IMPLEMENTING LEARNING ANALYTICS INTERFACE FOR MOOCs

by

VIGNAN VENNAMPALLY

Supervisor (s):

External:

Prof. TV Prabhakar
Computer Science and Engineering
IIT Kanpur

Internal:

Prof. Aparajita Ojha
Computer Science and Engineering
IIITDM Jabalpur.

ACKNOWLEDGMENT

It has been a precious and productive time for me to work at Media Labs , IIT Kanpur. I will always be grateful to all for creating this opportunity and for your support, help and encouragement which will always be reminded in my life.

I take this opportunity to express my profound gratitude and deep regards to Prof. TV Prabhakar (Professor, IIT Kanpur) who has guided me in all aspects. Also, I would extend my gratitude to Mrs .Revathy KT (Project Engineer, IIT Kanpur) for her constant support, help and motivation.

I would also like to take advantage of this opportunity to express my profound gratitude and thankfulness to Prof. Aparajitha Ojha (Professor, PDPM IIITDM Jabalpur) for creating me this wonderful opportunity. I perceive this opportunity as a big milestone in my career development and a great experience.

Date :- 12/11/2019 **Vignan Vennampally**

DropOut Prediction in MOOC’s

1.Introduction:

Problem Statement:

One of the current issue with all MOOC managing platforms is dropout of users from a particular course. Since this MOOC’s are free of cost users dropout from the platform without intimating to the instructor. As the completion rate of MOOC’s is very small it is very important to predict the dropouts in order to increase the retention rates.

If the instructor of a particular course can predict the dropouts, he can send interventions in the form of emails, messages etc and can understand the reasons for the dropout. So that he can take necessary measures like improving the course content, improving the study material etc..

I have gone through a lot of literature work to know about the previous works and contributions in this domain which used clickstream data as a measure to predict dropouts. Since there are many reasons why a particular students leaves the platform many research papers concentrated on determining the factors that mostly contribute to the dropout. Along with Predicting DropOut if the model can predict the probability of DropOut ,we can send interventions quickly to those users having higher probability as they are going to Dropout quickly when compared to other users.

After thorough research it has been found that 3 factors contribute a lot in predicting the dropouts

1. Lecture Views
2. Quiz Performance
3. No of absent days.

Solution:

If we could construct a Machine learning model which takes input('features') as clickstream data and predict the output as Binary ('DropOut' or Not) this issue can be solved. The probability of DropOuts can be predicted by using ML algorithms like SVM, Decision Tree. The major challenge here is to construct a Dataset with both 'features' and 'lables' intially to train the model.

2.Data Set:

1. For this model the data of a course **'Special Theory of Relativity'** containing 18,502 users is used. The mookit_certificates table of MySQL database consists of 'Lecture Percentage', 'Quiz Percentage' columns . 'No of absent days' is determined from the LastLogin(or Lastaccess) data present in UserInfo collection present analytics_relativity database of Mongo dump.

2. The 'Lecture Percentage' and 'Quiz Percentage' of all the users are directly obtained by converting mookit_certificates table present in MySQL database to Csv.

Note:

- My SQL Database can be converted to csv file from MySQL Workbench directly by exporting the data.
- MongoDB command to export data to csv file is:

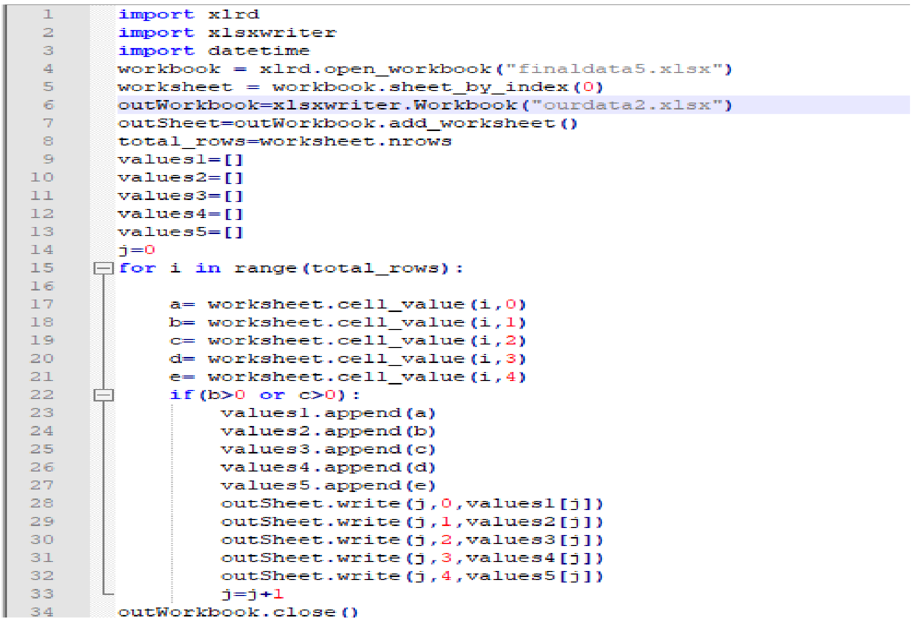
```
mongoexport --db analytics_relativity --collection userInfo type=csv fields
uid,LastLogin --out filename.csv
```

3. The course duration was 8 weeks starting from **'19 Dec 2019'** to **'08 March 2019'**. We assumed to be predicting this model on **'03 March 2019'.** Therefore a threshold of 14 days is selected i.e users whose Last Login date is before **'19 Feb 2019'** are considered as DropOuts.

Note: This threshold value of 14 days is selected after going through many research papers which clearly explained the reason and there is no compulsion to follow the same rule and can be changed as per the needs.

4. The entries having '0' timestamp as lastLogin are removed as they act as outliners. However this does not affect the model prediction because there are only a few users having good LecturePercentage, QuizPercentage and '0' as their LastLogin. And we can Indirectly consider these users as Dropouts.

5. Data is further Processed by considering only those entries having either ‘Lecture Percentage’ or ‘Quiz Percentage’ greater than ‘0’. Because users having ‘Lecture Percenatge’ and ‘Quiz Percentage’ equal to ‘0’ are not considered for prediction as they are DropOuts Indirectly.



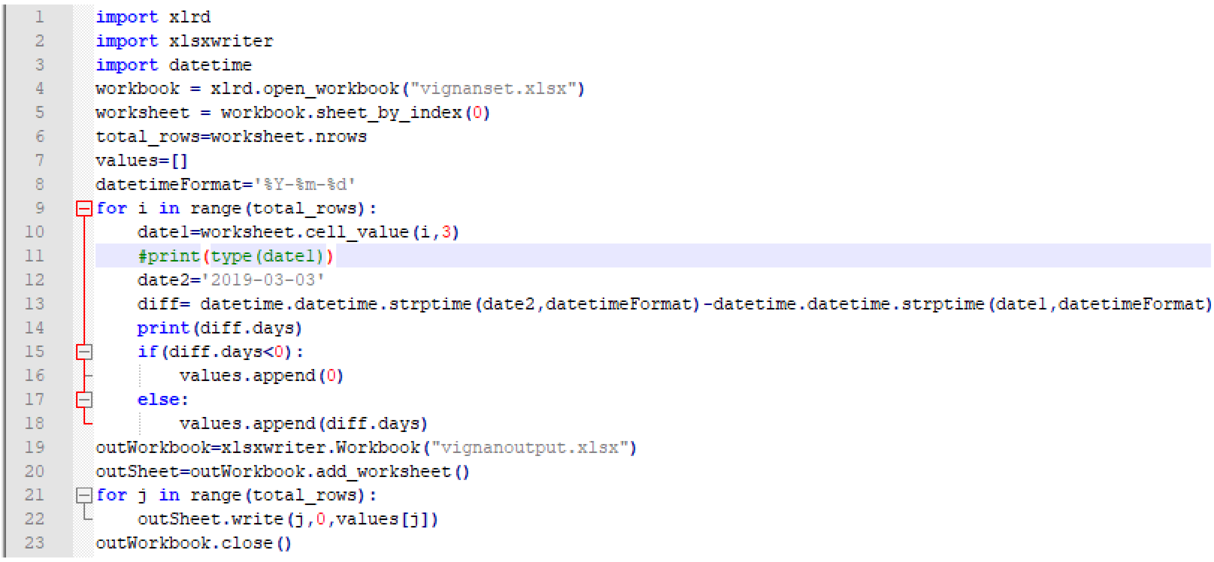
```
import xlrd
import xlsxwriter
import datetime
workbook = xlrd.open_workbook("finaldata5.xlsx")
worksheet = workbook.sheet_by_index(0)
outWorkbook=xlsxwriter.Workbook("ourdata2.xlsx")
outSheet=outWorkbook.add_worksheet()
total_rows=worksheet.nrows
values1=[]
values2=[]
values3=[]
values4=[]
values5=[]
j=0
for i in range(total_rows):

    a= worksheet.cell_value(i,0)
    b= worksheet.cell_value(i,1)
    c= worksheet.cell_value(i,2)
    d= worksheet.cell_value(i,3)
    e= worksheet.cell_value(i,4)
    if(b>0 or c>0):
        values1.append(a)
        values2.append(b)
        values3.append(c)
        values4.append(d)
        values5.append(e)
        outSheet.write(j,0,values1[j])
        outSheet.write(j,1,values2[j])
        outSheet.write(j,2,values3[j])
        outSheet.write(j,3,values4[j])
        outSheet.write(j,4,values5[j])
        j=j+1
outWorkbook.close()
```

6. Users who are still active but having poor ‘Lecture Percentage’ and ‘Quiz Percentage’ are considered as DropOut only. Clearly ‘Lecture Percentage’ and ‘Quiz Percentage’are given importance compared to ‘No of absent days’ of a user.

7. If any of the ‘Lecture Percentage’, ‘Quiz Percentage’ is quite low and the other is a high value that particular user is not a dropout as some students watch only Lectures while some attend Quizzes directly and score good marks.

8. Python code is written in order to convert the ‘ LastLogin’ present in timestamp to ‘No of absent days’ compared with date of model prediction. In this way all the 3 ‘features’ required for our ML model are created. The ‘No of absent days’ feature consists of No of absent days of each user from the date of Model prediction.



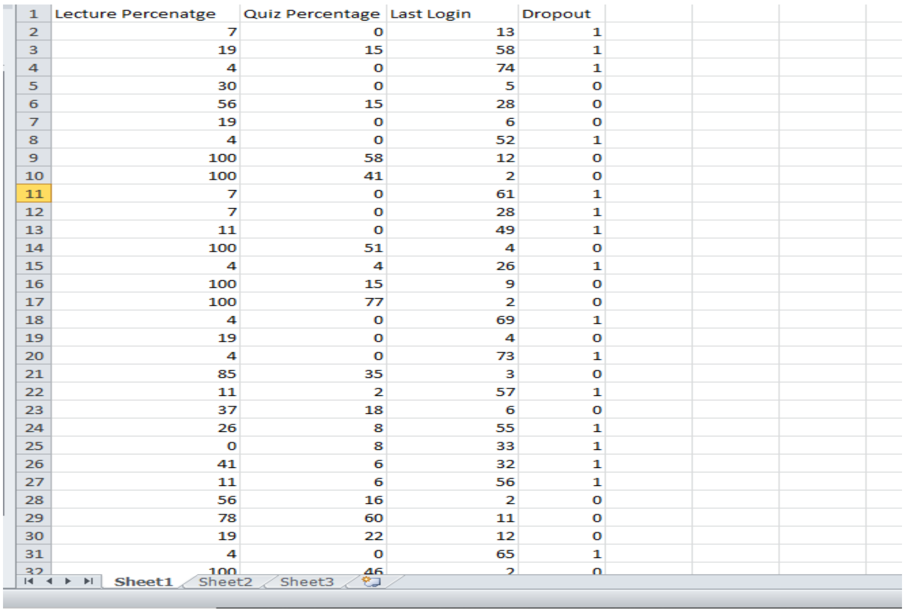
```
import xlrd
import xlsxwriter
import datetime
workbook = xlrd.open_workbook("vignanset.xlsx")
worksheet = workbook.sheet_by_index(0)
total_rows=worksheet.nrows
values=[]
datetimeFormat='%Y-%m-%d'
for i in range(total_rows):
    date1=worksheet.cell_value(i,3)
    #print(type(date1))
    date2='2019-03-03'
    diff= datetime.datetime.strptime(date2,datetimeFormat)-datetime.datetime.strptime(date1,datetimeFormat)
    print(diff.days)
    if(diff.days<0):
        values.append(0)
    else:
        values.append(diff.days)
outWorkbook=xlsxwriter.Workbook("vignanoutput.xlsx")
outSheet=outWorkbook.add_worksheet()
for j in range(total_rows):
    outSheet.write(j,0,values[j])
outWorkbook.close()
```

9. For any Machine Learning model to be trained both ‘features’ and ‘labels’ are required initially to train the model. But in this Dataset ‘labels’ data is not available.

10. We usually Represent Dropout by ‘1’ and Not a Dropout by ‘0’. If we can determine the users who had completed the course successfully from the certificates data, we can determine whether a particular user is DropOut or Not. But for this course, certificates are only give to the users satisfying a particular attendance criteria and Quiz Percentage criteria. So based on the values of each ‘feature’ the user is assigned ‘labels’ as ‘1’(DropOut’) or ‘0’(Not a Dropout). This assigning of labels is done manually as we are not using any particular threshold for every ‘feature’ .(**Note:** This assigining of ‘labels’which is done manually is initially required to train the model not every time we use the model for prediction).We are assigning this label as ‘1’or ‘0’ based on the 3 features values. For the ‘No of absent days’ feature we have considered a user as DropOut if he is absent since 14 days from the date of DropOut prediction.

11. This framing of dataset is very important initially to construct a Machine Learning model and it took good amount time in the whole process.

The Dataset after proper formation looks like below:



1	Lecture Percenatge	Quiz Percentage	Last Login	Dropout
2	7	0	13	1
3	19	15	58	1
4	4	0	74	1
5	30	0	5	0
6	56	15	28	0
7	19	0	6	0
8	4	0	52	1
9	100	58	12	0
10	100	41	2	0
11	7	0	61	1
12	7	0	28	1
13	11	0	49	1
14	100	51	4	0
15	4	4	26	1
16	100	15	9	0
17	100	77	2	0
18	4	0	69	1
19	19	0	4	0
20	4	0	73	1
21	85	35	3	0
22	11	2	57	1
23	37	18	6	0
24	26	8	55	1
25	0	8	33	1
26	41	6	32	1
27	11	6	56	1
28	56	16	2	0
29	78	60	11	0
30	19	22	12	0
31	4	0	65	1
32	100	46	2	0

3.Model:

1. The DropOut prediction model should have two basic features:

i. Determining the DropOut users correctly.

ii. Sending Interventions as quickly as possible.

If we can determine the probability of dropout of each user we can send Interventions quickly to those students having higher probability of dropout.

2. The DropOut prediction is an classification problem as the output is binary. In this approach we will train our Dataset on different Machine Learning models to find out which models suits to our Dataset best.

3. For any ML model to avoid bias and data leakage both 'PreProcessing' and 'cross-validation' must run in a parallel manner. **'pipeline'** technique is used for this purpose. A 10-fold Cross-validation is done to avoid overfitting of data.

4. Since the class variables of this dataset are not balanced in number **'AUC'**(Area under ROC) is used to determine the best fit ML algorithm. AUC can decrease the standard error compared with other traditional metrics like Precision, Recall, F1 score.

5. Using Algorithms like SVM, DT the probabilitiy of DropOut can also be predicted. When the respective ML models are trained and applied on test data the AUC scores are as follows:

ALGORITHM	AUC Score	Avg AUC Score (50 iterations)
Logistic Regression	0.97134	0.97735
SVC linear Kernel	0.97682	0.97893
SVC Rbf Kernel	0.98381	0.97955
DecisionTree Classifier	0.96955	0.97068
Randon Forest Classifier	0.976811	0.97780
KNeighbors Classifier	0.98522	0.97750
NaiveBayes Classifier	0.97365	0.97565
XGBoost Classifier	0.97995	0.97945

6. Each time we run the model different data is trained, tested as we have used 10-Fold Cross-Validation and so AUC scores changes accordingly for each run.

7. The avg AUC scores of 50 runs is calculated to find the best fit model. From the above table data it has been found that SVC(both linear and rbf) are giving the best AUC scores. Also SVC can be used to predict the probability of dropouts.

8. This model is used for prediciting Dropouts of **'Detection, Diagnosis and Management of plant diseases'** course containing 5,533 users started on October 15,2019 and the results along with their probability of dropouts are attached with this file.

4.Conclusion:

With the available resources I tried to build the best model satisfying all the conditions of the users. I have taken measures that during training and testing of data there is no data leakage. Dataset formation from MongoDB, MySQL dump is an important aspect in the whole process as data should be carefully processed and outliners should be removed if needed.

THANK YOU