Conversion Prediction Using Logistic Regression

Ren Wang

Jan 15, 2019

Executive Summary

- **Problem:** We developed a model that predicts whether a user will convert (buy) or not using logistic regression.
- **Method:** We first selected multiple features for the model. Then we used dataset with "train = true" to generate a logistic regression model, and also took the Likelihood Ratio Test. We used dataset with "score = true" to take the ROC test in order to choose the best thresholding value. Finally, we predict on the dataset with "test = true".
- **Results:** We have the prediction accuracy 91.43% on the testing set.
- **Evaluation:** We observed that all the three features are significant for the model, and our model can predict well in the testing dataset.

Findings from your exploratory data analysis

- When the conversion is “True”, usually num_impressions or avg_relevance is high.
- Most of users correspond to more than one session.
- The adds of sessions in the history have impact on the conversion of the current session.
- The number of sessions in the history ”seems” affect the current conversion.
- The impact of search number is not clear.

Feature engineering efforts

- ads_impress = num_impressions or avg_relevance is high. This feature measures the total relevance of ads in the current session.
- impress_prev = $\sum_{i=1}^{k}$avg_relevance(i)$\times$num_impressions(i)/(diff_day(i) + 1). The same user in the current session has interacted with k sessions before the current session. Here avg_relevance(i), num_impressions(i), diff_day(i) are the relevance of ads, number of ads of the i-th session, and the number of days of the i-th session ahead of the current session (time affects the impact). This feature measures the impact of the total relevance of ads in the previous session to the user of the current session.
- view_hist = k. Number of previous interactions. This measures the interest of users to the sessions.
- num_bought. Number of products bought before. (we did not use this feature in the final model)
- num_search. (we did not use this feature in the final model)
- The number of "true" and number of "false" of the conversion are

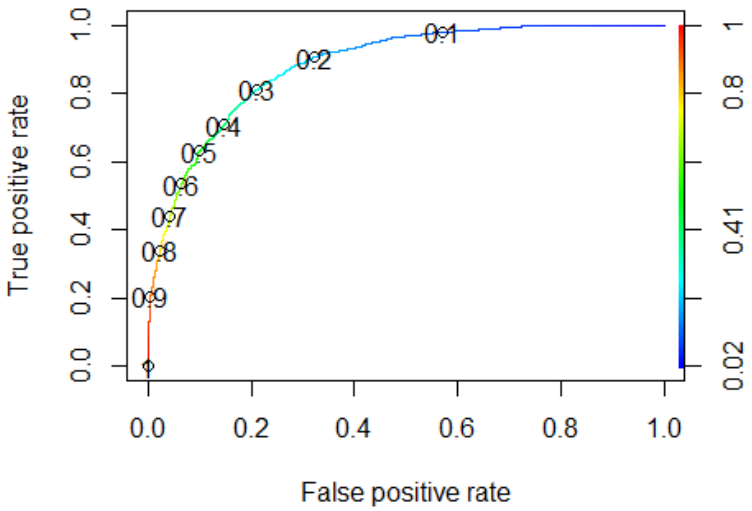
Model forms tried and the final model

- Model $\sim$ ads_impress + impress_prev + view_hist + num_bought + num_search.
- Model $\sim$ ads_impress + impress_prev + view_hist + num_search.
- Model $\sim$ ads_impress + impress_prev + num_search.
- Model $\sim$ ads_impress + impress_prev.
- All the above models except combined ads_impress and impress_prev as the same variable.
- Model $\sim$ ads_impress + impress_prev + view_hist (final model). y = 0.70995$\times$ads_impress + 0.54470$\times$impress_prev + 3.03041$\times$view_hist - 10.59789. We choose the thresholding of probability as 0.3. (obtained from ROC)

Justification of the model

- I used the Likelihood Ratio Test and the summary of the model to check whether the variable is significant.
- I checked the confusion matrix and the ROC curve.
- num_bought, num_search are not significant. The coefficients in the final model are all significant when using the training data. The view_hist is not significant when using the dataset with "score = true".
- The percentage of the false negatives is large compared with the percentage of the false positives.

Performance metrics for the model



(a)

Actualvalue	Predictvalue FALSE	TRUE
FALSE	11744	826
TRUE	460	1970

(b)

Figure: (a) ROC curve. (b) The confusion matrix of the prediction.

Ideas for future development

- Keep collecting the data (the range of the dates is small in the current dataset).
- Collect the time the user stayed in each session.
- Do A/B testing to see whether a feature affect the conversion.