Analyzing and Predicting Purchasing Intention with E-Commerce Data

Yuqing Zhang

Background

User behavior modeling in e-commerce

One of the essential goals for e-commerce companies is to increase *purchase conversion rates*, i.e. the percentage of users who complete the purchase at online stores. To achieve this goal, much efforts have been devoted to analyzing and modeling the behaviors of webpage users, especially with statistical and machine learning methods.

H2O platform

H2O is an open-source platform developed by H2O.ai, which implements scalable machine learning algorithms, and can efficiently handle large volumes of data. The backend of H2O is written in Java, which makes it easy to deploy models to production. H2O has APIs in both R and Python, and is supported on popular cloud computing platforms, including Amazon AWS, Microsoft Azure, and Google Cloud.

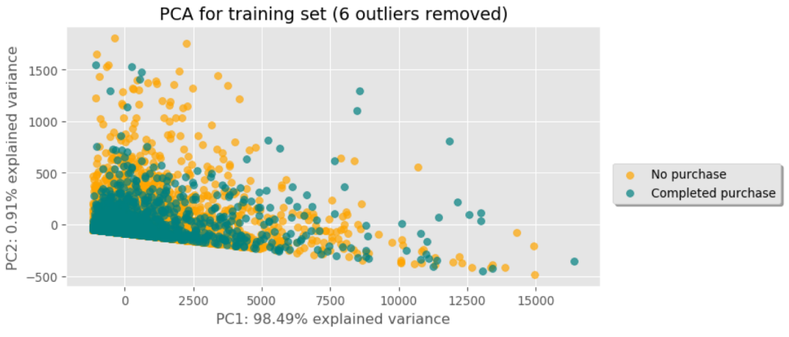
Auto ML

Automated machine learning has been a hot topic in artificial intelligence in recent years. For the same type of problems, it aims to automate the training process, especially for model selection, feature selection, and parameter tuning. And as a result, we are able to try a large number of models within a short timeframe, which makes it easier to find the optimal model for the data. The AutoML function in H2O trains a collection of pre-specified algorithms all at once.

Data

Samples

- 12330 users, 1908 (15.5%) completed purchase
- Held out 30% as "new" data, explore on the remaining 70% samples (training set)
- 6 outliers detected and removed from training



Features

7 Categorical Features

User profiles
- Visitor Type
- Region

Session information
- Browser
- Operating Systems
- Traffic Type
- Weekend
- Month

10 Numerical Features

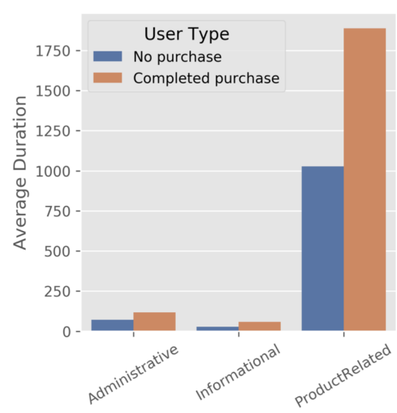
Number and time spent on 3 types of webpage
- Administrative
- Informational
- Product related

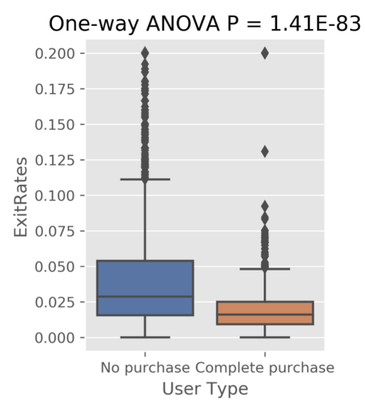
Google Analytics metrics
- Bounce rate
- Exit rate
- Page value
- Special day

Patterns of User Behaviors

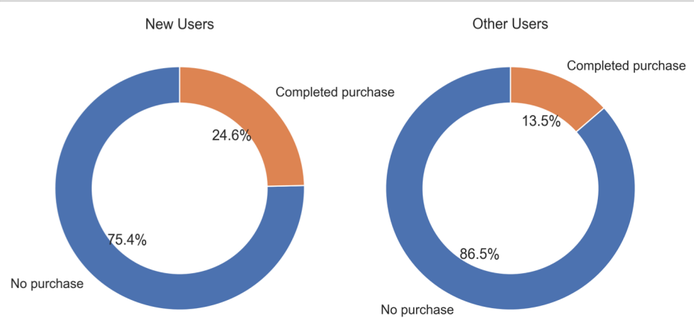
Users who purchase interact more with the webpages

On average, user who purchase visit more pages, and spend more time on these pages. They also have significantly lower bounce rates, exit rates, and higher page values.





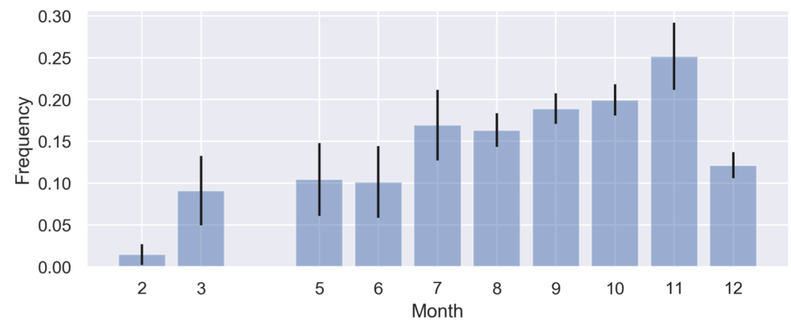
New users are more likely to complete purchase



It is commonly believed that returning users are more valuable to e-commerce companies than new users. However in this dataset, new users are significantly more likely to complete the purchase.

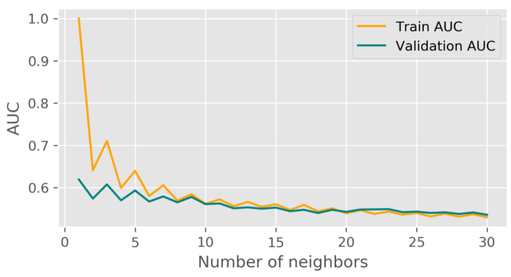
Users are more likely to complete transactions in November

The barplot shows the proportion of users who complete transactions in each month available in the training data. In November, users are significantly more likely to purchase, possibly due to the upcoming holiday seasons. Also, we noticed that data from January and April are missing.



Predictive Models

Scikit-learn



Select K for KNN with hold-out validation

K Nearest Neighbors
- Used K = 5
- Train AUC: 0.64
- Val AUC: 0.59
- Test AUC: 0.58

Logistic Regression
- L2 ridge regularization
- Train AUC: 0.68
- Val AUC: 0.68
- Test AUC: 0.66

Random Forest
- Parameters tuned via cross-validation
- Train AUC: 0.91
- Val AUC: 0.73
- Test AUC: 0.73

H2O

0.92 AUC

Best performing model

XGBoost (H2O AutoML)

Confusion matrix of XGBoost prediction

	False	True	Error	Rate
False	2821.0	269.0	0.0871	(269.0/3090.0)
True	161.0	448.0	0.2644	(161.0/609.0)
Total	2982.0	717.0	0.1162	(430.0/3699.0)

AutoML Leaderboard top-3 models

model_id	auc	logloss	mean_per_class_error	rmse	mse
XGBoost_2_AutoML_20191021_220104	0.924836	0.24189	0.16618	0.268566	0.0721277
XGBoost_1_AutoML_20191021_220104	0.924468	0.241069	0.151248	0.269428	0.0725913
XGBoost_grid_1_AutoML_20191021_220104_model_2	0.9235	0.242138	0.181113	0.267351	0.0714764

Summary of Insights

User behavior patterns in e-commerce data

1. Those who interact more with the webpages are more likely to complete purchases.
2. New users are more likely to purchase than returning users in this dataset.
3. Users tend to complete transaction in November than in other months, possibly due to the upcoming holidays.

Predictive modeling

1. In prediction tasks, ensemble models usually have an edge over single learning methods.
2. H2O handles imbalance data beautifully, but in terms of feature selection, it is not flexible enough yet.