Customer Segmentation, Churn Prediction and its implications using ML

Iyer Vinod, Joshi Mrinmayee, Kothari Adithya, Sharma Siddhant, Singh Sudipta, Taparia Raghav, Yang Wang
Purdue University, Krannert School of Management
iyer108@purdue.edu; kothari8@purdue.edu; joshi155@purdue.edu; sharm486@purdue.edu; singh912@purdue.edu; rtaparia@purdue.edu, yangwang@purdue.edu

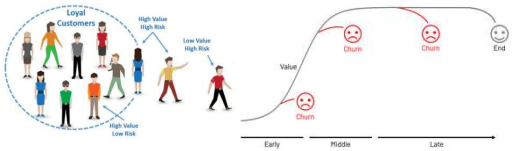
ABSTRACT

Customers enrolled in loyalty programs are the most important customers to national scale retailers, who are expected to have high repeat purchase rates. Consequently, losing these customers would result in a significant revenue impact. Our study is focused on predicting who are the customers that are likely to churn and when they are expected to churn within their customer cycle. For successful results, this study required identifying current high-value customers, establishing behavioral patterns and KPIs, and training our ML model to predict customers that are expected to churn within a set duration in time.

INTRODUCTION

Why is Churn important? Data from a national chain retailer shows that **70% of sales generated from engaged customers (30% of total customers)**

A business can generate more revenue by increasing customer loyalty, upselling existing customers, or gaining more new customers. Almost all revenue-generation methods involve an initial cost, and research shows that gaining new customers is more expensive than keeping existing ones. Thus, churn prediction is a useful tool for determining the actual return on investment for a certain product or service. Even though all businesses are at risk of losing customers, conducting an attrition analysis significantly reduces the risk. As soon as the churn-probable customers are identified, the next step is to strategize the marketing actions needed to keep them. The focus here is the retail industry.



How do we define churn? Let's look at some scenarios -
1. Understanding the type of churn: It's difficult to understand if a customer has churned voluntarily, silently or involuntarily.
2. What exactly constitutes a churn event? Can a customer who purchases very sporadically be labelled as a churned customer? Or can a customer who purchase consecutively in the last 3 months but did not purchase anything this month be labelled as a churned customer? The concept of 'churn' is very fuzzy.
3. Low Churn Rate: Considering a business is in a good shape, customer churn becomes a relatively rare event, so what constitutes churn in this case?

RESEARCH OBJECTIVES

- What are the most important factors affecting churn?
- Predict likelihood of a customer to churn
- Identify red flags and initiate interventions before customer churns

LITERATURE REVIEW

Various supervised algorithms like Logistic Regression. Naïve Bayes', Decision Tree, Random Forest, and Extreme gradient boosting have been used to predict churn.. Also, RFM analysis, K-means clustering have been used previously to segment users into churned and not churned customers.

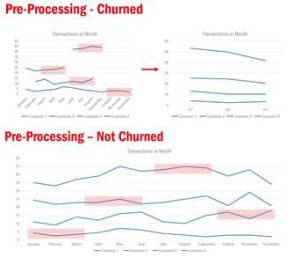
	NB	LR	DT	RF	GB	BPN
Annapurna et al., 2017					✓	
Ali Tamaddoni Jahromi et al., 2014		✓	✓		✓	
Abou el Kassem et al., 2020		✓				✓
T.Vafeiadas et al., 2020		✓	✓			✓
Our Study, 2021	✓	✓	✓	✓	✓	

Table 1. Literature Summary

METHODOLOGY

Fig 2. Methodology

- Added variables defining **last leg of customer's journey**
- Incorporated importance of **Grocery shopping** in predictor variable
- Added variables to define **customer experience** at supercenters



Churned Customers
Customers who have transacted for 5 months within a 6-month timeframe and did transact for 3 consecutive months are classified as churned customers.

Not-Churned Customers
Customers who have transacted at least once every month from their first transaction during the entire duration of data are classified as Not-Churned customers.

STATISTICAL RESULTS

A. Comparing different algorithms for accuracy

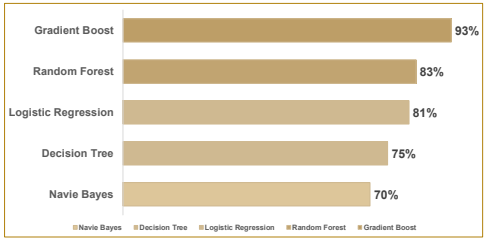


Fig 3. Model Results

Out of the 5 classification models, Gradient boost performed the best with 93% accuracy. While the remaining algorithms were significantly lower, Random forest and Logistic Regression were the next best algorithms with 83% and 81% accuracy, respectively.

B. Variable Importance Score

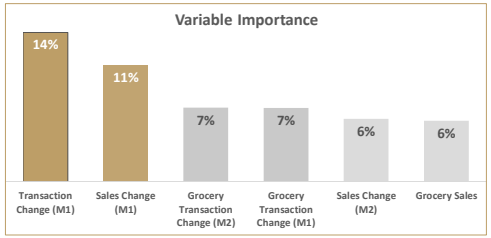


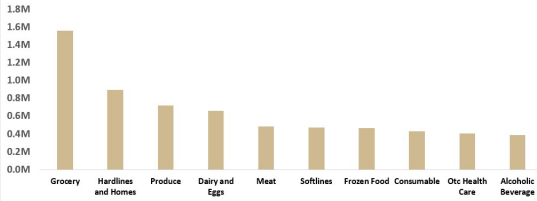
Fig 4. Importance Score

The top 6 variables in descending order of importance to predict churn are shown in the figure : The month before the churn (M1 was the most important amongst the 2 months). Grocery sales and pharmacy visits were the most important while comparing different categories.

EXPECTED IMPACT

- The number of transactions in the month M1 (month before the churn) is the crucial variable in determining churn. Therefore, it can be utilized to formulate strategies of intervention to increase engagement
- After the users at risk of churning have been marked, the future step is to strategize the marketing actions required to improve the possibilities of the churn-probable users staying engaged. Moreover, marketing costs can be optimized by providing promotions and coupons to only valuable customers
- Average monthly revenue for churned customers is highest for Grocery sales ($1.6M) followed by Hardlines & Home ($0.9M). Overall, $6.9M can be saved through churn prediction in top ten product categories

C. Average Monthly Revenue Lost by category



CONCLUSIONS

- Customers who have been shopping at Groceries and Staples at the stores indicate loyalty to the company
- Mitigating the challenge of churning customers can increase annual revenues by 60%
- The shopping patterns in the categories in the most recent 2 months would help us identify if the customer is on their way to churn or disengage
- Once these shopping patterns are identified for a customer, intervention strategies can be placed to potentially mitigate the loss of X dollars in monthly sales for that customer

ACKNOWLEDGEMENTS

We would like to thank Professor Yang Wang and our industry partner for this opportunity, their guidance, and support on this project.