2020

A Discrimination Aware Model to Predict Childhood Literacy Levels

Kate Byrne
Technological University Dublin

Recommended Citation

Byrne, K. (2020). A Discrimination Aware Model to Predict Childhood Literacy Levels. *A dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of M.Sc. in Computing (Data Analyticsm*. doi:10.21427/cmtz-1m13

A Discrimination Aware Model to Predict Childhood Literacy Levels

Kate Byrne

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (Data Analytics)

16-06-2020

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Kate Byrne

Date: 16-06-2020

Abstract

It is illegal in Ireland to discriminate in the provision of education on the basis of multiple characteristics including gender, race and religion. While the increased use of machine learning models can open multiple avenues to identify early intervention strategies in education, caution must be exercised to ensure that any intervention does not discriminate with respect to a protected class. Poor literacy in childhood can have long term effects as the child ages, including on employment and mental health outcomes. Early intervention is key in mitigating this. In this dissertation, a model was created that predicted the outcome of a literacy test at age 9 based on information about the individual child at ages 9 months, 3 years and 5 years, including their development, parental education levels and literacy, early exposure to books and reading, and early educational abilities. Each of these areas had been suggested in current literature to contribute to or be a risk factor for childhood literacy. As is particularly common in survey data, there is missing data. This was dealt with through deductive imputation, exclusion, and automatic imputation. The best performing model as measured by a minimal mean squared error was produced when data was deductively imputed and, where that was not possible, excluded. It was then investigated whether the resultant best performing model discriminated based on gender, race or religion. To achieve this, synthetic sets of 'twins' were created who were identical in every feature apart from the protected characteristic. These populations were created from the original data that was used to create the model. The best performing model, which minimised the mean squared error, was shown to explain 33.1% of the variance in literacy scores between children. It was also shown to discriminate based on religion and ethnicity with a weak effect. A model using

deductive imputation followed by automatic imputation performed less well and was shown to discriminate based on religion with a weak effect and discriminate based on ethnicity with a weak to medium effect. The overall experiment showed that it was possible to create a model to partially explain the variance in a measure of literacy in 9-year-old children using features from earlier in their childhood. However, this model displays some discrimination based on ethnicity. Although the effect of the discrimination observed is weak, caution should be exercised in the implementation of any real-world interventions based on similar models.

Keywords: Discrimination, Elastic Net, Childhood Literacy

Acknowledgments

I would to thank my supervisor Lucas Rizzo for his guidance and feedback and thank my family for all their support and cups of tea. I would also like to thank my employer, AIB for their invaluable support.

Contents

List of Figures

List of Tables

List of Acronyms

ASQ	Ages and Stages Questionnaire
DPRT-R	Drumcondra Primary Reading Test - Revised
ESRI	Economic and Social Research Institute
GUI	Growing Up in Ireland
ICQ	Infant Characteristics Questionnaire
LSAC	Growing Up in Australia: Longitudinal Study of Australian Children
MSE	Mean Squared Error
PCG	Primary Caregiver
SCG	Secondary Caregiver
SD	Standard Deviation
SDQ	Strength and Difficulties Questionnaire
SE	Standard Error
SSIS	Social Skills Improvement System Rating Scales

Chapter 1

Introduction

1.1 Background

As machine learning becomes more widely used by governments, businesses and non-profits, there is an increased opportunity to accelerate and improve decision making processes. However, while it may appear that the risk of human error or poor judgement is removed by automated decision making, there is the potential to either introduce bias that was not previously present or to perpetuate bias that already exists (Pedreshi, Ruggieri, & Turini, 2008). Additionally, the black box nature of some models means that discrimination against protected groups can be even harder to identify than before (d'Alessandro, O'Neil, & LaGatta, 2017; Žliobaitė, 2017). The definition of discrimination used throughout this dissertation is the unjust treatment of different groups, covered by the Equal Status Act as outlined below, not the ability to differentiate between two groups, as is a common meaning of discrimination used in machine learning[1].

One such area that has the potential to be affected by the proliferation of machine learning is education. Poor literacy in childhood can have long term effect on the life of the individual (Law, Rush, Schoon, & Parsons, 2009; Schoon et al., 2002; Wallace et al., 2015), so early intervention is key and has the potential to be very impactful. In the past, early intervention programmes for children at risk of literacy

[1] https://www.lexico.com/en/definition/discrimination

problems were largely based on qualitative research in behavioural and social sciences (Shonkoff, 2010). Adoption of machine learning can create opportunities for greater insight. According to Ireland's Equal Status Acts 2000-2015 (Government of Ireland, 2000), it is illegal to discriminate in the provision of goods and services, education and accommodation on the basis of any of the following grounds; race, membership of the travelling community, gender, religion, age, disability, marital status, family status or sexual orientation. Therefore any model designed to aid early intervention must be cognizant of discrimination.

The Growing Up in Ireland study is a longitudinal study of children in Ireland. Among its aims are to identify factors that lead to deprivation of any kind and to allow evidence based research to inform government policies. A weighting is provided to balance the dataset. However, not all protected characteristics are included in the weighting. The dataset has been used in previous machine learning and predictive studies (Crowe, O'Sullivan, Cassetti, & O'Sullivan, 2017; Murray & Egan, 2014; Hughes, Gallagher, & Hannigan, 2015), but apart from inclusion of the weighting, there has been no specific examination of discrimination using the dataset.

1.2 Research Project/problem

This work aims to investigate whether a model predicting literacy levels in schoolchildren at age 9 is discriminatory with regards to gender, religion or ethnicity. To achieve this, a model to predict the literacy ability of a child at age 9 will be created using measurements of the child's development, early educational experiences, and information about the child's family. It will be investigated whether this model predicts significantly different literacy abilities for synthetic individuals who are identical except for differing gender, ethnicity or religion, indicating a discriminatory model. The effect of any differences found will also be calculated.

Research Question

Does an elastic net model to predict literacy levels in children at age 9 based on measurements about the child's background, household, development and early education at ages 9 months, 3 years and 5 years discriminate across gender, ethnic or religious background in a population of children living in Ireland who were born in 2007/2008?

1.3 Research Objectives

The research objectives are to

- Review the current literature on the development of child literacy and risk factors for literacy difficulties
- Review the current literature on identification of discrimination in machine learning models
- Review the current literature on machine learning models with multiple potentially correlated predictors
- Obtain a suitable dataset
- Manipulate data into the format required by the chosen machine learning model and handle missing data
- Design and train such a model to predict literacy abilities
- Evaluate the predictions obtained from the implemented model
- Create synthetic child populations who differ on the protected characteristics
- Statistically compare the predictions performed by the same model using the original data and the synthetic data in order to identify any discrimination present. Calculate the effect size of any discrimination found.

1.4 Research Methodologies

Reviews will be carried out of previous literature in the areas of literacy development, risk factors in childhood literacy and discrimination in machine learning models. A dataset will then be identified that has available measurements suited to answer the research question. This work is therefore secondary research as the data being analysed has previously been collected by a third party. A suitable machine learning model will be chosen based on the target variable type and after examination of the predictive variables. A series of hypotheses will be presented regarding whether there is a difference in predicted literacy levels, as measured by DPRT-R logit score, between two synthesised individuals who differ only in protected characteristic. This is therefore empirical research. Quantitative methods will be used to both evaluate the strength and accuracy of the model and evaluate whether the model discriminates on the basis of gender, race or ethnicity. This research is inductive as it is bottom up, beginning with a theory and concluding with an observation.

1.5 Scope and Limitations

The scope of the dissertation is to build a model that predicts the reading ability at age 9, as measured by their score on the Drumcondra Primary Reading Test Revised (DPRT-R), of children who were born in 2007/2008, who grew up at least partially in Ireland and who took part in the Growing Up In Ireland (GUI) longitudinal survey wave 5 (2017/2018)[2]. Predictors will be taken from earlier waves of the same longitudinal dataset in areas that have been identified in previous literature as potential contributory or risk factors to childhood literacy, including child development, early education and family context. The model will be evaluated for discrimination based on the child's gender, religion, and the ethnicity of the primary caregiver (PCG), which acts as a proxy for the ethnicity for the study child.

As with any research study, a number of limitations are presented. A weighting is provided to make the dataset more representative of the general population. However,

[2]Growing Up in Ireland https://www.growingup.ie/

some features of interest are excluded from the weighting and so the reweighted model may not be representative of the general population with regard to those features. As no new children can join the study that were not present in Wave 1, the dataset excludes all children that were not present in Ireland at the commencement of the study, and so excludes more recent immigrants. Literacy is measured in the English language only, even if the child speaks another language at home or the school is a Gaelscoil (Irish language school). Additionally, the model does not take into account potential discrimination in the DPRT-R test itself, e.g. cultural differences that are unrelated to literacy ability. This work only investigates one measure of direct discrimination, no indirect discrimination is measured and may be present.

1.6 Document Outline

In chapter 2, Literature Review, the current research in child literacy and its contributing and risk factors and discrimination aware machine learning will be outlined. The Growing Up in Ireland dataset will be described and models suitable for multiple potentially correlated predictors will be reviewed.

Chapter 3 will cover the design and implementation of the experiment, the methods used for dealing with missing data, and the methods used to test for discrimination in the model. Evaluation methods for both the model and existence of discrimination will also be described.

Chapter 4 will outline the results of the experiment, including the evaluation of missing data, and results of the discrimination tests.

Chapter 5 will give an overview of the work carried out and a discussion of the experiment and its implications.

Chapter 2

Literature Review

2.1 Introduction

This chapter covers the current research on child literacy, how it develops and any known or suspected contributing or risk factors. A short description is also included of the development of child literacy theory and frameworks for teaching. The current literature on discrimination, and discrimination specifically in machine learning is then described, including types of discrimination, ways of measuring and ways to counteract it. The Growing Up in Ireland study is then described, including its aims, how it was constructed, and previous machine learning studies carried out using the dataset. Finally, current literature in predictive methods are described, focusing on methods suitable for the dataset proposed here, including multiple, potentially correlated predictors.

2.2 Literacy

Throughout recent history, there have been several different frameworks of childhood literacy education. In the first half of the 20th century it was thought that children under the age of 6 and a half should not be formally taught to read as they had not reached the required maturity level and the children would be pressured before they were ready and become discouraged. It was thought that children should be

identified as having 'readiness to read' through a series of tests and observations prior to commencing reading instruction. However in the 1960s, it was observed that many children showed interest and ability in learning to read before that age and so a new framework, 'emergent literacy', was proposed, which said that children gradually increase their literacy skills from a very young age and so should be continually exposed to learning materials and exercises suitable for their maturity level in order to aid their progress. It was also acknowledged that children develop at different rates, and so the level of educational materials and methods used should depend on the individual child's progress and maturity, not their age (Saracho, 2017).

Several learning methods in the emergent literacy framework are thought to have a positive influence on children's literacy development. Dialogic reading is a method where the adult uses interactive behaviours when reading to a child instead of reading the text directly. The adult should ask open ended questions about the story or images, ask follow up questions to the child's answer, praising and encouraging child's participation, linking the story to the child's interests or experience. Print referencing involves focusing on the print in a storybook, recognising letters, where on the page you should start reading and tracking the print while reading. Literacy-enriched play should be encouraged, enabling play scenarios that involve literacy, e.g. pretending to be a family going shopping with a list, playing 'school' etc (Justice & Pullen, 2003). There are also many teacher led methods including a curriculum based on phonological awareness, rhyming and identifying, blending and segmenting the components of words (Justice & Pullen, 2003).

Interventions to aid these methods involve training the caregiver on how best to incorporate these methods, providing materials like books or props for play such as blackboards and chalk or notebooks and pens, or additional help in teacher led methods. Interventions should be sensitive to cultural influences in raising children (Manz, Hughes, Barnabas, Bracaliello, & Ginsburg-Block, 2010). As the factors contributing to literacy ability are complex and interlinked, there are likely many influencing factors and therefore potential interventions that have not been identified.

Many reasons have been proposed to explain why a child might be at risk of

developing literacy skills at a slower rate than their peers, including diagnosed physical and learning difficulties, family literacy issues and environmental factors such as school or home environment that might result in a lack of the interventions mentioned above. Developmental language disorders are defined as language difficulties that are not associated with a diagnosed biological cause, and are thought to affect approximately 7% of school aged children (Armstrong et al., 2018). The effects of language difficulties in childhood can have far reaching effects. It is widely considered that success in literacy has a strong influence on success in future schooling and later life (Saracho, 2017) and early reading difficulties have been associated with adverse outcomes in educational achievement, employment and mental health in adulthood (Law et al., 2009; Schoon et al., 2002; Wallace et al., 2015). As many of the risk factors for literacy difficulties are also risk factors for poverty, poor academic achievement, mental health issues unemployment issues later in life (Schoon et al., 2002), caution should be used when determining causation. A risk factor here is defined as a factor that indicates that an individual is more likely to have issues learning to read. This is not a guaranteed outcome, and the risk factor can generally not be said to be the cause of the issue, correlation is only established, not causation. Ability in learning to read is a complex and highly individual process so there is no clear cause, and in reality there are likely many interlinking causes.

In previous studies, risk factors for speech and language delay have been shown to include male gender, family history of language difficulties, maternal non-English native language, maternal mental health distress (Taylor, Christensen, Lawrence, Mitrou, & Zubrick, 2013), low parental education (Law et al., 2009; Wallace et al., 2015), being in a single parent household, overcrowded housing, no pre-schooling (Law et al., 2009), communication skills and motor, social and adaptive skills, early temperament and social competence (Armstrong et al., 2018, 2016), availability of books in the house and if the child is read to (Justice & Pullen, 2003).

Models to predict literacy levels generally obtain R^2 in the region of 10-40%. A model based on pre-literacy and socio-emotional skills could predict 9% of variance in children's decoding skills (a precursor to literacy). A more detailed version of this

model could explain 29% of the variance (Pentimonti, Murphy, Justice, Logan, & Kaderavek, 2016). Another study showed that 16% of variance in later vocabulary skills could be explained by assessments at 4-9 years old (Armstrong et al., 2018). Literacy is a very complex ability and has many contributing factors. Due to the variation in time to reach early childhood language milestones, there is even evidence to show that tests related to language performed under 18 months are not good predictors of later literacy levels (Duff, Nation, Plunkett, & Bishop, 2015).

While there are several studies that have looked at the link between literacy and the protected characteristics studied here (Manz et al., 2010) and several studies using machine learning to predict literacy outcomes (Armstrong et al., 2018; Taylor et al., 2013), studies have not been done to look at whether these predictive models discriminate on protected characteristics.

2.3 Discrimination

Discrimination in the provision of goods and services, education and accommodation on the basis of race, membership of the travelling community, gender, religion, age, disability, marital status, family status or sexual orientation, is illegal in Ireland under the Equal Status Acts 2000-2015 (Government of Ireland, 2000). There have been reports of discrimination against migrant children in Irish Schools (Darmody, Byrne, & McGinnity, 2012). It is, however, legal to take positive action to promote equality for disadvantaged persons or cater for special needs of individuals. Discrimination may be direct, e.g. refusing admission to members of a certain race, or indirect, e.g. preference for school admission being given to children of parent who went to the school themselves, excluding children of parents who immigrated the the country as adults.

Since longitudinal studies like the Growing Up in Ireland dataset are used to inform policy on all aspects of children's lives, including early intervention programmes, we must ensure that discrimination is not occurring. The overall aim of discrimination aware machine learning is to create a model that maximises the accuracy while

minimising the level of discrimination. To do this, the level of discrimination needs to be measured and counteracted if it occurs.

2.3.1 Discrimination in Machine Learning

Recent developments in machine learning have resulted in quicker and more efficient decision making. However, there is the potential to either introduce bias that was not previously present or to perpetuate bias that already exists (Pedreshi et al., 2008). Additionally, the black box nature of some models means that discrimination against protected groups can be even harder to identify than before (d'Alessandro et al., 2017; Žliobaitė, 2017).

Discrimination can be either direct, where the model makes a decision based on the protected class, or indirect, where the protected class is excluded as a predictive attribute but the model still disadvantages members of the protected class (d'Alessandro et al., 2017; Calders & Verwer, 2010; Barocas & Selbst, 2016). Redlining is a famous example of indirect discrimination that occurred in many parts of the United States, where credit or other opportunities was denied to residents based on which neighbourhood they lived in, regardless of the financial circumstances of the individual, with race apparently excluded from the decision. It was however found that the neighbourhoods were largely racially segregated and the neighbourhoods denied credit were predominantly non-white, so credit decisions were indirectly made on the basis of race (Squires, 2003; Pedreshi et al., 2008).

Direct discrimination is measured using situation measures, which identify if individuals in the dataset have been discriminated against and how this is distributed across the whole dataset. They don't measure the magnitude of the discrimination (Žliobaitė, 2017). This can be counteracted by the suppression stage in pre-processing. If a difference in the outcome predicted by a model is observed between two individuals with identical features except for the protected class, direct discrimination can be said to have occurred (the 'twin test'). These two complementary individuals can be synthetically created. Removal of the protected attribute from the model can eliminate the risk of direct discrimination, but not of indirect discrimination (see the redlining

example above). Indeed, the protected attribute could be required to ensure that indirect discrimination is not taking place (Žliobaitė & Custers, 2016). Indirect discrimination can be said to have occurred if the difference in predictions across groups of individuals is larger than can be justified by their non-protected characteristics (Žliobaitė, 2017; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012).

Identifying and measuring indirect discrimination has proven to be more difficult. While there has been no consensus in the data community about a single measure (d'Alessandro et al., 2017; Žliobaitė, 2017), several have been proposed, including, statistical measures (Calders & Žliobaitė, 2013; Žliobaitė, 2017), absolute measures (Calders, Karim, Kamiran, Ali, & Zhang, 2013; Žliobaitė, 2017) and unexplained differences (Žliobaitė, 2017; Kamiran, Žliobaitė, & Calders, 2013).

Discrimination can occur at the pre-processing, in-processing and post-processing stages of a machine learning process. There are relevant methods for creating a discrimination-aware process at each stage (d'Alessandro et al., 2017; Dwork et al., 2012; Žliobaitė, 2017). There have been multiple alternative methods proposed and some specific packages have been created (Bellamy et al., 2019; Beutel et al., 2019; Calmon, Wei, Vinzamuri, Natesan Ramamurthy, & Varshney, 2018; Romei & Ruggieri, 2013; Veale & Binns, 2017; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Luong, Ruggieri, & Turini, 2011; Hu & Chen, 2018; Yeom & Tschantz, 2018; Kamiran, Karim, & Zhang, 2012).

Issues in pre-processing can occur with the dataset itself. The dataset can be biased, which can be counteracted by supressing protected attributes and attributes highly correlated with the protected attribute. There can also be sample bias, i.e. over-representation or underrepresentation (Kamishima, Akaho, Asoh, & Sakuma, 2012). Several methods are suggested to counteract these issues including massaging (Kamiran & Calders, 2009), reweighing (Calders, Kamiran, & Pechenizkiy, 2009) and stratified sampling (Kamiran & Calders, 2012). In-processing issues can occur with model misspecification, where the model does not correctly account for everything it should and thus gives rise to discrimination. A machine learning model is said to be discriminatory if one of the following occurs: 1) people with similar non-protected characteris-

tics receive different predicted outcomes or 2) differences in predicted outcomes across groups are larger than those that would be expected due to their non-protected attributes (Žliobaitė, 2017). Post-processing involves auditing of the results and deferral to human judgement (d'Alessandro et al., 2017).

Statistical measures detect the presence or absence of discrimination, but not its magnitude or the distribution within the dataset. They test the null hypothesis that there is no difference between the protected and non-protected groups using the appropriate standard statistical test, which depends on the type of data being tested (Calders & Žliobaitė, 2013; Žliobaitė, 2017). Absolute measures use only the protected characteristics and the predicted outcome to calculate the magnitude of the discrimination. It assumes that all individuals are identical aside from their protected characteristics which is generally not the case, so it is generally used in conjunction with other measures (Calders et al., 2013; Žliobaitė, 2017). Since there may be valid reasons for differences in outcomes for protected groups e.g. members of a protected group may have a lower average income than those not in the protected group, which would explain why individuals not in the protected class would be more likely to be offered credit. Issues like this are taken into account by separating differences into explained and unexplained differences using conditional measures and the unexplained differences are taken to be discriminatory (Žliobaitė, 2017; Kamiran et al., 2013).

2.4 The Growing up in Ireland Study

Growing up in Ireland (GUI): National Longitudinal Study of Children is an Irish government funded study carried out jointly by Trinity College Dublin and the ESRI (Economic and Social Research Institute) and is carried out under the Statistics Act (1993). It provides input into the National Children's Strategy, a major national plan for children published by the Department of Health and Children in 2000. It is managed by both the Department of Children and Youth Affairs and the Central Statistics Office. It is the first study of it's kind to take place in Ireland. The overarching aim of the study is to investigate the many factors that contribute to or undermine the

well-being of children growing up in Ireland and to allow evidence based research to inform national policies addressing challenges in childhood. The longitudinal nature of the study will allow the investigation of long term effects of factors as the children develop into adults. Further, they aim to describe the life of an Irish child and determine what is typical and what is not, and to gather children's opinions about their lives. The study also aims to identify factors that lead to social disadvantage or educational difficulties and provide evidence for the creation of policies and services for children and families (Thornton, Williams, McCrory, Murray, & Quail, 2013)

2.4.1 Topic and Question Selection

The study used the Bronfenbrenner framework to ensure that all critical areas of influence on a child's life and development were included in the study (Bronfenbrenner & Morris, 2007). Topics included those internal to the child such as gender, health, ethnicity, physical, social and psychological development and temperament, the microsystem around the child, which is the individuals and systems that the child directly interacts with such as parents and caregivers, immediate family and peers. Topics covered in this area involve parent education, health, stress and marital relationship, size of household and family structure, parenting and attachment style and childcare and relationships with peers. The study also explores the mesosystem of the child, which includes the factors that influence the individuals in the child's microsystem including the parent's work life balance and maternity leave policies, parental relationships with other family members, involvement with the community etc. The exosystem includes the institutions and systems that directly affect the microsystem and mesosystem such as the health, social welfare education and religious systems. Finally, the macrosystem involves global forces such as national policies, cultural beliefs, economic climate and socio-historic setting of the study (Thornton et al., 2013).

Expert panels from a wide range of relevant areas were consulted on the content of the questionnaires and study methodology. Some questions were derived from similar longitudinal studies such as the Growing up in Australia study and the Millennium Cohort Study (Thornton et al., 2013). Questions were further narrowed down based

on several criteria, including the importance of the topic to the welfare of children and whether it is feasible to action an item through public policy, whether accurate information could be ethically collected, whether the variable can be reliably measured and statistical considerations, such as whether the variable was sufficiently frequent in the population to be measured in this sample size (Thornton et al., 2013). To this end, information is gathered in the form of interviews and questionnaires directly from primary and secondary caregivers, childcare and schools, and the child themselves when they are old enough to participate. Interviews with the child (when appropriate), primary and secondary caregivers who live with the child were conducted in person by the interviewer using a laptop. Caregivers living separately to the child, additional out of home carers (i.e. childminder or creche), the teacher of the child and principle of the school all received postal questionnaires (Thornton et al., 2013), see details in Table 3.1 (p. 21).

Questions asked in the surveys were a combination of factual questions, e.g. age of caregiver, income of family, opinion based questions e.g. the parent's evaluation of the child's health or safety of the local area, and scale measurements. Parental evaluation of the child's health and development, while not replacements for assessment by a medical professional, have been found to be a valid measure (Thornton et al., 2013). Scale measures are often preferable to single questions because of their reliability and validity and scope to capture more complex concepts like a child's development that may be multi-faceted. A test in considered reliable if the same person gets a similar score if retested at a different point in time for a value that would not be expected to change and if the measure is internally consistent, i.e. similar questions should be answered similarly. The validity of a test measures the ability of the test to correctly assess the concept that it is trying to measure, there should also be consistency with other valid measures of the same concept (Thornton et al., 2013).

2.4.2 Previous research using GUI

One of the aims of the GUI dataset is to inform policy about children's lives, including early intervention programmes. Because of this, we need to ensure that the dataset

or any models using the dataset are discrimination aware. While there have been numerous studies performed on the Growing Up in Ireland infant cohort dataset, including a classification tree analysis on weight and dental status in early childhood (Crowe et al., 2017), a cluster analysis of infant sleeping patterns and maternal health (Hughes et al., 2015), and a statistical analysis of the link between reading to infants and cognitive development (Murray & Egan, 2014) none so far have examined the link between the above mentioned developmental milestones and literacy in childhood. Additionally, fairness of the GUI dataset and underlying bias against vulnerable groups does not appear to have been considered in these or other studies on this dataset. Since one of the purposes of the collection of this dataset is to inform government policy, an examination of the fairness of the dataset is crucial.

2.5 Predictive Modelling

Regularisation is a method to deal with a large number of potentially correlated predictors by controlling the impact of each variable. Lasso (least absolute shrinkage and selection operator) and ridge are two such types of regularisation. Lasso deals with these by grouping together correlated variables, selecting one and disregarding the rest by setting a penalty value on these to 0. In contrast, ridge regression keeps all variables, and lowers the impact of all correlated variables as a group, so strongly correlated predictors tend to be included or excluded from the model together. λ is the regularisation parameter and is the value of the penalty introduced. Multiple values for λ are tested and the best is chosen. Lasso is less sensitive to the training set that ridge regression and so is less prone to overfitting.

Elastic net is a method that incorporates both lasso and ridge regression via a tunable hyperparameter α, the elastic net mixing parameter, which controls the balance between lasso and ridge regression methods (Zou & Hastie, 2005). $\alpha = 1$ is equivalent to Lasso and α=0 is equivalent to ridge regression. It is particularly useful when the number of predictors is much larger than the number of observations. It has previously been used in models related to biomarkers and genomic selection

(Eliot, Ferguson, Reilly, & Foulkes, 2011; Ogutu, Schulz-Streeck, & Piepho, 2012). The *Glmnet* package in *R* was developed (Hastie & Qian, 2014) to use the elastic net.

2.6 Conclusions

This chapter covered the current research on child literacy and described factors that were shown to contribute to or be a risk factor for literacy, in particular literacy in childhood. Different measures of discrimination and the background to the GUI project were also discussed. Finally, the elastic net model was introduced, which will be detailed in the following chapter. Chapter 3 will also describe how the GUI survey was implemented, which of the features of interest are available in the dataset and how the 'twin test' will be used to test for discrimination.

Chapter 3

Experimental Design and Methodology

3.1 Introduction

This chapter describes the GUI study, how it was implemented and any issues that were created in the resulting dataset. The selection of features for inclusion is also described. Selection of the predictive variables was informed by current literature on child literacy as outlined in chapter 2. The protected variables were selected from those covered by the Equal Status Act. The method for data preparation is explained, including methods for dealing with missing data and all data cleaning steps that have to occur. The methods for building and evaluating the model are also described. Finally, the method for creating the synthetic datasets testing a model for discrimination is outlined.

3.2 The dataset

3.2.1 Data Collection

The dataset was obtained through the Irish Social Science Data Archive[1]. The study has two cohorts, infant and child, which consist of nationally representative samples of 11,134 and 8,568 individuals, respectively. As this is a longitudinal study, the same cohort of participants are interviewed at multiples stages throughout their lives. Participants in the infant cohort were born between 1st December 2007 and 30th June 2008. To date (as of mid-2020), they have been involved in 5 waves of data collection, when the study children were 9 months (September 2008 - April 2009), 3 years (December 2010 - July 2011), 5 years (March - September 2013), 7/8 years (Spring 2016) and 9 years old (June 2017 - February 2018). Participants in the child cohort were born between 1st November 1997 and 31st October 1998 and to date have participated in 4 waves of data collection, when the children were 9 years (August 2007 - May 2008), 13 years (August 2011 - March 2012), 17/18 years (April 2015 - August 216) and 20 years old (August 2018 to June 2019) (Thornton et al., 2013). While improving literacy is a life long process, the time period from birth to 8 years old is considered the most significant in literacy development (Saracho, 2017). Therefore the infant cohort was chosen for this analysis.

The interviews were planned to be carried out in the period September 2008 to end of April 2009. In order for the infants to be 9 months old at the time of the interview, details were collected of the 41,185 infants born between 1 December 2007 and 30th June 2008 out of the approximately 70,000 children born in Ireland in 2007. These details were collected from the Child Benefit Register. Child Benefit is a monthly payment made by the Irish Government to the primary caregiver for each child under the age of 16 years. In order to obtain a sample that was representative of the general population, the data was stratified by marital status, county of residence, nationality and number of children in claim and systematic selection based on random start and constant sampling fraction was used. For wave 1, families were interviewed in the

[1]http://www.ucd.ie/issda/data/growingupinirelandgui/

infant's 10th month, i.e. children born 1st - 31st December 2007 were interviewed in September/October 2008 (Thornton et al., 2013). This pattern was continued in subsequent waves (McNamara, O'Mahony, & Murray, 2020).

3.2.2 Response Rates and Weighting

The initial response rate to the wave 1 survey was 58.2%, with multiple reasons for lack of response; families did not want to participate, were unavailable to participate during the required dates, agreed to participate but subsequently withdrew or refused followup, the address provided was inaccurate, family was unable to participate due to language difficulties, and in a rare number of cases, the child had died since the initial contact information was collected. The remaining 41.8% who responded amounted to 11,134 participants. This was 27% of the total number of children born in Ireland in the relevant time period, 41,185 (Thornton et al., 2013), a high proportion compared to similar longitudinal studies (compared to 11% in the Growing Up in New Zealand study (Morton et al., 2015), ∼4% in the Millennium Cohort Study UK (Plewis, Calderwood, Hawkes, Hughes, & Joshi, 2007) and ∼3% in the Growing Up in Australia Study (Edwards et al., 2012)). In subsequent waves, attrition occurred due to emigration in addition to the above reasons. No new families were added the the study to replace those who failed to respond. Families might be missing for some waves and respond at later waves (McNamara et al., 2020). In contrast to the other waves, wave 4 was a shorter postal survey instead of an in person interview, see Table 3.1 below. Despite followups by mail and phone, wave 4 had a much lower response rate than other waves, which is expected for a postal survey (McNamara, Murray, & Williams, 2019).

Analysis from the creators of the GUI dataset found that response rates, both to the study overall and to each subsequent wave, do not vary uniformly across the population; demographic groups that experience some social disadvantage including families with non-married caregivers, non-national infants and their families and other social disadvantages (income, educational attainment, social class etc), tend to have a lower response rate than those without those factors. To rebalance the dataset, in wave 1 a weighting was created that gives higher weights to children in demographics

that are underrepresented and lower weights to children that are in demographics that are over represented in order to bring the proportions up to those found in the general population taken children under 1 year, taken from the 2006 Irish Census and the Child Benefit Register (Thornton et al., 2013). In additional waves, the data was reweighted so the distribution was in line with wave 1 (McNamara et al., 2020). The data was reweighted using a minimum information loss algorithm using the GROSS program that was developed for the ESRI (Thornton et al., 2013), using 11 main characteristics; family structure, including whether the family is a lone or two parent family and number of people in the family, mother's age, mother's principal economic status, e.g. working for payment, working in the home etc., father's principal economic status, family's social class, mother's education, household tenure, i.e. whether the family own the house, rent from a private landlord or state or voluntary body, child's gender, region of the country they are resident in, mother's marital status, mother's nationality and mother's residency status in Ireland. While some of the factors that we could be concerned about regarding discrimination feature here, namely, child's gender and parent's marital status, several are not, including child's race, religion and membership of the travelling community. This is a concern as it is unclear whether the data is balanced with regard to these features.

The surveys in waves 1, 2, 3 and 5 consist of a main and a supplementary questionnaire, which contains questions of a more sensitive nature. Some respondents chose not to complete the supplementary questionnaire. It's requested that the main and supplementary questionnaire are completed by both the primary caregiver (PCG) and secondary caregiver (SCG). The main questionnaire is always filled in by the primary caregiver, if not, the study child is not included in that wave. In some cases, the questionnaires are not filled in by the SCG; they may not be present or may have refused. Wave 3 additionally has specific surveys for the teacher of the child, if the child has started primary school. Wave 4 was a reduced postal survey and so consisted of a single part to be answered by the PCG only. Wave 5 additionally consists of responses from the child's teacher and the principal of their school. These, however, are not included in this analysis.

Wave	**Number of Surveys**	**PCG Main**	**PCG Supplementary**	**SCG Main**	**SCG Supplementary**	**Teacher Survey on Child**
1	11,134	11,134	10,998	8,632	8,526	N/A*
2	9,793	9,793	7,577	9,706	7,505	N/A*
3	9,001	9,001	8,853	6,751	6,648	8,373
4	5,344	5,344	N/A**	N/A**	N/A**	N/A**
5	8,031	8,031	7,914	5,440	5,371	N/A*

Table 3.1: Total number of respondents of each wave 1-5 followed by the number of PCG, SCG and teachers that responded to the main and supplementary parts of the questionnaires. (*) The teacher questionnaire was only included in wave 3. (**) Wave 4 was a limited postal survey with only a single questionnaire filled out by the PCG.

Eight thousand and 31 families responded to wave 5. Of those, 7,750 completed the our target variable of interest, the Drumcondra Primary Reading Test - Revised (DPRT-R). 3.5% of caregivers requested that their children not sit the test so the score is absent in these cases (McNamara et al., 2020), see Table 3.2. This missing data could be handled by being excluded or imputed, each of which could raise issues. As seen above, missing data tends to occur non-uniformly throughout the dataset, so excluding these families could introduce bias into the dataset. The data could be reweighed with the new reduced dataset, but as some fields have been collapsed (e.g. age of PCG) and shielded (e.g. member of the travelling community rolled into all Irish), without access to the full dataset it would not be possible to ensure that the data would be balanced with regard to all required fields. Alternatively, only the participants that responded to wave 5 would be included, and the weighting created for this set would be also applied to wave 1-3 to ensure a balanced dataset. The missing 3.5% of data in the DPRT-R field would then be imputed, though this could raise issues of reliability.

The responses to wave 5 will be considered the complete set of responses.The

Wave	**Num of Respondents**	**Num that Responded to Wave 5**	**Num of missing participants in Wave 5**	**Perc missing participants in Wave 5**
1	11,134	8,031	0	0%
2	9,793	7,768	263	3.27%
3	9,001	7,699	332	4.13%
4	5,344	4,983	3,048	37.95%
5	8,031	8,031	0	0%

Table 3.2: Total number of surveys where at least the PCG completed the main questionnaire to waves 1-5. Of the respondents to waves 1-4, how many also responded to wave 5, how many missing rows this will correspond to and what is the percentage of rows will be missing in each wave.

weighting in wave will be applied to all fields of the combined dataset to remove bias. Section 3.3.1 (p. 34) describes the number of missing participants due to lack of response in that wave.

3.2.3 Feature Selection

Two sets of features will be selected from the surveys; the target variable and the predictor variables, which can be either potentially discriminatory or non-potentially discriminatory. Potentially discriminatory features will be chosen from those that are listed in Ireland's Equal Status Acts 2000-2015 (Government of Ireland, 2000) that apply to children, i.e. race, membership of the travelling community, gender, religion and disability. There are additional groups that are not specifically listed in the equal status acts but nonetheless represent differing needs in the affected children, i.e. children with a parent in prison are more likely to struggle academically, have additional adverse experiences and experience discrimination (McLeod, Johnson, Cryer-Coupet, & Mincy, 2019; Dallaire, Ciccone, & Wilson, 2010; Turney, 2018); children experiencing homelessness can suffer disruption to their education and increased issues with aca-

demic, social and emotional development (Keogh, Halpenny, & Gilligan, 2006; Chow, Mistry, & Melchor, 2015); children that have been in foster care system. The predictor features will be used in a model to attempt to predict the target feature. The model will then be checked to see whether it discriminates based on any of our potentially discriminatory features, gender, ethnicity or religion.

Target Variable

The Drumcondra Primary Reading Test - Revised (DPRT-R) is taken as a measure of literacy. It is a standardised reading test that has been developed specifically for group administration in Ireland[2] and is generally taken as a valid method for assessing a child's verbal ability with respect to the Irish National School curriculum (Thornton et al., 2013). It was originally developed in 1993 by the Educational Research Centre, which develops standardised tests specifically for the Irish population and conduct research on education in Ireland. The test was subsequently revised in 2006 to incorporate changes made to the Primary School English Language curriculum in 1999.

There are 6 levels of the test, corresponding to the level of schooling for each child; 1st to 6th class in Irish primary schools. It is captured for the first time in the GUI study in wave 5, where interviewers administered the test that corresponded to the child's year in school. As the children were generally 9 years old and in 3rd class in Wave 5, most took the level 3 test, but some took level 2 or 4. The test is always administered in English, even if the child attends a Gaelscoil (Irish language school) (Thornton et al., 2013).

While the full DPRT-R test covers both reading and comprehension, only the reading part was administered as part of the GUI survey. It consists of 40 questions, where the child was asked to choose the meaning of an underlined word in a sentence from multiple choice answers. The child was scored 1 point for each correct answer, giving a total score between 0-40. However, it is preferable to use the logit score which is based on the expected a posteriori scoring, derived from the difficulty and discrimination of each item, so has been weighted for the difficulty of the questions

[2]Educational Research Centre - Overview http://www.erc.ie/about/overview/

answered correctly and the level of test that the child is taking. It is therefore possible to compare across children and cohorts[3] (Thornton et al., 2013). For responses rates to the DPRT-R logit questions in wave 5, see Table 3.2.

Predictive Features

Predictive features will be chosen from those that have been linked to literacy and development in previous studies or that could potentially be the cause of discrimination against the study child. No predictive features were selected from wave 5 as these will be measured at the same time as DPRT-R and are therefore not useful in a model that attempts to predict future DPRT-R levels based on current behaviour. No features were selected from wave 4 as the response rate was relatively low, see Table 3.1.

The surveys are divided into several different sections that cover different aspects of the child's life. Initially in each study, background and personal information about the child and household is collected, including gender of the study child and primary caregiver, age of the primary caregiver, makeup of the household and type of accommodation. As previous studies have indicated that the child being from a single parent household, and having a young mother have worse educational outcomes (Thornton et al., 2013), these variables will be included. Additionally, prohibited discrimination can occur on the basis of gender, ethnicity and religion, so these variables will also be included. The ethnicity of the child is not available in the survey, so the ethnicity of the PCG is taken as a proxy, however, it is unknown whether the child is of the same ethnicity as the PCG. Male gender is both a risk factor for poor literacy (Taylor et al., 2013), and a potential discriminatory variable. Additional questions were asked about family context, that is, the way the family interacts with the outside world. Caregivers were asked about the level of support that they received from their family and friends, and the Parental Stress Scale was used to assess the positive and negative aspects of parenthood. It has four sub-scales; parental rewards, parental stressors, lack of control and parental satisfaction. Previous research has shown that parental stress may affect the child's ability to regulate emotion, which may have a negative

[3]https://www.ucd.ie/issda/data/guiinfant/frequentlyaskedquestions/

effect on child outcomes (Thornton et al., 2013). The variables extracted from this section, and their corresponding variable name, are shown in Table 3.3.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
Gender of study child	aphc02a	-	-	-
PCG ethnicity	apsd53	-	-	-
And what about child. Does he/she belong to any religion?	apsd55a	-	-	-
Child's religious denomination	apsd55b	-	-	-
Age of PCG	-	bphc01b	-	-
Does the PCG have a partner living in the household	adid04	-	-	-
PCG parental stress	-	-	bpc3_stress	-
Number of people in household	aphc00	-	bpc3A4	-
How many separate bedrooms are in the accommodation?	apsd19	-	bpc3J4b	-

Table 3.3: Survey questions on the topic of the family, household and personal information about the child and PCG. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

Socio-demographic information was collected, including language and literacy abilities of the caregivers, religion and ethnicity, and measures of deprivation. The Basic Deprivation Scale is a widely accepted measure of poverty which consists of 11 measurements of poverty across multiple areas including food, clothing, furniture, debt and social life (Thornton et al., 2013). Features were selected for this model that measured the PCG's literacy and education, as low parental education has been linked to poor literacy outcomes in children (Law et al., 2009; Wallace et al., 2015; Taylor et al., 2013), as has parental stress levels and if the mother is a non-Native English speaker (Taylor et al., 2013) and poverty (Schoon et al., 2002). See Table 3.4 for full list of

questions included.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
PCG Is English your native language?	apsd45a	-	-	-
PCG Read aloud from a children s storybook in English?	apsd46	-	-	-
PGC Read and fill out forms in English?	apsd47	-	-	-
PCG Highest level of educational achievement	apsd43a	-	-	-
PCG current economic status	apsd20a	bpsd20a	p1empw3	-
Family's Social Class	adsd56a	bdsd56a	b3_hsdclass	-
Degree of ease or difficulty is the hsd able to make ends meet?	apsd42j	-	-	-

Table 3.4: Survey questions on the topic of the family education and social class. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

Reading to the child, listening to the child read, and the availability of books in the home have been suggested in previous studies to promote child literacy (Justice & Pullen, 2003), Speaking to the child has also has been shown to encourage the acquisition of vocabulary (Thornton et al., 2013). Questions around these topics were included, see Table 3.5.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
Do you talk to child while you are busy doing other things?	apfc04	-	-	-
About how many children's books does child have access to in your home now, including any library books?	-	-	bpc3E7	-
How often would you (PCG) visit the library with child?	-	-	bpc3E3ac	-
How often would you (PCG) listen to child read?	-	-	bpc3E3ad	-
How often would you (PCG) read to child?	-	-	bpc3E3ae	-

Table 3.5: Survey questions on the topic of the literacy in the household and access to books. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

The child's general health including medical issues surrounding physical and intellectual disabilities, visual and hearing issues and issues with using their hands and arms were included in the questionnaire. For wave 1, the variables also include prenatal care and birth. The PCG was asked whether they had concerns about the child's health or language development. Although not a replacement for assessment by a medical professional, this has been shown to be a valid measure (Thornton et al., 2013). While there is a lot of overlap between the areas of health and development, they have been separated here for ease of explanation. Several variables were also included that measured the development of the child. In wave 1, the Ages and Stages Questionnaire (ASQ) and Infant Characteristics Questionnaire (ICQ) scales were used to evaluate early infant development. The ACQ is a parent reported measure of child development that covers five developmental domains, communication, gross motor, fine motor, problem solving and personal/social and is an internationally recognised

measures of child outcomes at this age. There are different sets of tests depending on the infant's age between 4 and 60 months. As the children in the study were approximately 9 months, the 8, 10 and 12 month studies were administered, see Table 3.6. The parents respond *yes*, *sometimes* or *no* to a series of questions about the child, such as, *Does the child pick up a toy and put it into his mouth?* which are awarded 10, 5 and 0 points respectively. These are summed to give an overall score, which was marked as pass/fail. If the 8 month test was passed, the 10 month was administered, if that was passed, the 12 month test was administered (Thornton et al., 2013). The ICQ measures the caregiver's perception of the child's temperament. A child's temperament can influence their relationship with their caregivers (Thornton et al., 2013).

ASQ Question	**8 month**	**10 month**	**12 month**
ASQ Problem Solving	adcd04b	adcd05b	adcd06b
ASQ Gross Motor	adpd04b	adpd05b	adpd06b
ASQ Fine Motor	adpd08b	adpd09b	adpd10b
ASQ Communication	aded09b	aded10b	aded11b
ASQ Personal-Social	aded13b	aded14b	aded15b

Table 3.6: ASQ Survey Questions and variables for tests administered at 8 months, 10 months and 12 months.

As the child ages, different measures of development are taken into account. In wave 2 and 3, two subtests of the British Ability Scales are preformed, the picture similarities and naming vocabulary tests. These tests are a good measure of the child's reasoning capacity, problem solving skills and English language vocabulary (McCrory, Williams, Murray, Quail, & Thornton, 2013; J. Williams, Thornton, Murray, & Quail, 2019). Measurement of child's motor skills in wave 2 were evaluated from simple observations. Gross motor skills were determined from whether the child could stand on one leg for two seconds or more and throw a ball overhead. Fine motor skills were determined from whether the child could draw a straight line and hold a pencil in a

pincer grip.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
Child's weight at birth	apcb05	-	-	-
Do you have any concerns about how child talks and makes speech sounds?	-	bpch48	bpc3C21	-
Picture Similarities	-	bdcd09d	b3_pspercentile	-
Naming Vocabulary	-	bdcd10d	b3_nvpercentile	-
ASQ	✓	-	-	-

Table 3.7: Survey questions on the topic of the child's development. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used. For the scales, the overall scale type is shown. - indicated that this measure is not included in the wave, ✓indicates that it is included.

In wave 2 and 3, questions from the Strengths and Difficulties Questionnaire SDQ were used to measure the child's psychological adjustment across a range of behavioural and social domains including emotions, conduct and behaviour, hyperactivity or inattention, problems with peer relationships and kindness to others. These were combined to give a total difficulties score. In wave 3 the SDQ was taken by the child's teacher if the child attended school (McCrory et al., 2013; J. Williams et al., 2019).

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
SDQ Caregiver				
Emotional subscale	-	-	b3_sdqemotional	-
Conduct subscale	-	-	b3_sdqconduct	-
Hyperactivity subscale	-	-	b3_sdqhyper	-
Peer problems subscale	-	-	b3_sdqpeerprobs	-
Prosocial subscale	-	-	b3_sdqprosocial	-
Total difficulties score	-	-	b3_sdqtotaldiffs	-
Impact score	-	-	b3_sdqimpact	-

Table 3.8: SDQ survey questions. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

Three additional measures of child characteristics were used, taken from the similar longitudinal study Growing Up in Australia: Longitudinal Study of Australian Children (LSAC), Sociability, persistence, which measures the child's self-regulation and reactivity, which measures the duration of a child's reactions (McCrory et al., 2013; J. Williams et al., 2019). Finally, the Social Skills Improvement System Rating Scales (SSIS) measure the child's ability to interact with adults and peers in the areas of assertion, responsibility, empathy and self-control (Murray, Williams, Quail, Neary, & Thornton, 2015). Questions measuring concerns the PCG had about the child's speech, the ASQ, LSAC and SDQ measurements were included in the model as it has been suggested that issues with these can lead to negative educational outcomes in children (W. Williams, Latif, Hannington, & Watkins, 2005; Armstrong et al., 2018, 2016).

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
LSAC temperament measure				
Persistence Subscale	-	-	b3_persistence	-
Sociability Subscale	-	-	b3_reactivity	-
Reactivity Subscale	-	-	b3_sociability	-
SSIS				
Assertion Subscale	-	-	b3_assertion	-
Responsibility Subscale	-	-	b3_responsibility	-
Empathy Subscale	-	-	b3_empathy	-
Selfcontrol Subscale	-	-	b3_selfcontrol	-

Table 3.9: LSAC and SSIS survey questions. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

There are different paths through the questionnaire depending on answers given. One such question in wave 3 is whether the child is in preschool, primary school, or neither. Identical questions were asked to parents on each path, and the answers saved to different variables depending on the path, see Table 3.10. These will be manually collated. The Elmen Childcare Scales were used to measure the quality of childcare from a parent's point of view. The Rich Environment & Activities Scale measures the richness of the environment in the child's school or preschool, with questions such as whether there are lots of creative activities, toys, books and music for the child. The Quality of childcare Scale measures the quality of care in the child's preschool (J. Williams et al., 2019). The scales and variable names for each path are shown in Table 3.10.

Study Question	**Preschool Path**	**School Path**
Have you availed of the free preschool year?	bpc3G28	bpc3G47a
How often has child complained about school/preschool?	bpc3G51a	bpc3G14a
How often has child said good things about school/preschool?	bpc3G51b	bpc3G14b
How often has child looked forward to going to school/preschool?	bpc3G51c	bpc3G14c
How often has child been upset or reluctant to go to school/preschool?	bpc3G51d	bpc3G14d
Rich Environment & Activities Scale Combined	bpc3_richenviron_g32	bpc3_richenviron_g52
Quality of Child Care	bpc3_qualchildcare_g32	bpc3_qualchildcare_g52

Table 3.10: Identical questions asked on the primary school and preschool questionnaire paths. This occurs in Wave 3 only.

As the study child is 5 years of age in wave 3, they will generally have started preschool or school, so the opinion of the teacher can be measured. In the Achievement Scales measure, the child's teacher is asked to assess the child in the following areas; disposition and attitude, language for communication and thinking, linking sounds and letters, reading and numeracy (Murray et al., 2015). Finally, questions were included around the child's educational experience to date, whether they has attended or were attending preschool, which has been suggested to positively affect educational outcomes (Law et al., 2009), their PCG's and teacher's evaluation of their skills and the child's own feelings about education.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
What class is study child in?	-	-	b3_TC4	-
Total Teacher Report				
Language	-	-	b3_TC8b_language	-
Linking	-	-	b3_TC8c_linking	-
Reading	-	-	b3_TC8d_reading	-
To child's teacher: In so far as your professional experience allows, please rate the Study Child's performance in English in relation to all children of this age (not just in their present class or, even, school):				
Speaking and listening	-	-	b3_TC9a	-
Reading	-	-	b3_TC9c	b5_tc12c
Writing	-	-	b3_TC9e	b5_tc12e

Table 3.11: Survey questions on the child's education performance as reported by their teacher. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

Potentially Discriminatory Variables

Membership of the travelling community is not available in this more general dataset. Due to low numbers of responses, this data was shielded to protect the participant's anonymity. Marital status and family status do not apply to children. Sexual orientation is unavailable in this dataset and may not apply to young children. As all the children are of a very similar age, discrimination based on age will be excluded. As disability is a much more complex issue, it is outside of the scope of this work and so will be excluded. Therefore gender, ethnicity (used as a proxy for race) and religion will be chosen as potentially discriminatory variables. Each of these chosen variables are shown in Table 3.12. The responses from wave 1 only will be used. It is assumed that these responses will remain consistent for each child in subsequent waves.

Study Question	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 5**
Gender of study child	aphc02a	-	-	-
PCG ethnicity	apsd53	-	-	-
And what about child. Does he/she belong to any religion?	apsd55a	-	-	-
Child's religious denomination	apsd55b	-	-	-

Table 3.12: Potentially discriminatory variables. - indicates that the variable was not present in a wave, or was present but not used. The variable name is provided where the variable was used.

3.3 Initial Data Preparation

The required variables will be extracted from each dataset and waves 1, 2, 3 and 5 joined to produce a single row per participant. The data will then be evaluated for missing data.

3.3.1 Missing Data

Data can be missing due to several reasons. Firstly, as mentioned above, an entire section of the questionnaire can be missing (Curran, Molenberghs, Fayers, & Machin, 1998), e.g. if the PCG, SCG or teacher did not respond to a wave. The overall figures for this can be seen in Table 3.1. Secondly, the participant may have refused to answer a question or didn't know the answer to a question in an otherwise complete survey (Fayers, Curran, & Machin, 1998). Thirdly, as is common in surveys measuring a range of life experiences, there may be missing data due to the survey path (Holman, Glas, Lindeboom, Zwinderman, & De Haan, 2004), e.g. wave 3 has multiple paths available depending on whether the child is currently attending preschool or school. There are also some questions where a lack of response indicated a *No* response, e.g. *Do you have any of the following concerns about your child*, where the PCG was asked to tick all that applied, and leave those that did not apply blank. The refined dataset

will be examined for missing data.

Missing data has two main effects. Firstly, loss of information in extreme cases could mean that there is insufficient data from which to draw conclusions. Secondly, the data could become imbalanced. If participants from particular groups are less likely to respond, the bias could be introduced into the dataset and the results could be misleading (Fayers et al., 1998). As the GUI dataset is relatively large, the focus here will be on the latter.

3.3.2 Manual Imputation

When data is missing due to the survey path, it is generally not true missing data as the value can be deduced from context in that or other questions. In the case where the a participant was asked to tick all that apply, a blank should indicate *No/Not Present*. In the case that there is missing data from the path of the questionnaire, the value could be clear from a previous question e. g. in the question pair *Does the child belong to a religion?* and *Which religion?*, if the child has no religion, the answer to the second question will be missing. This can however be manually imputed with an additional value; *4=No Religion*. If a response is *Refusal* and *Don't Know*, this will be transferred to the second question. The ASQ tests will be collapsed into a single measure per ASQ type, see Table 3.6. Multiple versions of the same questions were asked in wave 3, depending on whether the child was currently attending preschool or primary school, see Table 3.10. These were also collapsed into a single question.

3.3.3 Additional Data Cleaning

Multiple questions are of the format *Does (a particular feature) apply to the study child?* with valid answers *1=Yes, 2=No*. Questions in this format will be reformatted to a binary answer set *1=Yes, 0=No*. Categorical variables will be one-hot-encoded. An additional variable will be created, bedroom density, the average number of people per bedroom in the accommodation. The original measures of number of people and number of bedrooms in the accommodation will be removed.

All rows with remaining missing values will be removed. The data will be split into 80% training and 20% test data, stratified on gender, ethnicity and religion to ensure that the training and test datasets are balanced with regards to these variables. The test and train sets will then be separated into a predictive fields and target field dataframes. It is not possible to apply the weighting created by the GUI analysts as some participants have been removed and so the dataset is no longer complete. The effect of data removal on each of our protected characteristics will be evaluated.

3.4 Initial Model Building, Training, and Evaluation

An elastic net model using 10-fold cross validation will be created and will be trained on the training data. K-fold cross validation is a method where the model is trained and tested multiple times on different splits of the same training dataset. The training data is randomly split into k equal subsets. In each case, one subset is reserved for testing, the model is trained on the other k-1 subsets and tested on the single reserved subset, and the evaluation metric is calculated. This is repeated for each of the k subsets and the average evaluation metric is calculated across the k values. It is therefore a more robust method of estimating accuracy. While any value of k can be used, k=10 is typical and will be used here. The penalty of the model is controlled by α and so will be varied and tested for multiple values. In the elastic net implementation, $\alpha = 0$ is the ridge penalty and $\alpha = 1$ is the lasso penalty so these will also be tested. As the target variable is continuous, the response type will be set to gaussian. For each value of α, the model will be used to fit the predicted data and the mean squared error will be calculated. Our target variable is standardised to have unit variance before the lambda sequence is computed. The resulting output predicted variables are unstandardised. The best fit will be the value of α that minimises the mean squared error.

3.5 Initial Test for Discrimination

3.5.1 Creating Synthetic Data

We will test for discrimination in our dataset using the 'twin test', where the predicted results are compared for two identical simulated participants who differ only in a protected characteristic. The current datset of predicted variables will be used, hardcoding the gender, ethnicity or religion to a specific value. Firstly, to perform the twin test based on gender, two datasets will be created. In the first set of data, all values in the gender field will be hardcoded to male. In the second, all gender values will be hardcoded to female. In this way, there are two identical populations created that differ only in gender. The same process will be repeated for ethnicity and religion.

It is not possible to reserve a portion of the dataset for the twin test testing as this would result in an imbalance in the remaining dataset.

3.5.2 Evaluating Discrimination

The twin test for each of the protected characteristics will be performed by using the best fit model. For each member of the protected variable (e.g. male and female), the best fit model will be used to predict values for the target, DPRT-R logit score. If any difference is observed in the predicted values, this difference will be tested for statistical significance.

In this case, a series of hypothesis tests will be created examining the difference between the mean DPRT-R logit scores values in each of the twin tests of the form

- H_0: There is no difference in predicted DPRT-R logit score between member of protected group A and member of protected group B.
- H_A: There is a difference in predicted DPRT-R logit score between member of protected group A and member of protected group B.

Each set of predicted values will be inspected for normality using a distribution plot, QQ-plot, skew and kurtosis. Skewness is the measure of symmetry. If a data

is skewed, it is not symmetric about the mean. Kurtotis measures the volume of outliers. If kurtotis is high, the data is heavy tailed when compared with the normal distribution. If the standardised skew and kurtotis fall within the accepted range of ± 2, it can be assumed that the distribution is normal. If either fall outside ± 2, the outliers in the data will be examined. If 95% of the standardised values lie between ± 3.29 (as the number of values is greater than 80), the distribution can be approximated to normal (Field, Miles, & Field, 2012).

Homogeneity of variance is then checked. The F test has a null hypothesis that the variances of the two samples are equal and an alternative hypothesis that the variances of the two samples are not equal. If the p-value of this test is $p<0.05$, there will be enough evidence to reject the null hypothesis and the variances of the samples will be considered to be different. If the p-value of this test is not less than 0.05, there will not be enough evidence to reject the null hypnosis and the variances of the samples will be considered to be equal.

A t-test will be used to compare the mean value of each set of predicted values. If the variances should not be treated as equal, the Welch two sample t-test will be used. If the variances should be treated as equal, the ordinary t-test will be used. In both cases, the t-test will be unpaired as these are different populations of simulated participants. For the t-tests, an α level 0.05 was chosen. If the p value of the t-test is found to be less than 0.05, there will be enough evidence to reject the null hypothesis that there is no difference in predicted DPRT-R logit score between member of protected group A and member of protected group B.

The effect size will also be considered. A difference may be statistically significant, but may give a small effect size. Cohens convention on effect size will be used, where $d=0.2$ is considered a small effect size, 0.5 is considered a medium effect size and 0.8 is considered a large effect size (Brase & Brase, 2001).

If a statistically significant difference is found in any of the twin tests, it can be said that the model discriminates based on one or more protected groups.

3.6 Further Data Preparation

3.6.1 Data Imputation

In the previous section, participants with any missing data were removed. An alternative to removal of missing data is imputation. It is necessary to impute all missing data including the target variable, as removal of any rows will result in the weighting being unusable. For that reason, additional fields from wave 5 relating to literacy will be added. These will be used for imputing the target variable only, they will not be used in creating the elastic net model. Data will be imputed before religion and ethnicity are one-hot-encoded to ensure that all fields within these are mutually exclusive. The quality of the imputation will then be evaluated. The imputations will then be checked with diagnostic plots, density plots to check whether all imputed values are realistic.

It is possible to pool the imputations, however as some variables will need to be combined or removed, and all fields multiplied by the wave 5 weight, each imputation will be extracted separately.

3.6.2 Further Data Cleaning

The further data cleaning outlined in Section 3.3.3 will be repeated on the dataset with imputed values. The weighting from wave 5 will also be applied to the dataset.

3.7 Model Creation and Evaluation of Discrimination

A model as described in Section 3.4 will be trained on the dataset with imputed values and weights applied. The synthetic data described in Section 3.5.1 will be used to test for discrimination. The same set is used as not to include additional variability to the test.

3.8 Software Used

The elastic net model was created using the *glmnet* package in *R*. Data was imputed using *R's MICE* package.

3.9 Conclusion

In this chapter, the GUI study and dataset was described. Additionally, the methods used to clean the data and deal with missing data were outlined. Finally, the methods to create and evaluate the elastic net model, and to test for and identify discrimination were described. In the next chapter, the implementation of these methods will be described.

Chapter 4

Results, Evaluation and Discussion

4.1 Introduction

In this chapter, all methods will be implemented as outlined in chapter 3. After initial data preparation and all possible data has been deductively imputed, all remaining rows with missing data are dropped. This data is not weighted due to the missing data. An elastic net model is created and evaluated. The best fit model is tested for discrimination using synthetic datasets generated from the original dataset. A second method for imputation is then used to recreate the full dataset and the weighting is applied. A model is again created, fitted and evaluated, and checked for discrimination as before. The results will then be discussed.

4.2 Data Processing

There are 3057 variables in waves 1 to 5 combined. Excluding wave 4, there are 2962 variables. The variables in wave 5 were examined for a measure of literacy in children. Each of the 2000 variables in wave 1 - 3 combined were considered for inclusion, taking into account whether there was evidence to support the variable's relationship to either literacy or our protected variables, and whether there was enough data available to warrant inclusion.

4.2.1 Loading the Data

The dataset was obtained through the Irish Social Science Data Archive[1]. Each wave is stored in a different file, available in .sas, .sas7bdat, .sav, and .dta formats. Each wave 1, 2, 3 and 5 was loaded into a dataframe in R. Each dataframe was then joined together using the unique id field, which corresponded to an individual participant. Next, all relevant data was extracted using the variable names listed in Tables 3.3, 3.4, 3.5, 3.7, 3.11, 3.10, 3.6, 3.9 and 3.12.

4.2.2 Missing Data

As described in Section 3.3.1, data can be missing for several reasons. If an entire section is missing in the GUI data, this is generally indicated by a response of *Nan* (not a number). If a participant refused to answer a question (response *Refusal*), or didn't know the answer (response *Don't Know*) to an otherwise complete survey, there were multiple ways in which this was coded. Therefore, all fields of interest had to be examined individually and manually corrected if needed. *Refusal* and *Don't Know* were generally indicated with a response of 8 and 9 respectively for questions with under 8 potential responses, and 88 and 99 respectively for questions with over 8 potential responses. This however is not a clear rule, in waves 2 and 3, *Refusal* was generally indicated with a response of 98 for questions with over 8 potential responses, as in the case for the measure of number of bedrooms in the household in wave 3 and the PCG's current economic status in wave 2. This is in contrast to wave 2, where missing data in PCG's economic status is indicated by a *Nan*. In a single case for the family's social class in wave 3, unknown values are indicated with a value of 666. There are also multiple cases where the lack of response to a particular question in indicated by *NA*, sometimes in addition to the typical 8/9 or 88/99, as in the case of the child's weight at birth in wave 1, the picture similarities and naming vocabulary tests in wave 2 and wave 3 and the SSIS, LSAC and SDQ measures in wave 3, missing data is indicated with *Nan*. There is no differentiation between *Refused* and *Don't Know*.

[1] http://www.ucd.ie/issda/data/growingupinirelandgui/

Not applicable answers to questions to the teacher about the child's ability in reading and writing in English in wave 3 and 5 are represented by the value 6. If there is a lack of response to the survey path, or if a question is not applicable this was generally also indicated with a *Nan*, indistinguishable from the *Nan* responses described above. Missing data arising from each of these three situations will be handled differently.

In wave 3, the child's teacher was asked to fill in a questionnaire. The teacher did not respond to the survey in 442 cases, which contributes to the overall number of missing data points, see Table A.7.

Variables who's missing data is accounted for by *Refusal* or *Don't Know* responses, and count of missing data, can be found in Appendix A.

Deductive Imputation

The questions asked to each participant often depends on the outcome to other questions. Non-applicable questions and answers tend to have a null value. However, this often does not indicate an unknown response, the response can often be discerned from context or from other questions. In the 'Family Personal' section, there is an initial question asking whether the study child belongs to a religion, and a subsequent question asking which religion, see Table 4.1. If the response to the initial question is *No*, the child is not a member of a religion, then the response to the second question, which religion, is *Nan*. These unknowns were manually imputed to an additional value 4 = *No Religion*. If the answer to the first question was *Refused* (8) or *Don't Know* (9), this was carried over to the second question. These true missing values will be examined in Section 4.2.2. The first question was then removed as all the information had been absorbed into the subsequent question and so it was unnecessary.

	Missing Data			To Impute	
Wave 1 Survey Question	**S**	**R**	**DK**	**N**	**%**
And what about child. Does he/she belong to any religion?	0	0	4	4	0.1%
Child's religious denomination	519	0	23	23	0.3%

Table 4.1: Child's religion collapsed to single question. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

There was a similar set of questions about the religion of the PCG. Additionally, there was a question asking how long ago the PCG moved to Ireland. If the PCG was born in Ireland, the response to this question was *Nan*. As the original valid responses ranged from 1=Within the last year to *5=More than 20 years ago*, a new value was manually added *6=Born in Ireland*. The question *Was the PCG born in Ireland* was then removed as it included no additional information.

In wave 3, there are two mutually exclusive paths through the survey depending on whether the child is currently in preschool or primary school, see Table 4.2. In many cases, an identical question was asked on each path. Such school/preschool pairs are *Did you avail of free preschool*, *How often has the child complained about school/preschool*, *How often has the child said good things about school/preschool*, *How often has the child looked forward to school/preschool*, *How often has the child been upset or reluctant to go to school/preschool*. In all cases, the school and preschool versions of the questions were combined into a single question, and the original question was dropped. Whether the child was in school or preschool is captured in the question *Has child started Junior Infants in primary school?*.

	Missing Data			To Impute	
Wave 3 Survey Question	**S**	**R**	**DK**	**N**	**%**
Did you avail of the free preschool year for the Study Child?	2058	0	0	0	0%
Have you availed of the Free Preschool Year for the Study Child?	5022	0	0	0	0%
Rich Environment & Activities Scale	2058	0	0	0	0%
Rich Environment & Activities Scale	5022	0	0	0	0%
Quality of Child Care	2058	0	0	0	0%
Quality of Child Care	5022	0	0	0	0%
How often has child:					
complained about preschool?	5021	0	0	0	0%
complained about school?	2059	0	0	0	0%
said good things about preschool?	5021	0	0	0	0%
said good things about school?	2059	0	0	0	0%
looked forward to going to preschool?	5021	0	0	0	0%
looked forward to going to school?	2059	0	0	0	0%
been upset or reluctant to go to preschool?	5021	0	0	0	0%
been upset or reluctant to go to school?	2059	0	0	0	0%

Table 4.2: Multiple paths collapsed to single question. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

As described above in Section 3.2.3, if a child passes the 8 month ASQ test, the 10 month test is administered. It this was also passed, the 12 month test was administered, see Table B.6. For each of the ASQ tests, the 8, 10 and 12 month tests were collapsed to a single value of the highest level of test passed.

	Missing Data			To Impute	
Wave 1 Survey Question	**S**	**R**	**DK**	**N**	**%**
ASQ Problem Solving 8mth	88	0	0	0	0%
ASQ Problem Solving 10mth	413	0	0	0	0%
ASQ Problem Solving 12mth	515	0	0	0	0%
ASQ Gross Motor 8mth	20	0	0	0	0%
ASQ Gross Motor 10mth	26	0	0	0	0%
ASQ Gross Motor 12mth	28	0	0	0	0%
ASQ Fine Motor 8mth	108	0	0	0	0%
ASQ Fine Motor 10mth	230	0	0	0	0%
ASQ Fine Motor 12mth	251	0	0	0	0%
ASQ Communication 8mth	29	0	0	0	0%
ASQ Communication 10mth	44	0	0	0	0%
ASQ Communication 12mth	75	0	0	0	0%
ASQ Personal-Social 8mth	34	0	0	0	0%
ASQ Personal-Social 10mth	99	0	0	0	1.3%
ASQ Personal-Social 12mth	145	0	0	0	0%

Table 4.3: Count of missing ASQ data in wave 1. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

There are other questions with similar path dependant answers. In the section on family education and literacy, there are questions around whether the PCG is able to read from a children's storybook and fill out forms in English and in their native language. If the PCG's native language is English, the response to the questions around native language is *Nan*. For PCGs who's native language was English, their responses to English language questions was replicated in the native language based questions, as English is their native language. No questions were removed as all

information was still required.

In the family context section, there is a question asking whether the PCG has had a morning/afternoon/evening out in the last fortnight, and the next question asks why. If the answer to the first question is no, then the answer to the second question is *Nan*. We are only interested in the case when they haven't had entertainment as they couldn't afford it, so cases when the answer to the first question is *No* are cast as 0. The first question is removed as it has no additional information.

Regarding the child's health and development, there are several sections where the PCG is asked to 'tick all that apply' regarding concerns they have about the child's health and development; whether the child has difficulty hearing, seeing, using their hands or any developmental delay. In these cases, a lack of response iS indicates by *Nan*, but indicates a *No* response and is cast as such. Additionally, in waves 2 and 3, there is a question asking whether the PCG has any concerns about how the child talks and makes speech sounds and further questions about specific concerns. The answers to these are *Nan* if no concern is present so will be manually cast to 0.

These questions were ultimately not included when developing the model to create a minimal model.

Additional Data Cleaning

After all possible data was manually imputed, what remains is true missing data. There are several methods for dealing with missing data. First, all rows with missing data were removed, leaving 4,428 rows. The religion and ethnicity variables were one-hot encoded as they are categorical variables. Other, two level variables were recoded to be binary. All now redundant variables were dropped, as described above. The variable bedroom density was created by getting the average number of people per bedroom in the accommodation. The original measures of number of people and number of bedrooms in the accommodation were then removed. All ordinal and categorical variables were converted to factors. The final dataset, data type, and all valid answers can be found in Appendix B. The data was split a 80% train and 20% test subsets stratified on gender, ethnicity and religion.

4.3 Initial Model

4.3.1 Model Creation

The Pearson correlations of the predictive features were inspected, see Figure C.1. As there are many predictors, some of which are correlated, an elastic net model was chosen.

The elastic net model was created using the *glmnet* package in *R*. The penalty of the model is controlled by α and will be varied from 0 to 1 in steps of 0.01. 10-fold cross validation is used.

4.3.2 Model Evaluation

The best model chosen was the value for α that minimised the MSE. In the case for the unweighted model with all rows containing nulls removed, this value was α=0.63, with value of MSE=0.5298002. The value of R^2 is 33.0874, indicating that the variables in this model explain 33.0874% of the variance in DPRT-R logit score between children. Though this could be considered low, it is a comparable value to other predictions of literacy (see section 2.2), possibly due to the complex and interlinked reasons behind language and literacy.

For each value of α, multiple values of the regularisation parameter λ are tested, which controls the strength of the penalty.

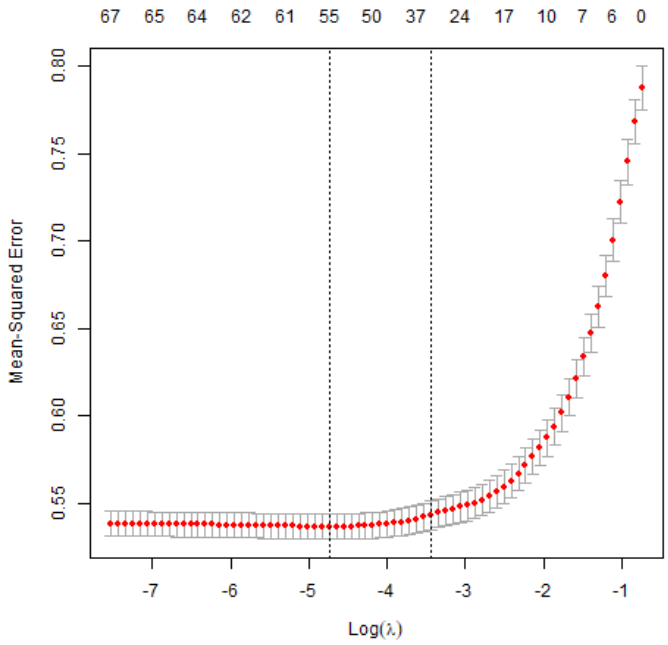


Figure 4.1: Values of the regularisation parameter λ for the initial model. The cross validation curve is shown in red. The upper and lower standard deviations are also shown.

4.4 Test for Discrimination

Synthetic data to test variations in gender, ethnicity and religion was created using the method outlined in Section 3.5.1

Each of these sets of data was used with best fit model to predict a value for DPRT-R for each simulated individual. The mean squared difference between each population set pair was then calculated.

4.4.1 Gender

The mean predicted DPRT-R logit scores for female participants (mean=0.435, sd=0.454) was lower than for male participants (mean=0.45, sd=0.454), with a mean difference of 0.00025

- H_0: Predicted DPRT-R logit scores for female participants are not lower than those predicted for male participants

- H_A: Predicted DPRT-R logit scores for female participants are lower than those predicted for male participants

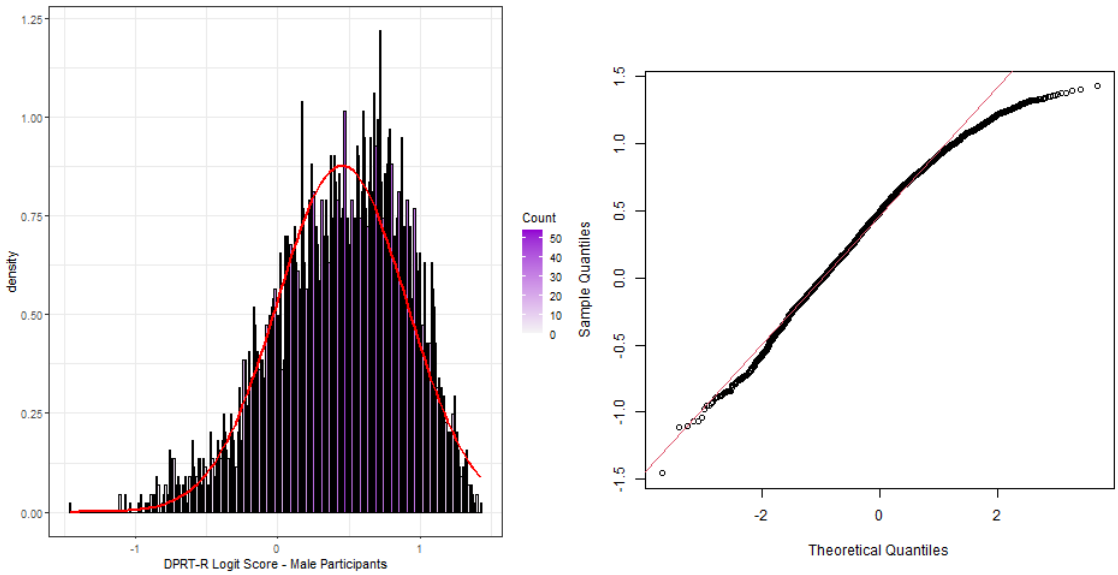
The male and female populations were then inspected for normality. Inspection of the histogram and corresponding normality plot shows that both the male and female distributions appear to conform to a normal distribution. To check this, the standardised normal scores of skew and kurtosis were inspected. Male participants had a standardised kurtotis score of -1.54 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.59 (skew=-0.46, SE=0.04). Female participants had a standardised kurtotis score of -1.54 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.59 (skew=-0.46, SE=0.04). While kurtotis is within the accepted range of ± 2 for both sets of data, skew is not.

Examining Figure 4.3, 0.11% of standardised data points are outside ± 3.29 in each, we can approximate the sample distributions as normal.

A F-test was used to see if the variances of the two samples could be considered to be equal. As the p-value was not less that 0.05 (P=1), there is not enough evidence to reject the null hypothesis and so the variances are considered as equal. Since the variances are equal, a two sample independent t-test was performed to see if the difference in average measure observed was statistically significant. The difference found between male (mean=0.44) and female (mean=0.49) score was found not to be significant (p=0.05091).

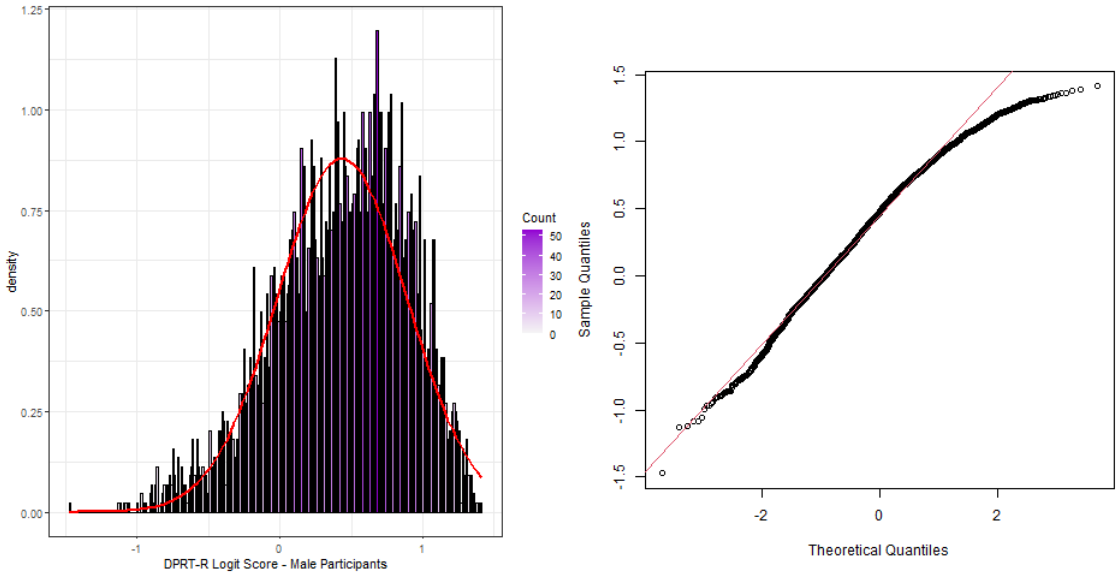
4.4.2 Religion

The mean predicted DPRT-R logit score for participants that are a member of the Catholic religion (mean= 0.438, sd=0.454) were lower than that for participants who had no religion, were members of other christian religions excluding Catholic, and all other religions. These all had identical predictions (mean=0.487, sd=0.454). There



(a) Male Distribution

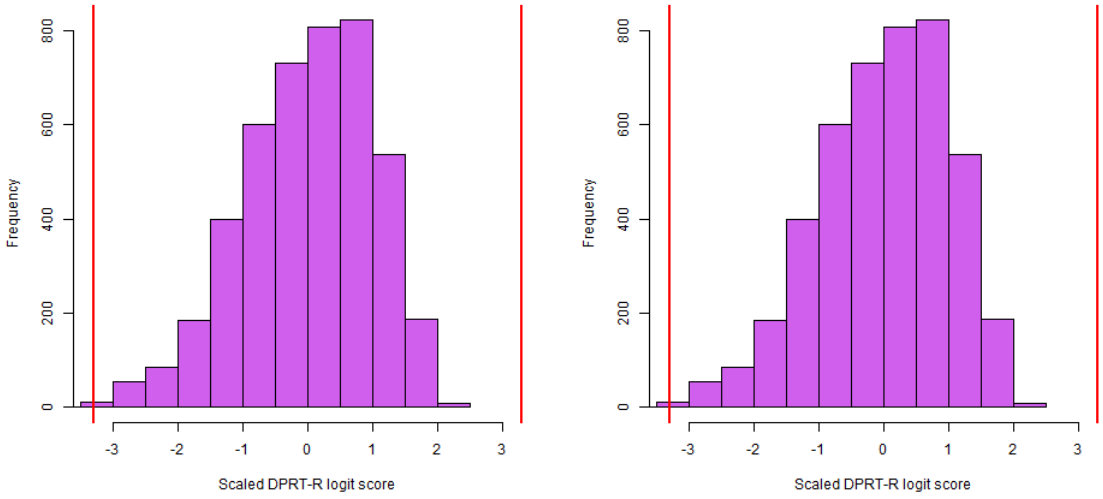
(b) Male QQ-plot



(c) Female Distribution

(d) Female QQ-plot

Figure 4.2: Distribution and QQ plots of DPRT-R scores for male and female simulated participants



(a) Outliers in male data

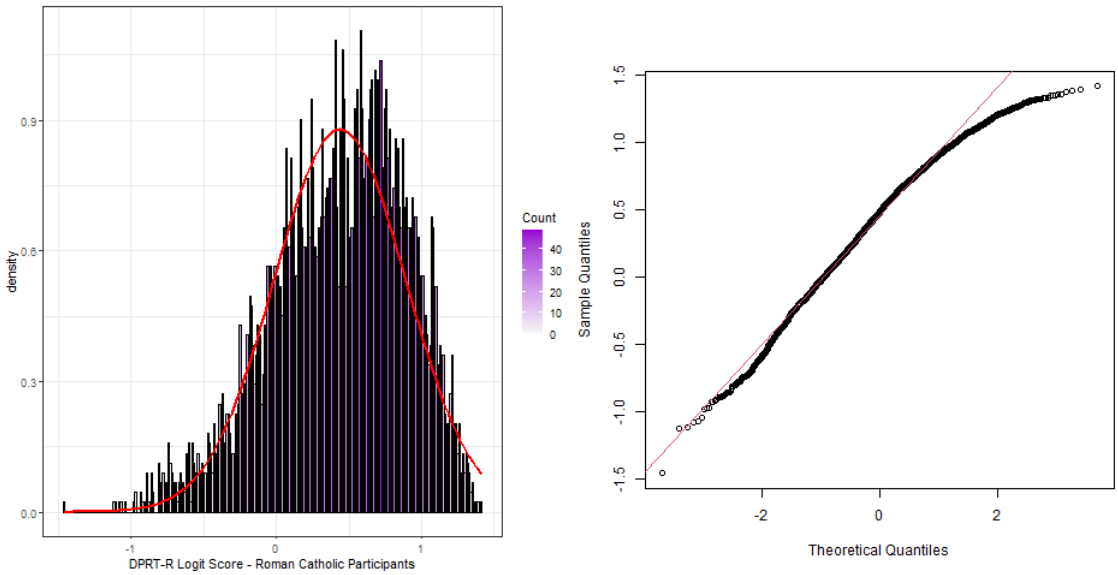
(b) Outliers in female data

Figure 4.3: Outliers in male and female scaled data

was a mean squared difference of 0.00238 between the two sets. As all other religions had identical predictions, they will be grouped under non-Catholic.

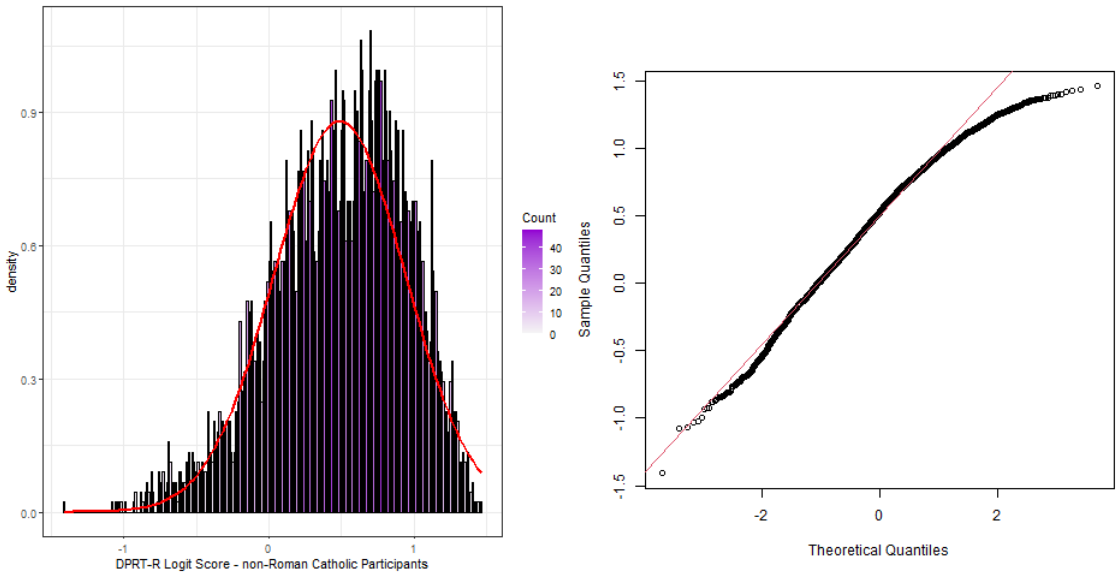
- H_0: Predicted DPRT-R logit scores for participants of the Catholic religion are not lower than those who are a not a member of the Catholic religion
- H_A: Predicted DPRT-R logit scores for participants of the Catholic religion are lower than those who are a not a member of the Catholic religion

The Catholic and non-Catholic were inspected for normality. Inspection of the histogram and corresponding normality plot shows that both Catholic and non-Catholic distributions appear to conform to a normal distribution. To check this, the standardised normal scores of skew and kurtosis were inspected. Catholic participants had a standardised kurtotis score of -1.48 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.69 (skew=-0.47, SE=0.04). Non-Catholic participants had a standardised kurtotis score of -1.48 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.69 (skew=-0.46, SE=0.04). While kurtotis is within the accepted range of ± 2 for both sets of data, skew is not.



(a) Catholic Distribution

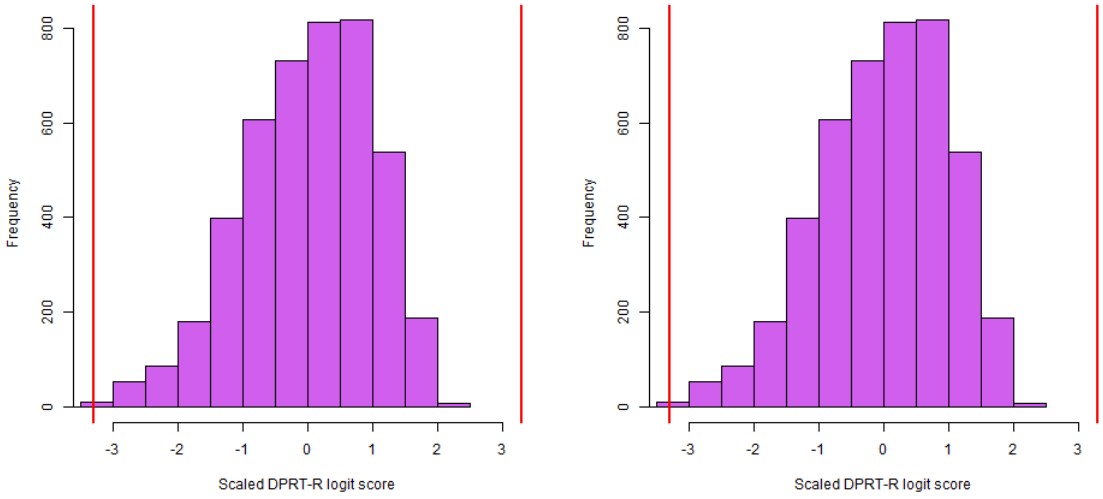
(b) Catholic QQ-plot



(c) non-Catholic Distribution

(d) non-Catholic QQ-plot

Figure 4.4: Distribution and QQ plots of DPRT-R scores for Catholic and non-Catholic simulated participants



(a) Catholic Participants (b) non-Catholic Participants

Figure 4.5: Outliers Catholic and non-Catholic scaled data

Examining Figure 4.5, 0.11% of standardised data points are outside ±3.29 in each, we can approximate the sample distributions as normal.

A F-test was used to see if the variances of the two samples could be considered to be equal. As the p-value was not less that 0.05 (P=1), there is not enough evidence to reject the null hypothesis and so the variances are considered as equal.

As the p-value was found to be p<0.05 (t=-5.0587, p=2.153e-07), there is enough evidence to reject the hypothesis that participants who are a member of the Catholic religion are not predicted to have lower scores that participants who are not members of the Catholic religion. Cohen's statistic was calculated and showed that there is a weak negative effect (d=-0.1075).

Ethnicity

The mean predicted DPRT-R logit score for participants who's ethnicity is white but of non-Irish origin (mean=0.475, sd=0.44) is higher than for every other ethnicity (white and of Irish origin, African or any other black background, Chinese or any other Asian origin, or all other origins including mixed origin), which all had identical

predictions (mean=0.454, sd=0.44). There was a mean squared difference of 0.00129 between the two sets. As all other ethnicities had identical predictions, they will be grouped under 'other ethnicities'.

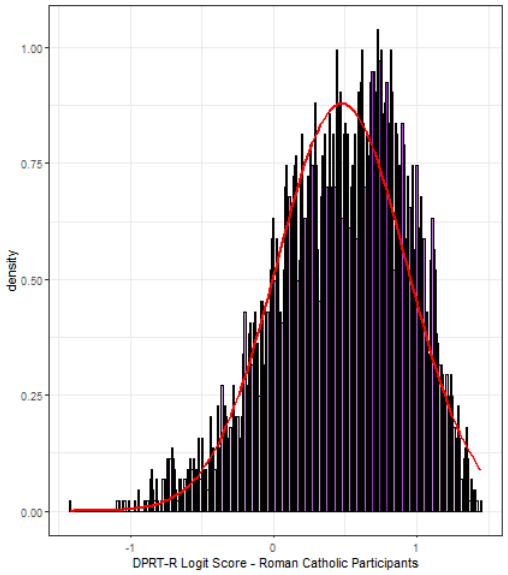
- H_0: Predicted DPRT-R logit scores for participants with white non-Irish origin ethnicity is not higher than those for all other ethnicities
- H_A: Predicted DPRT-R logit scores for participants with white non-Irish origin ethnicity is higher than those for all other ethnicities

Each group was inspected for normality. Inspection of the histogram and corresponding normality plot shows that both distributions appear to conform to a normal distribution. To check this, the standardised normal scores of skew and kurtosis were inspected. Participants with white of non-Irish origin ethnicity had a standardised kurtotis score of -1.46 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.66 (skew=-0.47, SE=0.04). All other ethnicities had a standardised kurtotis score of -1.46 (kurtotis=-0.11, SE=0.07), and a standardised skew of -12.66 (skew=-0.47, SE=0.04). While kurtotis is within the accepted range of ± 2 for both sets of data, skew is not.

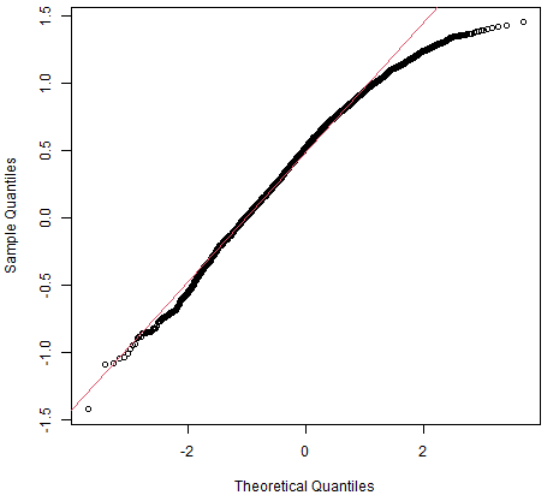
Examining Figure 4.7, 0.11% of standardised data points are outside ± 3.29 in each, we can approximate the sample distributions as normal.

A F-test was used to see if the variances of the two samples could be considered to be equal. As the p-value was not less that 0.05 (P=1), there is not enough evidence to reject the null hypothesis and so the variances are considered as equal.

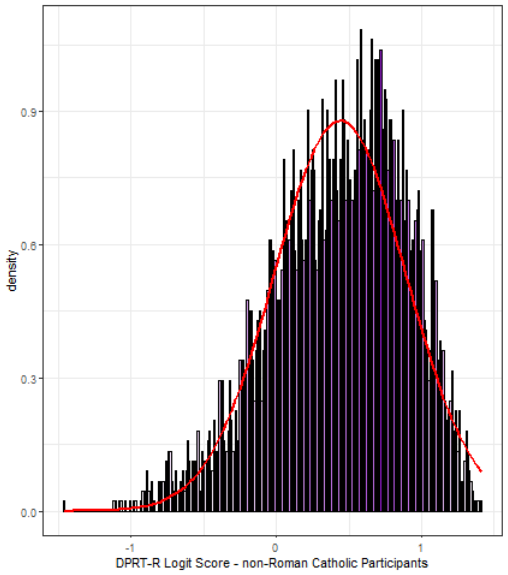
As the p-value was found to be $p<0.05$ (t=3.7255, p=9.808e-05), there is enough evidence to reject the hypothesis that participants who are of white non-Irish ethnicity are not predicted to have higher scores that for participants of all other ethnicities. Cohen's statistic was calculated and showed that there is a weak positive effect (d=0.0792).



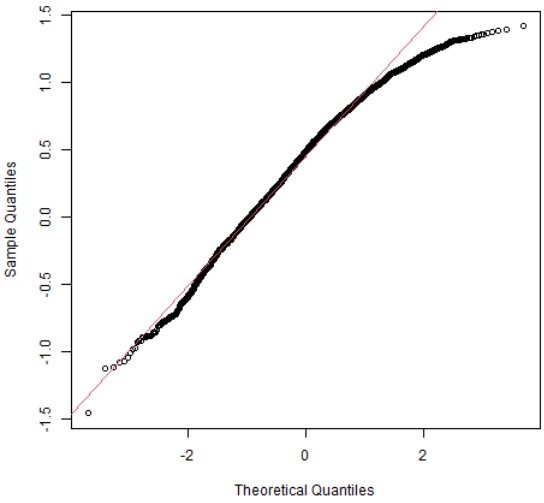
(a) Ethnicity white of non-Irish origin Distribution



(b) Ethnicity white of non-Irish origin QQ-plot

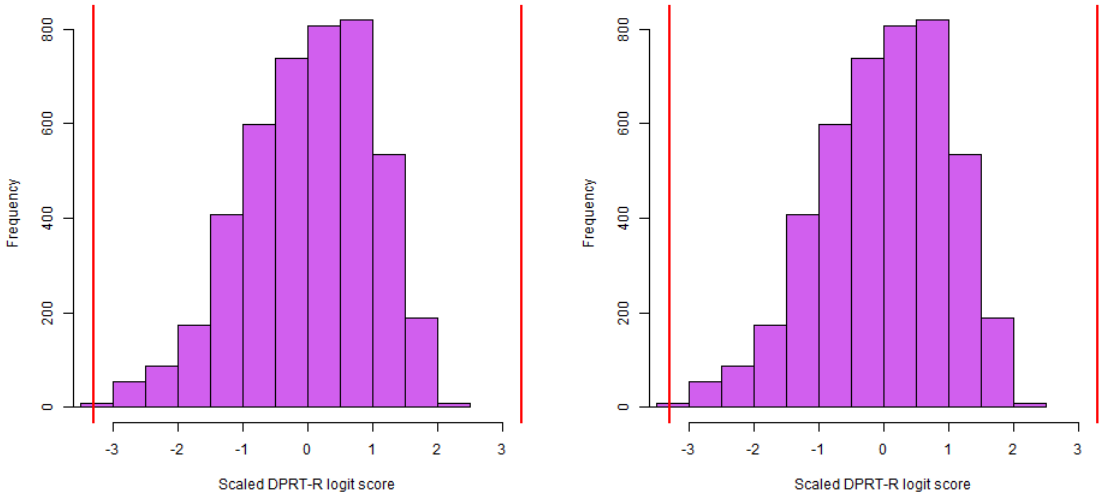


(c) All other ethnicities Distribution



(d) All other ethnicities QQ-plot

Figure 4.6: Distribution and QQ plots of DPRT-R scores



(a) Ethnicity white of non-Irish origin

(b) All other ethnicities

Figure 4.7: Outliers in scaled data

4.5 Model with Imputed Data

4.5.1 Imputing Data

As an alternative to removing all rows with a null value as in Section 4.2.2, the missing data can be imputed. The full dataset that includes all participants who responded to surveys in waves 1, 2, 3 and 5 was split into 80% train and 20% test sets, stratified on gender, ethnicity and religion. The weighting in wave 5 was additionally included when splitting the data so that the correct weight could be applied to its corresponding row. The weighting however was not included in the imputation. Train and test datasets were separately imputed to ensure that the test set was a true test set with no influence from the training set. Missing values were all cast to *NA*. The missing data was imputed using the *MICE* package in *R* (Buuren & Groothuis-Oudshoorn, 2010). Generally, the data from multiple imputations can be applied to a model and the results pooled. However, this is not supported for the glmnet method for elastic net models, so a single imputation for each of the train and test datasets was performed.

4.5.2 Further Data Cleaning

After the data was imputed, similar cleaning to 4.2.2 was applied. Religion and ethnicity were one-hot-encoded and the bedroom density variable was created. All unrequired fields were dropped. The weighting was applied to each field. This has the additional effect of converting every variable to a numeric variable.

4.5.3 Model

An elastic net model was created as described in Section 4.3.1. The value for α that obtained the lowest MSE (5.126078) was α=0.58, mid way between ridge and lasso. The value for R^2 is 27.4880, indicating that the variables in the model explain 27.488% of the variance in DPRT-R scores between children. Thought the MSE is slightly lower than the first model tested, the variance is also lower.

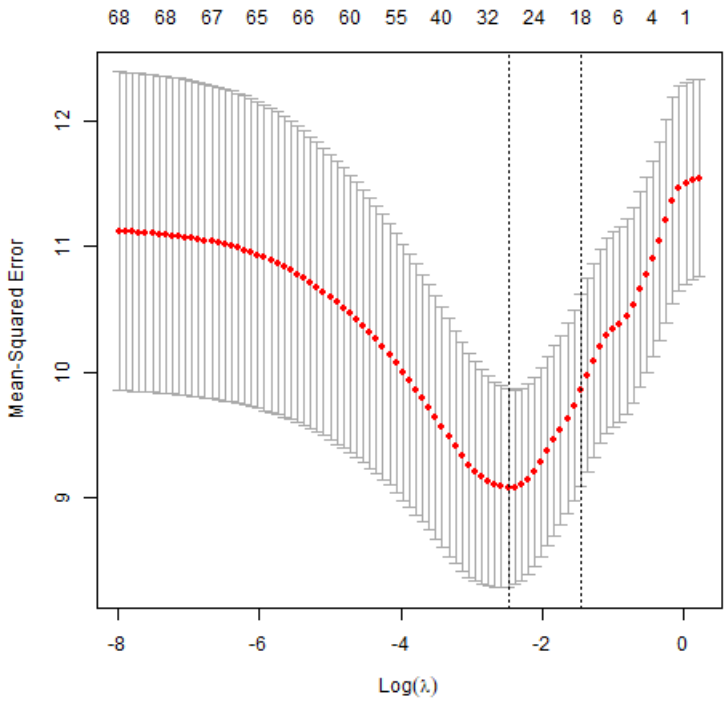


Figure 4.8: Values of the regularisation parameter λ. The cross validation curve is shown in red. The upper and lower standard deviations are also shown.

4.5.4 Test for Discrimination

As before, this model will be tested for discrimination in gender, ethnicity and religion. The same synthetic data was used as in Section 3.5.1 to minimise the additional variability introduced into the data. Each of these sets of data was fitted to the best fit model for the extended dataset to predict a value for DPRT-R for each simulated individual. The mean squared difference between each population set pair was then calculated.

Gender

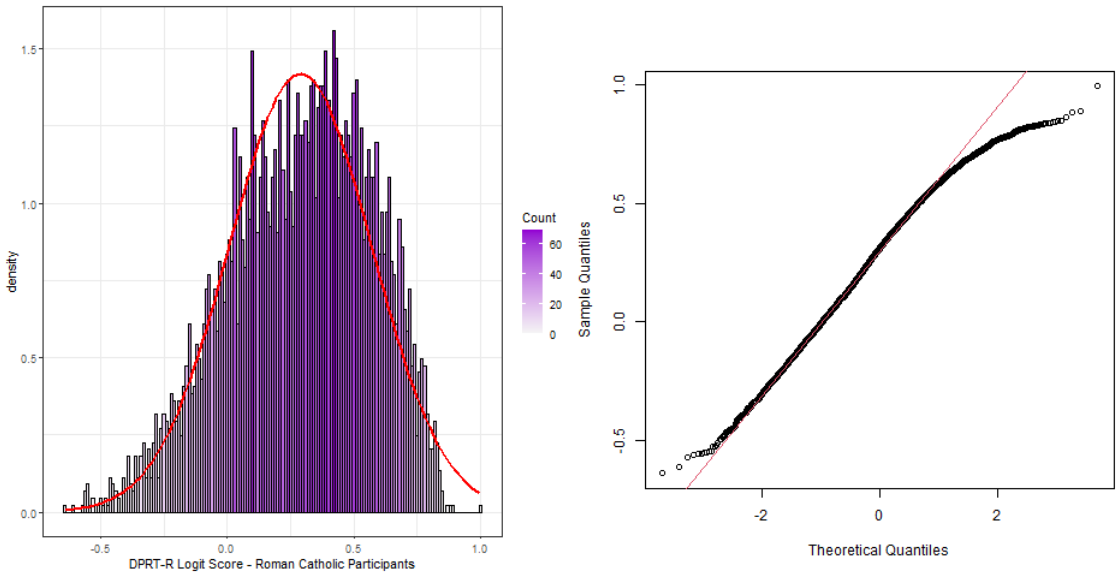
The mean predicted DPRT-R logit scores for female participants and male participants was equal (mean=0.292, sd=0.282) , as was the score predicted for each simulated 'twin'. We can conclude that there is no direct discrimination on the basis of gender in this model for this simulated data.

Religion

In contrast to the previous model, the mean predicted DPRT-R logit score for participants that are a member of the Catholic religion (mean= 0.289, sd=0.281) were higher than that for participants who had no religion, were members of other christian religions excluding Catholic, and all other religions. These all had identical predictions (mean=0.234, sd=0.281). There was a mean squared difference of 0.00306 between the two sets. As all other religions had identical predictions, they will be grouped under non-Catholic.

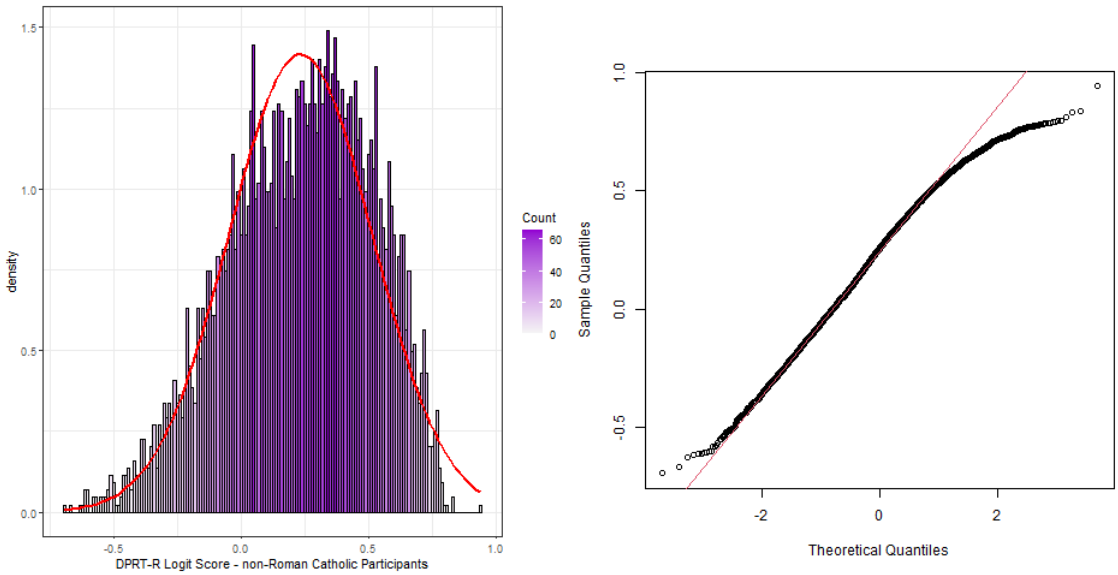
- H_0: Predicted DPRT-R logit scores for participants of the Catholic religion are not higher than those who are a not a member of the Catholic religion
- H_A: Predicted DPRT-R logit scores for participants of the Catholic religion are higher than those who are a not a member of the Catholic religion

The Catholic and non-Catholic groups were inspected for normality



(a) Catholic Distribution

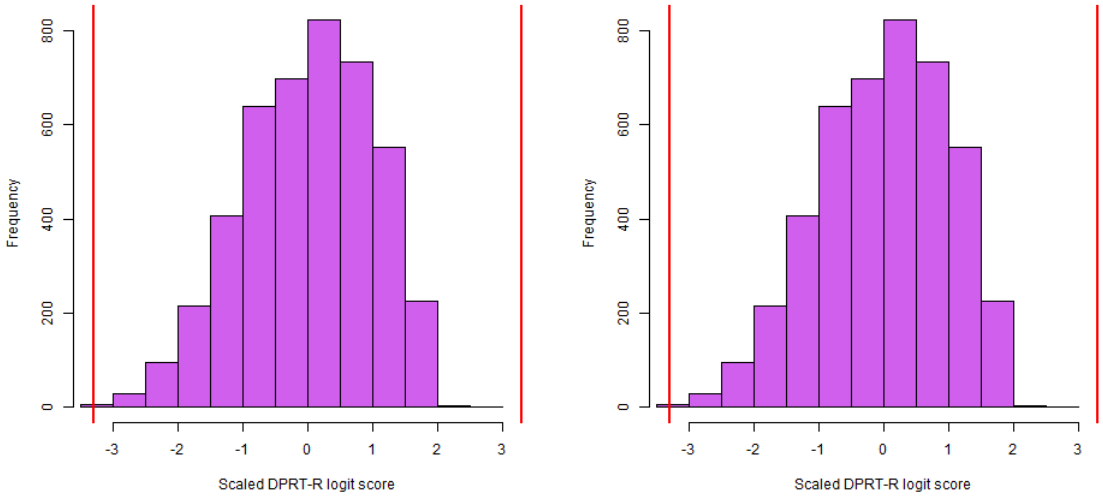
(b) Catholic QQ-plot



(c) non-Catholic Distribution

(d) non-Catholic QQ-plot

Figure 4.9: Distribution and QQ plots of DPRT-R scores for Catholic and non-Catholic simulated participants



(a) Outliers in Catholic Participants (b) Outliers in non-Catholic Participants

Figure 4.10: Outliers Catholic and non-Catholic scaled data

Inspection of the histogram and corresponding normality plot shows that both Catholic and non-Catholic distributions appear to conform to a normal distribution. To check this, the standardised normal scores of skew and kurtosis were inspected. Catholic participants had a standardised kurtotis score of -5.81 (kurtotis=-0.43, SE=0.07), and a standardised skew of -8.92 (skew=-0.33, SE=0.04). Non-Catholic participants had a standardised kurtotis score of 5.81 (kurtotis=-0.43, SE=0.07), and a standardised skew of -8.92 (skew=-0.33, SE=0.04). Neither skew nor kurtotis were within the accepted range of ± 2. Outliers will therefore be examined.

Examining Figure 4.10, 0.02% of standardised data points are outside ± 3.29 in each, we can approximate the sample distributions as normal.

A F-test was used to see if the variances of the two samples could be considered to be equal. As the p-value was not less that 0.05 (P=1), there is not enough evidence to reject the null hypothesis and so the variances are considered equal.

As the p-value was found to be $p<0.05$ (t=9.2436, p=2.2e-16), there is enough evidence to reject the hypothesis that participants who are a member of the Catholic religion are not predicted to have lower scores that participants who are not members

of the Catholic religion. Cohen's statistic was calculated and showed that there is a weak negative effect (d=0.196).

Ethnicity

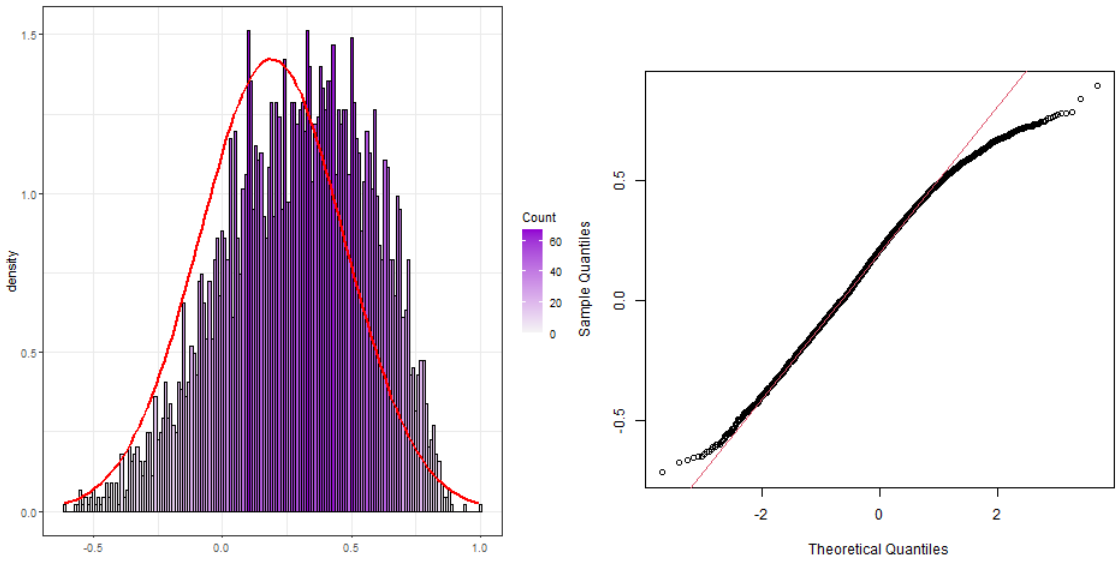
Again in contrast to the previous model, the mean predicted DPRT-R logit score for participants who's ethnicity is of African or other black background (mean=0.191, sd=0.28) is lower than for every other ethnicity (white and of Irish and non-Irish origin, Chinese or any other Asian origin, or all other origins including mixed origin), which all had identical predictions (mean=0.294, sd=0.28). There was a mean squared difference of 0.00129 between the two sets. As all other ethnicities had identical predictions, they will be grouped under other ethnicities.

- H_0: Predicted DPRT-R logit scores for participants with ethnicity of African or other black background is not lower than those for all other ethnicities
- H_A: Predicted DPRT-R logit scores for participants with ethnicity of African or other black background is lower than those for all other ethnicities

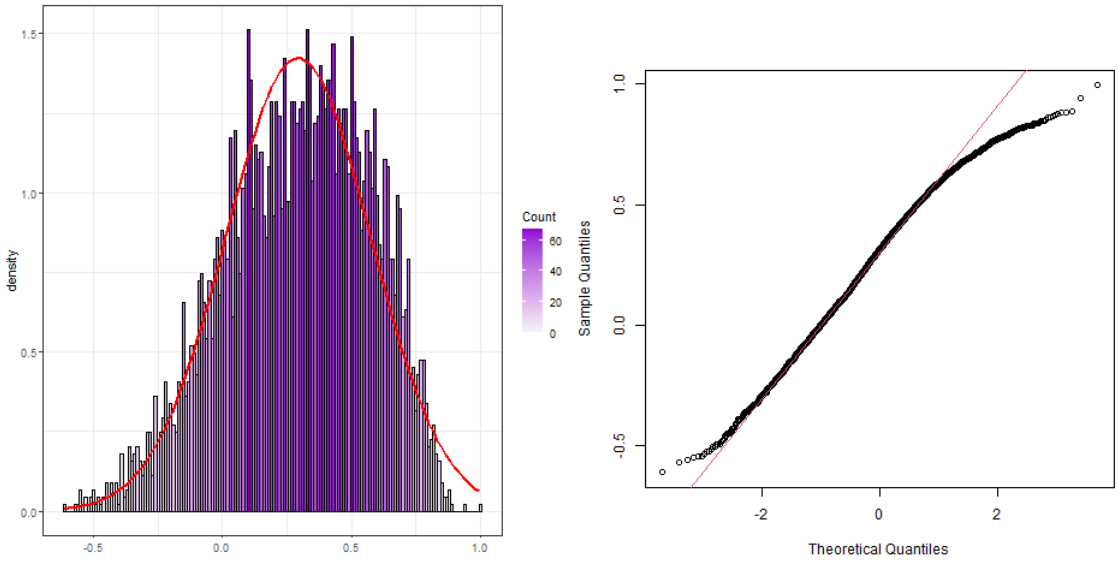
Each group was inspected for normality. Inspection of the histogram and corresponding normality plot shows that both distributions appear to conform to a normal distribution. To check this, the standardised normal scores of skew and kurtosis were inspected. Ethnicity white of non-Irish origin participants had a standardised kurtotis score of -6.05 (kurtotis=-0.45, SE=0.07), and a standardised skew of -8.51 (skew=-0.31, SE=0.04). All other ethnicities had a standardised kurtotis score of -6.05 (kurtotis=-0.45, SE=0.07), and a standardised skew of -8.51 (skew=-0.31, SE=0.04). While kurtotis is within the accepted range of ± 2 for both sets of data, skew is not.

Examining Figure 4.12, none of the standardised data points are outside ± 3.29 in each, we can approximate the sample distributions as normal.

A F-test was used to see if the variances of the two samples could be considered to be equal. As the p-value was not less that 0.05 (P=1), there is not enough evidence to reject the null hypothesis and so the variances are considered as equal.

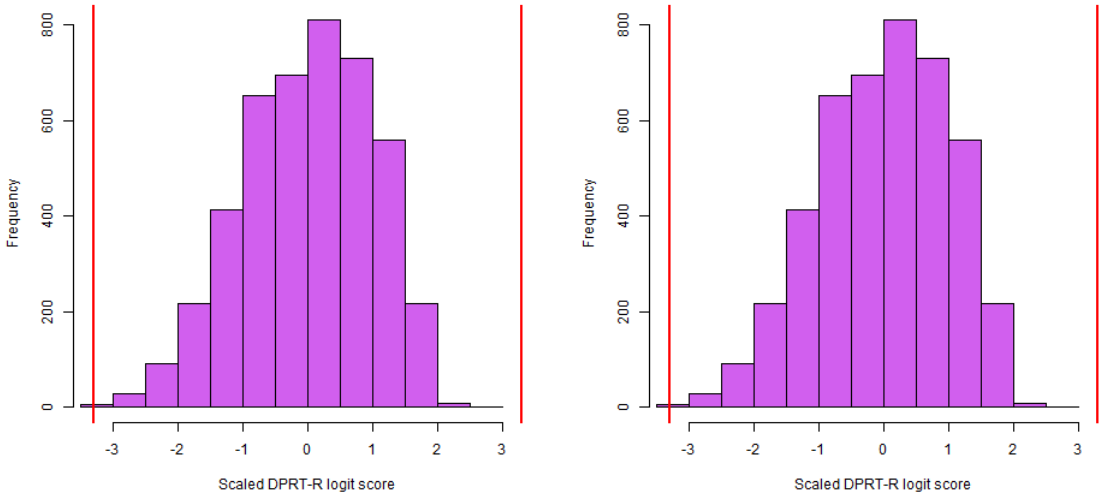


(a) African or other black ethnic background (b) African or other black ethnic background



(c) All other ethnicities (d) All other ethnicities

Figure 4.11: Distribution and QQ plots of DPRT-R scores



(a) African or other black ethnic background (b) All other ethnicities

Figure 4.12: Outliers in scaled data

As the p-value was found to be $p<0.05$ ($t=17.28$, $p=2.2e\text{-}16$), there is enough evidence to reject the hypothesis that participants who are a member of the Catholic religion are not predicted to have lower scores that participants who are not members of the Catholic religion. Cohen's statistic was calculated and showed that there is a small to medium positive effect ($d=0.3673$).

4.6 Conclusion

In this chapter, the proposed experiments were implemented. Two models were created, one trained on a dataset with all truly missing data removed, and a second trained on a full dataset with missing data imputed. The first model produced the lower mean squared error and higher variance (33%). Weak discrimination was also found regarding religion and ethnicity. However, only one imputation was used, so a more accurate dataset could have been obtained with multiple imputations.

Chapter 5

Conclusion

5.1 Introduction

In this chapter, the overall dissertation will be discussed. The experimental results and any limitations of the mode will be evaluated. Additionally, the original aims of the project will be evaluated for success. The contributions and impact of this work will also be discussed.

5.2 Research Overview

The aim of the experiment was to create a model that could predict the literacy abilities of a 9 year old child based on features of their life when they were 9 months, 3 years and 5 years old. A second aim was to check whether this model exhibits direct discrimination based on the protected classes of gender, race and religion. Features that were shown from a review of the current literature to contribute to or be a risk factor for childhood literacy were included in the elastic net model. Methods of identifying and mitigating discrimination were reviewed and a 'twin test' was chosen to test for the existence of discrimination.

Two methods for handling unavoidable missing data were tested. Each of the two resultant models were evaluated for prediction ability and whether they discriminated on gender, race or ethnicity. If significant discrimination was found, the effect was

calculated.

As part of this research, the following objectives were achieved:

- The current literature was reviewed in the areas of child literacy development and risk factors of literacy difficulties, identification of discrimination in machine learning models
- A review was carried out of machine learning methods suitable for a model with many potentially correlated predictors
- Access was obtained to a dataset from which the variables of interested were extracted and manipulated into format required by chosen machine learning model, including dealing with missing data
- Two different methods of dealing with missing data were tried
- A model was designed and trained to predict literacy abilities and evaluated on the ability to predict literacy levels with the lowest mean squared error
- A synthetic child population was created that differ on the protected characteristics of gender, religion and ethnicity
- Any differences found in predicted DPRT-R score between protected characteristics was evaluated for statistical significance and strength of effect.

5.3 Problem Definition

The research question asked was: does an elastic net model to predict literacy levels in children at age 9 based on measurements about the child's background, household, development and early education at ages 9 months, 3 years and 5 years discriminate across gender, ethnic or religious background in a population of children living in Ireland who were born in 2007/2008?

Based on the results of the experiment, it can be concluded that no discrimination exists based on gender, but discrimination with a weak effect exists based on religion and ethnicity.

5.4 Design/Experimentation, Evaluation & Results

The dataset was very complex, with over 4000 columns, each of which is described in PDF documents. There are multiple different types of missing data, requiring different treatments, sometimes indicated by the same value, or with a value that is a valid answer in a related question. Because of this, each variable had to be manually examined. Several more variables which had already been examined and partially cleaned had to be discarded in favor of a slimmed down feature set due to time constraints. These additional variables could potentially have improved the model.

First, all observations with truly missing data were removed. It was found that a model could be created that used features about a child at ages 9 months, 3, and 5 years to predict literacy at age 9, explaining 33% of the variance between children. This is in line with other models to predict literacy, as covered in the literature review. This model was then tested for discrimination. Discrimination was found for religion (Catholic children were predicted to have lower scores than all other religions or children that had no religion) and ethnicity (white children of non-Irish background were predicted to have a higher score than all other nationalities). In both of these cases, the size of the effect was small.

Automatic imputation was then tried. It resulted in a model that performed less well. The best model has a higher mean squared error than the previous model. It could explain 27.5% of the variance in scores. The model also showed stronger discrimination. Discrimination was found in religion (in contrast to the previous model, Catholic children were predicted to have higher scores than all other religions or children that had no religion). The effect of this discrimination is weak. There was also discrimination found based on ethnicity (children who's ethnicity is of African or other black background were predicted to have a lower score than every other ethnicity). The effect of this was found to be small to moderate.

5.5 Contributions and impact

While many studies have aimed to predict childhood literacy levels, none appear to look at discrimination in this context. Additionally, the GUI dataset has been used in multiple predictive models, including predicting reading ability, but discrimination has not been accounted for. The creators of the dataset provide a weighting to rebalance the dataset to the general population level, but not all factors are included in this weighting.

It is vital to consider discrimination in models to predict potential interventions. If a decision was made to stage an early intervention on the basis of a predicted score using a model that exhibited discrimination there could be multiple effects. If a child was predicted to get a higher score than reality, they could be denied intervention that an identical member of another class received. If a child was predicted to get a lower score than reality, they could have an unnecessary intervention. There is anecdotal evidence of immigrant children being singled out for extra language help when it is not required, perpetuating a feeling of exclusion. While there are always errors in any model, due to the Equal Status Act, it is illegal to discriminate based on any protected characteristic.

5.6 Future Work & Recommendations

As the discriminatory variables were included in the model to measure direct discrimination, they could be excluded in order to repeat the experiment while measuring indirect discrimination. Methods to counteract discrimination could be investigated and a new weighting could be developed where all protected variables were considered.

Membership of the travelling community is shielded to protect anonymity as part of the generally available dataset. A researcher could get access to the further dataset and repeat the analysis with this additional protected group.

The features could be evaluated for their influence on literacy. This was not an aim of this dissertation. The GUI dataset is incredibly rich and multiple predictive models could be created on any aspect of a child's life.

The next wave of the GUI dataset should be released this year. The study could be expanded to include features from wave 5 to predict an outcome in wave 6. Additionally, any models developed could be tested on data from similar international longitudinal studies. Many questions from the GUI study were based on those in the Longitudinal Study of Australian Children (LSAC)[1] and the Millennium Cohort Study, Britain[2] so many variables would be directly comparable. The Growing Up In Scotland[3] and Growing Up in New Zealand[4] studies would also be of interest as they are for similarly sized countries to Ireland and so similarities can be examined.

[1]https://growingupinaustralia.gov.au/

[2]https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

[3]https://growingupinscotland.org.uk/about-gus/

[4]http://www.growingup.co.nz/en.html

References

Armstrong, R., Scott, J., Copland, D., McMahon, K., Khan, A., Najman, J. M., ... Arnott, W. (2016). Predicting receptive vocabulary change from childhood to adulthood: A birth cohort study. *Journal of Communication Disorders*, *64*, 78–90. doi: 10.1016/j.jcomdis.2016.10.002

Armstrong, R., Symons, M., Scott, J. G., Arnott, W. L., Copland, D. A., McMahon, K. L., & Whitehouse, A. J. (2018). Predicting language difficulties in middle childhood from early developmental milestones: A comparison of traditional regression and machine learning techniques. *Journal of Speech, Language, and Hearing Research*, *61*(8), 1926–1944. doi: 10.1044/2018_JSLHR-L-17-0210

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*(3), 671–732. doi: 10.2139/ssrn.2477899

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. doi: 10.1147/JRD.2019.2942287

Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., ... Chi, E. H. (2019). Putting fairness principles into practice. In *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (p. 453-459). doi: 10.1145/3306618.3314234

Brase, C. H., & Brase, C. P. (2001). *Understanding basic statistics: Concepts and methods*. Houghton Mifflin College Division.

Bronfenbrenner, U., & Morris, P. A. (2007). The bioecological model of human development. *Handbook of child psychology*, *1*. doi: 10.1002/9780470147658.chpsy0114

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 13–18). doi: 10.1109/ICDMW.2009.83

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining* (pp. 71–80). doi: 10.1109/ICDM.2013.114

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. doi: 10.1007/s10618-010-0190-x

Calders, T., & Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society* (pp. 43–57). Springer. doi: 10.1007/978-3-642-30487-3_3

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2018, Oct). Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, *12*(5), 1106-1119. doi: 10.1109/JSTSP.2018.2865887

Chow, K. A., Mistry, R. S., & Melchor, V. L. (2015, July). Homelessness in the elementary school classroom: social and emotional consequences. *International Journal of Qualitative Studies in Education*, *28*(6), 641–662. doi: 10.1080/09518398.2015.1017855

Crowe, M., O'Sullivan, M., Cassetti, O., & O'Sullivan, A. (2017). Weight status and dental problems in early childhood: classification tree analysis of a national cohort. *Dentistry journal*, *5*(3), 25. doi: 10.3390/dj5030025

Curran, D., Molenberghs, G., Fayers, P. M., & Machin, D. (1998). Incomplete quality of life data in randomized trials: missing forms. *Statistics in medicine*, *17*(5-7), 697–709. doi: 10.1002/(SICI)1097-0258(19980315/15)17:5/7⟨697::AID-SIM815⟩3.0.CO;2-Y

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, *5*(2), 120–134. doi: 10.1089/big.2016.0048

Dallaire, D. H., Ciccone, A., & Wilson, L. C. (2010, July). Teachers experiences with and expectations of children with incarcerated parents. *Journal of Applied Developmental Psychology*, *31*(4), 281–290. doi: 10.1016/j.appdev.2010.04.001

Darmody, M., Byrne, D., & McGinnity, F. (2012, May). Cumulative disadvantage? educational careers of migrant students in irish secondary schools. *Race Ethnicity and Education*, *17*(1), 129–151. Retrieved from `https://doi.org/10.1080/13613324.2012.674021` doi: 10.1080/13613324.2012.674021

Duff, F. J., Nation, K., Plunkett, K., & Bishop, D. (2015, July). Early prediction of language and literacy problems: is 18 months too early? *PeerJ*, *3*, e1098. Retrieved from `https://doi.org/10.7717/peerj.1098` doi: 10.7717/peerj.1098

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). doi: 10.1145/2090236.2090255

Edwards, B., et al. (2012). Growing up in australia: the longitudinal study of australian children: the first decade of life. *Family Matters*(91), 7.

Eliot, M., Ferguson, J., Reilly, M. P., & Foulkes, A. S. (2011). Ridge regression for longitudinal biomarker data. *The International Journal of Biostatistics*, *7*(1). doi: 10.2202/1557-4679.1353

Fayers, P. M., Curran, D., & Machin, D. (1998). Incomplete quality of life data in randomized trials: missing items. *Statistics in medicine*, *17*(5-7), 679–696. doi: 10.1002/(SICI)1097-0258(19980315/15)17:5/7⟨679::AID-SIM814⟩3.0.CO;2-X

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using r/andy field, jeremy miles, zoë field.* London; Thousand Oaks, Calif.: Sage,.

Government of Ireland. (2000). *Equal Status Act.* (Retrieved from `http://www.irishstatutebook.ie/eli/2000/act/8/enacted/en/print.html`)

Hastie, T., & Qian, J. (2014). Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed September*, *20*, 2016.

Holman, R., Glas, C. A., Lindeboom, R., Zwinderman, A. H., & De Haan, R. J. (2004). Practical methods for dealing with'not applicable'item responses in the amc linear disability score project. *Health and quality of life outcomes*, *2*(1), 29. doi: 10.1186/1477-7525-2-29

Hu, L., & Chen, Y. (2018). Welfare and Distributional Impacts of Fair Classification. Retrieved from `http://arxiv.org/abs/1807.01134`

Hughes, A., Gallagher, S., & Hannigan, A. (2015). A cluster analysis of reported sleeping patterns of 9-month old infants and the association with maternal health: Results from a population based cohort study. *Maternal and child health journal*, *19*(8), 1881–1889. doi: 10.1007/s10995-015-1701-6

Justice, L. M., & Pullen, P. C. (2003). Promising interventions for promoting emergent literacy skills: Three evidence-based approaches. *Topics in early childhood special education*, *23*(3), 99–113. doi: 10.1177/02711214030230030101

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *2009 2nd international conference on computer, control and communication* (pp. 1–6). doi: 10.1109/IC4.2009.4909197

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. doi: 10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 924–929). doi: 10.1109/ICDM.2012.45

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, *35*(3), 613–644. doi: 10.1007/s10115-012-0584-8

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 35–50). doi: 10.1007/978-3-642-33486-3_3

Keogh, A. F., Halpenny, A. M., & Gilligan, R. (2006, November). Educational Issues for Children and Young People in Families Living in Emergency Accommodation-An Irish Perspective. *Children & society*, *20*(5), 360–375. doi: 10.1111/j.1099-0860.2006.00015.x

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research*. doi: 10.1044/1092-4388(2009/08-0142)

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 502–510). doi: 10.1145/2020408.2020488

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interven-

tions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? *Early Childhood Research Quarterly*, *25*(4), 409–431. doi: 10.1016/j.ecresq.2010.03.002

McCrory, C., Williams, J., Murray, A., Quail, A., & Thornton, M. (2013). *Design, Instrumentation and Procedures for the Infant Cohort at Wave Two (3 years)* (Tech. Rep.). The Economic and Social Research Institute. Retrieved from `https://www.growingup.ie/pubs/BKMNEXT253.pdf`

McLeod, B. A., Johnson, W. E., Cryer-Coupet, Q. R., & Mincy, R. B. (2019, May). Examining the longitudinal effects of paternal incarceration and coparenting relationships on sons educational outcomes: A mediation analysis. *Children and Youth Services Review*, *100*, 362–375. doi: 10.1016/j.childyouth.2019.03.010

McNamara, E., Murray, A., & Williams, J. (2019). *Design, Instrumentation and Procedures (including Summary Literature Review, Pilot Report and Findings) for Cohort '08 at Wave Four (7/8 years)* (Tech. Rep.). The Economic and Social Research Institute. Retrieved from `https://www.growingup.ie/pubs/20190404-Cohort-08-at-5years-design-instrumentation-and-procedures.pdf`

McNamara, E., O'Mahony, D., & Murray, A. (2020). *Design, Instrumentation and Procedures for Cohort '08 of Growing Up in Ireland at 9 Years Old (Wave 5)* (Tech. Rep.). The Economic and Social Research Institute. Retrieved from `https://www.growingup.ie/pubs/Cohort08at9-Design-Report-2020-1.pdf`

Morton, S. M., Ramke, J., Kinloch, J., Grant, C. C., Carr, P. A., Leeson, H., ... Robinson, E. (2015). Growing up in new zealand cohort alignment with all new zealand births. *Australian and New Zealand journal of public health*, *39*(1), 82–87. doi: 10.1111/1753-6405.12220

Murray, A., & Egan, S. M. (2014). Does reading to infants benefit their cognitive development at 9-months-old? an investigation using a large birth cohort survey. *Child Language Teaching and Therapy*, *30*(3), 303–315. doi: 10.1177/0265659013513813

Murray, A., Williams, J., Quail, A., Neary, M., & Thornton, M. (2015). *A Summary Guide to Wave 3 of the Infant Cohort (at 5 years) of Growing Up in Ireland* (Tech. Rep.). The Economic and Social Research Institute. Retrieved from https://www.growingup.ie/pubs/Summary-Guide-_Infant-Cohort_Wave-3.pdf

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *Bmc proceedings* (Vol. 6, p. S10).

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining - KDD 08.* ACM Press. Retrieved from https://doi.org/10.1145/1401890.1401959 doi: 10.1145/1401890.1401959

Pentimonti, J. M., Murphy, K. A., Justice, L. M., Logan, J. A., & Kaderavek, J. N. (2016). School readiness of children with language impairment: predicting literacy skills from pre-literacy and social–behavioural dimensions. *International Journal of Language & Communication Disorders*, *51*(2), 148–161.

Plewis, I., Calderwood, L., Hawkes, D., Hughes, G., & Joshi, H. (2007). Millennium cohort study: technical report on sampling. *London: Centre for Longitudinal Study, Institute of Education*.

Romei, A., & Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. *Knowledge Engineering Review*, *29*(5), 582–638. doi: 10.1017/S0269888913000039

Saracho, O. N. (2017). *Literacy and language: new developments in research, theory, and practice.* Taylor & Francis. doi: 10.1080/03004430.2017.1282235

Schoon, I., Bynner, J., Joshi, H., Parsons, S., Wiggins, R. D., & Sacker, A. (2002). The influence of context, timing, and duration of risk experiences for the passage from childhood to midadulthood. *Child development*, *73*(5), 1486–1504. doi: 10.1111/1467-8624.00485

Shonkoff, J. P. (2010). Building a new biodevelopmental framework to guide the future of early childhood policy. *Child development*, *81*(1), 357–367. doi: 10.1111/j.1467-8624.2009.01399.x

Squires, G. D. (2003). Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, *25*(4), 391–410. doi: 10.1111/1467-9906.t01-1-00168

Taylor, C. L., Christensen, D., Lawrence, D., Mitrou, F., & Zubrick, S. R. (2013). Risk factors for children's receptive vocabulary development from four to eight years in the longitudinal study of australian children. *PLOS one*, *8*(9), e73046. doi: 10.1371/journal.pone.0073046

Thornton, M., Williams, J., McCrory, C., Murray, A., & Quail, A. (2013). *Design, Instrumentation and Procedures for the Infant Cohort at Wave One (9 months)* (Tech. Rep.). The Economic and Social Research Institute. Retrieved from `https://www.growingup.ie/pubs/BKMNEXT252.pdf`

Turney, K. (2018, June). Adverse childhood experiences among children of incarcerated parents. *Children and Youth Services Review*, *89*, 218–225. doi: 10.1016/j.childyouth.2018.04.033

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 1–17. doi: 10.1177/2053951717743530

Wallace, I. F., Berkman, N. D., Watson, L. R., Coyne-Beasley, T., Wood, C. T., Cullen, K., & Lohr, K. N. (2015). Screening for speech and language delay in children 5 years old and younger: a systematic review. *Pediatrics*, *136*(2), e448–e462. doi: 10.1542/peds.2014-3889

Williams, J., Thornton, M., Murray, A., & Quail, A. (2019). *Design, Instrumentation and Procedures for Cohort '08 at Wave 3 (5 Years)* (Tech. Rep.). The Economic and Social Research Institute. Retrieved

from https://www.growingup.ie/pubs/20190404-Cohort-08-at-5years-design-instrumentation-and-procedures.pdf

Williams, W., Latif, A., Hannington, L., & Watkins, D. (2005). Hyperopia and educational attainment in a primary school cohort. *Archives of disease in childhood*, *90*(2), 150–153.

Yeom, S., & Tschantz, M. C. (2018). Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. Retrieved from http://arxiv.org/abs/1808.08619

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, *28*(PART 2), 1362–1370.

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183–201.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. doi: 10.1007/s10618-017-0506-1

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

Appendix A

Missing Data

	Missing Data			To Impute	
Wave 1 Survey Question	**S**	**R**	**DK**	**N**	**%**
Gender of study child	0	0	0	0	0%
PCG ethnicity	0	0	19	19	0.3%
Does the PCG have a partner living in the household	0	0	0	0	0%
Number of people in household	0	0	0	0	0%
How many separate bedrooms are in the accommodation?	0	0	15	15	0.2%
PCG current economic status	0	0	2	2	0%
Family's Social Class	0	0	0	0	0%
Degree of ease or difficulty is the hsd able to make ends meet?	0	0	6	6	0.1%
Child's weight at birth	0	0	87	87	1.2%

Table A.1: Count of missing data in wave 1. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 1 Survey Question	**S**	**R**	**DK**	**N**	**%**
PCG Is English your native language?	0	0	1	1	0%
PCG Read aloud from a children s storybook in English?	0	0	3	3	0%
PGC Read and fill out forms in English?	0	0	7	7	0.1%
PCG Highest level of educational achievement	0	0	4	4	0.1%
Do you talk to child while you are busy doing other things?	0	0	1	1	0%

Table A.2: Count of missing data in wave 1. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 2 Survey Question	**S**	**R**	**DK**	**N**	**%**
Age of PCG	0	0	0	0	0%
PCG current economic status	0	1	3	4	0.1%
Family's Social Class	0	0	0	0	0%
Do you have any concerns about how child talks and makes speech sounds?	0	1	24	25	0.3%
Picture Similarities	0	0	151	151	2%
Naming Vocabulary	0	0	354	354	4.7%

Table A.3: Count of missing data in wave 2. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 3 Survey Question	**S**	**R**	**DK**	**N**	**%**
PCG parental stress	0	0	103	103	1.4%
Number of people in household	0	0	0	0	0%
How many separate bedrooms are in the accommodation?	0	0	2	3	0%
PCG current economic status	0	0	23	23	0.3%
Family's Social Class	0	0	528	528	7%
About how many children's books does child have access to in your home now, including any library books?	0	0	3	3	0%
How often would you (PCG) visit the library with child?	0	0	2	3	0%
How often would you (PCG) listen to child read?	0	0	3	4	0.1%
How often would you (PCG) read to child?	0	0	1	2	0%
Do you have any concerns about how child talks and makes speech sounds?	0	0	2	2	0%

Table A.4: Count of missing data in wave 3. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 3 Survey Question	**S**	**R**	**DK**	**N**	**%**
SSIS - Assertion Subscale	0	0	20	20	0.3%
SSIS - Responsibility Subscale	0	0	20	20	0.3%
SSIS - Empathy Subscale	0	0	20	20	0.3%
SSIS - Selfcontrol Subscale	0	0	20	20	0.3%
LSAC temperament measure - Persistence Subscale	0	0	13	13	0.2%
LSAC temperament measure - Sociability Subscale	0	0	11	11	0.1%
LSAC temperament measure - Reactivity Subscale	0	0	6	6	0.1%
SDQ Emotional subscale - Caregiver	0	0	2	2	0%
SDQ Conduct subscale - Caregiver	0	0	2	2	0%
SDQ Hyperactivity subscale - Caregiver	0	0	3	3	0%
SDQ Peer problems subscale - Caregiver	0	0	3	3	0%
SDQ Prosocial subscale - Caregiver	0	0	3	3	0%
SDQ Total difficulties score - Caregiver	0	0	3	3	0%
SDQ Impact score - Caregiver	0	0	3	3	0%

Table A.5: Count of missing data in SSIS, LSAC and SDQ scales. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 3 Survey Question	**S**	**R**	**DK**	**N**	**%**
What class is study child in?	0	0	45	471	6.3%
Picture Similarities	0	0	55	55	0.7%
Naming Vocabulary	0	0	70	70	0.9%

Table A.6: Count of missing data in educational variables. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

	Missing Data			To Impute	
Wave 3 Survey Question	**S**	**R**	**DK**	**N**	**%**
Total Teacher Report					
Language	0	0	33	459	6.1%
Linking	0	0	69	495	6.6%
Reading	0	0	55	481	6.4%
In so far as your professional experience allows, please rate the Study Child in terms of a range of competencies in relation to all children of this age (not just in their present class or, even, school):					
Speaking and listening in English	0	0	162	588	7.8%
Reading in English	0	0	812	1238	16.5%
Writing in English	0	0	1017	1443	19.2%

Table A.7: Count of missing data in teacher's response in wave 3. S indicates that data is missing due to the survey path. R indicates the participant refused to answer that question. DK indicates that the participant didn't know the answer to the question. N is the total number to impute. % is the percentage missing of all data.

Appendix B

All Model Features After Cleaning

Appendix contains all the fields included in the model after all data cleaning steps have been completed.

Wave	**Question**	**Data Type**	**Valid Responses**
1	Study child is male	B	1=Yes, 0=No
1	Child's ethnicity is any white background excl. Irish	B	1=Yes, 0=No
1	Child's ethnicity is African or other black background	B	1=Yes, 0=No
1	Child's ethnicity is Chinese or any other Asian background	B	1=Yes, 0=No
1	Child's ethnicity is any other background inc. mixed	B	1=Yes, 0=No
1	Child has no religion	B	1=Yes, 0=No
1	Child's religion is other Christian excl. Catholic	B	1=Yes, 0=No
1	Child's religion is other excl. all Christian	B	1=Yes, 0=No
1	Single parent household	B	1=Yes, 0=No
1, 3	Bedroom Density	N	Numeric

Table B.1: All family background features, data types and valid responses included in model after all data preparation. B indicates a binary data type. N indicates a numeric data type.

Wave	**Question**	**Data Type**	**Valid Responses**
1	PCG Is English your native language?	B	1=Yes, 0=No
1	PCG Read aloud from a children s storybook in English?	B	1=Yes, 0=No
1	PGC Read and fill out forms in English?	B	1=Yes, 0=No
1	PCG Highest level of educational achievement	O	1=No formal education, 2=Primary education, 3=Lower secondary, 4=Upper secondary, 5=Technical or vocational qualification, 6=Both upper secondary and Technical or Vocational qualification, 7=Non Degree, 8=Primary Degree, 9=Professional qualification (of Degree status at least), 10=Both a Degree and a Professional qualification, 11=Postgraduate Certificate or Diploma, 12=Postgraduate Degree (Masters), 13=Doctorate
1, 2, 3	PCG current economic status	O	1-10
1, 2, 3	Family's Social Class	O	1=Professional/managerial, 2=Other non-manual/skilled-manual, 3=Semi-skilled/unskilled manual, 7=All others gainfully occupied and unknown, 8=Never worked at all - no class

Table B.2: All family socioeconomic features, data types and valid responses included in model after all data preparation. B indicates a binary data type. O indicates an ordinal data type.

Wave	**Question**	**Data Type**	**Valid Responses**
1	Degree of ease or difficulty is the hsd able to make ends meet?	O	1=With great difficulty, 2=With difficulty, 3=With some difficulty, 4=Fairly easily, 5=Easily, 6=Very easily
1	Do you talk to child while you are busy doing other things?	O	1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Always
1	Child's weight at birth	O	Rounded Grams
1	ASQ Problem Solving Max Test Passed	O	0=Fail, 1=8 month, 2=10 month, 3=12 month
1	ASQ Gross Motor Max Test Passed	O	0=Fail, 1=8 month, 2=10 month, 3=12 month
1	ASQ Fine Motor Max Test Passed	O	0=Fail, 1=8 month, 2=10 month, 3=12 month
1	ASQ Communication Max Test Passed	O	0=Fail, 1=8 month, 2=10 month, 3=12 month
1	ASQ Personal-Social Max Test Passed	O	0=Fail, 1=8 month, 2=10 month, 3=12 month
2	Age of PCG	O	Rounded age in years
2, 3	Do you have any concerns about how child talks and makes speech sounds? Would you say no, yes a little or yes a lot?	O	1=No, 2=Yes a little, 3=Yes a lot

Table B.3: ASQ and family background features, data types and valid responses included in model after all data preparation. O indicates an ordinal data type.

Wave	**Question**	**Data Type**	**Valid Responses**
2, 3	Picture Similarities	N	Percentile
2, 3	Naming Vocabulary	N	Percentile
3	PCG parental stress	N	Scale 1-30
3	About how many children's books does child have access to in your home now, including any library books?	O	1=None, 2=Less than 10, 3=10 to 20, 4=21 to 30, 5=More than 30
3	How often would you (PCG) visit the library with child?	O	1=Never, 2=Hardly ever, 3=Occasionally 4=One or two times a week, 5=Everyday
3	How often would you (PCG) listen to child read?	O	1=Never, 2=Hardly ever, 3=Occasionally 4=One or two times a week, 5=Everyday
3	How often would you (PCG) read to child?	O	1=Never, 2=Hardly ever, 3=Occasionally 4=One or two times a week, 5=Everyday
3	SSIS - Assertion Subscale	N	Scale
3	SSIS - Responsibility Subscale	N	Scale
3	SSIS - Empathy Subscale	N	Scale
3	SSIS - Selfcontrol Subscale	N	Scale
3	LSAC temperament measure - Persistence Subscale	N	Scale
3	LSAC temperament measure - Sociability Subscale	N	Scale
3	LSAC temperament measure - Reactivity Subscale	N	Scale

Table B.4: Development and literacy features, data types and valid responses included in model after all data preparation. O indicates an ordinal data type. N indicates a numeric data type.

Wave	**Question**	**Data Type**	**Valid Responses**
3	SDQ Emotional subscale - Caregiver	N	Scale
3	SDQ Conduct subscale - Caregiver	N	Scale
3	SDQ Hyperactivity subscale - Caregiver	N	Scale
3	SDQ Peer problems subscale - Caregiver	N	Scale
3	SDQ Prosocial subscale - Caregiver	N	Scale
3	SDQ Total difficulties score - Caregiver	N	Scale
3	SDQ Impact score - Caregiver	N	Scale
3	Rich Environment & Activities Scale	N	Scale
3	Quality of Child Care	N	Scale
3	Total Teacher Report - Language	N	Scale
3	Total Teacher Report - Linking	N	Scale
3	Total Teacher Report - Reading	N	Scale

Table B.5: SDQ and teacher report features, data types and valid responses included in model after all data preparation. N indicates a numeric data type.

Wave	Question	Data Type	Valid Responses
3	Did the study child attent free preschool?	B	1=Yes, 2=No
3	What class is study child in?	O	1=Junior Infants, 2= Senior Infants, 3=First class, 4=Other
To child's teacher: In so far as your professional experience allows, please rate the Study Child in relation to all children of this age (not just in their present class or, even, school):			
3	Speaking and listening in English	O	1=Highest , 2=Middle, 3=Lowest
3	Reading in English	O	1=Well above Average, 2=Above Average, 3=average, 4=Below average, 5=Well below average
3	Writing in English	O	1=Well above Average, 2=Above Average, 3=average, 4=Below average, 5=Well below average
3	How often has child complained about school/preschool?	O	1=More than Once a week, 2=Once a week or less, 3=Not at all
3	How often has child said good things about school/preschool?	O	1=More than Once a week, 2=Once a week or less, 3=Not at all
3	How often has child looked forward to going to school/preschool?	O	1=More than Once a week, 2=Once a week or less, 3=Not at all
3	How often has child been upset or reluctant to go to school/preschool?	O	1=More than Once a week, 2=Once a week or less, 3=Not at all

Table B.6: School related features, data types and valid responses included in model after all data preparation. B indicates a binary data type. O indicates an ordinal data type.

Appendix C

Feature Correlation

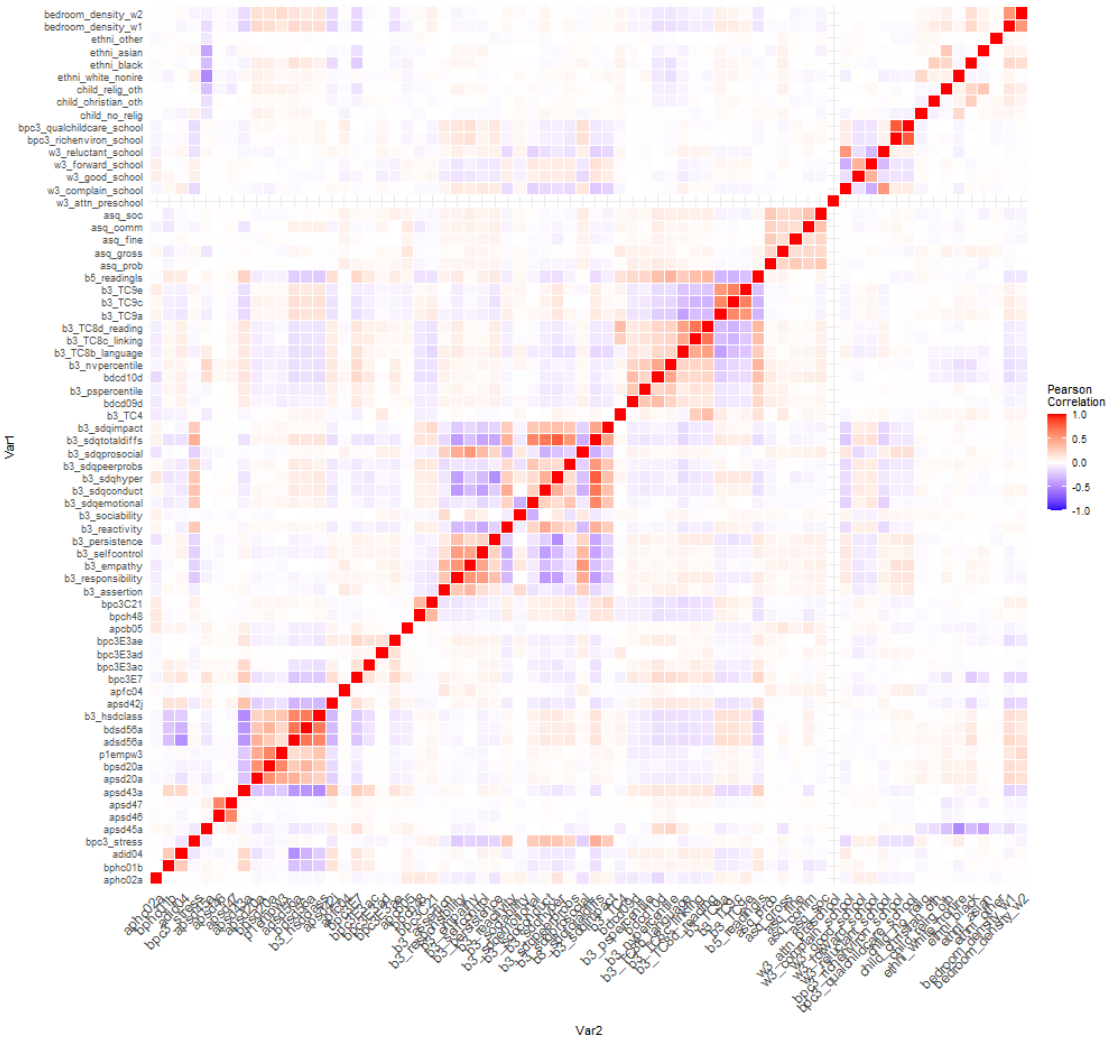


Figure C.1: Pearson correlation heatmap of all variables