Progress Report 1 – 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)

October 21, 2022

1. Problem Definition and Progress Overview

Prediction of commercial insurance payments for surgical procedures is a project sponsored by Johnson & Johnson (J&J). It is necessary to quantify customer economics, including the revenue (payment amounts), to comprehend the value that JJMD devices produce. About 10% of Americans participating in employer-sponsored health insurance plans have claim-level data, including payment amounts, in the IBM MarketScan® Commercial Claims and Encounters (CCAE) database. The claim counts for surgeries within some metropolitan statistical areas (MSAs) are either too low (e.g., 50) or unavailable for inference, even though this data is an effective source for assessing national trends in surgical care for this population. Therefore, novel approaches are needed to fill up knowledge gaps in healthcare economics at the local level. Thus, in this project, we propose a novel machine-learning approach to address payment information gaps at the MSA level. Furthermore, we aim to improve the results given by the AutoML platform, DataRobot, which was previously used in this project.

According to our knowledge, no research was done to examine the potential of using machine learning (ML) to forecast the costs of JJMD surgeries at certain commercial facilities. Our study required a novel strategy to estimate procedure-specific commercial reimbursement inside MSAs where these data are lacking or insufficient. With the help of the AutoML platform DataRobot (DataRobot, Inc., Massachusetts, USA) [1] and a novel approach, this research combined the features from several datasets (Census, Medicare, and IBM) for prediction.

- Impact and value to the business: To comprehend the value of client compensation, which eliminates the need to buy expensive data sets.
- Application and Scalability: It is simple to add new surgical categories by expanding the input parameters and procedure codes. As required, models can also be used for distributed analysis.

2. Exploratory Data Analysis

A series of preprocessing steps were conducted on the datasets shown below used in this project. To build the final dataset, we combined data from four different sources:

- 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.
- 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.

All these data sources were eventually merged into one master data file that was further used for exploratory data analysis and model building. We show the various EDA done below.

Figure 1 displays the columns along the percentages of the missing value. 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.

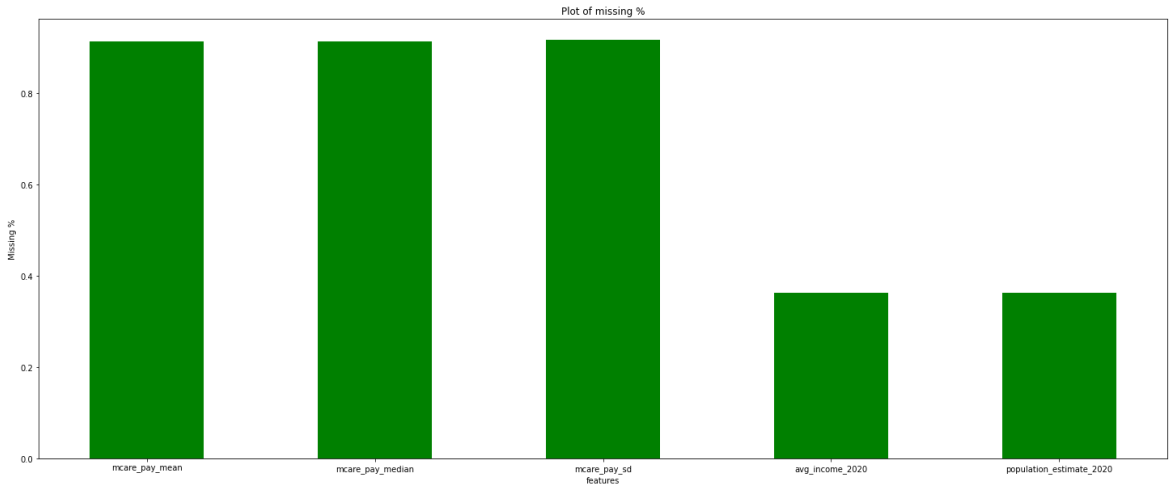

Fig 1. Missing value analysis of variables

The choropleths below (Figure 2 and Figure 3) show the distribution of hospitals and life expectancy, respectively, spread across all the states of the US. We see that Texas has the highest number of hospitals in the US while Wyoming has the least. Also, we observe that average life expectancy is higher in New Hampshire, Massachusetts and New York, while it is low in Georgia and Alabama. These analyses help us during the modelling as anticipating the insurance amount spent on surgeries will correlate with the average life expectancy and number of hospitals in regions.

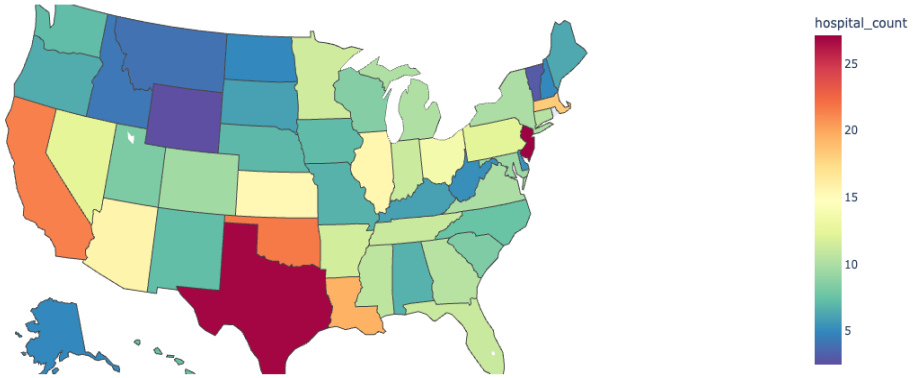

Fig 2. Choropleth plot of count of hospitals in the US

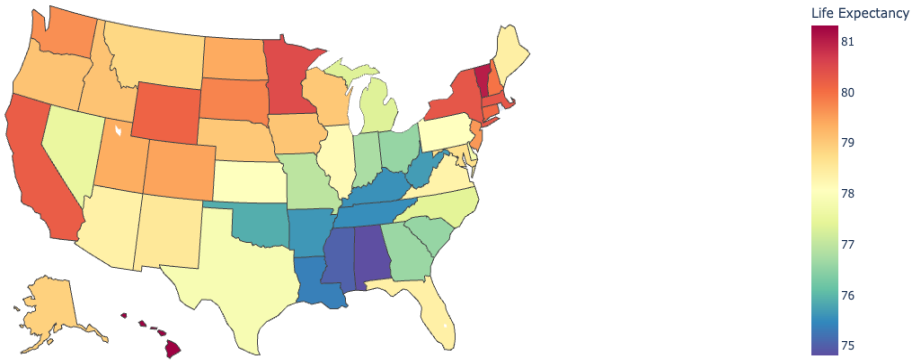

Fig 3. Choropleth plot of average life expectancy across the states of US

The plot below (Figure 4) shows us that our target variable is skewed to the right. It is affected by the mean, which is more towards the left. To make the model function better, we try to transform our target variable to be normally distributed. We tried two transformations, box-cox and log to establish this. We can see from the plot below (Figure 5) that our target variable has converged to a normal distribution by using the Box-Cox transformation. We also tried to convert our target variable to be normally distributed by doing a log transformation (Figure 6). But the Box-Cox transformation does a better job than the log transformation.

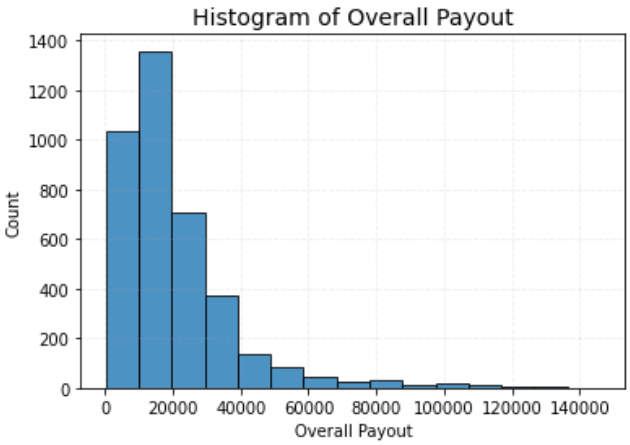

Fig 4. Histogram depicting the target variable – overall payout

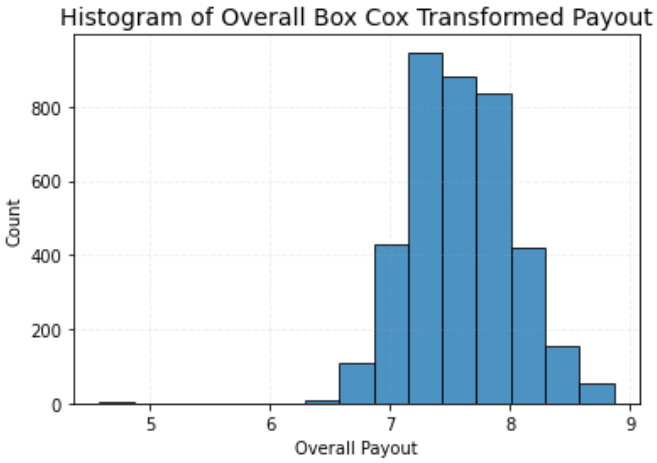

Fig 5. Box-Cox transformed histogram of Overall Payout

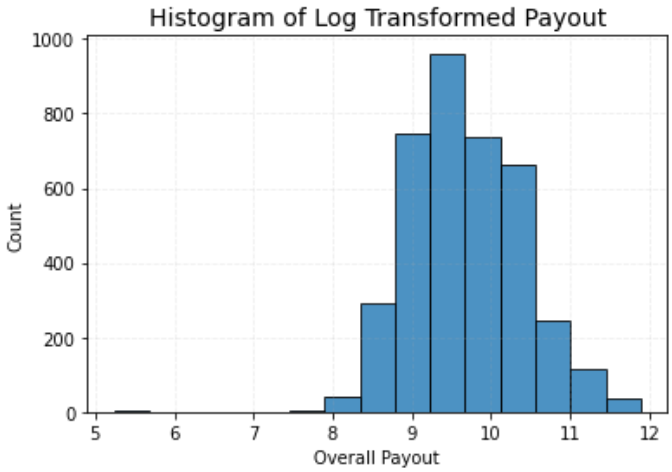

Fig 6. Log Transformed histogram of Overall Payout

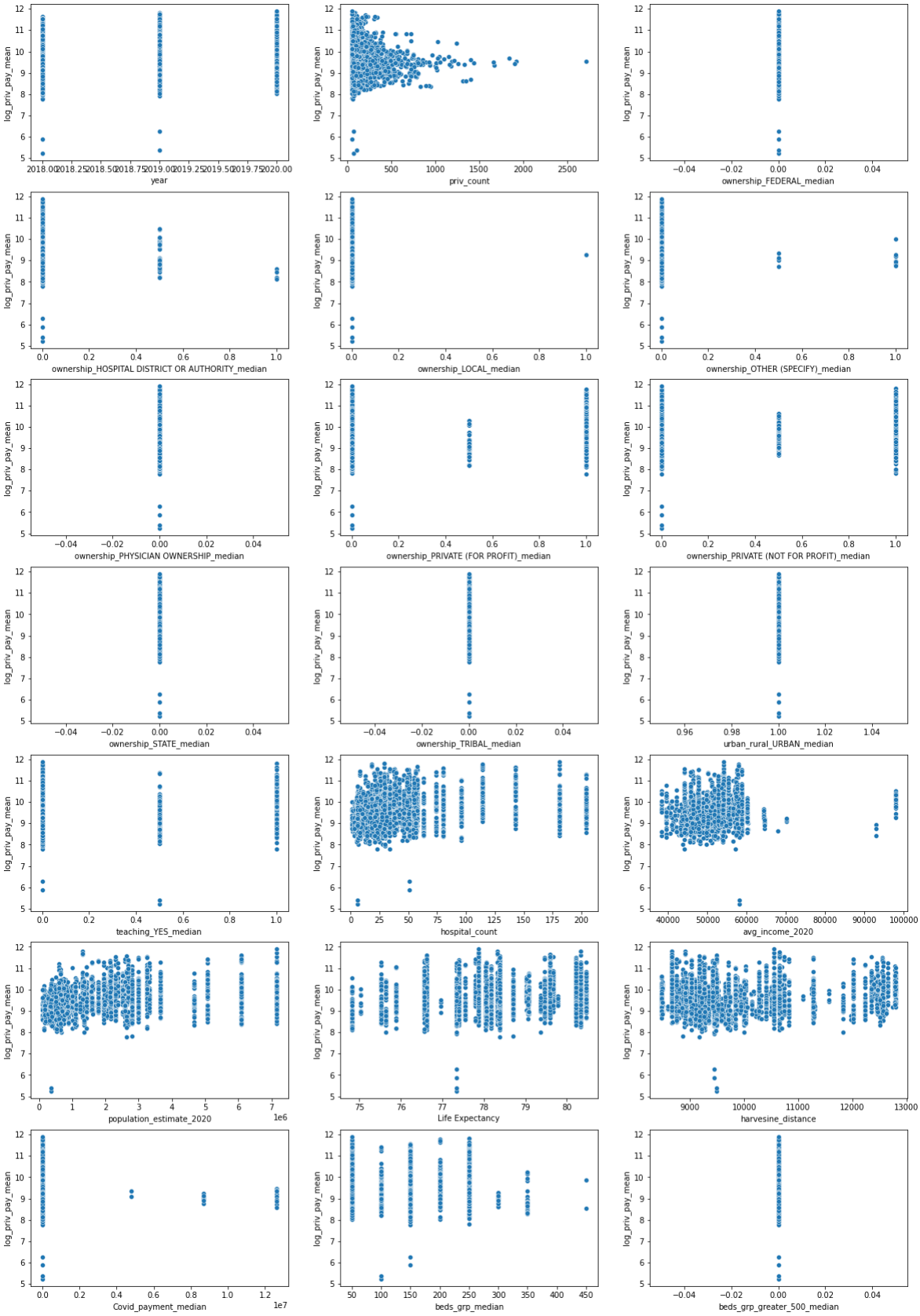

Fig 7. Scatter plots to understand the correlation between variables

The plot above (Figure 7) has 21 scatterplots in a single plot. Each scatterplot is plotted between a continuous variable and the log of the target variable. These plots help us to check and visualize the linear relationship between dependent and independent variables.

3. Methodology

3.1. Our Approach

After the initial exploratory data analysis was completed, we realised that most of our data set was empty, or in other words, it was sparse. To correctly predict our target variable 'priv_pay_median' with respect to a certain MSA, we needed to build our own features. We used data from various sources such as census, economic, average income, and life expectancy data to build these features. These data were combined with the initial dataset with respect to the MSA column, and later additional hospital data was also provided by Johnson & Johnson, which was also combined. Linear Regression was considered a baseline model, and a Decision Tree Regressor was also implemented. We also created a pipeline for the machine learning models to eradicate the repeated task of cleaning and encoding data.

3.2. Validation and Testing

- Linear Regression: Next step was to build a baseline model; we first started by building a Linear Regression model. The model returned an R2 of 0.804 on the validation set and an R2 of 0.837 on the test dataset. MAPE for the validation set was 2.7%, and the test set was also 2.6%. We looked at the coefficients of the Linear Regression model to determine the feature importance.
- Decision Trees: For our next model, we built a Decision Tree. We could achieve a MAPE of 1.59% on the validation set and a MAPE of 1.6% on the test set. The R2 score for the validation set was 0.926, and the R2 score for the test set was 0.935. We also looked at the feature importance of the decision tree model.

4. Goals and Next Steps

We are currently at about 40% of our project: we have combined hospital data with life expectancy and average income to create a master dataset that contains MSA-level data for hospital surgery expenditures and corroborates it with life expectancy and income in that MSA. Also, please note the current baseline is a work in progress, as the team received some data updates this week. We will incorporate changes during the next phase and might need to update the baseline model if required.

- Testing and Validation: As per the discussion with the team, we need to create a custom testing and validation pipeline since we have only ~4K observations in the training dataset. Thus, to fully utilize the dataset, we need to have data split as 90% training, 5% testing and 5 % as validation. The team must repeat the entire process at least 10 times (to be tested) to calculate a better estimate of the error – MAPE/R2.

- Unsupervised Learning: The J&J team will provide additional data about the procedure-level features such as costs, the severity of the procedure, etc. We will try to explore the unsupervised clustering approaches or might use the assigned cluster numbers as a variable in the supervised model in the next phase.

- Check for ASC payout amount – After the finalized testing pipeline, we have to deep-dive into cases where the ASC procedure type has higher payouts than the in-patient and out-patient payouts. Also, we can create a binary ordinal variable for these procedure types or use the capping method to validate the results.

- Dealing with Missing data: Since we have features that have a lot of missing data. We can try to impute NA values using advanced techniques like autoencoders.

- Advanced Modelling Techniques: In the next phase, we will test out the advanced modelling techniques not limited to:

    - Tree-based – Boosting and Bagging
    - Regularized Training – Ridge, Lasso and Elasticnet
    - Hyperparameter Tuning – Random, Grid Search and Bayesian.
    - Neural Networks – MLP

5. Contributions

- Mahesh Jindal: Contributed to data preprocessing and combining raw data from different sources. Built the master data file that is being used for further model building.

- Prerit Jain: Contributed to EDA, the validation and testing of the built model in this report.

- Parth Gupta: Contributed to EDA, building the baseline machine learning model explained in this report.

- Ayush Baral: Contributed in EDA, feature selection and building of new features for the master dataset.

- Rahulraj Singh (Team Captain): Set up meetings and milestones, and manage progress. Contributed to collecting and preparing life expectancy and income data.

6. References

[1] DataRobot. [n. d.]. Data Robot: Automated Machine Learning for Predictive Modeling. Retrieved 05-Aug-2021 from https://datarobot.com