David Wang

501208639

Tamer Abdou, Ph.D.

June 5, 2023

**Literature Review and Data Description**

**Literature Review**

**Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks**

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)

This paper used the same dataset from the UCI Machine Learning Repository. The authors sought to create a model that would predict a customer’s purchasing intention and the likelihood that they would abandon the website. This was done through creation of two modules - the first of which would predict visitor purchasing intention from the dataset. The second module would focus on using sequential clickstream data of customers to train a neural network that would predict a customer who is likely to leave the site. The main research goal was to look into the feasibility of predicting purchasing intention gained from clickstream data as well as session information data.

This paper sought to use decision tree (C4.5 and random forest) classifiers, multilayer perceptron and support vector machines to predict user shopping behavior. These methods were chosen based on their use in past research. Upon running the models, the researchers found that they achieved poor model performance due to the large imbalance between negative and positive class samples. The researchers addressed the imbalance of class samples by dividing the dataset into training and test groups and then oversampling the dataset in order to look for stronger model performance. Next, the features were ranked using different methods, including correlation, mutual information, and mRMR methods. These were chosen as they sought to apply filter-based feature selection over wrapper algorithms that would require learning algorithms. Afterwards, feature selection was conducted using the MLP algorithm, which selected the top features to keep, with the model performing the best when using the top 6 features with the mRMR method.

**Real-Time Prediction of Online Shoppers’ Purchasing Intention Using Random Forest**

Baati, K., Mohsil, M. (2020)

This paper used a subset of the same data obtained from the UCI Machine Learning Repository. In this article, the researchers goal was to focus on the session and user information (non-clickstream data) in order to create a model that would predict users with high purchasing intention as soon as they connected to the website. Their model would then allow for presentation of marketing offers to entice visitors who demonstrated high purchasing intention and were likely to complete a transaction.

The data used was a subset containing all the categorical attributes as well as the numeric attribute SpecialDay. Specific algorithms were chosen as much of the dataset was categorical, as the researchers used the Naive Bayes Classifier, as well as the C4.5 and random forest decision tree classifiers. Similarly, the researchers discovered the issue of poor model performance caused by class imbalance and resolved it by oversampling using the SMOTE (Synthetic Minority Oversampling Technique) methodology. The researchers concluded that the random forest decision tree classifier performed the best after oversampling.

**Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks**

Diamantaras, K., Salampasis, M., Katsalis, A., Christantonis, K. (2021)

The goal of this paper was to improve and analyze the usage of e-commerce platform to personalize web page content presented to different customers. It additionally aimed to create a real time model of user purchasing behavior, which could then lead to more successful transactions with the application of different marketing or promotional offers.

The LSTM-RNN (Long Short Term Memory Recurrent Neural Network) was chosen as the model based off of previous literature in which the method had been used previously to learn user session sequences. The researchers concluded that using the LSTM-RNN method, they were able to demonstrate a model that could successfully predict user purchasing behavior.

**Data**

The data consists of 12,330 entries, with 18 attributes. Of these 18 attributes, 10 are numerical and 8 are categorical. The dataset contains 84.5% (10,422) samples where a shopper did not go on to purchase an item, and 15.5% (1,908) samples that led to a purchase. The data has no missing values, and duplicate rows are coincidental as data was collected from different users.

**List of Attributes**

**Administrative - Numeric**

This attribute represents the number of ‘Administrative’ web pages the user visited in a shopping session, such as pages regarding account management. This attribute has a large number of zeros (46.8%) and high positive correlation with the attribute ‘Administrative\_Duration’.

**Administrative\_Duration - Numeric**

This attribute represents the amount of time (in seconds) spent on ‘Administrative’ pages. This attribute has a large number of zeros (47.9%) and high positive correlation with the attribute ‘Administrative’.

**Informational - Numeric**

This attribute represents the number of ‘Informational’ web pages the user visited in a shopping session, such as pages regarding contact information. This attribute has a large number of zeros (78,7%) and high positive correlation with the attribute ‘Informational\_Duration’.

**Informational\_Duration - Numeric**

This attribute represents the amount of time (in seconds) spent on ‘Informational’ pages. This attribute has a large number of zeros (80.5%) and high positive correlation with the attribute ‘Informational\_Duration’.

**ProductRelated- Numeric**

This attribute represents the number of related product web pages the user visited in a shopping session. This attribute has a high positive correlation with the attribute ‘ProductRelated\_Duration’.

**ProductRelated\_Duration- Numeric**

This attribute represents the amount of time (in seconds) spent on related product pages. This attribute has a large number of zeros (6.1%) and high positive correlation with the attribute ‘ProductRelated’.

**BounceRates - Numeric**

This attribute represents the percentage of visitors of a site who enter a website and leave (“bounce”) without triggering any other requests. This metric is collected by Google Analytics. This attribute has a large number of zeros (44.8%) and high positive correlation with ExitRates.

ExitRates - Numeric

This attribute represents the percentages of browsing sessions in which a specific page is the last in the session. This metric is collected by Google Analytics. This attribute has high positive correlation with BounceRates and high negative correlation with ProductRelated.

**PageValues - Numeric**

This attribute represents the average number of pages visited by the customer before completing a session. This attribute has a high number of zeros (77.9%) and is collected by Google Analytics.

**SpecialDay - Numeric**

This attribute is a custom attribute that represents the closeness to a special day, ranging from 0 to 1.0. This attribute has a high number of zeros (89.9%).

**Operating System - Categorical**

This attribute represents which operating system was used by the customer. It has 8 categorical values.

**Browser - Categorical**

This attribute represents which web browser was used by the customer. It has 13 categorical values.

**Region - Categorical**

This attribute represents the geographic region from which the session took place. It has 9 categorical values.

**TrafficType - Categorical**

This attribute represents the traffic source from which the customer arrived at the website. There are 20 categorical values.

**VisitorType - Categorical**

This attribute classifies customers as “New Visitor”, “Returning Visitor” and “Other”. “Returning Visitors” comprise 85.57% of values.

**Weekend - Categorical**

This attribute is a boolean value that indicates whether the session took place on a weekend. 76.74% of sessions took place on a weekday.

**Month - Categorical**

This attribute is a categorical value indicating the month in which the session took place.

**Revenue - Categorical, Class label**

This attribute indicates whether a customer interaction ended with a transaction.

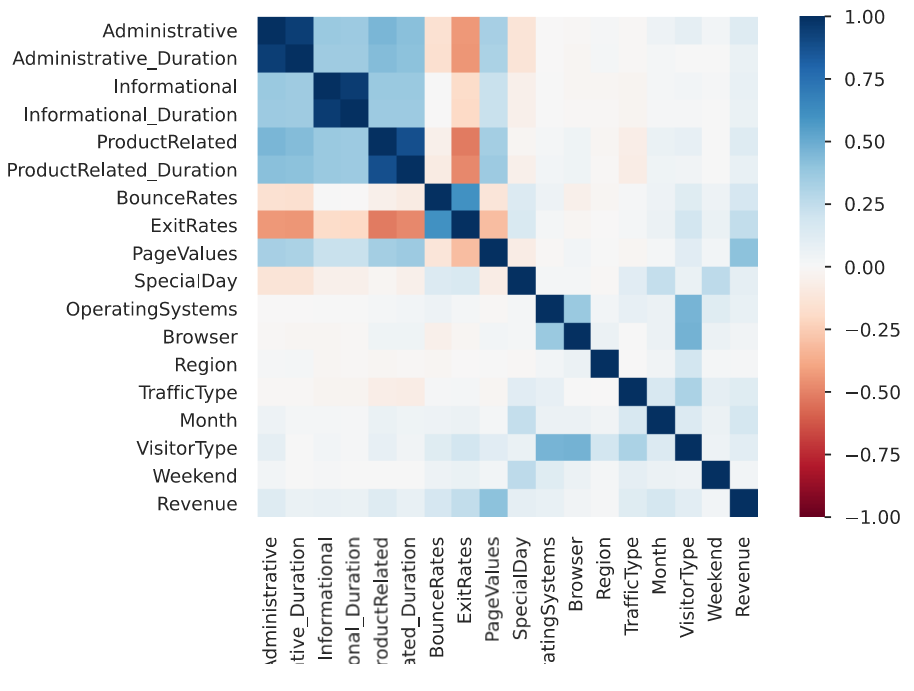
**Numeric Attributes**

| **Attribute** | **Mean** | **Min** | **Max** | **SD** |
| --- | --- | --- | --- | --- |
| Administrative | 2.32 | 0 | 27 | 3.32 |
| Administrative\_Duration | 80.82 | 0 | 3398.75 | 176.70 |
| Informational | 0.50 | 0 | 24 | 1.26 |
| Informational\_Duration | 34.47 | 0 | 2549.38 | 140.64 |
| ProductRelated | 31.73 | 0 | 705 | 44.45 |
| ProductRelated\_Duration | 1194.74 | 0 | 63973.52 | 1912.25 |
| BounceRates | 0.02 | 0 | 0.2 | 0.04 |
| ExitRates | 0.04 | 0 | 0.2 | 0.05 |
| Page Value | 5.89 | 0 | 361 | 18.55 |
| Special Day | 0.06 | 0 | 1.0 | 0.19 |

**Categorical Attributes**

| **Attribute** | **Number of Variables** |
| --- | --- |
| Operating System | 8 |
| Browser | 13 |
| Region | 9 |
| Traffic Type | 20 |
| Visitor Type | 3 |
| Weekend | 2 |
| Month | 12 |
| Revenue | 2 |

**Data Correlation**

****

There is high positive correlation between the pageview and the pageview duration attributes (Administrative & Administrative\_Duration, etc.) We also see that there is high positive correlation between the BounceRates and the ExitRates attributes. We see that there is also strong negative correlation between the ExitRates and the Administrative attributes, and between the ExitRates and the ProductRelated attributes.

**Research Question**

I plan to look into how website design can be optimized to promote increased successful transactions. Much of the previous literature focuses on supplementary marketing techniques or promotions to increase transactions, however, I believe that the clickstream data within the dataset combined with the user session data could be combined to learn about customer purchasing behavior and how websites and e-commerce platforms could be subsequently designed to promote increased transactions. I also noticed that there are many attributes within the dataset that are often filtered out by different feature selection algorithms. I would also like to look into applying dimensionality reduction techniques such as PCA (Principal Component Analysis) to try and summarize some of the different attributes into attributes that may represent the dataset better as a whole and lead to greater success in predicting user shopping behavior. I plan to use similar modeling techniques as those previously used in literature, including decision trees (C4.5, random forest), as well as SVM and MLP models.

**Methodology**

1. Get a baseline model performance for each of the methods (C4.5, Random Forest, SVM, MLP
2. Split the data 70:30 into training and test datasets. Apply SMOTE oversampling method to look and improve model performance due to differences in revenue class.
3. Feature selection
   1. Apply feature selection algorithms (Correlation, Random Forest, mRMR) and apply the MLP algorithm to determine how many attributes to use
   2. Apply PCA dimensionality reduction techniques to try to condense the number of attributes involved. Then rerun the feature selection algorithms and apply the MLP algorithm to see whether any of the PCA attributes have been included.
4. Compare model performance results across the feature selection used, as well as the predictive model used
5. Draw conclusions that can be made about user purchasing behavior and how website design can be modified to promote successful transactions

**References**

Baati, K., & Mohsil, M. (2020, May 6). *Real-time prediction of online shoppers’ purchasing intention using Random Forest*. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256375/

Diamantaras, K., Salampasis, M., Katsalis, A., & Christantonis, K. (2021, January). *Predicting shopping intent of e-commerce users using LSTM recurrent ...* Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks. https://www.researchgate.net/publication/353397684\_Predicting\_Shopping\_Intent\_of\_e-Commerce\_Users\_using\_LSTM\_Recurrent\_Neural\_Networks

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018, May 9). *Real-time prediction of online shoppers’ purchasing intention using Multilayer Perceptron and LSTM recurrent neural networks - neural computing and applications*. SpringerLink. https://link.springer.com/article/10.1007/s00521-018-3523-0