



How to Use Coursera Data to Improve MOOC Quality and Learner Experience

Presented by: Melody Wu
April 19, 2021



01

Background and
Problem Statement

02

Data Collection +
Pre-Processing

03

Course Info EDA and
Enrollment Modeling

04

Course Review EDA and
Classifier Modeling

05

Recommendations
and Conclusions

01

Background + Problem Statement



 **180M**
Students

 **950**
Universities

 **16.3k**
Courses

 **1180**
Microcredentials

 **67**
MOOC-based degrees

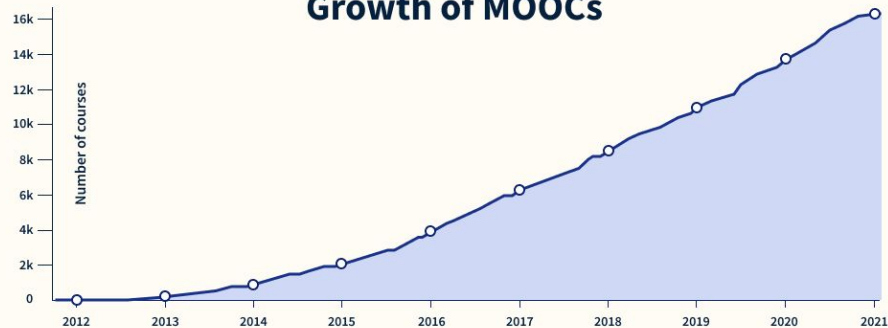
 class central

By the Numbers: MOOCs in 2020
Statistics do not include China


 class central

By the Numbers: MOOCs in 2020
Statistics do not include China

Growth of MOOCs



Only 3% learners, among all MOOC participants, completed their courses in 2017-18, according to Dr. Justin Reich and Dr. José A. Ruipérez-Valiente from Massachusetts Institute of Technology (MIT).



understand learner experience and improve MOOC quality

Understand **course characteristics** and explore the general landscape



Investigate the characteristics that can **predict enrollment**

Examine the **most discussed topics and sentiments** among MOOC learners



Leverage **learner reviews** to identify areas of improvements

The screenshot shows a Coursera video player interface. At the top, the Coursera logo is on the left, and a search bar with the text "What do you want to learn?" is in the center. To the right of the search bar is a user profile icon. Below the search bar, the breadcrumb "AI For Everyone > Week 1 > Non-technical explanation of deep learning (Part 1, optional)" is visible. The video player itself shows a man in a light blue shirt sitting at a desk. Behind him is a monitor displaying a red and white slide titled "What is AI" with the subtitle "Non-technical explanation of deep learning (Part 1, optional)". To the left of the man is a shelf with various items. Below the video, there are icons for "Save Note", "Discuss", and "Download". A language dropdown menu is set to "English", and a "Help Us Translate" link is to its right. At the bottom, a video transcript is visible, starting with "0.03 The terms deep learning and neural network are used almost interchangeably in AI. And even though they're great for machine learning, there's also been a bit of hype and bit of mystique about them. This video will demystify deep learning, so that you have a sense of what deep learning and neural networks really are. Let's use an example from demand prediction. Let's say". On the left side of the video player, there is a list of video topics with checkmarks and durations:

- ✓ Video: Week 1 Introduction 7 min
- ✓ Video: Machine Learning 6 min
- ✓ Video: What is data? 11 min
- ✓ Video: The terminology of AI 9 min
- ✓ Video: What makes an AI company? 7 min
- ✓ Video: What machine learning can and cannot do 7 min
- ✓ Video: More examples of what machine learning can and cannot do 8 min
- ▶ Video: Non-technical explanation of deep learning (Part 1, optional) 7 min
- ▶ Video: Non-technical explanation of deep learning (Part 2, optional) 3 min

11 subjects

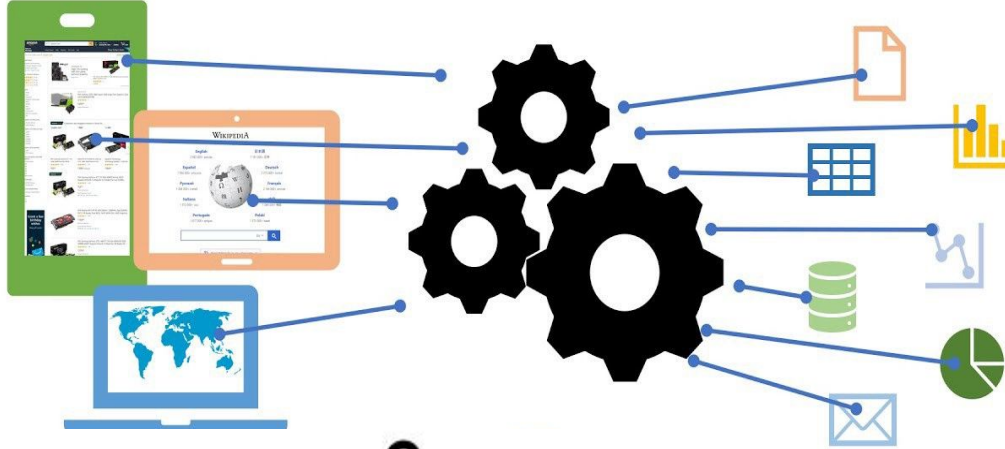
Data Science, Business, Computer Science, Personal Development, Health, and more

50+ online degrees

Online MBA, Data Science, Public Health, Management, Data Analytics

20+ MASTERTRACK CERTIFICATES

University of Michigan, UIUC, University of Chicago, HEC Paris



02

BeautifulSoup

Data Collection +
Pre-processing



Selenium

COURSE INFO

348 Data Science courses with information of views, ratings, providers, length, outcomes, etc.

The image shows a stack of three identical course enrollment cards for Coursera Plus. The top card is in focus, showing the following details:

- Start your 7-day free trial** (repeated three times on the stack)
- 5-course Specialization**: Python for Everybody Specialization (5-course series)
- Price**: \$49 / month after free trial. A button says "Start free trial" with "Cancel anytime" below it.
- Get access to 3,000+ learning programs**: Coursera Plus
- Monthly subscription includes Python for Everybody Specialization plus additional benefits:**
 - ✓ Advance or change your career with access to courses, Specializations, Guided Projects, and Professional Certificates.
 - ✓ Save money while learning unlimited skills and subjects each month.
 - ✓ Explore the course catalog at your own pace.
 - ✓ Showcase your knowledge with course assessments and certificates.
- Buttons**: "Start free trial" (blue) and "Learn more" (blue link).

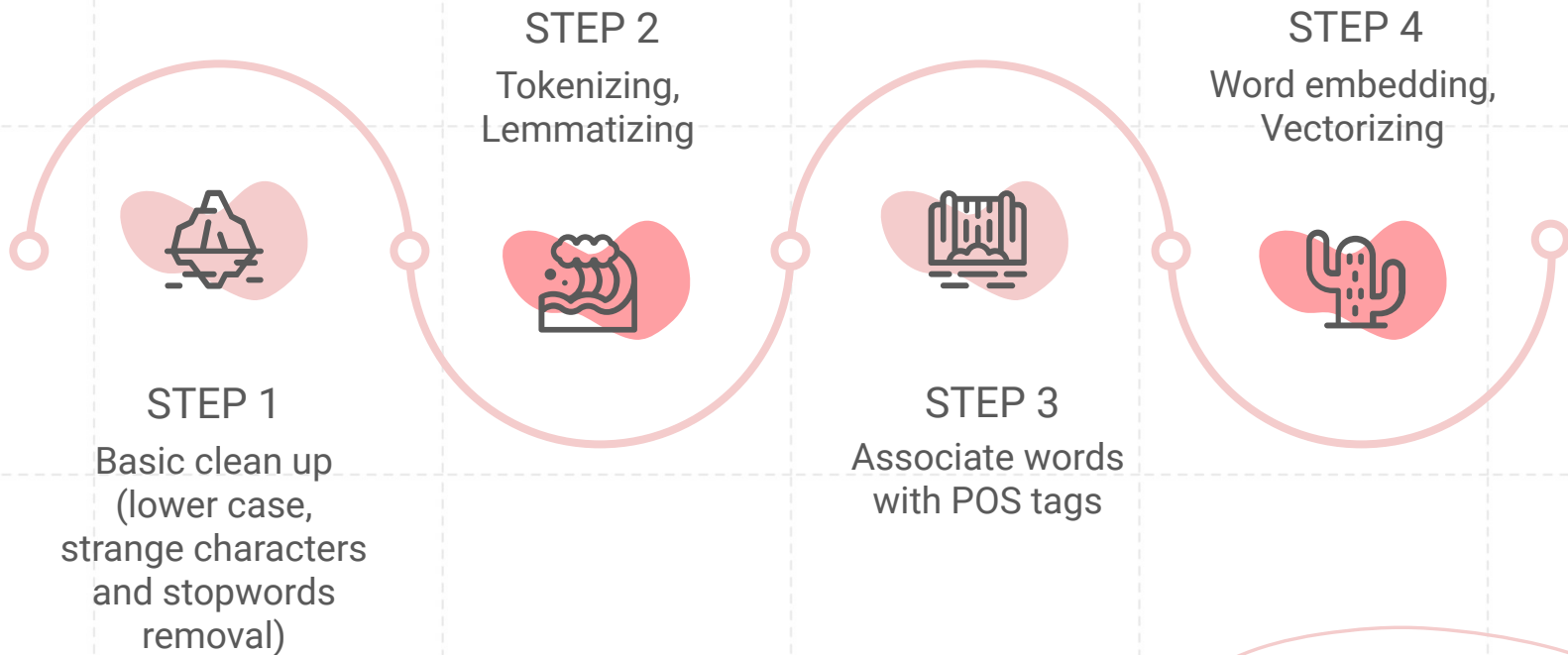
REVIEWS

12,000+ reviews from Machine Learning taught by Andrew Ng

The image shows a stack of three user reviews for Coursera. The top review is in focus, showing the following details:

- User**: ashton (5.0/5.0 - Coursera User)
- Rating**: 5 stars
- Title**: perfect for quarantine
- Text**: Coursera made my learning so much easier than just reading random books and trying to learn! Teachers here are really willing to help. I would recommend this platform to anyone who's reading.
- User**: Valerie F. (5.0/5.0 - Coursera User)
- Rating**: 5 stars
- Title**: Flexible!!
- Text**: Great courses, flexible timetables, not too pricey as well. Also I got a certificate after finishing!
- User**: Braiden (4.6/5.0 - Coursera User)
- Rating**: 4.6 stars
- Title**: worthwhile courses
- Text**: the learning process is a bit complicated as you get too many details at the start and get confused

TEXTUAL DATA PREPROCESSING





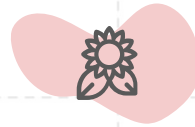
03

Course Info EDA and Enrollment Modeling



100%

100% Online



100%

Flexible Deadlines

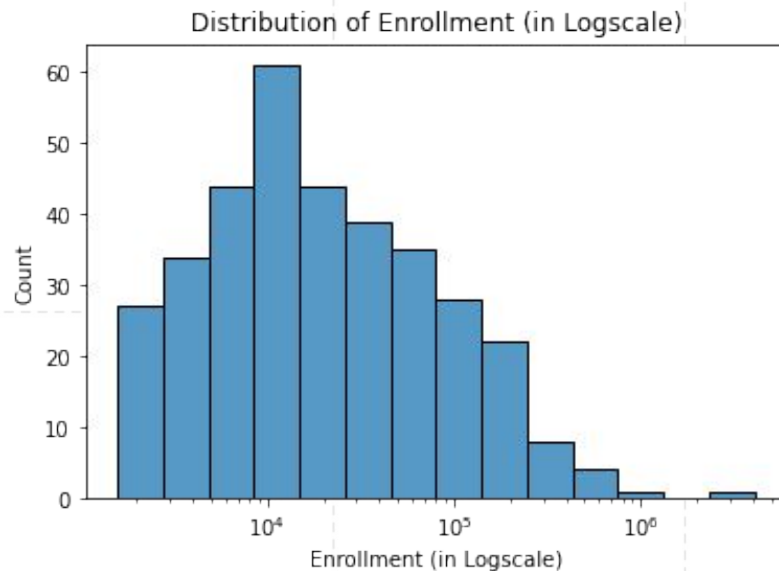


99%

Shareable Certificates

1,585 ~ 4,047,733

Enrollment is highly skewed!





61% of the Data Science Courses
are offered by higher education institutions

64% of these courses
are part of specializations (serieses of thematic courses)



4.6 (out of 5) average ratings
among all Data Science Courses, leftward skewed

CAREER OUTCOMES



31% started a new career



21% got a tangible career benefit



32% got a pay increase



Machine Learning Using SAS Viya

SAS

COURSE



4.7 (75)

1 COURSE

19 HOURS

INTERMEDIATE

This course covers the theoretical foundation for different techniques associated with supervised machine learning models. In addition, a business case study is defined to guide participants through all steps of the analytical life cycle, from problem understanding to model deployment, through data preparation, feature selection, model training and validation, and model assessment. A series of demonstrations and exercises is used to reinforce the concepts and the analytical approach to solving business problems.

[SHOW ALL](#)

Instructors

Instructor rating 🌟 4.58/5 (15 Ratings) ⓘ



Jeff Thompson

Senior Analytical Training
Consultant

Education

👤 5,758 Learners

📖 1 Course



Catherine Truxillo

Director, Analytical Education

Education

👤 5,758 Learners

📖 1 Course



Learner Career Outcomes



67%

got a tangible career
benefit from this course



67%

got a pay increase or
promotion



Shareable Certificate

Earn a Certificate upon completion



100% online

Start instantly and learn at your own
schedule.



Advanced Data Science Capstone

IBM

COURSE



4.6 (353) | 12K students

 **Advanced**

This project completer has proven a deep understanding on massive parallel data processing, data exploration and visualization, advanced machine learning and deep learning and how to apply his knowledge in a real-world practical use case where he justifies architectural decisions, proves understanding the characteristics of different algorithms, frameworks and technologies and how they impact model performance and scalability.

[SHOW ALL](#)

Instructor

Instructor rating  4.13/5 (22 Ratings) 



Romeo Kienzler

Chief Data Scientist, Course
Lead

IBM Watson IoT

 **299,745** Learners

 **8** Courses



Learner Career Outcomes



67%

started a new career
after completing these
courses



57%

got a tangible career
benefit from this course



Shareable Certificate

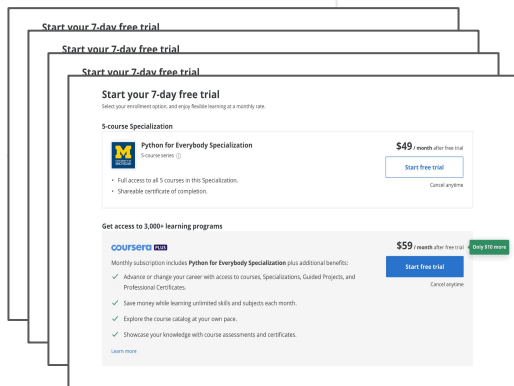
Earn a Certificate upon completion



100% online

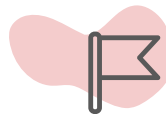
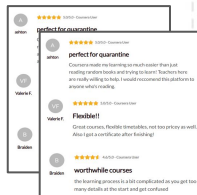
Start instantly and learn at your own
schedule.

COURSE INFO



Enrollment Estimator

REVIEWS



Review Classifier

ENROLLMENT ESTIMATOR

- Baseline RMSE: **235,563**
- Tested models (7): Linear Regression, Kneighbors, Decision Tree, Random Forest, AdaBoost, XGBoost, SVR.
- Shortlisted: Linear Regression, Decision Tree, XGBoost
- Final Model: Linear Regression + PolyNomial Features + LASSO Regularization
 - R-square: **0.70**
 - RMSE: **37,546**
- Contributing regressors: number of ratings, number of reviews, length, providers, difficulty levels (**do not include ratings!**)



Machine Learning

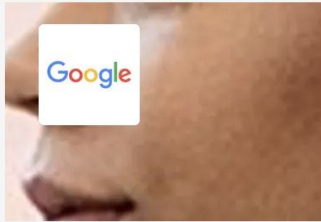
Stanford University

COURSE

★★★★★ 4.9 (158,250)

1 COURSE

55 HOURS



Foundations: Data, Data, Everywhere

Google

COURSE

★★★★★ 4.8 (3,110)

1 COURSE

12 HOURS

BEGINNER

PLUS



Ask Questions to Make Data-Driven Decisions

Google

COURSE

★★★★☆ 4.7 (799)

1 COURSE

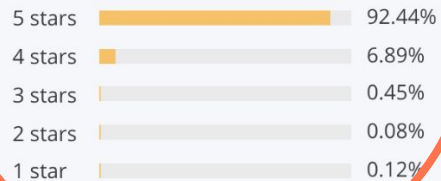
9 HOURS

BEGINNER

PLUS

Reviews

4.9 ★★★★★
40505 reviews



TOP REVIEWS FROM MACHINE LEARNING



by PT Sep 1, 2018

Sub title should be corrected. Since I'm not that good in English but I know when there're mis-traslated or wrong sub title. If you fix this problems , I thin it helps many students a lot. Thanks!!!!

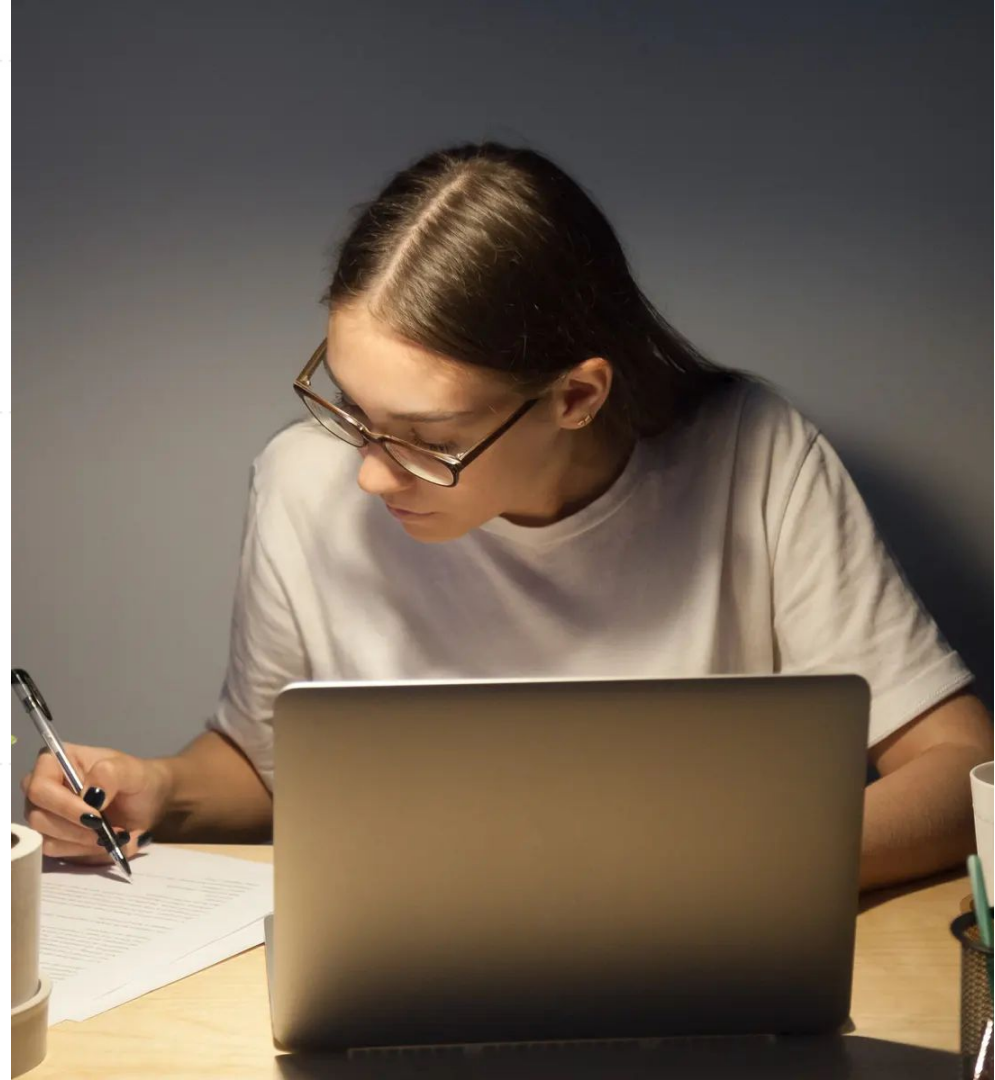


by AF Mar 17, 2021

I want to thank you very much for such a great course in any aspect especially from professor Ng . I just want to suggest that it would be great if there was a final project for the end of the course.

04

Course Review EDA and Classifier Modeling





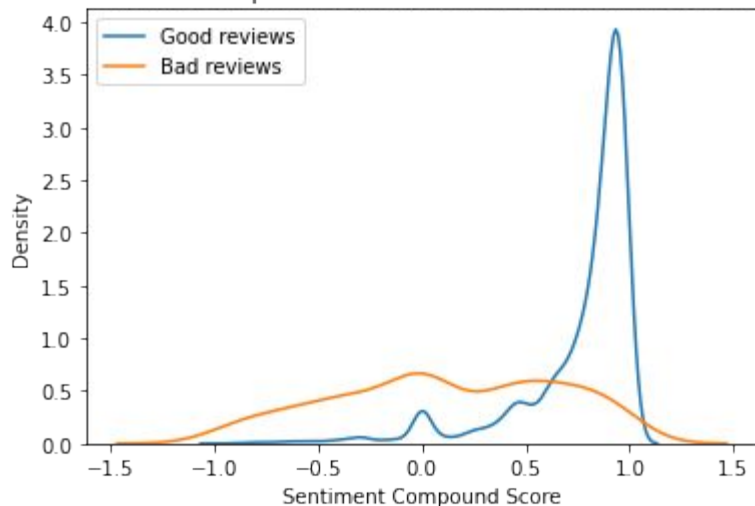
Raing 1, 2, and 3



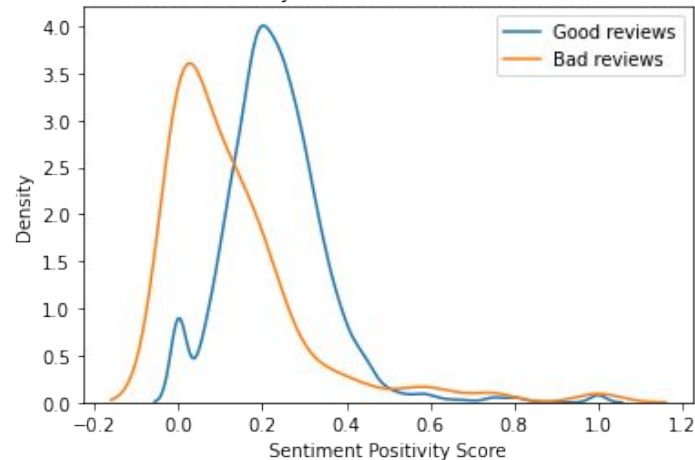
Rating 4 and 5

Sentiment Analysis

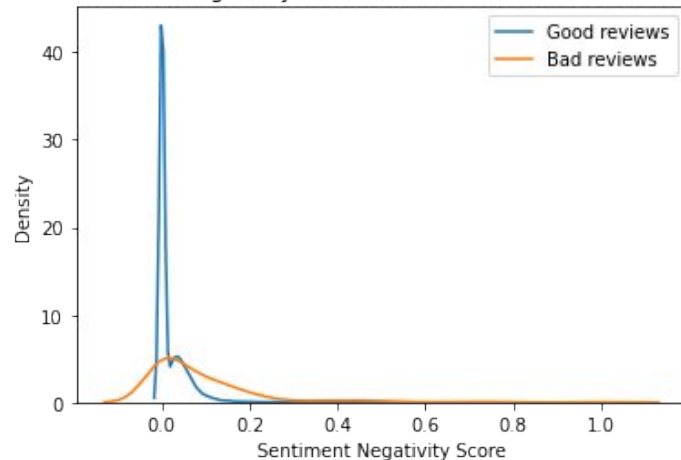
Sentiment Compound Score: Good reviews vs. Bad reviews



Sentiment Positivity Score: Good reviews vs. Bad reviews



Sentiment Negativity Score: Good reviews vs. Bad reviews



Topic Modeling

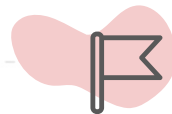
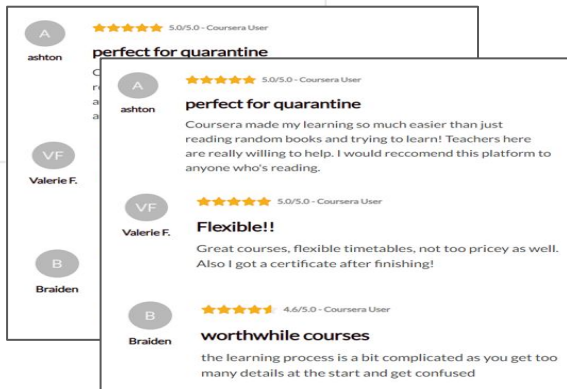
- Topic 1: general comments on the course
 - "course" + "great" + "assignment" + "program" + "time" + "ng" + "exercise" + "lecture" + "easy" + "lot"
- Topic 2: positive feedback on the instructors
 - "ng" + "course" + "thank" + "coursera" + "thanks" + "andrew" + "much" + "professor" + "i" + "best"
- Topic 3: comments on background and course prerequisites
 - "science" + "data" + "computer" + "python" + "benefit" + "scientist" + "thought" + "update" + "evaluate" + "language"
- Topic 4: course contents
 - "course" + "machine" + "learn" + "ml" + "great" + "good" + "use" + "algorithm" + "work" + "concept"
- Topic 5: suggestions and recommendations for fellow students
 - "course" + "machine" + "learn" + "learning" + "good" + "great" + "start" + "best" + "beginner" + "concept"
- Topic 6: reviews given by spanish speakers
 - "que" + "curso" + "un" + "la" + "muy" + "para" + "los" + "el" + "con" + "excelente"

COURSE INFO



Enrollment
Estimator

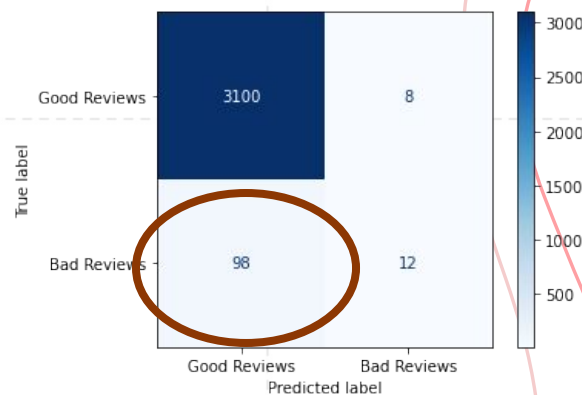
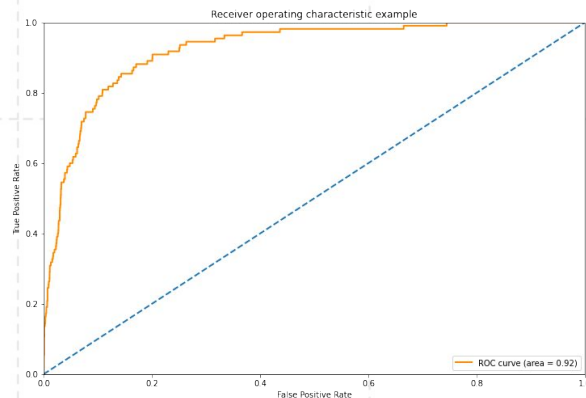
REVIEWS



Review
Classifier

REVIEW CLASSIFIER

- Baseline Models:
 - Without adjusting class imbalance: **0.97**
- Without adjusting class imbalance, Neither logistic regression nor tree-based models outperform the baseline. These models all have low recall rates of the positive class (**only 11% of the bad reviews being identified as bad reviews**).





Misclassified Reviews



Been asked to re-take all assignments *after* paying for a certificate! I will never pay for a Coursera course again, and **I would not recommend** my friends to do so. (True == Bad, Pred == Good)



Course content was good **but i felt few lectures were too boring** to constantly concentrate on. (True == Bad, Pred == Good)



I feel **this course is too old for 2020** and As programmer who has some linking for Python **I find this course quite impractical** ... My verdict is that **this is not suited for python intermediates** and only good if you get your toe wet in ... (True == Bad, Pred == Good)

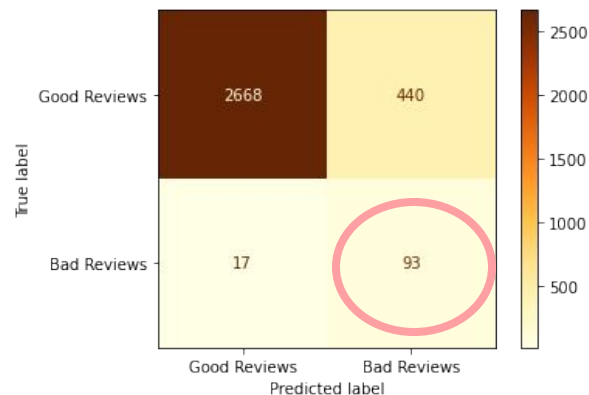
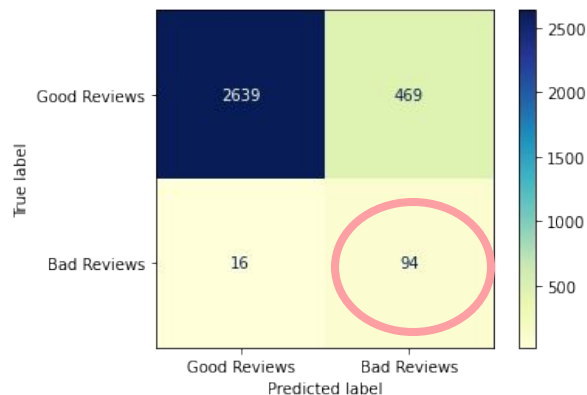
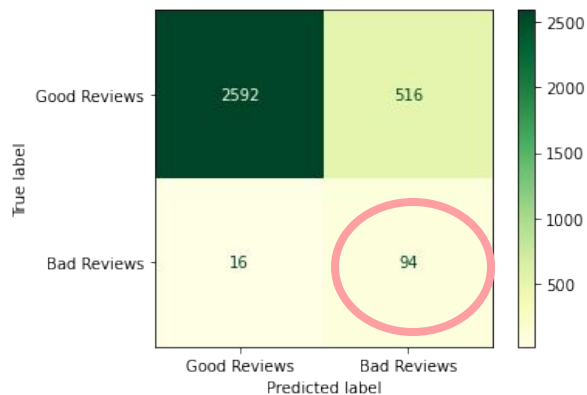


Mainly heavily maths based, **not practical** to apply to real industrial experience. Some more not using R/Python **is a big minus** for this course. (True == Bad, Pred == Good)

REVIEW CLASSIFIER

- Adjusting class imbalance:
 - Undersample More Frequent Class
 - Oversample Less Frequent Class
 - SMOTE (Synthetic Minority Over-sampling Technique)
- Baseline Models: accuracy rate = 0.50
- Testing: Logistic regression with LASSO Regularization:
 - 0.834 (Under)
 - 0.849 (Over)
 - 0.858 (SMOTE)
- Contributing regressors: length (characters (-), words(+)), sentiment compound scores (-), word_andrew(-), word_concept(-), word_learn(-), word_would(+), word_python(+), word_good(+), word vectors

Adjusting Class Imbalance



Misclassified Reviews



I **would be much better** if it would have included the more layman explanation. (True == Good, Pred == Bad)



... About the tests, I **sometimes found the text a little confusing**, but ... (True == Good, Pred == Bad)



I **am sad** to see after accomplishing 11 weeks course there wasn't any certificate given. (True == Good, Pred == Bad)



A very good course and introduction to this topic overall. The videos are perhaps **a bit old and some of the material might benefit from being updated**, ... (True == Bad, Pred == Good)



Course has become a bit outdated. (True == Good, Pred == Bad)



05

Recommendations and Conclusion

RECOMMENDATIONS

- Use **topic modeling** to label reviews for learners and for identifying areas of improvements for course instructors/developers
- Redesign UI to present more meaningful information to the learners
- Develop outcome and review **dashboard for course instructors/developers**
- Design **a rating recommender** - when the learner write a review, recommend a bad (1, 2 or 3) or good (4 or 5). Since the model is over 'critical', in the long run, there will be more 1, 2, or 3 ratings and learners can better use the over ratings to differentiate course quality.

NEXT STEPS

- Design a better web scraper
- Expand the research to other subjects and courses
- Conduct learner side research (etc. user logs, surveys, observations, interviews)
- Prototype the analytics dashboard (taking the url and give insight of a course)
- Develop a user-based recommender system



THANKS!

Do you have any questions?

melody.jwu@gmail.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**