Research Article

Churn Detection and User Classification via Machine Learning in the Food and Beverage Sector

Deniz Altay Avcı[1*], Gürkan Şahin[2], Murat Kan[3]

[1*] Department of Research and Development, adesso Turkey, Orcid ID: https://orcid.org/0009-0006-6443-6008, e-mail: deniz.avci@adesso.com.tr
2 Department of Research and Development, adesso Turkey, Orcid ID: https://orcid.org/0009-0003-7739-6760, e-mail: gurkan.sahin@adesso.com.tr
3 Department of Research and Development, adesso Turkey, Orcid ID: https://orcid.org/0009-0005-9647-4730, e-mail: murat.kan@adesso.com.tr
* Correspondence: denizaltayavci@gmail.com; +905318353606

Received 17 September 2024
Received in revised form 15 November 2024
In final form 26 December 2024

Reference: Altay Avcı, D., Şahin, G., & Kan, M. (2024). Churn detection and user classification via machine learning in the food and beverage sector. The European Journal of Research and Development, 4(4), 1-16.

Abstract

In the modern business world, detecting and predicting customer behavior is one of the key factors for any operation to achieve their goals. Cu0073tomer churn is one of these behaviors of interest, which makes churn detection and prediction a hot topic in the Machine Learning domain. The customer data that was studied is obtained from a global food and beverage company's operations in Turkey: Their gift-based mobile application rewards customers who buy their products, and the user data of many sorts is stored within its database. In this study, the unlabeled customer data of a large scale was analyzed and classified via the combination of various supervised and unsupervised ML methods such as K-Means Clustering, Random Forest, Support Vector Machines, Logistic Regression, XGBoost. Then, a score-based churn detection & prediction algorithm is developed after picking the best performing models based on their performance metrics.

Keywords: *Customer Churn, Customer Segmentation, Machine Learning, K-Means Clustering, Silhouette*

1. Introduction

Understanding the underlying factors that cause churn is crucial if a company wants to detect churners. It is even more critical if a company wants to take a step further and predict possible future churners. These factors are usually complex, multi-dimensional and sector-dependent, which makes the usage of state-of-the-art data science and machine learning methods a must. Thus, it is no surprise that ML-based churn analysis methods are widely used in various sectors ranging from banking, e-commerce to SAAS [1-9]. In this study, a popular global food and beverage company's operations in Turkey is investigated. Specifically, their gift-based mobile promotion application is at the center of the work. The user-specific data from the database of the subject application is utilized to detect previous and potential churners using various Machine Learning methods. First, the users are assigned to different clusters with certain meanings by utilizing a k-means based unsupervised clustering algorithm. Then, a set of supervised machine learning algorithms and feature impact analysis is used to verify and describe these clusters. Finally, a score criterion is developed to rank the users according to their churn potential. Besides performing exceptionally well in theoretical metrics, verifications from the company's officials show that the algorithm can accurately detect both the present and potential churners with minimal exceptions. The striking points of this study can be stated as the usage of real-life, large-scale data and the high-precision, computationally efficient classification results which can be directly put into use by the company.

2. Materials and Methods

2.1. Related Work and Literature Review

Emphasizing the importance of customer churn prediction in e-commerce, a study (Xiahou, X., & Harada, Y. 2022) [1] introduces a novel approach that merges k-means customer segmentation with support vector machine (SVM) prediction, leveraging B2C e-commerce customers' shopping behavior characteristics. In another study (Kumar, A., 2023) [2], exploratory data analysis was applied to shopping mall data, leading to customer segmentation through k-means clustering. Authors of a study (Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H., 2015) [3] investigate three hybrid models for developing an accurate churn prediction model, employing two phases: clustering and prediction. In another work (Bose, I., & Chen, X., 2009) [4], two-stage hybrid models, integrating unsupervised clustering techniques and decision trees with boosting, are applied to two datasets, and their performance is evaluated based on top decile lift. A study (Khine, S. T., & Myo, W. W., 2023) [5] introduces a customer churn prediction model for the banking industry, utilizing K-means and Multi-Layer Perceptron (MLP) on the Kaggle Churn-Modelling dataset. In another study (Liu, Y., Fan, J., Zhang, J., Yin, X., & Song, Z., 2023) [6], the main contribution to the subject lies in the preprocessing, feature extraction, and processing of the multidimensional dataset provided by the telecom operator. Subsequently, the k-means algorithm is employed to cluster different consumer groups, allowing for the analysis of factors pertinent to each group and the provision of targeted suggestions. In another work (Çallı, L., & Kasim, S., 2022) [7], authors trained supervised ML algorithms such as Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and K-Nearest Neighbor to perform churn detection by using the customer data of a SAAS company. The main attraction of study was the feature selection section, in which they employed techniques such as the calculation of information gain, the Chi-Square and the Gini Index. In another work (Bagul, N., Berad, P., Khachane, C., & Surana, P., 2021) [8], Recency-Frequency-Monetary (RFM) model was developed and and a K-Means Clustering algorithm was utilized to perform churn detection using the customer data of a company operating in the retail sector. A Study (Olaniyi, Abdulsalam & Arowolo, Micheal & Jimada-Ojuolape, Bilkisu & Yakub, Saheed, 2020) [9], shows a case which utilized K-means Clustering and Support Vector Machines to perform churn detection by using the customer data of a bank.

2.2. **Problem Statement and General Approach**

This study mainly aims to create a practical ML-based model that classifies the customers of a food and beverage company to detect and predict current and potential churners. To train the algorithms, a portion of the data from the gift-based mobile promotional application of the company, which has over 14 million unique users, is used. In this app, by using the cameras of their mobile phones, users can scan the codes that are located within the packaging of the products that they have bought to earn various gifts such as free products, mobile data packages and funds that can be used in online mobile games. Various data ranging from transactions to app usage habits are available in the app's database. Details of the data at hand will be discussed further in the following section. The approach that is used can be separated into four main parts. First, the unlabeled customer data is manually analyzed to be able to decide on which features to be used in training. The decided portions of the data, which will be referred to as the "Feature Vector", is then preprocessed before moving into training. Furthermore, manual labeling is carried out according to the known churn characteristics to be able to verify and assess the results of the upcoming unsupervised clustering algorithms. Secondly, the customers are divided into different numbers of clusters with different meanings and attributes by using unsupervised clustering algorithms such as K-means Clustering and Hierarchical Clustering. The results of these clustering algorithms are then assessed to decide on the best performing algorithm. The Elbow Method is used to decide on the optimal number of clusters. In the third section, various supervised algorithms such as Support Vector Machines (SVMs) with different kernels, Random Forest, Logistic Regression, and XGBoost are trained using both the manually labeled observations and the previous section's clustering results. The results from these supervised methods are compared and used to evaluate the unsupervised algorithm's performance. Also, a decision tree that produces the same results as the unsupervised clustering is fitted to the data in order to make feature impact analysis via SHAPley values possible. The SHAP analysis is then utilized to understand the meanings of different customer clusters. Finally, a score system based on silhouette method is developed for quantifying and ranking the churn potentials of customers. To sum up, this work aims to combine supervised and unsupervised ML methods to automatically classify customers and to detect current and potential churners with high accuracy. The algorithm can then be used by the company to take customer-specific actions which will enable them to prevent future customer churn along with the benefits of having the reasoning behind current churners. A flowchart that describes the stages of the implementation is provided below in Figure 1.

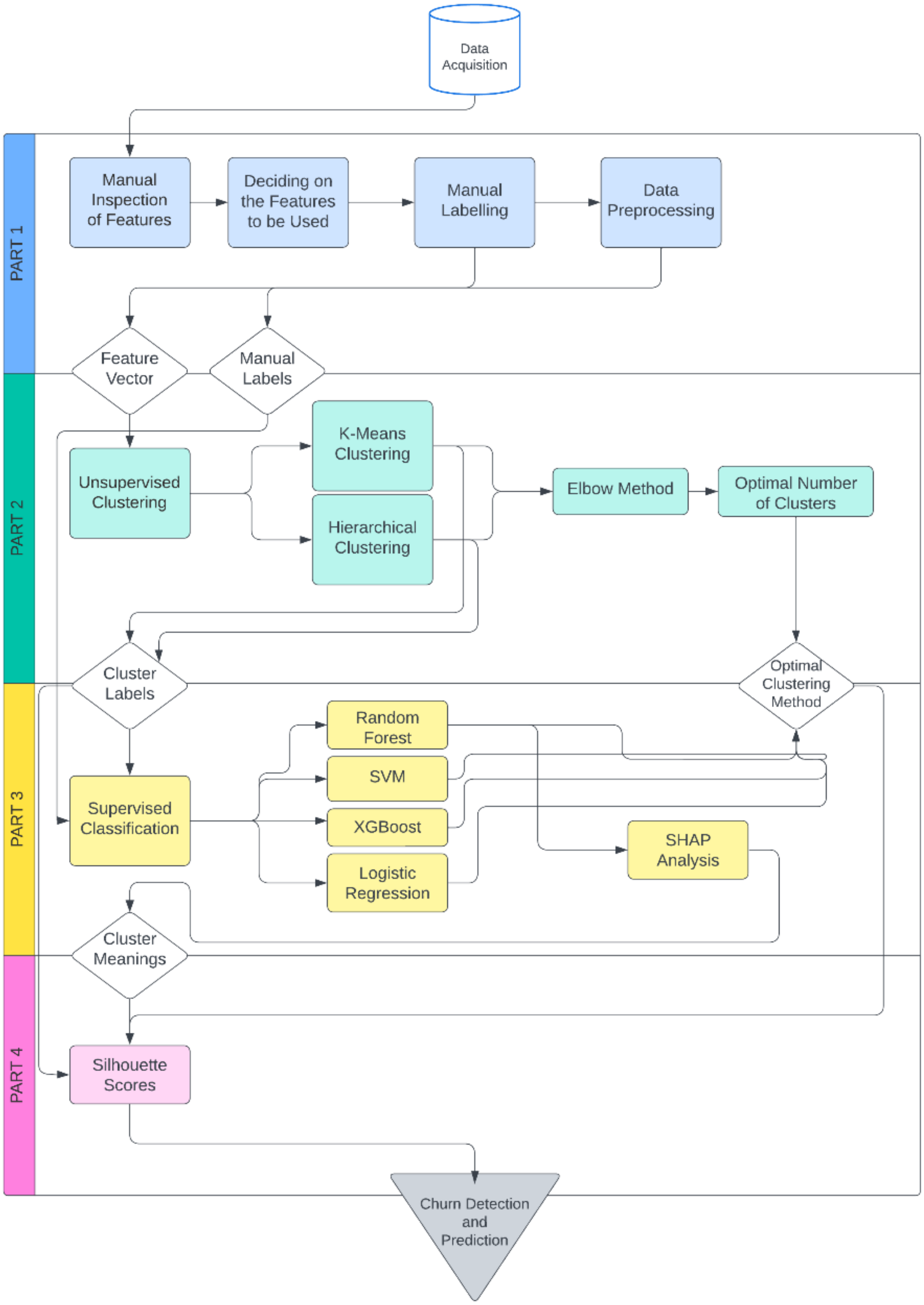

Figure 1: The Flowchart for the General Approach to the Problem.

2.3. Data Acquisition and the Nature of the Data at Hand

Table 1: The Features that were Used and Their Descriptions & Examples

Feature	Description	Examples
F1	Days Since App Removal	12, 95, 0
F2	Days Since Last App Usage	86, 194, 256
F3	Days Since Account Registration	92, 266, 387
F4	Total Codes Scanned	0, 125, 47
F5	Codes Scanned / Day	0, 1.74, 0.36

In the mobile application, users can log in with their credentials to see the ongoing campaigns featured by the company, can view the gifts that they have acquired, and can scan the codes inside the packaging of the products that they have bought using their mobile phone's camera to acquire new gifts. During their time and interactions within the app, both the metadata such as user ID, contact info and the usage data is generated and collected within the application's MongoDB based database under different subsections. During this study, the usage data and interaction habits that are in the various subsections of the database will be used to train our algorithms. No extra data collection will be carried out and no metadata will be used to avoid any privacy problems. Not all the data that is collected within the database is relevant to the goals of the study. Therefore, after a manual and thorough feature engineering process, the features in Table 1 were selected. Here, F1, F2, and F3 have a total of 11.8 million observations in the database, which is the number of total unique users. F4 and F5, on the other hand, have 142.7 million observations, since a customer usually scans way more than just one code. Therefore, all the codes that each unique user has scanned are gathered to form F4 and therefore F5. There are, however, many users that have not scanned any code at all. The strikingly low constituent count and its mean distance from other clusters are indicators that Cluster 3 could be the "outlier" cluster. Besides the features provided in Table 1, there were other potentially useful features such as chatbot messages and app reviews. However, these features were excluded from the feature vector since they were extremely sparse. For the preprocessing, the conversion of datetime data into integers by using the present day's date, generation of F5 from other features, and finally the normalization of the feature vector is done, before moving on.

2.4. Methodology

Since the company did not use any previous churn detection, there was no labeled data that could be used as the ground truth while training or testing a ML model. Therefore, two approaches were taken to overcome this issue. The first method was generating manual labels for potential and current churners by setting thresholds according to the expected user segments. The second stage was to apply unsupervised clustering to the data to gain an insight about the potential clusters that would form. After inspecting the observations manually and before applying unsupervised clustering, the expected outcome was a simple, 2-D scatter plot consisting of the first two PCA components and the labels generated by the unsupervised algorithm. In this assumption, the PCA axes were expected to represent the following: Activeness vs. Frustration. Activeness is positively affected when a user has scanned more codes and has been recently active. Frustration is positively affected when a user has deleted the app long ago and/or has not been active for a long time. From here, four main segments of users were expected: Apathetic users, Loyal users, Current Churners, and the Potential Churners. The expected clusters after the feature inspection and before applying clustering are provided below in Figure 2. Then, two unsupervised clustering algorithms were implemented: The K-means Clustering algorithm and the Agglomerative Hierarchical Clustering Algorithm. To decide on the number of optimal clusters, the elbow method was used for both algorithms. The details of these implementations will be discussed further in the unsupervised approach section. Also, manual labels for the expected clusters were generated by applying carefully adjusted thresholds or intervals for each feature of each observation and for each of the expected clusters. Then, different supervised classification methods were implemented. These were SVM with polynomial and RBF kernels, Random Forest, Logistics Regression, and XGBoost. To train these supervised algorithms, a subset of the labels that were obtained from the unsupervised algorithm were used. To test them, manually labeled observations were used, since they represented the clearest examples of each of the user segments that we expected, especially for the potential churners and current churners. From these results, the precision, recall and the F1 scores for each algorithm were obtained. By doing so, the verification of the results of the unsupervised algorithm was made possible despite the lack of any ground-truth data to test upon. Next, a random forest decision tree was fitted upon the previously verified clustering labels. This tree was then used for feature impact analysis via SHAP analysis which helped us provide meanings for the clusters. Finally, with the help of the fitted tree, the Silhouette Scores for the potential churners cluster were generated, which provided the numerical values that represent churn probability.

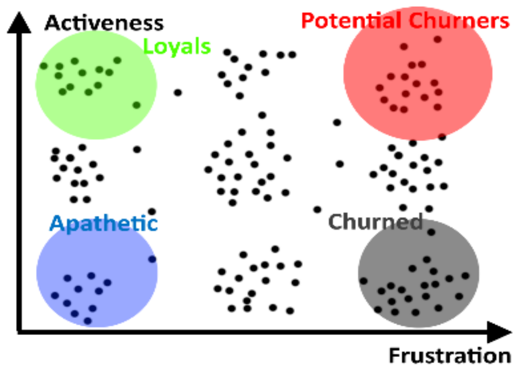

Figure 2: The Expected Clusters and Their Meanings.

2.5. The Detailed Approach

For all of the algorithms, the same dataset was used. It has approximately 156K observations, each belonging to a unique user, and each with five features that were described in Table 1. Before moving on to ML algorithms, approximately 98K of these users were taken as a subset and were labeled manually. The labels within this subset were selected by using thresholds and ranges for each feature that were defined after exhaustive inspection of the whole dataset. While deciding on these ranges and thresholds, RFM based reasoning was made. App usage and removal data was used as a Recency indicator, Codes scanned per day was used as a Frequency indicator, and the total number of codes scanned was taken as a monetary indicator. This process gave 20K customers that can be used as the ground truth for a potential churner. Moreover, 8K observations for apathetic users, 0.45K observations for loyal users, and 70K observations for long term churned customers. These manually labeled observations will be used to test the accuracy of the unsupervised approach, by using them as test samples on the supervised algorithms, which will be trained by another subset of the labels generated by the unsupervised algorithm itself.

2.5.1. Unsupervised Approach

Since no labeled observations to be used as ground truth data during training and testing were available, starting with unsupervised methods was the logical option. For the algorithms, K-Means Clustering and Agglomerative Hierarchical Clustering were implemented. Among these two, K-Means clustering produced more meaningful results. Moreover, the large size of our dataset rendered the Agglomerative Hierarchical Clustering method considerably more computationally costly when compared to the K-

Means Clustering algorithm. Therefore, the K-means Clustering algorithm was the algorithm of choice moving forward. To use k-means clustering, firstly, the desired number of clusters has to be determined, which is denoted by "k". To decide the integer k, a method called the "Elbow Method" is used. This method runs the k-means algorithm with different k values and plots "Within Cluster Squared Error" for each k value. The resulting elbow method graph is provided below in Figure 3. By looking at the WSSE for each cluster, it can be seen that the elbows occur at k=2 and k=4. However, setting k=2 would give only a "yes" or "no" type of answer whether a customer has churn potential or not. Moreover, WSSE is still considerably high at k=2. Therefore, it was decided to set k=4 to be able to obtain more detailed information and a lower theoretical error. After setting k=4 and obtaining the results from the K-means Clustering algorithm, a set of plots were generated to make the results easier to understand. The first of these plots were obtained after applying Principal Component Analysis to the dataset. After applying PCA, the first two principal components were utilized to produce a 2-D representation of our raw data. Then, the cluster labels obtained from the K-Means Clustering algorithm were scattered on top of this plot. The resulting graph is provided below in Figure 4.

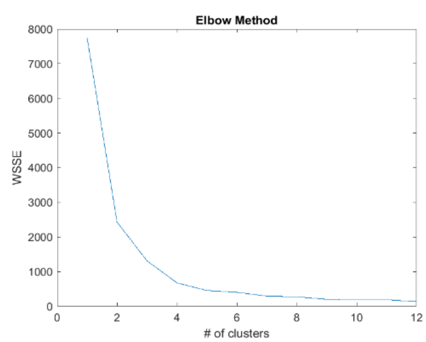

Figure 3: The Elbow Method for the K-means Algorithm.

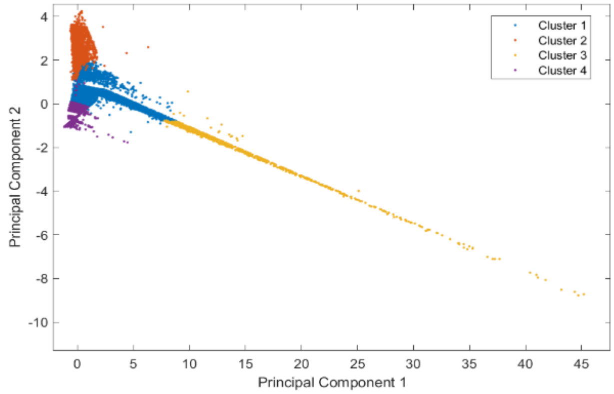

Figure 4: The Scatter Plot of the Clusters over the First Two Principal Components.

At first sight, the scatter plot may not seem to agree with the initial expectation that was based on the Activeness vs. Frustration perspective. However, as will become evident in the upcoming sections, the outliers and the ranges are the main reason behind that, and not the classification performance itself. The number of constituents for each cluster is as follows: 39.495K for Cluster 1, 14.924K for Cluster 2, 0.750K for Cluster 3, and 100.933K for Cluster 4. The strikingly low constituent count and its mean distance from other clusters are indicators that Cluster 3 could be the “outlier” cluster.

Another way to display the users is to plot them according to their registration date vs. the total number of codes they scanned. The resulting plot is provided below in Figure 5. A different way to visualize the results is to scatter the clusters across the #codes scanned and days since the last app usage axes. From a RFM perspective, this plot can also be thought of as a Recency (Last App Usage) vs Monetary (Total Codes Scanned) plot. Or from our initial perspective, an Activeness (Total Codes Scanned) vs. Frustration (Last App Usage) plot. It is provided below in Figure 6:

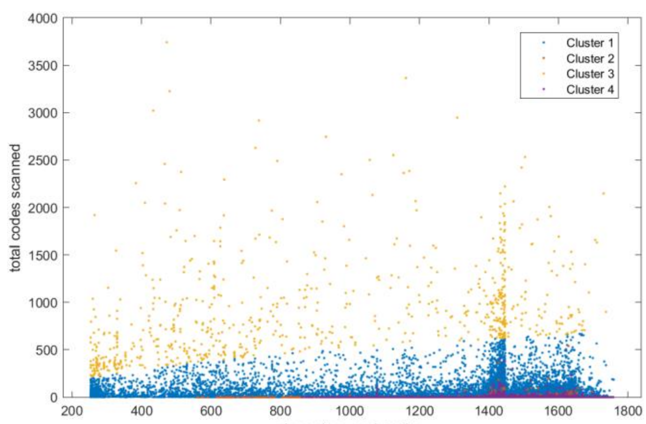

Figure 5: Days Since Registration vs. Total Codes Scanned.

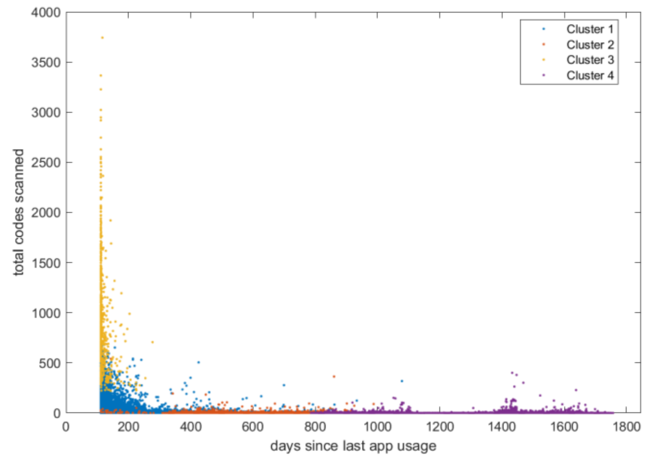

Figure 6: Days Since Last App Usage vs. Total Codes Scanned.

As evident, Cluster 3 users have scanned a high number of codes. There are also many users which have scanned an abnormally high number of codes. After consulting with the company, it was learned that some users abuse the gift system by scanning mass amounts of codes before the products are sold to actual customers. These users are called "fraud committing users". Hence, this cluster can be divided into two subsections and be used for fraud detection as well as the initial prediction for containing the most active users. Regardless, these findings support the initial suspicions that Cluster 3 is the outlier cluster. By looking at the same plot, it can also be inferred that Cluster 2 users are the "apathetic" users: They have been registered for over 1.5 years, but they did not scan a meaningful number of codes. Furthermore, it can be seen that the account tenure has no correlation to the activeness of a customer. Although Cluster 2 users are mostly apathetic, some still use the app which, over time, may become "less valuable" churners, since they did not provide much value to the company even while they were active. Moreover, it can also be inferred that Cluster 4 users are users that had low to no activity both in the present day and in the past. Cluster 1 users are the currently active or "recently gone" & mid to high level active users. These carry the most imminent and churn potential. Furthermore, it should be noted that the database that was provided lagged behind the present day by approximately 100 days. Therefore, in both Figure 5 and 6, the most recent day is capped at a minimum of 134 days, which enables the operation of shifting each day by this amount to simulate an up-to-date database.

2.5.2. **Supervised Approach**

The lack of ground truth data was a big problem for the project. Utilizing K-means Clustering provided valuable insight to the data, but no means to verify the results that we obtained during the unsupervised approach were available. Therefore, it was decided to make use of the manually labeled samples. First, by using a portion of the labels resulting from the unsupervised algorithm, various supervised ML models such as Support Vector Machines, Random Forest Classifier, Logistic Regression and XGBoost were trained. Then, to test these models, half of the manually labeled samples were used, since they were considered to be more robust due to the handpicked boundary conditions that were carefully applied while deciding on their labels. For each of these algorithms, the Accuracies and the F1 Scores were obtained, which are provided below in Table 4. As can be seen, all of the models have performed exceptionally well. This ensures that the unsupervised clustering results that were obtained via K-means clustering can be trusted to be used in a churn detection scheme.

Table 2: The Accuracies and the F1 scores for Various Supervised Algorithms.

Algorithm	Accuracy (%)	F1 Score (%)
SVM (Polynomial Kernel)	96.45	96.67
SVM (RBF Kernel)	96.35	96.57
SVM (Linear Kernel)	96.31	96.52
Random Forest	96.28	96.52
XGBoost	96.23	88.92
Logistic Regression	96.24	96.45

2.5.3. Feature Impact Analysis via SHAP Values to Obtain Cluster Meanings

Now that the resulting unsupervised labels could be numerically validated, next was to fit a decision tree that would completely encapsulate the unsupervised algorithm's results. This was a must, since the SHAP value analysis requires a decision scheme to be able to carry out the feature impact analysis. As mentioned above, up to now, most of the cluster meanings were made upon inspection. Although they were confirmed by the supervised side algorithms, to be able to numerically and rigorously evaluate and discuss the unsupervised clustering results, employing SHAP Value Analysis was still the best option. For Cluster 1 Outcomes, more days since last app usage and less days since app removal provided a larger & positive impact. The combined meaning is that such a user has not deleted the app but has not been using it recently. However, they have been active in the recent past. Thus, we say that they have **Churn Potential.** For Cluster 2 Outcomes, more days since app removal, less days since last app usage, and less codes scanned provided larger & positive impact. The combined meaning is that such a user is older, less active, and has "churned" in the recent past. **They are lost recently, but not necessarily valuable due to their near to no activity.** For Cluster 3 Outcomes, less days since last app usage, more total codes scanned, more codes scanned / day provided larger & positive impact. The combined meaning is that this cluster contains **the most active users**. But there is a large number of users in this cluster that are **potential fraud committers due to the abnormal codes scanned / day values.** Suspicion of Cluster 3 being the outlier cluster is confirmed by the low SHAP value magnitudes, as well as the knowledge that the active users form only a small portion of total users. For Cluster 4 Outcomes, low codes scanned / day, more days since app removal, more days since registration provided a larger & positive impact. The combined meaning is that these are long-gone and apathetic users **(Less valuable and possibly unrecoverable since they have churned a long time ago).**

2.5.4. Churn Scores and the Churn Detection Criterion

To calculate the churn scores, or any cluster's score for that matter, a commonly used and widely accepted method is the "Silhouette Method". Assume that the data have been clustered by any technique, into k clusters. $a(i)$ is defined as the mean distance between i and all other data points in the same cluster. $a(i)$ can be interpreted as a measure of how well i is assigned to its cluster. On the other hand, the mean dissimilarity, $b(i)$, is defined to be the *smallest* mean distance of i to all points in any other cluster. The "neighboring cluster", or the "next best fitting cluster" of i is the cluster with the smallest dissimilarity. Then the Silhouette Value $s(i)$ of a datapoint i is defined such that for $s(i)$ to be close to 1, it is required that $a(i) \ll b(i)$. As $a(i)$ is a measure of how dissimilar i is to its own cluster, a small value of $a(i)$ means it is well matched. Furthermore, a large $b(i)$ implies that i is badly matched to its neighboring cluster. Thus, a $s(i)$ value close to 1 means that the data is appropriately clustered. The heatmap for the potential churners' cluster (Cluster 1) is provided below in Figure 7. As can be seen the users that have been recently active but haven't scanned many codes have the highest churn potential. On the other hand, the users that have not used the app recently (since they churned a long time ago) and the ones that have recently been scanning many codes have a less churn potential. Plotting against days since app removal, which is provided below in Figure 8, it can also be seen that those who haven't removed the app but at the same time have been scanning less codes have the most churn potential. Also, those who have deleted the app not so long ago (since they might still be recovered) and have scanned a considerable number of codes are seen as potential churners as well.

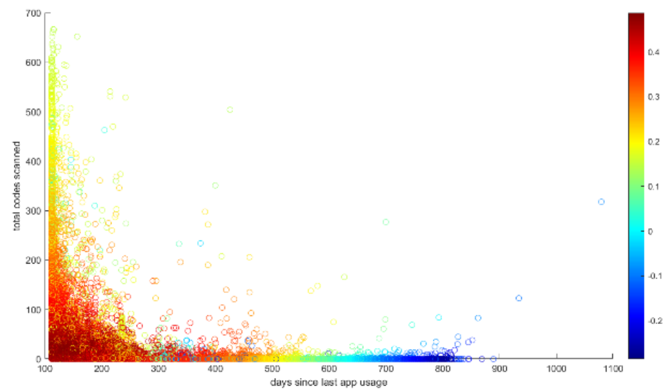

Figure 7: The heatmap w/ Total codes scanned vs. Last app usage.

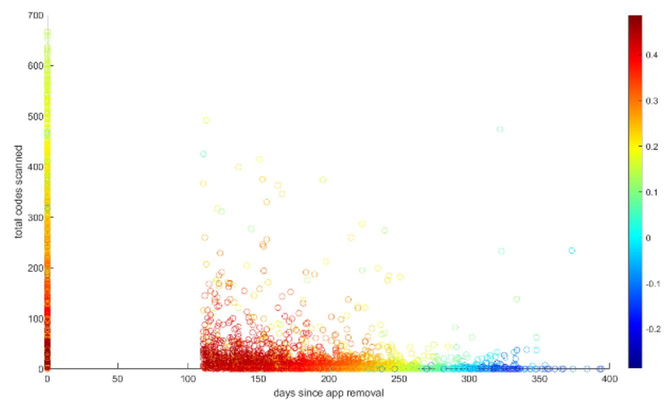

Figure 8: The heatmap w/ Total codes scanned vs. days since app removal

3. Results

Utilizing ML methods to develop an algorithm that detects and predicts customer churn, which was the primary goal of the study, was successfully achieved in many aspects. The first of these aspects was accuracy. The accuracy of the model was verified through the cross-validation process that was carried out between the manually labeled observations and the observations that were labeled by the algorithm. The second of these aspects was the usage of state-of-the-art techniques. After the literature review, the knowledge regarding the methods that were popular in previous studies in the churn detection topic was obtained. Then, it was made sure that not only the approach that we took included these methods, but also that it included new techniques so that it brought an innovative approach to the topic: Working with both unsupervised and supervised methods to overcome the lack of ground truth data problem was the biggest of the innovations.

4. Discussion and Conclusion

To begin with, the results were not only explained via inspection, but also via a numerical feature impact analysis process that made sure that the meanings of the resulting customer clusters had a rigorous meaning. Also, on top of the "yes" or "no" answer to the customer churn probability, a score system was developed that provides the company with valuable information regarding the level of urgency of a customer's churn possibility.

As future work, more features such as geolocational data, chatbot conversations, app reviews and complaints, age and gender distributions can be used while training the algorithm and/or interpreting the results. Moreover, a graphical user interface can be developed for ease of usage. Also, additional supervised and unsupervised classification methods can be added for an even more thorough comparison between them. Additionally, fraud detection can easily be implemented by making use of the outliers within Cluster 3. Finally, as the algorithm gets tested in the field, minor changes can be implemented according to the company's requirements.

5. Acknowledge

We would like to thank the people in the subject company for providing us with the interesting problem and the comprehensive, real-life data that made this study possible.

References

References should be cited in the text by name and year in parentheses.
Example:

[1] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458-475.

[2] Kumar, A. (2023). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. In *Advancing SMEs Toward E-Commerce Policies for Sustainability* (pp. 248-270). IGI Global.

[3] Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid data mining models for predicting customer churn. *International Journal of Communications, Network and System Sciences*, *8*(05), 91.

[4] Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce*, *19*(2), 133-151.

[5] Khine, S. T., & Myo, W. W. (2023, February). Mining Customer Churns for Banking Industry using K-means and Multi-layer Perceptron. In *2023 IEEE Conference on Computer Applications (ICCA)* (pp. 220-225). IEEE.

[6] Liu, Y., Fan, J., Zhang, J., Yin, X., & Song, Z. (2023). Research on telecom customer churn prediction based on ensemble learning. *Journal of Intelligent Information Systems*, *60*(3), 759-775.

[7] Çallı, L., & Kasim, S. (2022). Using machine learning algorithms to analyze customer churn in the software as a service (SAAS) industry. *Academic Platform Journal of Engineering and Smart Systems*, *10*(3), 115–123.

[8] Bagul, N., Berad, P., Khachane, C., & Surana, P. (2021). Retail Customer Churn Analysis using RFM Model and K-Means Clustering. *International Journal of Engineering Research and Technology*, *10*(3).

[9] Olaniyi, Abdulsalam & Arowolo, Micheal & Jimada-Ojuolape, Bilkisu & Yakub, Saheed. (2020). Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms. 48-54.