CSP 571 - Data Preparation and Analysis
Spring 2023
Project - Proposal & Outline

Title: Finding the pattern behind the online shoppers purchasing intention.

Team members:

Karthik Kumar Kaiplody (A20523668)
Naveen Raju Sreerama Raju Govinda Raju (A20516868) (Team Leader)

Professor:

Jawahar Panchal

1. Project Proposal

1.1 Description:

Our objective is to analyze trends in the online shoppers purchasing intention dataset using exploratory data analysis techniques, and build machine learning models to predict the purchasing intentions of visitors to a store's website. We plan to approach this as both a clustering and classification problem, grouping similar customers based on their purchasing behavior and predicting whether a new customer is likely to make a purchase based on their browsing and purchasing behavior. By applying these techniques, we hope to gain insights into customer behavior and optimize marketing strategies to increase sales.

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?

2) Project Outline

2.1 Literature review

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

1. Real-time prediction of online shoppers purchasing intention using multilayer perceptron and LSTM recurrent neural networks:

This paper presents a study on predicting the purchasing intention of online shoppers using machine learning algorithms based on user behaviour on an e-commerce website. The authors compare the performance of two machine learning algorithms, multilayer perceptron (MLP) and long short-term memory (LSTM) recurrent neural networks, using a variety of evaluation metrics. The results show that both algorithms can achieve high accuracy, but LSTM outperforms MLP in terms of recall and F1-score.

The study also identifies the most important features for predicting purchasing intention, such as time spent on the website, number of pages viewed, and bounce rate. Overall, the study demonstrates the potential of machine learning for improving the effectiveness of e-commerce websites and increasing sales.

2. Data Clustering: A Review:

The paper provides an overview of the field of data clustering, which is the process of grouping similar data points together. The authors discuss various clustering methods, such as hierarchical clustering, partitioning clustering, density-based clustering, and grid-based clustering. They also examine the strengths and weaknesses of these methods and provide examples of real-world applications of clustering, such as image segmentation and customer segmentation. The paper concludes with a discussion of current research directions and challenges in the field of data clustering.

2. A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised:

The paper explores the performance of supervised and unsupervised machine learning algorithms in detecting credit card fraud. The study evaluates several classification algorithms, including Logistic Regression, Random Forest, SVM, Naive Bayes, and K-Nearest Neighbors, as well as clustering algorithms such as K-Means and DBSCAN. The results of the study suggest that supervised learning algorithms generally outperform unsupervised methods in detecting credit card fraud, with Random Forest showing the highest accuracy among the evaluated algorithms. However, the authors note that a combination of supervised and unsupervised methods may offer the best results. Overall, the study provides insights into the effectiveness of different machine learning approaches for credit card fraud detection, and highlights the importance of selecting the appropriate algorithm for the specific application.

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	Categorical	Operating system of the visitor.
Browser	Categorical	Browser of the visitor.
Region	Categorical	Geographic region from which the session has been strated by the visitor.
Traffic type	Categorical	Traffic sources by which the visitor has arrived at the website.
Visitor type	Categorical	Visitor type as "New visitor", "Returning Visitor" and "Other"
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.

A) Classification

- We plan to use following machine learning models:

1. Logistic Regression: It is a popular algorithm used for binary classification problems. It is a simple and efficient algorithm that models the probability of an event occurring given a set of input features. It works by fitting a linear decision boundary to the input data and then using a sigmoid function to map the linear output to a probability value between 0 and 1.

2. Support Vector Machine, (SVMs): SVM is a powerful algorithm that can be used for both binary and multi-class classification problems. SVMs work by finding the optimal hyperplane that separates the data into different classes. The algorithm tries to maximize the margin between the hyperplane and the closest data points, which helps to improve the generalization performance of the model.

3. Random Forest Classifier: It is a popular ensemble learning algorithm that combines multiple decision trees to make predictions. Each decision tree in the random forest is trained on a subset of the input data and a random subset of the input features. This helps to reduce overfitting and improve the accuracy and generalization performance of the model.

4. Naive Bayes: It is a probabilistic algorithm that is based on Bayes' theorem. It is a simple and efficient algorithm that can be used for both binary and multi-class classification problems. Naive Bayes assumes that the input features are conditionally independent given the class label, which simplifies the computation of the posterior probabilities.

5. XGBoost Classifier: It is a popular gradient boosting algorithm that is often used for classification tasks. It is an ensemble learning algorithm that combines multiple weak models (decision trees) to create a strong predictive model. The algorithm works by iteratively adding decision trees to the model, and each new tree is trained to correct the errors of the previous trees. XGBoost also includes regularization techniques to reduce overfitting and improve the generalization performance of the model.

- For classification we try to focus on following metric such as Precision, Recall, F1 score, AUC curve to measure performance the models

B) <u>Clustering</u>

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 based on the similarity between data points. K-means clustering is a simple and fast algorithm which is suitable for both small and large datasets.

2. **Hierarchical clustering:** This is another unsupervised learning algorithm that clusters data points into a tree-like structure based on the similarity between them. Hierarchical clustering can be either agglomerative (bottom-up) or divisive (top-down).

3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** As name says this is a density-based clustering algorithm that groups data points into clusters based on their density. DBSCAN is particularly useful for datasets with irregular shapes and noises.

4. **Gaussian Mixture Models (GMM):** This is a probabilistic model-based clustering algorithm that assumes that the data points are generated from a mixture of Gaussian distributions. GMM is flexible algorithm that can capture complex patterns in the data.

- Some of metrics for clustering we are planning to use are, silhouette score, calinski-harabasz index, davies-bouldin index, dunn index, rand index.

3 <u>Software packages</u>

- 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:

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

[2]. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks [https://link.springer.com/article/10.1007/s00521-018-3523-0]

[3]. Data Clustering: A Review [https://dl.acm.org/doi/pdf/10.1145/331499.331504]

[4]. A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised [https://arxiv.org/pdf/1904.10604.pdf]

[5]. Real-Time Prediction of Online Shoppers Purchasing Intention Using Random Forest [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256375/]