DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Strategical Guideline Proposal: A Cluster Analysis Approach

Prepared for client:

ABCDEeats Inc.

Company:

Group 99

Consultants:

Hugo Fonseca, 20240520

Ricardo Pereira, 20240745

Jéssica Vicente, 20230744

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1.1. INTRODUCTION

The following work was performed for client ABCDEats Inc. following the need for a well-grounded strategical framework for the marketing department. Prepared by our most excellent consultants, the work provides solid foundation for future marketing strategy and investment horizon.

Section 2.1 starts with addressing issues such as handling duplicates, filling missing values, and treating outliers in ways that preserve the natural variability of the dataset. In Section 2.2, we describe the three perspectives taken: Recency, Frequency, and Monetary Value (RFM); temporal consumption patterns (Time); and customer preferences across diverse cuisines (Diversified Customers).

To improve clustering quality, we implemented as outlier-detection algorithm DBSCAN and Isolation Forest, described in Section 3.1, with hyperparameter tuning tailored to each perspective. Section 3.2 evaluates the clustering algorithms—K-Means, Gaussian Mixture Models (GMM), and Ward’s Hierarchical Clustering—using as performance metrics the Silhouette score and the Calinski-Harabasz index to determine the optimal number of clusters and their respective characteristics. The results are visualized in Section 4, where we use dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). Section 5 discusses the cluster results in detail, connecting them to actionable recommendations for business strategies. Section 6 describes further recommendations which are relevant to management regarding issues noted by our consultants.

2. PREPROCESSING

2.1. UNIVARIATE PREPROCESSING

2.1.1. Data Quality and Other Issues

We dropped every duplicated entry (13 instances) and kept the original. We filled (727) null values in customer_age with random integers valued between the first and the third quartile. Out of the (106) null values of the first_order column, 104 of these ordered once, and 2 others ordered twice, but all of them had a last_order value of 0, so we might as well just assume these orders took place on the first day of the study period by setting the null values equal to 0. For every instance *i*, we set HR_0 as the difference between the amount of orders as obtained by the sum in the DOW colums and the orders as obtained by the sum in the HR colums excluding HR_0: $HR\_0_{\,i} = \sum_{n\,=\,0}^{6} DOW\_n_{\,i} - \sum_{m\,=\,1}^{23} HR\_m_{\,i}$. We relabeled customer_regions as A, B, C, and so on where population is respectively decreasing.

2.1.2. Univariate Outliers Treatment

For the numerical features customer_age, vendor_count, product_count, first_order, last_order, we identified outliers using a threshold of 2.5 times the IQR below the first quartile or above the third quartile. We chose this more-conservative way of classifying outliers to decrease the effect of univariate outlier treatment and to leave more room for multivariate outlier treatment. This is only to deal with very extreme cases. These outliers were replaced with random values sampled from the 5th–10th percentile (for lower outliers) or the 90th–95th percentile (for upper outliers) of the feature’s distribution. This way we manage to both keep their "tail status" while keeping some degree of noise, thereby avoiding overfitting to specific thresholds.

3. MULTIVARIATE PREPROCESSING

3.1. APPROACHES AND TREATMENTS

3.1.1. Perspective 1: Recency, Frequency and Monetary Value (RFM)

We included the RFM perspective because it is a well-established marketing analysis approach to identify a firm's most valuable customers based on the nature of their spending habits. For this approach, we encapsulated 3 variables: Lifetime, which is the difference between last and first order; Frequency, which is the total amount of orders per customer; and Monetary Value, which is the total expenditure per customer. Because there are a significant share of customers who only ordered once, and are therefore not that much relevant, we manually classified these as outliers for all treatment. Treatment 1 consisted on only standardizing the variable set. Treatment 2 consisted in adding 1 unit to all three variables (because log function is not defined when the input is zero), taking the log, and then standardizing it. For treatment 3 we transformed the variables into percentile scores – for example, the median and top spender would have a value of 50 and 100 for Monetary Value, respectively. We found no use in standardizing this treatment since all variables were now on the same scale, ranging from 0 to 100.

3.1.2. Perspective 2: Time

The Time approach attempted at finding temporal consumption patterns among customers. We created a variable named Workdays and Weekend, which, for a given customer, represents the percentage of orders made from Monday to Friday inclusive and Saturday to Sunday out of all orders, respectively, multiplied by the total expenditure. We made the same for Lunch and Dinner variables, which, for a given customer, are percentage of orders made from 8AM to 2PM and from 4PM to 9PM, respectively, multiplied by total expenditure. The 3 treatments for time were the same as in RFM.

3.1.3. Perspective 3: Diversified Customers

This perspective involved first getting the percentage spent per cuisine out of the total expenditure. We then multiplied each percentage by product count, such as to get a weighted amount of orders per cuisine. We then selected the three cuisines with the most amount of average product count: Asian, American and Other. We then extracted customers who consumed all three catergories (which amounted to 969 customers, which is not that much, but should prove useful to study the behaviour of these customers and how they shall interact with other clusters).

3.2. OUTLIER-DETECTION ALGORITHMS AND HYPERPARAMETER-TUNING

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that uses a simple minimum density level estimation. It is based on a threshold for the number of neighbors, min_samples, and a value for a radius epsilon, eps (and uses euclidean distance). DBSCAN classifies instances as either 'core', 'non-core' and outliers. However, we are only interested in DBSCAN's outlier detection applications, so we will only address its outputs as either inliers ('core' and 'non-core') or noise (outliers). Regarding min_samples, Sander et al. suggest setting it to twice the dataset dimensionality, i.e., min_samples = 2 · dim. So, for RFM, which has three features, we set min_samples = 6. We computed the optimal epsilon such that we would obtain a given target percentage of noise. How much percentage points do we want to 'clean'? E. Schubert et al. suggests that the desirable amount of noise will usually be between 1% and 30%. So, we will allocate 10% to noise detected by DBSCAN and save 10% for Isolation Forest, which totals 20% of the dataset, thus assuming some degree of 'noisiness', but not to the extent to which we would take advantage of the full 30%. We applied the 20% for all treatments of all perspectives to simplify our analysis. So, for example, for the Diversified Customers perspective, a 10% noise level implies $\varepsilon \approx 0.2711$.

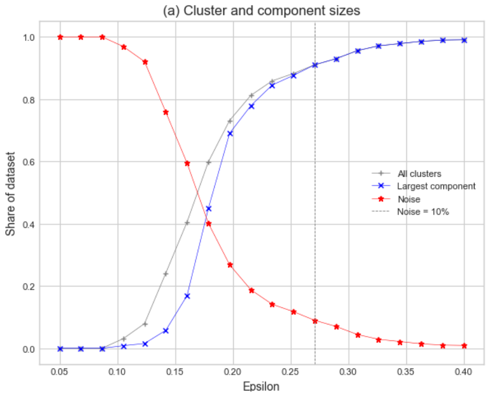
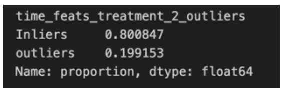


Figure 1 – Visual Representation of ε choice (Diversified Customers)

The original paper defines **Isolation Forests**, or iForests, as ensembles of iTrees where anomalies are instances characterized by shorter average path lengths. iTrees are *proper binary trees* – each node in the tree has exactly zero or two daughter nodes, that randomly partitions data, thus producing noticeable shorter paths for anomalies. Fewer instances of anomalies result in a smaller number of partitions – shorter paths in a tree structure, and instances with distinguishable attribute-values are more likely to be separated in early partitioning. This model has three main hyperparameters: the number of trees, t; the sub-sampling size, ψ; and the contamination level. The original paper shows that the average path length of x_i converges when t increases. This observation carries a useful heuristic for the choice of t: we should choose the ensemble size to be a point in which the average path length (or, equivalently, anomaly score, which embeds the average path length) either **is** or **begins to be** relatively stable. Note that a higher t costs computational efficiency, meaning that we are interested in finding an efficient allocation of t such that anomaly scores stability is maximized while also minimizing computational 'expenditure'. To find the optimal level of trees, we trained a series of iForest models, varying the number of estimators over a logarithmic scale (just as in the original paper), ceteris paribus, and extracted the anomaly score of 50 randomly sampled inliers and 50 randomly sampled outliers. The results are shown in figure 2. The asymptotic behavior of the score curves is plotted by the horizontal dotted lines to indicate the stabilizing score tendencies, and the vertical green dotted line represents our choice of trees. Finally, the original paper found that a small sub-sampling size, ψ, provided high Area Under Curve (AUC) and low processing time, and a further increase of ψ was not necessary. The authors found an optimal ψ when $\psi = 0.00045 * DataSet_{size}$ or $\psi = 0.0018 * DataSet_{size}$. We do not have a data for 'true outlier' classification, so we cannot use AUC, but we followed similar proportions as the original paper did: $\psi = 0.002 * DataSet_{size}$. Finally, we apply the iForest algorithm to the inliers from DBSCAN, which is 90% of the dataset. Therefore, if we want to have another 10% of the dataset being classified as outliers, conditional on being inliers of DBSCAN, the contamination level must be set to approximately 11.11% (conditional probability). This way we obtain a dataset with approximately 80% inliers and 20% outliers, as intended, **assuming outliers are classified solely on this section's described method (and no manual outlier detection)**. A summary of chosen metrics for outlier detection algorithms lies on following table:



```
time_feats_treatment_2_outliers
Inliers     0.800847
outliers    0.199153
Name: proportion, dtype: float64
```

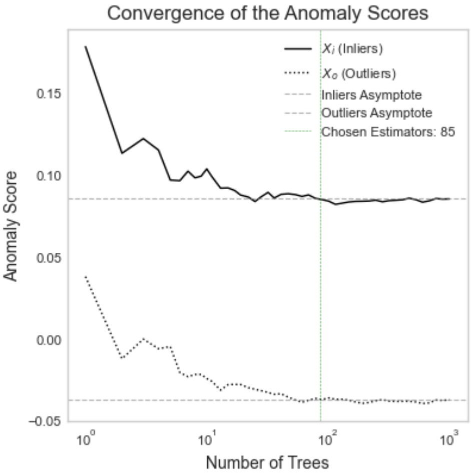


Figure 2 - Convergence of Anomaly Scores as an indicator of optimal number of trees (Diversified Customers);

	RFM (1)	RFM (2)	RFM (3)	Time (1)	Time (2)	Time (3)	Diversified Customers
ε (DBSCAN)	0.1237	0.0668	0.0716	0.3105	0.1741	0.0201	0.5204
t (iForest)	60	150	40	150	200	70	85

Detected outliers were not discarded from our analysis in general terms but put aside for the purpose of obtaining better, less deviated centroid/cluster results, and, on a posterior phase, were class-classified using the obtained centroids. By excluding 'badly-behaved' instances of the centroid computations, we should get centroids that better reflect cluster group 'core'/unique characteristics.

3.3. TREATMENTS, ALGORITHMS AND N CLUSTER CHOICES

For each outlier-cleaned treatment of the clustering approaches, we evaluated the performance of three clustering algorithms—K-Means, Gaussian Mixture Model (GMM), and Hierarchical clustering (Ward's Linkage)—using the number of clusters (n) as the independent variable, ranging from 2 to 6. K-Means is a centroid-based algorithm that partitions the dataset into k clusters by minimizing inertia (Within-Cluster Sum-of-Squares, WCSS), assuming spherical cluster shapes. GMM extends K-Means by modeling data as a mixture of Gaussian distributions, allowing clusters to take elliptical shapes. Hierarchical clustering constructs a dendrogram by iteratively merging or splitting clusters, and we specifically used Ward's linkage, which minimizes variance within merged clusters and is effective for detecting spherical structures. Clustering performance was evaluated using two metrics: the Silhouette score (ranging from −1 to 1, with higher values indicating well-defined and separable clusters) and the Calinski-Harabasz (CH) index (unbounded, with higher values indicating compact and well-separated clusters). it is important to note that the Silhouette score can favor fewer clusters, and the CH index may be sensitive to the dataset's sample size. To improve comparability, all variables were standardized prior to clustering.

For the RFM approach, the plots consistently indicate that n=2 clusters is optimal, as both the Silhouette score and the CH index reach their maximum at this point on average. Among the three algorithms, K-Means and GMM generally outperform Ward's Hierarchical clustering. Specifically, for n=2 clusters, K-Means achieves the highest Silhouette and CH index scores across all three treatments, except for Treatment 2, where its performance is similar to GMM. Based on these metrics, K-Means looks like the most suitable model. Consequently, we proceed with Treatment 1, which involves standardization and outlier treatment, and select n=2 clusters with K-Means, as it achieves the highest global scores for both metrics.

For the Time perspective clusters, the performance trends are different. The scores for Treatments 1 and 2 exhibit more concave shapes, suggesting the existence of n=4 clusters. Both treatment 1 and 2 exhibit an elbow-like pattern when n=4 clusters for KMeans and GMM, but particularly so for GMM, and this is what we will value the most, regardless of the fact that KMeans scores slightly above for n=4. GMM scores higher for treatment 1, so that is what we will use in our analysis, along with 4 clusters.

For the Diversified Customers set, where only one treatment was applied, n=4 clusters is clearly the optimal choice. Based on the Silhouette and CH index metrics, K-Means outperforms the other models and is therefore selected as the clustering algorithm for this dataset.

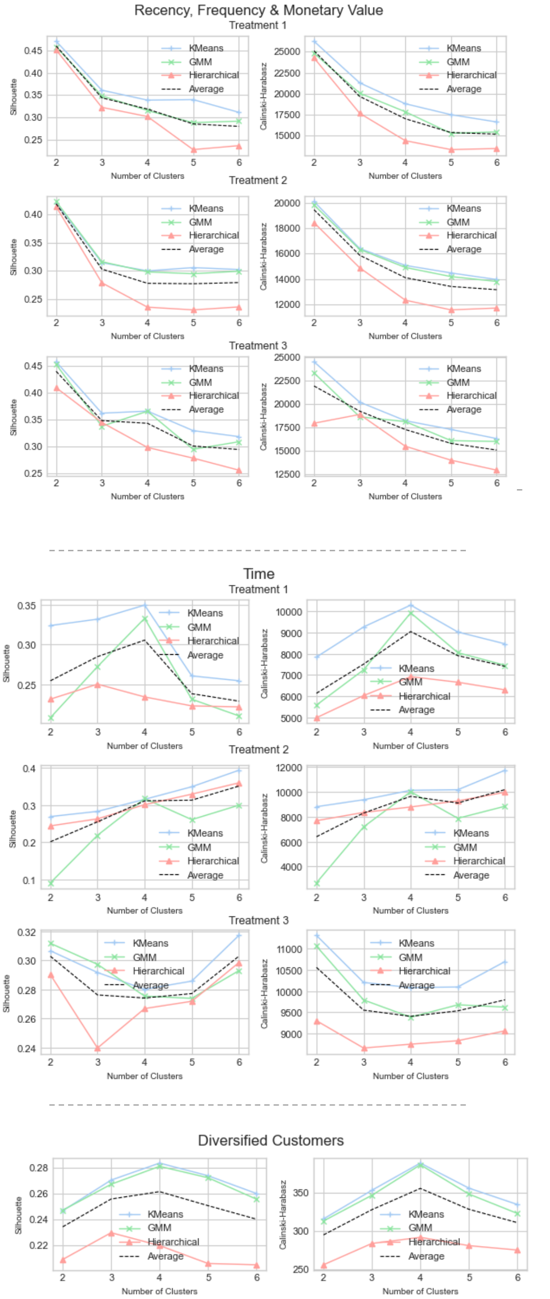


Figure 3 - Silhouette Scores and Calinski-Harabasz Indexes for performance comparison between KMeans, Gaussian Mixture Models, and Hierarchical with Ward's Linkage method for varying n clusters for different perspectives and treatments;

4. CLUSTER RESULTS

We ('fit'/) trained the models using the outlier-free dataset and (hit 'predict'/) then we classified the global dataset based using the obtained centroids. The results for the clustering of the global dataset, in the case of RFM and Time, and of the training set in the case of Diversified Customers, are shown on the right. On the left column, plot a Principal Component dimensionality reduction, which captures linear relationships in the data. On the right we have a Uniform Manifold Approximation and Projection (UMAP), which captures nonlinear relationships in the data.

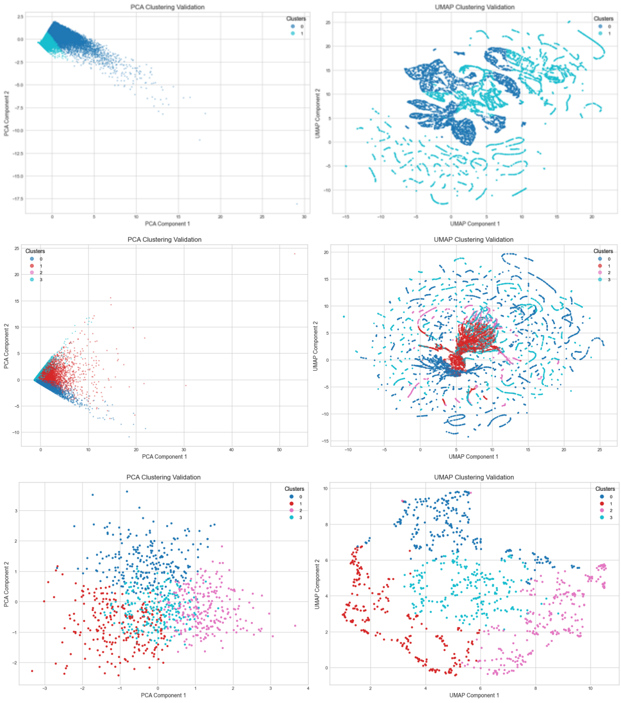
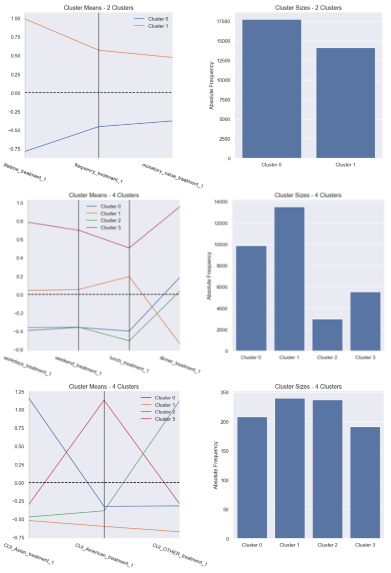


Figure 4 - Principal Component and UMAP dimensionality reduction for the final clustering results of the RFM, Time and Diverse Customers approach, in descending order respectively.

The plots show interesting results, as RFM (top row) has a significant degree of cluster separability for both decomposition algorithms, while Diverse Customers' (bottom row) quality deteriorated in PCA and Time (mid row) only carries decent results for PCA. This hints that while RFM (top) can be both linearly and non-linearly separable, Diverse Customers (bottom) has a non-linear nature and Time (mid) has a linear nature. In a way, Diverse Customers (bottom) has a similar structure for PCA and UMAP: cluster 3 is in the center; 1 is on the lower left; 0 is at the top and 3 is at the lower right.

The cluster centers indicate that, for RFM, there are two types of customers: High value (Cluster 1) and Low-Value (Cluster 0). High-Value customers order consume over a long time span (high lifetime), consume frequently (high frequency) and spend a lot (high monetary value. The contrary follows for Low-Value customers.

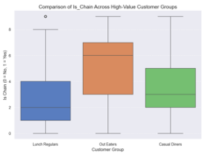


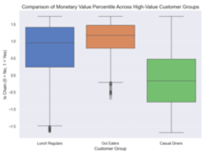
The cluster centers of Time, indicates 4 customer types, but Cluster 0 and Cluster 2 have very similar patterns, so we merged the two into one. That being said, we were left with three main customer types: the Out-Eaters (Cluster 3); the Lunch-Regulars (1); and the Casual Diner (0 and 2). Out-Eaters eat out regularly both from Monday to Friday and on the weekends. Lunch Regulars have average consumption patterns for workdays and weekends, but only for lunch, because they score lower for dinner. Casual Diners sporadically go out for dinner, but not much else.

The Diversified Customers show that, among the customers who consumed all these three cuisines, there are those who consume equally of all, and those who consume more of a particular type. This will become more relevant once we look at how these fit into the other cluster approaches.

5. EXPLORATION AND STRATEGICAL RECOMMENDATIONS

By combining the results from approach RFM and Time, we were able to find the following patterns (we found no distinctions in age between groups. See Conclusion paragraph 2 for a brief summary of what was performed in this section):





- **High-Value Out-Eaters**: 47% of this group comes from region B (the second most populated area). They show a clear preference for chain restaurants, which sets them apart from other high-value groups. This group is also concentrated at the top of the monetary value percentile ranks, making them the most valuable segment. They tend to dine out both on workdays and weekends, spending above average. Notably, 31% of customers in this group prefer miscellaneous cuisines, which complements their preference for chain restaurants.
- **High-Value Lunch-Regulars**: 70% of this group is from region A (the most populated area). A significant 58% prefer Asian cuisine, indicating a clear preference for specific dining experiences during lunch.
- **High-Value Casual Diners**: 48% of this group is based in region C (the third most populated area). While relatively small, this group enjoys occasional dinners out, with 27% of members favoring American cuisine.

Based on these characteristics, we recommend that ABCDEats Inc. implements region-specific marketing strategies to best exploit to each group’s preferences:

- **Region B (Out-Eaters)**: Since this region has a large population of high-value Out-Eaters who prefer chain restaurants and dine out on both weekdays and weekends, we suggest launching a "Loyalty Reward" program. Customers who visit a restaurant for five consecutive days could receive a free meal as a reward.
- **Region A (Lunch-Regulars)**: As Lunch-Regulars dominate this region and have a strong preference for Asian cuisine, we recommend introducing a lunch-specific discount for Asian meals.
- **Region C (Casual Diners)**: With a preference for occasional dinners and a liking for American cuisine, we suggest implementing a time-specific promotion like a "Happy Hour" discount during specific evening hours. This could give ABCDEats a competitive advantage, especially against other restaurants unaware of this group's preferences.

6. OTHER RECOMMENDATIONS

We believe demographic indicators like income level, occupation, or household size might reveal subtle patterns in dining behavior, and we believe ABCDEats could benefit from collecting some of these variables into their database. For example, customers with higher disposable incomes may be more inclined toward more expensive dining experiences. Additionally, we believe some of the features included, such as ‘Asian Cuisine’ may be an over-generalization, so it might be useful to reconsider some of the utilized labels. Additionally, further research should be done comparing the clustering performance for different subsets of the dataset.

7. CONCLUSION

We started our study by addressing some of the issues discovered in the exploratory data analysis (EDA) phase, such as removing duplicates, filling nulls and relabeling regions. We then identified univariate outliers using a threshold of 2.5 times the interquartile range (IQR) below the first quartile or above the third quartile, following a more conservative approach, under the aim of addressing only the most extreme cases, thereby maintaining the integrity of the dataset while preparing it for further analysis. These outliers were replaced with random values sampled from the 5th–10th percentile (for lower outliers) or the 90th–95th percentile (for upper outliers) of the feature's distribution. This way we manage to both keep their "tail status" while keeping some degree of noise, thereby avoiding overfitting to specific thresholds.

Before getting into the clustering itself, we first cleaned the data from multivariate outliers in order to better train the chosen algorithms. We used two outlier-detection algorithms: DBSCAN and iForest. For each, we justified in an objective and clear way the criteria for the choice of hyperparameters, where, for iForest, **we replicated a visualization from the original paper and further formed a heuristic based on the paper's insights which was not explicit in the paper – the number of trees should be set in such a way that it stabilizes anomaly scores, or average path lengths if that data available, and but small enough to minimize computational complexity**. For the choice of which treatments and which clustering algorithms, we extensively looked for the best combination, only to find that barely any treatment yielded the best results and that KMeans tended to outperform other algorithms.

We delved into multivariate preprocessing (section 3), in which 3 perspectives were formulated to later aid in the profiling phase (section 5). We went with Recency, Frequency and Monetary Value, which later allowed us to distinguish between high-value and low-value customers based on how long, how frequently and how much spent the customer registered on the dataset. We coupled this view with a temporal perspective, which allowed us to find the main behaviours of customers: those who go out to eat frequently (Out-Eaters), those who go regularly for lunch (Lunch-Regulars) and those who like to treat themselves with going out for dinner sometimes (Casual Diners). Using the RFM outcomes, we combined the two and extracted the high value groups within temporal clusters. We then used a dataset with customers who ordered from the three biggest cuisine types to form taste types and used these to figure out each high-value customer type's preference, further providing concrete, actionable measures for ABCDEats to take. That is, we found that the three top populated regions all have a different high-value group that can be exploited for economic purposes: region A has high value customers who love lunch and Asian meals; region B has the most valuable customers, which are those who eat out regularly, both lunch and dinner, eat whatever and go to a lot of chain restaurants. Region C has Casual Diners, which are dinner lovers. We tailored a specific business strategy for each region: a "Loyalty Reward" program for region A; a lunch-specific discount for Asian meals in region B; and a "Happy Hour" discount during specific evening hours for region C.

BIBLIOGRAPHICAL REFERENCES

Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. Isolation Forest

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann

Aggarwal, C. C., (2015). *Data Mining – The Textbook*, Springer

Géron, Aurélien (2017). *Hands-On Machine Learning – Concepts, Tools, and Techniques To Build Intelligent Systems*, O'Reilly