COURSE RECOMMENDATION SYSTEM

Amulya Mysore	Dharani Sakary Pirangi	Sujitha Metla	Saiteja Sirikonda
1212263102	1211203627	1211198544	1211246826
amysore2@asu.edu	dsakaryp@asu.edu	smetla@asu.edu	ssirikon@asu.edu

INTRODUCTION

Recommendations are ubiquitous helping in day-to-day activities. With the increase of content in World Wide Web and knowledge-building happening in every possible way, retrieving relevant information by users has become a challenging task. Recommendation systems play a vital role in processing the data and providing content useful for the users. Recommendations for users can be provided by looking at previous activity of user and finding interesting information (Content-based Filtering) or by looking at users similar to him/her (Collaborative Filtering).

As technology advances alongside real world problems, there is need for people to learn relevant techniques to solve problems. Universities are therefore providing these new advances in technologies and useful courses to students. However, universities need some quantitative background information on providing resources to students. Also, students in their first year may not have clear understanding on which career path to choose based on their skill. Existing Recommendations approaches cannot be directly applied to learning environments as they are characterized by specific characteristics such as learning style, skill, interests of the user. Therefore, there is a need for background knowledge specific to this domain for recommendations to be effective.

The Course Recommendation System we developed recommends resources to be allocated by universities to accommodate needs of students, course streams that can be considered by students to excel in their field of interest, skill, recommendations to future students based on statistics of current students.

PROBLEM DESCRIPTION

With the increase in the population and globalization, there has been an increase in the awareness for technology and innovation among people thereby causing an increase in the demand for education. In the past decade there has been an enormous rise in the number of times people searched for technologies on search engines like Google. On an average it has been reported that the keyword " New Technology" has been searched 90500 times in a day[1]. As new technologies continue to evolve day by day, the number of people interested in exploring and thereby impacting the global community have increased at an enormous scale.

Over the past decade there has been a exponential rise in the number of students who are looking for new learnings or knowledge sharing mediums. To satisfy these enormous requests, several cross discipline programs have been introduced through either universities or online courses termed as MOOCs - Massive Open Online Courses. The statistics of the enrollments are quite a figure. Here are some of them.

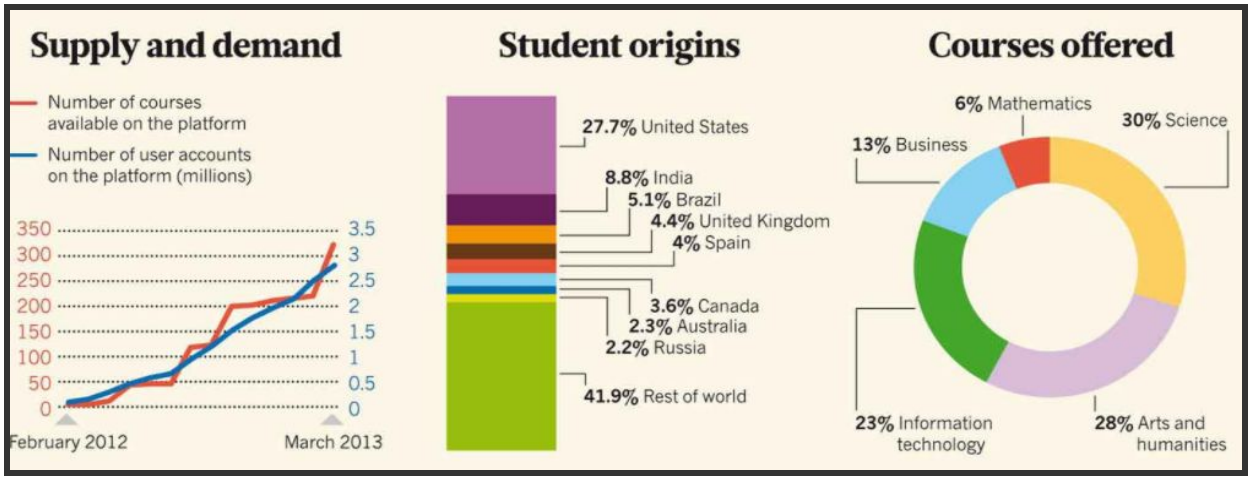


Figure 1 - MOOCs[2]

Coursera is one of the leading companies in Mountain View California offering MOOC's on a global scale has reportedly introduced 328 different courses from 62 universities in 17 countries. The platform has 2.9 million registered users dispersed among 220 countries across the globe.

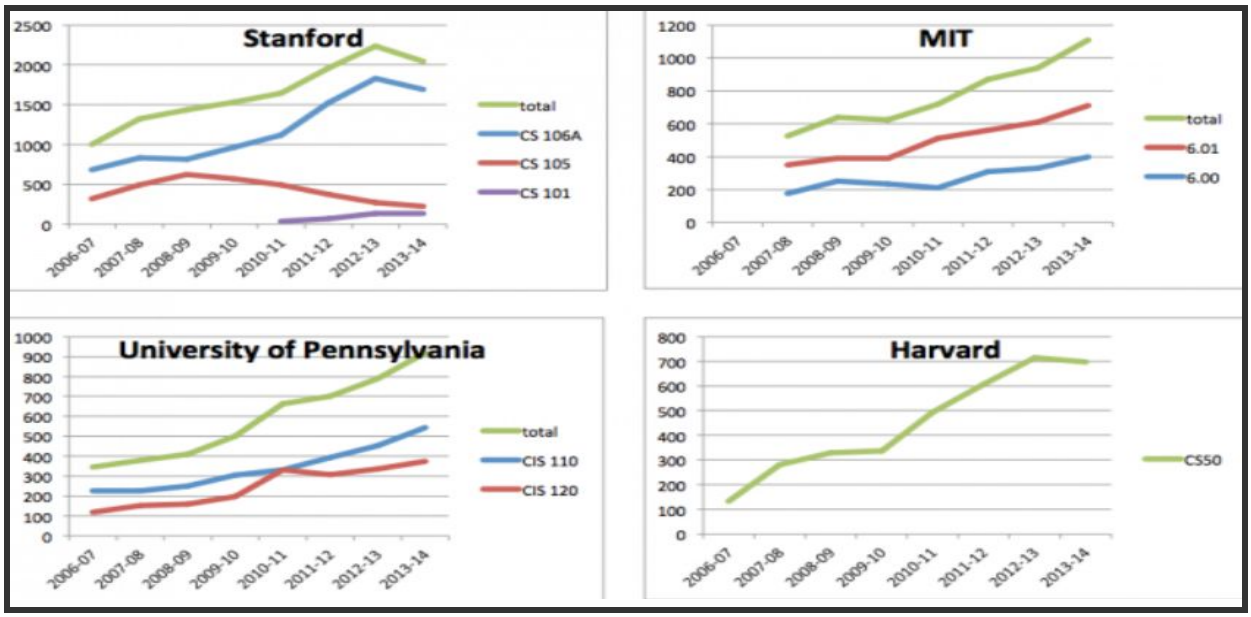


Figure 2 - Increase of Enrollment in universities[3]

It can be inferred from the above figure that there has been an significant hike in the number of programs offered by various universities to accommodate the new talent.

Even though there is a rise in the number of offered courses , there is no metric which lets the student know in which field he would excel based on his previous academic performance and also there is no available metric to calculate the degree of satisfaction of the student after taking a course in a particular domain which may provide a room for improvisation.

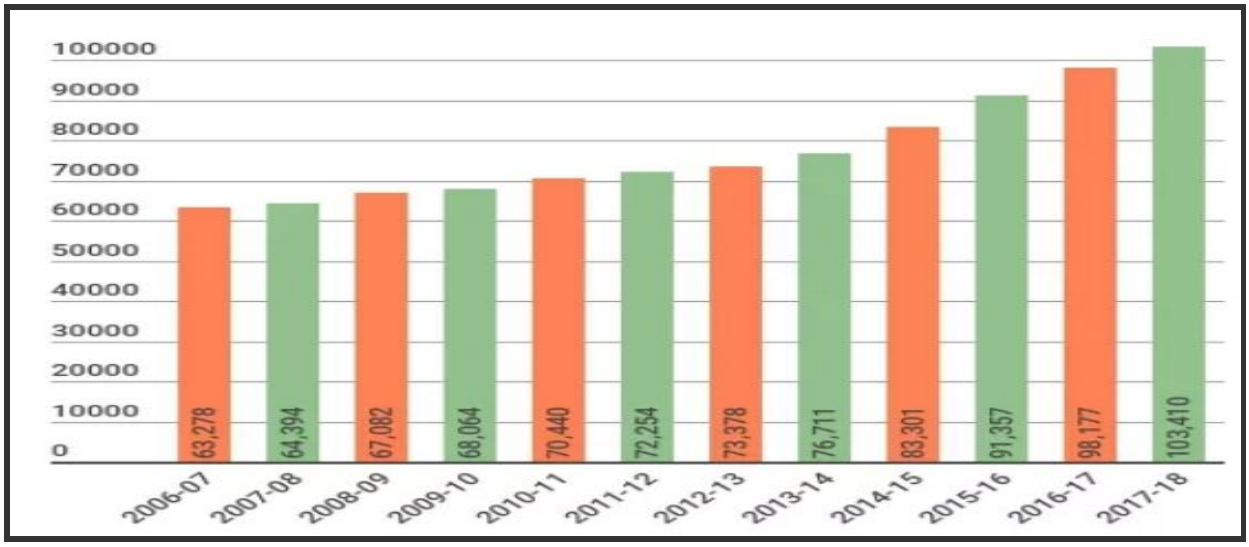


Figure 3 - Rise of Enrollment in ASU[4]

Our model helps to make recommendations to the universities on whether or not they should increase/decrease the number of seats of a particular course in the upcoming semester based on the performance of the students and their satisfaction in the present semester. It also tells the university about the level of difficulty it can introduce in a particular course based on the performance of the current students. It also recommends the students their future career path based on their previous academic performance. Courses will also be recommended to students based on other similar student's performance.

METHODOLOGY

Dataset

In order to implement the model, we have used the data from the "HarvardX-MITx Person-Course Academic Year 2013 De-identified Dataset, Version 2.0"[5]. This dataset contains the data of the first year students. These data are aggregate records, and each record represents one individual's activity in one course. The dataset contains the following attributes, of which few relevant attributes have been identified and used to calculate a satisfaction metric.

1. course_id: ID to identify a course
2. userid_DI: Random Unique ID to identify a student
3. registered: 1 if a student is registered for a course, otherwise 0
4. viewed: 1 if student has opened the course, otherwise 0
5. explored: 1 if student has gone through the course, otherwise 0
6. certified: 1 if student has passed the course with at least grade 50%, otherwise 0
7. final_cc_cname_DI: regional origin of the student
8. gender: gender of the student
9. grade: final grade in the course, ranging from 0 to 1
10. start_time_DI: date of course registration
11. last_event_DI: date of last interaction with course
12. nevents: number of interactions with the course based on tracking logs
13. ndays_act: number of unique days student interacted with course

14. nplay_video: number of times video relevant to course is played
15. nchapters: number of chapters with which the student interacted
16. nforum_posts: number of posts to the Discussion Forum

Data Cleaning

The data obtained from the dataset was incomplete and had some data (invalid) which is not useful for decision process. So the following steps have been performed to clean the data before analysis.

Handling Invalid Data

1. The records with the attribute Grade having values as zeros, were considered as invalid data. Hence these records were removed from the dataset in the cleaning phase.
2. The records with the attribute Grade having values as blank cannot be used for analysis, hence, these records were removed from the dataset in the cleaning phase.
3. There were some records where user had taken course twice and was not certified in both attempts. These were considered duplicate records and have been eliminated. Unique data are identified with UserId and certified attributes.
4. There were several records in which the attribute last_event_DI that signifies the last time the user has interacted with the course, has a value which is lesser than the value of the attribute start_event_DI that signifies the value of the first time the user has interacted with the course. These were assumed to be cases where student interacted with course before start date and did not enroll. Such records were eliminated.
5. There were some records which have blank value for the attribute last_event_DI which gives the last time the user has interacted with the course. These records have been removed, since it depict that the user hasn't interacted with the course at least once after registering.

Handling Incomplete Data

The attribute nplay_videos represents number of play video events within the course. This attribute has been identified with blank values due to the instances where students didn't watch video. These were replaced with 0.

Data Analysis

Feature Deduction

Based on the given features from the dataset, start_event_DI, value of the first time the user has interacted with the course and another feature last_event_DI, value of the last time the user has interacted with the course, dimensionality reduction was performed using a mathematical operation and a new feature has been deduced named 'total_days' which represents the total time invested by the student to finish a course.

$$total\_days = last\_event\_DI - start\_event\_DI + 1$$

Correlation Between Grade and other Attributes in the DataSet

A linear regression model has been developed in R to find if the attribute has a positive or a negative impact on the Final Grade of a student. Based on the significance of the model obtained, the best model was considered by removing certain attributes that did not have any

impact on the grade. The final model obtained is as follows

grade vs (*nevents*, *ndays_act*, *nplay_video*, *n chapters*, *nforum_posts*, *total_days*)

```
Call:
lm(formula = grade ~ nevents + ndays_act + nplay_video + nchapters +
    nforum_posts + total_days, data = filteredData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07936 -0.10210 -0.02523  0.04952  0.93525

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.557e-02  1.412e-03  -46.43   <2e-16 ***
nevents       1.686e-05  5.831e-07   28.92   <2e-16 ***
ndays_act     4.077e-03  7.633e-05   53.41   <2e-16 ***
nplay_video  -3.536e-05  2.353e-06  -15.03   <2e-16 ***
nchapters     2.203e-02  1.562e-04  141.08   <2e-16 ***
nforum_posts -2.213e-02  1.412e-03  -15.68   <2e-16 ***
total_days    3.850e-04  1.283e-05   30.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2149 on 68933 degrees of freedom
  (451 observations deleted due to missingness)
Multiple R-squared:  0.6076,    Adjusted R-squared:  0.6076
F-statistic: 1.779e+04 on 6 and 68933 DF,  p-value: < 2.2e-16
```

Figure 4 - Linear Model of grade vs attributes

It can be observed from the above model constructed that all the attributes considered here have a very high significance. The attributes ndays_act, nevents, nchapters and total_days have a positive coefficient. This means that with an increase in the value of these attributes, there will be an increase in the Grade. The remaining attributes have a negative coefficient, thus with an increase in their value there will be a decrease in the value of the Grade. However, when plotted against other attributes independently, nforum_posts does not have any correlation with the grade. Therefore, this attribute is ignored.

Data Distribution

We have plotted the values of Different attributes, in order to see their Distribution. The Distributions are as follows :

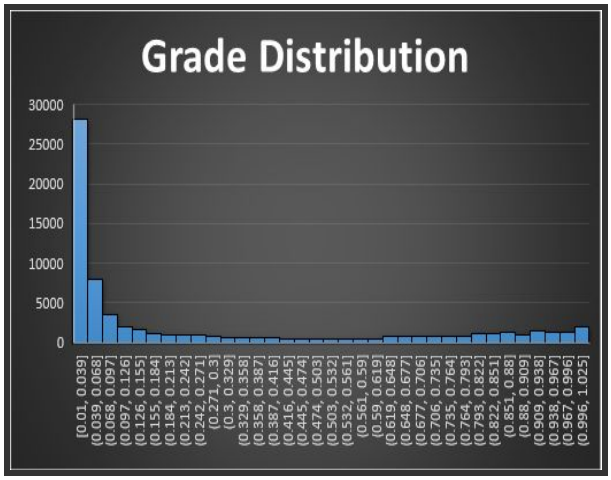


Figure 5 - Grade Distribution

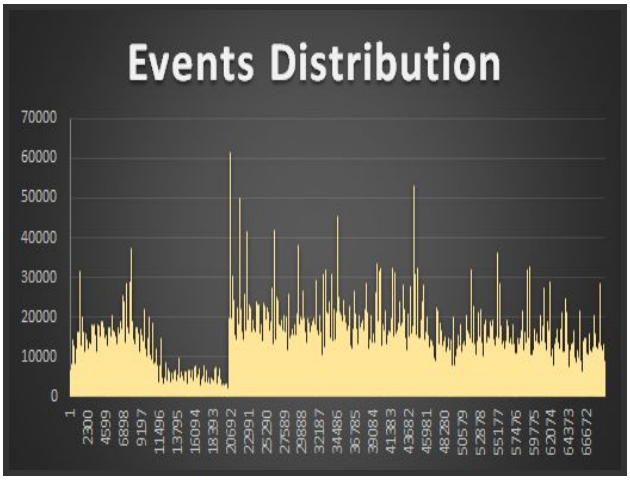
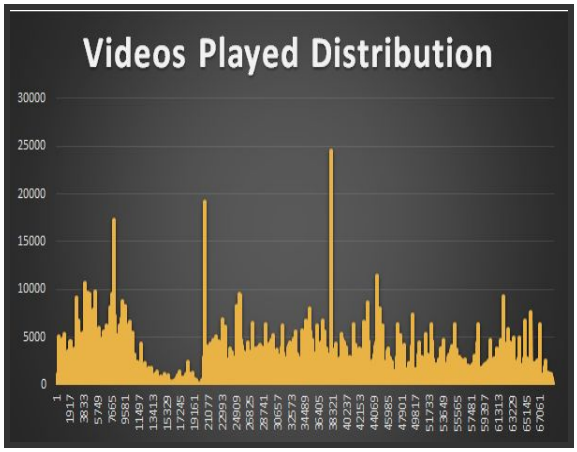


Figure 6 - Events Distribution

The Grade Distribution from Figure 5 shows the maximum number of students scored a grade between [0.01-0.039]. From the figure it can be observed that most of the population has been distributed among the first three bins, which if considered doesn't add any value to our analysis.

The events attribute follows a normal distribution of data as seen in Figure 6, with the maxima at the value 20692. Most of the inconsistent records with invalid values for start_event_DI (having negative timestamp), nevents fall into these bins and hence handled accordingly to promise fair distribution.



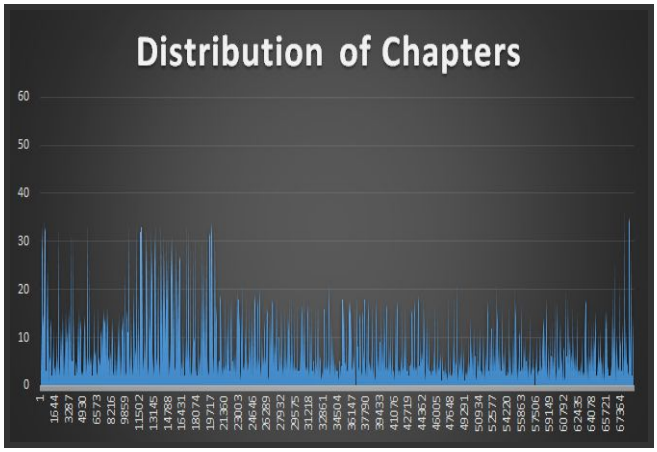


Figure 7 Videos Played Distribution *Figure 8 Chapters Distribution*

Nchapters feature which describes the number of chapters in a particular course contributes to an uniform distribution seen in Figure 8.

nplay_video feature which corresponds to the number of times a video has been played has a negative coefficient in our linear regression model from which we can infer that this may be an indication that the course material is of higher difficulty. Distribution of nplay_video can be seen in Figure 7.

New Feature Deduction, Satisfiability Metric

Satisfiability metric, measured per student, depicts the scale on which a student is satisfied with either the course structure or his performance and other attributes.

$$Satisfiability_{per\ day} = \frac{nevents + ndays + nchapters - nplayvideos}{total\ days} + (100 * grade)$$

We came up with this formula based on the results from the Linear Regression Model we built. We decided to measure the Satisfiability metric per day because each student, for a course in the same semester, has different start_dates and end_dates and some students continue to study the subject beyond the length of the semester, thereby making it difficult for us to quantify it.

We scaled the Satisfiability metric to lie between 1 and 1000, because we wanted to limit the maximum value of the Satisfiability metric in order to handle some anomalous data, where a student would take the course for just one day, making the denominator 1 and end up completing hundred's of events. Thereby leading to an absurdly large satisfiability metric. Then, the values of the Satisfiability metric is normalized to lie in between 0 and 10.

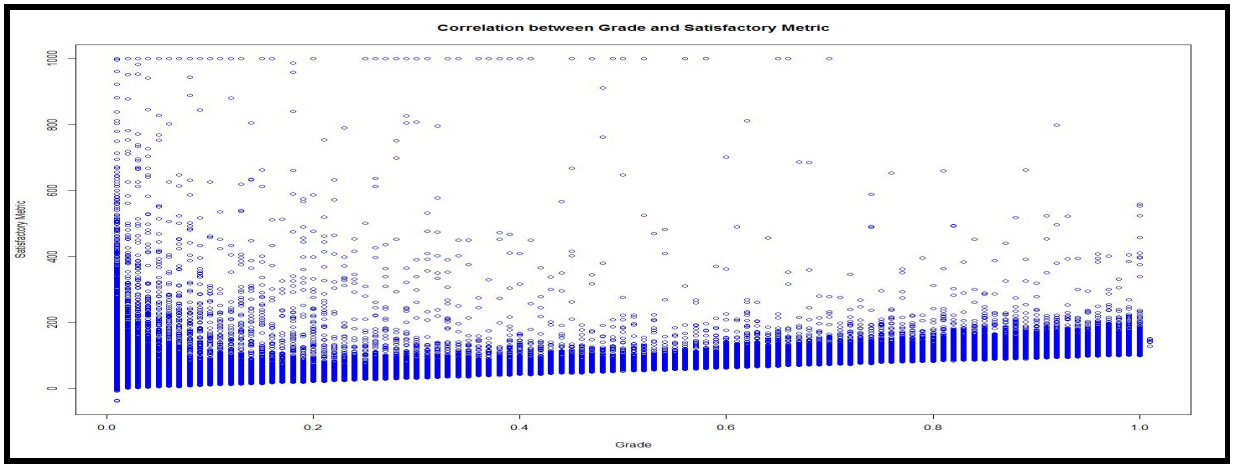


Figure 9 - Grade vs Satisfiability Metric

We noticed from the Figure 9, on plotting the correlation between the final attribute grade and our satisfiability metric, hence, we can conclude that our Satisfiability Metric fared pretty well and we made sure that the computation in Task - 1 and Task - 2 would work and give the required recommendations even if the definition of the Satisfiability Metric changes. The Satisfiability Metric has been scaled to get the values between 1 and 1000 and from weighing Student's surveys (traditional method) and then coming up with the right thresholds would be more informative to both Students and Universities.

OUR WORK:

As part of the Course project, we implemented three major recommendation systems that would be beneficial for both University/Institute and Students together.

Task 1:

For a course offered in a particular semester in a University, based on the average Satisfiability metric of the students enrolled in the university, we can recommend if the University should improve the quality and resources of the course. Universities could follow this procedure to filter out the courses, which are not upto the mark and make efforts to improve by making some changes, such as, change the curriculum of the course to include more topics in demand, change the instructor and if its satisfaction metric is too low, that course can be dropped from the list of courses offered by the university. Similarly if the course is on the higher side of the Satisfiability metric, the course needs more promotion to attract students, increase the number of TA's and allot more resources for the course.

Broadly, based on the average Satisfiability metric of the students enrolled in a particular course, we came up with a threshold:

1. If the course has a Satisfiability metric greater than the Threshold, then we can give the following suggestions:
   - Increase the number of enrollments in the course
   - tougher course material/make it more challenging and fun learning experience for the students
   - Provide for video recording and making them accessible to public (YouTube etc.)
2. If the course has a Satisfiability metric less than the Threshold, then we can give the

following suggestions:

- Change course Instructor
- Reduce the difficulty of the course
- Increase the number of TA's and recitation sessions for the course, etc.,
- Increase the Student - Instructor interaction mediums, for example, use of Piazza.

In order to come up with the Threshold, based on which we make the decision, we took the average Satisfiability metric of students for course and took the 60th percentile of the resulting distribution.

Task 2:

In this task, we provide recommendations for a student, on what major to choose based on his/her performance in their freshman year of college. We also provide courses in which he/she performed poorly compared to the rest of the class as part of self evaluation. We assumed that, the recommendations will be more effective for students in the top 95th percentile and the bottom 40th percentile in all the courses.

To summarize, For a particular student, based on the courses he took and his *Satisfiability metric*:

1. Recommend the subjects, he/she can choose to major in.
2. Provide subjects he/she performed poorly for self-evaluation.

For a user, we recommend the majors and improvement areas, for which his/her Satisfiability metric lies in the **95th percentile** or above and if it lies in **40th percentile** or below respectively.

Task 3:

In this task, we recommended courses to students based on the performance of the previous students. Since, we had limitations finding the perfect dataset, with a separate test set and training set, we decided to do K-Cross validation on the existing dataset and calculate the accuracy using the F1 measure as shown below.

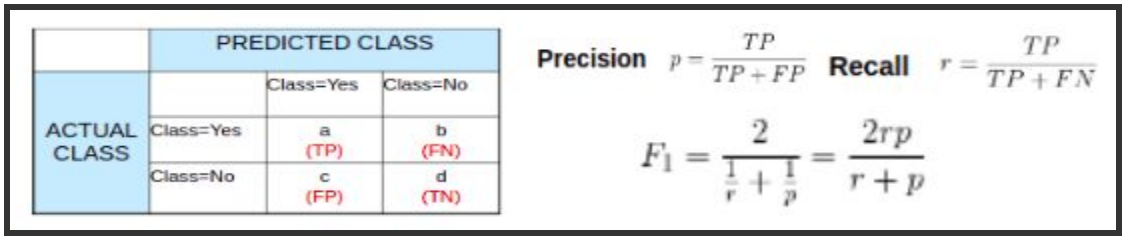


Figure 10 - Confusion Matrix, Calculation of F1 measure using precision and recall

We decided to use Collaborative Filtering with Nearest Neighbors and the similarity metric we used to compute the nearest neighbor is Euclidean Distance. The similarity decreases as the Euclidean Distance increases, therefore making it the Anti-Similarity Metric. We researched with other similarity metrics like, Pearson Correlation Coefficient, Spearman Correlation Coefficient, Getis-Ord Similarity metric and Cosine Similarity Metric and from [6] we found that in most cases, the Euclidean Distance gives the least accuracy than all the other similarity metrics (Since, the distance is very susceptible to comparatively high values, if they are not normalised). We decided to find the minimum accuracy and postulate that, given accurate data,

our algorithm is going to perform better than the accuracy obtained from Nearest Neighbor calculation with Euclidean Distance as the Similarity Metric (anti).

To perform the recommendation, we took all the numeric fields in the dataset, plus, the number of days between Start_date_DI and the Last_date_DI and gave to our Recommendation machine as the input. We considered K = {2k + 1, for k ∈ [0,6]} Nearest Neighbors and ran it through 5-fold & 10-fold Cross Validation averaged the accuracy attained for each fold and plotted the final accuracies for each K.

RESULTS

Task 1

In the task 1 mentioned in the above section with the help of our satisfiability metric we came up with the threshold, we took the average of the Satisfiability metric of students in each course, and took the 60th percentile. A graph was plotted to find out the course satisfaction.

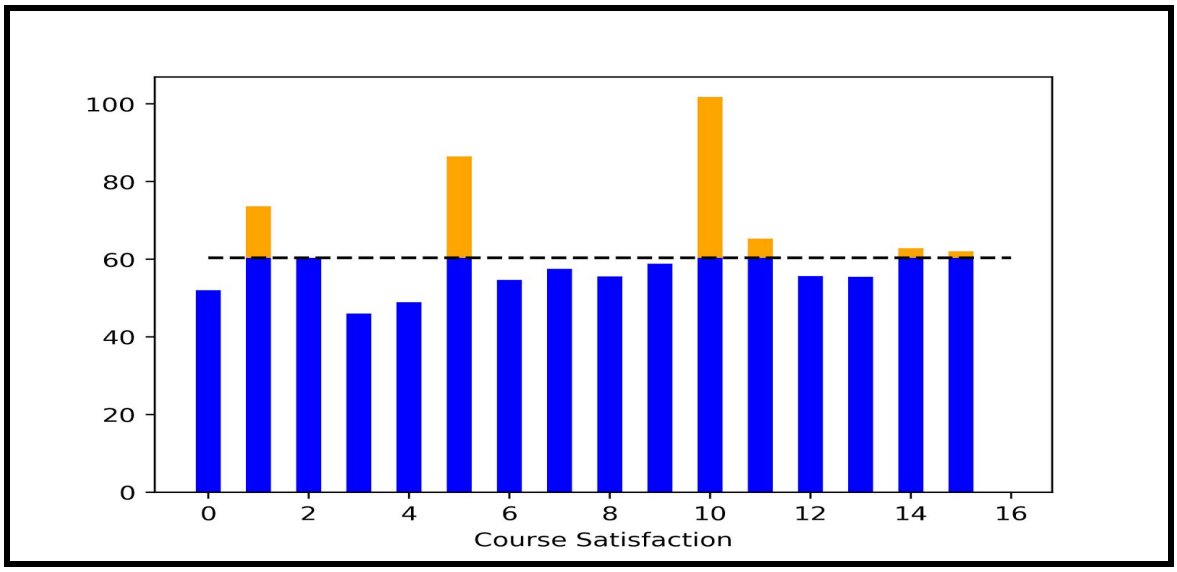


Figure 11 - Course Satisfaction with threshold at 60th percentile

The dotted line in the above graph represents the threshold value. It can be inferred from the above graph that all the courses with values above the threshold value have a very high value of Course Satisfaction. But some of the subjects whose course satisfaction is below the given threshold need improvement. Hence it can be deciphered that the university should think about ways on improving the course such as redesigning the course structure or changing the course instructor.

Task-2

In the Task 2, plotted the distribution of grades of students in each of the subjects present in our dataset. In Figure 12, the 95th percentile and the 40 percentile have been plotted. If the student's satisfiability lies in the 95th percentile for a particular subject, then we recommend that he can consider continuing with the specific domain or major. If the student's satisfiability lies in the 40th percentile, it can be deduced that the specific subject might be a wrong fit for him/her and he/she should re consider when choosing a major.

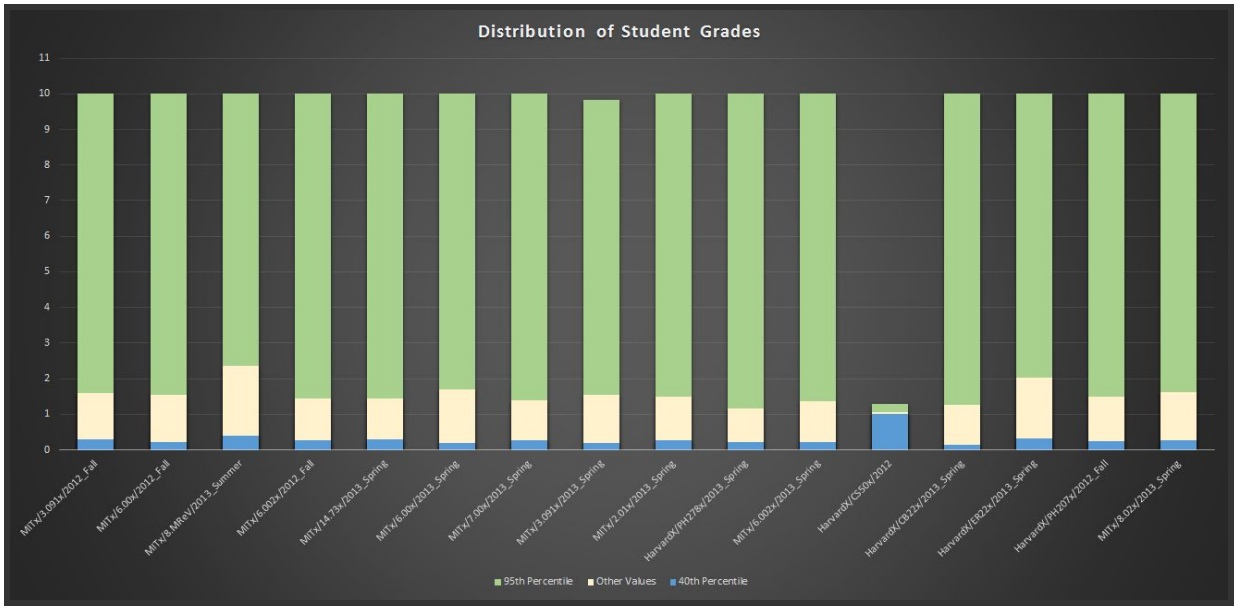


Figure 12 - Student satisfiability with 95th percentile recommending students to continue

Task 3

In task 3, we have implemented the collaborative filtering approach using the K-Nearest Neighbours algorithm. Since there was no separate test set, we have implemented K-fold cross validation with the K values 5 and 10. During the implementation of the K-Nearest Neighbours algorithm we considered all the odd values ranging from 3 to 13. The figure below shows the Accuracy of our algorithm for various values of K, highest of 74.8% at K = 9.

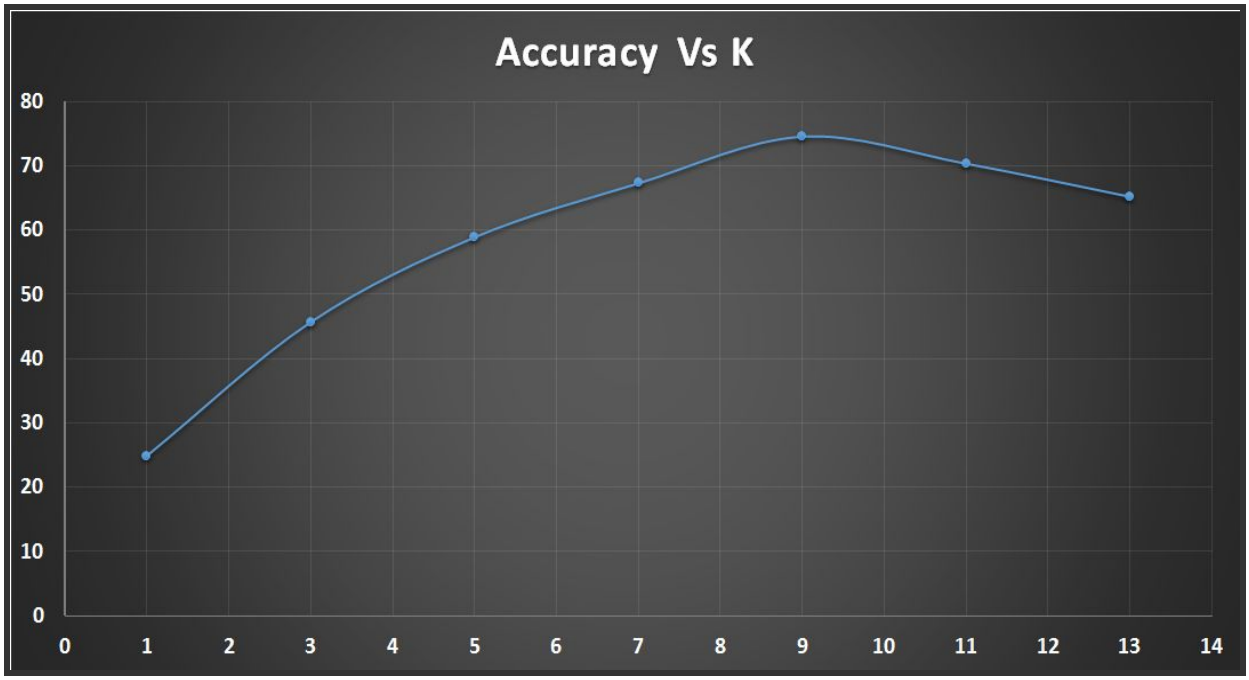


Figure 13 - Accuracy Vs K neighbors

CHALLENGES FACED

As part of the project, some of the challenges we faced are as follows:

1. Figuring all the inconsistencies in the data (dates etc.,). Since, we had a huge dataset, the inconsistencies in the dataset were not obvious at first-look. We had to figure them out as we went on with the implementation and come up with a suitable way to handle it.
2. Handling the inconsistencies of the Dataset. We had to come up with a series of very subtle Data Cleaning steps to handle these inconsistencies on the go.
3. Coming up with a Satisfiability measure, which serves the purpose of rating and making sure it was valid. This was probably the biggest challenge of them all. After trying out a few combinations, we fixed on the one mentioned above. The signs were given based on the weights achieved in the Linear Regression built in Figure 4.
4. We had to perform Multivariate Linear regression to see what all attributes contribute for the grade. We wanted to know what numerical attributes from the given list (nevents, ndays_act, nvideos etc.) actually influence the grade achieved the student and by how much.
5. Calculating the accuracy of a recommendation system. Though there are various measures used to calculate the accuracy of a recommendation system like Mean Square Error (MSE), Mean Average Precision (MAP), Root Mean Squared Error (RMSE) etc. we went with the F1 measure as it is closer to the larger value among Precision and Recall [12].
6. Coming up with the right visualizations to represent the data. Apart from all the calculations, we also had to figure out what visualization correctly depicts the distribution of the data. We used, Scatterplots, bar graphs, line graphs and 3D histograms as a part of the project.
7. Coming up with the thresholds for Task - 1 & 2 etc., In order to take the decisions about the quality of the courses, we had to assume some thresholds, which can be changed to their actual values based on explicit student inputs.

CONCLUSIONS

As a requirement for any recommendation system, we needed a way in which each student can rate the course he/she has taken based on his performance. Since, it was not implicitly given in the dataset, we decided to calculate the rating using a Satisfiability metric formula we came up with as a part of the project. The formula was decided based on the Linear Regression model in Figure 4 and we used the Satisfiability metric in Task - 1 and Task - 2.

As a part of the project, we implemented 3 tasks:

1. Task 1 - For a University
   We recommend if a given course is upto the mark as the University standards. If Yes, we suggest some measures to keep its standards high such as, increasing the number of seats, increasing the number of TA's or recitation sections etc. If No, then we come up with recommendations for the betterment of the course such as the Changing the instructor, revising the curriculum, improving the assignments & projects in the group etc. We do this by taking the average Satisfiability metric of each course and coming up with a threshold to decide if the course is upto the mark or not.

2. Task 2 - For a Student
   We recommend the major (Career Advice) a student should choose, based on the courses he took and his performance in the respective courses and We also provide the courses he performed poorly compared to the rest of the class based for self evaluation. We considered students in 95^{th} percentile and above & students in 40^{th} percentile and below for this task.
3. Task 3 - For a new Student
   We provided course recommendations based on the student data from previous semesters. We used Collaborative Filtering with Nearest Neighbor and Euclidean Distance as the Similarity Metric (anti). We considered $K = \{2k + 1, \text{for } k \in [0,6]\}$ Nearest Neighbors and ran it through 5-fold & 10-fold Cross Validation averaged the accuracy attained for each fold and plotted the final accuracies for each K.

FUTURE WORK

One of the major challenges of our project was to come with a ratings measure for each student enrolled in a course as it was not given in the dataset. Usually, the Satisfiability metric is computed by giving weights in elaborate surveys, which was infeasible for our project and so we had to come up with our metric and justify it. Our recommendation system can be made much more accurate and robust if we somehow had the ratings given in the dataset. This could be done by scraping course data and asking students to finish the survey for sufficient number of courses over a period of 2 or 3 years for more accuracy.

The other difficulty was the amount of information given in the dataset we had, even though it was the best dataset available for online to capture the student performance in a course at a university, the dataset was still incomplete. Our recommendation system would've been more powerful if we ontological details of each student like learning speed, age, gender, skill level etc. which increases the accuracy of Collaborative Filtering by 30% [7] or give the entire transcript of each student during their time at the university to perform Hybrid - Recommendations, Knowledge - based recommendations etc.

CONTRIBUTIONS

Name	ASU ID	Contributions	Contribution%
Amulya Mysore	1212263102	Literature Review, Data Analysis,Multivariate Linear Regression Task 1, Task 2, Documentation	25
Dharani Sakary Pirangi	1211203627	Data Cleaning, Data Visualization, Correlation, Documentation, Task 1, Task 3	25
Saiteja Sirikonda	1211246826	Data Analysis, Feature Extraction, Task 2, Task 3, Presentation	25
Sujitha Metla	1211198544	Literature Review, Data Analysis, Task 3, Task 2, Feature Extraction, Accuracy, Documentation	25

ACKNOWLEDGEMENT

We thank our Prof. Hanghang Tong for giving us this opportunity to learn about various types of Recommendation systems such as Collaborative Filtering, Content Based Filtering, Hybrid Recommendation System, knowledge based recommendation system etc and their latest trends.

We also had the opportunity to learn Data Analysis in R and Excel, learned Python libraries like numpy, scikit, pandas, matplotlib etc. and learn various Machine learning techniques as a part of the project.

REFERENCES

1. https://www.wordstream.com/popular-keywords/technology-keywords
2. https://www.pinterest.com/pin/394698354817044630/
3. https://www.geekwire.com/2014/analysis-examining-computer-science-education-explosion/
4. https://www.azcentral.com/story/news/local/arizona-education/2017/09/13/asu-enrollment-hits-more-than-100-000-first-time/646653001/
5. MITx and HarvardX, 2014, "HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0", doi:10.7910/DVN/26147, Harvard Dataverse, V10; HMXPC13_DI_v2_5-14-14.csv [fileName]
6. Wu, FaQing, et al. "An effective similarity measure for collaborative filtering." *Granular Computing, 2008. GrC 2008. IEEE International Conference on*. IEEE, 2008.
7. Tarus, John, Zhendong Niu, and Bakhti Khadidja. "E-Learning Recommender System Based on Collaborative Filtering and Ontology." *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 11.2 (2017): 225-230.
8. http://users.csc.calpoly.edu/~dekhtyar/466-Fall2010/lectures/lec11-1.466.pdf
9. https://www.nature.com/news/online-learning-campus-2-0-1.12590
10. http://www.data-mania.com/blog/recommendation-system-python/
11. https://medium.com/@m_n_malaeb/the-easy-guide-for-building-python-collaborative-filtering-recommendation-system-in-2017-d2736d2e92a8
12. https://www.quora.com/How-can-I-measure-the-accuracy-of-a-recommender-system/answer/Alvin-Thai-2?srid=p2R6
13. https://www.coursera.org/learn/recommender-systems-introduction