# **Predictive Analytics for Customer Churn Prediction in Retail**

Abstract—Customer churn is one of the most challenging problems for digital retailers. With significantly higher costs for acquiring new customers than retaining existing ones, knowledge about which customers are likely to churn becomes essential. This paper reports a case study where a data-driven approach to churn prediction is used for predicting churners and gaining insights about the problem domain. The real-world data set used contains approximately 200 000 customers, describing each customer using more than 50 features. In the pre-processing, exploration, modeling and analysis, attributes related to recency, frequency, and monetary concepts are identified and utilized. In addition, correlations and feature importance are used to discover and understand churn indicators. One important finding is that the churn rate highly depends on the number of previous purchases. In the segment consisting of customers with only one previous purchase, more than 75% will churn, i.e., not make another purchase in the coming year. For customers with at least four previous purchases, the corresponding churn rate is around 25%. Further analysis shows that churning customers in general, and as expected, make smaller purchases and visit the online store less often. In the experimentation, three modeling techniques are evaluated, and the results show that, in particular, Gradient Boosting models can predict churners with relatively high accuracy while obtaining a good balance between precision and recall. Index Terms—digital retailing, customer churn prediction, RFM analysis, correlations, feature importance, top probabilities.

1. INTRODUCTION

With the fast advancement of e-commerce, digital retailing has become an effective tool for enhancing the economy and serving customers worldwide [1]. However, when customers are dissatisfied with certain products or services, they stop doing business with the related companies and switch to other sellers or service providers for better offers and qualities. It is called customer defection or churn [2]. Accordingly, the more customers leave, the less growth that retailer makes. It significantly influences businesses as it lowers earnings and revenues [3]. At the same time, acquiring new customers is challenging and costly compared to customer retention [4], [5], especially in the very competitive retailing industry. Therefore, many retailers invest in customer retention as it is the biggest revenue driver [6]. On the other hand, marketers must effectively decide when and how to expose individual The current research is a part of the Industrial Graduate School in Digital Retailing (INSiDR) at the University of Boras, funded by the Swedish ˚ Knowledge Foundation, grants nr. 20160035, 20170215 customers to retention actions to avoid customer churning. However, this is not an easy task and the marketers, therefore, need support to conduct an in-depth analysis of the company’s data set on customer churns to identify critical markers and characteristics, which can be used for churn prediction in the marketers’ work with customer retention. This paper aims to provide a data-driven approach that retailers can apply to customer churn data to gain insights and more effectively predict which customers are at risk of being lost, with the purpose that fewer customers will churn through more timely and precise retention activities. It consists of exploratory data analysis, pre-processing, model building on different data segments, and evaluation. It also examines the probability estimates of predictions and provides reasoning on which segment of customers could be much more critical to understanding the factors leading to churns.

II. BACKGROUND In the digital retailing era, churners are those customers who will leave a particular shopping or service platform and switch to another. Since 2000, studies showed customer churn was mainly from telecommunication services [7]–[12], even though several other churners exist, such as in online gaming [13], multimedia streaming [14], banking [15], advertisements [16], and insurance agencies [17]. Besides, retailers have recognized that they cannot afford customer loss, and it was better to focus on being keepers rather than losers by following reactive approaches, where retailers wait for customers’ cancellation requests and then offer generous plans for retention, or by following proactive approaches where churners are predicted in advance, and plans are offered to customers [4]. However, in both cases, different strategies are made, including analyzing churn factors [10], giving better services [2], defining valuable customers and targeting the right audience [16], and finally, paying attention to customer engagement and dealing with complaints [18]. On the other hand, customer shopping behaviors may change based on the economy or other reasons, affecting customer retention and churning. Therefore, from a management and direct marketing perspective, another challenge is keeping track of customers without considering how frequently they shop online and how monetary values change over time [19]. Finally, several studies for predicting churns are found from a machine learning perspective, including surveys [2]

TABLE I CHURN DATA SET Feature Description Processing # D-Type General gender male, female, unknown one-hot 3 {0, 1} country customer’s country one-hot 4 {0, 1} payment method card, bank transfer, post, ... one-hot 3 {0, 1} shopping recency segment (1), (2), (3), and (above) of number of months since last order engineered 4 {0, 1} visits recency segment (1), (2), (3), and (above) of number of months since last visit engineered 4 {0, 1} shopping frequency segment (1), (2), (3), and (above) of number of purchases engineered 4 {0, 1} visits frequency segment (5), (10), (15), and (above) of number of online visits engineered 4 {0, 1} shopping monetary segment (1000), (2500), (3500), and (above) of spending in SEK engineered 4 {0, 1} Main RFM recency of purchases number of days since the last purchase from the online store 1 int. recency of visits number of days since the last identified visits 1 int. frequency of purchases total number of orders made over the past two years 1 int. frequency of visits total number of visits made over the past two years 1 int. monetary of purchases total purchases value made over the past two years 1 float Detailed RFM recency of purchases per category number of days since recent purchase per category 7 int. frequency visits per period number of online visits over 10, 30, 60, 90, , , 330, 365, 730 days 14 int. frequency purchases online number of purchases made through the online store 1 int. frequency purchases offline number of purchases made in physical store 1 int. frequency purchases per period number of purchases made in the 1st, 2nd, 3rd, and 4th past half-year 4 int. frequency purchases per payment number of purchases made based on the used payment method 6 int. frequency units ordered number of ordered units 1 int. frequency units sent number of ordered units. It should be ≥ # units ordered 1 int. frequency units returned number of returned units. It should be ≤ # units ordered 1 int. frequency units outlet number of outlet units. It should be ≤ # units returned 1 int. monetary discount total discount made in all purchases 1 float monetary per category purchase value made per category 7 float monetary per (category, period) indicator of shopping from 7 product categories, over 4 six-month periods one-hot 28 {0, 1} Class churn indicator a churned customer is the one that has not purchased for one year 1 {0, 1} algorithms [4], optimization [8], predictive modeling [18], and unsupervised learning [5]. Each used different data from ours and probably from each other, making results incomparable.

III. METHOD

A. Data The original data set has over 200,000 rows, with 83 attributes plus the target variable. Each row represents a unique customer’s past online store visits and shopping information mainly taken from the past two years. We organized it according to the recency, frequency, and monetary concepts (RFM) [19], i.e., the interval between the time of the latest purchase or online visit and the present, the number of purchases or online visits, and the monetary value, respectively, during a particular period. It consists of the following features: 1) General features: They are nominal attributes describing the customer’s general information, such as the gender, country, default payment mode, and customer shopping segment, i.e., whether the customer has made one, two, three, or four or more purchases over the past two years. 2) Main RFM features: Here, the number of days since the last purchase and since the last identified visit are the main R attributes. In the same way, the main F attributes are the number of customer purchases and online-store visits. The total value of all orders from the last two years is the main M attribute. 3) Detailed RFM features: These attributes give a more detailed description of the customer’s shopping history and behavior. For example, several R attributes represent the total number of days since the last purchase per category, i.e., Home textile, Furniture, Electronics, Shoes, Ladies, Children, and Men. Similarly, F attributes include the number of identified visits online in various periods, i.e., last ten days, last month, last 60 days, etc., the number of online and offline orders, the total orders made each six months from the past two years, and the number of orders made using each payment method. Other F attributes are the number of ordered, sent, returned, and outlet items. Finally, the M attributes include order discounts, order values per category, and binary indicators for customer purchases from each category every six months for the past two years.

B. Exploratory data analysis The data shape is 210708×84, representing a combination of numerical and categorical attributes. The number of churners is 106828, i.e., almost 50%, meaning that the data set is fairly balanced. Table II gives a statistical summary of the main RFM attributes, and Fig. 1, 2, 3, 4, and 5 show a graphical representation of Churn distribution over these attributes. As an example, we see from the table that 75% of customers have made at most five purchases and spent up to 4155 SEK. A median customer has made three purchases, visited the online store 15 times, and spent 1446 SEK. From Fig. 1, a key takeaway is that the proportion of churners is very different when looking at the number of purchases. Specifically, a large majority will churn for customers with just one previous purchase, i.e., not make another purchase in the next year. On the other hand, with four or more purchases, the picture is the complete opposite. Other patterns, i.e., churners have made few online visits, spent less money as shown in TABLE II A STATISTICAL SUMMARY OF THE MAIN RFM ATTRIBUTES F-purch. F-visits M-SEK R-purch. R-visits count 210708 210708 210708 210708 210708 mean 4.50 53.98 3532.84 143.81 260.77 std 6.10 116.53 6277.91 105.50 369.19 min 1 0 0.10 1 1 25% 1 2 508.30 50 23 50% 3 15 1446.00 123 97 75% 5 56 4155.00 230 262 max 724 4101 420287.90 365 1076 median 3 15 1446 123 97 TABLE III THE TOP X% FEATURE-TARGET CORRELATIONS. Top % n. Ftrs Avg. |Corr.| Top % n. Ftrs Avg. |Corr.| 1% 1 0.323 20% 16 0.284 5% 4 0.323 30% 23 0.274 10% 8 0.301 40% 31 0.265 15% 12 0.292 50% 39 0.25 TABLE IV THE TOP 20% OF MOST CORRELATED FEATURES WITH THE TARGET Feature Corr. Feature Corr. Frq. purchases -0.32 Rec. purchase Ladies 0.28 Frq. visits -0.31 Mon. discount -0.28 Mon. Ladies 3rd6-M -0.31 Frq. units ordered -0.28 Frq. purchases online -0.30 Mon. purchases -0.27 Mon. Ladies 4th6-M -0.30 Frq. purchases 3rd6-M -0.26 Frq. purchases Bank -0.30 Mon. Ladies 1st6-M -0.26 Frq. purchases 1st6-M -0.29 Frq. purchases 4th6-M -0.26 Frq. units sent -0.28 Rec. purchases 0.26 Fig. 2 and 3, and the time since the last purchase and visit are relatively long as shown in Fig. 4 and 5. Table III shows the features with the highest absolute correlation with the target. For example, the top 5% featuretarget correlations imply four features with an average absolute correlation of 0.323. However, the table shows that most features are only weakly correlated with the target. In a more descriptive analysis, Table IV shows the top 16 features with the highest correlation with the target. For instance, the frequency (purchases, visits, purchases online, purchases over the past first, third, and fourth six-month periods, purchases paid by bank transfer, and ordered and sent units) are negatively correlated with the churn class, meaning that the larger the attribute, the less likely the customer will churn, and vice versa. Similarly, the recency attributes have positive target correlations, so the longer the period since the last purchase, the more likely a customer is to churn. One potentially interesting finding is that spending specifically in the Ladies category seems to be a clear indicator that a customer will not churn.

C. Pre-processing and feature engineering In order to develop effective predictive models and provide reliable results, pre-processing the original data set is essential. Steps taken include data profiling, cleaning (e.g., Fig. 1. Churn class distribution over purchase frequency Fig. 2. Churn class distribution over visits frequency Fig. 3. Churn class distribution over purchase monetary Fig. 4. Churn class distribution over purchase recency Fig. 5. Churn class distribution over online store visit recency handling missing values and removing unnecessary columns), and encoding the attributes. At the same time, new features were engineered by grouping customers based on their main RFM attributes’ medians shown in Table II. For example, customer online visits recency is segmented into four groups, i.e., five visits, ten, fifteen, and above. Fig. 6 shows the segments of engineered features, and Table I describes the data set after pre-processing.

D. Modeling techniques Three machine learning algorithms are used to predict churns and compare the performance. One is a strong ensemble technique, Gradient Boosting (GB), an effective classification technique that is more accurate than tree-based classifiers in predicting churns [9], [16]. The number of estimators used is 100, with a maximum depth of individual regression estimators of 3. The second technique used is Decision Trees (DT). DTs are interpretable, making it possible to understand the logic behind the predictions and even analyze the model to gain insights into the underlying domain, see, e.g., [11], [12], [15]. In this study, the trees are induced using the Gini index and no specified depth setting, so the algorithm will often continue to split the data until a leaf is pure. The last technique is Logistic Regression (LR), which is commonly used by businesses for churn prediction and similar tasks, in particular when the data is imbalanced [2], [17]. The settings used are the Saga solver and the Elasticnet regulation. E. Modeling validation and evaluation metrics The current study uses a stratified 10-fold Cross-Validation (CV) in model building and performance analysis. Before starting the cross-validation procedure, 10% of the data set was put aside as an extra hold-out for final testing. Metrics used for the performance evaluation are; accuracy (ACC), precision (PRE), recall (REC), and the area under the ROCcurve (AUC). In many scenarios, maximizing either accuracy or precision would be the most important goal. For churn prediction, however, the cost of losing a customer that could have been retained is generally much higher than the cost of an unnecessary retention activity. Consequently, a false negative is a much more severe mistake than a false positive. So, for churn prediction, recall is often more important than precision. Still, a too-low precision must, of course, be avoided unless a majority of all customers will be targeted by retention activities. Balancing these criteria, we, in the current study, decided to look for customers that are very likely to churn while striving for an acceptable recall. For the results see Table X and the accompanying analysis and discussion in Section IV-D.) IV. RESULTS AND DISCUSSION A. Feature importance Since the GB model can provide feature importance as a tool for understanding and gaining insights, it is analyzed the importance of features using training data. Fig. 7 shows that the Frequency of visits feature is of the highest importance, followed by the Frequency of visits over 730 days and the Frequency of purchases. It further explains Fig. 1 and Fig. 2, where the number of visits and purchases increase, churns are less likely, and vice versa. Therefore, such a GB model’s feature importance explains churn indicators from the features used in the data set. Fig. 6. Overview of engineered features over 20 different segments of RFM. TABLE V THE ENTIRE DATA’S MOST IMPORTANT 16 FEATURES. ∗ Feature ∗ Feature 1 Frq. visits 9 Frq. purchases Bank 2 Frq. visits over 730 days 10 Frq. purchases Ladies 3 Frq. purchases 11 Frq. units sent 4 Frq. visits seg. (5 visits) 12 Payment method postal 5 Frq. visits seg. (above 15) 13 Frq. purchases 1st 6-M 6 Frq. purchases seg. (above 3) 14 Frq. units ordered 7 Frq. purchases offline 15 Rec. visits 8 Rec. purchases 16 Frq. purchases online Fig. 7. The top 16 feature importance ranked in tables V, VI, and VII. Table V shows the important features found by the GB model. Most of these features are overlapped with the targetcorrelated features found in Table IV. For example, Frequency of visits and the related segments, Frequency of purchases and the related segments, Frequency of purchases (made online, made in the 1 st past six-month period, paid by the bank transfer, and of the Ladies category), and finally, Frequency of units, ordered, and sent) features. B. Results from Cross-Validation Table IX shows the performance results. For the GB model trained on all instances (i.e., 1-4+ purchases), the accuracy and the precision are around 80%, and the recall is almost 78%. This, consequently, should be considered as the baseline. Nevertheless, a retailing company could potentially perform better by only acting on a subset of all predicted churners. These kind of strategies are further investigated in IV-D.

# Bibliography

] B. L. Dey, D. Y. (December 2019). “Digital consumer culture and. *” International Journal of Information Management,*, 102057, 2020.