Project Proposal
Introduction to Machine Learning

R.J. Klaasse Bos

November 5, 2017

1 Introduction

1.1 Online Education

In recent years online education has taken a flight: universities throughout the world have published their Massive Online Open Courses (MOOCs) on platforms such as Coursera and edX. While aforementioned platforms offer academics an online stage other platforms can be described as marketplaces where basically anyone can become an online instructor. One example of such a platform is *Udemy* which offers 55,000 courses, just hit the 15-million students mark and is still growing rapidly (Figure 1).

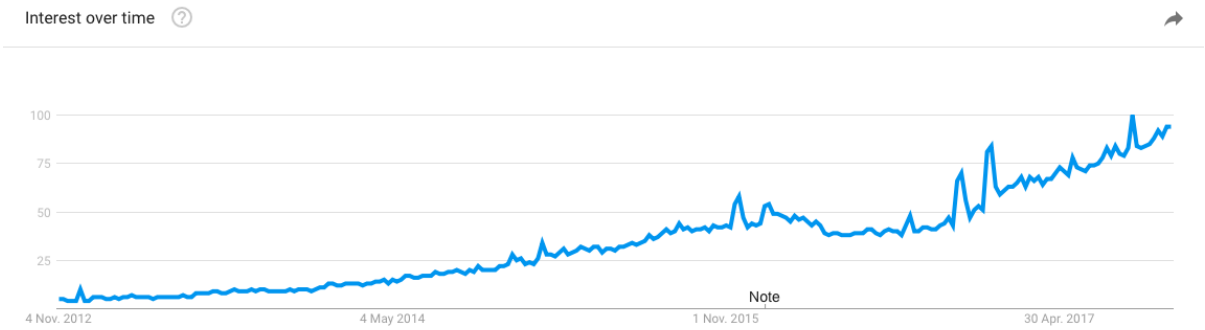

Figure 1: Google Search Trends for "Udemy" (2012-2017)

This rise in popularity also comes with new challenges. Due to network effects and increased student demand platforms attract more and more creators which makes it for teachers even more challenging to differentiate from the many already existing courses. Creating a course from scratch requires a significant upfront time investment while the final pay-off remains uncertain. Therefore teachers can benefit from insights into what course characteristics are positively related to their expected student base by estimating the total addressable market. From a commercial standpoint platforms can also benefit from these insights. For example, to decide what courses to promote on their website, since the screen space is very limited given the total number of courses. Further, Udemy has an interest in keeping students satisfied which becomes directly apparent from the many course reviews. By analysing and finding patterns in the abundance of reviews,

they can support teachers in improving their current courses and developing even better ones in the future.

1.2 Platform motivation

The reason Udemy was picked over any of the other platforms is because their course pages contain relatively many features compared to its competitors (e.g. features such as number of courses previously created by the teacher). Important to note here is that the platform dynamics are entirely different than for example on Coursera. That is because the large majority of Udemy courses are paid, whereas all Coursera's video lectures can be viewed for free.

1.3 Project motivation

In reality data scientists spend their majority of their time (70%) cleaning and preparing data. However, in the Data Science & Entrepreneurship pre-master courses we have solely worked with cleaned datasets so far. To become familiar with a larger part of the data pipeline: from (real-world) data collection to drawing conclusions I would like to work on this alternative project which involves additional steps such as data scraping and data cleaning. Furthermore, in other courses I have primarily applied supervised classification algorithms to predict a categorical outcome (e.g. employee attrition). As will follow from the posed predictive questions in the next section, this project allows me to become more knowledgeable on the use of regression algorithms.

Besides that, I have specifically chosen the topic of online courses for several reasons. First, I am really enthusiastic about how MOOCs democratize education and allow students to learn from the most famous experts in their field regardless of their location, age and financial position. Second, I have taken many online courses myself and thus have relevant domain knowledge which may help with identifying relevant features. Third, I find the consumer decision-making process very intriguing, especially when combined with quantitative research methods. For example, in my *bachelor thesis* I have identified predictors of Kickstarter crowdfunding success.

2 Research Design

2.1 Research Question

This project has been structured in such a way that both descriptive and predictive questions will be addressed. In fact, the results of the descriptive questions can contribute to building a better predictive model. Below follows an overview of 7 descriptive questions as well as 2 target features for predictive modelling:

Descriptive Analysis

1. Is there a positive relationship between the number of hours of video and the number of students enrolled in the course (since they get more value for their money)? Does this relationship also hold for free courses?

2. Is the average star-rating for free courses higher than for paid courses since students have different expectations for a free course (i.e. they are more lenient)?

3. Do courses with a relatively high star-rating attract significantly more students?

4. Is there no relationship between the original course price and the number of students enrolled due to Udemy's extreme discount policy?

5. Are there significantly fewer students enrolled in courses of an intermediate or expert level (since they exclude beginners)?

6. Do English spoken courses attract significantly more students compared to non-English spoken ones?

7. Is there a positive relationship between the average star rating across all courses and the number of courses created by the teacher (because he/she gains more experience over time)?

Predictive Analysis

1. *Number of Students Enrolled*

2. *Course rating*

   Note that predicting the course rating ($0.0 \leq \#stars \leq 5.0$) can be approached from two angles: 1) a regression problem to predict on the decimal accurately or 2) a classification problem in which multiple bins are created (e.g. $4.5 \leq bin_n \leq 5.0$). Both approaches will be considered after which the most appropriate technique is determined.

2.2 Project Type

Although general internet marketing strategies have been extensively covered in literature, very little research has been conducted in the field of online education marketplaces. Existing papers mainly focus on the best practices for MOOCs, the usability of platforms its user interfaces and the challenges of MOOCs (e.g. high drop-out rates). In that sense, this project can be characterised as an exploratory rather than a literature study.

2.3 Project Scope

One's judgement about the quality of an online course is very subjective and related to infinite many factors. This project focuses solely on directly scrapeable attributes from a Udemy course page. In other words, essential characteristics of a well-designed course such as the quality of the sound, video, presenter slides and the overall teaching style will be out of scope. Neither external effects such as course promotion on social media by the teacher will be taken into account. For an overview of all available features see *Appendix A*.

Moreover, the dataset exclusively consists of Udemy courses from the following subcategories: Data & Analytics, Web Development, Mobile Apps, Programming Language, Game Development, Databases, Software Testing and Software Engineering.

2.4 Extension

As a possible extension to the project scope, text mining techniques can be applied to Udemy course reviews. This decision will be dependent on time availability as well as technical feasibility (it turns out scraping the course ratings and the corresponding reviews is more complex than for example course related features).

In particular, the following two questions can be interesting to investigate:

1. *To what extent can the star-rating of a course review be predicted based on the review description?*
2. *What phrases in a course review are related to a high and low star-rating?*

3 Learning Objectives

- Data mining end-to-end process, starting from translation of the business problem to data mining task(s)
- Transforming raw data to a representation that can be understood by data mining techniques
- Evaluation of data mining output, model performance optimization and avoiding overfitting.
- Comparing performance of different techniques and making valid conclusions about the performance of the models and their utility for addressing the identified business problem.

4 Deliverable & Grading

Similar to *Introduction to Data Science (JBP010)* the deliverable for this course will consist of a Jupyter Notebook which will be structured as follows:

1. **Problem Explanation**

2. **Data Collection**

   An explanation of which features have been scraped and their respective definitions.

3. **Data Cleaning**

   Turning the raw data collected in the previous step into useful features.

4. **Descriptive Analysis**

   Answers to descriptive questions 1 to 7.

5. **Predictive Analysis**

   An evaluation to what extent it is possible to predict the total number of students enrolled and the average course rating.

6. **Conclusion**

   Summarize main findings of descriptive and predictive analysis and elaborate on practical implications for students, teachers and MOOC-platforms.

As will become clear in the next section, a subset of above tasks (at least steps 1 to 4) will be submitted for feedback before the next part of the Introduction to Machine Learning course takes place (V1). One and a half week after that the final deliverable (steps 1 to 6) will be handed in which makes up 100% of the total grade for this course. Note that the deadline has been chosen in such a way that it is in line with the pace of the regular course stream.

5 Timeline

Deadline	**Activity**
3rd of November (Fri)	Start course Introduction to Machine Learning
5th of November (Su)	Send project proposal
26th of November (Sun)	Hand in preliminary version (V1)
5th of December (Tue)	Hand in final version (V2)

Appendix A - Scraped Features

Feature	**Definition**
course_name	The name of the course.
course_category	The category of the course (e.g. Data & Analytics).
course_level	An indication of the course-level (e.g. Intermediate Level).
teacher_name	The name of the teacher for a specific course.
last_updated	The most recent date on which the course has been updated by the teacher.
paid_course	Whether the course is paid or free.
original_price	The original course price in euro's.
course_language	The language spoken in the course.
captions	Whether (custom) captions are available.
num_lectures	The number of lectures in the course.
hours_video	The number of hours of video in the course.
num_articles	The number of articles in the course.
num_supplemental_resources	The number of supplemental resources the course.
num_students_course	The number of students enrolled for a specific course.
num_students_total	The number of students enrolled for all courses by a specific teacher.
best_seller	Whether the course has been classified as a best-seller by Udemy.
top_responder	Whether the teacher has been classified as a top-responder by Udemy.
course_rating	The average star-rating for a specific course.
course_rating_total	The average star-rating for all courses by a specific teacher.
num_reviews	The number of reviews for a specific course.
num_reviews_total	The number of reviews for all courses by a specific teacher.
num_courses_total	The total number of courses created by a specific teacher.