Conversion Rate Prediction and Improvement

Zehao Hui
hzh98@bu.edu
U77561194

Abstract

Conversion rate has become one of the hot spots that many companies pay attention to nowadays. Increasing conversion rate means increasing profits in most cases. In the case of website data records, I hope to give reasonable suggestions to improve the conversion rate by analyzing a series of features. Machine learning predictive models based on these features are also built to predict whether customers will convert. The report will give specific data analysis, visualization pictures, specific results and conclusions of the prediction model.

1 Introduction

For many companies, especially Internet companies, they will attract others to log in to their website homepage and register by advertising or occupying search engines. The ratio of actual consuming behaviors to pageviews is called the conversion rate. And my project is to build a model and predict the conversion rate from an existing data set. Besides, some data analysis will be done to give some instructions on how to improve conversion rate.

This program can provide companies with the means to analyze customer conversion rates. Companies can infer user preferences by providing backoffice data from websites or apps, so as to find ways to improve user conversion rates. During the project, the dataset is normalized so that additional analysis parameters can be added more easily. In this way, the project can be applied to more situations.

The main goal of this project is to build an appropriate model to assist the company in analyzing its own customer conversion rate. The main task is to search relevant data from the Internet and social media, standardize the data, build a model and complete the analysis.

2 Data

The dataset is from Github, provided by Piyushkumar Jain's project. This dataset counts the visit information and conversion results of a site. The dataset contains the following data columns.

- country : user country based on the IP address
- age : user age. Self-reported at sign-in step
- new_user : whether the user created the account during this session or had already an account and simply came back to the site
- source : marketing channel source
  - Ads: came to the site by clicking on an advertisement
  - Seo: came to the site by clicking on search results
  - Direct: came to the site by directly typing the URL on the browser
- total_pages_visited: number of total pages visited during the session.
- converted: 1 means they converted within the session, 0 means they left without buying anything.

3 Data Analysis

3.1 Country

For viewers from different countries, I counted the number of people who were converted and those who weren't. In addition to that, I calculated the conversion rate for each country separately. Here is the resulting graph.

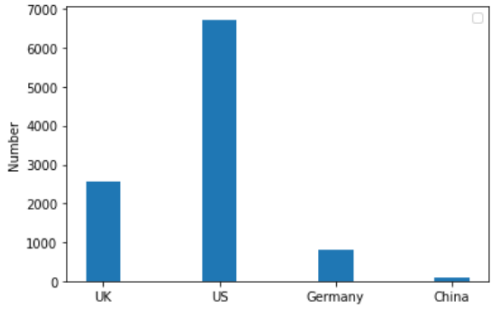

Figure 1: Converted=1 for Country

For those who finally converted, 2550 comes from the U.K., 6732 comes from the U.S, 816 comes from Germany, and 102 comes from China.

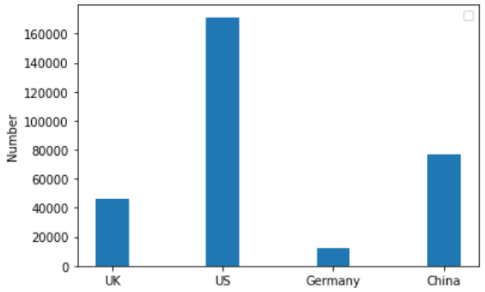

Figure 2: Converted=0 for Country

For those who not converted, 45900 comes from the U.K., 171358 comes from the U.S, 12240 comes from Germany, and 76499 comes from China.

Those who comes from the U.K., 5.26% converted. Those who comes from the U.S., 2.86% converted. Those who comes from Germany, 6.25% converted. Those who comes from China, 0.13% converted.

3.2 Age

For age, I also calculated the average age of the converted and non-converted populations. Here is the resulting graph.

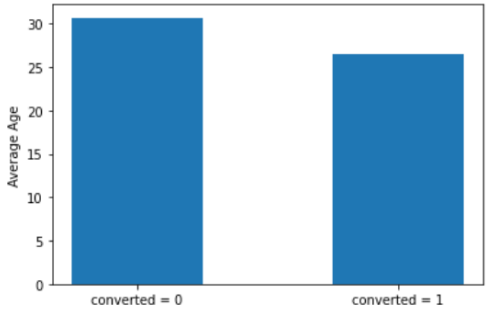

Figure 3: Average Ages

For those who didn't convert, their average age is 30.70 years old. For those who converted, their average age is 26.55 years old. The average age of the converted is 13.52% lower than that of the non-converted.

3.3 Source

For viewers from different sources, I do the same job as for country. Here is the resulting graph.

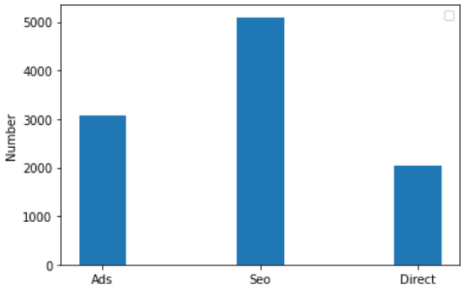

Figure 4: Converted=1 for Source

For those who finally converted, 3068 comes from advertisement, 5095 comes from Search Engine Optimization, and 2037 comes directly by themselves.

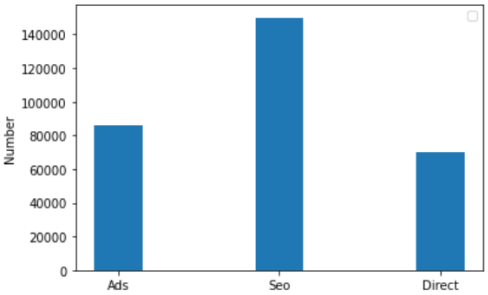

Figure 5: Converted=0 for Source

For those who finally not converted, 85678 comes from advertisement, 149936 comes from Search Engine Optimization, and 70383 comes directly by themselves.

Those who comes from advertisement, 3.46% converted. Those who comes from Search Engine Optimization, 3.29% converted. Those who comes directly, 2.81% converted.

3.4 Total Page Visited

For this part, I do the same job as Age. Here is the resulting graph.

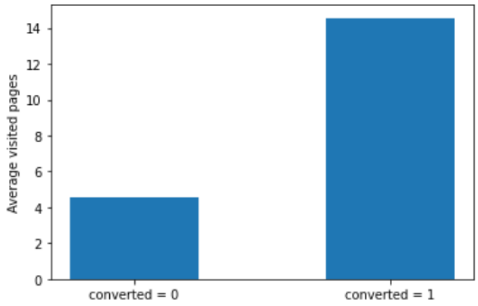

Figure 6: Average Pages

For those who didn't convert, they read an average of 4.55 pages. For those who converted, they read an average of 14.55 pages. The average pages visited of the converted is 219.78% higher than that of the non-converted.

4 Prediction

In the model prediction part, I tried a variety of different learning models. In addition to the first time, I applied ensemble learning to try to further improve the accuracy. Because the complexity of the dataset is relatively low, the various models perform relatively well. Here is the specific accuracy table and the confusion matrix corresponding to each model.

Model	Accuracy
SVC	0.98495677
LinearSVC	0.96097406
Naive Bayes	0.97556398
Logistic Regression	0.98504111
Ensemble Learning	0.98586190

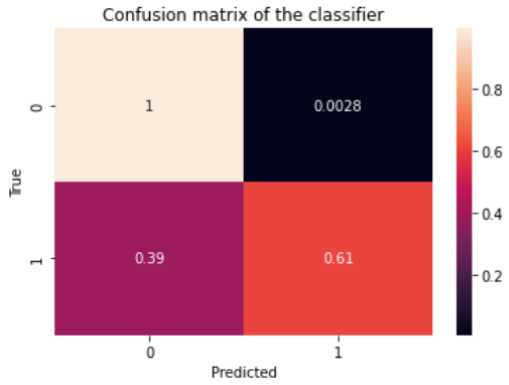

Figure 7: Confusion Matrix for SVC

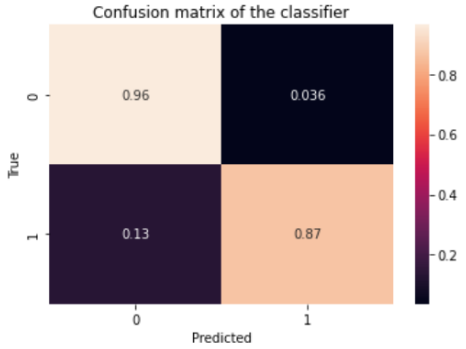

Figure 8: Confusion Matrix for LinearSVC

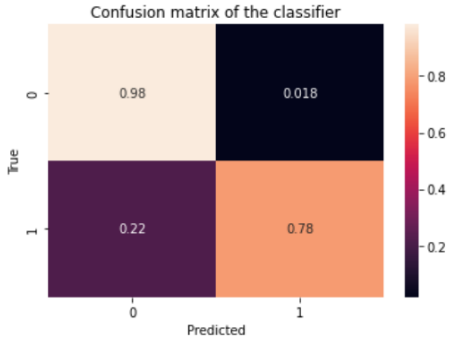

Figure 9: Confusion Matrix for Naive Bayes

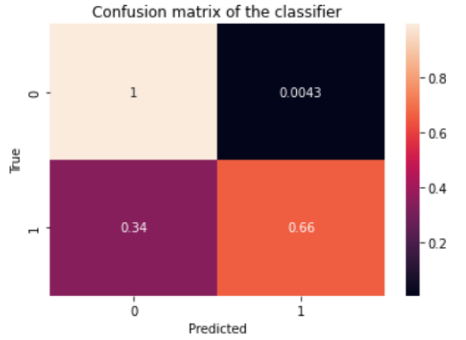

Figure 10: Confusion Matrix for Logistic Regression

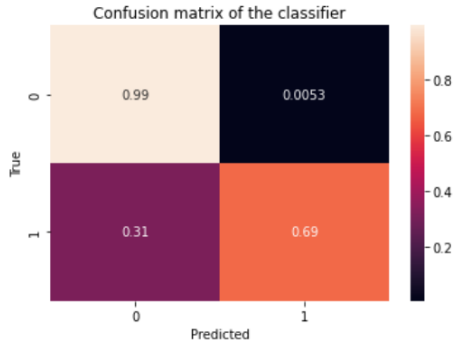

Figure 11: Confusion Matrix for Ensemble Learning

In ensemble learning, I use Votingclassifier, and the ensemble is soft voting. It can be found that the accuracy of ensemble learning is the highest among them, proving its effectiveness. By looking at the confusion matrix, it can be found that the model is significantly less accurate when predicting the category "converted = 1" than when predicting "converted = 0". This may be because the amount of data "converted = 1" is relatively small, resulting in poor prediction results. This result is also related to the model having fewer features.

5 Conclusion

1. Most of the people who are finally converted have viewed more pages, doing a good job in web page optimization will help to improve conversion rate.
2. This site is relatively more popular with young people, adding popular elements is a good way to attract more people.
3. Advertising brings the best conversion rate, increasing the investment in advertising is expected to have a good result.
4. The conversion rate of visitors from Germany and the UK is significantly high. We can study the reasons for this and apply it to improve the conversion rate of visitors in other countries. This needs some more data about these two countries' visitors.
5. Regarding the prediction of conversion rate, ensemble learning has the best performance, but part of the prediction is still not good. More data and features might help this.