Appendix I: Decision Tree Feature Importance

```
Feature Importance:
Feature 1: units_reimbursed: 0.1898
Feature 2: number_of_prescriptions: 0.3906
Feature 4: State Name: 0.0372
Feature 10: quarteravg_nadac_per_unit: 0.2389
Feature 11: ndcpackagecode: 0.0395
Feature 12: proprietaryname: 0.0233
Feature 75: year_quarter_encoded: 0.0046
Feature 71: dosageformname_TABLET: 0.0025
Feature 16: GDP in billions: 0.0043
Feature 5: unemp_rate: 0.0105
Feature 8: poverty percent, all ages: 0.0116
Feature 67: dosageformname_POWDER: 0.0017
Feature 9: median household income: 0.0105
Feature 15: CPI: 0.0044
Feature 60: dosageformname_INJECTION: 0.0003
Feature 55: dosageformname_CAPSULE: 0.0021
Feature 6: CountEnrolled: 0.0282
```


Feature Importances
Feature
Feature 6: CountEnrolled
Feature 55: dosageformname_CAPSULE
Feature 60: dosageformname_INJECTION
Feature 15: CPI
Feature 9: median household income
Feature 67: dosageformname_POWDER
Feature 8: poverty percent, all ages
Feature 5: unemp_rate
Feature 16: GDP in billions
Feature 71: dosageformname_TABLET
Feature 75: year_quarter_encoded
Feature 12: proprietaryname
Feature 11: ndcpackagecode
Feature 10: quarteravg_nadac_per_unit
Feature 4: State Name
Feature 2: number_of_prescriptions
Feature 1: units_reimbursed
Importance