Predicting Medicare Provider Prescribing Patterns

Sara Murley

May 2025

1 Research Question

The study seeks to answer the question: **What factors predict whether a Medicare Part D provider will prescribe a high proportion of brand-name drugs?**

2 Background

The cost of prescription drugs in the United States has been a long-standing issue, with drug prices far exceeding those in other developed nations [Mulcahy et al., 2024]. The financial burden of prescription drug costs strongly impacts older adults, including the more than 1 in 3 Medicare beneficiaries reporting incomes below 200% of the Federal Poverty Level (FPL) [Leonard et al., 2023]. In 2022, high costs forced one in five U.S. adults over 65 to skip or delay filling a prescription [Gumas et al., 2024]. These cost-coping strategies can have particularly serious consequences for older people who rely on these medications to control chronic conditions.

Medicare Part D is a voluntary benefit for people with Medicare provided through private insurance plans that contract with the federal government to cover payments for prescription drugs. In 2024, 53 million of the 67 million Medicare beneficiaries were enrolled in Part D plans [Cubanski and Damico, 2024]. Gross Part D expenditures have been increasing, reaching over $216 billion in 2021 and representing 15% of net total Medicare spending [Cubanski and Neuman, 2023, Cubanski, 2024]. The problems with Medicare Part D have become particularly pressing as the high cost of prescription drugs has increasingly threatened the financial stability of both the federal program and the senior citizens who rely on it.

Prescription drug spending has increased more than sevenfold from 1980 to 2018, driven by the development and use of many drugs proven to have significant health benefits and successfully manage chronic conditions [Hayford and Austin, 2022]. However, the growth has slowed since the mid-2000s due to the growing availability of generic alternatives. Generic drugs contain the same active ingredients as their brand-name counterparts but are often sold for a fraction of the price as their manufacturers do not face the steep research and development costs. According to a study done at the Johns Hopkins Bloomberg School of Public Health, the Medicare Part D program would have saved $977 million in a single year if all branded prescription drugs requested by prescribing clinicians had been substituted by a generic option [Socal et al., 2021].

Most Part D plans are designed to encourage utilization of low-cost drugs through benefit design and out-of-pocket costs [Dusetzina et al., 2020]. Furthermore, all 50 states have some

policy encouraging and regulating the use of generic substitution, though the design and features of these laws vary widely [Rome et al., 2022]. Despite government efforts at the federal and state level to promote generic drug dispensing, significant variability remains in prescribing patterns among Medicare Part D providers. In 2017, prescribing clinicians and patients together requested brand-name prescription drugs over generics 30 percent of the time [Socal et al., 2021]. Furthermore, patients would have saved $161 million if providers had preferred generic drugs [Socal et al., 2021]. The high costs remain although brand-name drug dispensing accounts for only 5% of Part D claims, emphasizing just how costly these drugs are [Socal et al., 2021].

The reasons behind differences in prescriber patterns are complex and attributable to numerous factors, including provider characteristics, patient demographics, and external influences such as pharmaceutical marketing and state regulations. Understanding these factors is crucial to designing more effective policies that promote cost-effective prescribing practices without compromising patient care. Employment of a predictive algorithm could also be used for targeted interventions to providers that are more likely to prefer brand-name drugs.

3 Literature Review

There is limited research into the factors that influence provider prescribing patterns. The key motivation for this analysis comes from an observational study from 2018 that examined the relationship between pharmaceutical industry payments to physicians and prescriptions for a select few high-cost brand-name drugs [Sharma et al., 2018]. The study found that a financial relationship between providers and pharmaceutical companies was associated with increased odds of prescribing high-cost drugs [Sharma et al., 2018]. The proposed study aims to expand on this research by analyzing more drugs and including more factors, as well as developing a comprehensive predictive algorithm.

There have been numerous studies performed internationally examining prescribing patterns. A study in Italy examined the factors influencing general practitioners' prescribing behavior, including policy, education and experience, demographic trends, and disease profiles [Tuncay et al., 2020]. Another study examined patterns in the UK [Stern, 1994]. In the US, studies have namely focused on patient preference and consumer purchasing patterns between brand-name and generic drugs or the dispensing of drugs to certain demographics [Howard et al., 2018, Kohli and Buller, 2013, Chen and Wu, 2008].

This study seeks to build on existing research into provider prescribing behaviors, using publicly available Medicare data as a basis to examine patterns in the US healthcare system broadly. Factors included in the study will draw on relationships that have been established through the studies discussed above, as well as insights from domain experts. The inclusion of numerous factors will serve to develop a comprehensive algorithm.

4 Data Sources

To develop a predictive model, the study will combine multiple publicly available datasets containing information on Medicare Part D prescribing behavior, provider characteristics, and pharmaceutical payments. The primary dataset will be the **Medicare Part D Prescribers** dataset from the Centers for Medicare & Medicaid Services. The dataset contains information of total prescriptions by physicians and other providers separated by brand

name, generic, and other drugs. The dataset also includes general information about each provider, including beneficiary information and specialty types.

To include pharmaceutical industry payments to providers as a factor in our model, the **2022 Open Payments** dataset from CMS will be included. The dataset will be combined with the original dataset using the recipient NPI (National Provider Identifier), a unique identification number that is assigned to the provider. The data is aggregated by provider, with total payments to a provider that are not related to research being the feature of interest.

The datasets are of high quality and include large amounts of observations. However, the data is pulled from numerous sources and may include incorrect or incomplete information, especially for smaller providers. There will also likely be challenges with integrating them, as Part D prescribers only account for some of the providers in the Open Payments dataset and the two datasets are not necessarily reflective of the same Medicare beneficiary populations. Furthermore, it will be crucial to check the identifiers used to merge the two datasets. There are limitations to the Part D Prescribers dataset, as noted by CMS, "Therefore, because the drug expenditures derived from the Prescription Drug Event data comprise only a piece of the payment process, it is not possible to directly attribute total drug costs at the prescriber or drug level to payments from the Medicare trust fund. Furthermore, these total drug costs do not reflect any manufacturer rebates." The Medicare payment process is very complicated, and the analysis proposed here will be limited in scope and not reflect total Medicare prescription drug spending. Furthermore, the prescription totals accounted for in this dataset reflect only the Medicare Part D population and are not necessarily representative of the overall prescribing behavior of the provider.

The datasets will be cleaned and combined to form a dataset aggregated by provider. The independent variables will include provider characteristics and payments from the drug manufacturer to the provider. The target variable is whether a provider prescribes less than the 25th percentile of proportions of generic drugs of the entire dataset, based on their prescribing activity for the year 2022.

4.1 Feature Selection

The Medicare Part D prescribers dataset contains information on prescription drugs provided to Medicare Part D beneficiaries by providers (identified by a unique NPI). The type and specialty of provider will be used in the model, as well as the RUCA code (the rural/urban classification of the zip code where the provider is located). The dataset is organized by NPI and includes overall drug utilization (claims, 30-day standardized fill counts and day's supply), drug costs, and beneficiaries. The dataset also includes drug utilization counts separated by brand drugs, generic drugs, and other drugs. In addition, beneficiary demographic and health characters are provided. The demographic counts will be divided by the total beneficiaries served by the provider to create proportions. The dataset also includes flags for the reason behind suppressing certain variables, which will be used to inform decisions about handling missing data.

The goal of this model is to predict providers that fall above the 75th percentile in proportion of non-generic drug prescriptions. In order to measure the volume of prescriptions, we will use the total claims of brand, generic, and other drugs filed by beneficiaries of that provider. This variable includes all relevant prescriptions and will be used to create a binary target variable for classification.

The open payments dataset will be used to add some additional features to the provider

characteristics. The dataset includes non-research related payments from companies to providers. Summary data from each provider about the payments they have received related to the promotion of a drug or biological will be important features in the predictive model.

The first step to preprocessing this dataset is to filter the data to only relevant payments. Payments that were transferred to a third-party entity at the request of or on behalf of a covered recipient will be removed, as they are not directly received by the provider. There is data to indicate each product listed in relation to each payment, including whether the product is a drug, device, biological, or medical supply. The payments of interest are for drugs and biologicals only.

To calculate the feature values, the dataset will be aggregated according to the behavior of each provider. Two variables will be used, the total payment amount in USD and the physician ownership indicator. The indicator reveals whether physician holds an ownership or investment interest in the applicable manufacturer. Thus, the payment amounts will be separated by whether the provider holds ownership or not. For rows missing the ownership value, a value of "No" will be imputed, as this is the most likely scenario according to the legal documentation of the dataset. For each provider, there will be two features, total payments received related to manufacturers in which the physician has vested interests, and the sum of any other remaining payments.

4.2 Descriptive Statistics

Table 1: Numeric Descriptive Statistics

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Tot_Clms	2221.45	3728.21	11	281716
Tot_Benes	272.36	523.92	11	204782
Bene_Avg_Age	70.34	4.53	11.93	93.85
Bene_Avg_Risk_Score	1.57	0.70	0.36	10.32
Bene_Age_LT_65_Prop	0.20	0.14	0	1
Bene_Age_65_74_Prop	0.43	0.09	0	1
Bene_Age_75_84_Prop	0.26	0.10	0	0.80
Bene_Age_GT_84_Prop	0.11	0.06	0	1
Bene_Feml_Prop	0.60	0.11	0	1
Bene_Male_Prop	0.40	0.11	0	1
Bene_Race_Wht_Prop	0.71	0.25	0	1
Bene_Race_Black_Prop	0.14	0.18	0	1
Bene_Race_Api_Prop	0.03	0.07	0	1
Bene_Race_Hspnc_Prop	0.09	0.17	0	1
Bene_Race_Natind_Prop	0.01	0.04	0	1
Bene_Race_Othr_Prop	0.03	0.02	0	0.61
Bene_Dual_Prop	0.34	0.21	0	1
Bene_NDual_Prop	0.66	0.21	0	1

4.3 Preprocessing and Final Dataset

Below are the preprocessing steps for both the public datasets used in this project. The two resulting datasets were joined on the unique provider NPI. Any providers that did

Table 2: Categorical Feature Counts (RUCA)

RUCA Code	Count
4	48,305
7	20,570
2	18,353
10	8,435
5	3,613
3	1,253
6	471
9	424
99	418

Table 3: Categorical Feature Counts (Types, Top 10)

Type	Count
Nurse Practitioner	127,625
Family Practice	81,974
Internal Medicine	70,280
Physician Assistant	68,748
Emergency Medicine	35,247
Dentist	17,280
Cardiology	13,933
Student in an Organized Health Care	11,937
Psychiatry	11,194

Table 4: Target Feature Counts

Category	Count
1 (High Non-Generic Prescriber)	454,717
0 (Normal Prescriber)	151,569

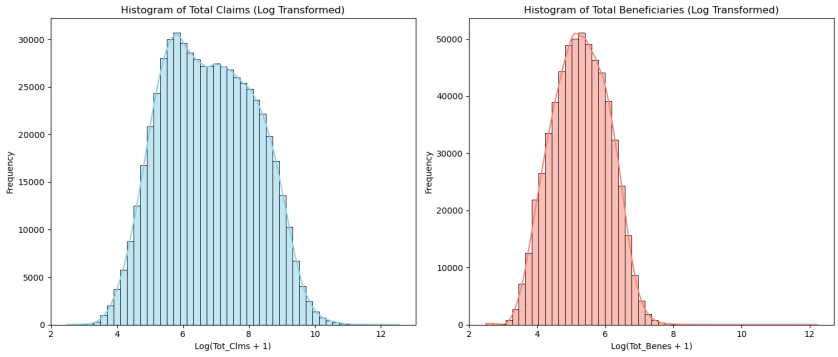

Figure 1: Histograms of Total Claims and Total Beneficiaries.

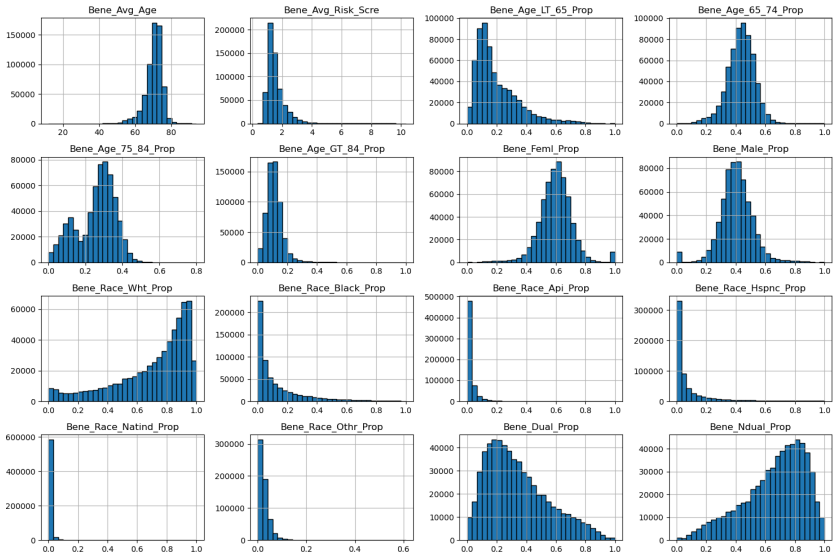

Figure 2: Numerical Feature Distributions

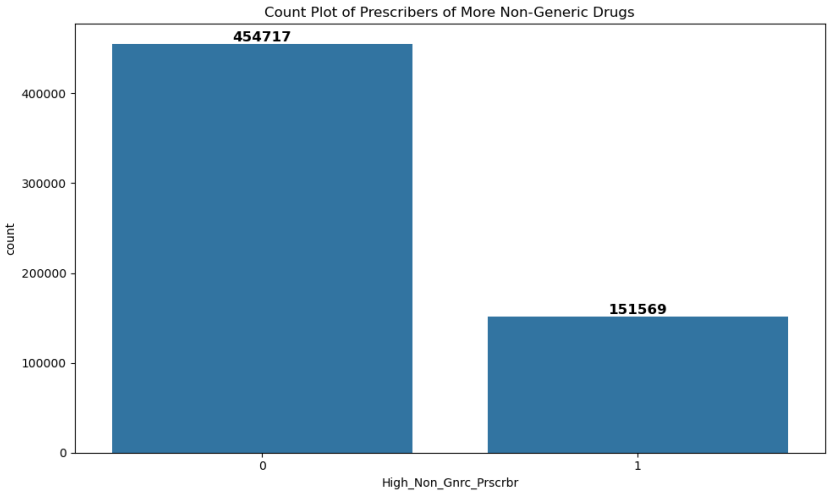

Figure 3: Distribution of Target Feature.

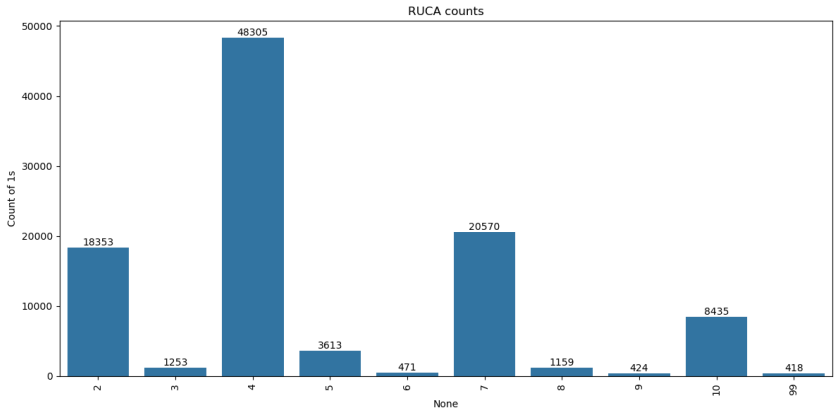

Figure 4: Counts for RUCA values.

not appear in the CMS Open Payments datasets will have a 0 imputed for their payment amounts.

Medicare Part D Provider Dataset

The first preprocessing step for this dataset is to handle missing data. Rows with crucial data missing from features that are not frequently missing throughout the entire dataset are removed to avoid the introduction of bias without a large sacrifice of information. This includes the RUCA code, total beneficiaries, and total claims data. The variables needed for the target features, total claims filed for brand, generic, and other drugs, also have suppression flags that indicate the reason for missing data. The values are as below:

$$\begin{cases} * : \text{Primary suppressed due to a value between 1 and 10} \\ \# : \text{Counter-suppressed because of primary suppression of another feature} \end{cases}$$

Using these flags, the first approach to handling missing data is to impute where possible. If all three values are suppressed, the rows are removed (this occurs when the total claims value is between 1 and 10). Such small amounts of total claims occur for small providers, likely individual practices, and can reasonably be excluded from the training of the model. If only one value is suppressed due to small amounts, the value can be calculated based on the other two values and the total claims. If two are suppressed, the imputed value will be an equal distribution of the remaining claims (claims in the total claims not accounted for by the known value) among the two unknown features.

A similar process will be followed for the beneficiary demographic proportions, by age, race, gender, and dual enrollment. The values are suppressed if the counts are between 1 and 10, and some are counter suppressed as a result. For each group (age, gender, race, and enrollment), the following process is followed to impute missing values:

1. Rows with missing values across all subgroups are removed.

2. Rows with exactly 1 missing value are imputed by calculating the remaining proportion out of 100%.

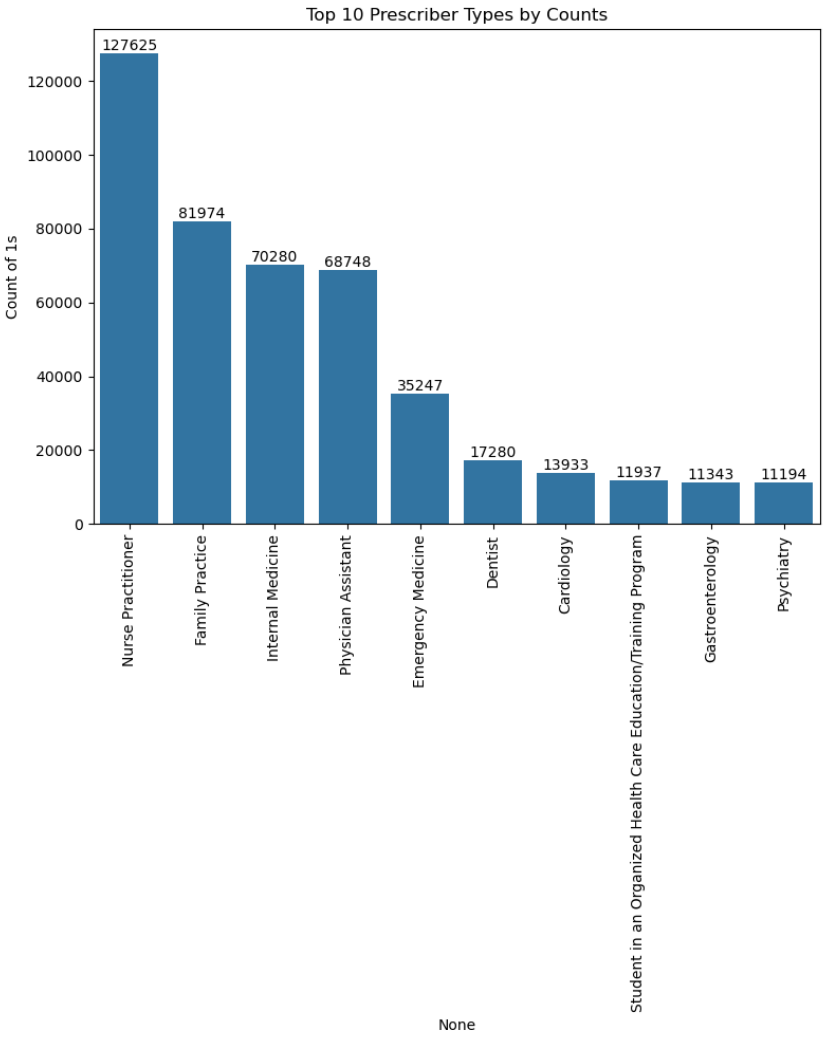

Figure 5: Top 10 Prescriber Types.

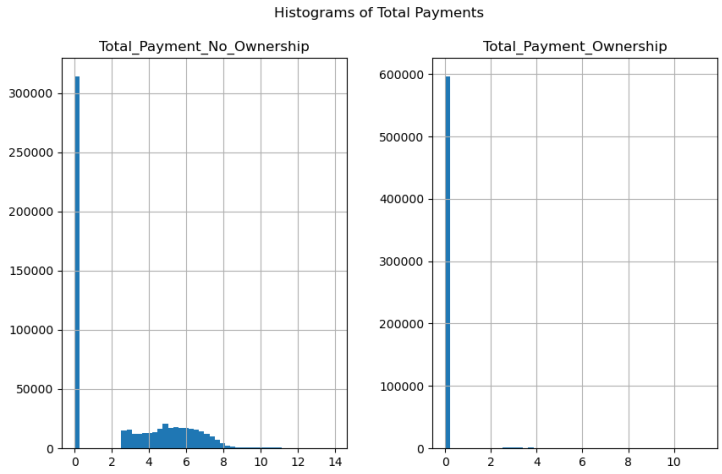

Figure 6: Histograms of Total Pharmaceutical Payments.

3. Rows with multiple missing values will have the remaining proportions distributed equally among them.

The above process will introduce a minimal amount of bias, given that data is missing because of small counts. Values that are still included are the most representative of the beneficiary characteristic, and thus imputing missing values with an approximation of small proportions should allow the features to be as accurate to reality as possible.

The next step for preprocessing the prescriber dataset is to handle categorical features. This includes the prescriber type and prescriber RUCA values, both of which will be converted to a series of dummy variables. The RUCA primary values are below. Secondary values (decimal points) indicate further information than the primary points (for example, smaller vs larger distance to the metropolitan area). To minimize the number of features introduced into the dataset from converting this to a series of dummy variables, all of the secondary values will be converted to the primary values by converting the float values to integers.

The next aspect of pre-processing is to calculate the target variable based on the number of claims filed for each type of prescription drug. First, the proportion of non-generic drug claims filed by the provider will be calculated. Then, a dichotomous variable will be created based on the 75th percentile of this proportion distribution. The final target variable will be categorical in type, with a value corresponding to the type of drugs most prescribed by this provider. Below is a map of values:

$$\begin{cases} \mathbf{0} : \text{Normal Prescriber} \\ \mathbf{1} : \text{High Non-Generic Prescriber} \end{cases}$$

CMS Open Payments Dataset

The preprocessing for this dataset is relatively simple. Since the features are aggregated by physician, there are no missing values (physicians with no received payments related to

Table 5: Primary RUCA Codes, 2010

Code	Classification Description
1	Metropolitan area core: primary flow within an urbanized area (UA)
2	Metropolitan area high commuting: primary flow 30% or more to a UA
3	Metropolitan area low commuting: primary flow 10% to 30% to a UA
4	Micropolitan area core: primary flow within an urban cluster of 10,000 to 49,999 (large UC)
5	Micropolitan high commuting: primary flow 30% or more to a large UC
6	Micropolitan low commuting: primary flow 10% to 30% to a large UC
7	Small town core: primary flow within an urban cluster of 2,500 to 9,999 (small UC)
8	Small town high commuting: primary flow 30% or more to a small UC
9	Small town low commuting: primary flow 10% to 30% to a small UC
10	Rural areas: primary flow to a tract outside a UA or UC
99	Not coded: Census tract has zero population and no rural-urban identifier information

ownership or non-ownership will have an imputed value of 0). The dataset does not contain categorical variables.

Table 6 contains a summary of the key features in the model, including the target variable of preferred drug type.

5 Methodology and Implementation

5.1 Preparing the Data

1. Data Splitting: First, the data is split into training, validation, and test sets. This is done in two steps: an 80-20 split for training and testing, followed by a 70-10 split of the training set into separate training and validation sets. This ensures that both the training and validation sets are representative of the overall data distribution, but the model can still be tested on unseen data (the test set).

2. Feature Scaling: The numerical features of the resampled training set, as well as the validation and test sets, are standardized using StandardScaler. This ensures that the model's learning process isn't biased by differences in the scale of features. The Scaler is fit to the training set and applied to the validation and test sets to avoid data leakage.

5.2 Model Selection

In this section, we outline the classification models selected for implementation and comparison, ranging from most interpretable to least interpretable and least complex to most complex. The goal is to identify a model that not only handles class imbalance effectively but also accurately predicts class assignments, particularly for the minority class. All four models will be compared for the initial results on the training set without hyperparameter tuning. The top-performing models will be selected to be tuned with hyperparameters, where the best model will be ultimately selected and investigated further.

Logistic Regression is a simple and highly interpretable model, which serves as our baseline for comparison with more complex models. It works well with imbalanced data when class weights are adjusted (using the class_weight='balanced' parameter in sklearn).

Table 6: Feature Descriptions

Description	**Feature Name**	**Values**
Provider ID	Prscrbr_NPI	National Provider Identifier (NPI) for the performing provider on the claim.
Number of Medicare Part D Claims, Including Refills	Tot_Clms	The number of Medicare Part D claims. This includes original prescriptions and refills.
Number of Medicare Beneficiaries	Tot_Benes	The total number of unique Medicare Part D beneficiaries with at least one claim for the drug. Counts fewer than 11 are suppressed and are indicated by a blank.
Average Age of Beneficiaries	Bene_Avg_Age	Average age of beneficiaries.
Average Hierarchical Condition Category (HCC) Risk Score of Beneficiaries	Bene_Avg_Risk_Scre	Average Hierarchical Condition Category (HCC) risk score of beneficiaries.
Proportion of beneficiaries per age group	Bene_Age_{}Prop	Proportion of beneficiaries within a specified age range.
Proportion of beneficiaries per racial group	Bene_Race_{}Prop	Proportion of beneficiaries per racial group.
Proportion dual enrollment	Bene_Dual_Prop	Proportion of beneficiaries with Medicare/Medicaid Dual enrollment.
Proportion not dual enrollment	Bene_Ndual_Prop	Proportion of beneficiaries without Medicare/Medicaid Dual enrollment.
Prescriber RUCA	Prscrbr_RUCA_{}	Rural-Urban Commuting Area Codes (RUCAs) dummy variables.
Provider Specialty Type	Prscrbr_Type_{}	Derived from the Medicare provider/supplier specialty code reported on the NPI's Part B claims.
Total Drug Payments (No Ownership)	Total_Payment_No_Ownership	Total non-research related payments for drugs/biologicals received by the provider in 2022 from manufacturers with which they have no vested interest.
Total Drug Payments (Ownership)	Total_Payment_Ownership	Total non-research related payments for drugs/biologicals received by the provider in 2022 from manufacturers with which they have a vested interest.
Target	High_Non_Generic_Prscrbr	Indicator for being above the 75th percentile of non-generic prescribers.

This model outputs probability scores, providing insight into the confidence of predictions, and the associated coefficients show how each feature influences the likelihood of predicting brand-name prescribing. Given its simplicity, Logistic Regression is a useful starting point to understand the relationship between features and the target variable before exploring more sophisticated techniques.

Decision Tree learns complex patterns and interactions within the data while being less sensitive to class imbalance than Logistic Regression. It is also highly interpretable, as the structure of the tree provides clear rules for decision-making. This model allows for easy visualization and understanding of how different features contribute to predictions, making it a valuable tool for gaining insights into the decision process. However, Decision Trees can be prone to overfitting without proper regularization.

Random Forest is an ensemble method that reduces overfitting by averaging the results of multiple decision trees, improving the model's generalizability. It is less sensitive to outliers compared to Decision Trees and provides valuable insights into feature importance. However, Random Forest is more complex and computationally expensive, and it is less interpretable than individual Decision Trees due to the aggregation of multiple trees. Despite its increased complexity, it can offer better performance in terms of handling imbalanced data and capturing intricate patterns.

XGBoost is a powerful ensemble method that excels in handling class imbalance through the use of custom evaluation metrics. It is highly efficient and scalable, making it suitable for large datasets. XGBoost automatically captures feature interactions and complex patterns, which makes it a powerful tool for improving prediction accuracy. However, it is also the least interpretable model due to its ensemble nature and the complexity of the boosting process. Despite this, its ability to achieve high predictive performance and handle imbalance makes it a strong contender for the final model choice.

5.3 Evaluation Metrics

To evaluate the performance of the models, we will consider several metrics.

- **Accuracy** represents the proportion of correct predictions, but it may be misleading in the case of imbalanced data, as it does not account for the distribution of the classes.
- **Precision** for Class 1 (brand-name) is the proportion of true positives (correctly predicted brand-name prescribers) among all instances predicted as brand-name prescribers.
- **Recall** for Class 1 (brand-name) is the proportion of true positives among all actual brand-name prescribers, providing insight into the model's ability to identify brand-name prescribers.
- The **F1-score** is the harmonic mean of precision and recall, balancing both metrics and providing a more comprehensive evaluation of model performance.
- The Area Under the Receiver Operating Characteristic Curve **(AUC-ROC)** measures the model's ability to discriminate between classes at various classification thresholds.

For hyperparameter tuning, we will primarily focus on the F1 score as our metric. The research goal is to understand the factors that predict whether a prescriber prefers brand name or generic drugs, so it is crucial that we capture as many positive cases as possible while still maintainng high precision.

5.4 Initial Results

The table below shows the performance of each model on the test data without hyperparameter tuning:

Model	Accuracy	Precision	Recall	F1	ROC_AUC
Logistic Regression	0.82	0.80	0.39	0.52	0.82
Decision Tree	0.77	0.54	0.56	0.55	0.70
Random Forest	0.84	0.82	0.47	0.60	0.88
XGBoost	0.79	0.56	0.78	0.66	0.88

Table 7: Performance of each model without hyperparameter tuning, with the maximum value in each column highlighted

Below are the confusion matrices:

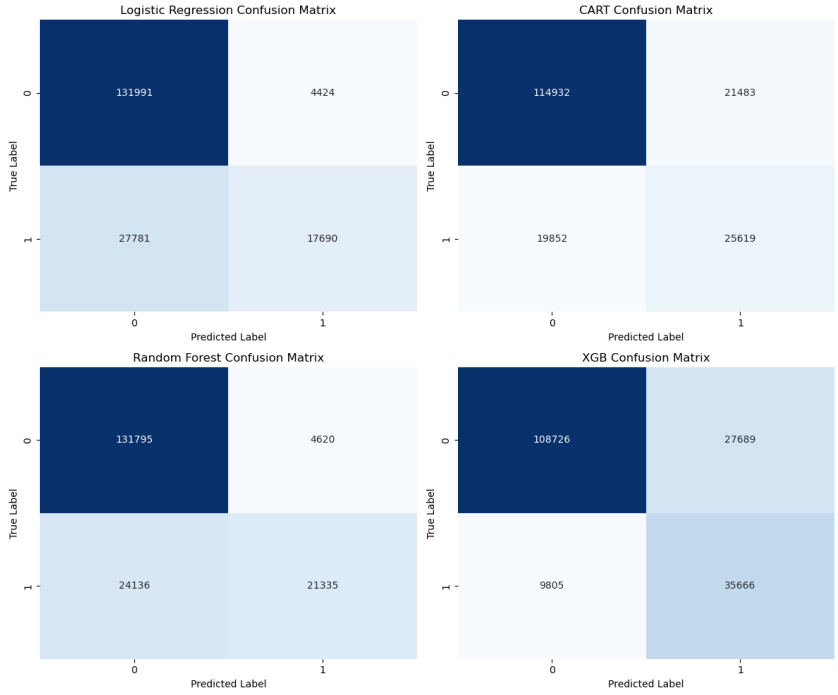

Figure 7: Initial Confusion Matrices.

Based on our overarching evaluation metric of the F1 score, the highest performing model is XGBoost. However, the Random Forest also performs well, especially for accuracy and precision. It is also useful to retain the Decision Tree as a baseline model and to explore model complexity. As such, we will proceed with hyperparameter tuning for these 3 models.

5.5 Hyperparameter Tuning

To optimize the performance of the machine learning models used in predicting prescribers with higher rates of brand-name drug prescriptions, I implemented a hyperparameter tuning process using Randomized Search with cross-validation. The models evaluated included Decision Tree, Random Forest, and XGBoost classifiers. Given the class imbalance in the dataset, I incorporated strategies such as class weighting and scale_pos_weight to improve recall for the minority class (prescribers favoring brand-name drugs).

The **Decision Tree** model was optimized by tuning hyperparameters such as max_depth, $min\_samples\_split$, $min\_samples\_leaf$, $class\_weight$, $max\_features$, and ccp_alpha. Randomized Search was performed over 10 iterations with stratified k-fold cross-validation (k=3). The best parameters identified were:

$$\begin{cases} max\_depth : 15 \\ min\_samples\_split : 20 \\ min\_samples\_leaf : 10 \\ class\_weight : \text{balanced} \\ max\_features : \text{None} \\ ccp\_alpha : 0.01 \end{cases}$$

The cross-validation F1 score on the training set was 0.505, and on the validation set, the recall was 0.721, with an F1-score of 0.504. Despite achieving reasonable recall, the Decision Tree model exhibited limited predictive power and was prone to overfitting.

For the **Random Forest** model, hyperparameters such as $n\_estimators$, max_depth, $min\_samples\_split$, $min\_samples\_leaf$, $class\_weight$, $max\_features$, and max_leaf_nodes were tuned. The best parameters found were:

$$\begin{cases} n\_estimators : 150 \\ max\_depth : 9 \\ min\_samples\_split : 10 \\ min\_samples\_leaf : 2 \\ class\_weight : \text{balanced} \\ max\_features : \log 2 \\ max\_leaf\_nodes : 20 \end{cases}$$

The mean F1 score from cross-validation was 0.588, while the validation recall reached 0.745, with an F1-score of 0.587. This model performed significantly better than the Decision Tree, offering a balance between recall and precision while mitigating overfitting.

XGBoost was optimized using a more extensive set of hyperparameters, including $n\_estimators$, max_depth, $learning\_rate$, $subsample$, $colsample\_bytree$, $gamma$, and $scale\_pos\_weight$. The best-performing model was found with:

$$\begin{cases} n\_estimators : 150 \\ max\_depth : 5 \\ learning\_rate : 0.2 \\ subsample : 1.0 \\ colsample\_bytree : 0.8 \\ gamma : 5 \\ scale\_pos\_weight : 3.0 \end{cases}$$

The model achieved a mean training recall of 0.779 and a validation recall of 0.780, with an F1-score of 0.649. XGBoost outperformed both Decision Tree and Random Forest in recall and overall performance, making it the preferred model for identifying prescribers with higher brand-name drug preferences.

5.6 Results After Hyperparameter Tuning

Below are the confusion matrices and tabular results for each of the best models:

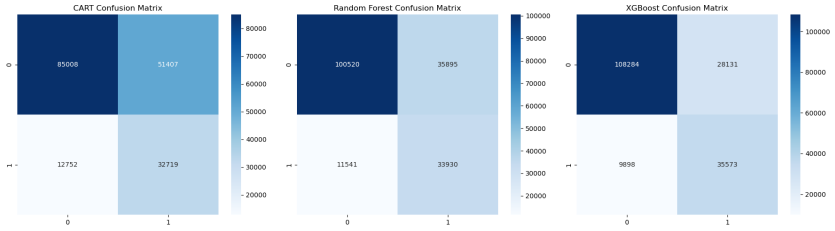

Figure 8: Final Confusion Matrices.

Model	Accuracy	Precision	Recall	F1	ROC_AUC
Decision Tree	0.65	0.39	0.72	0.50	0.70
Random Forest	0.74	0.49	0.75	0.59	0.82
XGBoost	0.79	0.56	0.78	0.66	0.87

Table 8: Performance of each model after hyperparameter tuning, with the maximum value in each column highlighted

Hyperparameter tuning played a crucial role in refining model performance. Although some metrics were lower after hyperparameter tuning, the models were simplified in complexity and improved significantly in recall. The Decision Tree, while interpretable, was outperformed by ensemble methods. Random Forest provided improvements in recall and F1-score, but XGBoost yielded the highest recall and overall balanced performance. Given the imbalanced nature of the dataset, recall was prioritized to minimize false negatives, ensuring that prescribers with higher brand-name prescribing tendencies were correctly identified. These results suggest that XGBoost, with optimized hyperparameters, is the most effective model for this classification task.

5.7 Model Interpretability

To examine each model more closely, we used a variety of model interpretation techniques.

5.7.1 SHAP Explainers

SHAP (SHapley Additive exPlanations) is a powerful tool for interpreting machine learning models by quantifying the impact of each feature on a model's predictions. Each feature in the dataset is assigned a SHAP value for each prediction. A positive SHAP value indicates that the feature increases the probability of the predicted class, while a negative value means it decreases the probability. The summary plot displays the distribution of SHAP values for all features. This can be used to identify key drivers of prescribing behavior and improve model transparency. Below are the SHAP summary plots for each of the three models:

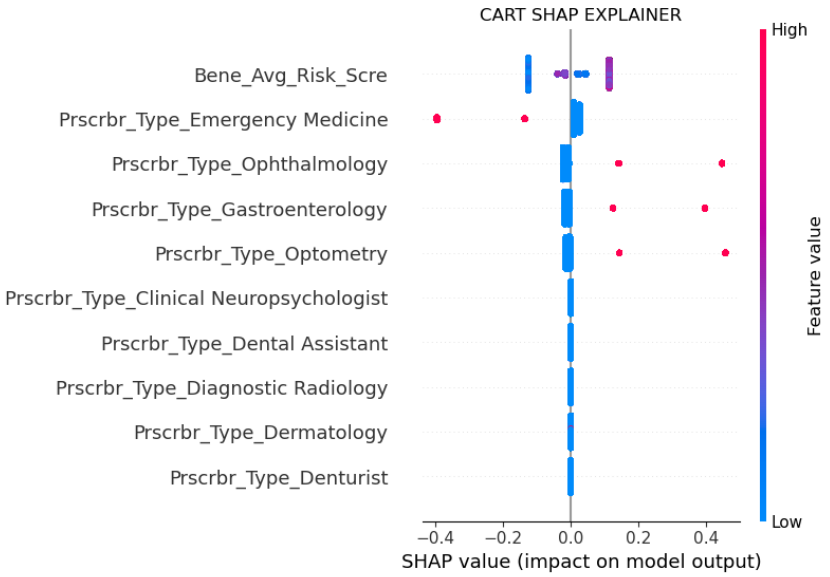

Figure 9: Decision Tree SHAP.

Based on the SHAP models, we can see that providers with beneficiaries that have higher risk scores tend to prescribe higher rates of brand-name drugs. Provider types are also highly influential in the models' decision making, with prescribers like ophthalmologists and gastroenterologists tending to prefer brand-name drugs, while prescribers in emergency medicine are more likely to prefer generics. Total payments from pharmaceutical companies are also highly influential in some of the models, including Random Forest and XGBoost, with higher payment amounts associated with higher prescribing rates for brand-name drugs.

5.7.2 Feature Importances

In order to understand each model's decision making process better, we also plotted the top 10 most important features. Contrary to the SHAP model, we cannot directly see which way the feature is influencing the decisions or how the influencing differs between each class. However, this can still provide key insights into key factors influencing prescribing behavior.

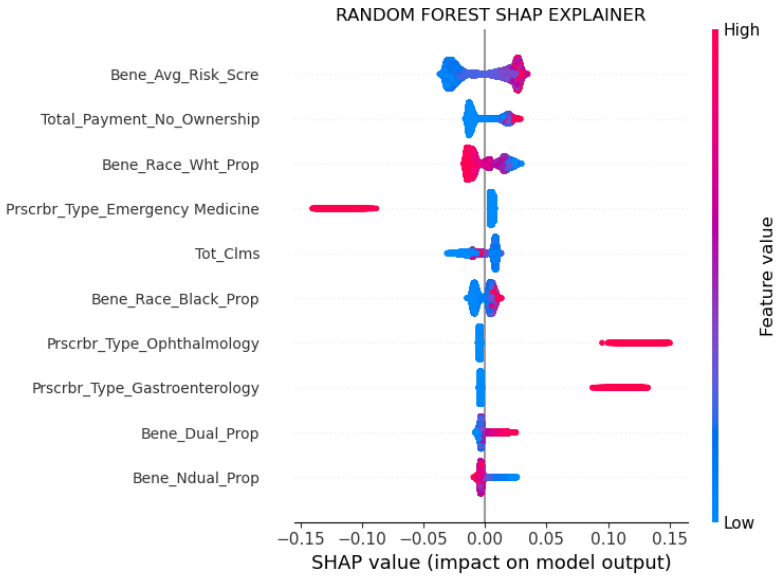

Figure 10: Random Forest SHAP.

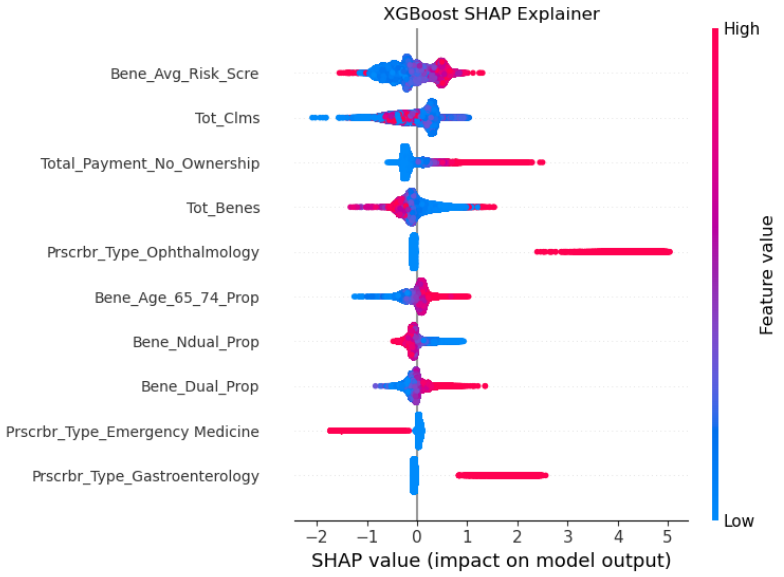

Figure 11: XGBoost SHAP.

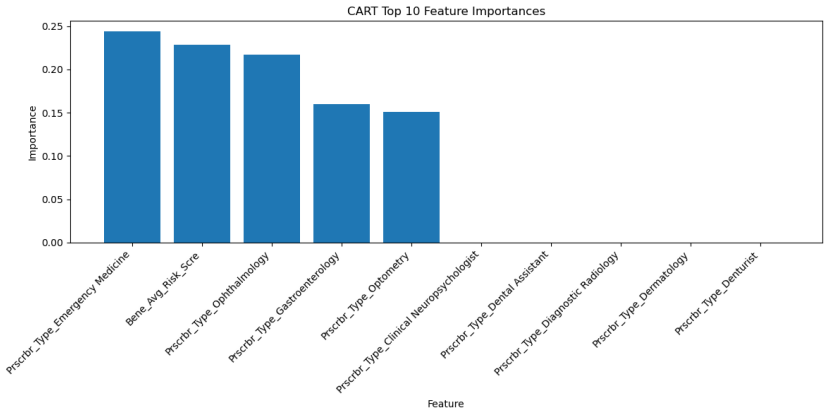

Figure 12: Decision Tree Feature Importances.

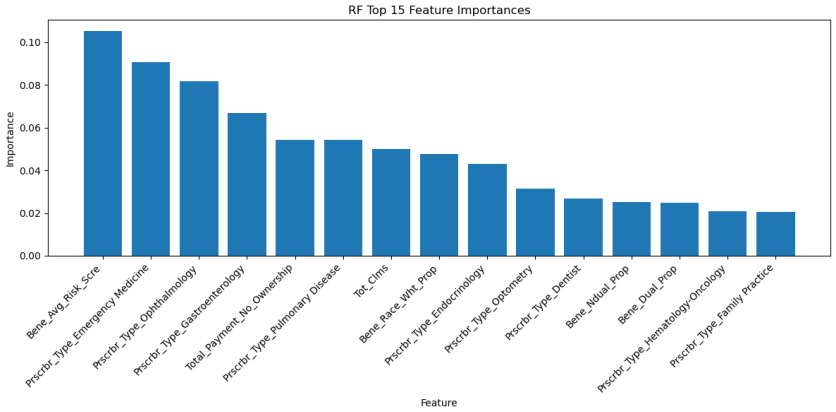

Figure 13: Random Forest Feature Importances.

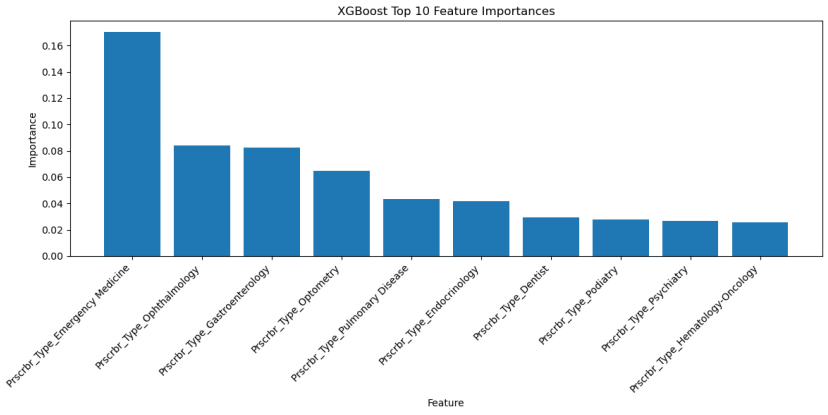

Figure 14: XGBoost Feature Importances.

There are a few consistent factors across models, including various prescriber types like emergency medicine, ophthalmologists, and gastroenterologists. Both the CART and Random Forest have more variety in the factors included in the model compared to XGBoost, which focuses more heavily on prescriber types. For Random Forest and CART, other key features include the average risk score of beneficiaries and payments from pharmaceutical companies.

5.7.3 Tree Diagram (Decision Tree)

For the most interpretable and least complex model, the CART Decision Tree, we can also visualize the tree and all of its decision nodes. The decision nodes are limited, with the first split according to the average beneficiary risk score. Then the decisions are made according the type of provider. The tree is shown below.

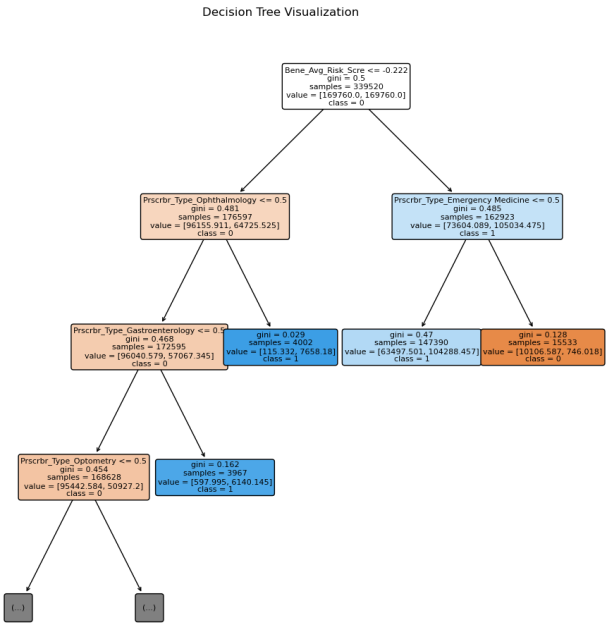

Figure 15: Decision Tree.

6 Conclusion and Analysis

6.1 Results and Recommendations

This study successfully implemented three machine learning models—Decision Tree, Random Forest, and XGBoost—to predict Medicare Part D prescribers with a preference for brand-name drugs. The XGBoost model emerged as the top performer, particularly excelling in recall (0.78) and F1 score (0.66). This aligns with the research goal of minimizing false negatives (missing prescribers that prefer brand-name drugs). While accuracy decreased slightly after tuning (from 0.84 to 0.78), this trade-off was justified to better capture the minority class of interest.

The research question was namely interested in discovering key predictive factors. The results suggest that high beneficiary risk scores, certain provider types, and pharmaceutical company payments were consistent predictors of brand-name prescribing across all models. SHAP analysis and feature importance rankings confirmed the influence of these factors, providing robust interpretability even for complex models like XGBoost. The SHAP models allow additional insight into which direction the factors are influencing the model.

The key substantive outcomes below can be used to inform policy intervention.

1. **High-risk patients are more likely to receive brand-name drugs.** Providers with patients who had higher average risk scores were more likely to prescribe brand-name drugs. This aligns with clinical intuition—sicker patients may require drugs that do not have effective generic alternatives. However, the model also suggests that this variable remains a strong predictor even when controlling for other provider and patient features. Policy interventions could include:

   - **Peer benchmarking:** Establish risk-adjusted prescribing benchmarks by specialty to identify outliers who prescribe brand-name drugs disproportionately, even after accounting for patient risk scores [CMS, 2024].
   - **Guideline audits:** Encourage Medicare to audit providers whose brand-name prescribing exceeds norms within high-risk populations, especially if generic equivalents are available [Shi, 2024].

   Further research could be done to analyze the causal mechanism behind providers with high-risk patients prescribing more brand-name drugs, given that there are many factors going into prescribing patterns. This analysis could also be done at the drug level to see whether high-risk patients are receiving newer, patented drugs or simply branded versions of older medications.

2. **Certain specialties have intrinsic prescribing tendencies.** Certain specialties, particularly ophthalmology and gastroenterology, were significantly associated with higher rates of brand-name prescribing. In contrast, specialties like emergency medicine were more generic-heavy. This may reflect differences in drug availability or prescribing culture across specialties. Some specialties may be more targeted by pharmaceutical marketing due to frequent use of expensive, branded treatments. Policy interventions could include:

   - **Specialty-specific guidelines:** Develop or update prescribing guidelines with an emphasis on cost-effectiveness for specialties identified as high brand-name prescribers.

- **Targeted CME:** Focus educational efforts on prescribing behavior and generic efficacy in specialties with higher brand-name rates. [Lublóy, 2014]
- **Incentive alignment:** Modify Medicare reimbursement models to reward generic prescribing more explicitly within certain specialties. [Commission et al., 2010]

This project could be altered to explore patterns at the drug level to see which drugs are most prescribed by different provider specialty types.

3. **Industry payments correlate with non-generic preferences.** Higher total payments from pharmaceutical companies were positively associated with higher brand-name prescribing. This relationship is intuitive, reinforcing concerns about industry influence on prescribing behavior. The fact that this dataset is publicly available reveals an interest by policymakers in openness and transparency into the relationships between providers and pharmaceutical companies. Policy interventions could include:

   - **Transparency thresholds:** Introduce stricter disclosure or cap requirements for providers receiving industry payments above a certain threshold.
   - **Conflict-of-interest monitoring:** Use similar models to this one within CMS to tag prescribers for audit based on high payments and high brand-name prescribing.
   - **Informed decision tools:** Equip patients with data (like the already existing CMS Open Payments website) that show whether their providers receive industry payments.

   Further research could investigate whether specific types of payments are more predictive of brand-name prescribing. Other studies could use causal inference techniques like difference-in-difference research designs to study prescribing behavior before and after providers begin receiving industry payments.

6.2 Evaluation

The models performed reasonably well, especially given the challenges of class imbalance and the complexity of the research question. The move from simple to more sophisticated models showed a clear progression in capturing the intricacies of prescribing behavior. The results were expected, with known drivers from previous research studies having significant predictive influence. XGBoost's ability to balance performance and interpretability makes it a suitable choice for this type of machine learning task, so it was expected to perform very well compared to the other selected models.

There are a few reservations regarding the model performance. The class imbalance (created by splitting at the 75th percentile) constrained the ability to optimize performance, and improvements post-tuning were moderate. Further experimentation with SMOTE or alternate evaluation metrics could yield gains. Other hyperparameters, especially within the gradient boosting library of algorithms, could also be effective in improving the model performance for the minority class. It may be useful in future projects or in further development of these models to explore feature engineering as well. Investigating additional features from other datasets or including interaction terms may help to improve performance or provide more insight into the research question.

The analysis adequately addressed the central question: What factors predict whether a Medicare Part D provider will prescribe a higher proportion of non-generic drugs compared

to other providers? The models consistently highlighted interpretable and policy-relevant variables—like specialty and industry payments—as significant. This work adds empirical rigor to discussions around prescribing incentives and could inform targeted policies aiming to reduce excessive brand-name prescribing, especially among high-risk specialties.

6.3 Reflections

The technical demands—merging datasets, handling imbalanced classes, tuning models, and interpreting SHAP plots—were steep but aligned well with the project's goals. There is a strong grasp of both predictive modeling and policy relevance, with acknowledgement of the project's limitations.

There are various future directions that this project could take. Methodologically, techniques mentioned above like oversampling, ensemble stacking, or other models can help improve predictive accuracy. Substantively, this analysis could be extended to include state-level variation, temporal trends, or integration of patient outcomes to evaluate impacts. The analysis could also be done at the drug level, looking at specific brand-name drugs that have high levels of Medicare Part D claims.

Various lessons were learned attempting to use complex machine learning models to tackle an intricate policy research question. First, this project underscores the importance of interpretability in complex models. The key insights from this project came from using model interpretability analyses like SHAP and feature importances to understand how the model is making decisions. Second, the challenges with handling imbalanced data. Deciding how to dichotomize the target variable was the first step in the project and heavily influences the results and application. Finally, clean preprocessing and thoughtful model selection significantly impact downstream performance and conclusions.

References

[Chen and Wu, 2008] Chen, A. Y. and Wu, S. (2008). Dispensing pattern of generic and brand-name drugs in children. *Ambulatory Pediatrics*, 8(3):189–194.

[CMS, 2024] CMS (2024). Benchmarking. `https://www.cms.gov/priorities/innovation/key-concepts/benchmarking`. Accessed: 2025-05-05.

[Commission et al., 2010] Commission, M. P. A. et al. (2010). Aligning incentives in medicare. In *Report to the Congress. Washington, DC: MedPAC. June*, volume 15.

[Cubanski, 2024] Cubanski, J. (2024). A current snapshot of the medicare part d prescription drug benefit. `https://www.kff.org/medicare/issue-brief/a-current-snapshot-of-the-medicare-part-d-prescription-drug-benefit/`. Accessed: 2025-02-01.

[Cubanski and Damico, 2024] Cubanski, J. and Damico, A. (2024). Key facts about medicare part d enrollment, premiums, and cost sharing in 2024. `https://www.kff.org/medicare/issue-brief/key-facts-about-medicare-part-d-enrollment-premiums-and-cost-sharing-in-2024/`. Accessed: 2025-02-01.

[Cubanski and Neuman, 2023] Cubanski, J. and Neuman, T. (2023). A small number of drugs account for a large share of medicare part d spending. https://www.kff.org/medicare/issue-brief/a-small-number-of-drugs-account-for-a-large-share-of-medicare-part-d-spending/. Accessed: 2025-02-01.

[Dusetzina et al., 2020] Dusetzina, S. B., Cubanski, J., Nshuti, L., True, S., Hoadley, J., Roberts, D., and Neuman, T. (2020). Medicare part d plans rarely cover brand-name drugs when generics are available: Study examines medicare part d coverage to quantify how often brand-name drugs are designated preferred over generics. *Health Affairs*, 39(8):1326–1333.

[Gumas et al., 2024] Gumas, E. D., Huffman, P., Papanicolas, I., and Williams II, R. D. (2024). How prices for the first 10 drugs up for u.s. medicare price negotiations compare internationally. https://www.commonwealthfund.org/publications/2024/jan/how-prices-first-10-drugs-medicare-negotiations-compare-internationally. Accessed: 2025-02-01.

[Hayford and Austin, 2022] Hayford, T. and Austin, D. (2022). Prescription drugs: Spending, use, and prices. Technical report, Congressional Budget Office.

[Howard et al., 2018] Howard, J. N., Harris, I., Frank, G., Kiptanui, Z., Qian, J., and Hansen, R. (2018). Influencers of generic drug utilization: a systematic review. *Research in Social and Administrative Pharmacy*, 14(7):619–627.

[Kohli and Buller, 2013] Kohli, E. and Buller, A. (2013). Factors influencing consumer purchasing patterns of generic versus brand name over-the-counter drugs. *South Med J*, 106(2):155–60.

[Leonard et al., 2023] Leonard, F., Jacobson, G., Collins, S. R., Shah, A., and Haynes, L. A. (2023). Medicare's affordability problem: A look at the cost burdens faced by older enrollees. https://www.commonwealthfund.org/publications/issue-briefs/2023/sep/medicare-affordability-problem-cost-burdens-biennial. Accessed: 2025-02-01.

[Lublóy, 2014] Lublóy, Á. (2014). Factors affecting the uptake of new medicines: a systematic literature review. *BMC health services research*, 14:1–25.

[Mulcahy et al., 2024] Mulcahy, A. W., Schwam, D., and Lovejoy, S. L. (2024). International prescription drug price comparisons: Estimates using 2022 data. Technical report, Assistant Secretary for Planning and Evaluation (ASPE), U.S. Department of Health & Human Services.

[Rome et al., 2022] Rome, B. N., Sarpatwari, A., and Kesselheim, A. S. (2022). State laws and generic substitution in the year after new generic competition. *Value in health*, 25(10):1736–1742.

[Sharma et al., 2018] Sharma, M., Vadhariya, A., Johnson, M. L., Marcum, Z. A., and Holmes, H. M. (2018). Association between industry payments and prescribing costly medications: an observational study using open payments and medicare part d data. *BMC health services research*, 18:1–8.

[Shi, 2024] Shi, M. (2024). Monitoring for waste: Evidence from medicare audits. *The quarterly journal of economics*, 139(2):993–1049.

[Socal et al., 2021] Socal, M. P., Bai, G., and Anderson, G. F. (2021). Factors associated with prescriptions for branded medications in the medicare part d program. *JAMA Network Open*, 4(3):e210483–e210483.

[Stern, 1994] Stern, P. (1994). *Patterns of pharmaceutical prescribing*. PhD thesis, University of London: London Business School.

[Tuncay et al., 2020] Tuncay, B., Pagano, S., De Santis, M., and Cavallo, P. (2020). Prescribing behavior of general practitioners for generic drugs. *International Journal of Environmental Research and Public Health*, 17(16):5919.