A Project Titled

Predicting Hospitalization using Demographic and Claims Information of Medicaid Members

Sponsored by

Payformance Solutions

Submitted By

Prashant Chaudhary
Satya Devi Julakanti
Lopamudra Penmetcha
Abhishek Singh
(M.S., University of Illinois at Chicago, Chicago)

Under the guidance of

Professor Kyle Cheek

For the partial fulfilment of the requirements
for the degree of Master of Science in Business Analytics
in the Graduate College of the
University of Illinois at Chicago, Chicago, Illinois

December 2019

Table of Contents

Chapter 1

Introduction

1.1. Healthcare Analytics

Healthcare Analytics addresses the rise in demand for high-quality, safe and affordable care. The correct combination of technology, skills, and strategy allows the healthcare providers to gain insights from health data thereby ensuring improvement in patient outcomes and experience while minimizing the cost for patients as well as insurance providers.

1.2. About the Sponsors

Payformance Solutions is a value-based reimbursement service provider dedicated to advancing payment transformation in the healthcare industry. Trusthub is a full, turnkey platform offered by Payformance Solutions which provides the technical tools and resources needed to build and drive scalable, value-based reimbursement ecosystems. The target markets for Payformance Solutions are the plans and providers who are focused on building and scaling value-based reimbursement programs.

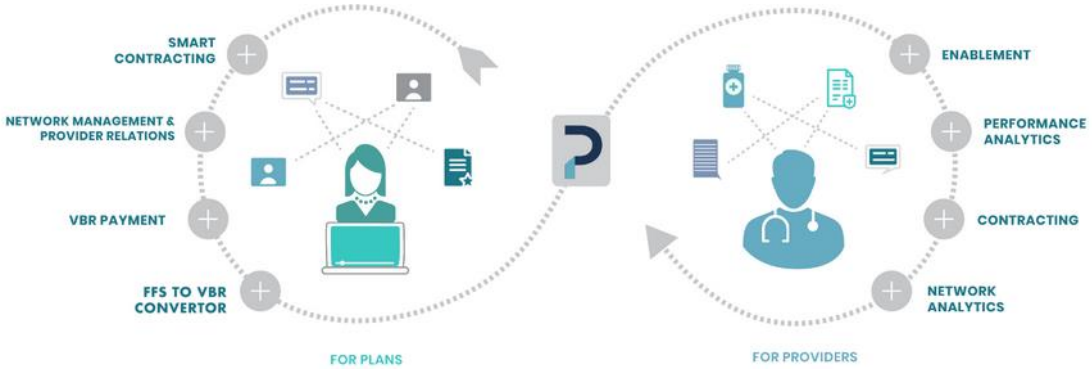


Figure1. Working of Payformance Solutions

Chapter 2

What's the Project?

2.1. Problem Statement

Medicaid members are hospitalized, resulting in higher costs and improper utilization, because chronic conditions are not properly managed, specifically Diabetes, but with adequate patient education and care coordination these at-risk cohorts can be more effectively managed with respect to cost and utilization.

Individuals with chronic conditions, such as Diabetes, should actively manage their health to ensure that conditions do not worsen and to minimize the risk of hospitalizations. In addition, lack of proper maintenance can result in the onset of new & worse conditions, known as comorbidities.

Comorbidities include kidney failure, blindness, loss of limbs, etc. The challenge is knowing if patients/members are actively managing their condition, or to what extent services are being used outside of a standard PCP setting. This can be influenced by many things such as:

1. Social/Economic Status
2. Access to care
3. Level of healthcare coverage
4. Volume of Physician interaction/outreach

Regular visits to a patients Primary Care Provider (PCP), in addition to a comprehensive diabetes care plan should result in improved condition management and a lessen the likelihood for the onset of newer conditions. These visits will also reinforce the proper education for condition maintenance. As a result, costs per patient with Diabetes should lower.

2.2. Objective of the Project

Develop a model to determine outcome/likelihood of different utilization types based on a selection of input variables based off their initial diagnosis and the frequency of office visits, including but not limited to PCP or specialists. Ideally, features of the model should allow factors to contribute to or against the likelihood of an outcome.

2.3. Diabetes

Diabetes is a group of diseases that result in too much sugar in the blood (high blood glucose). There are two types of diabetes – Type I and Type II

Type I Diabetes:

People with Type 1 diabetes don't produce insulin. You can think of it as not having a key. This is a condition people are typically born with. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The loss of beta cells is caused by an autoimmune response. The cause of this autoimmune response is unknown.

Type II Diabetes:

People with Type 2 diabetes don't respond to insulin as well as they should and later in the disease often don't make enough insulin. You can think of this as having a broken key. This is a condition that develops over a patient's life due to poor health choices. This form was previously referred to as "non-insulin dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is a combination of excessive body weight and insufficient exercise.

Type 2 diabetes is much more common than Type 1. According to the 2017 National Diabetes Statistics Report, there are 30.3 million people in the United States with diabetes, about 1 in 10 people. Among all these people living with diabetes, 90 to 95 percent have Type 2 diabetes.

2.4. Medicaid

Medicaid in the United States is a federal and state program that helps with medical costs for some people with limited income and resources. Medicaid also offers benefits not normally covered by Medicare, including nursing home care and personal care services. The Health Insurance Association of America describes Medicaid as "a government insurance program for persons of all ages whose income and resources are insufficient to pay for health care. Medicaid is the largest source of funding for medical and health-related services for people with low income in the United States, providing free health insurance to 74 million low-income and disabled people (23% of Americans) as of 2017. All Americans who meet the Medicaid eligibility criteria are guaranteed coverage. Following are some of the categories of individual and the coverage they are eligible for:

1. Low-Income Families:
    - Pregnant Women: Pre-natal care and delivery costs.
    - Children: Routine and specialized care for childhood development (Immunization, dental, vision, speech therapy)
    - Families: Affordable coverage to prepare for unexpected (Emergency, dental, hospitalizations, antibiotics)
2. Individuals with Disabilities:
    - Child with Autism: In-home therapy, speech/occupational therapy.
    - Cerebral Palsy: Assistance to gain independence (personal care, case management and assistive technology).
    - HIV/AIDS: Physician services, prescription drugs.
    - Mental Illness: Physician services, prescription drugs.
3. Elderly individuals:
    - Medicare beneficiary: help paying for Medicare premium and cost sharing
    - Community Waiver Participant: community-based and personal care
    - Nursing Home Residents: care paid by Medicaid since Medicare does not cover institutional care.

2.5 About the Dataset

We were given a dataset consisting of diabetes patients from Michigan State who were enrolled in Medicaid program. The dataset specifically had three different data:

- Member Information
- Claims Summary
- Claims Detail

The details for each of the tables is as follows:

1. Member Information:
    - This table consists of data for the members of the Medicaid program in Michigan state who are suffering from diabetes.
    - There are 20,000 records in this table having 20 features explaining the demographic details of the members.
    - The list of features along with their description for the member information table is available in Appendix A.
2. Claims Summary:
    - This table consists of summary of the claims data that is submitted by the 20,000 members available in the member information table.
    - There are approximately 1.8 Million records in this table having 13 features explaining the overview of the submitted claims.
    - The list of features along with their description for the Claims Summary table is available in Appendix A.
3. Claims Detail:
    - This table consists of detailed overview of the 1.8 Million claims available in Claims Summary table.
    - There are approximately 4.9 Million records in this table having 27 features explaining the detailed description of the each submitted claims.
    - The list of features along with their description for the Claims Detail table is available in Appendix A.

Chapter 3

Methodology

Implementation of this project is done using the following steps:

1. Exploratory Data Analysis
2. Date Pre-Processing
3. Learning the Models
4. Evaluating the models for improvement.

3.1. Exploratory Data Analysis

We tagged the members as Emergency, Inpatient or Readmission using the following conditions:

- Emergency: Claims containing revenue codes between '0450-0459' or '0981'.
- Inpatient: Claims which are 'not subacute' & type of bill beginning with '11*'
- Readmission: Inpatient members who are admitted again within 30 days of their previous discharge of their hospital.

Based on the above-mentioned conditions we identified the claims as emergency, inpatient or readmission. Following are the number of claims which were submitted as emergency, inpatient or readmission:

Table 1. Count of Claims submitted into all the three categories

Category	*Total # of Claims*	*Total # of UNIQUE Claims*
Emergency	123,165	81,384
Inpatient	164,550	17,298
Readmission	4,128	

Appendix B shows some charts that depict characteristics of the given dataset.

3.2. Data Pre-Processing

Data was spread across three tables; Member Information, Claims Summary and Claims Detail. The database diagram showing the connection between the tables using primary key and foreign key is as shown below:

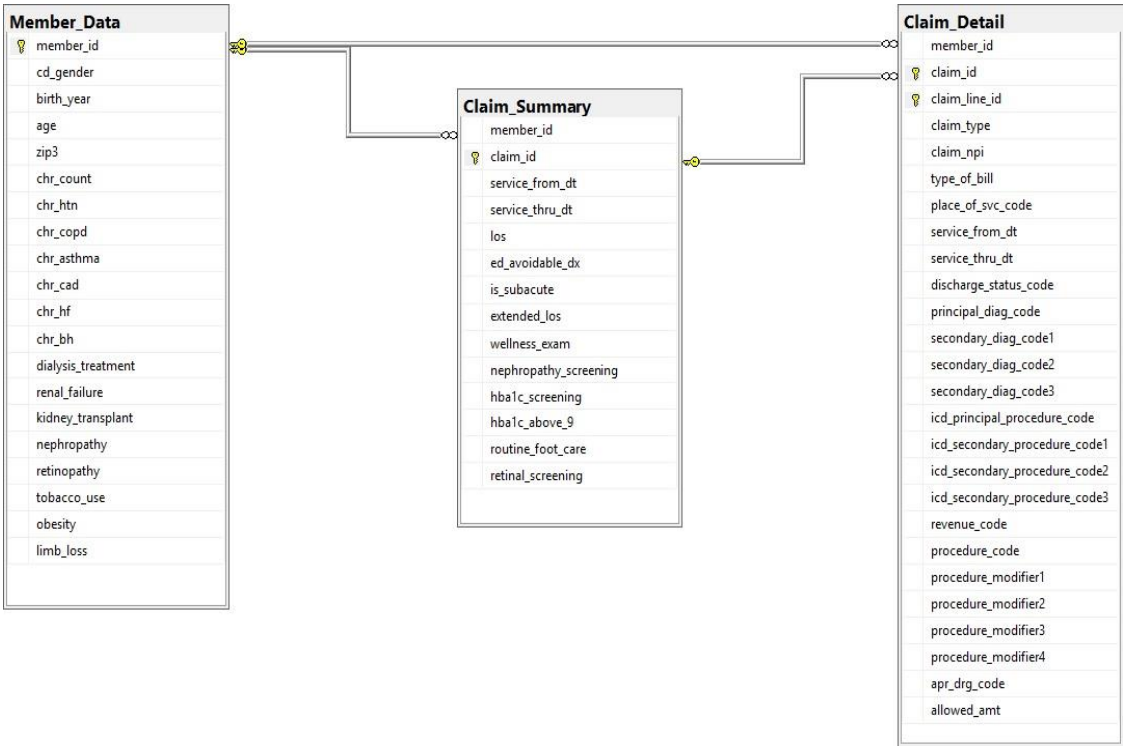


Figure 2. Database Diagram

For the purpose of modelling, we developed three target variables; EM_Counts (# times patient was admitted in emergency), Em_Member (Whether a patient was admitted as emergency member or not [1/0]), INPD_Member (Whether a patient was admitted as inpatient member or not [1/0]) and RM_Member (Whether a patient was admitted as readmission member or not [1/0]).

Once the target variables were created in the Member Information table, we aggregated the claims data of 1.8 million and 4. Million respectively to have only 1 record pertaining to each member in the member information table. This claims data was then joined with Member information data using SQL, so that we have one full dataset ready to build our models. To decide on the importance of each feature, and whether this feature is necessary or not, we went ahead to perform feature engineering.

3.3. Learning the models

We performed model building in 2 phases; the first phase consisted of building a model to predict whether a member will be admitted as Inpatient or readmission member or not in the near future. INPD_Member, EM_Member and RM_Member were the target variables used for the modelling in this phase. The second phase was to build a model to predict the # times a member will visit an Emergency department in the near future. Considering the fact that emergency visits tend to be much costlier, healthcare insurance companies are more destined towards working on this problem so as to reduce their cost. This problem focused on predicting how we can reduce the # times a person visits the emergency.

For the first phase of modelling, we split the data into train and test to perform evaluation at a later stage. We used Random forest, Support Vector Machine and K-NN algorithm for the purpose of training the model. It is evident from the fact that random forest proves to be beneficial since we have limited number of data points and small set of features (approximately 25) for modelling. We used accuracy as the performance metric to decide on the performance of the model.

For the second phase of modelling, we built a simple regression model to decide on how many times a member can visit the emergency in near future. We found the intercept of the regression line and the important features that are impacting how many times a person is going to visit an emergency department

3.4. Evaluating the models for improvement.

In both the phases, we tried evaluating the model for their improvement so as to have better accuracy and a better prediction can be obtained. The evaluation in the first phase of the modelling didn't lead to much change in the performance of the model. However, in the second phase we tried to perform Ordinary Least Square function to improve the performance, which in turn resulted in the results getting worst.

Chapter 4

Results

The results of the first phase of the modelling is as summarized below:

Table 2. Model Accuracies for all the three categories of hospitalization

Algorithm	Category of Hospitalization		
	Emergency	Inpatient	Readmission
Random Forest	76%	78%	82%
Support Vector Machine	75%	76%	81%
K-Nearest Neighbour	78%	77%	80%

The list of important features which are impacting the model and their scores is shown in Appendix C.

The Second phase of modelling involved building a regression model. The equation for the regression line is as shown below:

1.374 +
(*Tobacco_*Use * 2.286) + (*Limb_ Loss* * 1.457) + (*Chr_ Asthma* * 0.996) +
(*Chr_ Count* * 0.599) + (*Nephropathy_ Screening* * 0.499) + (*D_ Gender* * 0.36) +
(*Routine_ Foot_ Care* *0.349) + (*Chr_ Cad* *0.253) + (*Hba1c_ Above_*9 *0.125) +
(*Ed_ Avoidable_ Dx* * 0.096) + (*No_ Claim_ Id* * 0.028) + (*Zip3* * 0.002) +
(*LOS* * -0.003) + (Chr_Copd * -0.046) + (*Age* * -0.082) +
(*Retinal_ Screening* * -0.082) + (*Chr_ Bh* * -0.13) + (*Chr_ Htn* * -0.155) +
(*Is_ Subacute* * -0.158) + (*Obesity* * -0.17) + (*Extended_ Los* * -0.175) +
(*Chr_ Hf* * -0.319) + (*Wellness_ Exam* * -0.381) + (*Retinopathy* * -0.501) +
(*Renal_ Failure* * -0.56) + (*Hba1c_ Screening* * -0.563) + (*Nephropathy**-0.825) +
(*Dialysis_ Treatment**-2.122) + (*Kidney_ Transplant**-2.127)

Here the number 1.374 which is the intercept indicates that a diabetic patient is 1.3 times prone to visit the emergency than a normal patient. The important features impacting the regression model along with their scores is shown in Appendix D.

Chapter 5

Conclusion and Future Scope

Emergence of Machine Learning and Analytics tools that can make useful inferences and predictions based on available data. The dramatically positive impact Analytics can have on the pressures health systems face to be more efficient and improve clinical outcomes. The project focused on predicting the patients visit to the emergency or whether a patient will be admitted as an emergency, inpatient or readmission member based on their demographic and historical claims submitted. The random forest proved to be the best model in predicting the outcomes and this is evident from the fact that we have less number of data points for any other model to work efficiently. Also, the regression model proved to be beneficial in predicting that a diabetic patient is prone to visit the emergency department 1.3 times more than a normal patient.

The data points that we were provided was linear and without any major pre-processing needs to be performed for model building. However, in a real-life scenario, the data in healthcare industry tends to be highly non-linear. Hence, the future scope of the project could include building a non-linear regression model. nonlinear regression is a form of regression analysis in which observational data are modelled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations. Poisson Regression and sinusoidal regression are some of the models that can be implemented as part of the future scope of the project

Appendix A

Data Dictionary

Description for Member Data Table

Column Name	Description	Constraints
Member_Id	Member Identification Number	Primary Key
Cd_Gender	Gender (M/F)	
Birth_Year	Year of Birth	
Age	Age	
Zip3	First 3 digits of member zip code	
Chr_Count	Count of chronic comorbidities	
Chr_Htn	Comorbidity - Hypertension	
Chr_Copd	Comorbidity - Chronic Obstructive Pulmonary Disease	
Chr_Asthma	Comorbidity - Asthmas	
Chr_Cad	Comorbidity - Coronary Artery Disease	
Chr_Hf	Comorbidity - Heart Failure	
Chr_Bh	Comorbidity - Behavioural Health	
Dialysis_Treatment	History of Dialysis Treatment	
Renal_Failure	History of Kidney/Renal Failure	
Kidney_Transplant	History of Kidney Transplant	
Nephropathy	History of Nephropathy	
Retinopathy	History of Retinopathy	
Tobacco_Use	History of Tobacco Use	
Obesity	History of Obesity	
Limb_Loss	History of Limb Loss (Amputations, etc.)	

Description for Claims Summary Table

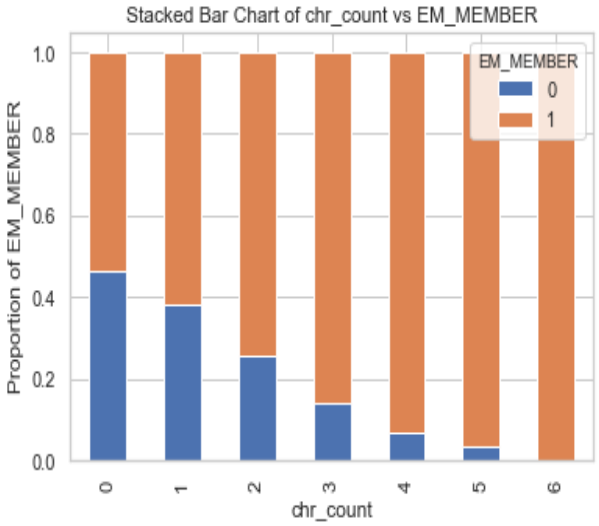
Column Name	Description	Constraints
Member_Id	Member Identification Number	Foreign Key
Claim_Id	Claim Identification Number	Primary Key
Service_From_Dt	Minimum from date on claim	
Service_Thru_Dt	Maximum thru date on claim	
Ed_Avoidable_Dx	Is the principal diagnosis considered avoidable if within an ER setting?	
Is_Subacute	Is the setting Inpatient Rehab Facility or LTAC?	
Extended_Los	Is the overall LOS greater than 30 days	
Wellness_Exam	Was an annual wellness exam conducted?	
Nephropathy_Screening	Was a nephropathy screening conducted?	
Hba1c_Screening	Was an HBA1C screening conducted?	
Hba1c_Above_9	Was HBA1C above 9?	
Routine_Foot_Care	Was a routine foot care screening conducted?	
Retinal_Screening	Was a retinal screening conducted?	

Description for Claims Detail Table

Column Name	Description	Constraints
Member_Id	Member Identification Number	Foreign Key
Claim_Id	Claim Identification Number	Composite PK
Claim_Line_Id	Row Number of Claim ID	Composite PK
Claim_Type	Type of Claim (I/P) (Institutional/Professional)	
Claim_Npi	Claim National Provider Identifier	
Type_Of_Bill	Institutional Bill Type	
Place_Of_Svc_Code	Professional Place of Service Date	
Service_From_Dt	Line from date	
Service_Thru_Dt	Line through date	
Los	Line length of stay	
Discharge_Status_Code	Status of member when discharge	
Principal_Diag_Code	Primary diagnosis code of member	
Secondary_Diag_Code1	Secondary diagnosis code of member	
Secondary_Diag_Code2	Secondary diagnosis code of member	
Secondary_Diag_Code3	Secondary diagnosis code of member	
Icd_Principal_Procedure_Code	Principal ICD Procedure Code	
Icd_Secondary_Procedure_Code1	Secondary ICD Procedure Code	
Icd_Secondary_Procedure_Code2	Secondary ICD Procedure Code	
Icd_Secondary_Procedure_Code3	Secondary ICD Procedure Code	
Revenue_Code	Line item revenue code	
Procedure_Code	Line item procedure code	
Procedure_Modifier1	1st modifier for procedure_code	
Procedure_Modifier2	2nd modifier for procedure_code	
Procedure_Modifier3	3rd modifier for procedure_code	
Procedure_Modifier4	4th modifier for procedure_code	
Apr_Drg_Code	DRG code for procedure code	
Allowed_Amt	Amount allowed for billing purposes	

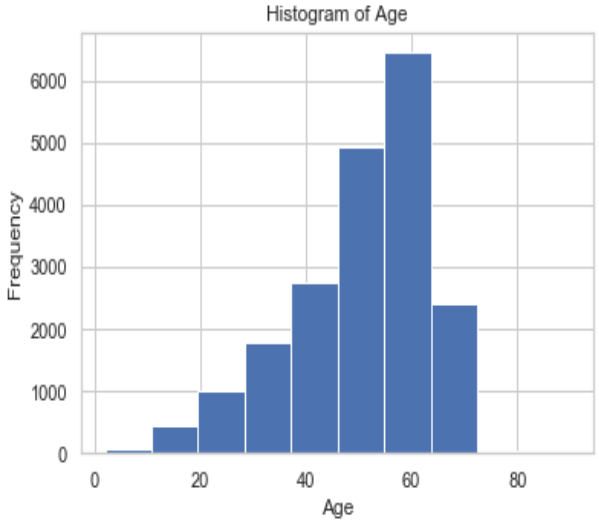
Appendix B

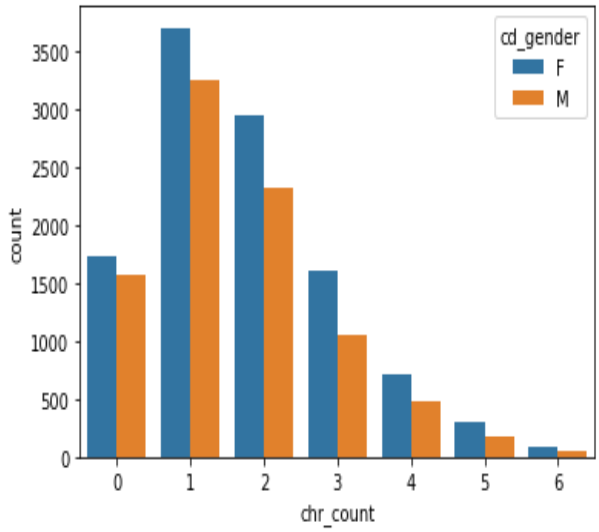
Exploratory Data Analysis



The stacked bar chart shows the number of patients who are identified as emergency and who are prone number of chronic diseases. To explain, the chart depicts more the number of chronic diseases, more is the chance that the patient will be admitted into emergency.

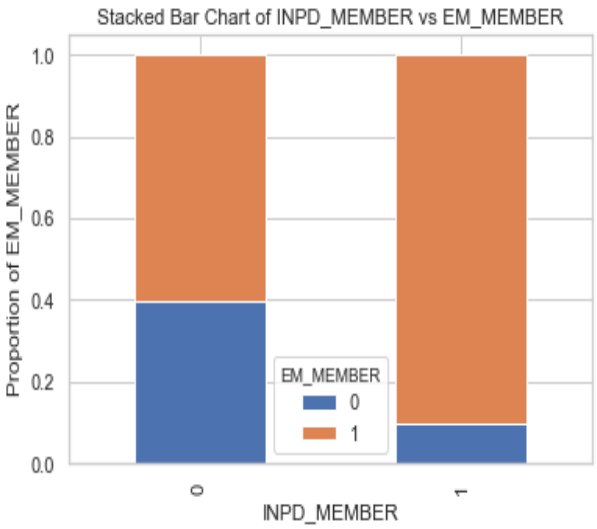
The histogram shows the distribution of members as per their age. Noticeably, there are approximately 9000 members who are aged more than 60

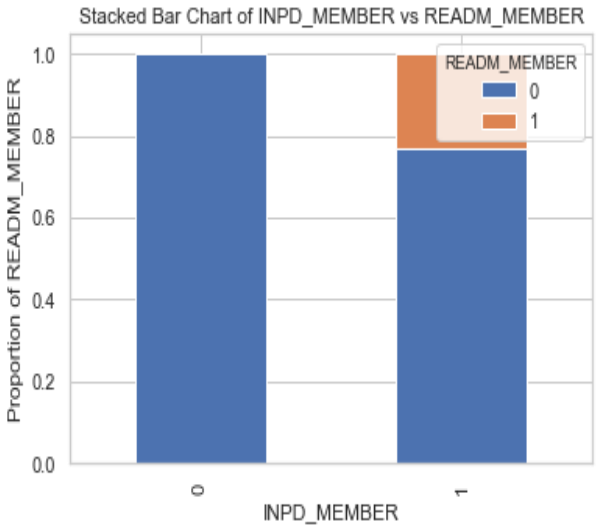




The chart alongside shows the distribution of members as per their gender and the number of chronic counts they have.

The chart alongside, shows the proportion of inpatient member who are admitted as emergency members. To explain, almost 90% of the inpatient members are admitted as emergency patients.





The chart alongside, shows the proportion of inpatient member who are admitted as readmission members. To explain, almost 22% of the inpatient members are admitted as readmission patients.

Appendix C

Feature Engineering for Classification Modelling

Important features and their scores for Emergency Hospitalization

Feature	Scores
LOS	0.301845
No_Claim_Id	0.196558
Age	0.092874
Chr_Count	0.092439
Ed_Avoidable_Dx	0.069815
Chr_Bh	0.047637
Chr_Asthma	0.042637
Hba1c_Screening	0.026708
Tobacco_Use	0.020306
Chr_Copd	0.017342
Chr_Cad	0.01231
Chr_Hf	0.011422
Zip3	0.009797
Wellness_Exam	0.008377
Nephropathy_Screening	0.008291
Obesity	0.007823
Renal_Failure	0.007622
Extended_Los	0.005724
Cd_Gender	0.004493
Chr_Htn	0.004403
Retinal_Screening	0.003203
Hba1c_Above_9	0.002077
Is_Subacute	0.001387
Dialysis_Treatment	0.00131
Nephropathy	0.000952
Limb_Loss	0.000832
Retinopathy	0.000811
Routine_Foot_Care	0.000764
Kidney_Transplant	0.000241

The features marked in Green are the ones which are most impacting the model and hence accommodate for the higher chances of predicting whether a patient will be admitted in emergency or not. The features marked in Red are the ones least impacting the model and hence do not contribute a lot for emergency hospitalization.

Important Features and their scores for Inpatient Hospitalization

Feature	Scores
Los	0.428524
No_Claim_Id	0.207539
Renal_Failure	0.091343
Chr_Count	0.052105
Chr_Hf	0.043476
Ed_Avoidable_Dx	0.040792
Chr_Cad	0.018823
Extended_Los	0.016712
Chr_Copd	0.013315
Age	0.01152
Wellness_Exam	0.010541
Limb_Loss	0.009042
Dialysis_Treatment	0.008433
Is_Subacute	0.008371
Tobacco_Use	0.006784
Hba1c_Screening	0.006765
Chr_Bh	0.005239
Nephropathy_Screening	0.00443
Chr_Htn	0.003939
Zip3	0.003275
Obesity	0.001857
Nephropathy	0.00149
Hba1c_Above_9	0.001367
Chr_Asthma	0.001296
Cd_Gender	0.001293
Retinal_Screening	0.001177
Retinopathy	0.000385
Routine_Foot_Care	0.000095
Kidney_Transplant	0.000074

The features marked in Green are the ones which are most impacting the model and hence accommodate for the higher chances of predicting whether a patient will be admitted as inpatient or not. The features marked in Red are the ones least impacting the model and hence do not contribute a lot for inpatient hospitalization.

Important Features and their scores for Readmission Hospitalization

Feature	Scores
No_Claim_Id	0.285221
Los	0.218661
Ed_Avoidable_Dx	0.099937
Renal_Failure	0.069856
Chr_Count	0.037937
Age	0.03384
Chr_Hf	0.033369
Hba1c_Screening	0.026106
Extended_Los	0.025515
Nephropathy_Screening	0.020594
Zip3	0.016314
Dialysis_Treatment	0.016056
Limb_Loss	0.015886
Wellness_Exam	0.013523
Chr_Cad	0.01199
Is_Subacute	0.009416
Hba1c_Above_9	0.009246
Chr_Bh	0.008771
Chr_Copd	0.00812
Tobacco_Use	0.007224
Retinal_Screening	0.006768
Chr_Htn	0.004982
Cd_Gender	0.004972
Obesity	0.004295
Chr_Asthma	0.003962
Nephropathy	0.003153
Retinopathy	0.002308
Kidney_Transplant	0.001479
Routine_Foot_Care	0.000499

The features marked in Green are the ones which are most impacting the model and hence accommodate for the higher chances of predicting whether a patient will be admitted as readmission or not. The features marked in Red are the ones least impacting the model and hence do not contribute a lot for readmission hospitalization. Interestingly, in all the three categories of hospitalization, LOS (Length Of Stay) and the No_Claim_ID (#of Claims submitted by the member) are the ones that play a crucial role for the purpose of prediction.

Appendix D

Feature Engineering for Regression Modelling

Important Features and their scores impacting the regression model

Intercept	1.374
Feature	Coefficient
Tobacco_Use	2.286
Limb_Loss	1.457
Chr_Asthma	0.996
Chr_Count	0.599
Nephropathy_Screening	0.499
Cd_Gender	0.360
Routine_Foot_Care	0.349
Chr_Cad	0.253
Hba1c_Above_9	0.125
Ed_Avoidable_Dx	0.096
No_Claim_Id	0.028
Zip3	0.002
Los	-0.003
Chr_Copd	-0.046
Age	-0.082
Retinal_Screening	-0.082
Chr_Bh	-0.130
Chr_Htn	-0.155
Is_Subacute	-0.158
Obesity	-0.170
Extended_Los	-0.175
Chr_Hf	-0.319
Wellness_Exam	-0.381
Retinopathy	-0.501
Renal_Failure	-0.560
Hba1c_Screening	-0.563
Nephropathy	-0.825
Dialysis_Treatment	-2.122
Kidney_Transplant	-2.127

The features marked in Green are the ones which are most impacting the model and hence accommodate for the higher chances of predicting how many time a patient will be admitted as emergency. The features marked in Red are the ones least impacting the model.