Prediction of Commercial Insurance Payments using Machine Learning

Mentors:
Cindy Tong, Kade Etter, Ziyu Tan

Team Introduction

Columbia University M.S Data Science Students (Graduating Dec ‘22)

Ayush Baral
(1+ Yrs of work exp.)

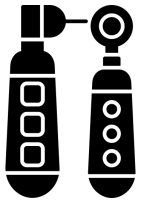
Mahesh Jindal
(2+ Yrs of work exp.)

Parth Gupta
(1+ Yrs of work exp.)

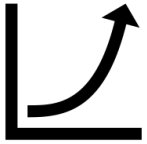
Prerit Jain
(4+ Yrs of work exp.)

Rahulraj Singh
(3+ Yrs of work exp.)

Current Problem



Understanding the dynamics of the commercial insurance payouts are **key to price** the Johnson & Johnson's (J&J) **Medical Devices** and **increase sales**.



These predictive tasks can be leveraged via **patients' geographical location** and **claims data**.

However, **predicting** for each surgical procedures and Metropolitan Statistical Area(MSA) is **time consuming , complex** and **often inviable** due to low frequency of occurrences.

Business Objective

Predict insurance payout for multiple surgery procedure at MSA level and save development time

Showcase the **procedure-level** performance (MAPE) and **eliminate** the **dependency** on **DataRobot**, an AutoML tool, to **save revenue**.

Enforce the **business requirement** – payout for **ASC site** should be **lower** than the **Inpatient care facility** at MSA and procedure level.

Data and Modeling (1/3)

Datasets

- MSA Data
- Hospital Data
- External Market Factors
- Medical Procedures

Data Preprocessing and Feature Engineering

Initial Preprocessing

- Performed Data Validation.
- Generated Aggregated raw datasets at MSA and Procedure Level.

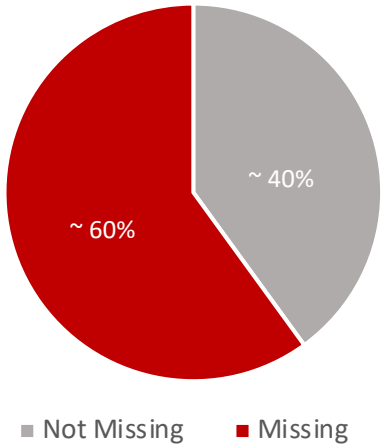
133K Reported Claims

51 Unique Procedures

30 Minimum surgeries for each procedure in MSA

Target Distribution

Target Variable (**Mean Payout**) Missing %



Feature Engineering

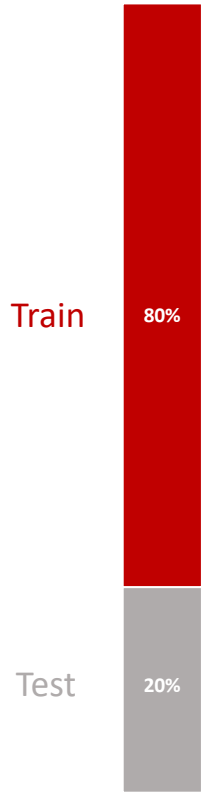
- Removed rows with missing target. Tried segmentation approaches.
- Used haversine distance to encode the geographical location.
- Log transformed the target variable.
- Added a custom weight feature to enforce monotonicity constraints.

Feature Selection and Data Split

Top Features

- Year
- MSA
- Ownership
- Avg. Income
- Population
- Life Expectancy
- Geographical Location
- #Beds in Hospital
- Facility Type
- Procedure Type
- Mean Payout
- Sites of Care
- Locality-specific
- Medicare reimbursement

DATA SPLIT



Data and Modeling (2/3)

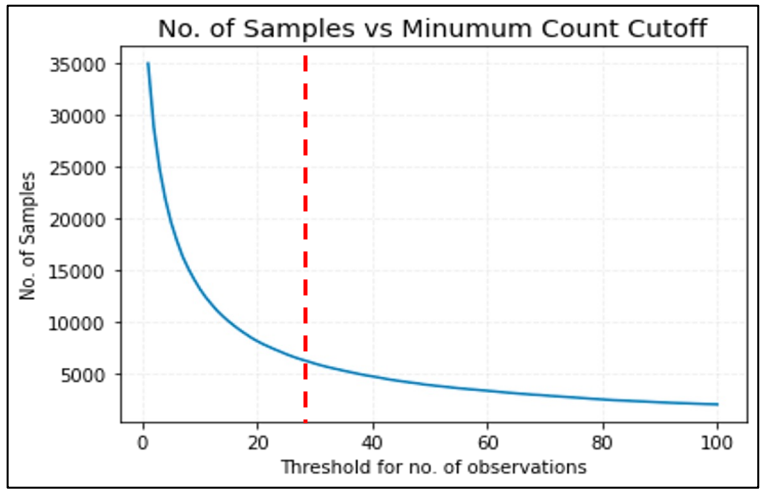
Assumptions for Minimum # of surgeries for each procedures in MSA

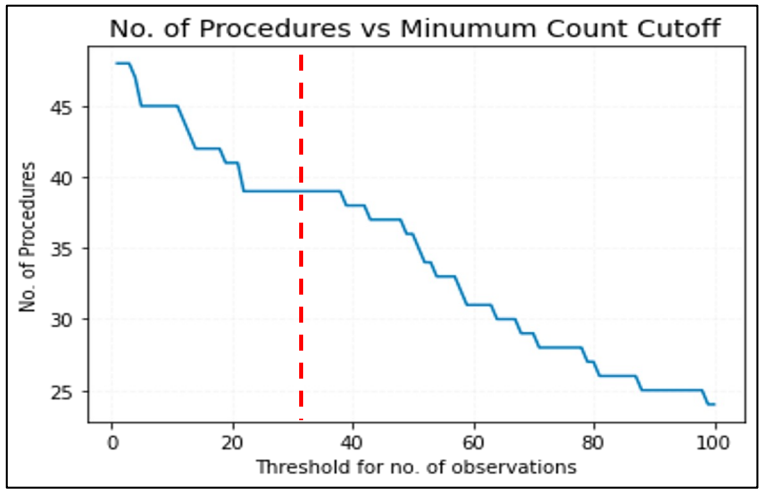
51

Unique Procedures

30

Minimum surgeries for each procedure in MSA





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 and Modeling (3/3)

Model Development and Training

Machine Learning Algorithms

- Linear Regression
- Decision Tree
- **XGBoost** ✓
- Random Forest
- Elastic Net

+

Decision Constraints

Imposed monotonic constraint while performing model training using Site Weight feature.

Model Evaluation and Tuning

Evaluation Metric

Used Mean Absolute Percentage Error (MAPE) as our Model Evaluation Metric.

Note: We have evaluated the MAPE on following two conditions:

- The overall MAPE on training and testing set.
- MAPE evaluated at procedure group level for training and testing datasets.

Model Selection

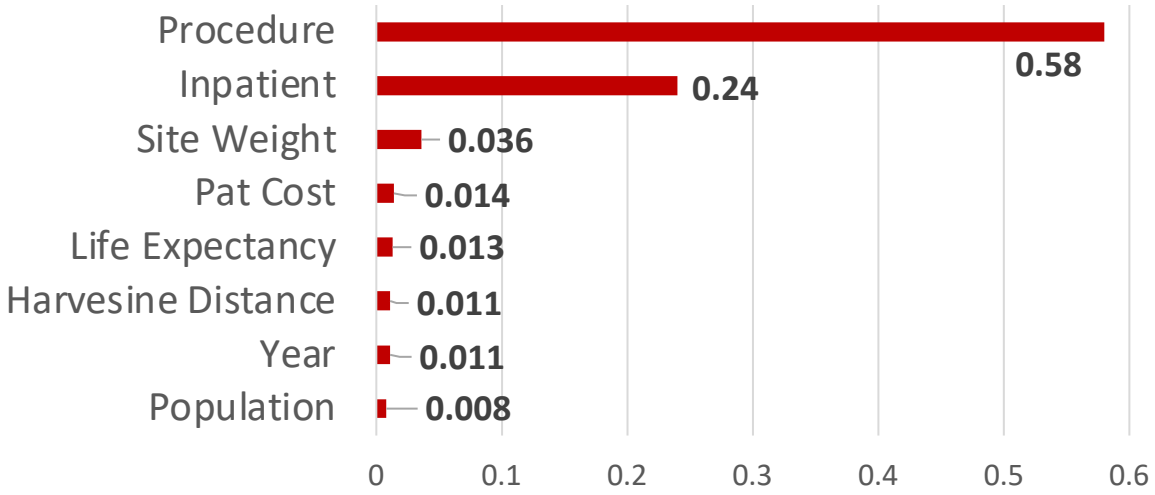
- Experimented with multiple data splits.
- Cross Validation
  - Grid Search
  - Random Search
- Bayesian Optimization for hyperparameter tuning.

Results

Performance of Models on Test Data

Models	MAPE*(Error)*
Linear Regression	30.86%
Elastic Net	28.94%
Decision Tree	24.04%
Random Forest	21.34%
XGBoost	**14.02%**

Top 8 Important Features



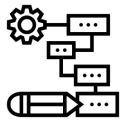
Performance of XGBoost vs DataRobot

Models	MAPE*(Error)*	Time
Data Robot	16.11%	~300 sec
XGBoost	**14.02%**	**~120 sec**

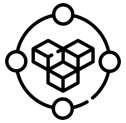
~15% improvement in MAPE (Error)

~2.5x reduction in Training Time

Business Impact



Highly robust model, eliminates the dependency on DataRobot leading to **saving licensing costs**.



Results and improved predictions from the model can results in **enhanced revenue from sales**.

End-to-end pipeline is **highly scalable** and can be **easily adapted** across multiple verticals within J&J.

Expected to **impact** J&J Data Scientists, medical devices pricing team, patients, and broader J&J group.

Future Scope

- **Exploring** the model **explainability** using **LIME** and **SHAP**.
- **Deployment** to the production and **monitor** for **drifts**.
- **Backtesting** the model using **iterative testing** as the test size is too small for analysis.
- Developing **optimal pricing** and **product penetration** strategies to **boost sales**.

Acknowledgement

We would like to thank our industry mentors Ziyu Tan, Cindy Tong, Kade Etter and our faculty advisor Adam Kelleher who offered this great opportunity to work on this interesting project.

We would also like to thank them of their guidance and valuable support throughout the project.