Under Armour Project: Member Lost Prediction

TegoChang

11/25/2021

Summary

In this report, we aim to predict if an existing member of Under Armour is likely to abandon the brand within the following six months. During the analyzing process, we have conducted exploring data analysis (EDA), model selection, assessment, and validation to construct a model with which we think can provide the best prediction to this topic. Our model explains 88% of the training dataset (AUC = 0.88) and predicts that 26% of the existing members might abandon Under Armour within the following six months.

Introduction

The questions we would like to answer in this analysis include 1) what factors might contribute to the churn rate of an existing user. 2) which members in the testing dataset are likely to abandon Under Armour in the following six months. Further, we would also pay attention to what additional methodologies could be applied to enhance the forecasting model.

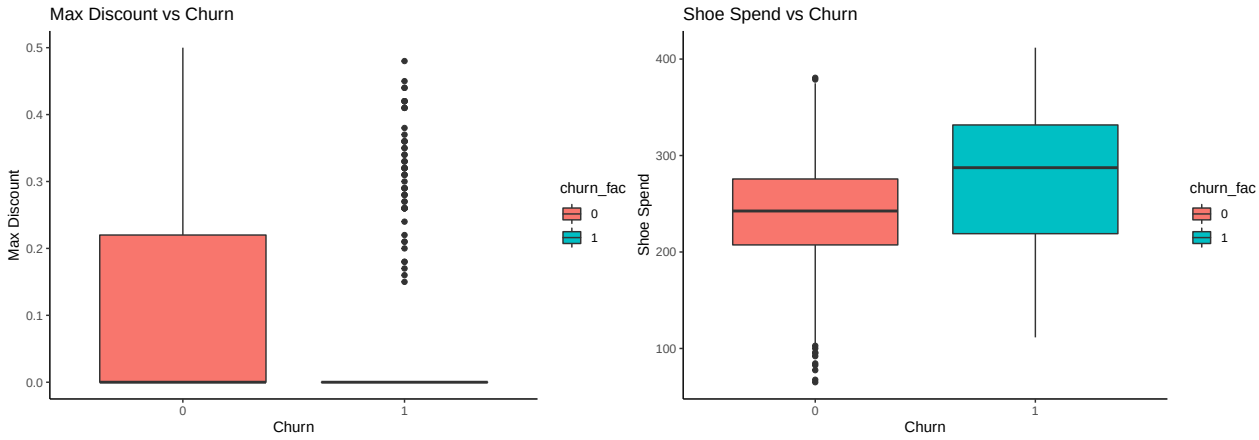
Data

In this statistical analysis, our inference and interpretation are according to the dataset "train.csv" with some parts of adjustments, including:

- We store two versions of the Churn variable, numeric and categorical, for analysis easiness.
- We set all the other variables to either numeric or factor based on their characteristics in the real world.

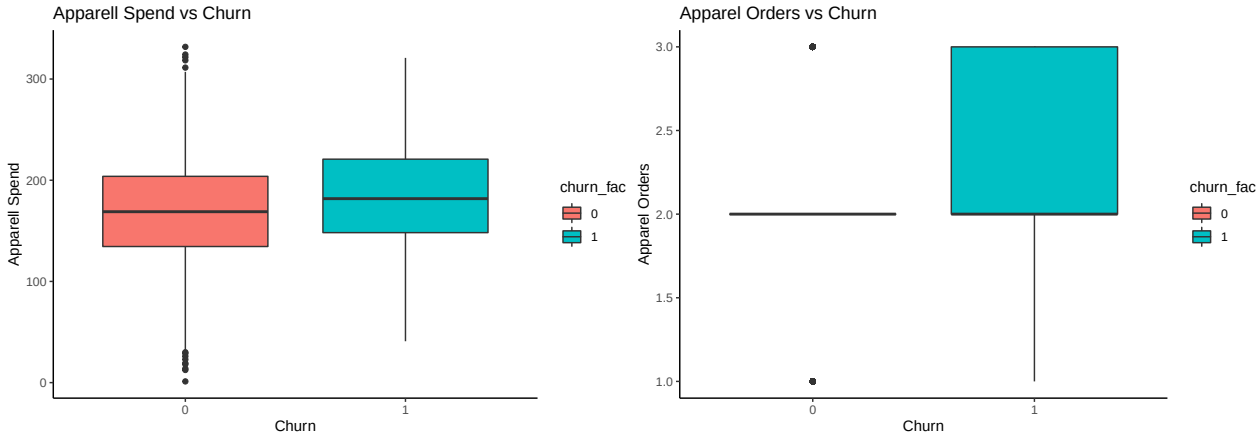
Based on the arranged dataset, we then start exploring data analysis (EDA). First, we check on all the numeric predictor variables' relation with the response variable, *churn*. We found that:

1. *max_discount* People who abandon the brand seems to get less discount.
2. *shoe_spend* People who abandon the brand pay more for shoes.



3. *custserv_calls* People who abandon the brand call custom service more.
4. *shoe_orders* More Shoe Order seems to have a higher Churn rate.

5. *apparell_spend* People who abandon the brand seem to pay more for Apparell. However, the difference is not obvious.
6. *apparel_orders* More Apparel Order seems to have a higher Churn rate.

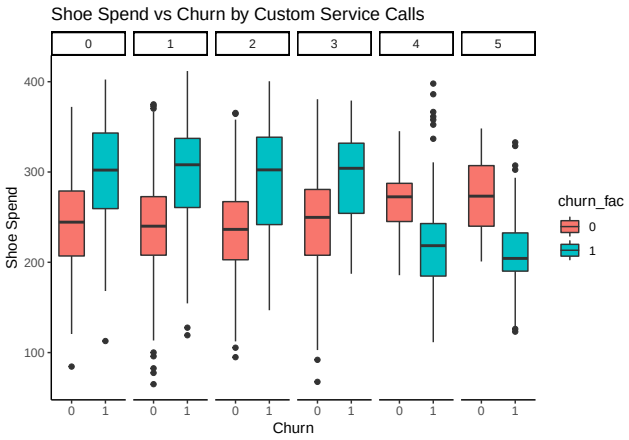


Then, we check on all the categorical variables' relation with the response variable, *churn*. We found that:

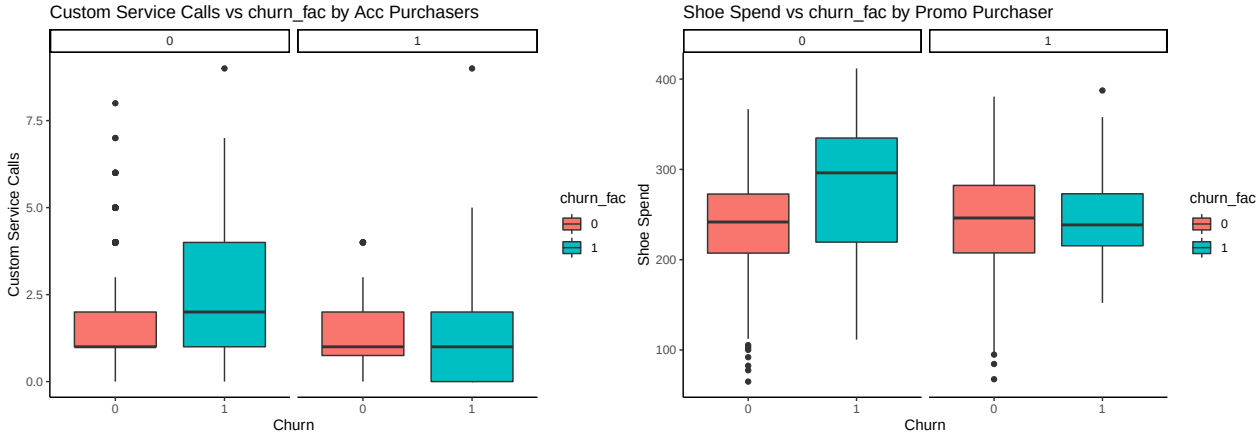
7. *acc_purchasers* ACC Purchaser has a higher Churn rate.
8. *promo_purchaser* Promo Purchaser has a lower Churn rate.
9. *gender* Male tends to have a little less Churn rate compared with Female.

At the last of EDA, we would like to know if there are interactions between the predictors. As the combinations of the predictors are too many to investigate one by one, we only pick up the pairs that we think are most likely to have mutual relationships in between. After investigations on several combinations, we decided that we might include the below three interactions into our model:

1. *shoe_spend* : *custserv_calls*



2. *custserv_calls* : *acc_purchasers*
3. *shoe_spend* : *promo_purchaser*



Modeling

As the response variable is binary in this research, we decide to build a logistic regression model to answer the questions of interest.

Model Selection

As our primary goal is to make predictions according to the testing dataset, test.csv, we consider to have fewer predictors included in our model in order to make better predictions. Thus, we decided to apply stepwise function with BIC as our model selection approach. The Null model only includes the two predictors, *shoe_spend* and *custserv_calls*, which we found most significant in the relationship with our response, *churn*; On the other hand, the full model includes the nine predictors, which we found indicating certain relationships with *churn*, and the three interactions mentioned in the previous sections.

Below lists the mathematical equation of the model suggested by stepwise BIC:

$$\begin{aligned} y_i = \beta_0 + \beta_1\ shoe\_spend_i + \beta_2\ custserv\_calls_i + \beta_3\ acc\_purchasers_i + \\ \beta_4\ promo\_purchaser_i + \beta_5\ apparell\_spend_i + \beta_6\ shoe\_orders_i + \\ \beta_7\ shoe\_spend_i : custserv\_calls_i + \beta_8\ shoe\_spend_i : promo\_purchaser_i + \\ \epsilon_i; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n. \end{aligned}$$

where y_i is log-odds of existing members abandoning the brand. Just to be safe, we also conduct an anova Chi-square test to confirm that the model selected by stepwise BIC is indeed statistically different from the

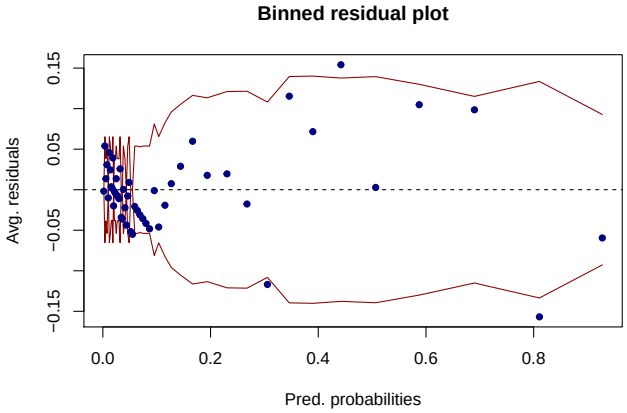
full model we mentioned above. The summary table of our proposed model is listed in below table. We can tell that all main effects and most of the interactions are statistically significant with a P-value a lot less than 0.05. This states that these predictors and interactions are very influential in predicting whether an existing member will abandon the branding shortly.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.21	1.00	-11.24	0.00
shoe_spend	0.02	0.00	5.59	0.00
custserv_calls_fac1	-0.44	1.17	-0.38	0.70
custserv_calls_fac2	0.88	1.18	0.75	0.46
custserv_calls_fac3	0.24	1.52	0.16	0.87
custserv_calls_fac4	12.78	1.44	8.85	0.00
custserv_calls_fac5	16.07	1.85	8.71	0.00
acc_purchasers1	2.22	0.18	12.23	0.00
promo_purchaser1	4.75	0.94	5.05	0.00
apparell_spend	0.01	0.00	6.84	0.00
shoe_orders	0.67	0.24	2.84	0.00
shoe_spend:custserv_calls_fac1	0.00	0.00	0.30	0.77
shoe_spend:custserv_calls_fac2	-0.00	0.00	-0.68	0.50
shoe_spend:custserv_calls_fac3	-0.00	0.01	-0.29	0.77
shoe_spend:custserv_calls_fac4	-0.04	0.01	-7.64	0.00
shoe_spend:custserv_calls_fac5	-0.05	0.01	-7.23	0.00
shoe_spend:promo_purchaser1	-0.02	0.00	-6.18	0.00

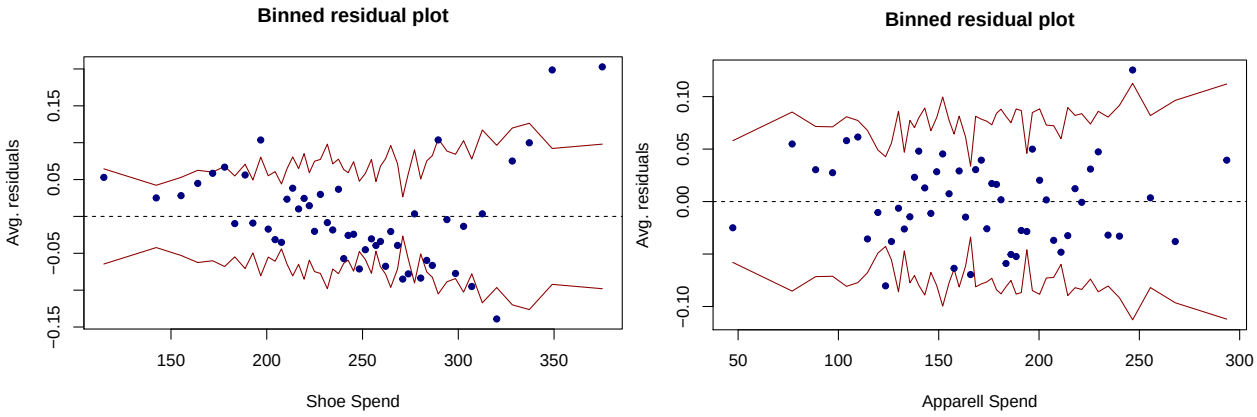
Table 1: Logistic Regression Results (Log Odds Scale)

Model Assessment

When diagnosing the model, we only plotted the binned raw residuals versus predicted probabilities, *shoe_spend*, and *apparell_spend* as the other binned plots of other variables like *custserv_calls* contain few data points for us to verify the assumption. In the first plot, residuals versus predicted probabilities, we found that there are only three data points outside the red line, the 2 Standard Error bands coverage. However, the distribution of the data points seems doesn't have a random pattern, which violates the assumptions of logistic regression. This could be an issue for us to investigate in the future.



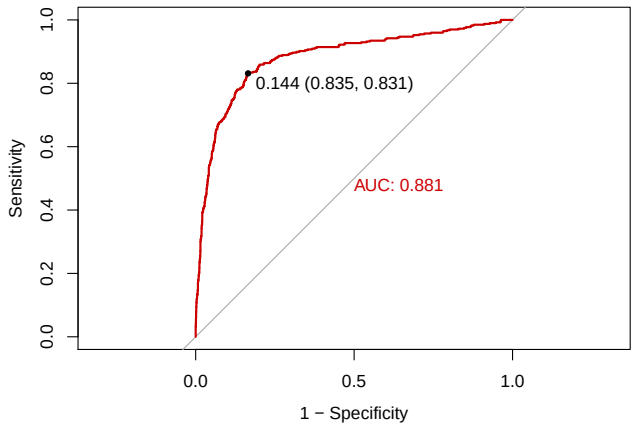
In the second figure, residuals versus *shoe_spend*, it has the similiar phenomenon as the first one: the data points within have a non-randomized pattern, and some points exceed the 2 Standard Error bands coverage. This indicates conducting transformations with *shoe_spend* might be necessary; In the last figure, residuals versus *apparell_spend*, the data points within has a randomized pattern and only few points exceed the 2 Standard Error bands coverage. This means we don't have to do any transformation with *apparell_spend*.



Model Validation

We then started to proceed with the validation part for our proposed model. We built the confusion matrix for our model with the threshold set to be 1) 0.5, which means that if the predicted probability exceeds 0.5, we consider the member will be lost and 2) the one suggested by the ROC curve, 0.144. The corresponding performance and the ROC curve are shown as below:

- Accuracy: 0.89
- Sensitivity: 0.44 & Specificity: 0.97
- The best threshold is 0.144 (Specificity: 0.83, Sensitivity: 0.83) and comes with AUC 0.88



Conclusion & Future Work

Based on our proposed model, we make predictions for the testing dataset, test.csv. According to the suggested threshold from ROC curve, we assume that if the predicted probability exceeds 0.144, we consider the member will be lost. The result shows that 175 among 667 members are expected to abandon the brand within the following six months, which is around 26%. These members are identified in the last column of test.csv, *churn*.

Hierarchical Model

In this research, we also noticed that the proposed model can be further constructed into a hierarchical model with *state* applied as a group variable. EDA for the relationships between *state* and some of the significant predictors are conducted as below figures. Though the interactions seem not very obvious, it could still be a direction that this research can be further extended.