Data Analysis

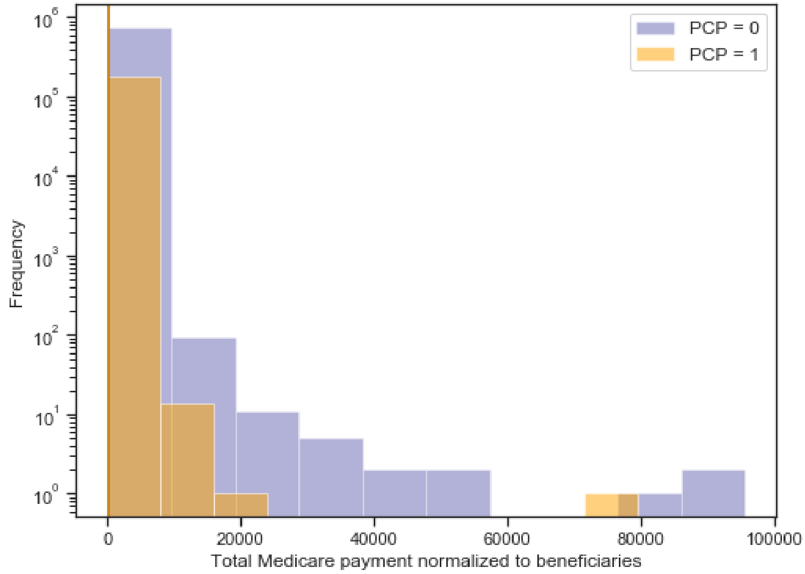
To understand the data and predict the class types describing provider types (classified by pcp) effectively, I first created the target feature 'pcp' and added it to the data before performing further analyses on it. The goal of the project is to predict the class type of providers (pcp = 0 or 1) from the existing information from the Medicare data released by CMS in 2014. The target variable pcp is 1 for providers that are either in Internal Medicine or Family Practice and is 0 for the rest of the provider types.

The data suggests that there are differences in the amount of payments and number of services provided by either classes of providers. Class 1 providers tend to older patients with health risks such as Alzheimer's whereas Class 0 providers have patients of more age groups (incl. very young patients) with slightly higher health risks pertaining to specific predispositions such as cancer, arthritis, and depression. In the following pages, I have explained these findings and how I have run my analyses and predictive models. Each step is also marked down in my code for better understanding.

From preliminary EDA:

1) Larger distribution of payments with higher values by Medicare to Class 0 providers

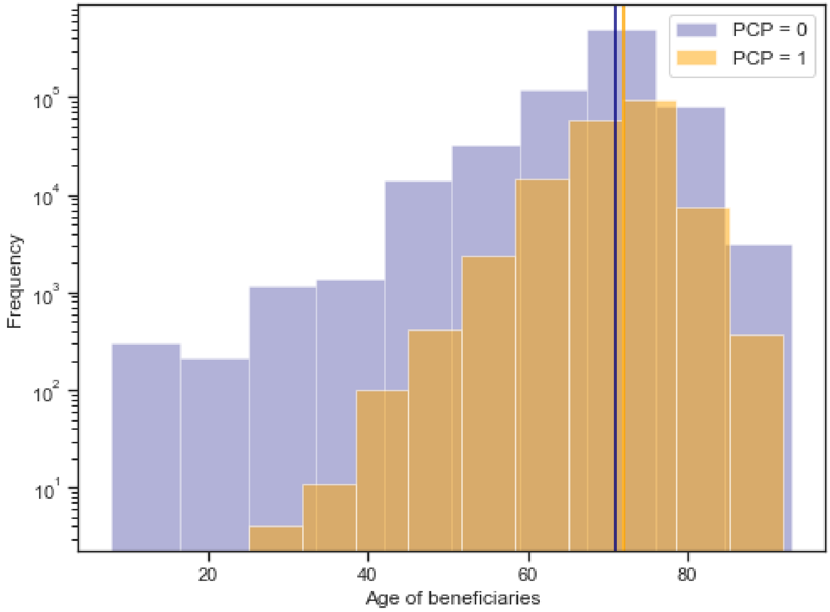
Some key differences in either classes, was that the payments ranged to higher values for providers in Class 0 (not Internal Medicine and Family Practice). Payments were more distributed for Class 0 while family practice or internal medicine having comparably lower normalized payments. From the data description, I found that the 25th and 50th percentile of Medicare payment is higher for Class 1. But the payments in the 75th percentile is higher for the other specialties i.e. Class 0 (as seen in the description table). Please note that I have normalized total payments and total services by the unique beneficiaries of the provider.



2) Class 1 providers had older beneficiaries

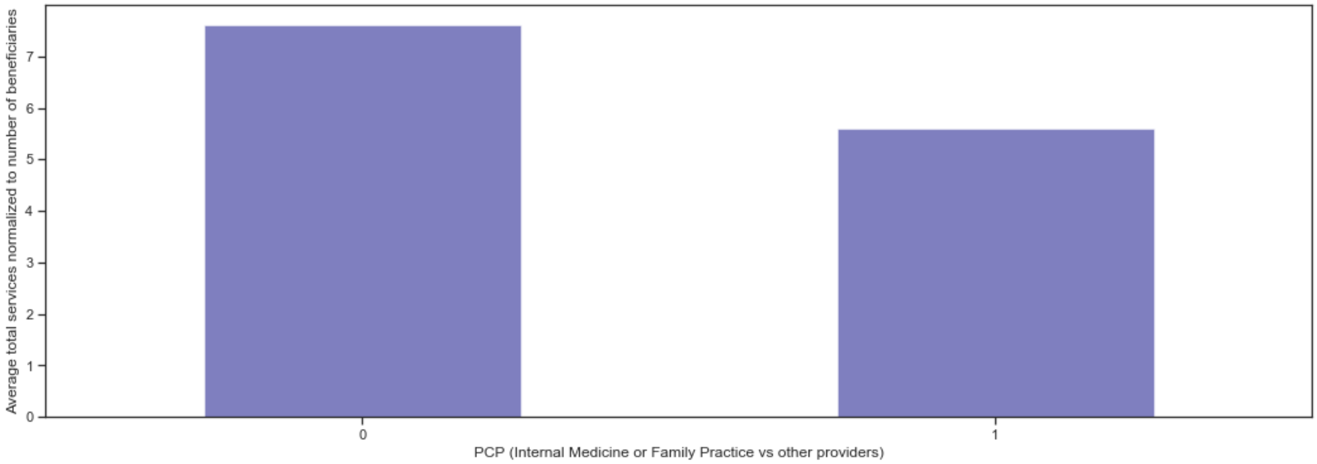
Class 0 providers had a higher frequency of beneficiaries belonging to younger age groups. This is interesting because Internal medicine is mostly for adults, but family practice is mostly for entire families (all age groups). But I found the range of age for Class 1 to be >30 which is a bit. Perhaps the average age in the family is skewed to a higher number because of older parents.

Exploring the age group of the beneficiaries a bit further, I also found that there are higher numbers of beneficiaries in the lower age groups for Class 0. This implies very young patients in that class (severe ailments requiring specialized treatments i.e. Class 0).



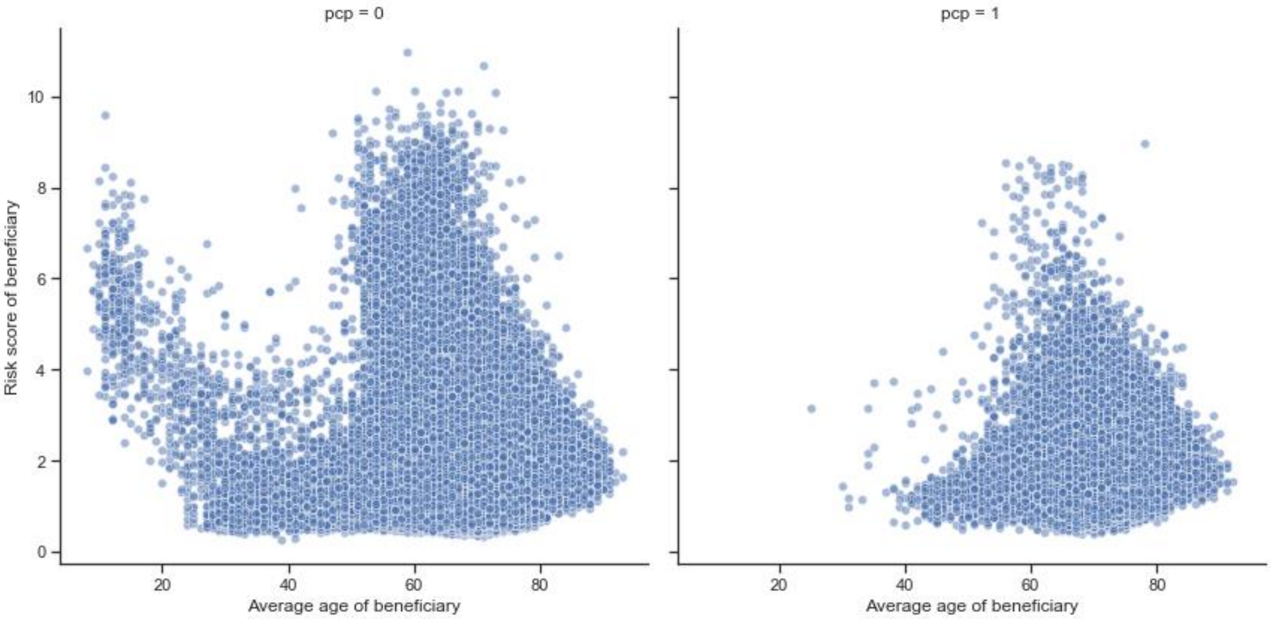
3) Class 0 providers offer larger number of services normalized to their unique beneficiaries

On average, the services rendered by the providers to each beneficiary is higher for Class 0. This is important because the volume of the services and the number of hcpcs codes of the providers may contain signal. But I found this piece of information in the document, "*the volume of procedures presented may not be fully inclusive of all procedures performed by the provider. Medicare allowed amounts and Medicare payments for a given HCPCS code/place of service can vary based on a number of factors, including modifiers, geography, and other services performed during the same day/visit*" (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf). However, for the scope of this project, I did not remove this feature from the model.



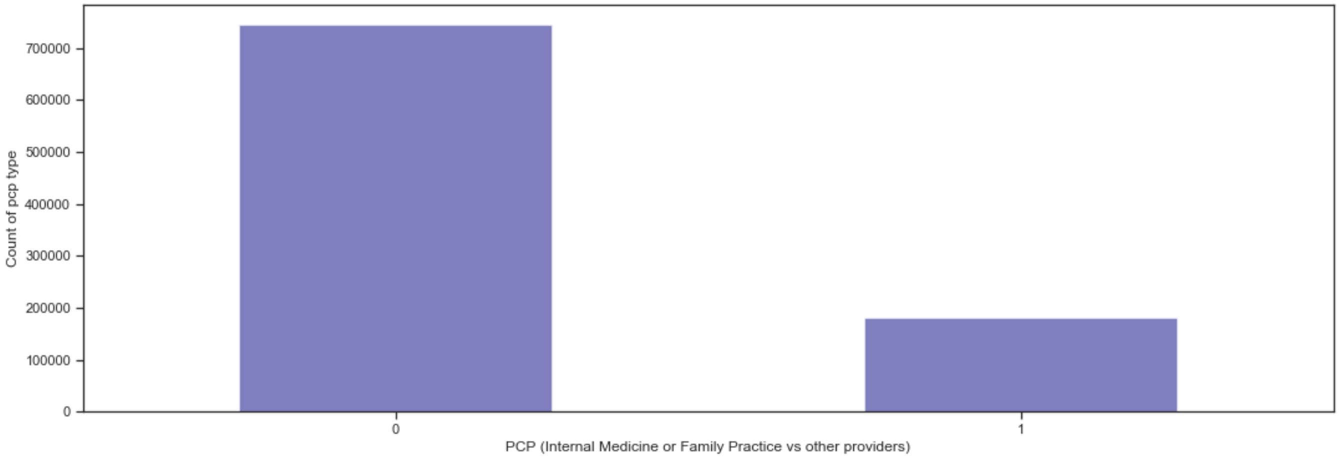
4) Average risk scores were bimodally distributed across age of beneficiary of Class 0 providers

The overall risk score of the beneficiaries were higher of Class 0 providers. Because of more specialized services in Class 0, it makes sense to find a higher average risk score in their beneficiaries. However, Internal medicine and Family practice providers have most beneficiaries between ages 40-80 with a large (0 – 8) span of risk score (average was at 72). But Class 0 providers have additional beneficiaries between ages 8-40 with a larger (0 – 10) span of risk. The severity of the risk of the patient must affect the type of care they need and thus the provider type.

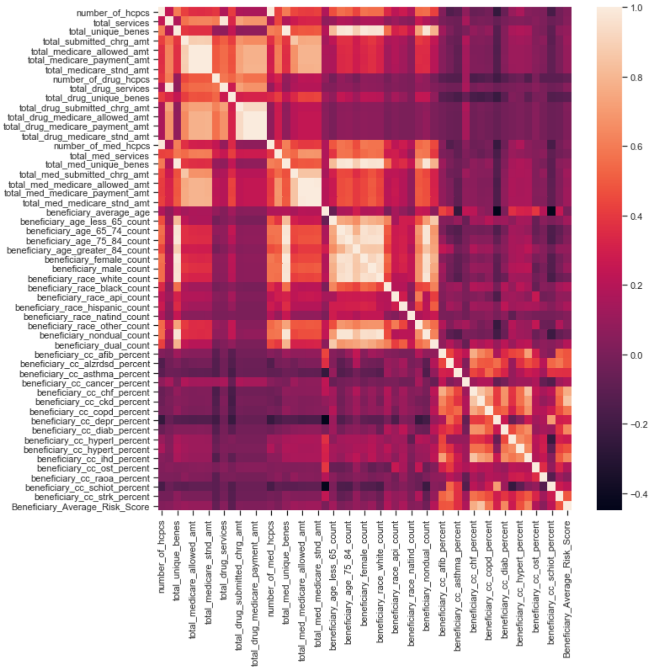


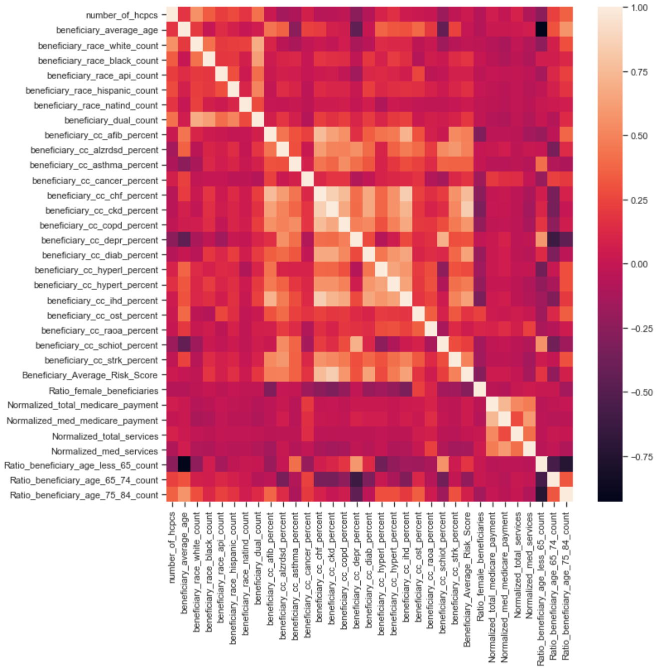
Feature selection, engineering, and balancing the classes

The variability in the total medical services and payment amounts for both classes were high (significantly high standard deviations comparable to mean values). The classes were imbalanced with different numbers of beneficiaries and providers for both classes. To handle the differences in the beneficiaries, I normalized the features (eg., total services or payment) to the number of unique beneficiaries of each provider. This controlled for the differences in the popularity of some specialties over some others. The graph below shows the class imbalance.



Upon preliminary exploration, I found some features to have very high correlations. There were 4 major groups with features that were highly correlated with one another. I chose to keep only one representative feature that were based on if they had the least missing values, or that what made the most intuitive sense to keep (e.g., total medicare payment instead of submitted charge amount or total_medicare_allowed_amt, etc.). Upon doing so I reduced correlations amongst features. Please check left (high correlations) vs. right (lower correlations) after feature selection and engineering. This is important to obtain a good machine learning model.





After feature selection, the data was ready to run through model algorithms. But before that, to handle the imbalances in the data, I down sampled the majority class (Class 0) so that it was 2x the minority. I did not perfectly balance the data in the training set because logistic regressions are more sensitive to drastic imbalance and this creates a more realistic training set. I measured the effectiveness of balancing the data by running a logistic regression on both balanced and imbalanced data.

Evaluating different machine learning models

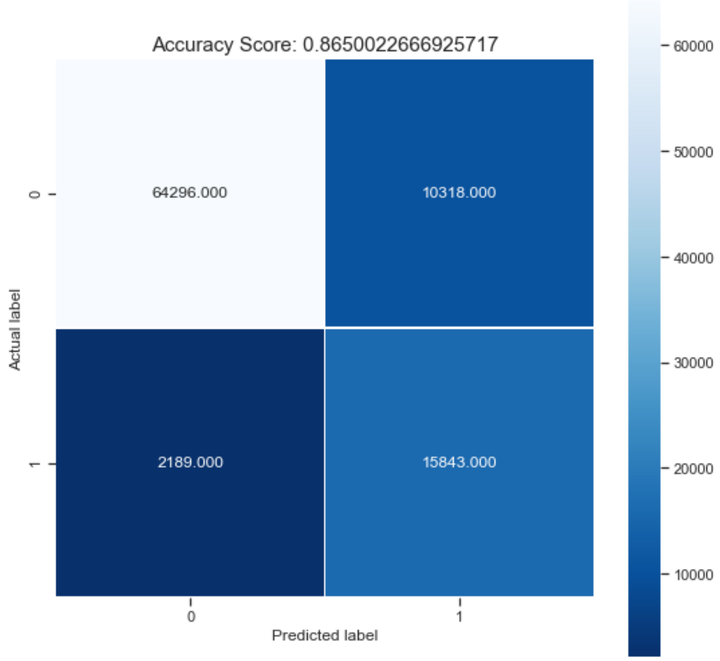
Because this project was a classification problem, I compared some typical classification models starting with logistic regression followed by a random forest classifier and a LightGBM model (Microsoft, 2017). I used LightGBM instead of its predecessor the XGBoost that uses a gradient boosting algorithm too because it used a method (gradient one-side sampling) to extract the most information from data very fast (https://medium.com/kaggle-nyc/gradient-boosting-decision-trees-xgboost-vs-lightgbm-and-catboost-72df6979e0bb). Light GBM splits the tree leaf-wise with the best fit instead of depth-wise or level-wise. The leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy.

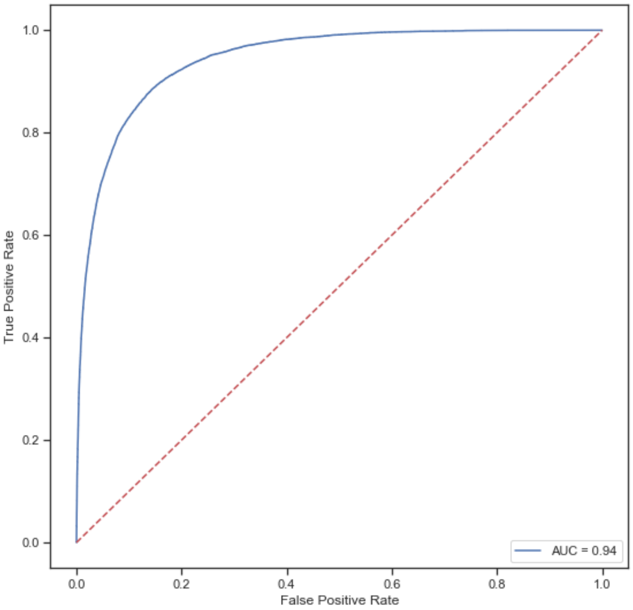
Performance metrics: Accuracy vs precision/recall?

In the logistic regression, balancing the data improved recall from 0.08 to 0.41 but suffered slightly in accuracy. Whether accuracy or precision/recall is more important to evaluate a model depends completely on what the model is going to be used for or its context. While recall expresses the ability to find the right class, precision expresses the proportion of the data points this model says belongs to one class vs the other. There is a trade-off that one needs to reach. For the scope of simply predicting the class type, having good accuracy (higher than baseline) and recall were reliable measures to evaluate the effectiveness of these models.

Light GBM classifier

After having compared all the machine-learning models, I selected Light GBM to evaluate the main summary points of the data. After performing a Grid Search to find good hyperparameters and 3 fold cross-validation, I found that the Light GBM model had an accuracy of 0.87, precision of 0.8, and recall of 0.88 which was better than the other models I used. Its auc was at a 0.94.





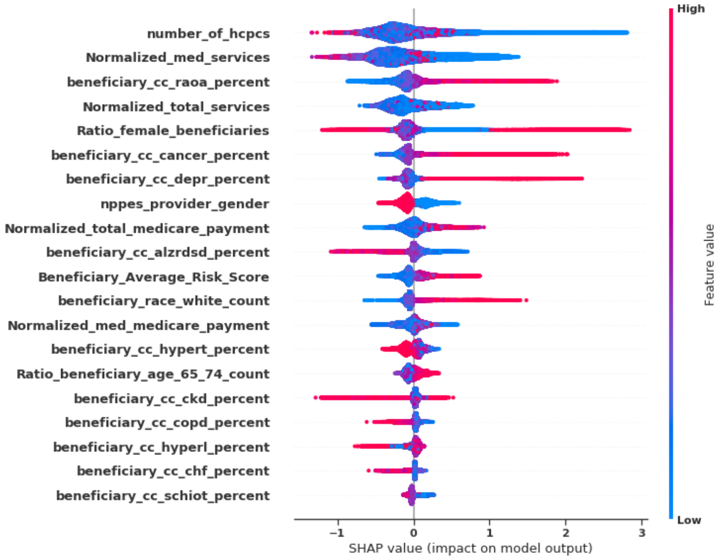
Key findings from the model

- Low number of hcpcs codes were positively correlated with Class 0 providers. This is interesting because it implies that Class 0 providers had a lower volume of medical service (they are specific). The opposite applies to Class 1 providers.
- Additionally, high values of specialized health risks like rheumatoid arthritis, cancer, depression in beneficiaries had a positive correlation with Class 0 providers and negative with Class 1. This makes sense because Internal Medicine and Family Practice could be broad fields.
- However, beneficiaries of Class 1 providers had an increased predisposition to Alzheimer's. Internal medicine or family practice could tend towards older patients who may stand higher odds of having dementia.
- Consistent with prior EDA, normalized Medicare payments were higher for Class 0 health providers. And higher risk score in beneficiaries were found to be positively correlated with

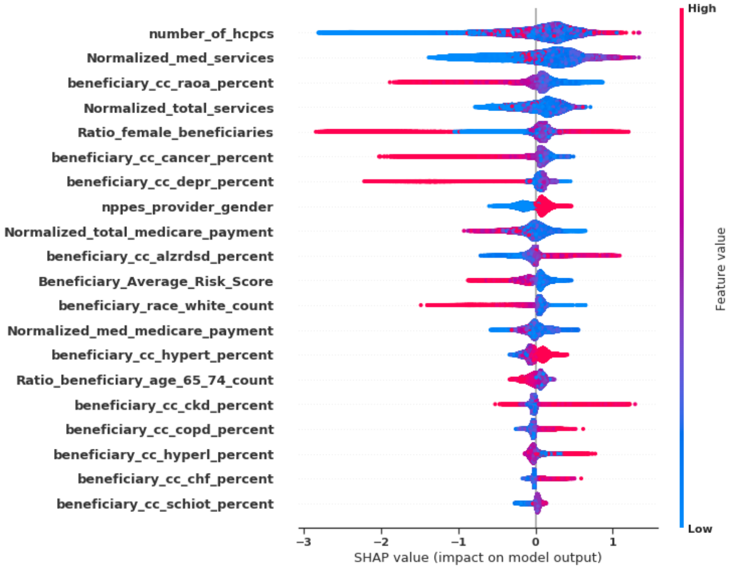
Class 0. Class 0 providers also provided lower normalized med services compared to Class 1 providers.

- Lastly, a surprising finding was that Class 1 providers were generally males (check feature ‘provider_gender’) and vice versa for Class 0.

Class 0



Class 1


number_of_hcpcs
Normalized_med_services
beneficiary_cc_raoa_percent
Normalized_total_services
Ratio_female_beneficiaries
beneficiary_cc_cancer_percent
beneficiary_cc_depr_percent
nppes_provider_gender
Normalized_total_medicare_payment
beneficiary_cc_alzrdsd_percent
Beneficiary_Average_Risk_Score
beneficiary_race_white_count
Normalized_med_medicare_payment
beneficiary_cc_hypert_percent
Ratio_beneficiary_age_65_74_count
beneficiary_cc_ckd_percent
beneficiary_cc_copd_percent
beneficiary_cc_hyperl_percent
beneficiary_cc_chf_percent
beneficiary_cc_schiot_percent
SHAP value (impact on model output)
High
Low
Feature value