Health Care Provider Fraud Detection associated with Medicare

Damayanti Naik, PhD

Data Science Career Track, Capstone Project, December 2020 Cohort

Springboard

Thanks to Springboard mentor:
Yadunath Gupta

Medicare

It is the U.S Federal government program which provides hospitalization insurance and voluntary medical insurance for elderly person over 65 years.

Medicare Fraud

- **Healthcare Provider Fraud is one of the biggest problem for Medicare.**
- **Government report claims, Medicare spending is increasing exponentially due to frauds in Health Care Providers' claims.**
- **It is an organized crime by Providers, Physicians and Beneficiaries acting together to make the claims, they use ambiguous diagnosis code to adopt costliest procedures and drugs.**
- **Insurance companies are badly impacted by this, increase their premium and as a result healthcare becoming costlier day by day.**

Different forms of Fraud

- Billing for services that were not provided.
- Duplicate submission of a claim for the same service.
- Misrepresenting the service provided.
- Charging for a more complex or expensive service than actually was provided.
- Billing for a covered service when the service provided in actual was not covered.

Solution (most probable)

Rigorous analysis of Medicare data has been able to classify/detect Healthcare provider and prevent in number of frauds.

Problem Statement

The goal of this project is to build a Machine Learning (ML) model to predict/classify a Health Care Provider as Potential Fraud or not based on the claims filed by the Provider.

To obtain the best model:

- **Different ML classifier models will be built.**
- **Models will be valuated using ROC AUC score (Receiver Operating Characteristic - Area Under Curve score).**
- **Model with highest score will be chosen for deployment.**

Data: Data was collected from Kaggle. There are four datasets: Beneficiary, Inpatient, Outpatient, Train

```
1 Beneficiary.head()
```

	BeneID	DOB	Gender	Race	RenalDiseaseIndicator	State	County	NoOfMonths_PartACov	NoOfMonths_PartBCov	ChronicCond_Alzheimer	...	Chr
0	BENE11001	1943-01-01	1	1	0	39	230	12	12	1	...	
1	BENE11002	1936-09-01	2	1	0	39	280	12	12	2	...	
2	BENE11003	1936-08-01	1									
3	BENE11004	1922-07-01	1									
4	BENE11005	1935-09-01	1									

A. Beneficiary data

This dataset contains Medicare beneficiaries KYC (Know Your Customer) details; their health conditions, gender, Insurance coverage, race and region (county and state) they belong to.

```
1 Inpatient.head()
```

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician	OtherPhysician	Admiss
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922	NaN	NaN	2009-
1	BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495	PHY318495	NaN	2009-
2	BENE11001	CLM68358								
3	BENE11011	CLM38412								
4	BENE11014	CLM63689								

B. Inpatient data

This dataset provides claims details filed by the Providers for Beneficiaries those treated in hospitals. It lists Physicians, admission and discharge dates, diagnosis and procedure codes, and Claim amount reimbursed.

```
1 Outpatient.head()
```

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician	OtherPhysician	ClmDia
0	BENE11002	CLM624349	2009-10-11	2009-10-11	PRV56011	30	PHY326117	NaN	NaN	
1	BENE11003	CLM189947	2009-02-12							
2	BENE11003	CLM438021	2009-06-27							
3	BENE11004	CLM121801	2009-01-06							
4	BENE11004	CLM150998	2009-01-22							

C. Outpatient data

This dataset provides claims detail filed by the Providers for Beneficiaries who visited hospitals but not admitted there for treatment. It includes data on Physicians, Diagnosis and Procedure codes, claims amount reimbursed.

```
1 Train.head()
```

	Provider	PotentialFraud
0	PRV51001	No
1	PRV51003	Yes
2	PRV51004	No
3	PRV51005	Yes
4	PRV51007	No

D: Train data

This dataset lists Providers as Potential Fraud or not.

Data Wrangling

Problem-1: Datetime columns were listed as object.

Solution: These are converted to Timestamp.

Problem-2: Renal disease indicator column listed two values “Y” and 0.

Solution: The “Y” was assigned to 1, then column was converted to numeric.

Problem-3: In ‘DOD’ column, most of the values (97%) were missing.

Solution: The entire column was dropped from the dataframe.

Problem-4: In Physicians columns (Attending Physician, Operating Physician, Other Physician), there were many null values.

Solution: Null values were replaced with 0 (assuming that, missing values represent, physicians were absent, data not missing). Entries with a value were replaced with 1, making column categorical.

Data Wrangling (Continues…)

Problem-5: For diagnosis/procedure columns, some entries were missing.

Solution: Missing values were replaced with 0 (assuming diagnosis/procedures were really not carried out, because for different treatments diagnosis/procedures differ). Entries with a value were replaced with 1.

Problem-6: Gender column values were 1 and 2.

Solution: The entries with 2 were replaced with 0.

Problem-7: In Inpatient dataset, in Deductible amount paid column, about 2% entries were missing.

Solution: Null values were replaced with median of the column, though it could have been replaced with mean/median/mode because all entries in that column were same.

Problem-8: Potential Fraud column had two values: “Yes” , “No”.

Solution: The “Yes” and “No ” were replaced with 1 and 0 respectively.

Exploratory Data Analysis (EDA)

- No. of Potential Fraud Providers : 9.3%

  No. of Non-Potential Fraud Providers: 90.7%

- Beneficiaries from state 40 and 48 are missing.

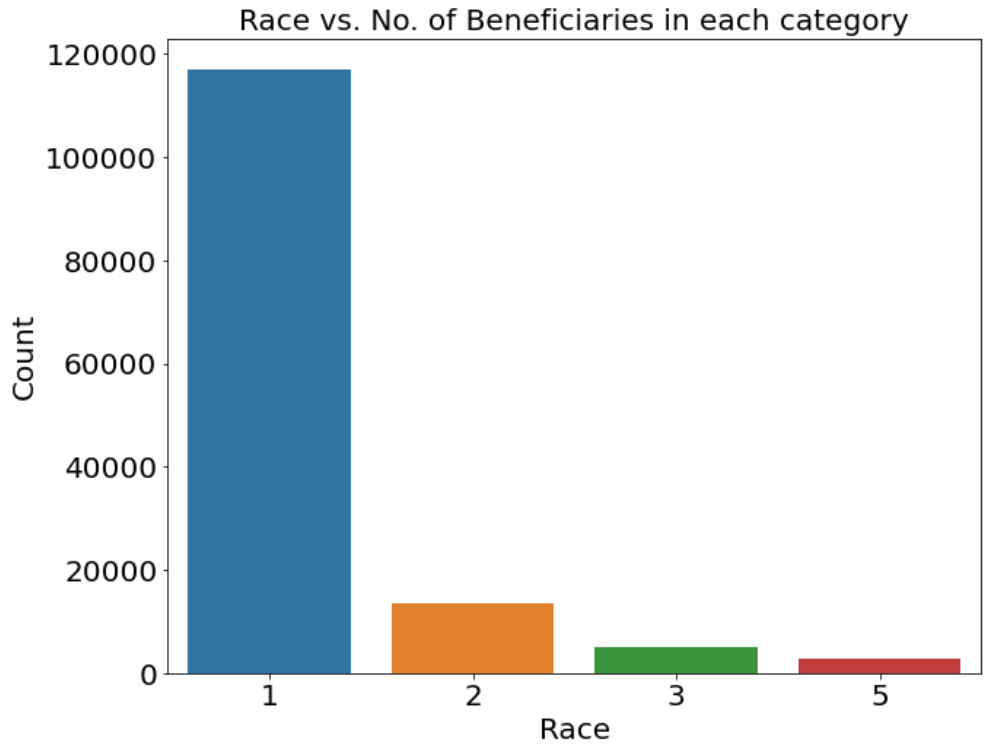
  Among all states:

  State-5 has maximum number of Beneficiaries (12,052)

  State-2 has least number (196)

- **In Race column, Category-4 is missing.**

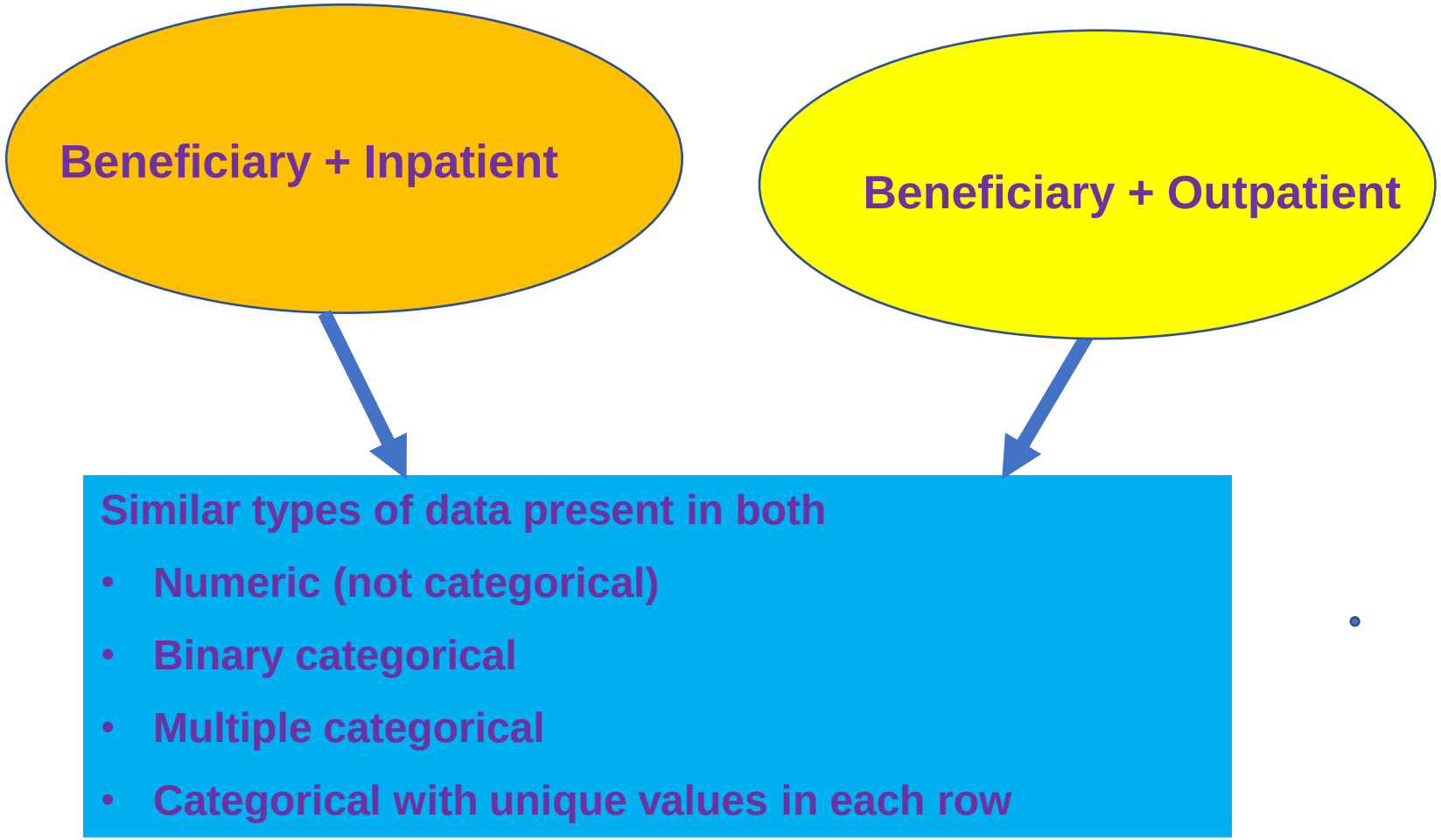
  The bar plot below shows the different categories with their Beneficiary counts:



- **County-200 has highest number of beneficiaries (3,943)**
  Few counties has only 1 beneficiary

Data Prepocessing

- Beneficiary dataset was merged with Inpatient and Outpatient dataset separately:



Data Prepocessing (Continued…)

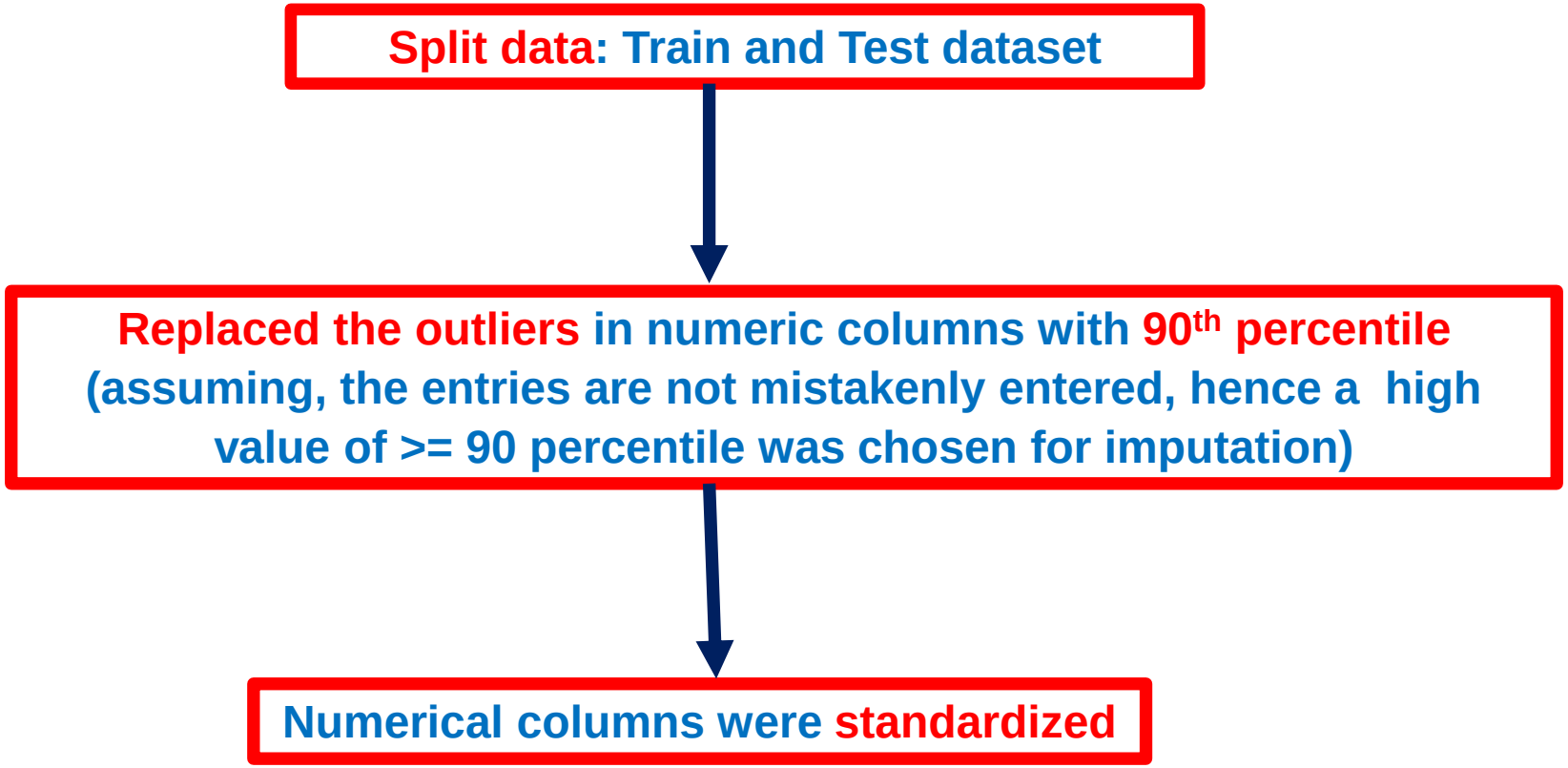
- Groupby data under each Provider:

  1. In numerical columns, data was grouped and their average was taken.
  2. In Categorical columns, data under each provider was handled in three different ways:
     - For binary category columns, fraction of one category was taken w.r.t to total.
     - For multiple categories columns, category with maximum occurrence was taken using mode, then dummies of those categories were obtained.
     - For columns with unique values, total number of entries were counted.

- Finally

  - All four groupby dataframes were merged together.
  - Then it was merged with Providers' information dataset (Train dataset) which contains whether Provider is Fraud or not.

Feature Engineering



ML models building, Evaluation and Optimization

MACHINE LEARNING MODELS

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Support Vector Classifier

Step-1: Models were built, all models' roc auc scores were > 64%, Logistic regression(LR) was highest with score ~ 70%.

Step-2: Applied SMORT(Synthetic Minority Oversampling Technique) to unbalanced data and made balanced, all Models rebuilt, Perfomance improved, LR was with highest roc auc score of 76%.

Step-3: GridSearchCV and RandomizedSearchCV were applied to LR, highest roc auc score obtained was 76%.

Step-4: To obtain Best Features, Select Best K was applied, many constant and quasi-constant features(columns), were observedand removed them. LR model rebuilt, performance almost remain same.

Step-5: PCA (Principal Component Analysis) was applied, LR was rebuilt. With 10 PCA components, best roc auc score was 76%.

Step-6: LR with PCA and without PCA had almost same roc auc score. Model with PCA was chosen as best for deployment based on run time (in future when applied to large dataset).

Models with Roc auc score:

Model	Roc_auc_score	Roc_auc_score (in Logistic Regression, applying PCA) (n_components = 10)
Logistic Regression	0.76	0.76
Decision Tree Classifier	0.71	
Random Forest Classifier	0.74	
Gradient Boosting Classifier	0.75	
Support Vector Classifier	0.75	

Winner: Logistic Regression

Logistic Regression with best hyperparameters:

Model	N_components	C	solver	Penalty	Max_iter	roc_auc_score
Logistic Regression	10	0.1	liblinear	L1	100	0.758

Future Recommendation

- **Collection of more useful data for model building:**

  In this project, ML models were built with < 1500 Providers, whereas there are thousands of providers. To build more efficient model, more amount of data is needed.

- **Making availability of Information regarding FDA approved diagnosis/procedure related to a disease:**

  There is no information listed regarding FDA approved diagnosis/procedure related to a disease, some resources related to that would be very useful for better model building.

Acknowledgement

I am very much thankful to Python community for providing very rich packages for data analysis and Machine Learning models building.

I am very grateful to my mentor Yadunath Gupta with whom I have in-depth discussion on this project, which helped me to complete the project taking into account many aspects of the data analysis and ML model building.

Thank you

Damayanti Naik (PhD)

dr.damayanti.naik@gmail.com

https://www.linkedin.com/in/damayanti-naik-phd/

Project report: https://github.com/damayantinaik/Health_care_Fraud_Detection_Medicare/tree/main/Final_report