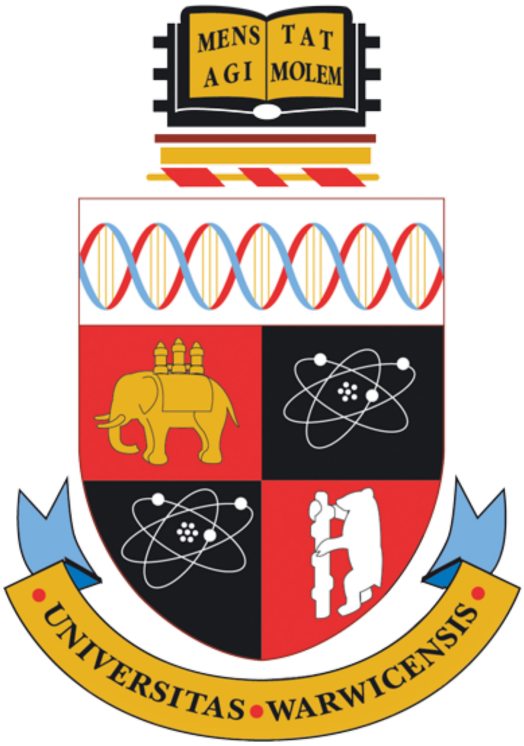
Labour Market Outcomes of Apprentices in the UK

Adam Gajtkowski
(Warwick ID No. 1534964)



Department of Statistics
University of Warwick
Coventry CV4 7AL
United Kingdom

Email: adam.gajtkowski@warwick.ac.uk

13 September 2019

Report submitted in partial fulfillment of the requirements for the degree of MSc in Statistics in the University of Warwick

Abstract

This project investigates the employment outcomes of apprenticeship starters. In particular, we focus on analysing the role of background characteristics of learners on their employment outcomes. For the purpose of the analysis we derive 4 data sets of apprenticeship starters in 2011, 2012, 2013, and 2014. We investigate the relationship between variables by conducting exploratory data analysis and running supervised and unsupervised machine learning algorithms. We find that characteristics related to prior employment have strong impact on employment outcomes. We note that gender, region, and ethnicity also have a strong impact on employment outcomes. We find that background characteristics of learners have limited predictive power for having both positive and negative employment outcomes. We find that among all tested statistical model, deep neural network performs best in terms of predictive power and different measures of errors. We conclude that background characteristics of individuals to some extent can predict employment outcomes, but they are not sufficient for creating accurate forecasts.

Acknowledgements

I would like to thank Prof. Jane Hutton for the academic guidance and support during my research. I also would like to thank the Institute for Apprenticeships and Technical Education for supplying high quality training, mentorship and friendly environment. In particular, I would like to thank my industry supervisors Dr Joseph Elliston and Dr Niall Goulding for supervision, guidance and help from the very beginning. Many thanks to all data scientists within the Institute for a friendly environment and great help.

Contents

Chapter 1

INTRODUCTION

1.1 Research Question

This project investigates the role of characteristics of individuals on their labour market outcomes. We attempt to answer the following research questions.

- What are the characteristics of 2013 apprenticeship starters?
- Which characteristics of learners are relevant for determining employment outcomes?
- Which characteristics have significant impact on the employment outcomes?
- To what extent is it possible to forecast employment outcomes using characteristics of individuals?
- Which statistical models perform the best for forecasting?

Department for Education (DfE) has recently investigated the impact of apprenticeship on employment outcomes in the publication called Outcome Based Success Measures (OBSM) released in 2018 [1]. The publication introduces some employment summary statistics. The publication also analyses the employment outcomes of those who are eligible learners and has completed the apprenticeships.

This project builds on the above publication by improving the data-cleaning process and conducting more advanced statistical analysis. The DfE publication does not mention the distribution of the employment outcomes and their characteristics. We build on this by including relevant histograms and contingency tables. Furthermore, the publication does not take into consideration missing values. We identify the missing values and investigate them. The institutional publication introduces some summary statistics. We perform multivariate exploratory data analysis and use more advanced statistical algorithms. In contrast to the above publication, we consider all apprenticeship starters, regardless of founding source. We do that, as we assume that the founding source has no influence on the employment outcomes.

The project also contributes by creating 5 new data sets including characteristics of apprenticeship starter and labour market outcomes of cohorts 2011, 2012, 2013, and 2014. We also create a data set with reduced match ratio of 66% which may be used for training algorithms as it has small number of missing values. We cannot use the methodology of merging data sets used by OBSM authors due to restricted access to the data sets they used.

We also contribute by testing and training machine learning algorithms. The results of our research can be used to create a machine learning pipeline for forecasting employment outcomes of potential apprenticeship candidates.

The project is divided into four chapters. The first chapter reviews relevant publications. Second chapter discusses the data quality, describes the procedure of data processing and conducts an exploratory data analysis. Fourth chapter builds up on the exploratory data analysis by investigating the performance of 6 statistical algorithms. The last chapter interprets EDA, model findings and suggests direction for the future research.

The above research questions are relevant as answering them may help the government to improve its current policy. Knowing which characteristics of learners are important for positive employment outcomes may help to target the right learners, establish correct level of founding and forecast the impact of apprenticeships on earnings of the current learners.

1.2 Literatue Review

1.2.1 Learners and Apprentices Survey 2018

The Lerners Survey publication [2] characterizes the further education and apprenticeship learners. The report is based on an online and telephone surveys. It describes both Further Education (FE) learners and apprentices 19 and over, sampled from the Individualised Learner Record (ILR). From the above description we may note that the survey is not specific enough to answer our research questions, as we consider only apprenticeship starters during our research.

The Learners Survey [2] emphasizes that the age within considered cohorts is very heterogenous. They state that although half of the apprenticeship are young people, there are many adults in 30s, 40s, 50s and 60s. The publication [2] also states that apprentices tend to live in relatively deprived areas, often has low initial incomes. Authors [2] mention that a large proportion of apprentices disd not grow up in the UK (14% apprentices [2]) and consistently 13% of apprentices [2] speak a language other than English as their main language. Overall, apprentices tend to come from relatively low socio-economic backgrounds [2].

Authors of the publication [2] mention that older apprentices (aged 25 and above) were already working for their apprentice employer, when starting the apprenticeship, while younger apprentices enrolled in an apprenticeship just after finishing the full-time education [2]. Also, only a third of apprentices were external candidates for the apprenticeship positions, while the rest was hired internally [2]. This is particularly visible for older applicants, where 84% of apprentices were already working for the apprenticeship employer [2]. Within this sub-population the most common reason for starting am apprenticeship was because the employer offered it to them.

The publication [2] concludes with evaluating of outcomes of further education and apprenticeship learners. Authors mention that those who successfully completed their learning had 20 percentage points [2] higher employment outcomes compared to before the start of an apprenticeship. The most common case was a move directly from studying into employment [2]. Authors also mention that 75% of those who have moved directly from full-time education into employment successfully completed the training [2]. That means that starting an apprenticeship has positive outcomes both for those with prior work experience and without prior work experience [2].

1.2.2 Outcome Based Success Measures Publication

The institutional publications Outcome Based Success Measures (OBSM) [1] released by the Department for Education focuses on summary statistics of employment outcomes. The publication is released annually from 2015. The publications focuses on description of impact of apprenticeships on future earnings, number of those who end up in employment and/or education, progression to more advanced apprenticeships [1]. It breaks down the employment outcomes by sector, level of apprenticeships, type of apprenticeship. The authors present their findings mainly by set of figures and by making brief comments.

The publication takes into consideration just specific subset of those who completed an apprenticeship, describing a background of an individual and a type of funding. In our project we focus on a broader question analysing those who have started an apprenticeship within given year. Furthermore, we consider all types of backgrounds while the above publication excludes OLASS learners [1]. We include OLASS learners in our models, but exclude them from the analysis due to sensitivity of this data. We also consider individuals with all sources of founding.

First, the OBSM publication analyses the sustained positive destination rates by academic year of completion. It states that in the academic year 2015/2016 53% of learners who completed the eligible learning aim were in further employment, 13% were both employed and doing further learning, 10% were just learning. The remaining 25% were neither studying nor employed. This figure does not mention if the 25% who were declared as neither studying nor employed are categorised as such because they were unemployed, or there is no data about their further career. It could happen that some learners moved abroad, thus there is no data. It is also possible that they could not report their further education, in case they study abroad or do online courses. We will improve this aspect of the publication by more detailed analysis of missing values.

Furthermore, the publication [1] comments on the above findings relative to previous years. This may be relevant, but we also need to consider exogenous year to year changes. Was there any change in the apprenticeship programmes? What were the economic conditions? We need to consider other factors which would possibly be outside of the model. After considering these factors, we would need to account for it when we compare different years. For that purpouse, when analysing earnings growth we deflate it, and consider earnings of apprenticeship starters after 2011 due to data quality

issue in the pre-2011 years.

Another figure produced in the OBSM [1] publication analyses sustained learning variable. It includes the analysis of further learning by the type of learning in the next year. It includes 4 main groups: any higher education course, apprenticeships, other sustained learning and any level 4 or higher FE course. The disadvantage of this division is lack of detail about breakdown by sector, by gender. In our project, we control for characteristics of individuals. We also analyse all educational levels.

Figure 3 [1] breaks-down destinations by level of learning for 2015/16. We may note that the level 4+ has the highest number of those who go to the employment, while other levels and traineeships have the lowest percentages. The lowest level qualifications have highest learning destinations. There is no information about those who did not report either employment or learning. In our thesis, we try to analyse it. We also do not take into consideration traineeships (figure 5, [1]), just apprenticeships as this is the focus of our thesis.

Figure 5 [1] shows the sustained positive destination rates by the level of apprenticeships for the year 2015/16. We may again see that the higher apprenticeship the better. Level 5 apprenticeships have 87% of subjects in employment while the intermediate level ones have just 68% of subjects in employment, and the 19% is studying. We will try to model this fact and check if the educational level has a predictive power on employment outcomes.

Figure 7 [1] display the advanced and higher level apprenticeships learning destinations (2015/16). It breaks down the destinations into 4 levels; any higher education course, any higher-level FE course, any higher-level apprenticeship, other sustained learning. This breakdown considers just two categories; the higher level 4 apprenticeship, and advanced apprenticeship. The drawback of this table is lack of consideration to other levels and lack of analysis of remaining students without the destination (it accounts for 13% in each case).

Figure 9 [1] focuses on earnings. Earnings are displayed for different level of apprenticeships and different years. Authors analyses the pace of growth of earnings. The analysis indicates a positive relationship between the level of apprenticeship and earnings, as well as number of years after completed apprenticeship and earnings. The pace of increase in earnings appears to be the highest in case of higher apprenticeships and is the lowest in case of level 2 apprenticeship. The analysis does not take into consideration individual

characteristics of learners.

Figure 10 analyses the median annualised earnings one year after study for advanced apprenticeships achieved in academic year 2015/2016. It shows the summary statistics of the distribution of earnings for sectors focused on advanced apprenticeships. Engineering, manufacturing and agriculture are 3 the highest paid sectors, while child development, wellbeing, service enterprises, and direct learning support are the three lower income areas. The highest number of apprentices is in Engineering, Business Management, Administration, Health and Social Care sectors. We would suggest including the break-down of apprentices by gender and their background characteristics to see if there is any significant difference. During our project we will not consider different sectors as it is not related to the research questions.

To sum up the above publication, it provides aggregated summary statistics of those who are eligible learners and completed an apprenticeship within given year. The OBSM publication highlights that those who enrol in higher level apprenticeships have higher level of employment and their earnings growth faster. Our thesis will extend the above analysis by providing detailed exploratory data analysis, conducting additional data cleaning and analysis of any patterns within missing values. Furthermore, we will implement advanced statistical techniques to analyse the outcomes, reduce the dimensionality of data and provide the interpretation of findings.

Chapter 2

DATA QUALITY AND EXPLORATORY DATA ANALYSIS

This chapter consists of 4 sections. Section 1 describes the process of creating a data set. Section 2 focuses on the processing of data set for analysis and machine learning algorithms. Section 3 describes the findings of exploratory data analysis (EDA) and the problem of outliers. Section 4 describes the multivariate EDA.

2.1 Data Set Preparation

2.1.1 Linking HMRC and Department for Education Data Sets

The project is based on the 2 governmental data sets. Longitudinal Education Outcomes (LEO) and Longitudinal Individualised Learner Record (LILR). The LEO learner's dataset is a collection of tables that has been created for the purpose of secondary analysis, from the employment, benefits, self-assessment and earnings data received from DWP/HMRC and the education data [5]. These tables include used employment, education, self-employment and earnings fields and several individual tables with data on earnings, benefits, self-employment and employment [3]. The DWP/HMRC data has been cleaned in preparation to produce the LEO learner's dataset.

Data cleaning includes quality assurance that all the entries are correct, sanity checks including eliminating wrong entries, checking start and end dates [5]. The LILR dataset is used from the Individualised Learner Records, which comes from the Education & Skills Funding Agency [4]. This dataset includes variables covering the characteristics of a learner, types of learning, dates, prior educational achievements [4].

In order to analyse the outcomes of those who have completed the apprenticeships we merge the two data sets, LEO and LILR.

First, it is helpful to define id's which allow us to merge these tables. RECORDID is a unique identifier of learners located both in LILR AIMS table and LEO FULL LOOKUP WITH AE ID table. External id is a unique key which allow us to identify given learner across LILR tables. External id is a combination of learner, provider, year and dataset id. EduKey is the LEOs unique learners' identifier. It allows us to track the earning outcomes of the learner across LEO tables.

We create a table ID13 which includes unique identifiers allowing us to access the individuals both from LEO and LILR tables for those who have started an apprenticeship in the academic year 2013/2014. We create this table in a single query. Within this query, we create a 3 temporary tables. It gives better structure to understand how merging process is done. First table selects the external id of all apprenticeship starters in the academic year 2013/2014. We get this information from the table LILR.AIMS. Second table selects the RECORDID of the 2013/2014 apprenticeship starters from the LILR.LEARNER table. To do that, we inner join LILR.AIMS table with the first table on the external id and get the RECORDID. Above manipulations allow us to access the external id and RECORDID of apprenticeship starters in the academic year 2013/2014. The 3rd table access the LEO data set. We select RECORDID and EDUKEY of all learners from the academic year 2013/2014. We do that by inner join of LEO tables FULL LOOKUP WITH AE ID and AE ID TO EDUKEY LOOKUP on AE ID, which is a unique identifier between these two LEO tables. Given the three above tables, we select unique external id and EduKey of apprenticeship starters in the academic year 2013/2014. We do that by inner joining the 2^{nd} LILR table (including external id and RECORDID) with 3^{rd} LEO table (including RECORDID and EduKey) on the RECORDID. This gives us all identifiers of learners which we can use to access all remaining LEO and LILR tables.

We create a table ID13 which includes 66% of all the apprenticeship starters. Although, it is possible to create a table which includes the 100% match between LEO and LILR (and we have done it), we then cannot fully ac-

cess the following tables LEO LEARNERS, LEO SELF-ASSESMENT, LEO COHORTS. This is possibly because of problem with data coverage and construction of the PMR data tables. PMR is an acronym for the Pupils Matching Reference. PMRs and EDUKEYs identify unique individuals, where a PMR is directly or indirectly matched to more than one EDUKEY [5]. Those with ILR but no PMR will not have a school level information including the region name, employment information such as days worked in a tax year, sustained employment, sustained benefit, sustained learning. In case of the 2013 apprenticeship starters cohort, there are 34% of such cases. For the purpose of training machine learning algorithms, we only consider 66% matched cohort. Otherwise we get a substantial amount of missing values, which have huge negative impact on the performance of machine learning algorithms. Furthermore, it is worth to note that those who were not matched are on average older (37 years old) than those who were matched (19 years old). We need to keep this limitation in mind when interpreting the findings of our models, predicting dependent variables, interpreting the EDA. We write the table with identifiers into the MS SQL Server and save it as a csv file called ID13 for easy access during further data derivation.

2.1.2 Creating Data Set

We create a full raw data set required for the analysis by accessing necessary variables from the given table and recording external id and EduKey next to it. We save these tables as a .csv files in a secure drive. Later we link all the files together using sql light in python. We do it in steps as the MS SQL Server crashes once we try to perform all operations at once. It is also faster to update one table when we want to include additional variable in our data set and compile it in SQL light, than to run big query in the MS SQL Server.

We first derive the outcomes of the individuals. The outcome is both the successful competition of apprenticeship in any year, as well as earnings 1 year, 2 years, 3 years after starting an apprenticeship.

We access the earnings of 2013/2014 apprenticeship starters by doing inner join between LEO EARNINGS APR 18 table with ID13 table on a field EduKey. We include both external id and EduKey next to earnings. We save derived 4 tables of earnings as a .csv file. We name earnings tables EARNINGS13, EARNINGS14, EARNINGS15, EARNINGS16, where 13 indicates that these are earnings for the 2013/2014 tax year.

Next, we access the outcome variable for the 2013 cohort. Derived

tables includes external id, EduKey, academic year and outcome. First, we access LEO AIMS table and inner join it on external id with the ID13 table. One lerner can appear several times in the table because they start several apprenticeships within given year. We take into consideration the outcome of only most recent apprenticeship, as indicated by the start date coming from LILR AIMS table. We further inner join above tables on external id with LILR AIMS table to include only those who have started an apprenticeship in the 2013/2014 academic year. We name outcomes tables outcome13 where 13 indicates the academic year 2013/2014.

The next table we use is LEO LEARNERS table. It includes variables external id, EduKey, academic year, days worked in a tax year, sustained employment, sustained benefit, sustained learning. We first join the LEO LEARNERS table with ID13 table on EduKey. Next, we join LILR AIMS 16/17 table with ID13 table on external id. We take into consideration only the entries from most recent start date in case of duplicates of the same learners. We inner join above tables with LILR aims table on external id to include only those who started apprenticeships in the academic year 2013/2014. We name the table LEO LEARNERS 13 table, where 13 indicates academic year 2013/2014.

We next use table including external id, EduKey, academic year and self-employment variables. We inner join LEO SELF-ASSESMENT table with the ID13 table on EduKey field. We then join LILR AIMS table with ID13 on external id. We take into consideration only entries from the most recent start date in case of duplicates of the same learners. Last, we inner join LILR AIM table with ID13 table on external id. We name this table SELF EMPL 13, where 13 indicates tax year 2013/2014.

Next table we derive includes variables external id, EduKey, academic year, gender, region, ethnic group major. We inner join LEO COHORTS table with ID13 table on EduKey field. We then inner join LILR aims 16/17 test table with ID13 table on external id field. We take into consideration only entries from the most recent start date in case of duplicates of the same learners. Last, we inner join LILR AIM table with ID13 table on external id. We name this table COHORTS 13, where 13 indicates tax year 2013/2014.

Another table we use includes variables external id, EduKey, academic year, completion, age at starting the apprenticeship, national level. We first inner join LILR AIMS 1617 test table with ID13 table on external id. We take into consideration only entries from the most recent start date in case of duplicates of the same learners. Last, we inner join LILR AIM table with ID13 table on external id. We name this table LILR AIM 13, where 13

indicates academic year 2013/2014.

The last table we use includes fields external id, EduKey, academic year, disability, learning difficulties, former prisoner status, prior attainment. We first inner join LILR LEARNERS 1617 FINAL table with ID13 table on external id. We then join LILR AIMS 1617 test table with ID13 table on external id. We take into consideration only the entries from most recent start date in case of duplicates of the same learners. Last, we inner join LILR AIM table with ID13 table on external id. We name this table LILR LEARNERS 13, where 13 indicates academic year 2013/2014.

In total, we have saved 11 .csv files which we compile together in SQL light. We load all above tables to the SQL light and perform a left join operation with the first table ID13 including external id and EduKey of apprenticeship learners of academic year 2013/2014. Left join assures that the match rate does not change when we perform joins even in case of missing values. We left join ILR tables, LEO LEARNERS, and COHORT tables on external id. We join EARNINGS tables on EduKey, as it allow us to track earnings of individuals across years.

After the above derivations, we save the output of the last SQL query as a .csv file. This file includes the raw data set which we will analyse in the EDA section. We call the file data 2013 raw.

For the purpose of model validation, we further construct data set including outcomes of the 2011, 2012 cohort with 99% match. We also create the data set of 2014 cohort with 99% match and forecast their employment outcomes. We do not consider the data before year 2011 due to different methodology in data collection and data quality issues.

The above methods and SQL queries has been checked by Data Scientist within the Institute for Apprenticeships and Technical Education. We have also performed quality assurance by checking if the data does not contain duplicates of either EDUKEY or external id. We made sure that the number of lerners within the derived data set is close to the number of apprenticeship starters.

The created raw data sets include following 21 variables (table 2.1). The first 18 variables, excluding external id and EduKey, are explanatory variables. Earnings for tax years 2013 to 2016 and outcomes are outcome variables.

Table 2.1: Variables of the 2013 apprenticeship starters data set [4] [3]

name	**categories**	**description**
external id	N/A	unique identifier of learner in the LILR tables
EduKey	N/A	unique identifier of learner in the LEO tables
gender	m, f	either male, or female
region	East Midlands, East England, London, North East (NE), North West (NW), South East (SE), South West (SW), West Midlands Yorkshire	name of the region where a learner attended school
enthicity	Any other ethnicity group (AOEG), Asian, Black, Chinese (CHIN), Mixed (MIXD), unclear (UNCL), white (WHIT)	self-declared ethnicity
days worked	N/A	number of days in employment within a tax year
sustained employment	1, 0	sustained employment one day in month October to March
sustained learning	1, 0	sustained learning one day in month in 6 months
sustained benefit	1, 0	sustained learning one day in month in 6 months
completion	1, 0	aim is completed in the academic year
age	N/A	age at the start of learning aim

name	categories	description
national level	0 = Entry Level, 1 = Level 1, 2 = Level 2, 3 = Level 3, 4 = Level 4, 5 = Level 5, 6 = Higher Level, 9 = Other Level, 22 = Not Known	national level of qualification for all aims
disability	22 = Missing (Not Applicable/ Not Known), 01 = Visual Impairment, 02 = Hearing Impairment, 03 = Disability Affecting Mobility, 04 = Other Physical Disability, 05 = Other Medical Condition (For Example Epilepsy, Asthma, Diabetes), 06 = Emotional/Behavioural Difficulties, 07 = Mental Health difficulty, 08 = Temporary Disability After Illness (For Example Post-Viral) or accident, 09 = Profound Complex Disabilities, 10 = Asperger's syndrome, 90 = Multiple Disabilities, 97 = Other, 98 = No Disability, 99 = Not Known/Information Not Provided	detailed disability category
learning diff	22= Missing (Not Applicable/ Not Known), 01 = Moderate Learning Difficulty, 02 = Severe Learning Difficulty, 10 = Dyslexia, 11 = Dyscalculia, 19 = Other Specific Learning Difficulty, 20 = Autism spectrum disorder, 90 = Multiple Learning Difficulties, 97 = Other, 98 = No Learning Difficulty, 99 = Not Known/Information Not Provided	detailed learning difficulty category
OLASS	1 = OLASS learner in custody, 2 = OLASS learner in the community, 3 = OLASS learner in custody and in the community, 9 = Not an OLASS learner, 22 = Missing not known	Offenders' Learning and Skills Service learner
prior attainment	N/A	level of qualification hold before starting learning aim

name	categories	description
self employment	0, 1	flag if the employee is self employed
outcome 13	1 = Achieved (non AS-level aims), 2 = Partial Achievement, 3 = No Achievement, 4 = Exam Taken/ Assessment Completed But Result Not Yet Known, 5 = Learning Activities Are Complete But The Exam Has Not Yet Been Taken And There Is An Intention To Take The Exam/Assessment, 6 = Achieved but uncashed (AS-levels only), 7 = Achieved and cashed (AS-levels only), 8 = Learning activities are complete but the outcome is not yet known, 9 = Study Continuing	outcome for the learning aim
earnings 13	N/A	yearly earnings for the tax year 2013/2014
earnings 14	N/A	yearly earnings for the tax year 2014/2015
earnings 15	N/A	yearly earnings for the tax year 2015/2016
earnings 16	N/A	yearly earnings for the tax year 2016/2017

2.2 Data Processing

This section focuses on pre-processing data. We conduct sanity checks and focus on handling missing values.

2.2.1 Missing Values Theory

There are three missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR) [6]. In case of MCAR and if the number of missing values is small relative to our data set, we may simply ignore those values and remove them from our dataset. If the data is missing at random, it means that there is some pattern, but it is expected - for instance women reporting their weight less often than man [6]. Then we may assume that it is desired feature and take a note of it during our analysis. If the data is MNAR we should include this fact in our model [6].

There are many statistical methods for missing data. They include maximum likelihood estimation, multiple imputation, full Bayesian methods, weighted estimating equations, missing outcomes versus covariates. For the purpose of this project we have decided to use missing outcomes versus covariates method [6]. We use this method because missing values in case of most of the data are not a large problem, as they contribute to as little as 10% of the whole records. Missing outcomes versus covariates method estimate the missing by the regression, treating the variable with large number of missing values as the dependent variable, and treating covariates as exploratory variables [6]. The advantage of this approach is that is it easy to implement, not computationally intensive, results are straightforward to interpret, and it gives better estimation than simple missing values estimators such as replacing missing values with mean or mode. Disadvantage of the missing outcomes versus covariates approach is that it assumes the linear relationship between covariates and dependent variable, it assumes the normality of the data and it states that the covariates are independent of each other [6]. This assumption is violated in many cases. However, due to relatively small numbers of missing values we will still use this method for continuous variables as it gives best results given the required workload.

We estimate the continuous missing values of variables earnings 13, earnings 14, earnings 15, earnings 16, age, days worked in a tax year by an experimental sklenar library called iterative imputer. As we conservatively assume that the missing values are missing not at random, for the missing values in categorical variables, we introduce the new category signalling that the given value was missing. This approach has the advantage that it allows us to forecast employment outcome of a new learner in case they have a missing value in some field. It may also allow us to capture some systematic relationship between having missing value and employment outcome [6]. In case of text

Table 2.2: Mising values within the 2013 apprenticeship starters cohort (%)

variable	gender			total
	male	female	missing	
self-employment	44	45	3	92
sustained benefit	40	36	3	79
disability	7	7	3	17
learning diff	7	7	3	17
earnings 2013	5	4	<1	10
prior attainment	3	3	3	9
earnings 2016	4	3	<1	8
sustained employment	2	2	3	7
earnings 2015	3	3	<1	6
ethnic group	1	1	3	5
earnings 2014	3	2	<1	5
days worked	<1	<1	3	4
outcome	<1	<1	3	3
age	0	0	3	3
completion	0	0	3	3
national level	0	0	3	3
region	0	0	3	3
sustained learning	0	0	3	3

variable, we denote the missing value as NA. For categorical variables having numerical categories, we input 77 as the missing value. We choose this value because it does not occur in the data set and it is easy to distinguish from others.

2.2.2 Missing Values

There is a substantial amount of missing values in some of the variables. Before we proceed to the data analysis it is necessary to manage missing values by either removing them, or estimating its value.

Table 2.2 shows the number of missing values within considered data set. We may note that the variables gender, region, ethnic group, days worked, sustained employment, sustained learning, completion, age, national level, have very small % of missing values. This could be because these data come

from LILR data set which has much higher match ratio than LEO. Also, these variables seem not to be sensitive. Hence, we may expect that the small amount of missing values is missing completely at random. We will confirm that in the modelling chapter. Variables disability, learning difficulties have similar percentage of missing values. We investigate this issue in the EDA section. We would not expect that these values are missing completely at random. Similar percentage of missing values in this case could be because of high correlation between both variables. Furthermore, those individuals may not want to disclose this information. Binary variable self-employed has 92% of missing values. This is quite surprising as we would not expect this variable to be sensitive information. 6% of learners are self-employed and 3% of learners are not self-employed. We would expect more not self-employed learners, as by definition apprenticeship is an on-job training. Staff within the Institute for Apprenticeships and Technical Education claimed that most probably only those who are self-employed are flagged in the data set. Variable sustained benefit has 79% of missing values. We suspect that it might be due to the problem with updating tables, as data about sustained benefit is easily available from the HMRC. Another interesting phenomenon is a substantial percentage (10%) of missing values in earnings for cohort 2013/2014 learners in 2013/2014 tax year. However, the percentage of missing values in earnings drops to 5% in 2014/2015 tax year and then gradually increases to 8% in 2016/2017. The initial high number of missing values may be caused that many of those who start an apprenticeship do not have earnings, as they join the programme directly from full-time education. Once they start an apprenticeship it decreases, as they receive salary as a part of a programme. It is hard to explain the increasing number of missing values between 2014/2015 to 2016/2017 tax year. One of the hypothesis may claim that because there is a substantial number of older people enrolled in apprenticeship [2] they retire or die and we cannot track their further earnings. The number of missing values may increase also due to time it takes to update these tables by the HMRC and later DfE.

2.2.3 Sanity Checks

Before we run the iterative imputer, we declare all categorical variables as categorical. Because the iterative imputer is based on multiple applications of the linear regression, it does not restrict the range of the dependent variable (unlike logistic regression). Because of that application of iterative imputer result in the out of range predictions. There are 1899 cases of variable earnings 13, earnings 14, earnings 15, earnings 16 where the filled missing values

Table 2.3: Outliers within variables of 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
days worked	7	6	0	12
age	<1	<1	3	3
earnings 2013	1	<1	<1	2
earnings 2016	1	<1	<1	2
earnings 2014	1	<1	<1	1
earnings 2015	1	<1	<1	1

of earnings are negative. We replace these negative values with 0. Furthermore, there are 4436 cases where missing value of age was imputed as below 10, and 20 cases where age is above 100 years. We replace these values with mode age 17. There are 8679 cases the number of days worked has been predicted to be above 366 or below 0. We replace these values entries with 0 for cases below 0 and 365 for cases above 366. Relative to the whole data set, this number of wrong entries in earnings and age is about 1% of the data set, while numbers of days worked is less than 3%.

For the purpose of the further analysis we also need to manage problem of negative values in the categorical data. The Naive Bayes classifier from sklenar library does not accept the negative entries, even if they are categorical values. Some categorical variables takes values -1 to indicate belonging to a given category. We replace -1 by 22 in categorical variables.

We save the file as data processed 2013.csv. We follow the same methodology of data processing for data sets including apprenticeship starters in year 2011, 2012, and 2014. We will use the processed data sets during further analysis, model fitting and forecasting.

2.3 Exploratory Data Analysis

We define an outlier as the observation located 1.5 inter-quartile range away from upper quartile, or from lower quartile. Table 2.3 shows that variables age and earnings has very small % of outliers. Variable days worked has a very high number of outliers (more than 20%). Most of these outliers

Table 2.4: Summary statistics of earnings (1000s GBP), age, and days worked 2013 apprenticeships starters

variable	**mean**	**std. dev**	10^{th} **centile**	25^{th} **centile**	50^{th} **centile**	75^{th} **centile**	90^{th} **centile**
days worked	352	50	349	366	366	366	366
age	35	17	18	19	30	45	59
earnings 2013	7.87	6.46	1.32	3.15	6.35	11.56	16.20
earnings 2014	10.72	6.87	3.17	5.91	10.12	14.55	18.63
earnings 2015	12.72	7.78	3.19	7.49	12.62	16.95	21.50
earnings 2016	14.43	9.36	3.36	8.43	14.42	19.16	24.54

are 1.5 inter-quartile range away from the lower quartile. That is because the median of days worked is 366 days and indeed majority of people were employed for 366 days of a year in 2013. The lower quartile of days worked is 366 days. That means that almost all workers who were not employed all the time during the year are counted as outliers. For that reason, we are not concerned about the % of outliers in the variable days worked, as it does not necessarily mean that the data is very heterogenous. It is just a feature of the data set.

2.3.1 Univariate EDA

This section describes the balance of categorical variables and distribution of continuous variables. It intends to give the overview of the created data set mostly in the form of summary statistics.

Table 2.4 describes the distribution of all continuous variables. We may note that the 2013 apprenticeship starters were mostly employed during all days of the year. The median of variable days worked within a year is 366 and the standard deviation is 50. It suggests relatively small variability. It is interesting to look at the summary statistics of earnings. The mean initial salary is £7870, and median is £6350. Given that majority of 2013 cohort work most of the year these are relatively small earnings. In fact, the minimum wage in the UK from October 2012 was £6.19 per hour, but the minimum wage for apprentices was £2.65 per hour. It is also worth to note that the minimum wage for those aged 16-17 was £3.68 per hour, and for those aged 18-20 was £4.98 per hour [11]. This could partially explain relatively small initial earnings of apprentices. We may make a rough estimation of earnings per hour by dividing mean earnings by mean days worked.

Then we get a salary of £22.34 per day. We can divide it further by 8 hours worked per day and we get £2.79 per hour. This suggest that the mean earnings of those who started apprenticeship in 2013/2014 are sensible given the minimum wage noted above. Of course the rough calculation we make has a limitation that we assume that given person works 322 days per year, which is not true as this variable relates to the number of days in employment per year. Some apprentices could start working late during a given tax year, and then their salary is respectively lower. Also, it is worth to note that the median age of the 2013 cohort is 30 and the standard deviation of age is 18. This suggest large variability in age. The Learners Survey publication [2] has mentioned that age of apprentices is highly variable and apprenticeship starters consists of substantial amount of older subjects. They also mention that the socio-economic background of those who start an apprenticeship tends to be lower relative to the population. This fact can explain the above deviation in wage distribution compared to the English population.

When looking at figure 2.2, we may note that the earnings are linearly increasing across years. While mean earnings in 2013 are £7870 they increase to £14,430 three years after starting an apprenticeship. We may also note that the variability of earnings increases as well, and this increase does not seem linear. The standard deviation of earnings in 2013 is £6460 and it increases to £9360 in a tax year 2016/2017. At the same time, mean earnings increase by about £2000 each year after starting the apprenticeship. Within all years, there is always some unemployed learners earning 0. We may note that while the 10^{th} centile is around £3000 across all years, while the 90^{th} centile earnings are gradually increasing.

Figure 2.1 shows the distribution of age among 2013 apprenticeship starters. We may note that this is a mix of two distributions, known as bi-modal distribution, with different variances and different means. The first peak around age 18 shows that there is a substantial amount of young apprenticeship starters. These are mostly students who join an apprenticeship programme directly after finishing the full-time education. It appears to have low variance and is leptokurtic. Another peak around age 40 is mesokurtic and is less visible. We may note that it has heavy right tail. That means that there is no clear age for those who start the apprenticeship at the later age. As the leraners survey publication [2] mentioned, these are mainly employees starting apprenticeship for their current employer. The above analysis

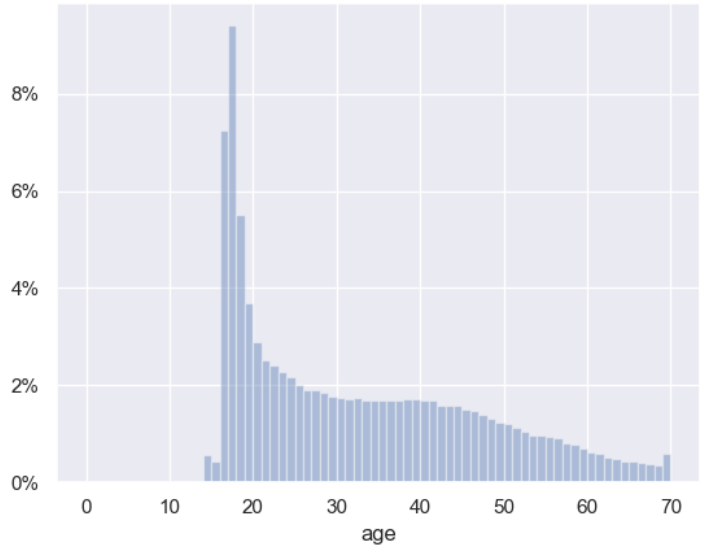


Figure 2.1: Age at the start of the learning aim, 2013 cohort of apprenticeship starters

suggests that the summary statistics in table 2.4 can be misleading, as it aggregates outcomes of different groups of learners.

Figure 2.3 shows the distribution of initial earnings. We may note that they are skewed to the right and peaks near £5000. This distribution is like the distribution of earnings within the population of England [27]. However, it is significantly moved to the left relative to the general population distribution. It can be related to the fact that those who enrol in an apprenticeships usually come from lower socio-economic backgrounds and that the minimum salary for apprenticeship is lower than the UK minimum salary. It is also worth to note that within this population there is a substantial number of students who just started their careers. As the figure 2.3 presents the initial earnings of 2013 cohort, there are many subjects who did not have any initial earnings, as they were enrolled in a full-time education before. Indeed, we may note that figure 2.4 presenting the earnings of 2013 cohort in the tax year 2016/2017 reminds more the distribution of earnings of the UK population [27], that is it is still skewed to the right, its peak is located near £16,000. However, it still has a substantial number of learners earning about £0. This might be also cause by the iterative imputer described in the previous section, who estimated some earnings to be negative, and during sanity checks we set it to be £0.

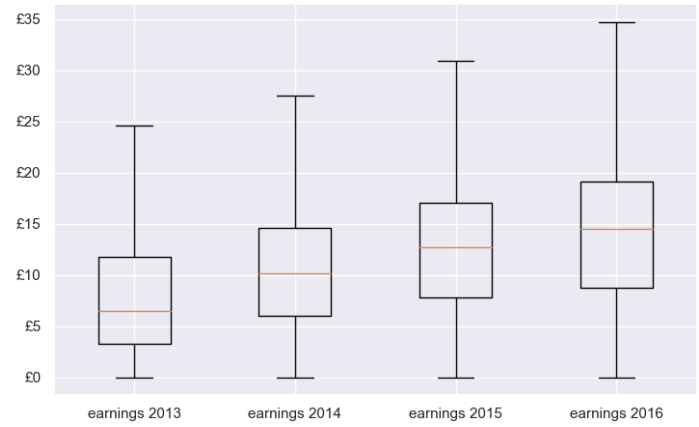


Figure 2.2: Time plot of distribution of earnings (1000s GBP) of 2013 cohort of apprenticeship starters. Deflated using ONS CPI with 2015 = 100. Learners having 0 earnings excluded.

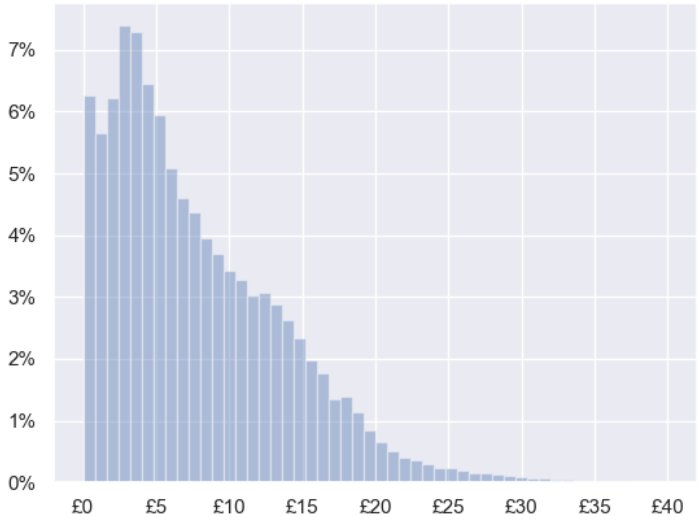


Figure 2.3: Distribution of earnings (1000s GBP) 2013/2014 tax year of 2013 apprenticeship starters.

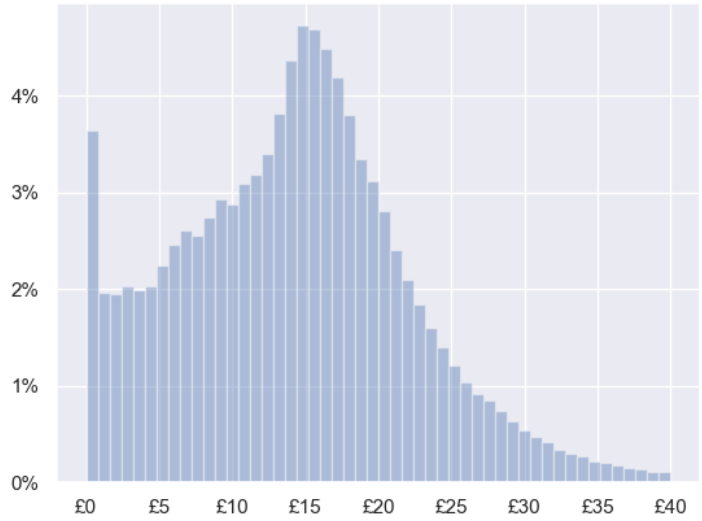


Figure 2.4: Distribution of earnings (1000s GBP) 2013/2014 tax year of 2013 apprenticeship starters

Table 2.5: Outcome counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
achieved	29	28	0	56
no achievement	11	10	0	22
study continuing	8	8	0	16
missing	3	< 1	< 1	3
partial achievement	< 1	< 1	0	1
achieved and cashed	< 1	< 1	0	< 1
achieved but uncashed	< 1	< 1	0	< 1
completed learning	< 1	< 1	0	< 1
exam takes	< 1	< 1	0	< 1
learning completed but outcome not known	< 1	< 1	0	< 1

Table 2.6: Region counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
North West	9	8	0	17
South East	6	6	0	13
West Midlands	6	6	0	12
Yorkshire and The Humber	6	6	0	11
East Midlands	5	5	0	10
South West	5	5	0	10
East of England	5	5	0	9
London	4	4	0	8
North East	4	3	0	7
missing	0	0	3	3

Table 2.5 shows that the variable outcomes is not well balanced. Most learners have positive outcome. 56% have achieved non AS-level aim, 22% have not achieved the learning aim and 16% of learners continue their studies. 1% means that the number is less than 1% of the 2013 cohort sample.

The gender variable is relatively well balanced. 50% of apprentice are male and 47% is female. Although number of missing values is small 3% , we will investigate in the next chapter which gender is missing more often.

Table 2.6 shows the number of apprenticeship starters in 2013/2014 academic year for a given region. This variable is relatively well balanced. We may note that each region consists of substantial amount of apprenticeship starters. Some regions (North West and South East) are more represented than others. This however seem to be proportional to the populations of these regions. For instance, North West has twice as much apprenticeship starters as North East. There is an obvious reason, as North East has a population of 2.66mln, while the population of North West is 7.29mln [21]. It is surprising that London has just 8% of starters, while the population of this region (8.9mln) is higher than North West. These suggests that there are some exogenous factors which impact the number of apprentices within given region other than its population. We suspect that this might be correlated with the fact that the apprenticeship starters usually come from a low socio-economic background. As London is relatively expensive city to live these

Table 2.7: Ethnic group major counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
White	44	41	0	85
missing	1	1	3	5
Asian	2	2	0	3
Mixed	1	1	0	2
Black	1	1	0	2
Any Other Ethnic Group	< 1	< 1	0	1
Chinese	< 1	< 1	0	1
Unclear	< 1	< 1	0	1

individuals tend to settle outside it.

Table 2.7 shows the number of apprenticeship starters in 2013/2014 academic year by ethnic group major. We may note that the highest number of apprenticeship starters is of White ethnicity. This is in line with the expectations, as this is dominant ethnicity in the UK. The remaining ethnicities are equally represented, except for Asian ethnicity which has 3% of apprenticeship starters. This is still relatively small compared to the total number of observations. We may note that there is a substantial number of missing values (5%).

The vast majority (93%) of apprenticeship starters are in the sustained employment, and only 1% are declared as not being in the sustained employment. It is quite interesting that the number of missing values is quite substantial, much higher than the number of those who are not in the sustained employment. The 6% of missing values may come from the fact that that those who are not in sustained employment are not willing to declare that. This may result in substantial number of missing values. We analyse this phenomenon further while fitting the logistic regression. In the previous section we have mentioned that we treat missing values in categorical variables as a separate category. This will allow us to check if there is any systematic pattern within the missing category. If the coefficient next to missing values category will not be statistically significant that would mean that there is no pattern within those who have missing values in sustained employment variable.

Majority of learners within 2013 cohort were in sustained learning at least one day in 6 consecutive months when they started the apprenticeship. 94% of apprenticeship starters were in the sustained learning, while only 3% are declared as not in the sustained learning. There is a small number of missing values 3%. The above distribution is in line with the expectation, as those who are starting apprenticeship usually receive in job training, that is why majority of apprenticeship starters should be in the sustained learning. The fact that relatively small number of apprenticeship starters are not in the sustained learning might be because the duration of apprenticeship was shorter than 6 months and they did not continue studying afterwards, or they dropped out of apprenticeship.

The variable sustained benefit indicates if the learner was claiming the sustained benefits at least one day in a month in 6 consecutive months. We may interpret this flag as being in unemployment in a long period of time. We may note that 21% of the apprenticeship starters were claiming the unemployed benefits. This is a substantial proportion. It could be because the starters were unemployed either before the apprenticeship, or in case of shorter apprenticeship, they could be unemployed shortly after finishing the apprenticeship. It is interesting to note that there are 0 entries of those who were not taking sustained benefits, but that the 79% of values is missing. We assume that this high number of missing values is due to wrong data entry. We would expect that majority of those who started an apprenticeship are not taking employment benefits. For that reason, during further data analysis and modelling we will assume that missing values in that case represent those who were not taking unemployment benefits.

To support this assumption, we checked that the mean number of days for those with missing sustained benefits flag is 355. We think that this is sufficient evidence to count those who have missing values as not taking sustained benefits, as they are working.

The completion variable shows if the aim has been completed in the given year. We may note that 68% has completed the learning aim within the given year and 29% has not. There is a very small (3%) number of missing values. Thus, majority of the apprenticeship starters have a positive outcome of their apprenticeship, and not including this variable as one of our features in the model should not impact the wage significantly.

Table 2.8 shows the national level of qualification for learning aims of 2013 starters. It corresponds to the qualification levels listed by the Regulated Qualifications Framework. For instance, qualification level 3 corresponds to A-level education. We may note that most of the apprenticeship

Table 2.8: National level counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
2	15	14	0	29
other level	13	12	0	25
1	11	11	0	22
3	8	7	0	15
Entry Level	2	2	0	4
missing	< 1	< 1	3	3
4	1	1	0	2
5	< 1	< 1	0	< 1
higher level	< 1	< 1	0	< 1
not known	< 1	< 1	0	< 1

starters start learning aim equivalent to qualification levels 2 and 3 (more than 60% of all starters). That could further explain the relatively low mean of earnings for 2013/2014 tax year. It is worth to note that less than 1% of apprenticeship starters started learning aims at the higher qualification level. This level corresponds to the bachelor's degree and higher.

Table 2.9 shows counts of learners with different disabilities. We may note that most learners (68%) have no disability. There is also a significant number of missing values (17%). This could be caused by learners not willing to admit that they have a disability, or due to sensitivity of the data.

We may note that these are the same records who both have missing value in disability indicator and learning difficulty indicator. That would suggest that there is some systematic phenomenon which fails to report the disability and learning problems. We would expect it to be caused either by not willingness to report disability/learning problems, or some technical issue. When we look at the average initial earnings of those with missing values, we discover that they are like the whole sample average (£7860). That could suggest, that this variable is missing at random. We also could not find any other variable which would allow us to distinguish learners with missing disability from the population.

Table 2.10 shows counts of learners with learning problems. We may note that the majority of learners do not have any learning problems (68%).

Table 2.9: Disability counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
no learning difficulties	35	33	0	68
missing	7	7	3	17
not known/information not provided	5	5	0	9
other	< 1	< 1	0	1
other medical condition	< 1	< 1	0	1
asparger's syndrome	< 1	< 1	0	< 1
disability affecting mobility	< 1	< 1	0	< 1
emotional/behavioural bifficulties	< 1	< 1	0	< 1
hearing impairment	< 1	< 1	0	< 1
mental health difficulty	< 1	< 1	0	< 1
multiple disabilities	< 1	< 1	0	< 1
not applicable/not known	< 1	< 1	0	< 1
other physical disability	< 1	< 1	0	< 1
profound complex disabilities	< 1	< 1	0	< 1
temporary disability after ilness	< 1	< 1	0	< 1
visual impairment	< 1	< 1	0	< 1

Table 2.10: Learning difficulty counts for 2013 apprenticeship starters (%)

variable	gender			total
	male	female	missing	
no learning difficulty	35	33	0	68
missing	7	7	3	17
not applicable/not known	5	5	0	10
dyslexia	1	1	0	2
moderate	< 1	< 1	0	1
autism spectrum disorder	< 1	< 1	0	< 1
dyscalculia	< 1	< 1	0	< 1
multiple learning difficulties	< 1	< 1	0	< 1
other	< 1	< 1	0	< 1
other specific	< 1	< 1	0	< 1
serve	< 1	< 1	0	< 1

The most common learning problem is dyslexia, which occurs within 2% of learners. There is substantial number of missing values, including 17% values which are missing directly and records flagged with 22, meaning that the variable is either non applicable, or not known.

As we explained above, large amount of missing values is related to disability variable.We have checked if there is any particular region where learning problems and disability variables have significant number of missing values but all regions have well-balanced variables. That means that these missing values are not related to any particular place. This finding might suggest that despite high number of missing values, these values are missing at random.

We may note that there are more self-employed learners (5%) than not self-employed (3%) learners. We found this fact interesting, as we would expect those who have started apprenticeship not to be self-employed. As we have commented earlier, 92% of values are missing, and these are most probably not self-employed learners.

Those who are not self-employed have substantially higher average earnings (£12,110) than those who are not self-employed (£7170). Learner having self-employment variable flagged as missing value have average earning of £7760, which is closer to those who are self-employed.

In order to characterise those individuals with missing values, we have computed the average number of days of being in employment for those who are self-employed and those who are not self-employed. Those who are self-employed work on average 327 days. Those who are not self-employed are employed on average 359 days. Individuals with missing employment status work on average 353 days. This suggests that in this aspect those with missing employment status are more similar to individuals not being in self-employment. We would expect majority of learners not to be in sustained employment, as apprenticeships offer on-job training.

2.3.2 Multivariate EDA

Figure 2.5 is called a heat map. It is the graphical representation of the correlations between variables. The brighter the colour, the more positive correlation. The darker is colour, the more negative correlation. We need to keep in mind that this correlation map does not distinguish between different categories in categorical variables. Through below analysis, we will comment on strength of the relationship, not its direction.

First, we are mostly interested in correlations between earnings and other variables. Based on that, we may see if there are any redundant variables, which might not be related to earnings. Furthermore, we are interested in the correlation between variables other than earnings, because of assumption of independence in some of our models. In regression it is called the multicollinearity problem. The heat map could help us to detect any possible cases.

We may note that the variables sustained employment, days worked in a tax year, sustained benefit, self-employment have visible correlation with the earnings 2016 variable. Other variables seem to be little or no correlated with earnings.

The most visible correlation between exploratory variables is between sustained employment and days worked in a tax year. Both variables have an impact on the dependent variable, but they also appear to be strongly related to each other. For that reason, we need to consider removing one of them, as the assumption of independence might not hold. Other strong correlations include relationship between variables sustained employment, sustained learning, completion, age, national level, outcome. Despite possible multicollinearity problem, we decide to keep these variables and check it in more detail when running multinomial logit model.

Figure 2.5: Correlations between variables of interest

Table 2.11: Dayes worked by different groups

variable	**mean**	**std. dev**	10^{th} **centile**	25^{th} **centile**	50^{th} **centile**	75^{th} **centile**	90^{th} **centile**
sustained employment 2013	360	25	365	365	366	366	366
not sustained employment 2013	215	115	61	209	363	363	366

Table 2.11 shows the average days worked by those who are in sustained employment . As we would expect given the above heat map, those in sustained employment work substantially more than those who are not in the sustained employment. When building our models, we may be forced to eliminate one of those variables, due to multicollinearity problem.

Table 2.12 shows the difference in distribution of earnings within different groups. We focus mainly on earnings during the tax year 2016/2017, but in case of variables gender and outcome we also analyse initial earnings. We analyse differences in gender, ethnicity, employment status, outcome, sustained employment, and age.

We may note that there is a significant difference between earnings of males and females both initially in 2013 and 3 years later in 2016. The earnings gap widens three years later to £3,880. Variability of earnings also increases both in case of males and females. We may note that males have in general higher variability of earning than females. We could interpret this finding as a signal that males earn more in general, and their earnings tend to be more variable on the positive side. Although the 10^{th} centile earnings are initially similar, this changes in a tax year 2016/2017 in favour of males.

Earnings varies a lot within different ethnicities. Highest earning ethnicities include White, missing, unclear and Chinese. White and Chinese ethnicities have similar variability of earnings, but missing and unclear ethnicity have much higher variability than any other ethnicity. That shows high heterogeneity of learners within these groups and possibly high mean due to substantial number of high-earning individuals. We may note that top 10% of earners within these 2 ethnicities earn at least £25,000, which is more than any other ethnicity. To contrast, Chinese earn on average £14,510 but the $90th$ centile earners get at least £24,320.

The lowest earning ethnicities include Asian, mixed and Black. They earn on average around £13,000. The standard deviation tends to be lower than other nationalities. We may note that the Black ethnicity learners have standard deviation of £4780, which is almost twice as low as any other na-

Table 2.12: Earnings by different groups (1000s GBP) and other continuous variables for 2013 apprenticeships starters

variable	**mean**	**std. dev**	10^{th} **centile**	25^{th} **centile**	50^{th} **centile**	75^{th} **centile**	90^{th} **centile**
male 2013	8.34	7.22	1.36	3.28	6.60	12.23	17.50
female 2013	7.37	5.53	1.28	3.01	6.10	10.91	15.00
male 2016	16.32	10.77	4.14	10.07	16.14	21.42	27.41
female 2016	12.44	7.06	2.70	7.25	12.79	16.88	20.82
unclear 2016	14.94	16.11	3.08	8.13	14.47	19.77	25.67
White 2016	14.57	8.73	3.65	8.82	14.58	19.23	24.56
Chinese 2016	14.51	8.95	2.39	6.93	14.91	21.10	24.32
ethnicity missing 2016	14.24	14.46	2.95	7.78	10.96	19.33	25.22
any other ethnic group 2016	13.12	8.71	2.03	6.34	12.70	18.73	23.54
mixed 2016	12.94	8.76	1.90	6.00	12.65	18.30	23.60
Asian 2016	12.72	8.78	2.07	5.78	12.04	18.02	23.69
Black 2016	12.64	4.78	1.42	4.78	10.96	17.71	23.71
not self employed 2016	19.72	19.11	5.75	11.92	19.94	24.66	30.11
missing employment 2016	14.56	8.50	3.72	8.91	14.60	19.13	24.43
self employed 2016	8.69	12.85	0.84	3.06	6.85	12.20	18.02
missing outcome 2016	14.44	9.56	3.34	8.86	14.42	19.13	24.60
positive outcome 2016	14.42	9.48	3.38	8.39	14.41	19.16	24.53
sustained emploument 2016	14.81	9.26	4.071	9.09	14.71	19.39	24.74
not insustained employment 2016	5.83	7.03	0	0.06	2.68	10.18	16.56
age less than 20 earnings 2016	14.46	10.30	3.37	8.42	14.44	19.16	24.53
age 20 - 30 earnings 2016	14.41	9.17	3.36	8.43	14.38	19.19	24.56
age 30 - 40 earnings 2016	14.40	8.75	3.30	8.45	14.43	19.13	24.50
40yo - 50yo earnings 2016	14.46	9.40	3.33	8.41	14.40	19.24	24.63
age older than 50 earnings 2016	14.40	9.21	3.43	8.43	14.42	19.10	24.45

tionality. That suggests that learners with Black ethnicity earn consistently less than any other nationality. To sum up above discoveries, ethnicity seem to be related to earnings and its variability. There are some ethnicities earning substantially lower, and substantially higher than the average. We need to be careful when interpreting these findings, as there are other possible exogenous variables related to ethnicity, which have indirect impact on earnings.

Those who are self-employed earns substantially less than those who are not self-employed. We may note that the difference is very large (£7090). This could be because providers offering full-time employment tend to offer higher wages. It is also interesting that variability of earnings of those who are self-employed is smaller. This suggests that those who are not self-employed earn consistently higher wages. Also, those who are not self-employed tend to have more variability in earnings, on the right tail of the distribution. Those who have employment status as a missing value mean, median is closer to the not self-employed workers.

Distribution of earnings of those with positive and negative outcomes is surprisingly very similar within all aspects of distribution. We would expect earnings of those who have positive outcome to have substantially higher earnings than those with negative outcome. This suggests that learning outcome has no impact on earnings.

Age at the start of the learning aim seem not to have an impact on the outcome variable. We may note that variability of earnings of those who start a learning aim at the age less than 20 is higher. The distribution of earnings of other age groups is very similar, suggesting that age has no significant role in determining employment outcomes.

2.3.3 Dimensionality Reduction

PCA

Principal component analysis (PCA) is an unsupervised machine learning algorithm reducing the dimensionality of data [7]. Conducting PCA means computing the PC loadings and PC scores [7]. We have attempted to explain relationship between variables in the EDA section. Due to high dimensionality of the data set, it is challenging to make any claims. PCA provides us a machinery to understand the relationship and impact of considered variables. During the PCA analysis we consider only explanatory variables. Let's

call this set of variables X, where $X = (X_1, .., X_1)^T$. During the PCA we attempt to find a set of 1-dimensional summaries of X obtained by linear projection. Let's define the linear projection as $v \in \mathbb{R}^{13}$ for $v^T X$. We find c which maximizes the $Var(v^T X)$, subject to v having a fixed norm [7]. All linear projections need to be uncorrelated [7].

Definition of Principal Components [7].

Let $X \in \mathbb{R}^p$ be a random vector with $[X^T X] < \infty$. The 1st principal component (PC) loading is a vector $v_1 \in \mathbb{R}^P$, $|v_1| = 1$, that maximises the variance of $v_1^T X$, or in other words

$$Var(v_1^T X) \geq Var(u^T X)$$

,

$$\forall |u| = 1$$

For $k = 2, ..., p$, the k-th principal component (PC) loading is the vector $v_k \in \mathbb{R}^p, |v_k| = 1$, that maximises $Var(v_k^T X)$ subject to $Cov(v_k^T X, v_j^T X) = 0$ for all $j = 1, ..., k-1$. In other words,

$$Var(v_k^T X) \geq Var(u^T X)$$

, $\forall u \in \mathbb{R}^p$, $|u| = 1$ such that $Cov(u^T X, v_j^T X) = 0$, for $j = 1, ..., k-1$.

The random variable $\mathrm{v}_k^T X \in \mathbb{R}$ is called the k-th principal component (PC) score [7].

Table 2.13: Principal Components

PC	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	0.0035	(0.0030)	(0.0217)	(0.9677)	0.2235	0.02115	(0.1059)	0.0480	0.0095	0.0049	0.0259	0.0752	0.0333	(0.0337)
2	(0.0094)	0.0140	0.0094	(0.0428)	(0.0411)	(0.0493)	(0.0202)	(0.1154)	(0.0172)	(0.0182)	(0.0476)	(0.2829)	(0.9473)	(0.0460)

Researchers [8] distinct three school of thoughts concerning choosing the number of principal components including subjective methods such as scree plots, distribution-based test tools such as Bartlett's test, and computational procedures such as cross-validation. Each of this approaches have drawbacks and benefits [9]. According to researchers [9] none of these schools of thoughts became a standard approach.

The figure 2.6 'shoulder rule' method aims to determine the point where adding another principal component does not explain much variance in comparison to previous principal components. We may identify shoulder on a 3^{rd} principal component. According to this method, we choose number of PCs

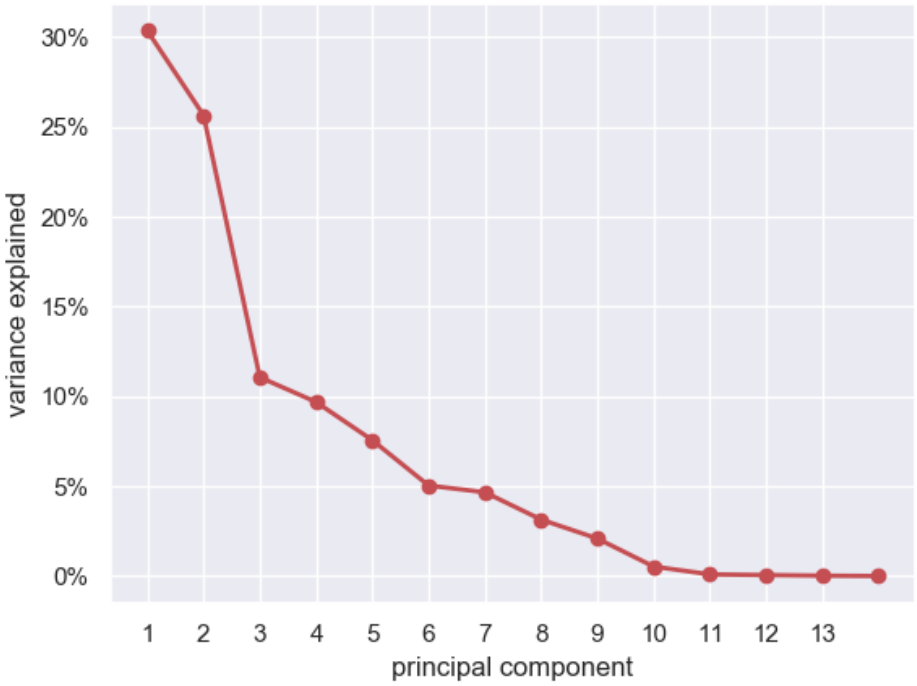


Figure 2.6: Variance Explained by the first 14 principal components for the 2013 cohort

just before the shoulder, so in the case of the above data, we choose 2 PCs. The first two PCs explain 56% of variance.

Another method suggests choosing the first n principal components which explain at least 90% of variance. According to this rule of thumb method, we should choose first 6 principal components which explain 94% of variance. We decided to use the elbow rule, as it does not require us to set the cut-off rate and it is clearer to interpret less PC loadings.

To interpret PC scores, we set a 0.2 threshold for variable to play significant role in the PC score. The first PC loading tells us that the first PC score is essentially the contrast between sustained employment and days worked in a tax year variables. The second entries of PC loading tells us that the second PC score is essentially average of variables prior attainment and OLASS learner status.

The first principal component emphasizes the importance of the prior employment as one of the indicators of ending up in one of the income groups.

As we have mentioned above, both sustained employment and days worked in a tax year are highly correlated, then the above finding is sensible. The second principal component explains the similar amount of variance as the first principal component, and it highlights the importance of prior attainment and OLASS learner status on employment outcomes. Weighted average means that both of those variables have similar weight. To conclude, both prior sustained employment, prior attainment, and OLASS status are important variables forming distinguishable clusters.

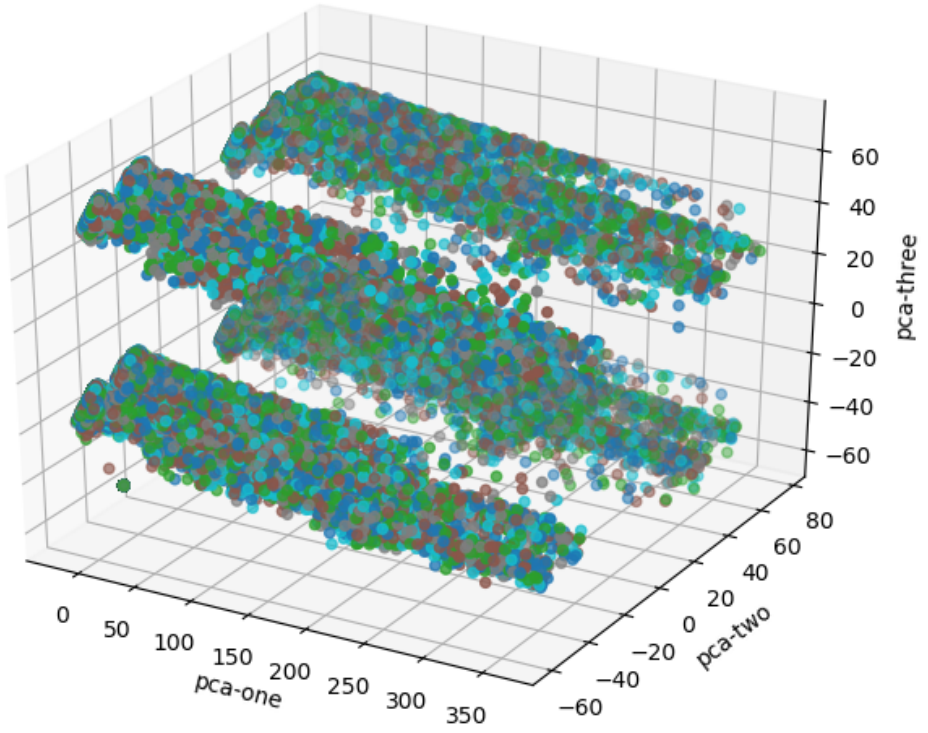


Figure 2.7: PC scores

We have attempted to create a plot with first two PC scores, but we could not obtain clear separability between clusters. Figure 2.7 shows the first 3 PC scores explaining 67% of variance. PC of 5 income groups have huge overlap, suggesting that this unsupervised learning algorithm does not perform well for the income class prediction problem.

t-SNE

t-distributed stochastic neighbour embedding (t-SNE) is a clustering algorithm allowing for nonlinear dimensionality reduction [10]. We suspect that our data set can have non-linear patterns, thus we want to test if t-SNE could be more relevant for the clustering by income group. We first reduce the dimensionality of our data to 6 features using PCA. As we have mentioned above 6 features preserve 94% of variance. We then take a random sample of 10,000 records. We reduce the dimensionality using PCA as it is recommended for better separation of the data. It also helps with computation time. We take a random sample to further decrease the computation time, as t-SNE is very computationally expensive [10].

Figure 2.8 shows that the t-SNE does not improve separability of income groups. We do not describe this algorithm in detail, as relative to the above PCA it does not offer improvement.

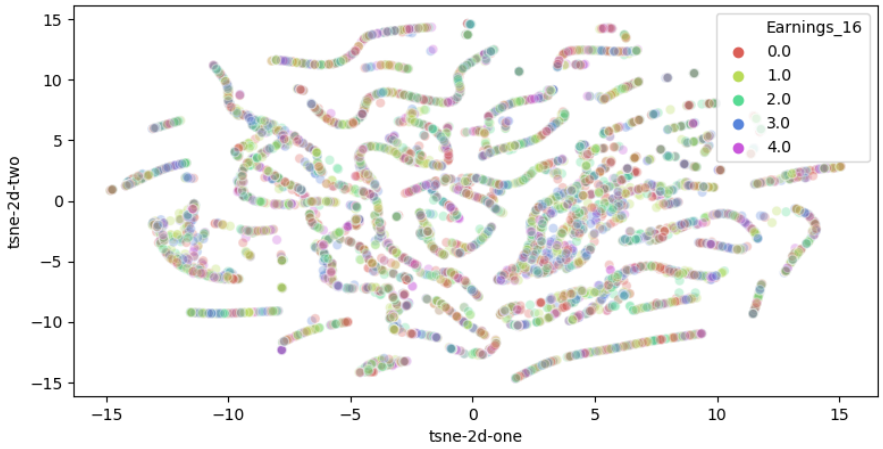


Figure 2.8: t-SNE

Chapter 3

MODELS

One of the aims of this thesis is to model the relationship between characteristics of individuals and forecast if they will be successful after starting an apprenticeship. In this chapter we introduce models used to forecast the earning group of learners three years after starting an apprenticeship. We explain methodology, parameters and we evaluate performance of the models. By the end of this chapter we choose the best model and validate its fit on 2011 and 2012 cohorts. We also attempt to forecast outcomes of 2014 cohort.

This chapter consists of 7 sections. Section 1 describes used models. Section 2 builds on feature selection from the previous chapter and selects features for the logistic regression model. Section 3 interpret findings of logistic regression. Section 4 compares the performance of models using ROC, precision-recall and confusion matrix. Section 5 analyses the performance of top best model. Section 6 validates best model on 2011 and 2012 cohorts with 99% of matched data. Section 7 forecasts employment outcomes of 2014 cohort with 99% matched data.

3.1 Description of Models

We have decided to fit 6 models to our data. The problem we are facing is a classification problem, with 5 earnings classes as outcome variables. We have chosen supervised machine learning models including logistic regression, CART decision tree, random forest, Naive Bayes, k-nearest neighbour, deep neural network.

3.1.1 Multinomial Logit

Logistic regression extends the ideas of linear regression [15]. In our case we attempt to predict one of 5 earnings classes. For that purpose, we use a multinomial logistic regression. It is a modification of logistic regression that generalizes to multiclass problem. We need to be aware that any high multicollinearity might make the interpretation of the coefficients of logistic regression harder. We have already encountered this problem when analysing correlation between explanatory variables.

In the multinomial logit model we assume that the log-odds of each response variable follow a linear model [7]

$$n_{ij} = log\frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i^{'} + \beta_j$$

where β_j is a vector of regression coefficients and α_j is constant [7]. Usually in the notation x includes the vector of ones representing constant, but in the case of the above equation we have written the constant α explicitly [7]. We note that the probability distribution of the response variable is multinomial instead of binomial as in case of the logistic regression [7]. The multinomial logit model can be also written in terms of original probabilities π_{ij}. Starting with the previous equation we can write the probability π_{ij} as

$$\pi_{ij} = \frac{exp\{\eta_{ij}\}}{\sum_{k=1}^{J} exp\{\eta_{ik}\}}$$

for $j = 1, ..., J$ [7]. We have decided to test this model as it has straightforward interpretation.

3.1.2 CART Decision Tree

Classifications and regression tree (CART) follow an idea like the divisive clustering [7]. Initially all learners are grouped as a single group. It forms the root of a tree. Then, that group is split into two nodes. This is typically done by setting the threshold on one of the predictors [7]. The predictors and threshold are chosen so that they separate the individuals from 5 earnings classes as much as possible. CART constructs binary trees using the feature and threshold that yield the largest information gain at each node. The information gain is computed based on the cross-entropy criterion [12].

$$Entropy(t) = -\sum_{j=1}^{n} p_j log_2 p_j$$

$$GAIN_{split} = entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i))$$

The disadvantage of this approach is that it tends to prefer splits that result in large number of partitions. This could lead to poor generalizability of our model, because the CART decision tree will essentially learn training data [7]. We usr this model as it is easy to understand and is fast to implement.

3.1.3 Random Forest

Random forest is an ensemble method of decision trees [12]. Ensemble methods in decision trees construct multiple, diverse predictive models from adapted versions of the training data. Later they combine the predictions of these models for instance by simple average voting (or weighted voting) [12]. In case of Random Forest model, we use the bagging sample. Bagging is a name for bootstrap aggregating. This means taking different random samples of the original data set. These are called bootstrap samples. We take them uniformly with replacement. We train the multiple trees on randomly sampled training data and randomly restrict the features used in each split. We validate model during training by out-of-bag validation [12]. The random forest algorithm is based on the methodology from CART decision trees.

Result: Set of Decision Trees
initialization;
for *for t=1 to T* **do**
build a bootstrap sample D_t from D by sampling $|D|$ data points with replacement;
select d features at random. The best split on these d features is used to split the node;
rain the model M_t on D_t without pruning;
return $\{M_t|1 \leq t \leq T\}$**end**

Algorithm 1: Random Forest algorithm [12]

During our research we also implement parameters tuning by grid search [12]. Parameter tuning is a process of finding the subset of a parameters of a random forest which could result in the best accuracy[12]. The higher is accuracy, the better is our model for forecasting the employment outcomes. We also consider different measures of errors. We have attempted to find two optimal parameters which is number of features (between squared and log_2

order of magnitude) and number of estimators between $[10, 500]$. We use this model as it usually performs better than a CART decision tree.

3.1.4 Naive Bayes

The Naive Bayes estimator is based on the classical Bayes formula

$$P[C|A] = \frac{P[A|C]P[C]}{P[A]}$$

. The estimator has 'naive' in its name as we assume independence between its attributes, here denoted by A. When we want to find the probability $P(A_1, A_2, ..., A_n|C)$ we may solve it using the equality $P(A_1, A_2, ..., A_n|C) = P(A_1|C_j)...P(A_n|C_j)$. This reduces the complexity of calculations, as we are no longer required to compute the conditional probabilities [7]. We use this model as it is easy to implement and has clear interpretation.

3.1.5 K-Nearest Neighbours

K-nearest neighbours (KNN) algorithm is a non-parametric method [7]. The classification is made based on the frequency of other classes in the neighbourhood of an object. Given object is classified by the plurality vote of its neighbours [7]. We need to make 2 choices before running the KNN. We need to choose a number of nearest neighbours based on which the classification decision will be made. We also should choose a distance matrix [19]. By default, sklearn library chooses the minkowski distance matrix [19] . We also make modification to use distance weights, and increase the number of nearest neighbours to 11, as we believe it will help us to improve accuracy of the classifier. It is also advised that once the number of features exceeds 10, we should reduce the dimensionality of our data [19]. We have tried to reduce the dimensionality using the PCA, so that 90% of variance is preserved. However, we did not get a better result. We got the same accuracy, but the training time has decreased. This is due to the curse of dimensionality, as the more features it is, the more complex calculation of measure it is. We use KNN as it requires just 2 arbitrary choices.

3.1.6 Deep Neural Network

An artificial neural network (ANN) is specified by the three components; architecture, activity rule, and learning rule [14]. Architecture includes variables involved in the network and their topology. Activity rule includes local rules defining how the activities of the neurons respond to each other. Learning rule specifies the way in which the parameters change with time [14]. Figure 3.1 shows the architecture of the built neural network.

In the previous chapter, we have focused on the EDA and unsupervised learning algorithms t-SNE and PCA. As we have shown, they did not perform well for clustering learners into income classes. We find that supervised learning algorithms perform better for our data set than unsupervised learning algorithms. The deep neural network performs best among all supervised learning algorithms.

Architecture

Universal approximation theorem states that "any continuous function $f : \mathbb{R}^D \to \mathbb{R}^k$ can be approximated uniformly (with respect to the Euclidean norm) on compact sets by the family of the feedforward networks with two layers, with linear activation function in the output layer and Heaviside units in the hidden layer" [14].

Furthermore, any function $f : \mathbb{R}^D \to \{0, 1\}$ of the form $f(x) =_{\cup_{l=1}^{L} C_l} (x)$ can be represented exactly by a feedforward network with three layers and Heaviside activation functions, where each C_l is a convex polytope [14].

We have used a more complicated architecture than necessary (figure 3.1), which consists of input layer, with 30 inputs, 3 hidden layers and output layer. We input the variables gender, region, ethnicity, number of days worked, sustained employment, sustained learning, sustained benefit, age, disability, learning difficulties, national level, OLASS learner, prior educational attainment, self-employment. We found that 3 hidden layers perform bests. First hidden layer includes 64 neurons, second, and third hidden layer includes 128 neurons. Output layer has 5 possible outputs, as we have 5 income classes.

We set the numbers of neurons to 128 and 64, as we found that they result in the best accuracy. The convention in machine learning, is that the number of neurons should be equal to the power of two. It is justified by the binary nature of the computations. The increase in computation speed can

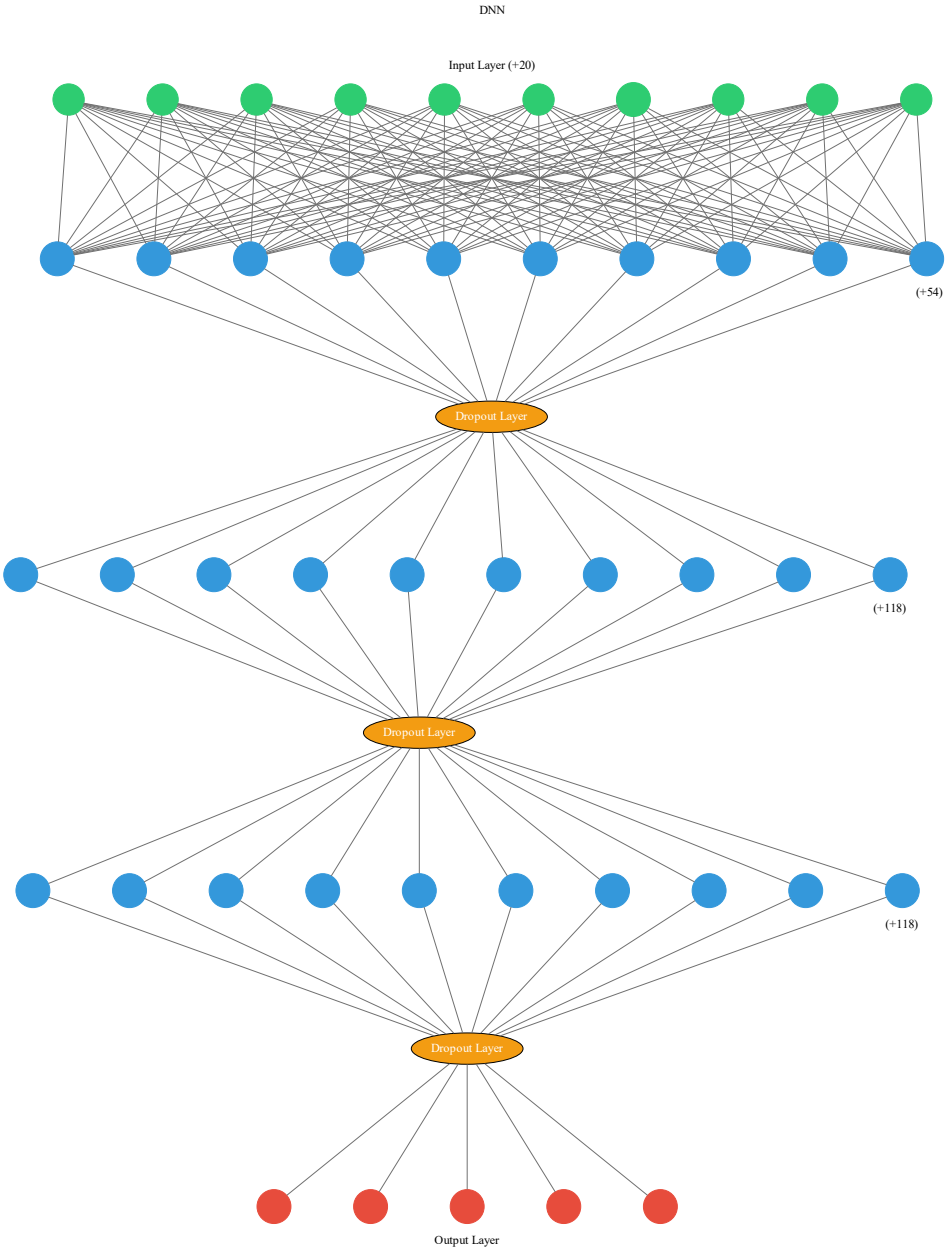


Figure 3.1: Architecture of the created nerual network

be visible when performing optimisation using GPU [13].

We have used 3 hidden layers, as by trial/error we found that 3 layers give the best accuracy. There is no fixed convention for how many layers to choose [14]. This depends on the particular data set. Usually, smaller data sets require less layers. Given the universal approximation theorem and the above claim, at most 2 hidden layers should be sufficient [14]. However, we keep 3 hidden layers, as it performs better on our data set in terms of accuracy.

Activity Rule

For the hidden layers, we have chosen Rectified Linear Units (ReLU) activation function, that is $g(a) = max\{0, a\}$. We have chosen ReLU, as its main benefit include reducing the probability of gradient vanishing [14].

Regularization is any modification of neural network that reduces its generalization error but not training error [14]. We have implemented both L2 regularization and dropout.

Within the created deep neural network, we use the ridge regression regularization, called L2 regularization. L2 adds the squared magnitude of a penalty term to the loss function. The L2 regularization element might be represented as [14].

$$\lambda \sum_{j=1}^{p} \beta_j^2$$

Then the cost function becomes [14].

$$\lambda \sum_{i=1}^{n} Y_i - sum_{j=1}^{p} X_{ij} {\beta_j}^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Compared to the L1 regularization, L2 does not shrink any of the features to 0, which allow us to have control over what we feed to the model [14]. The regularization is a way of feature selection, which can be compared to the backward elimination we used in the logistic regression [14]. This deep neural network does not give us any measure of certainty, thus we need to implement other algorithms allowing us to influence the weights of features.

Dropout is another popular regularization technique. We apply it to prevent the over-fitting. Dropout procedure consists of randomly setting a fraction rate of input units to 0 at each update during training time [18].

Within out network, we randomly drop 25% of the weights in the input layer and 50% in the inner layers. We use these values and combine it with L2 regularization as Srivistava et al.[18] mention that this approach may give the best regularization results.

3.2 Feature Selection for the Logistic Regression

This data set is challenging to model using multinomial logit algorithm due to large number of records and heterogeneity of dependent variables. We have run the multinomial logistic regression using StatsModel python library [22]. We did not use Sklearn library, as it has limited functionality [19]. It does not allow to choose the starting values, optimizers, and does not compute standard errors. During fitting the model we encountered numerous problems including singular matrix errors and convergence problems. Singular matrix is a matrix which cannot be inverted. We encountered this problem when iterating the optimizer for too many times. Convergence problem occurred when we set the maximum number of iterations above 30, which is the default option. Convergence problems occur even at 100-200 iterations with some subsets of features. Convergence issue means that values of coefficients were jumping significantly with each iteration of the optimizer. We have attempted to overcome these problems by removing highly correlated features, experimenting with different optimisers, changing the starting values, and sampling subsets of data (for instance 2000 observations). We also attempt to overcome this problem by reducing dimensionality using PCA. We decide not use this method, as the main aim of this logit model is clear interpretability of findings.

Convergence problems results in nan's in some of the standard errors. Singular matrix means that we are unable to compute standard errors. In cases when we encounter singular matrix error, we reduce number of iterations. For cases of lack of convergence, we experiment with different estimators and different starting values. We are not able to solve the convergence problem in some of the steps. We also experimented with aggregating categories in learning difficulties and disability variables. We found that even aggregated variables cause convergence problem. When we removed the variables learning difficulty and disability, sustained learning and prior-attainment we managed to get convergence. Figure 3.2 shows how gender and self-employment coefficients converge after 8 iterations.

We decide to select features based on the backward elimination. We choose this approach as StatsModel library has no function of automatic feature selection. Backward elimination is simple to implement manually. In backward elimination, we start with all the available features (eliminating those which are highly correlated, as they could possibly violate the independence assumption). We then compute their coefficients and standard errors. We look at the features which are not significant, and among these, we eliminate least significant feature. We then re-estimate the coefficient and standard errors of remaining features. We repeat elimination and re-estimation until all features are significant. We need to keep age variable, as otherwise we encounter convergence problem (table 3.1). Under normal circumstances we would eliminate it, as it is not significant.

As a result of the backward elimination, we have obtained the following features, coefficients and standard errors in table 3.1. The final algorithm converges, it has nonrobust standard errors, its pseudo R-squared is 0.09. We have just reported coefficients and standard errors for the lerners located in the 4^{th} income class, as we are mostly interested in features which increasse probability of ending up in this high income class, that is earners above £20,480. The highest earning bin includes earners above £986,820, which is why we decided to comment on coefficients of those who have earnings not that far away from mean. We comment on findings relative to the reference income category 1, that is earnings between £0 and £6910.

3.3 Interpretation of Coefficients of the Logistic Regression

The log odds shown in table 3.1 have the reference levels of variables which have the largest number of individuals, that is male, North West, White ethnicity, missing sustained benefit, missing self-employment. All log-odds are interpreted relative to the reference income class 1.

We may note that being female decreases the probability of being in income class 4 by 56%, relative to being male. Missing coefficient of missing gender variable is not significantly different to 0. That suggests that missing gender is MCAR.

Being in a region East of England increases the probability by 27% and being in region South East increases the probability by 23% region relative to North West region. Surprisingly, being in London region only increases

Table 3.1: Multinomial Logit Model, 4^{th} income class. Reference income class 1. (*) We need to keep this variable due to convergence problem.

variable	**coefficient**	**standard error**	**exp(coefficient)**
missing gender	-0.3430	9.3e+14	0.7096
female	-0.8314	0.021	0.4354
East of England	0.2385	nan	1.2693
South East	0.2055	nan	1.2281
East Midlands	0.1402	nan	1.1505
South West	0.1251	0.030	1.1333
Yorkshire and The Humber	0.1011	nan	1.1064
West Midlands	0.0683	nan	1.0707
London	0.0584	0.032	1.0601
North East	0.0087	0.027	1.0087
missing region	-0.3427	9.99e+14	0.7099
unclear	-0.0612	0.054	0.9406
Chinese	-0.1280	0.268	0.8799
missing ethnicity	-0.3966	0.038	0.6726
any other ethnic group	-0.4279	0.109	0.6519
Mixed	-0.4542	0.041	0.6349
Asian	-0.4894	0.036	0.6130
Black	-0.7263	0.038	0.4836
not self employed	-0.0674	nan	0.9348
self employed	-2.4246	0.040	0.08851
sustained benefit	-1.2539	0.014	0.2854
days worked	0.0273	nan	1.0277
age*	0.00003208	0.02	1.0000

the probability by 6%. We would expect London to have more high-paid positions relative to other regions due to higher maintenance costs. Being in West Midlands, Yorkshire and The Humber, South West, East Midlands increases the probability by about 10%. Missing values are not significant. Overall, we may note that the region might have an influence on being high-earning individual, and the changes in probabilities are large. We would expect that this is the case mostly due to regional differences in salaries and differences in industries. London region would apprear to be an exception.

Ethnicity appear to have a big influence on earnings. Learners of White ethnicity have the highest probability of ending up in a high erning group. Being Chinese decreases the probability of being high earner by 12% relative to White ethnicity. Those with missing ethnicity, any other ethnic group, Mixed and Asian have lower probability of being high earners by about 30%. Those of black ethnicity have lower probability of being high-earners by 51%. We need to keep in mind that these results do not mean that enthnicity has a direct impact on being high earner, but some ethnicities correlate with high-eraning jobs/industries. It is also worth to note that the sample is predominantly of white ethnicity, and other ethnicities are not well-represented. This might create less accurate estimates for those of other than white ethnicities.

Learners who are not self-employed have similar probability of being high-earners as those who have missing values in variable self-employment. On the other hand, those who are self-employed have 91% lower probability of being high-earners than those who have missing self-employment. This suggests that prior employment status have important impact on earning category, and that those who have missing employment status are most probably not self-employed, further supporting hypothesis posed in the EDA section.

Those who are on sustained benefits are 71% less likely to be high earners relative to those who have missing variable sustained benefits. This suggests that those with missing sustained benefits are probably not on sustained benefits, as they are either employed, or in further education. This again supports our hypothesis posed in the EDA section.

We may note that working one more day increases the probability of being higher earner relative to income group 1 by 3%. We don't know if that is significant, as the standard error has a nan value.

It is important to note that we have included age as one of our variables. It is not significant, but excluding it causes convergence problems.

Overall, the result of logistic regression has limited value due to con-

vergence problems and singular matrix cases. We may also note that the pseudo R^2 value is relatively small (0.09). However, we may conclude that gender, ethnicity, sustained benefits, self-employment have significant impact on being high earning learner.

The above analysis has 3 main limitations. First of all, not all standard errors have been computed. We may not be sure that some coefficients of regions are statistically significant. We also did not include robust standard errors. As we deal with survey data, they tend to be more heterogenous than experimental data and using robust standard errors (robust to heteroscedasticity) is recommended. This option is not available within used library, as well as R. STATA program has an option to specify robust standard errors, but we have no access to this software. Third, we didn't manage to get convergence with interactions. We would recommend including interaction between gender and region, ethnicity, employment status and benefit. We would also recommend including interaction between level of studies and mentioned variables, as [2] Learners and Apprentices Survey 2018 publication mentioned that there are differences in trends of earnings of learners undertaking different apprenticeship levels. We did not manage to obtain convergence when experimenting with these interactions.

3.4 Comparison of Performance of Models

There are numerous metrics used to evaluate statistical models. Most popular include accuracy, loss, confusion matrix, precision-recall curves, ROC curves, F-measures which are functions of precision-recall, macro and micro-averaging, Hamming loss, 0/1 loss [12]. We have decided to describe and compare accuracy (this is the most common metrics and we can use is as out data are well balanced between groups), ROC curves as it shows us the trade off between true positives and false positives, and precision-recall as it describes accuracy of algorithms taking into consideration both false positives and false negatives. We also analyse confusion matrix data in form of histograms for best performing classifier to identify the distribution of errors.

3.4.1 ROC Curve

Receiver Operating Characteristics (ROC) curve has been developed in 1950s for signal detection theory to analyse noisy signals. It has been used to characterise the trade off between positive hints and false alarms. Nowadays it

is widely used in the machine learning research for model evaluation [12]. We compate true positives (the data points which has been correctly classified) with false positives (the data points within given income class that have been incorrectly assigned as positives). False positive relates to type 1 error. The closer is ROC curve to the upper left corner the better is our algorithm. Another way to evaluate performance is to measure an area under ROC (AUROC) curve. The higher AUROC, the better is our algorithm.

We have plotted the ROC curves for all 6 classifier we use and for the 5 income classes considered. Figure 3.3 and Figure 3.4 shows ROC curves for classification of 1^{st} and 5^{th} income class. ROC curves for income class 2, 3, 4 are located in appendices. We may note that according to ROC all classifiers perform better than guess at random. The random guess would represent a straight dotted blue line from lower left corner to upper right corner. Across all classifiers, constructed deep neural network (figure 3.1), performs better than any other algorithm. Logistic regression is the second best algorithm.

We may note that our algorithms performs best in case of first and fifth income class, as the ROC curve is closest to the upper-left corner, compared to other classes. It means that both very low earners and very high earners have distinct features which are detectable by the algorithms. We can see that middle income classess (class 2, 3, 4) are hard to detect. All classifiers are much closer to the random guess.

3.4.2 Precision-Recall Curve

Precision-Recall curve shows the relationship between ratio of items that are correctly classified as positive and ratio of correct items that are classified of positive.

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

fp (false positive) is a type 1 error, and fn (false negative) is a type 2 error. tp (true positive) are those who are correcly classified to given income class, tn (true negative) are those who are not correctly classified to a given income class.

Figure 3.5 and figure 3.6 shows precision-recall curves for income class 1 and 5 respectively. Precision-recall curves for income classes 2, 3, 4 are located in appendicies. We may note that all of our algorithms are making

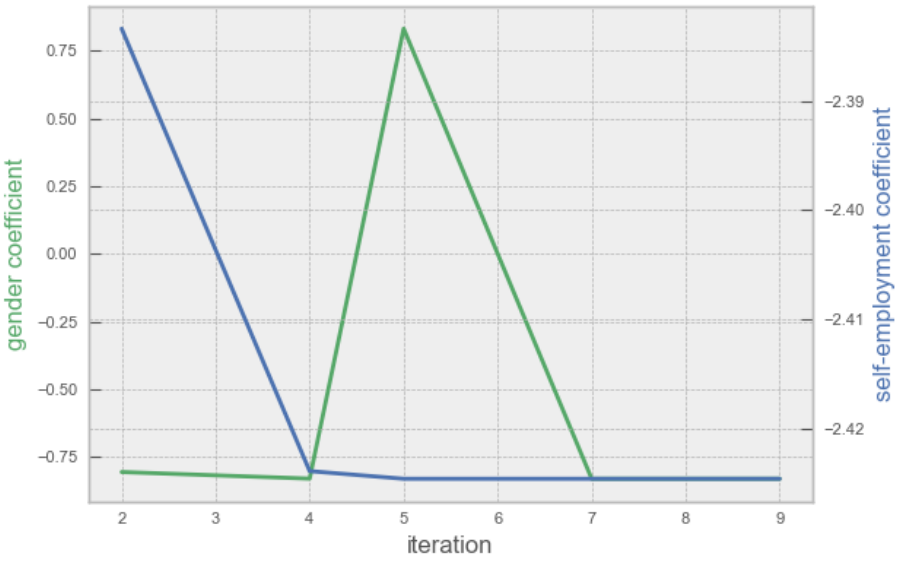


Figure 3.2: Convergence of coefficients of female gender and self-employment

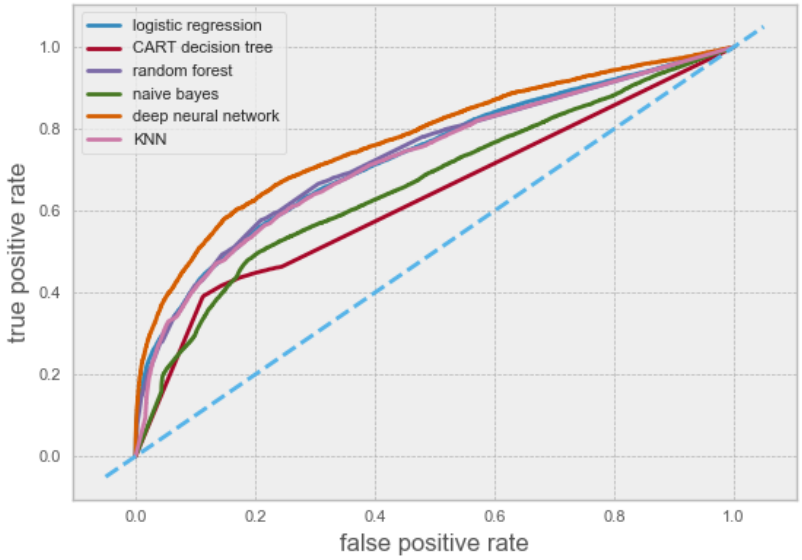


Figure 3.3: ROC curve for income class 1

significant amount of errors for all income classes. DNN seem to outperform not-greatly all classifiers within income class (1, 2, 3, 4). DNN performs significantly better in terms of precision-recall for the 5^{th} income class.

3.5 Analysis of Performance of the Deep Neural Network

3.5.1 Training DNN

We have already outlined the architecture and activity rules of the DNN in section 3.1.6. Here, we focus on learning rule and evaluation of its performance. There are no rules for selecting optimization algorithms. The most standard optimizer for the neural networks is adam. Adam optimizer is an extension of a stochastic gradient descent algorithm used to update weights in an iterative manner based on training data. In contrast to stochastic gradient descent, adam is an adaptive algorithm. It changes learning rate based on lower order moments. It is said that it outperforms other optimization algorithms [26].

During research of our data set, however, we found that RMSprop optimizer outperforms adam. It converges faster and results in higher accuracy, lower error. RMSprop is second most popular optimization algorithm. RMSprop is also based on the adaptive learning rates.

With trial/error approach, we have set batch size to 2000. When we set lower batch sizes such as 700, algorithms converge faster to lower accuracies, wheras large batch size of 4000 are slower and do not result in improved accuracy. We found size between 2000-3000 to be optimal.

Figure 3.7 shows that both the accuracy and loss converges very fast. Later, the rate of learning is much slower. We have iterated through data set 800 times. Figure 3.7 shows first 200 iterations for clarity. We may note that while the loss function converges relatively well after 25 epochs, the accuracy increases significantly up to 50 epochs and has much smaller increases after 50 epochs. After 800 epochs, we have managed to get accuracy on test set of 35.23%. That is significantly higher than any other algorithm.

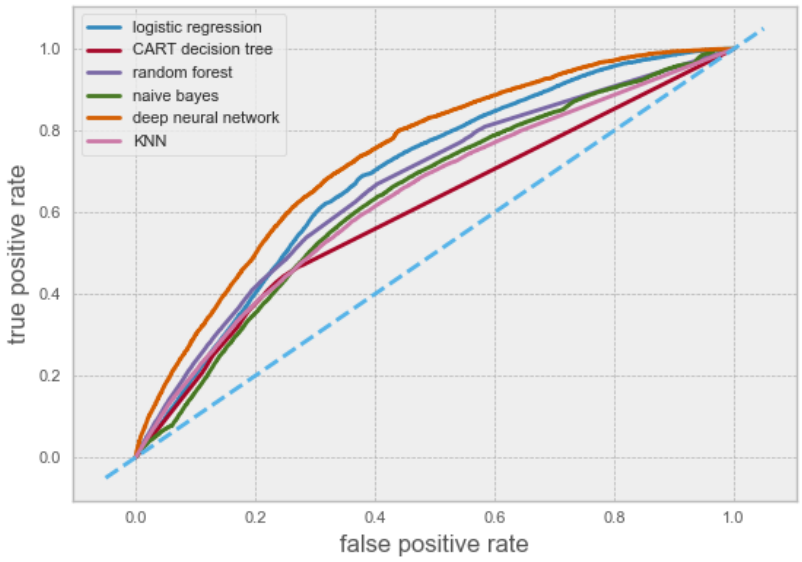


Figure 3.4: ROC curve for income class 5

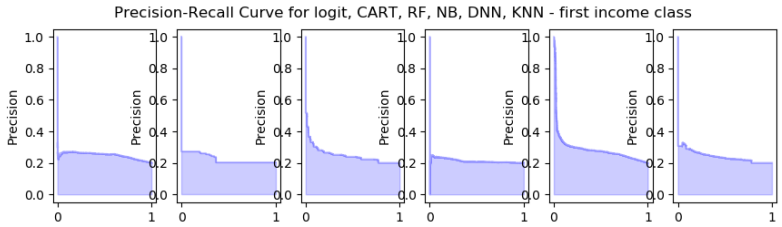


Figure 3.5: Precision-recall curve for income class 1

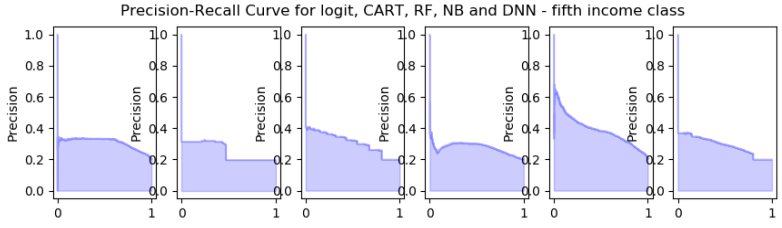


Figure 3.6: Precision-recall curve for income class 5

3.5.2 Errors DNN

We may note that the DNN performs very well in detecting low-earning and high earning individuals, (figure 3.8). DNN detects more than 55% of group 1 learners and more than 50% of group 5 learners. We check the distribution of errors within all classifiers and all of them have relatively high accuracy for high income and low income earners. Within low earners, we may notice that the DNN tends to classify more than 15% of learners to income class 3 and 5 (figure 3.8). The distribution of errors is relatively better for error detection of highest income class (figure 3.9). There is a singificant number (more than 15%) of learners misclassified to class 3 (see appendices). All other classes have less than 10% of misclassified cases. We could interpret this phenomenon as that there are some features of individuals which makes them very likely to become high earner, but some middle-income individuals have these features. On the other hand there are features which are correlated with low-earning class, but substantial amount of learners from low-income class have features of both middle-income and high-income earners.

DNN makes a lot of mistakes for detection average earning individuals. It seems to be almost impossible to detect fourth and second income classes (see appendices). We may note that DNN tends to miclassify these as a neighbouring classes. Furthermore, class two is misclassified frequently as a class 5. This shows that non of the above algorithms is well-suited for class 2 and class 4 detection. DNN classifies class 3 with accuracy above 35%. This result seem good, however when we look at the distribution of errors, we may note that there is almost equal number of learners classified to income class 5. About 15% of learners is classified to class 1 showing that DNN make serious mistakes for income class 3 detection as well.

The above findings show that it is possible to detect high earning and low earning individuals with high accuracy. DNN makes less serious errors in case of these two classes. We found it hard, in case of all algorithms, to detect 3 middle income classes.

We also run the DNN with discretisized income into 3 classes. We find that the errors in income were still unevenly distributed. The accuracy increase to 53% (as we have 3 income classes, random guess in that case would be 33%), but the distribution of errors does not improve. Based on that we conclude that this is a limitation of the considered features. To some extent they are helpful for predicting employment outcomes, but they are not deterministic. This is escpecially true for middle earning individuals.

Within our research, we treat income as an outcome variable. We have also decided to check how initial income impacts the future income. Once we treat 2013 earnings as a dependent variable, we have managed to get an accuracy of 66% using DNN. It violates the assumption of independence in case of logit model and Naive Bayes model, but it is worth to note that initial income is a strong indicator of future earnings. If the sole purpose of this exercise would be to create accurate machine learning pipeline, we would recommend including more prior employment related variables, such as earnings history.

3.5.3 Cost Sensitive Learning

The distribution of errors of the DNN suggests that the algorithm makes some serious misclassifications such as misclassifying second income class as fifth income class (25% error rate), or many misclassifications of third income class as either first, or second income class.

We decide to implement the cost sensitive learning which can improve the distribution of errors. Ideally, we want to see decreasing error rate, the further from given class. We have specified fixed cost of misclassification using class weight argument in the model.fit function of sklearn library [19]. Class weights are most often used in case of badly balanced data set, for instance for fraud detection [23]. We decide to use this approach, as we assume that misclassifying to neighbour income classes is less serious than misclassifying to distant income classes.

We first specifiy a fixed weight of misclassification, 10 for class 1 and class 5 and weight 1 for other classes. We assume that misclassification of learner as high-earner, or low-earner is more serious that any other misclassification. Figure 3.10 shows that this cost-sensitive learning gives better results than equal fixed weights. Accuracy of the model does not change, but the loss decreases by 1% point. The distribution of errors remains like the one for DNN without cost sensitive learning.

We have also specified weights specific for each case of misclassification. Misclassification to income group 1 and income group 5 have always fixed weight of 8, and the misclassification of groups nearby given income group increases. For instance, for income group 3, we have specified weights to be 2 for misclassification to income groups 2 and 3. We have tested the increases in weight both by the factor of $2X$ and 2^X. We do not find a significant improvement in accuracy and loss relative to the fixed weights of

8 for income groups 1,5 and 1 for others.

The above finding shows that the limited predictive power of specified features is the main obstacle of correctly classifying income groups 2, 3, and 4. Even when we attempt to change misclassification weights, we are not able to improve the distribution of error for these classes.

3.6 Model Validation

For the purpose of model validation, we create data sets including characteristics of earners and their employment outcomes for the 2011 and 2012 cohort. We create raw and processed data sets for each year. For the purpose of this exercise, we decide to create a full data set including missing values in LEO LEARNERS, LEO SELF-ASSESMENT, LEO COHORTS. We have already highlighted this problem in the data set preparation section. Previously, we have reduced our data set to rows which did not have any systematic missing values. Otherwise our machine learning models performed significantly worse. When we validate the trained neural network, we take into consideration fully matched data set, covering around 500,000 observations for each year. We believe it is more proper approach, as in reality policymakers will try to estimate the employment outcomes of both mached and not fully matched learners.

We find that the accuracy for the 2011 academic year apprenticeship starters is 23.19%. The validation accuracy increases to 25.37% for 2012/2013 academic year cohort. We also evaluate the accuracy on 2013 cohort population (for 99% match), and we get accuracy of 28.16%, compared to valuation accuracy on subset of training data set which is 35%. The accuracy of 2011 and 2012 cohorts is significantly worse than expected, given validation accuracy of above 35% on the 2013 cohort. We suspect two sources of worse performance. One could be heterogeneity between cohorts. 2011 cohort might have significantly different characteristics impacting employment outcomes than 2013 cohort. Then the generalizability error cohorts other than 2013 is quite large. We may note that this effect is visible, as accuracy increases by about 2% points from 2011 to 2012 cohorts. Another possible source of worse performance can come from the fact that we derive data sets including large numbers of missing values. As we have mentioned, DNN does not perform well on the data set with large number of missing values for 2013 cohort. We suspect that this is a major factor resulting in poor performance of DNN on the population sample with filled missing values. We could see that the val-

idation accuracy on whole data set of 2013 cohort (with significant number of missing values) reduces from 35% to 28%. We need to keep in mind that it is incorrect to evaluate the accuracy on the data we trained, but in case of 2013 cohort, we want to check to what extent missing values (missing 33% of matches in LEO tables) impact the prediction accuracy.

3.7 Forecast

Given limitations of the above model, we try to forecast outcomes of the 2014 learners cohort. This section intends to simply highlight prediction, and does not go into detail about individuals. We have created a 2014 apprenticeship starters data set including all of the considered variables. As the employment outcomes for the tax year 2017 has not yet been published in the departmental data set, we intend to predict the distribution of income. We have made a forecast using DNN, as it is the best performing algorithm. Figure 3.11 shows the predicted amount of 2014 apprenticeship starters within each income group 3 years after starting an apprenticeship. We may note that the DNN predicts that about 45% of 2014 apprenticeship starters will be low earners, and about 30% of the apprenticeship starters will be between lower-middle and upper-middle income class. Given that previously we have discretised outcome variable into 5 equal groups, we would expect the distribution of forecasts to be close to uniform distribution.

We need to keep in mind the limitations of this forecast. In the previous section we have highlighted that the two sources of bias are likely to affect the above distribution. The accuracy of the above forecast is limited, as there is a substantial number of missing values because of LEO tables. Furthermore, we assume that the characteristics of individuals impact employment outcomes in the same way as 2013 cohort. We can be more sure that the income 1 and income 5 group forecasts are accurate, as we have seen that the accuracy of these income groups is about 50% and errors are less serious when we analysed the performance of the DNN, whereas the middle groups predictions tend to be much less accurate.

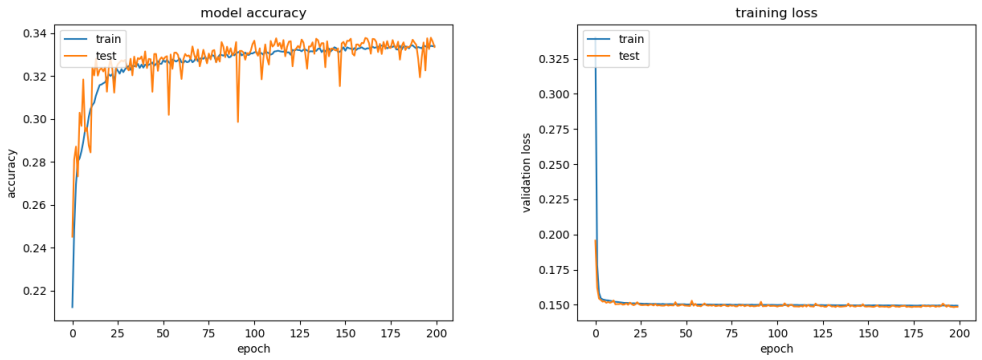


Figure 3.7: Accuracy and loss of the deep neural network

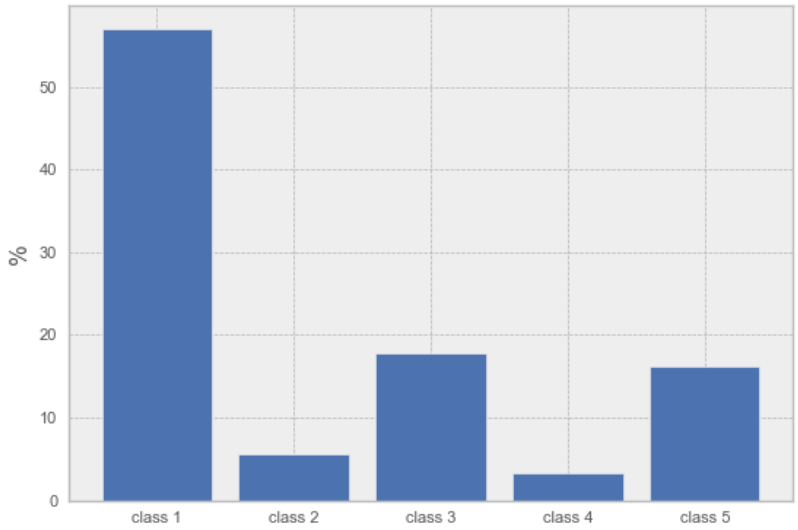


Figure 3.8: Assignment of income classess by the DNN for income group 1

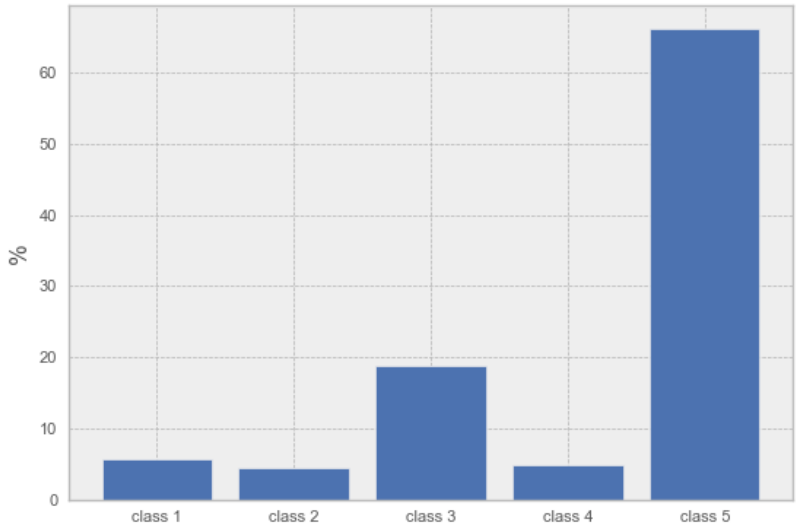


Figure 3.9: Assignment of income classes by the DNN for income group 5

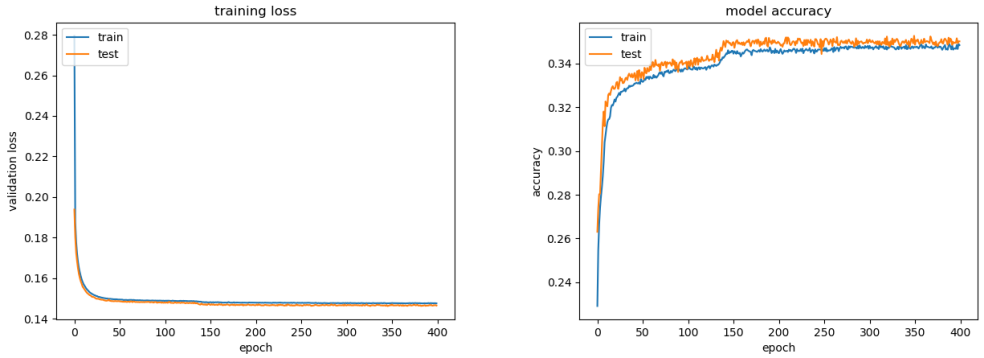


Figure 3.10: Accuracy and loss of the deep neural network with cost sensitive learning

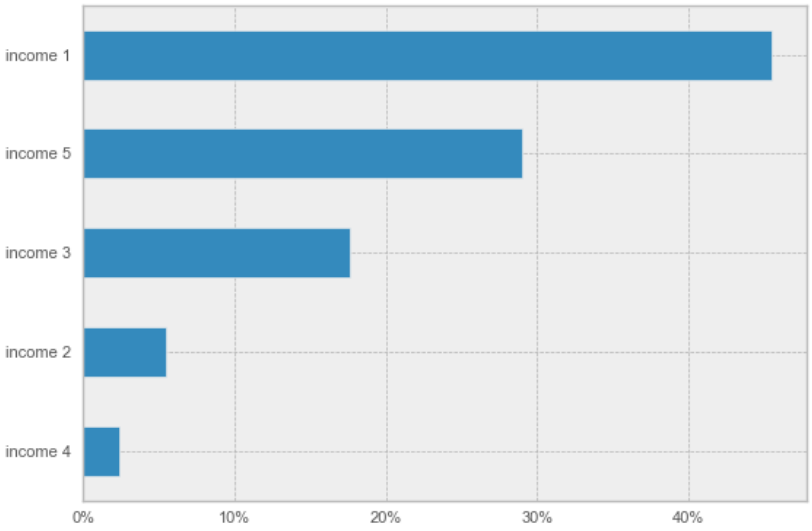


Figure 3.11: Forecast of the distribution of income groups 3 years after starting an apprenticeship. 2014 apprenticeship starters cohort

Chapter 4

INTERPRETATION AND CONCLUSION

4.1 Interpretation of Findings

The aim of this thesis is to analyse the factors which might have either direct or indirect impact on the labour market outcomes. We have defined labour market outcome as earnings three years after starting an apprenticeship. We have discretised the earnings variable into 5 income groups and define those with positive labour market outcomes as those within income groups 4 and 5.

In the section 1.1 we have posed 5 research questions. Firstly, we have asked about the characteristics of the 2013 cohort of learners. We extensively described it in Chapter 2 during univariate and multivariate EDA sections. We have found that on average 2013 apprenticeship starters are older than 30 years old, low earning individuals. They are employed for most of the days during given year. Gender and regions are well balanced. Vast majority of starters are of white ethnicity and are enrolled on low level apprenticeships. Relative to the English population, they have low earnings both at the start and in the end of apprenticeships. Overall, we can characterise this group as heterogenous in age, with low prior attainment and low initial income. This characteristic suggests that these starters should have on average worse employment outcomes relative to the general population.

Second research question relates to the relevance of features for determining employment outcomes. We also want to detect among all relevant

features, which are the most important. We have attempted to answer this question by producing heat-map showing correlations between earnings 3 year starting the apprenticeship and background characteristics. We have found that positive employment outcomes are strongly correlated with number of days worked in a tax year, sustained employment, sustained benefit, self-employment. Other variables seem to have little, or no correlation. This suggests that the major factor impacting positive earnings are characteristics directly related to prior employment. Those who are working more days in a tax year, are in sustained employment and are not taking benefits have higher earnings than others. It is interesting to note that these variables are defined for a tax year starting in April 2013. According to Leaners and Apprentices Survey 2018 [2] a substantial number of learners come straight from full-time learning. That means that their earnings are initially around 0 and they are not in sustained employment, do not work many days in a tax year. Furthermore, The Learners and Apprentices Survey 2018 publication mentions that also substantial amount of apprenticeship starters enrol in apprenticeships for their current employer. That would suggest that background characteristics of learners have no strong impact on employment outcomes, but that work experience and prior employment have more significant impact on future earnings. Furthermore, we have run unsupervised machine learning algorithm PCA to distinguish important features. Although the results of PCA are not very informative, they have shown that sustained employment, days worked in a tax year are important variables, followed by prior attainment and OLASS learner status. This supports our hypothesis that prior work experience is more important factor than any other feature. Finally, we can detect important feature deciding about positive employment outcomes when interpreting the coefficients of logistic regression. In chapter 3 we have focused on the interpretation of coefficients of group 4 income learners. These coefficients are very similar for income 5 group learners. We may note that the logistic regression emphasizes the importance of gender, ethnicity region, sustained benefit, and days worked in a tax year. This again appears to support the hypotheses that prior employment is important. However, we found that both gender and ethnicity has much bigger impact on probability of ending-up in a high-income group. Overall, we may summarise the above findings with conclusion that employment related features are the most important features, followed by gender, ethnicity, prior attainment and OLASS learner status.

The last two research questions relate to forecasting the income groups and selecting best performing models. We have found that both the deep neural network and logistic regression perform best among all income groups.

We have managed to obtain accuracy of more than 35% on the test set. This is significantly better than if we were to guess the income group randomly. For most of the purposes, when we are interested only in forecast, we would recommend using DNN. In case when more interpretability is needed, we would recommend using logistic regression. Despite its lower accuracy, logit model is easier to understand and implement. This benefit requires a trade off in accuracy.

4.2 Discussion about Models

In this section, we intend to discuss the performance and suitability of tested machine learning models. We have tested 6 models in total on derived data set. These models include multinomial logit model, CART decision tree, random forest, k-nearest neighbour, Naive Bayes, and deep neural network.

We have found that DNN followed by multinomial logit perform bests for all income classes. We have decided to use the multinomial logit model as it restricts the range of a dependent variable to interval [0, 1]. Its output gives us measure of uncertainty related to the given classification. Multinomial logistic regression can also detect the variables with no predictive power (as in linear regression we have standard errors), can combine both discrete and continuous predictors, together with their non-linear combinations. This is particularly useful in our case as the explanatory variables are both continuous (age, days) and discrete. Finally, multinomial logit model does not impose any distributional assumptions on predictors. In chapter 2 we have analysed the distribution of explanatory variables and we could clearly see that they were not normally distributed. On the other hand, the drawback of logistic regression can be overfitting. That means that it may not perform well for forecasts, as the generalizability error will be relatively high compared to other classifiers. Once we test deep neural network on 2011 and 2012 cohorts, we may note that generalizability is still a problem despite implemented regularization techniques.

CART decision tree and its extension random forest have both the advantage of fast implementation. CART decision tree is easy to interpret and visualise and both of these models captures complex non-linear and non-monotonic parameters, similarly to logistic regression. We have decided to implement CART decision tree as it might be easier to visualise and explain for policy makers. On the other hand, CART decision tree requires many parameters. We have attempted to print it and the decision tree was so

large that the variables were not visible. Decision tree is also sensitive to the choices of variables, meaning that it may not be the best for forecasting. Furthermore, it is sub-optimal to detect monotonic patterns and hard to detect interactions. As we assume that there is monotonic relationship between number of days worked and employment outcome, we suspect that it may be one of the reasons why it does not perform that well. We have decided to use random forest, as despite less interpretability and more complex structure it has better performance in case of our data set and is also more resilient to noise. This is particularly visible in survey data, such as ours. In our multinomial logit model, we should use robust standard errors to account for heteroscedasticity. Unfortunately, the SM library does not offer this option. In this aspect random forest may perform better than multinomial logit. We have also noticed that random forest model is significantly more time consuming compared to CART decision tree and multinomial logit models. This issue arises both during training (especially when we tune parameters) and during forecasting

We have also decided to test Naive Bayes classifier because it is easy to understand, evaluate, and Naive Bayes assumption works generally well in practise. In case of our data set it has one of the worst predictions. This could be because it assumes that variables are normally distributed. Also having too many features can result in spurious results. Naive Bayes also does not allow for interactions between explanatory variables. We could notice that the Naive Bayes was one of the worst performing classifiers. We suspect it is due to the strong assumption of normality and independence. EDA reviled that earnings and days worked in a tax year are not normally distributed. High correlations visible on the heat maps shows that the naive Bayes assumption can be severely violated. Furthermore, the data set has more than 10 features, which seems to be problematic in case of Naive Bayes. Summing up, this algorithm does not give any improvement both in accuracy and in understanding on relationship between background characteristics of learners and employment outcomes.

K-nearest neighbour is an intuitive and easy to implement algorithm. It requires relatively few users' choices. We only needed to specify the amount of nearest neighbour (11) and distance matrix (minkowski). It performs well relative to other classifiers, as it can capture non-linear and non-monotonic patterns. In contrast to Naive Bayes it can detect the interactions. Similarly, to CART decision tree it is not optimal to detect monotonic patterns. It also requires a good choice of k. When we were experimenting with different k, we have found that in case of our data set it is particularly sensitive to this choice, as ranging k from 3 to 11 ranged accuracy from 0.28 to 0.32. We did

not find it to be particularly sensitive to the choice of measure matrix in case of our data.

Our best performing algorithm is a deep neural network (DNN). The advantage of DNN is that compared to other algorithms it gives the best results both in terms of accuracy and error rates. Relative to other algorithms we have a lot of control over its architecture, activity rule and learning rule. This allow us to create the neural network which performs particularly well for our problem. This is different relative to other used algorithms where we have less control. DNN can also handle multi-dimensional data sets and detects more important features. On the other hand, it is only useful for forecast. We cannot interpret its findings and its complex architecture makes it impossible to get insight into how the decisions are made. Due to many arbitrary choices it may be the case that not enough experimentation with its settings may result in poor results. During research, we have obtained accuracy ranging from 0.21 to 0.35. The variability of results of other algorithms was much lower.

4.3 Limitations and Recommendation for Future Research

Our research is limited because we have analysed and trained our models based on the 66% matched 2013 cohort of apprenticeship starters. As we have mentioned it may be possible to construct the 99% match ratio when deriving missing LEO tables from other low level data sets. We have not attempted to do that due to lack of access to required tables and time constrained. We have checked that the summary statistics of available tables are similar on both 66% match and 99% match ratio. We suggest that future research may focus on deriving the missing tables and then reconstructing above project to see if the results change significantly.

Furthermore, during our research we have focused solely on the background characteristics of individuals. Future researchers may want to incorporate the initial income and further prior-employment statistics, as well as sectors as one of the variables. Although it violates the independence assumptions of some our models, such as logistic regression, it can be used with deep neural network to accurately predict the employment outcome. We have done some experimenting with this variable, and we have found that it is possible to train DNN and obtain overall accuracy of 70% for income class

prediction. It can be useful if the only aim of an algorithm is the forecast accuracy.

We also suggest focussing further on developing cost sensitive learning algorithms. We have found that the DNN makes some serious mistakes when classifying income classes 2, 3 and 4. It would be more beneficial to train an algorithm which is less accurate but makes less serious mistakes. This could be based on the distance measure, and the higher the distance of misclassification, the higher the higher the weight of this mistake.

We could also compare the employment outcomes of those who have finished the apprenticeship on a given level to those who have finished equivalent level of education and those who have stopped at the earlier level of the apprenticeship. This could give us better overview of how effective apprenticeships are compared to alternative routes.

Finally, the multinomial logit model we have implemented is quite basic. We would advise future research to focus on deeper investigation of convergence problem, obtaining convergence when including gender and level of learning aim interactions, as well as including robust standard errors.

Bibliography

[1] Assets.publishing.service.gov.uki (2019), *Outcome Based Success Measures.*, [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment data/file/748305/FE learners outcome based success measures.pdf [Accessed 2 Jul. 2019].

[2] A. Thornton, C. Witso, R. Menys, S. Hingley, P. Alexander (2018), *Learners and Apprentices Survey 2018*, Social Science in Government.

[3] Department for Education, internal document, *LEO Learners Dataset Apr18 Metadata 20181120*, [Accessed 2 Jul. 2019].

[4] Department for Education, internal document, *LILR documentation and definitions*, [Accessed 2 Jul. 2019].

[5] Schools LEO team (2018), *The LEO learners dataset: a guide for users*, Department for Education.

[6] J. G. Ibrahim, H. Chu, M. H. Chen, S. Hingley, P. Alexander (2012), *Missing data in clinical studies: issues and methods*, J Clin Oncol, 30(26):3297-3303.

[7] S. Tavakoli (2018), *ST323/ST412 Multivariate Statistics (2018-2019)*, University of Warwick.

[8] I. T. Jolliffe (2018), Component Analysis, Second Edition, Springer.

[9] Y. Choi, J. Taylor, R. Tibshirani (2017), Selecting the number of principal components: Estimation of the true rank of a noisy matrix, Ann. Statists, Volume 45, Number 6 (2017), 2590-2617.

[10] L. Maaten, G. Hinton (2008), *Visualizing Data using t-SNE*, Journal of Machine Learning Research 1 (2008) 1-48.

[11] Office for National Statistics (30.09.2013) National Minimum Wage to rise from 1 October 2013, Available at: https://www.gov.uk/government/news/national-minimum-wage-to-rise-from-1-october-2013 (Accessed: 2.09.2019).

[12] Y. He (2018), Data Mining CS909, University of Warwick.

[13] A. Graves (2012), Supervised Sequence Labelling with Recurrent Neural Networks, Springer.

[14] P. Jenkins (2018), Advanced Topics in Data Science, Artificial Neural Networks, University of Warwick.

[15] T. Hastie, R. Tibshirani, J. Friedman (2017), The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Springer.

[16] Alan. Beaulieu (2009), Learning SQL, O'REILLY.

[17] M. Kukar, I. Kononenko (1998), Cost-Sensitive Learning with Neural Networks, ECAI 1998.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever (), Dropout: A Simple Way to Prevent Neural Networks from Overfitting , Journal of Machine Learning Research 15(1):1929-1958.

[19] scikit-learn developers (2019), scikit-learn user guide, Release 0.22.dev0.

[20] Office for National Statistics (2019), Consumer price inflation, UK Statistical bulletins, Office for National Statistics. [Accessed 5 Sep. 2019].

[21] D. Clark (2019), Population of regions in England in 2018, [Last Edited Jun 27, 2019].

[22] statistical models library developers (2019), Statsmodel user guide, v0.10.1 release.

[23] R. Moser (2019), Fraud detection with cost-sensitive machine learning,[online] Available at: https:// towardsdatascience.com/fraud-detection-with-cost-sensitive-machine-learning-24b8760d35d9. [Accessed at Jun 27, 2019].

[24] M. Ma'amari (2018), Deep Neural Networks for Regression Problems ,[online] Available at: https://towardsdatascience.com/ deep-neural-networks-for-regression-problems-81321897ca33. [Accessed at Jun 27, 2019].

[25] D. P. Kingma, J. Lei Ba (), Adam: A Method For Stochastic Optimization , ICLR 2015.

[26] A. S. Walia (2017), Types of Optimization Algorithms used in Neural Networks and Ways to Optimize Gradient Descent,[online] Available at: https://towardsdatascience.com/types-of-optimization -algorithms-used-in-neural-networks-and-ways-to-optimize-gradient-95ae5d39529f. [Accessed at Jun 27, 2019].

[27] HMRC (2013), Personal income by tax year,[online] Available at: https://www.gov.uk/government/collections/personal-income-by-tax-year. [Accessed 5 Sep. 2019].

[28] W. Yung, J. Karkimaa, M. Scannapieco, G. Barcarolli, D. Zardetto, J. Sanchez, B. Braaksma, B. Buelens, J. Burger (2018), The use of machine learning in official statistics.

[29] Xiao-Li Meng (2019), From Statistically Significant... to Significantly Statistical IMS Bulletin Volume 48 Issue 4.

Chapter 5

Authorship Notes

This analysis has been produced whilst on secondment to the Institute for Apprenticeships and Technical Education, an executive non-departmental public body sponsored by the Department for Education.

The views expressed in this report are the authors' and do not necessarily reflect those of the Institute for Apprenticeships and Technical Education or the Department for Education.

Appendices

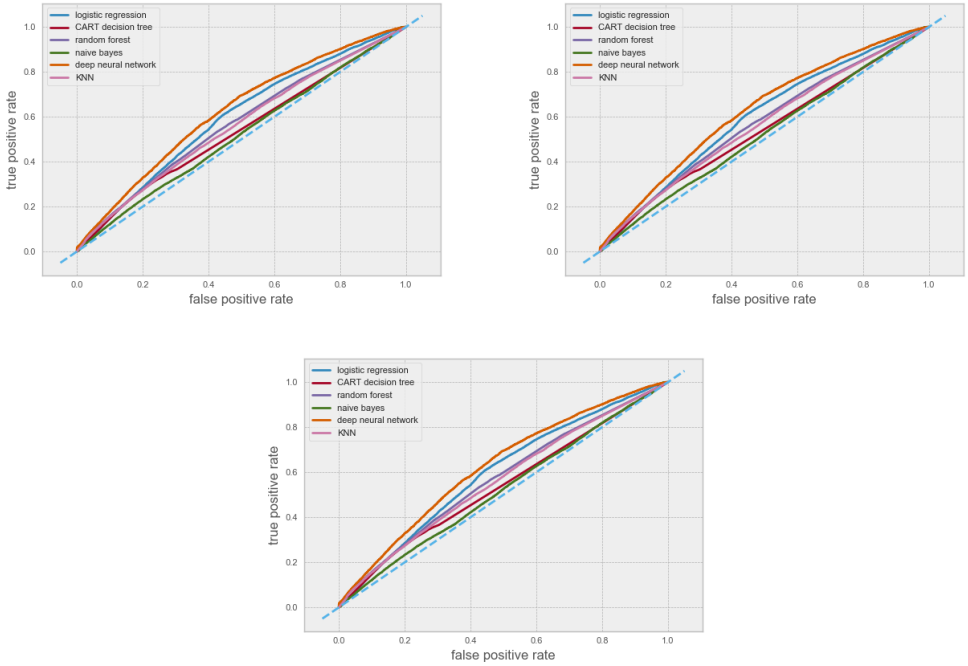


Figure 1: Class 2, 3, and 4 Income Groups

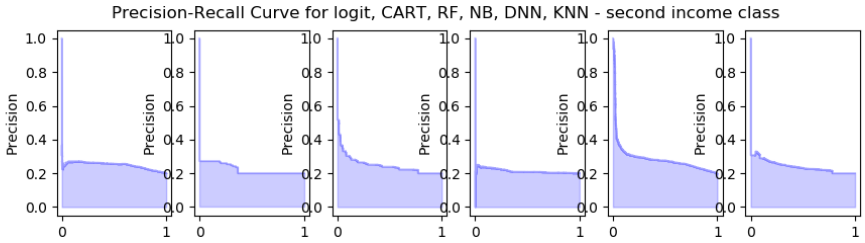


Figure 2: Comparison of performance of models for the second income class - Precision - Recall

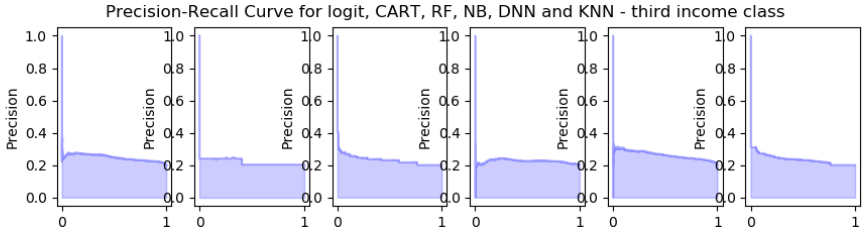


Figure 3: Comparison of performance of models for the third income class - Precision - Recall

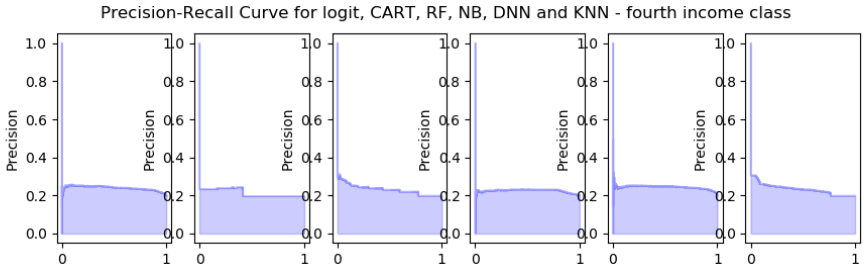


Figure 4: Comparison of performance of models for the fourth income class - Precision - Recall

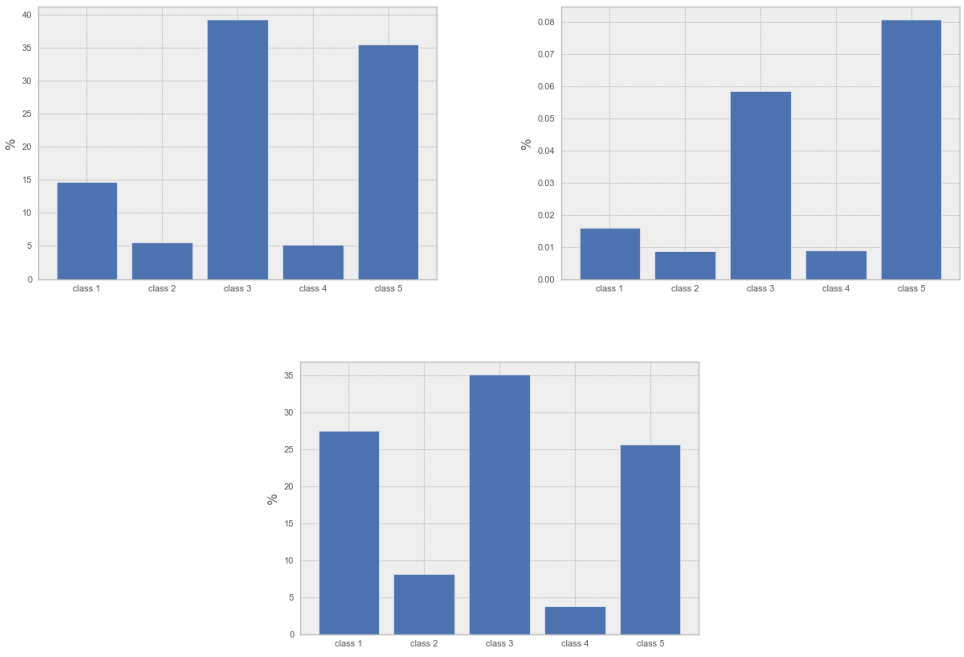


Figure 5: Distribution of classifications of income groups 2, 3, and 4