Improving Purchase Prediction with Seasonality Features

Capstone Project 2018 – Public Project Summary

Aurel Pasztor

Introduction – Project Goal

The industry partner of this project is one of the largest independent marketing platform companies in the world. The company's software enables one-to-one marketing interactions between online retailers and their customers across several digital channels. The performance of these interactions is measured by the conversion to purchase in the marketer's web shop. Based on past customer behaviour, the platform's different features predict probabilities of future purchases of the retailer's individual customers.

The goal of this project is to see if seasonality-based features can improve the company's existing model that predicts purchase probability for a contact in the next 14 days.

The Data

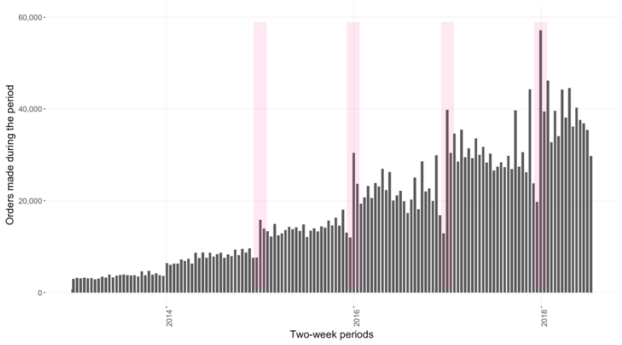
I have acquired purchase records of an e-commerce web store. The data includes all contact (i.e. customer) purchases over the last 5.5 years that is close to 8 million item purchases by 910 thousand customers.

The data includes:
- 7,936,808 purchase records
- start date: 2013-01-01
- end date: 2018-07-26

One purchase record includes:
- contact_id (n= 909,542)
- order_id (n= 2,623,387)
- product_id (n= 1074)
- purchase_date
- quantity
- sales_amount

Data Exploration

I set two-week periods as the base time unit for the analysis. Then I look at purchase frequencies by period over time and notice **seasonal spikes in the first periods of every year**.



Order frequencies by period

The sum of purchases also increases in these periods but there is no indication that a particular product is driving these sudden increases in the beginning of January. When examining contact level data, I also find that most contacts only made one order during the 5.5 years. The median order frequency by contact is 1.

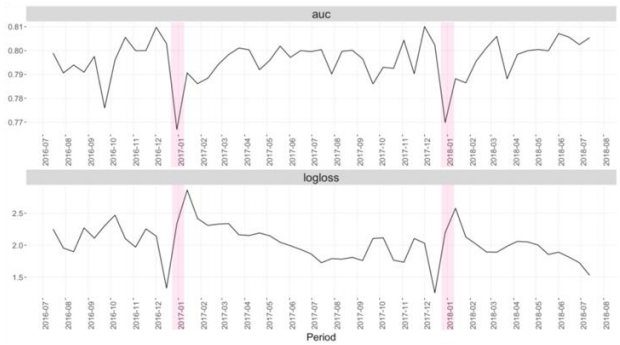
The Model

The company is using a logistic regression model for the binary classification problem of predicting whether a customer is going to make a purchase within the next 14 days or not.

The base model currently used in production by the company includes three features:

- frequency: how many orders have the customer made until now
- recency: how many days have passed since the last order
- monetary: how much has the customer spent until now

I have predicted 53 subsequent two-week periods with test sets built on the previous 52 two-week periods' purchase history then plotted the AUC metrics and computed their averages. I also computed logloss I optimize for AUC as it maximizes the model's ability to discriminate between classes whilst the logloss is a calibration statistic.



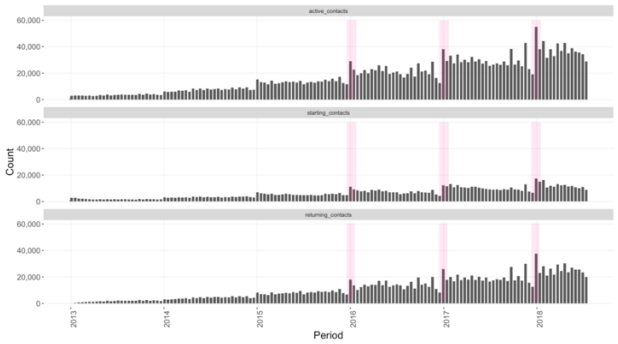
Model AUC and logloss over time

The plots show that **the base model provides a fairly good AUC, that is 0.796 on average but January spike periods show comparatively large drops**. This means the model is relatively poor in predicting purchase behaviour in January spike periods. To improve the average AUC and "smoothen out" the drops, I analyse what is happening in spike periods.

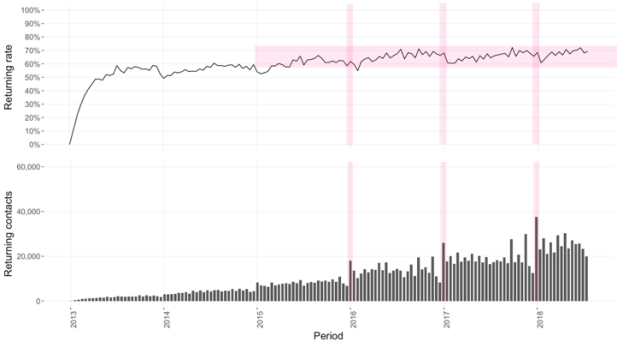
Spike periods

I examine the constitution of active contacts over time. Active contacts in any period are either returning (who have bought before) or starting (who just made their first order) contacts.

I find that there is no difference between increase patterns of starting and returning contacts in spike periods. They both show significant increases but because returning contacts make up on average 65% of active contacts in each period from 2015 (including spike periods) it can be said that **January spikes are mainly driven by sudden purchases of returning contacts**.



Active, starting and returning contacts by period



Rate of returning contacts

I examine the purchase patterns of returning contacts of spike periods and find that most of them made at least 3 orders before, but not regular purchases in other spike periods. They tend to have a "recency pattern" (i.e. are active in a shorter period) and fade away by time. This hints that recency may be the best indicator of a next purchase.

Adding features and experimenting with training parameters

Based on the results I introduce several new features to the model in several phases:

Feature name	Description	Significant
active	a dummy for contacts active in the train period (made purchases in the last 26 months)	yes
previous year active	a dummy for contacts active one year ago in the same period	no
Absolute average recency	average time a returning contact takes to make a new purchase	yes
Absolute average recency active	a dummy for returning contacts	yes
Active_90d	contacts that made an order in the last 90 days	yes
Active_180d	contacts that made an order in the last 180 days	yes
Active_270d	contacts that made an order in the last 270 days	no
Active_365d	contacts that made an order in the last 365 days	yes

The feature that aimed to capture the yearly seasonality pattern (*previous year active*) was not significant. I have also experimented with shifting the train label period to the previous year's spike period and use the preceding data for training to capture the previous year's spike effect, but it did not prove to be useful: average AUC has not increased, and the drops remained. Eventually, I have found that recency based dummy variables on short periods (0-6 months) were significant just as dummies on longer terms (12 and 26 months). **Adding these recency-based features has improved average AUC to 0.840 but could not eliminate seasonal drops.**

Conclusion

Adding additional features and adjusting training parameters had a slight improvement on average AUC but could not improve the model's ability to predict seasonal spikes in purchase frequency. I found that the seasonal spikes are driven by returning customers but there is no good indicator for which returning customer will buy in any given spike period.
I found that the best feature for future purchase behaviour is *recency* and dummy variables capturing recency can incrementally improve model performance.