Predicting frequent attenders to urgent care centres using machine learning

Lara Sophia Camille Chammas

A final year dissertation presented for the degree of

Bsc Biological Sciences

Department of Life Sciences

Imperial College London

June 8th 2018

Predicting frequent attenders to urgent care centres using machine learning

Lara Sophia Camille Chammas

Abstract

Reducing frequent attendance to urgent care centres, general practices, and emergency departments is of great importance to healthcare systems around the world. While frequent attenders are only a small proportion of the population, they represent a disproportionate number of visits. Using admission data from two West London urgent care centres, a supervised machine learning model was created to classify patients as either a frequent or non-frequent attender. Frequent attenders are defined as patients who attend the urgent care centre four or more times a year. The best model had a 61% accuracy rate and a 73% sensitivity rate. Taking the model's four most important variables, a logistic regression model was generated to predict the probability of a patient becoming a frequent attender. The regression model was based on the patient's ethnicity, age, gender, and income quintile. Probability of being a frequent attender rises with age, and decreases with an increase in income. Females are more likely to be frequent attenders until age 48, when males become the most likely to be frequent attenders.

Contents

1 Introduction

1.1 The problem of frequent attenders

Emergency departments are facing a worrying trend: a large proportion of their visits originate from a small proportion of patients. These patients are referred to as *frequent attenders*. Studies show that frequent attenders represent approximately 4% - 8% of the patient population, but account for 21 - 28% of visits (LaCalle and Rabin, 2010). Repeat attendance to emergency departments are considered unnecessary by medical experts and incur large costs as they are more expensive to maintain than primary care facilities (Buja et al., 2015). Emergency department frequent attendance has become a such a large issue, that the Royal College of Emergency Medicine has issued a "Best Practice Guideline" for managing frequent attenders. It states that "there should be a process of identifying frequent attenders in all Emergency Departments, in order to enable implementation [of the management guidelines]" (Hayhurst et al., 2017). Many hospitals, general practices, and other healthcare institutions have conducted retrospective analysis in-order to understand both the characteristics of the frequent attender population and the possible causes for their multiple visits (Lucas and Sanford, 1998).

There is no universally agreed attendance profile which defines frequent attenders. This lack of definition has resulted in variance by paper in deciding the number of visits required for a patient to be recognized as a frequent attender. Palmer et al. (2014) and Lucas and Sanford (1998) define the cut-off point for frequent attendance as patients who attend four or more times per year. Moore et al. (2009) defines the cut-off as patients who attend 10 or more times per year. Savageau et al. (2006) utilizes multiple definitions in order to analyse the differences between patients with 1, 2, 3, 4, 5-9, and 10+ visits per year.

The characteristics of frequent attenders are not universal, but there are some common features which regularly appear in the literature. Frequent attenders tend to be older than non-frequent attenders, and females disproportionately attend more often than males (Buja et al., 2015; Locker et al., 2007; Neal et al., 1998). In the US, people with low incomes were more likely to be frequent attenders (Savageau et al., 2006; Xu et al., 2009).

1.2 Aims

Most of the current research into frequent attenders is retrospective, aiming to describe the characteristics of the population. In this project, I take a new approach to frequent attender analysis. First, I create a supervised machine learning model that can classify patients as either frequent or non-frequent attenders. Second, I create a logistic regression model, using the assumptions of the machine learning model, to predict the probability of a patient becoming a frequent attender based on their age, gender, ethnicity, and income quintile. Classification systems and predictive models have not been used before in frequent attendance studies, but show promise as a tool to help hospitals quickly identify potential frequent attenders for more effective patient management.

1.3 What is machine learning?

Machine learning is a statistical and computer science technique which utilizes a computer's ability to find hidden structures and patterns within data (Alpaydin, 2014). There are two main types of machine learning: supervised and unsupervised learning.

Supervised learning is the training of a machine to predict the outcomes of independent variables. The machine is given both independent and dependent variables to create an algorithm that defines their relationship. Once the algorithm is built, the machine can generate the outcomes of different data-sets. There are two types of supervised learning techniques: classification and regression. Classification is used for discrete outcomes, for example labelling emails as normal or spam. Regression is used for continuous outcomes, for example predicting housing prices. Unsupervised learning involves feeding only independent variables (without pre-defined outcomes) to machine learning algorithms for data exploratory purposes; that is, to see how a machine might classify the unlabelled variables. Unsupervised learning techniques include clustering algorithms and neural networks (Bastanlar and Ozuysal, 2014; Deo, 2015; Vanschoren et al., 2013).This project utilizes only supervised machine learning.

Every supervised algorithm has its unique properties, but the process underlying supervised machine learning remains the same. The first step is to partition the data into a training and a testing set. The training set is then fed to the machine as a basis for model

building. Once the model is built, the testing set is given to the model to make predictions on. Predictions are then compared to the actual outcomes of the data to deduce the model's accuracy. As there is no "best algorithm", machine learning typically involves testing, tinkering, and tuning many different algorithms to find the one with the highest predictive performance (Bastanlar and Ozuysal, 2014).

Machine learning is quickly being adopted by many industries and fields. Classification is used by online retailers for product recommendations and by email providers to detect emails as spam (Kim et al., 2001; Chakraborty and Mondal, 2012). Geneticists use unsupervised techniques to annotate genomes (Hoffman et al., 2012). Plant biologists have also started employing machine learning to stream-line data analysis, such as high throughput stress phenotyping (Singh et al., 2016). In this paper, I apply the supervised machine learning algorithm random forests to the field of healthcare in-order to classify patients as frequent or non-frequent attenders.

2 Materials and Methods

2.1 Data preparation

I obtained a data-set comprised of 225,826 recorded visits by 126,779 unique patients to the urgent care centres (UCC) of Hammersmith Hospital and Charring Cross Hospital between November 2009 and December 2012. The data for each visit consists of 19 fields. Four fields refer to the visit itself: arrival date, arrival time, visit episode ID, and hospital site ID. Eight fields refer to the patient: patient ID, age, gender, ethnicity, if the patient is registered to a general practitioner (GP), if the patient is registered with the primary care centre at the hospital, the patient's Index of Multiple Deprivation, and their income quintile. The Index of Multiple Deprivation (IMD) is a measure of relative deprivation for small areas in England. It ranks each post code from 1 to 32,844. 1 is the most deprived area and 32,844 is the least deprived area (Department for Communities and Local Government, 2015). The income quintile was generated by data analysts at Charring Cross hospital by dividing the IMD into five equal sized groups. 1 represents the most deprived 20% of people while 5 represents the least deprived 20%. It is used as a proxy for income level and is assigned to

each patient based on their post-code.

The remaining seven fields refer to the clinical diagnosis of the patient. They are: triage stream, if the patient was referred to a specialist during the visit (and if so which one), the outcome of the visit (e.g. discharged home or sent to emergency department), the NHS standardized clinical code (e.g. M111), a clinical code descriptor (e.g. eczema), the broader category of the clinical code (e.g. skin/tissue disease), and the clinical code type (e.g. diagnosis).

I began my analysis by cleaning the data in several steps. First, I simplified the variable levels for patient ethnicity by reducing the number of ethnicity levels from 18 to six through combination. For example, combining "Asian British Bangladeshi" and "Asian British Indian" into "Asian". I then reduced the number of CCD levels from 25 to 15, for example by combining "other therapeutic procedures" and "operations, procedures, sites" into the category of "procedures". I removed any patients with missing personal information. The removal of these patients resulted in the loss of only 0.8% of visits and 0.5% of patients. I then added an additional column to the data-set titled "Index", which is the visit number that the admission represented for the patient.

I created two different data-sets to use during model building. The first data-set reflects the first visit of each patient. The second data-set reflects the second visit of each patient. In each data-set, each patient is represented only once. This ensures that when the data is partitioned into a training and testing set that there is no patient overlap. This prevents the model's accuracy from being artificially inflated, and makes sure that the outcome reflects the model's true predictive power on unseen data. As not all patients have a second visit, the second data-set is smaller than the first data-set. The first data-set contains all 126,086 patients while the second data-set consists of only 45,291 patients. I created both data-sets due to my hypothesis that the difference in clinical fields between visits may be a predictor of frequent attendance and enhance the predictive power of the model.

2.2 Classification of patients

I classified each patient by his or her number of visits. To do this I used four classification schemes. The first scheme, classification 1, classified each patient as either frequent

attender or non-frequent attender based on their average attendance rate over the three years. Patients who attended on less than four times per year were classified as non-frequent attenders while patients who attended four or more times a year were classified as frequent attenders. The second scheme, classification 2, classified patients as non-returners if they had a total visit count of one and as returners if they had a total visit count greater than one. The third scheme, classification 3, gave patients one of three different classifications based on their average attendance rate. Patients were classified as non-frequent returners if they attended on average less than one-time, occasional returns if they attended on average one to three times, and frequent returners if they attended on average four or more times a year. The fourth scheme, or classification 4, classified patients as non-frequent attenders if they attended on average less than three times a year and classified patients as frequent attenders if they attended on average three or more times a year.

I used four different classification schemes because machine learning requires defining the dependent variable differently to generate models with the highest predictive power. As there is no universally agreed upon definition for frequent attender, using different classification schemes allows for me to see if there was higher predictive ability for different attendance behaviours. I used four as the initial cut-off rate because it was the most commonly used definition in the literature (Buja et al., 2015). After initial model building using the first three schemes, I tried using a cut-off attendance rate of three times (classification 4) to increase the population size of frequent attenders to see if it would improve model accuracy.

2.3 Choosing a machine learning algorithm

The machine learning algorithm I chose was random forests because they are powerful and easy to use. Random forests are a supervised machine learning technique which can be used for classification or regression. Random forests are an ensemble algorithm because they use the median (regression) or mode (classification) prediction of multiple decision tree models to make predictions. The first step in classification by random forests is to build decision trees. To build a tree, the algorithm randomly samples with replacement the training data-set, generating a subset of data for tree building. At each node of the tree, it randomly selects a

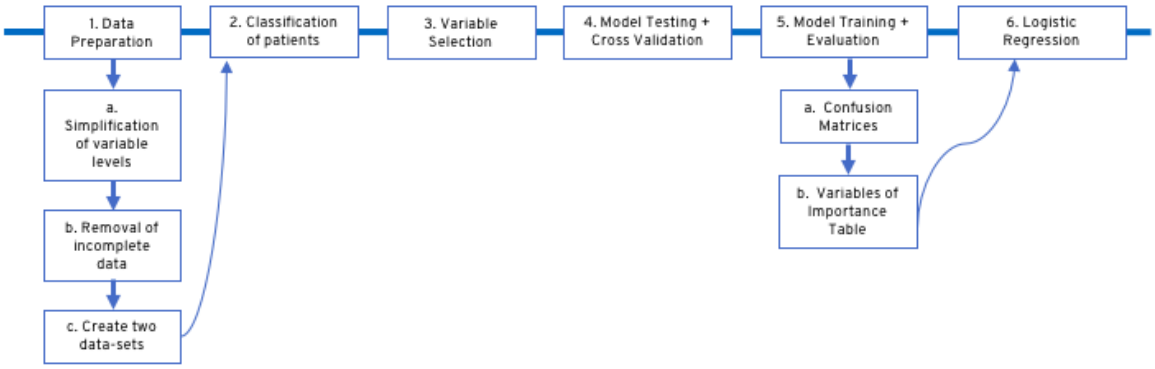


Figure 1: A flow chart diagram of the main steps of my methods.

certain number of predictor variables from all input predictor variables. From this selection, the machine uses the variable with the most predictive power and creates a binary split in the node. At the next node, another set of predictor variables is randomly selected from all the predictor variables and the binary split is created again using the best variable. This continues until the tree generates a classification. This iterative process results in a large number of trees that are independent and built using different subsamples of the training data and predictor variables. This technique has many names including bagging, bootstrap aggregating, or out-of-the-bag estimation. It helps to reduce the variance in the data by averaging many models, as well as prevent overfitting by the model. Once all trees are built and have generated a prediction, the classification with the majority vote becomes the "final prediction" classification. For example, if 100 trees vote frequent attender and 50 trees vote non-frequent, then the final predicted classification is frequent attender (Breiman, 2001; Friedman et al., 2001; Louppe, 2014).

Random forests are simple and effective because they require minimal data preparation before model training; they can handle non-normalized data, multicollinearity between variables, and are relatively robust to outliers and noise. They can also use both numerical and categorical variables, thus removing the need to create dummy variables for categorical variables. I also chose to use random forests because they generate an interpretable estimation and ranking of the model's important variables which provides insight into how it works.

2.4 Variable Selection

Preliminary model testing showed that using 11 out of the 19 independent variables gave the highest model accuracy. The final selected variables were patient age, gender, ethnicity, income quintile, clinical code descriptor, clinical code type, triage stream, GP registration status, hospital primary care registration status, referred specialty group, and visit outcome.

2.5 Model Training and Testing

Using the machine learning wrapper `caret` in `R`, the random forest models were trained and tested (R Core Team, 2013; Kuhn, 2008; Liaw and Wiener, 2002). Before training, I partitioned each data-set into a training (75%) and testing (25%) set, as per the standard partition rule. The model was trained using stratified 10-fold cross validation repeated 3-times. This means that the training data-set is divided into 10 subsamples (folds) of equal size. This division is stratified by classification. Stratification ensures that each fold has the same proportion of classification labels. During training, one fold is removed and held as the validation data, while the other nine folds are used as training data. This procedure is then repeated 10 times, using each subsample exactly once as the validation data. Once the first round of cross-validation is complete, it is repeated twice more for a total of three repeats. The mode prediction is then selected as the final classification. K-fold cross validation is a standard for many machine learning algorithms because it enhances model accuracy by training and then validation the model on all of the training data. To deal with the imbalance between my classifications, I employed down-sampling during cross-validation training. Down-sampling fixes class imbalances by randomly sampling the majority class to match the size of the minority class. This forces the model to see both classes equally during training, preventing the formation of a classification accuracy bias (Vanschoren et al., 2013).

To test the power of random forests in predicting frequent attendance, I explored eight different models. Each model was a different combination of classification schema and data-sets. Table 1 shows these combinations. After training each of the eight models on the training data-set, I tested their efficacy by inputting the unseen testing data-set. This gives an accurate estimate of the model's classification accuracy. Next, I investigated how each variable affected the model by using logistic regression. This was carried out in `R` using a

generalized linear model with a binomial family distribution and the `R` package `effects` to produce graphs of the results (R Core Team, 2013; Fox, 2003).

Model Name	**Model Description**
Model 1A	Predicting classification scheme 1 using the first visit
Model 1B	Predicting classification scheme 1 using the second visit
Model 2	Predicting classification scheme 2 using the first visit
Model 3A	Predicting classification scheme 3 using the first visit
Model 3B	Predicting classification scheme 3 using the second visit
Model 4	Predicting occasional & frequent returners from classification scheme 3 using the second visit
Model 5A	Predicting classification scheme 4 using the first visit
Model 5B	Predicing classification scheme 4 using the second visit

Table 1: A table defining which classification scheme and data-set each of the eight models used.

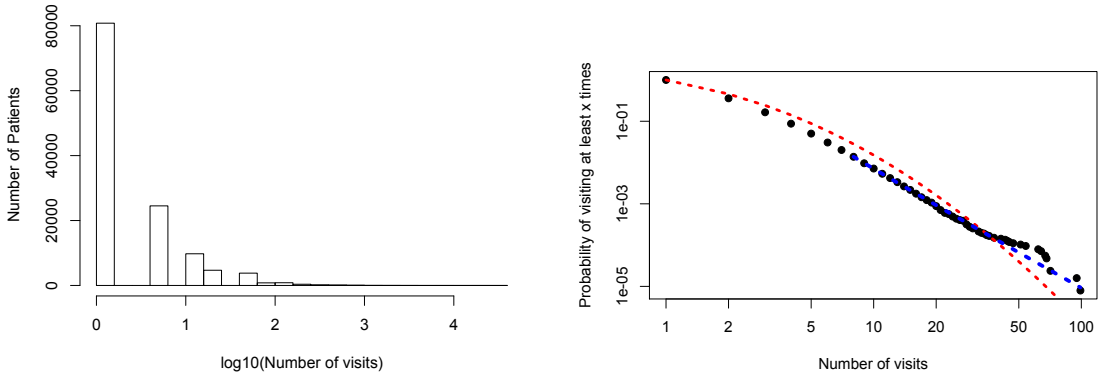
3 Results

3.1 Frequent attenders make up 0.42% of patients but 4.2% of visits

A total of 126,086 patients attended the UCCs between November 2009 and December 2012 resulting in 223,945 visits. Of the total visits, 56% were returns. Using the most common definition of frequent attenders (classification scheme 1), 0.42% of the patients are frequent attenders, as defined by an average attendance rate of four or more times per year. They are responsible for 4.2% of the visits to the urgent care centers. 1.3% of patients attended on average three or fewer times per year, creating 5.6% of visits. 6.7% of patients attended on average two or fewer times per year, creating 17.4% of visits. 91.3% of patients went on average one or fewer times per year, creating 71% of visits. 64% of patients had an attendance total of one over the three years. The highest total count of UCC visits by one patient over the three years was 99 times.

To investigate the distribution of attendance rates, I plotted a histogram of the log of total visit count versus the frequency of patients with that visit count. The distribution of attendance is clearly heavy tailed, or right skewed, as shown in Figure 2a. This distribution indicates that patient attendance may follow a power law distribution relationship (Mitzenmacher, 2004). To investigate further, I plotted the distribution as a complementary cumulative distribution function (CCDF), as shown in Figure 2b in black. The CCDF, however, suggests that it may be either a power law distribution or a

log-normal distribution. This is because while the CCDF of a power law distribution will appear as a straight line, the CCDF of a log-normal distribution may also appear nearly straight for a large proportion of its distribution (Newman, 2005). As the graph has a curvature towards the bottom right, I tested both distributions for their fit to the data using the `r` package `powerRlaw` (R Core Team, 2013; Gillespie, 2015; Prado et al., 2016). As shown in Figure 2b in blue, the power-law distribution gives an excellent fit to the right-hand tail of the distribution ($\geq$ 8 visits), but a poor fit to the left-hand tail. The log-normal distribution gives a better fit over all, as shown in red. Both power-laws and log-normal distributions can be the result of multiplicative processes. For my data, this would mean that the probability of a patient returning is independent of how many times he or she has returned before (Newman, 2005; Mitzenmacher, 2004; Limpert et al., 2001). However, the data suggests instead that the processes that cause patients to return more than eight times is different than the processes that cause them to return fewer.



(a) (b)

Figure 2: (a) Histogram of the logarithm of the total visit count to the urgent care centers versus the number of patients with that visit count. (b) Complementary cumulative frequency distribution of the frequency of visit count is plotted in black. A fitted power law distribution is plotted in blue. A fitted log-normal distribution is plotted in red. The power-law distribution gives an excellent fit to the right-hand tail with a visit count greater than eight. The log-normal distribution gives a better fit over all.

3.2 Random forests can modestly predict frequent attenders

To test the power of random forests to predict frequent attendance, I explored eight different models. Each model was a different combination of classification schema and data-sets. Table

1 shows these combinations. After model testing, Model 1A and Model 1B had the highest accuracy rates at 61% and 62% respectively, as shown in Table 2. Table 3 shows the confusion matrixes for Model 1A and Model 1B. Confusion matrices shows the number of predicted classifications for each data-point by the model as compared to the reference classification of the data-point. From the confusion matrix, the model's accuracy, sensitivity, specificity and other related metrics can be calculated. These equations are shown in Appendix A2.

Between the two models, I selected Model 1A as the better model because it had a higher sensitivity rate of 73%, compared to Model 1B's sensitivity of 53%. Sensitivity, or the "true positive rate", was selected as the main accuracy metric as the aim of this project is to predict the positive class of frequent attenders to UCC. This priorities the accuracy of predicting the positive class over predicting the negative class. However, Model 1A's overall accuracy rate of 61% is very modest; it represents only 11% greater accuracy than model prediction by random chance.

Model 1A Results	
Accuracy	61%
95% Confidence Intervals	0.60, 0.61
Sensitivity	73%
Specificity	61%
False Positive Rate	39%
Balanced Accuracy	67%

(a) Model 1A

Model 1B Results	
Accuracy	62%
95% Confidence Intervals	0.61, 0.63
Sensitivity	59%
Specificity	62%
False Positive Rate	38%
Balanced Accuracy	61%

(b) Model 1B

Table 2: The results of Model 1A and Model 1B on classification of patients in the unseen testing data set. Sensitivity is the true positive classification rate. Specificity is the true negative classification rate. Balanced accuracy is the predicted accuracy rate for the model if the classes were balanced in the dataset. The equations for these metrics can be found in Appendix A.2.

3.3 Patient age, gender, income quintile and ethnicity are the most important classifiers of frequent attendance

Random forest algorithms produce a "variables of importance" table after model training. This table ranks the variables by their mean decreasing Gini impurity (MDG), showing the most importance variables for the model when it is making predictions. The Gini impurity is the sum of the probability of an item (p_i) being chosen multiplied by the probability of

		Reference	
		Frequent	Non-Frequent
Prediction	Frequent	97	12,374
	Non-Frequent	35	19,015

(a) Model 1A

		Reference	
		Frequent	Non-Frequent
Prediction	Frequent	78	4,224
	Non-Frequent	54	6,966

(b) Model 1B

Table 3: The confusion matrix shows the model's classification results. The rows represent the model's class predictions while the columns represent the reference classes. The top right corner represents the count of true positive classifications and the bottom left corner represents the count of true negative classifications. The top left corner represents the count of false positives and the bottom right corner represents the count of false negatives. Frequent attender is the positive classification. A generalized confusion matrix can be found in Appendix A.2.

miscategorizing that item (p_k). The equation for the Gini impurity for a set of items (p) with J classes is $I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k$.

When each decision tree is built, the Gini impurity is used to decide the best variable split for each node. The best variable split is one which has the lowest Gini impurity. After each split, the new daughter nodes will have a decreased impurity value. Once the tree is built, all of the Gini impurity decreases are added for all the nodes which use the variable of interest. Then, the sum is averaged for all the trees in the forest to result in the MDG (Louppe, 2014). `Caret` then scales the mean decreasing Ginis for all the variables and ranks them as a percentage of the maximum MDG. Appendix Table A1 shows Model 1A's variables of importance. The top four variables of importance for Model 1A are patient age, gender, income quintile, and ethnicity.

3.4 Income quintile and the interaction of age and gender have an effect on the probability of being a frequent attender

Using the top four variables of importance from Model 1A — age, gender, ethnicity, and income quintile — I created a logistic regression model. I did this to understand the direct effects of each variable on the probability of being a frequent attender. The results of the logistic regression, shown in Table 4, also provides insights into how Model 1A classifies

frequent attenders.

The logistic regression modelled the interaction of age and gender plus the main effects of ethnicity and income quintile. Ethnicity had no statistical significance on the model. This means ethnicity does not influence the probability of a patient becoming a frequent attender. Income quintile and the interaction of age and gender did have a statistical significance, meaning that changes to these variables will affect the probability of a patient becoming a frequent attender. The effects of income quintile and the interaction of age and gender on the probability of being a frequent attender are plotted in Figure 3.

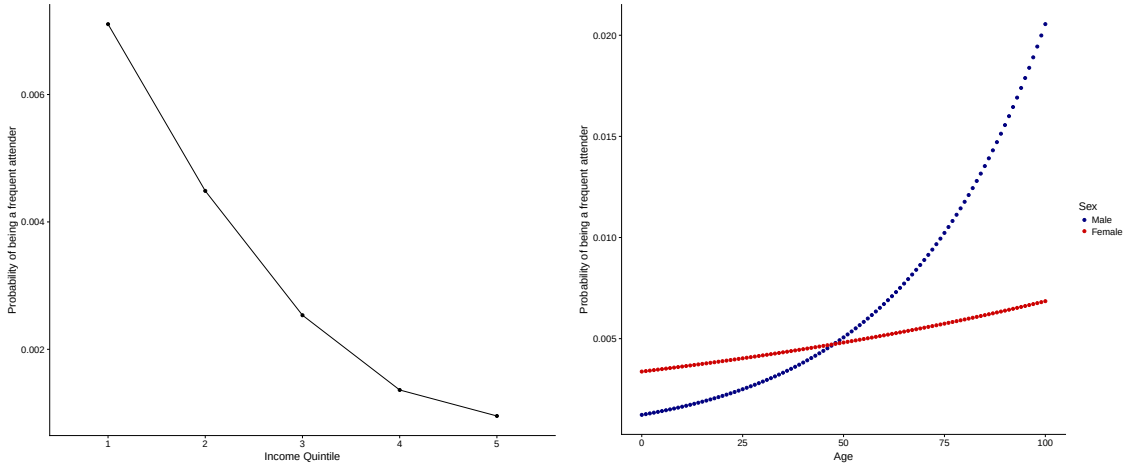
Variable	**Coefficient Estimate**	**Pr(>|z|)**
Intercept	-5.176	$<$ 2e-16***
Age	0.008	0.009**
Gender:Male	-1.003	3.08e-06***
Ethnicity: Black/African/Carribbean/Black British	0.166	0.355
Ethnicity: Mixed/Multiple Ethic Groups	0.138	0.619
Ethnicity: Not Stated	-0.679	0.009**
Ethnicity: Other Ethnic Group	-0.268	0.228
Ethnicity: White	0.022	0.888
Income Quintile: 2	-0.474	9.45e-07***
Income Quintile: 3	-1.060	3.25e-13***
Income Quintile: 4	-1.693	1.66e-14***
Income Quintile: 5	-2.040	2.87e-06***
Age*GenderMale	0.0213	2.87e-06***

Table 4: Logistic regression output showing the coefficents of each variable in the logistic model and the significance of the coefficient on the model output. The coefficient represents the change in the log odds ratio of being a frequent attender with an increase of one unit of the independent variable of interest. For an example, if a patient has an income quintile of 2, their log odds of being a frequent attender decreases by 0.474 units than if they had an income quinitile of 1. Signicance codes: 0 '***' 0.001 '**' 0.01 '*'

3.5 Poor, old men are most likely to be frequent attenders

The income quintile with the highest probability of being a frequent attender is 1, which represents the lowest fifth, or 20%, of incomes. The probability of being a frequent attender then decreases as income rises. This is shown in 3A. For example, the probability of a 36 year old (patient population average age) male with an income quintile of 1 for being a frequent attender is 0.0064. If his quintile is 5, his probability of being a frequent attender decreases to 0.0010.

There is a large effect on the probability of being a frequent attender based on age and gender. As shown in Figure 3B, the probability of being a frequent attender has a slight rise with age in females. However, for males, the change in probability as age increases is almost exponential. Females have a higher probability of being a frequent attender than males until age 48, when the probability of a male being a frequent attender becomes the highest. This is shown in Figure 3A. For a 36 year old female patient, the probability of being a frequent attender is 0.0044. The probability of a 36 year old male being a frequent attender is lower at 0.0035. However, for a male at age 65, the probability of being a frequent attender rises sharply to 0.0074 and at age 80 rises again to 0.011. For females, the probability of being a frequent attender at age 65 is 0.0052 and at age 80 is 0.0057. This is a much smaller change in probability as compared to males. To further illustrate this effect, a male who is 80 years old and has an income quintile of 1 has a probability of 0.021 of being a frequent attender. The probability of a female of the same demographics being a frequent attender is lower at 0.011.



(a) The effect of income quintile on the probability of being a frequent attender

(b) The effect of age and gender on the probability of being a frequent attender

Figure 3: (a) and (b) are effector graphs which show the effect of a change in an independent variable on the probability of being a frequent attender. The graphs are generated by changing the variable of interested in the logistic regression while keep the remaining variables constant. (a) shows the effct of income quintile on the probability of being a frequent attender. (b) shows the effect of the interaction of age and gender on the probability of being a frequent attender. The blue line is males and the red line is females.

4 Discussion

Machine learning is used in many fields to find hidden structures in data for analysis and predictive modelling. Its applications include such diverse instances as Gmail's "smart reply" function to cancer diagnosis and prognosis prediction (Kannan et al., 2016; Kourou et al., 2015). To tackle the problem of frequent attendance to urgent care centres, I attempted to use machine learning to classify patients as frequent or non-frequent attenders based on their admission data. To my knowledge, this is the first-time machine learning is being applied to this problem.

The best classification model I generated had a 61% accuracy in classifying patients as frequent or non-frequent attenders from the data of their first visit. This accuracy is modest, and is too low for the model to be used in hospitals as a tool for pre-emptively identifying patients who may become frequent attenders. However, I have also found that age, gender and income quintile are important determinants for frequent attendance at the Urgent Care Centres at Hammersmith Hospital and Charring Cross Hospital. The typical frequent attender is most likely female, until age 48 when males become the most likely to be frequent attenders. Frequent attenders are generally older in age and have lower incomes. These characteristics match the characteristics fond by other previous works on frequent attender populations (Palmer et al., 2014; Buja et al., 2015; Neal et al., 1998).

A limitation of this study, and a likely cause of the poor model accuracy rate, is the small data size of frequent attenders. I only had 530 frequent attenders, compared to the 125,556 non-frequent attenders, as defined by an attendance rate of four or more times per year. In an attempt to increase the amount of data, I reduced the cut-off attendance rate from four visits to three in classification scheme 4. This did increase the number of frequent attenders from 530 to 1,210, but this did not improve the model's accuracy, as shown in Appendix Table A6. This indicates that an even larger data-set may be needed, as random forests become more accurate with more training data (Breiman, 2001; Friedman et al., 2001; Louppe, 2014). However, this could also mean that there might not be a "hidden pattern" in the data that classifies patients as frequent attenders or non-frequent attenders accurately: there might only be factors that increase or decrease a patient's probability of being a frequent attender. While I was able to identify these factors using a logistic regression, the highest probability

increase was 0.010. This is very small and points to the hypothesis that there may not be logical pattern behind frequent attendance. This would explain why it has been hard for hospitals to reduce their attendance rates (Pope et al., 2000; Phillips et al., 2006).

Another way to improve model accuracy is to collect more fields of information on patients and their visits to the UCC. While age, gender and income were identified as the most important predictors of frequent attendance, there could be more influential predictors not yet identified. Previous work on frequent attenders have studied the effect of drug and alcohol dependence, the existence of mental health issues, homelessness, and other social factors on attendance rates (Vedsted et al., 2004; Savageau et al., 2006). Model 1A's variables of importance table identified some potential areas of interest for future analysis. This includes being triaged to the minor injuries unit or being discharged home for a routine GP follow-up. Besides collecting more data and more predictors, other machine learning algorithms, such as neural nets, may have a higher predictive accuracy.

It is generally agreed that there two types of frequent attenders. The first type is those whose frequent attendance is constant and prolonged over a long period of time (e.g. years). The second type is those whose frequent attendance is periodic or short lived. Their frequent attendance is due to a specific medical or psychosocial event, causing a spike in attendance over a short time (Pope et al., 2000). Future work on frequent attenders might focus on comparing these two groups. I found that the probability of returning for a future visit is not easily modelled by a power law or log-normal distribution. A log-normal distribution fits the data overall, but a power-law distrbution fits the data with eight or more visits extremely well. This indicates that there may be different processes behind return rates that are less than eight and greater than eight. Further research might identify these processes and uncover the causes behind return visits. This would help find new ways to reduce repeat visits.

To decrease the attendance rates of frequent attenders, many hospitals and general practices have started to implement management plans. These range from personalized care plans for flagged individuals, to standardized plans for clinicians to follow during patient visits. The results of these studies have been mixed. Many have shown a reduction in the number of return visits by the managed patients, but others have not. (Pope et al.,

2000; Okin et al., 2000; Skinner et al., 2009; Phillips et al., 2006; Spillane et al., 2008). As each hospital implemented a different management plan, direct comparisons cannot be made, but it appears that interventions which focused on creating individual plans for patients by an interdisciplinary team involving both clinicians and social workers resulted in the most desirable outcomes (Pope et al., 2000; Okin et al., 2000; Bristow and Herrick, 2002). This supports the common notion that frequent attendance is created by a multitude of both social and physical problems which cannot be solved solely with treatment at urgent or emergency care (Ng et al., 2015; Spillane et al., 2008).

In 2012, West Middlesex University Hospital in collaboration with the West London Mental Health Trust created intensive, individual care management plans for seven of the hospital's most frequent attenders. These care plans focused on caring for the mental health of these patients and ensuring continuous and effective care, as opposed to reactive, sporadic care. West Middlesex University Hospital and West London Mental Hospital Trust found a large reduction in the attendance rates of all patients as well as large improvements in health and other wellbeing outcomes. Due to the success of these management plans, the project has been scaled to eight London boroughs to treat 108 patients. This has resulted in a total annual reduction of over 1000 emergency department attendances (Ng et al., 2015; West London Mental Health NHS Trust, 2015).

In conclusion, while my model did not classify frequent and non-frequent attenders with high accuracy, I did identify important variables for predicting return visits. Using these variables, I was able to model frequent attenders more accurately and better understand the characteristics of the frequent attender population. As this was the first-time machine learning has been applied to this problem, my work points the way for using machine learning techniques to predict patient behaviour. With improvements in data collection, a model could be developed for hospitals to use to streamline the identification of at-risk patients. This would free clinicians to focus on delivering effective care, rather than repeatedly treating the same maladies.

5 Acknowledgements

I would like to thank my supervisor Armand Leroi, for not only facilitating this dream research project, but for providing constructive guidance that has helped my project to reach its full potential. Thank you as well to Geva Greenfield for providing the data for me to analyze and the insights into the issue of frequent attenders to urgent care.

References

Alpaydin, Ethem (2014), *Introduction to machine learning*, MIT press.

Bastanlar, Y. and M. Ozuysal (2014), *Introduction to Machine Learning*, Vol. 1107 of *In: Yousef M., Allmer J. (eds) miRNomics: MicroRNA Biology and Computational Analysis. Methods in Molecular Biology (Methods and Protocols)*, Humana Press, Totowa, NJ, pp. 105–128.

Breiman, Leo (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Bristow, Darlene P. and Charlotte A. Herrick (2002), 'Emergency department case management: The dyad team of nurse case manager and social worker improve discharge planning and patient and staff satisfaction while decreasing inappropriate admissions and costs: A literature review', *Professional Case Management* **7**(6).

Buja, Alessandra, Roberto Toffanin, Stefano Rigon, Camilla Lion, Paolo Sandonà, Daniela Carraro, Gianfranco Damiani and Vincenzo Baldo (2015), 'What determines frequent attendance at out-of-hours primary care services?', *European Journal of Public Health* **25**(4), 563–568.

Chakraborty, Sarit and Bikromadittya Mondal (2012), 'Spam mail filtering technique using different decision tree classifiers through data mining approach-a comparative performance analysis', *International Journal of Computer Applications* **47**(16).

Deo, Rahul C (2015), 'Machine learning in medicine', *Circulation* **132**(20), 1920–1930.

Department for Communities and Local Government (2015), The english index of multiple

deprivation (IMD) 2015 - guidance, Technical report, Department for Communities and Local Government.

Fox, John (2003), 'Effect displays in r for generalised linear models', *Journal of Statistical Software, Articles* **8**(15), 1–27.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani (2001), *The elements of statistical learning*, Vol. 1, Springer Series in Statistics New York.

Gillespie, Colin S. (2015), 'Fitting heavy tailed distributions: The poweRlaw package', *Journal of Statstical Software* **64**(2), 1–16.

Hayhurst, Catherine, Simon M. Smith and Duncan Chambers (2017), Frequent attenders in the emergency department, Best practice guideline, The Royal College of Emergency Medicine.

Hoffman, Michael M, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes and William Stafford Noble (2012), 'Unsupervised pattern discovery in human chromatin structure through genomic segmentation', *Nature methods* **9**(5), 473–476.

Kannan, Anjuli, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young and Vivek Ramavajjala (2016), Smart reply: Automated response suggestion for email, *in* 'Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM, New York, NY, USA, pp. 955–964.

Kim, Jong Woo, Byung Hun Lee, Michael J. Shaw, Hsin-Lu Chang and Matthew Nelson (2001), 'Application of decision-tree induction techniques to personalized advertisements on internet storefronts', *International Journal of Electronic Commerce* **5**(3), 45–62.

Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadis (2015), 'Machine learning applications in cancer prognosis and prediction', *Computational and Structural Biotechnology Journal* **13**, 8 – 17.

Kuhn, Max (2008), 'Building predictive models in r using the caret package', *Journal of Statistical Software, Articles* **28**(5), 1–26.

LaCalle, Eduardo and Elaine Rabin (2010), 'Frequent users of emergency departments: The myths, the data, and the policy implications', *Annals of Emergency Medicine* **56**(1), 42–48.

Liaw, Andy and Matthew Wiener (2002), 'Classification and regression by randomforest', *R News* **2**(3), 18–22.

Limpert, Eckhard, Werner A Stahel and Markus Abbt (2001), 'Log-normal distributions across the sciences: Keys and clues', *AIBS Bulletin* **51**(5), 341–352.

Locker, Thomas E, Simon Baston, Suzanne M Mason and Jon Nicholl (2007), 'Defining frequent use of an urban emergency department', *Emergency Medicine Journal* **24**(6), 398–401.

Louppe, Gilles (2014), Understanding Random Forests: From Theory to Practice, PhD thesis, University of Liège.

Lucas, Raymond H and Sandra M Sanford (1998), 'An analysis of frequent users of emergency care at an urban university hospital', *Annals of Emergency Medicine* **32**(5), 563–568.

Mitzenmacher, Michael (2004), 'A brief history of generative models for power law and lognormal distributions', *Internet mathematics* **1**(2), 226–251.

Moore, L, A Deehan, P Seed and R Jones (2009), 'Characteristics of frequent attenders in an emergency department: analysis of 1-year attendance data', *Emergency Medicine Journal* **26**(4), 263.

Neal, R. D., P. L. Heywood, S. Morley, A. D. Clayden and A. C. Dowell (1998), 'Frequency of patients' consulting in general practice and workload generated by frequent attenders: comparisons between practices.', *British Journal of General Practice* **48**(426), 895–898.

Newman, MEJ (2005), 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**(5), 323–351.

Ng, Audrey, Vivek Nadarajan, Sian McIver, Catriona Reid, Emma Schofield and Amrit Sachar (2015), 'Frequent attendances to a London emergency department: a service improvement project embedding mental health into the team', *London Journal of Primary Care* **7**(4), 70–77.

Okin, Robert L., Alicia Boccellari, Francisca Azocar, Martha Shumway, Kathy O'Brien, Alan Gelb, Michael Kohn, Phyllis Harding and Christine Wachsmuth (2000), 'The effects of clinical case management on hospital service use among ED frequent users', *The American Journal of Emergency Medicine* **18**(5), 603–608.

Palmer, Erin, Denise Leblanc-Duchin, Joshua Murray and Paul Atkinson (2014), 'Emergency department use', *Canadian Family Physician* **60**(4), e223–e229.

Phillips, Georgina Ann, David S. Brophy, Tracey J. Weiland, Antony J. Chenhall and Andrew W. Dent (2006), 'The effect of multidisciplinary case mangement on selected outcomes for frequent attenders at an emergency department', *The Medical Journal of Australia* **184**(12), 602–6006.

Pope, Dorothy, Christopher M. B. Fernandes, France Bouthillette and Jeremy Etherington (2000), 'Frequent users of the emergency department: a program to improve care and reduce visits', *Canadian Medical Association Journal* **162**(7), 1017.

Prado, Paulo Inácio, Murilo Dantas Miranda and Andre Chalom (2016), 'Fitting species abundance models with maximum likelihood quick reference for sads package'.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Savageau, Judith A, Martha McLoughlin, Alina Ursan, Yan Bai, Matthew Collins and Suzanne B Cashman (2006), 'Characteristics of frequent attenders at a community health center', *The Journal of the American Board of Family Medicine* **19**(3), 265–275.

Singh, Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh and Soumik Sarkar (2016), 'Machine learning for high-throughput stress phenotyping in plants', *Trends in Plant Science* **21**(2), 110–124.

Skinner, J, L Carter and C Haxton (2009), 'Case management of patients who frequently present to a Scottish emergency department', *Emergency Medicine Journal* **26**(2), 103.

Spillane, Linda L., Eileen W. Lumb, Daniel J. Cobaugh, Susan Riley Wilcox, John S. Clark and Sandra M. Schneider (2008), 'Frequent users of the emergency department: Can we intervene?', *Academic Emergency Medicine* **4**(6), 574–580.

Vanschoren, Joaquin, Jan N. van Rijn, Bernd Bischl and Luis Torgo (2013), 'OpenML: Networked science in machine learning', *SIGKDD Explorations* **15**(2), 49–60.

Vedsted, Peter, Per Fink, Henrik Toft Sørensen and Frede Olesen (2004), 'Physical, mental and social factors associated with frequent attendance in Danish general practice. a population-based cross-sectional study', *Social Science & Medicine* **59**(4), 813–823.

West London Mental Health NHS Trust (2015), 'A&E programme reduces unnecessary attendances', online.

Xu, K. Tom, Brian K. Nelson and Steven Berk (2009), 'The changing profile of patients who used emergency department services in the United States: 1996 to 2005', *Annals of Emergency Medicine* **54**(6), 805–810.e7.

Appendix A Appendix

A.1 Model 1A's variables of importance

Variables of Importance	**Scaled MeanDecreasingGini**
Age	100.000
Gender:Female	13.956
Quintile:2	12.509
Ethnicity: White	12.210
Quintile:3	11.317
PfHReg: Registered	10.141
Stream: Minor Injuries	9.652
Outcome: Discharged Homt; routine GP follow up	9.233
Ethnicity: Black/African/Carribbean/Black British	8.787
Quintile:4	8.349
Clinical Code: Injury and Poisoning	7.545
Clinical Code Type: Diagnosis	7.494
Stream: See and Treet	7.435
Clinical Code: Sympton Investigation	7.181
Stream: GP Priority	6.903
Outcomes: Emergency Department	6.630
Ethnicity: Other Ethnic Group	6.166
Clinical Code Type: Process of Care	6.156
Clinical Code: Musculoskeletal/Connective Tissue	5.894

Table A1: Model 1A's variables of importance ranked by mean decrease gini which is scaled to age.

A.2 Confusion Matrixes and Accuracy Metrics

		Reference		
		Frequent	Non-Frequent	
Prediction	Frequent	True Positive	False Positive	**Total Predicted Positive**
	Non-Frequent	False Negative	True Negative	**Total Predicted Negative**
		Total Positive	**Total Negative**	

Table A2: A generalized confusion matrix showing what each cell means. Reference is the predifined classification given to the data-points. Prediction is the classification given to the data-point by the model. True positives are positive classifications that the model predicted correctly. False positives are positive classifications that the model predicted incorrectly because the data was actually negative. True negatives are negative classifications that the model predicted correctly. False negatives are negative classifications that the model predicted incorrectly because the data was actually classified as positive.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Positive + Total\ Negative}$$

$$Sensitivity = \frac{True\ Positive}{Total\ Positive}$$

$$Specificity = \frac{True\ Negative}{Total\ Negative}$$

$$FalsePositive = \frac{False\ Positive}{Total\ Negative}$$

$$BalancedAccuracy = \frac{\frac{True\ Positive}{Total\ Positive} + \frac{True\ Negative}{Total\ Negative}}{2}$$

A.3 Results of Model Building

Model 2 Results	
Accuracy	54%
95% Confidence Intervals	0.53, 0.54
Sensitivity	64%
Specificity	48%
False Positive Rate	52%
Balanced Accuracy	56%

Table A3: Model 2 results on predicting classification scheme two, using the first visit, on the unseen testing dataset. Returner is the positive classification and non-returner is the negative classification. Sensitivity is the true positive classification rate. Specificity is the true negative classification rate. Balanced accuracy is the predicted accuracy rate for the model if the classes were balanced in the dataset.

Model 3A Results	
Accuracy	42%
95% Confidence Intervals	0.40, 0.43
Sensitivity	44%
Specificity	73%
False Positive Rate	27%
Balanced Accuracy	59%

(a) Model 3A

Model 3B Results	
Accuracy	44%
95% Confidence Intervals	0.43, 0.44
Sensitivity	47%
Specificity	74%
False Positive Rate	26%
Balanced Accuracy	61%

(b) Model 3B

Table A4: (a) Model 3A's results on predicting classification scheme three, using the first visit, on the unseen testing dataset. Frequent attender is the positive classification as compared to ocassionally frequent and non-frequent attender. (b) Model 3B's results on predicting classification scheme three, using the second visit, on the unseen testing dataset. Frequent attender is the positive classification as compared to ocassionally frequent and non-frequent attender. Sensitivity is the true positive classification rate. Specificity is the true negative classification rate. Balanced accuracy is the predicted accuracy rate for the model if the classes were balanced in the dataset.

Model 4 Results	
Accuracy	48%
95% Confidence Intervals	0.46, 0.50
Sensitivity	70%
Specificity	47%
False Positive Rate	53%
Balanced Accuracy	58%

Table A5: Model 4's results on predicting classification scheme three, using the second visit, on the unseen testing dataset. Frequent attender is the positive classification and ocassionally frequent attender is the negative classification. Sensitivity is the true positive classification rate. Specificity is the true negative classification rate. Balanced accuracy is the predicted accuracy rate for the model if the classes were balanced in the dataset.

Model 5A Results	
Accuracy	60%
95% Confidence Intervals	0.59, 0.60
Sensitivity	59%
Specificity	60%
False Positive Rate	40%
Balanced Accuracy	60%

(a) Model 5A

Model 5B Results	
Accuracy	58%
95% Confidence Intervals	0.57, 0.59
Sensitivity	58%
Specificity	60%
False Positive Rate	40%
Balanced Accuracy	59%

(b) Model 5B

Table A6: (a) Model 5A's results on predicting classification scheme four, using the first visit, on the unseen testing dataset. (b) Model 5B's results on predicting classification scheme four, using the second visit, on the unseen testing dataset. Frequent attender is the positive classification and non-frequent attender is the negative classification. Sensitivity is the true positive classification rate. Specificity is the true negative classification rate. Balanced accuracy is the predicted accuracy rate for the model if the classes were balanced in the dataset.