**IE 7374 - Final Project Description**

Group 8

Chushi Chen, Ziwei Duan, Binghui Lai and Weikun Zhuang

**Introduction & Problem Definition**

The increased e-commerce usage has generated great potential in the market over the past few years. It is crucial for companies to devise smart marketing and promotion strategies to retain as many shoppers as possible in a shopping session on their e-commerce website. Creating and presenting customized promotions to online shoppers is key to a successful business.

Our team is interested in predicting online shoppers’ purchasing intentions. What will influence a shopper’s decision to make purchases and on the contrary, what made shopping cart abandonment happen? Are there any other interesting facts to be discovered? In order to answer these questions, we will create several classification models to predict if a browsing session will convert to a purchase. We hope that insights from implementing these models will help us to understand a little more about online shoppers’ behavior and potentially shed some light on increasing profits for businesses and improving e-commerce site quality.

**Data Source and Description**

The dataset we’ll be using in this project comes from UCI machine learning repository. The original source[[1]](#footnote-0) of this dataset is a paper[[2]](#footnote-1) about prediction of online shoppers’ purchasing intention. Both sources are cited below.

This dataset contains 12330 observations and 18 attributes (including one target variable). Each observation is a session that belongs to a unique user who browses an e-commerce site in a 1-year period. Among these 17 predictors, there are 10 numerical and 7 categorical variables. Numerical variables include information about the *number of different types of pages visited and total time spent in each of the categories* (‘‘Administrative’’, ‘‘Administrative Duration’’, ‘‘Informational’’, ‘‘Informational Duration,’’ ‘‘Product Related,’’ and ‘‘Product Related Duration”), some *analytical metrics for each page in the site measured by “Google Analytics”* (‘‘Bounce Rate,’’ ‘‘Exit Rate,’’ and ‘‘Page Value’’), and *how close is the site visiting time to a special day* (e.g. Mother’s Day, Valentine’s Day). For categorical variables, we have “Month”, “Operating systems”, “Browser”, “Region”, “TrafficType” , “Visitor Type”, and if it is “Weekend”. Lastly, the target variable “Revenue” is a binary categorical variable with 84.5% (10,422) negative and 15.5% (1908) positive that ends up shopping.

**Project Roadmap**

Data Preprocessing and Visualization

In the preprocessing stages, we will first focus on exploring if the dataset has any missing values and NULL values. If so, we will take the necessary steps (e.g. imputing using median, dropping columns with too many NULL values). Next, we will explore each numerical variable and inspect if there are any outliers using the IQR values. For categorical attributes, we will create dummy variables using one-hot encoding if necessary. We will also encode the target variable to 0/1 binary variable. For exploratory analysis, we will create bar plots and boxplots to see distributions of numerical variables and create stacked bar plots for categorical variables. Correlation between variables will also be explored. Highly correlated columns will be removed.

Machine Learning Algorithms

We plan to implement three models for this classification project,

* A logistics regression model as baseline. Linear Regression is robust and widely-used in the industry. We will run this model with multiple different settings: using closed-form solution/gradient descent, with/without regularization term, and tune hyperparameters to find the optimized combination.
* A gaussian discriminant analysis. We will use a generative model as opposed to logistic regression - a discriminative model. The GDA will perform very well if each variable is normally distributed, and it’s computationally more efficient than logistic regression.
* A neural network model using tensorflow. We plan to implement a feed-forward neural network to capture possible non-linearity within the dataset.

Performance Evaluation

After applying our models, we will pick out the optimal one by analyzing performance evaluation metrics such as sensitivity/specificity, precision/recall, F-1, and AUROC. We will also conduct sensitivity analysis by exploring propensity threshold values on success class (We are more interested in the “purchase” behavior and thus “Revenue” =1 is the success class) and how they affect model performance. Furthermore, costs and revenues associated with shoppers’ behavior could be explored by plotting lift and gain charts.

Reflections and Conclusions

Last but not the least, we will reflect on the whole model fitting process and summarize insights from the results to hopefully answer the questions defined previously.

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. URL to dataset: <https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset> , accessed Feb. 28th, 2022 [↑](#footnote-ref-0)
2. Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018) [↑](#footnote-ref-1)