**Predicting Cost of Medicare Beneficiaries**

**Project Report**

**Group 3**

Sonal Patel, Wasim Meghani, Manoj Mareedu, Ronit Singhal, Sai Saranya Chaganti

**Table of Contents**

[Executive Summary: 3](#_Toc153143716)

[Project Motivation/Background: 4](#_Toc153143717)

[Data Description: 5](#_Toc153143718)

[Preprocessing: 7](#_Toc153143719)

[Visualizations: 10](#_Toc153143720)

[Models: 12](#_Toc153143721)

[Multiple Linear Regression Model 12](#_Toc153143722)

[KNN Model 13](#_Toc153143723)

[Decision Tree Model 15](#_Toc153143724)

[Random Forest Model 17](#_Toc153143725)

[Conclusion: 19](#_Toc153143726)

[References: 21](#_Toc153143727)

# Executive Summary:

For our project, we will create a predictive model to estimate the expected level of costs associated with an individual beneficiary's healthcare utilization. Predicting healthcare costs for Medicare beneficiaries provides many potential benefits in optimizing resource allocation and improving the overall efficiency of the healthcare system. This can also provide valuable insight for healthcare organizations in future healthcare planning and decision-making processes. By leveraging predictive models and data analytics, healthcare providers can gain valuable insight into potential future cost implications and guide organizations in resource allocation, risk assessment, and cost management.

# Project Motivation/Background:

In today’s healthcare landscape, predicting the healthcare costs of Medicare beneficiaries can be valuable to healthcare businesses and hospitals as it contributes to a more efficient, patient-centered, and cost-effective healthcare system. As the industry shifts towards value-based care and patient-centric models, understanding and estimating healthcare expenditures are pivotal in optimizing care delivery, financial planning, and resource allocation within healthcare systems. Predictive models drive resource optimization in healthcare systems. They can plan efficient staffing, supply-chain management, and service planning by anticipating high-cost patients. Further, the ability to anticipate healthcare costs fosters proactive patient engagement and collaborative care planning. Providers can leverage insight into the potential expenses to develop tailored and cost-effective treatment strategies, improving overall health outcomes. Leveraging robust data analytics and machine learning, predictive models allow healthcare organizations the insight needed to facilitate evidence-based decision-making that aligns with their strategic goals to reduce costs while optimizing care in order to improve health outcomes.

# Data Description:

The data used in this project can be classified as Medicare claims data. It was acquired from the Centers for Medicare and Medicaid Services (CMS). The dataset is titled “Synthetic Medicare Enrollment, Fee-For-Service Claims and Prescription Drug Events.” The data contains both Inpatient and Outpatient Medicare Claims ranging from 2015-2022. The dataset contains a total of 552,139 records, of which 494,073 records are outpatient claims, and 58,066 records are inpatient claims. Each row in the dataset represents a beneficiary and his total charged amount for the corresponding medical condition treatment. For our dataset, we selected various attributes from the original dataset, which include Beneficiary ID, Claim ID, Provider Number (also identified state of the provider), Claim Total Charge Amount, Claim Principal Diagnosis Code, CLM\_E\_POA\_IND\_SW1 (whether the diagnosis code was present at the time of arrival to the hospital) which has ‘Y’ and ‘U’ as its values, PRVDR\_STATE\_CD which is a unique code for each State. The target variable we seek to predict is the estimated total healthcare cost for each beneficiary. Predictor variables include principal diagnosis code, and number of claims for each beneficiary.

Before we created the models, it was necessary to pre-process the data. We selected the required features from both inpatient and outpatient datasets. We created and considered another dataset which has PRVDR\_STATE\_CD, State, Meadian\_Income (for corresponding State). The median income column gives the median annual income in that respective state. The following are the datasets we considered for our analysis: The BENE\_ID, CLAIM\_ID, CLAIM\_TOTAL\_CHARGE, Median\_Income are numerical columns, PRVDR\_NUM, PRNCPAL\_DGNS\_CD columns are alphanumeric. State variable is categorical, CLM\_E\_POA\_IND\_SW1 is binary categorical column. The ‘Y’ values for CLM\_E\_POA\_IND\_SW1 column denote that the diagnosis code was present at the time of arrival to the hospital. ‘U’ value denotes that the documentation is insufficient to state that the diagnosis code was present at the time of arrival to the hospital.

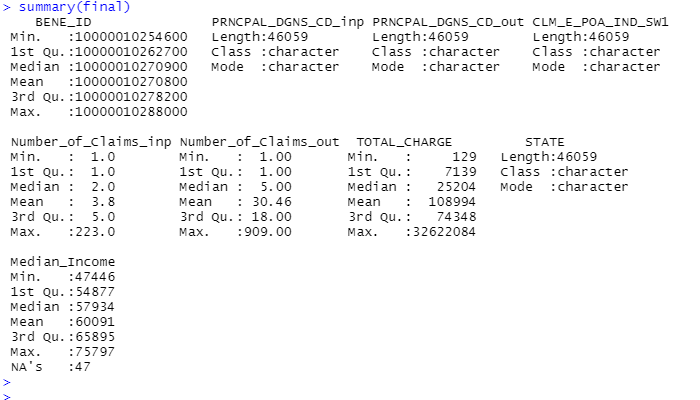
# Preprocessing:

In our data analysis project, we embarked on a comprehensive preprocessing journey to amalgamate information dispersed across various sources and create a cohesive dataset for subsequent analysis. Our primary goal revolves around predicting the anticipated healthcare utilization costs for individual beneficiaries, leveraging features from both the CMS synthetic and RIF datasets. To achieve this, we applied data mining techniques, specifically focusing on multiple linear regression analysis and classification models, including KNN, Decision Tree, and Random Forest. These methods fall within the broader scope of multivariate analysis and supervised learning.

Our initial step involved reading the inpatient and outpatient records from 'inpatient.csv' and 'outpatient.csv' files. We meticulously selected pertinent columns from both datasets and engineered two additional predictors: the number of inpatient claims and outpatient claims for each beneficiary and principal diagnosis code group. Using the 'groupby' function and aggregation with '.sum()', we computed the total charges for inpatient and outpatient services for each beneficiary and principal diagnosis code group. The next crucial step was horizontally merging these aggregated datasets into a unified dataframe, which we aptly named 'merged\_data.' Subsequently, we conducted a thorough check for duplicates and removed them. The consolidated and cleaned dataset was then saved as 'Patient\_Claim\_Data.xlsx.'

In R, we imported the 'Patient\_Claim\_Data.xlsx' and introduced an external dataset named 'Median\_Income.xlsx,' which captured median income information. The two datasets were merged using R programming, ensuring a seamless integration of median income details into our final dataset. Rigorous data cleaning procedures followed, encompassing the removal of duplicate entries and handling a minimal number of missing values (47 NA’s) using na.omit() function.

A pivotal variable, 'TOTAL\_CHARGE,' was then introduced, encapsulating the combined inpatient and outpatient charge values for each beneficiary. Given the extensive range of unique principal diagnosis codes for both inpatient and outpatient services, we opted for frequency encoding for these codes, effectively replacing them with their respective frequencies in the dataset. 'fastDummies' was employed to generate dummy variables for the 'CLM\_E\_POA\_IND\_SW1' column. Our final flat file has 46059 records in total.



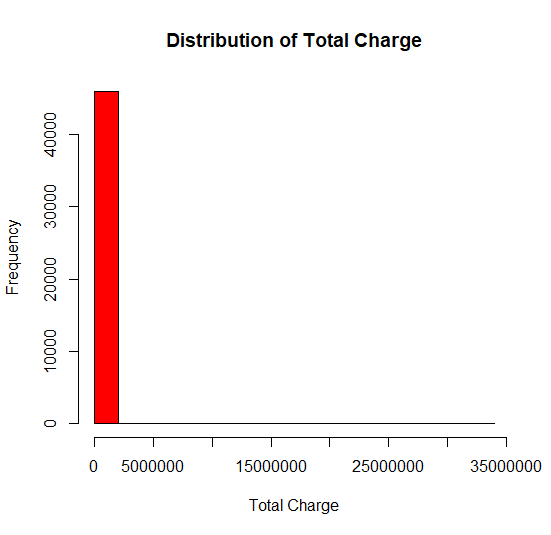
*Fig.1. Statistical summary of the final flat file before creating dummy variables.*

For our classification models, we categorized the 'TOTAL\_CHARGE' into five distinct classes to create a new target variable labeled 'TC\_Class.' These classes were defined based on the distribution of total charge values: very low (<$1,000), low (>$1,000, <$10,000), moderate (>$10,000, <$100,000), high (>$100,000, <$1,000,000), and very high (>1,000,000). The decision to employ Decision Tree and Random Forest models stemmed from the dataset's inherent presence of outliers, as these models exhibit robustness against outlier impact. The training and test set split for the classification model adhered to a 70:30 ratio.

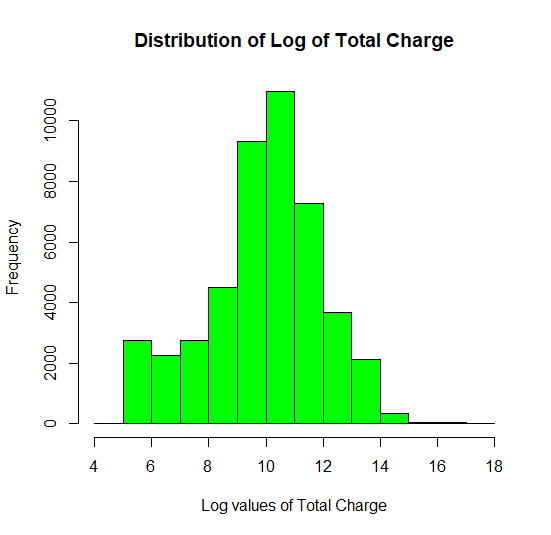
In summary, our preprocessing efforts laid a solid foundation for subsequent analysis, ensuring a refined and comprehensive dataset for the prediction of healthcare utilization costs.

## Visualizations:

Visualized distribution of the log-transformed TOTAL\_CHARGE variable is shown by the histogram. Confusion matrix Provides a summary of the true positive, true negative, false positive, and false negative prediction counts for classification models. The decision tree model is structured.

****

*Fig. 2. Distribution before applying Logarithm.*

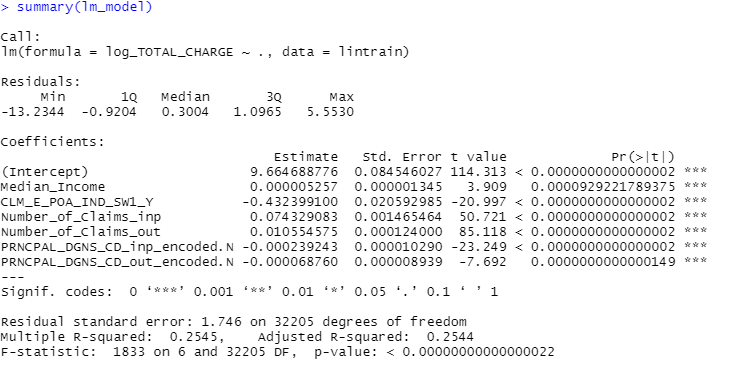
****

*Fig. 3. Distribution after applying Logarithm.*

# Models:

## Multiple Linear Regression Model

After partitioning the data into train and test datasets, we built a multiple linear regression model with log\_TOTAL\_CHARGE as target variable. The model’s performance is not satisfactory. The R-Squared value is 0.2545 which is low, and the RMSE Value is 1.7 which is high. From the Linear regression model, we observed that all the predictors are statistically significant. Poor performance of the model can be due to lack of sufficient predictors (unobserved factors).

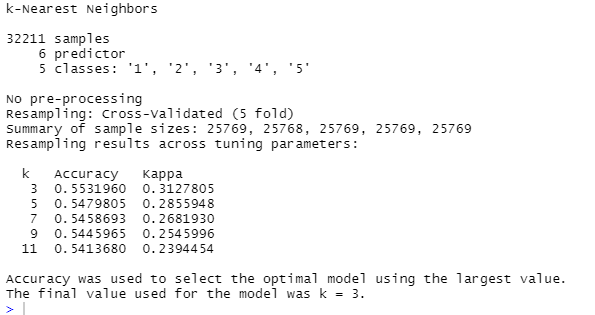


*Fig. 4. Linear regression Model Summary*

For our objective we can further preprocess data and classify the target variable ‘TOTAL\_CHARGE’ into 5 classes, which are very low, low, moderate, high, very high. Through this process we can use classification models to predict which class of total charge a particular beneficiary fall into.

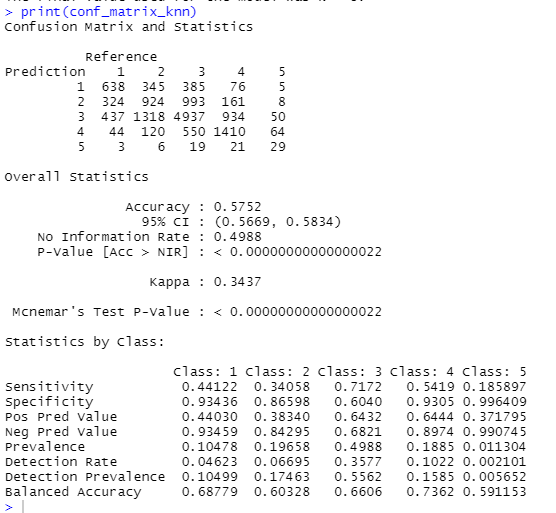
## KNN Model

We have built a KNN model with grid search (to find an optimal k-value which is the number of nearest neighbors). We used 5-fold cross validation. The optimal k - value chosen is k = 3. The KNN model achieved an accuracy of 57.52%, demonstrating its ability to classify total charge classes.



*Fig. 5. KNN Model summary*

The model is particularly sensitive in identifying the very low and low total charge classes, while showing lower sensitivity for higher classes. Despite the imbalanced nature of the dataset, the Kappa statistic of 0.3437 suggests fair agreement beyond chance in the model's predictions. Area Under the Curve (AUC) value for KNN model is 0.74458.

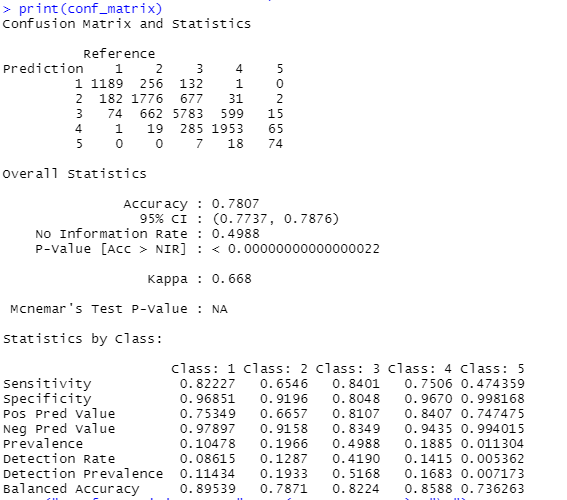


*Fig. 6. Confusion Matrix for KNN Model*

The low accuracy for KNN models is due to extreme outliers. We included the outliers without removing preprocessing because we choose to consider extremely high total charge value classes as well in our analysis. To overcome the outlier limitation, we chose a decision tree model.

## Decision Tree Model

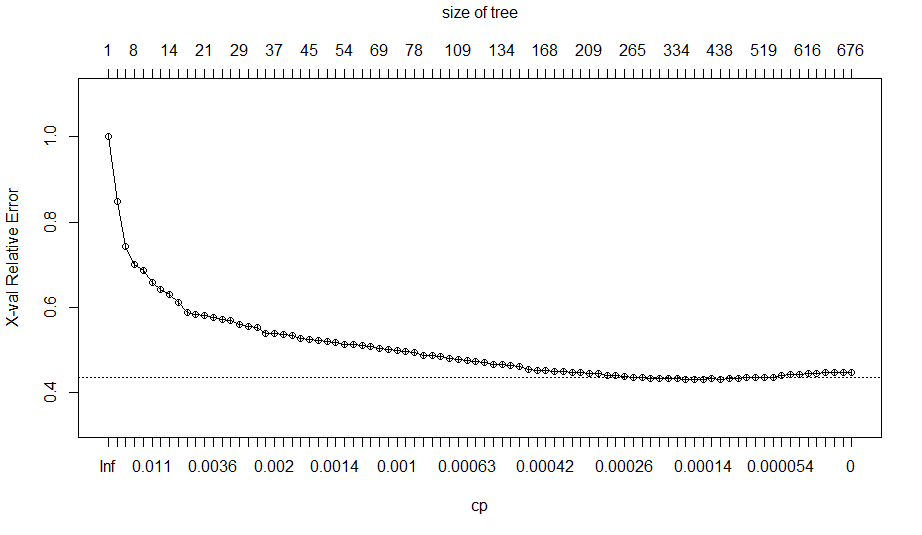
To overcome the outlier issue we built a decision tree using R. Based on the confusion matrix, the accuracy for the model was calculated to be 78.07%. The AUC was found to be 0.9397 (or 94%). This indicated that the model had a good performance, strong ability to distinguish between the five classes and at predicting the class of total charge for a new beneficiary. The model exhibits strong sensitivity in identifying classes 1, 3, and 4, indicating its ability to accurately predict these categories. However, the lower sensitivity in class 5 suggests room for improvement, while high specificity across all classes highlights the model's proficiency in avoiding false positives. The Kappa statistic of 0.668 indicates substantial agreement beyond chance, emphasizing the overall reliability of the decision tree model.



*Fig. 7. Confusion Matrix for Decision Tree Model*

The complexity parameter is used to control the size of the decision tree by penalizing the model for adding too many branches. It is a regularization parameter that helps prevent overfitting, ensuring that the tree does not become too complex and specific to the training data, which might not generalize well to new, unseen data.

Complexity Parameter:

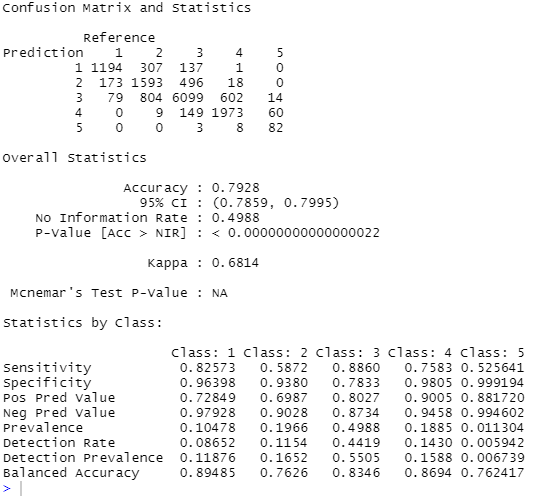


*Fig. 8. Complexity parameter plot*

The above cp plot shows the overall prediction error, for different values of the complexity parameter.

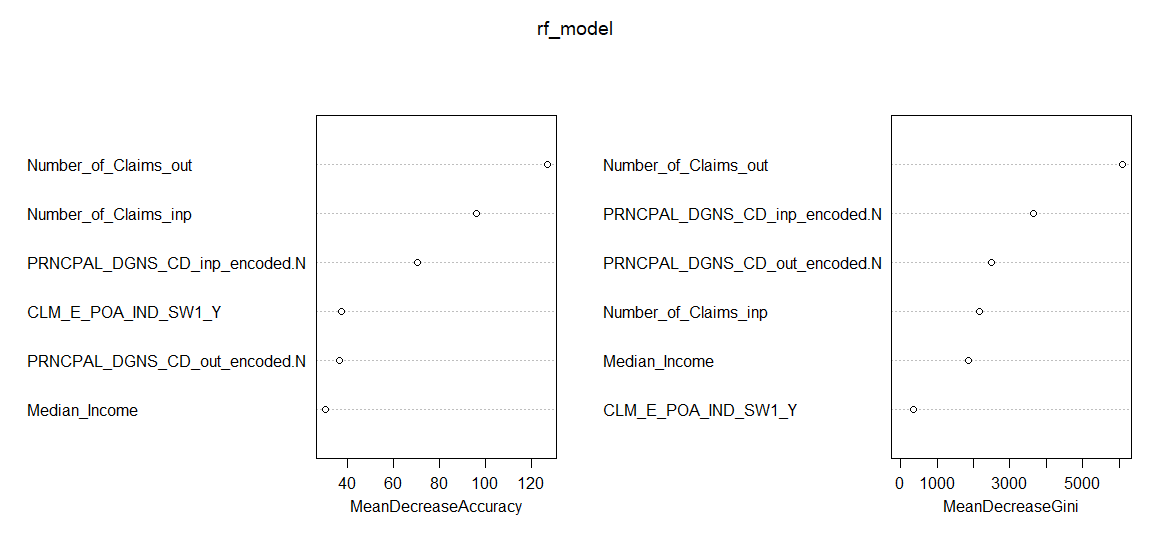
## Random Forest Model

Finally, to consider a more robust model which is not sensitive to outliers, we considered a random forest model. It provides the result by aggregating multiple decision trees. We considered 100 estimators (100 decision trees). The confusion matrix shows that the model achieved an overall accuracy of 79.28%, indicating a good performance in correctly predicting the classes. Notably, the model demonstrates high sensitivity (true positive rate) for class 3, suggesting effective identification of higher expense levels. However, lower specificity for class 2 indicates a tendency for false positives in predicting moderate expense levels. The balanced accuracy reflects a generally well-balanced trade-off between sensitivity and specificity across all classes. The Area under the curve (AUC) value for the random forest model is 0.9398.



*Fig. 9. Confusion Matrix for Random Forest Model*

Using varImpPlot() function we observed the importance of all the predictors in the model. The variable "Number\_of\_Claims\_out" exhibits the highest importance, contributing significantly to both accuracy and Gini index in the Random Forest model. Additionally, "PRNCPAL\_DGNS\_CD\_inp\_encoded.N" and "PRNCPAL\_DGNS\_CD\_out\_encoded.N" are meaningful predictors, emphasizing the importance of frequency-encoded principal diagnosis codes in predicting total charge classes.



*Fig. 10. Variable importance plot.*

We used the ‘randomForest’ library for building random forest models. The ensemble approach of random forest can be beneficial while predicting on new data. Considering all the metrics, random forest is the best model chosen.

# Conclusion:

The comparison of KNN, Decision Tree, and Random Forest models reveals that the Random Forest model outperforms the others with the highest accuracy (79.28%) and a Kappa value of 0.6814, indicating substantial agreement. The Random Forest model demonstrates superior sensitivity, and specificity across multiple classes, resulting in a well-balanced performance. While the Decision Tree model performs well, the Random Forest's ensemble approach provides enhanced predictive capabilities for the complex nature of the dataset, as evident from the high AUC value of 0.93. Additionally, the Random Forest model shows superior balance in recall, specificity, and precision, making it the preferred choice for predicting total charge classes in healthcare utilization.

These predictive models can be greatly beneficial to both hospitals and insurance firms such as CMS, allowing them to accurately predict the class of total charge of a new beneficiary based on factors such as disease/diagnosis code, median income, and number of previous inpatient and outpatient claims. Using this information, organizations can anticipate the range of total charges for a new beneficiary and allocate resources in a more efficient manner, thus, contributing to better financial planning and management. This includes planning for staffing, medical equipment, and other financial resources to meet anticipated demand. Additionally, hospitals can utilize these models to identify individuals at higher risk levels based on higher anticipated costs, and target certain populations with a specific disease through interventions and preventative measures to improve outcomes and reduce overall costs. Other business applications of predicting the level of total cost for new beneficiaries include pricing strategies, risk management, strategic planning, quality improvement, and waste reduction. Accurate predictions can help hospitals and insurance firms to make informed decisions that enhance both financial performance and patient outcomes.

# References:

* <https://data.cms.gov/collection/synthetic-medicare-enrollment-fee-for-service-claims-and-prescription-drug-event>
* <https://data.cms.gov/sites/default/files/2023-05/d51e1218-68c3-4c7c-9598-0b81f22fe903/User%20Guide%20-%20CMS%20Synthetic%20RIF%20Files%20May%202023_AM508_v2.pdf>
* **Frequency encoding index:**
  + [**Frequency Encoding Index.docx**](https://1drv.ms/w/s!ApZsJNhedAAeqkmS9tgWmkh9osvz?e=x1UNrP)
* **DATASETs:**
  + [**Median\_Income.xlsx**](https://1drv.ms/x/s!ApZsJNhedAAeqi-przKfafvmkCpS?e=sDZbLO)
  + [**Patient\_Claim\_Data.xlsx**](https://1drv.ms/x/s!ApZsJNhedAAeqiTeDLVHUqLXBRJa?e=VTofyq) **(Merged file)**