CAPSTONE PROJECT 2

Predictive Modeling of Fraud Charges

Julia Sheriff

SPRINGBOARD

Data Science Career Track

1. Introduction

The purpose of this capstone project is to use predictive modeling to determine fraudulent Medicare providers. After acquiring the data and conducting exploratory data analysis, I proceeded to use logistic regression, random forest regression, and extreme gradient boosting (XGB) trees classification to find the top-performing models. An XGB Tree model utilizing random under-sampling performed the best, with 86% overall accuracy and 87% recall for the target class: fraudulent providers. Alternative methods of collapsing claims data to the provider level could potentially improve model accuracy by preserving more nuances in the data that could indicate fraudulent providers.

2. Approach

2.1 Data Acquisition and Cleaning

The data is from a Kaggle competition, "Healthcare Provider Fraud Detection Analysis", provided by Rohit Anand Gupta at "https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis". Four healthcare datasets were used to model fraud, which described Medicare beneficiaries and their claims. The competition provided a testing and training set for each of these four datasets. Models were developed and tested using the training set. The independent variables provided were on the beneficiary level and claims level, but in order to predict provider fraud, I needed provider-level independent variables. After initial data wrangling and inferential statistics, I created appropriate independent variables on the provider level. The original datasets were clean, but the variables needed to be transformed to build models. Creating other independent variables for modeling could generate alternative, and perhaps superior models.

The beneficiary dataset described patient demographics, health, and health insurance (Table 1). I used string parsing to remove letters from beneficiary id. I also created binary variables describing chronic health conditions, gender, and death.

Table 1: Beneficiary Data Variables

Demographic Variables	Health Variables	Insurance Variables
'DOB', 'DOD', 'Gender', 'Race'	'RenalDiseaseIndicator' 'ChronicCond_Alzheimer' 'ChronicCond_Heartfailure', 'ChronicCond_KidneyDisease', ' ChronicCond_Cancer', ‘ChronicCond_ObstrPulmonary', 'ChronicCond_Depression', 'ChronicCond_Diabetes', 'ChronicCond_IschemicHeart', ' ChronicCond_Osteoporasis', 'ChronicCond_rheumatoidarthritis', 'ChronicCond_stroke'	'BeneID', 'NoOfMonths_PartACov', ' NoOfMonths_PartBCov', 'IPAnnualReimbursementAmt', 'IPAnnualDeductibleAmt', 'OPAnnualReimbursementAmt', 'OPAnnualDeductibleAmt'

Two datasets were provided on the claim level: inpatient claims and outpatient claims. International disease classification (ICD) codes were used to describe medical diagnoses and procedures. Both datasets described claim reimbursements, deductibles, ICD diagnosis codes, ICD procedure codes, and physicians. The inpatient dataset also described admissions, discharges, and diagnostic groups. I used string parsing to extract numbers from claim id, beneficiary id, and physician id. In order to indicate which dataset described inpatient claims, and which described outpatient claims, I made a binary variable: “in_out”. Claim start date, claim end date, admission date, and discharge date were transformed to datetime variables. A new variable, “duration” was calculated from the difference in admission and discharge dates. Because multiple diagnosis and procedure codes aren’t always necessary to describe a claim, there was no claim information for outpatient procedure code 5, outpatient procedure code 6, and inpatient procedure code 6. I dropped procedure codes 4, 5, and 6 because of the lack of data. A higher numeric label for a procedure or diagnostic code (ex: procedure code 6) indicates a more ancillary health procedure or diagnosis than a lower numeric label (ex: procedure code 1).

Table 3:Inpatient and Outpatient Claim Variables

Claim Information	Charges and Payments	Diagnosis Codes	Procedure Codes	Physicians

BeneID, ClaimID, ClaimStartDt, ClaimEndDt, Provider, DischargeDt*, AdmissionDt,* DiagnosisGroupCode*	InscClaimAmt Reimbursed, DeductibleAmtPaid ClmAdmitDiagnosis Code, ClmDiagnosisGroupCode	Claim Diagnosis Codes 1 – 10	Claim Procedure Codes 1-6	AttendingPhysician OperatingPhysician OtherPhysician

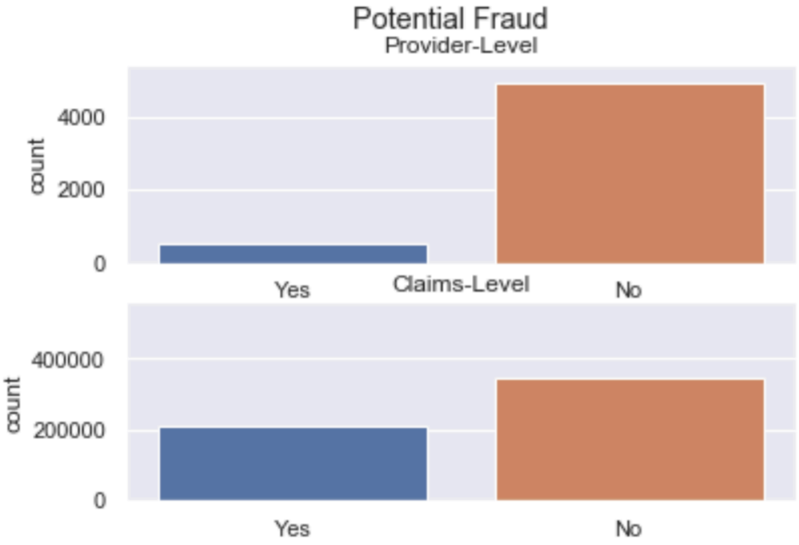
- inpatient variable only

I appended inpatient and outpatient claims, and then merged the set with beneficiary claims, using beneficiary as the key. The fourth dataset (dimensions: 5410, 2) classified providers as potentially fraudulent, or not potentially fraudulent. I merged this fraud dataset with the combined dataset, using provider as the key. Before generating independent variables to predict potential fraud, I explored the data to develop an approach.

2.2 Storytelling and Inferential Statistics

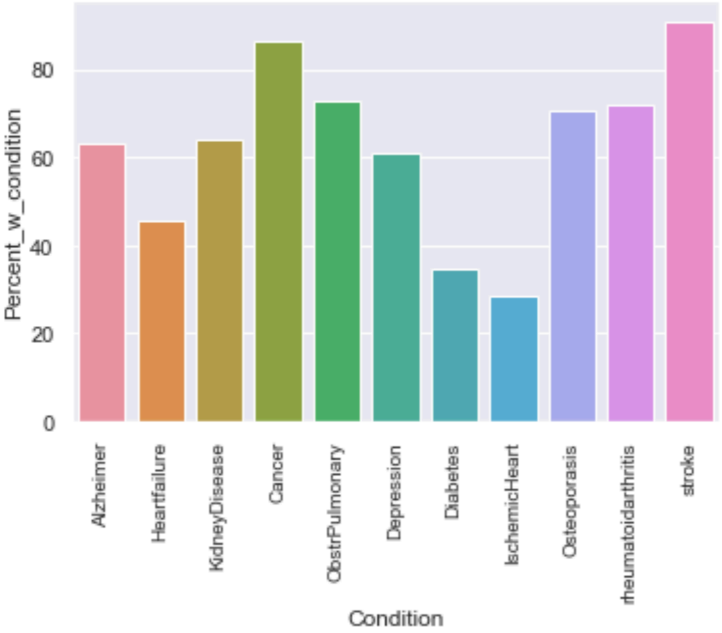
When comparing data on the claims versus provider level, the percent of fraudulent claims (9.35%) was far lower on the provider level than on the claims level (38.12%) (Figure 1). This suggests that most fraudulent claims come from a few providers who practice fraud.

Figure 1: Potential Fraud on the Provider and Claims Level



Regarding the beneficiary population, the median beneficiary age was 74, and most patients had 12 months of Part A and Part B Medicare Coverage. The most common chronic health conditions were stroke (91%), cancer (86%), and pulmonary obstruction (73%) (Figure 2). The mean inpatient stay was four days.

Figure 2: Chronic Health Conditions in Patient Population



In general, I explored data with violin and boxplots to check the normalcy of values for variables such as health codes, identification numbers, charges, and reimbursements. In order to do this for ICD codes, I used string parsing to remove letters from odes to check if the numerical values were typical. While I did discover some extreme values, I wanted to preserve them in case they could help indicate fraud, since fraud is irregular in nature. I did not find any instances which seemed so extreme as to indicate a flaw in the data.

2.3 Generating Modeling Variables

Most of the provider level variables I generated for modeling were measures of central tendency of the full dataset's variables, as described at the end of section 2.1. I calculated the mean reimbursements, deductibles, coverage, and chronic health conditions on the claims level, per provider. Aside from chronic health conditions, I also generated median values from those same variables. For race, state, and county, I calculated the percentage of claims in each individual category (ex: race = 5, state = 40) per provider. I also calculated the diversity in the number of physicians for each provider, as well as the number of potentially duplicate claims for the same beneficiary and physician. Because my data was becoming quite wide, I also wrote alternatives to the variables above in the event that I ran into difficulties with modeling due to my data dimensions. Regarding reimbursement, I calculated the maximum average reimbursement per attending physician per primary diagnosis code. For inpatient claims, I used "diagnosis group code" and for outpatient claims, I used "diagnosis code 1" as the most indicative diagnosis. I also compared the number of diagnosis and procedure codes assigned per claim over each provider. The final dimensions of the data were (5,410 by 16,888). Due to the specificity of the columns in the model dataset, it was difficult to get a general sense of the set through further inferential statistics.

2.4 Baseline Modeling

I used logistic regressors, random forest classifiers, and extreme gradient boosted decision trees (XGB Trees) classifiers to predict potentially fraudulent providers. Some fundamental measures that describe classification model performance include precision, recall, and overall accuracy. When considering which metrics were most important in identifying potentially fraudulent providers, I chose recall of the potentially fraudulent class and overall accuracy. This implies that the type 1 error of identifying a non-fraudulent provider as fraudulent is better than the type 2 error of not identifying a fraudulent provider as fraudulent. Class one recall describes the percent of potentially fraudulent providers who are classified as potentially fraudulent in the model. Baseline models had low class one recall scores and accuracy scores above 85% (Table 4).

Table 4: Baseline Classification Models, Test Results

Type	Class 1 Recall	Class 0 Recall	Accuracy Score
Logistic regression	.35	.93	.878
Random Forest	.32	.99	.928

Classifier			
XGB Trees	.45	.98	.92

The best model above was the baseline XGB Trees classifier because the class 1 recall was significantly higher than the other two models, with similar overall accuracy. One concern was that the logistic regression model had 100% overall accuracy in predicting fraud with the training data, and 87.8% accuracy in predicting fraud with the test data. This difference could indicate over-fitting, or idiosyncrasies in the data. For the purposes of this analysis, this issue was treated as an over-fitting problem, which was addressed with cross-fold validation.

Because 9.31% of the providers were potentially fraudulent in the dataset, the low class 1 recall can be attributed to this being an imbalanced classification problem. The algorithms use more than 9 times more data in class 0 than in class 1 to develop their models, so they are more sensitive to how features impact non-fraudulent providers than potentially fraudulent providers. To address the class imbalance, I tried various under-sampling and over-sampling techniques to build a model with higher class 1 recall.

2.5 Extended Modeling

2.5a Tuned Logistic Regression Models

Applying cross-validation and parameter tuning to logistic regression models improved class 1 recall significantly (+10%) from the baseline models, with a smaller decrease in accuracy (-2.3%)(Table 5). The cross validation algorithm searched over different parameters for ‘solver’ and ‘C’. While class 1 recall improved, it was still relatively low.

Table 5: Best Logistic Regression Models without Resampling

Penalty	Class 1 Recall	Accuracy Score
Lasso	.55	.897
Ridge	.55	.897

2.5b Under-sampling and Over-sampling Methods

Under-sampling improved class 1 recall with tuned logistic regression models (Table 6). Oversampling yielded worse overall results. The same top-performing tuned logistic regression models performed the best overall with random under-sampling. Random under-sampling worked better than cluster centroid and near miss sampling when considering overall model performance. Near miss under-sampling had the best class 1 recall score, but had a poor accuracy score due to a the misclassification of many non-fraudulent providers.

Table 6: Best Logistic Regression Models with Under-sampling

Penalty	Sampling Method	Class 1 Recall	Accuracy Score
Lasso Penalty	Random Under-Sampling	.75	.885
Ridge Penalty	Random Under-sampling	.75	.885
Ridge Penalty	Near Miss	.98	.554

2.5c Random Forest and XGB Trees Models

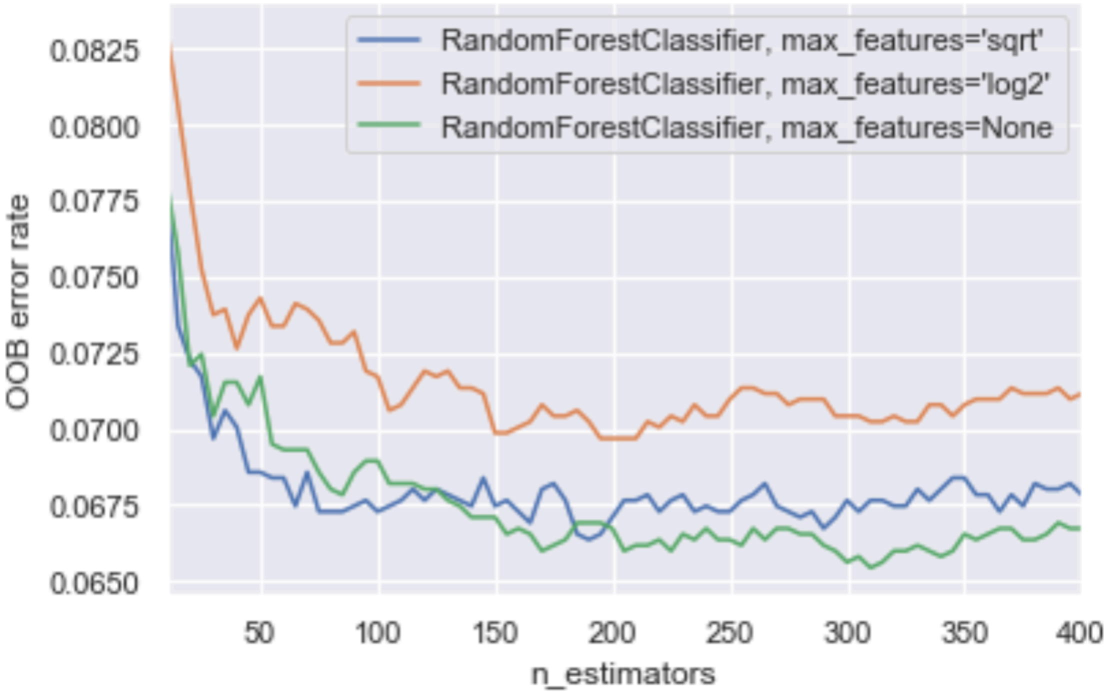
Random under-sampling also improved class 1 recall significantly in XGB Trees classifiers (+42%) and random forest models (+49%), when compared to the baseline models (Table 7).

Table 7: Best Decision Tree Models

Type	Sampling Method	Class 1 Recall	Accuracy Score
Random Forest	None	.39	.93
Random Forest	Random Under-sampling	.81	.853
XGB Trees	None	.45	.93
XGB Trees	Random Under-sampling	.87	.86

The top performing XGB Trees model had the following tuned parameters: 'max_dep th': 3, 'min_child_weight': 1. Expanding the range of the parameter grid could further improve the model. The top performing random forest model had no maximum nu mber of features, and used 600 estimators. According to the graph below (Figure 3), models with no maximum features have an out-of-bag error rate that stabilizes som ewhat at 300 estimators. With fewer than approximately 301 estimates, model accu racy will be worse. Having a higher number of estimators generally improves results , but is also more computationally expensive.

Figure 3: Number of Estimators and OOB Error in Random Forest Models



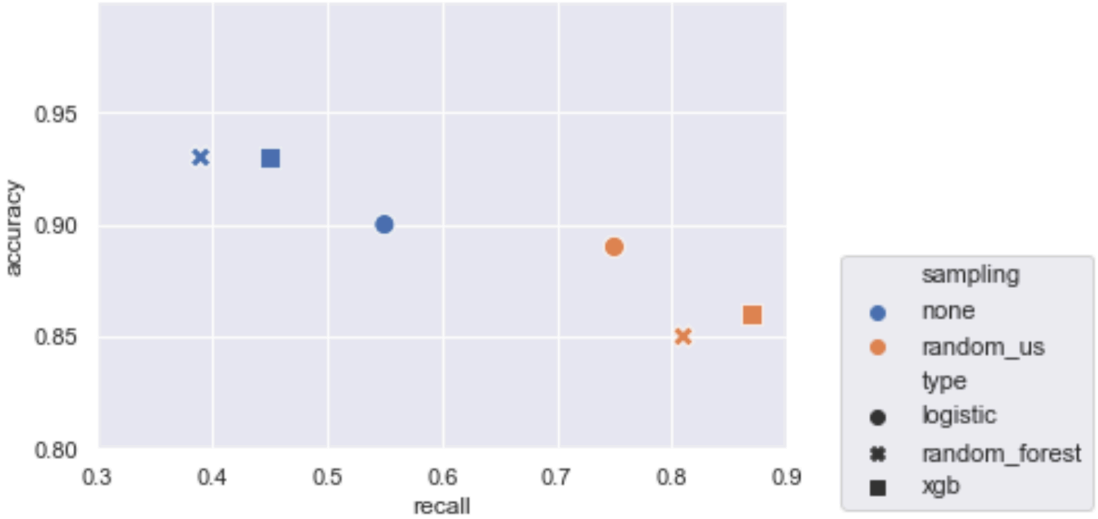
2.6 Analysis of Result

The top-performing model was an XGB Trees classifier that used random under-sampling, because it had the best overall performance with a recall of 87% and accuracy of 86%. The random forest classifier that used random under-sampling had the same accuracy, but a 4% lower score for recall. Another significant model was a logistic regressor, which also used random under-sampling. Tuned logistic models using lasso or ridge penalties both had a recall score 10% lower than the XGB classifier (75%) and accuracy score 4% higher than the XGB Trees classifier (89%). For a stronger performance in recall, the XGB Trees classifier with under-sampling is superior, but for stronger accuracy, the logistic model with random under-sampling is better (Table 8). Across the top models, there is an inverse relationship between recall and accuracy, and higher recall with the implementation of random under-sampling (Figure 4).

Table 8: Top-Performing Models using Random Under-Sampling

Type	Recall	Accuracy
Logistic	.75	.89
Random Forest	.81	.85
XGB	.85	.85

Figure 4: Impact of Under-Sampling in Parameter-Tuned Models



Mean Insurance Claim Amount Reimbursed was important in predicting potential fraud across all models. Mean duration was one of the most important features in random forest models and XGB models. Diag_41401, diag_40390, and diag_5990 had high feature importances in both the random forest and XGB models (Table 9). The most important features in classifying potentially fraudulent providers in the random forest model were counts per diagnosis group, mean duration of stay, median duration of stay, and inpatient to outpatient ratio (top five listed in Table 10). For the logistic regression model, top features were maximum mean insurance claim amount reimbursed amongst attending physicians per provider for various MS-DRGs (Table 11). The most significant feature importances for the logistic regression have importance scores that are higher than those for the tree regressors; the tree regressors have less hierarchy in feature importances than the logistic regression model.

Table 9. Top Feature Importances from Best XGB Trees Model

Importance	Feature	Description	ICD-9
.067	diag_41401	Counts of diagnosis 41401	Coronary atherosclerosis of native coronary artery
.045	diag_40390	Counts of diagnosis 40890	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified
.029	diag_5990	Counts of diagnosis 53081	Urinary tract infection, site not specified
.025	diag_4280	Counts of diagnosis 2449	Congestive heart failure, unspecified
.021	diag_4149	Counts of diagnosis 2762	Chronic ischemic heart disease, unspecified

Table 10. Top Feature Importances from Best Random Forest Model

Importance	Feature	Description
.011	diag_40390	Counts of diagnosis 40390
.010	Mean_Duration	Mean Inpatient Stay
.008	diag_53081	Counts of diagnosis 53081
.008	diag_2449	Counts of diagnosis 2449
.008	diag_2762	Counts of diagnosis 2762

Table 11: Top Feature Importances from Best Logistic Regression Model

Importance	Feature	Description	DRG
.337	in_c_854	Maximum mean insurance claim amount reimbursed amongst attending physicians per provider for various MS-DRGs	INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W CC
.230	in_c_290		ACUTE & SUBACUTE ENDOCARDITIS W/O CC/MCC
.233	in_c_225		CARDIAC DEFIB IMPLANT W CARDIAC CATH W/O AMI/HF/SHOCK W/O MCC
.228	in_c_983		EXTENSIVE O.R. PROCEDURE UNRELATED TO PRINCIPAL DIAGNOSIS W/O CC/MCC
.226	in_c_251		PERC CARDIOVASC PROC W/O CORONARY ARTERY STENT W/O MCC

3. Conclusions and Future Work

Because the results from this analysis were dependent on the training data, creating a different training data set from the original data provided could significantly impact results. Here are some alternative features:

- labels for groups of related diagnoses or procedures

- Mean Average (instead of Maximum) Insurance Claim Amount Reimbursed Amongst Attending Physicians per Provider for various diagnostic codes
- Mean Insurance Claim Amount Reimbursed Amongst Attending Physicians per Provider for various procedure codes

Because inpatient and outpatient claims are different in nature, perhaps analyses of inpatient, outpatient, and combined providers could yield more accurate results. 3,318 providers provided only outpatient services, 398 providers only provided inpatient services, and 1,694 providers provided combined services (Figure 5). 40,474 claims were inpatient claims, and 517,737 were outpatient claims (Figure 6). Having more data on inpatient claims and providers could assist in making better predictions on that subset of providers.

Figure 5: Counts of Inpatient, Outpatient, and Combined Providers

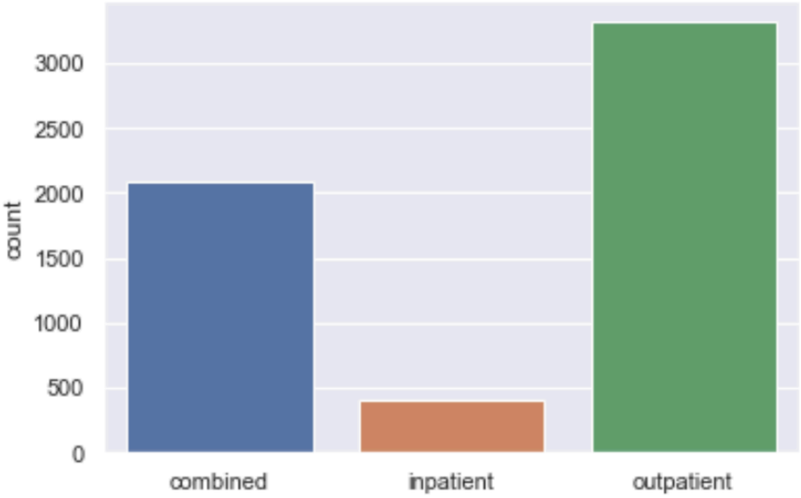
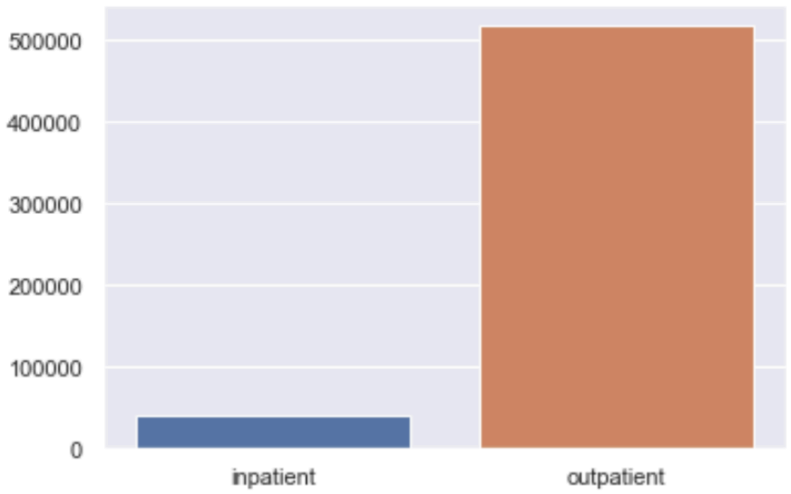


Figure 6: Counts of Inpatient, Outpatient, and Combined Claims



Additionally, any information that is specific to the provider could be useful. For example:

- Provider charges per claim
- Provider operational costs
- Provider date established
- Provider legal history
- Provider client reviews

The gap between the data provided and the question presented provided challenges in this analysis. Having alternate data or methods of bridging this gap could help improve model performance.