Background Introduction

For many companies, especially Internet companies, they will attract others to log in to their website homepage and register members by advertising or occupying search engines. The ratio of actual signups 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.

This program can provide companies with the means to analyze customer conversion rates. Companies can infer user preferences by providing back-office 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.

1 Motivation

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 Analysis

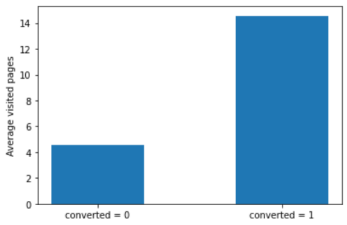

Figure 1: Average pages

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.

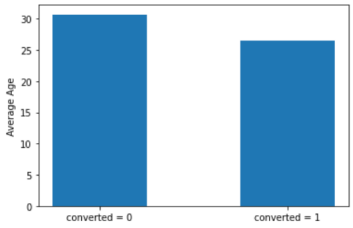

Figure 2: Average Age

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. For those who finally converted, 2550 comes from the

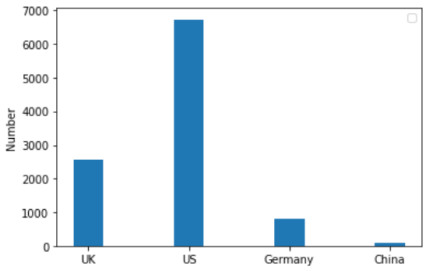

Figure 3: Converted visitors

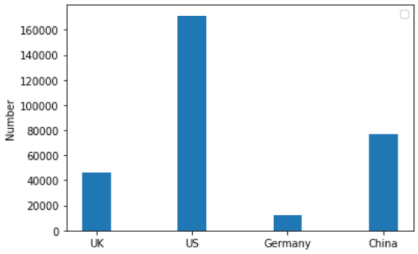

Figure 4: Unconverted visitors

U.K., 6732 comes from the U.S, 816 comes from Germany, and 102 comes from China. 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. For those who finally converted, 3068 comes from advertisement,

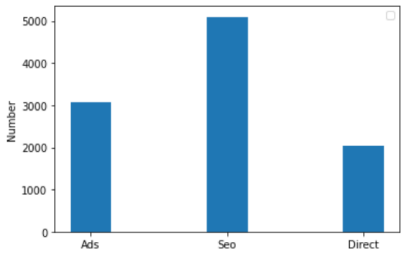

Figure 5: Converted visitors

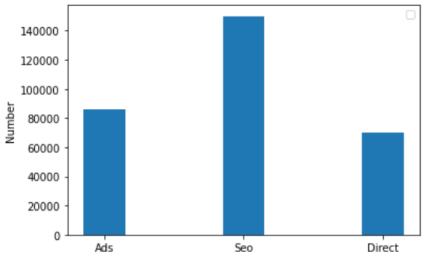

Figure 6: Unconverted visitors

5095 comes from Search Engine Optimization, and 2037 comes directly by themselves. 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 Prediction

I picked several different models to train and test, and selected the best performing model. Among the four models, logistic regression has the highest accuracy, but LinearSVC has

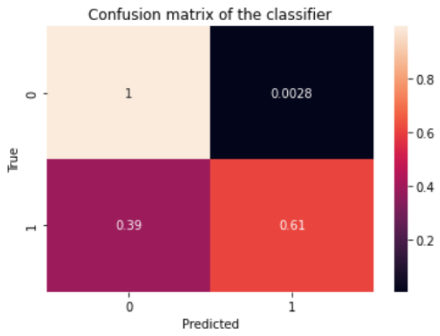

Figure 7: SVC

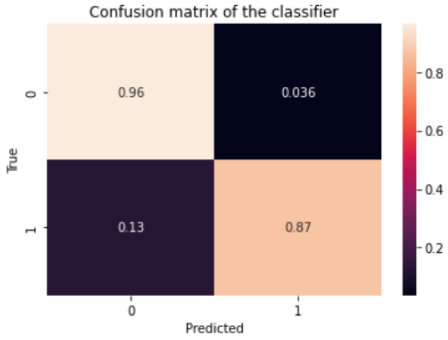

Figure 8: LinearSVC

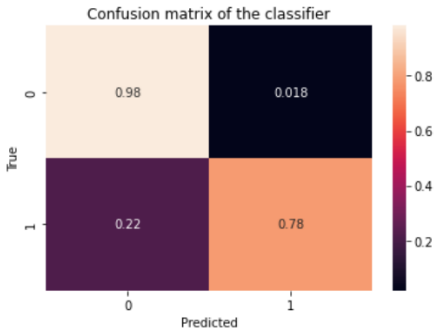

Figure 9: Naive Bayes

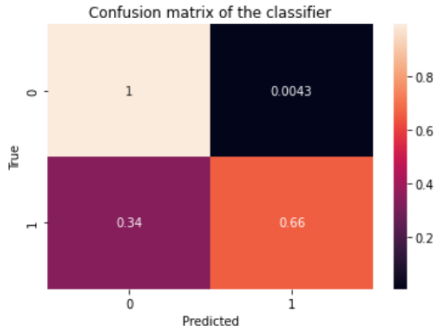

Figure 10: Logistic Regression

better accuracy when predicting data with the label "Converted=1". It is reasonable to predict that the data with the label "Converted=1" is more meaningful and valuable to the project, so LinearSVC will be given priority.

4 Ensemble Learning

I did further work and tried random forest and votingclassifier for ensemble learning, where votingclassifier has good results, here is the confusion matrix.

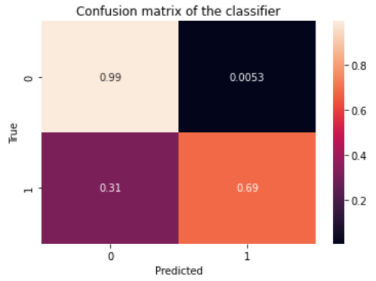

Figure 11: Average pages

The final accuracy rate is slightly higher than the previous four models.

5 Conclusion

1.Most of the people who are finally converted have viewed more pages, and doing a good job in web page optimization will help this.

2. This site is relatively more popular with young people, adding popular elements can help increase conversion rates.

3. Advertising brings the best conversion rate, and increasing the investment in advertising is expected to have good results.