Capstone I - Final Report :
Predicting the registrant success (rate of EdX (HarvardX / MITx) Courses

Sukanya Chandramouli
April 2018

Introduction

This report provides a summary of my work on the Capstone project (modeling the completion rates of EdX Course Registrants). A joint research team from Harvard and MIT released the dataset containing details of course-by-course patterns of student interactions with courses from Harvard and MIT during the first year of edX. This dataset is available here

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147&version=10.0

Project Goal

The goal of this project is build a model that can predict course completion rate of the registrants using the historical data. Such an analysis would be help in tracking the progress and also in providing early interventions to enable the participants to successfully complete the course.

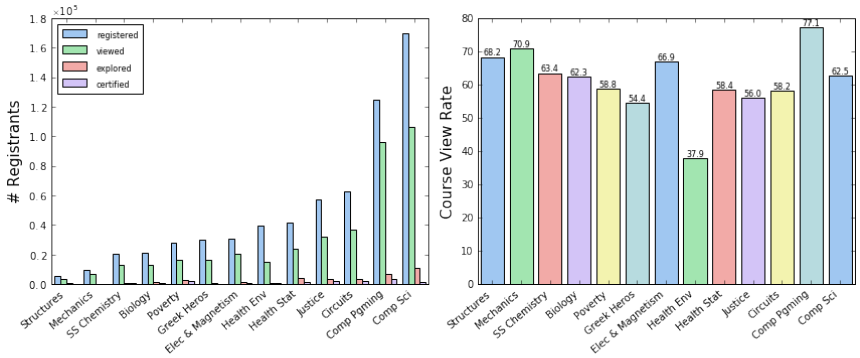
Tools Used

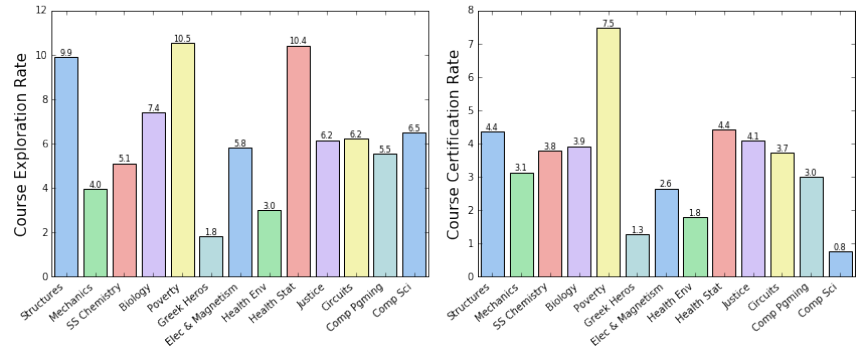
Pandas	Data Wrangling and Manipulation
Scikit learn	Classifiers – Logistic regression , Random Forest , GridSearch
Scikit learn Metrics	Metrics - Classification Report , Confusion Matrix , Precision-Recall Curve, ROC Curve, cohen_kappa
Matplotlib & Seaborn	Data visualization
Imbalanced Learn	SMOTE

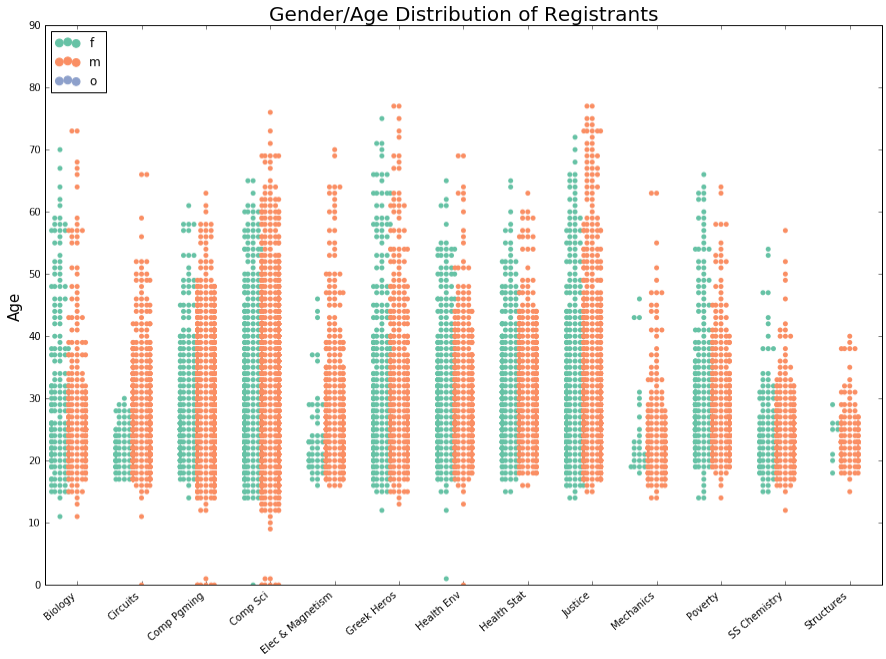
Data Wrangling & Exploratory Data Analysis

With over six hundred thousand entries in the dataset, there are roughly 17000 registrants who have completed the course. The dataset does not capture the intent of the registrants – whether the user plans to complete the course certification or plans to get access to the course/teaching material. Given that more than half the participants did not even click through the material, special considerations are to be made while interpreting the data.

Before building the prediction model , we need to clean the dataset for processing. The entries in the dataset which have the user information (year of birth, education, gender) incomplete are filled with median values (by course).Composite fields are split to get individual details such as institution, course ,semester and year. Finally the NaN/incomplete entries in the administrative data are filled with zeros .







Observations

From the preliminary data analysis and visualization we see that 2-10% of the registrants have completed fifty percent of the course material. The number of users completing the course certification varies from 0.8% to 7%. Computer courses have the highest enrollment rate and lowest certification rate. Enrollment in Computer Courses make up more than 40% of the dataset.

Users with secondary, bachelors and masters degree have a higher enrollment and certification rate.

Prediction/Classification Model

The dataset includes both demographic information provided by the user (such as the date of birth, educational background etc) and the administrative information (course interaction details of the registrants) provided by the institutions. The administrative information has aggregate statistics such as the the number of interactions, chapters reviewed, grades at the end of the course.

Administrative Data

- *Course and User identification*
- *Viewed,explored indicates user has viewed the content or completed 50% of course*
- *country / region of the user*
- *number of interactions with the course*
- *number of days student interacted with course*
- *number of play video events within the course*
- *number of chapters completed by the student*

User Data

- *level of education of the user*
- *year of birth.*
- *gender*

The dataset has the final grade of the user and certified field to indicate the user has completed the course certification. In this project we use the 'certified' variable to build a binary classifier. Though there are several different machine learning models for the binary classification, I have considered the logistic regression and Random Forest model to build the classifier.

Logistic Regression

I started building the prediction model using logistic regression. Given the huge size (over 640000 entries) with an imbalanced class (about 17000 course/certification completions) the dataset has been split to 50% training and 50% test.

For the first prediction (logistic regression) model , I have used ONLY the demographic information (age , education level etc) to predict the course completion rate.

For the second prediction model , I have used only administrative information (provided by the institutions which include the course interaction details) to predict the course completion rate.

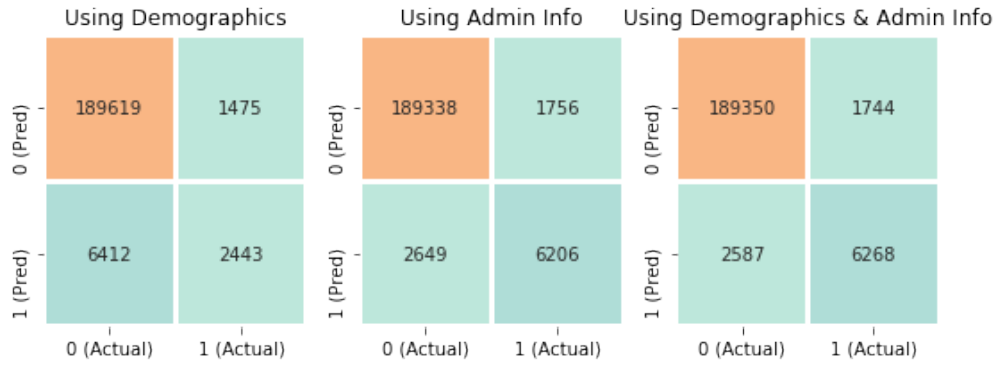
For the third prediction model , I have used both demographics and administrative information to predict the course completion rate.

Since we are dealing with in an imbalanced class , instead of looking of looking at general accuracy as metric , we look at the True Positive, True Negative, False Positive, False Negative rates and measures based on these to evaluate the performance of the model

Precision = True Positive / (True Positive + False Positive)
Recall = True Positive / (True Positive + False Negative)

f1-score = 2 * Precision * Recall / (Precision + Recall)

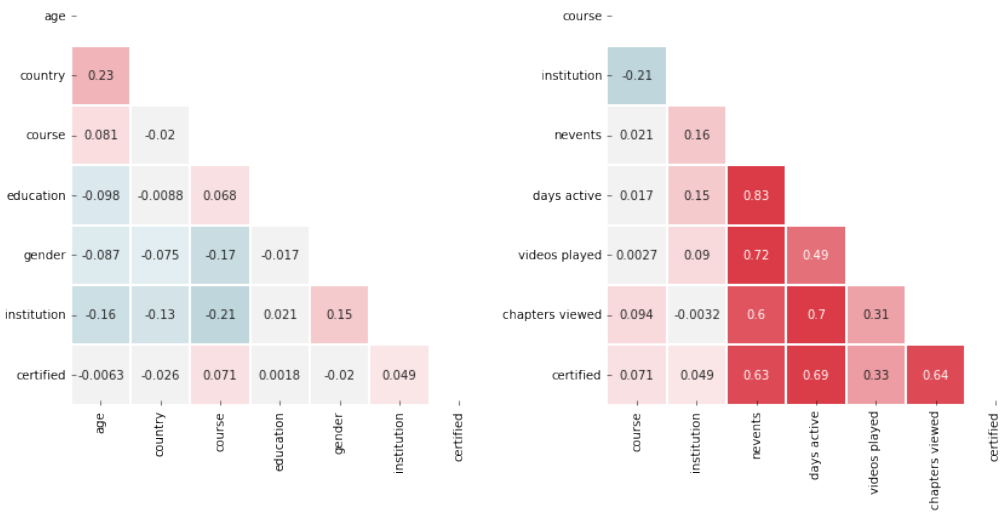
Lets now look at the confusion matrix and precision, recall and f1 scores from the different logistic regression models.



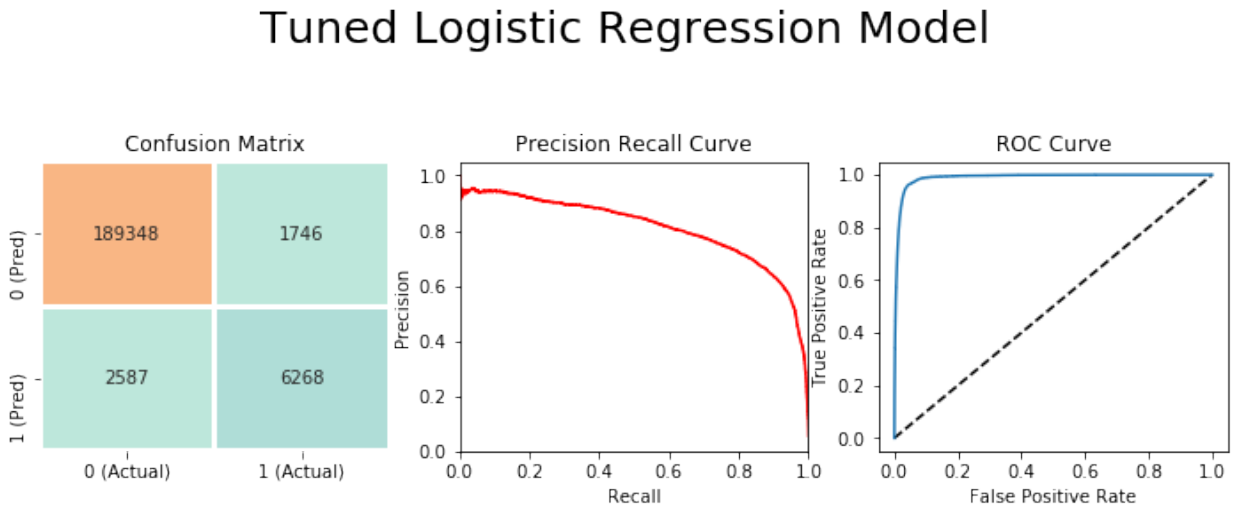
	Class	Precision	Recall	f1-score
Logistic Regression using demographics	0	0.97	0.99	0.98
	1	0.62	0.28	0.38
Logistic Regression using administrative info	0	0.99	0.99	0.99
	1	0.78	0.70	0.74
Logistic Regression using both demographics and administrative info	0	0.99	0.99	0.99
	1	0.78	0.71	0.74

As we can see from the confusion matrix , the version of the model using only the demographic information performs poorly for the minority class (which is our target class). When we use the administrative information which has the course interactions, we see a significant improvement in predicting the certification rate.

We can relate the model performance to the correlation between the certification rate and demographics and course interaction details. As we can see the from the correlation heat map below , there is very little correlation between the user specified info (age/gender/education) and the course completion rates. As one would expect, the course completion shows high correlation with factors such as the number of interactions , number of chapters viewed etc.



As a next step of model performance tuning, we explore the parameters of the logistic regression model (C and penalty) to improve the model performance and minimize over fitting. Using cross-validation and GridSearch we arrive at the optimal setting for the parameters

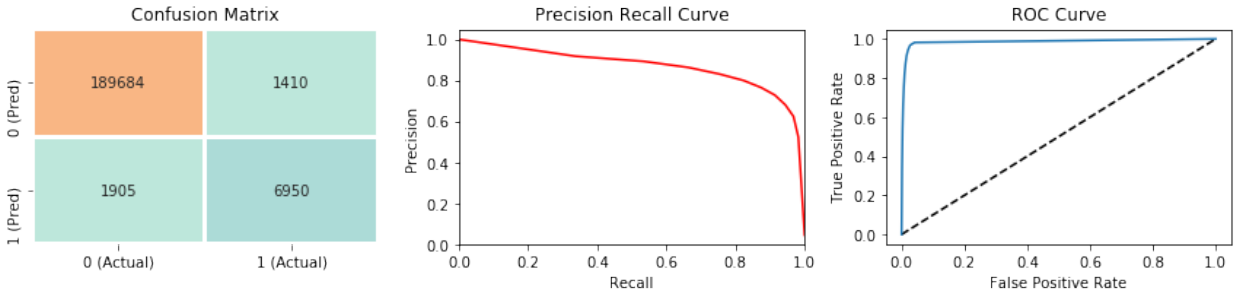


Random Forest Model

As Tree Ensembles have better performance with non linear features and large dataset, I have built the classifier (to predict the registrant success rate) using Random Forest Model . As we can see from the precision, recall and f1-score this model performs significantly better than the Logistic Regression

	Class	Precision	Recall	F1-score
Random Forest model using both demographics and administrative info	0	0.99	0.99	0.99
	1	0.83	0.78	0.81

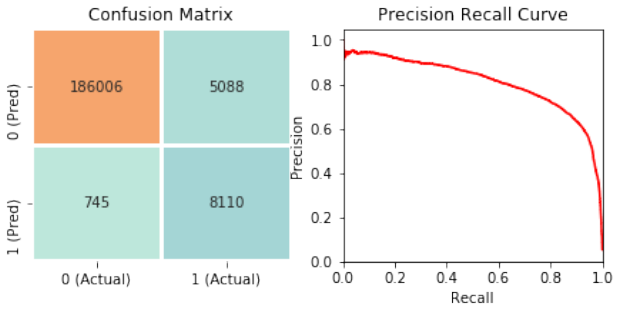
Random Forest Model



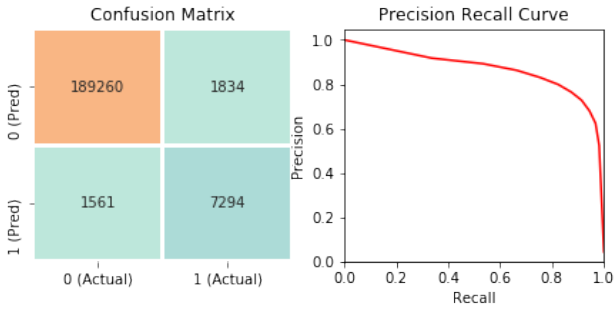
Addressing the class imbalance

In our data set with more than six hundred thousand registrants, we have less than 2.8% of registrants with course completion. To counteract the class imbalance, I have used oversampling of the minority class (using SMOTE) so it has more effect on the machine learning algorithm.As we can see from the confusion matrix below , we see an increase in True Positive . We also see a significant increase in the False Negative for Logistic Regression , and a moderate increase of False Positive and False Negative for the Random Forest Model. Here again the performance of the Random Forest is better than the Logistic Regression.

Logistic Regression with SMOTE



Random Forest with SMOTE



Conclusion

In summary, using the historic data of course interactions provided by the institutions, we can predict the success (certification) rate of the course participants. For this prediction model to be of use , we need the time series data of course interactions. Such a model would require institutions data such as the number of chapters / videos viewed in a week , number of course interactions , chapters / videos viewed at periodic intervals.so that early interventions can be made to enable the participants to complete the course in a timely manner .

We could also extend the model building to other classifiers such as Linear SVM, KNN classifier,Decision Trees,Naïve Bayes and more advanced models such as Gradient Boosted Decision Trees (GBDT) and fine-tuning the hyper parameters for better fit. We could also try Deep Learning methods to build more complex , multi layer model to improve the performance of the classification model.