Introduction

For this project, the head of Data Science for the BBC, Gabriel Straub, tasked us with analysing viewer data for their online viewing platform. This project was undertaken as part of London Business School's Data Mining Class. I completed it working in a team of three with two classmates - Julia Schmitt and Giuseppe Baldini. Our task was to categorise viewers into different segments, and then create an algorithm which predicted how long viewers would watch in the upcoming month.

1) Customer Segmentation Analysis.

First of all, we created a lot of new variables for the data to better segment the customer basis, e.g. time spend watching shows, percentage of nearly completed shows, time of the day or whether a person watches during the weekend.
Then, we created a segmentation of the customer base with different amounts of clusters and variables. Based on the sum of squared errors for each of the numbers of clusters, thus the "Elbow-Curve" (see Exhibit 1), we decided that the optimal number of clusters is 5 as more would not add much more explanatory power and also would make the segmentation more confusing. Also, we thought it was important to be able to easily explain each cluster as a distinctive customer group. If two clusters were not distinguishable, then there were too many – a method that also resulted in an optimal number of 5 clusters/customer segments.
The variables that were used are "Time spent watching each genre", "Number of total shows watched", "Total time watched", "Time of the day", and "Engagement in shows (watched more than 60% or 20 minutes)".
This resulted in the following clusters (see Exhibit 2):

- Segment 1: Diurnals (1173 Observations)
- Segment 2: Engaged Watchers (564 Observations)
- Segment 3: Sundowners (1628 Observations)
- Segment 4: Night-Owls (931 Observations)
- Segment 5: Power-Watchers (118 Observations)

Segments 1, 3 and 4 are very similar based on the time watched in each genre and in total. Their level of engagement is also relatively similar. The major differences between those segments are the time of the day in which the representatives of the segments watch the shows and the total time they watch shows. Segment 1 are heavily watching during the day, Segment 3 in the evening and Segment 4 are most active during the night.
Segments 2 and 5 don't really differentiate themselves by the time of the day they watch TV but rather in the amount of time they spend on the iPlayer. Segments 2 and 5 are watching much more TV than the other three segments – their total time watched and number of shows is much higher, especially for Segment 5.
The time watched in each genre also differs between those segments: Segment 2 customers watch much more Children, Comedy, Drama, Entertainment and Factual than Segments 1, 3 and 4. Segment 5 are watching every single genre much more than any other segment, except for Children and Learning which is on the same level as Segment 2.
Another factor that differentiates Segment 2 from the rest is the level of engagement which is much higher than in the other segments.
Power-Watchers watch on average 33 hours throughout the January-February dataset. This is feasible for dedicated people, suggesting that the data is not erroneous.

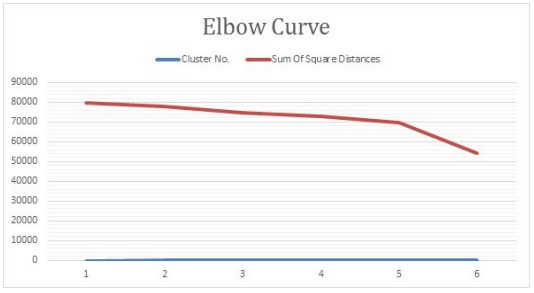

Exhibit 1: Elbow curve (Cluster No. 1-5: 2, 3, 4, 5, 6 Clusters, Cluster No. 6: 20 Clusters)

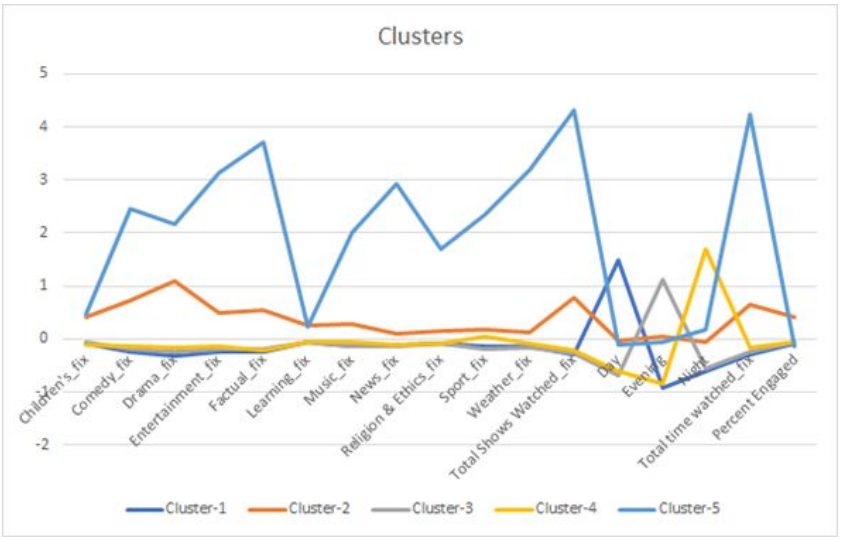

Exhibit 2: Customer Segmentation

2) Predictive Modelling

As we want to predict the views per customer based on whether the respective customer will watch at all in the following month, we first need to predict if a customer will watch or not. Then, we can take these predicted "watching customers" and perform the prediction on those. It would not make sense to predict the views per customer for those who wouldn't even watch in the following month. Therefore, doing the classification first saves time and effort and makes the prediction more accurate and efficient.

For the classification analysis, we used a logistic regression model as the output has to be either 0 or 1 (0 for non-watchers and 1 for customers who will watch the next month). Based on only the significant variables (variables with a p-value of <0.05), the classification model resulted as seen in Exhibit 3. As seen in Exhibit 4, the AUC for the Training- and Validation data are almost the same, indicating that the model is generalizable to data other than the training data. Thus, it is not overfitted. The sensitivity of the model (correctly predicted "Watchers") for the validation data is 58.5519% (cut-off point 0.5) and the specificity (correctly predicted "Non-Watchers") is 85.5478%. Depending on the cost of false predictions, the cut-off value could be set differently, changing the sensitivity and specificity of the model.

The model in Exhibit 3 shows that the more a customer watched shows in the genres on the iPlayer in January, the higher the likelihood that this customer is going to watch in February as well. However, the higher the total time viewed on the iPlayer actually reduces the likelihood of a person to watch in the next month. Interaction variables have not turned out to be significant.

Input Variables	Coefficient	Std. Error	Chi2-Statisti	P-Value	Odds	CI Lower	CI Upper
Intercept	-0.6104816	0.05719	113.94	1.3E-26	0.54309	0.4855	0.60751
Number of shows watched	0.15604449	0.01729	81.4847	1.8E-19	1.16888	1.12994	1.20916
Time in Children's	1.1362E-07	5.3E-08	4.58963	0.03217	1	1	1
Time in Comedy	1.6064E-07	4.7E-08	11.7269	0.00062	1	1	1
Time in Drama	1.258E-07	1.8E-08	51.45	7.3E-13	1	1	1
Time in Entertainment	1.5027E-07	4E-08	13.8426	0.0002	1	1	1
Time in Factual	9.5684E-08	2.3E-08	17.9292	2.3E-05	1	1	1
Time in Music	2.507E-07	7.1E-08	12.4365	0.00042	1	1	1
Time in Sport	1.3291E-07	2.9E-08	21.3169	3.9E-06	1	1	1
Sum of time_viewed	-1.038E-07	1.6E-08	40.6265	1.8E-10	1	1	1

Residual DF	2707
Residual De	3174.43
# Iterations	4
Multiple R^2	0.15026

Exhibit 3: Logistic Regression Model to Classify Watchers

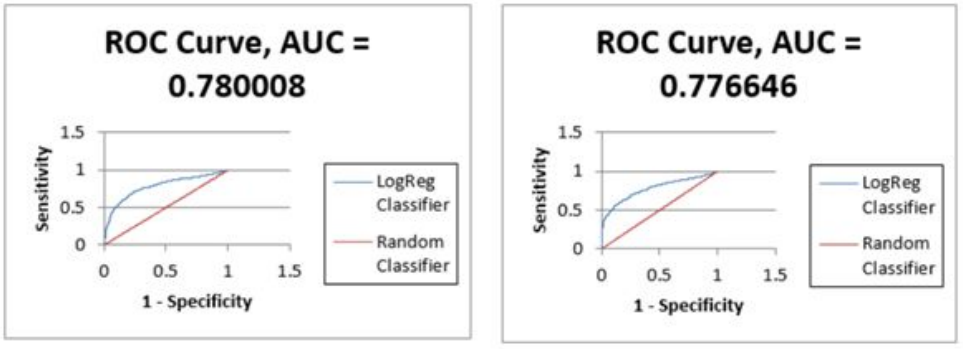

Exhibit 4: AUC for Training Data (left) vs. Validation Data (right)

Next, we want to see how many programmes and how many minutes each customer will watch in February, given that the customers are going to use the iPlayer. As we have the real data for the people who watch in February, we are going to use that as it is more accurate than our prediction and the prediction is also dependant on the chosen cut-off value that we cannot set due to not knowing the effect of false predictions.

First, we wanted to see how many shows the customers would watch, thus setting the output variable to “February Watchings” (Number of shows watched in February). Taking the insignificant variables out (p-value >0.05), we get the model that is shown in Exhibit 5 with an adjusted R^2 of 73.41% which seems to be predicting the use relatively well.

The variables that have a strong explanatory power on the number of shows watched in February are the times spent in each genre, the number of shows watched in January, how many times a person watched something in January as well as the total time spent on the iPlayer. The total time and the number of shows watched in January, however, have a negative effect on the number of shows watched in February (the higher those numbers, the fewer shows the customer is predicted to watch in the next month). Due to the expression of the times watched in milliseconds, those numbers are relatively high, thus the coefficients are correspondingly lower than for the number of times and shows watched in January.

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	2.473019748	0.301152937	8.21184	3.48818E-16	1.882478021	3.063561475	423898.5
Number of times watched in January	1.060727872	0.080353935	13.2007	1.75275E-38	0.903158923	1.218296821	993696.8
Number of shows watched in January	-0.455518047	0.104622282	-4.35393	1.39233E-05	-0.66067568	-0.250360416	2013.118
Time in Children's	3.90395E-06	8.10628E-07	4.815954	1.55464E-06	2.31436E-06	5.49354E-06	11245.88
Time in Comedy	3.01104E-06	8.11472E-07	3.71059	0.000211395	1.41979E-06	4.60228E-06	5461.027
Time in Drama	3.36988E-06	8.02684E-07	4.198262	2.78606E-05	1.79587E-06	4.94389E-06	387.1372
Time in Entertainment	3.51326E-06	8.04728E-07	4.36577	1.31955E-05	1.93524E-06	5.09128E-06	0.011516
Time in Factual	3.53297E-06	8.06123E-07	4.382671	1.22193E-05	1.95222E-06	5.11373E-06	1463.825
Time in Learning	6.21789E-06	8.9319E-07	6.961449	4.3121E-12	4.46641E-06	7.96938E-06	7465.193
Time in Music	3.31728E-06	8.25293E-07	4.01952	6.00836E-05	1.69894E-06	4.93563E-06	59.66233
Time in News	3.60666E-06	8.14948E-07	4.425638	1.00383E-05	2.0086E-06	5.20472E-06	4712.27
Time in Religion & Ethics	7.12945E-06	9.37334E-07	7.606085	4.00552E-14	5.29139E-06	8.9675E-06	8239.134
Time in Sport	3.31798E-06	8.04948E-07	4.121977	3.88172E-05	1.73953E-06	4.89643E-06	411.8671
Time in Weather	1.66508E-06	7.97359E-07	2.08824	0.036879619	1.01506E-07	3.22865E-06	9185.334
Sum of time_viewed	-3.37551E-06	8.02825E-07	-4.20454	2.71037E-05	-4.9498E-06	-1.80122E-06	2725.327

Residual DF	2441
R^2	0.73562
Adjusted R^2	0.734104
Std. Error Estima	12.41626
RSS	376312.9

Exhibit 5: Linear Regression Model to predict Number of Shows watched in February

Furthermore, we were able to raise the R^2 up by adding interaction terms of the variables (see Exhibit 6). After a couple of trial-and-error regressions, we found a model with similar variables (adding "% of nearly completed shows" and eliminating "Time in Religion") and 22 significant interaction terms between most of the variables that are in the model. However, the complexity of the model also goes up, but the increase in R^2 and Adjusted R^2 of about 5% can justify the increased complexity of the model.

However, there are limits on the applicability of these analyses. First of all, the data that was used was partly incomplete and had to be severely adjusted. There are also views which times or genre were not measured. Therefore, the data is not as exact as desired. Furthermore, there are only two months of data, whereas a longer timeframe of data could help uncover more unusual trends in customer behavior, especially as some customers were new to the iPlayer in the second month. Another issue could be that the data on the shows available on the iPlayer differs for each show due to the respective publication dates, i.e. a show could have been introduced recently and therefore has less views. Additionally, for logistic regression, it is possible that the model is not ideal. The data could be better suited to a pattern from a Decision Tree algorithm.

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	3.192676015	0.390128128	8.18366	4.40E-16	2.427656155	3.957695875	423898.5
Number of times watched in J	0.812618801	0.12931604	6.283975	3.9E-10	0.559037139	1.066200463	993696.8
Number of shows watched in	-0.56638518	0.167810583	-3.37515	0.000749	-0.895452526	-0.23731783	2013.118
Time in Children's	-1.4349E-06	5.87197E-07	-2.4437	0.014608	-2.5864E-06	-2.8347E-07	11245.88
Time in Comedy	-1.6172E-06	5.81245E-07	-2.78225	0.00544	-2.75696E-06	-4.7738E-07	5461.027
Time in Drama	-1.4146E-06	5.64308E-07	-2.50681	0.012248	-2.52119E-06	-3.0804E-07	387.1372
Time in Entertainment	-1.5508E-06	5.77754E-07	-2.68427	0.007318	-2.68379E-06	-4.179E-07	0.011516
Time in Factual	-1.4327E-06	5.68771E-07	-2.51891	0.011836	-2.54801E-06	-3.1735E-07	1463.825
Time in Learning	-9.1314E-06	1.01642E-06	-8.98384	5.14E-19	-1.11245E-05	-7.1382E-06	7465.193
Time in Music	-1.8138E-06	6.08002E-07	-2.98317	0.002881	-3.00603E-06	-6.2151E-07	59.66233
Time in News	-1.4838E-06	5.69104E-07	-2.60721	0.009184	-2.59975E-06	-3.6779E-07	4712.27
Time in Sport	-1.8444E-06	5.68712E-07	-3.2431	0.001198	-2.95961E-06	-7.2918E-07	507.0285
Time in Weather	-2.8668E-06	5.90593E-07	-4.85402	1.29E-06	-4.02487E-06	-1.7086E-06	10338.07
Sum of time_viewed	1.93628E-06	5.60869E-07	3.452285	0.000565	8.36446E-07	3.03611E-06	797.8233
%nearly completed shows	-1.8918445	0.904957576	-2.09053	0.036674	-3.66641667	-0.11727233	65.45776
IV NTW & TCO	9.19929E-08	2.21874E-08	4.146186	3.5E-05	4.84847E-08	1.35501E-07	1938.635
IV NTW & TD	8.97385E-08	1.92812E-08	4.654206	3.43E-06	5.19292E-08	1.27548E-07	74.19046
IV NTW & TE	1.10748E-07	2.37411E-08	4.664806	3.26E-06	6.41926E-08	1.57303E-07	1702.064
IV NTW & TF	9.84367E-08	1.95206E-08	5.042712	4.93E-07	6.01579E-08	1.36715E-07	3613.206
IV NTW & TL	9.43351E-07	7.61665E-08	12.38538	3.34E-34	7.93992E-07	1.09271E-06	22845.13
IV NTW & TM	1.47383E-07	3.52532E-08	4.180697	3.01E-05	7.82533E-08	2.16512E-07	1353.411
IV NTW & TN	1.02853E-07	1.97241E-08	5.214612	2E-07	6.41756E-08	1.41531E-07	4025.352
IV NTW & TS	6.93984E-08	2.11191E-08	3.286044	0.001031	2.79849E-08	1.10812E-07	654.5423
IV NTW & STW	-8.3676E-08	1.84596E-08	-4.53295	6.1E-06	-1.19875E-07	-4.7478E-08	62.0805
IV NTW & NCS	-2.77825388	0.568921244	-4.88337	1.11E-06	-3.893877238	-1.66263053	3068.304
IV NSW & TCH	1.02381E-08	4.33966E-09	2.359184	0.018394	1.72822E-09	1.87479E-08	2554.473
IV NSW & TCO	-1.0493E-07	2.60495E-08	-4.02801	5.8E-05	-1.56009E-07	-5.3846E-08	141.5216
IV NSW & TD	-1.0188E-07	2.29486E-08	-4.43961	9.42E-06	-1.46884E-07	-5.6882E-08	227.9424
IV NSW & TE	-1.2055E-07	2.80395E-08	-4.29917	1.78E-05	-1.7553E-07	-6.5563E-08	1572.175
IV NSW & TF	-1.0851E-07	2.3144E-08	-4.68844	2.91E-06	-1.53894E-07	-6.3125E-08	3268.547
IV NSW & TL	-9.2893E-07	8.36567E-08	-11.1041	5.61E-28	-1.09298E-06	-7.6489E-07	17526.39
IV NSW & TM	-1.5872E-07	4.30229E-08	-3.68927	0.00023	-2.43089E-07	-7.4358E-08	584.4852
IV NSW & TN	-1.1211E-07	2.36973E-08	-4.73097	2.36E-06	-1.5858E-07	-6.5642E-08	3303.501
IV NSW & TS	-7.0742E-08	2.50386E-08	-2.82531	0.004762	-1.19841E-07	-2.1643E-08	524.557
IV NSW & STW	9.09531E-08	2.20044E-08	4.133403	3.7E-05	4.78036E-08	1.34102E-07	1277.894
IV NSW & NCS	4.05324492	0.66799155	6.067809	1.5E-09	2.74335013	5.363139709	1994.678
IV STW & NCS	-7.5693E-07	1.11436E-07	-6.79249	1.38E-11	-9.75444E-07	-5.3841E-07	5855.415

Residual DF	2419
R^2	0.784318
Adjusted R^2	0.781108
Std. Error Estimate	11.26548
RSS	306997.7

Exhibit 6: LRM to predict Number of Shows watched in February with interaction variables

3) Exploratory Data Analysis on Channels

The long tail hypothesis says that most of the channel's views/revenue comes from a small fraction of TV shows. Typically, it is referred to as the 80/20-rule: 80% of the channel's views come from 20% of the channel's shows.

This is because in some genres, a few programs are very popular and therefore generate most views while other programs are just there to serve a few people who have a different taste in shows. Thus, the latter shows get less views – resulting in long tails of the viewing distributions.

A very strong tendency to a long tail can be seen in the views of all shows taken together across the genres. Dividing the shows into genres, however, makes it apparent that the long tail is stronger in some genres and less in others. Interestingly, it can also be seen that this hypothesis applies to the customer basis: Most of the views are generated by a small number of customers.

As shown in Exhibit 7, the long tail is stronger in genres like Factual and Music. This suggests that the mainstream channels have the majority of watchers, while niche channels are relatively less popular. Genres like Religion & Ethics, Comedy, Weather, Drama, News and Learning have a lesser long tail. For religion, channels are obviously segregated by religion, which means that several will be popular. For weather, some channels may cater to different regions, which means multiple ones can be popular. For news, channels can be targeted to a wide array of people depending on what party they follow, left/right wing, etc. For learning, viewers may be looking for specific topics to learn, segregating the channels.

Entertainment, Sport and Children's are in-between. This is likely because big entertainment and sport events can be extremely popular, gathering big audiences, but there are still other sports and events that have respectable sizes.

Genres that are very dependent on the individual's' tastes like Comedy, Drama and especially

Religion in which different people like to watch completely different shows tend to have a lesser long tail due to the different shows that are available and the diverging tastes. News and Learning also has a lesser long tail due to people wanting different information and political views.
The large number of shows for these channels enables BBC to give suitable content to more segregated audiences. Factual and Music – the genres with the strongest long tail – are apparently the genres in which the people who are browsing shows have very similar tastes, making a few programs very popular while the rest only get a handful of views per month.

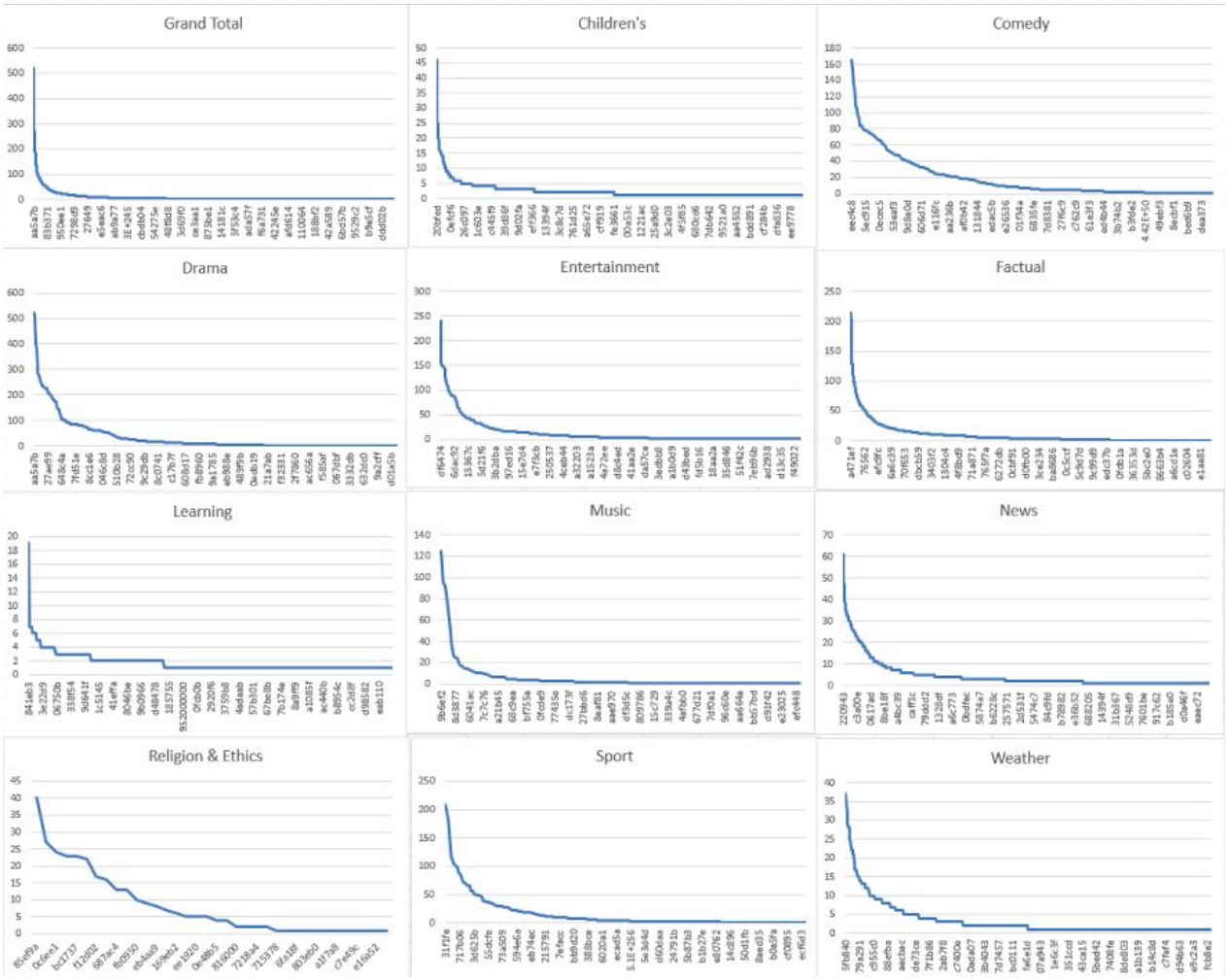

Exhibit 7: Long tails in different genres