Analysing customer satisfaction from broadband performance data over time and determining service related customer clusters

P. Grant
ez18352@bristol.ac.uk

M. Nelhams
oa18502@bristol.ac.uk

J. Lynch
fk18538@bristol.ac.uk

April 10, 2021

1 Introduction

Broadband issues and customer experience have never been more timely to address, where working remotely from home is now commonplace. The increased online traffic alongside 'growing use of data-hungry activities such as video streaming' [1] is impacting performance yet the need for reliable, strong, and stable broadband remains paramount to customer satisfaction. Not providing a broadband service of sufficient quality can have damaging consequences to consumers like disconnecting from an important conference call or missing the streaming of a loved one's life event. Therefore it is essential that customers that experience poor quality service are identified, even when they do not complain.

Internet service providers (ISPs) are strongly incentivised to optimise customer satisfaction, because it is the meta-heuristic that dictates the quality and longevity of their relationship. Reducing customer churn (the propensity that 'a customer ceases his or her relationship with a company') [2] is essential to ensure a high market share and customer equity for subscription based services. To maintain customer loyalty, ISPs must meet or exceed the service performance expected by its customers with a consistently high quality experience [3]. The financial benefits of achieving this are numerous - it is cheaper to retain existing customers due to 'lost revenue and the marketing costs involved with replacing those customers' [2]. Additionally, the home broadband market presents fierce competition, given the value for the fixed broadband market was six billion pounds in 2018 [4], therefore marginal gains in market share corresponds to thousands of pounds markedly boosting profitability. Furthermore, there remains significant room for improvement in broadband services. A recent 2019 cross-sector customer service quality examination by the telecommunications watchdog Ofcom showed that broadband ranked the lowest sector with an overall satisfaction of 83% compared to mobile, gas, electricity, landline with the latter reaching the highest satisfaction of 94% [5]. In an age of viral marketing which employs an 'internet-based word-of-mouth approach' [6], not meeting customer needs can lead to poor publicity resulting in far reaching adverse effects. As Ofcom states: 'It's never been simpler to switch broadband' [7], therefore a customer centred approach is imperative.

Existing literature surrounding the analysis of broadband related customer satisfaction is limited, particularly in the UK. The only empirical broadband study found [8] showed that discrepancies between the advertised speed and delivered transmission speeds are a major factor in customer satisfaction and retention. Whilst there exist many avenues for research, the objective of this report is to offer market insights to broadband companies looking to improve customer satisfaction and reduce churn by responding to the following questions of interest:

- How can customer experience be measured?
- How does customer satisfaction change over time?
- Can customers be clustered using their broadband performance?

There exists plenty of publicly available information regarding the performance, complaint and satisfaction data of various broadband customers around the globe [9]. The Ofcom data for the UK contains nine complete years of individual customer performance data and then nine further years of separate customer complaint data. The Ofcom performance data interests Sky Broadband, because of its vast number of individual features, large variance and it contains various popular broadband providers such as 'Virgin Media', 'BT' and 'TalkTalk'. In this report, the Ofcom public dataset is used to analyse how their broadband customers can be clustered; their customer satisfaction can be indirectly measured; how the customer satisfaction varies over time and to identify customers that leave without complaining to reduce customer churn.

The two models featured within the report use the Ofcom performance dataset, however, performing different variations of data cleaning and dimensionality reduction. The first simple method is a form of supervised machine learning (ML) with the goal of creating a predictive model, so when it is given a broadband speed

it can predict a customer satisfaction level. This model maps download speed directly to complaint data and because of this, it is the simpler of the two models. The second model uses multiple clustering algorithms, which is a form of unsupervised machine learning. The goal of this model is to best cluster the customer data with the expectation of identifying clear patterns and clusters of broadband users. The knowledge of these patterns could be used within industry as a method for targeting certain groups of customers to improve customer satisfaction. The code for the report can be accessed through the following GitHub repository [10].

2 Single-input supervised model

To measure customer satisfaction, the customers who are dissatisfied and complaining need to be identified. All of the features within the dataset must pertain to an important aspect of the customers' expectations, which could potentially cause complaints and worsen their experience. Using the number of monthly complaints as an intermediate measure for the customer satisfaction is not perfect, however, since it is expected that many dissatisfied customers will not complain, because they may not want to, it should be sufficient to produce a functioning representation of the population.

2.1 Data preprocessing

Data preprocessing is the act of preparing a dataset for manipulation [11]. This step is crucial before applying any mathematical techniques, to ensure the result is reliable. To obtain a useful result during data analysis, irrelevant/useless data is removed, preventing the ML techniques incorrectly identifying patterns or fitting to anomalous data.

The important data preprocessing for the simple model is managing the missing entries in feature columns, because the model cannot work with missing/null entries [12]. There are various approaches to processing the data, each focused on conserving different properties of the data [13]. A key concern is that the mean and the variation of the data should not be drastically altered by any data preprocessing, since this affects the variance and bias of any potential models. To process the missing values, all of the empty feature entries were replaced with their respective feature mean. By doing this, the feature means are conserved absolutely, however, the variance is reduced. Reducing the variance is an acceptable downside, since all the Ofcom data already has an exceedingly high natural variance.

2.2 Mapping the customer data to the number of complaints

A simplified model should predict a customer's satisfaction given only one feature from their broadband performance, which is under the assumption that there exists a correlation between customer complaints and customer broadband performance. If the customer complaints data was paired with customer information data then using supervised learning or a statistical model to predict the complaints and hence their customer satisfaction would be trivial. However, these two datasets were measured independently and thus the inputs and outputs are not paired. Instead, the mean number of complaints per 100,000 customers for a month can be paired with the relevant recorded month of customer information statistics.

The most well-documented feature, which contains the most datapoints regarding the general broadband customers, is their mean download speed over 24 hours (Mbits/seconds). Since the customer data is divided by the individual customers, but the complaints data is an average per month, the monthly mean of all the individuals can be used to pair monthly download speed with monthly complaints. Figure 6 in the appendix portrays a negative correlation for the number of complaints over the past ten years, whereas Figure 7 in the appendix portrays a positive correlation for the mean broadband download speed over the same period; indicating that customers are complaining less with increased download speeds. This provides a basis for the assumption there exists a correlation between customer complaints and customer broadband performance.

Comparing the complaints against the corresponding download speeds should produce a negative exponential curve since there is no upper limit for the customer download speed, which can only be explained via a negative exponential curve. There is also a lower bound of 0 (Mb/s) in the model since it is impossible for the download speed to be negative. Furthermore, the number of complaints should tend to zero as the download speed tends to infinity. The following variables are defined: mean number of complaints per $100,000 = \overline{\mathbf{y_c}}$, mean 24 hour download speed $= \overline{\mathbf{x_D}}$. Therefore the predicted graph for complaints against download speed should follow Equation 1.

$$\overline{\mathbf{y_c}} = Ae^{-b\overline{\mathbf{x_D}}} \tag{1}$$

Equation 1 is solved using least-squares regression, resulting in Equation 2. Least-squares regression optimises the equation such that it reduces the sum of the squares of the residual values [14]. The data evidently follows the expected correlation as evidenced in Figure 1.

$$\overline{\mathbf{y_c}} = 10.118e^{-0.01\overline{\mathbf{x_D}}} \tag{2}$$

Tracking the download speed against the complaints, produces a negative exponential correlation as shown below in Figure 1.

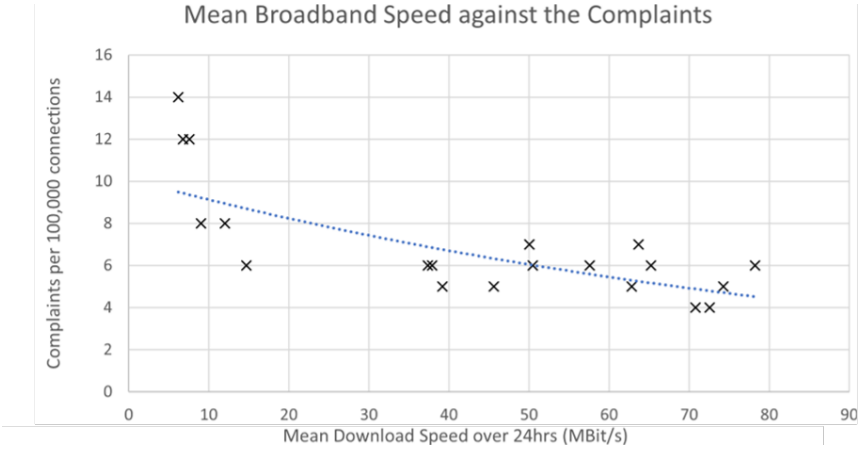

Figure 1: The mean number of complaints decreases exponentially with the mean download speed from their respective month. The data is gathered as an average of each Ofcom recorded broadband and the line of best fit was produced using Least Squares exponential regression, displayed as the dotted line.

2.3 Predicting the percentage customer experience from the complaint data

There currently exists no publicly available Ofcom report containing explicit customer experience levels, meaning that there exists no explicit values of satisfaction that can be manipulated [1]. This leads to the use of customer complaints as a substitute measure. Assuming that users will only complain when they are demotivated and dissatisfied, the complaints can be used as an indirect measure of the customer experience and from that, percentage satisfaction levels can be discerned using one-dimensional clustering algorithms.

Given the number of customer complaints for a month, it is expected that there is some reasonable discrete function that can estimate the customer satisfaction percentage $\overline{\mathbf{y_s}}\%$ given the number of complaints per $100,000$, $\overline{\mathbf{y_C}}$. The processed Ofcom data remains particularly noisy, and it has only 114 datapoints, which is typically considered small. For this reason, the Fisher-Jenks breaks algorithm was chosen as it is a reliable statistical-based approach that works for high-variance clustered one-dimensional data [15]. The results of the Fisher-Jenks algorithm being applied to the experimental data are illustrated in Figure 2.

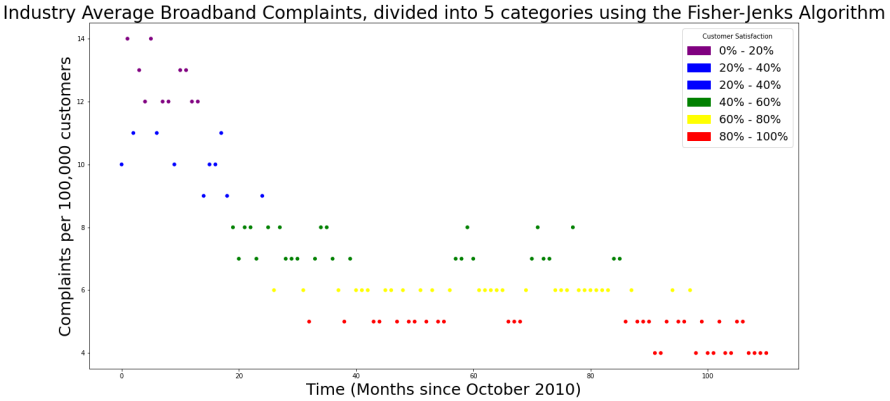

Figure 2: The figure shows the industry average number of complaints per $100,000$, for every broadband, and that the number of complaints decreases exponentially over time. Each colour represents how it can be divided into percentages using the one-dimensional Fisher-Jenks algorithm [15].

The number of divisions m can be any arbitrary integer that factorises 100 and is less than the number of individual data points. For the data, setting the hyper-parameter to $m = 5$ divides the data evenly according to the means and variances, which is a heavily desired feature for fair categorisation. Therefore, for any given number of monthly complaints scaled to per $100,000$ customers, the percentage customer satisfaction interval can be calculated. The average customer percentage satisfaction estimations are plotted in appendix Figure 8.

2.4 Validating the simple model using hypothesis testing

The download speed and individual complaints are random variables, therefore the mean values of the data should be used for predictions. Assuming the means $\overline{\mathbf{x_D}}$ and $\overline{\mathbf{y_c}}$ both follow the law of averages, then hypothesis testing (t-tests) can be used to measure the significance level of the data to 95%, which is a research and industry standard [16]. The null hypothesis for the data is that there is no negative exponential correlation between mean complaints and mean download speeds. Equation 3 shows how to calculate the experimental test p-value t^* from a given correlation coefficient r [17], where n is the number of degrees of freedom. For this given Ofcom dataset, there are 18 months of recorded performance data, so $n = 18$.

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3}$$

The null hypothesis for this representation is that there is no exponential correlation between the download speed and the number of complaints. For the experimental data, the correlation coefficient for the negative exponential curve is calculated as $r = 0.5899$ and therefore $t^* = 3.0995$. Comparing this to the t-distribution value for 95% significance $t_{18} = 2.101$, $t^* > t_{18}$, so the null hypothesis can be rejected with 95% significance, which means that the model itself is justified by the Ofcom data to 95% significance. The 95% significance level was used, because it is an industry standard and the report is non-medical, therefore greater significance is not required [17]. This model will remain valid for the future, so long as the mean customer download speed continues to increase every season, as it has in the past ten years as shown by the appendix Figure 6. However, Ofcom consistently update their consumer statistics, so the model will need to be renewed monthly to maintain a high level of accuracy. The results clearly outline the positive fact that customer satisfaction is quantifiable and increasing every quarter since at least 2010 and Ofcom has concluded similar results [1].

3 Customer clustering and unsupervised modelling

One of the principal aims of this report is to identify if different clusters of broadband users exist. Customers that exist in the same cluster are thought to share commonalities in their service experience based upon certain features in the data. Conversely, customers that are dissimilar with respect to their service tend to be in distinct clusters. Therefore, clustering may reveal the underlying structure in the data, allowing telecommunication marketers to understand key differentiators that partition the broadband market into smaller and more discernible segments. For this investigation, the latest available broadband performance data will be used from the measurement period of November 2019 [1] comprising of 3465 entries and 77 attributes.

3.1 Data preproccessing and feature selection

The data cleaning approach that was taken for the clustering model was to conserve as many features as necessary that may explain divisions in clusters with less emphasis on those that impact customer satisfaction. For this unsupervised learning task, peak performance metrics were chosen over their non-peak counterparts, since customers are assumed to only complain when their service is the worst. The purpose of removing non-peak features is to avoid having many almost perfectly colinear features as depicted in Figure 13, which do not supplement the existing information. Without their removal, they can harm the model's performance to separate customers based on critical features. Therefore, the only other features that were removed were all the weight features (ISP, NAT and rural), market classes and unit IDs. The weight attributes were removed to avoid adding noise and due to their lack of explanatory power. They were not measured characteristics of the system and their description was not given by literature elsewhere. Market Class was removed because its categorical values appeared to be arbitrary and indeterminate. Likewise, unit ID served no purpose since it was not an explanatory variable, but it was instead a variable to distinguish the customers.

The next data cleaning step was to handle the missing datapoints, given clustering algorithms cannot work with missing features [12]. Any entry that was missing a value for the 'Web loading Peak' feature was also missing all streaming service data, which comprised a large proportion of the feature space. As a result, these samples were not deemed to contribute sufficient information to merit their use. These records were a small number of the British Telecom's FTTP (fibre to the premise) or FTTC (fibre to the cabinet) connections, suggesting an issue in the data collection process. The only categorical data that was missing values was 'Distance from exchange' which applied to DSL (digital subscriber line) connections solely. Despite DSL being a small proportion of the dataset, given its importance for many ISPs, a dummy categorical value was added, 'not DSL', to represent other connection types.

Of the remaining 3054 entries, features were mostly well populated with the most incomplete feature still containing 2641 values, which is over 85%. Subsequently, the missing numerical values were imputed using Sklearn's 'Simple Imputer', assigning the mean value of each feature respectively. Given complete features, scaling can be done; this was necessary because 'ML algorithms don't perform very well when the input numerical attributes have very different scales.' [12]. The scale difference was evident with features such as 'Web loading Peak', which ranged from 30 to over $15,000$ and it dwarfed other features like 'Jitter downstream Peak', which ranged from 0.02 to 7.58. For this reason, all of the features were normalised, using Sklearn's MinMax scaler, so that their values lay between 0 and 1.

Lastly, the objective was to encode categorical variables numerically, so that they could be interpreted by a ML algorithm. An ordinal encoding was not appropriate because 'ML algorithms will assume that two nearby values are more similar than two distant values' [12] which would induce bias. Instead, one-hot encoding was used, which creates a dummy attribute for each possible categorical value that can only take on the values of 1 or 0, True or False. Whilst encoding did increase the number of attributes from 73 to 123, thus computational complexity, it allowed the information to be encoded in an unbiased format. The encoded features were then combined with the scaled features.

3.2 Dimensionality reduction

The more dimensions there are to a dataset, the greater the risk that it is sparsely populated since training samples are more likely to be scattered further away from one another. Consequently, the predicting power of clustering diminishes as 'any new instance will likely be far from training instances, making predictions much less reliable than in lower dimensions, since they will be based on much larger extrapolations' [12].

Initially, linear principal component analysis (PCA) was used, to project the data onto a lower dimensional hyperplane. Given a d-dimensional dataset and a target number of q components, PCA achieves this by a sequence of q projections ordered, from largest to smallest, by the amount of variance that they encode where $q \leq d$, for $d, q \in \mathbb{Z}^+$. As reference [18] explains for: observations $x_1, x_2, ..., x_N$, location vector μ, a $p \times d$ matrix $\mathbf{V_q}$ of orthogonal unit vector columns and a q vector of parameters λ, this is equivalent of minimising the least squares reconstruction error presented by Equation 4.

$$\min_{\mu, \{\lambda_i\}, \mathbf{V_q}} \sum_{i=1}^{N} \|x_i - \mu - \mathbf{V_q}\lambda\|^2 \tag{4}$$

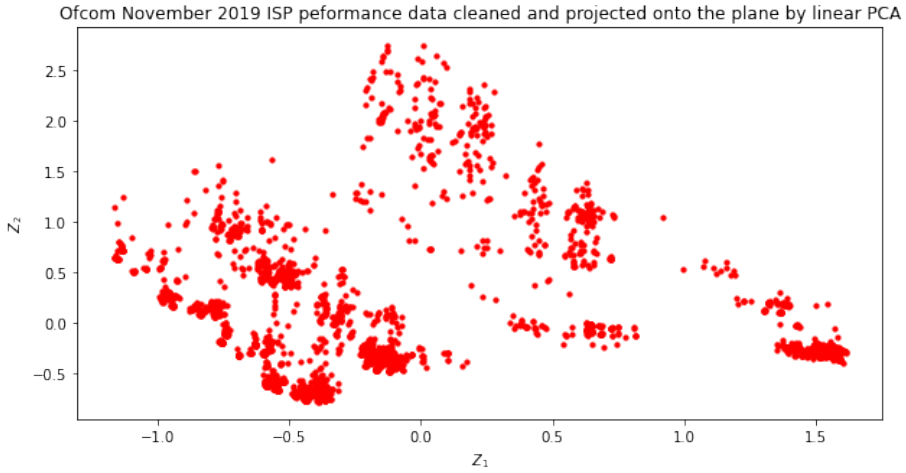

Figure 3: The arrangement of points after linear PCA reduction down to two components, with coordinates of projection Z_1 and Z_2, accounting for 26.1% of the explained variance. There appears to be a number of more densely populated regions that stretch across the graph. However, in many cases these regions appear to be ill-defined with a lack of salient clusters and noise adjoining densely populated regions.

The linear PCA projection, in Figure 4, lacked easily identifiable clusters, so other non-linear dimensionality reduction methods were used for comparison. A popular method named UMAP was used, which assumes that the data lies on a uniformly distributed Riemann manifold that is locally connected and then it projects this to its nearest fuzzy topological representation [19]. The UMAP projection as shown by appendix Figure 10 was very fragmented making it unfavourable, yet nonetheless, it showed some structure. Kernel PCA was also applied, with various kernels, to reduce the projection to two dimensions after having used linear PCA to reduce 123 dimensions to 35 dimensions whilst preserving 95% of the variance. Kernel PCA works by 'expanding the features by non-linear transformations, and then applying PCA in this transformed feature space' [18]. The effect of this is shown in the appendix by Figure 11 where there are, from top left to bottom right, the Gaussian; the polynomial; the sigmoid kernels; and lastly the cosine kernel in Figure 4. The first three of these kernels did not show improvement on linear PCA, the cosine kernel appears to separate denser areas of clustering more clearly with less noisy points interfering between clusters. Due to this, clustering algorithms are more likely to discern the existing customer segments therefore this projection is preferred to test their performance.

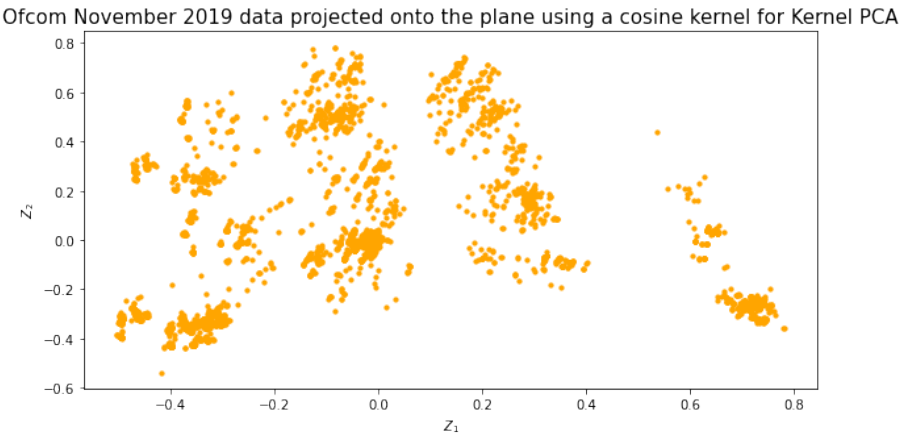

Figure 4: The cleaned clustering data projected onto the plane using Kernel PCA with a cosine kernel and coordinates of projection Z_1 and Z_2.

3.3 Clustering algorithms comparison

Given the relative strength and weaknesses of clustering algorithms in different applications, the suitability was tested for some of them based on their fitting success to the data. Figure 5 shows the following algorithms applied to the data, from top left to bottom right: Gaussian mixture model [20], DBSCAN [21], K-means [22], spectral clustering [23] and HDBSCAN [24]. All of the clustering algorithms used Euclidean distance as their metric. To the right of each graphic is the respective colour bar denoting the colour and label of each cluster, where clustering labels are indexed from 0 for all except DBSCAN and HDBSCAN which have an extra label, -1, to represent any outliers. All of the algorithms have had their hyperparameters tuned, through trial and error, to best capture the existing patterns in the data. Observing the graphs briefly, it can be seen that most algorithms performed their best with 6 or 7 clusters with this being explicitly programmed with all but the density based approaches (DBSCAN and HDBSCAN).

Examining spectral clustering first, it forms semi-irregular clusters that stretch over the remote regions and clearly separates groupings in the data, and therefore it doesn't perform well enough. K-means improves upon this by clearly separating areas with high density, but it still suffers from some non-intuitive cluster boundaries despite an optimal number of clusters, K. Furthermore, whilst the non-deterministic nature of K-means can be weakened by repeated initialisation, its bias towards spherical regions of similar size and density means that it doesn't generalise to the data well. Gaussian mixture model (GMM), the only parametric model, assumes the data is derived from a specified number of Gaussian distributions each with unique mean and standard deviation. GMM improves upon K-means by modelling covariance and thus ellipsoidal formations in the data. However, the expectation algorithm (maximum likelihood), for which GMM tries to maximise, means it struggles to distinguish the distinct clusters nearby- as shown in turquoise.

A more flexible approach is to use DBSCAN, which clusters data based on the continuous regions of high density, using a distance parameter, epsilon, which defines the neighbourhood region of every instance. The graph for DBSCAN shows that it finds some more pertinent and meaningful clusters in this data, like the separation of the pink and green cluster, whilst other clusters are too optimistically clustered, like those in brown and dark blue. In spite of this, DBSCAN was particularly sensitive to changes in epsilon, so further optimisation was not possible; changes would result in large cluster fusing and excessive outliers. A solution to this was to use HDBSCAN, which combines a cluster hierarchy tree to DBSCAN that allows clusters to have independent epsilon values and it deals effectively with noise thus preventing bridges in low density areas from fusing clusters. After defining a minimum cluster size, 3% of data is used, the instances are assorted based on density via a hierarchy tree and the clusters are separated based on their stability. As a

result, the DBSCAN algorithm seems to generalise best to the data and make the most meaningful clustering so these labels are used to understand what separates them.

3.4 Analysing clusters

This cluster analysis uses the means of numerical features indexed by cluster labels and their respective technologies are shown in Figure 14, included in the appendix for readability. Note that the cluster label '-1' is ignored since it relates to six outliers of no particular importance, which are likely to be anomalous data.

The clarity of each cluster's significance varies. More notable clusters include that labelled zero, in dark blue, in Figure 5, which comprised of all 528 of Virgin's cable customers. Not only is cable technology exclusively supplied by Virgin, but it is also the only technology they offer. This combined with the fact that cable is the best technology with respect to performance features like peak download speed, of $202Mb/s$, in appendix Figure 14, means that it stands out clearly. The green cluster with label one consists of 112 FTTP customers from various providers. This cluster separates nicely, because it's the second-fastest technology with an average speed of 196 Mbps and it has the lowest average jitter, web loading and start-up delays across all medias at peak times, shown in the appendix Figure 14. Conversely, cluster label three, seen in red, is made up primarily of ADSL1 and ADSL2, so it's characterised by the lowest download speeds of 8 and 15 Mbps and the highest latency of 32 and 26 milliseconds respectively. There appears to be impurity in cluster two, because it contains 34 FTTP connections that are most likely misclassified, since they don't share any similar measurements along any of the features in appendix Figure 14. In retrospect, the cause of this is likely due to the dimensionality reduction methodology. The points encircled in blue in appendix Figure 12 show a moon shaped cluster that could all be FTTP connections. Given the kernel chosen, this structure was not conserved, thus some points have been misclassified as label two instead of label one.

Cluster two should be a key concern for broadband suppliers since it contains customers who could have a higher tendency to complain due to the low mean and high variance of ADSL1 and ADSL2 connections, which leads to performance issues particularly for those in the lower quartile. The price-to-performance ratio for their service is likely to cause a complaint.

Clusters three, five and six, respectively shown in peach, yellow and brown in Figure 5 are all associated by the fact that they comprise of only FTTC customers on a 76 Mbps download package. Cluster six is unique from the other two, because it contains only BT customers in England located in urban regions. There's some suggestion that these customers receive a slightly more reliable connection than clusters three and five since disconnection rate is almost half, 0.18 compared to 0.33, and latency is marginally lower, 10 ms vs 13.5 ms. Nevertheless, given these differences are unlikely to be noticeable or impact customer services and the differences in important features such as the mean peak downloads speeds and peak media delay times are negligible, it's likely that these clusters experience similar customer satisfaction rates. The obvious distinction between clusters three and five is that the former pertains to only rural customers and the latter pertains to the remaining urban customers on this package speed. Lastly, cluster four contains customers on slower FTTC and FTTP packages than was previously seen. 24 customers on a 76 Mbps FTTP package and 819 FTTC customers, of which 634 have a 35 Mbps package a further 171 have either a 50, 52 or 55 Mbps package and the remaining 11 noisy instances lie above or below theses package speeds. The FTTP customers in cluster four receive a significantly better average service, over all features, than their FTTC counterparts with the same package speeds at peak times. However, FTTC connections in cluster four, receive the poorest broadband performance of its kind due to package restrictions and peak time performance losses.

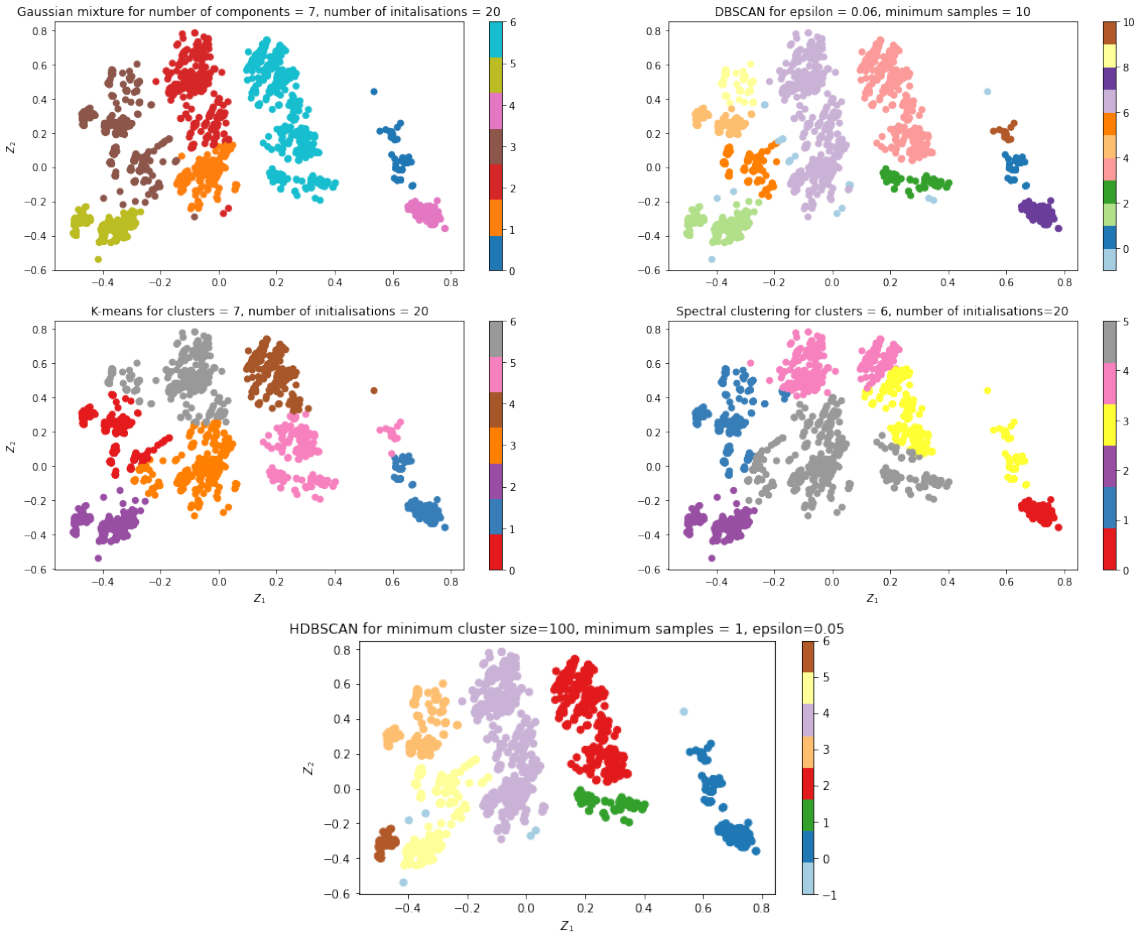

Figure 5: Performance of different tuned clustering algorithms (from top-left to bottom-right): GMM, DBSCAN, K-means, Spectral clustering and HDBSCAN on the dimensionaly reduced data.

4 Discussion

The simple complaints-based model provides an effective insight into the general customer satisfaction, however, it is constrained by its initial assumptions. The first assumption in using only the download speed to predict customer satisfaction is that the download performance perfectly explains all of the nuances and attributes of the customer experience. It is likely that this assumption is only partially true, but with more independent variables and the implementation of a neural network, the model accuracy can be improved. Secondly, the mathematical model assumes that all dissatisfied customers will complain. Complaints data contains no insight as to the nature of the complaints, therefore there is some expected loss in accuracy. Furthermore, there may be false-positive complaints, these would be considered complaints regarding non-performance related properties, such as billing or customer service. Lastly, the performance data is assumed to be entirely representative of the broadband population, however, there are only 3,465 recordings listed in 2019 for example, which is far fewer than the expected number of households with home broadband [25]. It is possible that the entire population may have different performance data trends and absolutely different clustering patterns and this assumption disadvantage extends to the data clustering done in this report, which could render the models invalid. With a larger data set, the models findings will be more reliable.

The single-input model also lacks the precise ability to predict any individual customer's satisfaction. Given

there's no labelled data available for customers satisfaction it's difficult to infer the reason or likelihood of churning. One suggestion is to introduce labels like the mean opinion score or MOS, a popular metric in the telecommunication industry [26], which takes the arithmetic mean of customer rankings of various service metrics. Adding a response variable such as MOS to the data, whilst highly subjective by nature, would open avenues for supervised ML such as neural networks [12]. Alternatively, future data collection could include labels for customer churn or complaints to perform classification. Given enough data, a neural network could be able to learn well the general customer attitudes over very varied internet service performance results which should relate to churn probability and help identify customers who churn without complaining. Without labels, clustering is the most informative ML procedure possible.

Given the clustering labels identified in this report, it is possible to use a classifier like decision trees to identify which cluster unseen data pertains to, but this is beyond the scope of this project. The information gained from these clusters alone is limited. However, if the labels could be interpreted in conjunction with complaint or churn data, it may help to identify customers who are unhappy with their service but choose not to complain. Broadband providers could target these churn susceptible groups to improve their service and maintain their market share. Although further inspection of the dimensionality reduction technique used is required to identify if it is a cause of clustering misclassification.

Previous research has shown that other account information such as creation date, the billing frequency, price of the package, the account balance and payment types are 'useful for predicting the customer behaviour for the next observation period.' [27] thus reducing churn. Data from other interactions with the customer is also pivotal given 'customer service and system reliability is most influential on perceived value and customer satisfaction' in mobile services [28]. A recommendation of this report is for further research into customer accounts and customer interactions data. With this, a broader and more robust model with a larger information base of informative features could be implemented that makes more accurate predictions possible. It follow that more reliable and extensive conclusions can be drawn from that model.

5 Conclusion

Measuring customer satisfaction for broadband consumers across the UK was a primary objective for the given report. A model for predicting the mean monthly customer satisfaction as a percentage was produced, with a tolerance for the precision of the percentage accuracy also included. The model constructs its predictions from the Ofcom performance and complaints data, evidenced by statistical t-tests, which were evaluated to a 95% significance level. Therefore there exists an exponential relationship between the download speed and customer complaints. Additionally, this model will need to be renewed monthly, for as long as Ofcom continue to publish new data.

The secondary focus for the project was to cluster the broadband consumers. Using the HDBSCAN algorithm [24], customers have been categorised into seven clusters of varying size and characteristics. Almost all clusters were pure or had a meaningful reason for separation. Cluster two is the most problematic since it contains ADSL1 and ADSL2 connections, who are likely to be dissatisfied with their broadband. This algorithm showed that combinations of more influential features like Technology, ISP, and Package directly impact the overall broadband performance e.g download speeds and latency at peak times and therefore the discrimination in much of the clustering. However, the relationship is more complex with less indicative features like their country and their geography (urban or rural) also contributing to their segmentation in certain cases.

References

[1] Ofcom. Uk home broadband performance, measurement period november 2019 [online]. Available from: https://www.ofcom.org.uk/research-and-data/telecoms-research/broadband-research/home-broadband-performance-2019. [Last Accessed: 20/11/20].

[2] Customer churn prediction and prevention. [online]. Available from: https://www.optimove.com/resources/learning-center/customer-churn-prediction-and-prevention. [Last Accessed: 18/10/20].

[3] Frederick Herzberg. Understanding customer experience [online]. Available from: https://hbr.org/2007/02/understanding-customer-experience. [Last Accessed: 29/11/2020].

[4] Emma Leech. Fixed broadband market size. Available From: https://www.ofcom.org.uk/__data/assets/pdf_file/0027/191673/fixed-broadband-market-size.pdf. [Last Accessed: 02/12/20].

[5] Market research open data - quality of customer service. [online]. Available from: https://www.ofcom.org.uk/__data/assets/file/0021/146244/csq-open-data-2018.csv. [Last Accessed: 24/11/20].

[6] M. Woerdl, Savvas Papagiannidis, Michael A. Bourlakis, and Feng Li. Internet-induced marketing techniques: Critical factors in viral marketing campaigns. *Journal of Business Science and Applied Management, Vol 3, Issue 1*, 2008.

[7] It's never been simpler to switch broadband – and get a guaranteed speed. [online]. Available from: https://www.ofcom.org.uk/phones-telecoms-and-internet/advice-for-consumers/advice/broadband-speeds-code-practice. [Last Accessed: 17/10/20].

[8] Torsten J. Gerpott. Relative fixed internet connection speed experiences as antecedents of customer satisfaction and loyalty: An empirical analysis of consumers in germany. *Management & Marketing*, 13(4):1150–1173, dec 2018.

[9] Worldwide broadband speed league 2020 [online]. Available from: https://www.cable.co.uk/broadband/speed/worldwide-speed-league/. [Last Accessed: 21/11/20].

[10] M. Nelhams and P. Grant. Skybroadband unsupervised algorithm [online]. Available From: https://github.com/MikeMNelhams/SkyBroadbandProject/blob/main/PythonScripts/SkyBroadband_UnsupervisedAlgorithm.ipynb. [Last Accessed: 27/11/20].

[11] Data preprocessing : Concepts [online]. Available From: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825. [Last Accessed: 02/12/20].

[12] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly UK Ltd., 2019.

[13] Ralph Neuneier Volker Tresp, Subutai Ahmad. Training neural networks with deficient data. Technical report, Siemens AG, 1994. [Last Accessed: 21/11/20].

[14] Stephanie Glen. Least squares regression line: Ordinary and partial [online]. Available From: https://www.statisticshowto.com/least-squares-regression-line/. [Last Accessed: 02/12/20].

[15] Rehan Ahmad. Jenks natural breaks — the best range finder algorithm. [online]. Available from: https://medium.com/analytics-vidhya/jenks-natural-breaks-best-range-finder-algorithm-8d1907192051. [Last Accessed: 04/11/2020].

[16] Tools for fundamental analysis [online]. Available from: https://www.investopedia.com/terms/t/t-test.asp#:~:text=Key%20Takeaways-,A%20t%2Dtest%20is%20a%20type%20of%20inferential%20statistic%20used,of%20hypothesis%20testing%20in%20statistics. [Last Accessed: 30/10/2020].

[17] Dr. Laura Simon and Dr. Derek Young. Hypothesis test for the population correlation coefficient [online]. Available from: https://online.stat.psu.edu/stat501/lesson/1/1.9. [Last Accessed: 30/10/2020].

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[19] Umap: Uniform manifold approximation and projection for dimension reduction. [online]. Available from: https://umap-learn.readthedocs.io/en/latest/. [Last Accessed: 23/11/20].

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[22] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967)., 1967.

[23] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[24] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg, 2013.

[25] Internet access – households and individuals, great britain: 2019 [online]. Available From: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2019. [Last Accessed: 27/11/20].

[26] Christos Tsiaras and Burkhard Stiller. A deterministic qoe formalization of user satisfaction demands (dqx). 2014.

[27] B. Q. Huang, M-T. Kechadi, and B. Buckley. Customer churn prediction for broadband internet services. In *Data Warehousing and Knowledge Discovery*, pages 229–243. Springer Berlin Heidelberg, 2009.

[28] Ying-Feng Kuo, Chi-Ming Wu, and Wei-Jaw Deng. The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior*, 25(4):887–896, jul 2009.

A EDI Statement

EDI issues are systemic and are becoming increasingly frequent in modern systems that integrate automated machine learning. Any AI system that interacts with people, or uses data in any way connected with people needs to be checked carefully. Respectively, the topic of measuring broadband performance and customer satisfaction is no different, since many practical machine learning approaches have been applied to the task. Fortunately, all public data is anonymous, contains no personal information of the customers and each customer is subsequently treated equally. The datasets and impacts of this report pose no ethical, diversity or inclusion implications, since the metrics and conclusions apply to any and all broadband customers equally.

For the cooperation, communication and teamwork aspect of the given project, there were no EDI problems visible or represented by the allocated group. All members were briefly trained for EDI by the university department and each member was provided a valid opportunity together to raise any current or potential problems that could occur regarding EDI and nothing was noted.

B Covid Mitigation Statement

The global pandemic is an evident obstacle for all given projects, especially current group projects, due to the issues that can arise regarding timetabling, communication and the effect of coronavirus on each individual's personal lives. For the majority of the group, no extenuating circumstances applied to this project due to coronavirus directly or indirectly. Communication was fluidly integrated on an online means, with timetabling exceptions made to accommodate for each group members' time zones.

The team suffered from the loss of one member who transitioned onto the technical project relatively early in the project. Given there was no warning of the event, the team morale was disrupted and the work had be assigned to him had to be redistributed to the remaining members which greatly impacted our project schedule. Furthermore it caused consistent strain through the remaining project development as there were fewer people to undertake the same project goals so more time and effort from the remaining members was needed.

C Additional Graphs and Figures

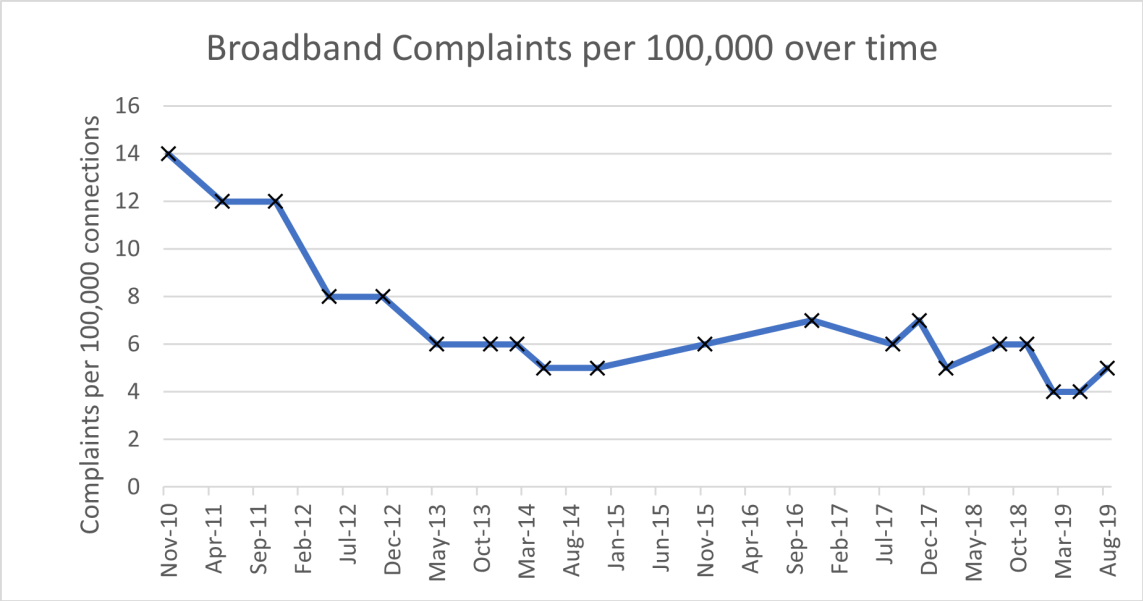

Figure 6: Displayed above is a line graph of how the industry average number of complaints per 100,000 for every broadband decreases gradually over the past ten years.

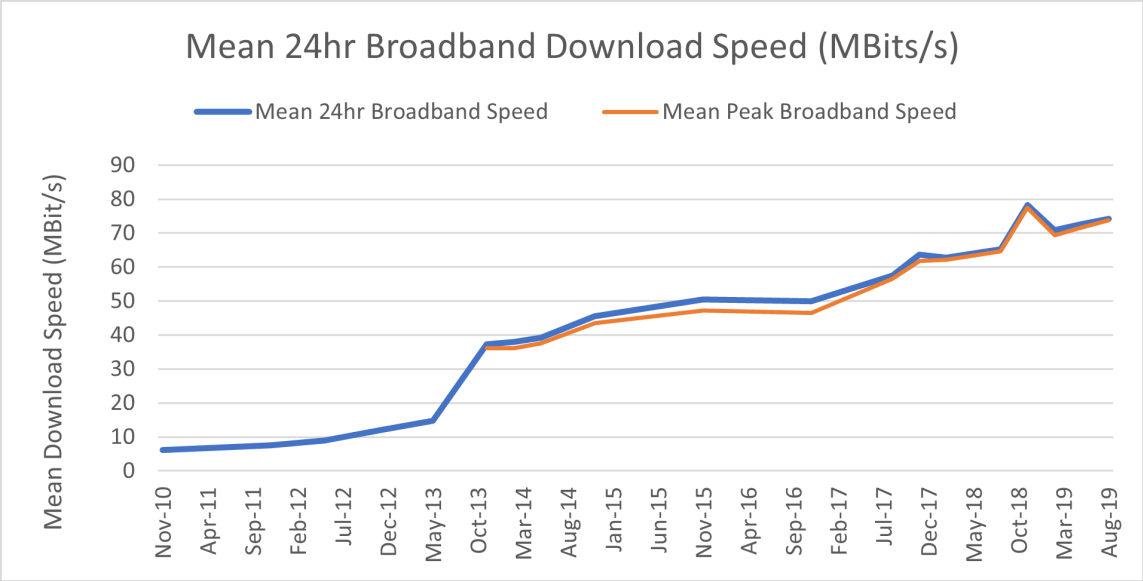

Figure 7: Displayed above is a line graph of how the industry average download speed (MBits/s) for every broadband gradually increased over the past ten years.

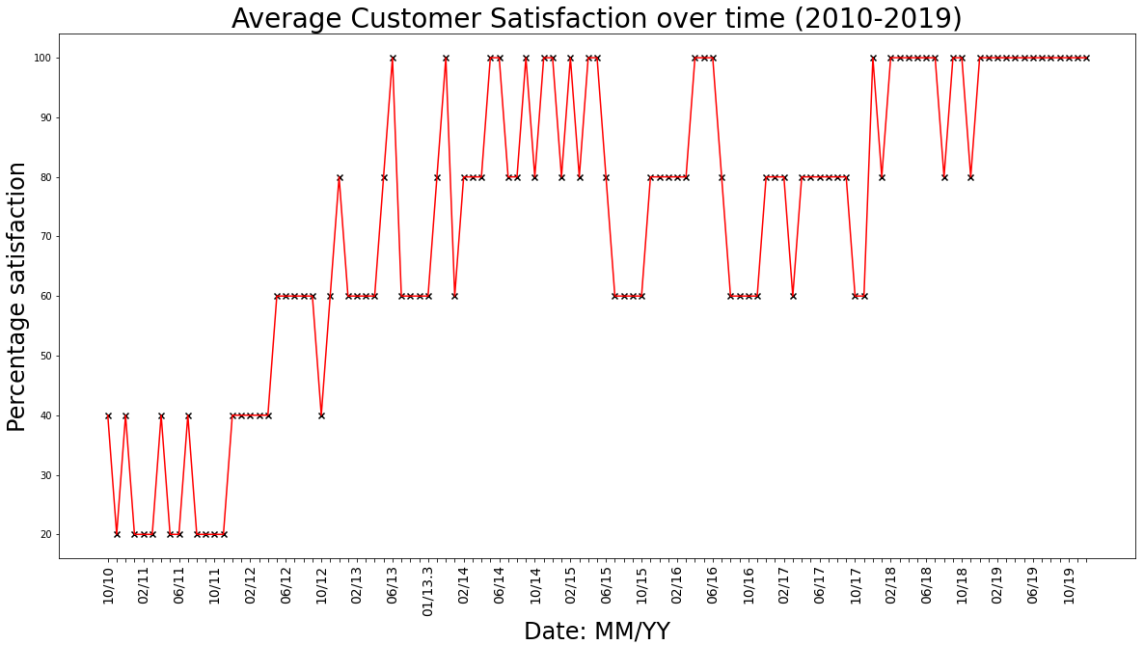

Figure 8: The industry average number of complaints per 100,000, for every broadband, decreases greatly over time and it can be divided using the Fisher-Jenks breaks algorithm [15]. Each colour represents a different 'bucket' category for dividing the data.

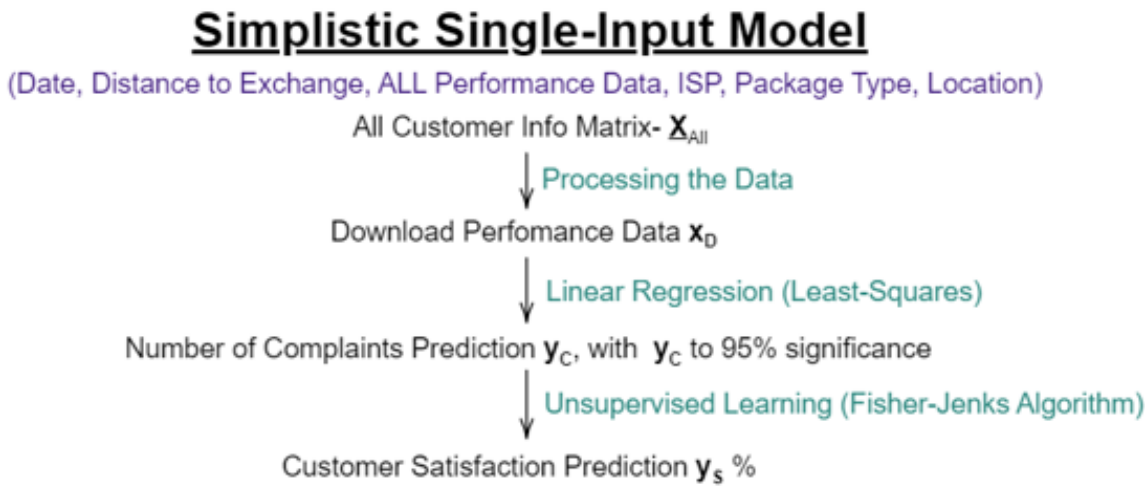

Figure 9: The finished simple single-input model is shown in the figure above. It begins with a large dataframe of the customer performances, then processes this data. Then linear regression is applied to the cleaned data. Lastly, the Fisher-Jenks breaks algorithm is utilised to predict the average customer satisfaction as a percentage for each month.

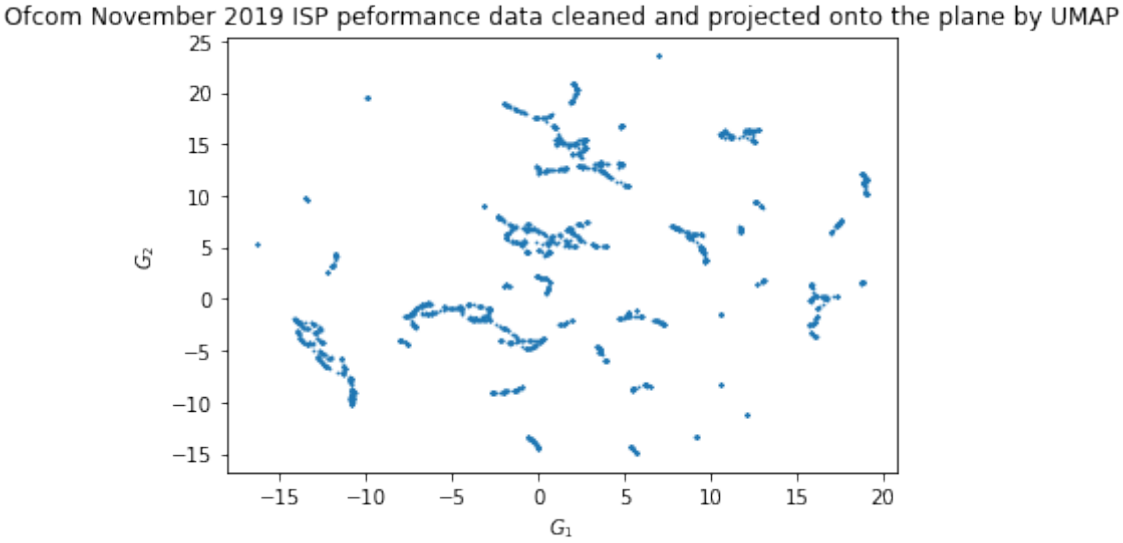

Figure 10: The cleaned Ofcom 2019 ISP performance data projected onto the plane by UMAP with coordinates of projection G_1 and G_2 is shown in the figure above.

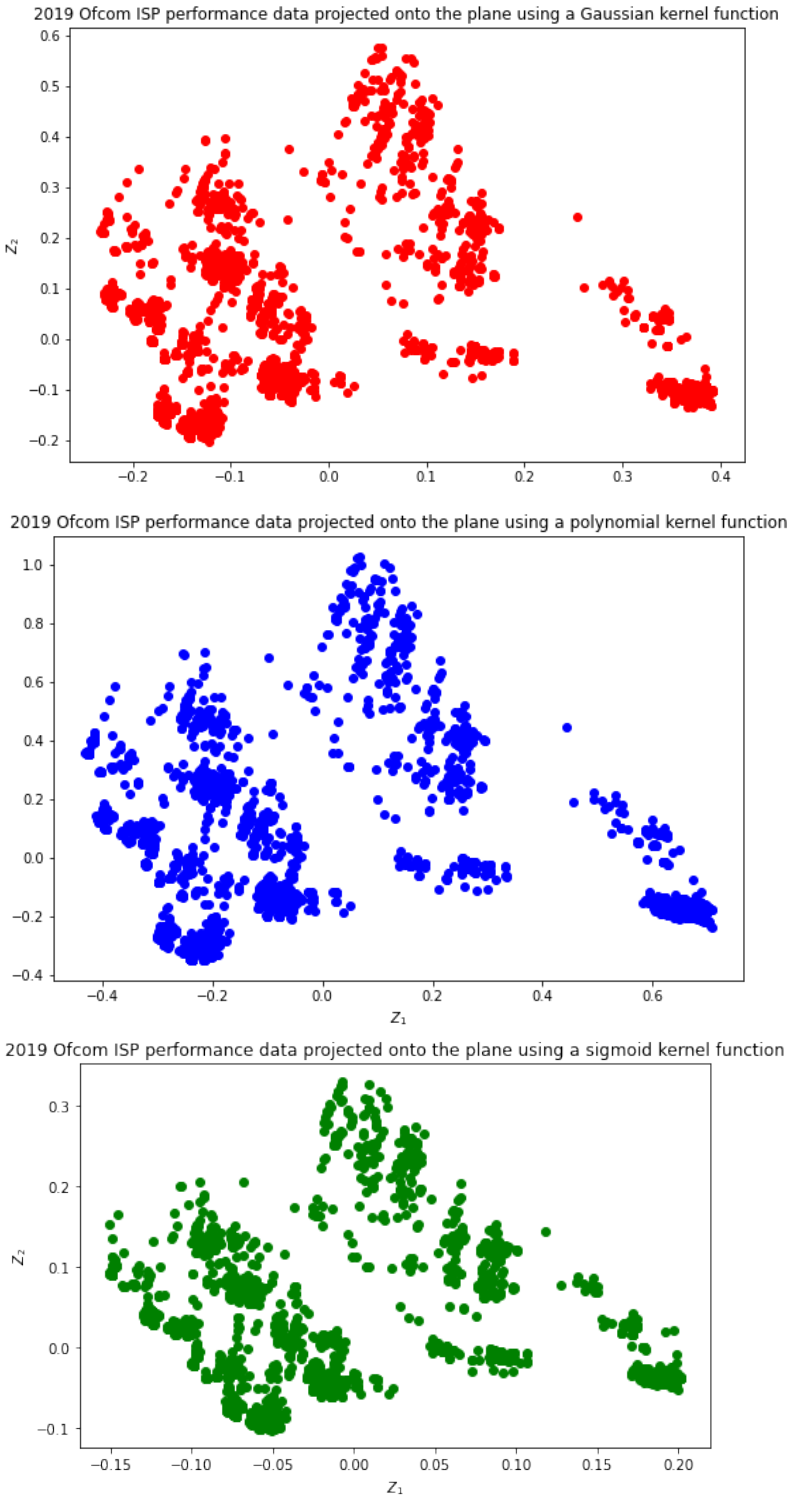

Figure 11: Displayed above is the cleaned clustering data projected onto two the plane using Kernel PCA with various kernels (from top left to bottom right): Gaussian, polynomial and sigmoid with with coordinates of projection Z_1 and Z_2.

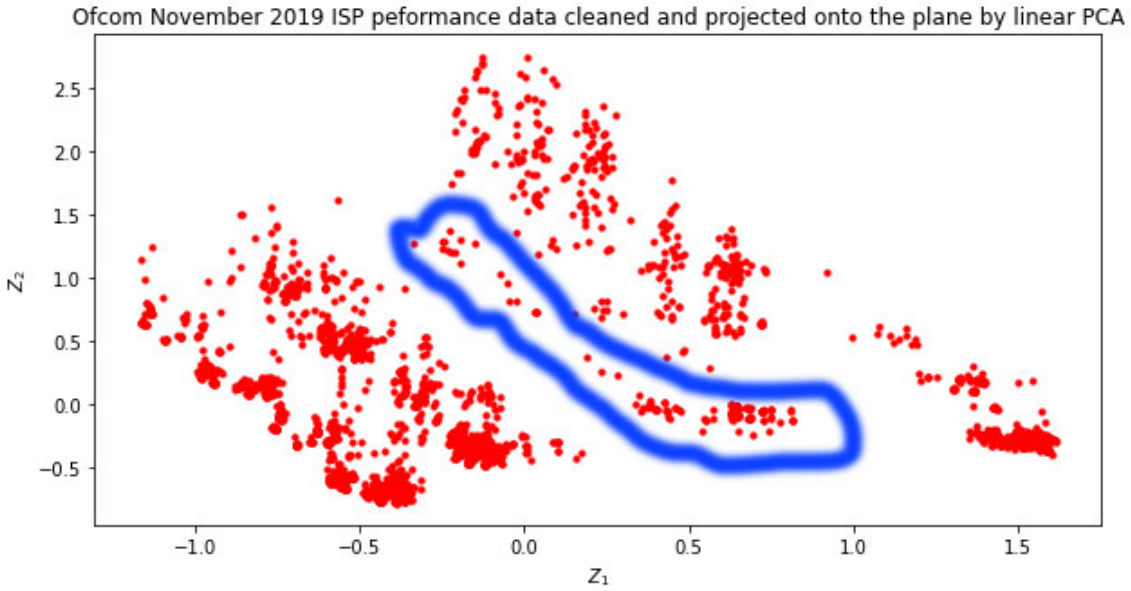

Figure 12: Highlighted are the points from the linear PCA of Ofcom 2019 ISP performance data that could explain the incorrect classification of FTTP instances in cluster label 2 after a cosine kernel PCA.

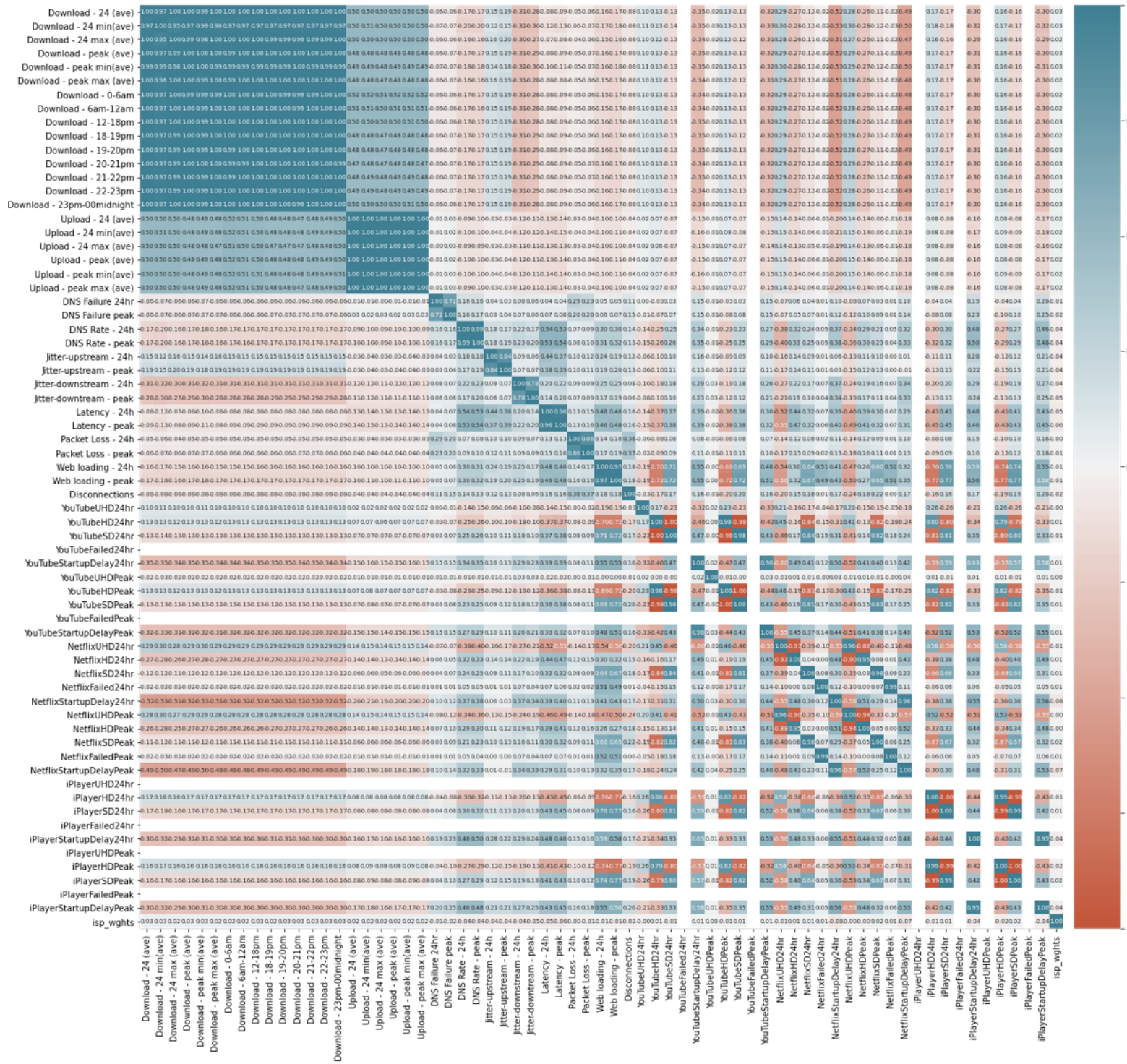

Figure 13: A correlation matrix showing the linear correlations between all features in the Ofcom 2019 ISP performance data is shown in the figure above. Blue signifies a strong positive correlation, whereas red signifies a strong negative correlation between the features.

Cluster_labels	Technology	PACKAGE (download by upload)	Download - peak (ave)	Upload - peak (ave)	DNS Rate - peak	Jitter-upstream - peak
-1	Cable	100.000000	79.740000	9.820000	12.330000	2.530000
	FTTC	76.000000	28.086667	8.020000	19.626667	0.343333
	FTTP	76.000000	65.820000	18.430000	15.610000	0.460000
0	Cable	212.708333	202.859754	20.397462	15.602140	2.387784
1	FTTP	211.142857	195.234286	58.030893	7.018929	0.337321
2	ADSL1	8.155556	2.729333	0.498889	37.769556	2.270000
	ADSL2	14.573379	9.917167	0.781297	27.902389	1.646370
	FTTP	200.058824	167.887941	80.330588	10.886471	0.340000
3	FTTC	76.000000	57.521116	15.203678	15.995826	0.352149
4	FTTC	42.619048	32.252418	7.290183	17.149683	0.613077
	FTTP	76.000000	71.124583	19.510000	9.435417	0.280417
5	FTTC	75.784091	58.636690	16.284034	16.235284	0.490696
6	FTTC	76.000000	57.303401	15.412267	13.994980	0.403320

Cluster_labels	Technology	Latency - peak	Web loading - peak	Disconnections	YouTubeStartupDelayPeak	NetflixStartupDelayPeak	iPlayerStartupDelayPeak
-1	Cable	20.910000	292.360000	0.140000	NaN	653.160000	NaN
	FTTC	16.806667	320.593333	0.300000	1155.096667	1136.416667	1405.910000
	FTTP	13.175000	175.135000	0.615000	538.650000	1012.835000	543.510000
0	Cable	18.510530	286.442121	0.221499	645.441540	463.048102	768.407672
1	FTTP	8.623393	106.801675	0.646518	545.543611	446.377917	368.996197
2	ADSL1	32.722889	2570.080222	3.751778	2656.124333	2239.506774	2669.834286
	ADSL2	26.670751	910.989151	1.199078	1558.088077	1879.805402	1959.993702
	FTTP	9.690588	146.842059	0.278529	575.776667	778.743704	492.305185
3	FTTC	13.312851	221.337686	0.335124	876.307909	1086.373957	886.993468
4	FTTC	15.110562	318.772262	0.334298	901.293926	1433.896443	1141.796953
	FTTP	7.807917	170.064496	0.330417	684.154783	848.440435	605.551818
5	FTTC	13.898224	194.742060	0.331307	836.383047	998.278521	877.171008
6	FTTC	10.839514	207.794939	0.180810	946.970826	1037.762428	808.324737

Figure 14: Above are the means of numerical features indexed by cluster label and technology in each label from the Ofcom 2019 ISP performance data used for clustering.