**Homework 11 - Data Competition**

**Predicting Customer Acceptance of Coupons**

Due December 5th, 11:59 PM

IST 5520, Fall 2022, Chen

***Instruction***:

The objective of this data competition is for students to apply various predictive modeling techniques to accurately predict whether a customer will accept coupons for local businesses recommended by an in-vehicle recommender system. Students work in a pair of two students to finish the data competition. If you want to find a partner, you can submit a post on our discussion board. Students are also allowed to finish the work individually. Please work as early as possible so that you have sufficient time to refine your final model by the due time.

Student pairs should work individually to finish the data competition. Any form of collaboration or help from other pairs/persons/parties is NOT allowed. Please submit your work by the due time.

***Data:***

This original data was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios including the destination, current time, weather, passenger, etc., and then asks the subject whether he will accept the coupon if he is the driver. The original data contains the following columns:

* destination: No Urgent Place, Home, Work
* passenger: Alone, Friend(s), Kid(s), Partner (who are the passengers in the car)
* weather: Sunny, Rainy, Snowy
* temperature:55, 80, 30
* time: 2PM, 10AM, 6PM, 7AM, 10PM
* coupon: Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50)
* expiration: 1d, 2h (the coupon expires in 1 day or in 2 hours)
* gender: Female, Male
* age: 21, 46, 26, 31, 41, 50plus, 36, below21
* maritalStatus: Unmarried partner, Single, Married partner, Divorced, Widowed
* has\_Children:1, 0
* education: Some college - no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School
* occupation: Unemployed, Architecture & Engineering, Student, Education&Training&Library, Healthcare Support, Healthcare Practitioners & Technical, Sales & Related, Management, Arts Design Entertainment Sports & Media, Computer & Mathematical, Life Physical Social Science, Personal Care & Service, Community & Social Services, Office & Administrative Support, Construction & Extraction, Legal, Retired, Installation Maintenance & Repair, Transportation & Material Moving, Business & Financial, Protective Service, Food Preparation & Serving Related, Production Occupations, Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry
* income: $37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 - $87499, $50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999, Less than $12500
* Bar: never, less1, 1~3, gt8, nan4~8 (feature meaning: how many times do you go to a bar every month?)
* CoffeeHouse: never, less1, 4~8, 1~3, gt8, nan (feature meaning: how many times do you go to a coffeehouse every month?)
* CarryAway:n4~8, 1~3, gt8, less1, never (feature meaning: how many times do you get take-away food every month?)
* RestaurantLessThan20: 4~8, 1~3, less1, gt8, never (feature meaning: how many times do you go to a restaurant with an average expense per person of less than $20 every month?)
* Restaurant20To50: 1~3, less1, never, gt8, 4~8, nan (feature meaning: how many times do you go to a restaurant with average expense per person of $20 - $50 every month?)
* toCoupon\_GEQ15min:0,1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 15 minutes)
* toCoupon\_GEQ25min:0, 1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 25 minutes)
* direction\_same:0, 1 (feature meaning: whether the restaurant/bar is in the same direction as your current destination)
* direction\_opp:1, 0 (feature meaning: whether the restaurant/bar is in the same direction as your current destination)
* accept:1, 0 (whether the coupon is accepted)

In the current dataset, all categorical variables have been converted into dummies. The “train\_data.csv” contains 3426 observations that are used to train classification models. The “test\_data.csv” contains a dataset of 8456 observations that are used to evaluate the performance of classification models. The test dataset has the same data structure as the training dataset.

***Task:***

Apply various predictive modeling techniques to build a classification model that performs very well for predicting if a person will accept the coupon. Below is a list of possible techniques that could help you tune your models:

1. Dimension reduction or feature selection
   1. Use PCA to reduce the dimensionality
   2. Use Random Forests to select features
   3. Use other methods to select features (for more details, refer to <https://scikit-learn.org/stable/modules/feature_selection.html>)
2. Tune hyperparameters. To increase the speed of hyperparameter tuning, you may parallelize the model training process. For more detail, refer to <https://scikit-learn.org/stable/computing/parallelism.html>.
3. Deal with imbalanced dataset

For more details, refer to <https://imbalanced-learn.org/stable/introduction.html>.

1. Assign class weights
2. Use regularization parameters such as C in SVMs, and alpha in neural networks. For the effect of regularization parameters on model performance, refer to an example shown at <https://scikit-learn.org/stable/auto_examples/neural_networks/plot_mlp_alpha.html>.
3. Use advanced models such as ensemble methods (for example, random forests and gradient tree boosting) and deep learning. For more details, refer to <https://scikit-learn.org/stable/supervised_learning.html>.
4. Other methods……

You can only use the training dataset to train the classification models without touching the test dataset. Using the test dataset as a validation set to refine your models is not allowed. You can normalize/transform the variables in the test dataset, but any resampling applied to the test dataset is not allowed. The test dataset is only used to evaluate the performance of your model. Your models need to be evaluated by AUC score [area under the ROC curve, using the method sklearn.metrics.roc\_auc\_score() to calculate].

Make sure that your jupyter notebook can be reproducible. Re-running the jupyter notebook will lead to the same results.

***Submit the Following Documents:***

1. Submit your jupyter notebook with detailed explanations. Your jupyter notebook should at least include the following contents:

* Specify your pair information in the jupyter notebook. Each pair only needs one submission.
* Specify the contribution by each member. Students without significant contributions to the work will not receive the full regular points and extra credit.
* Data transformation and/or preprocessing if there are any.
* The final model trained from the training dataset.

***Note:*** I know you need to try multiple models and try many parameters. Please only submit the final model with the best parameters you find from your parameter tuning. Keep those time-consuming model-tuning processes in another jupyter notebook and submit it to the homework as a supplemental document. This ensures that your jupyter notebook with the final model can be re-run by the instructor in a short time period. If your final model needs a long time to run, the instructor will not be able to verify your results and thus cannot grant full points to your submission.

* Test the performance of your final model using the test dataset. At least AUC score needs to be reported.

***Note:*** The jupyter notebook should display the results in consecutive sequence starting from 1. Make sure you click the “Kernel -> Restart & Run All” menu to re-run your jupyter notebook after you finish your code. A jupyter notebook that does not show the results in consecutive sequence will receive a zero point.

1. Submit an HTML report generated from your jupyter notebook for your final model. On jupyter notebook, click “File -> Download as -> HTML” to generate the HTML report.
2. If you have used hyperparameter tuning, please submit your jupyter notebook containing the hyperparameter tuning process as a supplemental document.
3. Submit a one-page Microsoft Word document that explains the logic of your predictive modeling.

***Grading:***

***Regular Points***

Submissions will be graded based on the AUC score of the final model. Correct work submitted by the due time will earn 40 points as regular homework. Errors in the submission will reduce the points.

***Extra Credit***

Students who finish the work by the due time can earn extra credit according to the following rule:

* Submissions ranked as top 4 in terms of the AUC score of the final model will receive additional 2% extra credit for the final grade;
* Submissions ranked 5th to 8th best models in terms of the AUC score of the final model will receive additional 1% extra credit for the final grade.