National College of Ireland

Exploration and Mining of Educational data for Analyzing Student's Engagement

MSc Research Project
Data Analytics

Sri Durga Meghana Yadav
x16102126

School of Computing
National College of Ireland

Supervisor: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Sri Durga Meghana Yadav
Student ID:	x16102126
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Simon Caton
Submission Due Date:	11/12/2017
Project Title:	Exploration and Mining of Educational data for Analyzing Student's Engagement
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	14th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploration and Mining of Educational data for Analyzing Student's Engagement

Sri Durga Meghana Yadav
x16102126
MSc Research Project in Data Analytics

14th December 2017

Abstract

In the recent years data mining and analytical tools are used to thoroughly observe and find unanticipated relationships with in the data. They are customized into useful, novel ways for the data owners by extracting insights. A lack of appropriate methods, findings for exploring the student-learning data using advanced feature selection methods has paved way for this idea. Events occurring in the most common sources of diverse learning platforms were identified where ever common engagements are observed facilitating Student participation.Completely anonymous data is used for analysis keeping ethical concerns in mind. LIWC is used for breaking down the text into numerical measures, helped to pop out any interesting patterns in the review comments. This research showcases different available advanced feature selection options while trying to exclude predictors with least impact on the dependant variable. Best subsets, Lasso, Ridge and Elastic net were evaluated with lowest MSE of 0.46 achieved for Ridge regression validated by k-fold technique. The data depicting most common student personalities is classified into 3 groups based on discrete learning events. The contribution of this work is an accurate prediction of linear models entities via validating different feature selection techniques. Student personalities were identified based on most common student participation platforms where both numerical and textual measures were considered for analysis.

1 Introduction

Data mining has made major advances in biomedical, medical, engineering, and business fields. The techniques are used to analyze largescale data and discover meaningful patterns such as natural grouping of data records, unusual records detection, and dependencies with in data. However, researchers have not yet adequately addressed approaches to explore for different relationships with in Student data, the propensity to display most common characteristics in Educational data including forms, e-learning platforms, social media etc. The scope for potential insights in educational data is high, but the ability of researchers to showcase the ground truth models is considerably low. This paper focuses on several available advanced approaches about the selection of right features from the educational data. Educational data mining (EDM) emerged in the last few years. As one of the most fundamental skill for all human beings, learning has always formed the basis

of many discussions to illuminate the ways or approaches of the learning process. In an attempt to comprehend learning, experts in such matters look at learning as a robust process in need of constant study. From their perspective of learning, it then becomes imperative that learning is approached from its functional analytics (Dietz-Uhler and Hurn; 2013). Therefore, in defining learning analytics, one can simply refer to it as the extreme inter-disciplinary effort that seeks for more ways of providing learning support (Ferguson; 2012). It also means that learning combines techniques and knowledge derived from data and psychometrics based on all relevant successful learning processes.

In every classroom learning endeavor, the instructor in charge must look for ways of ensuring that there is ample participant engagement if proper learning is the main goal. In this case, based on the study's hypothesis, the instructor's efforts can be evaluated based on the study patterns of the students, who in this case are the participants in the research project at hand. After the instructor takes the participants through the learning activity based on a proper assessment and implementation of the teaching methods, a pattern is supposed to emerge because of the dynamics involved when different participants learn something similar. In gauging the learning analytics at play, the study will assist in looking into ways of selecting important features, explore and try to identify interesting patterns that can shape a participant's behavior and the engagement based on the discrete learning events (Xing et al.; 2016). In the learning analytics study, the instructor will be expected to engage the participants in a learning activity on weekly or monthly basis and provide guidance. The study results will also form an important inference in making conclusions about the general principles and specific research questions.

Through Learning Analytics (LA) application the Georgian State University claims to have eliminated the achievement gap between students from financially challenged backgrounds and their peers. In US students take 20% more classes than that are needed to graduate. Through offering help with the course selection thus it can cut down tuition costs and accelerate graduation rates and retention at universities. At Nottingham Trent University statistics depicts that 27 % of the first-year students are said to have changed their behavior in learning due to fact of implementing learning analytics (Clow; 2013).

Figure 1: Trend - Learning Analytics

Learning Analytics field needs attention, if we could see the LA trend figure for this year 2017 in Figure 1, it has a spike only in the month of April and seeks attention for rest of the time as well. Learning Analytics generally deals with the development of methods that tap educational sources of data to enhance learning process in the institutions.

Now, allow me introduce you to the organization of this paper. Section II consists of literature review and related work which has been done previously in the field of Learning Analytics. And, we will discuss about available techniques in this area of research, for predictive analysis while critically evaluating the available approaches. Section III will introduce the process model, datasets, and we will give a brief description of the feature selection process. It will also focus on the pre-processing steps of this project. Section IV will outline the implementation of different ML methods for feature selection and also take you through unsupervised Clustering and PCA methods. Section V will include the results and evaluation of the outcomes along with some case-studies. In our last section, we will conclude our project work followed by a discussion of future work possibilities.

1.1 Research Questions

How much MSE can be achieved while selecting features using advanced model selection methods? How many clusters can be formed after/before selecting only useful features based on previous research question?

1.2 Purpose

The purpose of this research is to identify most suitable method which showcase better MSE while analyzing a small data set with many columns. We make an estimate and try to reduce penalty by bias variance trade-off in relation to high-dimensionality. The level of difficulty and complexity varies from one data to the other (Thai and Polly; 2016). And so the techniques vary, from basic to mathematical and probabilistic demonstration. They all present a different aspect of Educational Datamining that is relevant to all types of Learners. Lets go through several dimensions of Learning Analytics (Cerezo et al.; 2016). Our goal is to predict the review score as a linear function of attributes (independant variables). For the purpose of interpretation it would be helpful to have a linear model involving small set of attributes

The Course in relation to this research project was delivered in Sep 2017 for 10 weeks through the Classroom Moodle platform and was organized into weekly modules. As all resources and activities were available at the Moodle, participants were encouraged to explore the modules in any order, use the content, and complete the activities, according to their own needs and interests, at any time during the course structured period Betts et al. (2000); Navot et al. (2006). In addition, participants had the option of completing assignments based on their intent or goal Mosteller and Tukey (1977); Sclater et al. (2015, 2016). These were intended to further develop engagement with peers, designed to be undertaken by (i) those wanting to apply the knowledge they gained in the design of their own learning activity and (ii) for those who wanted to obtain a pass grade in the course.

The assignments were open from the beginning of the at least a week before of the course and participants could complete and submit assignment any time before the time provided on the Moodlein keeping with the flexibility around being able to engage with course content at their own pace. However, the limitations of the structured submissions involved time required specific dates to be set for when assessments were submitted by and when peer assessment could begin and end. Hence, the full flexibility of the course design was partly compromised by these deadlines. While the course was designed with

nonlinear flexibility in mind, the course instructors chose to engage with participants on a weekly basis, providing guidance about how to navigate the material for those new to the topic area. Characterization of engagement or participation can be our area of interest for Case-studies.

What is learning analytics(LA)? Learning analytics is the science and art of gathering, processing, interpreting and reporting data related to the efficiency, effectiveness and business impact of development programs designed to improve individual and organizational performance and inform stakeholders. It is also a diverse field that encompass statistics, machine learning, visualization, and ethics. This research in learning analytics is based in Technology-Enhanced Learning. LA is essential in providing useful suggestion to policy makers and learners by analyzing educational data. Why measure learning? The primary focus of learning analytics is to determine whether a learning experience is effective or not. The way effectiveness is defined and measured varies. In its simplest form, effectiveness might relate to knowledge and skills gained during training. A more complex measure might be the impact of training on a business metric such as sales growth or revenue.

What? This dimension refers to the data collected (i.e., the kind of data that the system gather, manage, and use for the analysis). These data can be from the virtual learning environment (VLE) like datacamp, from institutional sources like moodle or from social media like Facebook and Twitter etc. Who? This dimension refers to the stakeholders (i.e., who is targeted by the analysis). They can be teachers, administrators of the educational institution, researchers, system designers, students, etc. Each one of them has its own perspectives, goals and expectations about the learning practice. Why? This dimension is related to the objectives of the analysis i.e., why does the system analyze the collected data (Atif et al.; 2013).

2 Related Work

Massive Open Online Courses(MOOCs) have gained a lot of popularity so that they can provide massive opportunities of online course participation through open access on the web-based platforms. Also, this has led to the need of scholars to incorporate privacy and security issues to the open data on the web-based platforms (Sin and Muthu; 2015). Gradient Learning Analytic System (GLASS) is a web-based platform that supports access and protects personal data, handle management, multi-user support and availability. Students at University of Maryland who have in a position to compare their virtual learning environment through a web-based system were 1.92 times likely to be awarded grade C or higher in respect to those who did not access a web-based system for comparison purposes.The diverse educational data has probed researchers to come up with an easy way to handle data.

2.1 Visualization

Artificial intelligence or machine learning has been employed to analyze broader spectrum of learner data that make it easier for one to realize students at risk and their academic potential thus helping the student to improve his or her performance. At Nottingham Trent University statistics depicts that 27 % of the first-year students are said to have

changed their behavior. This behavioral change has been triggered by data on their analytic dashboards. The data available in the analytic dashboard has enabled some students to compete with others that has resulted to highest engagement score. Dashboard data has enabled tutors to view a better profile of the student and target interaction accordingly. It has also given tutors a quick picture of their material group which will help support them and track their progress. In addition, dashboard data helps tutors know which students were not engaged with their course thus they would contact them using mails (Clow; 2013).

2.2 Machine Learning Methods

Critical Evaluation for model/approach selection: In this section let's explore different available machine learning approaches or possibilities and critically evaluate which method actually suits our analysis and also highlight the methods with in the discussion. **Time series analysis** encompasses finding the sequential relationship between diverse learning behaviors to find out students engagement. This methodology is crucial to those researchers who wish to analyze students learning pattern so as to find out why some students have better learning performance than others (Chatti et al.; 2012). It is also an essential reference for a teacher or tutor to give feedback to respective students via a web-based system. **Clustering** involves grouping a set of data items that belong to the same group that depict more similar features than those in other groups, thus this method is useful to identify a student with similar characteristics for recommending goals for individual students. We will use this method k-means for our analysis as it is more relevant to our research idea. **Decision tree** is the modification of the relationships between data items by classifying model in the form of a tree structure for easy visualization and conceptualization. This method can be applicable in predication for new cases based on the model derived from the existing cases (Romero and Ventura; 2013). As we are not going to deal with a classification problem we are not going to consider this model for our analysis. For instance, it is applicable in a case where one wants to determine whether a certain student will drop out of an online course based on a model built by closely monitoring the online learning behaviors of the students who dropped out and those who completed the online courses successfully (Baker and Inventado; 2014). This method needs timely data, so we can ignore this method for our analysis

Association rule mining entails finding the relationships instances of learning behaviors and the structure of learning contents. Therefore, it gives useful information for researchers to determine learning path content through linking relevant literature. In addition, this method is equally useful in providing references for recommending personalized learning content to the students whose performances meet the premise of rules of association. In recent years several researches about learning analytics in education have been contacted in different countries of the world (Baker and Inventado; 2014). For instance, between 2015 and 2016 learning analytics on European education policy was carried out and it gathered evidence of implementing learning analytics in education sector. The study was focused on the use and the process of implementing learning analytics in any tier of education. The two sources of the study are: inventory of tools, practices and policies from all tiers of the education system including informal and non-formal learning respectively. Secondly, they presented evidence on five case studies that provide deeper understanding into current and recent practices in the piloting of learning analyt-

ics focusing on how it was easy to understand and the possible constraint of implementing learning analytics. This is also

2.3 Ethical Concerns

Collecting information from the public sites doesn't fall under ethical violation (Holland and Holland; 2014) In addition, others argue that when learners post a message in a public forum there is an assumption the content will be read and archived. However, some users may feel uncomfortable with their message being read by researcher and it may be better to get their informed consent for this an advance. For example, a learner may feel violated if they saw post de-conceptualized and highlighted in a publication, thus there is unusual problem with quoting people here too in that some opt their commend to be anonymized while others feel they should be recognized as the author of the publication. In addition, (Sclater and Bailey; 2015) added more concern pertaining obtaining informed consent from student for use of their data from external site. Firstly, the institution has no control over the diverse data protection policies of those web-based platforms. Secondly, it may be impossible to authenticate the student identity accordingly.

To avoid these issues we have completely anonymised student data by generating MD5 hashes for every username. (McDowell et al.; 2014) highlights that the main potential threats to privacy of huge data. Invasion of private communicators an individuals right to private communication may need to be established may need to be re-established in this digital era. Contrary, collection of data and their use can encounter a number of challenges including location and interpretation of data, informed consent, privacy and de-identification of data, classification and management of data.This approach highlights the need for transparency and acknowledgement that student identification is a transient and temporal. Generating MD5 hashes for every username was marked as high priority and accomplished initially at the start of the data collection phase.

3 Methodology

Data analysis is a process through which we can discover meaningful patterns and insights from data. Analytics can be applied to any business data where there is need to predict, interpret and improve business processes. Areas that are most explored in this field are web analytics, predictive analytics, behavioral analysis, marketing optimization etc., Earlier, data was present in smaller volumes and sizes, so the data analysis was done manually. But, with the advent of globalization and user generated data, it has become difficult for any company (or) organization to manually analyze data (Ester et al.; 1996). Therefore, there is requirement for Data Mining techniques which can help automate the processes, finding important patterns from huge and complex data. This technique is also known as discovering knowledge from databases (Fayyad et al.; 1996) (or) KDD. This project uses KDD technique as shown in the below Figure 2 for data processing and predicting.

3.1 Data Integration, Cleaning and Transformation

Before we deep dive into data sources, it is necessary to know about MD5 hash. It's a cryptographic function which takes piece of information (in our case it's email) and

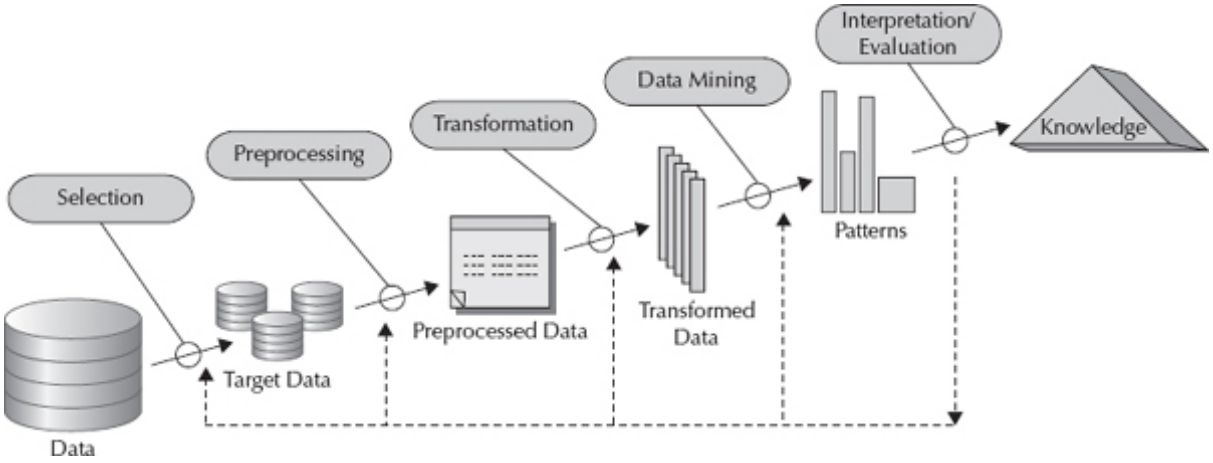


Figure 2: The KDD Process

converts it into long string. Every time the same input detail provided, the resultant string would be same. In other words, this can be considered as an encryption feature. By using R studio digest package we have created Md5 hashes for user-names (emails) to keep the anonymity and avoid any ethical concerns. The data sources are considered by keeping an eye on the most common platforms students are used to and have a good engagement levels. For example, online discussion forums are always available but interaction of students was observed as quite low. So, data sources are picked by sorting and identifying top levels of student interactivity. Now let's deep dive more about the data sources;

1. **Datacamp** is one of the bigger e-learning platforms on which you can learn different courses interactively. The raw data was collected from datacamp logs and it was then encrypted with MD5 hashes. This is a high dimensional data with more columns than rows. Cleaning has been done to replace the null values and missing date values of "/" with relevant values. The data is then transformed (rows to columns and vice-versa) using Excel. Every course in the datacamp has a start date, completed date, grade and %completion, Last active date time-stamp. So, a new column is manually generated using excel double click functionality which categorizes different tasks of courses into single column. Finally, from the transformed data new calculated fields were created for every unique user; Sum(Grade) sum(%Completion), sum(Courses Enrolled),sum(Courses Completed). This serves as a Learning Dimension for understanding student's data.

2. **Moodle** is one of the largely used Learning Management System (LMS) platforms on which you can learn different courses interactively. Moodle logs every click that students make for navigational purposes (Schoor and Bannert; 2012). Log files can be filtered by course, participant, day, and activity, and they can be shown or saved in files with the formats: text format (TXT), open document format for office applications (ODS), or Microsoft excel file format (XLS). The raw data was collected from Moodle Activity logs and it was encrypted with md5 hashes. This is also a high dimensional data with more columns than rows. Cleaning has been done to replace the null values and missing date values of "/" with relevant values. The data is then transformed (rows to columns and vice-versa) using Excel. Every activity in the Moodle has an activity date and it's

own name along with date time-stamp. So, a new column is manually generated using excel double click functionality which categorizes different tasks of activities into single column. Finally, from the transformed data new calculated fields were created for every unique user; Activity_Completed, Activity_Not_Completed fields. This data is not about actual student learning but, how student's look at the data and utilize the data for their learning purposes.

3. **Reviews and Scores**: Student's were provided with a peer-reviewing system called Easychair. One student is assigned to review 4 other papers and asked to write their comments and also scores in relation to the peer-review. This is a very interesting review comments data (after parsing on to LIWC) which actually describes how many average words he/she writes and % of positive words, % negative words used, number of commas, articles used etc;. Easy chair has a very huge log which contains more than 10 tabs in an excel sheet but could consider it as the best peer-review management tool experienced so far (Künzle and Reichert; 2009). As usual the data was anonymized before being provided to me for my research. Text analysis package **Linguistic Inquiry and Word Count** (LIWC) was used as mentioned before to extract important information from the review comments. The reason for choosing LIWC is; it has been found that machine learning approaches often perform better only for prediction tasks (Komisin and Guinn; 2012). All the LIWC entities are extracted using the review comments for every unique user.

4. **Assessment Scores** These are the final evaluated scores that were given to the students after examining their continuous assessments and also reviewing peer-reviews. So, these scores are checked by 4 peer-reviewers and one Lecturer. We can rely on this data. This was anonymized excel file with all the details. Our area of interest is Review Scores column which was computed based on different sub-section scores ("Quality + References + Objective + critical Review"). The computed score is the sum of sub section scores out of 60 total score.In Table 1 all the variables are listed out and briefly explained as follows;

3.2 Statistical Analysis

A Linear regression model is good only when all the assumptions are satisfied and they can be summarized as follows:

- Linearity: There should be linear relationship between the predictor and the response predictors. If this relationship is not identified then log, polynomial or exponential transformations can be your choice to handle this scenario.
- Non-correlation of errors: If the residuals are highly correlated, there is always a risk of creating a poorly fitted model.
- Homoscedasticity: Normally distributed data and similar difference in variance among residuals, the difference in errors variances should be constant across different input vlaues. Violations of this assumption can create biased coefficient estimates, leading to too high or two low values of statistical significance which in turn results increasing more false positives and false negatives. This violation is referred to as heteroscedasticity

Variable	Description	1
Username	generated MD5 hash	1,2,3
Grade	Total grade	1
Completion	Total % of Courses completed	1
Courses_Enrolled	Total sum of Courses Enrolled	1
Courses_Completed	Total sum of Courses completed	1
Last_Active	Last active date	1
Activity_Completed	Total sum of Activity completed	2
Activity_Not_Completed	Total sum of activities not completed	2
Reviewer.No.	Reviewer number	3,4
Submission.No.	Submission number in easy chair	3,4
Review_Score	Total Review Score	4
WC	Total WordCount	3
WPS	Total words per sentence	3
Track #	Track number numeric	3
Track Shot Name	Track Short Name Categorical	3
Average_Six_Letter_Words	Avg of Six letter Words	3
Articles	Average number of articles used	3
Negative	Average negative words %	3
Positive	Average positive words %	3
Comma	Average commas used	3
Participation	Total participation score in all files	-

Table 1: Variables where 1 - Datacamp, 2- Moodle, 3 - Reviews and Scores, 4 - Assesment Scores

- No collinearity: No linear relationship between two independant variables and there should be no correlation identified among the various features. This may lead to biased estimates. For the time being, correlation matrix is developed to identify such case for our research.
- Presence of outliers: Outliers can severely effect the skewness and they must be removed prior to fitting a model using linear regression.

All the data is merged based on unique username md5 hashes and submission #. The Structure of the Combined data can be analyzed using str() function in R. Our statistical tests follow here; the relationship between different continuous variables can be described using correlation matrix as shown in the Figure 3 below and also visualization Distribution. The nature of many continuous variables well explained by a normal distribution. The mean showcases where the distribution's highest level is present and the standard deviation, showcases how much far or close the distribution actually is. We could see there is a strong negative correlation between Activity_Completed and Activity_Not_Completed. All the derived variables from datacamp are correlated to each other. There are also some very weak correlations. We can actually kick out some weakly correlated variables from the data but our intention to use advanced feature selection technique will help us to do that. This is just statistical understanding that we already have acquired now prior to our analysis.

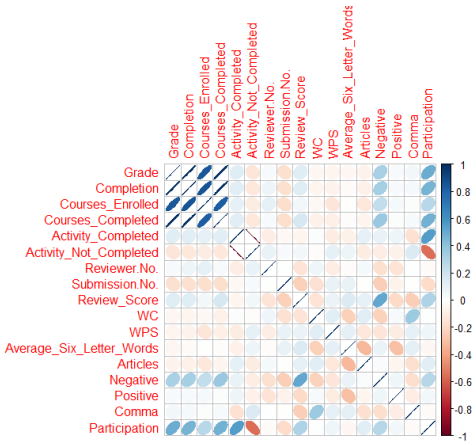


Figure 3: Correlation Matrix

4 Implementation

4.1 Feature Selection and Regularization

Some of the date fields and text fields are removed ans data is standardized using scale function in R. Feature Selection and regularization are the methods and techniques known to exclude useless or unwanted predictor variables. Most of the data sources have numerous random features in relation to the number of observations called as the high-dimensionality issue. Advanced techniques such as best subsets, stepwise feature selection can consume loads of time to interact with anky kind of super or hybrid computers. There is another way which is called as regularization where the coefficients are limited and restricted to the values around zero (Micchelli and Pontil; 2005). There are a number of methods in parallel and conjugation to these methods of regularization but our focus is Ridge regression, LASSO, and elastic net. This is known for combining the advantages of both models into one. In other words of regularization,the application of restricted/limited forfeit is to be in line with a concept of minimizing RSS value. This forfeit contains a lambda value with normalized beta coefficients and weights. Regularization method helps to resolve multi-collinearity issues (Woodbury and Beck; 2013).

4.2 Model Comparison

First of all we divide the data into training set and test set. Best subsets are also available options in parallel to step wise methods like backward and forward, where the former one adds feature one after the other until maximum variance is explained by dependant variable and the later one adds all features at once and removes variables which explain minimum variance of independant variables. A best subset is an alternative to previous stepwise and an object created on training data would choose the features and these features will be modelled on test set data. This can be evaluated using Mean Squared Error (MSE). In our case, we have less number of rows than features. Different statistical measures like Aikake's Information Criterion (AIC), Mallow's Cp (Cp), Bayesian Information Criterion (BIC), and the adjusted R-squared are available for feature selection. In general Cp will hold up more features than BIC (Foster and George; 1994). BIC will tend to take on a small value for a model with a low test error, and so generally

we select the model that has the lowest BIC value. There are functions like which.min() and which.max() functions to differentiate between statistical measures mentioned above. Deviance is the negative two times the maximized log-likelihood that plays the role of RSS for a broader class of models. Also, the advanced feature selection models that we are going to experiment are explained as follows;

- Best Subsets: The actual search is to achieve 2p for heavy data sources but it's not a good choice to attempt. Using R we try to fit lonely model for each and every value of lambda which makes it more productive. This model simply predicts the sample mean for each observation.

- Ridge regression: The normalization term is the sum of the squared weights, referred to as an L2-norm. As the value of lambda rises, the coefficients restrict or limit themselves around the values of zero but not completely become zero valued. The advantages are more predictability score but limitation is not reaching the zero for any features can cause false or noisy results.

- LASSO : It applies the L1-norm instead of the L2-norm and it is the addition of absolute values of all the weights of the features. This limited or restricted penalty can cause the high eligibility of reaching zero. It can be considered as an added advantage when compare to ridge and also aids high accuracy of the model with good results.

- Elastic net: It fulfills the disadvantages of both LASSO and ridge by clustering the features into units that was not possible with other methods. LASSO will show positive nature in selecting feature from cluster of correlated values and avoiding rest of the sample. For the elastic net, a new parameter alpha is in relation to lambda with preferable values of 0 and 1. Lambda will manage the range of error. The zero alpha is ridge regression and one alpha is LASSO.

The Flow diagram for the implementation is shown as below as Figure 4;

4.3 Principal Component Analysis (PCA)

The methods that we have discussed so far have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, X1, X2, . . . , Xp while we also have a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as dimension reduction methods. It is a most common dimensionality reduction technique. PCA drops the number of dimensions in a data source by projecting the data onto a lower-dimensional subspace. a large datasourc is shrinked into a single line. Every case in the datasource can be represented as a single entity instead of grouping them. Similarly higher-dimensional data sources greater than 2 can also be shrinked to one or two dimensions depending upon the variables. It is known to solve the curse of dimensionality problem while working a compressor without missing or unlinking the importance of actual information. Usually they are used prior to the actual model building. Using simple visualization techniques the high dimensional data can be shrinked down into lower dimensions with aiding quick visibility based on the concepts of variance and value decomposition. The former one is the meaure for strength of correlation between pairs. Eigen values and eigen vectors are used to project data into components (Adler and Golany; 2001).

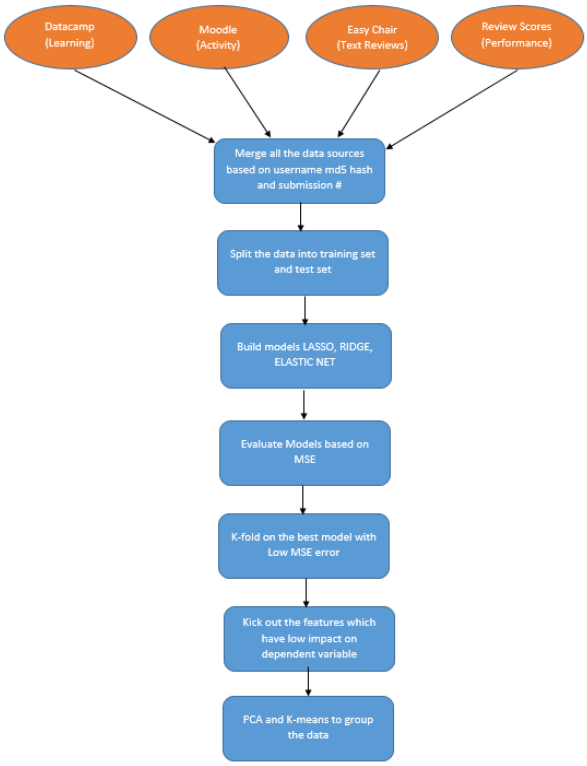


Figure 4: Implementation Methodology

4.4 Elbow Method

This method is used to determine the optimal number of clusters in k-means clustering. For the various values of the K, cost function is created by the elbow method plots. As you know, if k increases the straight line pattern will decrease. Each and every cluster will have some instances nearer to the centroids. However, the improvements in straight line pattern will reduce when k increases. The value of k will be better in straignt line pattern and the place where tilted-ness reduces is known as the elbow. Once we reach this point we don't need to dive into more clusters and stop our analysis at this cluster value (Ajin and Kumar; 2016).

4.5 K-means - Clustering

Using K-mean clustering and MRF feature selection, (Saeb et al.; 2015) performed opinion mining on Thai restaurant reviews. They assessed that, in comparison to Self-Organizing Map, Fuzzy C-Means, and Hierarchical Clustering, K-Mean clustering has a better performance. However, (Ajin and Kumar; 2016) concluded that the clustering algorithms suffer from stability issues, high computational time requirements and do not perform well for all characteristics of Big Data. These issues can be resolved by using ensemble clustering, relying on advance hardware, and selecting the best programming languages for handling specific clustering algorithm. (Ng et al.; 2002) experiments on clustering for sentiment analysis revealed some interesting findings. On balance dataset, k-means clustering performs exponentially better than on unbalanced dataset. Newly

design weighting models also perform better than traditional weighting models. Adjective and adverb words extraction offer improvements on clustering performance. However, stemming and stopword removal has the opposite effect.

5 Evaluation

All the evaluations were based on the methods discussed before in the methodology phase. We choose the MSE to evaluate different methods. Handling high-dimensional data is made easy with the advent of advanced feature selection methods.We have modern methods like elastic net originated from Stanford university. Its a dynamic combination of Ridge and Lasso regression and known for doing fantastic job of dealing with high dimensional data. Ridge regression takes number of variables that are highly correlated and divides the co-efficients among them. Lasso regression takes a number of highly correlated variables and get rids of all them and entire co-efficient to one of them. Hence, ridge is good at smoothing up predictions and lasso is greater for variable selection. Combination of both is Elastic net. The underlying code is in FORTRAN. So, its said to be incredibly fast.

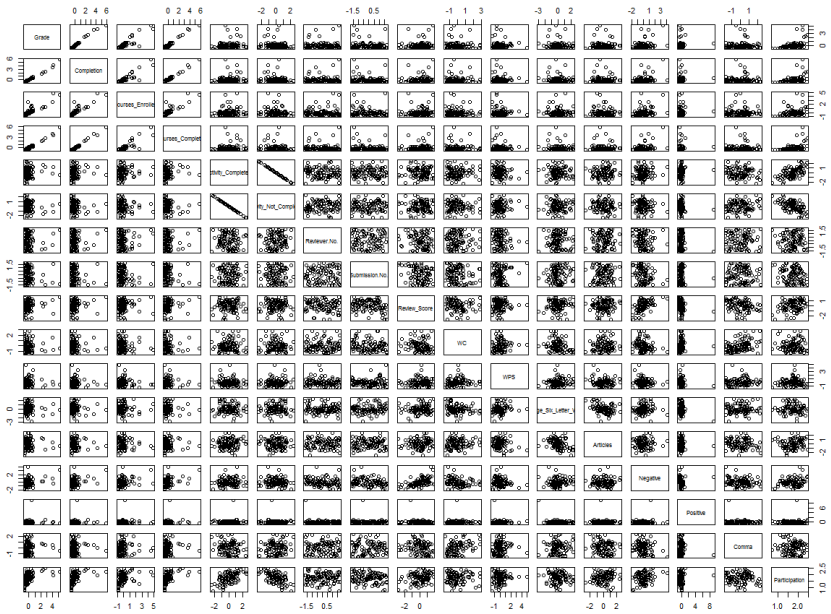


Figure 5: Scatter Plot Matrix

If we create scatterplot matrix with many variables,sometimes difficult to understand the relationship between variables. While normalizing the data, there can be situations in a data frame where input figures for one feature ranges from one to hundred or figures for other features range from one to ten thousand. In such conditions, having different numeric ranges for different random variables might impact predictability score. Our goal is to have better predictability score and not have a dominated feature impacting the predictability score due to random numeric input figures within the data. Thus, we may need to scale variables so as to normalize and make them fall or have common numeric range. Although there are techniques like Min-Max normalization technique, z-score standardization, we would simply use scale function in R. We can also check for

the linear relationship between variables, other assumptions just checking out the scatter plot as shown in the Figure 5 above;

In this plot we can observe there is linear relationship between variables; the variables which follow a diagonal path from lower left to upper right showcases the strong linear relationship especially the Datacamp variables have strong linear relationship with positive correlation. While all other variables which do not have scattered points like a diagonal are not associated with each other but well balanced with adequate dispersion. The positive field looks like it is not associated and might create problems while analyzing. The intention is to understand the data before starting the analysis. As our aim of this research is to explore different advanced feature selection options, for the time being, we would not be eliminating the non-associated variables directly from our analysis. If we drill down more into the field positive, Figure 6 showcased above; Here if we observe there

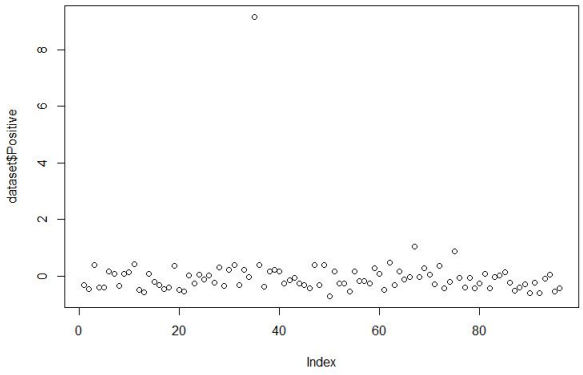


Figure 6: LIWC Positive field plot

is an issue with only one data point for positive field, which is above 8. We can remove the outlier from the dataset as null values may not work with glmnet package. So, now if we go back to the scatter plot matrix it has given more much more clear information but it hasn't highlighted about outliers. Every plot has it own advantages and disadvantages. Depending upon the requirement we need to explore and decide the one that best suits the analysis. The dataset has to be divided into training set and test set.

5.1 Best subset

The best subset model is built using training data and using BIC as the measure, the R output showcases a value of 6 which has lowest BIC value. This was achieved using which.min() function. We can have glance at the performance of different subset combinations as follows in the Figure 7 Actual Model object is also plotted Figure 7; From this plot we can notice completion, positive, comma, participation and Activity_Not_Completed, Activity_Completed are highlighted in darker color. So, these 6 features have the lowest BIC values. As a check for linearity we have to plot the fitted values of OLS with actual values of training set for this 6 features, to make sure the variance is constant. After reviewing the plot we can accept linearity fit and variance is not constant but can be ignored for the moment. Now we fit this data to test set. The fit is not exactly linear

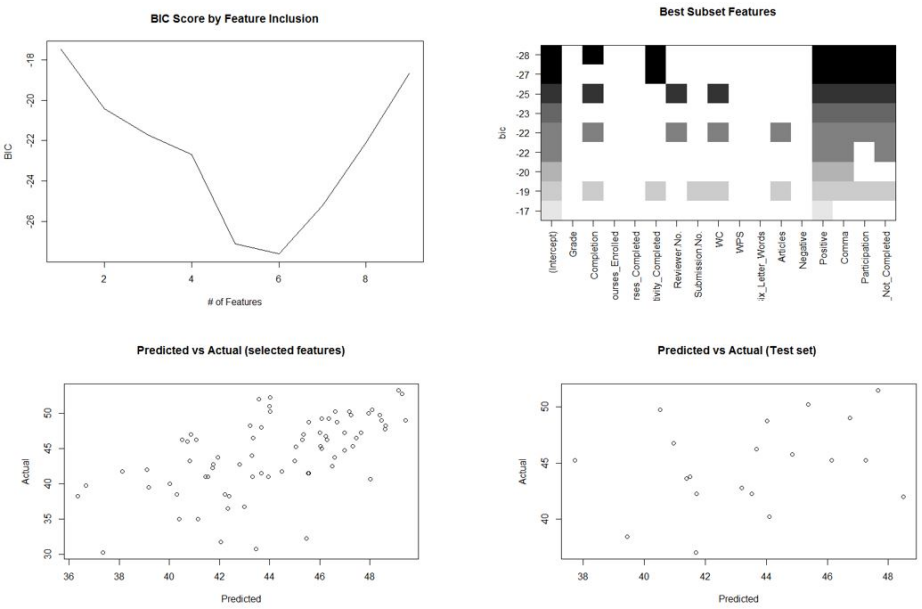


Figure 7: Best Subset

because of the presence of outliers at the edges. The calculated MSE is 0.673. This value serves as basis for our analysis.

5.2 Ridge Regression

In this method, we will have all the features present for analysis. We use glmnet package in R with an alpha value as zero and run ridge regression on training set. The optimal values of nonzero coefficients, percent deviance explained and value of Lambda are displayed using print function as follows;. These are top 2 and bottom 3 optimal values added for reference. From these values the lambda for test set is 0.0563 rounded to 565. Let's evaluate the results using some plots.

Df %Dev Lambda
[1,] 16 1.545e-36 565.30000
[2,] 16 3.091e-03 515.10000

[98,] 16 5.547e-01 0.06809
[99,] 16 5.553e-01 0.06204
[100,] 16 5.558e-01 0.05653

In the plots of Figure 8 the first plot contains an extra axis exactly equal to the number of features in the model. To dig in deep we can explore co-efficient values with lambda values in the next plot. We can notice here for the reduction in lambda values the co-efficient distances increased. We specify the lamba value as 0.06 and fit the model again. By comparing the plots prior to this we can notice that as lambda falls, the coefficients rise and the fraction deviance rises as well. If the lambda was chosen to be zero then our model would be equal to ols because of lack of shrinkage penalty. We can apply this on the test set. The Actual Vs Predicted is almost similar to subset model with outliers at the edges. We can explore, dig deeper for this outliers and try to underrstand if there is

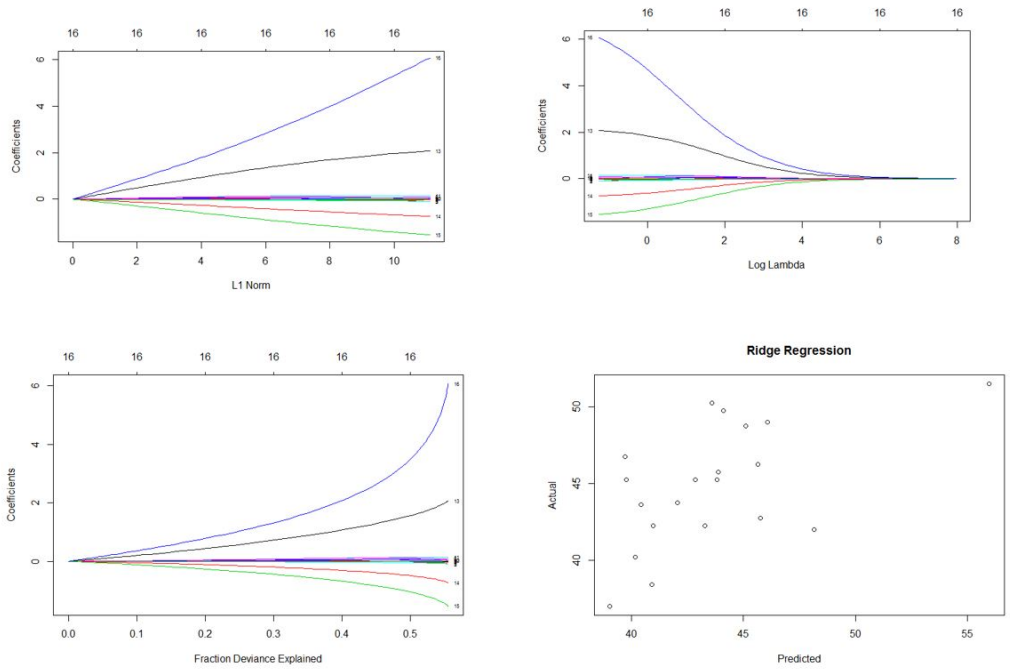


Figure 8: Ridge Regression

any noise or wrong calculated fields in the data source. The MSE for this model is 0.547 and a better value is achieved than the previous subset model.

5.3 LASSO

In this method, we will have all the features present for analysis similar to Ridge. We use glmnet package in R with an alpha value of one and run LASSO regression on training set. The iteration stopped at 85 and df is varying with varying values of lambda. This time we will choose the lambda when df changed i.e; 0.002336 with less features (15). We plot results for LASSO Figure 9 similarly like the one's did for ridge regression. The lines correspond to labels correspond to the actual feature numbers. There are many fields which are closer to the zero and even less than zero. So, the MSE for this model has actually got worsen than the previous model with a value of 0.586. In the case of the lasso method, the L1 shrinkage penalty has the effect of kicking and pressurizing some of the coefficient estimates to be exactly equal to zero especially when the tuning parameter is sufficiently large. As in ridge regression, selecting a good value of for the lasso is critical; cross-validation is again the method of choice to reduce error.

Df %Dev Lambda
[59,] 15 0.55880 0.0025630
[60,] 15 0.55910 0.0023360

[84,] 16 0.56110 0.0002504
[85,] 16 0.56110 0.0002282

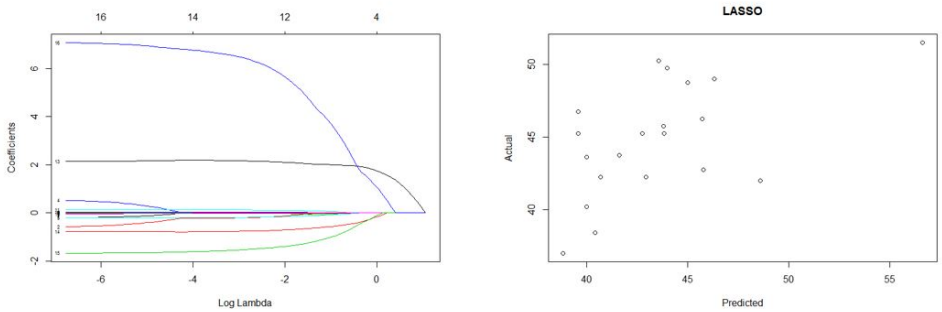


Figure 9: LASSO Regression

5.4 Elastic net

We focus on finding out optimal lambda mix and alpha by utilization of caret package in R. Once the parameters are selected, they are applied on to the test data set. Alpha ranges from 0 to 1 and lambda values are chosen based on previous iterations of Ridge and Lasso. For Lasso, chosen lambda was 0.0023 and Ridge was 0.06. We will choose range of values in between these limits. The considered values range from 0.02 to 0.002. LOOCV is chosen as a re-sampling method.

glmnet

76 samples
16 predictors
No pre-processing
Re-sampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 75, 75, 75, 75, 75, 75, ...
Re-sampling results across tuning parameters:
alpha lambda RMSE Rsquared MAE
alpha lambda RMSE Rsquared MAE
0.0 0.000 0.8813462 0.3453395 0.6500147
0.0 0.002 0.8813462 0.3453395 0.6500147
0.0 0.004 0.8813462 0.3453395 0.6500147
0.0 0.006 0.8813462 0.3453395 0.6500147
0.0 0.008 0.8813462 0.3453395 0.6500147
0.0 0.010 0.8813462 0.3453395 0.6500147
0.0 0.012 0.8813462 0.3453395 0.6500147
0.0 0.014 0.8813462 0.3453395 0.6500147
0.0 0.016 0.8813462 0.3453395 0.6500147
0.0 0.018 0.8813462 0.3453395 0.6500147
0.0 0.020 0.8813462 0.3453395 0.6500147
0.2 0.000 0.9193438 0.3257838 0.6822545
0.2 0.002 0.9123496 0.3306230 0.6770626

1.0 0.014 0.8669636 0.3582895 0.6463997
1.0 0.016 0.8646141 0.3586484 0.6450910
1.0 0.018 0.8626573 0.3586822 0.6440800
1.0 0.020 0.8609231 0.3585583 0.6433073

From the data interpretation our alpha is 0 and lambda is 0.02 which means it is a ridge regression with s = 0.02. The test set is validated with these values. The calculated MSE is 0.571 which is lower than ridge regression MSE again. The predicted plot Figure 10 for elastic net is as follows;

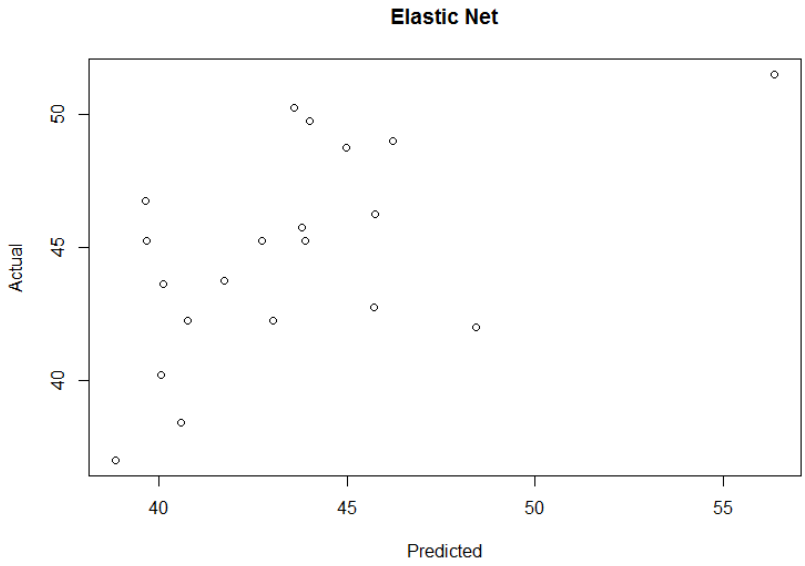


Figure 10: Elastic net

5.5 Cross validation

The k-fold cross-validation technique is the most widely used technique for performance estimation of different models. The over-fitting issues can be reduced with this technique. As we have the data divided into training and test sets, k is the number of iterations used for assessing the performance of model on test set. The average iterations prorate to form a unique accuracy which is noted for the best performance of the built model. The log(Lambda) versus Mean-Squared Error along with the number of features Figure 11 is as follows;. In our case, we are going to test different model's performance using k-fold again. As we have noticed out of all the available options ridge regression has outperformed other models with MSE of 0.547. The default alpha is one so we need to provide alpha value zero to perform k-fold on ridge regression. Surprisingly this methodology didn't kick out any feature from the analysis after k-fold. The final mean value after k-fold validation is 0.46. No feature was kicked out or selected in the process and ridge has shown highest MSE value.

5.6 Case study - Clustering and PCA

The major idea while forming the research question was to have differentiate and remove variables that explain least variance of dependant variable. After identifying and understanding that each and every variable is important and contribute % variance for dependant variable, it is important to reduce the variables using PCA and perform cluster

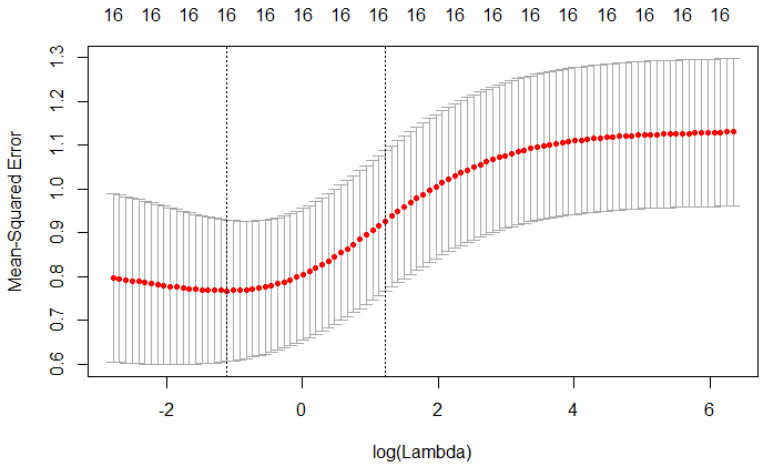


Figure 11: K- fold Cross validation

analysis. This will allow me not only to explore the data but identify students into different groups. Elbow method was used to identify optimum number of clusters. In our case, the optimal number identified was 3. We fit the k-means model with value 3 for clustering the data into similar groups. The following Figure 12 showcases the visualizations in relation to fitting k-means to the student's data. The clustering data is exported to a CSV file to investigate.

Also evaluated this method in rapid miner to have summary of results quicker; where 36% data belongs to cluster 0, 32.8% data belongs to cluster 1, 31.2% data belongs to cluster 2 respectively and the centroid table is populated as Figure 12 follows;
Cluster 0: 35 items Cluster 1: 31 items Cluster 2: 30 items Total number of items: 96

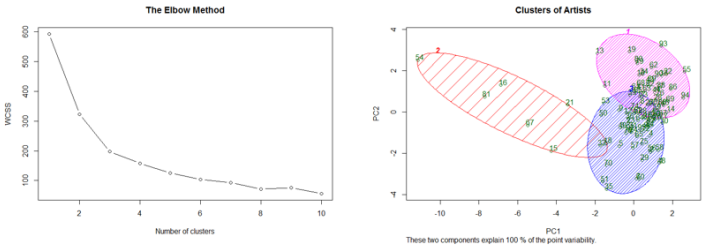
5.7 Results and Discussion

The final results are populated here in the below Table 2 provided.

Model	MSE
Best subset	0.673
Ridge	0.547
LASSO	0.586
Elastic net	0.571
K-fold Ridge	0.46

Table 2: MSE Results table

We have started analyzing student data using different methodologies. Out of all the models best subset, ridge, LASSO, Elastic net. Ridge has shown very good performance. The BIC will always be interested on low values for low errors. So generally we select the model that has the lowest BIC value. Cross validation has more advantage(Willmott and Matsuura; 2005; Ephraim and Malah; 1984). As the Ridge regression is proned to



Attribute	cluster_0	cluster_1	cluster_2
Grade	2.485	1.967	2.925
Completion	2.635	2.185	3.398
Courses_Enrolled	4.257	3.613	6.367
Courses_Completed	2.229	1.613	2.733
Activity_Completed	25.743	24.774	23.433
Activity_Not_Completed	24.257	25.226	26.567
Reviewer.No.	22.971	79.774	74.833
Submission.No.	64.057	92.613	22.833
Review_Score	45.181	43.045	44.431
WC	42.127	42.786	46.673
WPS	17.304	17.257	16.893
Average_Six_Letter_Words	27.681	28.149	27.566
Articles	9.645	10.008	9.539
Negative	3.262	2.841	3.147
Positive	1.485	1.095	1.119
Comma	2.061	2.028	1.987
Participation	1.755	1.718	1.699

Figure 12: Elbow method, Clusters and Centroid Table

seek estimates that has smaller RSS value but the relative impact is controlled by lambda and depends on the lambda value chosen. As we have already scaled and standardized the data before starting the analysis ridge regression showed best performance. It is said to be better than OLS method (Zheng and Dagnino; 2014). Although lasso and elastic net has better advantages than Ridge, one more reason for better performance of ridge regression is that, in our case there might be no variable which explains less variance of the dependant variable. Every in dependant variable has some % contribution to the variance of dependant variable.

6 Conclusion and Future Work

The research idea has showcased many options available for analyzing a high-dimensional data. Discussion around learning analytics were highlighted and reasons for being LA as an attention have been discussed as well. Different machine learning methods were critically evaluate before choosing the models for analysis. Statistical analysis were also presented. Most of the methods to transform or understand a high-dimensional data source were presented. The research questions pertaining to the field of Learning analytics and discussions around different dimensions are discussed. Based on the available methods we were able to partially satisfy the research question and with MSE achieved. However we couldn't find any variables kicked out using this methods. Although the data is standardized format and most of the assumptions satisfied, we identified and dealed with a dataset which is very much interesting. Every independant variable in the dataset has lead to a minimal or acceptable percentage of variance. This has cause other models to have high MSE value while the model which puts all the variables into selection process has won. Depending upon the data and similarities we may experience changes. There are many standardized rule followed methods which may also not perform very

well when it comes to change in data set or domain on which it was tested. This is a very interesting insight that researchers need to concentrate upon, explore and look for a better solution. In terms of standards, our dataset differs from all other datasets which actually gave a good experience to learn new things and lead me to dig deeper into the literature for understanding the same.

Future work: There are other methods which were discussed in the critical evaluation of machine learning models phase. If we had categorical predictors/dependant variable we may have explored how different methods/approaches might have applied to achieve maximum accuracy in classification. We might also dig deeper into the contexts of imbalanced data. There are many hybrid models which can be evaluated on our current data source if the dependant variable is divided into categorical data type. There is also much literature on comparing different methods and identify advanced methods for data in relation to learning analytics. We might try to fill this research gap and explore different neural network or deep-learning models if we have daily/hourly activity data of students with a larger or optimum sample size as deep learning models require a significant amount to data to learn effectively from the data source.

References

Adler, N. and Golany, B. (2001). Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to western europe, *European Journal of Operational Research* **132**(2): 260–273.

Ajin, V. and Kumar, L. D. (2016). Big data and clustering algorithms, *Research Advances in Integrated Navigation Systems (RAINS), International Conference on*, IEEE, pp. 1–5.

Atif, A., Richards, D., Bilgin, A. and Marrone, M. (2013). Learning analytics in higher education: a summary of tools and approaches, *ASCILITE-Australian Society for Computers in Learning in Tertiary Education Annual Conference*, Australasian Society for Computers in Learning in Tertiary Education, pp. 68–72.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics, *Learning analytics*, Springer, pp. 61–75.

Betts, J. R., Reuben, K. S. and Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California.*, ERIC.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P. and Núñez, J. C. (2016). Students' lms interaction patterns and their relationship with achievement: A case study in higher education, *Computers & Education* **96**: 42–54.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U. and Thüs, H. (2012). A reference model for learning analytics, *International Journal of Technology Enhanced Learning* **4**(5-6): 318–331.

Clow, D. (2013). An overview of learning analytics, *Teaching in Higher Education* **18**(6): 683–695.

Dietz-Uhler, B. and Hurn, J. E. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective, *Journal of Interactive Online Learning* **12**(1): 17–26.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32**(6): 1109–1121.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., *Kdd*, Vol. 96, pp. 226–231.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI magazine* **17**(3): 37.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges, *International Journal of Technology Enhanced Learning* **4**(5-6): 304–317.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression, *The Annals of Statistics* pp. 1947–1975.

Holland, J. and Holland, J. (2014). Implications of shifting technology in education, *TechTrends* **58**(3): 16.

Komisin, M. C. and Guinn, C. I. (2012). Identifying personality types using document classification methods, *FLAIRS Conference*.

Künzle, V. and Reichert, M. (2009). Towards object-aware process management systems: Issues, challenges, benefits, *Enterprise, Business-Process and Information Systems Modeling*, Springer, pp. 197–210.

McDowell, G. S., Gunsalus, K. T., MacKellar, D. C., Mazzilli, S. A., Pai, V. P., Goodwin, P. R., Walsh, E. M., Robinson-Mosher, A., Bowman, T. A., Kraemer, J. et al. (2014). Shaping the future of research: a perspective from junior scientists, *F1000Research* **3**.

Micchelli, C. A. and Pontil, M. (2005). Learning the kernel function via regularization, *Journal of machine learning research* **6**(Jul): 1099–1125.

Mosteller, F. and Tukey, J. W. (1977). Data analysis and regression: a second course in statistics., *Addison-Wesley Series in Behavioral Science: Quantitative Methods* .

Navot, A., Shpigelman, L., Tishby, N. and Vaadia, E. (2006). Nearest neighbor based feature selection for regression and its application to neural activity, *Advances in Neural Information Processing Systems*, pp. 996–1002.

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems*, pp. 849–856.

Romero, C. and Ventura, S. (2013). Data mining in education, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3**(1): 12–27.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P. and Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study, *Journal of medical Internet research* **17**(7).

Schoor, C. and Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining, *Computers in Human Behavior* **28**(4): 1321–1331.

Sclater, N. and Bailey, P. (2015). Code of practice for learning analytics, *Joint Information Systems Committee (JISC)* .

Sclater, N., Berg, A. and Webb, M. (2015). Developing an open architecture for learning analytics, *EUNIS Journal of Higher Education* .

Sclater, N., Peasgood, A. and Mullan, J. (2016). Learning analytics in higher education, *London: Jisc. Accessed February* **8**: 2017.

Sin, K. and Muthu, L. (2015). Application of big data in education data mining and learning analytics–a literature review., *ICTACT journal on soft computing* **5**(4).

Thai, T. and Polly, P. (2016). Exploring the usefulness of adaptive elearning laboratory environments in teaching medical science, *Data Mining and Learning Analytics: Applications in Educational Research* pp. 139–155.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate research* **30**(1): 79–82.

Woodbury, K. A. and Beck, J. V. (2013). Estimation metrics and optimal regularization in a tikhonov digital filter for the inverse heat conduction problem, *International Journal of Heat and Mass Transfer* **62**: 31–39.

Xing, W., Chen, X., Stein, J. and Marcinkowski, M. (2016). Temporal predication of dropouts in moocs: Reaching the low hanging fruit through stacking generalization, *Computers in Human Behavior* **58**: 119–129.

Zheng, J. and Dagnino, A. (2014). An initial study of predictive machine learning analytics on large volumes of historical data for power system applications, *Big Data (Big Data), 2014 IEEE International Conference on*, IEEE, pp. 952–959.