edWisor - Data Science

Assignment - 03

NOTE:

- I had created single pdf file instead of separate files for both R and Python code. So it will make the evaluation easy.
- Also, I can create R code file but I prefer jupiter Note book instead of R studio for better visualization.

Problem Statement

Answer the following questions to the best of your knowledge including the concepts taught to you in the level. You can find the IMDB_data in the folder containing this file.

Ques 1. Write R code using data "IMDB_data" to

a. Load CSV in R by skipping second row.
b. Extract the unique genres and its count and store in data frame with index key.
c. Convert the required data types
d. Sort the genre by its name
e. Create new variable whose values should be square of difference between imdbrating and imdbvotes.

Logical Approach

Loading CSV, skipping 2nd row, Finding and Extracting Unique genre & its count :

1. Remove all the previous object store (optional).

2. Set the current working directory (optional).

3. Get the current working directory (optional).

4. Ignore the warnings given by R language(optional).

5. Reading and Loading the IMDB data (.csv file)

6. Skipping the second row using read.csv()[-2,]

7. Extracting unique Genres using unique(df$column_name)

8. Store in dataframe with index key.

9. Chnaging the name of the columns.

10. Showing the obtained results.

Converting data types and Sorting genre

1. List out all column names for reference.

2. Check the class of required variables, if they are factor convert them in numerical vallues.

3. Using **as.numeric(levels(df$coulmn_name))[df$coulmn_name]** convert factor into numeric values

4. Store values and Calculate square of difference between imdbrating and imdbvotes. After conversion simple substraction using (-) minus sign and Squaring using (^) powering sign.

5. Adding that row in the dataset and show the results.

Coding in R language

Programatical Approach

In [1]:

```
#Step 1: Remove all the objects stored

rm(list=ls())

#Step 2: Set current working directory
setwd("D:/edWisor/2. PREDICTIVE ANALYSIS USING R AND PYTHON/R/BASIC_ANALYTIC

#Step 3 : Current working directory
getwd()

options(warn=-1)
```

'D:/edWisor/2. PREDICTIVE ANALYSIS USING R AND PYTHON/R/BASIC_ANALYTICS_AND_PROGRAMMING'

In [2]:

```
#Step 4: Reading and Loading IMDB csv file

IMDB_data = read.csv("IMDB_data.csv", header = T)[-2,]

#Showning that csv is loaded by skipping second row.

head(IMDB_data[1:3,])
```

	Plot	Title	imdbVotes	
1	Despite his tarnished reputation after the events of The Dark Knight, in which he took the rap for Dent's crimes, Batman feels compelled to intervene to assist the city and its police force which is struggling to cope with Bane's plans to destroy the city.	The Dark Knight Rises	2679	imdb.com/images/M/MV5BMTk4ODQzNDY3Ml5BMl5BanBnXkFtZTc
3	Based on the novel written by Stephen Chbosky, this is about 15-year-old Charlie (Logan Lerman), an endearing and naive outsider, coping with first love (Emma Watson), the suicide of his best friend, and his own mental illness while struggling to find a group of people with whom he belongs. The introvert freshman is taken under the wings of two seniors, Sam and Patrick, who welcome him to the real world.	The Perks of Being a Wallflower	1270	imdb.com/images/M/MV5BMzIxOTQyODU1OV5BMl5BanBnXkFtZT

	Plot	Title	imdbVotes	
4	Mike Lane is a thirty-year old living in Tampa,Florida. By day he works as a roofer whilst at night, as Magic Mike, he is the star attraction of the Kings of Tampa, a group of male strippers. Secretly he wants out in order to further a projected furniture-making business but his credit rating precludes a bank loan for this despite his considerable savings. One night Adam, a teen-aged work-mate of Mike, follows him to the club and, when one of the acts is unable to go on,he is prevailed upon to strip - becoming a huge hit. However success goes to his head and his foolish actions not only threaten to jeopardize his sister Brooke's relationship with Mike but Mike's ambitions as well.	Magic Mike	2580	imdb.com/images/M/MV5BMTQzMDMzOTA5M15BMl5BanBnXkFtZ

In [3]:

```
#Step 5: Extracting unique Genres

unique_genre = unique(IMDB_data$Genre)

#Step 6: Unique count of Genres
unique_genre_cnt = length(unique_genre)

#Printing the the count value
cat("Count of Unique Genres is : ", unique_genre_cnt)
```

```
Count of Unique Genres is :  249
```

In [4]:

```
#Step 7: Creating another Data frame var

#Removing 1st row which contains 0 Unique Genre and 0 count.

IMDB_df = table(IMDB_data$Genre)[-1]

#Step 8: Store in dataframe with index key.

index_key  = data.frame(IMDB_df, 1:unique_genre_cnt)


```

In [5]:

```
#Step 9: Chnaging the name of the columns.

colnames(index_key)<-c("Unique_Genre","Count","index")
```

In [6]:

```python
#Step 10: Displaying the result.
index_key
```

Unique_Genre	Count	index
Action	41	1
Action, Adventure	3	2
Action, Adventure, Biography, Drama, History, Musical, Romance	1	3
Action, Adventure, Comedy, Crime, Musical	1	4
Action, Adventure, Comedy, Drama	1	5
Action, Adventure, Comedy, Drama, Fantasy, Musical, Romance, Sci-Fi	1	6
Action, Adventure, Comedy, Drama, Musical, Thriller	1	7
Action, Adventure, Comedy, Horror	1	8
Action, Adventure, Comedy, Romance, Thriller	1	9
Action, Adventure, Crime, Drama	1	10
Action, Adventure, Crime, Drama, Family	1	11
Action, Adventure, Crime, Drama, Thriller	5	12
Action, Adventure, Drama	11	13
Action, Adventure, Drama, Fantasy	2	14
Action, Adventure, Drama, Sci-Fi, Thriller	1	15
Action, Adventure, Drama, Thriller	1	16
Action, Adventure, Fantasy, Horror, Thriller	1	17
Action, Adventure, Fantasy, Musical	1	18
Action, Adventure, Romance, Sci-Fi	1	19
Action, Adventure, Thriller	4	20
Action, Adventure, Western	2	21
Action, Biography, Crime, Drama, Thriller	1	22
Action, Biography, Crime, Sport, Thriller	1	23
Action, Biography, Drama, History	1	24
Action, Biography, History	1	25
Action, Comedy	17	26
Action, Comedy, Crime, Drama	1	27
Action, Comedy, Crime, Drama, Musical, Romance	1	28
Action, Comedy, Crime, Drama, Thriller	2	29
Action, Comedy, Crime, Musical, Romance	1	30
...	...	...
Fantasy, Musical, Sci-Fi	1	220
Fantasy, Romance	2	221
History	10	222

Unique_Genre	Count	index
History, Musical	1	223
Horror	115	224
Horror, Adventure	1	225
Horror, Mystery, Thriller	29	226
Horror, Romance, Thriller	4	227
Horror, Thriller	68	228
Music, Musical	2	229
Musical	21	230
Musical, Comedy, Romance	1	231
Musical, Drama	1	232
Musical, Drama, Action, War	1	233
Musical, Drama, Adventure, Comedy	1	234
Musical, Drama, Family	1	235
Musical, Drama, Romance	1	236
Musical, Mystery, Thriller	1	237
Musical, Romance	3	238
Musical, Romance, Action, Drama	1	239
Musical, Romance, Drama	1	240
Musical, Romance, Drama, Family	1	241
Mystery	21	242
Mystery, Romance	2	243
Mystery, Thriller	27	244
Romance, Sci-Fi	1	245
Sci-Fi	31	246
Sport	14	247
Thriller, Mystery	1	248
War, Drama, Romance	1	249

In [7]:

```
colnames(IMDB_data)
```

```
'Plot'   'Title'   'imdbVotes'   'Poster'   'imdbRating'   'Genre'   'imdbID'   'Year'   'Language'
```

In [8]:

```
cat('Class of imdbVotes:', class(IMDB_data$imdbVotes), '\nClass of imdbRatin
```

```
Class of imdbVotes: factor
Class of imdbRating: factor
```

In [9]:

```
imdb_rg = as.numeric(levels(IMDB_data$imdbRating))[IMDB_data$imdbRating]
imdb_votes = as.numeric(levels(IMDB_data$imdbVotes))[IMDB_data$imdbVotes]
```

In [10]:

```
diff = imdb_rg - imdb_votes
comparison = diff^2

comparisons  = data.frame(comparison)
colnames(comparisons)<-c("comparison")
```

In [11]:

```
IMDB_data[c('imdb_Rating', 'imdb_Votes', 'diff','Square of diff')] <- c(imdb
head(IMDB_data[8:13])
```

	Year	Language	imdb_Rating	imdb_Votes	diff	Square of diff
1	2012	English	75	2679	-2604	6780816
3	2012	English	71	1270	-1199	1437601
4	2012	English	51	2580	-2529	6395841
5	2012	English	68	1807	-1739	3024121
6	2012	English	68	1828	-1760	3097600
7	2012	English	60	1538	-1478	2184484

In [12]:

```
nullSafe <- function(comparisons) {
  if (!exists(deparse(substitute(comparisons))) || is.null(comparisons)) {
    return(NA)
  } else {
    return(comparisons)
  }
}

nullSafe(my.nonexistent.var)
```

```
<NA>
```

Observation:

- Here we, can see that the imdbVotes and imdbRation are not in numeric value they are in factor type so, won't able to perform Arithematic operation but after conversion to numeric value using as.numeric() we will able to perform operation.

- We store the difference between imdbVotes and imdbRation in 'diff' variable and create another column having the square of different of the two columns.

- We can clearly, see the 'diff' and 'Square of diff' column which we include in the existing dataset.

Ques 2. Write Python code using data “IMDB_data” to

a. Load CSV in R by skipping second row.
b. Extract the unique genres and its count and store in data frame with index key.
c. Convert the required data types
d. Sort the genre by its name
e. Create new variable whose values should be square of difference between imdbrating and imdbvotes.

Coding in Python

Logical Approach

Loading CSV, skipping 2nd row, Finding and Extracting Unique genre & its count :

1. Importing required libraries like os, warnings, numpy, pandas, matplotlib etc.,

2. Change the current working directory (optional).

3. Set the current working directory (optional).

4. Get the current working directory (optional) using **os.getcwd()**.

5. Ignore the warnings given by R language(optional).

6. Reading and loading CSV data file using pandas and builtin method **pd.read_csv()**

7. Skipping 2nd row using attribute inside pd.read_csv(skiprows=[2])

8. To Extract unique values of imdb['Genre'] we use unique()

9. For counting unique values use **values_count()**, convert series into data frame we use **to_frame()**

10. using **rename()**renaming columns and Using **sort_values(ascending=True)**, we sort the genre in alphabetical manner and reset_index() for resetting the index values starting from 0. Now display the results.

Converting data types and Sorting genre

1. List out all column names for reference.

2. Check the class of required variables, if they are object type convert them in numerical vallues.

3. Using **pd.to_numeric()** convert object into numeric values

4. Store values and Calculate square of difference between imdbrating and imdbvotes. After conversion simple substraction using (-) minus sign and Squaring using (**) powering sign.

5. Adding that row in the dataset and show the results.

Programming Approach

In [1]:

```
#Set working directory

import os
import warnings
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
warnings.filterwarnings("ignore")

os.chdir('D:/edWisor/2. PREDICTIVE ANALYSIS USING R AND PYTHON/R/BASIC_ANALY
```

In [2]:

```
os.getcwd()
```

Out[2]:

```
'D:\\edWisor\\2. PREDICTIVE ANALYSIS USING R AND PYTHON\\R\\BASIC_ANALYTICS_AND
_PROGRAMMING'
```

In [3]:

```python
# Load csv
imdb_data = pd.read_csv('IMDB_data.csv', sep=',', encoding='unicode_escape',
imdb_data
```

Out[3]:

	Plot	Title	imdbVotes	Poster	imdbRating
0	Despite his tarnished reputation after the eve...	The Dark Knight Rises	2679	http://ia.media-imdb.com/images/M/MV5BMTk4ODQz...	75
1	Based on the novel written by Stephen Chbosky,...	The Perks of Being a Wallflower	1270	http://ia.media-imdb.com/images/M/MV5BMzIxOTQy...	71
2	Mike Lane is a thirty-year old living in Tampa...	Magic Mike	2580	http://ia.media-imdb.com/images/M/MV5BMTQzMDMz...	51
3	When Bond's latest assignment goes gravely wro...	Skyfall	1807	http://ia.media-imdb.com/images/M/MV5BMjAyODkz...	68
4	Against medical advice and without the knowled...	Silver Linings Playbook	1828	http://ia.media-imdb.com/images/M/MV5BMTM2MTI5...	68
...	...	...	...	...	...
3384	Vikram Rathod, A sincere Police man joins with...	Vikramarkudu	1,158	NaN	7.5
3385	After an eventful train trip on the Visakha/Ra...	Visakha Express	43	NaN	4.5
3386	Mahi is a beautiful wealthy heiress, but her f...	Yamadonga	1,316	NaN	7.3
3387	Yevade Subramanyam has wrapped its shooting an...	Yevade Subramanyam	188	NaN	7.9
3388	Sathya (Allu Arjun) returns from a coma to tak...	Yevadu	3,966	NaN	4.6

3389 rows × 9 columns

In [4]:

```python
# To Extract unique values of imdb['Genre'] we use unique()
imdb_df =  imdb_data['Genre'].value_counts().to_frame().reset_index().rename
imdb_df.sort_values('Unique_Genre', ascending=True).reset_index(drop=True, i
```

Out[4]:

	Unique_Genre	count
0	Action	41
1	Action, Adventure	3
2	Action, Adventure, Biography, Drama, History, ...	1
3	Action, Adventure, Comedy, Crime, Musical	1
4	Action, Adventure, Comedy, Drama	1
...	...	...
244	Romance, Sci-Fi	1
245	Sci-Fi	31
246	Sport	14
247	Thriller, Mystery	1
248	War, Drama, Romance	1

249 rows × 2 columns

In [5]:

```python
imdb_data['Genre']      = imdb_data['Genre'].astype('category')
imdb_data['Year']       = imdb_data['Year'].astype('category')
imdb_data['Language']   = imdb_data['Language'].astype('category')
imdb_data['imdbVotes']  = pd.to_numeric(imdb_data['imdbVotes'],errors='coerc
imdb_data['imdbRating'] = pd.to_numeric(imdb_data['imdbRating'],errors='coer
imdb_data['imdbID']     = pd.to_numeric(imdb_data['imdbID'],errors='coerce')
```

In [6]:

```
diff =  imdb_data['imdbVotes'] - imdb_data['imdbRating']
imdb_data['Square'] =  diff**2
imdb_data.head(5)
```

Out[6]:

	Plot	Title	imdbVotes	Poster	imdbRating	Genre	i
0	Despite his tarnished reputation after the eve...	The Dark Knight Rises	2679.0	http://ia.media-imdb.com/images/M/MV5BMTk4ODQz...	75.0	Action, Thriller	
1	Based on the novel written by Stephen Chbosky,...	The Perks of Being a Wallflower	1270.0	http://ia.media-imdb.com/images/M/MV5BMzIxOTQy...	71.0	Drama, Romance	
2	Mike Lane is a thirty-year old living in Tampa...	Magic Mike	2580.0	http://ia.media-imdb.com/images/M/MV5BMTQzMDMz...	51.0	Comedy, Drama	
3	When Bond's latest assignment goes gravely wro...	Skyfall	1807.0	http://ia.media-imdb.com/images/M/MV5BMjAyODkz...	68.0	Action, Thriller	
4	Against medical advice and without the knowled...	Silver Linings Playbook	1828.0	http://ia.media-imdb.com/images/M/MV5BMTM2MTI5...	68.0	Comedy, Drama, Romance	

Observation:

- Here we, can see that the imdbVotes and imdbRation are not in numeric value so, won't able to perform Arithematic operation but after conversion we will able to perform operation.
- We store the difference between imdbVotes and imdbRation in 'diff' variable and create another column having the square of different of the two columns.
- We can clearly, see the Square column which we include in the existing dataset.

Ques 3. Define problem category for below problem statement “A chemist wants to find some interesting patterns in which patients are behaving upon administering the drug”

The chemist type problem statement falls in Unsupervised Learning problem category. In the above problem the chemist will not focus on pre-determined attributes, nor does it predict a target value. Rather, chemist wants to find hidden structure and relation among data.

Unsupervised learning refers to techniques that find patterns in unlabeled data, or data that lacks a defined response measure. In Unsupervised learning, we have less information about objects.

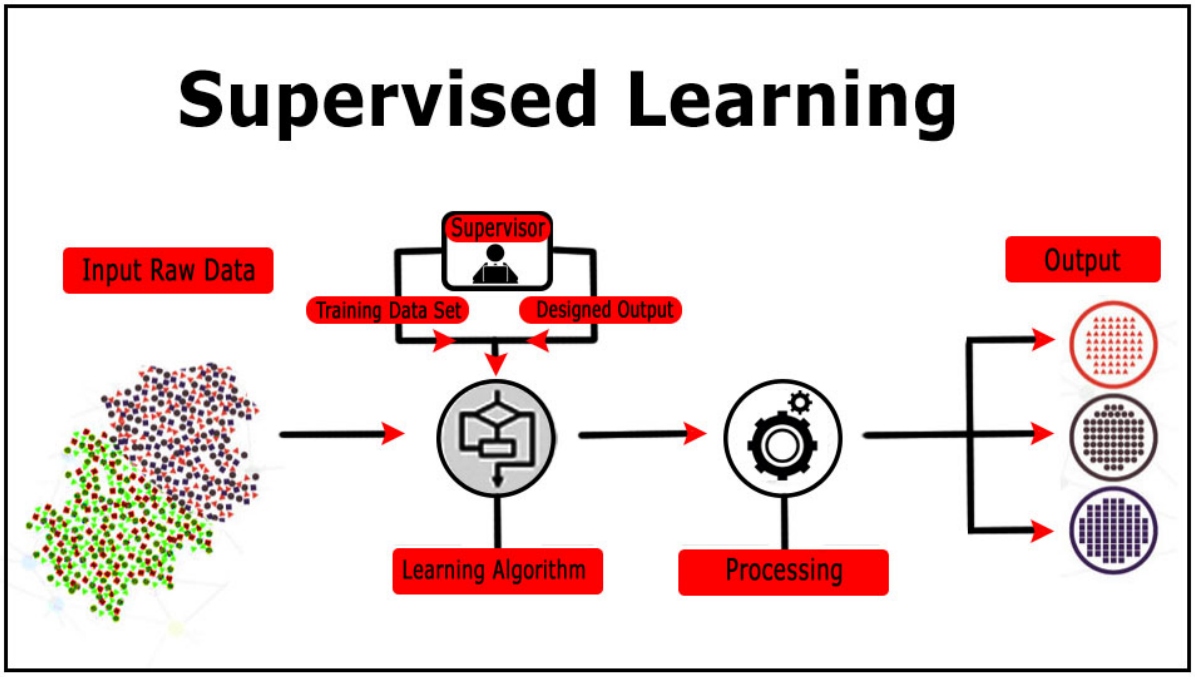
So here we try to find some similarities between groups of objects and include them in an appropriate cluster which is called as segmentation clustering. Some object can differ hugely from all clusters, in this way we can say that these objects are anomalies.

Supervised learning: Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

Supervised learning classified into two categories of algorithms:

Classification: A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.

Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. Unsupervised learning



Unsupervised learning: Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

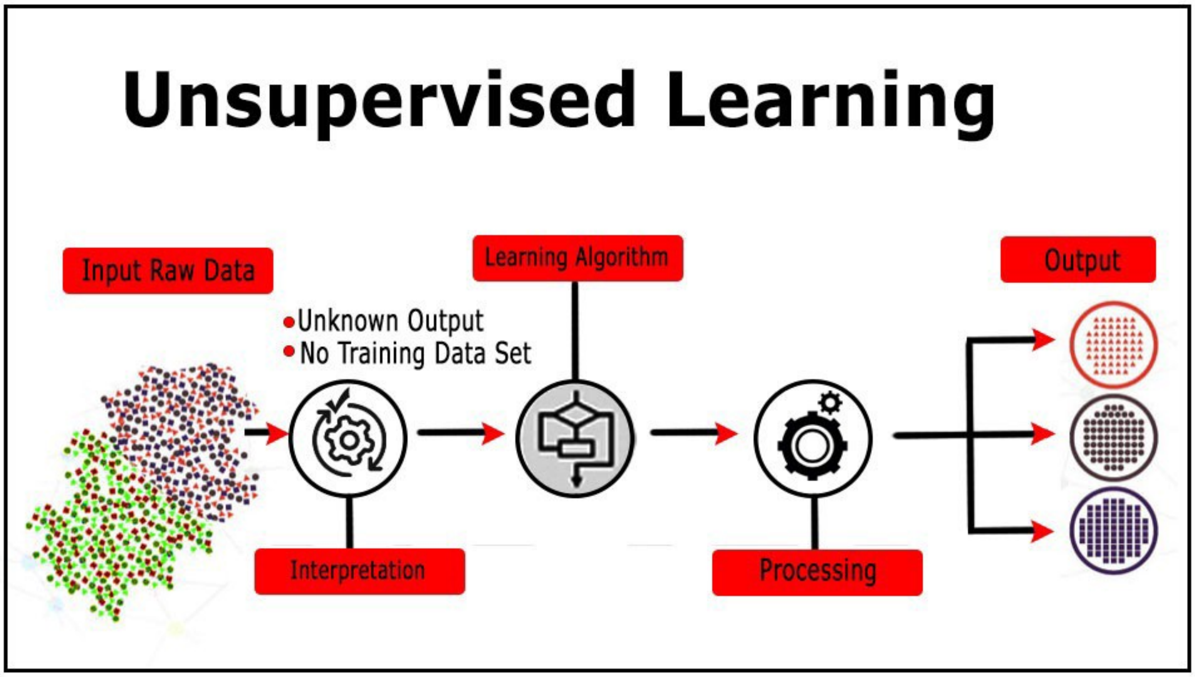
Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore machine is restricted to find the hidden structure in unlabeled data by our-self. For instance, suppose it is given an image having both dogs and cats which have not seen ever.

Thus the machine has no idea about the features of dogs and cat so we can’t categorize it in dogs and cats. But it can categorize them according to their similarities, patterns, and differences i.e., we can easily categorize the above picture into two parts. First first may contain all pics having dogs in it and second part may contain all pics having cats in it. Here you didn’t learn anything before, means no training data or examples.

Unsupervised learning classified into two categories of algorithms:

Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.



Observation: In the above problem the chemist will not focus on pre-determined attributes, nor does it predict a target value. Rather, chemist wants to find hidden structure and relation among data. So the problem is falls under unsupervised Learning.

Ques 4. How will you select suitable machine learning algorithm for a problem statement

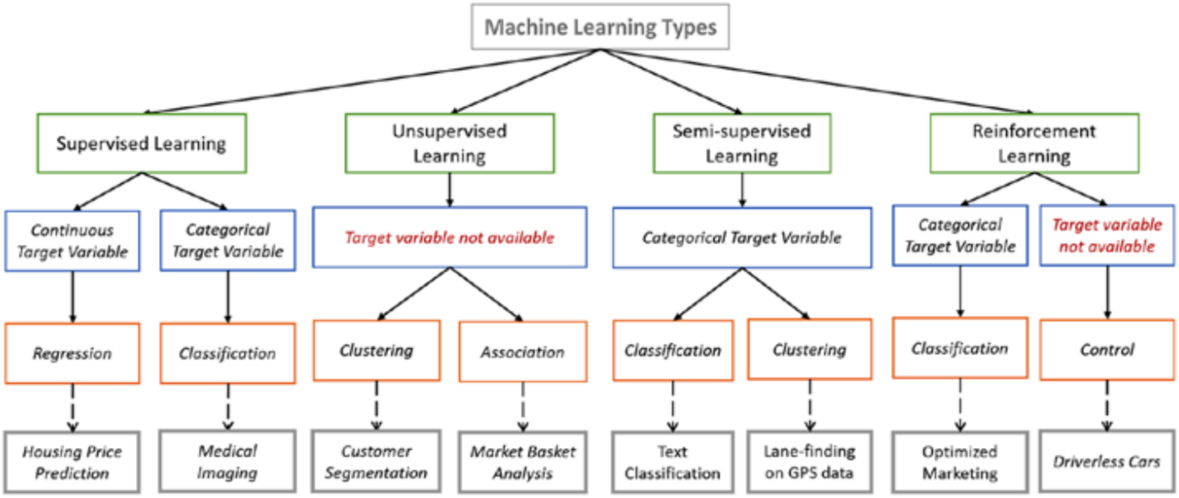
In order to select the Appropriate Machine Learning Algorithm we have to focus on your Domain knowledge, Problem statement, Data Understanding, feature Engineering and business Understanding. We have to focus on Client Business Requirement & needs.

We have to preprocess the data, if there is any unuseful data, missing values or and unuseful content is present we have to filter the useful data from unuseful data. Feature Engineering and Univariate, Bivariate or multi variate, cdf, pdf and other hypertuning techniques also played an important role for understating of data.

1. Data Understanding : The algorithm we use do depend on the data we have. Besides the 'no free lunch theorem', the approach we follow , depends on the data. No machine learning method is really going to completely solve any serious real case problem. The idea is to design an approach that resolves most of the important points and provides the most useful inference about the data so best choices are most often made.
2. tradeoff : Any approach is a tradeoff between accuracy and processing resources. Given infinite resources of time and computational power you can fully solve any problem and its entirely pointless.

3. Noise: We do find noise always. we have to limit it to the greatest possible extent.
4. Pipelines and bottlenecks: Machine learning implementations are typically a pipeline of multiple modules wrangling with data. Based on the data and the desired learning objectives, select algorithms for each of these modules that work well together and with each other.
5. Scaling: build a prototype quickly and test for subset of data and if things are looking promising, forge ahead, otherwise try something else. The reason for this is that data is almost never what you think it is.

Feature descriptors do not actually provide what they are designed to provide, and consequently a given machine learning algorithm, which seems appropriate in theory, will perform very poorly on the actual data. It's a good idea to try different related ML algorithms and see how they perform with scalability and other metrics like accuracy, robustness, computational cost, stability, etc. and then move forward with the algorithm that seems best suited to the data.



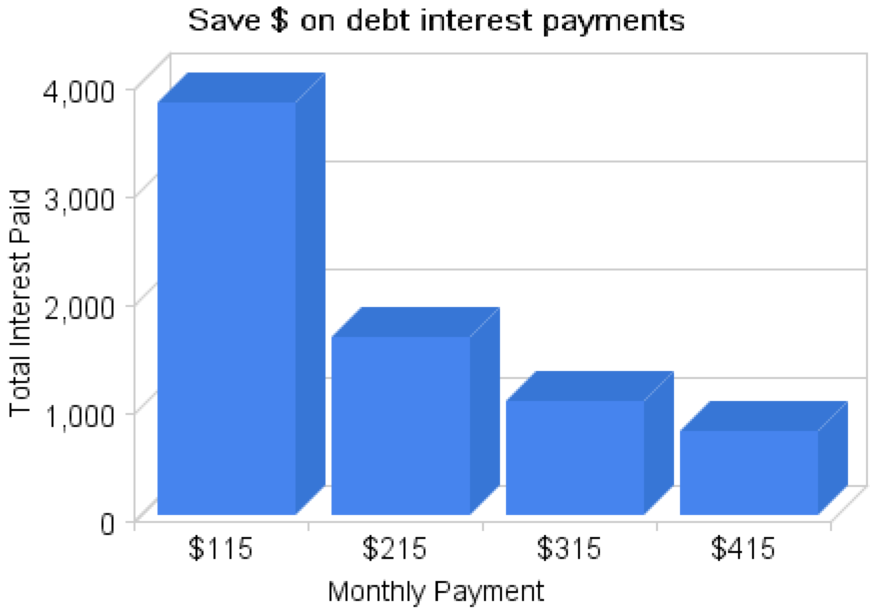
Ques 5. Define one problem statement for Education industry?

Problem Statement: Universities have started facing increasing pressure to beat the inflation on tuition fees, the burden of student cumulative debt money, outstanding credit card debt that is to be collected and rejection of state funds.

Solution: Predictive analytics, basically follows a wide range of statistical methods from predictive modeling, machine learning, artificial intelligence and data mining that evaluate and analyze the facts from the present as well as the past to make predictions about future or any unknown events.

Predictive analytics is used in Customer relationship management (CRM), healthcare, cross selling, fraud detection, risk management, direct marketing, underwriting, etc. In the field of education, the desire for stronger actions and efficient operations is leading colleges and universities to alter their methods of operation.

These situations demand well-organized business practices to be more efficient in the competitive market.



Since the online systems like Learning Management System and Student Information System are the driving functions in educational institutes, a colossal amount of data about students, workers, sponsors, and graduates get stored during each and every transaction that is being done.

This series of online analysis applications goes beyond reporting and monitoring.

This data can further be used in analysis and estimation of the future. Institutes today make decisions based on these estimation results that are produced based on events already occurred. Later the actuals are compared with the estimated results to see if the improvement exists as expected. The common areas of improvement are

- Student performance
- Student retention
- Effective management of enrollment
- Institutional progression
- Management of financial aid
- Campus security enhancement

Working: Predictive analysis provides perceptions on events even before it occurs. Hence the outcome of the event can be influenced. For example, Google map uses GPS and location information about our current location and destinations and suggests the best possible route and all possible alternate routes to reach the destination safer and faster.

In a similar way, the data of the student is used to analyze the factors that can affect the student from reaching the goals and deadlines. The progress of the student is estimated at the initial point itself.

Any hurdle that may cross the way can be arbitrated and eradicated in a proactive manner so as to helps the students attain the goals with ease. This whole process seems tedious and tough, but is helpful.

Reason: The reason why any educational institute should go for predictive analysis depends and varies from institute to institute. But the common expectation is to have a foreseen vision at each and every step of this entire process. Hence the path to be taken is known and understood beforehand and it helps in meeting the objectives of this entire process. The end results are very efficient and powerful in terms of students and faculty management.

Retention of students: Predictive Analysis is used to identify the factors that lead to the success and failure of academics results. This helps in retaining the students through graduation.

Management of enrollment: The prediction of the future helps in targeting the optimal mix of students to be enrolled in the institute. It also helps in bringing down the enrollment costs and improving operation efficiencies of the institute.

Finance management: The entire process of financial aid is optimized as the future is predicted based on the current financial situation of the institute. This also helps in detecting and preventing any potential fraudulent claims that may occur.

Growth of the institute: Predictive analysis also helps in predicting potential donors and estimates the amount of finance the institute may receive.

Security of the institute: Since the current data can be used, predictive analysis is also used in predicting the probability of any dangerous or criminal activities that may occur for which the institute can proactively take preventive measures.

Some More aspect where Machine Learning are useful:

1. Grading students

Machine learning can grade students by removing human biases. Some recent examples are use of the supervised Machine Learning for text classification to predict students' final course grades in some course and exhibited the potential of using ML classified messages to identify students at risk of course failure. In addition, there are aims to improve the assessment of problem solving in education by employing language technologies and computational-statistical machine learning methods to grade students' natural language responses automatically. Great example of use machine learning for grading students is by comparing their actions to a model of expert behaviour.

2. Improving student retention

As we said before, by identifying "at risk" students early, schools can detect and contact those students and help them to be more successful. Student retention is an essential part of many enrolment systems. It affects almost all segments of university or school metrics: reputation, financials, ranking. Specially, student retention has become one of the most important things for managers in higher education institutions. There are few studies, which developed models to predict and to introduce the reasons behind student's number decreasing.

3. Predicting student performance

Probably a major benefit of machine learning (regarding number of studies in scientific databases) is its ability to predict student performance. By "learning" about each student, the technology can identify weaknesses and suggests ways to improve, such as additional practice tests. This seems to be very hot research trend; there are lot of studies in recent years in this area, as we said before. For example, study employs the machine learning approach called the Recursive Clustering technique to group the students of the programming course into groups based on their performance in the prerequisite courses, co-requisite and current course work result. Students present in the lower groups will be taken into consideration since they are highly prone to fail. In another interesting study in this category, authors have proposed a new model to categorize students into three categories to determine their learning capabilities and to help them to improve their studying techniques. They have chosen the state of the art of machine learning approach to classify student's nature of study by selecting prominent features of their activity in their academic field. They have chosen a data driven approach where key factors that determines the base of student and classify them into high, medium and low ranks. Yet another study in this category, but from other perspective brings

4. Testing students

The machine learning based assessment provides constant feedback to teachers, students and parents about how the student learns, the support they need and the progress they are making towards their learning goals. Authors in study introduced a system for training of students' ability to construct correct proofs in propositional or predicate logic. In addition to common techniques, including presentations supported by slides and exercises they used animations, which were based on carefully selected demonstrative examples and their step-by step solutions. In order to test students' knowledge, they prepared a questionnaire that captured the entire process of a logic proof construction. A student constructed a proof and then answered questions from the questionnaire. They described the design of the questionnaire and discussed its dis/advantages. At the end, they then applied frequent subgraph mining together with supervised machine learning algorithms to perform an automatic evaluation of correctness of the proofs.

Conclusion

The aim of this study was to evaluate the current state of the art in the application of machine learning in education area. The amount of studies (papers and articles) was large, so only some of studies, which we found as good representatives, were mentioned in results this study.

This study shows that there are significant different ways to benefit from machine learning application in education area. As we stated in introduction section, one of our goals was try to classify studies in the field of machine learning application in education area. Based on our survey, the papers reviewed under category marked as A research a ways how machine learning can grade students by removing human biases (fairly grading) Reviewing studies under category marked as B, showed how machine-learning algorithms can help schools or faculties to reach out to students and get them the help they need to be successful as early as possible. Student retention is an essential part of many enrolment management systems.

It affects university rankings, school reputation, and financial wellbeing. Student retention has become one of the most important priorities for decision makers in higher education institutions, so there are lot of studies in that category. Reviewing studies under category marked as C, showed us how major benefit of machine learning (regarding number of studies in scientific databases) is its ability to predict student performance. By “learning” about each student, the technology can identify weaknesses and suggests ways to improve. According to our survey, this is most interesting area of machine learning application to researchers.

There are lot of studies in recent years in that category, and lot of machine learning models were provided to predict student performance on different parameters. We would say that this category is definitely the trend. Reviewing studies under category marked as D, showed some models how machine learning can help move away from standardized testing. Machine learning based assessment provides constant feedback to teachers, students and parents about how the student learns, the support they need and the progress they are making towards their learning goals. As we have found earlier, research has been made over several relevant databases, but of course, not all were involved, so this can be considered as limitation of study. In addition, there is a possibility that some of the relevant studies may be skipped by chance. In the future, we plan to implement own machine learning model for suggesting potential student to enrol or not to enrol on University College algebra, Study of Software Engineering, based on different parameters. As we have rich database with lot of information of students on previous years, we believe that study would be of help to support our admission office as help in student enrolment process.