Predicting Individuals at Risk of Developing Mental Health Conditions in Barking and Dagenham

[redacted]

Course: MSc Business Intelligence and Analytics

Supervisor: Professor Thierry Chaussalet

Table of Contents

Aims

The purpose of the project is to predict which individuals living in the London Borough of Barking and Dagenham are likely to develop a mental health problem in the future.

This will be achieved by using techniques commonly referred to use as ‘risk stratification’ or ‘risk prediction’. Specifically, risk stratification is used to ‘case find individuals and develop proactive approaches to intervention’. It does this by using ‘patients’ past characteristics to make predictions about future risk’ (NHS England 2017). Ultimately, such tools are based on statistical and machine learning models, typically using regression techniques, but increasingly using more modern and sophisticated methods such as neural networks (NHS England 2017). As part of this project a range of different techniques will be tried and compared, covering traditional statistical methods and state of the art machine learning.

Being able to make accurate predictions about these outcomes can have significant benefits for the individuals affected, the health and social care organisations who treat them, and the professionals working in the field. One reason for this is because conditions such as depression and anxiety are often linked to other health problems and comorbidities (Bhakta, Sau 2017). Risk stratification methods are useful in this context since they ‘allow for individual prediction, early diagnosis and treatment selection’ (Bzdok, Meyer-Lindenberg 2018). As such, proactively identifying individuals who are at risk of developing these conditions may mean that preventative action can be taken to reduce or even eliminate the risk. Moreover, it may also help to reduce the chances that these individuals go on to develop other potentially serious health conditions in the future.

Such an approach can help to facilitate the model of early intervention that is widely supported across the NHS, local government and other community providers. Successfully implementing this approach would have significant benefits for patients by improving their quality of life and reducing their future risk of harm, while also benefitting the health and social care organisations by reducing overall levels of demand and cost. Nevertheless, it is important to note that risk stratification tools are not a replacement for professional judgement. Instead they are a tool to help guide professionals to be more targeted, and to provide insight that supports the decision-making process.

The dataset to be used has been developed as part of a project led by Care City London, an innovation centre for healthy ageing that works across four London boroughs including Barking and Dagenham. The dataset combines and matches information from various sources about over 115,000 individuals living in Barking and Dagenham, covering primary care activity, prescription data, hospital activity data, community mental health services and social care services. It includes information about a range of health conditions, both physical and mental, as well as certain demographic and socioeconomic characteristics. The breadth of variables available in the dataset and the large sample size make it a very suitable dataset to use for this kind of prediction problem.

Objectives

There are four overarching objectives which together will realise the overall project aim.

The primary objective of the project is to build a model which can make predictions at an individual level. The target variable being predicted will be the risk of any given individual developing a mental health condition at some point in the future. The problem proposed can be considered as either a binary classification problem, or a probability estimation problem. For this reason, several models will be trained and tested to explore both possible outputs. This is likely to include models such as logistic regression, CART and a random forest. These models will be built in either R or Python.

All the models will be compared for accuracy to determine which one is the most suitable for the purpose of prediction. Therefore, one of the key objectives will be to maximise the performance of the models (in terms of accuracy, error, precision, sensitivity and specificity). This will be achieved in several ways, not limited to feature engineering and parameter tuning. The performance of all the models will be evaluated on the testing data, and the most suitable model will be chosen. Using the chosen model, a prediction will be made for all the individuals in the dataset.

A third objective of the project is to communicate the results of these predictions to mental health and social care practitioners in Barking and Dagenham Council and the NHS, so that they can be discussed as part of case meetings and used to inform decisions about treatments and interventions. The information will be passed to the relevant professionals either at the individual level or in a summarised format, depending on the specific requirements of the stakeholders and restrictions placed by information governance rules. The information will need to be easy to interpret and provide the correct level of detail required by the users.

The final objective of the project is to compare the utility of traditional parametric models used in statistics (e.g. logistic regression) to more modern nonparametric machine learning models (e.g. a decision tree or random forest) for the purposes of risk stratification and risk prediction. To achieve this objective, the performance measures of each of the models will be used and compared. In addition, this objective will require further analysis of each of the models, looking at more subjective measures of performance such as levels of generalisability and interpretability. These measures are important when considering the nature and context of the problem. For example, in many cases professionals might not only wish to know the predicted outcome, but also the risk factors that are most associated with that outcome. In this scenario there might be a trade-off between the predictive power of the model and its level of interpretability. This issue is discussed further in the next section.

Background and Context

There are several themes that emerge from the literature and which provide context for this project. Broadly these can be summarised as; understanding mental health conditions and their risk factors, the role and purpose of risk stratification and risk prediction techniques in health and social care, and the relative advantages of parametric and nonparametric approaches for these prediction problems. Each of these three themes is discussed below.

Mental health conditions and risk factors

Mental health conditions are an epidemic problem in modern society, and it is estimated that 1 in 6 people in the UK suffer from a mental health condition at any given time (BBC 2018). In Barking and Dagenham, the data shows that approximately 10,000 people are suffering from depression or another mental health condition, and this only includes those with a formal diagnosis. These conditions can have profoundly negative impacts on people's quality of life and create a significant demand on the health and social care system. Being able to proactively predict and possibly prevent people from developing such conditions should be considered a priority from a public health perspective.

As noted by Bhakta and Sau, mental health conditions such as depression and anxiety often occur in people who have other comorbidities (Bhakta, Sau 2017), and studies have shown that having a mental health problem can increase the risk of developing other health conditions such as cardiovascular disease (Chesney, Ilyas and Patel 2017). It is difficult to confidently assess the direction of relationship in all cases, as it is also reasonable to assume that those with other health conditions are more likely to develop a mental health problem. Nevertheless, regardless of the direction, it seems clear that there is a relationship between mental health problems and other health conditions, which makes it even more important that preventative action is taken.

In order to understand why mental health conditions develop in individuals it is important to take a holistic viewpoint, considering both health and social factors. Bhakta and Sau identify 11 key variables in their research, namely age, sex, marital status, employment status, personal income, pain at multiple sites, recent bereavement, diabetes mellitus, hypertension, visual impairment, and hearing impairment (Bhakta, Sau 2017). In order to build a model accurately predicts mental health outcomes, it is therefore important to take a wide range of factors into consideration.

Risk stratification and risk prediction in health and social care

Risk stratification is currently used across many areas of health and social care, and particularly in large integrated care systems, for example the NHS vanguards. Many of the vanguards have developed bespoke models for predicting unplanned emergency admissions and use their models to 'case find individuals and develop proactive approaches to intervention' (NHS England 2017). In addition to prediction, risk stratification tools can also support 'early diagnosis and treatment selection', by clustering individuals with the same symptoms into groups based on other relevant shared characteristics (Bzdok, Meyer-Lindenberg 2018). Bzdok and Meyer-Lindenberg note that psychiatry has 'typically relied on human-conceived diagnostic groups' and ignored variation between patients with the same diagnosis. By using machine learning methods for risk stratification, the heterogeneity of patients is recognised and used to inform more effective and personalised treatment and intervention (Bzdok, Meyer-Lindenberg 2018).

Parametric vs nonparametric approaches for risk stratification

The final theme that comes out of the literature and which will be explored in this project is the relative advantages of parametric and nonparametric models for risk stratification in the health and social care context.

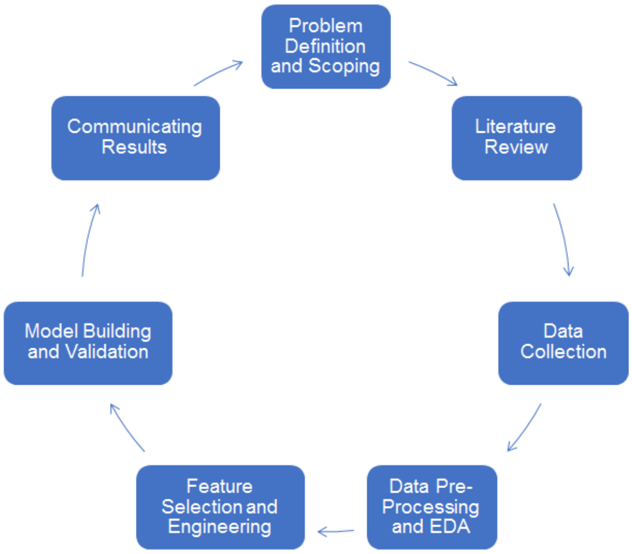
König, Kruppa and Ziegler provide a simple distinction between the two methods and how they relate to risk stratification. On the one hand you can opt for a model that classifies patients into risk groups (e.g. high, low or medium). The technique to do this is classification and typically uses nonparametric approaches (König, Kruppa, Ziegler 2012). On the other hand, you can opt for a model that estimates the probability level of the risk. The technique for this is called risk prediction and usually uses parametric approaches (König, Kruppa, Ziegler 2012). Indeed, there is an ongoing debate as to which approach provides more practical utility to health and social care professionals; probabilistic prediction through nonparametric machine learning approaches, or the identification of specific risk factors through traditional parametric statistical approaches (Bussas, Király, Mateen et al 2016).

Generally, the literature seems to favour the nonparametric approach, though this might be because of its novelty, or potential bias towards a specific approach by the authors. Indeed, the only advantage of parametric approaches that is described are their ability to both identify and measure the strength of association between input variables and the target variable (Bussas, Király, Mateen et al 2016). This advantage shouldn't be overlooked though, as the formal nature of parametric modelling process and the ability to understand the relationship between all the different variables makes these models much more interpretable. In the context of health and social care this can be very beneficial as it allows practitioners to develop a more thorough understanding of the conditions they are treating, and the risk factors that contribute to them. Knowing this information can help these professionals to identify possible 'root causes' and develop solutions to tackle these. On the other hand many drawbacks of parametric approaches are noted including their tendency to overfit the data (Bussas, Király, Mateen et al 2016), their tendency not to generalise well (Bussas, Király, Mateen et al 2016), that variables which are strongly associated often do not make good predictors (König, Kruppa, Ziegler 2012, Bzdok, Meyer-Lindenberg 2018), their lack of ability to make judgements for specific individuals or groups (Bzdok, Meyer-Lindenberg 2018), and the burden of having to specify the model and the relationship between the variables in advance (König, Kruppa, Ziegler 2012).

On the other hand, nonparametric machine learning approaches are widely praised for not making many prior assumptions (Bzdok, Meyer-Lindenberg 2018), their ability to generalise and extrapolate to other datasets (Bussas, Király, Mateen et al 2016, Bzdok, Meyer-Lindenberg 2018), that they often use data that is routinely collected and therefore more reproducible for other researchers or practitioners (Bzdok, Meyer-Lindenberg 2018), and their ability to make predictions for multi-class target variables (Bzdok, Meyer-Lindenberg 2018). The one drawback of these approaches that is noted is that they often do 'not allow consistent estimation of probabilities' (König, Kruppa, Ziegler 2012).

Methodology and Task Schedule

The methodology for the project will follow best practice in machine learning and predictive analytics. The diagram below sets out the high-level steps that are proposed:



In practice it is expected that many of the steps will occur in parallel, as the models will be built iteratively, assessed in terms of performance, presented to stakeholders for feedback, and continuously tuned and improved where possible to further improve accuracy.

Problem Definition and Scoping

In this preliminary phase the problem to be addressed and scope of the project will be defined alongside stakeholders from Barking and Dagenham Council, Care City and the NHS. This phase is already underway and consultation with stakeholders has begun. The remainder of the phase will be spent ensuring that stakeholders agree on the objectives and outcomes of the project.

Literature Review

This phase has also already begun and will continue as the project progresses, and potentially as the scope changes. The literature has been broken down into three broad categories:

- Risk stratification in Health and Social Care - this covers academic articles but also official guidelines around building predictive models, particularly in the NHS. These articles will be helpful for learning good practice and ensuring that the model has a practical benefit
- Predictive Analytics in Mental Health - this explores previous research in predicting mental health outcomes or specific conditions (e.g. depression or anxiety). These articles will be particularly helpful in determining the features to be used in the model
- Machine Learning Approaches to Prediction - this covers a broader range of articles and textbooks where machine learning approaches have been used for prediction and will be used to ensure that models are built and validated using best practice approaches. They will also be useful for assessing the advantages and disadvantages of parametric and nonparametric approaches

Data Collection

The data collection phase is a key component to ensuring the success of the project. The first challenge is confirming the availability of the data. This has already been agreed with the data owners and now requires discussion as to the mechanism by which the data will be accessed. One this is agreed, there may be further work required to enhance the dataset, particularly with demographic and socioeconomic data that is available in the council. It will be within my ability to collect this data; however, I will be reliant on others to match the new variables to the dataset, anonymise it and send it back to me. For this reason, I have allowed a period of six weeks for this process.

Data Pre-Processing and Exploratory Data Analysis

Once the dataset is ready and available for analysis, I will begin the pre-processing and exploratory data analysis (EDA). These two phases will be completed in parallel as the EDA will draw out a lot of the nuances in the data. One of the key tasks during this phase will involve exploring any collinearity between the features. The reason for this is to try and avoid any confounding effects, which can create bias and lead to an overestimation of the predictive power of models (Bzdok, Meyer-Lindenberg 2018). An initial logistic regression model will also be built during this phase, which will provide some insight as to which features are most strongly associated with the target variable.

Feature Selection and Engineering

The next phase involves choosing the features that will be used in the models. It is proposed that a mixture of methods are used, including drawing on the associations found from the EDA and logistic regression model, formal techniques for dimensionality reduction (e.g. principal component analysis), as well as drawing on the literature review. For example, Bhakta and Sau identify several socio-demographic and health related factors that are associated with depression and anxiety. Using this pre-existing research can save time in the modelling process.

The feature selection and engineering will be an ongoing process, and as the models are built features may be adapted or created to try and enhance the accuracy of the models.

Model Building, Tuning, Assessment and Validation

Model building will involve partitioning the data into training and testing sets. The models will be built using the training dataset, and then validated for performance using unseen data from the testing dataset. A resampling procedure will be used for this purpose, for example k-fold cross validation.

Several different models will be built using different statistical and machine learning techniques. It is expected that approximately five models will be developed, including at the very least a logistic regression and a decision tree model. It is hoped that more advanced models can be built, such as a hybrid model using ensemble methods. The hyperparameters of the models will also be tuned to optimise performance. All model building will be carried out in either R or Python.

Communicating Results

The final step in the process is to communicate the results to stakeholders, and to write up the project report. This phase will also include putting together the prediction outputs for the mental health practitioners to take forward.

A GANTT chart with a full list of tasks is included in Appendix 2.

Risks and Ethical Considerations

There are several risks and ethical considerations to the project that should be noted:

Data Access

Firstly, there is a risk around data access. This includes not receiving the data in time or having difficulty getting access to the data on my own workstation. Generally, these risks are low as I already have familiarity with the dataset and have a good relationship with the data owners. I have also left a six-week period to resolve any issues with access which is hopefully sufficient.

Input Data and Features

Linked to the above there is also a risk that the data itself is insufficient to build a good model. This includes not having enough variables, or more likely not having variables of a high enough quality to build a good predictor. As noted by NHS England, 'having a breadth of different indicators is important' (NHS England 2017). This risk shouldn't be high as the dataset contains over 250 variables and includes a mixture of sociodemographic and health related factors. It may also be possible to add more variables to the dataset based on council data that I have access to.

The second risk is not having enough observations. As noted by Bzdok and Meyer-Lindenberg, the best prediction usually comes from datasets where the number of observations is greater than 1,000,000 (Bzdok, Meyer-Lindenberg 2018). The dataset in question contains approximately 115,000 records which is still a very significant number. This is much better than other studies, for example Bhakta and Sau's used data from only 520 patients to train their model.

A potentially more problematic consideration is the computational power required to train more complex models on a dataset of this size, for example a random forest. In such situations a sample of the data might have to be taken.

Model Building and Bias

A further risk that must be considered is the potential for model bias. Care must be taken in choosing the variables to be used, understanding that 'choosing features based on their level of association does not been they will be good predictors' (König, Kruppa, Ziegler 2012). Similarly, it is important to be aware of possible confounding effects which have the potential to 'inflate predictive performance and lead to bias' (Bzdok, Meyer-Lindenberg 2018). A related concern is around overfitting which is a common problem when using machine learning methods. One mechanism for reducing this problem is to use techniques such as regularisation.

Generalisation and Reproduction

The next concern to be aware of is the level of generalisability and reproducibility of the model. Linked to the point above about the number of observations, it is important to have a good sample size for a model to be generalisable (Bhakta, Sau 2017). Similarly, the training data should be representative of the population if the model is to be applicable more widely (Bussas, Király, Mateen et al 2016, Bzdok, Meyer-Lindenberg 2018). On this point it is important to recognise that the model will be built using only data from Barking and Dagenham. As such, it is only possible to be confident of its predictive value for individuals who live in the borough. It may be the case that it can generalise to other London boroughs which have similar demographics in terms of age, ethnicity and socioeconomic factors.

However, it will be challenging to adapt the model for use more widely. Unfortunately, there is no external dataset available to test and validate the model further on, which would be helpful in determining its generalisability.

Ethics and GDPR

The final risk of note is around ethics, and specifically the protection of individuals' rights under GDPR. To this end all the data will be fully anonymised before it is shared and before any analysis or processing is carried out. As such, there will be no way of personally identifying any of the individuals in the dataset.

Expected Outcomes

There are four expected outcomes of the project, which link directly to the four project objectives. These are:

A tangible set of models built in R or Python which can make binary predictions about whether an individual is likely to develop a mental health condition or not. The criteria for assessing that this outcome has been met are:

- An R / Python script exists with code used to build the models
- A set of predictions for the target variable have been produced for all the individuals in the testing dataset
- The dataset and code can be provided to a third party with knowledge of R / Python. This third party can run the models and produce the same set of predictions

A model that is accurate, has a low error rate, and has good precision, specificity and sensitivity. It is difficult to know exactly how well the model will perform in advance, however the work of Bhakta and Sau could be used as a comparator, and specifically the metrics for accuracy (89%), precision (89%) and sensitivity (89%). This sets a very high benchmark for the model to achieve.

The ability to provide key decision makers and professionals in the council with a list of anonymised individuals, or summary level information, about those individuals with the highest level of risk so that judgements can be made about possible interventions or treatment. This outcome will be assessed through feedback from key stakeholders in the council on the question ‘To what extent does the information provided to you from the model meet your requirements for professional practice?’. A simple scale can be used for responses:

- 1 - the information received meets all my requirements
- 2 - the information received meets most of my requirements
- 3 - the information received meets some of my requirements
- 4 - the information received meets none of my requirements

A judgement on whether parametric or non-parametric models are more suitable to questions of risk stratification and risk prediction. This outcome is the most subjective of the four. It can be measured based on whether there is a clear section in the final report that provides an analysis of the merits of different kinds of models in the context of predicting outcomes in the health and social care sector. This analysis should arrive at a clear judgement of which approach is preferred and a recommendation for future work in this area.

Annotated Bibliography

NHS England (2017) – "Risk stratification: Learning and Impact Study"

This is a study produced by NHS England exploring the impact of the various risk stratification tools being used by the new care model vanguards. It provides an explanation of what risk stratification is and its use in 'case finding' individuals for early intervention. It explains the different types of risk stratification and makes a case for predictive modelling over other techniques such as threshold modelling. It is explained that most tools currently in use are designed to predict unplanned emergency hospital admissions and argues that it would be beneficial to predict other events and outcomes as well. One of the key lessons from this study is about the implementation of the model, and how to ensure it has some practical benefit. In addition to the usual measures of model performance it introduces the concept of 'impactability' which looks at whether the risk that is identified is preventable, and the probability that an intervention will be effective. These are crucial elements to consider if the model is going to yield practical benefits, even though their judgement lies outside of the scope for prediction. The article also touches on some of the risks around data size and shape, and anonymisation.

König, Kruppa, Ziegler (2012) – "Risk estimation and risk prediction using machine-learning methods"

This article introduces machine learning for risk prediction and provides a comparison between different methods. One important distinction that is made is between classification and risk prediction, whereby the latter provides a probability score rather than a single class outcome. It is argued that classification uses non-parametric approaches, whereas risk prediction uses parametric approaches. A comparison of the two approaches is provided and the authors seem to favour nonparametric approaches. This is because parametric approaches such as logistic regression require all the variables and their relationships to be specified in advance, which is practically very difficult. Moreover, it is argued that such models can show strong associations between certain features and the target variable, but these features may still be poor predictors. On the other hand, the authors do recognise that nonparametric models also have drawbacks, noting that they do not always 'allow consistent estimation of probabilities'.

Bzdok, Meyer-Lindenberg (2018) – "Machine Learning for Precision Psychiatry: Opportunities and Challenges"

This article explores the potential for using machine learning approaches in the field of psychiatry and mental health. It stresses that traditional approaches to psychiatry tend to rely on 'human-conceived diagnostic groups' which are constructed based on specific conditions and symptoms rather than the individual differences of patients. Noting that 'drug treatment choices are only successful in every second patient', the authors emphasise the importance of personalising diagnosis and treatment and recognising heterogeneity between individuals. They propose machine learning prediction as a mechanism to support this position. They provide an overview of the benefits of machine learning, arguing that it makes minimal prior assumptions, generalises well to external datasets, is good at dealing with multi-class problems, and tends to use routinely collected data rather than data collected from specific experiment design, hence making the models more reproducible. One interesting potential problem raised is around longitudinal analysis, and the role that machine learning plays in predicting at different time intervals in the future. The authors only touch on this question but it is an important consideration, as most machine learning models will be restricted to making predictions within a certain timeframe. This differs from more traditional models used in epidemiology, for example survival analysis, which can model the time to a specific event.

Bhakta, Sau (2017) – "Predicting anxiety and depression in elderly patients using machine learning technology"

This study is very similar to the one that is being proposed, as it uses a range of classification techniques to predict a binary outcome, namely 'anxiety and/ or depression present' and 'anxiety and depression absent'. To do this, the authors use a range of input variables covering socio-demographic characteristics and health conditions. Ten different classifiers are built covering a range of different methods, including logistic regression, bayesian networks, random forest, sequential minimal optimisation and others. All the models are compared in terms of performance and the most accurate one is chosen, which is the random forest (accuracy of 89%). The authors follow a similar methodology to the one proposed here, for example using feature selection methods and resampling techniques, in this case tenfold cross-validation. The main differences between this study and this proposal is that this was focussed on elderly patients with the specific conditions of depression or anxiety. The dataset is also smaller, with only 510 patients in the training dataset and 110 patients in the testing dataset. Nevertheless, overall this study serves as a good template for the project.

Bussas, Király, Mateen et al (2016) – "Machine Learning in Falls Prediction; A cognition-based predictor of falls for the acute neurological in-patient population"

This article is not specifically about mental health but does provide a useful overview of using machine learning methods for prediction in a health and social care context. Like other articles it provides some analysis of different model types, on this occasion distinguishing between 'descriptive modelling', meaning traditional statistical inference, and 'predictive non-linear models'. It argues that the former is good at identifying relationships, however, may overfit the data and often do not generalise well to other datasets, which is less of a problem for the non-linear models. However, it also points out that the predictive models only generalise well if the population at hand is similar, which often may not be the case. Another useful insight from this paper concerns model performance measures. Specifically, it is pointed out that models with low specificity are unhelpful in a healthcare setting since they increase the number of patients and therefore the level on demand on a set of already very stretched resources.

Appendix 1 - References

BBC (2018), Mental health: 10 charts on the scale of the problem, 04/12/2018, available at: https://www.bbc.co.uk/news/health-41125009

Bhakta, Sau (2017), "Predicting anxiety and depression in elderly patients using machine learning technology", *Healthcare Technology*, vol 4, Issue 6, pp.238-243

Bussas, Király, Mateen et al (2016), "Machine Learning in Falls Prediction; A cognition-based predictor of falls for the acute neurological in-patient population", *CoRR*, abs/1607.07751

Bzdok, Meyer-Lindenberg (2018), "Machine Learning for Precision Psychiatry: Opportunities and Challenges", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol.3, issue 3, pp.223-230

Chesney, Ilyas and Patel (2017), "Improving life expectancy in people with serious mental illness: should we place more emphasis on primary prevention?", *The British Journal of Psychiatry*, 211(4), pp.194-197

König, Kruppa, Ziegler (2012), "Risk estimation and risk prediction using machine-learning methods", *Human Genetics*, October 2012, volume 131, issue 10, pp.1639-1654

NHS England (2017), "Risk stratification: Learning and Impact Study", *Operational Research and Evaluation Unit, NHS England*, July 2017, available at: https://imperialcollegehealthpartners.com/wp-content/uploads/2018/07/ORE__Risk_stratification_learning_and_impact_study.pdf

Appendix 2 - GANTT Chart

	Start Date	End Date
Problem Definition and Scoping		
Define the problem to be addressed, and the target variable being predicted	15/04/2019	03/05/2019
Agree the data requirements and features to be collected	15/04/2019	03/05/2019
Agree the objectives of the project	15/04/2019	03/05/2019
Agree the benefits of the project including cost savings	15/04/2019	31/05/2019
Agree the performance criteria for the model	15/04/2019	03/05/2019
Agree the key stakeholders and plan to discuss the agreed scope with them	15/04/2019	03/05/2019
Finalise project plan	15/04/2019	03/05/2019
Literature Review		
Literature Review - Risk Stratification in Health and Social Care	01/04/2019	30/06/2019
Literature Review - Predictive Analytics in Mental Health	01/04/2019	30/06/2019
Literature Review - Machine Learning Approaches to Prediction	01/04/2019	30/06/2019
Data Collection		
Gain agreement on use of Care City dataset	01/04/2019	03/05/2019
Review dataset and propose additions of new features	03/05/2019	10/05/2019
Data collection of new features	10/05/2019	24/05/2019
Joining of new features to dataset	24/05/2019	14/06/2019
Data Pre-Processing and EDA		
Carry out data cleaning (missing data, outliers etc)	14/06/2019	21/06/2019
Pre-process data for model building (reshaping, normalisation, type conversion etc)	14/06/2019	21/06/2019
Exploratory analysis of features and relationships, including collinearity	21/06/2019	28/06/2019
Test for association between features and target variable	21/06/2019	28/06/2019
Logisitc regression model	28/06/2019	05/07/2019
Production of EDA outputs (tables, plots etc) and slides for presentation to stakeholders	28/06/2019	05/07/2019
Feature Selection and Engineering		
Feature selection methods / dimensionality reduction	05/07/2019	12/07/2019
Choose features based on literature review	05/07/2019	12/07/2019
Choose features based on exploratory analysis and confoudning factors	05/07/2019	12/07/2019
Change and / or create features or proxies to try and improve model accuracy	05/07/2019	02/08/2019
Model Building, Tuning, Assessment and Validation		
Data partitioning	12/07/2019	12/07/2019
Train all models	12/07/2019	02/08/2019
Test all models	12/07/2019	02/08/2019
Calculate performance measures for each model	12/07/2019	02/08/2019
Tuning of hyperparameters	12/07/2019	02/08/2019
Communicating Results		
Creation of tables and plots	12/07/2019	09/08/2019
Creation of powerpoint slides summarising results	09/08/2019	16/08/2019
Write up of final project report	30/06/2019	16/08/2019
Production of output to be shared with professionals to take forward	09/08/2019	16/08/2019