Predicting Physician Specialty Using Medicare Prescription Drug Data

Francesca Abulencia

Brown University

1 Introduction

The Centers for Medicare Medicaid Services (CMS) is a US agency that administers Medicare - a national health insurance program for Americans aged 65 and older and for select people under 65 with certain disabilities. Medicare is divided into 4 parts that cover different healthcare services. Part A covers inpatient hospitalizations, Part B covers outpatient visits (doctor's appointments, urgent care, medical equipment), Part C (also known as Medicare Advantage) covers aspects of parts A, B, and D but is offered through a private insurance company, and Part D offers prescription drug coverage.

In 2017, Medicare improper payments were estimated to be about $52 billion - out of the $705.9 billion total expenditures reported for that year [1,2]. Unfortunately, Medicare has often been a target by scammers in order to obtain millions of dollars from reimbursements. In September 2022, a physician pleaded guilty to providing unnecessary medical equipment and genetic tests for 2,184 Medicare beneficiaries from 2019 to 2021, billing Medicare more than $6.2 million. Thus, Medicare fraud and abuse is a serious concern, costing taxpayers millions of dollars [3]. Automated solutions to detect fraud is an active field of research.

A 2016 study conducted by Bauder et al. used a Medicare dataset that contains information on the procedure claims associated with a provider and the amount paid for the procedure to predict medical provider specialties in an effort to detect Medicare fraud [6]. Based on their methodology, if the predicted specialty matches the provider's actual specialty, then the provider is assumed to be practicing within the norm of their specialty. If the prediction and actual specialty does not align, the provider could either have unique patients or there are anomalous and potentially malicious insurance claims associated with that provider. This method can flag certain providers whose practicing habits may need to be investigated. The researchers in this study used a Naive Bayes classifier that successfully predicted 7 physician specialties with an F-score over 0.9 and 18 specialties with an F-score between 0.5 to 0.9. A similar study by Hancock et al. used Catboost and XGBoost on the same Medicare procedure claims data and their results yielded a mean AUC of 0.9080 using Catboost and a mean AUC of 0.8616 for XGBoost.

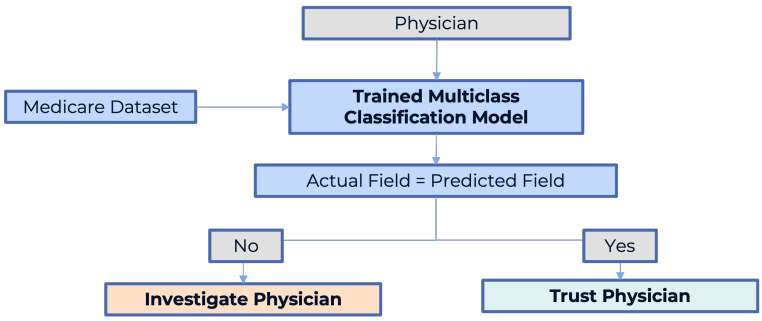

Figure 1. Workflow for Medicare Anomalous Claim Detection. Adapted from Bauder et al. 2016.

The Medicare claims dataset from the previous studies are known as the Medicare Provider Utilization and Payment Data, and these are publicly available from the CMS website [4,5]. Another data source that will be interesting to analyze is the Medicare Part D Prescribers Dataset. This data, also publicly available on the CMS website, is based on information gathered from CMS administrative claims data for Medicare beneficiaries enrolled in the Part D program available from the CMS Chronic Condition Data Warehouse. This dataset describes the prescription drugs provided to Medicare Part D beneficiaries by physicians and other health care providers. More specifically, it contains drug utilization information such as the total number of prescriptions, total drug cost, total beneficiaries associated with a brand name drug and generic name drug for all Medicare patients and for those

that are 65 years and older.

The purpose of this project is to develop a multi-class classification model that can predict physician specialties using the Medicare prescription drug data. The 2020 Medicare Part D raw dataset contains 25,209,72 observations and 22 features.

2 Exploratory Data Analysis

Each row in the dataset represents a distinct combination of NPI and drug name where multiple rows can belong to the same NPI. To simplify data analysis, specialties were included only if there are more than 5,000 providers within that specialty. Non-physician specialties were also excluded in the analysis. Second, columns with redundant information were excluded such as physician first and last name, FIPS code for states, and flags that indicate the reason for missing values. The information contained in these columns are conveyed by other columns. After applying these simplifications, the dataset was truncated to 17,972,550 observations and 16 columns. From the 16 columns, 14 are considered features. The remaining two columns are reserved for the target variable - physician specialty, and the other column specifies the group structure of the data. 6 of the columns are categorical, and the remaining 10 features are continuous. The number of unique categories in the categorical variables are shown in Table 2.

Table 1. Number of Unique Categories in Categorical Features

Variable	Name	Unique Counts
Prscrbr_Type	Specialty	26
Prscrbr_NPI	NPI	598,684
Prscrbr_City	Physician City	10,583
Prscrbr_State_Abrvtn	Physician State	61
Brnd_Name	Brand Name Drug	2,930
Gnrc_Name	Generic Drug	1693

To explore the target variable, the balance of each class was calculated and is shown in Figure 2. From this data, we see that the classes are highly imbalanced and this distribution must be taken into account when splitting the data into train, validation, and test sets.

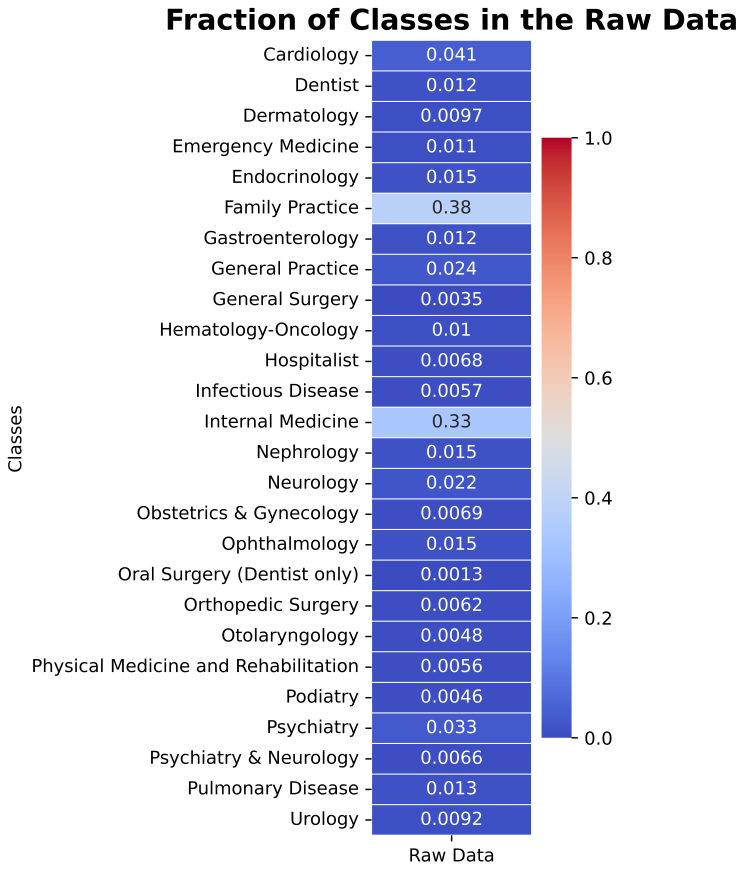

Figure 2. The 26 classes of the target variable is imbalanced in the raw dataset

The number of unique physicians within each specialty was also investigated. Figure 3 shows the 5 most and least frequent specialties determined by the number of physicians that practice each specialty. The distribution of physicians in each specialty must also be taken into account for the data splitting step.

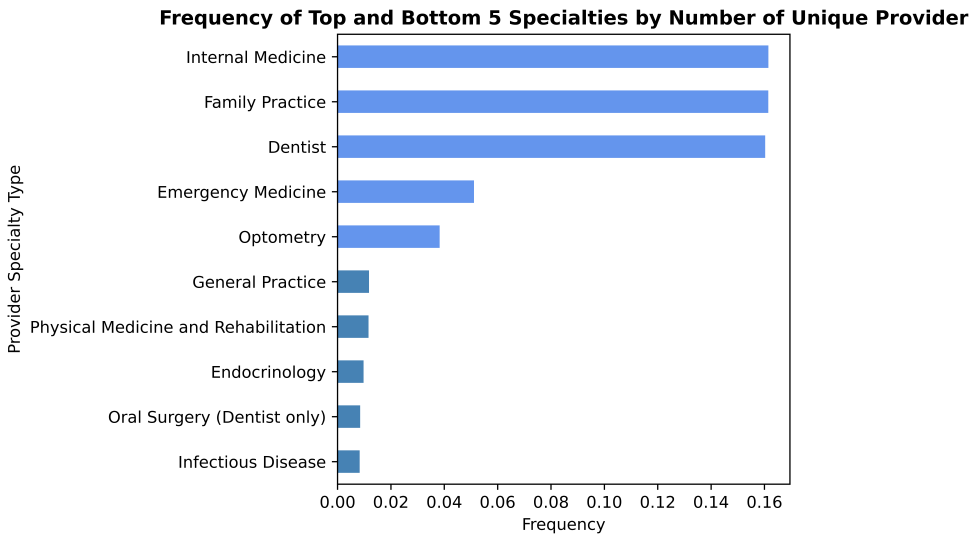

Figure 3. The number of unique NPIs associated with each specialty is not uniformly distributed in the dataset

The distribution of claims per drug was also examined by specialty. The specialties with the 3 highest and lowest median total claims is shown in Figure 4. Certain specialties, such as Cardiology, Ophthalmology, and Oral Surgery exhibit higher median number of claims associated with a single drug prescribed in contrast to specialties like Emergency Medicine, Obstetrics Gynecology, and Hospitalists.

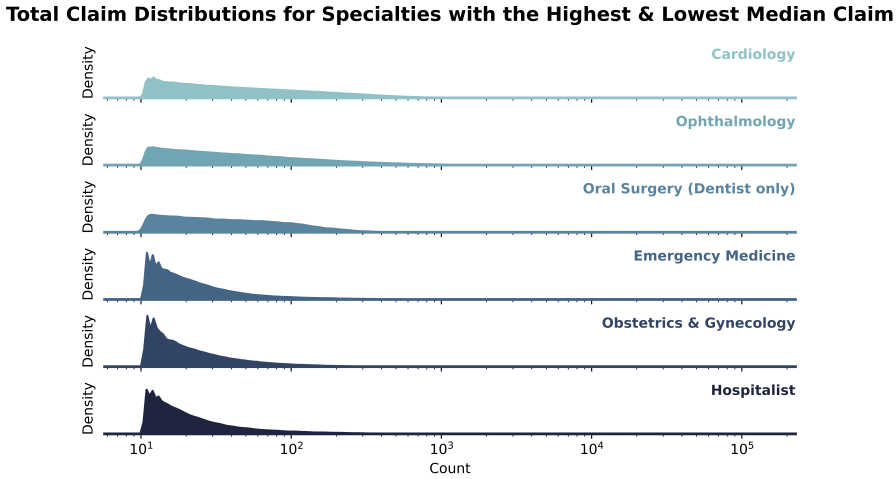

Figure 4. The total claims per drug vary by specialty. Cardiology, Ophthalmology, and Oral Surgery are the specialties with the highest median total claim per drug.

The relationship between total drug cost and total number of claims per drug for each specialty was also assessed. The specialties with the 3 highest and lowest average drug cost per claim ratio is shown in Figure 5. Specialties like Pulmonary Disease, Endocrinology, and Ophthalmology have higher total drug cost and lower claims, signaling that the drugs prescribed within that specialty are more expensive.

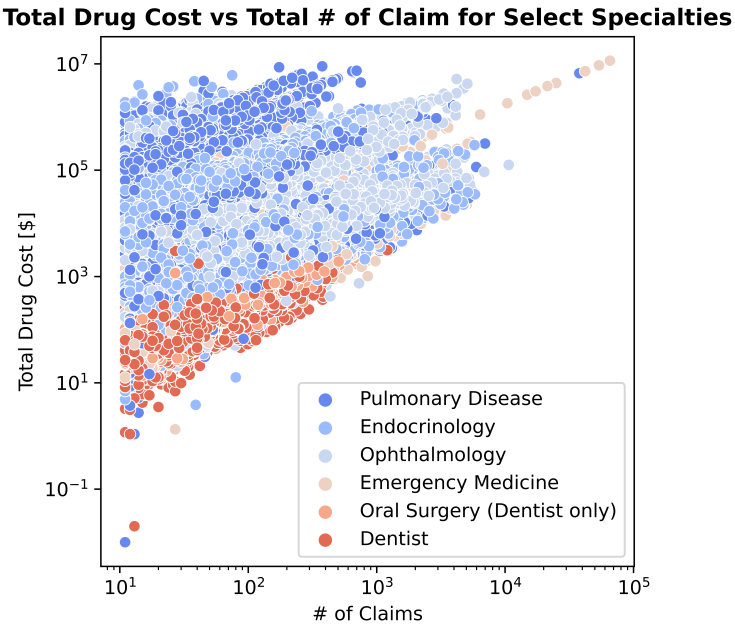

Figure 5. Median total claim per drug vary by specialty.

Missing data is present in 6 out of 14 features, with GE65_Tot_Benes, or Beneficiaries (Age 65+), lacking 89% of

data.

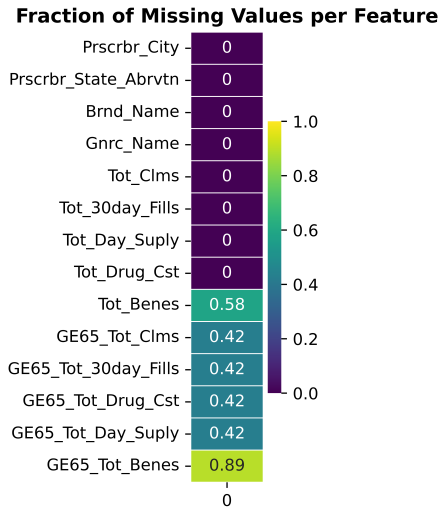

Figure 6. 6 out of 14 features contain missing numeric data.

3 Methods

Because the data has group structure and imbalanced classes, the initial strategy was to use StratifiedGroupKFold to sub-sample and split the data into the train, validation, and test sets. However, applying two rounds of StratifiedGroupKFold for train/validation/test split generates multiple folds that became too computationally expensive for model training. Thus, observations from 3 specific physician specialties were subsampled and used as a starting point for this project. The 3 specialties - Ophthalmology, Nephrology, and Endocrinology, were selected because they have equal weights in the raw dataset. The subsampled data contains 6,417 observations. This is an appropriate size that ensured there are more rows than features after performing OneHotEncoding of the categorical features in the train/validation/test set.

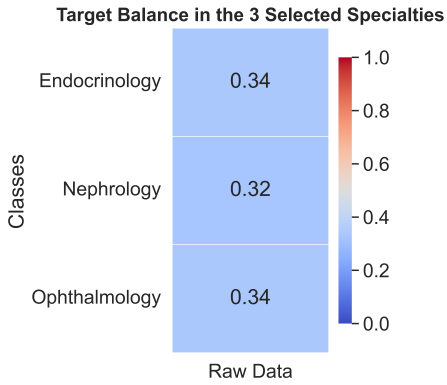

Figure 7. The 3 selected specialties are equally distributed in the raw dataset.

A 60-20-20 train/validation/test split was performed on the data. First, GroupShuffleSplit was applied to split the data into a 20% test set and 80% other set. The other set was then split using GroupKFold with 4 splits to generate a 60% train and 20% validation set. The 4-fold split is used for a 4-fold cross validation.

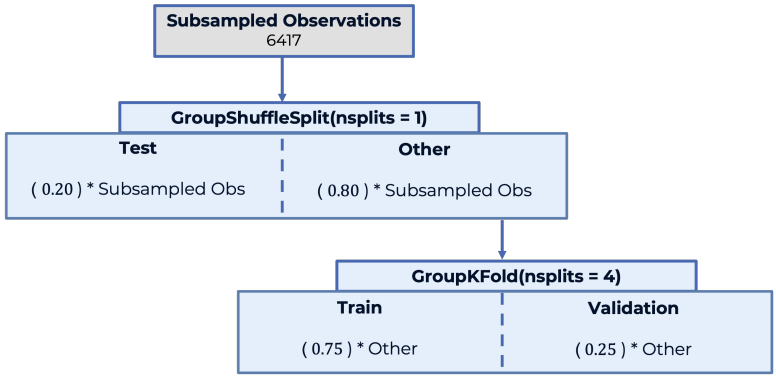

Figure 8. Schematic for a 60-20-20 train/validation/test split

For data pre-processing, a pipeline for continuous features was created with IterativeImputer to impute missing values using a Linear Regression estimator and StandardScaler. A pipeline for categorical features was developed with OneHotEncoder. For certain linear models, StandardScaler was applied after OneHotEncoding to normalize coefficients for later analysis of feature importance.

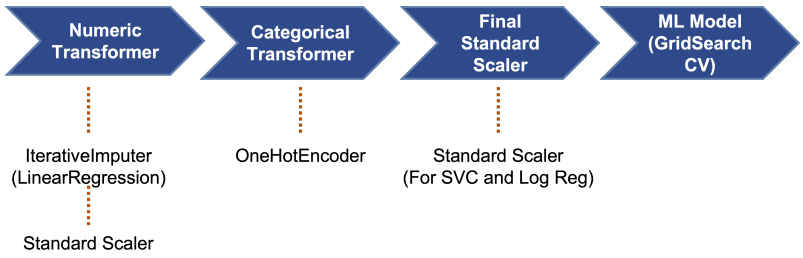

Figure 9. Using a pipeline for preprocessing categorical and continuous data

Five models (XGBoost, Random Forest, Support Vector Machine and Logistic Regression with and without L1 penalty) were implemented with GridSearchCV for hyperparameter tuning with cross-validation.

Table 2. ML Models and their Corresponding Hyperparameters

ML Model	Hyperparameter	Values
XGBoost Classifier	max_depth	3, 10, 30, 100, 300
	learning_rate	0.01, 0.1
	colsample_by_tree	0.5, 0.7, 0.9
	n_estimators	300, 500
Random Forest Classifier	max_depth	1, 3, 10, 30, 100, 300
	min_samples_split	0.01, 0.1
	n_estimators	1, 3, 10, 30, 100
Support Vector Classifier	gamma	0.01, 0.1, 1, 10, 100
	C	0.01, 0.1, 1, 10, 100
Logistic Regression	penalty	none, l1
	C	0.001, 0.01, 0.1, 1, 10, 100, 1000

Uncertainties from splitting and non-deterministic models were addressed by performing the model training and testing over 5 random states. The evaluation metric chosen is accuracy since the true positive and true negatives

are very important. The false positive and negative rates are not as important since we are not concerned which specialty a physician is misclassified as. The target classes are also balanced, making accuracy an appropriate metric.

4 Results

Based on the mean accuracy of the models on the test set, the XGBoost Classifier performed the best at 0.78 accuracy. The results are shown in Figure 10. While all models performed better than the baseline accuracy of 0.34, logistic regression without penalty performed the worst out of the 5.

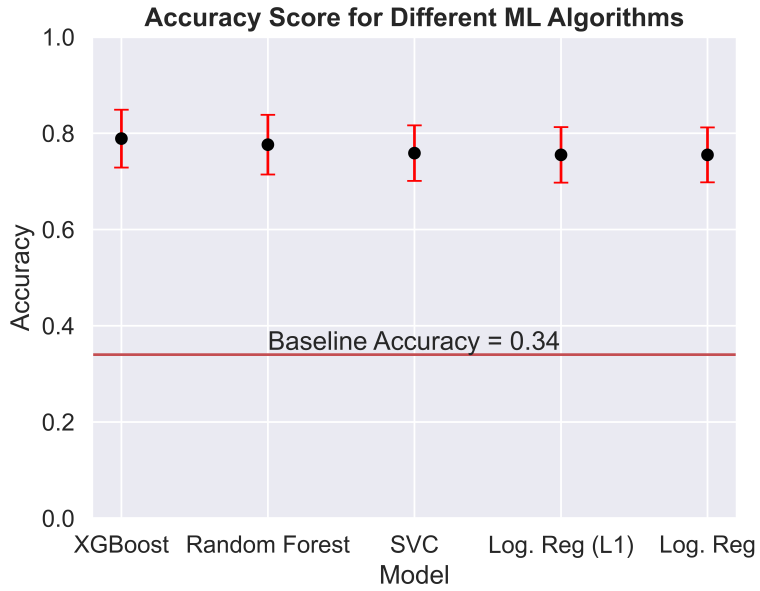

Figure 10. The XGBoost Classifier performed the best in terms of accuracy.

Looking at the normalize confusion matrices for all models on Figure 11, the Ophthalmology specialty consistently has the highest accuracy compared to Nephrology and Endocrinology. This is not surprising since Endocrinology - a specialty involving hormones, and Nephrology - a specialty involving the kidneys, can have a overlapping prescribing habits because of connected function of kidney and hormones. In contrast, Ophthalmology - a specialty concerning the eyes, can have different prescribing habits from hormone and kidney disorders.

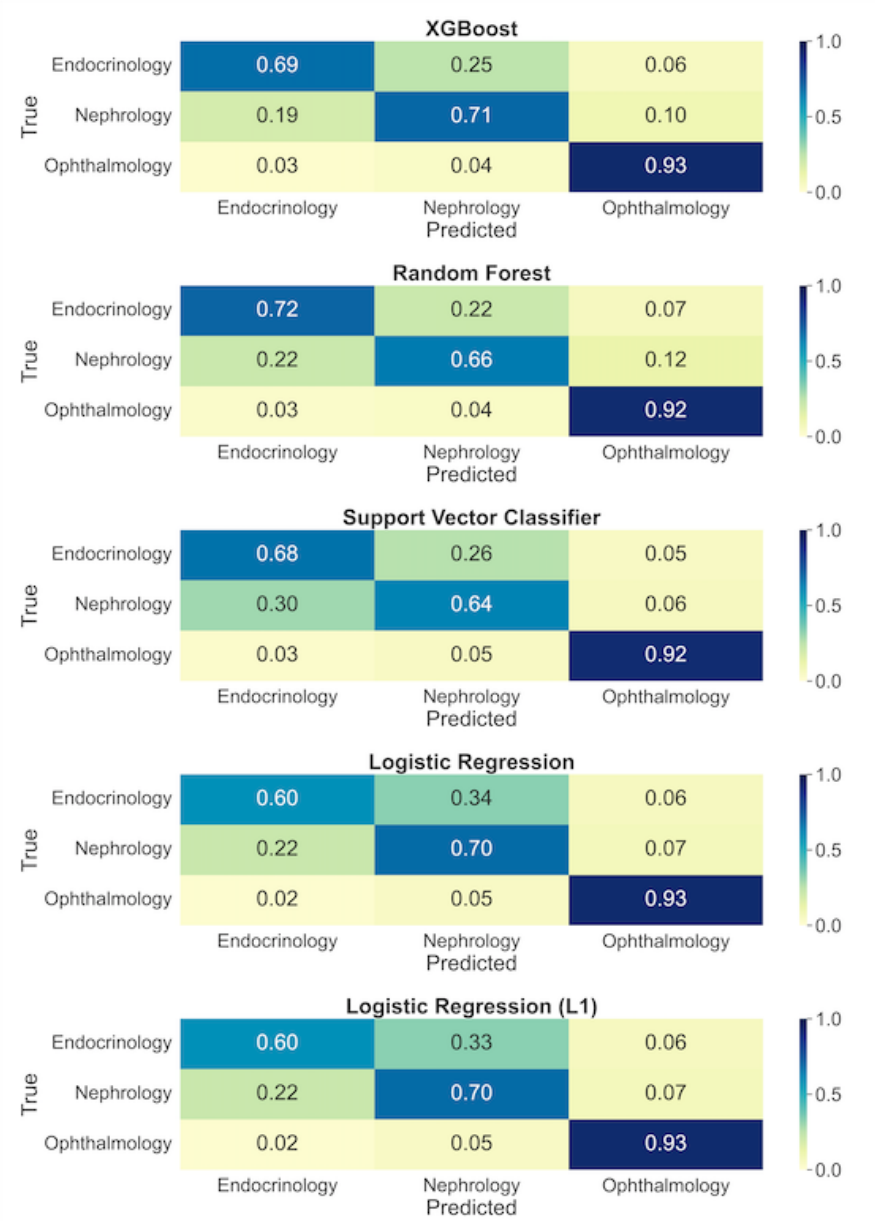

Figure 11. Normalized confusion matrices show high accuracy for identifying Ophthalmology specialty.

To study which features are most important in the best performing model, different metrics were explored. First, feature importance scores from the XGBoost model were studied. Figures 12 and 13 show the top 5 most important features according to the weight and total gain metric.

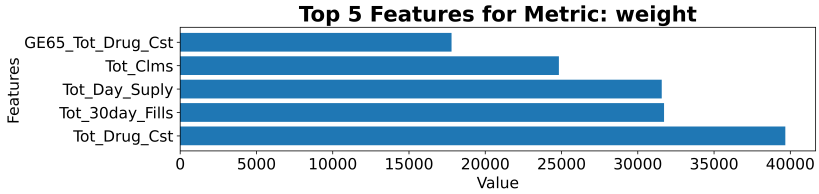

Figure 12. Total drug cost is the most important feature according to XGBoost weight metric.

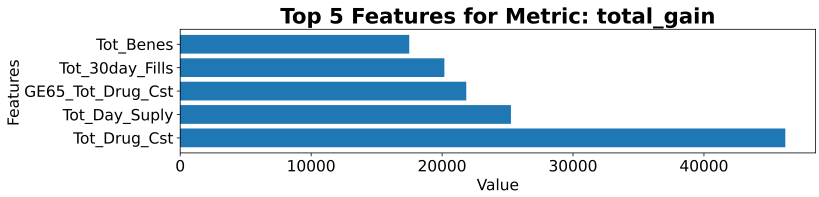

Figure 13. Total drug cost is the most important feature according to XGBoost total gain metric.

Using permutation feature importance, the difference in test score was calculated as a result of randomly shuffling a feature's value. Based on Figure 14, the model is dependent on total drug cost as it introduces the largest decrease in model score.

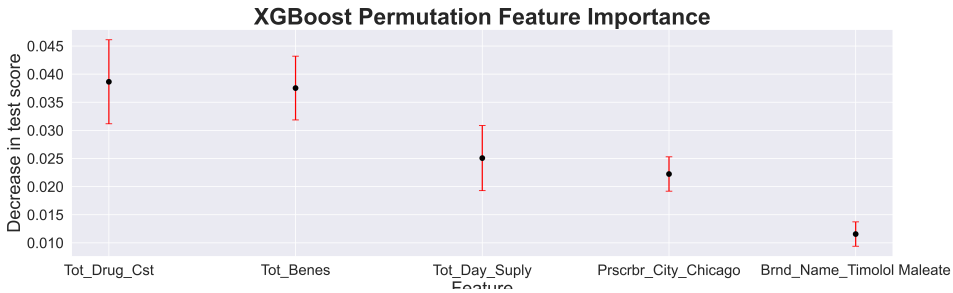

Figure 14. Total drug cost is the most important feature based on permutation feature importance.

SHAP global feature importance was also calculated in Figure 15 which reports the mean SHAP value for class 0 (Endocrinology), class 1 (Nephrology), and class 2 (Ophthalmology). While total drug cost is ranked the 2nd most important feature, it has appeared consistently in multiple metrics, confirming its critical role in the XGBoost model prediction.

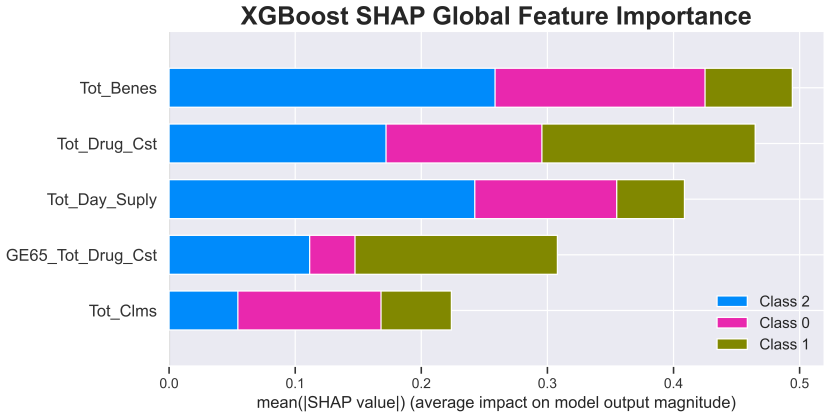

Figure 15. Total beneficiaries ranks the highest in SHAP Global Feature Importance.

Local feature importance was studied to see which features influence the prediction for a single observation. A SHAP force plot is shown in Figure 16. 3 plots for each class is generated for one observation. For this specific physician (with actual specialty of Endocrinology), the predicted probability is as follows: 0.83 for Endocrinology vs 0.17 for 'Not Endocrinology', 0.39 for Nephrology vs 0.61 for 'Not Nephrology', and 0.27 for Ophthalmology vs

0.72 for 'Not Ophthalmology. Each class is treated like a binary variable, and the high predicted probability for Endocrinology aligns with the actual specialty. The Gnrc_Name_Metformin Hcl (a type of drug) had the highest influence on increasing the prdicted probability towards Endocrinology.

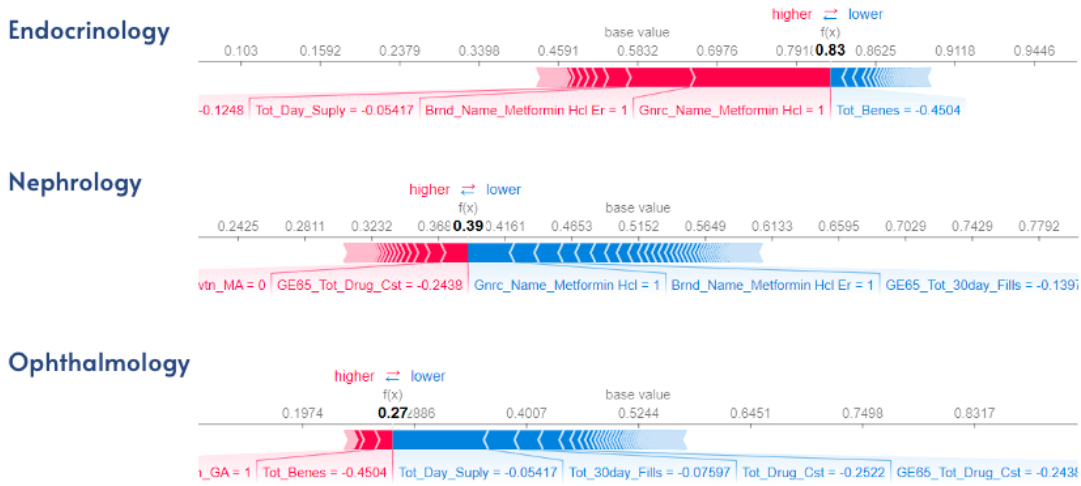

Figure 16. SHAP local feature importance for a single observation

5 Outlook

To improve this project further, strategies to increase model performance and interpretability can be undertaken like expanding the search space for hyperparameters to improve the test score. A specific weak spot in the XGBoost model lies in the tuning of n_estimators. In the future, early stopping should be implemented instead. Incorporating early stopping with GridSearchCV proved to be difficult but can be addressed by manually looping through the hyperparameters. Additionally, more models like Logistic Regression with L2 penalty and the Reduced Features Model can be tested to see if accuracy can be improved. In terms of model explainability LIME can be used to explore other important features. Lastly, more data and more physician specialties should be included in the analysis to capture a more representative sample from the raw dataset.

6 Github Repository

https://github.com/fabulenc/CMS-Medicare-Provider-Specialty-ML-Project

7 References

[1] https://www.cms.gov/newsroom/press-releases/cms-office-actuary-releases-2017-national-health-expenditures
[2] https://www.gao.gov/products/gao-18-660t
[3] https://www.justice.gov/usao-wdmo/pr/physician-pleads-guilty-making-false-statements-more-2000-medicare-medicaid-patients
[4] https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service
[5] https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider-and-drug
[6] R. A. Bauder, T. M. Khoshgoftaar, A. Richter and M. Herland, "Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims," 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), 2016, pp. 784-790
[7] J. Hancock and T. M. Khoshgoftaar, "Performance of CatBoost and XGBoost in Medicare Fraud Detection," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 572-579