**CSP 571 - Data Preparation and Analysis**

**Spring 2023**

**Project - Proposal & Outline**

**Project title: Finding the pattern behind the online shoppers purchasing intention.**

**Team members:**

Karthik Kumar Kaiplody (A20523668)

Naveen Raju Sreerama Raju Govinda Raju (A20516868)

**Professor:**

Jawahar Panchal

1. Project Proposal

1.1 Description:

Our goal is to perform exploratory data analysis to analyse trend in the data and build machine learning model to predict purchasing intentions of a visitor who visit a store’s website. We are planning to approach this as both clustering and classification problem.

1.2 Questions that the project seeks to address:

1. To find which subset of the 10 numerical and 8 categorical attributes of a data set has an strong impact to predict purchasing intention of customer.
2. To find which attributes are highly correlated so as to reduce number of predictor variables, to make model simpler.
3. To identify which attributes has an positive and negative impact on customer purchasing intention.
4. What are the distinct purchasing patterns among the online shoppers and how do these patterns differ based on demographic and behavioural factors?
5. Project Outline

2.1 Literature review

Few literature references has been found and mentioned at the end will be analyzed in depth to understand and solve our problem statement better.

2.2 Data Set

2.2.1 Data Source

The data set used that is being used in this project was obtained from the UC Irvine Machine Learning Repository.

Data set contributes:

1. C. Okan Sakar

Department of Computer Engineering, Faculty of

Engineering and Natural Sciences, Bahcesehir University,

34349 Besiktas, Istanbul, Turkey

2. Yomi Kastro

Inveon Information Technologies Consultancy and Trade,

34335 Istanbul, Turkey

2.2.2 Data set description

The dataset consists of feature vectors belonging to 12,330 sessions.The dataset consists of both numerical and categorical attributes.The 'Revenue' attribute can be used as the class label.

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| Administrative | Numerical | Page category |
| Administrative Duration | Numerical | Total time spent in this page |
| Informational | Numerical | Page category |
| Informational Duration | Numerical | Total time spent in this page |
| Product Related | Numerical | Page category |
| Product Related Duration | Numerical | Total time spent in this page |
| Bounce rate | Numerical | The bounce rate for a web page is the percentage of visitors who navigate away from the site after viewing only that page, without interacting with the page or visiting any other pages on the site. So, it's not just about the visitors who enter the site from that page, but rather about visitors who land on the page and then leave without taking any further action. |
| Exit rate | Numerical | The exit rate for a web page is the percentage of visitors who leave the site after viewing that page as the last page in their session. Unlike bounce rate, which only takes into account the visitors who leave after viewing a single page, the exit rate includes visitors who may have viewed multiple pages on the site before leaving after viewing the specific page in question.  To calculate the exit rate for a specific web page, you would divide the number of exits from that page by the total number of page views for that page. |
| Page value | Numerical | The Page Value feature is a metric in Google Analytics that represents the average value of a page that a user visited before completing an e-commerce transaction or a goal conversion on a website. It is calculated by dividing the total value of all transactions or goal completions by the number of unique page views for a particular page or set of pages. |
| Special day | Numerical | The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. |
| Operating system | Numerical |  |
| Browser | Numerical |  |
| Region | Numerical |  |
| Traffic type | Numerical |  |
| Visitor type | Categorical | New\_Visitor or Returning\_Visitor |
| Weekend | Categorical | True if it is either Saturday or Sunday. Or else it is False. |
| Month of the year | Categorical | Jan, Feb,…..,December |
| Revenue | Categorical | True is customer purchased anything or else it is False |

2.2.3 Data pre-processing

* Duplicate observations must be removed
* Missing values must be found and replaced with suitable values
* Conversion of categorical values to numerical values

2.3 Data set analysis

* Analyse co-relation of attributes of dataset
* Understanding each attributes data distribution
* Bi variate analysis of each predictor attribute with response attribute.
* Multi variate analysis between different predictor variable and response attribute
* Analysis of outliers

2.4 Model selection

* Here we try to approach the problem as classification as well as clustering task.

1. Classification

* We plan to use following machine learning models: Logistic Regression, Support Vector Machine, Random Forest Classifier, Naive Bayes, XGBoost Classifier
* For classification we try to focus on following metric such as Precision, Recall, F1 score, AUC curve to measure performance the models

1. Clustering

One of the question we are focused on to answer from this study is understanding the distinct purchasing patterns among the online shoppers using their demographics with which clusters they fall into.

Some of the popular machine learning algorithms we are planning to explore would be,

1. K-means clustering: This is widely used unsupervised learning algorithm that partitions a dataset into K-clusters

3 Software packages

* R Studio is an Integrated Development Environment (IDE) for the R programming language. It is a an open-source IDE that provides a user-friendly interface for data analysis, visualization, and statistical computing.
* CRAN (Comprehensive R Archive Network) host R packages, documentation, and other resources for the R programming language.

4 References resource

https://jurnal-ppi.kominfo.go.id/index.php/jppi/article/view/341

https://link.springer.com/article/10.1007/s00521-018-3523-0