Airman Qualification Prediction

Project Final Report

Nasri Binsaleh, nasri.binsaleh-1@ou.edu

Uma Maheshwar Reddy Jangalapalli, umamaheshwarreddy.jangalapalli@ou.edu

Faculty Supervisor – Dr. Beattie Matt J, Dr. Danala Gopichandh.

Company & Sponsor – Data Institute for Societal Challenges – Dr. David Ebert

Table of Contents

Introduction

The primary goal of this project is to comprehend medical records to forecast the pilot's license renewal and its term. The data containing the medical records of patients were collected by IBM Truven Health MarketScan® Research database and purchased by DISC, OU. Gathering insights from these historical medical records of patients can be valuable for Federal Aviation Administration (FAA) to assist in improving their pilot licensing process. It could be a helpful tool to determine who should have their pilot's license renewed and for how long, depending on their historical medical records. A pilot must file an application and go through the entire renewal procedure, which includes a physical test and necessitates additional time and money when their license expires. An essential step in this process is evaluating their medical exams to assess their potential for renewal. In this project, we will evaluate foundational data science concepts of statistics, machine learning, and deep learning to develop tools that provide insights into predicting the potential for the renewal of a pilot's license to assist the FAA or corresponding authorities.

Objectives

The technical objectives of the project include analyzing and understanding medical records, evaluating existing databases and approaches, and identifying critical features for accurate prediction. The existing codebase will be validated and adjusted before modifying ML techniques, addressing class imbalance issues, and redesigning feature engineering. The project will involve developing and comparing various predictive models and creating a software tool with an interface to display visualization of factors affecting prediction. The iterative process of feature assimilation, modeling, and hyperparameter finetuning will be necessary for successful completion.

In terms of individual learning objectives, the project will provide an opportunity to understand metadata and data dictionaries to process valuable data. It will also provide experience with real-time data and its impact on daily lives. Evaluating efficient database structures for storing data is another learning objective. The project will enhance knowledge of the modeling pipeline, including data preprocessing, model selection, training, and hyperparameter tuning. The use of visualization tools to present findings and receive feedback is also a learning objective. Finally, the project will involve exploring the incorporation of prediction results into a simple web application with an interactive graphical user interface.

Data

Ingestion

The dataset was acquired by OU Data Institute for Societal Challenges (DISC) from IBM Truven Health MarketScan® Research database. This database person-specific clinical utilization, expenditures, and enrollment across inpatient, outpatient, prescription drug, and carveout services. These data were gathered from health plans, government, large employers, and public organizations. Although this database covers patients with no restriction on what their occupation may be, we will apply machine learning on this data to predict with pilot's data.

Currently, the database tables are stored as CSV files on hard drive. Then some parts of the data are read into python and stored there as pandas dataframe. However, with larger dataset, this may pose a memory problem. Thus, we are also looking into storing these data tables in a database system such as Azure SQL, or PostgreSQL. This way, it may be more efficient to clean up and organize the data as well.

Exploration

The MarketScan® data has several databases including:

- Commercial Claims and Encounters Database
- Medicare supplemental
- Health and Productivity Management Database
- Benefit Plan Design Database
- Medicaid Database
- MarketScan Lab

These databases that we acquired have medical records between the years 2017 to 2020. After inspecting the nature of each database, we narrowed down the scope of the project to only the Commercial Claims and Encounters database, as others are less relevant and would require too much memory and computing resources.

The Commercial Claims and Encounters database contains the following data tables:

- Inpatient Admissions Table
- Facility Header Table
- Inpatient Service Table
- Outpatient Services Table
- Outpatient Pharmaceutical Claims Table
- Enrollment Table
- Prescription Drug Table
- Red book Table

The columns of each table were inspected to look for relevant information that can be useful to train the prediction model. The columns of each table across the years were also inspected to make sure that we have the correct reference keys. This can be seen in the Appendix section. After inspecting the columns of the tables, it was determined that the majority of the relevant information is in the Inpatient Admissions Table. Other tables contain information that may not be useful in predicting diseases. And for the Inpatient Admissions Table, the useful information, for now, are the following columns: *'ENROLID', 'YEAR', 'AGE', 'DX1', 'DX2', 'DX3', 'DX4', 'DX5', 'DX6', 'DX7', 'DX8', 'DX9', 'DX10', 'DX11', 'DX12', 'DX13', 'DX14', 'DX15', 'SEX'*. From the mentioned list of columns, DX1 to DX15 are the disease diagnosis codes in ICD-10-CM format. Narrowing down the amount of information to be loaded into the program is beneficial as irrelevant data will take up unnecessary amount of memory.

Since we are trying to see the development of cardiac disease from one year to the next, we decided to use the data from 2019 and 2020. The number of records in 2019 and 2020 are 1,133,288 and 984,798 records respectively.

Some data exploration can be seen below. First is the age distribution in the dataset of both years.

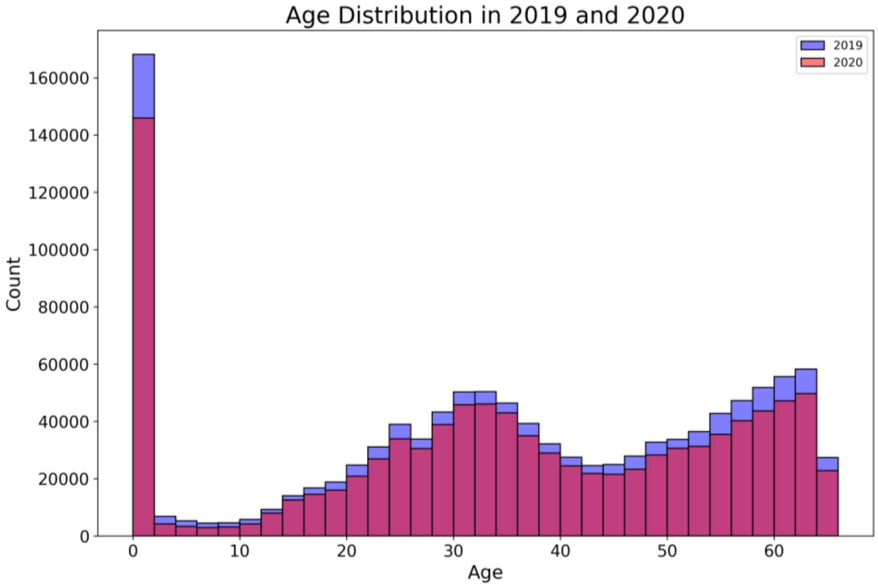


Figure 1 Age Distribution in 2019 and 2020

It can be seen that there is a great number of patients with age between 0- to 2-year-old, this must be taken into consideration, and they may be eliminated from the data because those infants are not qualified to be a pilot by default. Next is the distribution of sex in both years.

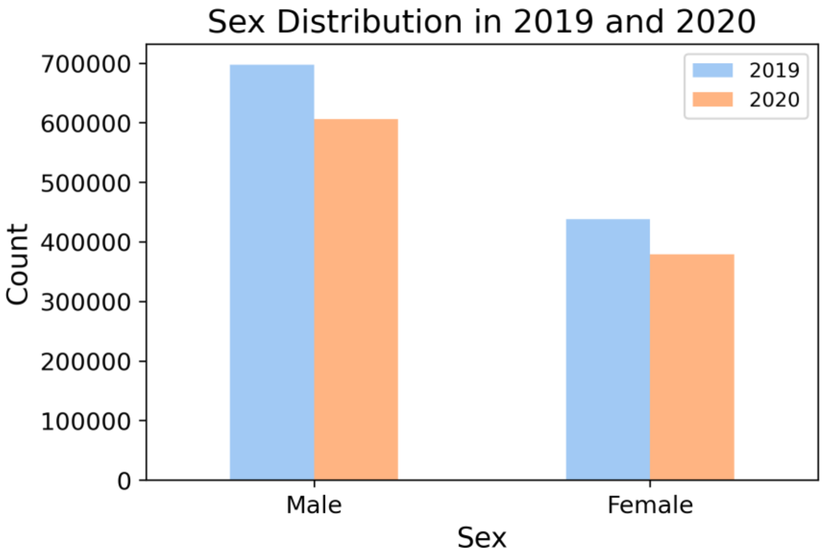


Figure 2 Sex Distribution in 2019 and 2020

In the above figures, 1 and 2 represents male and female sex respectively. It can be seen that the data contains more records from female than male. Just some information to consider. We can also see the Age VS. Sex density plots below.

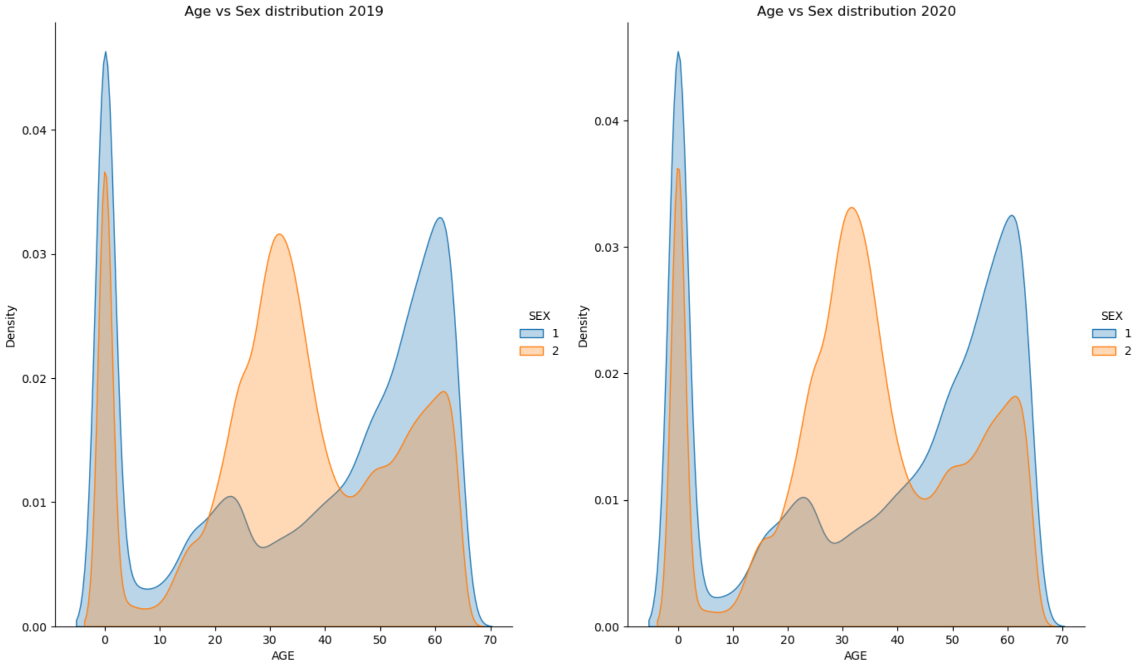


Figure 3 Age VS Sex Distribution in 2019 and 2020

Some of the most visited patients were also inspected to see if there is any interesting information.

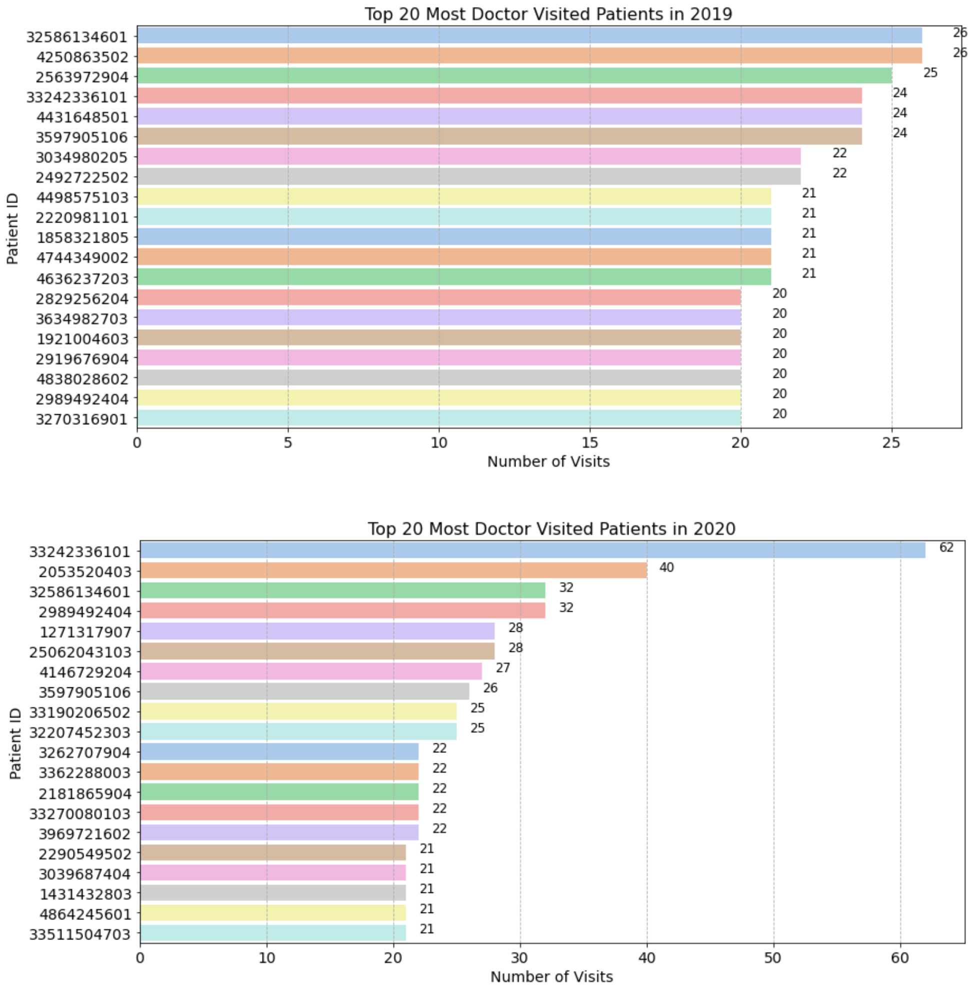


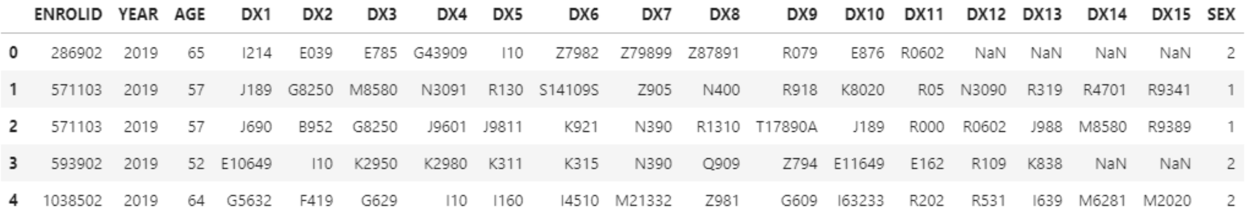
Figure 4 Top 20 Most Visited Patients in 2019 and 2020

However, it seems like it is not so important.

Preparation

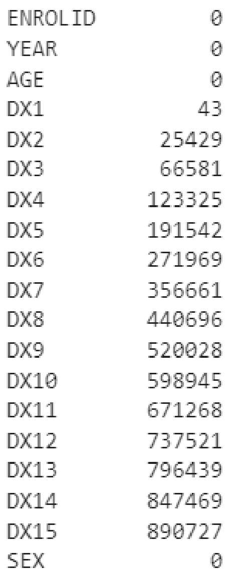
After loading the data into python data frame, the data frame looks like the table below,

Data before Feature Engineering:



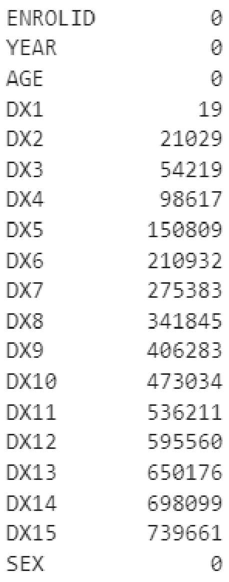
	ENROLID	YEAR	AGE	DX1	DX2	DX3	DX4	DX5	DX6	DX7	DX8	DX9	DX10	DX11	DX12	DX13	DX14	DX15	SEX
0	286902	2019	65	I214	E039	E785	G43909	I10	Z7982	Z79899	Z87891	R079	E876	R0602	NaN	NaN	NaN	NaN	2
1	571103	2019	57	J189	G8250	M8580	N3091	R130	S14109S	Z905	N400	R918	K8020	R05	N3090	R319	R4701	R9341	1
2	571103	2019	57	J690	B952	G8250	J9601	J9811	K921	N390	R1310	T17890A	J189	R000	R0602	J988	M8580	R9389	1
3	593902	2019	52	E10649	I10	K2950	K2980	K311	K315	N390	Q909	Z794	E11649	E162	R109	K838	NaN	NaN	2
4	1038502	2019	64	G5632	F419	G629	I10	I160	I4510	M21332	Z981	G609	I63233	R202	R531	I639	M6281	M2020	2

In the dataset for 2019 and 2020, there was no missing values for *ENROLID*, *AGE*, and *SEX*, however, many missing values appear in the diagnosis codes *DX* because not all 15 of the diagnosis code columns would be filled with diagnosis code. This could be because for some patients, just a few diagnoses code is enough to diagnose the patient. The number of missing values in the data tables can be seen below,



```
ENROLID         0
YEAR            0
AGE             0
DX1            43
DX2         25429
DX3         66581
DX4        123325
DX5        191542
DX6        271969
DX7        356661
DX8        440696
DX9        520028
DX10       598945
DX11       671268
DX12       737521
DX13       796439
DX14       847469
DX15       890727
SEX             0
```

Figure 5 Missing Values for 2019

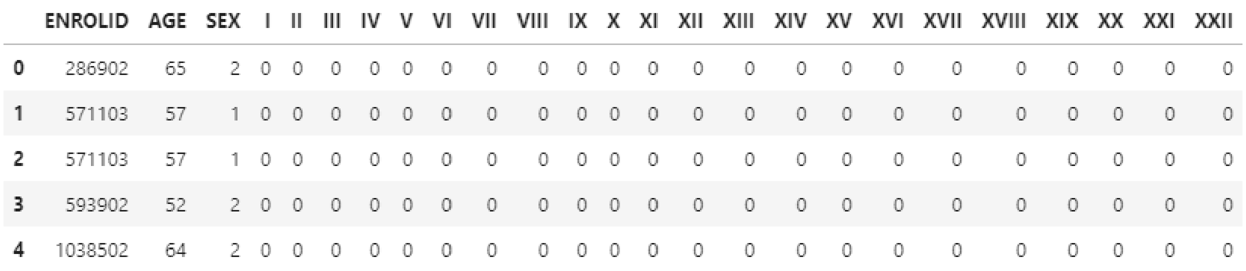


```
ENROLID         0
YEAR            0
AGE             0
DX1            19
DX2         21029
DX3         54219
DX4         98617
DX5        150809
DX6        210932
DX7        275383
DX8        341845
DX9        406283
DX10       473034
DX11       536211
DX12       595560
DX13       650176
DX14       698099
DX15       739661
SEX             0
```

Figure 6 Missing Values for 2020

These missing values were imputed by replacing them with '999', a value that would not interfere with interpretation of the disease codes.

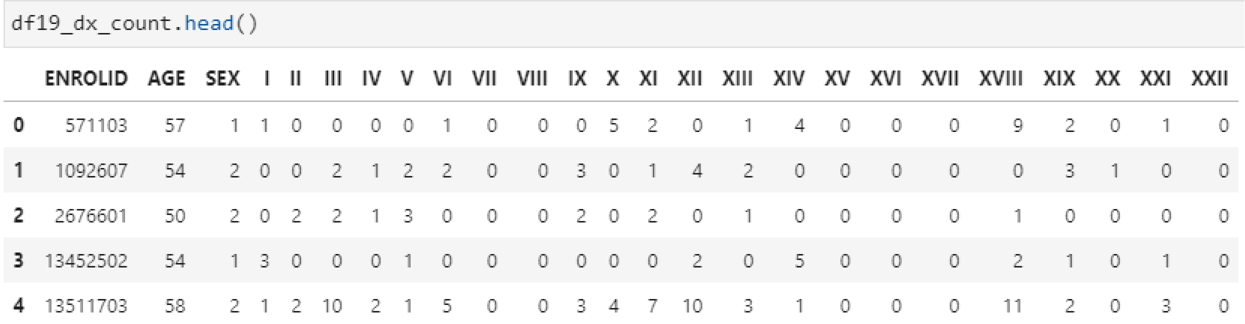
After the missing values were imputed, the next step was to create new features from the data. Since the diagnosis codes were recorded in the ICD-10-CM format, it may be very hard to train the model with that raw string data, as the possible combination of ICD-10-CM codes are nearly endless. Therefore, the features were transformed by counting the diagnosis codes and categorize them into their disease categories. In ICD-10-CM format, the codes can be divided into 22 major categories. Thus, new tables were created with new features as their disease counts. The example of the tables can be seen below,



	ENROLID	AGE	SEX	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	XXII
0	286902	65	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	571103	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	571103	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	593902	52	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1038502	64	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

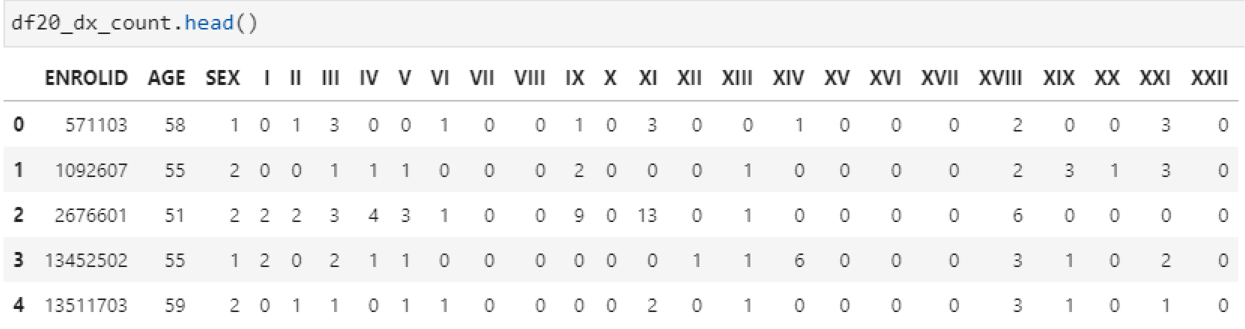
The roman numbers in the columns represent each disease category or chapter as described in a python ICD-10-CM package here: https://pypi.org/project/icd10-cm/. There is a table like the one above for both 2019 and 2020 data. Since we are interested in exploring the development of diseases related to the circulatory system, our focus would be on chapter IX, Diseases of the circulatory system. This also means that the patients must appear across both 2019 and 2020. Thus, inner join between the two tables was performed to get the data for patients with *ENROLID* that appeared in both years. Now that the two tables have the same patients in them, the disease counting can begin. icd10-cm python package from pypi.org was used to aid in categorizing the ICD-10-CM codes. For each instance in the table, if the code belongs to, for example, chapter X, then the value in column 'X' in the above table is incremented for that instance. The example of the final disease counts can be seen below.

Data after Feature Engineering:



```
df19_dx_count.head()
```

	ENROLID	AGE	SEX	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	XXII
0	571103	57	1	1	0	0	0	0	1	0	0	0	5	2	0	1	4	0	0	0	9	2	0	1	0
1	1092607	54	2	0	0	2	1	2	2	0	0	3	0	1	4	2	0	0	0	0	0	3	1	0	0
2	2676601	50	2	0	2	2	1	3	0	0	0	2	0	2	0	1	0	0	0	0	1	0	0	0	0
3	13452502	54	1	3	0	0	0	1	0	0	0	0	0	0	2	0	5	0	0	0	2	1	0	1	0
4	13511703	58	2	1	2	10	2	1	5	0	0	3	4	7	10	3	1	0	0	0	11	2	0	3	0



```
df20_dx_count.head()
```

	ENROLID	AGE	SEX	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	XXII
0	571103	58	1	0	1	3	0	0	1	0	0	1	0	3	0	0	1	0	0	0	2	0	0	3	0
1	1092607	55	2	0	0	1	1	1	0	0	0	2	0	0	0	1	0	0	0	0	2	3	1	3	0
2	2676601	51	2	2	2	3	4	3	1	0	0	9	0	13	0	1	0	0	0	0	6	0	0	0	0
3	13452502	55	1	2	0	2	1	1	0	0	0	0	0	0	1	1	6	0	0	0	3	1	0	2	0
4	13511703	59	2	0	1	1	0	1	1	0	0	0	0	2	0	1	0	0	0	0	3	1	0	1	0

Since chapter IX is the diseases of the circulatory system, we will also convert that column into a response variable for prediction.

From the disease count table, more exploratory analysis can be done such as diagnosis distribution. Some sample diagnosis distributions can be seen in the figure below.

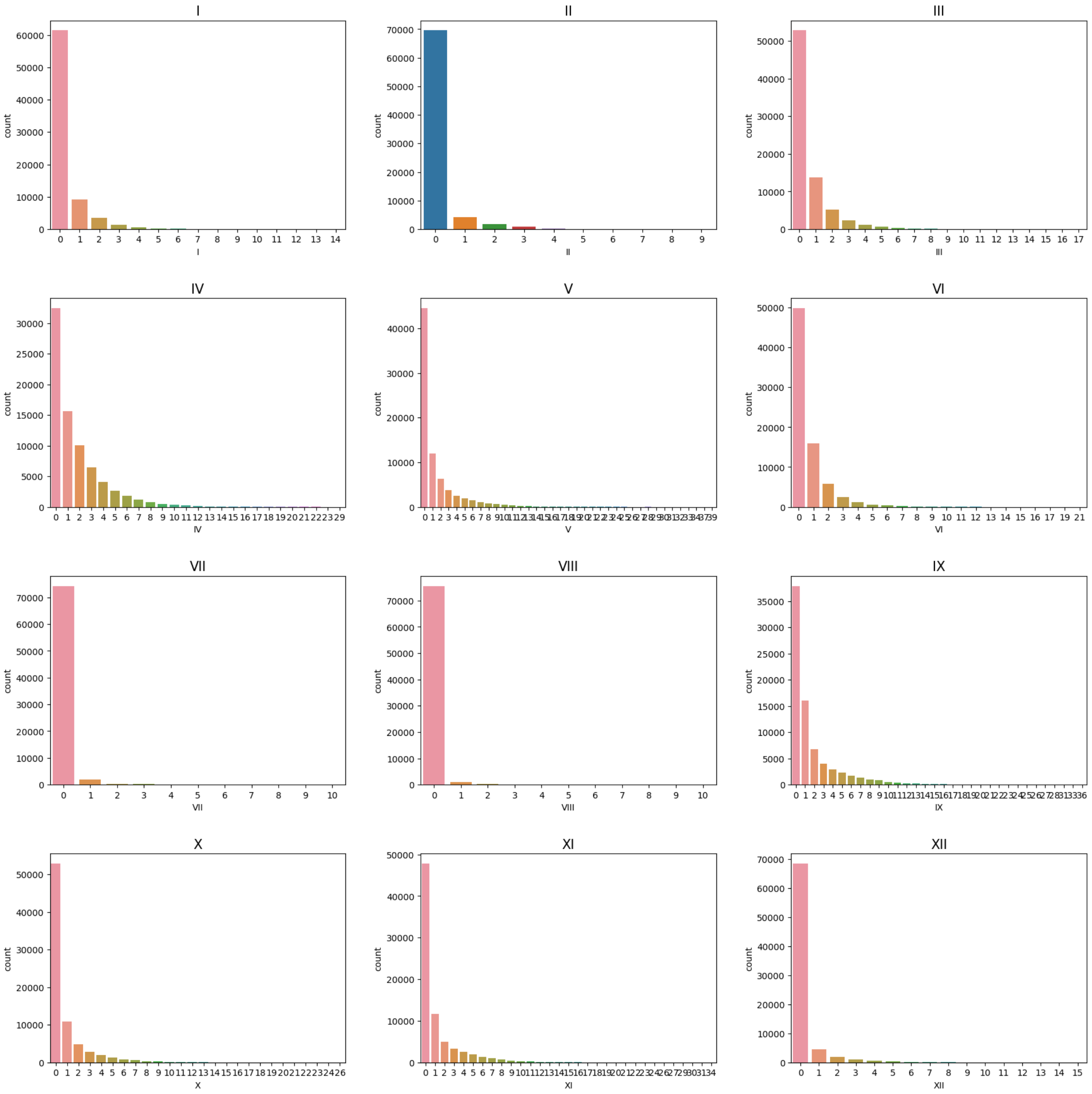


Figure 7 Distribution of each Diagnosis Chapter

The distribution of our focus chapter (IX) can also be seen in the figure below.

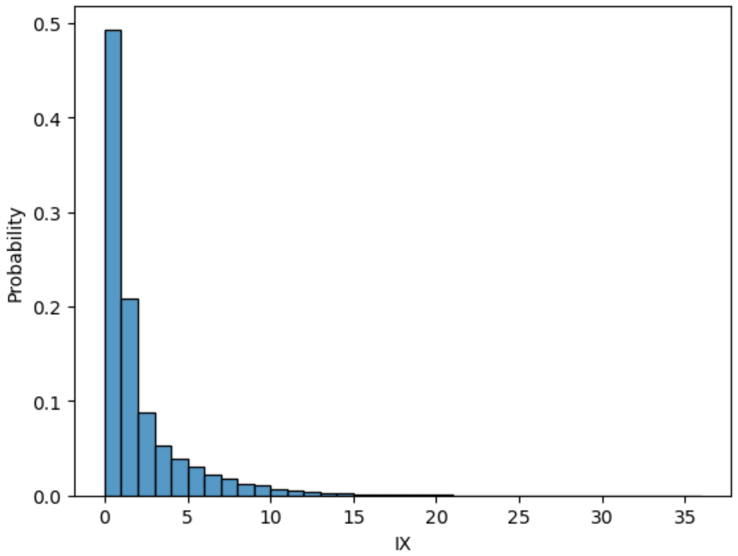


Figure 8 Distribution of Disease Count for Diseases of the Circulatory System

The correlation between chapter IX and other features of the table can also be plotted below to inspect which features are closely related to chapter IX. The correlation plot can be seen below.

Correlation Matrix:

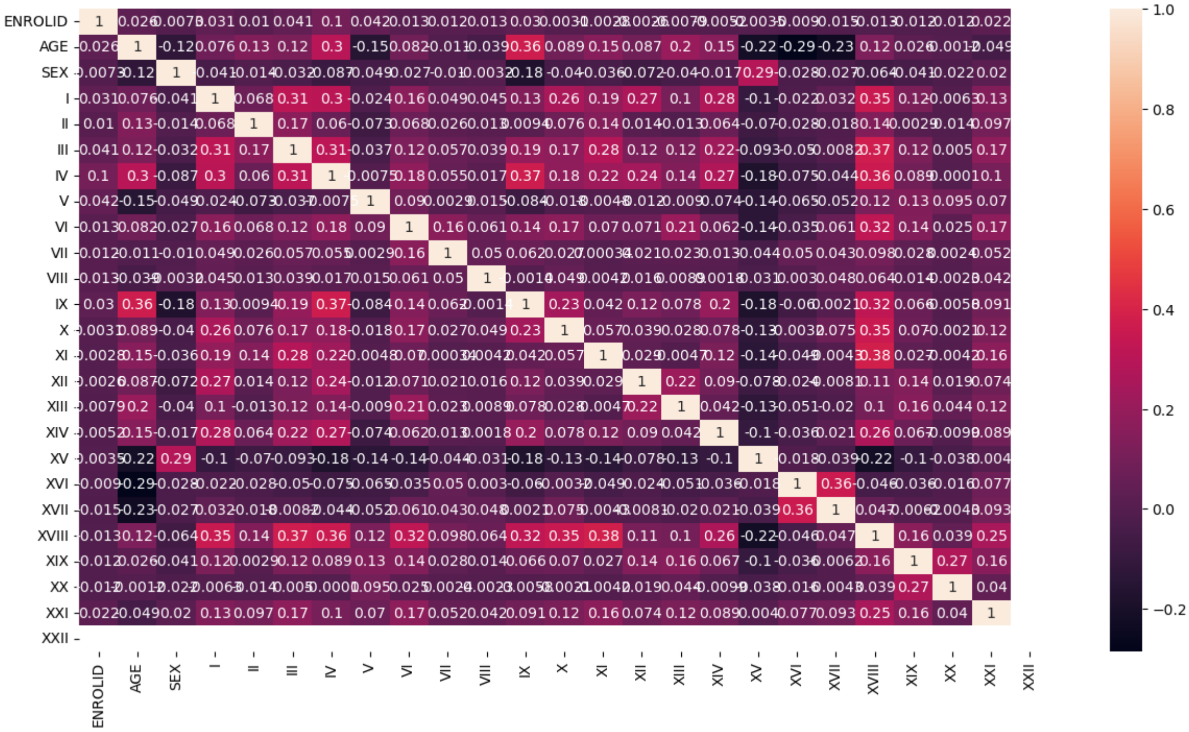


Figure 9 Correlation Matrix

Correlation Between Chapter IX and other Features:

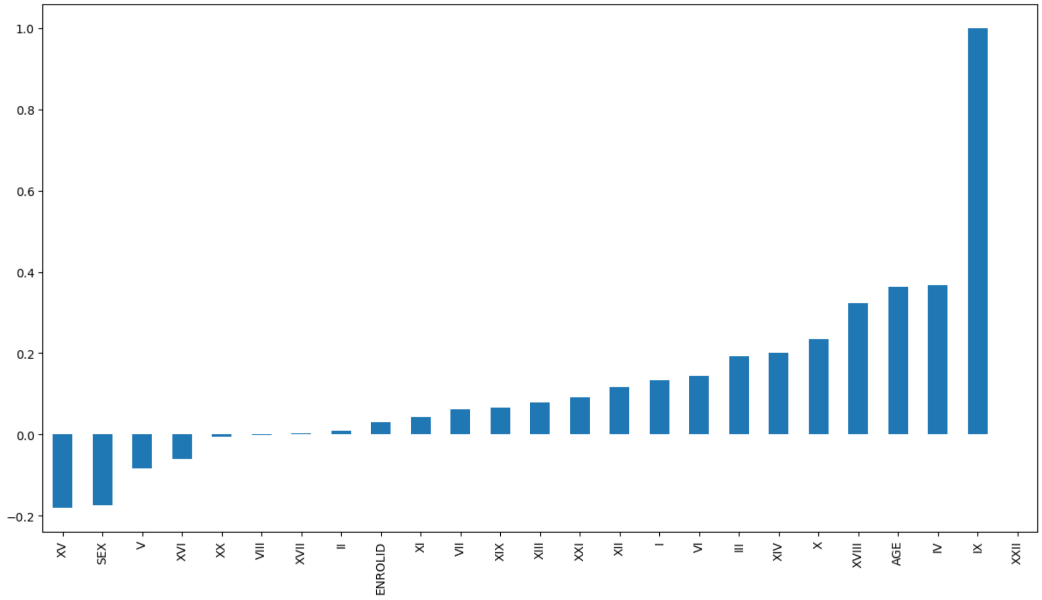


Figure 10 Correlation Between Chapter IX and other features

It can be observed that the top three features that are the most related to chapter IX are chapter IV, Age, and chapter XVIII. Chapter IV are Endocrine, nutritional and metabolic diseases and chapter XVIII are Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. Several features have negative correlation including Sex, chapter XX, XVI, V and XV. Chapter XX, XVI, V and XV corresponds to External causes of morbidity and mortality, Certain conditions originating in the perinatal period, Mental and behavioral disorders, and Pregnancy, childbirth and the puerperium respectively. The correlation matrix and their corresponding values can be seen below. Note that correlation with chapter XXII is NaN (missing) because there was no count for diagnosis codes that belongs to that chapter in the dataset.

Methodology

Techniques

The modeling technique used in this project was supervised machine learning, since the models learn from labeled training data to predict the class labels of new, unseen data. More specifically, classification models were created to predict if a pilot is at risk of having a disease of circulatory system. After chapter IX has been converted to a response variable (a label), we can used that as a training and testing label.

Our chapter IX data were transformed into binary class labels, where 0 represents the absence of a circulatory system disease and 1 indicates the presence of the disease. It was observed that the distribution of class labels was well-balanced, as depicted in the figure below.

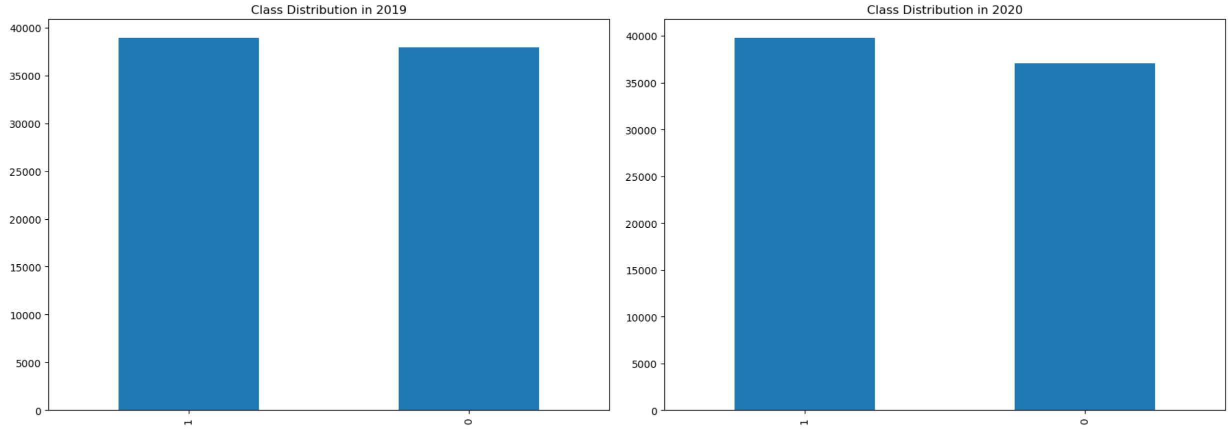


Figure 11 Class Label Distribution in Both Years

However, this is not the final data that would be used. Since we want to predict if the individual is at risk of having a disease related to circulatory system in the future (next year in this case), the two datasets need to be combined. We are simulating the present by using the 2019 data, and treating 2020 as the future, in order to analyze potential trends and make predictions based on the available data. So, a little more data transformation was done by paring 2019 features with 2020 labels. A clearer picture can be seen below.

Picture of way of combining 2019 and 2020 datasets:

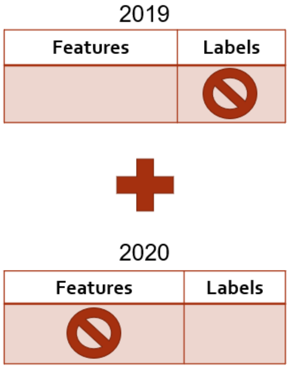


Figure 12 Combining 2019 and 2020 data.

Due to the enrollment criteria, the dataset size was reduced to only include patients who were enrolled in both 2019 and 2020. Furthermore, to model disease progression from one year to the next, only patients with no circulatory disease (as indicated by an IX count of '0') in 2019 were considered, resulting in a further reduction in dataset size. After merging the datasets from both years, the class distribution was plotted and revealed a class imbalance issue, as shown in the figure below. Specifically, the number of instances in the positive class was significantly smaller than the number of instances in the negative class, which may pose a challenge for modeling and prediction tasks.

Analysis of merged dataset:

Imbalance class distribution before SMOTE:

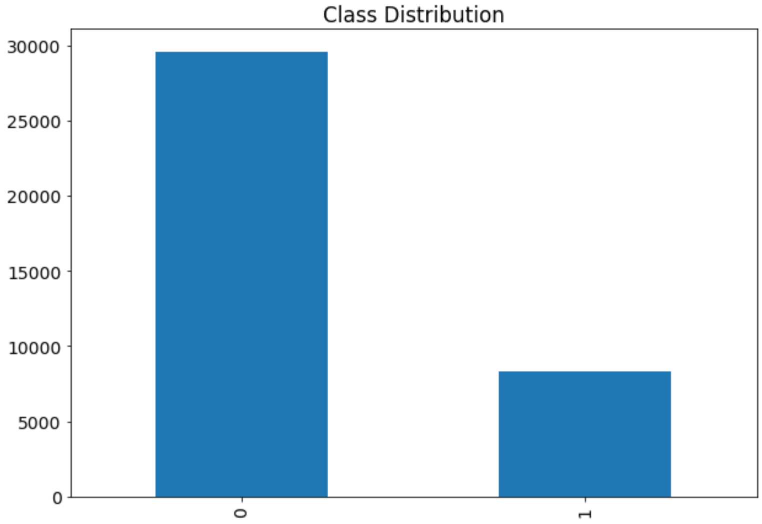


Figure 13 Class Distribution of the merged dataset

One way to tackle this challenge is by utilizing SMOTE (Synthetic Minority Over-sampling Technique). SMOTE is a widely used algorithm for addressing class imbalance by oversampling the minority class using synthetic samples [1]. In SMOTE, synthetic samples are generated by interpolating between existing minority class instances, creating new instances that lie along the line segments connecting pairs of instances [1]. By doing so, the number of instances in the minority class is increased, while preserving the distribution of the minority class.

The dataset was partitioned into training and testing sets using a 70:30 ratio. Before training the classification models on the training set, SMOTE was applied to address class imbalance. Specifically, SMOTE was applied only to the training set to generate synthetic examples of the minority class such that the distribution of the two classes was balanced. The class distribution of the training data can be seen in the figure below.

Balanced class distribution after SMOTE:

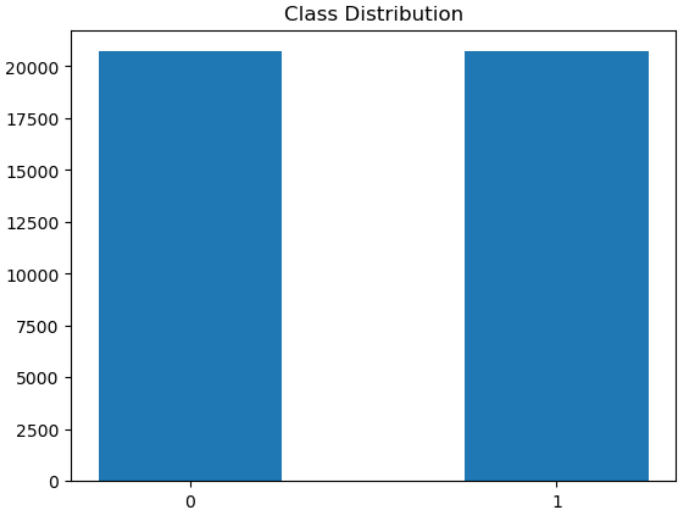


Figure 14 Class Distribution of SMOTE training set

Next, The classification models were then trained on the augmented training data using the 5-fold cross-validation scheme. Finally, the models' performance was evaluated on the testing set to assess their generalization ability.

To assess our models, we employed various metrics, including the confusion matrix, which provides valuable insights into the number of true positives, true negatives, false positives, and false negatives. This matrix can also be leveraged to derive other crucial metrics such as the False Positive Rate (Type I error), False Negative Rate (Type II error), Accuracy, Precision, and Recall. The Accuracy score is an appropriate metric to use when the class labels are balanced. However, when dealing with imbalanced data, such as in our case, a better metric to use is the F1 Score. As a binary classification metric, the F1 Score effectively combines both precision and recall into a single score.

Additionally, the Receiver Operating Characteristic (ROC) is another useful metric that can be employed in evaluating unbalanced data. The ROC curve plots the true positive rate against the false positive rate, and the area under the curve can provide an excellent measure of a model's performance. Precision-recall curve is also a useful metric for evaluating machine learning models, especially in cases of imbalanced datasets where the positive class is rare. The precision-recall curve is a graph that plots the precision (positive predictive value) against recall (true positive rate) at different probability thresholds. A high precision indicates that the model returns very few false positives, while a high recall indicates that the model returns most of the true positives. The area under the precision-recall curve (AUPRC) is another metric that can be used to compare models. A higher AUPRC indicates better performance, and this metric is particularly useful when the positive class is rare, as it puts more emphasis on recall than precision.

In this project, various classification models were employed, including Support Vector Machines (SVM), Neural Networks, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting. To explore the effectiveness of these models, we compared their performance with the mentioned metrics. Additionally, we fine-tuned the hyperparameters of each model using techniques such as GridSearchCV to obtain optimal results. GridSearchCV works by searching through a pre-defined set of hyperparameters and evaluating the performance of the model using each set of hyperparameters. It then selects the set of hyperparameters that results in the best performance based on a specified evaluation metric, such as accuracy, precision, recall, or F1 score.

In Machine Learning, SVM is a powerful algorithm that is often used for classification tasks. SVM works by finding the optimal hyperplane that separates data points into different classes while maximizing the margin between the classes. SVM is particularly useful when dealing with non-linearly separable data, as it can use kernel functions to map the data to a higher-dimensional space where it is easier to separate. SVM can also handle high-dimensional data and is relatively insensitive to outliers [2][3].

Neural Networks algorithms are inspired by the structure and function of the human brain. They consist of interconnected nodes or neurons that are organized into layers. Neural Networks can learn complex patterns and relationships in the data by adjusting the weights between neurons during the training process. Neural Networks are particularly useful for tasks such as image and speech recognition, where the input data has complex features and relationships [4]. Neural

Networks can be a good model for diagnosing disease because they can learn to identify complex patterns and relationships in medical data that may be difficult for human experts to recognize.

Logistic Regression is a simple yet powerful algorithm that is used for binary classification tasks. It works by modeling the probability of an instance belonging to a certain class using a logistic function. Logistic Regression is easy to implement and interpret. It is also relatively robust to noise and outliers [5].

KNN is a non-parametric algorithm that is used for classification tasks. KNN works by finding the k-nearest neighbors to a new data point and assigning the class label based on the majority vote of the neighbors. KNN is particularly useful when the decision boundary between classes is non-linear and complex. KNN is also robust to noise and outliers but can be computationally expensive for large datasets[6].

Random Forest is an ensemble learning method that combines multiple decision trees to improve the accuracy and stability of the classification. Random Forest works by randomly sampling the training data and features to create multiple trees, and then aggregating their predictions. Random Forest is particularly useful when dealing with high-dimensional data and complex decision boundaries. Random Forest is also robust to overfitting and can handle missing data [7].

Bagging is an ensemble learning technique, often referred to as bootstrap aggregation, is frequently used to lessen variance within a noisy dataset. In bagging, a training set's data is sampled at random with replacement. These weak models are subsequently trained independently after several data samples have been collected. Bagging is used to generate random forests more effectively, improving the precision of predictions[14].

Gradient Boosting is another ensemble learning method that is used for classification tasks. Gradient Boosting works by iteratively training weak models and combining their predictions to form a strong model. Gradient Boosting is particularly useful when dealing with imbalanced data and when the decision boundary is complex. Gradient Boosting can also handle missing data and is relatively robust to noise and outliers [8].

XGBoost (Extreme Gradient Boosting) was also explored in this project. It is an extension of Gradient Boosting that incorporates a regularization term in the objective function to prevent overfitting. XGBoost can handle missing data, is computationally efficient, and has been shown to achieve state-of-the-art performance on various machine learning tasks [9].

Procedure

Support Vector Machines (SVM)

In this project, we employed the Support Vector Classification (SVC) algorithm from the scikit-learn (sklearn) library, which provides an implementation of SVM. To optimize the performance of the classifier, we used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of a regularization parameter (C) equal to 1.0, a radial basis function kernel (rbf), and a gamma coefficient set to 'scale'.

Neural Networks (NNs)

For our Neural Networks model, we utilized the PyTorch library along with the skorch library to integrate PyTorch NNs into scikit-learn for easier GridSearchCV hyperparameter tuning. Our initial NN model consisted of two layers with 50 hidden neurons and a simple forward function. After hyperparameter tuning, we added a dropout rate of 0.1 to prevent overfitting and a weight constraint of 2.0 to avoid exploding gradients during training. We used the binary cross-entropy loss function (BCELoss) as our loss function, Adagrad as the optimizer, and rectified linear unit (ReLU) as the hidden layer activation function. In addition to these hyperparameters, we also tuned the learning rate, setting it to 0.01. The number of neurons were also tuned to 500 neurons per hidden layer. Finally, we set a maximum of 1000 epochs and a batch size of 100. After extensive experimentation using GridSearchCV, this combination of hyperparameters was optimal for our model.

Logistic Regression

For our logistic Regression model, we utilized the scikit-learn(sklearn) library that provides implementation of logistic. To optimize the performance of the classifier, we used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of a regularization parameter (C) equal to 110, norm of the penalty to l2 and solver to optimize model as liblinear.

K-Nearest Neighbors (KNN)

For our K-Nearest Neighbors model, we utilized the scikit-learn(sklearn) library that provides implementation of KNN. We used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of number of iterations(n_neighbors) to 2, weights to uniform, algorithm to ball_tree and leaf size to 20.

Random Forest & Bagging

For our Random Forest with bagging model, we utilized the scikit-learn(sklearn) library that provides implementation of Random Forest using bagging. We used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of number of iterations(n_estimators) to 100, maximum features used(max_features) to 10, maximum samples used (max_samples) to 100.

XGBoost

For our XGBoost model we utilized the xgboost library that provides implementation of xgboost. To optimize the performance of the classifier, we used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of a learning rate to 0.5, maximum depth of tree(max_depth) to 50 and number of iterations(n_estimators) to 500.

Gradient Boosting

For our Gradient Boosting model, we utilized the scikit-learn(sklearn) library that provides implementation of Gradient Boosting. To optimize the performance of the classifier, we used the GridSearchCV function, which allowed us to explore different hyperparameters and select the best combination based on cross-validation. After this process, we identified the optimal set of hyperparameters for our model, which consists of a learning rate to 0.05, maximum depth of tree(max_depth) to 11 and number of iterations(n_estimators) to 800.

Results and Analysis

The primary objective of this project is to develop a model that can accurately predict the risk of circulatory diseases. As the dataset used in this study is unbalanced, the evaluation criteria for the models are selected to account for this imbalance. The following performance metrics will be used to assess the models' performance:

- F1 score
- Area Under the ROC Curve (AUCROC)
- Area Under the Precision-Recall Curve (AUPRC)
- Confusion matrix

The AUCROC score is a widely used performance metric that summarizes the overall ability of the model to distinguish between positive and negative classes. The score ranges from 0 to 1, where a score of 0.5 indicates random guessing, and a score of 1 indicates perfect discrimination. A score greater than 0.5 and less than 1 indicates that the model is able to distinguish between positive and negative classes with some degree of accuracy. In other words, the higher the AUCROC score, the better the model's ability to distinguish between the two classes. Conversely, a score less than or equal to 0.5 suggests that the model is not able to discriminate between the two classes and is performing worse than random guessing. In this case, the model requires further investigation and improvement to be useful for practical applications.

As the dataset is imbalanced, the F1 score is a more reliable performance metric as it equally weights both the false positives and false negatives. In particular, the F1 score is useful for assessing the accuracy of the anticipated positive predictions, which is often of particular interest in the context of disease diagnosis or risk assessment.

In this project, our primary objective is to accurately predict the risk of circulatory diseases, which requires minimizing the number of false negatives (FN) and maximizing the number of true positives (TP). False negatives correspond to cases where the model incorrectly predicts that an individual does not have the disease when they actually do, which can have serious consequences in terms of patient health outcomes. Therefore, we will focus on minimizing the FN rate and increasing the TP rate when evaluating the models using the confusion matrix.

Comparison Area under curve of ROC and Precision Recall curve

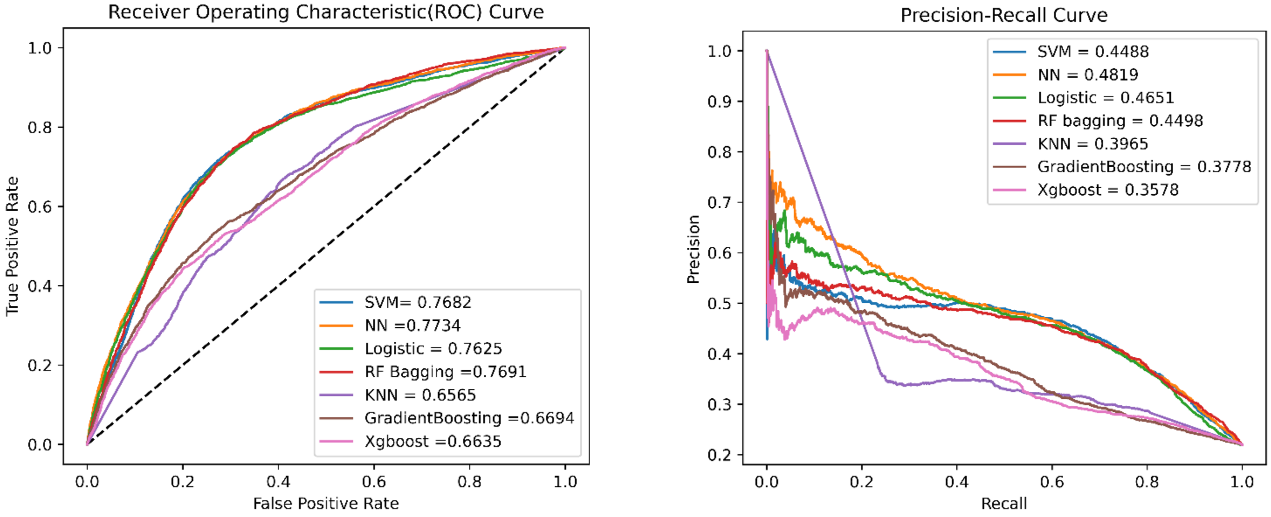


Figure 15 Receiver Operating Characteristics Curve (ROC) and Precision-Recall Curve

On the left, we have an ROC curve that illustrates the performance of different models. Based on the curve, the top-performing models are neural networks, SVM, logistic regression, and Random Forest, as indicated by the larger area under their respective curves, which can be seen in the legend. Moving to the right, we have the precision-recall curve, which provides a more apparent distinction between the model performances. The neural network emerges as the best-performing model based on its higher area under the curve, as shown in the legend. Overall, the ROC curve and the precision-recall curve offer complementary insights into the model performance, allowing us to select the optimal model for our use case..

Overall models Performance

Table 1 Model Results

Model	**Package**	**Hyperparameter**	**Accuracy**	**F1 Score**	**ROC AUC**	**PRAUC**
SVM	sklearn.svm.SVC	C: 1.0, kernel: rbf, gamma: ‘scale’	0.72	0.67	0.768	0.449
Neural Networks	PyTorch	Layers: 2, Hidden Neurons: 23 -> 500 -> 1, Optimizer: Adagrad, Loss function: BCELoss, Dropout rate: 0.1, Learning Rate: 0.01,	0.69	0.64	**0.773**	**0.481**
Logistic Regression	sklearn.linear_model	'C': 0.01, 'penalty': 'l2’, 'solver': 'liblinear'	0.71	0.66	0.763	0.465
Random Forest	sklearn.ensemble	bagging max_features : 10, n_estimators : 100, Max_samples = 100	0.72	0.66	0.769	0.450

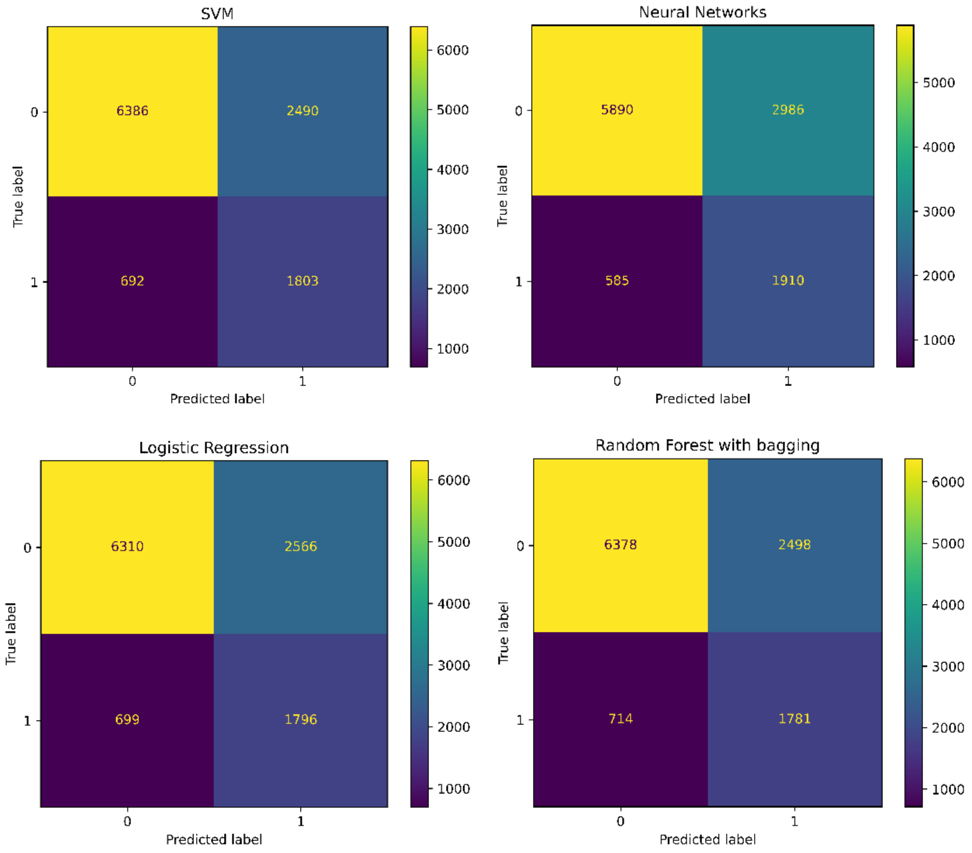
Model	Package	Hyperparameter	Accuracy	F1 Score	ROC AUC	PRAUC
KNN	sklearn.neighbors	algorithm='brute’ leaf_size=10, n_neighbors=4, weights='distance'	0.66	0.58	0.655	0.40
GradientBoosting	sklearn.ensemble	learning_rate : 0.05, max_depth : 11, n_estimators : 800	0.72	0.61	0.668	0.376
Xgboost	xgboost	'learning_rate': 0.5, 'max_depth': 50, 'n_estimators': 500	0.72	0.61	0.663	0.358

The table above presents the results of our experiments. For each model, these are the packages that we used. And their hyperparameters after grid search hyperparameter tuning. Their accuracy, f1-score, area under the ROC, and area under the precision-recall curve can also be seen in this table.

From the above table, SVM, Neural Network, Logistic Regression, and Random Forest with bagging have the highest AUCROC, AUPRC and F1 score. It can also be said that the best performing model is neural networks followed closely by SVM or Random forest. Logistic Regression also performed surprisingly well when compared to others.

Next, the models are compared and understand if they have the lowest False Negative in the confusion matrix.

Confusion matric for each model



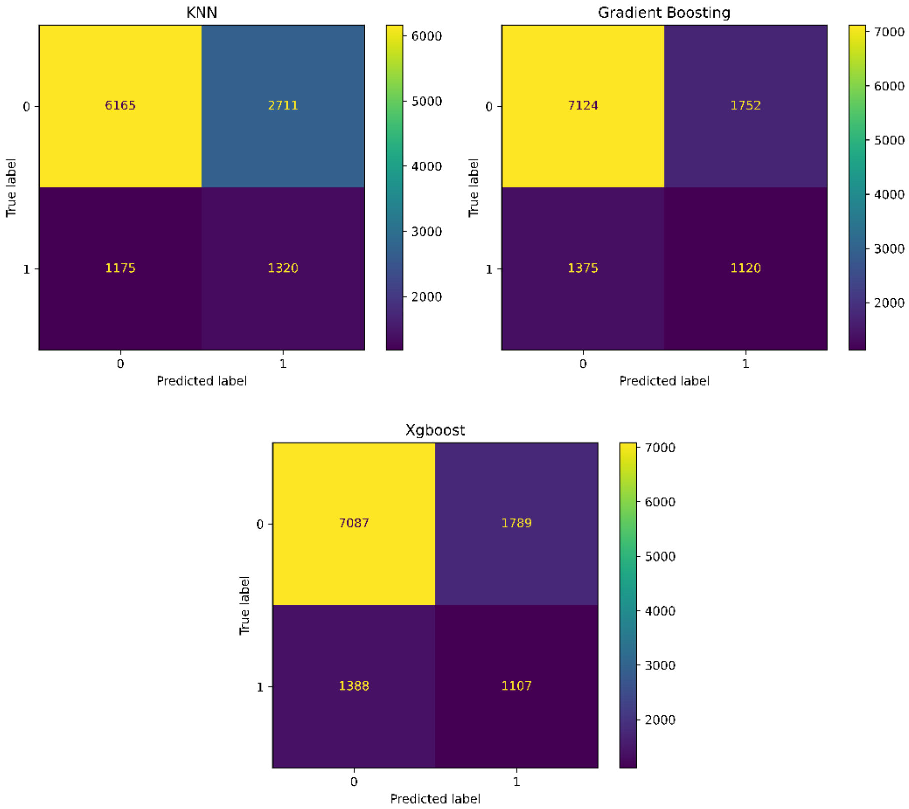


Figure 16 Confusion Matrix for each model

Figure 16 shows that the top 4 models selected have fewer False Negatives, which is a crucial metric for our problem. A closer analysis of the confusion matrices indicates that the neural network model stands out with a higher True Positive (TP) rate and a lower False Negative (FN) rate. This outcome is highly desirable as False Negatives can lead to missed diagnoses, delayed treatments, and potentially harmful health outcomes for pilots. Although the neural network model has a higher False Positive (FP) rate, this is a minor concern, as it is generally safer to have a False Positive than a False Negative when dealing with medical conditions. In the case of aviation, the consequences of a missed diagnosis could be severe, potentially endangering not only the pilot's life but also the passengers and crew onboard. Therefore, the neural network model's higher FP rate is an acceptable trade-off given the critical importance of reducing the risk of False Negatives.

Overall, the neural network model's superior performance in reducing False Negatives and improving True Positive rates makes it the most effective model for predicting the risk of circulatory diseases in pilots.

Deliverables

Based on our analysis, we can leverage machine learning techniques to predict whether individuals are likely to transition from a healthy state to an unhealthy one or vice versa, based on one safety-critical task related to the diagnosis category of circulatory diseases. Although our dataset does not solely pertain to pilots, applying this concept to the FAA's objectives aligns with their interests in identifying whether a medical issue is present or absent. To identify the diagnosis category or medical condition that affects performing safety-critical tasks in aviation environments, domain expertise is necessary. However, in our experiment, we focused on exploring the probability of individuals transitioning between health statuses. Despite the relatively low model accuracy, we can improve it by incorporating more details from individual medical records, such as weight, height, blood pressure, and other relevant features available in the IBM database. Our primary objective is to propose a concept that can be further refined and enhanced with additional data and insights.

Some of the problems we have encountered in this project.

During the project, there were several challenges that we encountered. Firstly, the IBM database contained numerous tables that were not solely focused on pilots. This made it difficult to track individuals' diagnoses in the circulatory system category from 2019 and determine whether or not these patients would move to the circulatory system in 2020. Finding the relevant table proved to be a challenging task.

Additionally, the large size of the data (100 GB) made it difficult to merge and concatenate useful tables that we thought would help in building the model to track individuals across two years. For instance, we tried merging inpatient and outpatient service tables, but we could only track individuals in one year. However, the Inpatient Admission, Prescription Drug, and Red Book tables were the only ones that could track individuals through different years and provide relevant information as predictors.

Lastly, deciding which features to include in the final dataset was also a challenge. The original dataset contained numerous features related to payments and health plans, which were not relevant to our project. As we were not medical experts, we relied on diagnosis codes to contribute the most to the project. We also included demographic variables such as gender and age to improve model accuracy. Despite these challenges, we were able to complete the project and obtain some meaningful insights.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321--357, 2002.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273--297, 1995.

[3] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1--27, 2011.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436--444, 2015.

[5] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant, *Applied logistic regression*. Wiley, 2013.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21--27, 1967.

[7] L. Breiman, "Random forests," \emph{Machine Learning}, vol. 45, no. 1, pp. 5--32, 2001.

[8] J. Friedman, "Greedy function approximation: A gradient boosting machine," \emph{Annals of Statistics}, vol. 29, no. 5, pp. 1189--1232, 2001.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in \emph{Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining}, pp. 785--794, 2016.

[10] Commercial Claims and Encounters Medicare Supplemental

[11] ICD-10-CM python package: https://pypi.org/project/icd10-cm/

[12] https://neptune.ai/blog/evaluation-metrics-binary-classification

[13] https://towardsdatascience.com/the-explanation-you-need-on-binary-classification-metrics-321d280b590f

[14] Bagging technique - https://www.ibm.com/topics/bagging

[15] Previous work - FAA-Pliot-License-Renewal-Forecasting/DSA Final Report.pdf at main · JUmaMaheshwarReddy/FAA-Pliot-License-Renewal-Forecasting · GitHub

Self-Assessment

Technical Project Objectives:

1. Analyze and understand medical records, evaluate existing databases and approaches, and identify critical features for accurate prediction.
2. Validate and adjust existing codebase before modifying ML techniques, addressing class imbalance issues, and redesigning feature engineering.
3. Develop and compare various predictive models and create a software tool with an interface to display visualization of factors affecting prediction.

Individual Learning Objectives:

1. Understand metadata and data dictionaries to process valuable data.
2. Gain experience with real-time data and its impact on daily lives.
3. Evaluate efficient database structures for storing data.
4. Enhance knowledge of modelling pipeline, including data preprocessing, model selection, training, and hyperparameter tuning.
5. Learn to use visualization tools to present findings and receive feedback.
6. Explore incorporation of prediction results into a simple web application with an interactive graphical user interface.

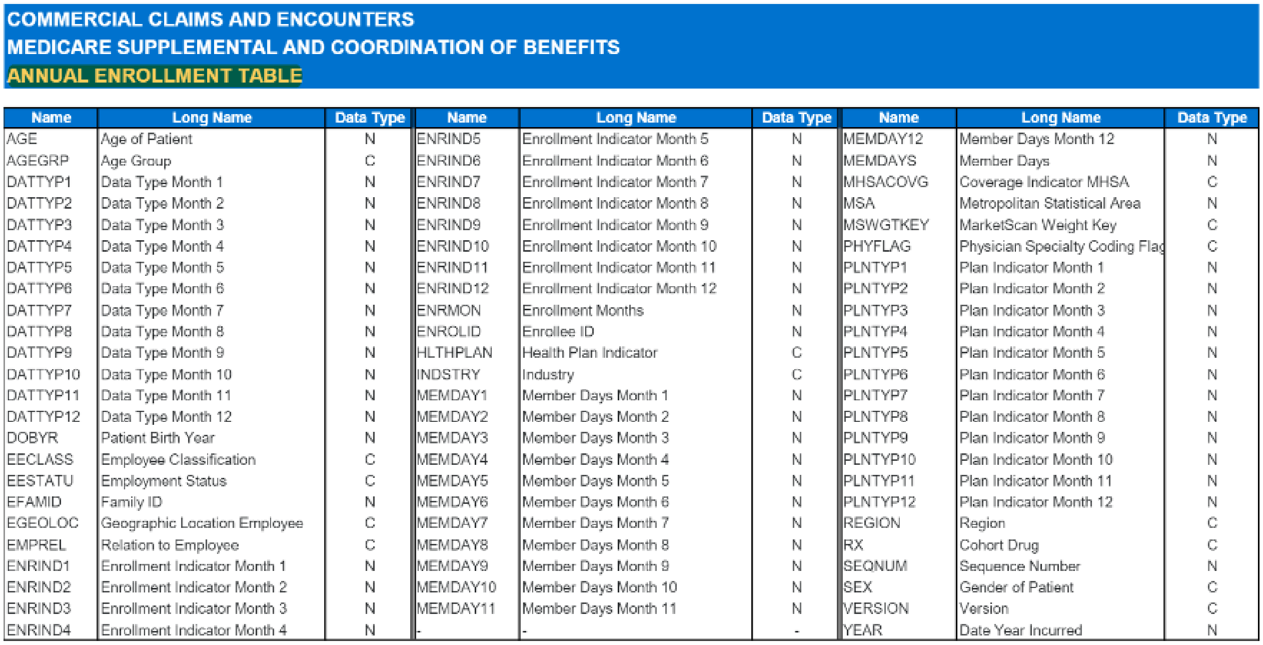
During our project, we successfully accomplished most of our initial objectives within the given timeframe of one semester. We analyzed and understood medical records, evaluated existing databases and approaches, identified critical features for accurate prediction, validated and adjusted existing codebase, and developed and compared various predictive models. Although we did not have time to develop a web application with a graphical user interface (GUI), we were able to meet our other objectives.

Throughout the project, we found several data science and analytics (DSA) skills to be particularly useful. Data cleaning and preprocessing were critical for ensuring that the data was ready for analysis. Exploratory data analysis (EDA) techniques helped us understand the data and identify important features. Feature engineering allowed us to create meaningful predictors for the predictive models, and model selection, training, and hyperparameter tuning helped us build and optimize accurate models. Finally, visualization skills enabled us to effectively communicate the results of the project to our stakeholders.

Although we had learned many of the necessary data science and analytics (DSA) skills throughout our graduate studies, we found that some of these skills were easy to forget without practice. As a result, we had to refresh our knowledge and learn some new skills independently to complete this project successfully. For example, we had to learn how to handle class imbalance issues in predictive modeling and how to redesign feature engineering to improve model performance. We also had to explore different approaches to hyperparameter tuning, such as using grid search or manually by hand, to find optimal model parameters. In addition, we had to learn how to use specific libraries, such as imbalanced-learn, scikit-learn and PyTorch, to implement these techniques effectively.

Furthermore, we had to learn how to create clear and effective visualizations to communicate our findings to stakeholders. We had to research and experiment with different types of charts, graphs, and other visualization tools, such as Tableau or Power BI, to determine which ones would best convey our results. Finally, we had to learn how to use Jupyter Notebook, GitHub, and other collaborative tools to work effectively as a team and manage our codebase. By learning these skills independently, we were able to apply our knowledge and complete the project successfully, demonstrating our ability to adapt to new challenges and solve complex problems.

Appendix



COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
ANNUAL ENROLLMENT TABLE

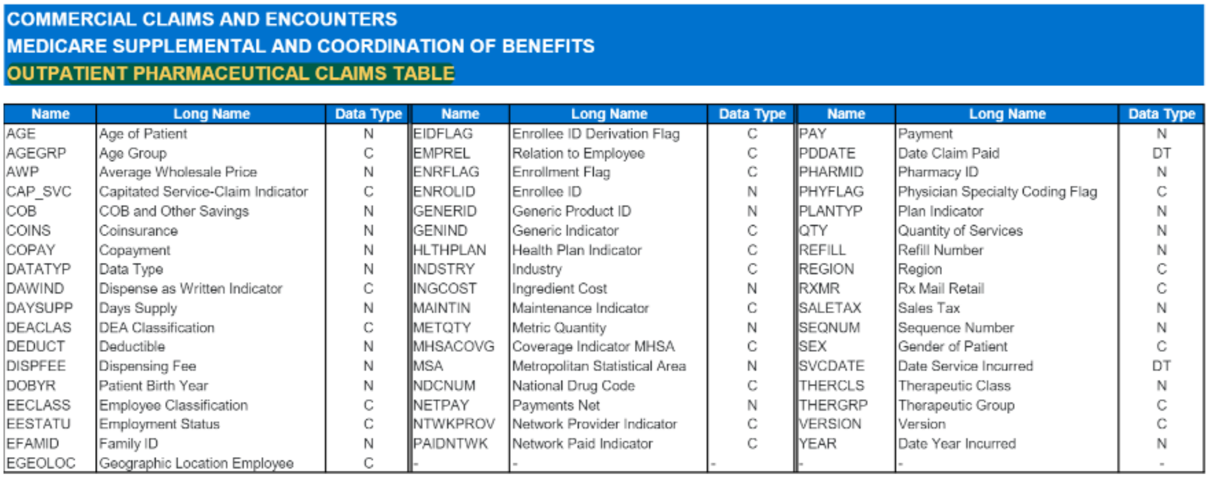
Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	ENRIND5	Enrollment Indicator Month 5	N	MEMDAY12	Member Days Month 12	N
AGEGRP	Age Group	C	ENRIND6	Enrollment Indicator Month 6	N	MEMDAYS	Member Days	N
DATTYP1	Data Type Month 1	N	ENRIND7	Enrollment Indicator Month 7	N	MHSACOVG	Coverage Indicator MHSA	C
DATTYP2	Data Type Month 2	N	ENRIND8	Enrollment Indicator Month 8	N	MSA	Metropolitan Statistical Area	N
DATTYP3	Data Type Month 3	N	ENRIND9	Enrollment Indicator Month 9	N	MSWGTKEY	MarketScan Weight Key	C
DATTYP4	Data Type Month 4	N	ENRIND10	Enrollment Indicator Month 10	N	PHYFLAG	Physician Specialty Coding Flag	C
DATTYP5	Data Type Month 5	N	ENRIND11	Enrollment Indicator Month 11	N	PLNTYP1	Plan Indicator Month 1	N
DATTYP6	Data Type Month 6	N	ENRIND12	Enrollment Indicator Month 12	N	PLNTYP2	Plan Indicator Month 2	N
DATTYP7	Data Type Month 7	N	ENRMON	Enrollment Months	N	PLNTYP3	Plan Indicator Month 3	N
DATTYP8	Data Type Month 8	N	ENROLID	Enrollee ID	N	PLNTYP4	Plan Indicator Month 4	N
DATTYP9	Data Type Month 9	N	HLTHPLAN	Health Plan Indicator	C	PLNTYP5	Plan Indicator Month 5	N
DATTYP10	Data Type Month 10	N	INDSTRY	Industry	C	PLNTYP6	Plan Indicator Month 6	N
DATTYP11	Data Type Month 11	N	MEMDAY1	Member Days Month 1	N	PLNTYP7	Plan Indicator Month 7	N
DATTYP12	Data Type Month 12	N	MEMDAY2	Member Days Month 2	N	PLNTYP8	Plan Indicator Month 8	N
DOBYR	Patient Birth Year	N	MEMDAY3	Member Days Month 3	N	PLNTYP9	Plan Indicator Month 9	N
EECLASS	Employee Classification	C	MEMDAY4	Member Days Month 4	N	PLNTYP10	Plan Indicator Month 10	N
EESTATU	Employment Status	C	MEMDAY5	Member Days Month 5	N	PLNTYP11	Plan Indicator Month 11	N
EFAMID	Family ID	N	MEMDAY6	Member Days Month 6	N	PLNTYP12	Plan Indicator Month 12	N
EGEOLOC	Geographic Location Employee	C	MEMDAY7	Member Days Month 7	N	REGION	Region	C
EMPREL	Relation to Employee	C	MEMDAY8	Member Days Month 8	N	RX	Cohort Drug	C
ENRIND1	Enrollment Indicator Month 1	N	MEMDAY9	Member Days Month 9	N	SEQNUM	Sequence Number	N
ENRIND2	Enrollment Indicator Month 2	N	MEMDAY10	Member Days Month 10	N	SEX	Gender of Patient	C
ENRIND3	Enrollment Indicator Month 3	N	MEMDAY11	Member Days Month 11	N	VERSION	Version	C
ENRIND4	Enrollment Indicator Month 4	N	-	-	-	YEAR	Date Year Incurred	N

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	EFAMID	EFAMID	EFAMID
4.	ENROLID	ENROLID	ENROLID
5.	MEMDAYS	MEMDAYS	MEMDAYS
6.	YEAR	YEAR	YEAR
7.	AGE	AGE	AGE
8.	DOBYR	DOBYR	DOBYR
9.	AGEGRP	AGEGRP	AGEGRP
10.	EMPREL	EMPREL	EMPREL
11.	PHYFLAG	PHYFLAG	PHYFLAG
12.	RX	RX	RX
13.	SEX	SEX	SEX
14.	HLTHPLAN	HLTHPLAN	HLTHPLAN
15.	ENRMON	ENRMON	ENRMON
16.	DATTYP1	DATTYP1	DATTYP1
17.	DATTYP2	DATTYP2	DATTYP2
18.	DATTYP3	DATTYP3	DATTYP3
19.	DATTYP4	DATTYP4	DATTYP4
20.	DATTYP5	DATTYP5	DATTYP5
21.	DATTYP6	DATTYP6	DATTYP6
22.	DATTYP7	DATTYP7	DATTYP7
23.	DATTYP8	DATTYP8	DATTYP8
24.	DATTYP9	DATTYP9	DATTYP9
25.	DATTYP10	DATTYP10	DATTYP10
26.	DATTYP11	DATTYP11	DATTYP11
27.	DATTYP12	DATTYP12	DATTYP12
28.	ENRIND1	ENRIND1	ENRIND1
29.	ENRIND2	ENRIND2	ENRIND2

30.	ENRIND3	ENRIND3	ENRIND3
31.	ENRIND4	ENRIND4	ENRIND4
32.	ENRIND5	ENRIND5	ENRIND5
33.	ENRIND6	ENRIND6	ENRIND6
34.	ENRIND7	ENRIND7	ENRIND7
35.	ENRIND8	ENRIND8	ENRIND8
36.	ENRIND9	ENRIND9	ENRIND9
37.	ENRIND10	ENRIND10	ENRIND10
38.	ENRIND11	ENRIND11	ENRIND11
39.	ENRIND12	ENRIND12	ENRIND12
40.	MEMDAY1	MEMDAY1	MEMDAY1
41.	MEMDAY2	MEMDAY2	MEMDAY2
42.	MEMDAY3	MEMDAY3	MEMDAY3
43.	MEMDAY4	MEMDAY4	MEMDAY4
44.	MEMDAY5	MEMDAY5	MEMDAY5
45.	MEMDAY6	MEMDAY6	MEMDAY6
46.	MEMDAY7	MEMDAY7	MEMDAY7
47.	MEMDAY8	MEMDAY8	MEMDAY8
48.	MEMDAY9	MEMDAY9	MEMDAY9
49.	MEMDAY10	MEMDAY10	MEMDAY10
50.	MEMDAY11	MEMDAY11	MEMDAY11
51.	MEMDAY12	MEMDAY12	MEMDAY12
52.	PLNTYP1	PLNTYP1	PLNTYP1
53.	PLNTYP2	PLNTYP2	PLNTYP2
54.	PLNTYP3	PLNTYP3	PLNTYP3
55.	PLNTYP4	PLNTYP4	PLNTYP4
56.	PLNTYP5	PLNTYP5	PLNTYP5
57.	PLNTYP6	PLNTYP6	PLNTYP6
58.	PLNTYP7	PLNTYP7	PLNTYP7
59.	PLNTYP8	PLNTYP8	PLNTYP8
60.	PLNTYP9	PLNTYP9	PLNTYP9
61.	PLNTYP10	PLNTYP10	PLNTYP10
62.	PLNTYP11	PLNTYP11	PLNTYP11
63.	PLNTYP12	PLNTYP12	PLNTYP12
64.	EECLASS	EECLASS	EECLASS
65.	EESTATU	EESTATU	EESTATU
66.	EGEOLOC	EGEOLOC	EGEOLOC
67.	INDSTRY	INDSTRY	INDSTRY
68.	MHSACOVG	MHSACOVG	MHSACOVG
69.	MSA	MSA	MSA
70.	REGION	REGION	REGION
71.	MSWGTKEY	MSWGTKEY	MSWGTKEY
72.			MEDADV1
73.			MEDADV2
74.			MEDADV3
75.			MEDADV4
76.			MEDADV5
77.			MEDADV6
78.			MEDADV7
79.			MEDADV8
80.			MEDADV9
81.			MEDADV10

82.			MEDADV11
83.			MEDADV12

CCAED(Prescription Drug D)



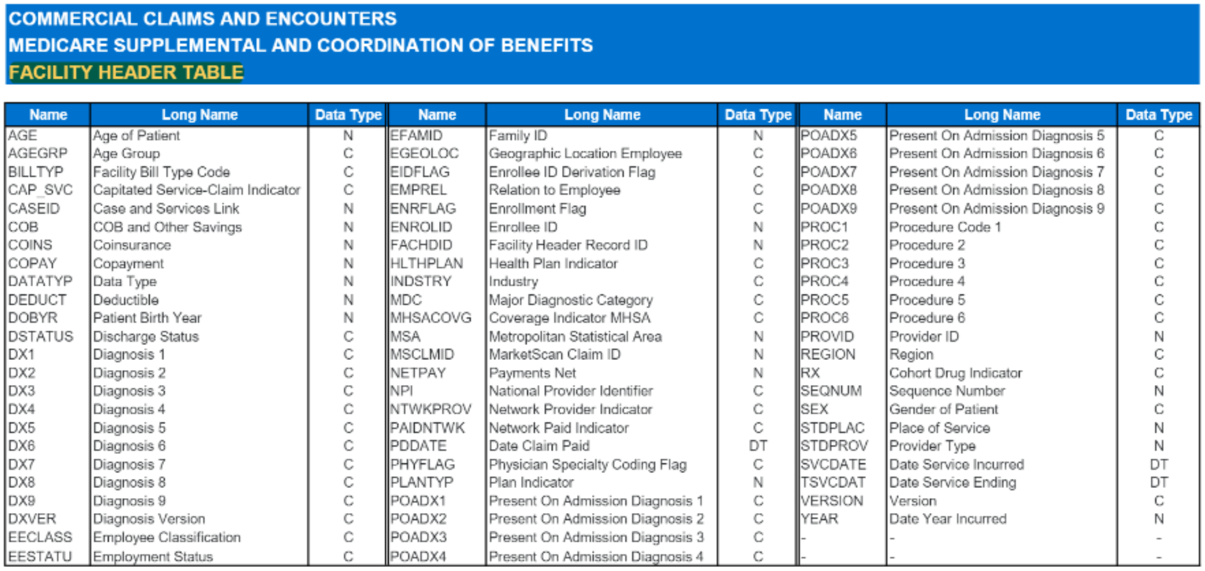
COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
OUTPATIENT PHARMACEUTICAL CLAIMS TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	EIDFLAG	Enrollee ID Derivation Flag	C	PAY	Payment	N
AGEGRP	Age Group	C	EMPREL	Relation to Employee	C	PDDATE	Date Claim Paid	DT
AWP	Average Wholesale Price	N	ENRFLAG	Enrollment Flag	C	PHARMID	Pharmacy ID	N
CAP_SVC	Capitated Service-Claim Indicator	C	ENROLID	Enrollee ID	N	PHYFLAG	Physician Specialty Coding Flag	C
COB	COB and Other Savings	N	GENERID	Generic Product ID	N	PLANTYP	Plan Indicator	N
COINS	Coinsurance	N	GENIND	Generic Indicator	C	QTY	Quantity of Services	N
COPAY	Copayment	N	HLTHPLAN	Health Plan Indicator	C	REFILL	Refill Number	N
DATATYP	Data Type	N	INDSTRY	Industry	C	REGION	Region	C
DAWIND	Dispense as Written Indicator	C	INGCOST	Ingredient Cost	N	RXMR	Rx Mail Retail	C
DAYSUPP	Days Supply	N	MAINTIN	Maintenance Indicator	C	SALETAX	Sales Tax	N
DEACLAS	DEA Classification	C	METQTY	Metric Quantity	N	SEQNUM	Sequence Number	N
DEDUCT	Deductible	N	MHSACOVG	Coverage Indicator MHSA	C	SEX	Gender of Patient	C
DISPFEE	Dispensing Fee	N	MSA	Metropolitan Statistical Area	N	SVCDATE	Date Service Incurred	DT
DOBYR	Patient Birth Year	N	NDCNUM	National Drug Code	C	THERCLS	Therapeutic Class	N
EECLASS	Employee Classification	C	NETPAY	Payments Net	N	THERGRP	Therapeutic Group	C
EESTATU	Employment Status	C	NTWKPROV	Network Provider Indicator	C	VERSION	Version	C
EFAMID	Family ID	N	PAIDNTWK	Network Paid Indicator	C	YEAR	Date Year Incurred	N
EGEOLOC	Geographic Location Employee	C	-	-	-	-	-	-

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	EFAMID	EFAMID	EFAMID
4.	ENROLID	ENROLID	ENROLID
5.	NDCNUM	NDCNUM	NDCNUM
6.	SVCDATE	SVCDATE	SVCDATE
7.	DOBYR	DOBYR	DOBYR
8.	YEAR	YEAR	YEAR
9.	AGE	AGE	AGE
10.	AWP	AWP	AWP
11.	CAP_SVC	CAP_SVC	CAP_SVC
12.	COB	COB	COB
13.	COINS	COINS	COINS
14.	COPAY	COPAY	COPAY
15.	DAYSUPP	DAYSUPP	DAYSUPP
16.	DEDUCT	DEDUCT	DEDUCT
17.	DISPFEE	DISPFEE	DISPFEE
18.	GENERID	GENERID	GENERID
19.	INGCOST	INGCOST	INGCOST
20.	METQTY	METQTY	METQTY
21.	MHSACOVG	MHSACOVG	MHSACOVG
22.	NETPAY	NETPAY	NETPAY
23.	NTWKPROV	NTWKPROV	NTWKPROV
24.	PAIDNTWK	PAIDNTWK	PAIDNTWK
25.	PAY	PAY	PAY
26.	PDDATE	PDDATE	PDDATE
27.	PHARMID	PHARMID	PHARMID
28.	PLANTYP	PLANTYP	PLANTYP
29.	QTY	QTY	QTY

30.	REFILL	REFILL	REFILL
31.	RXMR	RXMR	RXMR
32.	SALETAX	SALETAX	SALETAX
33.	THERCLS	THERCLS	THERCLS
34.	DAWIND	DAWIND	DAWIND
35.	DEACLAS	DEACLAS	DEACLAS
36.	GENIND	GENIND	GENIND
37.	MAINTIN	MAINTIN	MAINTIN
38.	THERGRP	THERGRP	THERGRP
39.	REGION	REGION	REGION
40.	MSA	MSA	MSA
41.	DATATYP	DATATYP	DATATYP
42.	AGEGRP	AGEGRP	AGEGRP
43.	EECLASS	EECLASS	EECLASS
44.	EESTATU	EESTATU	EESTATU
45.	EGEOLOC	EGEOLOC	EGEOLOC
46.	EIDFLAG	EIDFLAG	EIDFLAG
47.	EMPREL	EMPREL	EMPREL
48.	ENRFLAG	ENRFLAG	ENRFLAG
49.	PHYFLAG	PHYFLAG	PHYFLAG
50.	SEX	SEX	SEX
51.	HLTHPLAN	HLTHPLAN	HLTHPLAN
52.	INDSTRY	INDSTRY	INDSTRY
53.			MEDADV

CCAEF(Facility Header F)



COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
FACILITY HEADER TABLE

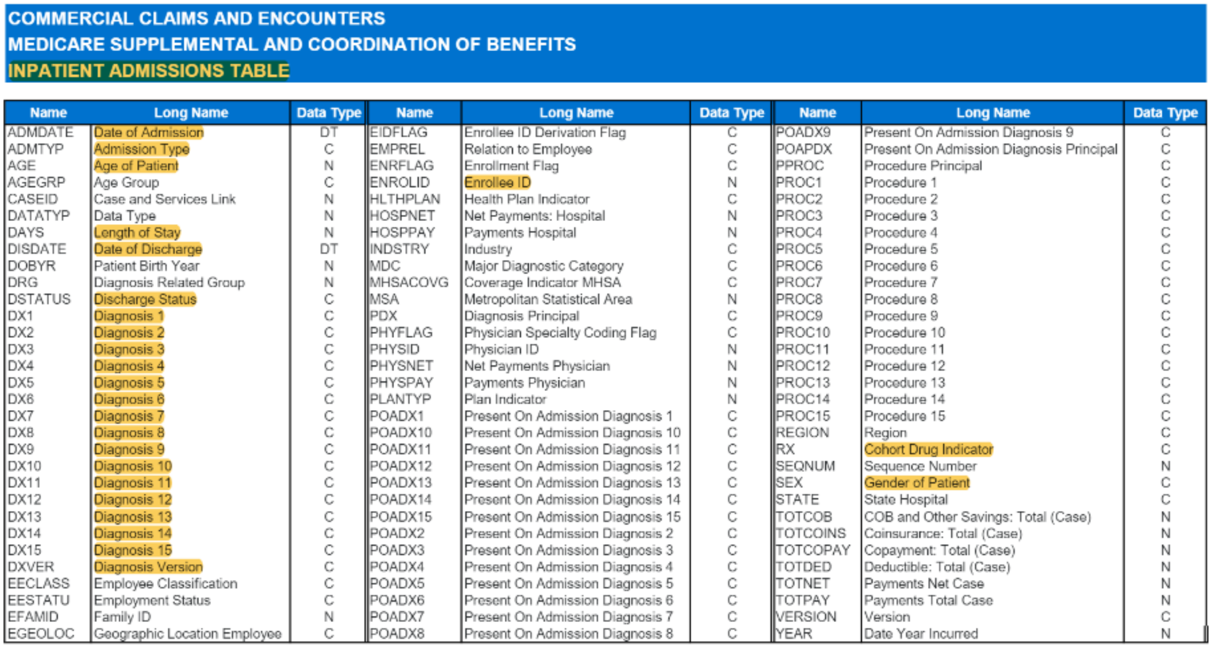
Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	EFAMID	Family ID	N	POADX5	Present On Admission Diagnosis 5	C
AGEGRP	Age Group	C	EGEOLOC	Geographic Location Employee	C	POADX6	Present On Admission Diagnosis 6	C
BILLTYP	Facility Bill Type Code	C	EIDFLAG	Enrollee ID Derivation Flag	C	POADX7	Present On Admission Diagnosis 7	C
CAP_SVC	Capitated Service-Claim Indicator	C	EMPREL	Relation to Employee	C	POADX8	Present On Admission Diagnosis 8	C
CASEID	Case and Services Link	N	ENRFLAG	Enrollment Flag	C	POADX9	Present On Admission Diagnosis 9	C
COB	COB and Other Savings	N	ENROLID	Enrollee ID	N	PROC1	Procedure Code 1	C
COINS	Coinsurance	N	FACHDID	Facility Header Record ID	N	PROC2	Procedure 2	C
COPAY	Copayment	N	HLTHPLAN	Health Plan Indicator	C	PROC3	Procedure 3	C
DATATYP	Data Type	N	INDSTRY	Industry	C	PROC4	Procedure 4	C
DEDUCT	Deductible	N	MDC	Major Diagnostic Category	C	PROC5	Procedure 5	C
DOBYR	Patient Birth Year	N	MHSACOVG	Coverage Indicator MHSA	C	PROC6	Procedure 6	C
DSTATUS	Discharge Status	C	MSA	Metropolitan Statistical Area	N	PROVID	Provider ID	N
DX1	Diagnosis 1	C	MSCLMID	MarketScan Claim ID	N	REGION	Region	C
DX2	Diagnosis 2	C	NETPAY	Payments Net	N	RX	Cohort Drug Indicator	C
DX3	Diagnosis 3	C	NPI	National Provider Identifier	C	SEQNUM	Sequence Number	N
DX4	Diagnosis 4	C	NTWKPROV	Network Provider Indicator	C	SEX	Gender of Patient	C
DX5	Diagnosis 5	C	PAIDNTWK	Network Paid Indicator	C	STDPLAC	Place of Service	N
DX6	Diagnosis 6	C	PDDATE	Date Claim Paid	DT	STDPROV	Provider Type	N
DX7	Diagnosis 7	C	PHYFLAG	Physician Specialty Coding Flag	C	SVCDATE	Date Service Incurred	DT
DX8	Diagnosis 8	C	PLANTYP	Plan Indicator	N	TSVCDAT	Date Service Ending	DT
DX9	Diagnosis 9	C	POADX1	Present On Admission Diagnosis 1	C	VERSION	Version	C
DXVER	Diagnosis Version	C	POADX2	Present On Admission Diagnosis 2	C	YEAR	Date Year Incurred	N
EECLASS	Employee Classification	C	POADX3	Present On Admission Diagnosis 3	C	-	-	-
EESTATU	Employment Status	C	POADX4	Present On Admission Diagnosis 4	C	-	-	-

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	DX1	DX1	DX1
4.	DX2	DX2	DX2
5.	PROC1	PROC1	PROC1
6.	FACHDID	FACHDID	FACHDID

7.	EFAMID	EFAMID	EFAMID
8.	ENROLID	ENROLID	ENROLID
9.	DOBYR	DOBYR	DOBYR
10.	YEAR	YEAR	YEAR
11.	AGE	AGE	AGE
12.	BILLTYP	BILLTYP	BILLTYP
13.	CAP_SVC	CAP_SVC	CAP_SVC
14.	CASEID	CASEID	CASEID
15.	COB	COB	COB
16.	COINS	COINS	COINS
17.	COPAY	COPAY	COPAY
18.	DEDUCT	DEDUCT	DEDUCT
19.	DX3	DX3	DX3
20.	DX4	DX4	DX4
21.	DX5	DX5	DX5
22.	DX6	DX6	DX6
23.	DX7	DX7	DX7
24.	DX8	DX8	DX8
25.	DX9	DX9	DX9
26.	DXVER	DXVER	DXVER
27.	MHSACOVG	MHSACOVG	MHSACOVG
28.	NETPAY	NETPAY	NETPAY
29.	NTWKPROV	NTWKPROV	NTWKPROV
30.	PAIDNTWK	PAIDNTWK	PAIDNTWK
31.	PDDATE	PDDATE	PDDATE
32.	PLANTYP	PLANTYP	PLANTYP
33.	PROC2	PROC2	PROC2
34.	PROC3	PROC3	PROC3
35.	PROC4	PROC4	PROC4
36.	PROC5	PROC5	PROC5
37.	PROC6	PROC6	PROC6
38.	PROVID	PROVID	PROVID
39.	SVCDATE	SVCDATE	SVCDATE
40.	TSVCDAT	TSVCDAT	TSVCDAT
41.	MDC	MDC	MDC
42.	DSTATUS	DSTATUS	DSTATUS
43.	REGION	REGION	REGION
44.	MSA	MSA	MSA
45.	STDPLAC	STDPLAC	STDPLAC
46.	STDPROV	STDPROV	STDPROV
47.	DATATYP	DATATYP	DATATYP
48.	AGEGRP	AGEGRP	AGEGRP
49.	EECLASS	EECLASS	EECLASS
50.	EESTATU	EESTATU	EESTATU
51.	EGEOLOC	EGEOLOC	EGEOLOC
52.	EIDFLAG	EIDFLAG	EIDFLAG
53.	EMPREL	EMPREL	EMPREL
54.	ENRFLAG	ENRFLAG	ENRFLAG
55.	PHYFLAG	PHYFLAG	PHYFLAG
56.	RX	RX	RX
57.	SEX	SEX	SEX
58.	HLTHPLAN	HLTHPLAN	HLTHPLAN

59.	INDSTRY	INDSTRY	INDSTRY
60.	MSCLMID	MSCLMID	MSCLMID
61.	NPI	NPI	NPI
62.	POADX1	POADX1	POADX1
63.	POADX2	POADX2	POADX2
64.	POADX3	POADX3	POADX3
65.	POADX4	POADX4	POADX4
66.	POADX5	POADX5	POADX5
67.	POADX6	POADX6	POADX6
68.	POADX7	POADX7	POADX7
69.	POADX8	POADX8	POADX8
70.	POADX9	POADX9	POADX9
71.			MEDADV

CCAEI(Inpatient Admissions I)



COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
INPATIENT ADMISSIONS TABLE

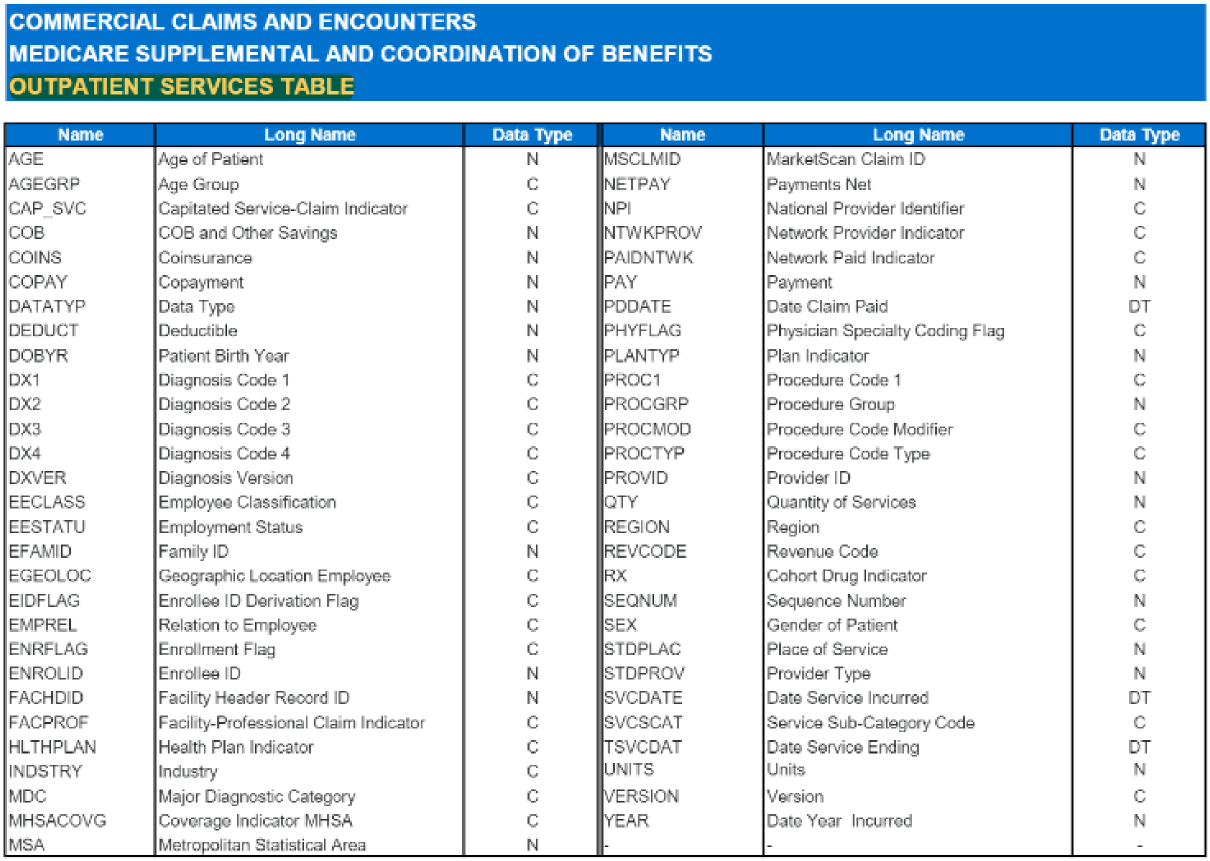
Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
ADMDATE	Date of Admission	DT	EIDFLAG	Enrollee ID Derivation Flag	C	POADX9	Present On Admission Diagnosis 9	C
ADMTYP	Admission Type	C	EMPREL	Relation to Employee	C	POAPDX	Present On Admission Diagnosis Principal	C
AGE	Age of Patient	N	ENRFLAG	Enrollment Flag	C	PPROC	Procedure Principal	C
AGEGRP	Age Group	C	ENROLID	Enrollee ID	N	PROC1	Procedure 1	C
CASEID	Case and Services Link	N	HLTHPLAN	Health Plan Indicator	C	PROC2	Procedure 2	C
DATATYP	Data Type	N	HOSPNET	Net Payments: Hospital	N	PROC3	Procedure 3	C
DAYS	Length of Stay	N	HOSPPAY	Payments Hospital	N	PROC4	Procedure 4	C
DISDATE	Date of Discharge	DT	INDSTRY	Industry	C	PROC5	Procedure 5	C
DOBYR	Patient Birth Year	N	MDC	Major Diagnostic Category	C	PROC6	Procedure 6	C
DRG	Diagnosis Related Group	N	MHSACOVG	Coverage Indicator MHSA	C	PROC7	Procedure 7	C
DSTATUS	Discharge Status	C	MSA	Metropolitan Statistical Area	N	PROC8	Procedure 8	C
DX1	Diagnosis 1	C	PDX	Diagnosis Principal	C	PROC9	Procedure 9	C
DX2	Diagnosis 2	C	PHYFLAG	Physician Specialty Coding Flag	C	PROC10	Procedure 10	C
DX3	Diagnosis 3	C	PHYSID	Physician ID	N	PROC11	Procedure 11	C
DX4	Diagnosis 4	C	PHYSNET	Net Payments Physician	N	PROC12	Procedure 12	C
DX5	Diagnosis 5	C	PHYSPAY	Payments Physician	N	PROC13	Procedure 13	C
DX6	Diagnosis 6	C	PLANTYP	Plan Indicator	N	PROC14	Procedure 14	C
DX7	Diagnosis 7	C	POADX1	Present On Admission Diagnosis 1	C	PROC15	Procedure 15	C
DX8	Diagnosis 8	C	POADX10	Present On Admission Diagnosis 10	C	REGION	Region	C
DX9	Diagnosis 9	C	POADX11	Present On Admission Diagnosis 11	C	RX	Cohort Drug Indicator	C
DX10	Diagnosis 10	C	POADX12	Present On Admission Diagnosis 12	C	SEQNUM	Sequence Number	N
DX11	Diagnosis 11	C	POADX13	Present On Admission Diagnosis 13	C	SEX	Gender of Patient	C
DX12	Diagnosis 12	C	POADX14	Present On Admission Diagnosis 14	C	STATE	State Hospital	C
DX13	Diagnosis 13	C	POADX15	Present On Admission Diagnosis 15	C	TOTCOB	COB and Other Savings: Total (Case)	N
DX14	Diagnosis 14	C	POADX2	Present On Admission Diagnosis 2	C	TOTCOINS	Coinsurance: Total (Case)	N
DX15	Diagnosis 15	C	POADX3	Present On Admission Diagnosis 3	C	TOTCOPAY	Copayment: Total (Case)	N
DXVER	Diagnosis Version	C	POADX4	Present On Admission Diagnosis 4	C	TOTDED	Deductible: Total (Case)	N
EECLASS	Employee Classification	C	POADX5	Present On Admission Diagnosis 5	C	TOTNET	Payments Net Case	N
EESTATU	Employment Status	C	POADX6	Present On Admission Diagnosis 6	C	TOTPAY	Payments Total Case	N
EFAMID	Family ID	N	POADX7	Present On Admission Diagnosis 7	C	VERSION	Version	C
EGEOLOC	Geographic Location Employee	C	POADX8	Present On Admission Diagnosis 8	C	YEAR	Date Year Incurred	N

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	EFAMID	EFAMID	EFAMID
4.	ENROLID	ENROLID	ENROLID
5.	DOBYR	DOBYR	DOBYR
6.	YEAR	YEAR	YEAR
7.	ADMDATE	ADMDATE	ADMDATE
8.	AGE	AGE	AGE
9.	CASEID	CASEID	CASEID
10.	DAYS	DAYS	DAYS
11.	DISDATE	DISDATE	DISDATE
12.	DRG	DRG	DRG
13.	DXVER	DXVER	DXVER
14.	HOSPNET	HOSPNET	HOSPNET
15.	HOSPPAY	HOSPPAY	HOSPPAY

16.	MHSACOVG	MHSACOVG	MHSACOVG
17.	PDX	PDX	PDX
18.	PHYSID	PHYSID	PHYSID
19.	PHYSNET	PHYSNET	PHYSNET
20.	PHYSPAY	PHYSPAY	PHYSPAY
21.	PLANTYP	PLANTYP	PLANTYP
22.	PROC	PROC	PROC
23.	TOTCOB	TOTCOB	TOTCOB
24.	TOTCOINS	TOTCOINS	TOTCOINS
25.	TOTCOPAY	TOTCOPAY	TOTCOPAY
26.	TOTDED	TOTDED	TOTDED
27.	TOTNET	TOTNET	TOTNET
28.	TOTPAY	TOTPAY	TOTPAY
29.	ADMTYP	ADMTYP	ADMTYP
30.	MDC	MDC	MDC
31.	DSTATUS	DSTATUS	DSTATUS
32.	REGION	REGION	REGION
33.	MSA	MSA	MSA
34.	DATATYP	DATATYP	DATATYP
35.	DX1	DX1	DX1
36.	DX2	DX2	DX2
37.	DX3	DX3	DX3
38.	DX4	DX4	DX4
39.	DX5	DX5	DX5
40.	DX6	DX6	DX6
41.	DX7	DX7	DX7
42.	DX8	DX8	DX8
43.	DX9	DX9	DX9
44.	DX10	DX10	DX10
45.	DX11	DX11	DX11
46.	DX12	DX12	DX12
47.	DX13	DX13	DX13
48.	DX14	DX14	DX14
49.	DX15	DX15	DX15
50.	PROC1	PROC1	PROC1
51.	PROC2	PROC2	PROC2
52.	PROC3	PROC3	PROC3
53.	PROC4	PROC4	PROC4
54.	PROC5	PROC5	PROC5
55.	PROC6	PROC6	PROC6
56.	PROC7	PROC7	PROC7
57.	PROC8	PROC8	PROC8
58.	PROC9	PROC9	PROC9
59.	PROC10	PROC10	PROC10
60.	PROC11	PROC11	PROC11
61.	PROC12	PROC12	PROC12
62.	PROC13	PROC13	PROC13
63.	PROC14	PROC14	PROC14
64.	PROC15	PROC15	PROC15
65.	AGEGRP	AGEGRP	AGEGRP
66.	EECLASS	EECLASS	EECLASS
67.	EESTATU	EESTATU	EESTATU

68.	EGEOLOC	EGEOLOC	EGEOLOC
69.	EIDFLAG	EIDFLAG	EIDFLAG
70.	EMPREL	EMPREL	EMPREL
71.	ENRFLAG	ENRFLAG	ENRFLAG
72.	PHYFLAG	PHYFLAG	PHYFLAG
73.	RX	RX	RX
74.	SEX	SEX	SEX
75.	STATE	STATE	STATE
76.	HLTHPLAN	HLTHPLAN	HLTHPLAN
77.	INDSTRY	INDSTRY	INDSTRY
78.	POAPDX	POAPDX	POAPDX
79.	POADX1	POADX1	POADX1
80.	POADX2	POADX2	POADX2
81.	POADX3	POADX3	POADX3
82.	POADX4	POADX4	POADX4
83.	POADX5	POADX5	POADX5
84.	POADX6	POADX6	POADX6
85.	POADX7	POADX7	POADX7
86.	POADX8	POADX8	POADX8
87.	POADX9	POADX9	POADX9
88.	POADX10	POADX10	POADX10
89.	POADX11	POADX11	POADX11
90.	POADX12	POADX12	POADX12
91.	POADX13	POADX13	POADX13
92.	POADX14	POADX14	POADX14
93.	POADX15	POADX15	POADX15
94.			MEDADV

CCAEO(Outpatient Services O)



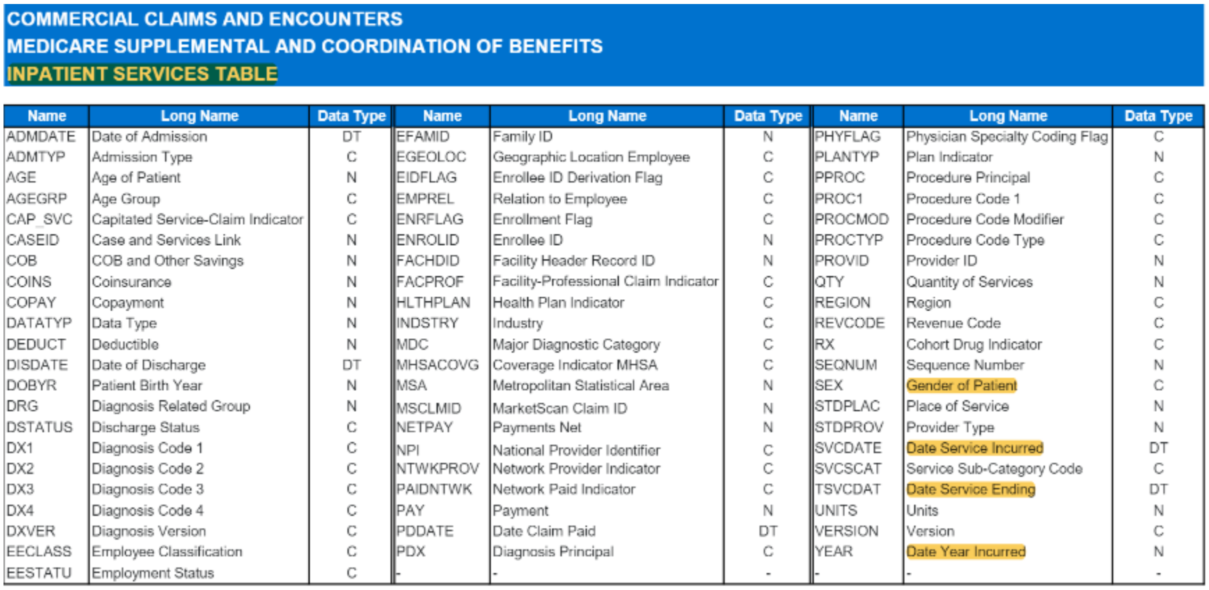
COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
OUTPATIENT SERVICES TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	MSCLMID	MarketScan Claim ID	N
AGEGRP	Age Group	C	NETPAY	Payments Net	N
CAP_SVC	Capitated Service-Claim Indicator	C	NPI	National Provider Identifier	C
COB	COB and Other Savings	N	NTWKPROV	Network Provider Indicator	C
COINS	Coinsurance	N	PAIDNTWK	Network Paid Indicator	C
COPAY	Copayment	N	PAY	Payment	N
DATATYP	Data Type	N	PDDATE	Date Claim Paid	DT
DEDUCT	Deductible	N	PHYFLAG	Physician Specialty Coding Flag	C
DOBYR	Patient Birth Year	N	PLANTYP	Plan Indicator	N
DX1	Diagnosis Code 1	C	PROC1	Procedure Code 1	C
DX2	Diagnosis Code 2	C	PROCGRP	Procedure Group	N
DX3	Diagnosis Code 3	C	PROCMOD	Procedure Code Modifier	C
DX4	Diagnosis Code 4	C	PROCTYP	Procedure Code Type	C
DXVER	Diagnosis Version	C	PROVID	Provider ID	N
EECLASS	Employee Classification	C	QTY	Quantity of Services	N
EESTATU	Employment Status	C	REGION	Region	C
EFAMID	Family ID	N	REVCODE	Revenue Code	C
EGEOLOC	Geographic Location Employee	C	RX	Cohort Drug Indicator	C
EIDFLAG	Enrollee ID Derivation Flag	C	SEQNUM	Sequence Number	N
EMPREL	Relation to Employee	C	SEX	Gender of Patient	C
ENRFLAG	Enrollment Flag	C	STDPLAC	Place of Service	N
ENROLID	Enrollee ID	N	STDPROV	Provider Type	N
FACHDID	Facility Header Record ID	N	SVCDATE	Date Service Incurred	DT
FACPROF	Facility-Professional Claim Indicator	C	SVCSCAT	Service Sub-Category Code	C
HLTHPLAN	Health Plan Indicator	C	TSVCDAT	Date Service Ending	DT
INDSTRY	Industry	C	UNITS	Units	N
MDC	Major Diagnostic Category	C	VERSION	Version	C
MHSACOVG	Coverage Indicator MHSA	C	YEAR	Date Year Incurred	N
MSA	Metropolitan Statistical Area	N	-	-	-

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	
2.	VERSION	VERSION	
3.	DX1	DX1	
4.	DX2	DX2	
5.	PROC1	PROC1	
6.	PROCTYP	PROCTYP	
7.	EFAMID	EFAMID	
8.	ENROLID	ENROLID	
9.	REVCODE	REVCODE	
10.	SVCDATE	SVCDATE	
11.	DOBYR	DOBYR	
12.	YEAR	YEAR	
13.	AGE	AGE	
14.	CAP_SVC	CAP_SVC	
15.	COB	COB	
16.	COINS	COINS	
17.	COPAY	COPAY	
18.	DEDUCT	DEDUCT	
19.	DX3	DX3	
20.	DX4	DX4	
21.	DXVER	DXVER	
22.	FACHDID	FACHDID	
23.	FACPROF	FACPROF	
24.	MHSACOVG	MHSACOVG	
25.	NETPAY	NETPAY	
26.	NTWKPROV	NTWKPROV	

27.	PAIDNTWK	PAIDNTWK	
28.	PAY	PAY	
29.	PDDATE	PDDATE	
30.	PLANTYP	PLANTYP	
31.	PROCGRP	PROCGRP	
32.	PROCMOD	PROCMOD	
33.	PROVID	PROVID	
34.	QTY	QTY	
35.	SVCSCAT	SVCSCAT	
36.	TSVCDAT	TSVCDAT	
37.	MDC	MDC	
38.	REGION	REGION	
39.	MSA	MSA	
40.	STDPLAC	STDPLAC	
41.	STDPROV	STDPROV	
42.	DATATYP	DATATYP	
43.	AGEGRP	AGEGRP	
44.	EECLASS	EECLASS	
45.	EESTATU	EESTATU	
46.	EGEOLOC	EGEOLOC	
47.	EIDFLAG	EIDFLAG	
48.	EMPREL	EMPREL	
49.	ENRFLAG	ENRFLAG	
50.	PHYFLAG	PHYFLAG	
51.	RX	RX	
52.	SEX	SEX	
53.	HLTHPLAN	HLTHPLAN	
54.	INDSTRY	INDSTRY	
55.	MSCLMID	MSCLMID	
56.	NPI	NPI	
57.	UNITS	UNITS	
58.			MEDADV

CCAES(Inpatient Services S)



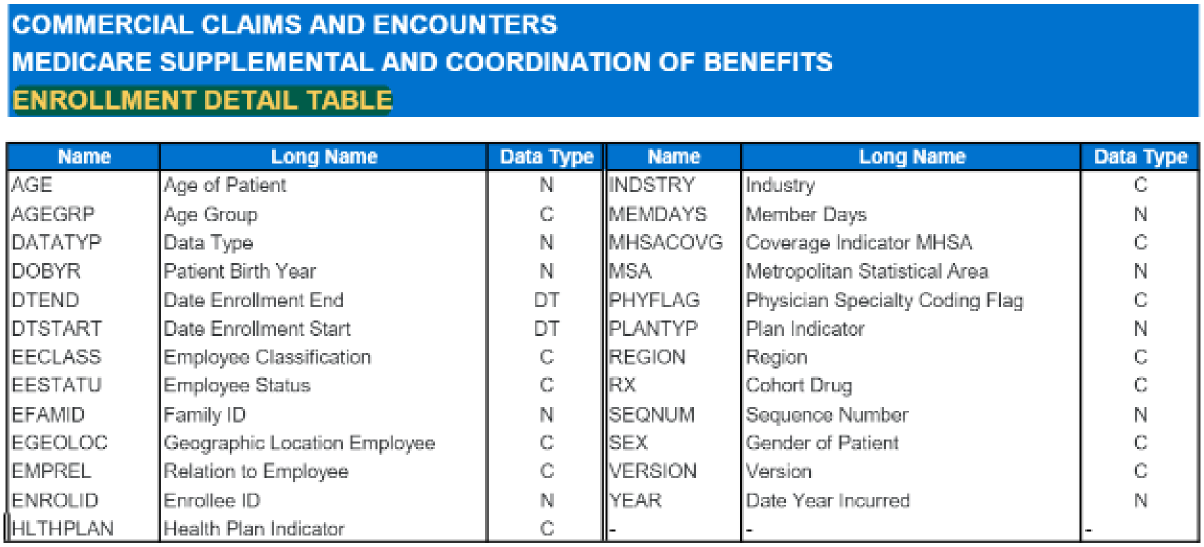
COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
INPATIENT SERVICES TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
ADMDATE	Date of Admission	DT	EFAMID	Family ID	N	PHYFLAG	Physician Specialty Coding Flag	C
ADMTYP	Admission Type	C	EGEOLOC	Geographic Location Employee	C	PLANTYP	Plan Indicator	N
AGE	Age of Patient	N	EIDFLAG	Enrollee ID Derivation Flag	C	PPROC	Procedure Principal	C
AGEGRP	Age Group	C	EMPREL	Relation to Employee	C	PROC1	Procedure Code 1	C
CAP_SVC	Capitated Service-Claim Indicator	C	ENRFLAG	Enrollment Flag	C	PROCMOD	Procedure Code Modifier	C
CASEID	Case and Services Link	N	ENROLID	Enrollee ID	N	PROCTYP	Procedure Code Type	C
COB	COB and Other Savings	N	FACHDID	Facility Header Record ID	N	PROVID	Provider ID	N
COINS	Coinsurance	N	FACPROF	Facility-Professional Claim Indicator	C	QTY	Quantity of Services	N
COPAY	Copayment	N	HLTHPLAN	Health Plan Indicator	C	REGION	Region	C
DATATYP	Data Type	N	INDSTRY	Industry	C	REVCODE	Revenue Code	C
DEDUCT	Deductible	N	MDC	Major Diagnostic Category	C	RX	Cohort Drug Indicator	C
DISDATE	Date of Discharge	DT	MHSACOVG	Coverage Indicator MHSA	C	SEQNUM	Sequence Number	N
DOBYR	Patient Birth Year	N	MSA	Metropolitan Statistical Area	N	SEX	Gender of Patient	C
DRG	Diagnosis Related Group	N	MSCLMID	MarketScan Claim ID	N	STDPLAC	Place of Service	N
DSTATUS	Discharge Status	C	NETPAY	Payments Net	N	STDPROV	Provider Type	N
DX1	Diagnosis Code 1	C	NPI	National Provider Identifier	C	SVCDATE	Date Service Incurred	DT
DX2	Diagnosis Code 2	C	NTWKPROV	Network Provider Indicator	C	SVCSCAT	Service Sub-Category Code	C
DX3	Diagnosis Code 3	C	PAIDNTWK	Network Paid Indicator	C	TSVCDAT	Date Service Ending	DT
DX4	Diagnosis Code 4	C	PAY	Payment	N	UNITS	Units	N
DXVER	Diagnosis Version	C	PDDATE	Date Claim Paid	DT	VERSION	Version	C
EECLASS	Employee Classification	C	PDX	Diagnosis Principal	C	YEAR	Date Year Incurred	N
EESTATU	Employment Status	C	-	-	-	-	-	-

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	DX1	DX1	DX1
4.	DX2	DX2	DX2
5.	PROC1	PROC1	PROC1
6.	PROCTYP	PROCTYP	PROCTYP
7.	CASEID	CASEID	CASEID
8.	DISDATE	DISDATE	DISDATE
9.	DOBYR	DOBYR	DOBYR
10.	YEAR	YEAR	YEAR
11.	ADMDATE	ADMDATE	ADMDATE
12.	AGE	AGE	AGE
13.	CAP_SVC	CAP_SVC	CAP_SVC
14.	COB	COB	COB
15.	COINS	COINS	COINS
16.	COPAY	COPAY	COPAY
17.	DEDUCT	DEDUCT	DEDUCT
18.	DRG	DRG	DRG
19.	DX3	DX3	DX3
20.	DX4	DX4	DX4
21.	DXVER	DXVER	DXVER
22.	FACHDID	FACHDID	FACHDID
23.	FACPROF	FACPROF	FACPROF
24.	MHSACOVG	MHSACOVG	MHSACOVG
25.	NETPAY	NETPAY	NETPAY
26.	NTWKPROV	NTWKPROV	NTWKPROV
27.	PAIDNTWK	PAIDNTWK	PAIDNTWK
28.	PAY	PAY	PAY
29.	PDDATE	PDDATE	PDDATE
30.	PDX	PDX	PDX
31.	PPROC	PPROC	PPROC
32.	PROCMOD	PROCMOD	PROCMOD

33.	PROVID	PROVID	PROVID
34.	QTY	QTY	QTY
35.	REVCODE	REVCODE	REVCODE
36.	SVCDATE	SVCDATE	SVCDATE
37.	SVCSCAT	SVCSCAT	SVCSCAT
38.	TSVCDAT	TSVCDAT	TSVCDAT
39.	ADMTYP	ADMTYP	ADMTYP
40.	MDC	MDC	MDC
41.	DSTATUS	DSTATUS	DSTATUS
42.	STDPLAC	STDPLAC	STDPLAC
43.	STDPROV	STDPROV	STDPROV
44.	EFAMID	EFAMID	EFAMID
45.	ENROLID	ENROLID	ENROLID
46.	PLANTYP	PLANTYP	PLANTYP
47.	REGION	REGION	REGION
48.	MSA	MSA	MSA
49.	DATATYP	DATATYP	DATATYP
50.	AGEGRP	AGEGRP	AGEGRP
51.	EECLASS	EECLASS	EECLASS
52.	EESTATU	EESTATU	EESTATU
53.	EGEOLOC	EGEOLOC	EGEOLOC
54.	EIDFLAG	EIDFLAG	EIDFLAG
55.	EMPREL	EMPREL	EMPREL
56.	ENRFLAG	ENRFLAG	ENRFLAG
57.	PHYFLAG	PHYFLAG	PHYFLAG
58.	RX	RX	RX
59.	SEX	SEX	SEX
60.	HLTHPLAN	HLTHPLAN	HLTHPLAN
61.	INDSTRY	INDSTRY	INDSTRY
62.	MSCLMID	MSCLMID	MSCLMID
63.	NPI	NPI	NPI
64.	UNITS	UNITS	UNITS
65.			MEDADV

CCAET(Enrollment T)



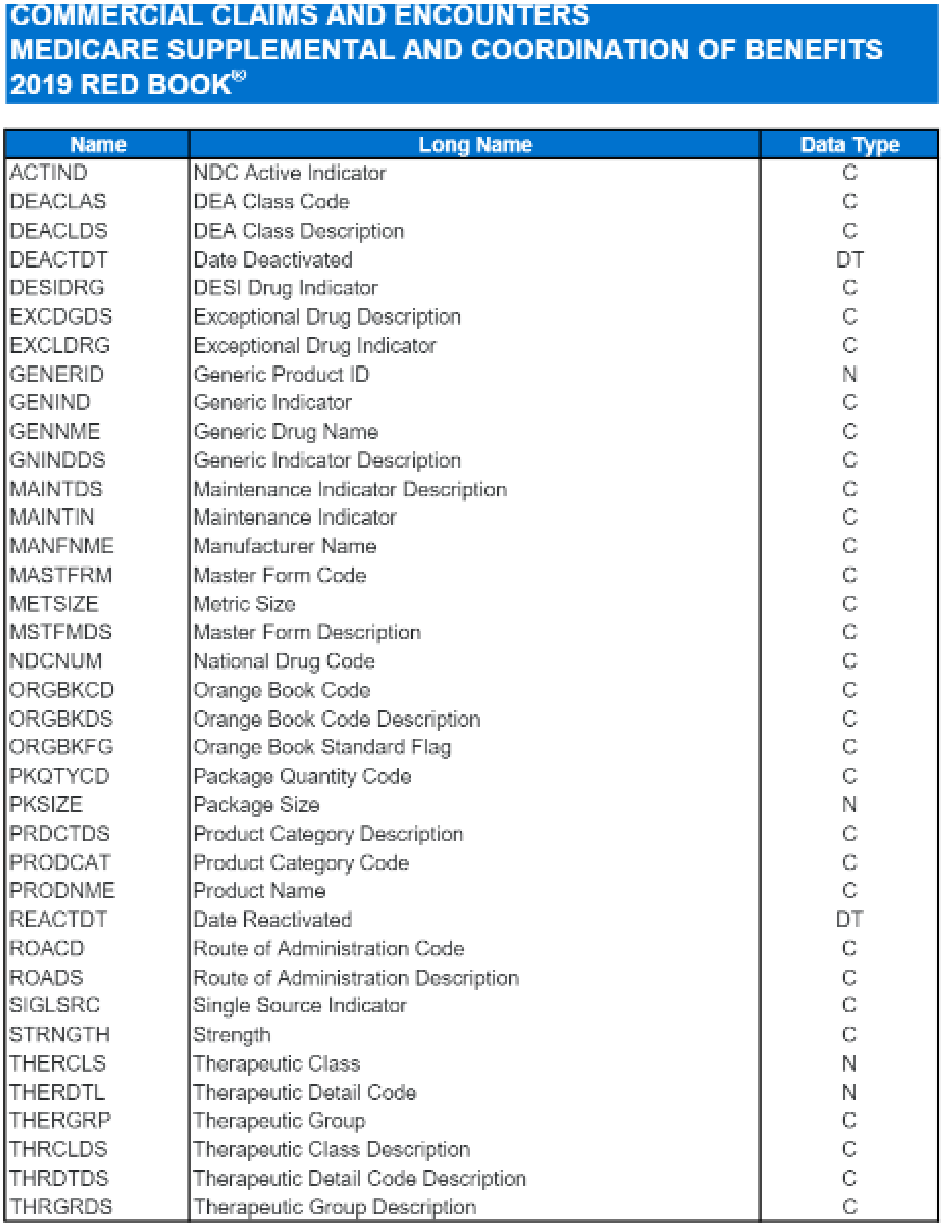
COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
ENROLLMENT DETAIL TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	INDSTRY	Industry	C
AGEGRP	Age Group	C	MEMDAYS	Member Days	N
DATATYP	Data Type	N	MHSACOVG	Coverage Indicator MHSA	C
DOBYR	Patient Birth Year	N	MSA	Metropolitan Statistical Area	N
DTEND	Date Enrollment End	DT	PHYFLAG	Physician Specialty Coding Flag	C
DTSTART	Date Enrollment Start	DT	PLANTYP	Plan Indicator	N
EECLASS	Employee Classification	C	REGION	Region	C
EESTATU	Employee Status	C	RX	Cohort Drug	C
EFAMID	Family ID	N	SEQNUM	Sequence Number	N
EGEOLOC	Geographic Location Employee	C	SEX	Gender of Patient	C
EMPREL	Relation to Employee	C	VERSION	Version	C
ENROLID	Enrollee ID	N	YEAR	Date Year Incurred	N
HLTHPLAN	Health Plan Indicator	C	-	-	-

No of Columns	181	192_a	200_a
1.	SEQNUM	SEQNUM	SEQNUM
2.	VERSION	VERSION	VERSION
3.	EFAMID	EFAMID	EFAMID
4.	ENROLID	ENROLID	ENROLID
5.	DTEND	DTEND	DTEND
6.	DTSTART	DTSTART	DTSTART
7.	MEMDAYS	MEMDAYS	MEMDAYS
8.	MHSACOVG	MHSACOVG	MHSACOVG
9.	PLANTYP	PLANTYP	PLANTYP
10.	YEAR	YEAR	YEAR
11.	AGE	AGE	AGE
12.	DOBYR	DOBYR	DOBYR
13.	REGION	REGION	REGION
14.	MSA	MSA	MSA
15.	DATATYP	DATATYP	DATATYP
16.	AGEGRP	AGEGRP	AGEGRP
17.	EECLASS	EECLASS	EECLASS
18.	EESTATU	EESTATU	EESTATU
19.	EGEOLOC	EGEOLOC	EGEOLOC
20.	EMPREL	EMPREL	EMPREL
21.	PHYFLAG	PHYFLAG	PHYFLAG
22.	RX	RX	RX
23.	SEX	SEX	SEX
24.	HLTHPLAN	HLTHPLAN	HLTHPLAN
25.	INDSTRY	INDSTRY	INDSTRY
26.			MEDADV

RedBook

RedBook has 37 attributes.



COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
2019 RED BOOK®

Name	Long Name	Data Type
ACTIND	NDC Active Indicator	C
DEACLAS	DEA Class Code	C
DEACLDS	DEA Class Description	C
DEACTDT	Date Deactivated	DT
DESIDRG	DESI Drug Indicator	C
EXCDGDS	Exceptional Drug Description	C
EXCLDRG	Exceptional Drug Indicator	C
GENERID	Generic Product ID	N
GENIND	Generic Indicator	C
GENNME	Generic Drug Name	C
GNINDDS	Generic Indicator Description	C
MAINTDS	Maintenance Indicator Description	C
MAINTIN	Maintenance Indicator	C
MANFNME	Manufacturer Name	C
MASTFRM	Master Form Code	C
METSIZE	Metric Size	C
MSTFMDS	Master Form Description	C
NDCNUM	National Drug Code	C
ORGBKCD	Orange Book Code	C
ORGBKDS	Orange Book Code Description	C
ORGBKFG	Orange Book Standard Flag	C
PKQTYCD	Package Quantity Code	C
PKSIZE	Package Size	N
PRDCTDS	Product Category Description	C
PRODCAT	Product Category Code	C
PRODNME	Product Name	C
REACTDT	Date Reactivated	DT
ROACD	Route of Administration Code	C
ROADS	Route of Administration Description	C
SIGLSRC	Single Source Indicator	C
STRNGTH	Strength	C
THERCLS	Therapeutic Class	N
THERDTL	Therapeutic Detail Code	N
THERGRP	Therapeutic Group	C
THRCLDS	Therapeutic Class Description	C
THRDTDS	Therapeutic Detail Code Description	C
THRGRDS	Therapeutic Group Description	C