The Comparisons of Machine Learning Approaches-- An Example of Online Shoppers Purchasing Intention

Ge Wu , Difei Chen

Abstract **-- This project is focused on online shoppers' purchasing intention. The data is provided on UC Irvine Machine Learning Repository, including various kinds of information about consumers' online shopping behavior and the final purchasing results. The sample size is 12330 with 17 features and a binary output label. After analyzing these 17 features, we decide to find an optimal model to solve this classification problem. In this study, we use the logistic regression model, Ridge model, neural network (FNN), Kernel SVM, and Random forest to model the data. In the evaluation process, we use the confusion matrix, ROC, AUC area, and accuracy rate to determine which model outperforms. Overall, FNN and Random forest outperform**

Index Terms **-- Machine learning, Classification, Probability, Data mining**

I. INTRODUCTION

In this digital age, E-commerce has become more and more popular. It makes retailing and shopping more efficient by showing that the business can run 24/7. Through collecting the various information and comparing the price, consumers can buy anything they want online. At the same time, the selling party can also utilize the information, such as the usage of advertisement, the amount of time consumers spent in the searching of such products, consumers' personal information, to analysis whether the consumer will buy their products. Besides, they can also find which factor has the strongest correlation with consumers' purchasing behavior and the potential customers.

Our group try to conduct this risk prediction through five different models. Then, we evaluate these models by confusion matrix, ROC, AUC area, and accuracy rate. Finally, FNN and Random forest outperform Basic logistic regression

Below is a brief introduction of the five models.

1. Basic Logistic Regression

Logistic Regression is commonly used for classification problems. It finds the relationship between input features and shows the predictions of binary data which is either zero or one.

2. Logistic Regression with Ridge Regularization

Ridge regression is a penalized version of LR which it maximized the likelihood function with a squared magnitude. It causes most of the theta to be zero, and only a small portion of non-zeros. Ridge regression is suitable for overfitting problem.

3. Feedforward Neural Network (FNN)

Feedforward neural network is commonly used for classification problems. It consists of a lot of neurons which located in multiple layers. Each layer's input is previous layer's output, and it uses different weights in each connection. They data always enters into the input layer and produce prediction output in the output layer. All the layers in between input layer and output layer are called hidden layers.

4. Kernel SVM

SVM is commonly used for classification problems. It transforms the training samples, so that a non-linear decision surface is able to project the data into a higher number of dimensions without increasing the complexity of the cost function. Kernel SVM can also lead to overfitting.

5. Random Forest

Random forest is an ensemble learning method for classification. It builds multiple decision trees and uses bagging and bootstrapping to produce a more accurate prediction.

II. TASK DESCRIPTION

We use the online shoppers purchasing intention data to predict whether the customer buys it or not. First, we do the data exploration to visualize each feature. Secondly, we do the data transformation and normalization. Then, we fit 5 different models. For each model, we calculate the confusion matrix, accuracy, recall and precision rate, ROC Curve and AUC area to evaluate the models.

III. MAJOR CHALLENGES AND SOLUTIONS

1. High AUC area but low accuracy

In the FNN model, we had a very high AUC area(0.92), which was better than other models'. But the accuracy of this FNN model was very low(0.15), which seemed to conflict with the AUC evaluation result.

Then we analyze the confusion matric and found that the number of false-positive predictions is very large (2827), which indicated that this model classified some negative labels as true labels. Besides, we also reviewed the definition and features of ROC and the accuracy rate. We realized that the

accuracy rate uses a fixed threshold, while the ROC Curve is changed with the unfixed thresholds.

threshold = 0.5	predicted purchase_ture	predicted purchase_false
purchase_ture	492	0
purchase_false	2827	11

	precision	recall	f1-score	support
purchase_false	1.00	0.00	0.01	2838
purchase_true	0.15	1.00	0.26	492
accuracy			0.15	3330
macro avg	0.57	0.50	0.13	3330
weighted avg	0.87	0.15	0.04	3330

Therefore, we found the problem was with thresholds so that we changed the threshold from 0.5 to 0.65. This means the probability from the sigmoid function should be larger than 0.65 to be recognized as positive.

Finally, with a threshold = 0.65, the accuracy rate was changed to 0.88, which was a comparatively high level in these five models.

threshold = 0.65	predicted purchase_ture	predicted purchase_false
purchase_ture	150	360
purchase_false	38	2782

	precision	recall	f1-score	support
purchase_false	0.89	0.99	0.93	2820
purchase_true	0.80	0.29	0.43	510
accuracy			0.88	3330
macro avg	0.84	0.64	0.68	3330
weighted avg	0.87	0.88	0.86	3330

2. Data Normalization

At first, our model cannot fit the data well. Then we found that our feature values were in different ranges. Some binary features were just between 0 and 1. But some features had very large values, such as 10000. If the features were not on the same scale and the weights of each feature were different, this led to a bad influence on fitting the models. Therefore, we used Max-Min scaling to transform all features in the [-1,1] range. Finally, our model did a better job than before.

IV. EXPERIMENTS

1. Dataset Description

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. Among the 12330 data points, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. There are in total of 17 variables as independent variables.

Y: Revenue: A boolean representing whether or not the user completed the purchase. If the user didn't complete the purchase, the dependent variable equals to zero. If the user completes the purchase, the dependent variable equals one.

X1) Administrative: the number of pages of this type (administrative) that the user visited.

X2) Administrative_Duration: the amount of time spent in this category of pages.

X3) Informational: the number of pages of this type (informational) that the user visited.

X4) Informational_Duration: the amount of time spent in this category of pages.

X5) ProductRelated: the number of pages of this type (product related) that the user visited.

X6) ProductRelated_Duration: the amount of time spent in this category of pages.

X7) BounceRates: The percentage of visitors who enter the website through that page and exit without triggering any additional tasks.

X8) ExitRates: The percentage of pageviews on the website that end at that specific page.

X9) PageValues: The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction.

X10) SpecialDay: This value represents the closeness of the browsing date to special days or holidays (eg Mother's Day or Valentine's day) in which the transaction is more likely to be finalized.

X11) Month: Contains the month the pageview occurred, in string form.

X12) OperatingSystems: An integer value representing the operating system that the user was on when viewing the page.

X13) Browser: An integer value representing the browser that the user was using to view the page.

X14) Region: An integer value representing which region the user is located in.

X15) TrafficType: An integer value representing what type of traffic the user is categorized into.

X16) VisitorType: A string representing whether a visitor is New Visitor, Returning Visitor, or Other.

X17) Weekend: A boolean representing whether the session is on a weekend.

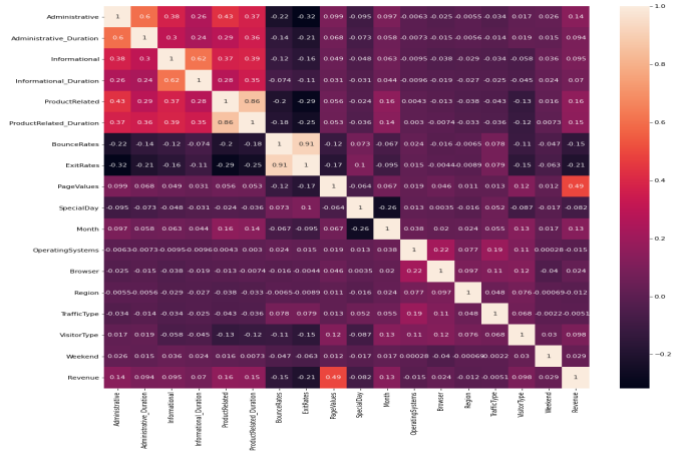
We randomly select 9000 of the data into training dataset and the rest as testing dataset.

2. Data Exploration

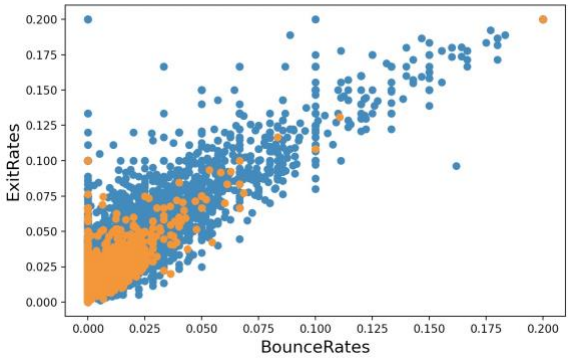
After we download the data from UCI website, we checked the shape (12330, 18), info, description of the data. We then checked for any missing values which is zero. In order to use the heatmap, we reassigned numbers to features: month, visitor types, weekend and revenue.

From the heat map, we can see that there isn't much correlation between features. Bounce rate and Exiterate has a correlation as 0.91. Productrelated and prodcuct related duration has a correlation as 0.86. Informational and informational_duration has a correlation as 0.62.Administraive and administrative_duration has a correlation as 0.6.

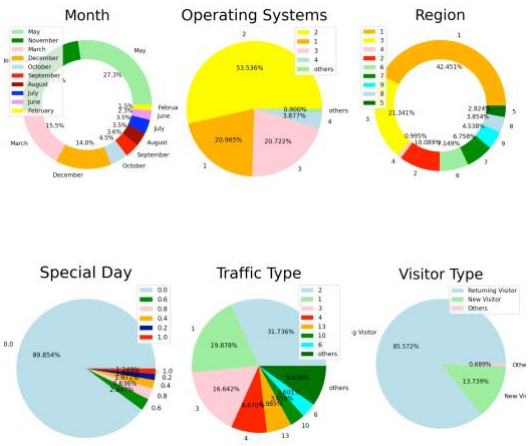
However, these three pairs of features are closely related since people who spent more time on these categories definitely viewed more pages of these categories. PageValues has a 0.49 correlation with revenue and we can see that from the correlation with revenue graph that PageValues has the highest correlation with Revenue. From the histogram, we observe that the data distribution is highly imbalanced.



We plot the BounceRates against ExitRates. The orange dots represents revenue=1 and the blue dots represents revenue = 0. As we can see from the figure above, the orange dots are the most concentrated when both rates are low which shows that consumer will purchase the product when Bounce Rates and Exit Rates are low.



From the plots of Administrative_Duration vs Revenue, Informational_Duration vs Revenue and ProductRelated_Duration vs Revenue, we find that the distributions of data are bell-shaped for both purchased and not purchased, however, there are more outliers in the not purchased distribution. From the plot of Page value vs Revenue, we find there are many outliers for both purchased-true part and purchased-false part, however, we find that the page value affects the purchased-true part greatly.



From the above charts, we get the following observations:

1) 89.85% of the user was browsing on the days that were not close to any special days.
2) March, May, November, December are the months of the year that users view the page more often.
3) 99% of the users are using the top four operating systems and the top operating system has a high share of 53.5%.
4) The company can focus more on users from region 1, 3 and 2 since 70% of the users come from these three regions.
5) The company should focus more on users who use traffic type 2, 1, and 3, since 69% of the users come from these three traffics.
6) Returning Visitor makes up 85% of the total visitors. The company should put more focus on returning visitors to increase revenue.

3. Evaluation metrics

1) Confusion Matrix (Accuracy Rate, Precision and Recall Rate)

Confusion Matrix		
	Y_prediction = 1	Y_prediction = 0
Y = 1	True-Positive (TP)	False-Negative (FN)
Y = 0	False-Positive (FP)	True- Negative (TN)

Accuracy Rate, Precision and Recall Rate		
Metrics	Equation	Evaluation Emphasis
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Precision	$\frac{TP}{TP+FP}$ & $\frac{TN}{FN+TN}$	Precision metric measures the fraction of positive/negative patterns that are correctly classified.
Recall	$\frac{TP}{TP+FN}$ & $\frac{TN}{FP+TN}$	Recall metric is used to measure the fraction of positive/negative patterns that are correctly classified.

2) AUC Area

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC from (0,0) to (1,1).

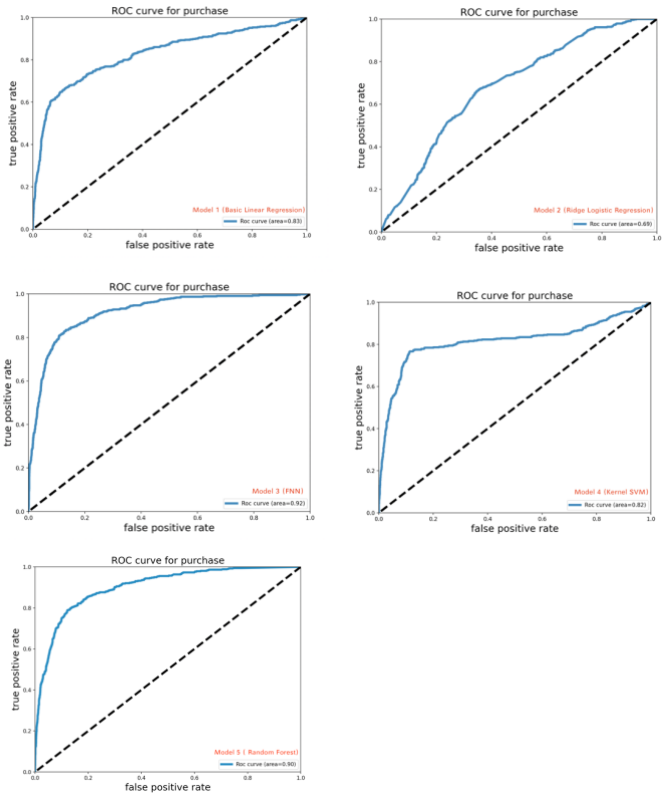
4. Major Results

1) Rates

	Accuracy	Precision_0	Precision_1	Recall_0	Recall_1
Model1 (Basic LR)	0.87	0.89	0.65	0.97	0.35
Model2 (Ridge LR)	0.88	0.95	0.59	0.91	0.72
Model3 (FNN)	0.88	0.89	0.80	0.99	0.29
Model4 (SVM)	0.85	0.85	0.90	1.00	0.02
Model5(Random Forest)	0.89	0.92	0.69	0.96	0.52

2) AUC Area and ROC Curve

	AUC Area
Model1 (Basic LR)	0.83
Model2 (Ridge LR)	0.69
Model3 (FNN)	0.92
Model4 (SVM)	0.82
Model5 (Random Forest)	0.90



5. Analysis

From the rate table, we can find that the Random Forest Model has the highest accuracy rate, which is equal to 0.89. Besides, the Feed Forward Neural Network and the Logistic Regression Model with Ridge Regularization have the second-highest accuracy, which is equal to 0.88. The regularization method is used to solve the overfitting problem. From the result, we can find the Logistic Regression Model with Ridge Regularization has a better accuracy rate than the Basic Logistic Regression Model.

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. For the negative output label, we can find that the Logistic Regression Model with Ridge Regularization has the highest precision rate and the SVM Model has the best recall rate. While for the positive output label, the SVM Model has the highest precision rate and the Logistic Regression Model with Ridge Regularization has the best recall rate.

From the AUC Area table and ROC Curve, we can conclude that the Feed Forward Neural Network Model has the largest AUC area, which is equal to 0.92. Besides, the AUC Area of the Random Forest Model is 0.9, which is the second-highest level. The high level of AUC Area means that the model can predict the output accurately with an appropriate threshold. Therefore, we set the threshold of the Feed Forward Neural Network Model to be 0.65, which means the output from the sigmoid function should be larger than 0.65 to be recognized as a positive label. After adjusting the threshold, the accuracy of the Feed Forward Neural Network Model has been improved. Besides, from the ROC Curve, we can find the Basic Logistic Model, FNN Model, SVM Model, and Random Forest Model have the same feature, which is that the true positive rate increases drastically when the false positive rate is at a very low level (between 0 and 0.1). However, when the false positive rate is larger than 0.1, the true positive rate increases slowly. This feature shows that when the false-positive risk is very high or handing false-positive risk is very expensive, it is a better choice to use these four models. What's more, as regards the Logistic Regression Model with Ridge Regularization, the true positive rate increases slowly with the changed false positive rate.

Overall, the Feed Forward Neural Network model and the Random Forest Model both have high values in accuracy rate and AUC Area. Therefore, we think these two models are better.

V. CONCLUSION AND FUTURE WORK

1. Conclusion

In this project, we analyze the online shoppers' purchasing intention and try to fit five different models to make a prediction. Through comparing and evaluating these models, we find that the Feed Forward Neural Network Model has the largest AUC Area and the Random Forest Model has the second-highest. Besides, in the accuracy rate aspects, the Random Forest Model has the highest score, and the Feed Forward Neural Network Model is the second-highest. Therefore, we can conclude that the Feed Forward Neural Network Model and the Random Forest Model are better than the other three models.

2. Future Work

First, we find some models, such as the Ridge, have a high AUC Area but are not good at accuracy rate. This indicates that we need to find the appropriate threshold to make the predictions more accurate. We can further divide the training samples into the training part and validation part to help us find the best threshold.

Second, some models can be improved. For example, in the Kernel SVM model, we only use the RBF kernel. We can also use the validation data to select the best C value and kernel (linear, poly, or RBF).

Third, it is necessary to update the data because this data only contains one-year period information. Besides, 80% of the outputs are negative, showing that this is an unbalanced data set. If we can keep paying attention to this online shopper purchasing intention problem and update this dataset to be more balanced while at the same time increase the sample size, we can fit better models.

REFERENCES

[1] The data set is provided on UC Irvine Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[2] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks