We would like to thank our industry mentors Cindy Tong, Kade Etter, and Ziyu Tan 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.

Data & Modeling

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

Data Preprocessing and Feature Engineering
- Initial Preprocessing
  - Performed Data Validation.
  - Generated Aggregated raw datasets at MSA and Procedure Level.
- Target Distribution
  - Target Variable (Mean Payout) Missing %
  - ~60%, ~40%
- Feature Engineering
  - Removed rows with missing target.
  - Used haversine distance to encode the geographical location.
  - Log transformed 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

Model Development and Training
- Machine Learning Algorithms
  - Linear Regression
  - Decision Tree
  - Random Forest
  - XGBoost
  - 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
- Top 8 Important Features
- Performance of XGBoost vs DataRobot

Business Impact

Highly robust model eliminates the dependency on DataRobot, saving license costs.

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

The 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 iterations as the test size is too small for analysis.

Developing optimal pricing and product penetration strategies to boost sales.