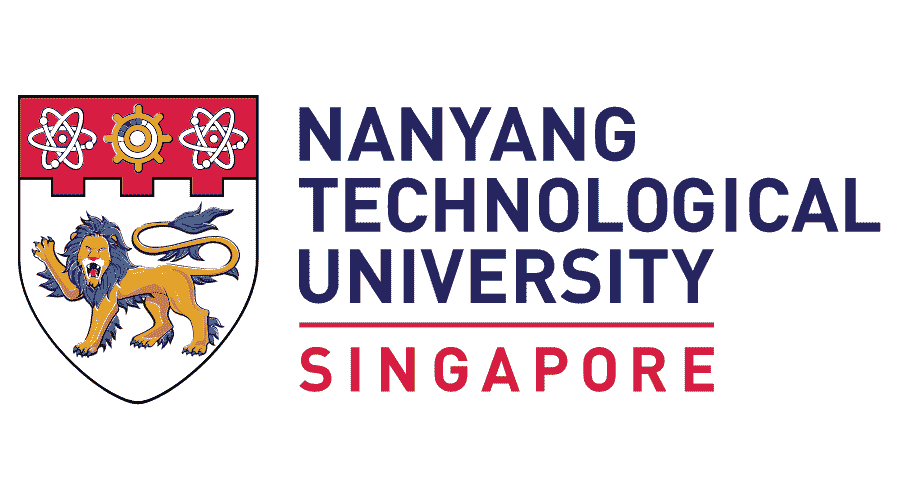
****

**BC2407 Analytics II:** Advanced Predictive Techniques

Academic Year 21/22 Semester 2

**Seminar Group 1 Team 1**

**Project Title:** Predicting the likelihood of readmission within 30 days

for diabetic patients for hospitals in Singapore

**Team Members:**

|  |  |
| --- | --- |
| **Name** | **Matriculation No.** |
| Faith Lee Kai Ling | U2022307C |
| Lau Shing Hung | U2021155C |
| Lek Jie Kai | U2021229C |
| Lim Yan Di | U2010984H |
| Muhammad Ihsan Bin Mohammad Azmi | U2021160B |

**Table of Contents**

[**1 Executive Summary**](#_slxuh8x4fr91)

[**2 Business Problem Analysis**](#_ex9efv1tyk9c) **1**

[2.1 Introduction](#_1fob9te) 1

[2.2 Current Operations](#_3znysh7) 2

[2.3 Business Problem Statement](#_2et92p0) 3

[2.4 Business Outcome Measures & Targets](#_tyjcwt) 4

[**3 Analytics Solution**](#_ngfajjr2zhkp) **6**

[3.1 Analytics Performance Measures and Targets](#_1t3h5sf) 6

[3.2 Initial Exploratory Analysis](#_obih7ilg0yjb) 7

[3.3 Data Cleaning](#_y3gzwjypv0pa) 8

[3.4 Data Preparation](#_4e05k938hjag) 9

[3.5 Method 1: Logistic Regression](#_wo63jt3j8g3i) 10

[3.6 Method 2: Neural Network](#_26in1rg) 11

[3.7 Method 3: Random Forest](#_m51b5gfsj1on) 13

[**4 Evaluation of the Models**](#_xsa4pme81v93) **15**

[4.1 Strengths of analysis models](#_1ksv4uv) 15

[4.2 Limitations of analysis models](#_nit9gvsux0gk) 15

[4.3 Comparison of performance of the models](#_z337ya) 17

[4.4 Recommendations](#_egecnys4e501) 18

[**5 Conclusion**](#_as0abuuquwe8) **20**

[**6 References**](#_x1g3i0iccyic) **21**

[**7 Appendix**](#_4i7ojhp) **24**

[7.1 Data Dictionary](#_2xcytpi) 24

[7.2 Exploratory Analysis](#_kaj9idibg8ds) 26

[7.3 Logistic Regression](#_u9e7mtbnf5pe) 27

[7.4 Neural Network](#_by2b89sc1lbk) 29

[7.5 Random Forest](#_22gvp9i407gb) 32

# 

# 1 Executive Summary

With a rapidly growing ageing population, increasing life expectancy, and greater chronic illness burden, Singapore’s healthcare expenditure is set to hit $49 billion in 2029. It is therefore important that unnecessary healthcare costs are identified and minimised so as to reduce the strain on our finite healthcare resources. Since the largest percentage of hospital costs comes from the cost of inpatient care, one area with huge potential cost savings is avoidable hospital readmissions. In particular, readmission of diabetic patients is pertinent in Singapore, since 1 in 3 Singaporeans are at risk of developing diabetes, and in general the readmission rate for diabetic patients is higher than for non-diabetic patients.

As such, our project aims to predict the likelihood of hospital readmission within 30 days for diabetic patients in Singapore, in order for hospitals to put in place preemptive measures to reduce their readmission rate. The current predictive models available in Singapore do not focus on diabetic patients, and hence we decided to make this demographic the focus of our project.

We used a dataset from the United States that contained 101,766 clinical database records of diabetic patients to demonstrate and generate the models. We cleaned and balanced the data in order to improve our analysis. To find the best model, we compared three methods to find the one that was the most accurate at predicting the likelihood of readmission.

1. Logistic Regression
2. Neural Network
3. Random Forest

Significant variables identified from the Logistic Regression Model were used for generating the Neural Network model. We used the false negative rate as a key indicator of accuracy for all our models.

From these models, we concluded that the Random Forest model was the most optimal due to its lowest false negative rate as compared to the other models and its useful in-house variable importance function, among other considerations. The optimal random forest model was identified through trial and error of the model parameters, where the optimal model has the RSF size of 8 and default number of trees of 500. Using random forest’s variable importance function, we found that the top 3 significant variables identified were number of inpatient visits, time in hospital, and age. The hospital can consider directing efforts into studying the relationship between these variables and readmission status.

In conclusion, hospitals in Singapore can use the Random Forest model to identify diabetic patients that are likely to be readmitted, for early intervention and prevention of avoidable readmissions.

# 

# 2 Business Problem Analysis

## 2.1 Introduction

In 2019, Singapore’s healthcare spending for both private and public sectors was estimated to be $19.6 billion, about 4.1% of the country’s GDP that year (ITA, n.d). With a rapidly ageing population, increasing life expectancy, and increasing chronic illness burden, this amount is set to continue to rise in the future, with experts expecting it to hit $49 billion in 2029 (ITA, n.d). It is therefore important that unnecessary costs are identified and minimised as much as possible so as to reduce the strain on our finite healthcare resources. As with most developed countries, the largest percentage of this comes from the cost of inpatient care (Goetghebeur et. al., 2003). Therefore, one area in healthcare with huge potential for reducing healthcare expenditure in Singapore is avoidable hospital readmissions.

Hospital readmissions are defined as the event where a patient is readmitted to the same department within a certain period of time for the same disease. The most commonly used time frame is 30 days. Readmissions within 30 days are more likely to be linked to the quality of care received at the hospital and how well discharges were coordinated. On the other hand, admissions more than 30 days after their initial admission are harder to trace back to the hospital, and are more likely to be related to factors that are beyond the hospital’s control like outpatient care, or personal choices and behaviour. It has also been found that readmissions are most preventable in the immediate period after discharge (Graham et. al., 2015).

Currently, Singapore’s all-cause 30 day readmission rate stands at about 15%–meaning about 15% of all patients that are discharged from a hospital in Singapore are readmitted again within 30 days (Khew, 2017). Frequent admitters to hospitals are high-cost patients and they contribute to bed shortage (Low et. al., 2018). By reducing preventable readmissions, Singapore’s healthcare expenditure can be reduced significantly and the limited healthcare resources in Singapore will not be wasted. Hospital readmission rates are also often used as a benchmark of quality for healthcare systems. The rate of hospital readmission is indicative of the quality of care the hospital provides to patients – higher rates imply that the hospital is not giving adequate care to patients. Reducing avoidable readmissions means that patients will be able to receive better quality care, while also reducing the financial pressure and psychological stress on patients (Shang et. al., 2021).

While there has been research done in the area of predicting hospital readmissions in Singapore, most predictive models tend to be for all patients in general (Low et. al., 2016, NUS, 2017). Thus, we decided to narrow down our scope and focus on a specific area that has not been explored yet–diabetic patients in Singapore.

Currently, over 400,000 Singaporeans have diabetes (Ow Yong et. al., 2021). The lifetime risk of developing diabetes is one in three, and the number of those with diabetes is projected to surpass one million by 2050 (Ow Yong et. al., 2021). An estimated 430,000 (14%) Singaporeans aged 18 to 69 years are also diagnosed with pre-diabetes, where their normal blood sugar levels are higher than normal but not high enough to be diagnosed as diabetes (MOH, 2017). This highlights the significant problem regarding diabetes in Singapore and we can expect this problem to worsen in the future.

A vast majority of diabetic patients require repeated hospitalisations due to poor disease control (Shang et. al., 2021). To make things worse, readmissions due to non-diabetes related causes have also been found to increase for patients that have diabetes. A journal published by the University of Michigan in the field of Clinical Diabetes and Endocrinology found that the 30-day readmission rate for hospitalised patients with diabetes is estimated to be 14.4% – 22.7%, much higher than the rate for all hospitalised patients in the United States (8.5% – 13.5%) (Ostling et. al., 2017).

From this we can see that it is very important that our healthcare system in Singapore is able to provide the necessary support for diabetic patients, especially when it comes to providing proper care for them and reducing their hospital readmission rates.

## 2.2 Current Operations

Being such a widespread problem, many organisations, both local and international, have developed measures to counter the large volume of readmissions for diabetes patients. Hospitals have attempted to use both quantitative and qualitative measures to tackle this issue:

Prediction Tools in Singapore:

In Singapore, a research team from the Department of Pharmacy at the National University of Singapore (NUS) Faculty of Science have come up with a web-based prediction tool to predict the risk of 15-day hospital readmission (NUS, 2017). The tool used patient data from Khoo Teck Puat Hospital and Singapore General Hospital to come up with a calculator that could identify high-risk patients and facilitate the administration of targeted interventions for these patients. The tool has an accuracy rate of about 65% (Health Xchange, n.d.), and researchers are currently working on improving the accuracy of the model while also collaborating with local hospitals to scale this up.

Other predictive models and tools have been developed by various teams in Singapore as well. Another team from NUS came up with the predictive tool GEMINI, that can be used to give the predicted chance of readmission as well as the risk factors for readmission for each patient (GEMINI Web Team, 2020). A model known as the Multiple Readmission Predictive Model was also developed by a team of researchers from Integrated Health Information Systems (IHiS), National Healthcare Group (NHG), National University Health System (NUHS) and Singapore Health Services (SingHealth), which predicts the readmission likelihood for patients in general (IHiS , n.d.).

However, all of these prediction tools are only meant for patients in general, not diabetic patients specifically. As such, the risk factors and variables considered are geared towards patients in general rather than variables specific to diabetic patients.

Model of Care (Patient Education):

The Health Management Unit (HMU) was started by Changi General Hospital (CGH) in 2010 to address gaps in diabetes care, focusing on patient education. The HMU model of care mainly consisted of a telehealth service supported by a Patient Relationship management (PRM) system. (Foo & Fock, 2013)

The system had tele-nurses that would periodically contact patients during stipulated hours, with predetermined scripts to facilitate patient education regarding diabetes, including topics such as dietary advice, lifestyle improvements, and hypoglycaemia management. Feedback from the patients called would also be requested to check that the patients understood the contents of the calls.

Calls were also used for collection and monitoring of clinical markers of patients, enabling early intervention should a problem be identified. These interventions would reduce the worsening of a patient’s issues, preventing unnecessary hospitalisation. Collected patient information would also be condensed into a short summary to be used by doctors during consultations.

Statin therapy (Lipid management):

Diabetes patients have a higher risk of developing cardiovascular diseases such as heart attack or failure, due to the higher formation of plaque, or buildup of cholesterol, in the blood vessels. In Singapore, close to 60 percent of those with Diabetes Mellitus die of Cardiovascular disease. Statins are hence commonly prescribed to diabetic patients to lower the body’s bad cholesterol, reducing the risk of cardiovascular conditions (Bragato, 2021). It was found that a prescription of a 90 day supply of statins was effective in decreasing the likelihood of 30-day hospital readmission (Chen et al., 2012).

However, public perception of statins in Singapore is negative, with misunderstandings surrounding its use. Common myths include the causing of muscle aches, liver damage, and worsening of diabetes. A survey conducted showed that 56% of participants believed that statin use was associated with higher risks of cancer (Lim et al., 2021). These misconceptions contribute to the fall in adherence to prescribed statins in the long term. Medical organisations in Singapore such as AVHC and NHC have hence started awareness efforts to counter the prevalent misinformation.

Overall, many of the current operations and measures we found were aimed at holistically improving quality of care or raising patient knowledge to reduce the likelihood of readmission. However, more can be done to apply analytics to predict and identify the diabetic patients with high risk of readmission. This would complement and enhance existing measures to better support and care for the community of diabetic patients in Singapore.

## 2.3 Business Problem Statement

Before we can begin our analytics process, it is important that we breakdown and properly understand the matter at hand and its implications. The following are the sub-problems that stem from diabetic patient readmissions:

Fall in quality of care:

Higher readmission rates inherently result in a higher volume of patients, and with the ongoing COVID-19 pandemic, Healthcare Workers (HCWs) end up being severely overworked. Hospitals are consistently working at full capacity, regularly having 100 percent of their beds occupied. Furthermore, in just the first half of 2021 about 1500 HCWs resigned, compared to the usual 2000 annually pre-covid (CNA, 2021). This strain in manpower combined with the larger number of patients would have a negative impact on the quality of care, as HCWs are stretched thin and have neither the time nor energy to properly manage their patients. Hence, having measures to reduce readmissions would be beneficial.

On the flipside, we must recognise that additional manpower is needed to implement new procedures. Haphazardly enacting operations could result in HCWs having to do unnecessary work, which would exacerbate the problem of overworking them.

Suboptimal use of resources:

Hospital readmissions are emblematic of the larger problem that is resource wastage. It has been found that accommodating for readmissions costs more than giving a patient higher quality care during the initial visit. Whether it is spending more time on patient education or providing larger amounts of recovery materials on the first discharge, a hospital would do well to allocate more resources the first time a patient is present, to prevent higher unnecessary costs later on.

However, it is important that resources are optimally allocated. Ideally, to fully minimise costs, we would want to apply readmission counter-measures exclusively to those who are at risk of coming back within 30-days.

Hence, by applying analytics appropriately, current measures could be better optimised. The following are the two overarching questions that we should aim to address and solve in our approach:

1. How can we **predict whether a patient is likely to be readmitted** based on their profile and current clinical markers?
2. What techniques can be used to **determine the significant variables** that heavily affect the chance of a patient’s readmission?

Our project aims to predict the likelihood of hospital readmission of diabetic patients in Singapore, in order for hospitals to put preemptive measures in place to reduce the risk of readmission, which in turn lowers the costs incurred.

It has been widely suggested that the best way to reduce preventable hospital readmissions, is for hospitals to first predict and identify the patients with the highest likelihood to be readmitted, then to apply intervention programmes targeting the patients needs (Low et. al., 2016).

As such, our group plans to use analytics to predict how susceptible a diabetic patient is to readmission, based on their patient profile. We aim to extract the key variables that affect the likelihood of readmission, and build a model that accurately identifies patients that are at high risk of readmission within 30 days.

With this, hospitals can identify the patients that are more likely to be readmitted early on and provide more targeted care for them, both during the hospitalisation phase and after-discharge. This will then decrease the likelihood of a particular patient needing to be readmitted. Examples include improving inpatient education, providing specialty care, and giving better discharge instructions, post discharge support and ensuring that they enter the appropriate post-discharge care facility (NUS, 2017).

## 2.4 Business Outcome Measures & Targets

We aim to identify diabetic patients who are at higher risk of readmission through the analysis of their given data and clinical markers. Hence, to test the strength and accuracy of our models, we can observe both test set performance, which we will be covering, and actual outcomes and improvements, which is not in the scope of this report.

Test set performance:

During the formation of our models, we will randomly split our data into train and test sets. By training our model using the train set and subsequently applying the model to the test set, we will be able to get an estimate of how well our model can make predictions. A metric that we would pay closer attention specifically for our project to would be the false negative rate (FNR) since it would indicate a high chance of misclassifying a patient who is actually high risk. By predicting that a patient will not be readmitted when in fact he will, will be detrimental as resources and care may not be catered to the patients who are likely to be readmitted.

Actual outcomes / improvements:

While the test set gives us a rough estimation, we would only really be able to determine the practicality of our models by applying them to current data and checking their performance. If the results are acceptable and the models are deemed accurate, we can then begin to selectively apply counter measures to patients, exclusively providing them to patients deemed by our model to be riskier. Over time, we could measure changes in the readmission rate and costs to the hospitals, to determine if our proposal is effective. The ideal outcome would be for both the readmission rate and hospital costs to fall after actual implementation.

Hence overall, applying our data to both the test and current data would allow us to measure the outcome of our project.

# 3 Analytics Solution

## 3.1 Analytics Performance Measures and Targets

The main goal of our project is to develop a reliable method to accurately determine the likelihood of readmission of diabetic patients within 30 days of their discharge, based on their previous data and clinical markers. To do this, our approach was to first develop multiple models, and then choose the model with the best accuracy metrics.

The predictive models that we will be utilising are

1. Logistic Regression
2. Neural Network
3. Random Forest

The variable *‘readmitted’* was used as the dependent variable Y, while the rest of the variables were used as independent variables.

As the Neural Network algorithm is unable to identify statistically significant variables on its own, we will use the statistically significant variables identified from Logistic Regression in our Neural Network model. We chose to use logistic regression’s significant variables because it is more similar to a neural network in how it uses numerical weights to derive a prediction, rather than the rule-based classification as in random forest. The performance of all 3 models will be assessed using a confusion matrix and by comparing the number of false negatives.

False Negative Rate

We will be focusing on the false negative rate as our model prediction accuracy metric. A false negative occurs when the model wrongly predicts that the patient will not be readmitted within 30 days, when they actually were.

False negatives carry greater consequences than false positives in this context. From the patient’s perspective, failure to predict a readmission correctly would cause them to be discharged without the sufficient care that they need that could lower their chance of a subsequent readmission. This could cause further health complications and worsen their condition. It also places a greater financial burden on the patient or their insurer. From the hospital’s perspective, frequent failure to predict a case of likely readmission will increase the number of preventable readmissions, putting a strain on hospital resources (Shang et. al., 2021). Furthermore, readmitted patients generally have a higher severity of illness (Wong et al., 2016) – which is more costly for hospitals and harmful to patient wellbeing.

On the other hand, a false positive is not as detrimental. While more resources would be directed to patients who are not as likely to be readmitted, this just means better quality care for more patients. The increase in short term spending will be offset by the long term cost savings of preventing avoidable readmissions (Upadhyay, 2019).

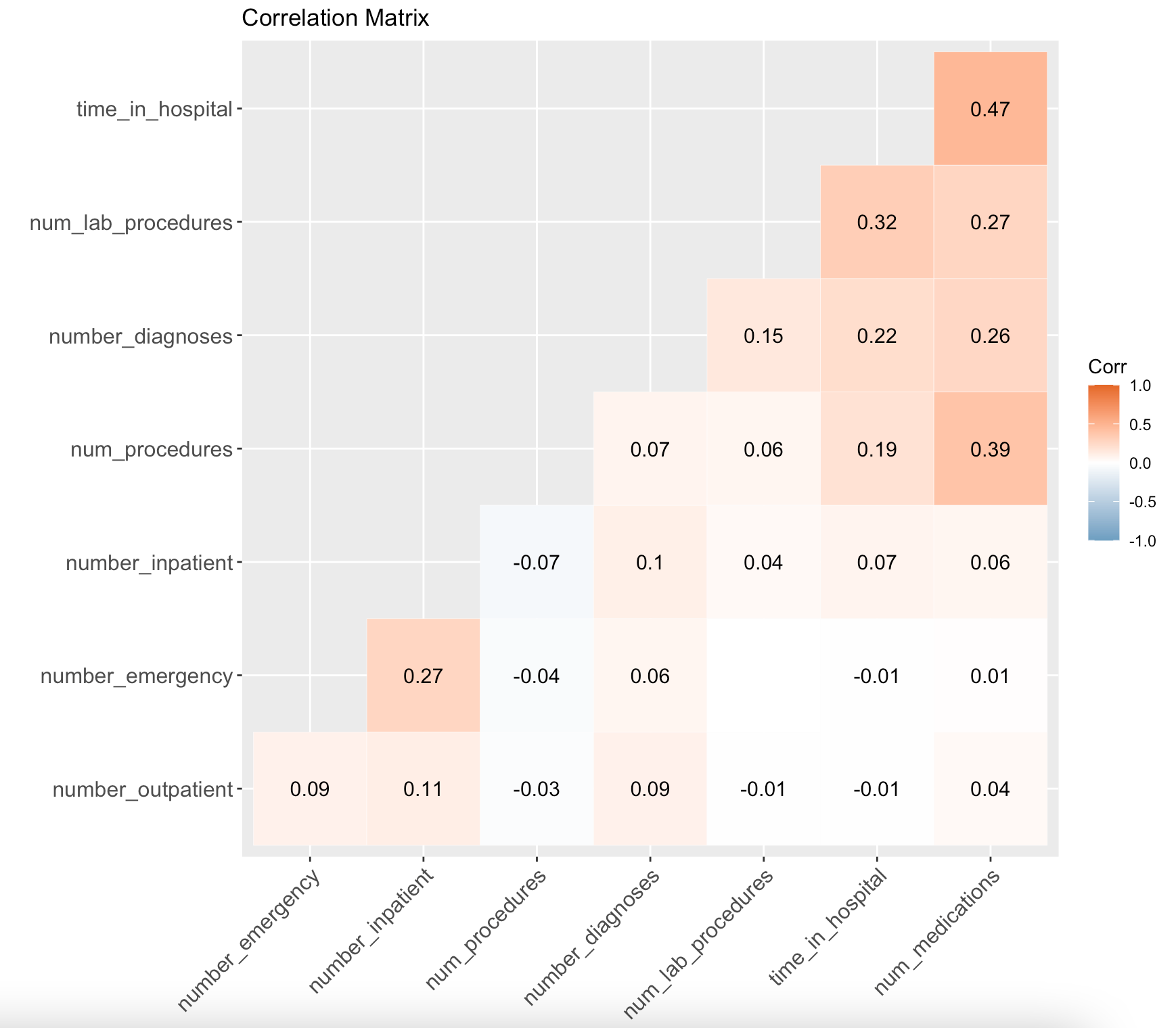
The overall accuracy rate is found by taking 100% minus the weighted average of the FNR and FPR. For a particular model, the overall accuracy may appear higher, but this could be due to a low FPR since the proportion of all the positive cases is greater than the negatives. Since we are more concerned with the FNR itself, we will choose the best model by giving the highest priority to a low FNR instead.

## 3.2 Initial Exploratory Analysis

Data Source

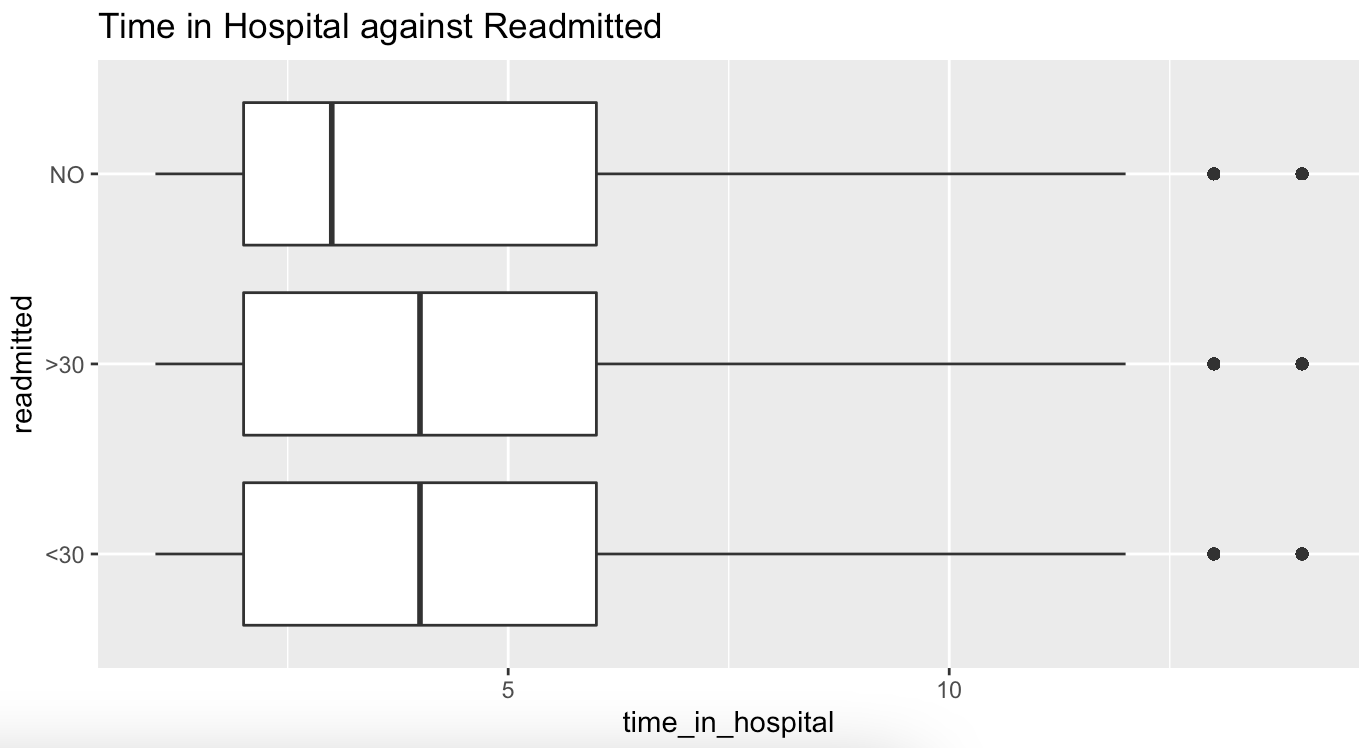
We used a dataset of 101,766 clinical database patient records for diabetic patients from across 130 hospitals in the United States to model our problem. The data was collected over 10 years (from 1999 to 2008). It includes 49 features of patient data, as well as the outcome of their hospital admission–readmitted within 30 days, readmitted after 30 days, or not readmitted.

Correlation between Numeric Independent Variables

****

*Fig 3.2.1: Correlation Matrix of Numeric Independent Variables*

We checked for collinearity between the numeric variables by looking at their correlation values. The highest absolute values of correlation were mostly between *‘time\_in\_hospital’* and the other numeric variables, which makes sense because the more time a patient spends in the hospital, the more the number of medications or procedures they are likely to have. However, since the correlations were not very high (meaning they were not close to |1|), we did not remove these variables.

****

*Fig 3.2.2: Boxplot of Number of Days Spent in the Hospital against Readmission Outcome*

We also observed that the median number of days that a patient spends in the hospital is higher for those who have been readmitted, both in less than and more than 30 days. This suggests that for patients that spend more time in the hospital, the likelihood of them being readmitted within 30 days increases. This trend could be because the number of days that a patient spends in the hospital could indicate the severity of a patient’s condition - more days means the patient’s condition is more severe. With a more severe condition, this could be why the patient was more likely to be readmitted.

After exploring the data, we proceeded to clean and prepare our data.

## 3.3 Data Cleaning

Selection of independent variables

From the initial 49 patient features in the dataset, we selected 20 of them for use as independent variables in this project. The variables that were not selected were largely due to reasons such as a huge amount of missing data.The independent and dependent variables used for our models can be found in Appendix 7.1.1.

Removing NA values in the dataset

The summary function was used to visually check which variables in our dataset contained missing data, NA values or any of its variations (eg. null, na, n.a etc.). Rows containing NA values were identified and removed. The entire row was removed instead of replacing the NA values with either the mean or mode because the number of NA values in the dataset was minimal.

Removing the rows where admission source ID and admission type ID are missing

The dataset contains two categorical variables called *admission\_source\_id* and *admission\_type\_id*. These came with coded values, 1-8 for admission source and 1-26 for admission type. These variables did not explicitly contain “NA” (or its alternatives) or blanks, these were instead coded into the levels, which meant that they were not removed in our initial NA removal. For *admission\_source\_id*, numbers 5, 6 and 8 meant “not available”, “null” and “not mapped” respectively. For *admission\_type\_id*, numbers 9, 15, 17, 20 and 21 meant “not available”, “not available”, “NULL”, “not mapped” and “unknown/invalid” respectively. As these values are the equivalent of NA values and do not contribute meaningfully to our model creation, we removed all the rows that contained these values in *admission\_source\_id*  and *admission\_type\_id.*

Remove rows that were outside specified range for continuous variables

To ensure that all the continuous variables had values that were logical (eg. time\_in\_hospital and number\_medications cannot be negative), we used a while loop for all the continuous variables to make sure the values fit into a range between 0 and the given maximum number in the provided data dictionary. All the rows that contained values that were not within the specified range were removed.

Cleaning the gender variable

The initial gender variable in the dataset consisted of 3 categories, “Male”, “Female” and “Unknown/Invalid”. There are only 3 patients that have a gender of “Unknown/Invalid”. Hence, these 3 rows were removed from the dataset.

## 3.4 Data Preparation

Changing categorical variables to factor type

The categorical variables in the data set - *'gender', 'age', 'change', 'A1Cresult', 'diabetesMed', 'insulin', 'glipizide', 'metformin', ‘glyburide’, ‘pioglitazone’, 'admission\_type\_id'* and *'readmitted'* - were converted to factor type using the factor() function as they were originally either string or numeric type.

Choosing the 5 medications to be included in our final dataset

As there are a total of 24 types of diabetic medications in our original dataset, for simplicity, only the top 5 most prescribed medications were identified and used in the models. This was done by using a for-loop to count the number of “No”s (which meant that the medication was not administered to the patient) for each medication, and then sorting the medications by the number of “No”s in descending order, and finally taking the 5 medications that had the least number of “No”s.

Reclassifying the categories of our *‘readmitted’* variable to a binary variable of 0 and 1

The dependent variable *‘readmitted’* initially had 3 levels, “<30”, “>30” and “NO”. We combined the outcomes “>30 days” and “NO” to represent the negative outcome “0”, and “<30 days” to be the positive outcome “1”. This is because our focus is on early readmissions within 30 days, since a readmission after 30 days is likely to be due to factors unrelated to the prior hospital stay.

Therefore, if the prediction for *‘readmitted’* is 1, it means that the patient was readmitted within 30 days (<30 days). If the prediction is 0, it means that the patient was not readmitted within 30 days, whether it is not at all or after 30 days.

Normalisation of continuous variables

The continuous variables in the dataset were normalised to a value between 0 and 1. This is useful as the data for continuous variables have varying scales and some of the machine learning models do not make assumptions about the distribution of our data. Hence, normalisation helps to transform the continuous variables to have similar distributions and this is especially important to the Neural Network.

Train-test split and balancing of imbalanced data

The models are used to predict the binary dependent variable ‘*readmitted’.* However, we realised that the responses under the *‘readmitted’* variableare not balanced, with 10125 “1”s, and 80082 “0”s. Having such unbalanced data could skew our model when training it and create a large amount of false negatives during prediction with the test set. To balance the train set, we first split the data into a train set and a testset using a 90-10 train-test split ratio. This resulted in a train set with 81186 rows and a test set with 9021 rows. Next, we sampled from the majority (“0”) in the train set, to get the same number of “1's'' and “0’s” in the train set. With this, we had a train set containing 18224 rows, and a test set containing the same 9021 rows. This balanced train set and test set was then used in all three predictive models.

The reason why we decided on a 90-10 train-test split ratio instead of the usual 70-30 is because of the size of the minority. When we used a 70-30 train-test split ratio, and then sampled from the majority to get an equal number of rows for the majority and minority, we ended up with a train set that was far smaller than the test set. As such, we chose to increase the size of the train set before balancing so that even after sampling from the majority and balancing the data, we would get a train set that was larger than the test set.

After cleaning and preparing the data, we moved on to generating our models.

## 3.5 Method 1: Logistic Regression

Our first analytical model is Logistic Regression with backward elimination. Logistic Regression is the estimation of the probability of a binary or categorical outcome variable. We used a confusion matrix to retrieve and examine the results of our prediction using the Logistic Regression Model.

The glm() function is first used on the train set to get the initial model using all the independent variables as input to the model. Next, the vif() function was used to check for multicollinearity. As the GVIF for all the variables was less than 5 (Appendix 7.3.2), none of the variables were removed for multicollinearity.

Next, the step() function was used on all the variables to perform backward elimination to retrieve the variables that would get the best-fit logistic regression model. The step() function identifies the variables by finding the combination of independent variables that will give the lowest Akaike Information Criterion (AIC) value.

The variables left in the model after backward elimination were found to be *‘age’, ‘A1Cresult’, ‘diabetesMed’, ‘insulin’, ‘metformin’, ‘glyburide’, ‘time\_in\_hospital’, ‘num\_procedures’, ‘num\_medications’, ‘number\_emergency’, ‘number\_inpatient’,* and *‘number\_diagnoses’*. The table in Appendix 7.3.3 shows the interpretation and details of the identified variables.

The updated model was then used to predict the readmission outcome for both the train set and test set. The predicted values were then compared to actual values using a confusion matrix, one for the train set (Appendix 7.3.1) and one for the test set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 5496 | 2512 | **FPR** | 31.4% |
| Readmitted within 30 days | 493 | 520 | **FNR** | 48.7% |
| **Overall Accuracy** | | 66.7% | | | |

*Fig 3.5.2: Test set Confusion Matrix for Logistic Regression Model*

In examining the test set confusion matrix, there is a false negative rate of 48.7%, which means that out of the number of diabetic patients that were actually readmitted, 48.7% of them were predicted by the model to not be readmitted.

Notably, the overall accuracy of the model is 66.7%, which is not very high. However, we have identified that this comes from a rather high amount of false positives. As mentioned previously, false positives are not as detrimental to the hospital as false negatives, thus we prioritise the false negative rate over the overall model accuracy.

Next we used the summary() function to identify the statistically significant variables to use for the neural network model, by looking at their p-value (values less than 0.1). The identified variables were *age, diabetesMed, insulin, num\_procedures, num\_medications, numer\_emergency, number\_inpatient* and *number\_diagnoses.*

## 3.6 Method 2: Neural Network

Neural network is a machine learning model that learns to perform certain tasks by studying labelled training examples. It models after a human brain where many nodes are interconnected. Data is fed in from the input layer to the hidden layers, where it is passed from one layer of nodes to the next, finally ending at the output layer. A working neural network will have layers of neurons with weights assigned. While using the model, input is passed through these weights and biases to produce a prediction at the output.

Our Inputs

Significant variables from our Logistic Regression model were used as inputs and fitted into the neural network model. As the neuralnet() package in R is unable to read categorical variables, we created dummy variables for every category of each categorical variable.

For example, for the categorical variable, insulin, it has {insulin\_No, insulin\_Steady, insulin\_Up} as possible values and it was converted as shown.

|  |  |  |  |
| --- | --- | --- | --- |
| Before converting | After converting | | |
| **insulin** | **insulin\_No** | **insulin\_Steady** | **insulin\_Up** |
| No | 1 | 0 | 0 |
| Steady | 0 | 1 | 0 |
| Up | 0 | 0 | 1 |
| Down | 0 | 0 | 0 |

*Fig 3.6.1: Example on creating Dummy Variables for Categorical Variables*

The categorical variables which dummy variables were created for are *gender, age, admission\_type\_id, admission\_source\_id, A1Cresult, metformin, glipizide, glyburide, pioglitazone, insulin, change, diabetesMed* and *readmitted*.

The Neural Network model will be trained using the train set, and then used to predict the patient’s readmitted status on the test set.

Construction of Neural Network

The accuracy of the neural network can be affected by parameters such as the number of hidden layers and the number of hidden nodes per layer. Weights of the paths are randomly assigned at the start and are revised by the Neural Network model to optimal values via backpropagation. The Neural Network with optimised weights can be found in Appendix 7.4.5.

The following combinations of the number of hidden nodes and hidden layers used in the neural network model were tested, producing the following results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of hidden layers and nodes** | **False Positive Rate** | **False Negative Rate** | **Overall Model Accuracy** |
| 1 hidden layer with 2 hidden nodes | 35.7% | 44.3% | 63.4% |
| 1 hidden layer with 1 hidden node | 36.3% | 42.8% | 62.9% |
| 2 hidden layers:  - 2 hidden nodes in 1st hidden layer  - 1 hidden node in 2nd hidden layer | 39.9% | 41.0% | 59.9% |
| 2 hidden layers, 1 hidden node per layer | 36.8% | 43.1% | 62.5% |

*Fig 3.6.2: Comparison on performance of Neural Network model on the test set for different number of hidden layers and hidden nodes*

Out of the combinations above, the model with 2 hidden nodes in the 1st layer and 1 hidden node in the 2nd layer was the optimal one with a False Negative Rate of 41%, which is the metric to be prioritised as explained earlier.

The optimal model was then applied to both the train set (Appendix 7.4.4) and test set, producing the following confusion matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2 hidden layers:  - 2 hidden nodes in 1st layer  - 1 hidden node in 2nd hidden layer | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 4809 | 3199 | **FPR** | 39.9% |
| Readmitted within 30 days | 415 | 598 | **FNR** | 41.0% |
| **Overall Accuracy** | | 59.9% | | | |

*Fig 3.6.3: Test set Confusion Matrix for Chosen Neural Network Model*

The best Neural Network model gives us a False Negative Rate of 41.0% and an overall model accuracy of 59.9%. While its overall accuracy is slightly lower compared to the other models, this is compensated by its relatively lower False Negative Rate.

## 3.7 Method 3: Random Forest

Random forest is an algorithm that makes use of multiple decision trees for classification. It works by bootstrapping – that is, sampling with replacement – the data to generate samples of size equal to the dataset used to train the model. The sample points not selected during bootstrap form the out-of-bag data, which are used as a natural test set for random forest. From there, each sample is used to create a tree independently. During the creation of each of these trees, random subset feature selection is used at every split. The randomness of every decision tree in a random forest ensures that each tree is trained on a slightly different set of observations in the dataset, thus each tree is split slightly differently. When making predictions, the predictions of each model are bagged (bootstrap aggregation). In other words, each individual decision tree is used to produce a result, and the result with the highest tally of votes becomes the model's prediction as a whole.

The use of separate samples for each tree combined with the random subsetting at every split introduces instability, resulting in different trees. That instability is balanced out by generating many of such trees. Hence, random forest has increased accuracy over its subsidiary CART due to the aggregation of these multiple models.

We used all the predictor variables from the dataset for the random forest model, because unlike neural network, random forest has the ability to derive variable importance. Variable importance in random forest is derived by randomly permuting a particular predictor variable's values (while keeping all other predictor variables constant), then finding out the mean decrease in accuracy, in which the greatest decrease signifies the highest variable importance.

The random forest model was created using the ‘RandomForest’ package in R, with the following default values: number of trees = 500, random subset feature size = floor(sqrt(20)) = 4. The default ‘ntree’ parameter works fine in this case; looking at the random forest plot at Appendix 7.5.1, we can see that the errors have converged and stabilised, indicating that the number of trees is sufficiently large to prevent model overfitting. The random forest model obtained an out-of-bag (OOB) error of 41.1%, which is indicative of how good of a fit the model is on the data. Fig 3.7.1 shows the confusion matrix of test set predictions, which in contrast, reveal the prediction accuracy of the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 4863 | 3142 | **FPR** | 39.3% |
| Readmitted within 30 days | 428 | 584 | **FNR** | 42.3% |
| **Overall Accuracy** | | 60.4% | | | |

*Fig 3.7.1: Test set Confusion Matrix for Random Forest Model (Initial)*

We decided to further refine the model, as the FNR was quite high at 42.3%. To do so, we had two approaches:

1. Choose specific variables to feed into the random forest formula
2. Tune the hyperparameters

For approach 1, we fed in the significant variables of logistic regression (hence using the same variables selected for neural network). We also tried only the top five most important variables of the original random forest model, as can be seen in Appendix 7.5.2; because from the sixth variable onwards, the mean decrease in accuracy is relatively smaller. We found that in both cases, the FNR actually increased. Hence, we decided to stick with using all the available variables.

In approach 2, we changed the ‘mtry’ parameter, which is the RSF selection size – the number of variables chosen to be considered at each split in the decision trees created. This is because the ntree – number of trees – looked to be sufficiently large such that the errors have stabilised. Appendix 7.5.3 shows the table of using different ‘mtry’ sizes and their corresponding testset error metrics. The lowest error was obtained when mtry = 8, hence that was the parameter used for the optimised random forest model. The optimised model had a marginally higher OOB error than the original untuned random forest model of 41.4%. Therefore, the optimised model fit the data slightly less well. However, the FNR decreased, as can be seen in Fig 3.7.2 below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 4679 | 3327 | **FPR** | 41.6% |
| Readmitted within 30 days | 406 | 606 | **FNR** | 40.4% |
| **Overall Accuracy** | | 58.6% | | | |

*Fig 3.7.2: Testset Confusion Matrix for Random Forest Model (Optimised)*

Therefore, random forest was able to yield the best predictive model from our analysis, albeit marginally better than the optimised neural network model.

Lastly, for both the random forest models created, the top five most important variables were the same, with ‘number\_inpatient’ being the most important variable by far (Appendix 7.5.2 and Appendix 7.5.4). Number inpatient is the number of in-patient visits the patient has had, although we do not know what the previous inpatient visits were for. This variable being identified as the most important makes logical sense, because the more in-patient stays a patient has had, the more severe their health condition is likely to be, and hence the greater the likelihood of that patient being readmitted. Hospitals could consider studying these variables to better understand the relationship between them and the chance of patient readmission.

# 

# 4 Evaluation of the Models

## 4.1 Strengths of analysis models

Logistic Regression

The first strength of Logistic Regression is that it does not require as much computational power as Neural Network and Random Forest. Running the glm() function was almost immediate despite our rather large dataset (90207 rows), while the other 2 models took much longer to run. It is also simple to develop logistic regression models as the glm() function can automatically create dummy variables.

Another strength of Logistic Regression is that the resulting model is simple and easy to interpret. It is generally quite easy for the layman to understand since it is just multiplying the values with the variable coefficients and adding them up. The coefficients for continuous independent variables are also often indicative of the importance of that particular variable, but only if the units are comparable, which we ensured by normalising the values of continuous variables during data preparation. We can also determine whether the independent variable has a positive or negative impact on the dependent variable by looking at the coefficient. Lastly, we are able to identify statistically significant variables using their associated p-value.

Neural Network

Its biggest strength is in being able to model complex relationships. Neural networks are able to approximate any continuously differentiable function, they are hence unique from almost every other machine learning algorithm in that respect.

Despite the neural network model being quite difficult to understand, its implementation is relatively hands off. All the training and fitting is done automatically by R, with just a few parameters such as the number of layers that needs to be adjusted. This is useful as parameters can be adjusted to fit different datasets well. In our case, the parameters that can be adjusted were used to fine tune the model to achieve better overall accuracy and false negative rates.

Random Forest

Random forest is very effective at reducing overfitting and variance. Aggregating the prediction of many differing trees to arrive at a final prediction allows random forest to have an increased accuracy compared to a single decision tree. Random forest is easy to use and understand, making it simple to implement with R. Furthermore, the default parameters generally perform well, which further adds on to its ease of implementation. Another strength is how the selection of bootstrap samples creates a natural testset for random forest to predict on, forming the out-of-bag estimate. The out-of-bag estimate can hence be thought of as the train set error. However, we still predicted on the test set for random forest, to standardise the procedures and be able to compare the prediction accuracies across all the models we developed. Lastly, random forest is able to yield the variable importance with the permutation method described earlier. The mean decrease in accuracy graph in Appendix 7.5.4 gives us a clear idea of the most important variables, in relation to the other variables as well.

## 4.2 Limitations of analysis models

Overall Limitations

All our models suffered from a relatively high train set error of around 40%, it followed that the test set errors were also around 40%. This could be due to underfitting, in other words, the predictor variables were simply not good predictors of a patient’s readmitted status.

Another limitation was the presence of a large amount of missing data for certain variables in the original dataset. As mentioned in our data cleaning, variables with huge amounts of missing data were omitted from the dataset used for training the model and these unused variables could have been significant in predicting readmission rates.

In addition, during data preparation, we corrected the class imbalance in the trainset by sampling from the majority (those who were not readmitted within 30 days). Hence, much of the dataset comprising of the negative class (patients who were not readmitted within 30 days) were left out from the trainset. Hence, the models may not have had enough data to “learn” from, as the dataset was downsampled, which could potentially contribute to a high error rate, especially for predicting the negative class.

Lastly, the dataset used contained data for diabetic patients from hospitals in the USA, which may not be representative of diabetic patients and their readmission rates in Singapore. Thus, the significant variables we identified using data from the diabetic patients in the USA may not be the same as the ones potentially identified for diabetic patients in Singapore.

Logistic Regression

One of the limitations of Logistic Regression is that it assumes a linear relationship between the independent and dependent variables. In the real world, this often does not hold true. Forcing variables that do not have a linear relationship into the model could therefore take away from the accuracy and performance of the model. Additionally, logistic regression is also prone to overfitting (Robinson, 2018), particularly when it is complex and there are many variables involved. In our case, our model has 11 variables, which is objectively quite a large number. However, since our dataset is rather large (90207 rows), the effects of overfitting due to the large number of variables have been reduced for our Logistic Regression Model.

Additionally, there is also the inherent limitation of using an algorithm to decide the significant variables, which is the absence of real-life situational knowledge and expertise (Flom, 2018). Due to this, it is often recommended that these models should also be supported by expert opinion. However, since we are not experts, we have no choice but to rely on what the algorithm tells us the significant and useful variables are.

Neural Network

Neural network is said to be a black box model; the functioning of the model cannot be well understood. Hence, when the model gives a probing solution or predictions, it does not give a clue as to why or how. There is thus no way to diagnose an issue other than to retrain with different parameters. Generally, we cannot use the Neural Network to identify significant variables. Thus, this is the reason why we had to first obtain the significant variables from Logistic Regression then fit those variables into our Neural Network model.

Not only that, the neural network model requires significant computational power to train, and took the longest time among the 3 models used. This problem arises particularly when the size of the dataset is very large as the computational times do not generally scale well with the amount of data, or the number of layers and nodes. Due to this limitation and our large data set, we had to downsize the training dataset for our Neural Network model (using sampling) to compute its predictions within a reasonable amount of time. Since neural networks also require more data than other machine learning algorithms in order to learn the abstract representations of the data, that further adds to the computational times.

Lastly, Neural Network has no variable importance feature. While we can still interpret the sign and magnitude of a variables’ effect on the prediction, we have to normalise the continuous variables first, like in logistic regression.

Random Forest

Random forest’s more hands off approach makes it easy to use, as a user can simply input the data, build the model, and make predictions. However, this black box nature of random forest also means that the model is relatively inflexible, since the user has little control over how the model is created and how predictions are made. Other than a few parameters like the total number of trees or the random subset feature size, options to tune the accuracy of the model are limited. It is also difficult to interpret the resulting random forest model because it consists of numerous trees, making visualisation of the model hard.

Additionally, random forest also has a moderately high computation time, especially when a large pool of data is used, since it has to form the many individual trees and aggregate their results. For our project, even when using a downsampled version of our data, we still took a significant amount of time to train and test our models. Hence, this can be an issue when the model has to be updated frequently to accommodate changes in demographics.

## 4.3 Comparison of performance of the models

The overall accuracy and false negative rates of the three models are shown below:

|  |  |  |
| --- | --- | --- |
| **Model** | **False Negative Rate** | **Overall Model Accuracy** |
| Logistic Regression | 48.7% | 66.7% |
| Neural Network | 41.0% | 59.9% |
| Random Forest | 40.4% | 58.6% |

*Fig 4.3.1: FNR and Overall Model Accuracy comparisons between all 3 models*

The model with a false negative rate far higher than the other two was logistic regression (48.7%). Therefore, we ruled out the logistic regression model as a possible recommendation to the hospital. Next, the neural network model generally had a slightly higher overall model accuracy as compared to random forest. However, the random forest model has a slightly lower false negative rate as compared to the neural network model. Therefore, the random forest and neural network models were almost similar in their performance.

Other than the false negative rate and overall model accuracy, we decided to also look at the precision rate, recall rate and F1 score of the 3 models to decide on the most optimal model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision Rate** | **Recall Rate** | **F1 Score** |
| Logistic Regression | 17.2% | 51.3% | 25.6% |
| Neural Network | 15.8% | 59% | 24.8% |
| Random Forest | 15.4% | 59.9% | 24.5% |

*Fig 4.3.2: Comparison of Precision Rate, Recall Rate and F1 Score between all 3 models*

The neural network and random forest models have almost similar precision rates, recall rates and F1 scores. Both of them also have a lower precision rate but a higher recall rate as compared to the Logistic Regression model. The F1 score is a more useful metric than overall accuracy since overall accuracy is generally not a good metric to use for data with class imbalance (Korstanje, 2021). In our case, the test set has a class imbalance. However, the Logistic Regression has only a slightly higher F1 score than the Neural Network and Random Forest model. Also, the Neural Network model has only a slightly higher F1 score than the Random Forest model. Therefore, it is difficult to conclude whether the Neural Network or Random Forest model is more optimal. Hence, we will focus on the false negative rate as the metric to determine the optimal model.

Even though the random forest model has only a slightly lower false negative rate as compared to the neural network model, we will recommend the random forest model as random forest is able to provide variable importance while neural network cannot. This will be useful for the hospitals to identify top variables that are significant in affecting a patient’s readmission rate within 30 days.

Hence, based on our random forest model, we recommend the hospital to take into consideration the following variables as they have the highest variable importance in predicting whether a patient will be readmitted within 30 days.

The 3 variables *“number\_ inpatient”, “time\_in\_hospital” and “age”* are the most significant variables in determining if a patient will be readmitted within 30 days. Hence, we can recommend hospitals in Singapore to give higher considerations and priority to these 3 variables. In addition, hospitals can apply their domain knowledge and monitor processes related to these 3 variables more closely. For example, hospitals can closely monitor the number of inpatient visits a diabetic patient goes for and provide more medical attention and resources to patients who have a higher number of inpatient visits.

However, it is noted that our models generated a low overall model accuracy, as the model with the highest overall accuracy is only 66.7%. Our low overall model accuracy is mainly contributed by false positives, where a large number of patients will be predicted by the model to be readmitted when in fact they will not. The low overall model accuracy could also be due to the dataset used and could be improved with the recommendations below.

## 4.4 Recommendations

Recommendations for dataset

**Different sets of data can be used in the model to cross check the accuracy of the variables.** A dataset from Singapore is preferred, as it is more applicable to hospitals in Singapore. With the use of more relevant and applicable data, this will increase the reliability of our models and increase its predictive power.

**Introduce new significant variables that were not included in the dataset used.** There might be other independent variables currently not present in the 20 independent variables in the dataset we used for the models that will be able to improve the predictive accuracy of our Random Forest model. For example, variables such as weight that have many missing rows in the original dataset should be more meticulously recorded and subsequently included in the predictive models, as the weight variable might be able to contribute to an increase in accuracy of predicting a patient’s readmitted status.

**Get hospital specific data on the patients frequency of readmission.** In addition, the dependent variable *“readmitted”* only focuses on whether or not the patient was readmitted within 30 days. The dataset has no variables on how many times a patient is readmitted within the 30 days. Patients with higher frequency of readmission will likely require more medical attention from the hospital, but may go undetected by the model. Hence, more specific data on the frequency of readmission within 30 days, or reason for readmission, can be collected by the hospital, and used by the model to better allocate medical attention and resources.

Recommendations for the model chosen

**Tune the hyperparameters of the Random Forest model in order to get the optimum value of each parameter**. It is recommended to run several tests of the Random Forest model with different settings of its hyperparameters and then get the best model which will improve the accuracy of our model. Some other hyperparameters that can be tuned are stated below.

1. **Tuning the maximum depth of the trees in the Random Forest model.** By default, trees are expanded until all leaves are either pure or contain less than the minimum samples for the split. However, this might cause the trees to overfit or underfit in some cases. Thus, we can try multiple numbers of max\_depth hyperparameters to find an optimal number which provides the best overall model accuracy.
2. **Changing the n\_estimators number in Random Forest.** The “number of estimators” hyperparameter sets the number of trees to be built before taking the maximum averages of predictions. The higher the number of n\_estimators set, the higher the number of trees, which will then give better performance. However, this will in turn result in higher computational time, and hence, it is recommended to choose as high a value as the processor can handle.

# 

# 5 Conclusion

In conclusion, we recommend hospitals in Singapore to use the Random Forest model to predict the likelihood of readmission within 30 days of diabetic patients, since the model has a relatively low false negative rate when compared to the Logistic Regression and Neural Network models. However, hospitals should not rely solely on the predictions of the Random Forest model as it has a relatively low overall accuracy due to its high false positive rate. Hence, further improvements to both the dataset and model can be first implemented via our recommendations (Section 4.4) in order to increase the model accuracy.

Furthermore, the model is trained using past data of diabetic patients from hospitals in the USA, which may not be representative of the demographics of diabetic patients and their readmission rate in Singapore. Variables deemed significant for the diabetic patients in the USA may not be significant to the diabetic patients in Singapore. Furthermore, readmission of diabetic patients within 30 days may also be affected by factors other than the independent variables in the dataset used to build our models. Should the model be implemented in a local hospital here, current comprehensive data of diabetic patients in hospitals in Singapore should be used during training to ensure that the predictions are made in the right context.

In addition, we realised that some of the patients appeared in our dataset more than once. This indicates that they might have a history of being constantly readmitted to the hospital. As such, we feel that to make our model more accurate, another variable that indicates if they are frequent readmitters can be included so that the hospital can direct more resources and pay more attention to these patients.

Moving forward, hospitals and other healthcare providers in Singapore can continue to do further research on the use of predictive and analytical models in their operations. Even beyond the predictive models suggested in this paper, other forms of Machine Learning and Artificial Intelligence can be further integrated into our healthcare systems, to not only predict hospital readmissions, but help hospitals in many other ways. There is also potential for them to cut costs and utilise their limited resources more efficiently to meet Singapore’s growing healthcare needs. This can reduce the burden on manpower and healthcare workers, without compromising on the quality of care.

# 6 References

Bragato, D. (2021). Statins and diabetes. Diabetes UK. Retrieved April 2, 2022, from <https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/treating-your-diabetes/tablets-and-medication/statins>

Chen, J. Y., Ma, Q., Chen, H., &amp; Yermilov, I. (2012). New Bundled World: Quality of Care and readmission in diabetes patients. Journal of Diabetes Science and Technology, 6(3), 563–571. <https://doi.org/10.1177/193229681200600311>

ITA (n.d.). *Singapore Healthcare*. International Trade Administration | Trade.gov. Retrieved April 2, 2022, from <https://www.trade.gov/market-intelligence/singapore-healthcare#:~:text=Government%20healthcare%20expenditure%20is%20estimated,to%20%2413.5%20billion%20in%202029>.

Flom, P. (2018, December 11). *Stopping stepwise: Why Stepwise selection is bad and what you should use instead*. Medium. Retrieved April 2, 2022, from <https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52df>

Foo, J. P., &amp; Fock, K. M. (2013). Diabetes care in Singapore: how an integrated approach can help meet the challenges ahead. Singapore Med J. Retrieved 2022, from <http://www.smj.org.sg/sites/default/files/5407/5407co1.pdf>

Graham, K. L., From Beth Israel Deaconess Medical Center and Harvard Medical School, Wilker, E. H., Howell, M. D., Davis, R. B., Marcantonio, E. R., M. Krumholz, K. D. and H., Reena Duseja, L. Hirsch, R., R. Marcantonio, K. L. G. and E., Hirsch, R. L., & Dharmarajan, K. (2015, June 2). *Differences between early and late readmissions among patients: A cohort study: Annals of internal medicine: Vol 162, no 11*. Annals of Internal Medicine. Retrieved April 2, 2022, from <https://www.acpjournals.org/doi/10.7326/M14-2159>

Goetghebeur , M. M., Hay, J. W., & Forrest, S. (2003, June). *Understanding the underlying drivers of inpatient cost growth: A literature review*. The American journal of managed care. Retrieved April 2, 2022, from <https://pubmed.ncbi.nlm.nih.gov/12817611/>

Khew, C. (2017, March 27). *New web tool to predict risk of early hospital readmission*. The Straits Times. Retrieved April 2, 2022, from <https://www.straitstimes.com/singapore/new-web-tool-to-predict-risk-of-early-hospital-readmission>

Lim, C. Y., Ho, J. S., Huang, Z., Gao, F., Tan, S. Y., Koh, W. P., Chua, T., Low, L. P., Tan, H. C., &amp; Yoon, S. (2021). Public perceptions and knowledge of cholesterol management in a multi-ethnic Asian population: A population-based survey. PLOS ONE, 16(8). <https://doi.org/10.1371/journal.pone.0256218>

Low, L. L., Liu, N., Wang, S., Thumboo, J., Ong, M. E. H., & Lee, K. H. (2016, December 9). *Predicting 30-day readmissions in an Asian population: Building a predictive model by incorporating markers of hospitalization severity*. PLOS ONE. Retrieved April 2, 2022, from <https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0167413#pone.0167413.ref003>

Low, L. L., Liu, N., Wang, S., Thumboo, J., Ong, M. E. H., & Lee, K. H. (2016, October 14). *Predicting frequent hospital admission risk in Singapore: A retrospective cohort study to investigate the impact of comorbidities, acute illness burden and social determinants of health*. BMJ open. Retrieved April 2, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5073633/#R8>

Low, L. L., Tay, W. Y., Ng, M. J. M., Tan, S. Y., Liu, N., & Lee, K. H. (2018, January). *Frequent Hospital admissions in Singapore: Clinical risk factors and impact of socioeconomic status*. Singapore medical journal. Retrieved April 2, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778257/>

*Multiple-readmissions-predictive-model //* . Healthcare\_Systems. (n.d.). Retrieved April 2, 2022, from <https://www.ihis.com.sg/Project_Showcase/Healthcare_Systems/Pages/Multiple-Readmissions-Predictive-Model.aspx>

National University of Singapore. (2017, March 27). *NUS Pharmacy team develops online “calculator” to predict risk of early hospital readmission*. NUS news. Retrieved April 2, 2022, from <https://news.nus.edu.sg/nus-pharmacy-team-develops-online-calculator-to-predict-risk-of-early-hospital-readmission/>

*News highlights*. Ministry of Health. (2017, November 14). Retrieved April 2, 2022, from <https://www.moh.gov.sg/news-highlights/details/ministry-of-health-studying-measures-to-better-support-persons-with-pre-diabetes#:~:text=About%20430%2C000%20>(or%2014%20per,to%20be%20diagnosed%20with%20diabetes.

Ostling, S., Wyckoff, J., Ciarkowski, S. L., Pai, C.-W., Choe, H. M., Bahl, V., & Gianchandani, R. (2017, March 22). *The relationship between diabetes mellitus and 30-day readmission rates - clinical diabetes and endocrinology*. BioMed Central. Retrieved April 2, 2022, from <https://clindiabetesendo.biomedcentral.com/articles/10.1186/s40842-016-0040-x>

Ow Yong, L. M., & Koe, L. W. P. (2021, February 8). *War on diabetes in Singapore: A policy analysis - health research policy and systems*. BioMed Central. Retrieved April 2, 2022, from <https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-021-00678-1#:~:text=In%20Singapore%2C%20over%20400%2C000%20Singaporeans,million%20by%202050%20%5B1%5D>

*Readmission prediction*. GEMINI - a Generalizable Medical Information analysis and Integration System. (2020). Retrieved April 2, 2022, from <https://www.comp.nus.edu.sg/~dbsystem/gemini/readmission.html>

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021, July 30). *The 30-days hospital readmission risk in diabetic patients: Predictive modeling with Machine Learning Classifiers*. BMC medical informatics and decision making. Retrieved April 2, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8323261/>

Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). *Readmission rates and Their Impact on Hospital Financial Performance: A Study of Washington Hospitals*. Inquiry : a journal of medical care organization, provision and financing. Retrieved April 2, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6614936/>

Khew, C. (n.d.).*Web Tools Spots Patients at Risk of Being Readmitted to Hospital*. HealthXchange.sg. Retrieved April 1, 2022, from <https://www.healthxchange.sg/news/web-tool-spots-patients-at-risk-of-being-readmitted-to-hospital>

Wong, E. G., Parker, A. M., Leung, D. G., Brigham, E. P., & Abarje, A. I. (2016). *Association of severity of illness and Intensive Care Unit Readmission: A systematic review*. Heart & lung: the journal of critical care. Retrieved April 1, 2022, from <https://pubmed.ncbi.nlm.nih.gov/26702501/#:~:text=In%20most%20studies%2C%20severity%20of,form%20of%20a%20meta%2Danalysis.>

Korstanje, J. (2021, August 31) Towards Data Science - The F1 score: All you need to know about the F1 score in machine learning. Retrieved April 2, 2022 from <https://towardsdatascience.com/the-f1-score-bec2bbc38aa6#:~:text=The%20F1%20score%20becomes%20especially,grid%20search%20or%20automated%20optimization>.

# 7 Appendix

## 7.1 Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Categorical** | | |
| **Dependent Variable** | **Levels** | **Remarks** |
| Outcome | 0, 1 | 0 = not readmitted or readmitted in more than 30 days  1 = readmitted within 30 days |
| **Independent Variable** | **Levels** | **Remarks** |
| Gender | Female, Male | - |
| Age | [0, 10), [10, 20), ..., [90, 100) | Grouped in 10-year intervals |
| Change of medications | Yes, No | - |
| A1C result | None, >8%, Other | Average blood sugar level over the past 3 months |
| Diabetes medications | Yes, No | - |
| Insulin (Type of medication for diabetes) | No, Steady, Up, Down | “Up”: dosage increased  “Down”: dosage decreased  “Steady”: no change to dosage  “No”: drug not prescribed |
| Glipizide (Type of medication for diabetes) | No, Steady, Up, Down | Same as above |
| Metformin (Type of medication for diabetes) | No, Steady, Up, Down | Same as above |
| Acarbose (Type of medication for diabetes) | No, Steady, Up, Down | Same as above |
| Examide (Type of medication for diabetes) | No, Steady, Up, Down | Same as above |
| Admission\_type\_id | - 1-8 types  - eg. Emergency, Urgent, ... | Types of admission |
| Admission\_source\_id | - 1-25 sources  - eg. Physician Referral, Clinic Referral, ... | Where the patient was admitted from |
| **Numeric** | | |
| **Variable** | **Range** | **Remarks** |
| Time in hospital | 1 - 14 days | - |
| Number of lab procedures | 1 - 132 procedures | - |
| Number of procedures | 0 - 6 procedures | - |
| Number of medications | 1 - 81 medications | - |
| Number of outpatient visits | 0 - 42 visits | - |
| Number of emergency visits | 0 - 76 visits | - |
| Number of inpatient visits | 0 - 21 visits | - |
| Number of diagnosis | 1 - 16 diagnosis | - |

*Appendix 7.1.1: Independent and Dependent Variables used from our dataset*

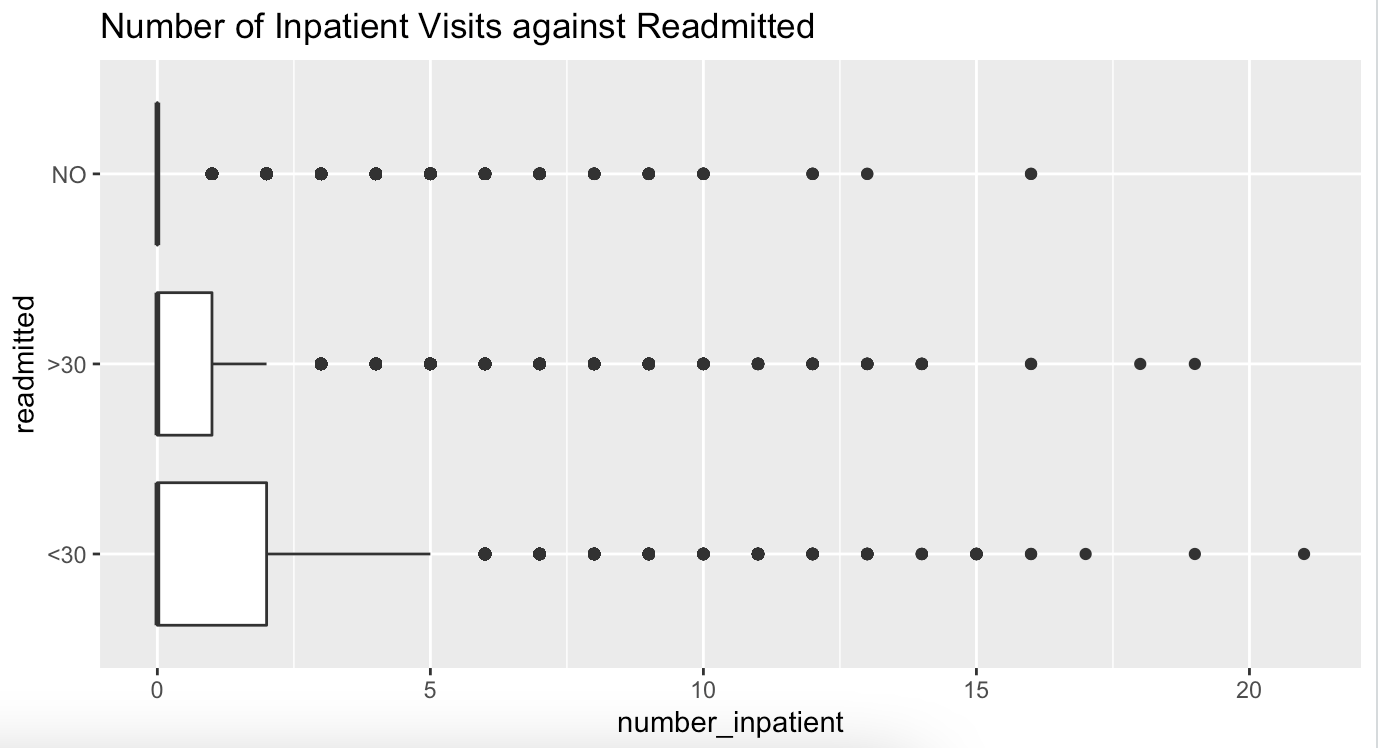
|  |  |
| --- | --- |
| **Admission Type ID** | |
| 1 | Emergency |
| 2 | Urgent |
| 3 | Elective |
| 4 | Newborn |
| 7 | Trauma centre |
| **Admission Source ID** | |
| 1 | Physician referral |
| 2 | Clinic referral |
| 3 | HMO referral |
| 4 | Transfer from a hospital |
| 5 | Transfer from a Skilled Nursing Facility |
| 6 | Transfer from another health care facility |
| 7 | Emergency room |
| 8 | Court/Law enforcement |
| 10 | Transfer from critical access hospital |
| 11 | Normal delivery |
| 12 | Premature delivery |
| 13 | Sick baby |
| 14 | Extramural birth |
| 18 | Transfer from another home health agency |
| 19 | Readmission to same home health agency |
| 22 | Transfer from hospital inpatient |
| 23 | Born inside this hospital |
| 24 | Born outside this hospital |
| 25 | Transfer from ambulatory surgery centre |
| 26 | Transfer from hospice |

*Appendix 7.1.2: ID Mapping for Encoded Variables*

## 

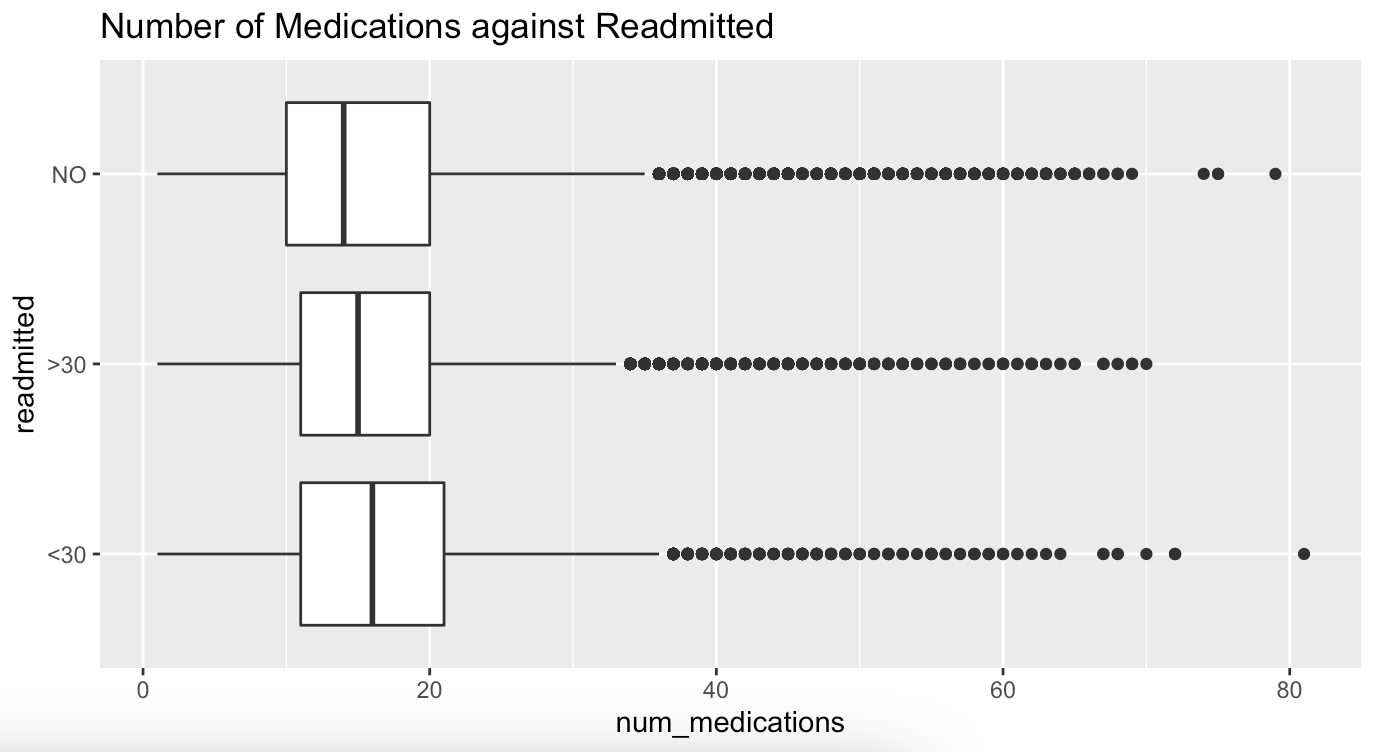
## 7.2 Exploratory Analysis

Other Interesting Plots



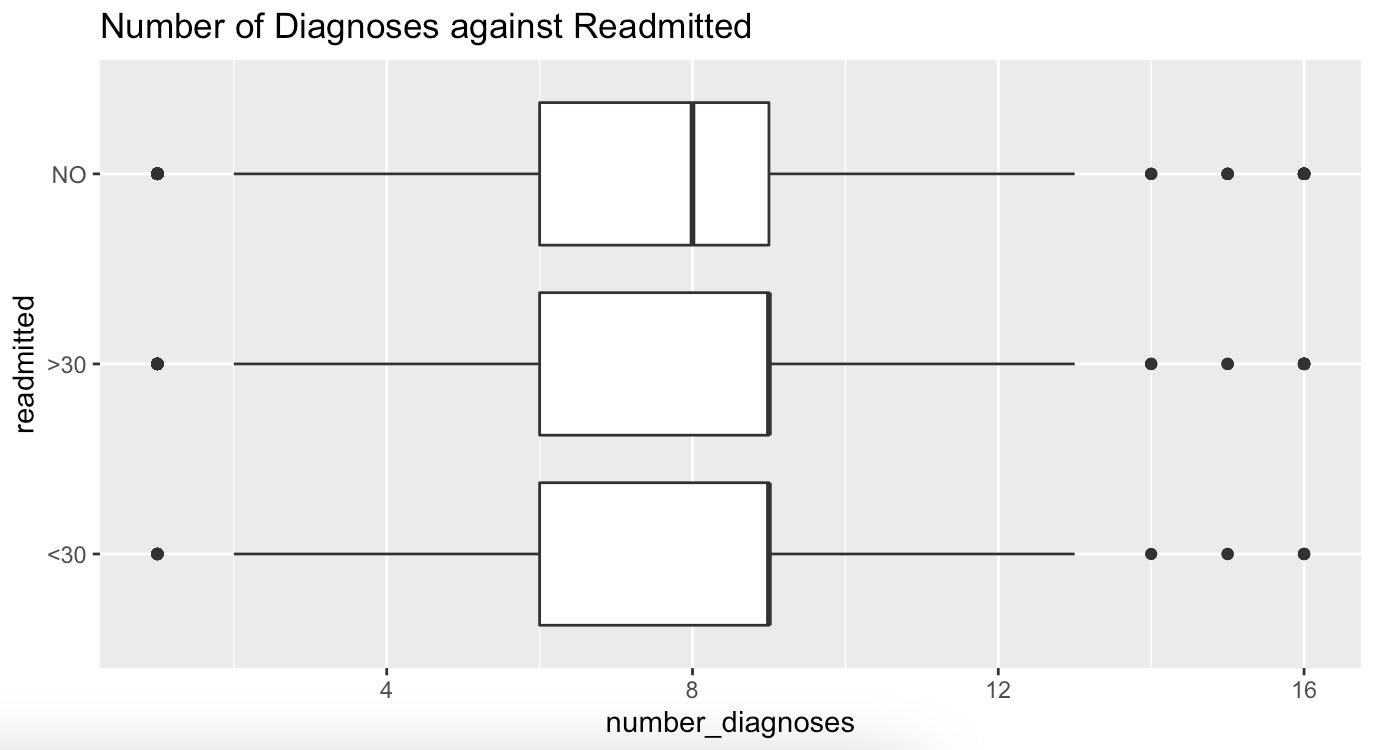
*Appendix 7.2.1: Boxplot of Number of Inpatient Visits against Readmission Outcome*

The 3rd quartile for the number of inpatient visits is the highest for patients readmitted within 30 days. Suggests that those who have multiple inpatient visits are more likely to be readmitted, likely due to a higher severity of their illness or condition.



*Appendix 7.2.2: Boxplot of Number of Medications against Readmission Outcome*

The 1st quartile, median, and 3rd quartile for the boxplot for those readmitted within 30 days is the highest. Implies that those who are prescribed more prescriptions are more likely to be readmitted. This could be because more medications indicate a more serious condition or illness, which is why they are more likely to be readmitted.



Appendix 7.2.3: Boxplot of Number of Diagnoses against Readmission Outcome

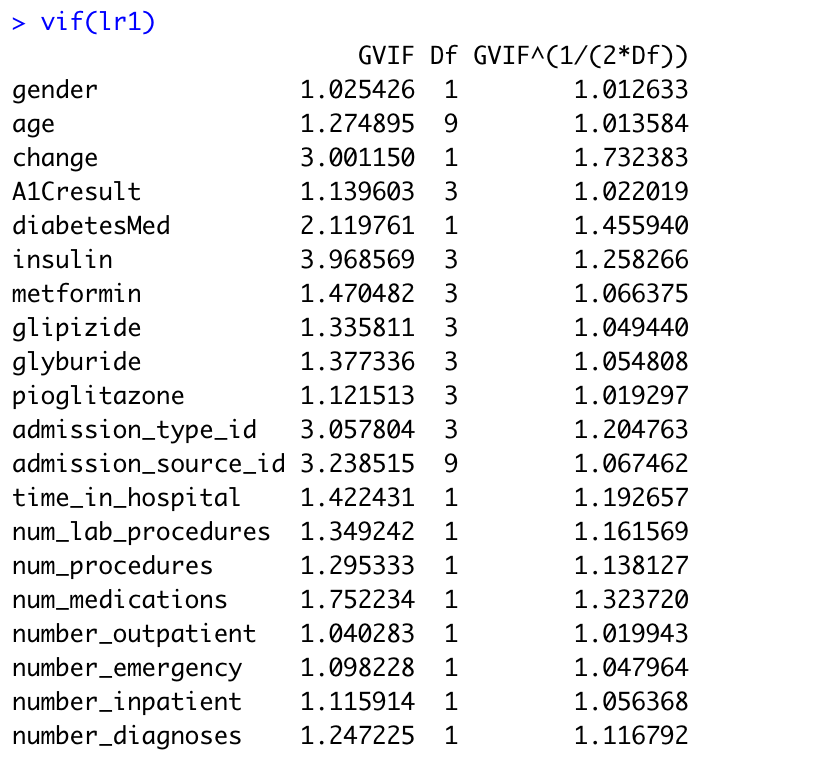
The median for the number of diagnoses is higher in the boxplots for readmitted within 30 days and readmitted in more than 30 days. Suggests that the likelihood of being readmitted is higher for those who have more diagnoses. This trend could be because a higher number of diagnoses indicates that the patient’s situation is more severe, and therefore would be more likely to be readmitted.

## 7.3 Logistic Regression

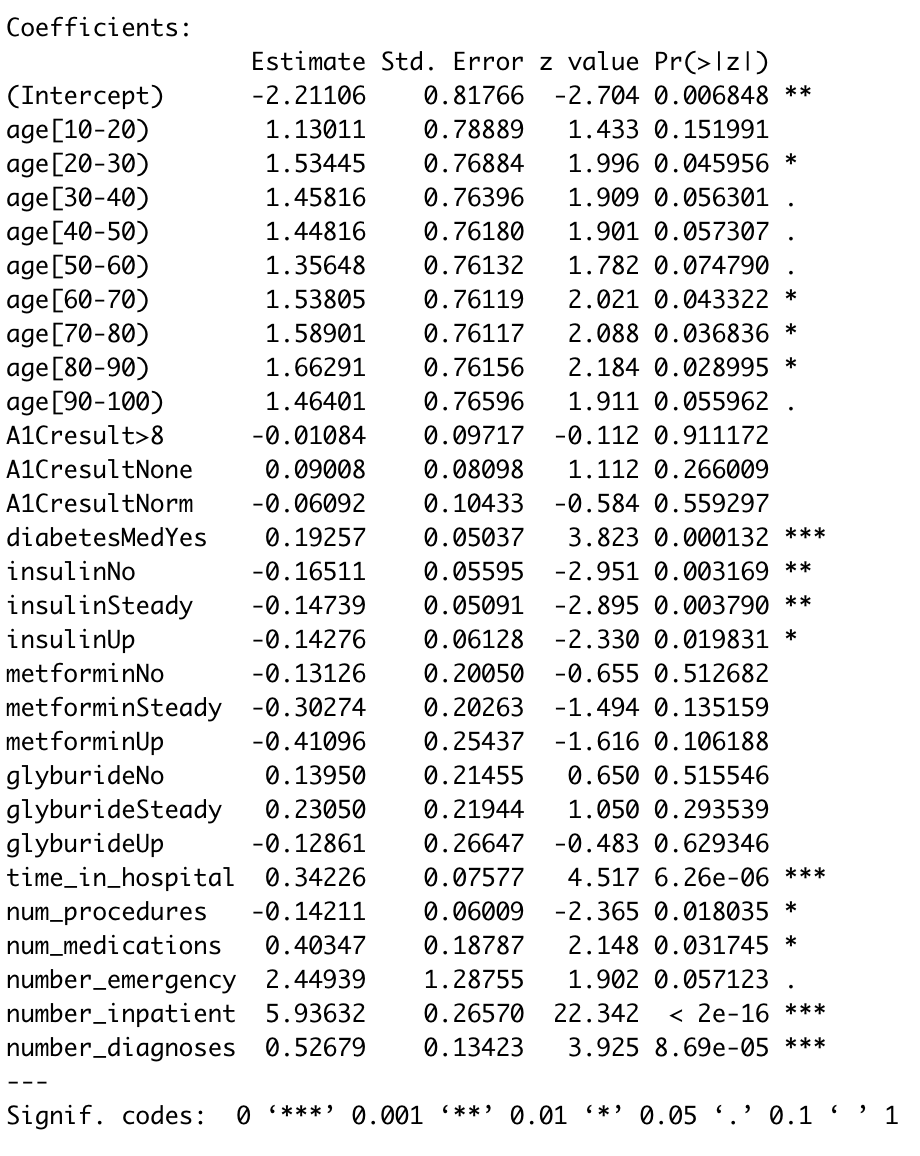
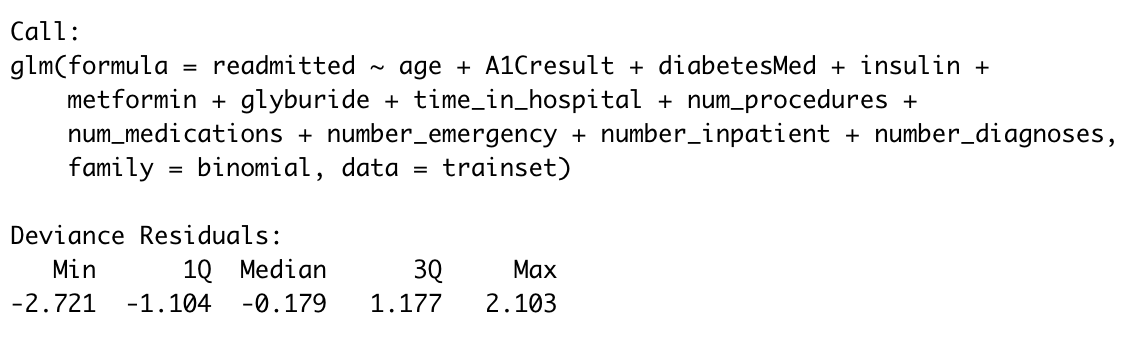
**Logistic Regression**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 6232 | 2880 | **FPR** | 68.4% |
| Readmitted within 30 days | 4545 | 4567 | **FNR** | 49.9% |
| **Overall Accuracy** | | 59.3% | | | |

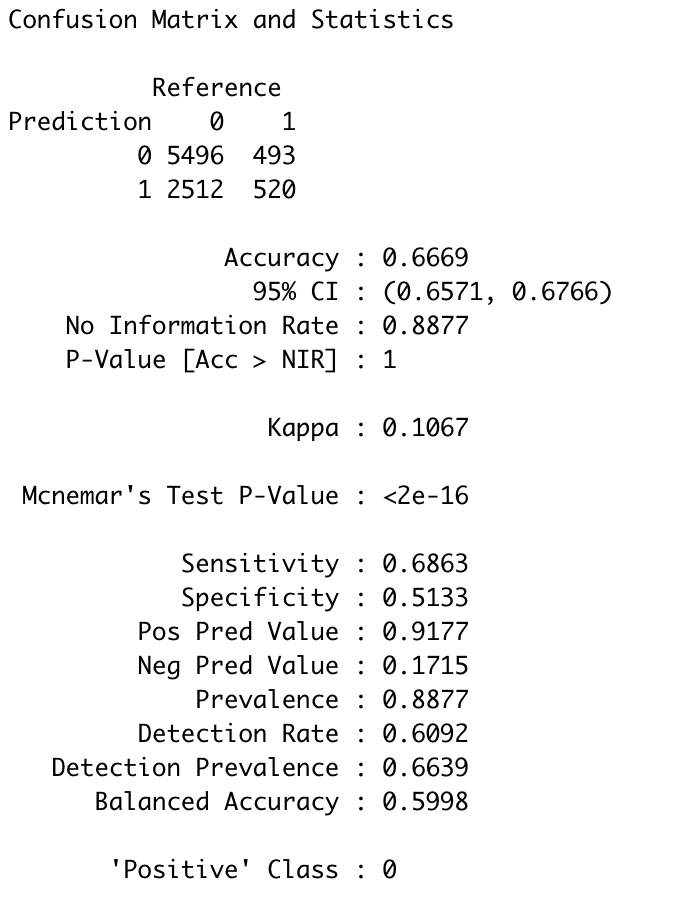
*Appendix 7.3.1: Trainset Confusion Matrix for Logistic Regression Model*



*Appendix 7.3.2: GVIF of all the variables*



*Appendix 7.3.3: Significant Variables generated from Logistic Regression after Backward Elimination*



*Appendix 7.3.4: Confusion Matrix from Logistic Regression*

## 7.4 Neural Network

|  |  |  |  |
| --- | --- | --- | --- |
| 1 hidden layer with 2 hidden nodes | | **Predicted** | |
| Will not be readmitted within 30 days | Will be readmitted within 30 days |
| **Actual** | Will not be readmitted within 30 days | 5153 | 2855 |
| Will be readmitted within 30 days | 449 | 564 |
| **Overall Accuracy** | | 63.4% | |
| **Overall Error Rate** | | 36.6% | |

*Appendix 7.4.1: Confusion Matrix for Neural Network Testset (*1 hidden layer with 2 hidden nodes)

|  |  |  |  |
| --- | --- | --- | --- |
| 1 hidden layer with 1 hidden node | | **Predicted** | |
| Will not be readmitted within 30 days | Will be readmitted within 30 days |
| **Actual** | Will not be readmitted within 30 days | 5099 | 2909 |
| Will be readmitted within 30 days | 434 | 579 |
| **Overall Accuracy** | | 62.9% | |
| **Overall Error Rate** | | 37.1% | |

*Appendix 7.4.2: Confusion Matrix for Neural Network Testset (*1 hidden layer with 1 hidden node)

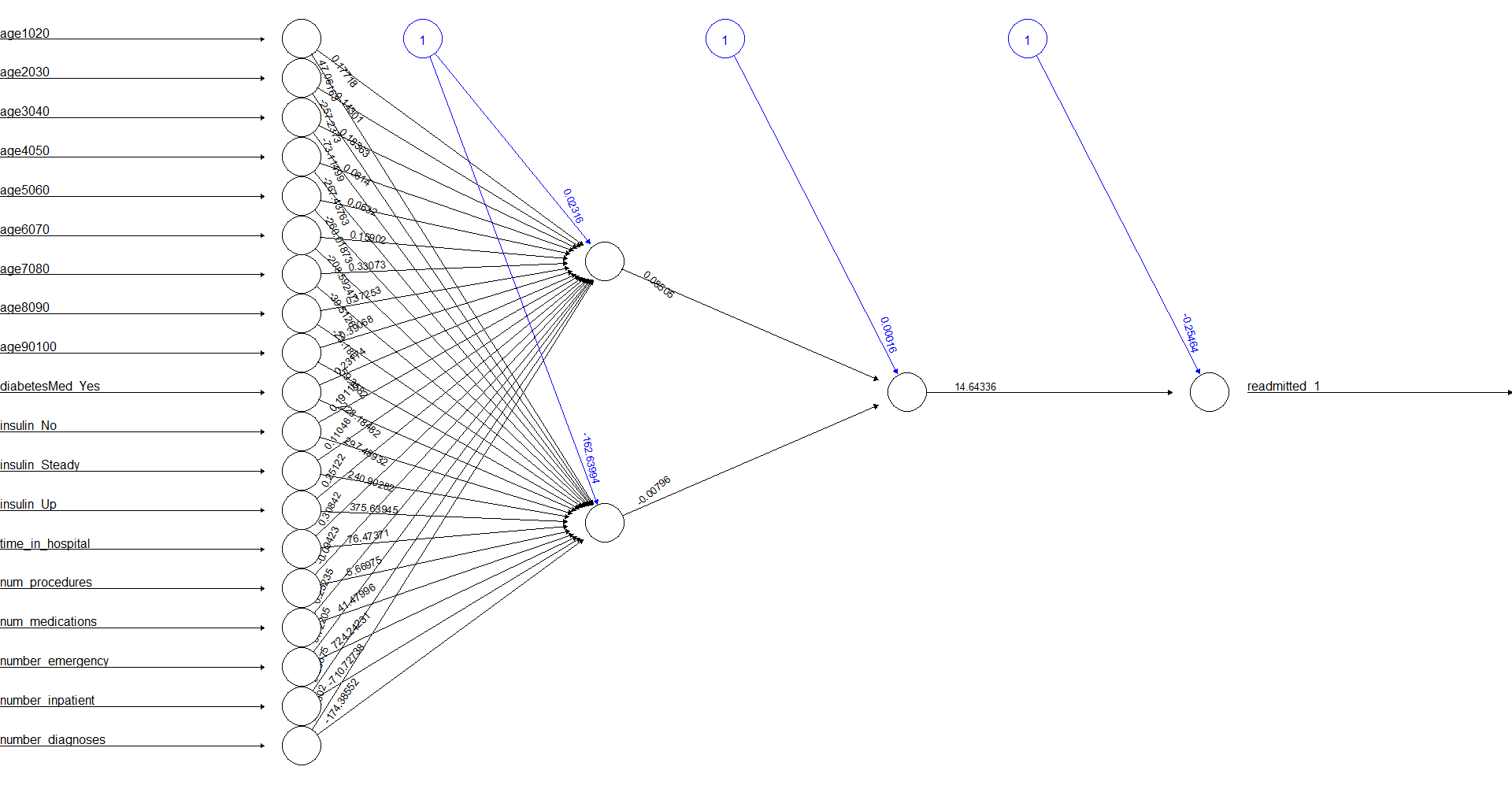
|  |  |  |  |
| --- | --- | --- | --- |
| 2 hidden layers, 1 hidden nodes in 1st layer, 1 hidden node in 2nd hidden layer | | **Predicted** | |
| Will not be readmitted within 30 days | Will be readmitted within 30 days |
| **Actual** | Will not be readmitted within 30 days | 5063 | 2945 |
| Will be readmitted within 30 days | 437 | 576 |
| **Overall Accuracy** | | 62.5% | |
| **Overall Error Rate** | | 37.5% | |

*Appendix 7.4.3: Confusion Matrix for Neural Network Testset (* 2 hidden layers, 1 hidden nodes in 1st layer, 1 hidden node in 2nd hidden layer)

**Neural Network**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2 hidden layers:  - 2 hidden nodes in 1st layer  - 1 hidden node in 2nd hidden layer | | **Predicted** | | **Class Error** | |
| Not readmitted within 30 days | Readmitted within 30 days |
| **Actual** | Not readmitted within 30 days | 3137 | 1918 | **FPR** | 37.9% |
| Readmitted within 30 days | 2022 | 2923 | **FNR** | 40.9% |
| **Overall Accuracy** | | 60.6% | | | |

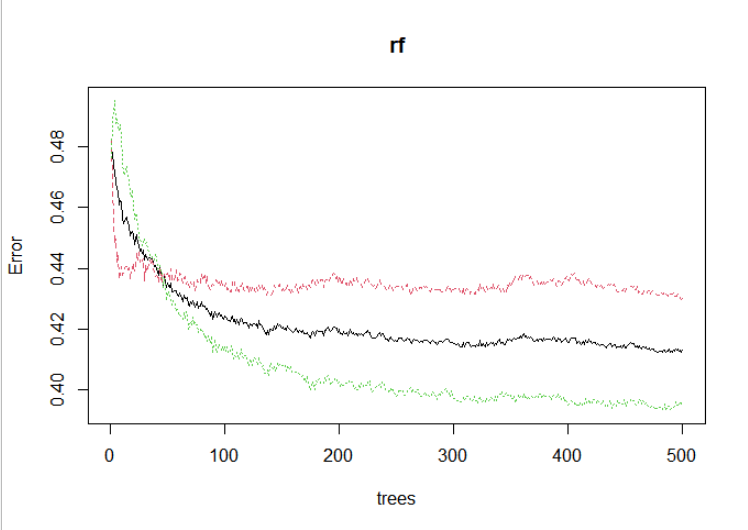
*Appendix 7.4.3: Trainset Confusion Matrix for Chosen Neural Network Model*

**

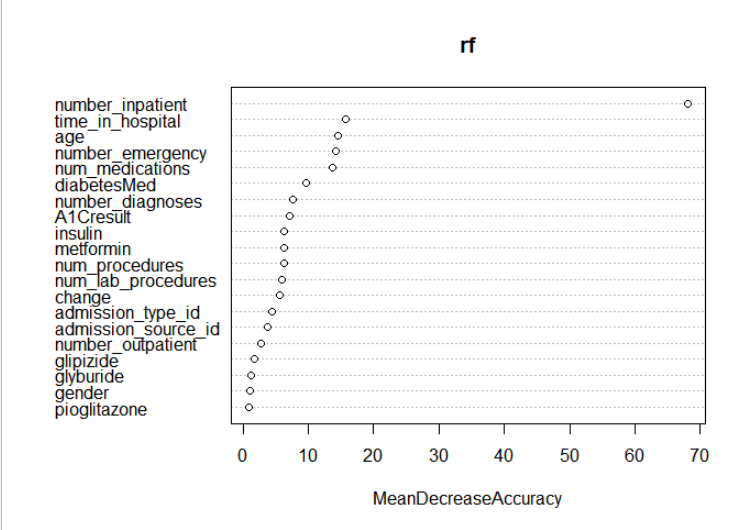
*Appendix 7.4.5: Neural Network model plot containing the final weights*

## 

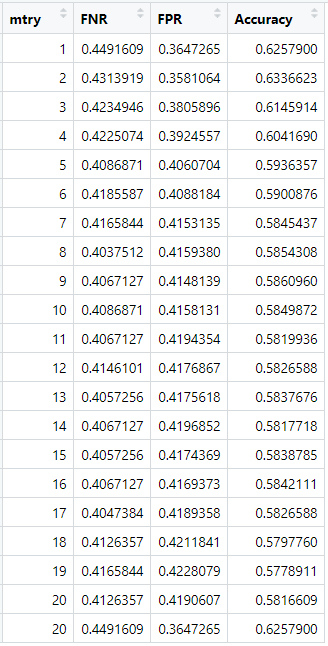
## 7.5 Random Forest



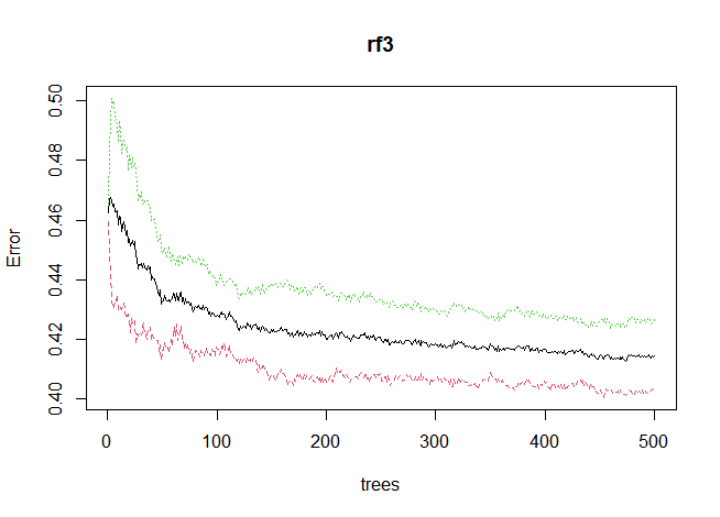
*Appendix 7.5.1: Random Forest Error Plot (Initial Model)*



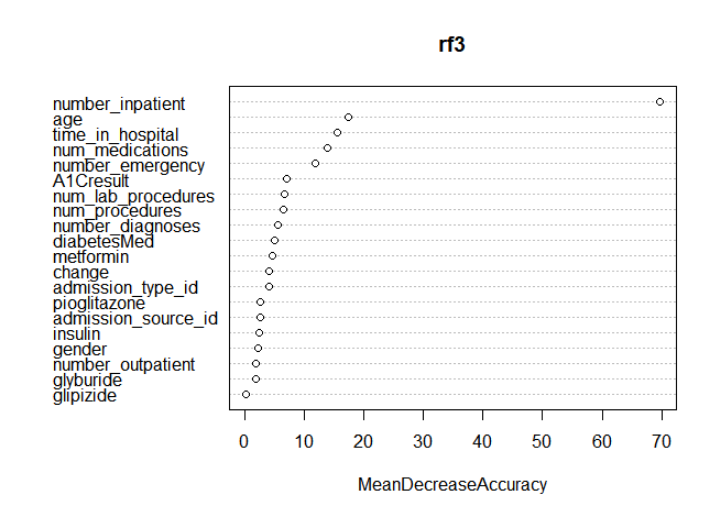
*Appendix 7.5.2: Random Forest Variable Importance Plot (Initial Model)*



*Appendix 7.5.3: Table of Different ‘mtry’ Sizes for Random Forest and Testset Error Metrics*



*Appendix 7.5.4: Random Forest Error Plot (Optimised ‘mtry’ Model)*



*Appendix 7.5.4: Random Forest Variable Importance Plot (Optimised ‘mtry’ Model)*