National College of Ireland

Which geographical, socio-economic and facility attributes predict hospital readmission from Skilled Nursing Facilities

MSc Research Project
MSc in Science of Data Analytics

Michelle Waters
Student ID: X17100020

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	……. …… Michelle Waters ……
Student ID:	…… x17100020……..……
Programme:	Masters in Science of Data Analytics Top Up **Year:** 2021
Module:	… Final Project Submission……..……
Supervisor:	… Jorge Basilio ……..……
Submission Due Date:	…23/09/21……..……
Project Title:	…… Which geographical, socio-economic and facility attributes predict hospital readmission from Skilled Nursing Facilities …
Word Count:	25 pages excluding cover, references…… **Page Count**……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …… Michelle Waters ……

Date: ……25/09/2021……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Which geographical, socio-economic and facility attributes predict hospital readmission from Skilled Nursing Facilities

Michelle Waters
X17100020

Abstract

There are more than 15,000 Skilled Nursing Facilities (SNFs) in the US which play a key role in the rehabilitation of patients post acute hospital care. Hospital readmissions are a significant problem for Health Service and are high at over 20% from SNFs and other post acute care facilities. Understanding the factors that contribute to readmissions is the aim of this study. This research builds on previous work but extends the factors that have been previously modelled to understand. *What makes one Skilled Nursing Facility more likely than another for patients to be readmitted to hospital?* and secondly, *What are the types of socio-economic and demographic attributes of the geographic location of a nursing home that might influence the likelihood of being readmitted to hospital?*

The scale and ambition of the work is unique and important and requires integration and examination of data from across states and levels to improve the generalisability. Support Vector Machine and Random Forest Classifiers were utilised to model predictive factors with Random Forest performing best with 73% Accuracy and 80% Sensitivity for 6 features. In agreement with the literature characteristics of the SNF and Patients are important predictors of readmission, but also identified as important are the local area patterns of hospitalisation and health services utilisation suggesting that norms within jurisdictions need also to be evaluated. This points to an opportunity for Federal Health Administrators to examine utilisation policies and practices across Geographies. Socio- demographic were not identified as the most important factors in this research.

1 Introduction

Hospital readmissions in the US are a significant cause of pressure on Healthcare systems with many being considered preventable. In 2019 alone readmissions from Skilled Nursing Facilities (SNF), which includes 15,331 Nursing Homes and Rehabilitation facilities providing services to Medicare patients cost in excess of €26 billion annually to the system. With nearly 22% of residents who were in short-term SNF care post an acute hospital discharge being readmitted to hospital during or within 30 days of their stay in these facilities. The Centre for Medicare and Medicaid Services (CMS)[1] report on measures associated with the quality of resident care in certified SNFs. Performance of the SNFs across these key measures along with other variables including staffing levels and health inspections has been shown to be linked to the number of patients who are re-hospitalized.

Modern strategic health planning is becoming increasingly reliant on evidence-based policies. Datasets such as those generated by CMS.gov can be used to inform strategic planning and contribute to the prevention of poor healthcare outcomes for patients and also

[1] CMS.gov, Centre for Medicaid & Medicare , https://data.cms.gov/provider-data/search?theme=Nursing%20homes%20including%20rehab%20services

providers of services. The Minimum Data Set (MDS) national database collects data on the quality of resident care measures. MDS assessments are taken in Skilled Nursing Facilities (SNF) at regular intervals on every resident in a Medicare or Medicaid-certified nursing homes. These facilities are increasingly used to care for patients who are beneficiaries of Medicare (after the age of 65) or Medicaid (determined by means) post acute care in hospital until they can return home. Information is collected about the beneficiaries health, physical functioning, mental status, and general well-being. These data are used by the SNF to assess each beneficiaries needs and develop a plan of care. Service provider information is also collected by CMS and includes facility address, ownership and operational data, including staffing types and ratios. These data are also important to analyse for their contribution to hospital readmission.

The MDS quality measures dataset aggregates averages for quality measures for each nursing home and it is proposed that these data can be used to predict patient outcomes including re-hospitalization and attendance at emergency outpatient departments. According to recent research by Fuller et al. (2019) the rate of potentially preventable readmissions and Emergency Department visits from nursing homes residents varies greatly both across and within states, with 5 states having more than 20% more than the national average for both. The aim of this research is to utilise SNF data from the Minimum Data Set including provider information, quality measures and also claims data which measures and ranks SNFs by Patient Readmission Rates. It is proposed that the Readmission Rates will be the target variable to be predicted and socio-economic and demographic attributes for the geographic area in which the SNF is located will be integrated into this data to determine if these factors can be included in a model to predict which nursing homes will have a higher rate of hospital readmissions.

The County Health Rankings[2] aggregates data from a wide range of federal and state primary resources including the long-running Behavioral Risk Factor Surveillance Survey (BRFSS)[3] to provide a holistic list of factors that are linked to Health Outcomes; 1) Length of Life and 2) Quality of Life. It has been suggested however that there is variation within counties for preventable hospitalisations, therefore overall County Health Rankings alone may not be sufficient to predict re-hospitalisation. Geographical Health data can also be found at a more local level than county using the Darthmouth Atlas Project[4] which provides analysis of patterns across smaller health care areas across the US based on CMS data

Using publicly accessible files for post acute care (PAC PUF)[5] which are anonymised datasets derived from the Minimum Data Set which the CMS publishes for purpose of research. This research will be supplemented with socio-economic and demographic data for counties and small areas. The study will seek to model factors that predict readmission to hospital during or after short-term care in Medicare Skilled Nursing Facilities. It is hoped that the model will identify attributes of Nursing Homes that lead to better patient outcomes. The primary data includes more than 15,000 nursing homes in the US and this research is unique in that it looks at the attributes of Nursing Homes instead of the attributes of individuals and seeks to broaden the factors that have been previously modelled. Key questions of this research are:

---

[2] Countyhealthrankings.org/County Health Rankings, https://www.countyhealthrankings.org/what-is-health

[3] Behavioral Risk Factor Surveillance System https://www.cdc.gov/brfss/index.html

[4] Dartmouth Health Atlas, https://data.dartmouthatlas.org/

[5] PAC PUF Post-Acute Care and Hospice Provider Data 2018 | CMS)

- What makes one Skilled Nursing Facility more likely than another for patients to be readmitted to hospital?
- What are the types of socio-economic and demographic attributes of the geographic location of a nursing home that might influence the likelihood of being readmitted to hospital?

The scale and ambition of the work is unique and important and requires integration and examination of data from across states and levels to improve the generalisability of the work. In the following section related work is presented which further highlights that hospital readmissions is a multi-factorial problem and supports examination of Social and Behavioral Determinants of Health (SBDoH) to better understand it. There is an opportunity to utilise this research also to better understand how to make researching this area more accessible to non-domain experts.

2 Related Work

This section presents the body of work that has been drawn upon and is relevant for this paper. The questions under research give rise to a broad examination of factors to be considered and therefore work has been reviewed across three main themes; Rehospitalisation studies, Skilled Nursing Facility studies and work that utilises socio economic and demographic factors to model outcomes. Work of particular interest that this study builds upon includes the papers previously introduced in the earlier section by Fuller et al. (2019) and Bartley et al (2020). In addition a report commissioned by the Dartmouth Institute for Health Policy and clinical practice in 2013 on the "Revolving Door of Hospital Readmissions" is an important reference for this work. (Dartmouth, 2019)

Bartley et al. examined factors associated with the rehospitalisation of post acute care (PAC) patients from SNFs. They focussed on the examination of the Quality Star Ratings from the CMS, Medicare Compare service to understand if an association could be found between the ratings which are an aggregate of results for performance and quality metrics such as staffing, quality measure ratings and facility inspections scores. They found that higher star rated facilities have lower readmission levels with quality ratings being most important and staffing and inspections ratings having no association. The study was small with no socio economic or demographic variables included. In similar research by Fuller et al. in 2019 they sought to understand the cause of variance in readmissions from SNFs and whether star ratings are a factor. This was a significantly larger study which investigated over 12,000 SNFs with CMS certification across the US over two years. This study used proprietary software to analyse deficiencies such as instances of inadequate care to study SNF performance at a more granular level. A recommendation of the research is for cross geographical comparison of SNF performance across states to better understand these location based factors. The need to understand the factors other than clinical, facility and patient profile factors has been the focus of work by the Darthmouth Institute for many years. Their research which is generated from Medicare data suggests that the causes of hospital readmissions are complex and are not well understood. They suggest that one important factor that is seldom studied is the local pattern of hospital utilisation and that areas that have higher underlying hospital utilisation will also have higher readmissions. They also counter-intuitively claim that more healthcare is not better and how in areas where there are more beds per capita there will be more unnecessary hospitalisations. Each of these studies points to further opportunities for research and in particular to extend the evaluation to include all of these factors together where possible.

2.1 Rehospitalisation Research

The importance of addressing hospital readmission is a long-standing problem. Understanding the factors that contribute to hospital readmissions from post acute sectors has been an on-going challenge for many years. Research conducted nearly 30 years ago by Corrigan and Martin (1992) using regression analysis highlights that the problems and methods of analysis remain largely similar today focussing in the main on the clinical presentation of patients and less on the social determinants of health. In a literature synthesis Gaugler et al (2007) investigated over 700 reports relating to SNFs and admissions again highlighting a large body of interest in understanding the relationships between patients and SNFs. In a Multivariate Statistical Analysis by a UK team of researchers, regression models were utilised to predict readmission to hospital using clinical and health profiles of patients. The model performed better on those most at risk, however the model tuned to identify moderate risk profiles performing less well on sensitivity and specificity measures.

Much of the research into readmission is focused on clinical health profiles of beneficiaries. Yoo et al. (2015) also utilised patient and facility data to predict the cost of readmission from SNF facilities. As part of their study they also looked at care interventions provided but found there to be no predictive value whilst the characteristics of the residents and the nursing home itself were important. A recent work by Howard et al. (2021) saw them utilise Machine Learning (ML) segmentation profiling methods to understand what effect the relationship between patient profile groups and the type of post acute care (PAC) settings they were discharged to has on readmissions. The PAC settings includes SNFs, Home Care and Inpatient Rehab care. The study was implemented in two stages, first using PCA and cluster approaches to segment, followed by regression analysis to develop models. It identified that those discharged to SNFs did best in all cases other than for very high risk patients groups.

2.2 Skilled Nursing Facility Studies

Other studies that provide insight into factors that could be relevant to understanding the differences between SNFs include a 2015 work by Lepore et al. which investigated nursing home characteristics and how they vary dependent on the proportion of Medicare versus Medicaid Beneficiaries resident in the nursing home. The study identified differences that suggest higher number of Medicare patients leads to higher investments in the environment and culture of care. A limitation of this study is how to define and measure improvement in culture to be generalisable. Newman et al (2014) focussed on SNF performance metrics including health deficiencies to model links with hospital readmission and from their study found that better staffing results in improved rates, whilst positive facility inspections were also important. Previous research had not found staffing to be important so again it is clear that there are difficulties in finding consensus on which are the most important factors to consider.

Flanagan and Boltz (2021) focussed on long term nursing home care rather than PAC but also utilised classification models to predict who will need long term care. Similar to other studies the authors analysed patient data and used explanatory logistic regression. The study could be improved by comparing against a second or third model. Jacobsen et al. (2017) in a qualitative based study sought to understand preventability of hospital readmissions from SNFs by interviewing patients who had been readmitted. Their study suggests a mix of hospital and/or SNF responsibility contributes to readmission and again

points to other factors outside of the SNF and Beneficiary characteristics that should be considered.

2.3 Geographic and Socio Economic and Demographic Focus

Jacobsen et al's paper is a reminder that the attributes of the SNF local area such as other interrelated health services must also be considered factors for investigation. Socio demographic analysis is widely used in research and underpins much of the work of health analysts as they try to identify factors causing or contributing to health issues. Health is a multifactorial problem and Social Behavioral determinants of Health (SBDoH) need to be considered. Kansagara et al (2011) conducted a study to summarise and synthesis research on models predicting readmission to hospital. Of the 26 models they reviewed few examined social determinants of health., they concluded that most of the models deigned for either comparative or clinical purposes perform poorly.

Tan et al (2020) reviewed the inclusion of Social Behavioral Determinants of Health and found that they have not been shown to improve health predictions. However there is a gap in the research studied that examines these factors as it relates to readmissions from SNFs. Wrathnall and Belnap's 2017 research focussed on modelling high cost patients using logistic regression and decision tree methods. Their risk adjusted model, utilised health records, clinical data and some socio-economic data as part of a retrospective study. Improvements in predictions of high cost and at risk patients were achieved and could potentially be valuable for a broader generalisation to SNF facilities. One of the best datasets to help explore SBDH is the County Health Rankings. Nagasako et al (2018) explored differences on populations health at a subcounty level utilising this data. Implementation of their design included PCA, Pairwise Correlation, Linear Mixed Modelling. In a similar method Nuti et al (2019) utilised unsupervised methods in the form of cluster analysis to segment groups without apriori outcome data, a good approach that could be supplemented with classification.

The literature points to an opportunity for including more factors in the model to identify Increased Risk of rehospitalisation. A study by Frizzel et al (2017) related to studying readmissions specifically for heart failure patients presents methodologies that will be similar to this work. Their analysis used multiple ML methods to understand the problem compared Decision Tree augmented with Bayesian Network, Random Forest Algorithms and Gradient Boosted Model compared with logistic regression. In the following section the approach, design and implementation of this research are detailed.

3 Research Methodology

3.1 Introduction

The research procedure followed for this project is primarily based on the CRISP_DM framework. CRISP-DM is an approach used in Data Science projects and is applied in an iterative way such that the original problem may be refined. The stages are Problem Definition, Data Exploration, Preparation, Modelling, Evaluation and Deployment. Figure 1 represents the key steps undertaken in this research.

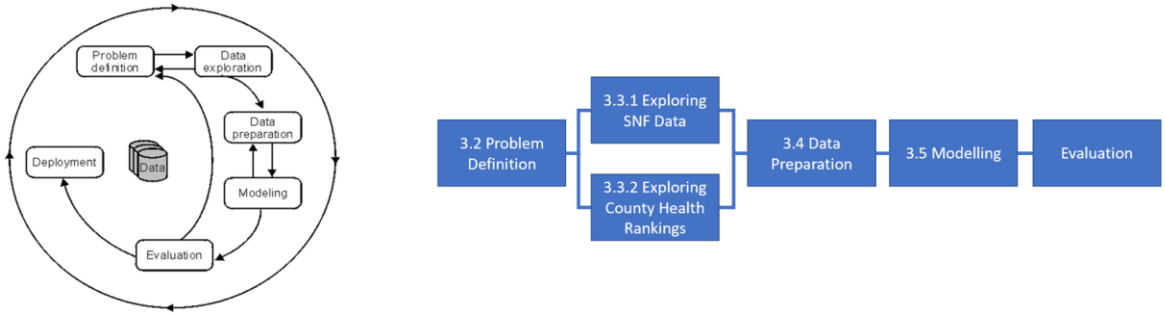


Figure 1 - Research Methodology CRISP_DM

The research design is to treat separately the SNF datasets and County Health Rankings for the initial selection and processing stages of the research as each have unique processing challenges and it also supports iterative learning.

3.2 Problem Definition

The first stage of the CRISP_DM approach is to identify the primary area of research and the associated dataset. This stage consists of identifying the target problem and the primary data or subset of variables or data samples, on which discovery is to be performed. For this research Primary Data came from the CMS MDS datasets relating to Skilled Nursing Facilities which includes the proposed target variable; 'Hospital Readmissions'. The secondary datasets were selected to include socio-economic and demographic factors that could be predictive of readmissions and includes dataset from the County Health Rankings & Roadmaps and the Dartmouth Health Atlas.

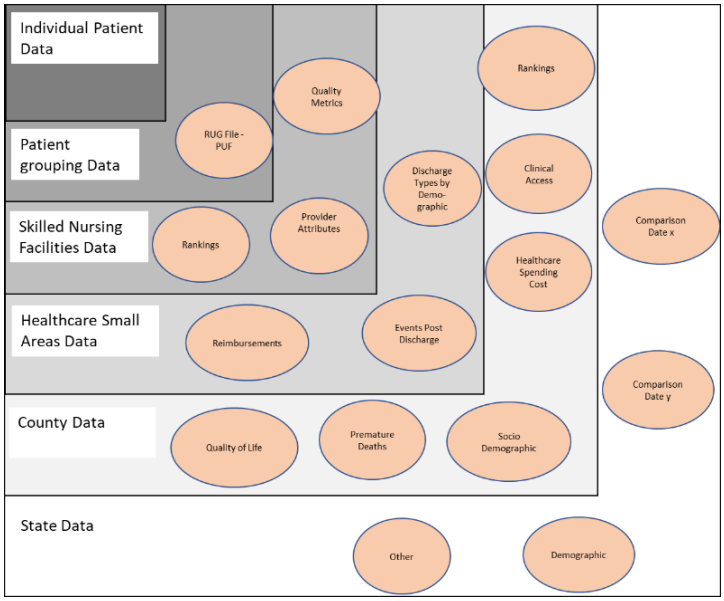


Figure 2 - The Skilled Nursing Facility Ecosystem

The problem of modelling for risk of readmission is defined by the volume and complexity of data available in the ecosystem and is both an opportunity and barrier to progressing broadly generalisable research and the time required to examine and access all available data. Figure 2 presents an ecosystem view of the levels of data that have been selected as having relevance for this research. A level of domain knowledge is important for understanding the relevance of factors and their interaction. Acquiring this knowledge is an

additional challenge of this research and the iterative approach promoted by CRISP_DM supports exploring variables to help make selection and processing decisions but is time consuming. A summary of the primary and secondary data follows.

3.2.1 CMS Medicare Provider Datasets

1) Skilled Nursing Facility Provider Data: These are the official datasets used on Medicare.gov provided by the Centres for Medicare & Medicaid Services. These datasets allow comparison of the quality of care provided in Medicare-certified skilled nursing facilities and nursing homes nationwide (Table 1)

Table 1 – SNF Attributes for Modelling

Skilled Nursing Home (SNF)Provider Info	Facility Information	Performance
• Type of Provider • Single or multiple facilities • Did provider change in the last 2 years? • Ownership • When first approved by Medicaid services?	• Number of certified beds • Average number of resident p/days • Age of the facility • Region • Town • State • Nursing ratio • Types of Staff/Services • Investment	• Fines in previous years • Reimbursement • Number of complaints • Events/Quality/Performance to date • Reimbursement of fines • % Readmissions to hospitals within 30 days (Target Variable) • % Emergency room visits within 30 days

2) Medicare Provider Utilization and Payment Data: Post-Acute Care and Hospice: The PAC PUF dataset includes 30,467 provider records with 122 variable and requires pre-processing to select data relevant only to SNFs. In order to examine hospital readmission rates both files will be utilised provide as comprehensive a profile of the 15,000 plus Skilled Nursing Homes across the US, including identifying demographics of patients, Management of the facility and Performance and Quality Results.

3.2.2 Dartmouth Health Atlas

The Data Health Atlas provides data on specific regions including hospital referral regions (HRR) and hospital service areas (HSA). The data available provides an opportunity to supplement county data with a more granular level of data where it is possible. The key datasets from Dartmouth include:

- Medicare average cost of reimbursements which has 13 features and nearly 40,000 records pre-processing. This data can be filtered by SNF and Hospital for county and local areas at HSA and HRR level.
- Post-Discharge events - Data includes 30-day readmission rates, emergency room visits within 30 days of discharge, and follow-up visits within 14 days of discharge for Dartmouth areas. There are nearly 60,000 records with 22 features.
- Medical Discharge Rates – Gives information on what the patient was being treated for and includes counts and ratio values. Pre-processing there are over 1 million records and 21 features.

3.2.3 The County Health Rankings & Roadmap

To examine whether variables that predict Community Health across counties in the US including socio-economic, demographic and behavioural data could also predict hospital

readmission from Skilled Nursing Facilities the County Health Rankings & Roadmap dataset was chosen. The Rankings use more than 30 measures that help communities understand how healthy their residents are today (health outcomes) and what will impact their health in the future (health factors). Pre-processing, this dataset has 3141 records for all counties across the US including aggregated data for each state. There are 245 variables and a challenge will be selection of the most relevant attributes to include. A subset of the measures are broken out by race and gender demographics, however on assessment of the data there are data sparsity issues which will need to be addressed if these measures are to be usable. *Outcome Measure* are Length of Life and Quality of Life. Some of the *Health Factors* are Health Behaviours, Clinical Care, Social & Economic Factors and Physical Environment.

3.3 Data Exploration

According to Witten and Frank (2005) dataset exploratory analysis should be conducted to understand the potential for using the most appropriate DM methods and identify important features and is an iterative process. Exploration to understand interactions and opportunities was conducted for all variables and included correlation analysis, plotting of relationships between the Target Variable which as part of this process was engineered as 'Increased Risk'. Dimension reduction techniques were also utilised where appropriate. Early exploration and processing was conducted in Excel and clearly redundant features were removed, this reduced the size of data to be read into R. Integration utilising geo-coding of Facility, PACPUF and Dartmouth data required matching records which in some cases were missing. Some of the challenges encountered in combining these data included:

- Aligning the years of data available
- Naming convention consistency
- Data Types and how many features/records matched
- Feature Engineering no feature existed to identify whether the SNF is part of a group or chain and this was engineered and could be important.
- Geographic Levels available and consistency

The SNF csv file of 11,589 records and 61 multivariate features and the Health County Rankings csv file of 2,952 records and 36 multivariate features were read into R where additional processing was conducted. This involved conversion of features to correct type such as factors, numerical and nominal. Encoding decisions were also made based on levels of heterogeneity and cardinality. State and Health Service Areas were encoded as character as cardinality is too high for analysis. The date of certification was transformed to 'Tenure' and a new engineered feature 'Occupancy' created as a function of the SNF 'Certified Beds' count and average number of residents per day.

3.3.1 Inspecting and Transforming SNF Data

There are a large number of variables and an initial focus is to decide if some can be identified at this stage to remove and also to understand which are most likely to be relevant to the research question and retained. Features which are sparse, are not normal, are duplicate, redundant or correlated may need to be removed. It is important to remove noise from the data so models can be trained successfully.

a) Descriptive Summary Review

Both R dfSummary and Describe() from the Hmisc library were used for inspection and evaluation of the variables. Each provide summary descriptives and together provide a good overview of the data. Using these functions information is provided for: data type, count, missing values, mean, standard deviation, coefficient of variance, distribution plots, inter-quartile range and for factor variables proportions and frequencies are given. Inspection highlighted a number of features with large numbers of missing values for minutes of care provided in physiotherapy (PT), occupational therapy (OT) and speech and language therapy (SLT) and although potentially relevant these features are too sparse to be suitable for imputation methods. Several numeric variables had skewed distributions with outliers whilst factor class imbalance looked to be a potential problem for some.

b) Correlation Evaluation

To better understand relationships and interactions of the data correlation was conducted for all numeric variables. Multicollinearity can be an issue in high dimensional space and plotting in a correlation matrix is an effective way of understanding what is contributing and how.

c) Missing Values Analysis

A further evaluation of sparsity was conducted and missing values were identified completely at random and subsequently imputed using multivariate imputation by chained equations (MICE) and a random forest approach.

d) Initial Insights and Next steps

The outputs of the summary descriptives, correlation and missing value analysis provides the following insight that informs the next steps required for exploration and processing.

- There are 46 numeric, 6 nominal and 8 factor variable types in the SNF integrated file
- 3 variables were removed due to missing values (PT, OT, SLT)
- Others were removed due to imbalanced proportions. The Special Focus Feature which relates to SNFs that are under review for quality (SFF) had just 0.1% of records classed as 'Y' and it was similar for the Sprinkler System 'N' class.
- It was observed that there is an overall lack of correlation of the data and that the potential target variable 'Standardised Readmission Rate' is weakly correlated with other independent variables.
- All potential target variables associated with readmission will need to be evaluated against each other and all data separately.
- The selected target needs to be made into a binary class to allow further analysis
- The factor variables should be plotted and assessed for correlation against the target
- It was expected that the 'Weighted Total Score' which is an index value representing SNFs performance on quality health metrics could be related to rehospitalisation's, however the correlation plot does not indicate any relationships. The raw data (SNF deficiencies file) feeding this index value is available and may also provide opportunity for increasing information gain through identifying components that can be utilized in the data as an alternative.
- Resource Utilisation Groups variables (RUGs) could be treated separately in a PCA analysis to reduce dimensionality but retain information on the type of care that is provided in the SNFs

e) Target Class Identification & Exploration

Modelling to understand the factors associated with readmissions to hospital from SNFs will be trained on the target variable and it is important that it is possible to identify correlation between independent variables and the target class. Secondly features that are highly correlated with the Target could cause confusion in modelling, therefore it is important to assess these criteria. The target feature should also be binary coded to Y/N and the threshold selected based on business logic and the literature.

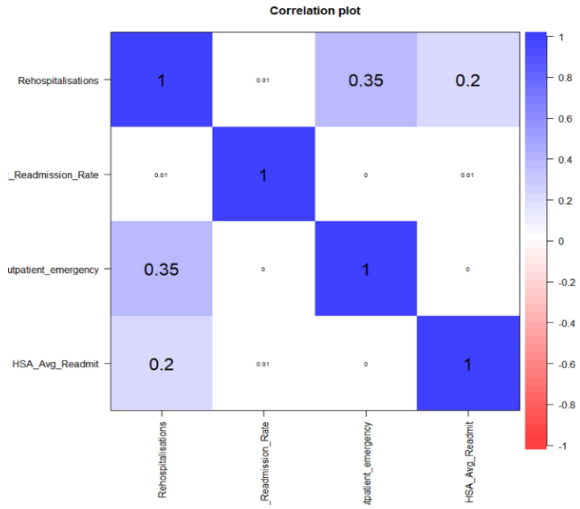
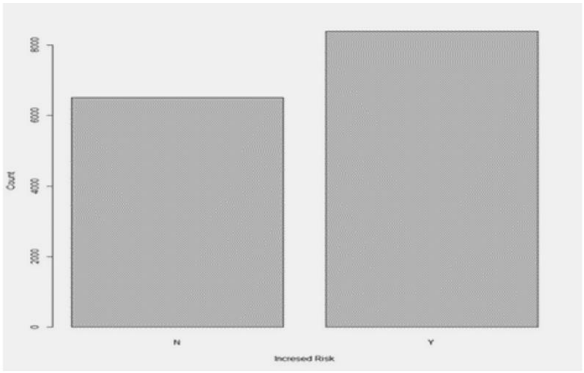


Figure 3 – Target Variable correlation Plot

The correlations plot in figure 3 includes the features originating from the original SNF MDS file, includes ‘Rehospitalization Rate’, ‘Standardized Readmission Rate’ and ‘Inpatient Emergency Rate’. Also included is the ‘HSA Average Readmittance Rate’, this relates to readmissions for the Hospital Service Area that the SNF sits in. It would have been expected that several of these factors would have been highly correlated however only Rehospitalization and Outpatient Emergency show some correlation. Given that ‘Standardized Readmissions is an adjusted standardized value processing applied seems to have affected the meaningfulness of the value for comparison against other unprocessed values in the dataset and therefore ‘Rehospitalizations’ which is a raw data value is better suited to be the Target Class.



Target Variable – ‘Increase Risk’		
***Yes** to Increased Risk*	56%	8389
***No** to Increased Risk*	44%	6510

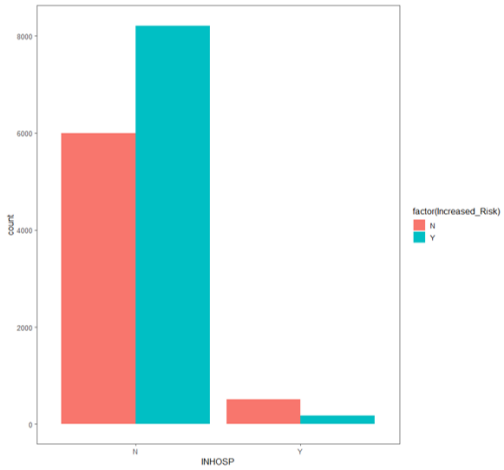
Figure 4 - Increased Risk Proportion

Readmission Rate will be removed whilst HSA_Avg_Readmission and Outpatient_emergency can be retained as there is no multicollinearity. Rehospitalization rate was then transformed to a binary class to be modelled against. The threshold for the new

variable ‘Increased_Risk’ is above the average value for rehospitalization in this data which is 19.74, this value aligns with figures of SNF average readmission published the CMS. The bar chart representing Increased Risk in figure 4 shows that 56% of SNFs fall into the high risk category and 44% do not.

Examining relationships against the target variable class enables further decisions to be made on what factors to retain or remove. Some of the categorical variable were initially examined including ‘Ownership’, ‘Certification’, ‘ResidentFamCouncil’, ‘In Hospital’ and ‘Change of Ownership in last 12 months’ amongst others. Some finding from this stage of analysis include:

- Fig 5 gives that when an SNF is in a hospital setting there is less risk of rehospitalisation
- Risk seems lower in non profit and government SNFs compared to for profit when the ownership feature is examined. Both Ownership and In-hospital assessed together for Increased Risk show that Government Ownership Types that are located within Hospitals have the lowest risk of Readmission. (Fig 6)



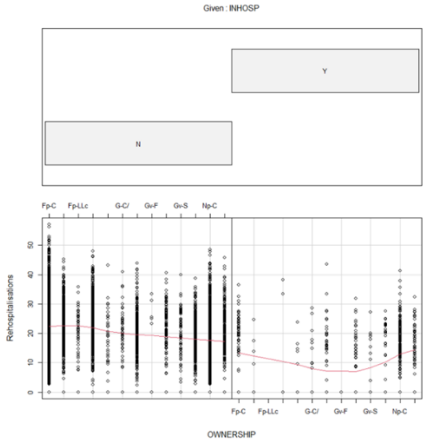


Figure 5 – Increased Risk for SNFs ‘In Hospitals’

Figure 6 – Rehospitalizations given ‘Ownership’ and ‘In Hospital’

- SNFs without any form of family of resident council seem to have lower risk of rehospitalisation. Interesting that this shows that resident & family councils seem to be unrelated or inversely related to rehospitalisation’s
- In analysis of risk dependent on the features family council and change of ownership in the last year the numbers are very low for changed ownership.
- There is a slight reduced risk of rehospitalisation’s from Continuing Care Residential Facilities (CCRC)
- The numbers of SNFs that have changed ownership are low, but from this group it does appear that it increases the risk of rehospitalisation
- From initial analysis it was decided that it would be most useful to transform/collapse categorical variables at this stage before further analysis would be conducted

f) Collapsing Categorical Variables

Issue with categorial cardinality can cause issues in classification modelling, particularly when using Decision Trees where splits can be biased to selecting factors with higher

numbers of levels. Another reason to reduce the number of levels of a factor using dummy coding or other methods is to enable linear modelling that require number values to calculate distance between datapoints. A risk of collapsing variables is loss of information and therefore decisions on how to split classes should be supported by the literature. The features selected for transforming are Ownership type, Resident/family council and Certification Type. The **Ownership feature** has 13 levels and is reduced to 3 macro-categories already defined: 'for profit' (70%), 'government' (6.8%) and 'Not for Profit' (24%).

Figure 7 – Increased Risk based on Ownership Type Collapsed

The **Resident/Family Council feature** was reduced to two levels - has a council (family and/or resident) is Yes (96%) and No (4%)

The **Certification** variable identifies if SNFs are certified for Medicaid, Medicare or both types of beneficiaries and this was binary transformed to show if SNFs are certified for Medicare (97.8%) or No (2.2%)

g) Evaluation of Numeric Features

The correlation plot highlighted some numeric variables in particular to be investigated and assessed for their importance to the research question. Correlation analysis has shown that the '**Beneficiary risk score'** which is based on Hierarchical Conditional Categories[6] which assign risk scores to patients based on scores weighted to illnesses and health conditions. Lower values are better and SNFs with higher average Beneficiary Risk score could also have an Increased Risk of readmission to hospital which would align with some of the conclusions from the literature reviewed. Histograms and boxplots were utilized for all variables to examine variable distributions and to explore individual relationships. Plots were also utilized to identify outliers and the base data file was then examined to understand the reason for the outliers. An evaluation of the effects of numeric variables on Increased Risk of Rehospitalization resulted in nearly 20 features removed from the analysis for reasons outlined here whilst other variables were identified for further processing. The remaining variables had varying degrees of relationship with the Target Variable and were not be eliminated at this time.

[6] Hierarchical Conditional Categories https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/Research-Reports-Items/CMS023176

- **No relationship with Increased Risk**, which was evident where the mean values for both classes of Increased Risk are the same, these variables included 'Certified number of Beds', Physio-therapy Hours, 'Average Service Days' and 'Average Pay' and the two engineered variables 'Occupancy' and 'SNF Chain/Franchise'. Five quality measures related to vaccinations and health outcomes for beneficiaries were also removed due to no relationship whilst the final quality measures 'Pneumococcal Vaccine' was retained. In total 11 variable were removed for the reason of no relationship.
- **Five Variables that are duplicate** in nature for example counts and percentile features both represented were removed including 'Adjusted Rate for payments' (average), 'Nursing Assistant' and 'Registered Nursing' hours per day were removed as they were consistent with the 'Total Nursing hours' relationship with Increased Risk with was retained. However 'Licensed Practitioner Nurse' hours affected Increased Risk differently – increasing risk and so were retained. 'Rehospitalisation' and 'Standard Risk Readmission' were also removed for reasons of duplication and collinearity.
- **Beneficiary Attributes** were retained which are the average score for the SNFs patient population and include 'Beneficiary Average Risk Score' as discussed, 'Number of Chronic Conditions', 'Average Age', 'Percentage Male' and 'Percentage White' and all saw mean differences for the positive and negative classes of the Target Value. It is interesting as can be seen in figure 8 that perhaps counter-intuitively the risk of readmission decreases for higher ages. The other plots are more aligned with expectations – when an SNFs beneficiary population has a higher count of chronic conditions, risk of readmission increases.

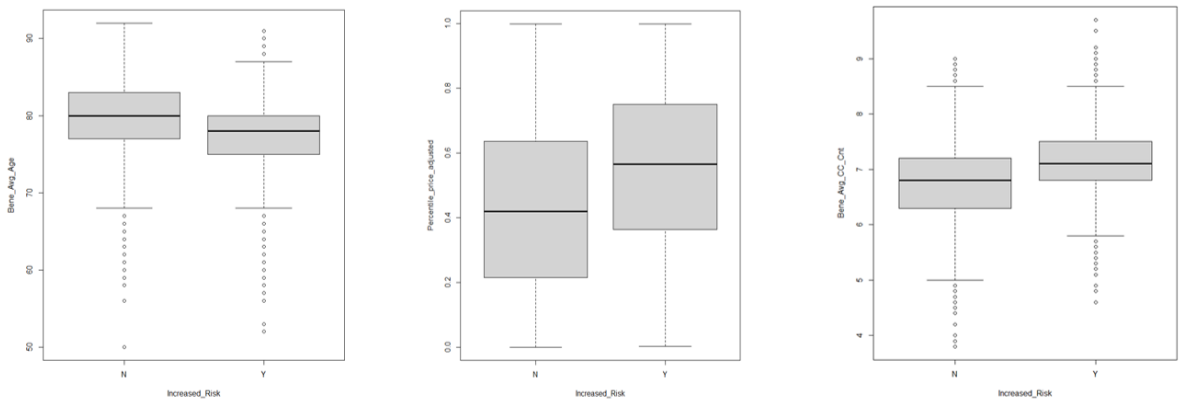


Figure 8 – 'Increased Risk' Bar charts for Age, Avg. Cost of Care & Chronic Conditions

- **Dimensionality Challenges.** 8 correlated features which have group membership Resource Utilisation Group (RUGs) and aggregate variable that are an index function of a set of features 'Deficiencies' were both identified for differing reasons to be candidates for Dimensionality Reduction analysis.
- **HSA features** which present average rates and values for the local area in which SNFs are located were retained as relationships were observed with the Target Class.
- **Outliers and Data Normality** identified were not input errors or anomalies. Some of the features in the data are of a heterogenous nature with wide ranges e.g. 'HSA Beneficiary Population' on inspection of the outliers for this feature it was identified that HSA area '45131' is the largest area of care and is located in Texas the second biggest state in the US. There are 95 skilled nursing facilities in this HSA and therefore it was not removed. Feature removal and missing value imputation were the approaches used to deal with issues of normality.

h) Dimensionality reduction RUGs

A correlation and normality plot in figure 9 was created to asses the RUG data for suitability for Principle Component Analysis. The matrix shows some highly correlated areas and therefore supported PCA analysis. A PCA script using R statistical package was employed on the data. The function prcomp supports scaling and rotation of the outputs and was used. PCA that can explain over 70% of variance in a dataset is considered a good result. Eigenvalues of <1 are retained and the scree plot here shows that 4 components should be retained representing a cumulative variance of 76%. (Joliffe, 1972). The original RUG features were removed from the dataset to be replaced by the PCA values.

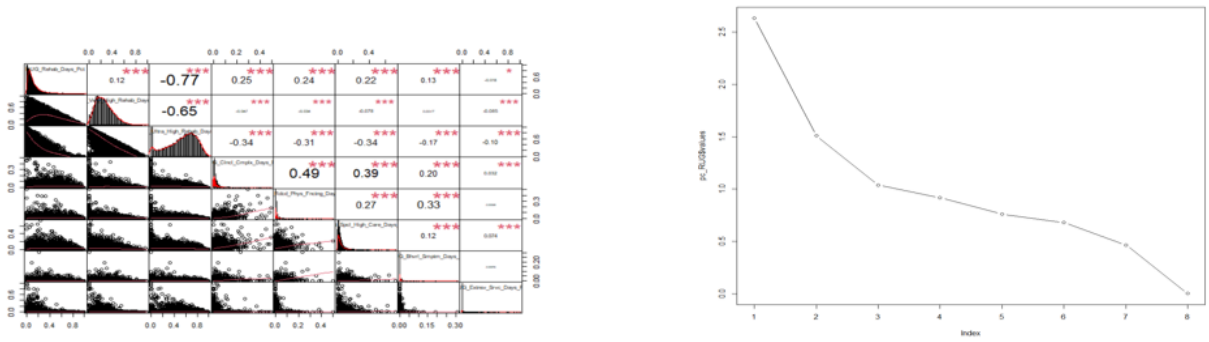


Figure 91 – RUG Correlation and PCA Analysis Scree Plot

i) Dimensionality Reduction SNF Deficiencies File

The Total Weighted Value which in aggregate represents performance of an SNF on the Health Survey and Inspection data and comprises of 31 Deficiency categories did not show any relationship with the target class and so a full evaluation of the input data has been undertaken to separate out the main components of variation which may be more important as factors in the model. After a first analysis two multicollinear features also aggregate values were removed, reducing the data to 29 features. Iterations to test for sufficient numbers of PCA starting at 4 and increasing to 8 were run. Only 54% of cumulative variance was represented in the top 8 PCAs and it was decided that these it would not be useful to combine a large number of PCAs for deficiencies into the main data. Instead the two summary features were added, 'Total Health Deficiency' and 'Total Fire Deficiency'.

3.3.2 Inspecting and Transforming County Health Rankings Data

It has been previously described that geographic and social and behavioral determinants of health are important to the research question however in total the original dataset compromises 245 features with varying degrees of missingness throughout. Analysis of these data mirrored the approach undertaken to understand the SNF however it was quickly evident that methods to capture variance in a reduced set of component would help manage the data. Pre-processing activity reduced variables to 36.

a) Descriptive Summary Review

The summary descriptive functions were again used for this analysis to support getting to know the data as a whole. The data comprises of *Outcome Measure* which are Length of Life and Quality of Life. Some of the *Health Factors* are Health Behaviours, Clinical Care, Social & Economic Factors and Physical Environment.

- The data consists of 32 numeric features, 3 nominal and 1 categorical variable.
- The dataset is relatively clean with few missing values, the variable with most missing values is for ‘Graduation Rate’ with just over 12% missingness, suitable for imputation.
- A correlation Plot for all numeric data was also produced.
- Histograms and Boxplots were created for a number of features to support decisions to retain or remove and to evaluate outliers
- The data is overall well distributed and normal and there are some strong zones of correlation in the data that could be captured within the variance of PCAs. ‘Mental Health’ and ‘Physical Health’ are strongly correlated. (Figure 10)
- Three other variable ‘Physician Rate’, ‘Dentist Rate’ and ‘Mental Health Provider’ are also correlated but on investigate they are in the moderate zone, just below 0.7
- Eight variables were selected to be removed based on lack of contribution.

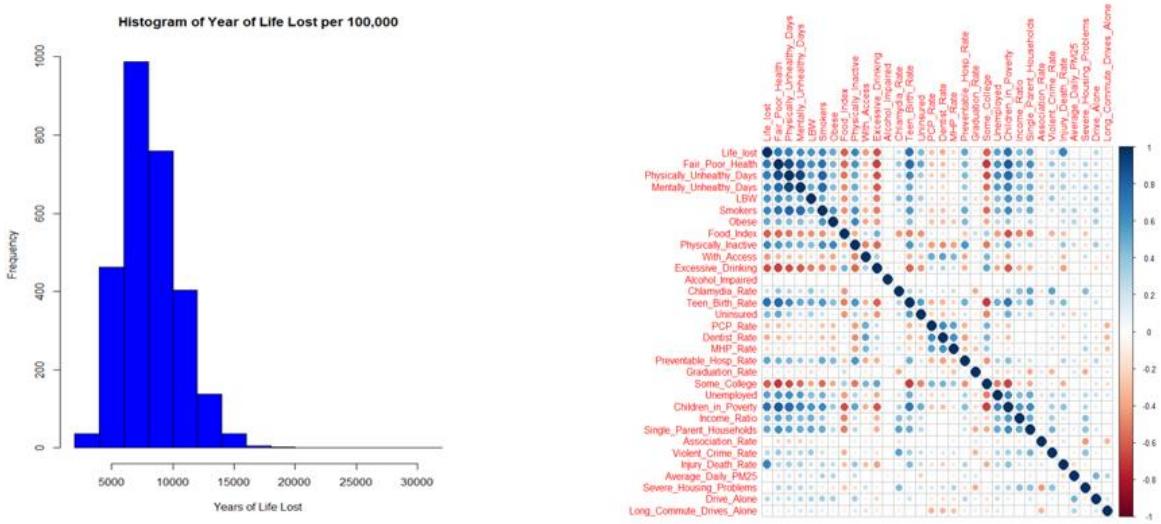


Figure 10 - Histogram of Year of Life Lost/correlation for County Health Rankings

b) Principle Component Analysis

The same methods as previously described were to identify the principle components of variance in this data. 25 features were included in the model, geographic identifiers were excluded. An MSA of 92% on the KMO test suggest the data is suitable for PCA. Five components were identified representing 70% of the cumulative variance. Analysis of the contributing features for each PCA resulted in the following segment descriptions.

- TC1 – Wide Range of Health and Socio Economic Issues
- TC4 – Poor Opportunities and Disadvantaged Youth
- TC2 – Good Health and Opportunities
- TC3 – Uninsured and perhaps working class
- TC4 – Overweight & Unfit

The County Health Rankings principle components were then subsequently integrated with the main SNF dataset ready for the next stage of analysis.

3.4 Data Preparation

The primary aim of this study is to build a model that will maximize sensitivity of predicting SNFs that are at risk of high rates of hospital readmissions. Some final processing steps were required to prepare the data for modelling. Post integration there were 36 features for 14,794 records, a very high number of dimensions relative to records.

- Matching of records to integrate the SNF dataset and County Health Ranking using geographic identifiers
- Encoding of the newly introduced features and components
- Evaluation and imputation of missing values was conducted using MICE package in R. The variable with the most missing values is 'Beneficiary is Male %' with 2562 missing. There are differing views on the proportion of missingness that can be imputed as discussed by Dong & Peng (2013) ranging from less than 10% up to 25-30%. At 17% of values missing for this feature it was decided that it and all features would be included.
- Factor variables were dummy encoded
- The Boruta package in R is a wrapper function used for selecting the most important features and uses bootstrapped decision trees in a random forest (Kursa & Rudnicki, 2015)
- Class Imbalance was reviewed but with 10% difference between classes it was decided rebalancing was not required
- The dataset was split into Training & Testing in a 70:30 ratio.

3.5 Modelling

Three models were selected based on their robustness to missing values, multivariate data types, multicollinearity and also have built in functions to support scaling and tuning and include Support Vector Machine (SVM), Random Forest and Elastic net logistic regression.
The models each have advantages and disadvantages and will be described in more detail in the next section. Each model was trained and tested on the same set of factors during tuning and iteration. Elastic Net was least successful and the Random Forest Model had the most predictive results. The final stage of this project is to evaluate and compare the models against the research questions, do they meet the research of understanding factors linked to readmission and incorporate multivariate data efficiently?

3.6 Evaluation

The primary aim of this study is to build a model that maximizes the sensitivity of predicting SNFs that are at risk of high rates of hospital readmissions. This is a binary classification problem and seeks to identify the best algorithm for the dataset building on previous research. Criteria included in the evaluation were simplicity, overall robustness and parsimony in addition to performance measures.

4 Design Specification

The aim of this research is to understand what makes it more likely that patients from any given Skilled Nursing Facility will be readmitted to hospital and a second aim is to understand what types of socio-economic and demographic attributes or other factors of the geographic location of an SNF will influence the risk of being readmitted to hospital? The design specification needs to support achieving both of these aims in the most efficient way possible. The research type is retrospective with secondary analysis design. It utilises data from several sources, therefore strategies for pre-processing and integration are required. The

output of this research is the classification of SNFs that have 'Increased Risk' of rehospitalisation taking into account a wide range of potentially contributory factors that are part of the 'SNF Ecosystem'. This makes the research of particular value given the multiple factors that are considered, however it is also challenging given the associated processing requirements of dealing with multiple datasets.

A mixed methods approach is appropriate given the requirements of the project including Inductive and Deductive steps. The initial focus is inductive and during this stage the learning aim is not predetermined. Exploratory analysis and unsupervised ML approaches can reveal patterns and trends to be further investigated and will constitute a significant part of the work in this project.

The second stage of this work is largely deductive and utilises selection and classification methods pointed to the defined the target class of the research. Deduction methods look to find explanation for the observed target outcome, in this case, the search for factors that will predict SNFs where there is Increased Risk of Readmission to hospital. To do this 2018 retrospective data is split into training and testing sets to build and test models for this prediction.

4.1 Data Challenges & Design Framework

4.1.1 Specific Data Challenges

- Domain Knowledge: time required to build knowledge on the subject area.
- Data availability: sourcing, finding, accessing the data needed for the project
- Pre-processed data including index or adjusted or standardised values which can be problematic for interpretation and relevance to the question and for how to proceed.
- Missing Values: for particular variables and also records not common across datasets.
- Consistency of naming conventions
- Multivariate data types, including numeric data, nominal values, date types, factors including those with high cardinality which can be difficult to transform.
- Volume of data and dimensionality issues: Eulers number 2.718 * n(factors) is the number of records to factors recommended.
- Data Scale Ranges: percentile ranges to absolute cost monetary values can lead to algorithms being more biased to the larger range.
- Normality Issues: skewed distributions with outliers for numerical data and highly imbalanced proportions for some categorical variables

4.1.2 Design Framework

Data Exploration and Feature Selection is very important for a number of reasons including stream-lining the modelling process and computational cost and avoiding the 'Curse of Dimensionality' which means as dimensions go up the sample space search required to find a pattern increases exponentially. This can lead to overfitting during training phases caused by too many noisy dimensions and in turn impacts generalisability, value, and cost of the model. Feature Selection and Feature Extraction are the two differing approaches utilised to reduce data to the most relevant and discriminative features to achieve the best results and both approaches will be used iteratively throughout.

a) Iterative approach: Start with a large amount of data and use statistical methods and plotting to make decisions on what to remove or retain. A manual and time and resource intensive stage to understand individual variables and relationships
b) Process two main datasets first separately before integrating and repeating the process
c) Utilise both Unsupervised and Supervised ML Methods

Table 2 – Framework Design

Utilise Exploratory Statistical Methods	Dimension Reduction Methods	Modelling Preparation and Implementation
• Inspect & Evaluate full dataset and run summary descriptive statistics • Review All Data for Missing Values, data issues • Identify important Variables • Clean and Process	• PCA • Boruta - Wrapper techniques are very effective as they evaluate all features to find the best set but are time	• Class Balancing Requirements • Scaling & Normalisation Requirements • Dummy Coding to transform categorical data to binary variables • Train & Test Sets (70:30) • Iterative • Parameter Tuning • Comparison of Models

This methods this design proposed at a high level fall into two categories, unsupervised and supervised methods and are described here.

4.2 Unsupervised Methods

Unsupervised ML Methods are used in datamining for 'Pattern Mining' and are ideal for exploratory analysis in which a linear relationship is sought but there is no predetermined target or dependent variable. The literature points to previous success in utilising unsupervised methods and therefore a number of these methods have been employed for this research and are described here.

a) Data Exploration

This work largely utilises summary descriptive statistics and visual plots to develop knowledge of the data and subject area. For continuous variables it is useful to examine individual variable parameters using visuals such as histograms and boxplots. Relationships can be explored via scatterplot and correlation matrices. Categorical variables in the dataset can offer good insight into the population under study and include boxplots and bar charts and conducted early and iteratively throughout the research can give early insight into which variables could need further investigation.

b) Correlation

Correlation is the most basic of unsupervised techniques and supports early identification of relationships that exist between variables. Creating a correlation matrix can serve two purposes. Firstly, it can provide an alert to potential multi collinearity issues which if not addressed could distort the results. It can also give early insight into potential patterns and groups.

c) Dimension Reduction

Feature Extraction in the form of Principle Component Analysis (PCA) is a is a statistical method used to find most of the variability in a dataset replacing the original dimensions with

a smaller number of component variables. It seeks to minimize the perpendicular distances from the data to the fitted model which is the Total Least Square value. This can help with issues of multicollinearity and also practically improve the manageability of a dataset however a challenge can be the interpretability of results as the original data is lost. PCA will be used in this case to support information gain without incurring large computational costs.

4.3 Supervised Methods

Supervised methods are associated with predictive models and require a target or apriori segments and will be the final step in this analysis where three different model types will be tested. Classification methods will be used for this prediction and models are selected for their robustness to multicollinearity.

a) Elastic Net

Elastic Net is a linear regression model which utilises regularisation parameters to improve the predictive power of the model without increasing bias. It builds upon and is more efficient than LASSO and Ridge regression by both reducing by reducing the impact of collinearity and also reducing dimensionality by shrinking some of the regression coefficients to zero. This model has the advantage of variable selection and regularization occurring simultaneously and which provides advantage for this research.

b) Support Vector Machine

SVM was selected as it can perform well on a limited number of examples and is robust against overfitting, especially in high-dimensional space as it utilises built in and embedded filtering. An SVM model is stable as small changes to the data do not affect the hyperplane. The polynomial SVM kernel is optimized towards learning non-linear patterns across a combination of features. Some of the disadvantages are that SVs do not scale well for larger datasets and can be difficult to tune in selecting the best kernel. The polynomial kernel is a more generalised version of the linear kernel and better for this data but can be less efficient.

c) Random Forest

Random Forest (RF) is an ensemble non-parametric approach and was selected because it is robust to handling missing values, no scaling is required and it has high performance and accuracy. One challenge of RF is the interpretability of results and the computational resources required.

A common problem for data scientists is how to manage the trade-off between model prediction accuracy and the volume of data required to achieve it. This is a particular challenge of this research and the design approach recognises the need for several iterations of data-processing to manage the high dimensionality, feature selection and integration requirements. Given this upfront workload, the design also seeks to minimise further additional transformation requirements associated with model building by selecting model that are more robust to issues of sparsity, normality and scaling requirements. Assumptions have been made that the built in Algorithms in R will be utilised.

5 Implementation

5.1 R Scripts

Excel was used in early stages of work for pre processing of large csv files which included data not relevant to the research question, outside of this the stage R was used for the full

project given it is designed to be a comprehensive statistical analysis package and has a wide range of libraries available for modelling and graphing that are relevant to this project. By using R for the full project a script can be run end to end. R is also powerful and faster than other languages which is important give the requirements of this work. High computational processes that affected performance during the implementation were using the Boruta selection package and RF.

5.2 Transforming Data

Processing and feature transformation requirements were heavy for this project with initial work required to ensure data consistency and enable integration. Adjustments made to standardise variables, differences in naming, missing values all provided challenges in generalising the variables for modelling. Processes were applied iteratively during exploration including dimension reduction, selection and engineering. Redundant, collinear and sparse features were removed

a) Processing & Cleaning Outputs

Following processing the SNF & County Health Rankings were integrated into the SNF_Base file for preparation for modelling. Missing values were identified as completely at random and subsequently imputed using multivariate imputation by chained equations (MICE) and a random forest approach. Data was encoded appropriately for ML Once the SNF_Base file had been cleaned and coded it was ready for the next stage of analysis and modelling

b) Dimension Reduction Outputs

Dimensionality reduction: was conducted at various stages of this analysis given the number of features, complexity and multifactorial nature of the problem. To retain variance without incurring the curse of dimensionality Health County Rankings, SNF Resource Utilisation Groups (RUGs) and SNF Survey of Health Deficiencies were each transformed using PCA methods. There were mixed results in introducing this stage. It supported additional exploitation but added further time and complexity to the project. With hindsight the analysis conducted on 'Deficiencies' did not provide return on investment, and there still remain questions as to whether these data could be Important to the question. In contrast PCA was valuable in reducing the County Health Rankings from over 30 variables to 5 components, enabling the full variability of the data to be brought into the analysis without the cost of a very large dimension increase.

c) Feature Engineering Outputs

Increased Risk, Facility Occupancy, Tenure, Number of SNFs in the Chain are some examples of features that were engineered to support the analysis. Counts were normalised to percentages and rates to allow for generalisability. The features engineered overall were not important to the question, outside of the target and 'Tenure'.

d) Boruta Selection Outputs

Boruta was applied to select the most important variables in the dataset. It is a selection wrapper technique based on RF and comes at a high computational as it checks each feature for importance. 36 features were examined in Boruta which performed 20 iterations in 8.395 mins and just two attributes were deemed unimportant. The feature importance graph in figure 18 was used to inform decisions on which features would be modelled in the next stage.

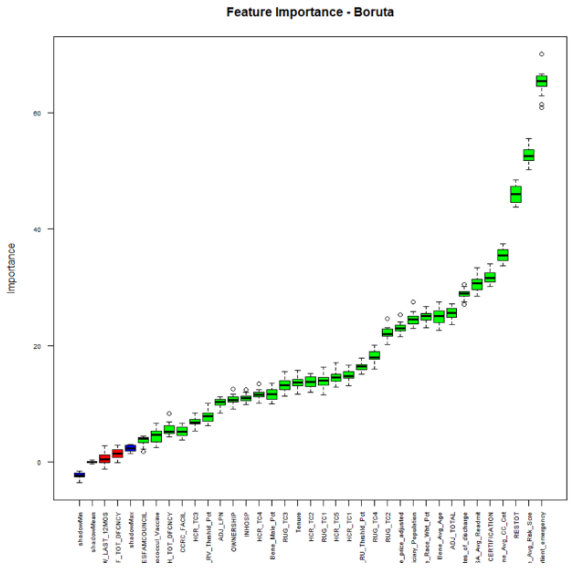


Figure 11 - Feature Importance - Boruta

Dataset for Modelling

Once feature selection was completed in Boruta the Base_SNF dataset used to build the first models in Elastic Net, SVM and Random Forest comprised of 14,607 records and 21 features. The positive class of 'Increased Risk' is the dominant class with 56% of SNFs classifying as True. It was decided based on the principles of parsimony and pragmatism and given already heavy processing requirements to retain this relatively balanced ratio. The dataset was split into training and testing 70:30.

5.3 Models developed

Models were built using the same data inputs and each was run though first using default settings and then tuning of parameters as required.

5.3.1 Implementing Elastic Net Regression

Elastic Net logistic regression was implemented as a regularised parametric linear model and was scripted using the Caret packages Glmnet and Gmodels in "R". To select the two tuning parameters (α and λ) required for the elastic net, a custom train control was created of fivefold cross-validation as proposed by (Zou and Hastie, 2005) where ten values of α ranging from 0 to 1 were tested and for each α value, 100 λ values were examined. Training and on 'Increased Risk' was conducted and an initial prediction to determine accuracy and sensitivity as a baseline developed to be followed up with testing against a random samples over 50 iterations. Caret automatically tests a range of possible alpha and lambda values, then selects the best values for lambda and alpha, resulting in the best model. Despite parameter tuning the performance metrics did not change as would have been expected. Several alterations were made to change the code and also to check the encoding of data types was correct for this model - as a linear model it requires numeric and binary inputs, however resolution could not be identified. It was decided to not progress further with this model as it is not suited to them data and additional time that would need to be given to find a solution which could involve additional processing or scaling of the data.

5.3.2 Implementing Support Vector Machine

SVM was implemented with a polynomial kernel as a non-linear parametric approach using the Parallelsvm() library and was tuned using an adaptive random search approach in a fivefold cross validation, cross-validation and resampling support finding the best cost parameter in SVM. An initial prediction to determine accuracy and sensitivity was generated as a baseline and them tested against a random sample over 50 iterations. Testing of models was conducted in a stepwise approach involving initially including all the most important features from the Boruta selection process to model and subsequently reducing the number of factors and assessing the impact on the model performance metrics. Figure 12 represents several runs of the SVM model that were conducted to include different factors and will be evaluated in the next section.

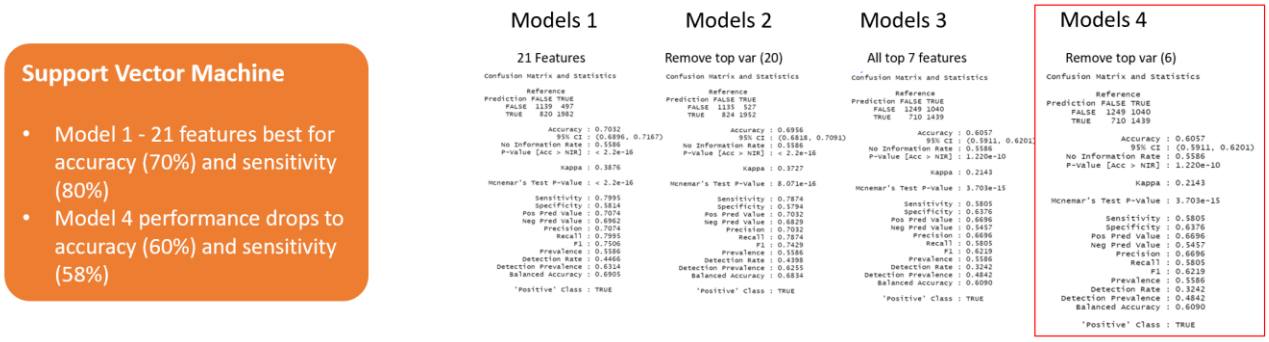


Figure 12 - SVM Analysis

5.3.3 Implementing Random Forest

The random forest model is a bootstrap aggregation approach that was built using 500 trees. Default parameters were initially used to baseline the model followed by fine tuning of parameters utilised a for loop to identify the right mtry for the model, which is the number of variables that are tested at each decision tree split. Models were built to model a range of factors in line with the approach for SVM and for comparison purposes. These models will be evaluated in the next section. The main challenge of implementing the RF model was the high computational cost, particularly when modelling 21 features. A large amount of working memory and time was needed to run the models and on occasion R was not able to complete.

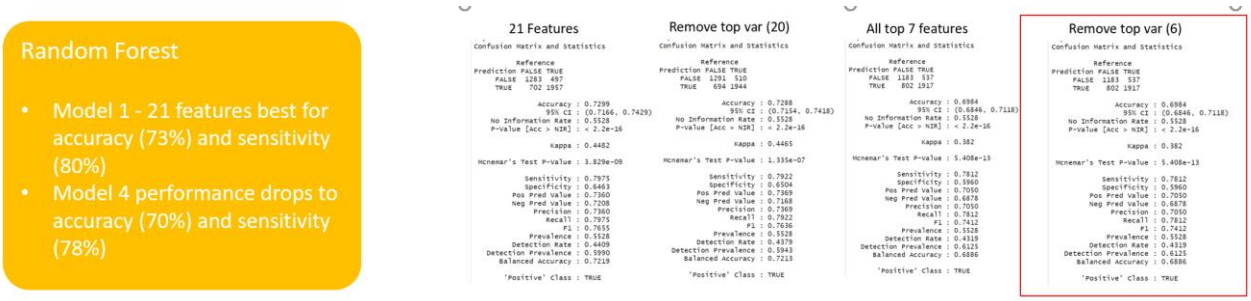


Figure 132 - Random Forest Analysis

6 Evaluation

The model building processes like all other stages of this research was iterative and decisions were made to remove through incremental testing of performance across a range of features. The first models were built using 21 features to understand which would contribute most,

thereafter features were removed and performance model metrics assessed for the size of gain that each feature contributed. It was decided to eliminate 'Beneficiary Risk Score' and retain 'Number of Chronic Conditions' as both measure similar factors.

6.1 SVM

SVM performed best with all 21 features retained with accuracy (70%) and sensitivity (80%), however the number of dimensions is high and the information gain for the lowest contributing variables is low. Model 4 has 6 features included but performance drops to accuracy (60%) and sensitivity (58%). The Kappa statistic is also low (McHugh, 2012)

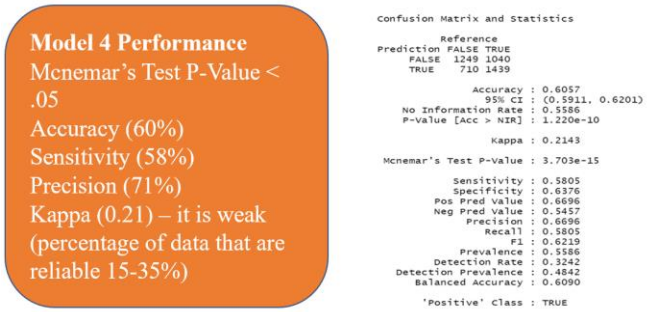


Figure 14 - SVM Performance

6.2 Random Forest

Although the first Model run in RF with 21 features had the best overall performance with Accuracy (73%) and Sensitivity (80%) a feature set of this size has high computational cost, particularly in RF and therefore is not extensible. In model 4 with just 6 features included, accuracy and sensitivity did not reduce significantly, therefore this model has been selected as best model for the problem.

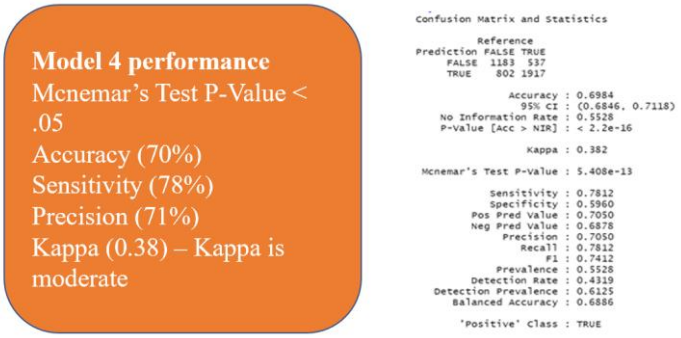


Figure 15 - Random Forest Performance

In total 6 were included with the following feature importance based on mean decrease in accuracy. These factors represent attributes of the SNF facility & performance, the patient health profile and the prevailing hospitalisation utilisation for the area where the SNF is located, Figure 17

- **Outpatient Emergency Visits:** percentage rate of patients visiting Outpatient Emergency
- **Certification Type:** Is the SNF a Medicare Certified facility, just 2.2% of SNFs are not and therefore are Medicaid

- **RESTOT** - Average total number of Residents (Beneficiaries) in the SNF which is an indicator of the size of the Nursing Home
- Average number of **Chronic Conditions** (CC) of Beneficiaries in the SNF
- The **Average HSA Readmittance** - rate for the HSA area where the SNF is located
- **Average hospital utilisation** - (rates of discharge) for the HSA area of the SNF

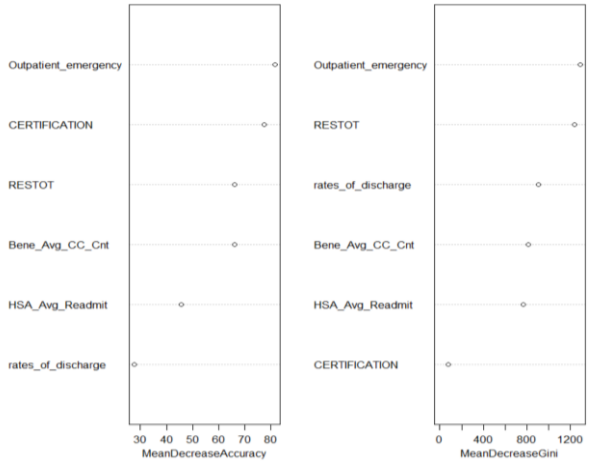


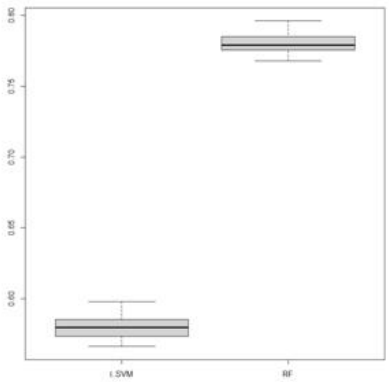
Figure 17- Feature Performance Random Forest

6.3 Comparison of Models

The key performance criteria for modelling Increased Risk is Sensitivity and Accuracy and figure 17 presents a plot comparing each models performance for this metric. Other important criteria are parsimony to create the simplest model with the least processing. Cohens Kappa value should also be considered, lower values suggest that data may be unreliable.

Evaluation Summary

- The training set is relatively balanced with True Increased Risk class dominant at 56%.
- A lazy or naïve model could achieve 56% by predicting all 'Increased Risk'
- The Elastic Net Model was eliminated as it was not the best model for the dataset
- An overall classification accuracy of 60% for the SVM model was achieved.
- An overall classification accuracy of 70% for the Random Forest model was achieved. Given the complexity and multifactorial nature of identifying 'Increased Risk' of Readmission to hospital from a SNF this is a good result.



```
> summary(model_sensitivities)
     ï..SVM             RF
 Min.   :0.5663   Min.   :0.7678
 1st Qu.:0.5740   1st Qu.:0.7754
 Median :0.5797   Median :0.7790
 Mean   :0.5799   Mean   :0.7795
 3rd Qu.:0.5851   3rd Qu.:0.7849
 Max.   :0.5978   Max.   :0.7961
```

Figure 18 - Summary of Sensitivities

- Sensitivity indicates if the model is able to correctly classify SNFs who have Increased Risk of Hospitalisation therefore the RF and SVM models were plotted for the average of 50 runs for comparison.
- Random Forest with 78% sensitivity is robust in identifying SNFs with Increased Risk of Readmission to hospital
- The SVM model is less sensitive, identifying just under two thirds of SNFs with increased risk at 58%

6.4 Discussion

An overall classification accuracy of 70% for the Random Forest model was achieved. Given the complexity and multifactorial nature of identifying 'Increased Risk' of Readmission to hospital from a SNF this is a good result. The model was tuned to optimise sensitivity in identifying all instances of increased risk in SNFs so that this can be further actioned upon. The sensitivity for this model is robust at 78%

This study sought to model factors relevant to the problem that previously have been studied separately. This supports better understanding of the types of features that are most important for predicting hospital readmissions. Robust modelling was selected and stepwise analysis to accommodate the analysis of a large number of features. The most important variables in the model are related to patient (beneficiary) and nursing facility characteristics as has been previously proposed in the literature but also and importantly, the characteristics of the small area (HSA) that the nursing home sits within, meaning for areas where hospital utilisation and readmission rates are high, SNFs in the area will also have higher readmission rate – 'A culture of hospitalisation'. This points to an opportunity for Federal Health Administrators to examine utilisation policies and practices across Geographies.

Each of the 6 features selected for the final model should be examined individually for meaning. Outpatient Emergency Visits was identified as the highest predictor of Increased Risk of Readmission and perhaps this is not unexpected as readmission and emergency visits are both examples of decisions or tendency to engage in hospital intervention. This feature was previously examined for multicollinearity as part of this work and it was not identified as an issue. Certification type was also identified as important and with just 2.2% of SNFs not being certified as Medicare Facilities this suggests that Medicaid only facilities are less likely to have increased risk. Medicaid is accessible to beneficiaries who are under the age of 65 who meet means test criteria. The selection of this feature may be a function of the difference in beneficiary profiles in the different SNF certification types. The average number of residents in nursing home was also selected as important and is a proxy for the size of the facility, the model suggests that higher risk may be associated with larger facilities? It has been discussed that SNFs that have patients with higher average numbers of chronic conditions also have Increased Risk and this aligns with the literature. The final features in the model that relate to HSA patterns have already been discussed and it is confirmation that a range of factors need to be considered together when examining predictive factors.

There have been several challenges with the research though out all stages of the research framework. Data access, selection, exploration and processing were all resource intensive. The ambition of the work to include factors across the entire SNF ecosystem with reference to literature meant that dimensionality was a challenge. Cohen's Kappa value is low for the

SVM model and moderate for RF which poses a question for the reliability of the data modelled. This should be a focus of future work and could likely be improved with more time given to normalising the data and examination of outliers. Some of the important features were adjusted or standardised values which also caused some problems. Suggestions for addressing these issues and opportunities are presented in the next section.

7 Conclusion and Future Work

This project sought to investigate two keys questions:
1)What makes one Skilled Nursing Facility more likely than another for patients to be readmitted to hospital?
2)What are the types of socio-economic and demographic attributes of the geographic location of a nursing home that might influence the likelihood of being readmitted to hospital?

To address the first part SNF rehospitalisation rates were binary classified for Increased Risk of readmission to hospital. Factors from the SNF ecosystem that could be predictive were modelled to test this classification using SVM and Random Forest Classifiers and the RF model with 6 features achieved 70% accuracy and 78% sensitivity, meaning 78% of positive cases of Increased Risk of readmission can be identified based on the factors included. The factors are in agreement with previous research that Patient and SNF characteristics are predictive of increased risk. The decision to include factors representing other broader aspects of the SNF ecosystem was validated as the HSA factors relating to geographical patterns of healthcare were identified important.

The second aim of the research was progressed by utilising data available on the social behavioural determinants of health accessible in the County Health Rankings. The main principle components were extracted and included in the model dimension space, however these components were not among the top contributing features. Although no predictive socio economic factors were identified the approach of focussing on a wider range of features did have some success with the previously described geographic patterns of healthcare proving important. It is believed that the approach is correct and with further focus could produce insights.

Hospital Readmissions is a significant and costly problem in US Healthcare and there has been much focus on how to identify and address causative factors, including those that relate to SNFs. This work is valuable as it presents an approach to model for multiple factors in the SNF eco-system. Previous research has highlighted that Readmissions is a multi-factorial problem that should be evaluated as a whole. Although challenges have been encountered during the work, it may serve as a template for other future work.

Opportunities for follow-up work are to refine the feature selection stages of the process through experimentation and iteration and to build a reduced and validated set of variables for future model building. To build and test other models which are perhaps less robust but have increased predictive power on the reduced set of variables that are shown to be important from this work. Another opportunity is to catalogue, shortlist and describe methods for identifying and integrating publicly accessible data sources that can be utilized to streamline multifactorial analysis of the problem given the challenge of navigating the mountain of related datasets and sites. Finally; Social, Behavioral Determinants of Health factors were not identified as important in this model. Future work should be conducted on

the importance of this factor but Readmissions may not be the outcome of focus, given other factors seem more directly relevant to understand this.

References

A. Field, J. Miles, and Z. Field, Discovering statistics using R. SAGE PublicationsSage CA: Los Angeles, CA, 2012Billings, J., Dixon, J., Mijanovich, T., & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. BMJ (Clinical research ed.), 333(7563), 327.

Corrigan, J. M., & Martin, J. B. (1992). Identification of factors associated with hospital readmission and development of a predictive model. Health services research, 27(1), 81–101.

County Health Rankings, Last accessed https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation (last accessed 10/08/2021).

Department of Health, "Principal Causes of Death: Numbers and Age Standardised Death Rates per 100,000 Population," 2018. [Online]. Available: https://health.gov.ie/publications-research/statistics/statistics-by-topic/causes-of-death/. (last accessed 31/05/2021)

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 222. https://doi.org/10.1186/2193-1801-2-222

Flanagan J, Boltz M, Ji M. A Predictive Model of Intrinsic Factors Associated with Long-Stay Nursing Home Care After Hospitalization. Clinical Nursing Research. 2021;30(5):654-661. doi:

Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK. Prediction of 30-Day All-Cause Readmissions in Patients

Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiol. 2017 Feb 1;2(2):204-209. doi: 10.1001/jamacardio.2016.3956. PMID: 27784047.

Gaugler, J.E., Duval, S., Anderson, K.A. et al. Predicting nursing home admission in the U.S: a meta-analysis. BMC Geriatr 7, 13 (2007). https://doi.org/10.1186/1471-2318-7-13

H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 67, no. 2, pp. 301–320, Apr. 2005.

I. T. Jolliffe, "Discarding Variables in a Principal Component Analysis. I: Artificial Data," Appl. Stat., vol. 21, no. 2, p. 160, 1972.

J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics. 2015.

Jacobsen, J., Schnelle, J. F., Saraf, A. A., Long, E. A., Vasilevskis, E. E., Kripalani, S., & Simmons, S. F. (2017). Preventability of Hospital Readmissions From Skilled Nursing Facilities: A Consumer Perspective. The Gerontologist, 57(6), 1123–1132. https://doi.org/10.1093/geront/gnw132.

Kansagara D, Englander H, Salanitro A, et al. Risk Prediction Models for Hospital Readmission: A Systematic Review. JAMA. 2011;306(15):1688–1698. doi:10.1001/jama.2011.1515.

Lepore MJ, Shield RR, Looze J, Tyler D, Mor V, Miller SC. Medicare and Medicaid Reimbursement Rates for Nursing Homes Motivate Select Culture Change Practices But Not Comprehensive Culture Change. J Aging Soc Policy. 2015;27(3):215-31. doi: 10.1080/08959420.2015.1022102. PMID: 25941947; PMCID: PMC4714704.

M. B. Kursa and W. R. Rudnicki, " Feature Selection with the Boruta Package ," J. Stat. Softw., 2015.

M. Kuhn, “Futility Analysis in the Cross-Validation of Machine Learning Models,” arXiv Prepr., May 2014.

Mairead M Bartley, Parvez A Rahman, Curtis B Storlie, Paul Y Takahashi, Anupam Chandra, Associations of Skilled Nursing Facility Quality Ratings With 30- Day Rehospitalizations and Emergency Department Visits, Annals of Long-term Care. 2020 March ; 28(1): e11–e17.

McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282.

Medicare.gov, https://www.medicare.gov/ (last accessed 10/08/2021)

Nagasako, Elna MD, PhD, MPH; Waterman, Brian MPH; Reidhead, Mathew MA; Lian, Min MD, PhD; Gehlert, Sarah PhD, MA, MSW Measuring Subcounty Differences in Population Health Using Hospital and Census-Derived Data Sets: The Missouri ZIP Health Rankings Project, Journal of Public Health Management and Practice: July/August 2018 - Volume 24 - Issue 4 - p 340-349 doi: 10.1097/PHH.0000000000000578.

Neuman MD, Wirtalla C, Werner RM. Association Between Skilled Nursing Facility Quality Indicators and Hospital Readmissions. JAMA. 2014;312(15):1542–1551. doi:10.1001/jama.2014.13513.

Nitesh V. Chawla, K. W. Bowyer, and L. O. Hall, “SMOTE: Synthetic Minority Over-sampling Technique,” J. Artif. Intell. Res., vol. 2009, no. Sept. 28, pp. 321–357, 2006.

Nuti SV, Doupe P, Villanueva B, Scarpa J, Bruzelius E, Baum A. Characterizing Subgroups of High-Need, High-Cost Patients Based on Their Clinical Conditions: a Machine Learning-Based Analysis of Medicaid Claims Data. J Gen Intern Med. 2019 Aug;34(8):1406-1408. doi: 10.1007/s11606-019-04941-8. PMID: 30887432; PMCID: PMC6667598.

P. Cortez and M. J. Embrechts, “Using sensitivity analysis and1visualization techniques to open black box data mining models,” Inf. Sci. (Ny)., 2013.

Richard L. Fuller, Norbert I. Goldfield, John S. Hughes, and Elizabeth C. McCullough, Nursing Home Compare Star Rankings and the Variation in Potentially Preventable Emergency Department Visits and Hospital Admission, Population Health Management, Apr 2019.144-152.

Sai T. Moturu, William G. Johnson, and Huan Liu, Predictive risk modelling for forecasting high-cost patients: A real-world application using Medicaid data, International Journal of Biomedical Engineering and Technology 2010 3:1-2, 114-132.

Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, Weiner J Including Social and Behavioral Determinants in Predictive Models: Trends, Challenges, and Opportunities, JMIR Med Inform 2020;8(9):e18084.

The Centre for Medicare and Medicaid Services, https://data.cms.gov/provider-data/

The Dartmouth Institute for Health Policy and Clinical Practice, Robert Wood Johnson Foundation, The Revolving Door: A Report on U.S. Hospital Readmissions (2013)

Witten, I.H. and Frank, E. (2005) Data mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann Publisher, Burlington.

Wrathall JA, Belnap T. Reducing Healthcare Costs Through Patient Targeting: Risk Adjustment Modeling to Predict Patients Remaining High-Cost. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2017;5(2):4.

Yoo JW, Jabeen S, Bajwa T Jr, Kim SJ, Leander D, Hasan L, Punke J, Soryal S, Khan A. Hospital readmission of skilled nursing facility residents: a systematic review. Res Gerontol Nurs. 2015 May-Jun;8(3):148-56. doi: 10.3928/19404921-20150129-01. Epub 2015 Feb 24. PMID: 25710452.