Erdős Institute

Data Science Bootcamp, May 2022.

Project: CoverMyMeds

Group: Drugs4All

Mentor: Kritika Singhal

Members: Diego Galdino and Wayne Uy

June 4th, 2022.

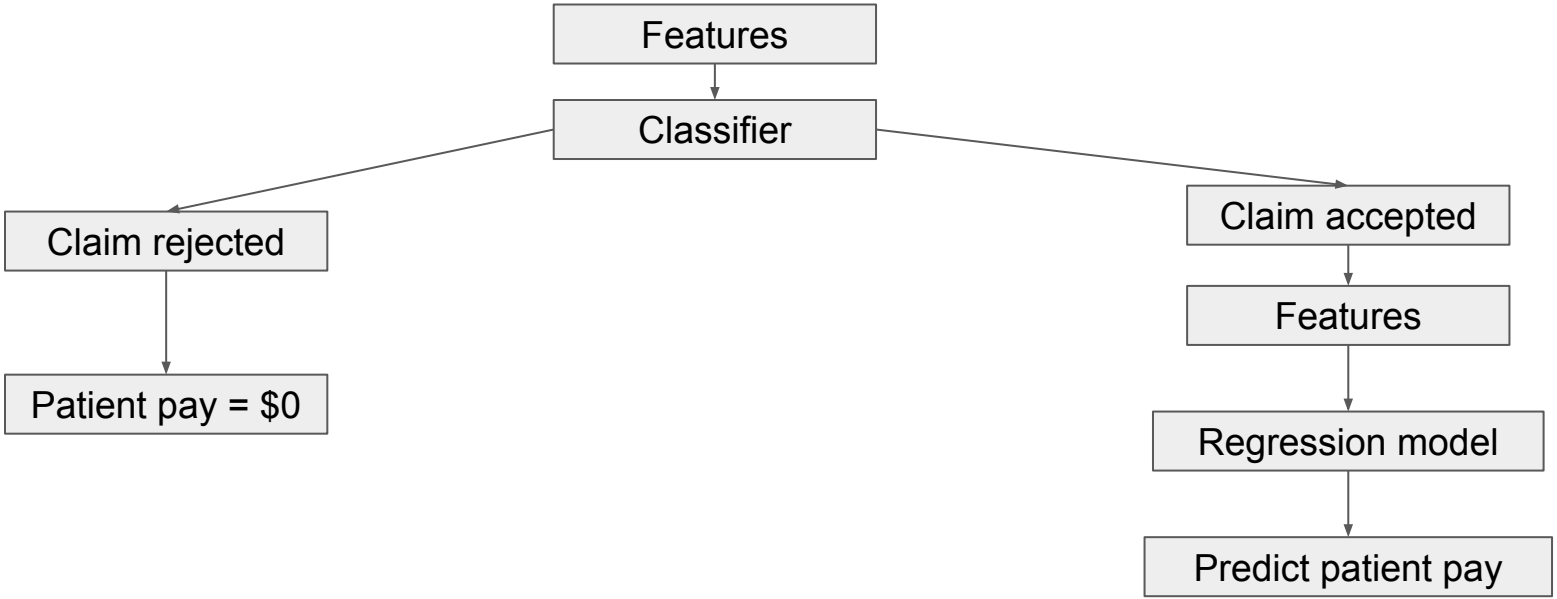
Problem definition

- Predict cost of patient pay for a particular medication
- Predict whether the claim will be accepted or rejected by insurance plan
- Group medications according to formulary status and copayment requirements

	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

Our predictive model: Two-stage prediction

- 1st stage: Predict if claim is accepted or rejected given features.
- If claim is rejected, patient pay is $0.
- 2nd stage: If claim is accepted, predict patient pay given features.



Our accomplishments: Predictive performance + App

Scores from Classification Models

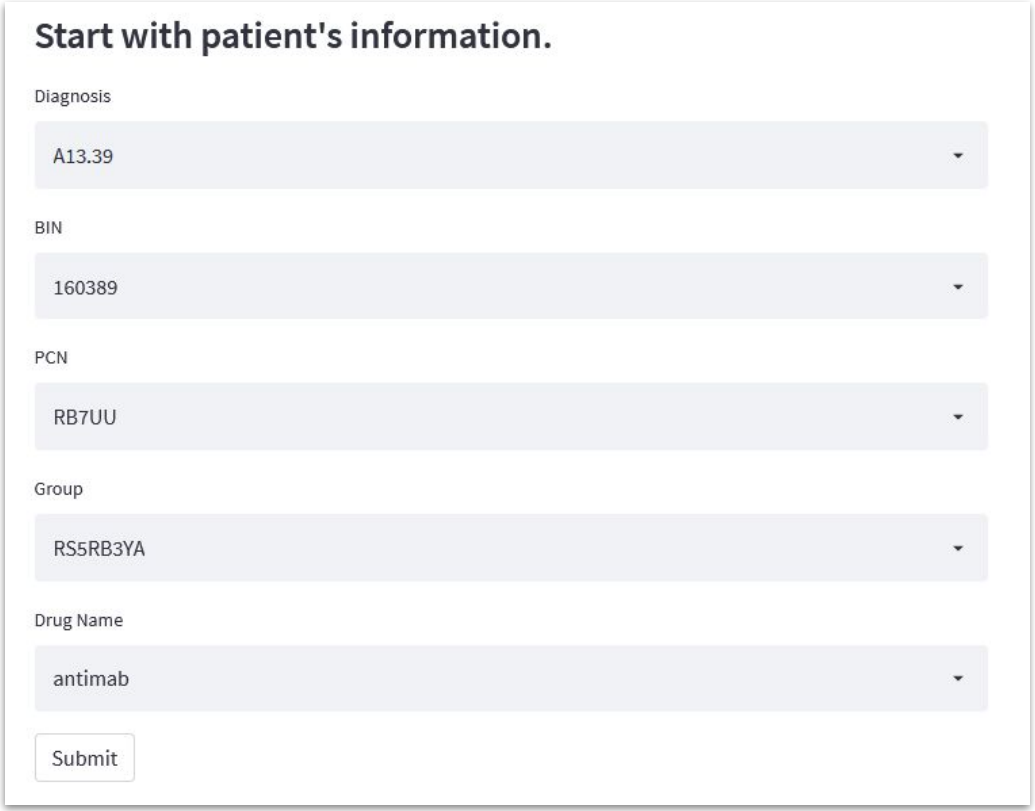
Model	Features	Negative LogLoss	Accuracy	Precision	Recall	F1	ROC_AUC
LightGBM Classifier	All	NR: -2.9018 Over: -7.4809 Under: -7.6471	NR: 0.9160 Over: 0.7834 Under: 0.7786	NR: 0.2875 Over: 0.2762 Under: 0.2730	NR: 0.0008 Over: 0.9753 Under: 0.9849	NR: 0.0017 Over: 0.4304 Under: 0.4274	NR: 0.8952 Over: 0.8954 Under: 0.8907
LightGBM Classifier	Plan-, drug-, and diagnosis-related	NR: -2.9025 Over: -7.6304 Under: -7.6474	NR: 0.9160 Over: 0.7791 Under: 0.7786	NR: 0.3032 Over: 0.2743 Under: 0.2740	NR: 0.0010 Over: 0.9922 Under: 0.9932	NR: 0.0021 Over: 0.4298 Under: 0.4295	NR: 0.8972 Over: 0.8970 Under: 0.8937

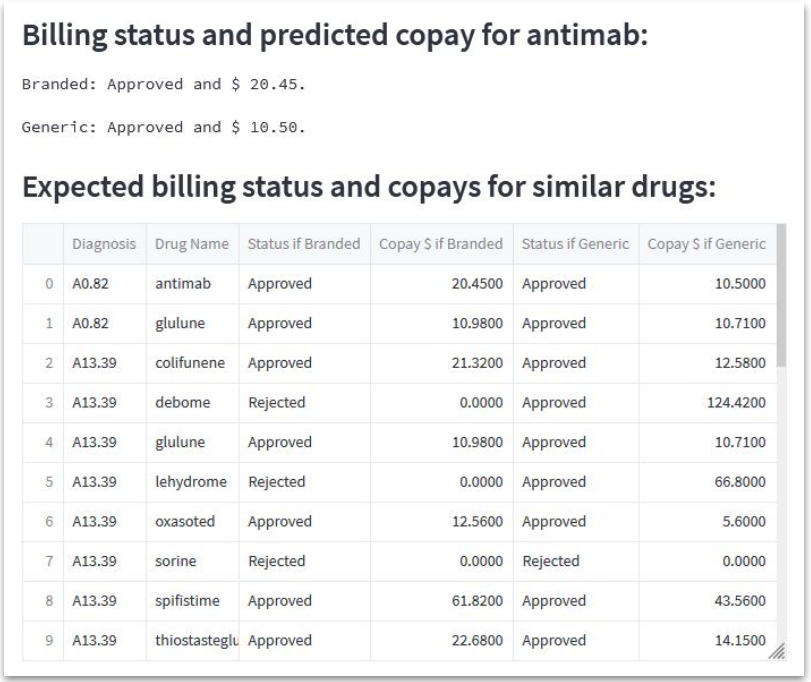
Our accomplishments: Predictive performance + App

Scores from Regression Models

Model	Features	MSE	MAPE
Baseline (Mean of copays)	All	1744.4081	1.2935
LightGBM Regressor (Mean copay)	All	6.7890	0.0236
LightGBM Regressor (Mean copay)	Plan-, drug-, and diagnosis-related	6.9942	0.0175
LightGBM Regressor (Max copay)	All	7.1812	0.0251
LightGBM Regressor (Max copay)	Plan-, drug-, and diagnosis-related	7.3717	0.0180

Our accomplishments: Predictive performance + App





	Diagnosis	Drug Name	Status if Branded	Copay $ if Branded	Status if Generic	Copay $ if Generic
0	A0.82	antimab	Approved	20.4500	Approved	10.5000
1	A0.82	glulune	Approved	10.9800	Approved	10.7100
2	A13.39	colifunene	Approved	21.3200	Approved	12.5800
3	A13.39	debome	Rejected	0.0000	Approved	124.4200
4	A13.39	glulune	Approved	10.9800	Approved	10.7100
5	A13.39	lehydrome	Rejected	0.0000	Approved	66.8000
6	A13.39	oxasoted	Approved	12.5600	Approved	5.6000
7	A13.39	sorine	Rejected	0.0000	Rejected	0.0000
8	A13.39	spifistime	Approved	61.8200	Approved	43.5600
9	A13.39	thiostasteglu	Approved	22.6800	Approved	14.1500

App URL:
https://share.streamlit.io/diego-galdino/erdos_bootcamp_may22_cover_my_meds/main

Classification: Predict if claim is accepted or rejected

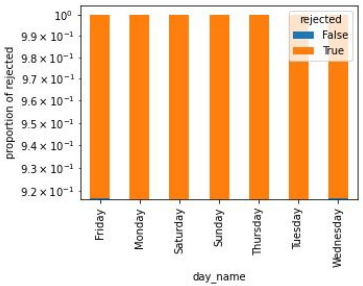
Data Cleaning:

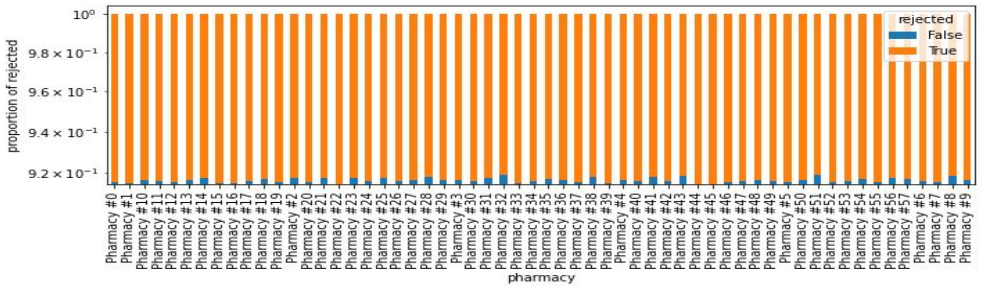
- Remove duplicate rows in the data set
- Remove rows whose with pcn and group both missing

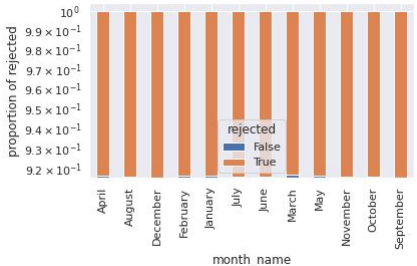
Feature Engineering:

- Extract month and day of week from date data
- Extract drug brand type and drug name
- Extract first letter of diagnosis
- Merge bin, pcn, group to identify insurance plan specifics.

Exploratory data analysis: Output does not vary significantly for these inputs

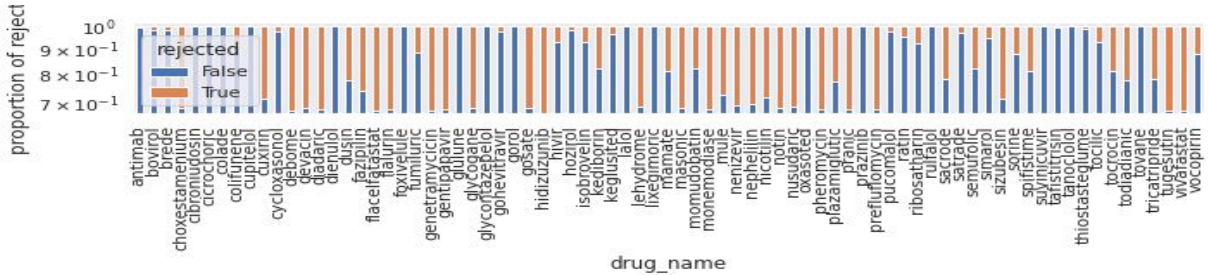


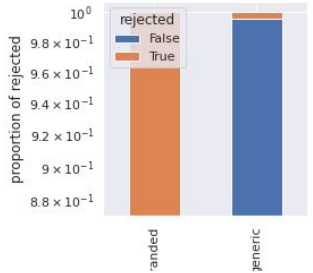


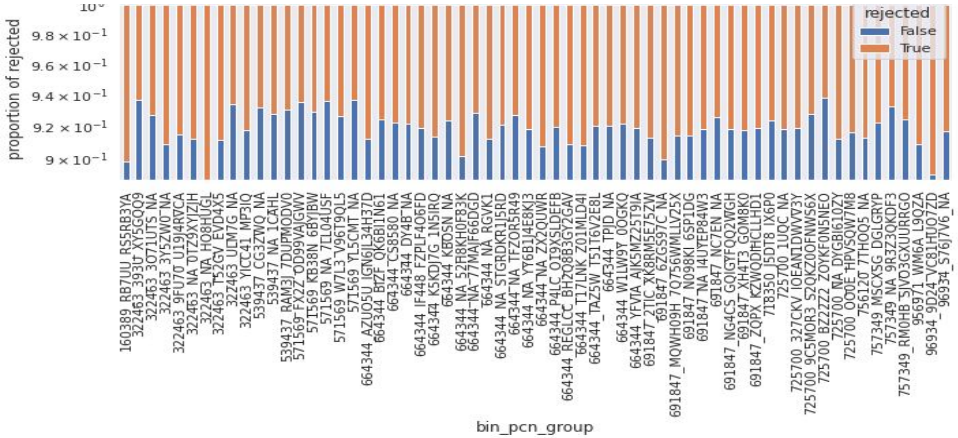


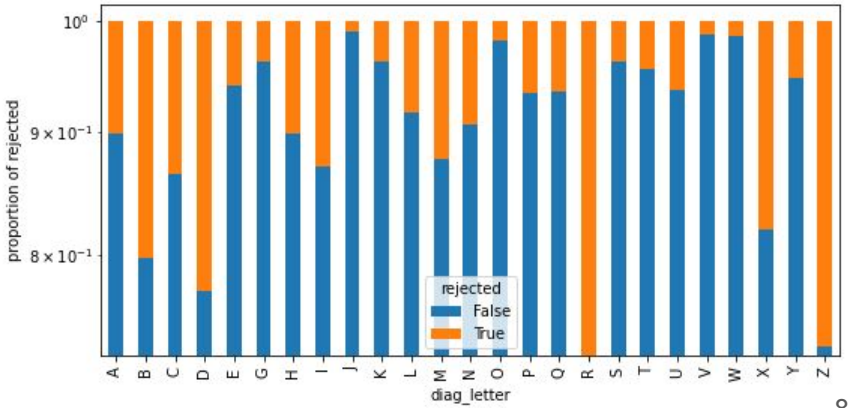
Classification: Predict if claim is accepted or rejected

Exploratory data analysis: Output varies significantly for these inputs









Classification: Predict if claim is accepted or rejected

Machine learning model specifications:

- Gradient boosting via LGBM
  - Native support for categorical features without need to one-hot encode
  - Balance between computation time, accuracy, interpretability
- Model claim accepted as 1 and claim rejected as 0
- Cross-validation was used to tune hyperparameters: number of trees, depth of tree, number of leaves, learning rate
- Address class imbalance by oversampling/undersampling techniques

Precision 0.7601, Recall 0.9922 on the test set

Confusion Matrix with test dataset into the best model:

	Approved	Rejected
Approved	139344	43985
Rejected	119	16552

Regression: Predict patient pay for accepted claims

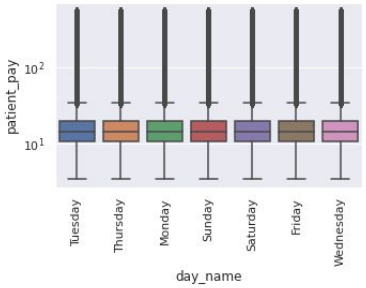
Data Cleaning:

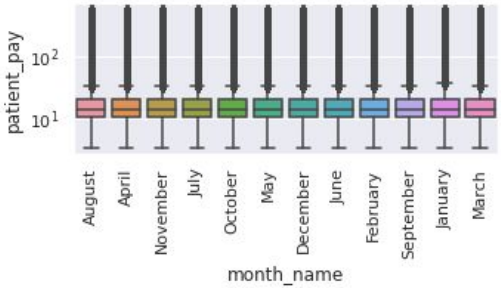
- Remove duplicate rows in the data set
- Remove rows whose with pcn and group both missing
- Take max/average of patient pay for rows with identical feature values in training set

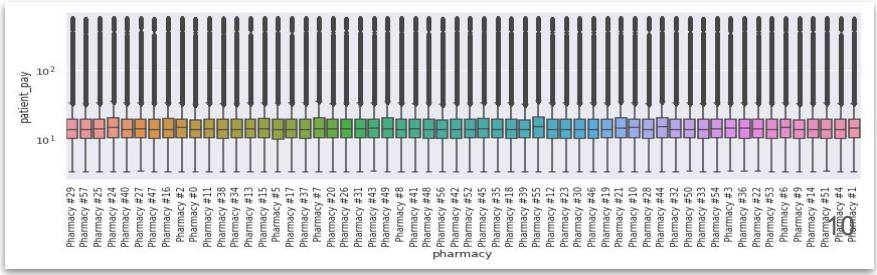
Feature Engineering:

- Extract month and day of week from date data
- Extract drug brand type and drug name
- Extract first letter of diagnosis
- Merge bin, pcn, group to identify insurance plan specifics.

Exploratory data analysis: Output does not vary significantly for these inputs

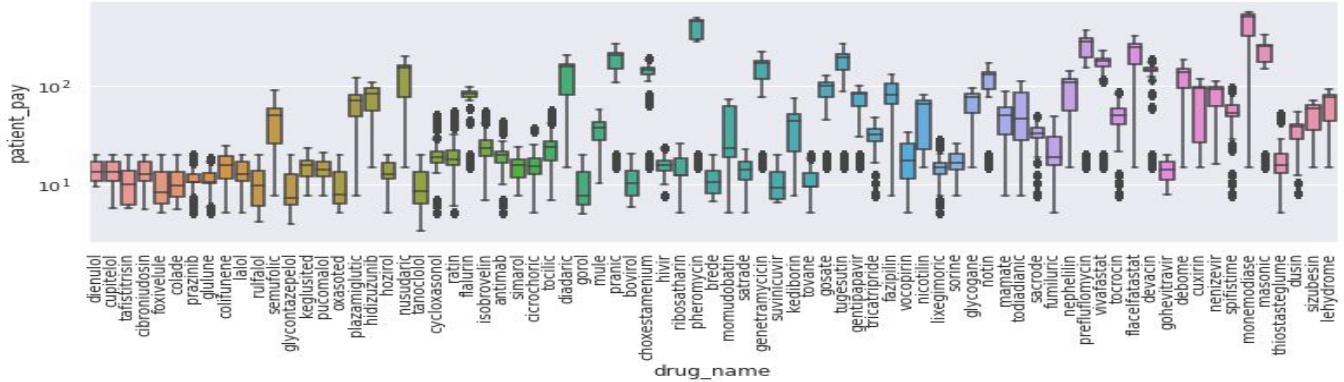


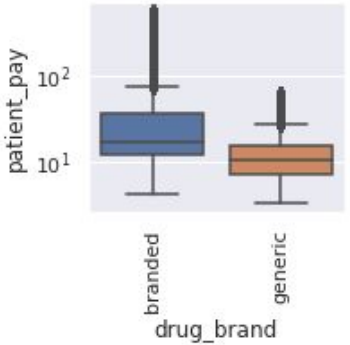


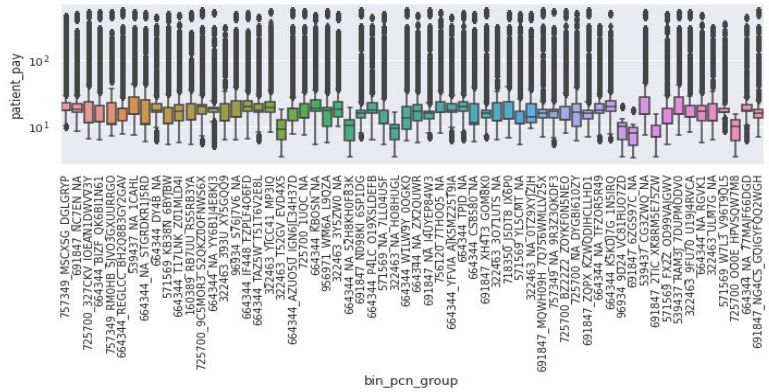


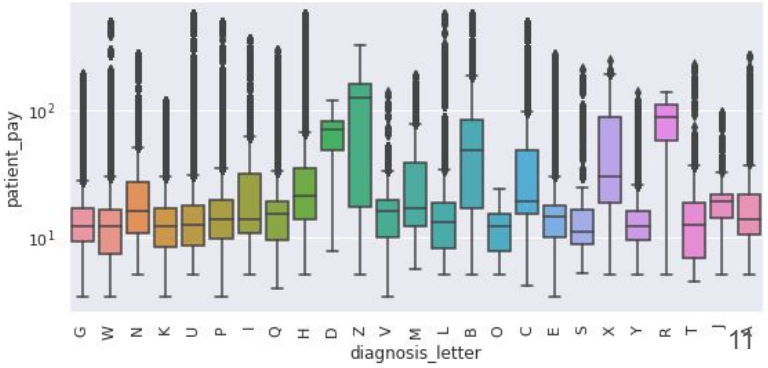
Regression: Predict patient pay for accepted claims

Exploratory data analysis: Output varies significantly for these inputs









Regression: Predict patient pay for accepted claims

Machine learning model specifications:

- Gradient boosting via LGBM
  - Native support for categorical features without need to one-hot encode
  - Balance between computation time, accuracy, interpretability
- Cross-validation was used to tune hyperparameters: number of trees, depth of tree, number of leaves, learning rate

Regression: Predict patient pay for accepted claims

Residuals and Predicted vs Observed plots for the final LightGBM regression model.

6.9942 MSE and 0.0175 MAPE on the test set