Business Intelligence Assignment 3

Group 40, Magdalena Breitenauer (51824556) (B)
e51824556@student.tuwien.ac.at

Group 40, Rastko Gajanin (11930500) (A)
e11930500@student.tuwien.ac.at

1 BUSINESS UNDERSTANDING

1.a Scenario

Udemy aims to identify the key characteristics of the most popular IT courses, such as popular titles, average rating, rating count, and price, that appeal to its customers to improve the course portfolio and enhance student retention. Course creators are interested in the potential revenue of their course. Thus, we want to predict the number of subscribers to be able to estimate revenues.

1.b Business objectives

The following business objectives will be considered:

(1) Identifying key features of popular courses: The characteristics and features of courses that are highly rated and have a large number of subscribers shall be analysed so that strategies for improving student retention can be developed. This way Udemy can improve it's course portfolio and increase customer satisfaction
(2) Analyzing the profit and course content (based on title) and extracting patterns that can be monetized
(3) Analyzing the most common terms in course titles
(4) Identifying the most popular courses and analyzing their prices and published time
(5) Analyzing the distributions of several important attributes (average rating, number of subscribers) and detecting patterns
(6) Number of subscribers prediction: Build a model that can predict the number of subscribers for a given course which can then later be used to estimate the sales profit for courses

1.c Business Success Criteria

The judgement of the fulfillment of these criteria should ideally be performed by Udemy's product management.

(1) Identify the courses that fall within the top 10% in both subscriber count and rating. Further examine the attributes of these courses
(2) Have a prediction for subscribers to estimate the approximate revenue for a course within 1000$ error range

1.d Data Mining Goals

(1) Examining the correlation between number of subscribers and other attributes such as price and rating
(2) Build a regression model to predict the estimated number of subscribers for courses
(3) Examining the total and average profit the courses made per year published

1.e Data Mining Success Criteria

(1) Successfully identify the courses accurately that are in the top 10% most subscribed and in top 10% best rated courses
(2) Predict the number of subscribers with MAE being at most 500 (i.e. On average the correct number of subscribers should be in the +-500 range of the prediction)
(3) Offer an overview of the most important discoveries defined in DM goals in a **visual** manner along with a discussion of each plot and

1.f AI Risk aspects

(1) Since the student data is not stored, there are no risks w.r.t. Privacy
(2) Potential Bias and Discrimination could be present for newer courses with no ratings and few subscribers. Furthermore, a great imbalance between the paid and free courses is present, which might also be a source of bias in the models

Moreover, it is important to ensure that the models and their predictions are transparent and explainable, so that stakeholders can understand how the predictions are being made and whether there is any bias in the system.

2 DATA UNDERSTANDING: DATA DESCRIPTION REPORT PRESENTING

2.a Attribute types and their semantics

Table 1 shows the data types of all attributes in the used dataset. Table 2 shows the most important statistics of all integer attributes for an overview of the attribute's value ranges and variances.

(1) id : The course ID of that particular course. (integer)
(2) title : Shows the unique names of the courses available under the development category on Udemy. (string)
(3) url: Gives the URL of the course. (string)
(4) is_paid : Returns a boolean value displaying true if the course is paid and false if otherwise. (bool)
(5) num_subscribers : Shows the number of people who have subscribed that course. (integer)
(6) avg_rating : Shows the average rating of the course. (float)
(7) avg rating recent : Reflects the recent changes in the average rating. (float)
(8) num_reviews : Gives us an idea related to the number of ratings that a course has received. (float)
(9) num_ published_lectures : Shows the number of lectures the course offers. (integer)
(10) num_ published_ practice_tests : Gives an idea of the number of practice tests that a course offers. (integer)

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Attribute	DType
id	int
title	string
url	string
is_paid	bool
num_subscribers	int
avg_rating	float
avg_rating_recent	float
rating	float
num_reviews	int
is_wishlisted	bool
num_published_lectures	int
num_published_practice_tests	int
created	datetime
published_time	datetime
discount_price__amount	float
discount_price__currency	string
discount_price__price_string	string
price_detail__amount	float
price_detail__currency	string
price_detail__price_string	string

Table 1: Attribute types

(11) created : The time of creation of the course. (datetime)
(12) published_time : Time of publishing the course. (datetime)
(13) discounted_ price_amount : The discounted price which a certain course is being offered at. (float)
(14) discounted_ price_currency : The currency corresponding to the discounted price which a certain course is being offered at. (string)
(15) price_ detail_amount : The original price of a particular course. (float)
(16) price_ detail_currency : The currency corresponding to the price detail amount for a course. (string)

Note: Composite columns such as price_detail_string and the *is_wishlisted* column which contains the same value for all observations have not been described.

2.b Statistical properties describing the dataset including correlations

The dataset contains around 22k observations of 20 variables, with the variable types being mostly numeric. Majority of these numeric variables are heavily right skewed, meaning they are very imbalanced. Further category imbalance can be seen in the *is_paid* variable where only 2% of the observations have category *False*, while the rest have *True*. After analyzing the datetime variables, we observe that the majority of the courses are created and published between 2017 and 2020. A further elaboration on these properties is contained in the Subsection 2.d.

2.c Data quality aspects

2.c.1 Completeness. Only the metadata of the courses themselves is contained in the dataset. Further exploration and data mining could be facilitated by including information about the students

Attribute	Count	Mean	Std	Min	Max
id	22853	1818466	927352	2762	3486006
num_subscribers	22853	3205.44	11051	0	564444
avg_rating	22853	3.95	0.875	0	5
avg_rating_recent	22853	3.937	0.888	0	5
rating	22853	3.937	0.888	0	5
num_reviews	22853	270.28	2048.8	0	188941
num_published _lectures	22853	34.92	48.652	0	699
num_published _practice_tests	22853	0.375	1.160	0	6
discount_price _amount	21024	486.26	234.100	455	3200
price_detail_ amount	22356	4445.518	3098.531	1280	12800

Table 2: Attribute statistics

(such as age, employment type, country etc.), number of students that finished the course or number of students that failed. However these properties would have implications on privacy and potentially introduce bias.

2.c.2 Timeliness. The dataset contains very recent information about the courses (namely in the time frame from 2010 to 2020), which is a good quality sign.

2.c.3 Relevancy. The information contained in the dataset can be considered *marginally relevant* for the analysis task at hand. There is plenty of room for extending the dataset with further data which could enhance the analysis results.

2.d Visual exploration of data properties and hypotheses

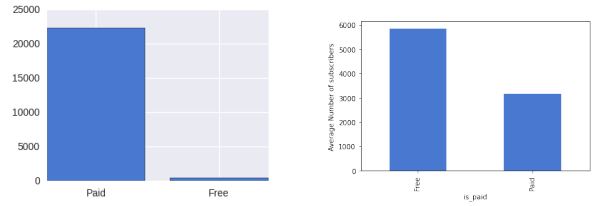
2.d.1 Correlation between the date and course popularity. We expected that there should be a correlation between the course creation date and number of subscribers. This however is not the case as visible in the correlation matrix heatmap in figure 2.

2.d.2 Unique (identifier) columns. The columns *id*, *topic*, *url* are unique for each observation and hence will not further be considered.

2.d.3 Constant column. Column *is_wishlisted* is constant (has no variance) and will therefore be discarded.

2.d.4 Paid vs Free courses. We observe a great imbalance in the *is_paid* attribute. Figure 1a shows that the amount of paid courses in the dataset is significantly higher than free courses. But figure 1b shows that there are on average more subscribers for free courses.

2.d.5 Correlation matrix. The correlations shown in the heatmap (Figure 2) are built only with numerical attributes and certain derived attributes. It should be noted that the Spearman rank correlation was used to measure the monotonic dependence between the variables instead of just linear dependence. We can observe that the majority of variables are uncorrelated, which means that features



(a) Distribution of the *is_paid* attribute

(b) Average subscribers for paid vs. free courses

Figure 1: Paid vs free courses.

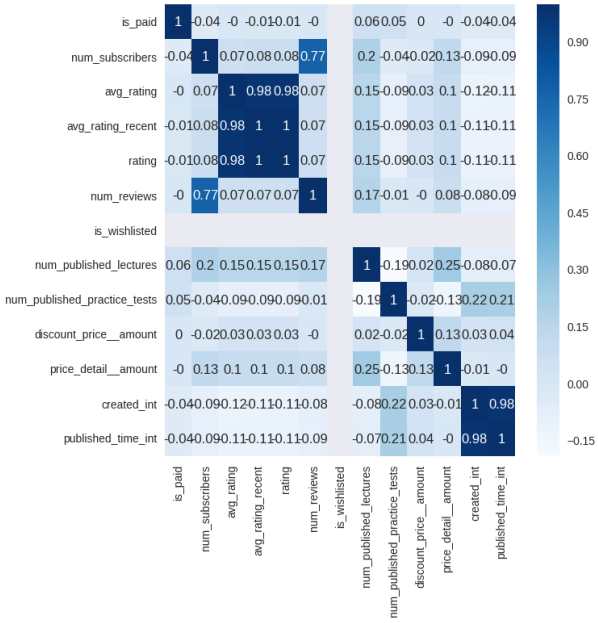


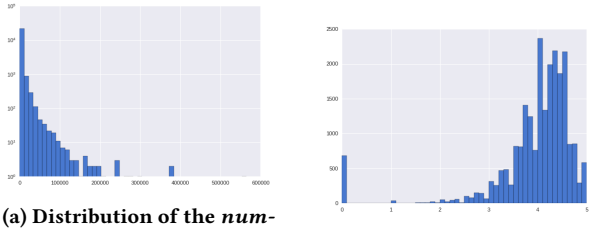
Figure 2: Heatmap of the Correlation matrix

are not *linearly* dependant on each other. Other forms of dependence (e.g. quadratic or monotonic) could further be inspected by using other metrics.

An unexpected occurrence is that the number of subscribers is not at all correlated to the average rating of the course.

The obvious high correlations such as the correlation between created and published date or the correlation between rating and avg_rating_recent indicate that these pairs of variables carry almost identical information. meaning one of them can be discarded.

2.d.6 Distributions of the features of interest. From the Figure 3a we observe that it is heavily right skewed. This will have to be addressed during the data preprocessing and modeling phase. On the other hand, Figure 3b shows that the average rating tends to be bell shaped with a reasonable amount of outliers at zero. An explanation for this could be that there is still plenty of courses with no ratings and therefore the average rating is set to zero.

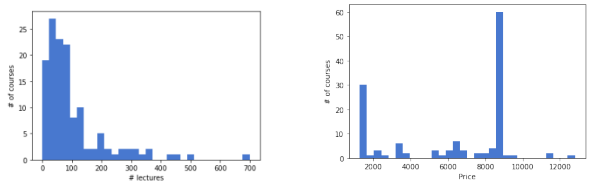


(a) Distribution of the *number of subscribers* variable (*y is represented on the log scale*)

(b) Distribution of the *average rating* variable

Figure 3: The distributions of the variables of interest

2.d.7 Most popular courses. To delve deeper into the characteristics of popular courses, we focused on those with both high subscriber counts and high ratings. We chose courses that fall in the top 10% for both subscriber count and rating. Figure 4a shows the amount of lectures the most popular courses have and figure 4b shows their prices. Figure 5 shows in which years they have been published.



(a) Amount of published lectures

(b) Prices of most popular courses

Figure 4: Most popular courses characteristics

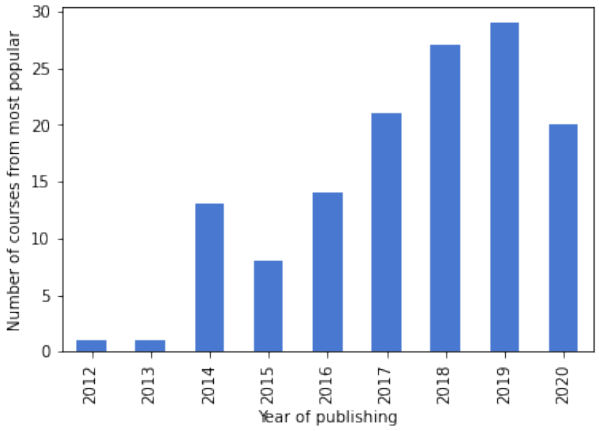
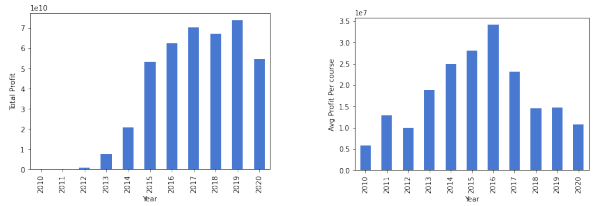


Figure 5: Most popular courses per year

2.d.8 Profit. We have analysed the profit which was constructed as product of number of subscribers and course price. The profit per year can be seen in figure 6a in total and in 6b in average. We can observe that the peak of the highest average profit is of courses

which were published in 2016. This is also interesting because we have seen before in figure 5 that the most popular courses were mostly published in 2018 and 2019. We have found out that some of the courses in 2016 that are famous are more expensive than on average.



(a) Total profit per year **(b) Average profit per year**

Figure 6: Profit of courses per year

2.d.9 Analysis of titles. For analyzing popular topics of the Udemy platform we have extracted the most common terms in the titles which are presented in figure 7. Business-related terms ("Business", "Management", "Trading") are very common which we didn't expect because the name of the dataset implies that it contains only IT courses. But also in IT those terms occur often because it's part of some e.g. Business Intelligence and API Management.

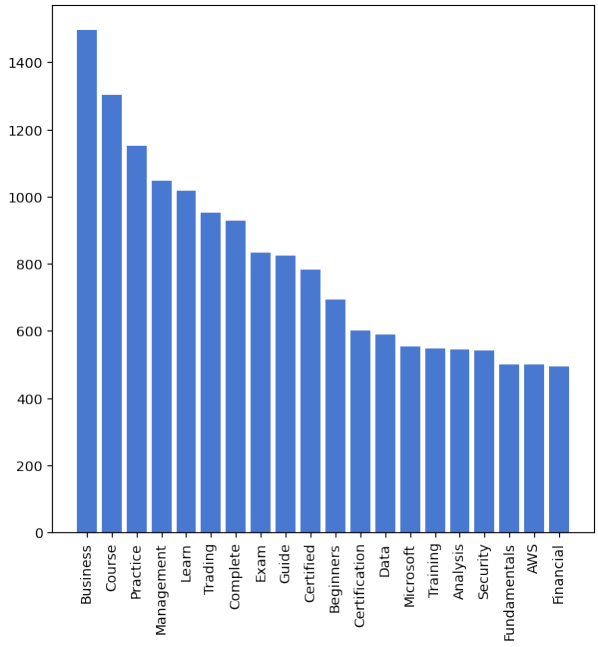


Figure 7: Most common terms in titles

2.d.10 Missing values. The missing values are only present in the price related attributes: price detail amount/currency/string and discount price amount/currency/string. After further investigation it has been established that there are two causes for this occurrence:

(1) The course is free, which implies that it has **no** price/discount price information
(2) The course is not free, **but** was never on discount, meaning there is no discount price information available.

2.d.11 PCA. After encoding the data and filling the missing values, PCA was carried out on the training set. This revealed that 65% of the variance was explained by the first 5 components and only 30% with the first one. The first two PCs "focused" mostly on the integer representation of the date related columns, which was not very surprising since these attributes have the largest number of unique values and therefore contribute the most to the general variance of the dataset.

2.e Ethically sensitive, minority classes or underrepresented data groups and potential bias

The used dataset does not contain any ethically sensitive properties because it does not include any user data but only meta data about the courses themselves. One potential bias though could be the date because recently added courses might be ranked lower.

2.f What potential risks and additional types of bias exist in the data? What questions would you need to have answered by an external expert in order to determine potential bias or data quality issues?

One potential bias would be the one towards paid courses. There could be discrimination against free courses since they are a minority class in this dataset.

From the data description it is not clear if some courses are getting updated (e.g. just by moving the published date) or a new course is created for a new semester. (as there are multiple title duplicates which have a created and published date once in October and once in January)

3 DATA PREPARATION REPORT

3.a Derived attributes

3.a.1 days_since_last_published. The attribute how long a course has a course already been published we calculate from the difference between the published date of the course and the max published date because we suppose that's the newest data we have.

3.a.2 is_beginner. We want to see the influence in the prediction whether the course is a course for beginners because we suppose that there will be more subscribers for beginner courses. Therefore all courses are marked as beginner courses which contain one of the following terms in the title:

(1) beginner
(2) fundamental
(3) basics
(4) start
(5) intro
(6) 101
(7) how to

(8) foundation.

3.a.3 is_advanced. We have decided for the same reason as for beginner's courses to also look at advanced courses. We are trying to get the advanced courses by looking for courses which contain one of the following terms in the title:

(1) advanced
(2) master
(3) expert
(4) professional

3.b External data sources

Some more meta data of the courses would have been interesting, the most important ones being:

(1) Content update frequency and last updated timestamp: For our scenario it would be beneficial to know how often a course is updated and the timestamp of the last update. It was not obvious for us whether or not the published date gets updated so if it's the last-published-timestamp or the first-published-timestamp.
(2) Number of subscribers over all days since the course has been published (or subscription-timestamps per course): This would have been interesting to analyse trends and predict how long a course is usually successful and when the peak is.
(3) Category (e.g. Web development, DevOps, Security,..): Some categories are more popular than others.
(4) Existence of course description or agenda: A binary attribute like "agendaProvided" would have been interesting to see if people are rather taking courses where an agenda is provided.
(5) Level of the course: It would be interesting to see if beginner courses are subscribed to more often. As we suspect this is the case, it could influence the prediction. It was possible to get it out of the title for most of them but it's not fully correct since some course titles are like "From beginner to pro", which is a full-level course and not a beginners-only.
(6) Course teacher: This could introduce some bias of course but just having the teacher ID in the model could have been an option
(7) Course format (does it include reference code on GitHub, are books or handouts provided, etc.)
(8) Course edition-number
(9) Number of times the course was on sale (discounted).

One - not recommended - option to get the meta data would be to crawl the Udemy website and extract further information about the courses. The open API from Udemy (Affiliate API) unfortunately is no longer supported and will be deprecated soon. However, to create the authentication credentials you had to create your API client. If the API would have been still supported it would have been interesting if we would have gotten some of the missing metadata and how much of a difference it would have made for the prediction.

Additionally, it could be interesting to know which other courses students of a particular course have registered for to show individual recommendations for users. This way we could perform customer segmentation and show customers other courses they might be interested in because of correlating properties or with association rules showing them courses that are often subscribed to in addition of another one. However, this is not relevant for our goal.

3.c Other pre-processing steps

3.c.1 Outlier removal.

- We have found 4 courses without subscribers but for some suspicious reason they have a rating (4 out of 215 without subscribers). Those 4 will be removed.
- The dataset includes 1 course which is listed as paiddiscourse (not free) but it doesn't include a price and the currency is NaN as well. This course will be removed as well.
- Another outlier is a course that has a published date before the created date.

3.c.2 Disregard outdated data. We have decided to focus on recent data and disregard coursed that were published before 2015 since we consider that data as irrelevant to predict the behavior of future customer interactions.

3.c.3 Splitting. In order to reduce the chance of accidental data leakage, the original dataset was split at this point into train, validation and test sets. Each of the subsequent steps was performed on each dataset separately.

3.c.4 Encodings.

- Date values such as *published_time* have been encoded with their year and month
- Boolean values such as *is_paid* have been converted to int (False -> 0, True -> 1)

3.c.5 Attribute removals.

- Highly correlating attributes removal: The three rating attributes are highly correlating (see figure 2) or almost identical so in consequence two of them can be removed. As we are only looking at the last few years it doesn't really make a difference anyway. We have selected the *rating* attribute
- Constant attributes removal: The two currency attributes *discount price currency* and *price detail currency* contain only two unique values: INR and NaN. The dataset has some NaN values because some courses are for free and not all have a discount. Thus we can disregard the currency attributes. We have also thought about converting the values to another currency (EUR or USD) to have a better understanding for the values but for training the model that's not relevant. The attribute *Is_Wishlisted* attribute also has no variance and will therefore be disregarded.
- Unique attributes removal: The attributes *id*, *title* and *url* are unique and can be removed.
- Price string attributes removal: The two attributes *discount_price__price_string* and *price_detail__price_string* can be disregarded because they are just the join of the price/discount value and the currency.
- Is_Wishlisted attribute removal: The attribute has no variance and will therefore be disregarded.

3.c.6 Setting the missing values. As discussed in the data description report, there are two possible causes for missing values. Each of them will be treated as follows:

(1) Free courses: The price and discount price will be set to 0.
(2) Paid course that was never on discount: discount price will be set to the original price. (I.e. no discount)

3.c.7 Scaling. Z-Score normalization was used on all columns to bring them to a zero mean and unit variance. This is needed because the distance based models were used. The Sklearn's standard scaler was fit on the training data and used for transforming all three datasets. Scaling the test and validation sets separately would cause a shift in the representations.

3.d Final dataset

In the following table 3 the first 6 attributes are unchanged from the initial dataset. The other 10 listed attributes have been either derived or in some way encoded or converted.

Table 3: Final dataset including attribute types

Attribute	Dtype	Origin
rating	float	initial
num_reviews	int	initial
num_published_lectures	int	initial
num_published_practice_tests	int	initial
discount_price__amount	float	initial
price_detail__amount	float	initial
is_paid	int	converted
created_int	int	converted
published_time_int	int	converted
is_beginner	int	derived
is_advanced	int	derived
days_since_last_published	int	derived
pub_year	int	encoded
created_year	int	encoded
pub_month	int	encoded
created_month	in	encoded

4 MODELLING

4.a Model selection

Since the nature of our business goal requires insight into the dependencies between the features, the explainability of the model will play the major role in model selection. In the following we consider the three algorithms that offer a relatively high degree of interpretability. We also provide justifications why each should or should not be used.

4.a.1 K Nearest Neighbors. Even though KNN is a simple model that is easy to understand it relies on one assumption that is in practice generally false: Namely, that all features are equally important. This means that even if the feature is not relevant for a prediction, it still has influence on it. Therefore we do not find this model to be appropriate for the task at hand.

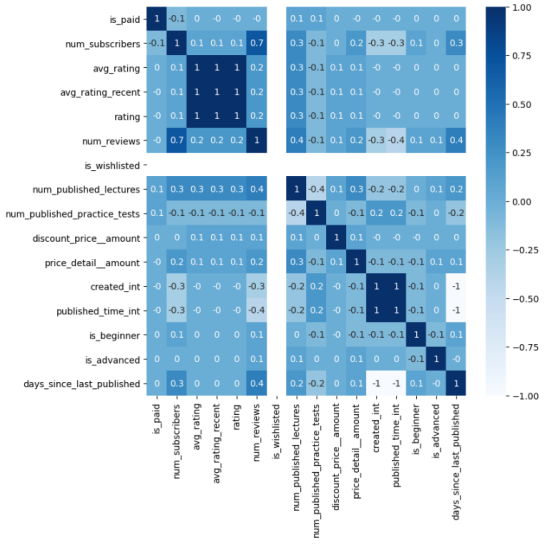


Figure 8: Heatmap of the Correlation matrix after data preparation

4.a.2 Linear Regression. Since the variables are in general not linearly dependent on each other (i.e. have low correlations) and many are skewed, the linear regression model would probably not be a good choice of the modeling algorithm because it assumes normal distribution of the data, even though it is considered as one of the models that are easily interpreted. One could also log transform the skewed variables and attempt to use this model then. However, since the features are generally not correlated with each other, we chose not to continue with this model.

4.a.3 Random Forests. As a third option we considered Random Forests. To get insights about how features influence the target variable we can analyze feature importance of the trained model. Furthermore, by their construction Random forests are robust against skewed and imbalanced features, which is a desired property w.r.t. this dataset.

4.a.4 Selection. The Random forests and KNN were fit with the following values of hyperparameters (for random forests the optuna framework was used and for KNN the GridSearchCV class provided by sklearn library and both hyperparameter configurations the error was calculated with 5-fold-cross-validation.):

- **RFs** - Criterion mse Number of trees between 50 and 200, maximal tree depth between 1 and 15, minimal number of samples needed for a split between 10% and 50% of all observations. At each iteration optuna used the mean of MAEs of all five folds. Best model has following hyperparameters: { 'n_estimators': 77, 'max_depth': 4, 'min_samples_split': 0.117 } and validation MAE **2621.19**
- **KNN** - Number of neighbors tried: 3,5,9,11,15. GridSearchCV used negative MAE as scoring criterion and five folds in CV. Best model has 11 neighbors and validation MAE **2897.79**

Linear regression model was fitted with default parameters (of which relevant for reproducibility: fit intercept = True, positive =

False) LinearRegression class (sklearn). This model had the following MAE on the validation set: **2731.15**

After fitting all three models with different values of hyperparameters, we established that the Random Forests and Linear regressor achieved the lowest MAE. Since interpretability of the models is the priority in this project, we have chosen the Linear regressor as the final algorithm, since it achieves comparable performance as the RFs with higher degree of interpretability.

4.b Hyperparameters

Since the linear regressor has been chosen as the most suitable model, it has no standard hyperparameters in it's base form. Furthermore the default parameter setting described above produced the coefficients presented in Figure 9.Therefore, we used LASSO regularization to remove any noisy attributes that might affect the model. We furthermore obtain "cleaner" coefficients of the linear model which are easier to interpret.

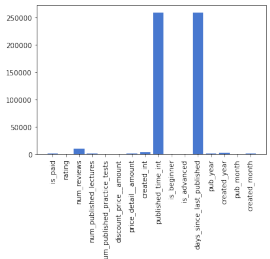


Figure 9: Attribute coefficients obtained with LR default settings

4.c Splitting the data

20% of the observations was dedicated for testing (as a test set). Since this dataset has a temporal dimension, the testing observations are **the most recently published observations.**

Of the remaining 80% a random third of the observations was used as a validation set and the rest was used as a training set. With that we have the 60-20-20 split rule.

The random splitting was performed with sklearn's train_test_split method and parameters: test_size=0.33, random_state=45.

4.d Training

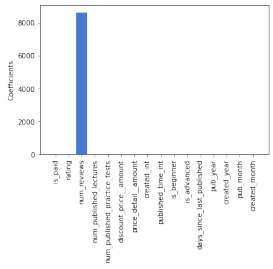
4.d.1 Baseline. Before training a baseline was established by predicting the target values of the validation set (y_valid) as the mean of all target values of the training set (y_train).

The baseline MAE achieved in this manner was: **4054.0**.

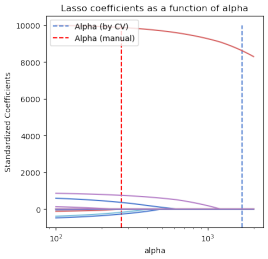
4.d.2 Hyperparameter optimization. Since the linear regressor has been chosen as the most suitable model, it has no hyperparameters in its base form. We will use LASSO regularization however to remove any noisy attributes that might affect the model.

The regularized model was fit with sklearns's LassoCV class. 10 folds were taken with maximum of 10k iterations and the random state set to 1. The optimal alpha parameter extracted from that process was **1678.07**. For that we only used the training set (without the validation set) and obtained MAE of **2931.15**. After examining the coefficients shown in Figure 10a of the model fitted with that lambda, we observe that only the *num_reviews* attribute is used. Since the best model found with CV ignores all attributes except num_reviews, we will select the alpha parameter manually to have at least 4 coefficients that are not zero. In order to select the lambda the MAE path plot (Presented in Figure 10b) of the LASSO regularization was inspected and we agreed upon setting the alpha to **270**. The plot also shows the position of the alpha parameter selected by CV.

After fitting a new model with manually selected alpha we reached MAE of **2751.63** and the coefficients shown in Figure 11a.

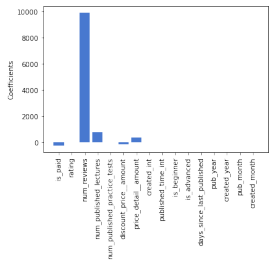


(a) Attribute coefficients of the LR model with alpha set to 1678.07 (found by CV)

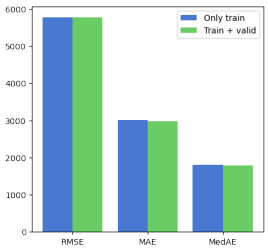


(b) LASSO MAE Path with two alphas marked

Figure 10: LASSO regularization



(a) Attribute coefficients of the LR model with alpha set to 270 (set manually)



(b) Performance of the two final models

Figure 11: Manual alpha coefficients and final model performance

4.e Results

From the coefficients shown in Figure 11a we can conclude that *num_reviews* attribute again has the highest significance. This is expected because we observed a relatively high correlation (around *0.7*) between that attribute and the target variable. On the second place w.r.t. magnitute of coefficients is the number of published lectures. Given that this attribute has a positive coefficient, the higher number of published lectures indicates a higher number of subscribers. Further non-zero coefficients show us that price has a

slightly positive impact on the prediction and the status "is paid" a negative impact. This is somewhat contradictory and might be an artefact of encoding the "is paid" variable and then scaling it, along with the fact that it is imbalanced. Finally, the magnitude of discount price amount coefficient is very small compared to other coefficients so it may be set to zero without affecting the model performance significantly.

5 EVALUATION

Three evaluation metrics were taken for this step.

(1) RMSE : Because it is the standard in majority of regression problems. It is however sensitive to outliers.
(2) MAE : Since RMSE is sensitive to outliers, MAE was chosen as a second option.
(3) MedAE : MAE is however not as robust against extreme outliers as MedAE, which is why this metric has also been employed.

5.a Training the final model

Final model is set to a Linear regression model with following parameters: alpha=270, fit_intercept=True, normalize=False, precompute=False, copy_X=True, max_iter=1000, tol=0.0001, warm_start=False, positive=False, random_state=None, selection='cyclic'

After retraining the final model **firstly only with training set** and **then with training and validation sets merged** and measuring performance with three different metrics: **RMSE, MAE, MedAE** the models achieved performance depicted in Table 4 resp. in Figure 11b

Metric	Only training data	Training + Validation data
RMSE	5776.0	5774.0
MAE	3009.0	2981.0
MedAE	1808.0	1782.0

Table 4: Final result scores

From the figure it can firstly be observed that both models achieve comparable performance. Since there are many outliers in the dataset we see that MAE and especially MedAE have lower values in comparison to RMSE.

5.b Performance of models fitted on the similar dataset from other sources

5.b.1 State-of-the-art performance. The used dataset has seen limited utilization. Our investigation revealed a single notebook, titled "Overview - Best rated courses," which appeared to primarily explore the data rather than make predictions. We have also looked into notebooks of closely related datasets.

We have found one conference paper of July 2022 by Lin, L. Zachari S., Dragan G. and Guanliang C. called "Popularity Prediction in MOOCs: A Case Study on Udemy" but the dataset used there was collected from the API, which was available back then, so it has more attributes. Therefore it includes a lot of other important features such as the availability of captions in various languages. We realize it's difficult to compare but as mentioned our chosen dataset hasn't been used in published articles.

5.b.2 Expected base-line performance. As a baseline we took predicting test targets as mean of targets of training and validation sets merged The performance this baseline achieves is: MAE = **3905.32**, RMSE = **6337.0**, MedAE = **3301.0**.

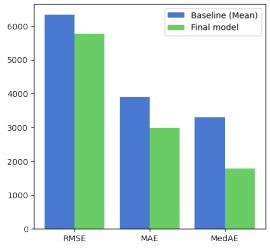
5.c Comparisons

5.c.1 Comparison with state-of-the-art performance. The prediciton of the published analysis has a RMSE of about 2 which is - needless to say - way better than our result. However, it's hard to compare them because we have used a simplier dataset. That's why we were especially interested in the most relevant features of this analysis which are shown in Figure 12. As visualized there, we notice that a lot of the important features for this prediction were just not available in our used dataset.

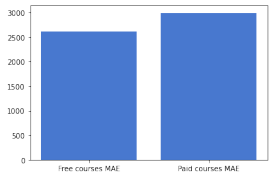
Feature category	Feature input	# Avg. monthly reviews		# Avg. monthly enrolments	
		RMSE	R^2	RMSE	R^2
	All feature	**1.9057**	**0.4377**	**2.7235**	**0.3576**
Content	w/o Title & Headline	* 1.9114 (−0.18%)	* 0.4343 (−0.78%)	* 2.7397 (−0.59%)	**•0.35(−2.13%)**
	w/o Content description	**•1.9266(−0.65%)**	**•0.4253(−2.83%)**	* 2.7352 (−0.43%)	* 0.3521 (−1.54%)
	w/o Target audience	* 1.9080 (−0.07%)	* 0.4363 (−0.32%)	* 2.725 (−0.06%)	* 0.3569 (−0.2%)
	w/o Prerequisites	1.9018 (0.12%)	0.4400 (0.53%)	* 2.7306 (−0.26%)	* 0.3543 (−0.92%)
	w/o Objectives	* 1.9060 (−0.01%)	* 0.4375 (−0.05%)	2.7201 (0.12%)	0.3592 (0.45%)
Structure	w/o # Lectures	* 1.9131 (−0.23%)	* 0.4333 (−1.01%)	* 2.7291 (−0.21%)	* 0.355 (−0.73%)
	w/o Lecture duration	* 1.9074 (−0.05%)	* 0.4367 (−0.23%)	* 2.7266 (−0.11%)	* 0.3562 (−0.39%)
	w/o # Articles	* 1.9143 (−0.27%)	* 0.4326 (−1.17%)	* 2.733 (−0.35%)	* 0.3532 (−1.23%)
	w/o # Assessment tests	1.9044 (0.04%)	0.4385 (0.18%)	2.7192 (0.16%)	0.3597 (0.59%)
	w/o # Practice tests	1.9036 (0.11%)	0.4398 (0.48%)	2.7194 (0.15%)	0.3596 (0.56%)
	w/o # Coding exercises	1.9047 (0.03%)	0.4383 (0.14%)	2.7232 (0.01%)	0.3578 (0.06%)
	w/o # Questions	* 1.9080 (−0.07%)	* 0.4363 (−0.32%)	2.7207 (−0.1%)	0.3592 (0.45%)
	w/o # Additional resources	* 1.9084 (−0.08%)	* 0.4361 (−0.37%)	2.7228 (0.03%)	0.3580 (0.11%)
Metadata	w/o Subject category	1.9029 (0.08%)	0.4393 (0.37%)	* 2.7269 (−0.12%)	* 0.356 (−0.45%)
	w/o # Instructors	1.9120 (−0.19%)	0.4340 (−0.85%)	**•2.7442(−0.76%)**	**•0.3479(−2.71%)**
	w/o Enrolment fee	**•1.9408(−1.10%)**	**•0.4168(−4.77%)**	* 2.7349 (−0.42%)	* 0.3523 (−1.48%)
	w/o # Captions	**•1.9445(−1.22%)**	**•0.4145(−5.30%)**	**•2.742(−0.68%)**	**•0.3489(−2.43%)**
	w/o Published date	* 1.9066 (−0.03%)	* 0.4372 (−0.11%)	* 2.7316 (−0.3%)	* 0.3538 (−1.06%)
	w/o Published year	1.9028 (0.09%)	0.4394 (0.39%)	2.7235 (0.0%)	0.3576 (0.0%)
	w/o Published_month	1.9047 (0.03%)	0.4383 (0.14%)	2.7233 (0.01%)	0.3577 (0.03%)
	w/o Instructional level	1.9027 (0.09%)	0.4394 (0.39%)	2.7226 (0.03%)	0.3581 (0.14%)
	All features marked *	1.9057	0.4377	2.7185	0.3600

Figure 12: Feature importance analysis of the state-of-the-art performance

5.c.2 Comparison with other notebooks. The notebooks which are working with a similar dataset we do came all to the same conclusion that utilizing a Random forests model is the most appropriate method for predicting the number of subscribers and they had a comparable RMSE result.



(a) Comparison of the Baseline (calculated as mean of training target column) vs Final model



(b) MAE on free courses vs on paid courses in test set

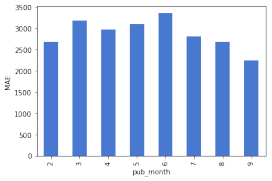
Figure 13: Comparison with the baseline (left) and Performance on protected attribute (right).

5.c.3 Comparison with baseline. In all three metrics we see an improvement over the baseline. It is noticeable that greater performance gain is present in MedAE metric. This implies that model achieves a significant improvement over the baseline even with many outliers present.

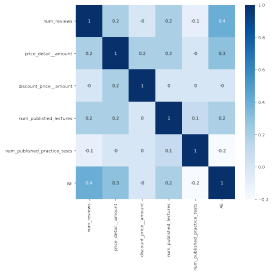
5.d Regression errors in different parts of data space

5.d.1 MAE vs publication month. In the Figure 14a the MAE is plotted for each month in the test set. We can notice that for all months we have similar MAEs. The lowest MAE is achieved for the last month. This is probably due to the absence or lower ratio of outliers in that month, which is caused by the lower number of samples.

5.d.2 Absolute error vs numerical attributes. The correlation matrix of the dataset that contains absoulte error made by the model for each observation of the test set is shown in the Figure 14b. The important numerical columns have been extracted (i.e. no date and no boolean columns). In the correlation matrix we observe that the highest correlation exists between AE and number of reviews. i.e. the higher the number of reviews the higher the MAE. We explain this by the fact that the model heavily relies on this attribute and it's results are therefore moderately correlated with it.



(a) MAE Per month in test data.



(b) Correlation matrix containing important numerical attributes and absolute error(AE).

Figure 14: Comparison with the baseline (left) and Performance on protected attribute (right).

5.e Success criteria fulfillment

(1) Criteria 1 (Features of popular courses): Identifying characteristics of popular courses was successful. We have seen that courses with 20 to 80 lectures are popular and that the price of the most popular courses is either very low at around 1000 INR which complies to about 11 EUR or more advanced and popular courses cost around 9000 INR which complies to about 100 EUR. We have also analysed frequent terms in popular courses and e.g. found that a lot of them are beginner-level courses.
(2) Criteria 2 (Subscriber prediction): Predicting the number of subscribers with MAE less than 500 was unfortunately not accomplished by the model. Since the performance of other models with a similar dataset is similar to the performance of our model, one can claim that the dataset simply does not have that much predictive power in general.
(3) Criteria 3 (Visualization of data exploration) The contents of Subsection "Visual exploration of data" contribute the most to this goal. All plots provided in that subsection and generally in Section "Data understanding" offer a visual view of the examined properties followed with a brief description, discussion and potential explanation of the occurrences

5.f Identification of "protected attribute"

As previously mentioned the protected attribute is the **is_paid** column. To examine if the models shows bias towards paid courses we calculated MAE for both classes separately in the test set. The estimation can be considered reliable because there are 161 free courses in the test set.

The resulting MAEs are presented in Figure 13b. From the figure we observe that the model actually achieves **slightly better** performance on the **minority** class (Free courses), which is a preferred characteristic and might imply that the bias is not as pronounced. On the other hand it might be the case that due to a larger number of samples belonging to the paid class, there are more outliers present which are misclassified and hence increases the MAE for that group.

6 DEPLOYMENT

6.a Business objectives fulfillment

The information gathered on the most in-demand courses can be utilized to enhance student retention and increase sales revenue in the future. Due to not meeting our success criteria we do not recommend deploying the model for predicting the subscriber count. We recommend trying it out with other datasets of udemy courses or trying to get the data that was used in the article of the "state of the art" comparison.

6.b Ethical Aspects

In the evaluation phase it has been discussed on the bias that model exhibits towards the majority class and has been established that the bias is insignificant, if present at all. Therefore data privacy regulations and securing the model and data against unauthorized access isn't as important as for others but nevertheless it's a good practise.

6.c Monitoring

The performance of the deployed model should be regularly monitored to ensure it continues to meet requirements and to detect any drift in its predictions. To improve the model's accuracy a feedback mechanism could be provided for users to provide feedback on the model's predictions. The model should be retrained on new data to improve its accuracy and adapt to changing data distributions.

It is recommended that the ratio between MAE of the free courses and MAE of the paid courses remains similar because this attribute contains a minority class. If a case occurs that free courses have

MAE more or equal to 1.5 times MAE of the paid courses, an intervention should be carried out in order to prevent further degradation of the performance wrt minority class.

6.d Reproducibility

Since reproducibility plays an important role in this project, special focus was placed on the seeding of randomized methods and documentation not only of used parameters but also implementations. Also, make sure the model is transparent and its predictions can be easily explained, to build trust and enhance decision-making.

7 SUMMARY

Through this project we managed to get an insight into the data mining/analysis process that is somewhat closer to the industry standard when compared to the projects of other courses. Two major tracks were pursued almost concurrently throughout the assignment: **Implementation** and **Documentation** track. We believe that, through this assignment, we learned to *exhaustively* analyse a relatively simple dataset and get an understanding of the topics relevant for a typical State of the art data mining process. Finally, when choosing a model the focus was heavily shifted to **interpretability** and **explainability** of the results rather than performance. Unfortunately we weren't able to add more features because the API is deprecated but in the end when we compared it to published analysis of Udemy courses we saw that a way better prediction could be made with better features.