CCT College Dublin
MSc in Data Analytics
Capstone Project
Evaluating current causal inference methods in data analytics, and demonstrating novel results in projects where they are added

Milo Moran sbs23081

Supervisor: James Garza

23rd February 2024

This report was written entirely by the author, except where stated otherwise. The source of any material not created by the author has been clearly referenced. The work described in this report was conducted by the author, except where stated otherwise.

Abstract

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

1HE/OHE = One Hot Encoding
ANOVA = Analysis of Variance
ATE = Average Treatment Effect
CIM = Causal Inference Method
CV = Cross Validation
DAG = Directed Acyclic Graph
DTC = Decision Tree Classifier
KNN = K Nearest Neighbours
ML = Machine Learning
StDev = Standard Deviation

SVM = Support Vector Machine
XGBoost = Extreme Gradient Boosting

1 Intro

Talk about feature analysis, Statistical testing talk about Dimensionality talk about classifiers Hyperparameter tuning Feature importance scores then explain causal vs predictive. then e

2 Literature Review

While correlations, predictions, and forecasts are all common tasks in data analytics, causal analysis is significantly less common.

2.1 Motivations

We often talk about correlation vs causation in theory but while implementing Data Science solutions towards solving business problems not much influence is given to validating causation amongst independent and dependent features. Global AI (2022)

Much of the novelty of this project comes from it's attempts to demonstrate the power of causal inference methods and their viability for pointing towards answers to causal questions in all data based quantitative research. The importance of an uptake of these methods is argued in Hernan, Hsu and Healy (Hernán et al. 2019). The article outlines the historical context whereby causal inference from observational questions has been suppressed by mainstream statistics, and claims that appropriate integration of causal inference and counterfactual prediction into data analytics

claim that there is a historic opportunity to redefine data analysis to naturally accommodate a science - wide framework for causal inference from observational data.

Gelman & Imbens (2013) considers the difference between forward causal questions seeking to learn effects of causes, and reverse causal inference which looks for the causes of effects -NOT VERY RELEVANT

Brady, Henry and Iris Hui. 2006. "Is it Worth Going the Extra Mile to Improve Causal Inference?"." Paper presented at the 23rd Annual Summer Meeting of the Society of Political Methodology

2.2 Frameworks

There exists two primary frameworks or models for causal inference. There is the Structural Causal Method (SCM) associated with Judea Pearl (Pearl 2009), and there is the Rubin Causal Model(RCM) associated with Donald Rubin, also known as the potential outcome framework(Imbens & Rubin 2015).

Ibeling & Icard

Aliprantis (2015) claims that while the two methods are often viewed as analogues, there is a distinction. This working paper demonstrated this

2.3 Methods

Several large scale surveys can be found that on CIM's are part of 'A Survey on Causal Inference'(Yao et al. 2021) is a review of the potential outcome framework of causal inference. In Table 3, the paper lists 4 tool boxes where various methods can be found and used, some in Python -DoWhy(Sharma & Kiciman 2020), CausalML(Chen et al. 2020) and EconML(Microsoft Research 2019), and one in R - causalToolbox(Künzel et al. 2019). Table 4 lists individual methods as well as providing sources for each. These tables will be used as a valuable reference for this research going forward Structural causal model, 102,105,107 Sharma & Kiciman (2020) describes the DoWhy library, and explains all the features it has to enable for end to end causal inference.

2.3.1 Assumption independent methods

What is assumed by each method? Does using methods independent of assumptions add value in terms of breadth of applicability or ease of use etc? Yao et al. (2021) divides methods based on whether they are independent of the assumptions of the potential outcome framework or not. In section 4, it describes the workarounds that researchers have used to infer causality in scenarios where the assumptions or parts thereof do not hold. This is important for this research project, as when dealing real world projects, it is unlikely that all three assumptions of the framework will be met, and being able to use causal inference methods in these situations is invaluable. When discussing future directions in section 7, the paper notes the need for more research into cases in which the assumptions of the causal model can be relaxed, as practical settings frequently do not follow all the assumptions and current methods may not always be appropriate. This outlines a potential hurdle for this research, as it may not be possible to accurately apply the causal model in some practical projects.

2.3.2 Model selection

Scott (2019) compares Heckit models, Propensity Score Matching, and Instrumental Variable models, demonstrating their performances in various scenarios through simulation. Conditions are varied in terms of selection observables, selection unobservables, and outcome unobservables, and the

Stuart (2010) summarises previous recommendations for selecting matching methods in section 6.1.3 as well as proposing guidelines for their use in section 6.2. In the former section the paper references balance as being a key factor for best method selection, and suggests 3 possible criteria for choosing a method:

> (1) the method that yields the smallest standardized difference of means across the largest number of covariates, (2) the method that minimizes the standardized difference of means of a few particularly prognostic covariates, and (3) the method that results in the fewest number of "large" standardized differences of means (greater than 0.25).

The paper references a method which automates the process (Diamond & Sekhon 2013) using a genetic matching algorithm to reduce bias and error of estimated causal effects. In section 6.2, it summarises 6 points of guidance for practice, how to decide what covariates to include, how to choosse distance measures, examining the implications of the estimand and choosing the appropriate one, implementing the appropriate matching method, and evaluating the covariate balance.

2.4 Evaluation

There are three general approaches to validation of causal methods: Firstly the face validity test, whereby the result is compared with the intuition of an expert in the domain. Second the placebo test, where the nature of the data allows for the separation into placebo and treated groups based on either time or selection within the sample. Thirdly, synthetic data is used for testing methods based on a known created treatment effect. Schuler et al.(2017) proposes Synth Validation, whereby simulations are used to test and select causal inference methods for use in given scenarios, allowing the most appropriate method to be used for that scenario. The paper notes that previous research has been inconsistent and has failed to find any one-size-fits-all methods that tests methods against hand crafted benchmark data

Parikh et al.(2022)

what large scale evaluations have been done and what was found?

(Dorie et al. 2019) An thorough survey of causal inference for time series analysis has already been conducted(Moraffah et al. 2021). It covers the various questions and methods as well as for evaluating the results of those methods. Do I need to go into massive detail on just how optimised, or just how much detail I will need to justify use of given methods?

2.5 Current advice and guidelines for real world applications and so on

Stuart(2010) provides a review of the state of matching methods across various disciplines, and attempts to bring together much of the research, as well as making suggestions as to where the literature should be headed. It

Kerzner (2022) is a guide to causal inference that thoroughly covers much of the overarching steps in implementing causal inference, and would make for a strong introduction to the subject for a researcher that already has a grasp of basic statistics. The article makes many practical recommendations for how to deal with common problems along with referencing best practice. While it is not well structured, this guide is an example of an excellent starting point to encourage the use of causal inference to the right audience.

2.5.1 Project selection?

What areas would be promising in terms of domain what types of data would be appropriate in terms of maximal variation of the typical case? where do I go to find them? kaggle is a bad idea right? (assuming I can source the sets.)

2.6 Lit Review conclusions

3

In choosing criteria for the selection of methods, the exact aims of the selection, and the proposed use cases, must be very clearly defined and justified. The criteria of "ease of use" includes ease of access, and so a massively important criteria is that the method needs to be open sourced. Three large "tool-boxes" in Python dedicated specifically to providing access to these kinds of methods are DoWhy, CausalML, and EconML. These libraries contain a large number of methods along with tutorials to assist users in applying causal inference methods as part of their research. This makes the methods available from these libraries an excellent starting point for evaluation in line with the goal of finding easy to use CIM's

Can I fully justify using only methods from these sources?

Other evaluation criteria involve

4 Methodology

4.1 Research Objectives

4.2 Primary Research

Experimental research has been chosen as the primary research methodology for this project. Experiments will be performed to test the hypothesis “Does performing ML prediction on a datasets and producing feature importance metrics provide more insight into causal relationships with the target variable than statistical testing alone?”

To test this hypothesis, ML algorithms will be used to make predictions and produce feature importance metrics. The most important features as per these metrics will then be used to infer causal relationships with the target variable. Statistical testing will also be used to determine the most correlated features with the target variable, and the causal inference will also be performed on these results. The ATE measurements from both domains will be compared to determine if there is significant difference between the two methods. ATE’s from features selected via purely statistical analysis will serve as the control group, in order to evaluate the performance of the ATE’s from the Feature importance methods.

For this experiment, all four factors will be satisfied. There will be concomitant variation between the independent variable describing applied/ not applied and the dependent variable describing novel results found/not found, there will also be temporal sequence of the states prior to and after the methods being applied. All the research provided by the literature review will comprise the theoretical support. Finally, there will be complete control of the system where nothing will be introduced that could influence the dependent variable aside from the treatment.

4.3 The Data

The dataset used for this project describes survey results from a survey of students in higher education at Near East University in Cyprus and was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019. It includes personal details, family details, education habits, and performance outcomes. The dataset was sourced from the UC Irvine Machine learning repository. The dataset was first introduced in the paper by Yilmaz & Sekeroglu (2019), in which the authors experimented with the data and used Radial Basis Function Neural Network to achieve accuracy of 70-88%. the dataset is licensed under the Creative Commons Attribution 4.0 International, which allows for the sharing and adaptation of this dataset for any purpose once it is appropriately credited. The dataset contains a total of 33 features and 145 observations. All the data are numerically encoded as values ranging from 0-9. There is no missing data. The first feature of the dataset is an identifier for the observations, each one relating to a unique StudentID. This feature has no relevance for analysis and is dropped immediately. The answers to the questions in fig. 1 correspond to features 1-30 of the dataset. The next feature, COURSE ID, assigns a number to the course the student was undertaking for each of the 9 courses covered. The last feature is GRADE, and represents the grade band that the students final grade fell into.

Personal questions	**Family questions**	**Educational questions**
Age	Mothers'Education	Weekly study hours
Sex	Fathers'Education	Reading (non-scientific)
High School Type	Number of Brother/Sister	Reading (scientific)
Scholarship Type	Parents'Relationship	Attendance to Seminar/Conference
Additional Job	Mothers'Job	Effect of Projects and Activities
Sports/Arts	Fathers'Job	Attendance to Lectures
Relationship		Taking notes
Salary		Writing/Listening
Transportation		Effect of in-class Discussions
Accommodation		Effect of Flip Classroom
		GPA of Last semester
		Expected CGPA at graduation

Table 1: Summary of questionnaire, adapted from Table 1 in Yilmaz & Sekeroglu (2019)

The data is based on personal data for these students, but attempts have been made to anonymise the data. According to the Full Guidance Note on Anonymisation and Pseudonymisation from the Data Protection Commission (2022), personal data that has been irreversibly anonymised ceases to be 'personal data' or require compliance to Data Protection law. To determine that this data is suitably anonymised, identifiability of the subjects must be ruled out. The data set contains no information about unique identifiers relating to the students such as names, phone numbers, student numbers, birthyears/days, or addresses. There is no information that allows for any student to be singled out. Numerical data such as age, salary, and grades are binned which makes identification more difficult A combination of the Course ID factor along with the information given related about the courses in the original paper may be enough to identify the course of some of the students, but the courses feature here doesn't correspond directly to those mentioned in the original paper, and this could be difficult. Some responses to the personal questions regarding family status may increase risk of linkage between values to identify students, but there is no other public data available related to the family status of students with which to corroborate this information, and reidentification in this manner would not be likely. As a result of all these factors, it appears that it is not reasonably likely for the identification of any of the subjects and given the nature of the topic, identification attempts are also unlikely. Therefore the data can be considered to be fully anonymised, and will not need to be treated as personal data for the purposes of this report.

4.4 Approach

The ultimate objective of the research project is to use causal inference to estimate the causal effect of different interventions, or features, on the final grade of the students based on the survey responses. The causal effect will inferred through the calculation of the ATE of various appropriate features based on feature analysis obtained through several methods. predictive ML models, and compare . . . To this end, the features that best predict the target variable are identified through several methods. Features themselves are investigated via several statistical methods and heuristics to enable better understanding. A subset of features is selected based on results and are then used for the training of machine learning models. The results of preliminary machine learning models are used for a range of comparisons. An XGBClassifier model is tuned for 2 different subsets of the data using 2 different hyperparameter searching methods respectively. The feature importance is

4.4.1 GRADE column and re-binning

In order to better understand and more appropriately treat the target variable, some further interepretation is needed. Turkish grade conversions were used to accompany the categories with percentage ranges between 0 and 100 as visible in Table 1 below. To determine the the spread of the variables a histogram of the variable was included. It shows quite an uneven distribution, with a DD grade being most frequent, and FF and CB grades being particularly underrepresented relative to the other categories

Category	Grade	Percentage Range
7	AA	90.00 - 100.00
6	BA	85.00 - 89.00
5	BB	80.00 - 84.00
4	CB	75.00 - 79.00
3	CC	70.00 - 74.00
2	DC	60.00 - 69.00
1	DD	50.00 - 59.00
0	FF	0.00 - 49.00

Table 2: Corresponding Categories, Grades and Percentage Ranges

There is a question of linearity involved with the treatment of this variable. The encoded 0-7 scale is not representative of the size of the differences between different grades on the 0-100 scale at all. However this isn't wholly inappropriate; the distribution amongst the different grades isn't linear, and the difficulty of increasing from one grade to the next is presumably non linear as well. For

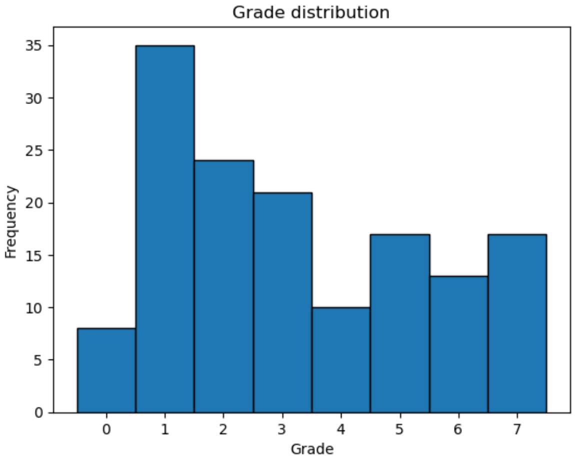


Figure 1: Histogram of the GRADE student feature

design of further research, it might be preferable to use deciles or quartiles of student of grades as bands to predict between.

One major issue with the full 8 category Grade column is highlighted by the initial decision tree algorithm runs. The amount of total points for classification assuming a 25% testing split of 36/145 observations across 8 categories is very low, with on average just over 4 occurences per category for testing, and often significantly less. Such low amounts make any predictions extremely sensitive to noise, and hinders predicting to any significant degree of accuracy. This issues was further compounded by the uneven distribution of the categories. One choice to help counteract this effect was to increase testing split sizes for the rest of the analysis up to .3, providing slightly more values in each category for testing.

This issue is the primary motivator for the re-binning of the target variable from 4 bins into 8. Another motivator is that this balances the distribution of the dataset significantly vs the 8 category distribution, as can be seen in fig.2. The new percentage bands for the 4 bins can be seen in table 3. As a goal of this analysis is to ascertain whether certain interventions increase or decrease grades, re-binning the values like this improves the interpretability, as while data and granularity are reduced, the overall ability to predict higher or lower is maintained.

4.4.2 Dimensionality

With such a high ratio of features to observations in the dataset, over 1:5 in the initial set, prior to the addition of even more via the One Hot Encoding of non-ordinal categorical features, there is a definite concern that the Curse of Dimensionality will effect model results. As a result, efforts are made in several stages to justify the removal of features, on the assumption that doing so will

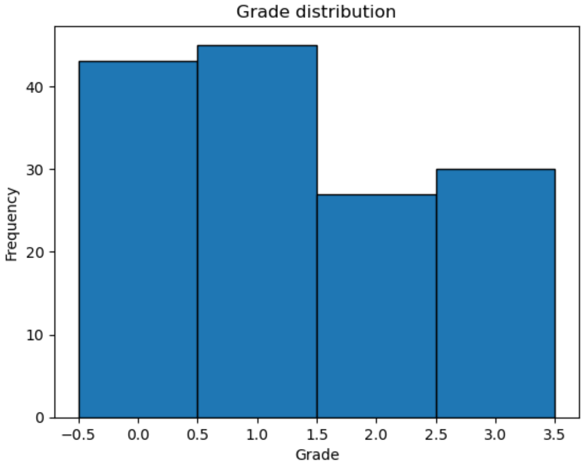


Figure 2: Histogram of the new grade feature created through rebinning

Category	**Percentage Range**
3	85.00 - 100.00
2	75.00 - 84.00
1	60.00 - 74.00
0	0.00 - 59.00

Table 3: Categories and Percentage Ranges after re-binning

reduce the negative effects of dimensionality. Further, to evaluate the trade off between the datasets that contain more and less features, models created using both subsets referred to as small, and large, are compared against each other throughout the analysis.

Models that are largely based on distance metrics, like KNN classifiers, are particularly susceptible to the curse of dimensionality, as given enough dimensions,. SVM models are also susceptible through (J. S Marron, Michael J Todd Jeongyoun Ahn (2007) Distance-Weighted Discrimination, Journal of the American Statistical Association, 102:480, 1267-1271, DOI: 10.1198/016214507000001120)

4.4.3 Feature management

As feature importance is intended to be used going forward, it is important that feature management not include any kind of irreversible feature transformation or reduction methods such as Principle Component Analysis(PCA). While the use of such methods would reduce dimensionality and collinearity by combining heavily correlated features, they would be a large hindrance to the interpretability of the individual features. Various qualities of the features were initially investigated to motivate treatment going forward. One important step was statistical testing for correlation between features. As the target GRADE feature is in the form of numeric categoricals representing binning of a continuous feature, it was decided to consider the feature as both numerical and categorical for the purposes of statistical testing. ANOVA tests, where the target is taken to be continuous, and Chi-Square tests, where the target was taken to be categorical, were both performed for each feature. The tests generated P values and tested the hypothesis that each feature was correlated with the target dependent variable to a 95% confidence level. The tests were accompanied with box plots and group bar plots respectively for the purposes of visualisation. As the independent variables were all treated as categorical by the tests, further encoding of the values was at this point not necessary. The results and interpretation of these results led to the creation of two subsets of the data. One removing many of the features, referred to as the small dataset, and the other removing significantly less features and referred to the larger dataset.

Certain columns were non ordinal categorical, and would need to be one hot encoded for the purposes of applying ML models. As 1HE greatly increases the number of features in a dataset, this was taken as an opportunity to limit increasing the dimensionality of the dataset by condensing categories together where appropriate. Categories were merged for three of the non ordinal categorical features, 9,10 and 15 (representing transportation methods, accommodation types, and mothers occupation respectively, combining things like ”Bike” and ”Other” into ”Bike and Other”, ensuring continued interpretability) on the basis that for some of the categories, there were not enough occurrences (¡3) for a model to appropriately learn a pattern for use in prediction going forward. Consolidating categories like this for features not already marked for 1HE may have proven beneficial, but it was only performed here in the case of the double utility in limiting dimensionality. Two of the binary categories were changed to yes(1)/no(0) categories to improve interpretation by models and readers. Finally, the non ordinal categorical features were One Hot Encoded to enable independent treatment by the prediction models.

4.4.4 Predictions

Initially, a Decision Tree classifier(DTC) was used to set benchmarks for further models. Models were run with both size datasets, and compared against the full 8 category grade feature as well as the re-binned 4 category grade feature. Heatmaps and cross validated accuracy scores were used in combination to evaluate model performance visually and quantitatively. Feature importance scores were produced. As these models were primarily investigative, and as there are relatively few parameters involved with DTC models, hyperparameter tuning was not performed.

Next, a SVM classifier model was used, selected as it could potentially produce evidence of the effect of dimensionality on the predictions. It was performed for both sized datasets, again using cross validated accuracy along with heatmaps to evaluate results. To investigate repeated patterns in the heatmaps, a heatmap for an alternative split of the data was also taken. While coefficient scores similar to feature importance scores do exist for SVM classifiers, they are only applicable for linear kernel SVMs, as other kernels transform the data before creating the model, as described here(BartozKP 2014). Hypereparameter tuning was performed based on recommendations for C and Gamma from (geeksforgeeks 2023) but with the inclusion of several extra kernel options, as the optimal feature space for a model many features was expected to be unpredictable. As there was only 100 possible values with 5 cross validating fits, the GridSearchCV algorithm was used, as it would excecute in an appropriately short amount of time(20sec) and is easily controllable in terms of its values.

Finally, an XGBClassifier method was used. Distributed (Deep) Machine Learning Community (2024) has compiled a list of many of the occasions where XGBoost models have placed first or second in machine learning challenges. Such a widely successful model is a promising choice in this situation and from some other knowledge of

As XGBClassifier models typically are tuned by changing many parameters using large grids or random search spaces, but these are too time consuming and computationally expensive for the scope of this project, so standard grid or random search algorithm methods were ruled out. Instead two different approaches combining several methods were used for hyperparameter tuning. Initially, on the smaller dataset, the method of learning curve adjustment outlined by Brownlee (2021) was used to adjust several important parameters. Based on manual manipulation and inspection of the learning curves, an apparent minimum between 4 hyperparameters was found. These were then selected as the basis for a gridsearchCV, with the remainder of the parameter grid selected based on suggestions from several tutorials. (Navas & Liaw 2022)(Toth 2024) The gridsearch included over 5000 fits, and was time intensive, but significantly less so than it would have had the learning curves not been investigated beforehand. For the larger set of the data, in an attempt to address some of the issues with the previous approach, a different method was selected: Bayesian hyperparameter optimisation. One advantage of Bayesian optimisation is that it can be explicitly scaled by running for more and more iterations, depending on time available, and it can approach closer and closer to global minima.

The Bayesian optimisation algorithm initially chooses random values from within the search space, but then proceeds to use one of two mechanisms to choose the next set of parameters by evaluating the previous ones. It uses exploitation, to select the points with the highest uncertainty, or exploration, to select a point from the region with the current best results. This informed hyperparameter selection allows for more efficient tuning, and typically finds better hyperparameters in less time compared to other tuning methods. From the tuned XGBClassifier models, feature importance plots were produced using an inbuilt method, and these were based on an f score which measures the frequency of the feature being split in the classification tree.

4.4.5 Causal inference

To implement Causal inference, the DoWhy library was used, as it provides a comprehensive framework for Causal inference via it's 4 steps of ”model”, ”identify”, ”estimate”, and ”refute”. One of the first steps of causal inference is the creation of a causal graph. A causal graph was created based on assumptions about interactions between the most important features and can be seen in full in fig 3. For simplicity, this graph was not created with the same rigour as would be appropriate for a true investigation into what causes Student Grades. Doing so would require acquiring much more domain knowledge to more accurately model the causal interactions between variables, and is outside the scope of this project. Several features were assumed to be independent of others (Gender, Age, Mothers Education). It was assumed that other variables caused the two unobserved variables Student Interest and Available Time, which in turn led to Weekly Study, Scientific Reading, and Taking Notes. It was assumed that all factors were partially causal to the students grade either directly or, in the case of Partner, indirectly. All the assumptions in the graph can be viewed in Appendix A: ”Causal Graph Assumptions”

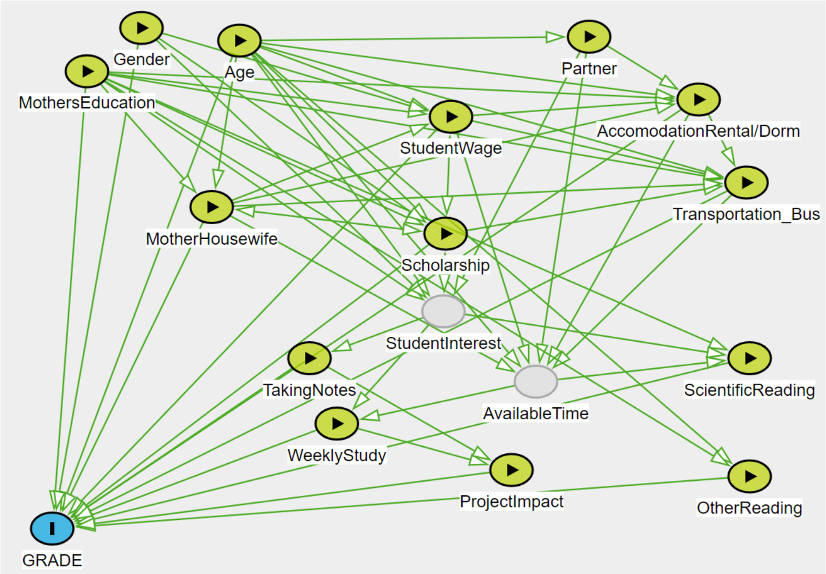


Figure 3: Directed Acyclic Graph for the Data

Using the graph, causal relationships were modeled, identified and estimated between the target and features highlighted through the use of feature importance methods. As no candidate instrumental variables were included in the causal graph, and there is no support in the dowhy library for using propensity score methods with nonbinary treatments, the linear regression method was the only one appropriate for estimation for the majority of the features. For the ”Partner” feature however, the propensity score methods can be used.

5 Findings

5.1 Feature Exploration Findings

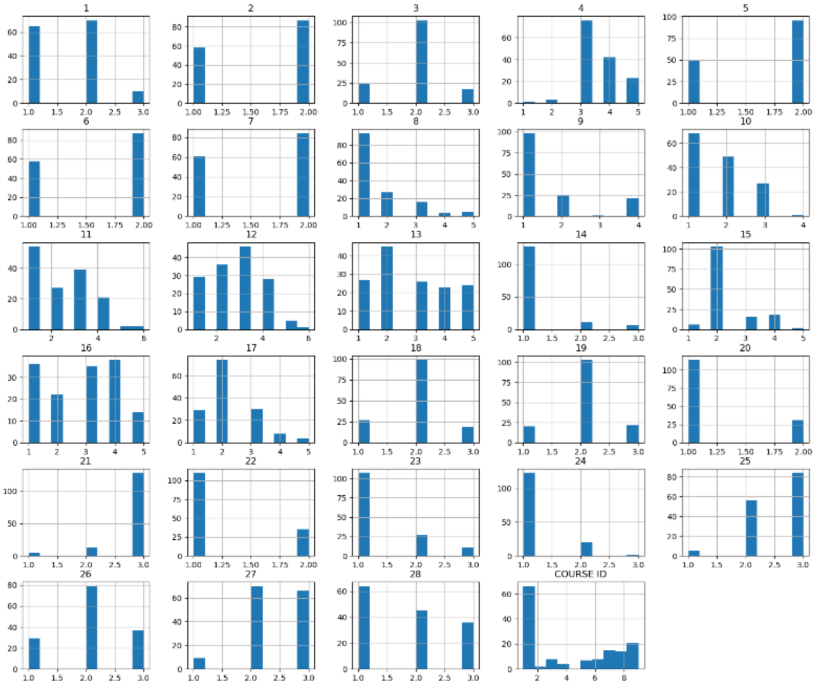
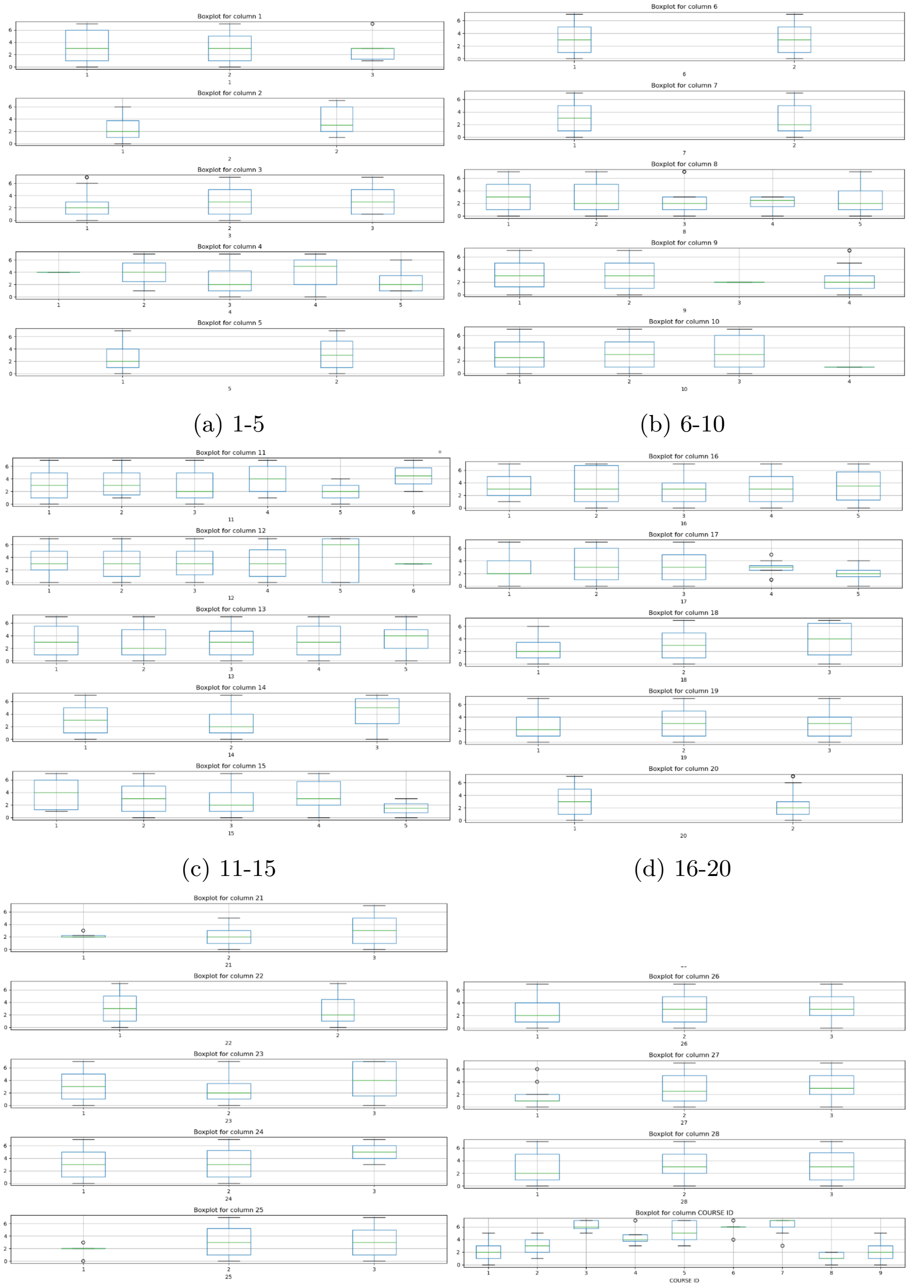


Figure 4: Frequency Distributions for all Features

Looking at the histograms of the data in figure 4, it's clear that a lot of features are unevenly distributed. From the ML point of view, difficulty can arise when lack of total observations of a category means that any rules learned by the model are more likely to be based on outlier observations and not generalise well to unseen data. As a result of this, some features are noted as highly unbalanced, and interpretation of models must account for potential bias of the model against the minority classes. For several of the features that undergo 1HE, minority categories have been combined, which significantly reduces total class imbalance, but this treatment was not applied to all important variables. Of the variables eventually selected for ML, 18 (Non-scientific reading), 19 (Scientific reading), and 21 (project impact) especially should all be cautiously interpreted.

From the ANOVA tests, only 3 features were accepted as being correlated with the target variable to any statistical significance: 2(Sex), 21(Project Impact), and COURSE ID. From the chi-square test, 3 more: 1(Age), 4(Scholarship), 11(Mothers Education) were found to be significant on top of those from the ANOVA test. 2 visualisations were produced, grouped bar plots, treating the target as categorical, and boxplots (figure5) treating the target as continuous. The grouped plots are not particularly valuable in comparison to the boxplots as they don't allow for any overall effect of the categories on the grade on the whole to be observed. The boxplots show that a vast majority of the categories are not very well correlated with the grade, and are distributed somewhat evenly between the categories, which fits with the results from the testing. The plots however do isolate some features, and more importantly some categories, as being better correlated with the target. Course ID especially has a very uneven plots, though it cannot be appropriately interpreted, and has some plots correlate with high grades whereas some skew towards lower ones. "High" and "Low" here have been defined differently, as the distribution of the grade feature itself is skewed, and both have been selected to capture a smallest categories with more than 1 or 2 different associated grades, as those are assumed to be too under-observed to accurately interpret. For example in feature 1, the third category, (1)Age:26+, visibly correlates with lower grades. Negative correlations, defined as having the plot max ≤ 4 can also be seen for (8)Total salary: USD 271-340 and USD 341-410, (11)Mothers education:MSc, (15)Mothers employment:Self-employed, (17)Weekly Study hours: 11-20 hours and More than 20 hours, (21)Project impact:Negative, and (27)Discussion improves interest:Never. Positive correlations, here defined as median ≥ 5, can be seen for (4)Scholarship:75% (11)Mothers education:PhD, (12)Fathers education:MSc, and (24)Preparation for mid term exams:Never. Some of these relationships are those that would be expected intuitively, like never taking notes and project work having a negative impacts on the success of a student, or parents being better educated, but some effects are quite counterintuitive and bring the interpretations into question: Do Mothers with Masters degrees reduce student grades? Do students who never prepared for their midterms succeed while those who studied the most did worse? While these effects may not be significant when combined with analysis of the entire dataset as a whole, another possible reason for these correlations is underobserved categories with outlier values influencing the plots. Of these categories with visible effects, Age), Total Salary(For the 341-410 category), Mothers Education(for both categories), Fathers education, Weekly Study Hours(for both categories), Project Impact , and Preparation for midterm exams, are all particularly underobserved. As a result, it can be said that the majority of correlated categories detected in the box plots are underobserved with observations. This isn't entirely unexpected, as rarer categories associated with edge of distribution events are likely to be related to edge of distribution grades, but the possibility that these values are caused by outliers that will not create generalisable rules remains. Evaluation of interactions with the target grade individually are inherently flawed as they don't account for combined effects. ML classifiers on the other hand are well equipped to deal with this problem.

As described in section 4.2.2, one intention of the statistical testing was to justify the removal of features from the data, thus reducing dimensionality. While this analysis was intended to only remove the least correlated features to contribute to predictions and causal inference, only 6 features passed the tests at this p value. Removing all other features from the prediction would be



(a) 1-5

(b) 6-10

(c) 11-15

(d) 16-20

(e) 21-25

(f) 26+

Figure 5: Boxplots for features 1-27 and Course ID

removing a lot of information from the dataset, and while correlation is low for the rejected variables, the trade off of reducing dimensionality at the expense of ability for models to learn interactions from subtle interactions between features is not worthwhile. Selection of the threshold value alpha is an important step in statistical testing, especially because it is an arbitrary value. Changing the value of α after already seeing the results of the tests in order to change whether the null hypothesis is accepted or rejected and the statistical significance of the data being tested is known as p-hacking, and is a major issue in research whereby researchers can manipulate results such that they are more in line with what the researcher expected to see. Changing the alpha value in the case of this analysis is possible, and would lead to more features being found to correlate with the target, but it could be misleading and potentially unethical to do so. Increasing the risk of false positives for the purposes of choosing features in this scenario however is not a major issue, as these statistical tests are being used for supportive purposes as opposed to conclusive ones. The solution for this is the creation of two datasets: the first one with the strict, original statistical test threshold of p¡ 0.05 referred to as the small dataset, and one with a much broader inclusion threshold of p ¡ 0.2. The results of predictions using the two datasets are compared to evaluate the differences and test the value of reduced dimensionality vs the inclusion of low correlation features. The larger dataset based around the 0.2 α includes the following features: 1(Age), 2(Sex), 4(Scholarship), 7(Partner), 8(Salary), 9(Transportation), 10(Accommodation), 11(Mothers' education), 15(Mothers' occupation), 17(Weekly Study), 18(Non-scientific Reading), 19(Scientific Reading), 21(Project Impact), 23(Midterm Study), and 25(Note Taking).

5.2 Prediction Results

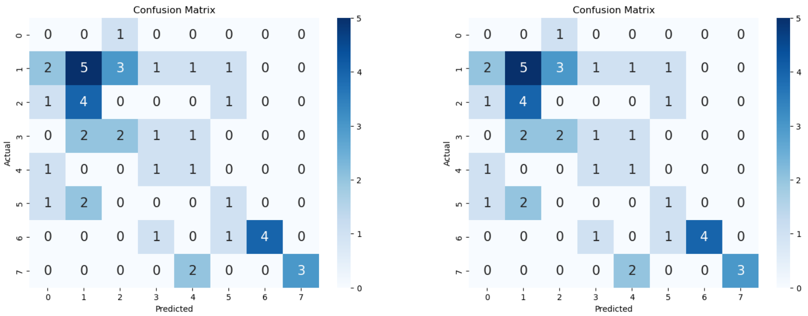
A significant result across all the predictions is that in most cases the standard deviations of the mean accuracy are frequently between 1/2 and 1/3 of the mean accuracy itself. This reveals that the models are quite sensitive to the split of the training and testing data, and implies that these models will likely generalise poorly. This is typical of predictions made with datasets that are lower in datapoints, as low numbers of observations for given categories will limit the ability of the models to learn and facilitate overfitting. These results are expected given the size and nature of the dataset and emphasise the importance of cross validation in evaluating accuracy for classifications based on smaller datasets.

Analyzing the heatmaps generally, shows that while the 4th(3) and highest grade category is fairly well predicted, the 3rd(2) is misclassified quite significantly. According to table 3, the third bin(Category 2) contains the least integer percentages, at 10 compared to 15, 16, and 60 for the other bins. The assumption of an uneven grade distribution seems appropriate but it may not be entirely true and result in this effect. The distribution of the 4 bin grade feature as can be seen in figure 2, where it can be seen that the third bin contains the least values, though not to the same degree that the total percentage points included at the bin varies from others. The smaller total number of points for the bin leads to less potential patterns and more outliers for models to learn, especially in cases where training testing splits reinforce this. However this effect is not seen at all for the 4th bin which is also underrepresented in the distribution(though not by as much), and other mechanisms likely explain this discrepancy. For the first two bins, there is a lot of misclassification between them. All these patterns in the heatmaps generally hold for all models as will be shown going forward.

5.2.1 Decision Tree Classifier

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
8 Grade bins (k=10)	0.303	0.141	0.275	0.078
4 Grade bins (k=10)	0.413	0.108	0.406	0.137

Table 4: Cross validated accuracies for Decision Tree Classifier model



Actual \ Predicted	0	1	2	3	4	5	6	7
0	0	0	1	0	0	0	0	0
1	2	5	3	1	1	1	0	0
2	1	4	0	0	0	1	0	0
3	0	2	2	1	1	0	0	0
4	1	0	0	1	1	0	0	0
5	1	2	0	0	0	1	0	0
6	0	0	0	1	0	1	4	0
7	0	0	0	0	2	0	0	3

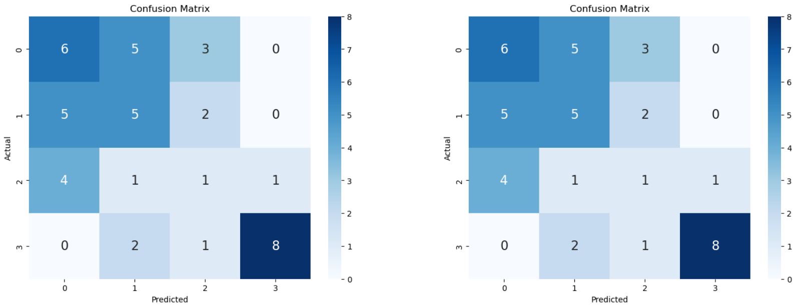
(a) Small Dataset

Actual \ Predicted	0	1	2	3	4	5	6	7
0	0	0	1	0	0	0	0	0
1	2	5	3	1	1	1	0	0
2	1	4	0	0	0	1	0	0
3	0	2	2	1	1	0	0	0
4	1	0	0	1	1	0	0	0
5	1	2	0	0	0	1	0	0
6	0	0	0	1	0	1	4	0
7	0	0	0	0	2	0	0	3

(b) Large Dataset

Figure 6: Heatmaps from DTC predicting 8 bin grades

The goal of implementing the DTC models was primarily investigative. Table 4 shows the Cross validated mean and standard deviation accuracies for the model in 4 domains. Contrasting the accuracies of the 8 bin and 4 bin results shows a marked increase. This is an expected result, based on there being less possible opportunities for mis-classification, and the overall simplification of the target variable. As a result of this, and because the trend of whether a grade is lower or higher due to the independent variables is more important than the explicit category, the 4 bin grade category is used for all further predictions. Differences between the mean CV accuracy between the two sizes of dataset are quite small, and considering the relative size of the standard deviation and



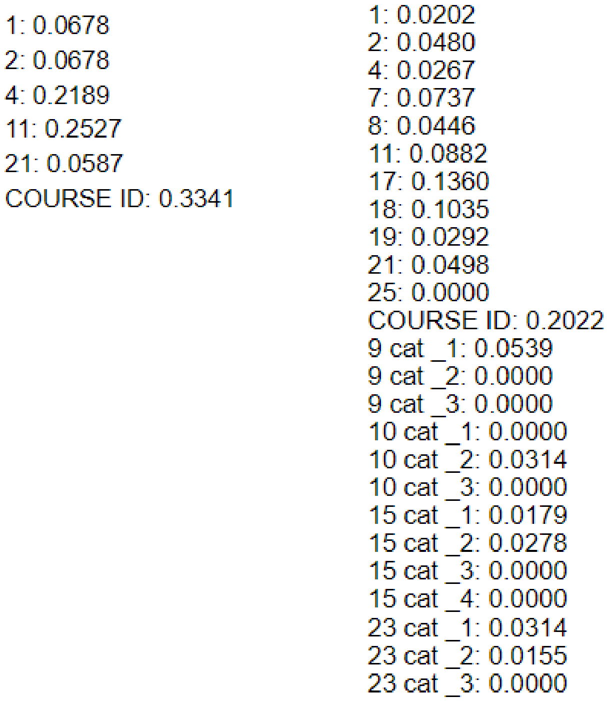
(a) Small Dataset

(b) Large Dataset

Figure 7: Heatmaps from DTC predicting 4 bin grades

the general sensitivity to the train/test split, the differences between the two sets of predictions can be taken as insignificant. This implies that there is little additional accuracy gained as a result of the inclusion of more features in the larger dataset with this model.

Feature importance of the (encoded) features created from the DTC classifier can be found in figure 8. This particular metric for feature importance gives each feature a score denoting it's contribution to the total predictions, with the combined scores adding to 1. For the small dataset model, features 4(Scholarship Type) and 11(Mothers' Education) were two of the most contributing to the predictions, while in the big dataset model, 17(Weekly Study Hours), 18(Non- Scientific Reading) are some of the highest, followed closely by again 11, and 7(Relationship). The Course ID feature dominates both models quite significantly however as mentioned before, the Course ID feature has been encoded for anonymisation purposes and will not be of much value for interpretation or causal inference, though it can be said that students from different courses received significantly different grades on average.



```
1: 0.0678
2: 0.0678
4: 0.2189
11: 0.2527
21: 0.0587
COURSE ID: 0.3341
```

(a) Small Dataset

```
1: 0.0202
2: 0.0480
4: 0.0267
7: 0.0737
8: 0.0446
11: 0.0882
17: 0.1360
18: 0.1035
19: 0.0292
21: 0.0498
25: 0.0000
COURSE ID: 0.2022
9 cat _1: 0.0539
9 cat _2: 0.0000
9 cat _3: 0.0000
10 cat _1: 0.0000
10 cat _2: 0.0314
10 cat _3: 0.0000
15 cat _1: 0.0179
15 cat _2: 0.0278
15 cat _3: 0.0000
15 cat _4: 0.0000
23 cat _1: 0.0314
23 cat _2: 0.0155
23 cat _3: 0.0000
```

(b) Large Dataset including 1HE categories

Figure 8: Feature importances extracted for from the DTC classifier model

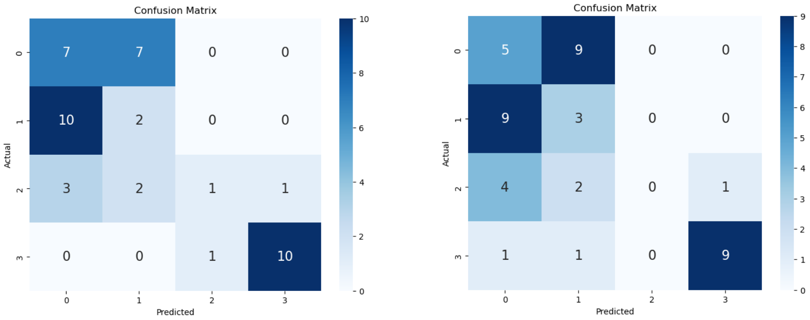
5.2.2 Suport Vector Machine

One goal of using the SVM model was to demonstrate the value of hypertuning models through comparison with untuned ones. As can be seen from table 5, the differences in accuracy are quite significant, around 34% & 46% increases in accuracy for the small and large dataset models respectively. Compared to the DTC models, the tuned SVM is notably more accurate, though the difference is not particularly large and the DTC models are completely untuned. Another observation is that the hyperparameter tuned models have significantly smaller standard deviations of accuracy. Through finding the most optimal parameters in the grid, the model has found parameters that produce the least poor results, leading here to a tighter clustering of accuracies. It isn't clear whether increasing k in this case would reduce or increase the stdev. There is not a significant difference between the error patterns of these classifications compared to either the DTC classifiers, or with each other. The only one that could potentially be significant is that the untuned models underpredicts category 2 more

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
Default parameters (k=10)	0.331	0.119	0.344	0.180
Hyperparameters tuned (k=5)	0.446	0.073	0.504	0.093

Table 5: Cross validated accuracies for SVM Classifier model

significantly than the others.

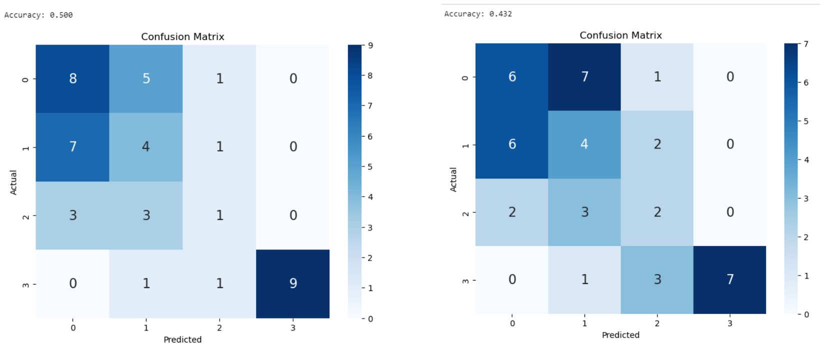


(a) Small Dataset

(b) Large Dataset

Figure 9: Heatmaps from SVM default parameters

For the hypertuned models, there is a moderately significant difference between the accuracies using the small vs the large dataset. It was intended for the SVM method to be used to test for the effects of dimensionality, hypothesising that models created with the larger data set might suffer in performance due to the curse of dimensionality. This was not in fact the case, as the larger dataset, with almost 4x the number of features, outperformed the smaller dataset by 1̃3%. This strongly implies that the curse of dimensionality has not come into effect here. The model predicts slightly better when using the extended dataset. It is unclear whether this is a function of simply the quantity of features involved in the prediction, or the quality of some of the features excluded from the smaller dataset. To evaluate the dependence on the training/testing split being used, the separate split of the data was compared with the arbitrarily chosen default split in figure **??**. While the alternate split seems to have a bias towards predicting category 1 as opposed to category 0, and seems to be slightly less accurate for category 3, it still follow the same broad pattern, in that it underpredicts category 2 and frequently misclassifies categories 0 and 1. As a result it can be assumed that these aspects of the heatmaps are not just a function of split.



(a) Small Dataset

(b) Large Dataset

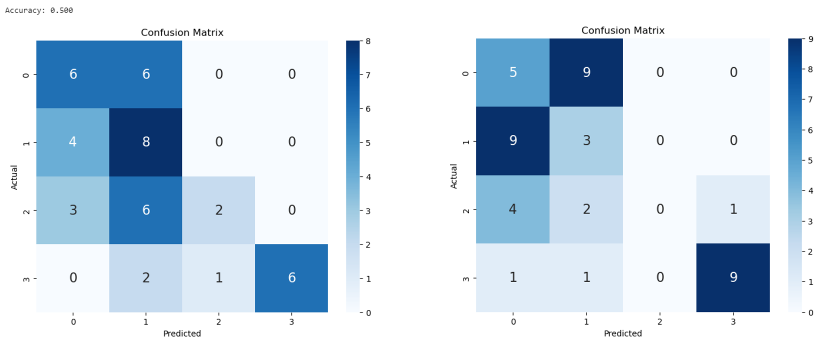
Figure 10: Heatmaps from SVM after Hypertuning. Accuracies for specific split included

5.2.3 XGBoost Classifier

With just the default parameters, and increase in accuracy vs the previous two type of untuned models can already be seen. Note that for the visualisation of larger dataset on the used split in 12, the majority of category 2 observations are predicted correctly, better than for any other model. XGBoost models have large numbers of parameters and hypertuning such models can be quite computationally/time intensive. As a result two strategies for hyperparameter tuning have been used. For the first instance, using the combination of the learning curve method to minimise the total parameters for gridsearch, was not successful at all in significantly increasing the accuracy of the model, with a very similar cross tuned accuracy to the default model. For the bayesian hyperparameter methods on the other hand, there was a noticeable difference. It is also important to mention that increasing the number of iterations for the bayesian model notably increased the maximum CV accuracy of the model, for n=10, best score = 0.448, for n = 100, best score = 0.483, and for n=200, best score = 0.497. There were issues trying to run the model for n=200 and it was very time consuming relative to the project, so further increasing the accuracy via running the model with higher n wasn't pursued.

	Small Dataset		Large Dataset	
CV Accuracy	Mean	StDev	Mean	StDev
Default parameters (k=15)	0.426	0.167	0.419	0.176
Learning Curve and GridSearchCV tuned (k=5)	0.432	0.164	-	-
Bayesian Optimiser tuned, n=200 (k=5)	-	-	0.497	-

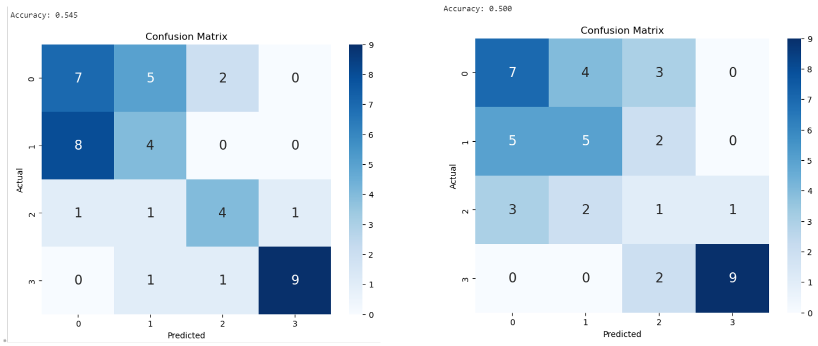
Table 6: Cross validated accuracies for XGBClassifier model



(a) Alternative split

(b) Default split

Figure 11: Heatmap of an alternate split vs the default split used for the SVM with the Large Dataset (no tuning)



(a) Large

(b) Small

Figure 12: Heatmap of the untuned XGB classifier models

The feature importance plots produced for the XGB model contain different numbers of features, and the plot with less features shows greater magnitudes for the f score, which are both expected results, but amongst the shared features there are notable differences in the order of features. For the purposes of modelling causality, the causal graph was created based on the most important features from the plot based on the larger dataset. A minimum of 50 F score is used as a threshold for feature inclusion. Features from the original that were not included in the creation of the ML models in the first place are also excluded. It is assumed that the excluded features don't have sufficient effect on the target variable to warrant inclusion. This assumption is very unlikely to be correct, however it serves an important purpose in allowing the causality of the scenario to be modeled without excessive detail in the causal graph. The incorrectness of this assumption does render the resulting causal inference itself significantly less valid, but the assumption is stated so the causal inference model can be evaluated on that basis.

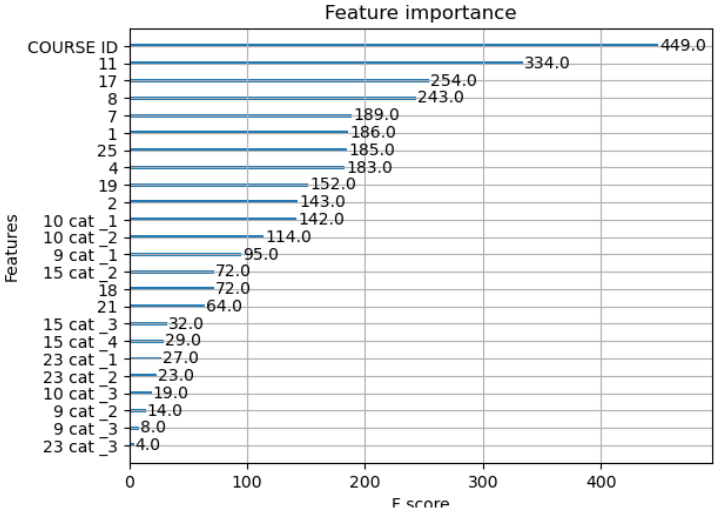


Figure 13: Feature importances from the hypertuned XGBClassifier on the larger dataset

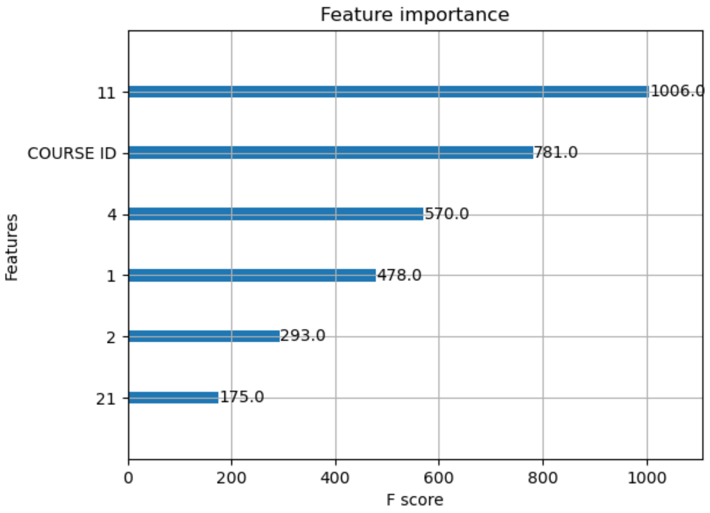


Figure 14: Feature importances from the hypertuned XGBClassifier on the smaller dataset

5.3 Causal Inference

For one of the target features used, WeeklyStudy, it was tested whether using the 8 bin or 4 bin Grade as target made any difference to the outcome of the causal inference. It was found that there was no difference when predicting the ATE using the 8 bin or 4 bin grade feature as the target. This result represents the total magnitude of the grade being cut in half, but the distances between the bins doubling proportionately. The results of the

Treatment variable	**Category #**	**Average Treatment Effect**
Mothers education	11	0.082
Weekly study	17	-0.076
Student wage	8	-0.161
Scholarship	4	0.079
Non-Scientific reading	18	0.383
Partner	7	0.018

Table 7: ATE for various features using the linear regression causal inference method

Linear Regression	**Propensity score weighting**
0.018	0.063

Table 8: ATE for the "Partner" feature comparing the linear regression and propensity score weighting

6 Conclusions

As an attempt to infer causality and to compare the causal effects for various features on the target student grade variable, the results of the project highlighted quite a lot of the limitations to the approach chosen to infer causality. Taking data created for one purpose and trying to use it for another purpose is inherently inferior to designing data collection methods with the goal of the research in mind and based on a significant amount of domain knowledge. For this project, the attempt was made to infer causality without specialist domain knowledge, relying instead largely on intuitions related to student performance and relationships between the collected variables. Also, the graph was designed based only on features deemed to be important from the prediction models, as opposed to considering the research question of "What variables have the strongest Causal effect" as the starting point for the approach and coming up with features based on that While causal inference is inherently imperfect, and all created causal models are based on assumptions that may not be correct, increasing the number of assumptions and using less well founded assumptions combines to significantly increase the error in the measurement of treatment affects, and reduce the validity of the causal model. It became apparent upon the analysis of results that the approach used was not satisfactory

In creating the causal DAG graph, many assumptions were made that are likely to be inaccurate. The first being that all the features were a priori assumed to be directly or indirectly considered to be causal to the target. Especially the fact that many are assumed to be both. The principles in Huntington-Klein(2021) Chapters 6 and 7 outline a few of the areas where the graph creation went astray: there wasn't significant revision of the graph to check for things like unimportance

while it is difficult to know what kind of ATEs would have been found if the causal graph had been modeled differently without

The backdoor linear regression model is not

To comply with the restrictions of DoWhy for propensity score matching on binary treatments only, the data could have been engineered to create binary treatments, allowing for the use of a more diverse range of scores. creating something like "Mother third level education: y/n" would have been appropriate in this regard, as would several others for the other treatment variables Alternatively, the nonbinary categorical potential treatments could be analysed using a threshold crossing approach, or discontinuity analysis, by creating several binary thresholds from the treatment feature and combining multiple causal inference instances created with them to analyse the resulting pattern.

Regression discontinuity design has a lot of research involved in it, and is a slightly different approach that uses a running variable that is increased and defines the threshold as it is increased

References

Aliprantis, D. (2015), 'A distinction between causal effects in structural and rubin causal models'.

BartozKP (2014), 'How to obtain features' weights'.
URL: *https://stackoverflow.com/a/21260848*

Brownlee, J. (2021), 'Tune xgboost performance with learning curves'.
URL: *https://machinelearningmastery.com/tune-xgboost-performance-with-learning-curves/*

Chen, H., Harinen, T., Lee, J.-Y., Yung, M. & Zhao, Z. (2020), 'Causalml: Python package for causal machine learning', *arXiv preprint arXiv:2002.11631* .

Data Protection Commission (2022), 'Anonymisation and pseudonymisation:full guidance note'.
URL: *https://www.dataprotection.ie/en/dpc-guidance/anonymisation-and-pseudonymisation*

Diamond, A. & Sekhon, J. S. (2013), 'Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies', *Review of Economics and Statistics* **95**(3), 932–945.

Distributed (Deep) Machine Learning Community (2024), 'Awesome xgboost'.
URL: *https://github.com/dmlc/xgboost/tree/master/demoawesome-xgboost*

Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. (2019), 'Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition', *Statistical Science* **34**(1), 43 – 68.
URL: *https://doi.org/10.1214/18-STS667*

geeksforgeeks (2023), 'Svm hyperparameter tuning using gridsearchcv — ml'.
URL: *https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/*

Gelman, A. & Imbens, G. (2013), 'Nber working paper series why ask why? forward causal inference and reverse causal questions'.

Global AI, P. C. 2022), 'Causal analysis utilizing causalml', Available at `https://medium.com/@publiciscommerce/causal-analysis-utilizing-causalml-bfabd8015860` (29/12/2023).

Hernán, M. A., Hsu, J. & Healy, B. (2019), 'A second chance to get causal inference right: A classification of data science tasks', *CHANCE* **32**(1), 42–49.
URL: *https://doi.org/10.1080/09332480.2019.1579578*

Huntington-Klein, N. (2021), *The Effect: An Introduction to Research Design and Causality*, 1st edn, Chapman and Hall/CRC.
URL: *https://doi.org/10.1201/9781003226055*

Ibeling, D. & Icard, T. (2023), 'Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions'.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Kerzner, S. (2022), 'A complete guide to causal inference', Available at `http://www.https://towardsdatascience.com/a-complete-guide-to-causal-inference-8d5aaca68a47` (29/12/2023).

Künzel, S. R., Walter, S. J. & Sekhon, J. S. (2019), 'Causaltoolbox—estimator stability for heterogeneous treatment effects', *Observational Studies* **5**(2), 105–117.

Microsoft Research (2019), 'Econml: A python package for ml-based heterogeneous treatment effects estimation'.

Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., Raglin, A. & Liu, H. (2021), 'Causal inference for time series analysis: problems, methods and evaluation', *Knowledge and Information Systems* **63**, 3041–3085.

Navas, J. & Liaw, R. (2022), 'Guide to xgboost hyperparameter tuning'.
URL: *https://www.anyscale.com/blog/how-to-tune-hyperparameters-on-xgboost*

Parikh, H., Varjao, C., Xu, L. & Tchetgen, E. T. (2022), 'Validating causal inference methods'.
URL: *https://proceedings.mlr.press/v162/parikh22a.html*

Pearl, J. (2009), *Causality*, Cambridge university press.

Schuler, A., Jung, K., Tibshirani, R., Hastie, T. & Shah, N. (2017), 'Synth-validation: Selecting the best causal inference method for a given dataset', *arXiv preprint arXiv:1711.00083* .

Scott, P. W. (2019), 'Causal inference methods for selection on observed and unobserved factors: Propensity score matching, heckit models, and instrumental variable estimation', *Practical Assessment, Research, and Evaluation* **24**(1), 3.

Sharma, A. & Kiciman, E. (2020), 'Dowhy: An end-to-end library for causal inference', *arXiv preprint arXiv:2011.04216* .

Stuart, E. A. (2010), 'Matching methods for causal inference: A review and a look forward', *Statistical Science* **25**, 1–21.

Toth, D. J. (2024), 'Binary classification: Xgboost hyperparameter tuning scenarios by non-exhaustive grid search and cross-validation'.
URL: *https://towardsdatascience.com/binary-classification-xgboost-hyperparameter-tuning-scenarios-by-non-exhaustive-grid-search-and-c261f4ce098d*

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J. & Zhang, A. (2021), 'A survey on causal inference', *ACM Transactions on Knowledge Discovery from Data* **15**, 1–46.

Yilmaz, N. & Sekeroglu, B. (2019), Student performance classification using artificial intelligence techniques, *in* 'International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions', Springer, pp. 596–603.

A Causal Graph Assumptions

```
"AccomodationRental/Dorm" -> AvailableTime
"AccomodationRental/Dorm" -> GRADE
"AccomodationRental/Dorm" -> Transportation_Bus
Age -> "AccomodationRental/Dorm"
Age -> AvailableTime
Age -> GRADE
Age -> MotherHousewife
Age -> OtherReading
Age -> Partner
Age -> Scholarship
Age -> StudentInterest
Age -> StudentWage
Age -> Transportation_Bus
AvailableTime -> ScientificReading
AvailableTime -> WeeklyStudy
Gender -> GRADE
Gender -> Scholarship
Gender -> StudentInterest
Gender -> StudentWage
MotherHousewife -> "AccomodationRental/Dorm"
MotherHousewife -> AvailableTime
MotherHousewife -> GRADE
MotherHousewife -> StudentWage
MotherHousewife -> Transportation_Bus
MotherHousewife <-> Scholarship
MothersEducation -> "AccomodationRental/Dorm"
MothersEducation -> GRADE
MothersEducation -> MotherHousewife
MothersEducation -> OtherReading
MothersEducation -> Scholarship
MothersEducation -> ScientificReading
MothersEducation -> StudentInterest
MothersEducation -> StudentWage
MothersEducation -> Transportation_Bus
OtherReading -> GRADE
Partner -> "AccomodationRental/Dorm"
Partner -> AvailableTime
Partner -> StudentInterest
ProjectImpact -> GRADE
Scholarship -> GRADE
Scholarship -> Transportation_Bus
Scholarship <-> StudentInterest
Scholarship <-> StudentWage
ScientificReading -> GRADE
StudentInterest -> ScientificReading
StudentInterest -> TakingNotes
StudentInterest -> WeeklyStudy
StudentWage -> "AccomodationRental/Dorm"
StudentWage -> AvailableTime
StudentWage -> GRADE
StudentWage -> Transportation_Bus
TakingNotes -> GRADE
TakingNotes -> ProjectImpact
Transportation_Bus -> AvailableTime
Transportation_Bus -> GRADE
WeeklyStudy -> GRADE
WeeklyStudy -> ProjectImpact
```