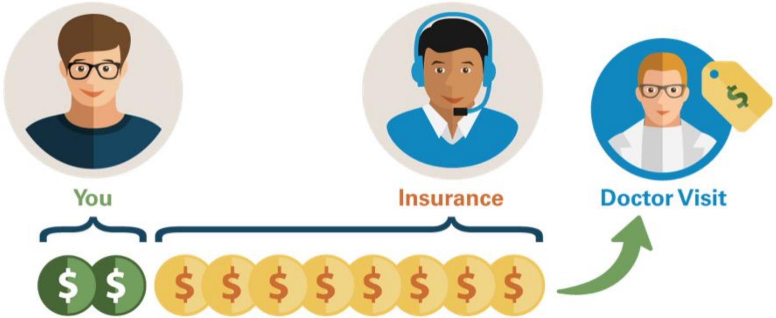
Patient Copayment Prediction



Mandy Cheung, Craig Franze, Shirali Obul, Charles Ruggiero
https://github.com/Erdos-Red-Eye/CMM_Patient_Expenses/

Problem Statement

- The primary goal of this project is to predict a patient copay for a prescribed medication given basic information including drug name and insurance number.
- Recent research suggest that this is a difficult problem facing physicians and the lack of accurate information impedes informed conversations about financial trade-offs between physicians and patients.[1]

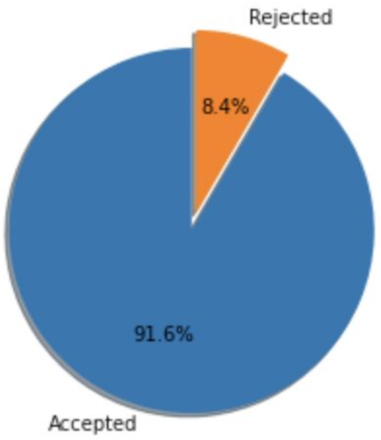
[1]Sloan C.E., Millo L., Gutterman S., BA, Accuracy of Physician Estimates of Out-of-Pocket Costs for Medication Filling, JAMA Netw Open. 2021;4(11):e2133188. doi:10.1001/jamanetworkopen.2021.33188

Data

- The data set consists of approximately 14 million simulated pharmacy transactions for the year 2022.

```
(13910244, 9)
```

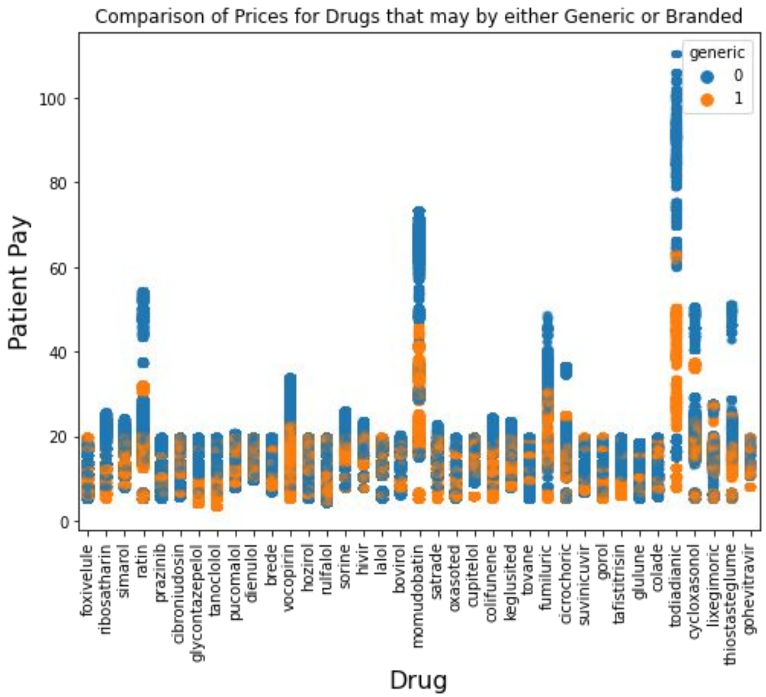
	tx_date	pharmacy	diagnosis	drug	bin	pcn	group	rejected	patient_pay
0	2022-01-02	Pharmacy #6	G99.93	branded tanoclolol	725700	1UQC	NaN	False	13.39
1	2022-01-02	Pharmacy #42	U60.52	branded oxasoted	664344	NaN	52H8KH0F83K	False	7.02
2	2022-01-02	Pharmacy #37	Q85.91	branded cupitelol	725700	1UQC	NaN	False	13.39
3	2022-01-02	Pharmacy #30	U60.52	generic oxasoted	571569	KB38N	6BYJBW	False	10.84
4	2022-01-02	Pharmacy #18	N55.01	branded mamate	664344	NaN	ZX2QUWR	False	47.00
...	...	...	...	...	...	...	...	...	...
13910239	2022-12-30	Pharmacy #42	U27.71	branded colifunene	322463	NaN	HO8HUGL	True	0.00
13910240	2022-12-30	Pharmacy #45	N59.44	generic tafistitrisin	664344	NaN	TFZOR5R49	False	6.28
13910241	2022-12-30	Pharmacy #54	W50.87	generic tanoclolol	691847	N098KI	6SP1DG	False	6.94
13910242	2022-12-30	Pharmacy #0	I68.27	branded prazinib	96934	S76J7V6	NaN	False	13.93
13910243	2022-12-30	Pharmacy #46	G99.93	branded bovirol	322463	3O71UTS	NaN	False	12.22

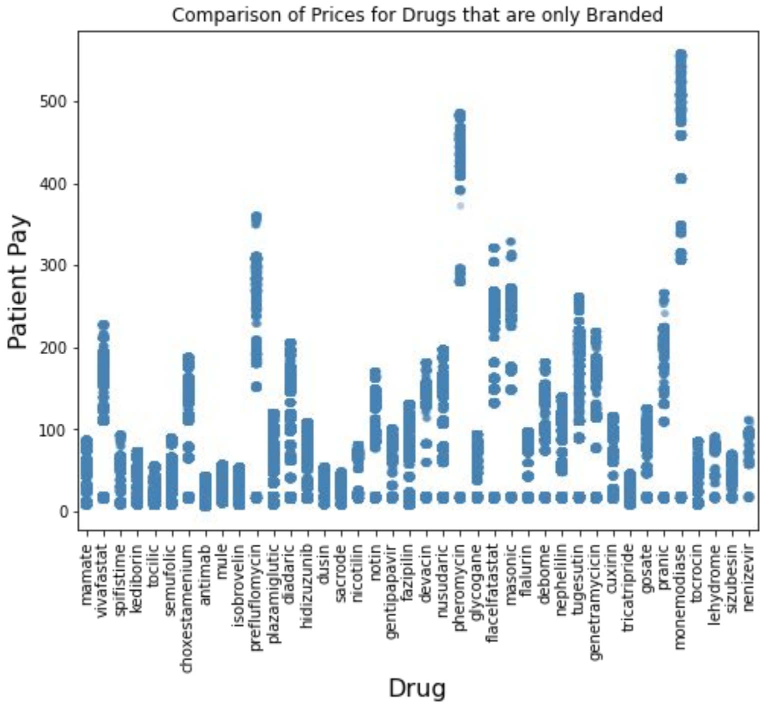


Data

tx_date – The date on which the pharmacy transaction was attempted
pharmacy – The particular pharmacy where the transaction was attempted
diagnosis – The diagnosis of the patient associated with the transaction
drug – The drug that the patient was prescribed that the pharmacy is attempting to bill
bin – The broadest identifier of a patient's insurance plan (banking identification number)
pcn – An identifier that more narrowly specifies a plan underneath the broader "bin"
group – Another identifier that more narrowly specifies a plan underneath the broader "bin"
rejected – Whether the billing transaction was rejected by the plan
patient_pay – The amount of copayment for which the patient is responsible

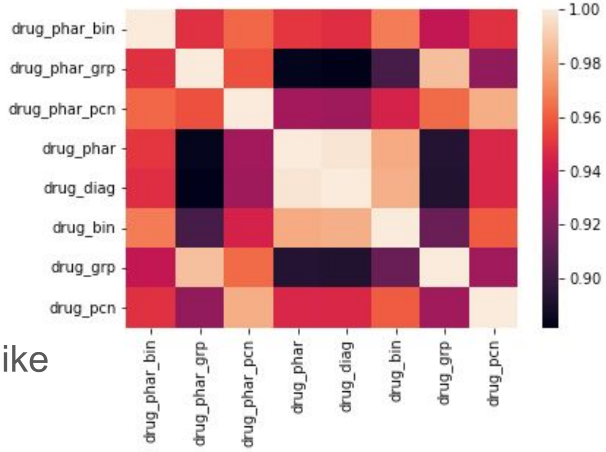
Exploratory Data Analysis





Challenge

The central challenge of this project is predicting a continuous variable (patient pay) with only categorical variables.



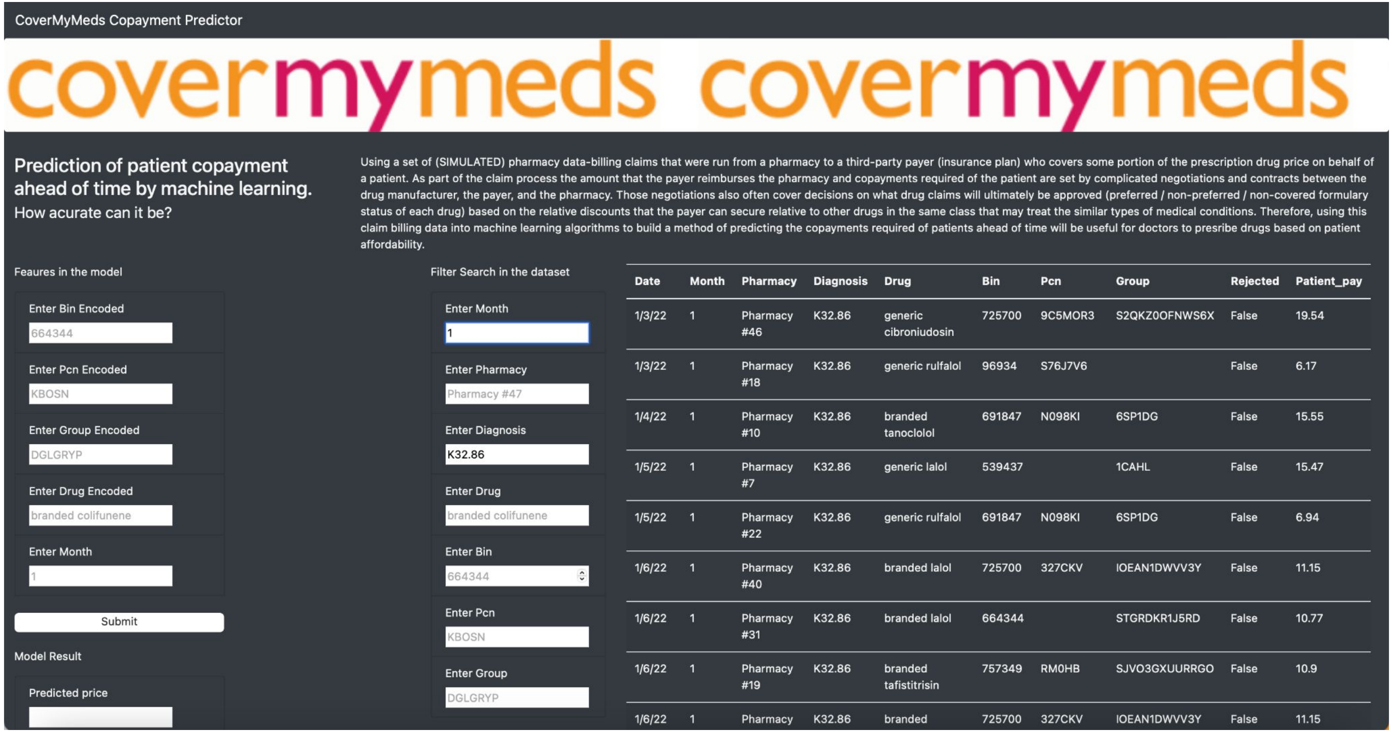
1. Feature engineering with *target encoding*
    a. Engineer 8 features by aggregating over various columns.
    b. Allows for impute claim entries when some of the features like drug, group, pcn or pharmacy are missing.
    c. Creates a highly correlated feature set.
2. *Ordinal encoding* with regressors that natively support categorical variables.
    a. Minimizes feature engineering
    b. Limits the possible regressors that may be used.

Models (CV Scores)

Model	**Categorical Variable Encoding**	**RMSE**	**Mean Absolute Percentage Error**
Gradient Boost	Feature aggregate with target encoding	11.989	3.917
Random Forest	Feature aggregate with target encoding	11.982	3.873
Histogram-based Gradient Boosting	Ordinal encoding (*month, bin, pcn, group, drug)*	2.157	1.410

Applications

A web application (prototype) allows for the prediction of drug prices.



Applications

Given a diagnosis, e.g. 'B45.03', along with the month, bin, pcn, and group numbers the models can be used to make a financial comparison of drugs that have been prescribed over the past year for the diagnosis.

month= 5, *bin*=725700, *pcn*=1UQC, *group*=NaN

```
option_cost('B45.03', 5, 725700, '1UQC', np.nan)
```

	drugs	predicted_pay
2	generic todiadianic	49.365954
4	branded todiadianic	92.863260
0	branded pranic	221.182812
1	branded masonic	271.147075
3	branded monemodiase	541.926317

Future Work

- Build separate models for the generic and branded drugs.
- Examine the outliers where the models perform poorly and see if there is a particular commonality amongst them that could explain this.
- Combine engineered features, such as the ones used in the Random Forest model, or rejection rate for example, with the HistGradientBoosting model.
- Develop the app further by integrating these improvements to help doctors easily obtain predicted prices of medications ahead of time.