Predicting participant completion rate in EdX / Open Courses

Sukanya Chandramouli

Data Science

April 2018

EdX Dataset

- Course Interactions Details from Harvard/MIT Open/Online Courses on EdX
- Dataset includes participant/interaction details for 13 courses offered during the Academic Year 2013-14
- 641138 Registrants in total
- 17687 Registrants Completed the course
- Dataset includes
  - *User Data (Age , gender , educational background)*
  - *Administrative Data – chapters viewed , videos viewed , days active ,#events*
  - *Grade , certified*

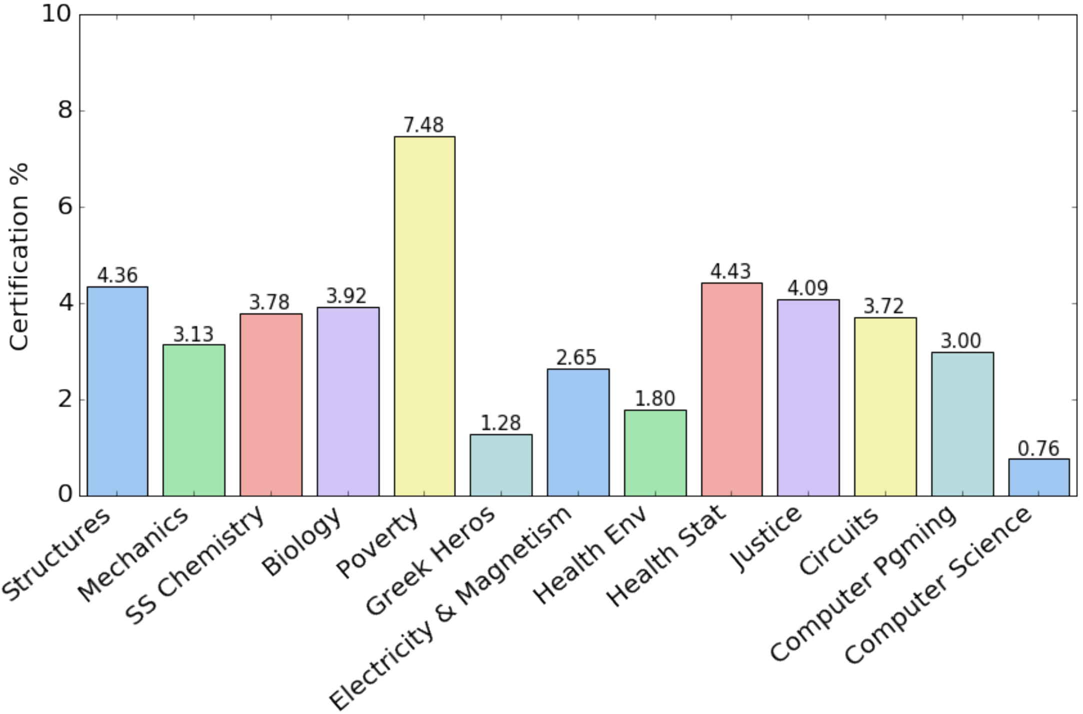
HARVARD
Dataverse

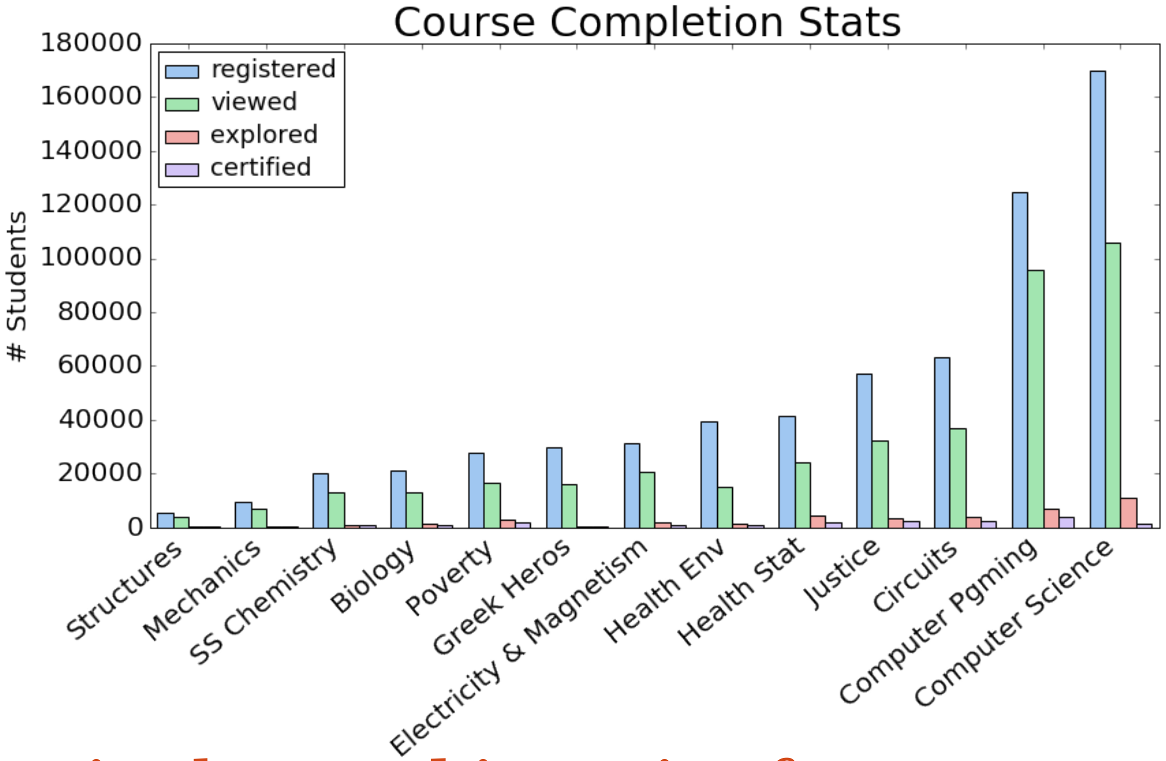
edX

Massachusetts
Institute of
Technology

HARVARD
UNIVERSITY

EdX Dataset



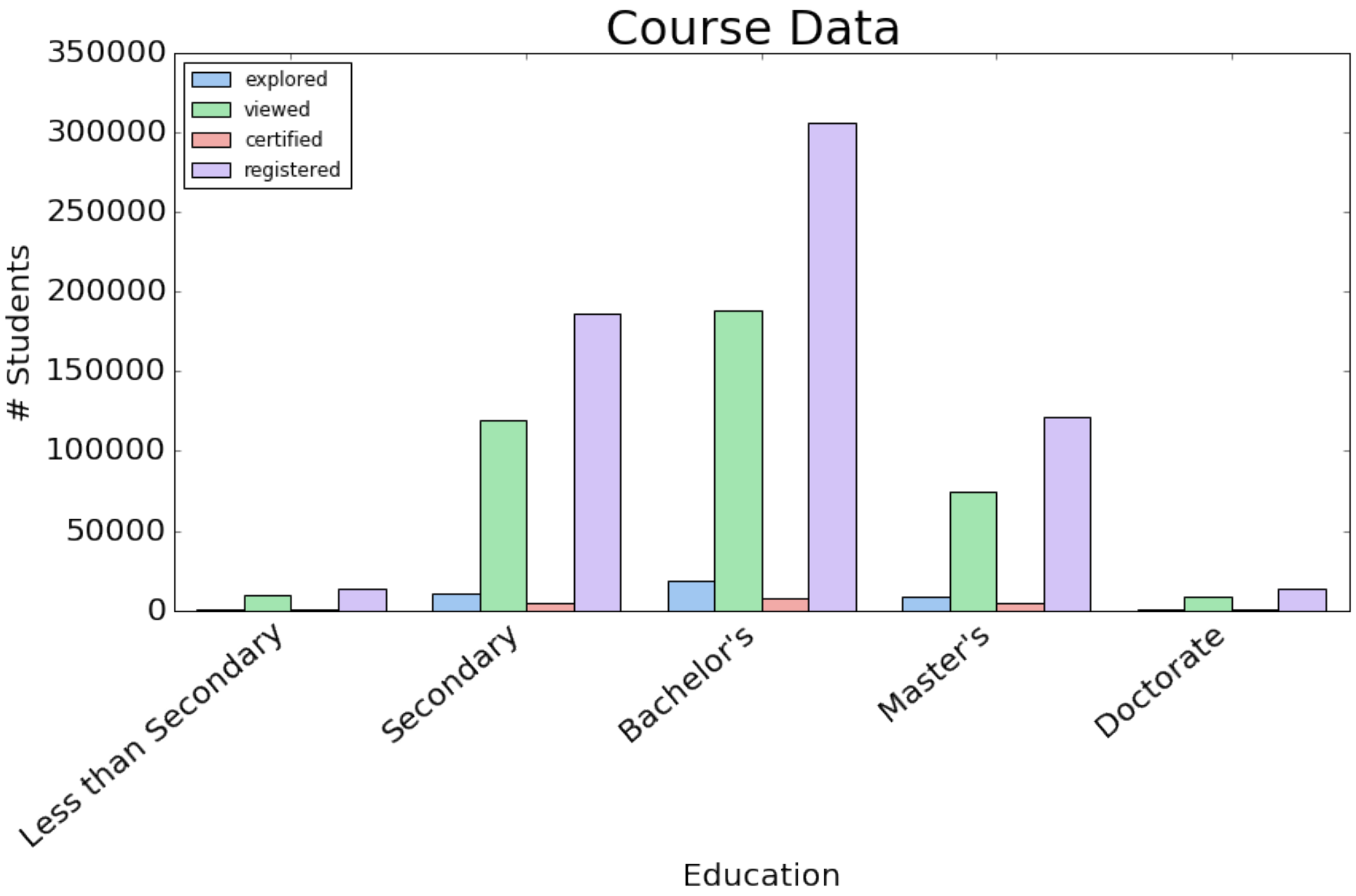


- **Course completion rate is quite low and it varies from 0.7% to 7.5%**
- **Can we predict the completion rates early on, to provide interventions for successful completion ?**

Project Goals

- Identify features from historic data
- Build a machine learning model to predict course completion rates
- Evaluate the performance of the Machine Learning (Classification) model

Data – Visual Analysis



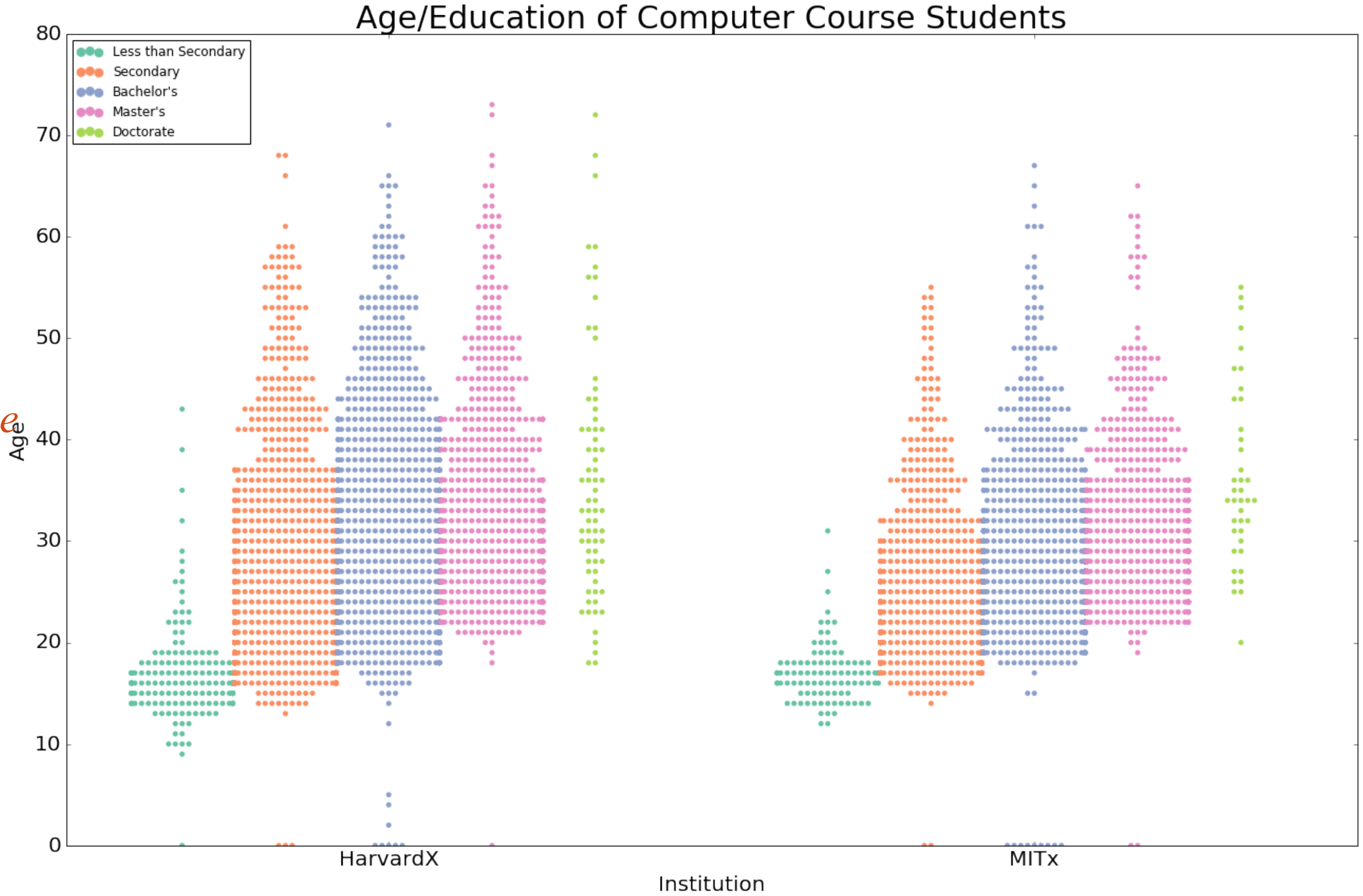
- *Users with secondary and bachelors degree have a higher enrollment*
- *The certification rate is low across different education levels*

Data – Visual Analysis

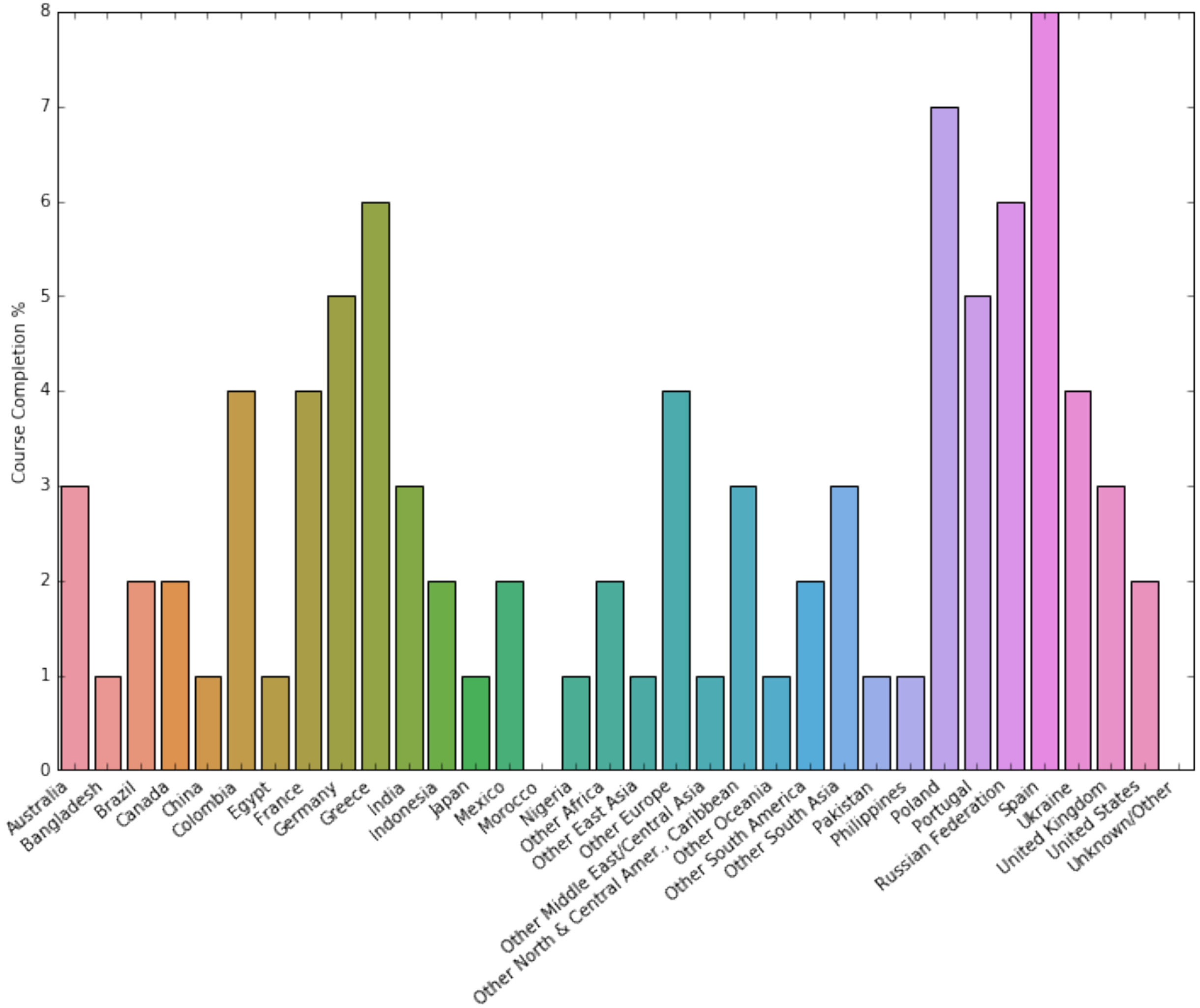
- *Computer courses have highest enrollment and make up more than 45% of dataset*

- *Users with secondary , bachelors and masters degree have a higher enrollment*

- *The certification rate is low across different education levels*

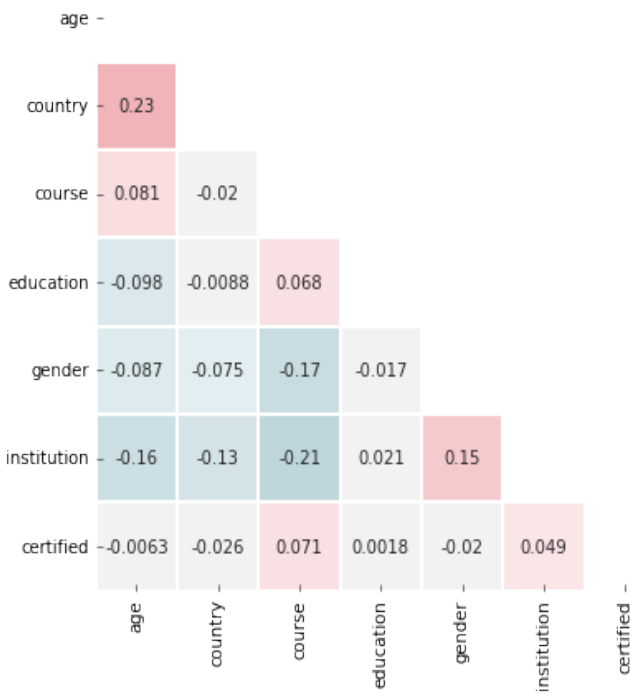


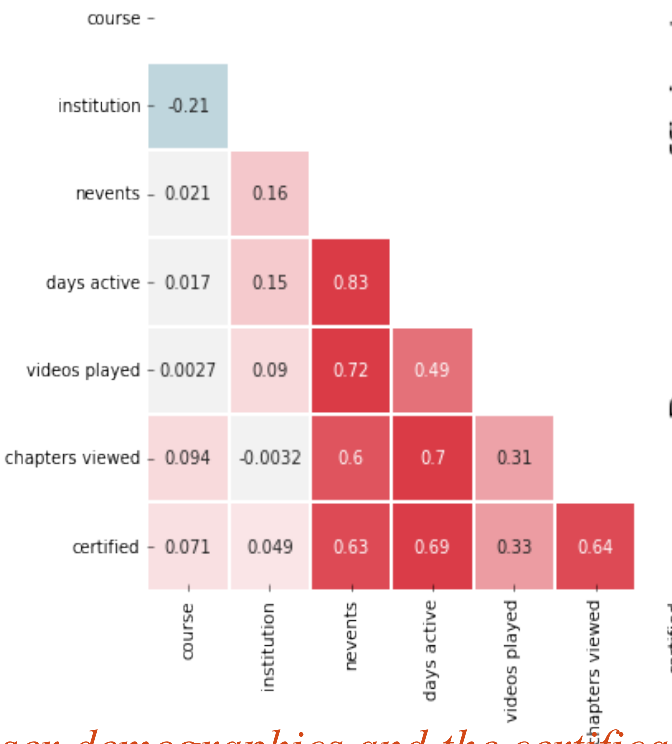
Data – Visual Analysis

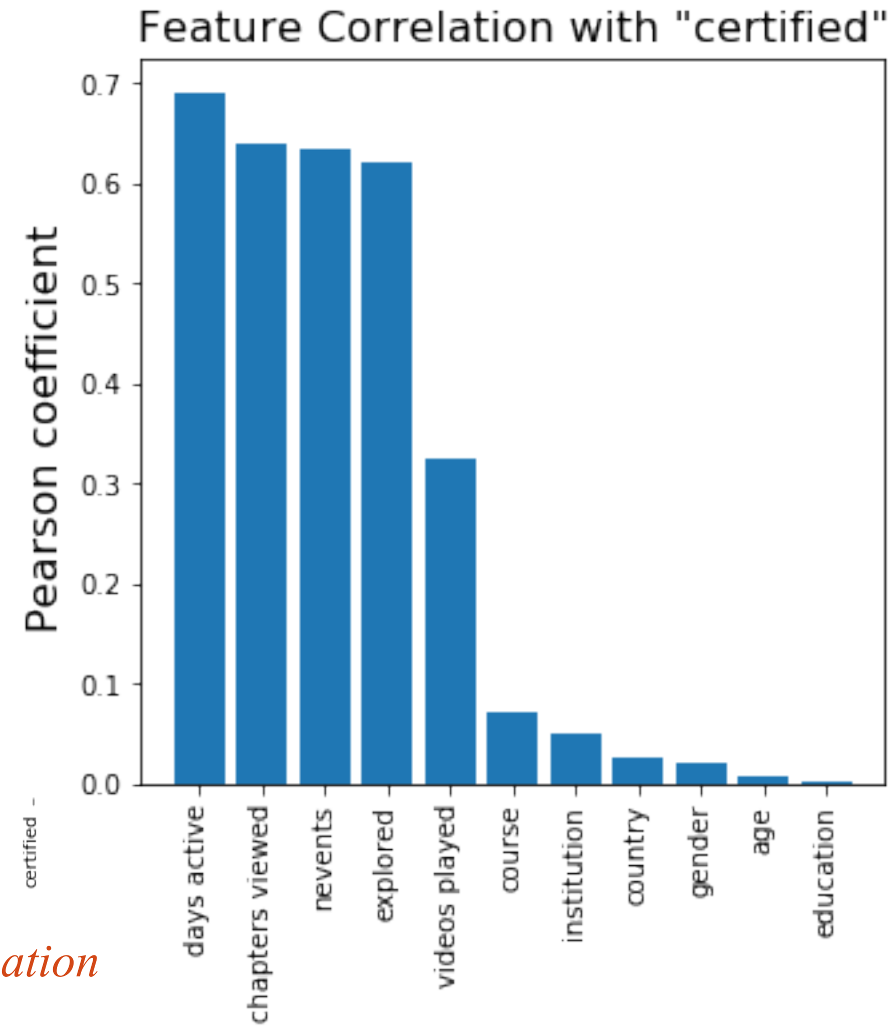


Certification rates across different regions

Feature Extraction





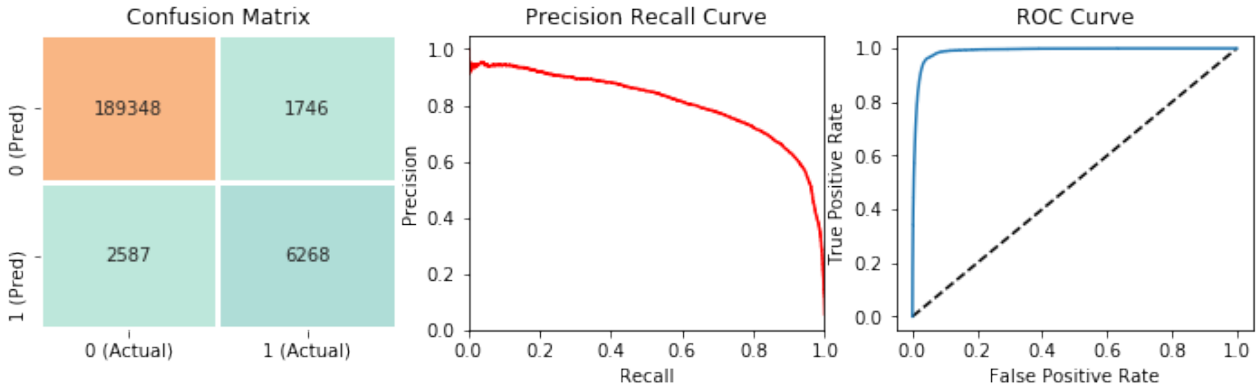


There is low correlation between user demographics and the certification rate

Using the Correlation Matrix/Pearson coefficient we can identify the key variables that affect the certification rate (right)

Classification Model

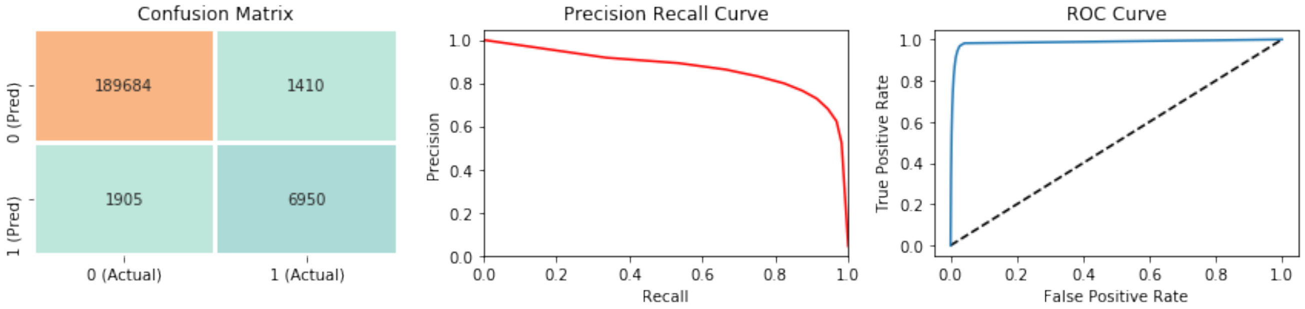
Tuned Logistic Regression Model



Confusion Matrix

	0 (Actual)	1 (Actual)
0 (Pred)	189348	1746
1 (Pred)	2587	6268

Random Forest Model



Confusion Matrix

	0 (Actual)	1 (Actual)
0 (Pred)	189684	1410
1 (Pred)	1905	6950

- *Dataset split into 50% training and 50% test set*
- *Logistic Regression model is built and fine tuned using GridSearch and Cross Validation*
- *Next we build the classifier using Random Forest Model*

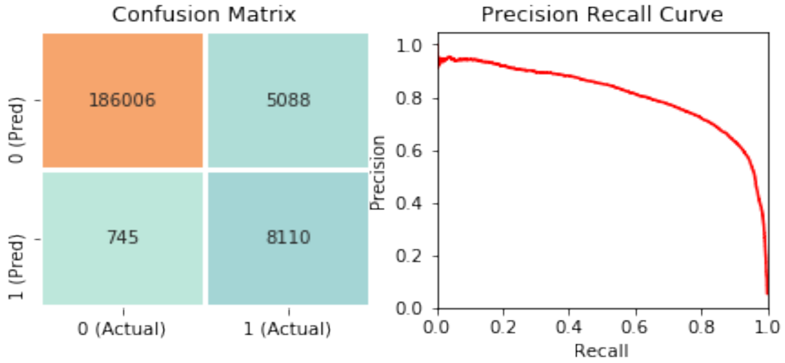
Comparison of Classification models

	Class	Precision	Recall	F1-score
Logistic Regression Model	0	0.99	0.99	0.99
	1	0.78	0.71	0.74
Random Forest Model	0	0.99	0.99	0.99
	1	0.83	0.78	0.81

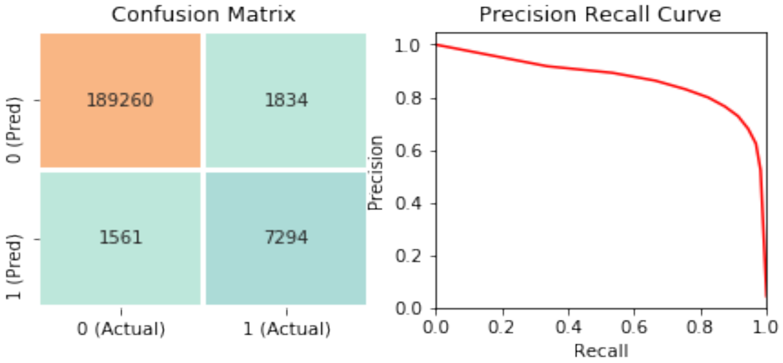
- *Both models have a good accuracy , Precision & Recall scores*
- *Random forest model performs better than the logistic regression (confusion matrix , precision , recall scores are better)*

Class Imbalance

Logistic Regression with SMOTE



Random Forest with SMOTE



- *SMOTE – Synthetic up-sampling of minority class (training set) to improve the class imbalance*
- *Increases True Positive rate but also increases False Negative rates*
- *Performance of the Random Forest Model better than Logistic Regression*

Summary

- Using historic data , we can predict the course completion rates with 80% accuracy
- Early interventions for course completion can be made if we have time-wise break of the course interaction details
- Classification model can be extended with time-series data
- Model performance can be improved with Deep learning methods using multilayer networks