Diabetes Readmission Predictive Model

Authors: Youssef Sadquy, MD Nafiz Rahman

Contents

Dataset Overview

The initial dataset for diabetic patients has 101,766 records and 50 features which captures comprehensive information on diabetic patients, their medical history, ethnic data, various hospital visit records and much more data.

Key Variables

The dataset contains variables that may be categorized into many groups, each group of data offering in depth insight into the patient's medical history and hospital's interaction.

Patient Identifiers: encounter_id and patient_nbr are unique identifiers for each patient. 'patient_nbr' is unique for each patient, additionally for each visit to the hospital a patient have been assigned unique different encounter_id.

Demographic data: Each patient's record contain data about their race, gender and age. These variables are very important to address distribution of diabetes and rate of readmission across different segments of population.

Medical history: The dataset contains different stages of diagnosis for each patient. These are diag_1, diag_2 and diag_3 which represents their initial diagnosis, secondary diagnosis and additional diagnosis respectively. It also contains num_medications (number of medications used by patient), num_lab_procedures (number of lab procedures on a patient) and different encounters with the hospital shown by number_outpatient, number_inpatient and number_emergency.

Treatment and Outcome variables: The dataset also contains admission_type_id, admission_source_id and discharge_position_id which provides insight into the specifics of reason for the patient's admission and the outcome of the admission.

Medications: The dataset has 18 different variables each indicating a specific medication. These variables provides data about the dose of that medication, if the patient takes the medication, if the dose is steady, or if does went up.

Target Variable: The dataset contains the main target variable for our model building analysis which is the 'readmitted' variable. This variable shows whether that patient has been readmitted within 30 days, after 30 days or has not been readmitted at all.

Some characteristics to be noted from the dataset are:

Age Group: In the dataset the age for each patient is given as a range of 10 years intervals.

Diagnosis Codes: The dataset includes a large number of unique ICD codes for each diagnosis of a patient.

Initial Observations

The broad overview of each patient could provide important insight towards the readmission rate. After initial observation of the data, there are multiple variables that could potentially influence the readmission rates, such as age, race, gender, weight, number of medications, number of hospital visits, outcome of hospital visits, number of procedures and number of lab procedures.

Data cleaning

Handling Missing Values

The dataset initially contained missing values shown by '?'. These were replaced with numpy.nan to represent the missing information across the dataset.

Dropping Columns with Excessive Missing Values

We identified columns that have more than 50% of their values missing and removed those columns. Because columns with lots of missing data offers limited insight over the total population and could skew the analysis.

Dealing with Low Variability Columns

Additionally we also identified columns where over 95 percent of the values were the same. These columns were removed as these offer limited insight due to the very low variability of data.

Age Transformation

The original dataset had age as category and a range of intervals of 10 years. The age column was transformed to take the midpoint value of the range. This ensures numerical representation of age and inclusion of age in model building analysis.

Filling Missing Diagnoses

We dealt with missing values in the diagnosis columns (diag_1, diag_2 and diag_3) by replacing them with number 0.

Row Removal for Missing Data

After we have removed columns with lots of missing data or column with low variability, we removed any rows that still contained missing values.

Outlier Detection and Removal

For numerical columns we have identified outliers that are 3 standard deviations away from the mean. Then we removed these outliers to only keep data within 3 standard deviations from the mean.

Duplicate Removal

Patient identifier (patient_nbr) is unique identifier for each patient. Multiple occurrences of it may occur which is multiple visits to the hospital for the same patient. Therefore any duplicate instance of patient_nbr was removed ensuring that each data for a patient is unique.

Resulting Dataset

After we finish the data cleaning process adhering strictly to the given instructions the dataset was significantly reduced in dimensions, however the quality of the dataset was enhanced. The shape of the dataset post cleansing was 17508 rows with 31 columns. On the resulting dataset after we do exploratory analysis to find any relation of the variables with our target variable.

Data Exploration

Age and Readmission Rates Analysis

We plotted bar plot to visualize rates and distribution of readmission rates across different age group. This will aid us to examine the hypothesis stating that age has higher impact on readmission rates.

Visualized Data Insights and Total Counts:

After thorough inspection of the graph it is evident that -

- **Age 25:** 1.19% not readmitted and 0.46% readmitted out of 288 total data points.
- **Age 35:** 2.63% not readmitted and 1.06% readmitted from a total of 645 data points.
- **Age 45:** 6.53% not readmitted and 3.09% readmitted with 1,684 data points.
- **Age 55:** 11.58% not readmitted and 5.86% readmitted, totaling 3,053 data points.
- **Age 65:** 14.11% not readmitted and 8.53% readmitted from 3,964 data points.
- **Age 75:** 15.18% not readmitted and 10.25% readmitted, the highest count at 4,452 data points.
- **Age 85:** 9.61% not readmitted and 7.09% readmitted from 2,923 data points.
- **Age 95:** 1.86% not readmitted and 0.99% readmitted, accounting for 499 data points.

Hypothesis Reevaluation:

The graph highlights a noticeable increase in both readmission and non-readmission rates as age increases. The peak of the increase is at age 75, after we can see a slight decline at 85 and massive decline at 95. The massive decrease in both readmission and non-readmission rate past age 75 is also due to less available data points. This trend confirms that older patients generally are more likely to be readmitted than the younger patients.

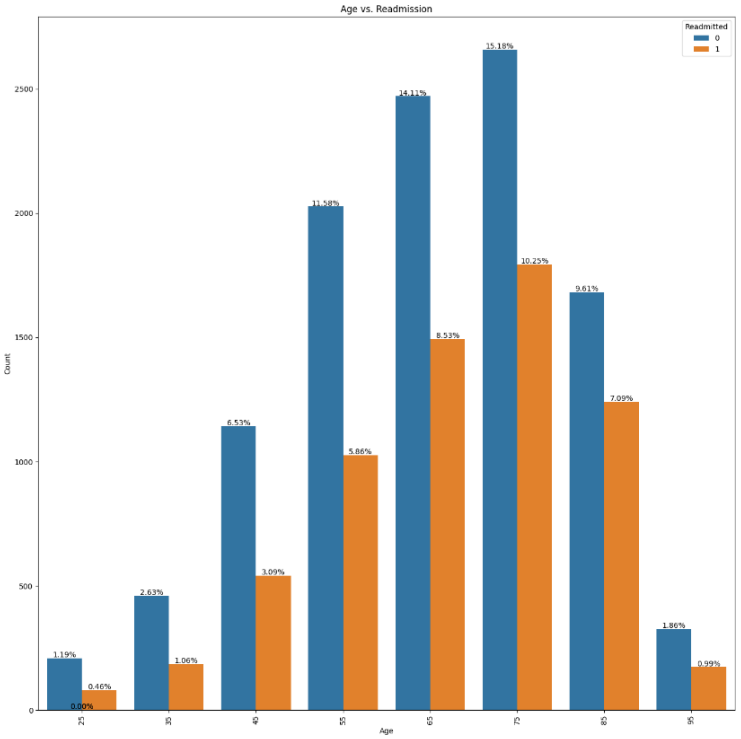

Figure 1. Age vs Readmission Bar Plot

Race and Readmission Rates Analysis

We also plotted race against readmission and non-readmission rates. This will aid us into further examining the hypothesis regarding race and readmission rates.

Visualized Data Insights and Total Counts:

- **Caucasian Patients:** Representing the largest group, 46.27% were not readmitted, and 29.19% were readmitted with 13,211 total data points for this race.
- **African American Patients:** 13.45% were not readmitted, and 6.63% were readmitted from a total of 3,516 data points.
- **Hispanic Patients:** Had 1.30% not readmitted and 0.59% readmitted, with a total of 331 data points.
- **Asian Patients:** Showed 0.55% not readmitted and 0.28% readmitted, from a total of 145 data points.
- **Other Race Patients:** Had 1.11% not readmitted and 0.63% readmitted, with 305 total data points.

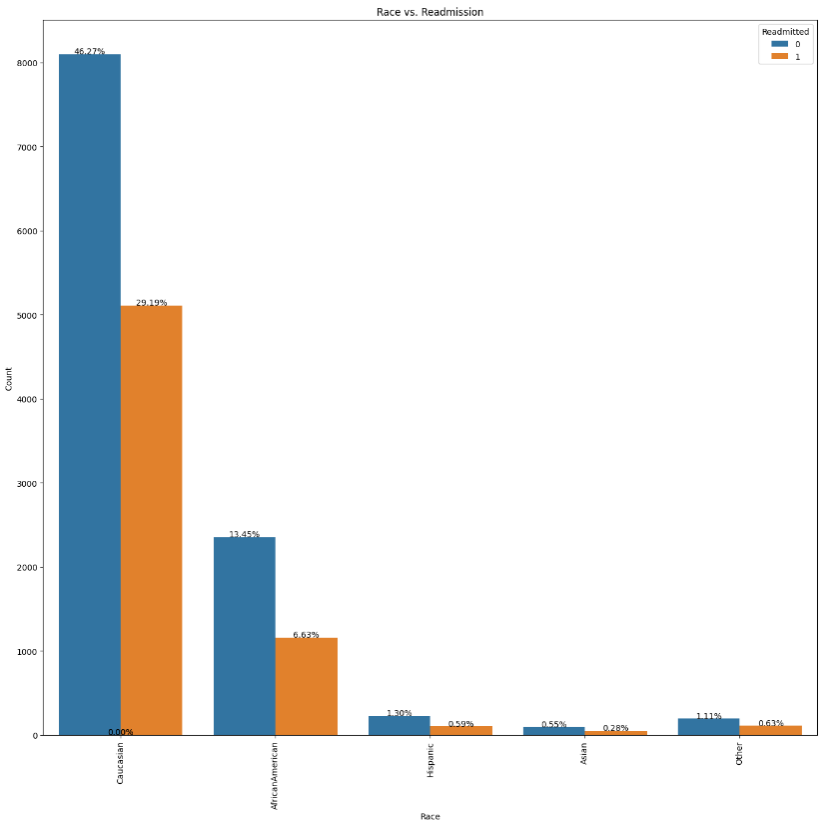

Figure 2. Race vs Readmission Bar Plot

Hypothesis Revaluation:

From the bar plot it is imminent that there is a clear disparity in readmission rates across different racial groups. As shown in the data insight Caucasian patients post data cleansing has the highest number of data entries which could also impact how it has the highest number of readmission and non-readmission rates. Next race group which is African American has a significant decrease of readmission rates from Caucasian patients, it is also important to note that the data points available for African American patient is just over one third of Caucasian patients. Although African American patients clearly have higher readmission rates compared to Hispanic, Asian and Other races, it is still significantly lower than Caucasian Patients. Therefore without more data points for African American it can be said with some degree of certainty that African American patients are more likely to be readmitted compared to other races, although not compared to Caucasian.

Gender and Readmission Rates Analysis

Given the clarified data from the gender-based readmission rates visualization and the final cleaned dataset's demographic breakdown, we can more specifically address the gender hypothesis in hospital readmission rates for diabetic patients.

We plotted gender against readmission rates to visualise how patient's gender may affect readmission and also to address the hypothesis for gender in hospital readmission rates.

Visualized Data Insights:

- **Female Patients:** Among the female patients, 33.39% were not readmitted, while 20.13% were readmitted.
- **Male Patients:** For male patients, 29.30% were not readmitted, and 17.18% were readmitted.
- **Total Counts Post-Cleaning:** The final dataset, after cleaning, contains data on 8,137 male and 9,371 female patients.

Hypothesis Revaluation:

Based on the graph and analysis we can see that females have a higher percentage of readmission (20.13%) compared to males (17.18%). Females also have higher percentage of non-readmission compared to males. This is also due to the proportion of female data points being slightly higher than males. Although we have more female data entries, for the context of this dataset we can say that women patients are more likely to be readmitted than men.

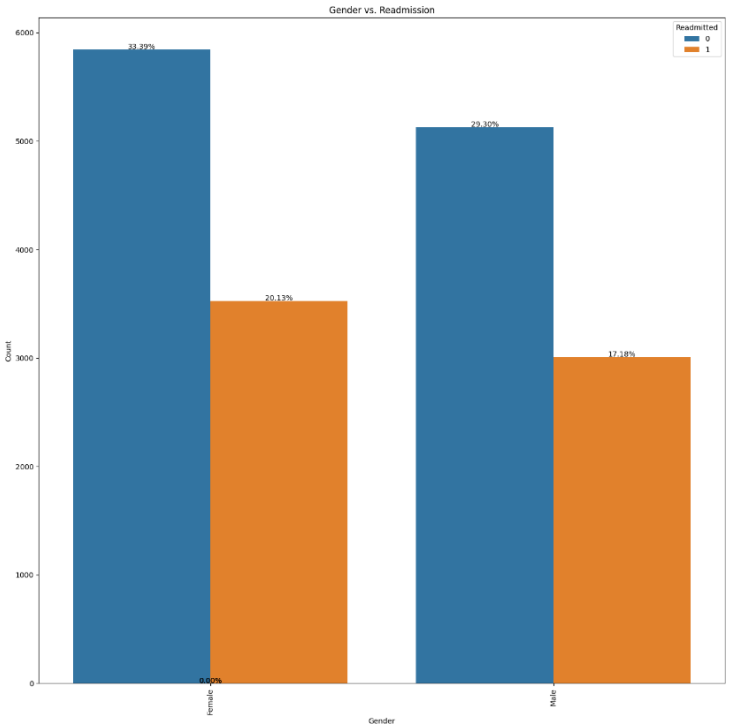

Figure 3. Gender vs Readmission Bar Plot

Diagnosis 1 and Readmission Rates Analysis

We categorized the ICD codes [1] following the provided guidelines. This helped us decrease the number of distinct ICD code values in each diagnosis to larger groups, then we plotted primary diagnosis (diag_1) against readmission rates.

Visualized Data Insights and Total Counts:

- The percentage provided on the bar plots reflects readmission and non-readmission rates between different diagnosis groups. Some key observations from the dominating group for readmission and non-readmission rates are given below.

Key Observations:

- **Disease of the Circulatory System:** Represents the largest group with 18.18% not readmitted and 12.31% readmitted out of 5,338 total data points.
- **Disease of the Respiratory System:** Has 8.30% not readmitted and 5.41% readmitted from a total of 2,400 data points.
- **Diabetes mellitus:** 4.70% not readmitted and 3.12% readmitted, indicating a significant readmission rate for diabetes-specific conditions among the total of 1,370 data points.
- **Diseases of the Digestive System:** Shows 5.77% not readmitted and 3.08% readmitted out of 1,549 data points.
- **Injury and Poisoning:** 4.11% not readmitted and 2.40% readmitted, suggesting notable readmissions related to injuries and poisonings among 1,141 data points.

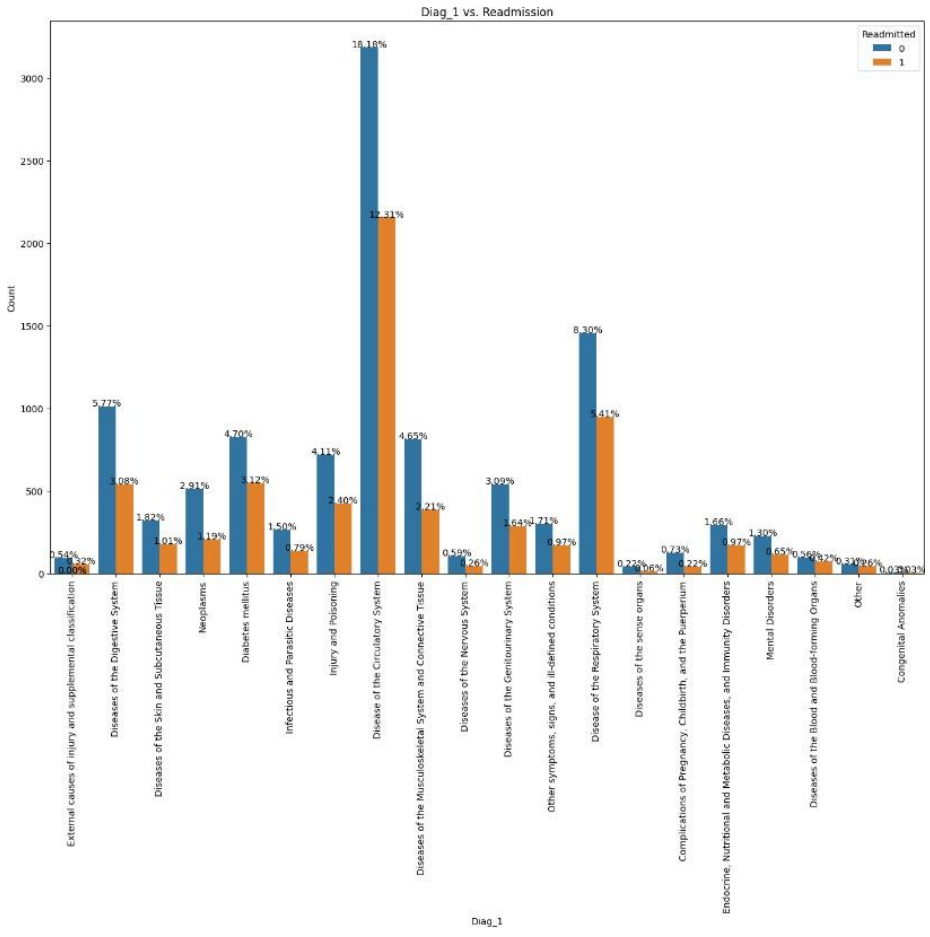

Figure 4. Diag_1 vs Readmission Bar Plot

Hypothesis Reevaluation:

The plot shows impact of various diagnosis categories on readmission rates. From the graph we can see that people with primary diagnosis of diseases under 'Disease of the Circulatory System' have the highest rate of readmission and non-readmission. This is also due to their dominant entry numbers in the dataset. Followed by 'Disease of Circulatory System' we can see 'Disease of Respiratory System', 'Disease of Digestive System' and 'Diabetes Mellitus' descending respectively for readmission and non-readmission rates. It is also important to note that the diagnosis with relatively low readmission also

have low data entry points. The graph shows us that various types of diagnosis have higher impacts on readmission rates which aligns with the hypothesis.

Model Performance Evaluation:

For the initial model we used a Random Forest Classifier as it is effective for high dimensional data. Below we will provide walkthrough steps that we took to build the model, provide relevant metrices for the model and evaluate the model using cross validation.

Model Building Steps:

- Initial step involved data cleansing following the given instructions which includes handling missing values, dropping columns with low variability or large number of missing values, dropping rows with any missing values, transforming age to midpoint value of the range, filling missing values in diagnosis columns and removing outliers from numerical features.
- Feature selection for the model was given in the instructions which included 'num_medications', 'number_outpatient', 'number_emergency', 'time_in_hospital', 'number_inpatient', 'num_lab_procedures', 'number_diagnoses', and 'num_procedures'.
- The dataset was split into a training set and a testing set. 70 percent is training set and 30 percent of the data is testing set.
- The Random Forest Classifier model was trained on the training set, random_state was set to 0 to ensure reproducibility of the score for evaluation.
- Relevant metrices to understand the performance of the model is printed which includes accuracy score, precision score, recall and F1 score.
- 10 Fold cross validation is done to also evaluate the model.

Confusion Matrix:

We plotted confusion matrix graph to see how the model performs at predicting readmissions and non-readmissions.

- **True Negatives (TN):** True Negatives of 2685 indicates that the model correctly predicted 'Not Readmitted' for 2685 cases.
- **False Positives (FP):** False Positives of 592 indicates that the model incorrectly predicted 'Readmitted' for 592 cases that were actually not readmitted.
- **True Positives (TP):** True Positives of 694 indicates that he model correctly predicted 'Readmitted' for only 694 cases.
- **False Negatives (FN):** False Negatives of 1282 indicates that the model incorrectly predicted 'Not Readmitted' for 1282 cases that were actually readmitted.

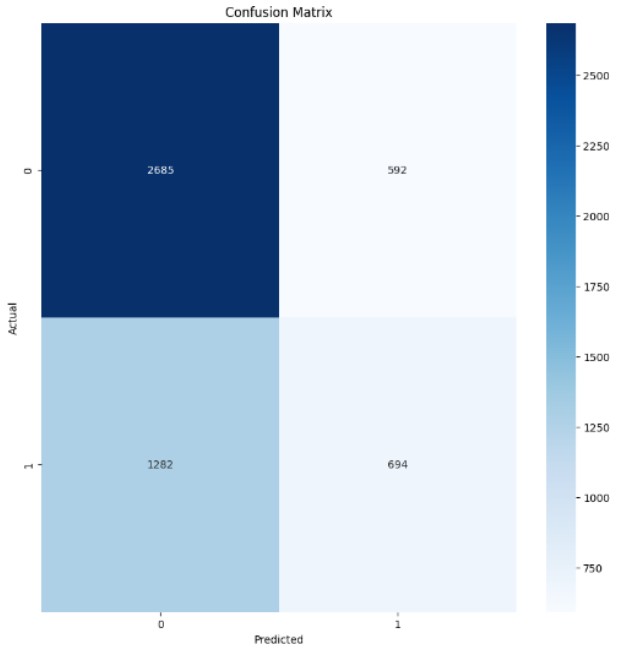

Figure 5. Model 1 - RandomForestClassifier Confusion Matrix

This insight from the confusion matrix suggests that the model is significantly better at predicting non-readmission compared to readmissions.

```
Accuracy: 0.6432514753474206
Precision: 0.5406850955414012
Recall: 0.3436234817813765
F1 Score: 0.42017326732673266
Cross-validation scores: [0.62678469 0.37492861 0.40719589 0.40159954 0.62696372]
Mean cross-validation score: 0.48749449247673543
```

Figure 6. Model 1 - Output of evaluation metrices.

- **Accuracy:** Accuracy of approximately 65.22% suggests that the model correctly predicts readmission stats for around two-thirds of the total cases. This is a measurement of overall performance.
- **Precision:** The precision score of 53.94% suggests that the model is correct slightly above half of the time at making prediction on readmission. This highlights the model's performance on the ability to identify positive readmission instances.
- **Recall:** The recall rate of approximately 34.68% suggests that the model identifies around a third of all actual readmission cases. This lower recall rate suggests potential area for improvement.
- **F1 Score:** The F1 score is approximately 42.22% which is the balance of precision and recall.
- **Cross-Validation Scores:** We performed 10-fold cross validation which provided us with a range of accuracies of 10 different subsets of the data, this highlighted some significant variance in the performance of the model. The mean cross validation score is 48.75% which is significantly lower than the model's accuracy suggesting overfitting of the model or imbalanced dataset.

Considerations for Improvement:

- **Feature Engineering:** Further feature engineering could introduce some useful features that the model could utilise to improve the performance. New features or transformations of existing features could aid the model in capturing underlying patterns of readmissions and non-readmissions.
- **Hyperparameter Tuning:** Currently the Random Forest model does not contain any tuning of the hyper parameter except of random_state to retain reproducibility of the score. Further hyper parameters could be used to optimize its performance.
- **Alternative Models:** Exploring different models could give us an overview of how the initial model performs comparing to different predictive models.
- **Balanced Data:** From the confusion matrix we can notice a clear imbalance in the data for readmission and non-readmission rates. This could cause bias in the data favouring the majority class. We can apply some balancing technique such as oversampling the minority class or under sampling the majority class.

The initial model serves as a baseline for the prediction of readmission. The outcome of the model provides guidance on which area to focus to refine the model's predictive abilities.

Improved Data Cleansing and Processing

We took some additional steps on top of initial cleaning to refine the dataset more appropriately for the model building.

1. **Identification and Removal of Unknown Gender Data**: For the improved data cleansing steps we identified three instances of 'Unknown/Invalid' records for gender column and removed them.

2. **Strategic approach to handle missing values:** To improve the data cleaning steps for the race column that had '?' entries we do not simply remove them, instead we replaced with 'Other' which will allow the inclusion of these entries. As 'Other' in race does not specify what race it is, we may add entries with '?' to the other category without the need of incorrectly predicting or guessing the race.

3. **Dropping population that are out of scope:** After a thorough inspection of IDs mapping document we found out that patient with discharge disposition id 11,19,20 and 21 passed away either at home or hospice. These population are out of scope and interest for the model as they cannot be readmitted again.

4. **Dropping irrelevant and high missing value columns**: We handled missing values carefully now. We did thorough inspection of the amount of missing value for each column.

Weight, medical_specialty and payer_code has the highest percent of missing values. Similar to initial cleaning we decided to drop weight and medical_specialty due to high missing values and we dropped payer_code due to irrelevance towards our target variable of readmission. We also dropped any rows that had all three-diagnosis data entry missing.

col	count_missing	percent_missing
weight	98569	96.86
medical_specialty	49949	49.08
payer_code	40256	39.56
race	2273	2.23
diag_3	1423	1.40
diag_2	358	0.35
diag_1	21	0.02

Figure 7. Percentage of missing values in each column

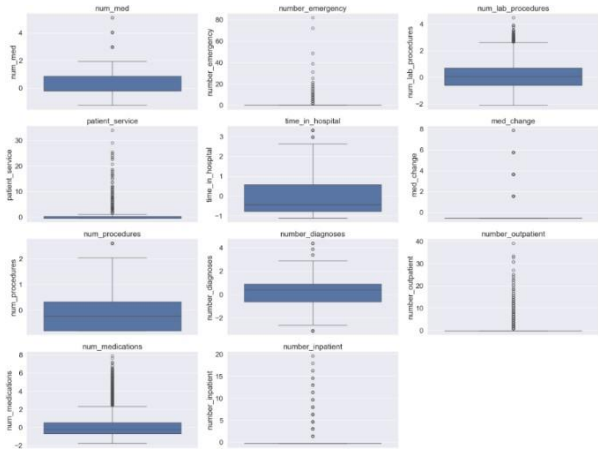

Figure 8. Outlier detection

5. **Outlier Detection and Removal:** Unlike the initial model using z-score to base the outlier we used Interquartile Range (IQR) method to treat outlier more effectively. As the dataset has lots of variables and dimensions we came to conlcusion that dealing with any noises in the dataset more aggressively is effective for the model improvement.

Feature Engineering

We implemented some feature engineering steps to use for the improved model.

1. **Creation of Composite Features:** New variables are introduced to aggregate important information in multiple columns. We took the sum of outpatient visits, emergency visits and inpatient visits to make a new feature for each patient called patient_service, essentially tracking interaction with the hospital for each patient. We also introduced a new feature called num_med that is an aggregation of all the number of medications taken by a patient. Additionally we created a new feature called med_change for each patient to track the number of times their medication was changed.

2. **Recoding of Diagnosis ICD codes**: We reduced the number of major categories for each diagnosis. We group the low number or any unique instance of the ICD codes into one category recoding it to 0. All other major categories were recoded to numerical from 1 to 8 in order to include in the model building process for better insight towards readmission rates.

3. **Narrowing down admission and discharge IDs:** We replaced certain IDs in admission_source_id, admission_type_id and discharge_position_id after a thorough investigation of the ID mapping document. The IDs were replaced with very similar counter part which can be categorised together for better insight and predictive efficiency.

4. **Log Transformation to skewed variables**: Skewness was not addressed in the initial processing of the dataset. For the improve feature engineering process we identified any skewness of the numerical features and treated them by applying log transformation to create new features.

5. **Standardisation of Numerical Features**: We standardise all numerical features in the dataset to have a mean of 0 and standard deviation of 1. This ensures that all features are scaled and no specific feature due to large scale may influence the model in decision making.

6. **Dummy variable for categorical variables**: We created dummy variables for all categorical variables and removed first level for dummy variable to avoid dummy variable trap. These dummy variables allow the categorical features to contribute significantly for the model building.

7. **Correlation Analysis:** With all the available features we print correlation matrix heatmap to visualise the correlation of other features with the target variable. From the correlation matrix it is clear that no specific feature now has very strong correlation (shown by lighter colours) with readmitted. The top feature somewhat showing slightly higher correlation than others with readmitted are patient_service_log, number_inpatient_log, number_inpatient, patient_service and

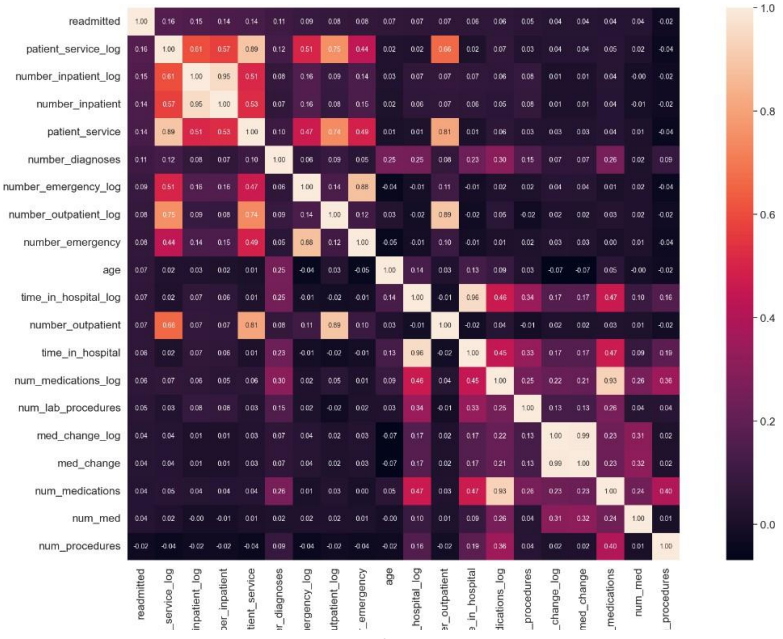

Figure 9. Correlation Heatmap

number_diagnosis. The rest of the feature show very weak correlations to the readmitted feature. But ultimately for tree-based model correlation of feature is not as important as for linear models. For our final improved model we will do Random Forest classifier model.

The advanced data cleansing and new feature engineering process was done to produce an improved dataset of the original with better quality providing effective insight to the models for improved analyses.

Model Building

Key Steps and Components:

1. **Data Splitting and features selection:** We split the data into features and the target variable. Initially we included all the features that was engineered, transformed, processed including dummy variables. Due to previously shown weak correlation towards the target variable on the correlation matrix we took all these extensive features denoted by feature_set to see how the model might react.

2. **Class Distribution check:** Before we continue to the modelling we checked the distribution of the target variable classes. As shown in the image it is clear that there is a significant imbalance present in the dataset. The imbalance of 1.88:1 suggests that in the final dataset there are nearly twice as many instances of patient's non-readmission compared to readmission. As in the initial model building, we did not address class imbalance we will try to fit Logistic regression model in the imbalanced dataset and analyse the outcome.

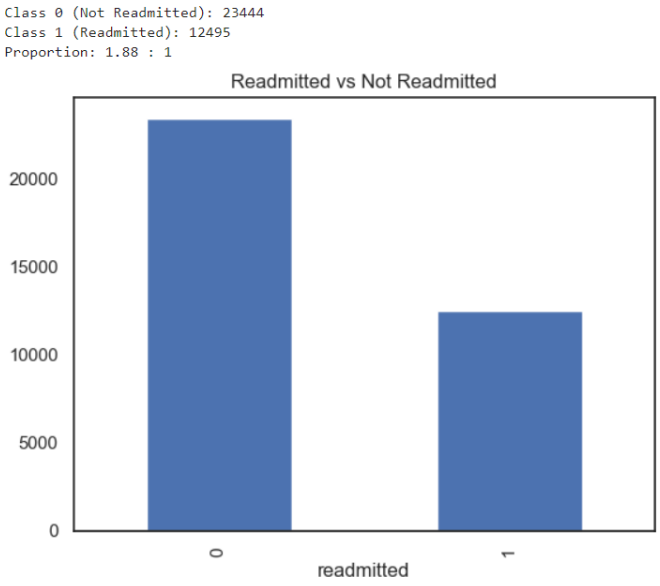

Figure 10. Bar plot visualising imbalanced dataset

Initial observations prior to model building suggest significant imbalance. Before this imbalance is addressed, we will plot logistic regression model to analyse the result as the in the initial model we did not address any class imbalance.

Logistic Regression

1. **Data Splitting:** Similar to the initial model building we split dataset into features (X) and target variable (y). Then we split 70 percent of the data for training and 30 percent for testing.

2. **Logistic Regression Model:** A logistic regression model is created with max_iter=6000 to ensure convergence is done during training of the model. The model is trained on the training split using fit method.

3. **Model Prediction and Evaluation:** To evaluate the logistic regression model we print the accuracy, precision, recall and f1-score of the model. Additionally we also perform 10-fold cross validation to evaluate the model.
4. **Confusion Matrix:** We generate a visualisation graph for the confusion matrix of the logistic regression model to better understand the true positives, true negatives, false positives and false negatives of the model.

Findings from the Outcome:

The findings of accuracy score, precision, recall and f1-score of the logistic regression model is given below:

- **Accuracy:** Accuracy of 65.63% indicates that the model correctly predicts readmission status for approximately 65-66 cases out of every 100 cases.
- **Precision:** Precision of 48.69% indicates that the prediction of a patient being readmitted is correct about 48.69% of the times.
- **Recall:** The recall of the model is very low at 10.61%. This indicates that the model unable to identify a lot of patients who should classified as readmitted.

This is a potential decrease of performance compared to the very initial model. This decrease is highly likely due to the unaddressed imbalance of data which will solve.

Analysis of the Confusion Matrix:

The confusion matrix that we plotted for the logistic regression provided us with some useful insight given below:

- **True Negatives (TN):** True Negatives of 6685 indicates that the model correctly predicted 'Not Readmitted' for 6685 cases.
- **False Positives (FP):** False Positives of 412 indicates that the model incorrectly predicted 'Readmitted' for 412 cases that were actually not readmitted.
- **True Positives (TP):** True Positives of 391 indicates that the models correctly predicted 'Readmitted' for only 391 cases.
- **False Negatives (FN):** False Negatives of 3294 indicates that the model incorrectly predicted 'Not Readmitted' for 3294 cases that were actually readmitted.

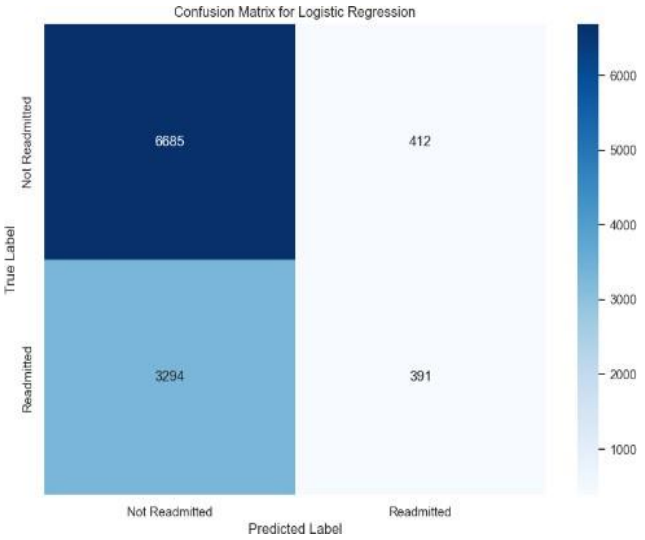

Figure 11. Model 2 - Logistic Regression Confusion Matrix

From the insight of confusion matrix we can see that there is a large number of false negatives compared to true positives which explains the recall score. Low recall score suggested that the model fails to identify many of the patients that needs to be readmitted.

Data balancing

Explanation of the Balancing Process:

1. **Identifying Majority and Minority Classes:** To properly balance the dataset we found and segregated the readmitted column into two groups. These two groups are majority class (df_majority - non-readmission) and minority class (df_minority - readmission).
2. **Analyse initial size and proportions:** We analysed the number of instances for both majority and minority classes which are 23,444 instances for majority and 12,495 for minority. The proportion is 1.88:1.
3. **Oversampling the minority class:** To address the issue of imbalance dataset we oversampled the minority class. Under sampling of the majority class to match the minority class in our case would have resulted in a loss of a big portion of the data. That's why we oversample the minority class by random sampling with replacement until the number of minority class is equal to the majority class.
4. **New balanced dataset:** We concatenated the oversampled minority class with the original majority class to produce a new balanced dataset (df_oversampled) that we will perform model building on.
5. **Splitting balanced dataset:** We split the balanced dataset into features (train_input_new) and target (train_output_new) for model training and evaluation on the new balanced dataset.
6. **Analyse size and proportion of data post balancing:** After we balance the dataset, we verify that the proportion of both readmission and non-readmission in the new dataset is 1:1 with equal count of 23,444 instances.
7. **Split of balanced dataset:** Finally we split the balanced dataset into training and testing sets, the same splitting ratio of 70 percent training and 30 percent testing was followed.

Justification for Balancing:

- **Mitigating model bias:** By balancing the dataset we mitigate the probability of the model being biased towards the majority class. Without balancing of the dataset the model will predict most instances to be the majority class. This will result in poor generalisation, specifically for the minority class.
- **Improving Recall:** Earlier when we plotted logistic regression model for the imbalanced dataset, we observed very low recall suggesting many false negatives. Balanced dataset will ensure that the model has equal number of readmission and non-readmission instances to learn from, thus reducing false negatives.
- **Impact on other metrices:** Balanced dataset will give an accurate overview of the accuracy as it will now give a straightforward overall correctness of the model's prediction on equal instance of readmission and non-readmission. Precision measures the quality of positive predictions and with balanced dataset the precision will be more meaningful. Similarly F1-score will be more reliable now.
- **Enhanced generalisation:** By ensuring that the model is provided with a balanced dataset where number of instances for both classes appear in an equal rate, the model is more likely

to generalise better and predict more accurately on unseen data. This will avoid any skew towards the majority class which often is the case on imbalanced datasets.

Random Forest Classifier on Balanced Data:

A Random Forest classifier model was trained on the balanced dataset. The hyper parameter for random forest was tuned to have n_estimators=100 and out-of-bag (OOB) score enabled.

Performance Metrics:

- **Accuracy:** A significant increase of accuracy is observed at 82.57% showing good improvement from the first random forest classifier and logistic regression. This is likely due to the balanced dataset.
- **Precision:** The precision of the model is 83.07% which means that when the model predicts readmissions, over 80% of the time it is correct.
- **Recall:** Recall is 81.37% suggesting that the model is much better at identifying true readmissions compared to the logistic regression on imbalanced dataset.
- **Cross-Validation scores:** 10-fold cross validation was done with different random subsets of the data. The model achieved a cross validation mean of 79.62% showing how the model's performance varies in different subsets of data. This mean cross validation score is close to the accuracy observed at 82.57% which is a good sign for the model.
- **OOB Score:** OOB score for the model was observed at 80.44% which is very close to the cross validation mean score and the accuracy of the model. This suggests that the model generalizes well.

Confusion Matrix Analysis:

The confusion matrix showed the following:

- **True Negatives (TN):** 5915
- **False Positives (FP):** 1162
- **True Positives (TP):** 5700
- **False Negatives (FN):** 1290

Feature Importance and Selection

In order to identify the most important features used by the Random Forest classifier we generated a list of most important features using built-in attribute feature_importances_ for Random Forest classifier. When random forest is trained using .fit(), internally list of important features which contributes to prediction is also calculated. [2]

Key Observations:

There are several key features that we extracted from the list of most important features used by random forest classifier.

- Features that are related to medical procedures and diagnoses features (num_lab_procedures, diag_1, diag_2, diag_3) are amongst the most important features.
- There are also other important features which are num_medications, age, number_diagnoses, time_in_hospital, num_procedures and many more ordered in a descending list of importance.

Feature	Importance
num_lab_procedures	0.083766
diag_1	0.075379
diag_2	0.072631
diag_3	0.069968
num_medications_log	0.060880
num_medications	0.060667
age	0.047899
number_diagnoses	0.044456
time_in_hospital	0.038994
time_in_hospital_log	0.038933
num_procedures	0.026400
num_procedures_log	0.026286
gender_1	0.018441
num_med	0.017528
additional_diag_1	0.012690
discharge_disposition_id_2	0.012631
race_1	0.012593
admission_source_id_7	0.011818
insulin	0.011710
secondary_diag_1	0.011436
race_2	0.011181
primary_diag_1	0.010520
metformin	0.010120
A1Cresult_1	0.010066
primary_diag_2	0.009326
additional_diag_4	0.008983
glipizide	0.008839
change	0.008600

Figure 12. List of feature set for RandomForest classifier in descending order of importance

Creation of a Refined Feature Set:

After analysing the list of important features used by random forest classifier model, we created a separate list of feature set to include only the features that contribute more than 1% to the prediction. We strategically reduce the dimension which will aid the model to focus more on the relevant features potentially improving performance on unseen data and reducing any potential overfitting.

Steps and Justification for the New Feature Set:

- **Dimensionality Reduction:** The complexity of the model is reduced by using only the topmost important features, this can improve the model training time and the interpretation of the model.
- **Improved Performance:** By reducing the feature set the model might become better at generalising on unseen data by focusing on the most relevant features.
- **Reduction in overfitting:** Eliminating features that are least of importance for the model will reduce the risk of overfitting.

After we created the new feature_set2 it was selected and the dataset was split again, this ensured that the model was trained and evaluated on the most important and relevant set of features.

Model Building with Refined Features:

Another Random Forest model was trained on the new refined feature set that was narrowed down based on the importance of the features. This feature set contained features that contributed more than 1% to the model's prediction.

Model Training and Evaluation:

The random forest classifier was retrained with same hyper parameters for the earlier random forest model to allow straight forward comparison of performance post feature selection.

Performance Metrics:

- **Accuracy:** The model achieved an overall accuracy of 82.20% suggesting a high overall rate of correct predictions.

- **Precision:** The precision of 82.95% suggests that the model is very likely to be correct when predicting readmissions.
- **Recall:** The recall of 80.79% suggests that the model is good at finding true readmissions from the dataset.
- **Cross-validation scores:** The 10 fold cross validation scores shows less variability in the accuracy of each fold and the mean score of 79.10% suggests that the model stable and good at generalising.
- **OOB score:** OOB score of 80.39% which is close to accuracy and cross validation mean score further confirms the reliability and performance of the model.

Confusion Matrix Analysis:

The confusion matrix that we plotted for the final random forest classifier model with refined feature set is given below:

- **True Negatives (TN):** True Negatives of 5916 indicates that the model correctly predicted 'Not Readmitted' for 5916 cases.
- **False Positives (FP):** False Positives of 1161 indicates that the model incorrectly predicted 'Readmitted' for 1161 cases that were actually not readmitted.
- **True Positives (TP):** True Positives of 5647 indicates that he model correctly predicted 'Readmitted' for only 5647 cases.
- **False Negatives (FN):** False Negatives of 1343 indicates that the model incorrectly predicted 'Not Readmitted' for 1343 cases that were actually readmitted.

Comparing the confusion matrix and the performance metrices with the previous random forest model we can conclude that there is almost little to no performance difference. This is because the refined feature set included all the necessary features for the model with the exclusion of features that could introduce noise or irrelevant patterns for the model.

Conclusion and Recommendations:

By refining the feature set based on importance and retaining the same performance on the balanced dataset we managed to build a strong random forest classifier model with significantly improved predictive performance.

K-Means Algorithm

For the final part of the analysis we use unsupervised modelling technique known as K-means clustering. By applying K-means clustering the data is partitioned into K distinct number of clusters where each data point belongs to only one cluster.

K-Means Algorithm Process:

1. **Feature selection for K-Means:** To use K-means clustering we selected a subset of features from the refined feature set. We selected top 5 numerical features to properly produce clusters and for increased clarity.

2. **Normalisation:** The selected features for K-means clustering were normalised to ensure that they were scaled properly to contribute equally to the clustering results, since the algorithm use distance metrics that may be biased if the scale of data differ from each other.

3. **Optimal Number of Clusters (K) using Elbow Method:** We use the elbow method to determine the optimal number for K. First we plotted the sum of squared errors (SSE) in the y axis for K ranging from 1 to 10 in the x axis. In the plot we identified the elbow pattern or the bottom point of an elbow which indicated the optimal number for K. In our case the optimal number for K is 4. When K is more than 4 the clusters will not result in any significantly improved modelling of data.

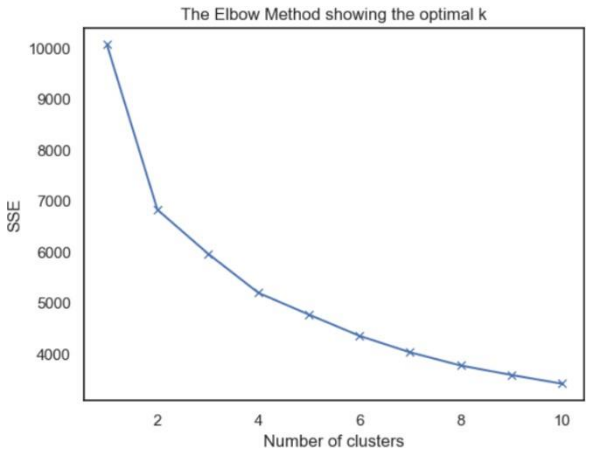

Figure 13. Elbow method plot for K-Means Model

4. **Analysing clusters with PCA:** We made use of Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, this will aid in visualising the clusters more clearly. The original dataset had more than three dimensions, that is not possible to visualise in a 2D plot, that is why use of PCA was necessary.

5. **K-Means Clustering:** We trained the K-means model with number of K set to 4, as we identified earlier plotting SSE against number of K and using elbow method.

6. **Cluster Mean Calculation:** For each feature within a cluster the mean value was calculated. This helped us extract valuable insight about characteristics of each cluster.

PCA Visualization:

To visualise the clusters we plotted two dimensional PCA plot, each cluster was given a distinct colour. This provided us with an initial impression of how well-separated and distinct the clusters were.

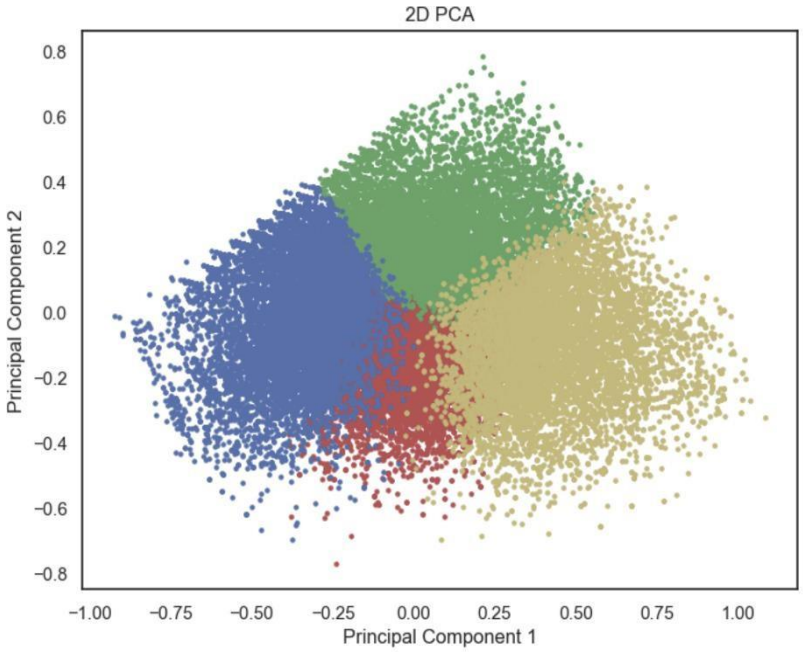

Figure 14. K-Means cluster visualisation

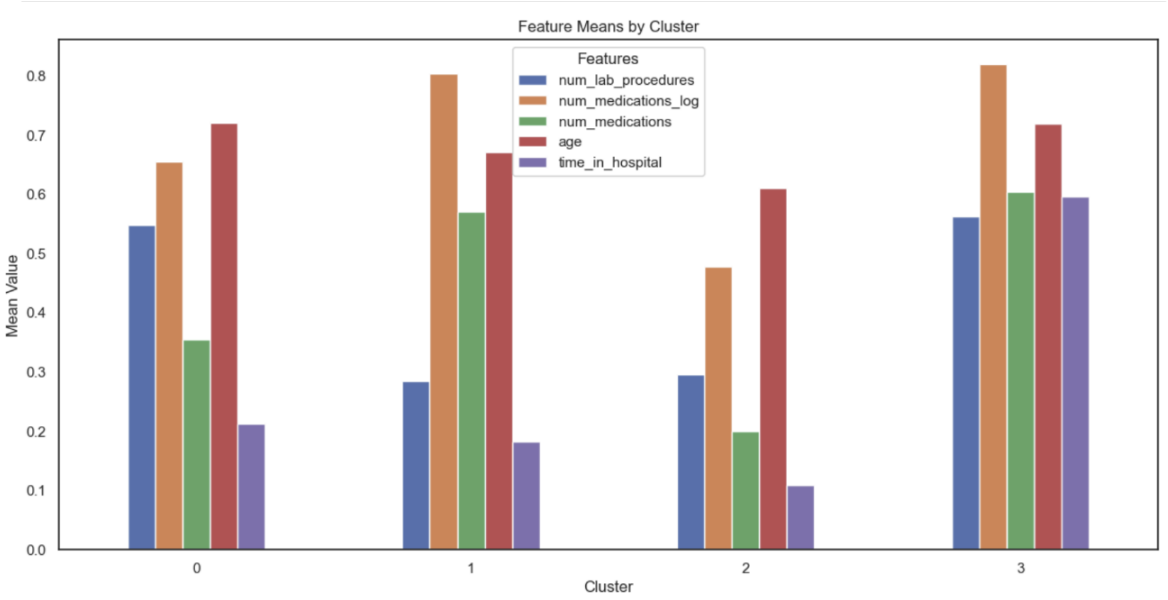

Figure 15. Feature means vs Cluster

Cluster Characterization:

The clusters were characterized by calculating the mean of the features within each cluster. Then we visualised the mean of the features for each cluster by plotting them on a bar plot.

Findings and Justification of the Clusters:

Cluster 0 (High Age and Moderate Medical Intervention):

- **Age:** This cluster has the highest mean for age, suggesting that it mostly consists of older patients compared to the other clusters.
- **Time in Hospital:** Time in hospital is relatively longer than cluster 1 and 2 but significantly lower than cluster 3.
- **Lab Procedures:** Cluster 0 has moderately high lab procedures, less than cluster 3 but more than cluster 1 and 2.
- **Medications:** The logarithmic of medications is relatively low compared to cluster 1 and 3.

Cluster 1 (Moderate Age, High Medication Usage):

- **Lab Procedures:** This cluster shows the lowest number for number of lab procedures.
- **Medications:** Patients from this cluster have the highest mean of log medication and moderately high number of medications.
- **Age:** Patients from the cluster have moderate age, higher than cluster 2 but lower than cluster 0 and 3.
- **Hospital Time:** They have a slightly lower mean time in hospital, potentially due to effective medication management.

Cluster 2 (Youngest Patients, Least Medical Intervention):

- **Lab Procedures:** Patients within this cluster have lower number of lab procedures mainly compared to cluster 0 and 3.
- **Medications:** Patients within this cluster require the least amount of medication suggested by mean of number of medications and logarithmic scale of medication.
- **Age:** This cluster contains the youngest patients out of all the clusters.
- **Hospital Time:** This cluster has the lowest mean time in hospital.

Cluster 3 (High age, intensive treatment):

- **Lab Procedures:** This cluster has a high number of lab procedures similar to cluster 1 indicating high monitoring needs.
- **Medications:** There's a high requirement for medications, both in terms of actual numbers and the logarithmic measure, suggesting a complex medication regime.
- **Medications:** This cluster contains the highest number of medications out of all the clusters.
- **Age:** The age for this cluster is very close to cluster 0, this suggests that these patients are also among the oldest from the population.
- **Hospital Time:** This cluster has the highest mean time in hospital indicating that the patients from this cluster maybe have long term serious conditions requiring more medicines.

Interpreting the Clusters:

The characteristics of each cluster can help healthcare providers tailor their approaches. For instance:

- Cluster 0: May benefit from various care programs, chronic disease management, and longer-term care planning as they represent older patients with moderate healthcare needs.
- Cluster 1: Might require a focus on complex care regimes, possibly with specialist consultations to manage complex treatment, due to being moderate age but high on medication usage.
- Cluster 2: Could be targeted for early discharge planning and education on self-health management. This is due to cluster 2 having the youngest patients with least medical interventions.
- Cluster 3: May require the highest care as this cluster suggests patients with high age and very intensive treatment in terms of medications taken and time spent in hospital.

Conclusion of Clusters:

Clustering provided valuable insights into patient grouping based on age, time in hospital, treatment and procedures. These clusters could inform hospital to correctly allocate resource for each cluster with patient management strategies, tailored care for patient and discharge planning for patients.

Reference

[1] Hindawi, “Table 2: Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records,” Table 2 | Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, https://www.hindawi.com/journals/bmri/2014/781670/tab2/ (accessed Mar. 15, 2024).

[2] “Feature importances with a forest of trees,” scikit, https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (accessed Mar. 16, 2024).