Johnson & Johnson - Prediction of commercial insurance payments for surgical procedures

Rahulraj Singh (rs4211), Prerit Jain (pj2383), Parth Gupta (pg2677), Mahesh Jindal (mj3038), Ayush Baral (ab5247)

Problem Statement & Goals

- In this project, we propose a novel machine-learning approach to address payment information gaps at the MSA and procedure level.
- Predicting MSA-level and procedure-level insurance payout.
- Furthermore, we aim to improve the results given by the AutoML platform, DataRobot, which was previously used in this project.
- Constraint - The predicted insurance payout for the **ASC** type procedures should be less than the payout for the **Inpatient** procedures for each MSA, group and year combination.

Datasets

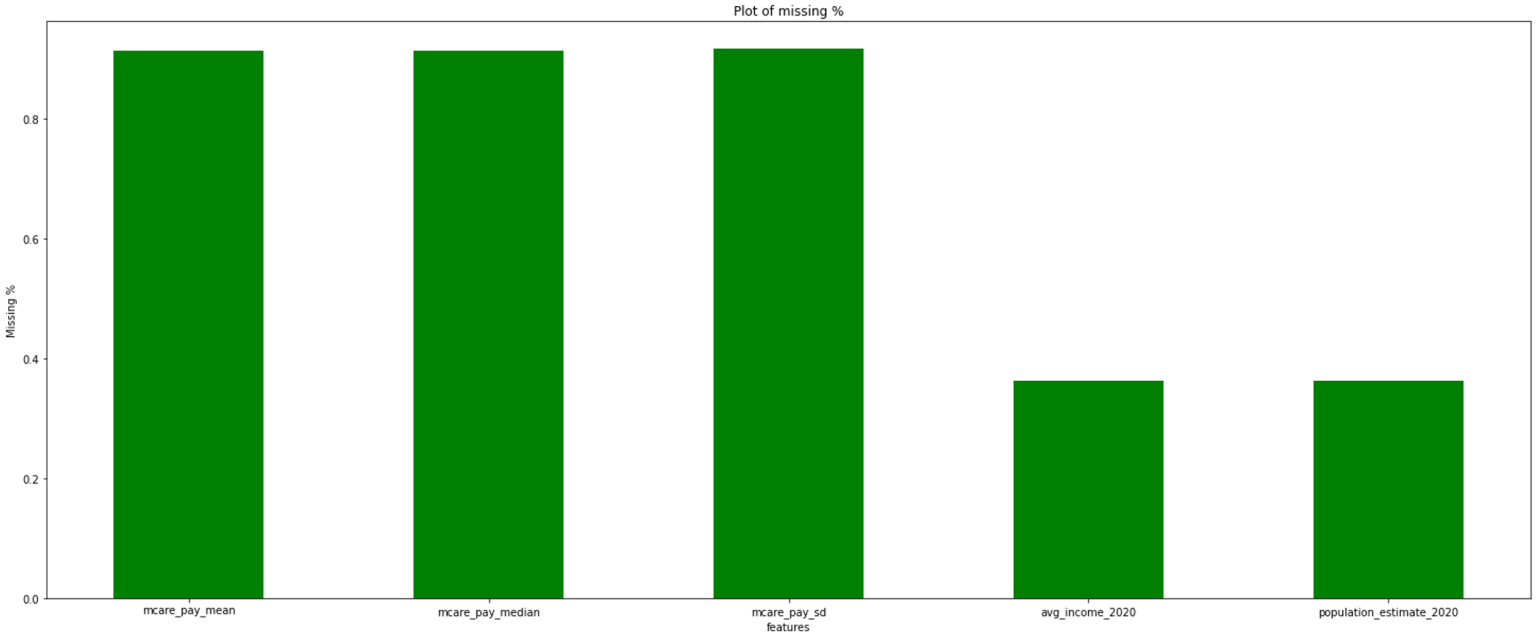
Existing Data:

- **Hospital Dataset**: This dataset contains data about all the hospitals across all regions of the US, holds their MSA codes and the amount paid to each of these hospitals throughout the COVID-19 pandemic.
- **MSA Data**: This dataset contains information about all MSAs across the US and reports their population estimates for three consecutive years (2010, 2011 and 2012). We further connect this with life expectancy data to combine life expectancy and population information.

Externally Added Data:

- **Average Income Data**: This dataset contains MSA-level average family incomes for households in that MSA.
- **Life Expectancy Data**: This county-level dataset gives the average life expectancy for all counties across the US. We further connect this with MSA data to combine life expectancy and population information.
- **Procedure Level Data** - This dataset contains the diagnostic statistics of the different procedures.

Missing Value Analysis



Analysis of Missing Values: Only columns having >1% missing values are displayed in the above chart. There are >80% missing values in mcare-related features (mean, median and standard deviation), and missing value indicators will be used instead. Avg Income and Population Estimated were imputed by the corresponding MSA-level mean values.

Feature Engineering

Variables created/transformed:

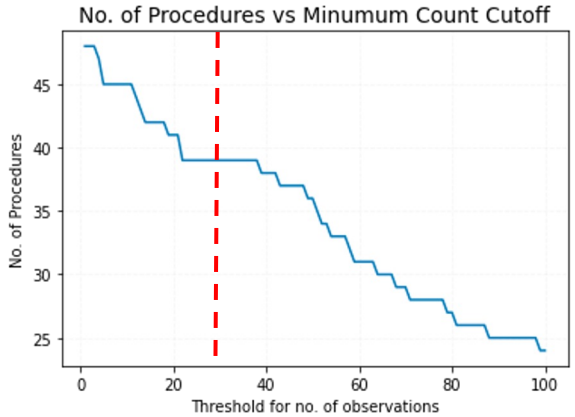
1. Haversine distance: used Lat/Long
2. Preference for site - 0 ASC , 1 Inpatient and Outpatient. (To tackle monotonicity constraints)
3. Cluster Numbers
4. Log Transformation of target variable.
5. Average Income
6. Population Estimate
7. Mean/Median/Mode/Occurrence

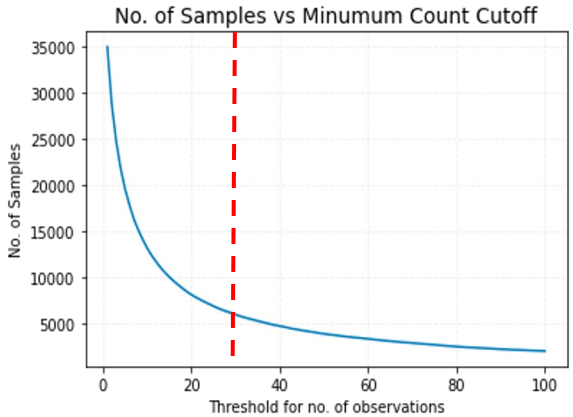
Also, we have applied following encodings to the variables:

1. Target Encoding: Group, MSA
2. One Hot Encoding: Site, Cluster Numbers
3. Standard Scaling: Numeric features

Assumptions

Cutoff value for no. of observations - There are 51 unique values in dataset.

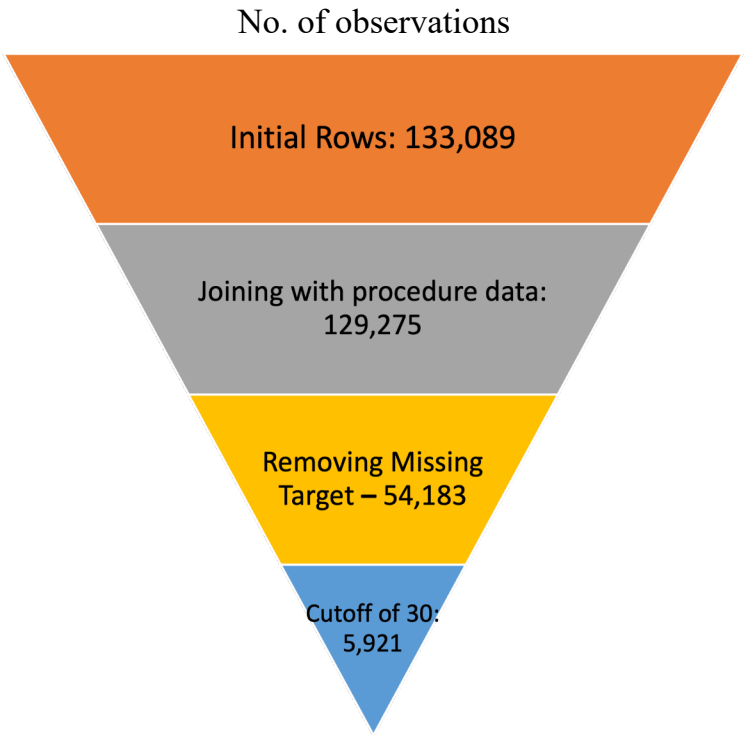


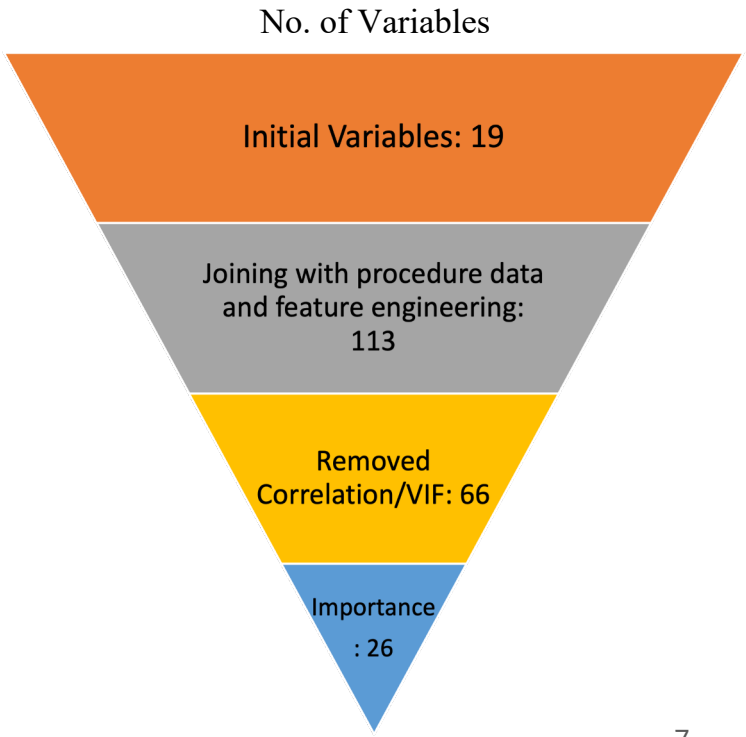


Cutoff of 30 was decided on based on 3 factors:

- No. of training samples remaining after filtering - we want large dataset.
- No. of unique procedures - ideally want the set to be exhaustive
- The threshold should be higher enough to represent the actual summary (mean/median/mode) and withstand the effect of outliers.

Data Funnel


No. of observations
Initial Rows: 133,089
Joining with procedure data: 129,275
Removing Missing Target – 54,183
Cutoff of 30: 5,921


No. of Variables
Initial Variables: 19
Joining with procedure data and feature engineering: 113
Removed Correlation/VIF: 66
Importance : 26

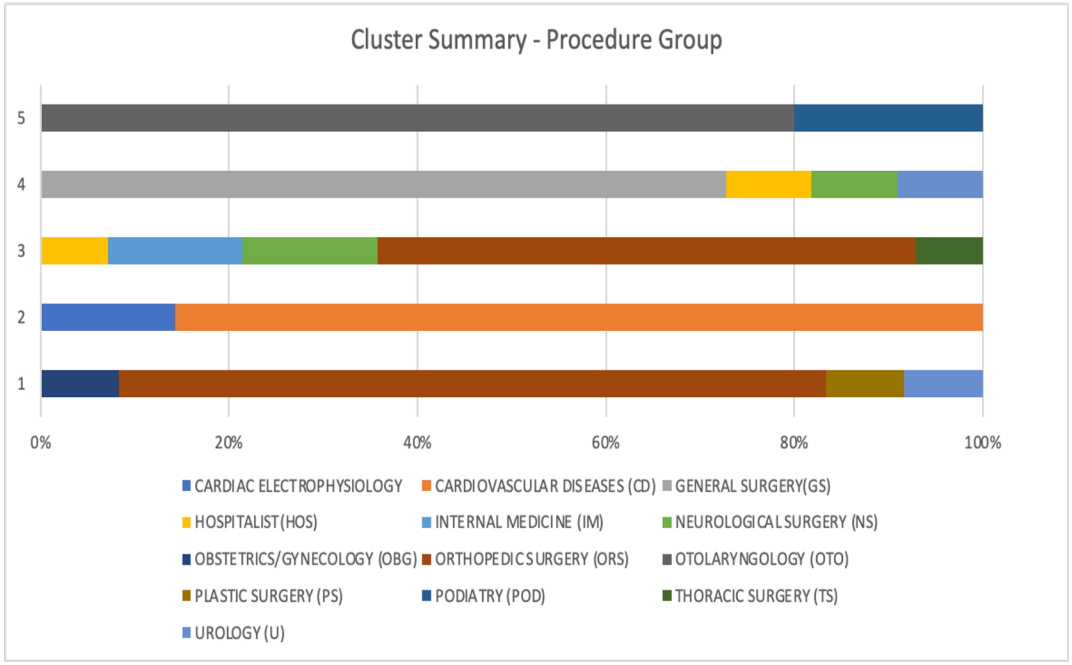
Clustering: K-Medoids

Methodology:

1. Computed gower distances for the input features present in cost discharge dataset.
2. Generated clusters by grouping data based on medoids.
3. Chosen the optimal *k value* with lowest inertia.

Cluster Number	Average of pat_cost	Average of discharge_to_home_pct	Average of IP_pct	No. of Groups
1	12,017.72	83%	27%	12
2	23,484.36	96%	18%	7
3	28,729.43	43%	95%	14
	17,47	84%	53%	

Clustering: K-Medoids



- **Cluster 1:** 75% Orthopedics. All PLASTICS(PL) and OBSTETRICS/GYNECOLOGY (OBG).
- **Cluster 2:** All cardiac related procedures-
  - CARDIAC ELECTROPHYSIOLOGY - 14%
  - CARDIOVASCULAR DISEASES (CD) - 86%
- **Cluster 3:** Not a very homogeneous cluster - Majority of orthopedics (55%) followed by NEUROLOGICAL SURGERY (NS)
- **Cluster 4:** Dominated by GENERAL SURGERY(GS) - 73%. Others include HOSPITALIST(HOS) and NEUROLOGICAL SURGERY (NS).
- **Cluster 5:** Dominated by ENT and DPM -
  - OTOLARYNGOLOGY (OTO)- 80%
  - PODIATRY (POD)- 20%

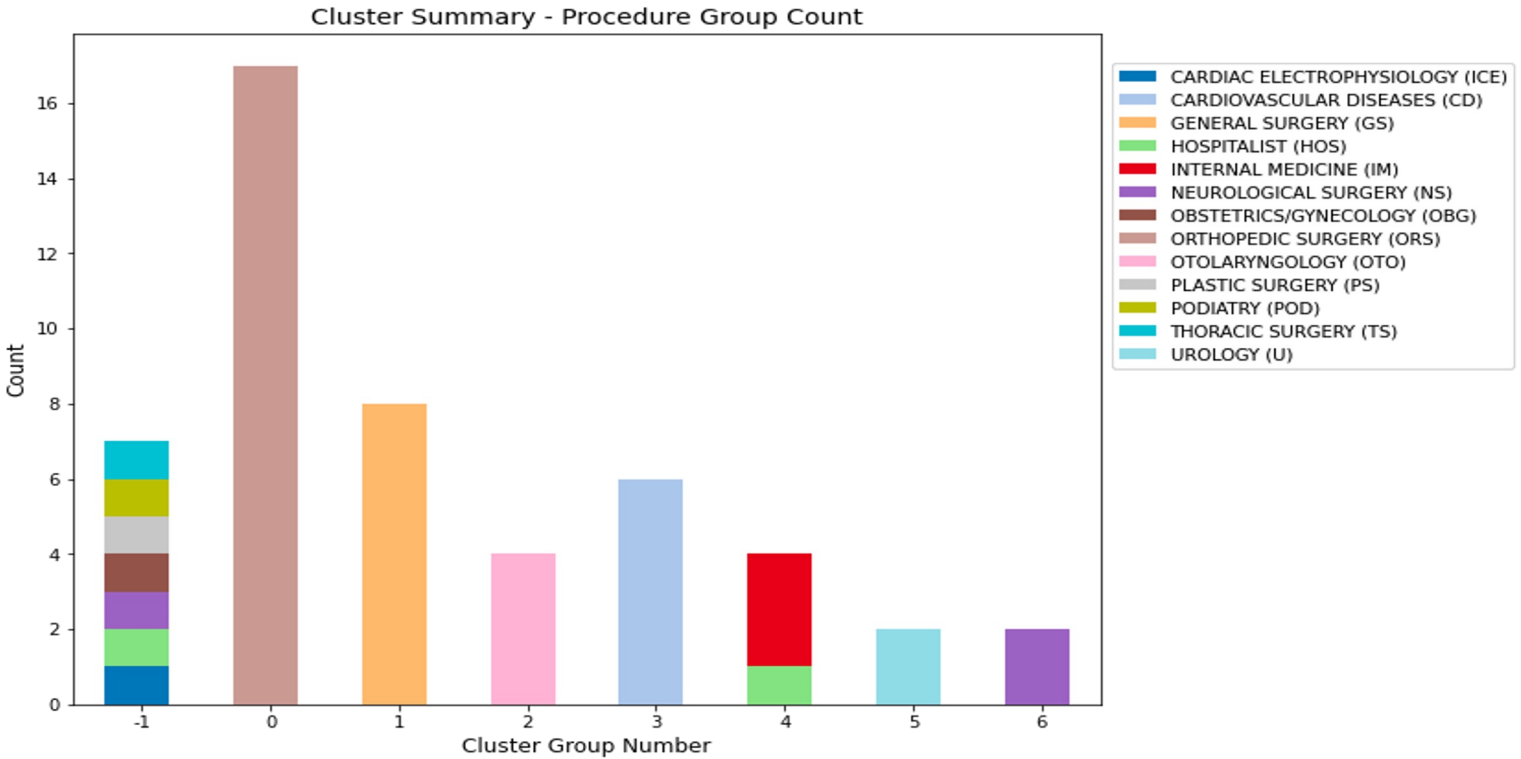
Clustering: Density Based Spatial Clustering

Methodology:

1. Computed gower distances for the input features present in cost discharge dataset.
2. Used Density Based Spatial Clustering with optimal hyperparameter values for ***epsilon*** and **min_samples**.

Cluster Number	Average of pat_cost	Average of discharge_to_home_pct	Average of IP_pct	No. of Groups
0	18273.59	62%	57%	17
1	17508.36	83%	49%	8
2	5670.52	98%	0%	4
3	24438.84	95%	20%	6
4	26883.03	56%	76%	4
5	11416.02	90%	50%	2
6	43448.35	48%	98%	2
-1	16228.02	78%	44%	7

Clustering: Density Based Spatial Clustering



Baseline Model & Performance

- Linear Regression: Our baseline model is Linear Regression with 80/20 split of the training and testing dataset. Also, 26 features are used to train the model.
- Results are as follows:
  - Training-set MAPE: 24.25%
  - Testing-set MAPE: 30.86%

Next, we have implemented various models - Elastic Net, Decision Tree, RandomForest, Gradient Boosting and XGBoost.

Model Variations and Results(1/3)

Performance of Models

Model	Dev MAPE	Test MAPE
Elastic Net	27.97%	30.35%
Decision Tree	20.76%	26.70%
RandomForest	19.15%	24.66%
Gradient Boosting	7.79%	16.96%
XGBoost	**6.44%**	**15.22%**

Performance of Models
(Cluster Number added as feature)

Model	Dev MAPE	Test MAPE
Elastic Net	27.86%	27.29%
Decision Tree	20.51%	23.25%
RandomForest	19.28%	21.50%
Gradient Boosting	7.02%	14.87%
XGBoost	**4.95%**	**14.68%**

Model selection was done using Grid Search CV with parameters cv = 10 and 80/20 split

After Analysing the above results we have decided to train XGBoost as our final approach.

Model Variations and Results(2/3)

Performance of Models
(Procedure level data added as features)

Model	Dev MAPE	Test MAPE
Elastic Net	27.71%	28.94%
Decision Tree	20.23%	24.04%
RandomForest	18.75%	21.34%
Gradient Boosting	6.51%	14.56%
XGBoost	**4.53%**	**14.21%**

Added **pat_cost, discharge_to_home_pct, phy_spec, IP_pct** as additional features.

Model selection was done using Grid Search CV with parameters cv = 10 and 80/20 split

Monotonic Constraints

- A common type of constraint in this situation is that certain features bear a monotonic relationship to the predicted response:
  - $f(x_1,x_2,\ldots,x,\ldots,x_{n-1},x_n) \leq f(x_1,x_2,\ldots,x',\ldots,x_{n-1},x_n)$ whenever $x \leq x'$ is an increasing constraint; or
  - $f(x_1,x_2,\ldots,x,\ldots,x_{n-1},x_n) \geq f(x_1,x_2,\ldots,x',\ldots,x_{n-1},x_n)$ whenever $x \leq x'$ is a decreasing constraint.
- XGBoost has the ability to enforce monotonicity constraints on any of the features.
- In our case we have created a variable site_weight = 0 if site is ASC, otherwise 1.
- We have enforced increasing constraint for the feature site_weight.

Model Variations and Results(3/3)

Performance of Models
(XGBoost model Monotonicity Constraints)

Model	Dev MAPE	Test MAPE
XGBoost	3.93%	13.71%

After enforcing this constraint in our model and comparing for (msa, group, year) for ASC and inpatient in the training + test dataset, we are predicting correctly for all 81 cases (unique combination of MSA, Group and Year).

Model selection was done using Grid Search CV with parameters cv = 10 and 80/20 split

Result Comparison with DataRobot Model

Procedure	Model MAPE	DataRobot MAPE
ankle_fix	19.79%	25.57%
ant_cerv_fusion	16.85%	21.31%
bariatric	9.70%	11.71%
breast reconstruction	12.69%	15.60%
bsp	21.18%	12.16%
bunionectomy	13.25%	27.99%
cardiac ablation	14.76%	16.84%
cardiac_ablaton_ice	11.45%	22.32%
colorect	15.85%	10.02%
fess	15.56%	24.07%
hepat	29.94%	32.23%
hernia	20.88%	11.73%
hysterect	12.82%	11.96%
lap appendectomy	9.20%	13.51%
mastectomy	15.04%	23.19%
post_tls_fusion	11.89%	8.96%
prox_tibia_fixation	4.53%	16.18%
radius/ulna internal fixation	11.83%	13.12%
robotic_assisted_surgery	21.66%	18.15%
rtc_slap_bank	18.05%	15.93%
septoplasty	13.70%	14.29%
tha	8.78%	6.81%
tka	13.03%	8.83%
tpa	27.33%	7.07%
	15.41%	***16.23%***

Next Steps

- Use deep learning models to impute missing values
- Hyperparameter tuning - Hyperopt, Bayesian Optimization.
- Confirm the following with J&J team:
  - Are datasets same for the DataRobot and dataset provided to us?
  - What is the performance of the DataRobot model on the training dataset?
  - No. of cases predicted amount for ASC > Inpatient.