Divide and Conquer

Identifying and understanding a customer/ customer base through means of clustering, classification, and variable importance analysis.

Introduction

The purpose of any business is to provide a benefit to society. The better a business understands its customers, the better they can provide for customers, drive sales, grow, and adapt to change. The impact of demographic/socioeconomic factors are well understood, but another factor is personality. Personality provides important context that can help a business better cater towards the requirements of their customer base.

Personality Psychology in context

As businesses adapt to an Internet-of-things framework, it will become simpler to attain data on individual personalities. In the field, psychologists use randomly sampled conversation recordings, visual stimulus, etc. to determine personalities (Bollich et al. 2016; Chen & Lin 2017). The binary model introduced in the Myers-Brigs personality test has fallen out of favour with psychologists now recognising traits as distributions (Haslam 2020; Stromberg & Caswell 2015). One model recognising this is 'The Big Five' which focuses on distributions in: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Reece 2009). A newer model, HEXACO, introduces a sixth factor which measures humility (Wiernik et al. 2020). It has been argued that fewer factor tests result in an oversimplification of personality, but recognised that they can still have utility depending on the goal (Wiernik et al. 2020).

In this case study, it is hypothesized that these multidimensional tests produce a datascape that can be broken into clusters that the business can use to inform its strategy; identifying the most typical personality types and better catering to them. Similar work has been applied to customer loyalty, routines, satisfaction, perceptions, and marketing strategy (Castillo 2017; Chittaranjan, Blom & Gatica-Perez 2011; Lin 2010). Previous outcomes have been achieved using k-means clustering, naive-bayes, support-vector machine, boosted tree, k-nearest neighbour

(**KNN**), neural network, and regression models (Gerlach et al. 2018; Shah & Modi 2021; Talasbek et al. 2020; Tinwala & Rauniyar 2021). With the goal of understanding groupings in the customer base and assigning new customers to a group, it was essential the final model had both explanatory power and predictive utility.

To achieve these goals, two methods were trialled to cluster the data: a KNN - DBSCAN algorithms presented in the literature, and a k-modes algorithm (Béjar Alonso 2013; Sander et al. 1998; Schubert et al. 2017). Logistics regression (**LR**) and random forest (**RF**) algorithms were used to create binary classifiers that predicted if a customer was in a given cluster and enabled an assessment of variable importance.

Problem Definition and Data Collection

The aim of this report is to provide methodology that can:

- Cluster a customer base into groups of common personality types
- Predict if a customer is a representative of a group
- Explain the prediction in a way that reveals the traits of that cluster

Data Source: The data, uploaded by 'BOJAN TUNGUZ' in 2020 was downloaded from Kaggle. There were 1,015,342 instances and 110 variables. Of these variables, only 50 represent responses to ‘the big five’ test and two represent lambert coordinates; these were the variables most relevant to this investigation.

Note: The data is ordinal on scales of 1 (disagree) to 5 (agree), meaning there is a middle value and potentially 1,015,342 interpretations of the scale.

Methodology

The five-factor test has been selected here because it is the most widely used, and a measure of altruism (as in the HEXACO test) is unlikely to provide meaningful insight. The five factors are also understood to be orthogonal to each other, although there is debate on this point (Biesanz & West 2004; Saucier 2002). In lieu of a trained psychologist, the data was strategically summed together to create a rudimentary score for each of the five distributions; a substantial and necessary variable reduction. Variable reduction techniques such as PCA would impede the

model's explanatory power without providing any great benefit due to the orthogonal nature of this data specifically. Accepting some small systematic error in exchange for explanatory power was deemed an acceptable trade off given the known issues of self-reporting (Vazire & Carlson 2010). This manipulation of the data results in a loss of information about more specific sub-traits, but is a good approximation of personalities. Missing values and outliers were removed, and duplicates were managed prior by the uploader using unique IP addresses. Outlier removal is particularly important as the LR in attempt 1 (see below) is sensitive to outliers; their removal should not impact the goals of the study (Lavalley 2008). Memory and relevance issues made it necessary that only Victorian values were extracted and clustered into personality groupings; these groupings represented the target variable. A density-based clustering algorithm was deemed appropriate, as clustering structure are unknown and rather abstract by nature (Béjar Alonso 2013; Bhattacharjee & Mitra 2021). The two clustering methods trialled were:

Attempt 1 (numerical) - A KNN algorithm was used on a range of k values to determine k-distances (the radius around a point for which k neighbours were within the circle); the optimal k, as determined by an elbow plot and silhouette score, was passed as 'min_samples' and the corresponding distance as the 'eps' hyperparameter in a DBSCAN; a model that identifies core points based on hyperparameters, and clusters points that are within the boundary of other core points ignoring all remaining points. All distance metrics were trialled, even cosine similarity.

Attempt 2 (categorical) - A k-modes algorithm on arbitrarily categorised data. K-modes takes k random samples and then groups points that have the most similarity between them. The process is repeated until cluster assignment is unchanged.

The data was representative of the entire population, which is perhaps applicable to an organization like Woolworths or Coles that deal with many types of people daily. However, the nature of the data is expected to change if you sampled from a local chemist, tech store, or Officeworks. The problem with the dataset is that every person is so well represented that it was difficult to find clusters in attempt 1 as the boundaries weren't all that clear, making the methodology off attempt two necessary.

Each instance was labelled with the cluster number assigned by the k-modes algorithm which would act as the target variable.

After splitting the data, reserving 20% as a testing set, supervised binary classification models were used to predict if a customer had a personality similar to cluster X. There are some difficulties to consider, such as the number of clusters, and how to interpret what each cluster represents (Talasbek et al. 2020). To address the first issue, the choice of density-based clustering meant a threshold needed to be set as to what constitutes a dense patch, the maximum number of possible clusters was initially defined by 5^{50} and then reduced to 5^5 because there are 5 variables and each can have one of 5 values. While the range of possible k values is [1, 5^5], the informative maximum is likely around 3^5 because the questions can be consolidated into five factors, each with three sections on a bell curve (agree, neutral, or disagree). The distributions are continuous, so arbitrary boundaries were created in the case of attempt 2: values lower than one standard deviation from the mean, values higher than one standard deviation from the mean, and values that lay between. An elbow curve was generated to find the most useful number of clusters with the lowest cost. As for cluster interpretation, the target variable was made binary so that variable importance related back to the personality traits associated with individual clusters. While RF and LR models were trialled, the later was most useful; describing magnitude and direction of importance, while the RF could only describe magnitude. LR is therefore most useful to leadership, management, and marketing teams in a business.

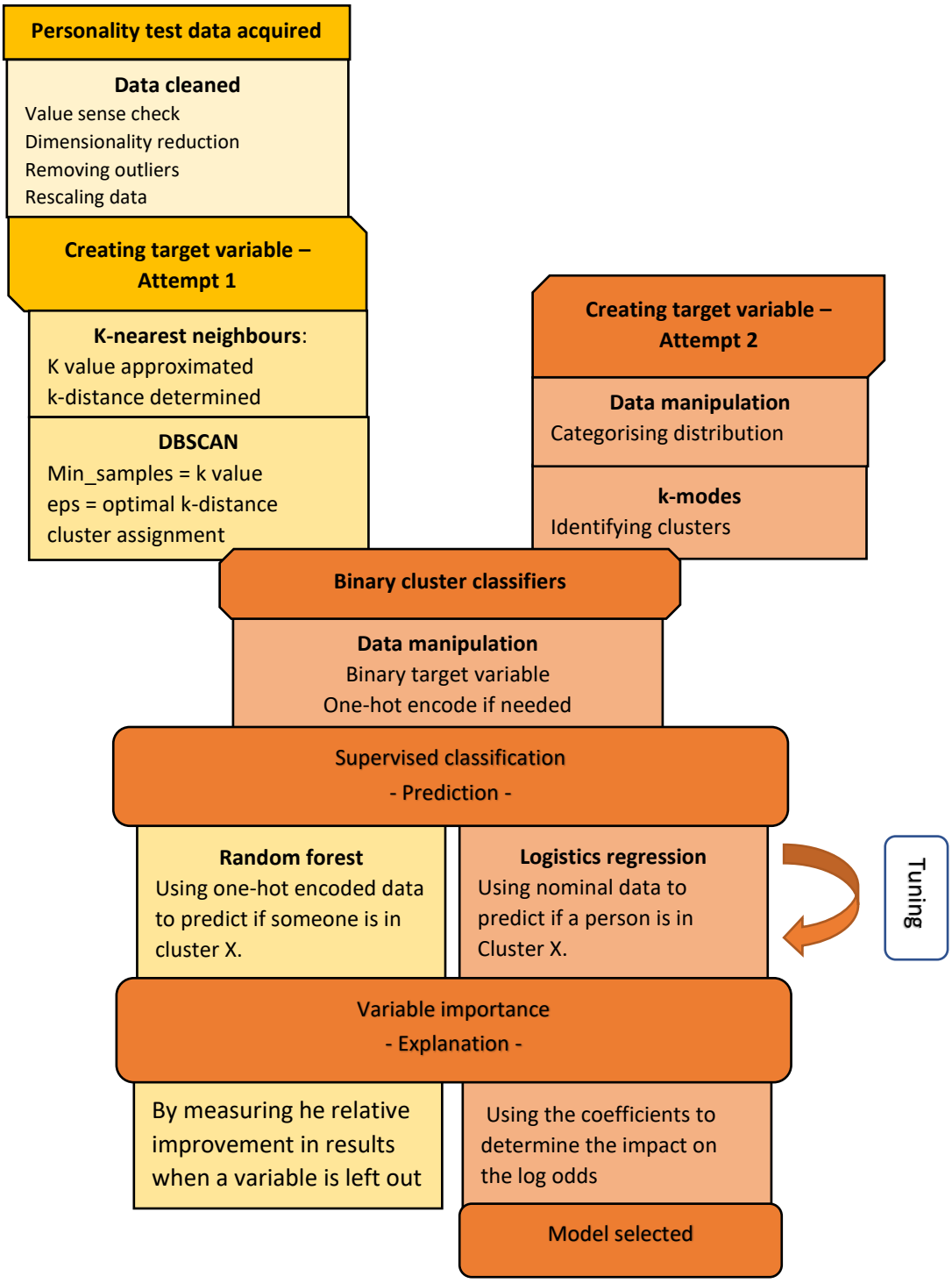

Figure 1: Brief summary of the steps taken.

Results & Discussion

Exploratory Data Analysis

Table 1: A summary of the Data.

Shape	Initial: (1015341, 110), Final: (10199, 16)
Skewness	Unimodal, slight right skew in agreeableness but otherwise normal
outliers	310 instances with values 1.5 x IQR either side of the IQR
Domain	Variable: extraversion , Min value: 10.0 , Max value: 50.0 Variable: neuroticism , Min value: 14.0 , Max value: 45.0 Variable: agreeableness , Min value: 10.0 , Max value: 46.0 Variable: concientiousness , Min value: 10.0 , Max value: 50.0 Variable: Openness_to_exp , Min value: 17.0 , Max value: 46.0

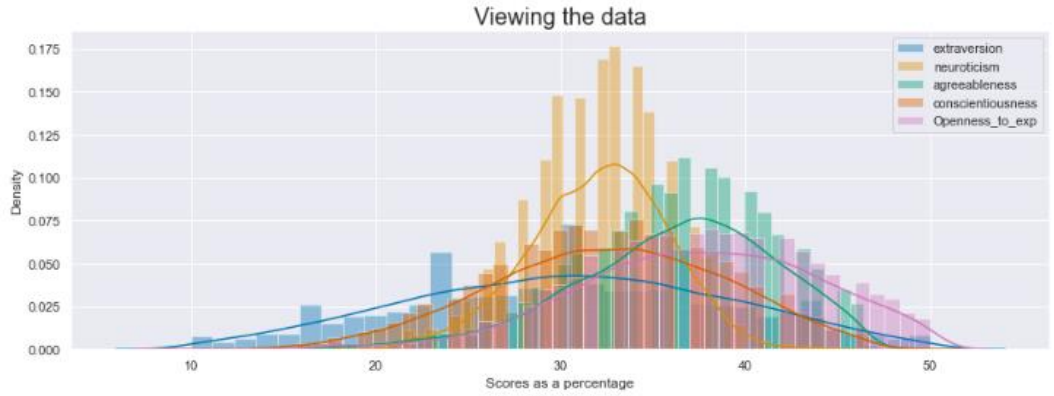

Figure 2: A visual demonstrating the shape, skew, and domain of the data prior to outlier management.

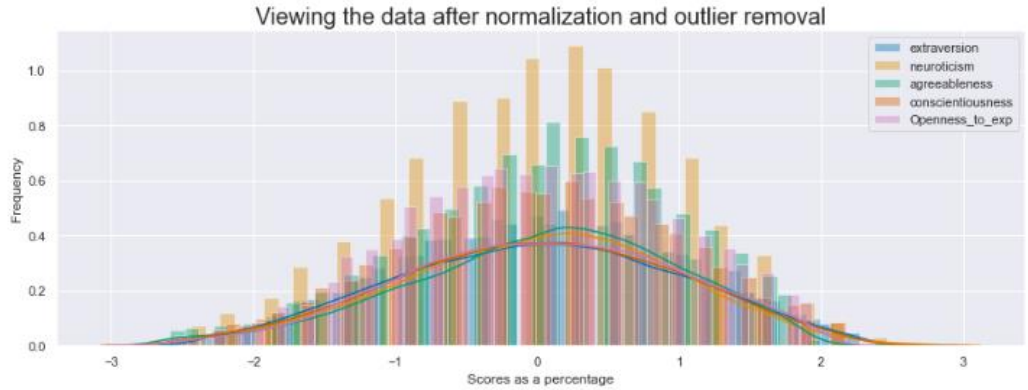

Figure 3: A visual demonstrating the shape, skew, and domain of the data after normalization and outlier management.

Variable interpretation:

Table 2: A table indicating how each variable value should be interpreted.

Variable	Lower end	Upper end
Extraversion	Introverted	Extroverted
Neuroticism	Strong Positivity	Strong Negativity
agreeableness	Self-serving	Very Agreeable
Contentiousness	Disorganized	Organized/meticulous
Openness to change	Unyielding/ conservative	Imaginative/ thoughtful

Clustering

Attempt 1 – k-nearest neighbours + DBSCAN

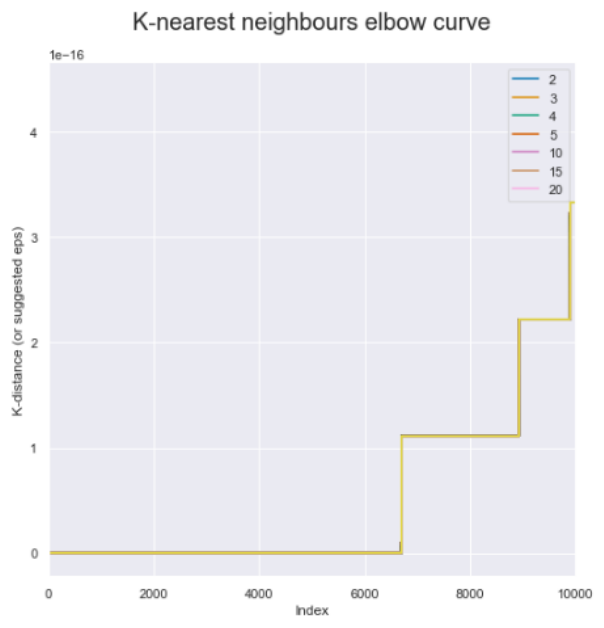

Figure 4: The elbow curve for the k-nearest neighbours generated using all possible values of k found no clear optimal distance to use as the 'eps' parameter in the DBSCAN.

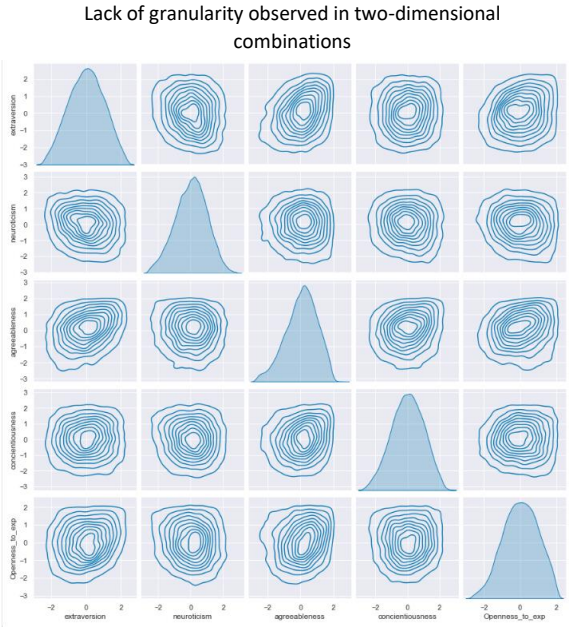

Figure 5: A pair plot indicating the difficulty in identifying clusters. In two dimensions, there are no substantial deviations between plots.

Attempt 2 – k-modes

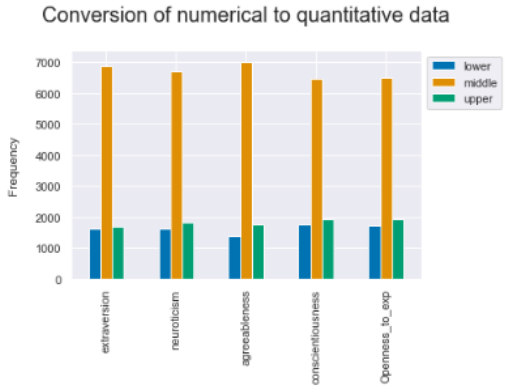

Figure 6: A visual summarizing the distributions observed in the data after conversion from numerical to categorical data.

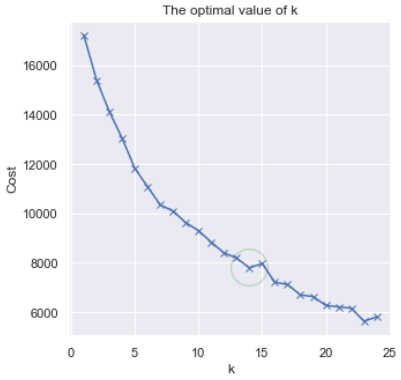

Figure 7: An elbow curve generated by iterating through a selected subset of k-values. K of 14 was selected.

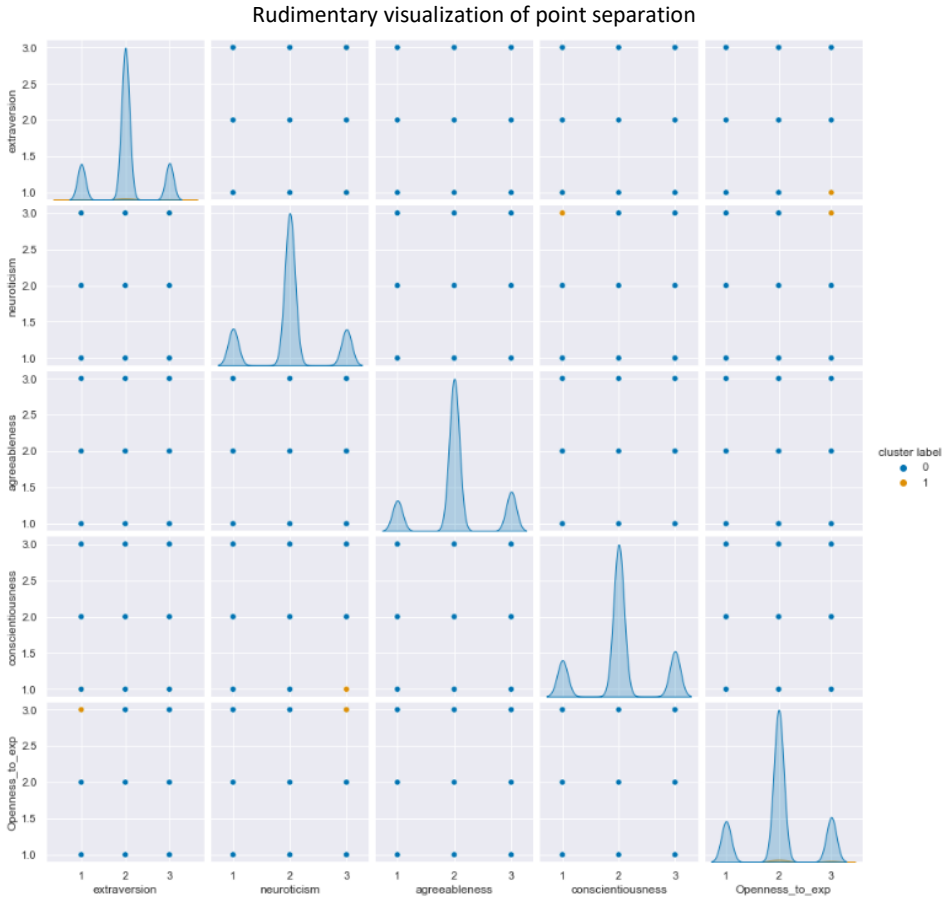

Figure 8: This plot is a visual aid only that demonstrates the clear separation created by converting to categorical data. (Cluster 13 in Orange.

Prediction

Note – all models aim to predict membership of cluster 13.

Random Forest

Table 3: The accuracy scores relating to the tuned RF model predicting a participant's membership of cluster 13.

Score type	Score (untuned)	Score (tuned)
k-folds cross validation score (k=5)	1.0	1.0
Accuracy score	1.0	1.0
ROC AUC score	1.0	1.0

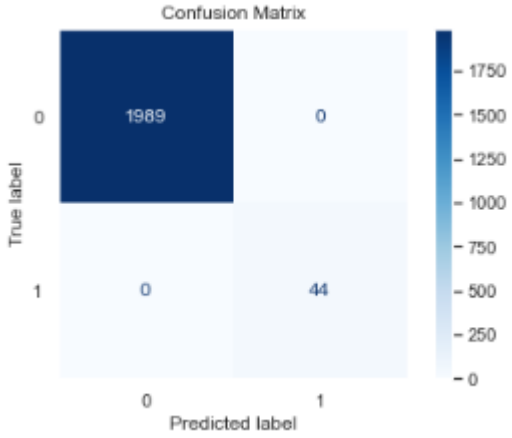

Figure 9: The confusion matrix from the tuned RF models predictions.

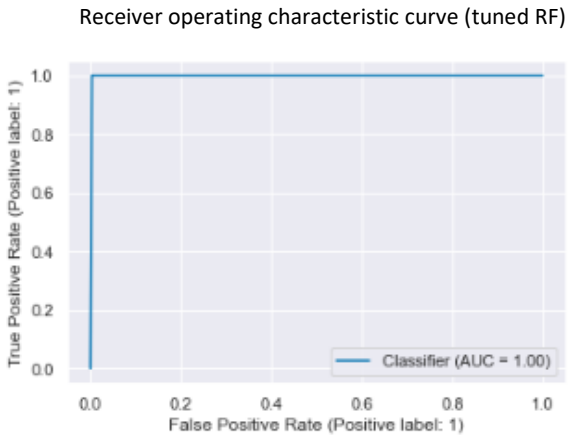

Figure 10: The ROC for the tuned RF models predictions.

Logistic Regression

Table 4: The accuracy scores relating to the tuned LR model predicting a participant's membership of cluster 3.

Score type	Score (untuned)	Score (tuned)
k-folds cross validation score (k=5)	0.99779	0.99779
Accuracy score	0.99852	0.99852
ROC AUC score	0.99428	0.99428

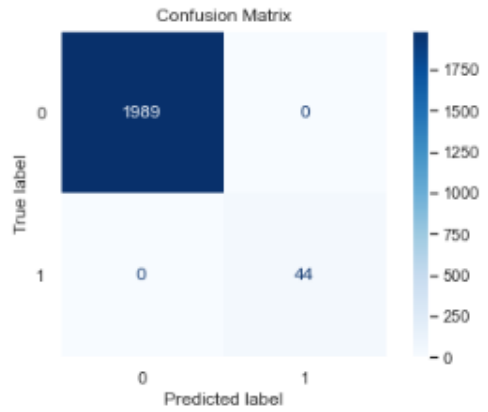

Figure 11: The confusion matrix from the tuned LR models predictions.

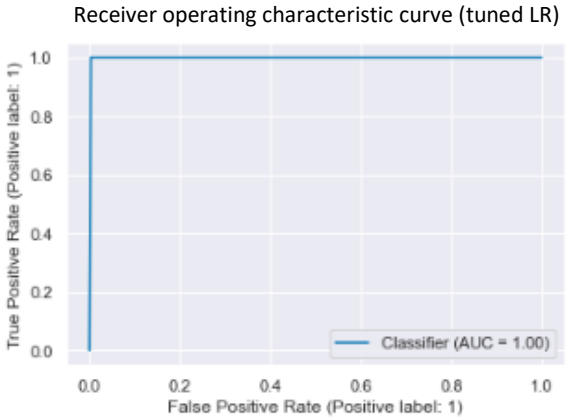

Figure 12: The ROC from the tuned LR models predictions.

Explanation

Random Forest

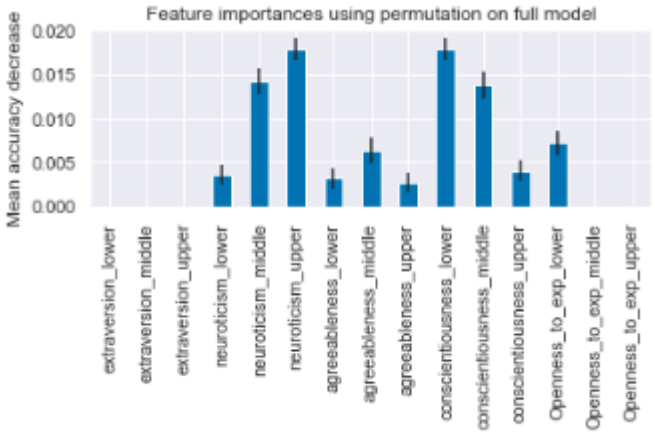

Figure 13: The variable importance identified in the RF model.

Logistic Regression

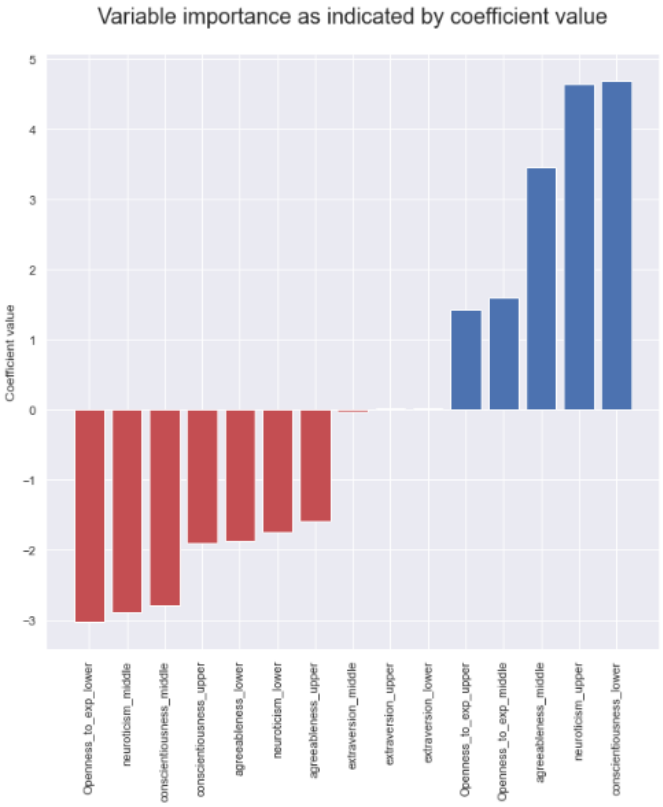

Figure 14: The variable importance identified in the LR model.

Viewing initial 2D scoring pairs coloured by the k-modes clusters

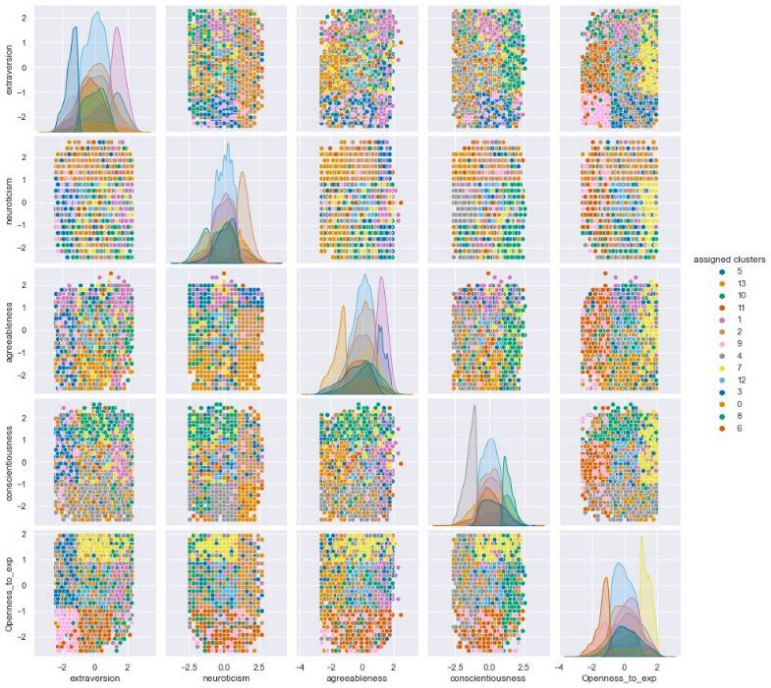

Figure 15: A visualization of the clustering structure identified in k-modes (k =14) in two dimensional combinations.

Discussion

It needs to be said that although attempt 1 was not applicable to this dataset, it is the preferable of the two. The DBSCAN looks for clusters of density in the five-dimensional datascape and identifies groups of interest as well as noise (Béjar Alonso 2013; Bhattacharjee & Mitra 2021; Schubert et al. 2017). Attempt 2 uses k-modes, clustering based on minimal difference between instances which creates locally meaningful clusters (ideal for this case study) but does not necessarily extrapolate into the real world well (Chaturvedi, Green & Caroll 2001). This also means that some variables may be more accurately represented than others and potentially misses some of the less clear personality traits in the cluster set. Figure 5 reveals the difficulty faced in attempt 1. The data is representative of the population which has a good distribution of all possible personality scores, this would not be seen in some businesses that inherently attract specific personality types; in that case the pair plot would reveal signs of granularity. That being said, the data needed to be clustered, and k-modes was an efficient way to accomplish that with the data in its nominal form. The code was written in a way that allowed users to input their preferred number of clusters; with this data, 14 was the optimum value.

With the target variable obtained, the binary classifiers could be constructed and tuned. In attempt 2, the predictors were one-hot encoded and the target variable made binary (in-cluster or out-of-cluster). Then two different classifiers were compared, the RF and LR. The RF model constructs decision trees using bootstrapping (which resamples the data with replacement) in order to construct a classifier which takes the majority vote to determine the class (Couronné, Probst & Boulesteix 2018). Although the data was unbalanced, the cross-validation scores and accuracy scores indicated that the minority class could be predicted well; the opposite of what would be observed if the imbalance was a concern (Zheng & Jin 2020). The LR model is robust to imbalanced data because it operates on probabilities; linking the linear regression formula and the logit function (Couronné, Probst & Boulesteix 2018). When tested on cluster 13, both of these models performed very well across all metrics. Larger clusters are better represented in the training and test sets and will theoretically be predicted more accurately. However, even the smallest clusters in the dataset, cluster 6, had comparable accuracy scores. The ability to predict an individual’s personality is not all that informative to

business strategy, however an understanding of the personality types represented by each cluster most definitely is (Ramadhanti et al. 2020; Woodcock, Green & Starkey 2011). For example, a business may want to target people that aren't open to change because they'll stick with your products for the long term, or perhaps a business wants to know what colour choices and ad campaigns would be most effective to the largest customer group, or perhaps a business wants to know if they can exploit the link between neurotic traits and poorer health (Jackson et al. 2010; Lahey 2009). The variable importance associated with the classifier enables the user to base decisions on the character of each cluster; most specifically when using LR for the reasons outlined above. It is also worth noting that the RF is known to have internal biases when it comes to variable importance (Strobl et al. 2008; Strobl et al. 2007). As an example, in the LR model, cluster 13 marked 'openness_to_exp_lower', 'neuroticism_upper, and 'conscientious_lower', and as having the largest coefficient magnitudes indicating they are the most important variables for that cluster (*Figure 14*); these were in agreement with those observed in the RF, but should more confidence be needed in this assertion, six alternate methods exist to determine variable importance of LR (Figure 13) (Thompson 2009). The first variable is negative indicating the probability of being in this group is lower if you are conservative, while the other two are positive, indicating you are more likely to be in the group if you tend to be more negative and/or disorganized (*Table 2*). With this information, inferences can be made about the group as a whole, as well as the individual.

Deep Learning

This case study is a good example of why deep learning isn't always useful. While it can cluster data using complex patterns and make extremely accurate predictions, it cannot always offer as much explanatory power (Ras, Van Gerven & Haselager 2018). The method outlined is likely more time and cost effective and retains that explanatory power. If any aspect of this would benefit from a deep learning algorithm, it would be:

1. In the data acquisition phase, a deep learning algorithm could be trained to interpret visual or audio recording and output more accurate trait scores (Mehta et al. 2020). These scores would be free from self-reporting errors and

provide a more accurate measure of each trait, however constructing such an algorithm would require great care as these algorithms can inadvertently propagate stereotypes and social injustices if patterns appear in the training data (Zhang & Han 2022).

2. The clustering technique as the current method accomplishes the task somewhat arbitrarily and in a manner that makes it difficult to determine if the optimal structure has been recognised. Whereas an autoencoder or an adaptive resonance theory model would identify more optimal structures while providing as much explanation (Károly, Fullér & Galambos 2018; Tinwala & Rauniyar 2021).

However, should a business have access and means, it would likely be better spent on other aspects of the business.

Conclusion

The process suggested here effectively found groups of common personalities in a customer base, assigned new customers a grouping, and provided an indication of the personality traits common to a specific cluster. An online quiz is a simple tool most businesses could implement; however, future models should aim to incorporate information about age, gender, and time taken per question to more accurately describe similar customer personality groups. Additionally, businesses should endeavour to find ethical ways to use audio and/or visual data that is free from bias relating to self-reporting and can be managed with natural language processing, facial recognition, etc. to do what psychologists are already doing at microscales.

References

Béjar Alonso, J 2013, 'Strategies and algorithms for clustering large datasets: A review'.

Bhattacharjee, P & Mitra, P 2021, 'A survey of density based clustering algorithms', *Frontiers of Computer Science*, vol. 15, no. 1, pp. 1-27.

Biesanz, JC & West, SG 2004, 'Towards Understanding Assessments of the Big Five: Multitrait-Multimethod Analyses of Convergent and Discriminant Validity Across Measurement Occasion and Type of Observer', *Journal of Personality*, vol. 72, no. 4, pp. 845-876.

Bollich, KL, Doris, JM, Vazire, S, Raison, CL, Jackson, JJ & Mehl, MR 2016, 'Eavesdropping on character: Assessing everyday moral behaviors', *Journal of Research in Personality*, vol. 61, pp. 15-21.

Castillo, J 2017, 'The relationship between big five personality traits, customer empowerment and customer satisfaction in the retail industry', *Journal of Business and Retail Management Research (JBRMR)*, vol. 11, no. 2.

Chaturvedi, A, Green, PE & Caroll, JD 2001, 'K-modes Clustering', *Journal of Classification*, vol. 18, no. 1, pp. 35-55.

Chen, Z & Lin, T 2017, 'Automatic personality identification using writing behaviours: an exploratory study', *Behaviour & Information Technology*, vol. 36, no. 8, pp. 839-845.

Chittaranjan, G, Blom, J & Gatica-Perez, D 2011, 'Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones', in IEEE.

Couronné, R, Probst, P & Boulesteix, A-L 2018, 'Random forest versus logistic regression: a large-scale benchmark experiment', *BMC Bioinformatics*, vol. 19, no. 1.

Gerlach, M, Farb, B, Revelle, W & Nunes Amaral, LA 2018, 'A robust data-driven approach identifies four personality types across four large data sets', *Nature Human Behaviour*, vol. 2, no. 10, pp. 735-742.

Haslam, N 2020, 'Bell-Shaped Distribution of Personality Traits', in Springer International Publishing, pp. 441-443.

Jackson, JJ, Wood, D, Bogg, T, Walton, KE, Harms, PD & Roberts, BW 2010, 'What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC)', *Journal of Research in Personality*, vol. 44, no. 4, pp. 501-511.

Károly, AI, Fullér, R & Galambos, P 2018, 'Unsupervised clustering for deep learning: A tutorial survey', *Acta Polytechnica Hungarica*, vol. 15, no. 8, pp. 29-53.

Lahey, BB 2009, 'Public health significance of neuroticism', *American Psychologist*, vol. 64, no. 4, p. 241.

Lavalley, MP 2008, 'Logistic Regression', *Circulation*, vol. 117, no. 18, pp. 2395-2399.

Lin, LY 2010, 'The relationship of consumer personality trait, brand personality and brand loyalty: an empirical study of toys and video games buyers', *Journal of product & brand management*, vol. 19, no. 1, pp. 4-17.

Mehta, Y, Majumder, N, Gelbukh, A & Cambria, E 2020, 'Recent trends in deep learning based personality detection', *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2313-2339.

Ramadhanti, AR, Bastikarana, RS, Alamsyah, A & Widiyanesti, S 2020, 'Determining Customer Relationship Management Strategy With Customer Personality Analysis Using Ontology Model Approach', *Jurnal Manajemen Indonesia*, vol. 20, no. 2, p. 83.

Ras, G, Van Gerven, M & Haselager, P 2018, 'Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges', in Springer International Publishing, pp. 19-36.

Reece, TJ 2009, 'Personality as a Gestalt: A cluster analytic approach to the Big Five'.

Sander, J, Ester, M, Kriegel, H-P & Xu, X 1998, *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169-194.

Saucier, G 2002, 'Orthogonal markers for orthogonal factors: The case of the Big Five', *Journal of Research in Personality*, vol. 36, no. 1, pp. 1-31.

Schubert, E, Sander, J, Ester, M, Kriegel, HP & Xu, X 2017, 'DBSCAN Revisited, Revisited', *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1-21.

Shah, V & Modi, S 2021, 'Comparative Analysis of Psychometric Prediction System', in IEEE.

Strobl, C, Boulesteix, A-L, Kneib, T, Augustin, T & Zeileis, A 2008, 'Conditional variable importance for random forests', *BMC Bioinformatics*, vol. 9, no. 1, 2008-12-01, p. 307.

Strobl, C, Boulesteix, A-L, Zeileis, A & Hothorn, T 2007, 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC Bioinformatics*, vol. 8, no. 1, 2007-12-01, p. 25.

Stromberg, J & Caswell, E 2015, 'Why the Myers-Briggs test is totally meaningless', *Vox*.

Talasbek, A, Serek, A, Zhaparov, M, Yoo, S-M, Kim, Y-K & Jeong, G-H 2020, 'Personality classification experiment by applying k-means clustering', *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 16, pp. 162-177.

Thompson, D 2009, 'Ranking predictors in logistic regression', *Paper D10-2009. Online available at http://www. mwsug. org/proceedings/2009/stats/MWSUG-2009-D10. pdf.(visited 2015, June 25)*.

Tinwala, W & Rauniyar, S 2021, 'Big Five Personality Detection Using Deep Convolutional Neural Networks'.

Vazire, S & Carlson, EN 2010, 'Self-knowledge of personality: Do people know themselves?', *Social and personality psychology compass*, vol. 4, no. 8, pp. 605-620.

Wiernik, BM, Yarkoni, T, Giordano, C & Raghavan, M 2020, 'Two, five, six, eight (thousand): Time to end the dimension reduction debate!'.

Woodcock, N, Green, A & Starkey, M 2011, 'Social CRM as a business strategy', *Journal of Database Marketing & Customer Strategy Management*, vol. 18, no. 1, pp. 50-64.

Zhang, J & Han, Y 2022, 'Algorithms Have Built Racial Bias in Legal System-Accept or Not?', in Atlantis Press.

Zheng, W & Jin, M 2020, 'The effects of class imbalance and training data size on classifier learning: an empirical study', *SN Computer Science*, vol. 1, no. 2, pp. 1-13.