**Online Shopping Research Proposal**

**Research direction**

Our research aims to find which factors contribute most to shopping behavior. Nowadays, online shopping become one of the most important parts of the business world. It is necessary to grasp the user’s buying habits and predict their shopping intention. Using these analysis results and prediction to improve shopping websites and applications can improve the sales and profit of these companies.

**Rational**

These data on users’ activities and shopping behavior have some relationships. And these relationships can be used to build a prediction model to predict users’ intentions. In that case, online shopping websites can improve their design and push more exact commodities and repo to get higher sales and profit. Meanwhile, customers will have a better shopping experience. Moreover, this analysis is a good reference for a merchant to improve their service and product. Shopping websites and applications also can refer to this model for their marketing strategy. Therefore, this research is meaningful and beneficial to many parties.

**Dataset**

The dataset is about customers’ scanning time, exiting rate, browsing pages, and other information which is relevant to their shopping intention. The dataset is highly consistent with our study. It includes almost every element we’d like to analyze.

**Data collection**

We collected this dataset from UCI Machine Learning Repository.

The download link is below.

<http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

**Data processing**

The dataset has 12330 observations without null values. So, we don’t need to do data cleaning.

But there are 18 variables, which means too many independent variables. We don’t need to use all of them. Especially, some variables may have multicollinearity with the dependent variable. So, we need to pick up several most relevant variables used in the model. The specific variables will be decided when modeling.

It is just a simple selection. Tools are not necessary. Just need to pick up the used columns out of the data frame.

**Model**

We would like to use 2 models to analyze the dataset.

**(1) Kaplan-Meier analysis**

Kaplan-Meier analysis can be used to calculate the probability of an event happening over time.

For this dataset, time variables could be the time users spent from the first login to their leave. The event is the purchasing action. We can also group by other factors to see whether different groups have a significant difference.

**(2) Cox Regression**

Cox Regression can help us to find the predictive variables and the event. There are lots of variables in the dataset, we can use the method to find the most relevant factors to purchasing.

The dataset has the property of time. It also includes event variables. It works for the requirements of Cox Regression.

**Expected outcomes**

For now, we assume that product related page may be the most relevant factor. And there is doubt that whether administration duration has a significant impact on the probability of an event occurring. But these need to be proven in the analysis process.

We kind of worry there may be a limitation of this dataset. It does provide the duration time and customers’ browsing time for us. But all these variables are generated from the website angle. We don’t get the information on customers’ profiles. What if the most important factor is the characteristics of customers rather than the website’s design? This is an unknown question to us.