TMall Repeat Buyer Prediction

Yunhe Cui | Shijia Gu | Jiawen Liang | Pengzi Li

Data Mining Final Project

Instructor: Claudia Perlich

May 07, 2019

BUSINESS UNDERSTANDING

TMall, formerly known as Taobao Mall, is the largest business-to-customer online retailer in China, operated by Alibaba. It is a platform that allows merchants to sell their products as well as virtual goods to customers. TMall was launched in 2010, and all the products on this website are sold either by the brand owner or the authorized distributor. In 2011, Alibaba was recombined into three companies, and TMall became one of them. After becoming an independent company, TMall never stops growing and expanding. In 2013, it took more than half of the Chinese business-to-customer market share. Now, it has more than 500 million users and 50,000 online stores, which include more than 70,000 brands.

In 2009, Alibaba started an online shopping festival called "The double eleven" because it will be held on November 11th every year. It is a same kind of promotion event as Black Friday in the US and or Boxing day in the UK. There were not many brands involved in this festival (only 27 brands) in the first year, and the discount was not very attracting. However, the daily sales for that Double 11 Event were over 50 million yuan, which was far more beyond the expectation. In 2010, there were 711 online stores participated, and the daily sales increased to 936 million yuan (CIW team, 2015). Buyers and sellers started to realize how the enormous the impact of this festival was. Double 11 then became the biggest online shopping event in China immediately. On November 11, 2014, in the fifth Double 11 festival, the day sales of the festival were 57.1 billion yuan (Millward, 2014). In 2018, the number reached 213.5 billion yuan (Arjun, 2018).

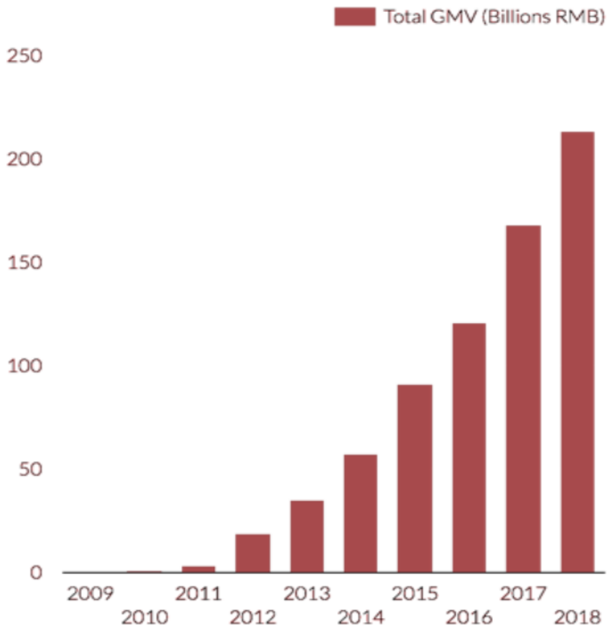


Figure 1. The histogram of daily sales on Double 11 festival, 2009-2018, (Ventures, 2018)

During this phenomenon shopping festival, all the sellers are trying to play all their cards to attract customers. Big promotion, discount, and giveaways are always involved in the Double Eleven festivals. Most of the sellers use coupons as one of their tools to attract customers. However, the buyers that attracted by coupons are mostly onetime buyers, who barely can benefit sellers in the long term. How to attract new customer and how to maintain customer loyalty became new challenges in front of all retail sellers. The more efficient a seller can attract new customers while maintaining the old customer, the more business value the seller will gain. Normally, sellers are expecting coupons will allure old customers to come back and become repeat buyers since most customers are lazy, price sensitive and incline to buy from stores they familiar with. However, if the merchants give coupons to every old customer, even though it will increase the number of return customers, it will also affect the profit negatively since there are some customers are coming back whether there are coupons or not. By building this model, we are planning to use the model to help merchants to predict potential repeated buyer so that they can efficiently reduce the promotion cost and gain more profit with loyal customer to get the "win-win" situation: the stores "hunted" new customers who may become returning customer and contributes to their future benefit, and the customers get the product at a relatively lower price.

DATA UNDERSTANDING

The data we are using for this project is provided by Tianchi Datahub which is owned by Alibaba, the parent company of TMall. The dataset contains anonymous user shopping log of six months, from May 11th to Nov 12th. In consideration of users' privacy, the dataset is sampled in a biased way, so the result that computed from this dataset would be different compared with the result of actual data from TMall. However, it will only have limited effect on the usability of the result.

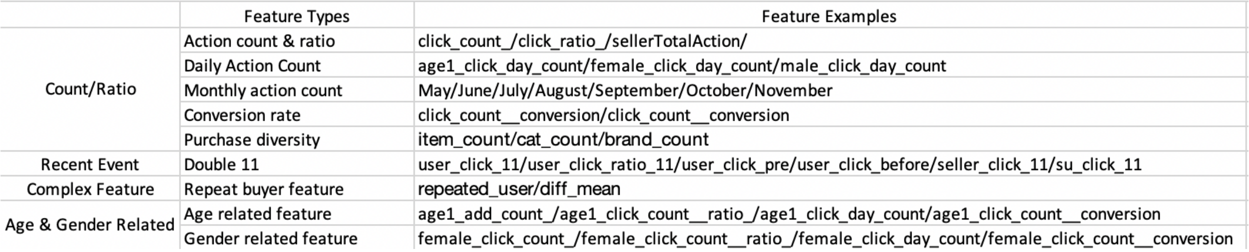
The dataset consists of four parts. The first part is called user data sample, which contains seven features:

- user ID: a unique id for the shopper;
- item ID: a unique id for the item;
- category ID: a unique id for the category that the item belongs to;
- seller ID: a unique id for the merchant;
- brand ID: a unique id for the brand of the item;
- timestamp: the date the action took place (format: MMDD, 511 ~ 1112);
- action type: it is an enumerated type {0,1,2,3}, where 0 is for click, 1 is for add-to-cart, 2 is for purchase and 3 is for add-to-favorite.

The second part is the user info, which only indicates the gender and age range of users. Gender contains women, men, and unknown. There are 8 different age ranges divided in the data, where 1 indicates the user with age smaller than 18, 2 for 18 years old to 24 years old, 3 for 25 to 29, 4 for 30 to 34, 5 for 35 to 39, 6 for 40 to 49, 7 and 8 represent 50 years old and above. The last two parts are the train set and the test set, where the train set contains the label indicating whether a user is a repeat buyer after the Double 11 Event and the test set only has user id feature with no label. In this case, we can predict labels and test our models using the test set. There are three different labels in the train set: 0, 1, and -1. 1 indicates the user is a repeat buyer and 0 indicates the opposite. -1 means the user is not a new buyer, so he or she should be excluded from the prediction.

DATA PREPARATION

Since our goal is to predict whether a new buyer would become a repeat buyer, this problem can be defined as a typical classification problem. The process of building models is not so much different from that of other classification problems. Instead, in our case, feature engineering is the most important part to build good models. As mentioned before, the original dataset only contains nine features, which is far from sufficient to train good classification models since our shopping history record dataset has over 500 million rows. Therefore, we made an intensive study of features, generated range from basic counts to complex features and classified into five groups, which respectively are: count/ratio features, product diversity features, repeat buyer features, Double 11 event features, and age/gender-related features. Table 1 provides a summary of the types of features contained in these profiles.



	Feature Types	Feature Examples
Count/Ratio	Action count & ratio	click_count_/click_ratio_/sellerTotalAction/
	Daily Action Count	age1_click_day_count/female_click_day_count/male_click_day_count
	Monthly action count	May/June/July/August/September/October/November
	Conversion rate	click_count__conversion/click_count__conversion
	Purchase diversity	item_count/cat_count/brand_count
Recent Event	Double 11	user_click_11/user_click_ratio_11/user_click_pre/user_click_before/seller_click_11/su_click_11
Complex Feature	Repeat buyer feature	repeated_user/diff_mean
Age & Gender Related	Age related feature	age1_add_count_/age1_click_count__ratio_/age1_click_day_count/age1_click_count__conversion
	Gender related feature	female_click_count_/female_click_count__ratio_/female_click_day_count/female_click_count__conversion

Table 1. Summary of feature groups

Count/ratio features are the basis for generating more complex features. Action counts are the number of click, add-to-cart, purchase, add-to-favorite actions for each id over the whole data period as well as in each month (monthly counts). Action ratio is defined as the proportion of a particular action type among all action types for each id over the whole period. Active day counts are the day count of a particular id has actions, which mainly used to measure active levels for different user/seller/su_id. Conversion ratio is defined as the ratio of a non-purchase action count for each id and purchase count over the whole data period.

Product diversity features, concerning a user and a seller**,** are the number of unique items, categories, and brands that the user took actions in each month as well as over the whole data period. For a merchant, product diversity features are defined similarly. For the "users and sellers" pair, they are the number of unique items, categories, and brands of the seller that were taken actions by the user in each month or over the user. The intuition behind product diversity features is that if there are more items in a merchant that appeal to the user, then the chance that the user repurchase from the merchant might be greater.

Repeat Buyer Features includes two parts: For users we constructed a binary label to annotated repeat users (1 for repeat buyer while 0 for non-repeat buyer) and a new feature containing the mean of the time interval of repeat purchases conducted by a repeat user in a particular merchant; for merchant, we calculated the number of buyers and repeat buyers for each merchant, then calculated ratio between them to measure the ability of constant attraction of customer. Shorter time intervals of repeat purchases conducted indicate that the user has a solid repeat purchasing habit; meanwhile, a large repeat customer base and higher returning rate reveals that the merchant is widely favored and has a good reputation, so the customers are more likely to come back again.

Double 11 Event Features are counts of clicks, add-to-carts, purchases, add-to-favorites based on unique user id, seller id and user-seller id (i.e. id of each user associated with a particular merchant) for three different time intervals (on the Double 11 Day, one week before Double 11 Day, and 5-month period before that week). The ratio of each action counts to the overall counts in every time interval are also calculated. If a user's buy-ratio is much higher on November 11 than in the other two time-intervals, then the user is most likely to be a one-time buyer.

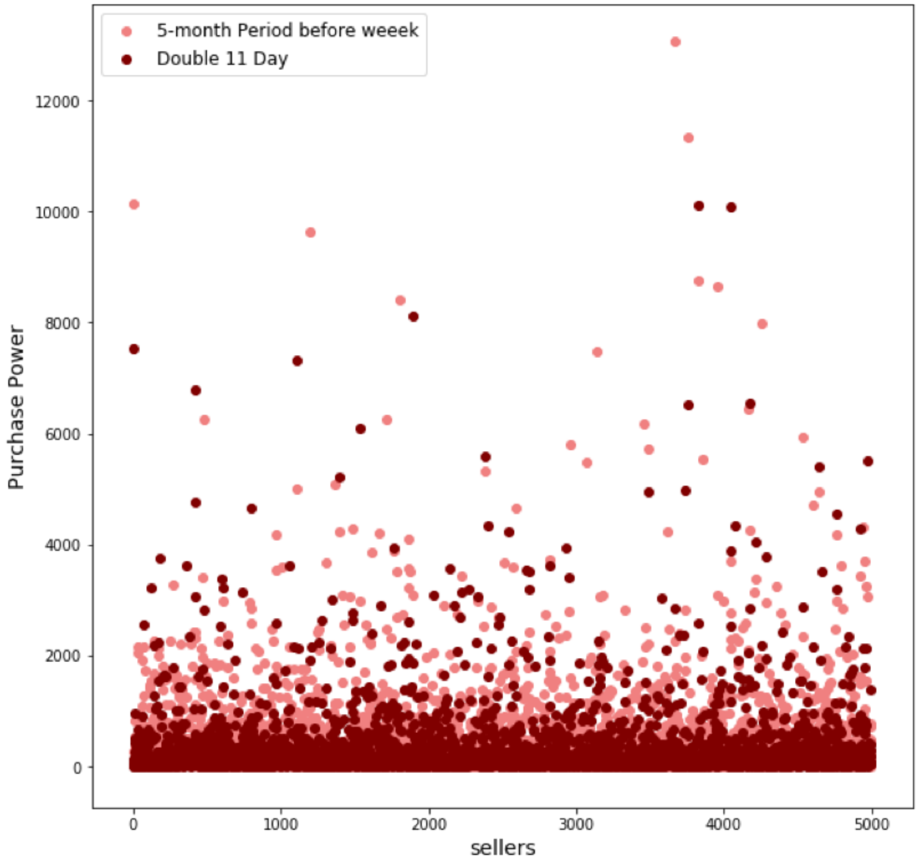


Figure 2. Distributions of the number of purchase actions on Double 11 Day and during the previous 5-month for each merchant

From the above scatter plot, the purchasing power within Double 11 Day nearly similar to that during the previous five-month, which implies the significant impact of Double 11 Event. Furthermore, if a user's buy-ratio is much higher on November 11 than in the other two time-intervals, then the user is more likely to be a one-time buyer.

Age/gender-related features measure different shopping habits and buying decisions within different user groups. For example, TMall, as an online retail website, is more widely used among younger people while order people favor traditional shopping methods. Another example is concerning genders: female consumers tend to be more astute than male customers as they are willing to invest time and energy necessary to research and compare products (Lewis, 2018). As such, we generated features to describe the different action pattern conducted by different user groups, where users are grouped based on their genders and age range defined in previous data preprocessing. These features include every action counts and ratios, total action counts, daily aggregation on daily action counts, as well as the purchase conversion rate for each age and gender group.

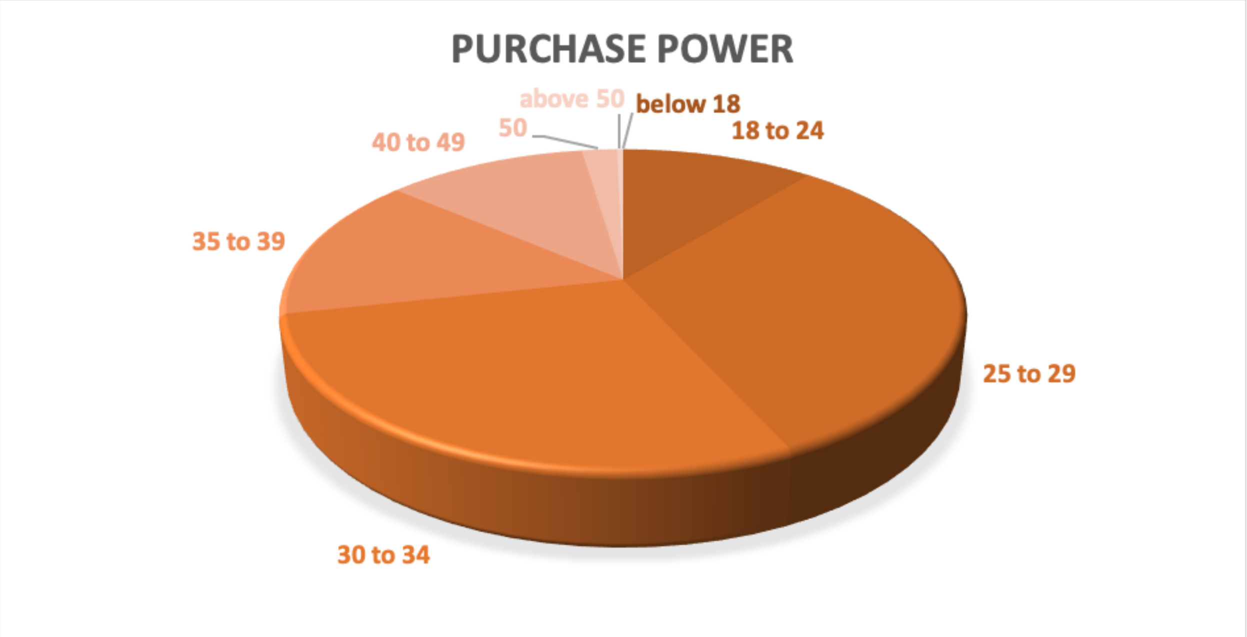


Figure 3: Pie chart of purchase power for each age group

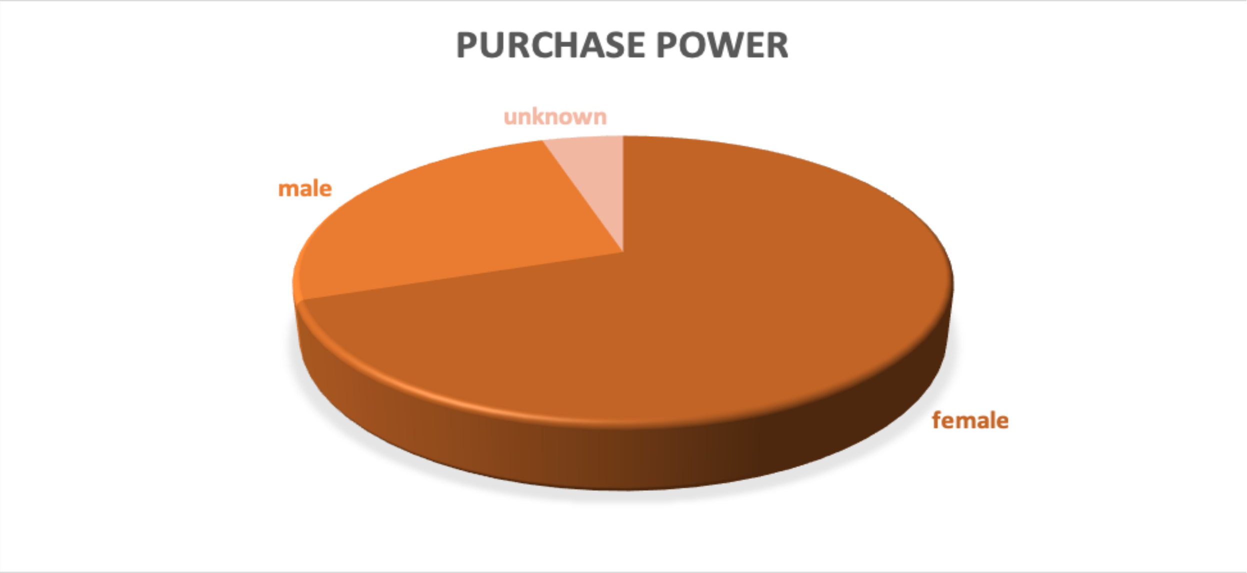


Figure 4: Pie chart of purchase power for each gender group

Figure 3 reveals that it is authentic that TMall is widely used among younger people from 18 to 39. Not surprisingly, according to Figure 4, the purchasing power of the female consumer is dominant in TMall online retail market.

After we performed feature engineering, we expand our data features from 9 to 353 which enables us to start a more meaningful model building and feature selection process.

MODELING

To select the best model for this project and achieve higher accuracy, we trained six different supervised classifier models with our 353 features by using DataRobot which is an automated machine learning platform.

Our target variable "label" is a dummy variable that indicates whether this user is a repeat buyer: 0 means no, 1 means yes. Since we have labels in our dataset, this project should be applied supervised models.

In this project, we employed four different algorithms: Naive Bayes Classifier, random forest classifier, support vector machine classifier, and logistic regression. After running all the model, we then selected the one that has the best model performance.

Naive Bayes Classifiers

First, we tried the Naive Bayes model since we want to create an interpretable model and understand how each attribute affects our predictions and it is simple and fast to use. Naive Bayes is an algorithm based on Bayes' Theorem, which has good interpretability. It assumes that all attributes are conditionally independent given the class and works well with high dimensions. However, the performance of Naive Bayes is highly correlated to the independence assumption, that is to say, Naïve Bayer will perform poorly if the premise is not met (Tufts, 2015).

The important features when running Naive Bayes classifier shown below:

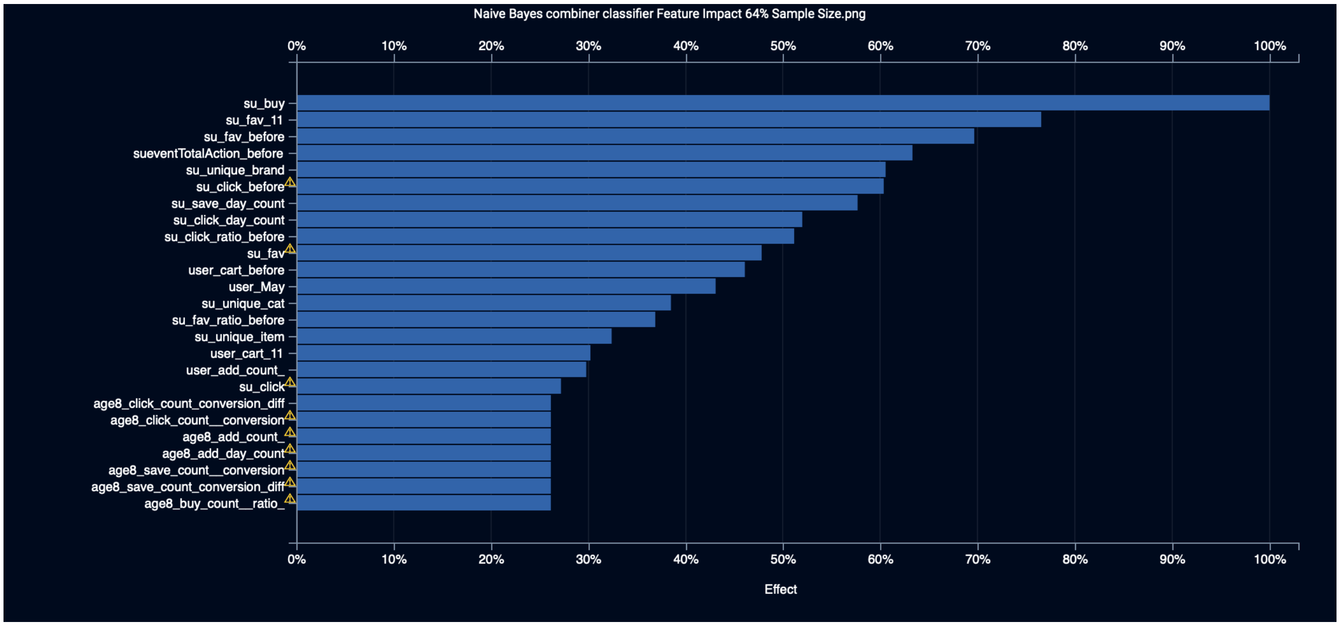


Figure 5. Naive Bayes: Feature Importance

The number of purchase of a user did in a particular merchant is the most informative feature, while the counts of a given user add items in a store to a favorite during double 11 and counts of a user save items of a particular store to favorite prior double 11 are also important features in this model.

Random Forest Classifier

Random Forest is a supervised learning algorithm which has a reputation on higher accuracy, and it is easy to implement. It is a tree-based ensemble model; we restrict our choice to a randomly chosen subset of features when building each tree. Random Forest does not require careful parameter tuning or huge amounts of training data, but it is a black-box classifier which lacks interpretability.

The important features when running Random Forest classifier shown below:

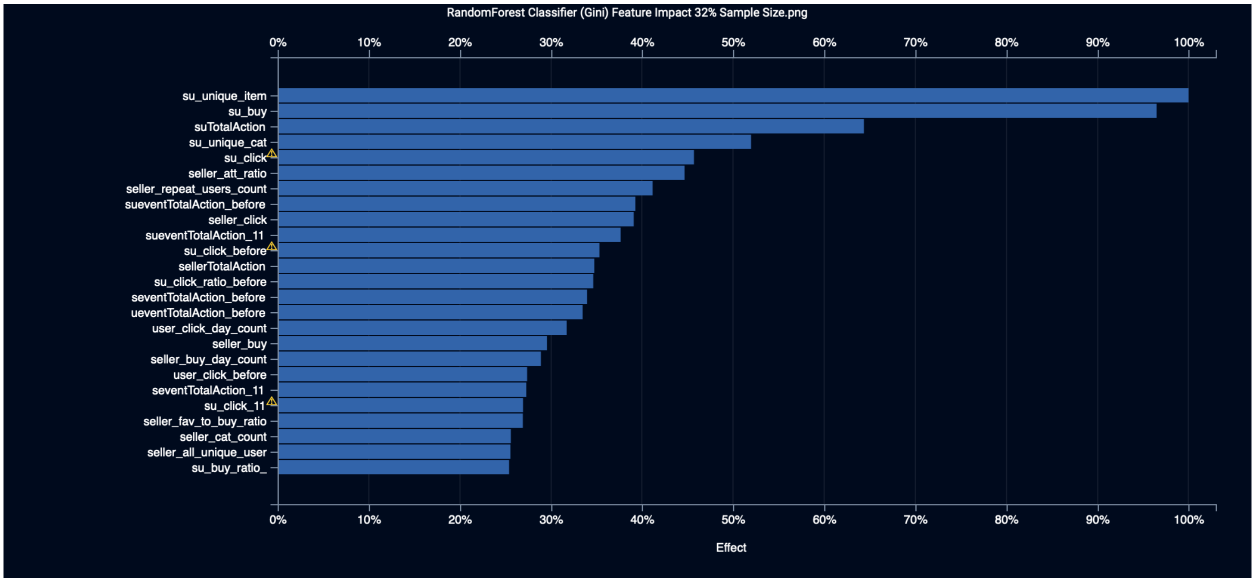


Figure 6. Random Forest: Feature Importance

The most informative feature in Random Forest classifier is the number of unique items for each user bought from a particular merchant. The number of purchases of a user did in a specific merchant, and the total action of a given user interactive with a particular merchant also has a significant impact on the target variable.

Support Vector Machine Classifier

Support Vector Machine (SVM) is an optimization-based prediction approach used primarily for binary classification which fits our target variable. SVM assumes real-valued attributes on the same scale. Thus, it is very important to pre-process the data before training the model. SVM has outstanding performance, and its theoretical guarantees about their generalization performance (accuracy for labeling test data) based on statistical learning theory. However, it is susceptible when choosing parameters: parameter choice would significantly affect model performance.

From Figure 7 shown below, the number of purchases of a user did in a particular merchant is again the most informative feature in the model. And the two product diversity features (number of unique items/categories for each user bought from a particular merchant) are the rank 2 and rank 3 importance feature when running SVM classifier.

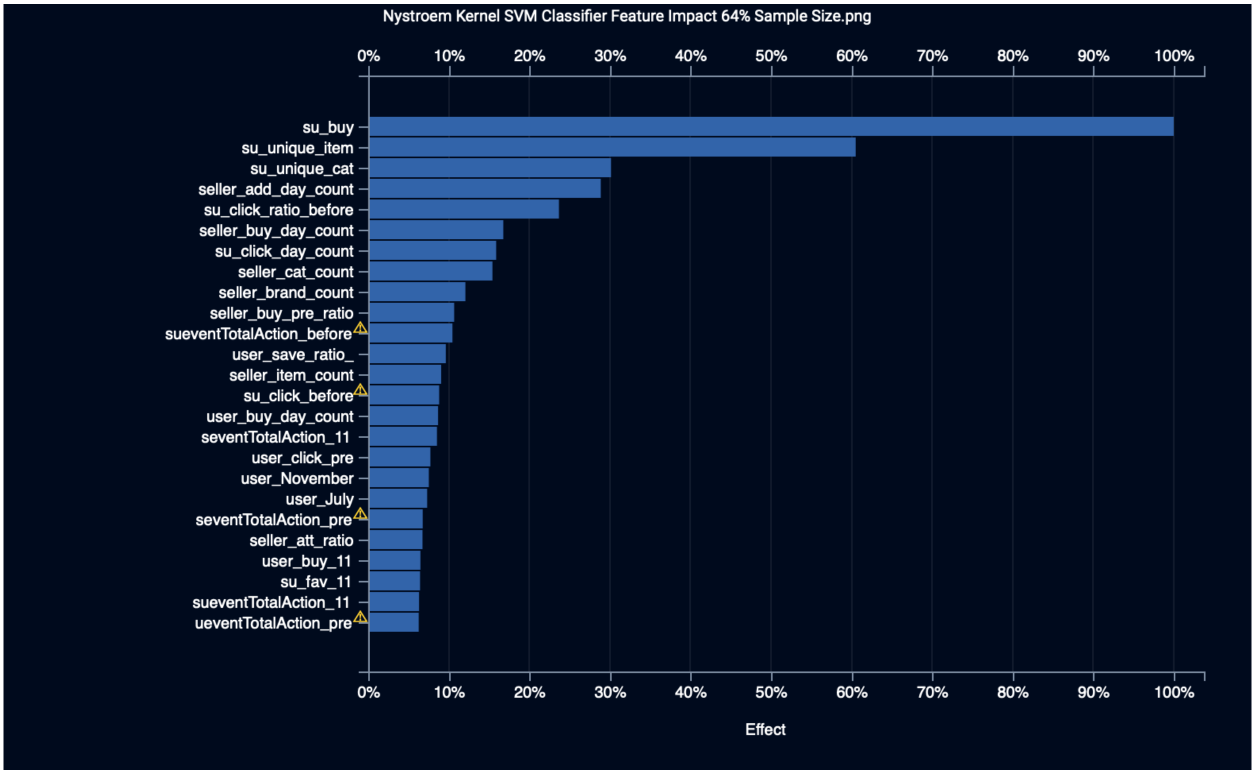


Figure 7. Support Vector Machine: Feature Importance

Logistic Regression

Logistic regression is an easily interpretable predictive algorithm that appropriate for applying when the dependent variable is a binary variable. But it does not have a requirement on independent variables: could either numerical or binary variable) (“What Is Logistic Regression?”, 2019). However, logistic regression cannot solve nonlinear problems since its decision surface is linear (Mayfield, 2018).

The most informative feature in the logistic regression model is still the number of purchases of a user (See Figure.8). But the rank 2 and 3 of the important feature changed this time: the ratio of a given user's click items in a store before double 11 and the number of days with purchase action performed by a user in a particular store.

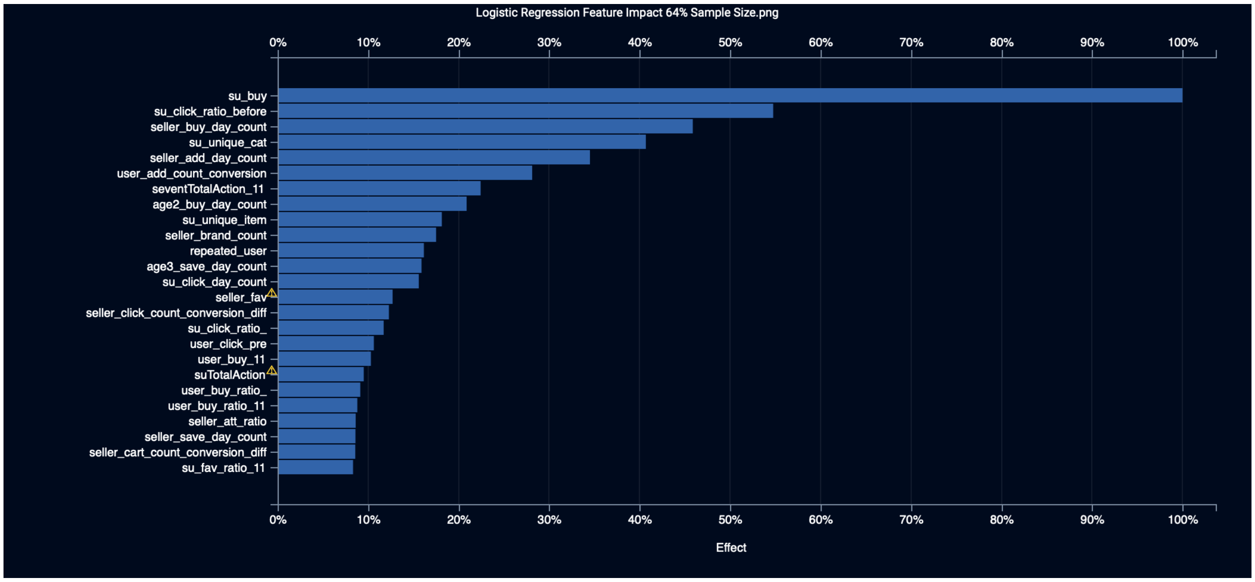
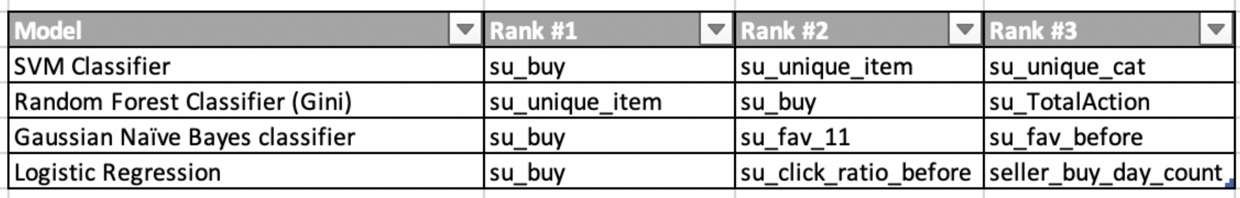


Figure 8. Logistic Regression: Feature Importance

Based on the features importance discussed above, the number of purchase action of a user made in a particular merchant is the number one important feature in three out of four of the models we performed and the product diversity features (the most popular item/category in the store) also plays crucial roles when running various models (See Table 2). We can then conclude that merchants in TMall e-commerce platform should pay more attention to users who purchased at least one item in the store, and the product diversity the user interested and purchased.



Model	Rank #1	Rank #2	Rank #3
SVM Classifier	su_buy	su_unique_item	su_unique_cat
Random Forest Classifier (Gini)	su_unique_item	su_buy	su_TotalAction
Gaussian Naïve Bayes classifier	su_buy	su_fav_11	su_fav_before
Logistic Regression	su_buy	su_click_ratio_before	seller_buy_day_count

Table 2. Summary of feature importance

Model Business Implication

In terms of business use, the promotion cost has a great significance in the e-commerce business, which accounts for loss of potential profit. As the e-commerce market competition becoming more and more fierce, sending the same "Great Deal! $25 off when you spent $150" email to all customers cannot let the merchant strike a balance between cultivate repeat buying habit and remain an acceptable profit rate. In this situation, targeting the right customer to send coupon become an urgent need. The ideal model would be able to classify customers into two categories based on their action log, potential repeat buyers and non-repeat buyers with a reasonably high accuracy. With the application of our model, the target coupon will be activated by a certain character of user action.

Evaluation

After the data glance with feature engineering and data cleaning finished, we decided to use user-merchant interactive features to run the above four models as our baseline model. We want to test whether the model performance will increase if we add more features to the models. Therefore, our advanced model includes all 353 features we extracted from the data: user-related features, merchant-related features and user-merchant interactive features.

The evaluation parameter we used in this project is the Area Under Curve (AUC) score. The curve that indicates in the name of AUC is Receiver Operating Characteristics (ROC) curve; it is a probability curve which plotted with True Positive Rate (TPR) against the False Positive Rate (FPR). It is one of the most important evaluation metrics for checking any classification model's performance. In our project, the goal is to accurately predict whether a given user is a repeat buyer or not, so the higher the AUC, the better the model is at predicting repeat buyer as 1 and one-time buyer as 0.

From the bar plot below, the baseline models generally have poor performances compare to the advanced models, which indicates the more features we put into the models, the better performances we will have. Among the advanced models, SVM classifier has the highest AUC score, while the naive Bayes classifier has the lowest AUC score. Therefore, in this project, we will choose the SVM classifier as our final prediction model.

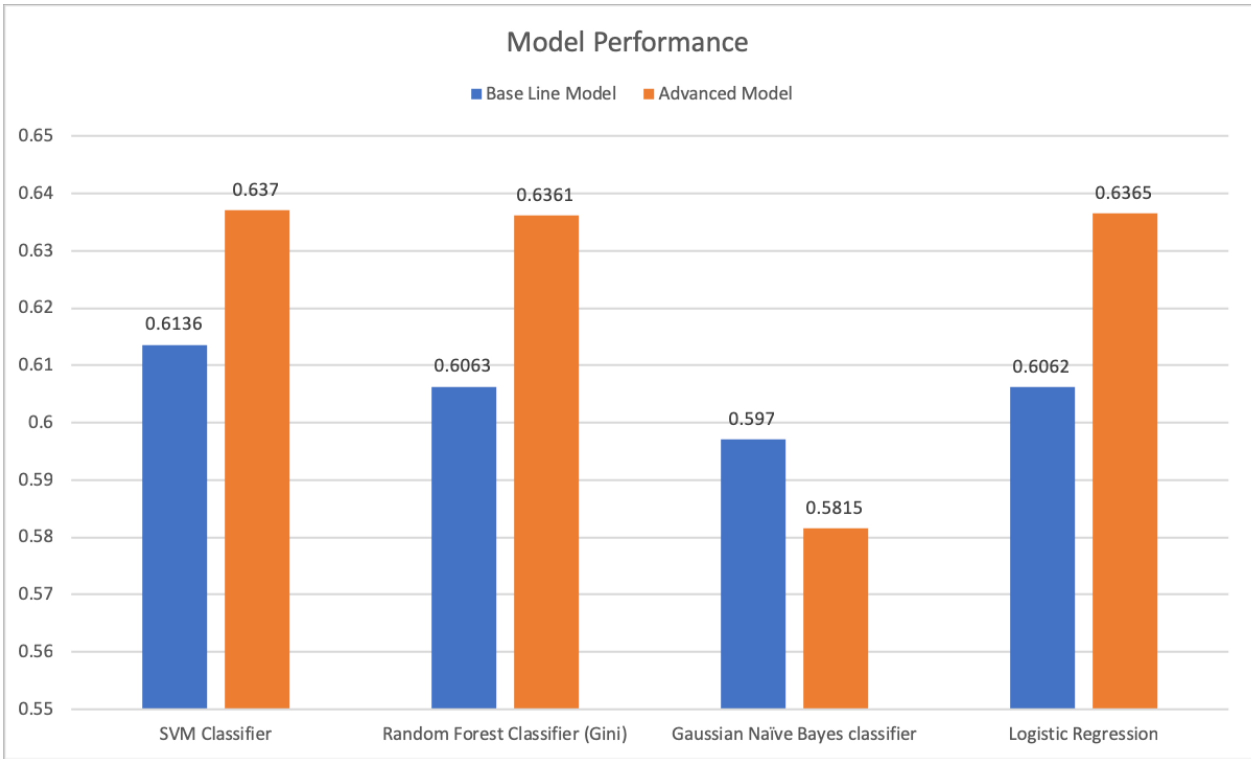


Figure 9. Model Performance (AUC score)

By using our developed model, merchants can allocate the coupons more efficiently. For example, if a seller gains five thousand new customers on the double eleven festivals, without our model, the seller will have to send coupons to all the five thousand new buyers, with no guaranteed repeat buyer conversion rate. In contrary, our prediction model (SVM classifier) can help the sellers to send coupons to the customers who need them most and possibly not come back without the coupon. Then we can track those customers who received coupons and calculate the repeat buyer conversion ratio.

We can also compute the return on investment (ROI) after the merchants using the model and see the improvements.

Concerns

The primary concern for all types of user behavior analysis is privacy. As online payment methods and online shopping platform constantly developing, people are becoming "transparent" with data gathering. Data analysts can easily summarize the behavior pattern of a certain person, make a precise prediction of his future action and generate a user portrait based on record dataset. Although people sometimes could benefit from the action tracing; for example, the customized recommendation list will enhance the shopping experience, ethical concerns are highlighted in this case. The online shopping accounts are generally connected with sensitive data such as home address and real name, it could be easy to identify an individual if those data are not desensitized correctly. Even worse, the probability of fraud will increase to a large extent if the credit card number and card verification value (CVV) is also recorded. Thus, how to strike a balance between blurring the sensitive data and gathering informative user data is challenging to the data scientist team for TMall.

Another concern for TMall repeat buyer analysis is that the training data we used might be biased and outdated. As mentioned in previous sections, TMall user group and user purchase habit changed dramatically with time goes. Ten years before, people are less likely to purchase virtual items such as service and electronic books than in recent days. Also, compared to the early TMall user group, we could easily identify that recent user group has a broader age distribution line (more age groups rather than young people only). There is an inevitable time lag between desensitized user data release.

Deployment

This data mining model could be used as the stepping stone of the coupon distribution optimization project and for future real-world e-commerce problem-solving. The sellers could use this model to identify potential repeat buyer and avoid giving the share of profit to the one-time-hunter (reduce non-targeted coupon) as much as possible. The step II coupon locator model could be used to explore the correlation between coupon value and repeat buying probability. The combination of these two models enables sellers to "hunt for" their loyal customer at the minimum cost.

Moreover, TMall should pay more attention (probably raise their budget) on shortening the turnover time between new purchase record collecting and desensitized user data publishing while maintaining a high data safety standard to avoid “garbage in, garbage out” problem.

Conclusion and Future work

In this project, we employed four different algorithms to predict the repeat buyers for TMall e-commerce platform. We generated a large number of features to capture the user behavior patterns and characters of merchants. However, none of these features is a single robust indicator of labels, so we need a large number of features to achieve an acceptable Area Under Curve (AUC) score. The best model we chose is Support Vector Machine Classifier which has a higher AUC score. Merchants in TMall platform should pay attention to the buyer who has purchased at least one items in the store and the product diversity in the store when considering sending coupons.

As for the data preparation, it was a tedious task to generate, clean and merge a large number of features. Exploring how to generate feature automatically will be time-saving when facing real-world business challenges. In the next step, we also want to construct an ensemble model which assign normalized weights by using grid search to the four classifiers and then use the weighted average probability as a final prediction to get a more precise and robust model.

Reference List

Arjun, Kharpal. "Alibaba Sets New Singles Day Record with More than $30.8 Billion in Sales in 24 Hours." CNBC. December 05, 2018. Retrieved May 05, 2019. https://www.cnbc.com/2018/11/11/alibaba-singles-day-2018-record-sales-on-largest-shopping-event-day.html.

Lewis, Michael. "Men vs. Women: Differences in Shopping Habits & Buying Decisions." Money Crashers. August 31, 2018. Retrieved May 06, 2019. https://www.moneycrashers.com/men-vs-women-shopping-habits-buying-decisions/.

Millward, Steven. "New Record for World's Biggest Shopping Day as Alibaba's Shoppers Spend $9.3 Billion in 24 Hours." Tech in Asia - Connecting Asia's Startup Ecosystem. November 11, 2014. Retrieved May 05, 2019. https://www.techinasia.com/china-alibaba-singles-day-2014-spending-total.

Mayfield, Philip. 2018. "The Logistic Regression Algorithm." *Machinelearning-Blog.Com*(blog). Retrieved April 23, 2018. https://machinelearning-blog.com/2018/04/23/logistic-regression-101/

Tufts, Chris. 2015. Cheat Sheets for Stats/ML/SP/DM. Contribute to Ctufts/Cheat_Sheets Development by Creating an Account on GitHub. Retrieved May 05, 2019. https://github.com/ctufts/Cheat_Sheets.

Ventures, Click. "Technology Lessons behind China's 30 Billion Retail Fiesta." Medium. November 21, 2018. Retrieved May 05, 2019. https://medium.com/swlh/technology-lessons-behind-chinas-30-billion-retail-fiesta-2fe9a0df457a.

“What Is Logistic Regression?” n.d. *Statistics Solutions*(blog). Retrieved May 8, 2019. https://www.statisticssolutions.com/what-is-logistic-regression/.

CIW Team. "Review of Sales Value on Double 11: 2009-2013." China Internet Watch. February 09, 2015. Retrieved May 05, 2019. https://www.chinainternetwatch.com/10400/review-sales-value-double-11-2009-2013/.