Online Shoppers Purchasing Intention Analysis

***Abstract*—In this study, we analysed online shopping behavioural patterns and how behaviours and attributes are associated with purchase intention using a dataset from UCI. With visualisation, we found that the purchase rate peaked in November. We also found that not-purchased customers have higher bounce rates and exit rates, while purchased customers have higher page values and percentage of special days, and the not-purchased group have a higher variance in terms of behavioural metrics. With subgroup comparison and Chi-square test, we found operation system 3, new visitors, and browser 3 have the relatively highest purchase rate. Finally, we trained a random forest model with a decent performance on the test set (accuracy score = 0.896, precision = 0.746, recall = 0.506, F1 score = 0.603). The findings help gain insight into customer behaviours and optimise business decisions.**

Keywords—online shopping, purchase prediction, digital marketing, e-commerce, random forest

I. INTRODUCTION

E-commerce usage has been increasing in the past two decades, and this trend has become more significant ever since COVID-19 and massive digitisation. Many people are now doing shopping or window shopping mainly on e-commerce platforms [1], and e-commerce has become one of the largest sectors within the IT industry. Among the website visitors, most of them are browsing rather than purchasing. According to digital marketing companies such as AdRoll and MarketingSherpa, no more than 2% of visitors end up buying at an online shopping site, while the other 98% solely do window shopping, and around 8% of them come back for a purchase later [2].

The purchase conversion rate is the key to e-commerce business and is one the most important metrics for e-commerce, sometimes even the North Star metric. Traditional market and user research tend to understand customers' purchase intentions based on survey or interview data and focus more on motivation, values and demographic characteristics. These approaches provide valuable insights to marketing and product design teams, but the timeliness is often a problem and thus sometimes cannot fit in agile development. Data science techniques and online behavioural metrics provide another more objective perspective for understanding customer behaviours. They can generate a more accurate and real-time result that enables more timely insights and responses, such as retargeting. For example, we can explore the association between page browsing and purchase and see if long-time reading or short-time glancing benefits more for buying, identify the best and worst customer groups in terms of conversion rate, and build predictive models to meet business needs.

II. ANALYTICAL QUESTIONS

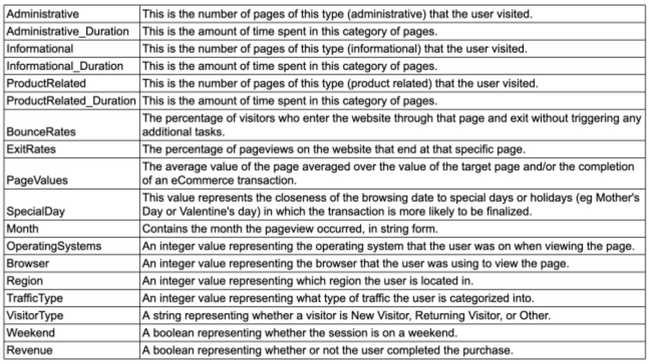
This study aims to understand the patterns of online purchases (e.g. What does it look like and how does it vary?) and try to predict online purchases using measurable behavioural variables. Specifically:

1. Is there any seasonal trend in online shopping purchase rates? This helps craft marketing campaigns and product design aligning with peak shopping periods, potentially boosting sales.
2. What is the difference between the profiles of purchased and not purchased visiting? This provides a basic understanding of customers and is a starting point for further, more detailed exploration.
3. What is the difference between user groups (e.g. different visitor types, regions, operating systems, and traffic types)? This helps target the right audience segments and optimise marketing efforts and ROI.
4. How does browsing influence purchase conversion rate? Is quick browsing more likely to lead to a higher purchase rate, or is it the other way around? This inspires and helps optimise the product page design.
5. Can we build an effective purchase prediction model using these features? How effective is it? What is the most important feature? It helps identify the key drivers of purchase behaviour and is valuable for marketing strategies and product design.

III. DATA

"The Online Shoppers Purchasing Intention Dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period." [3] It contains 12330 records and 18 columns, 9 of which are behavioural features (e.g., page view duration), 3 of which are time-related features (e.g., weekend or not), 5 of which are user attributes features (e.g., operating systems) and 1 of which is the target variable (Revenue). This dataset does not contain any missing values. The description of each column is shown in TABLE Ⅰ, and the descriptive statistic is shown in TABLE Ⅱ.

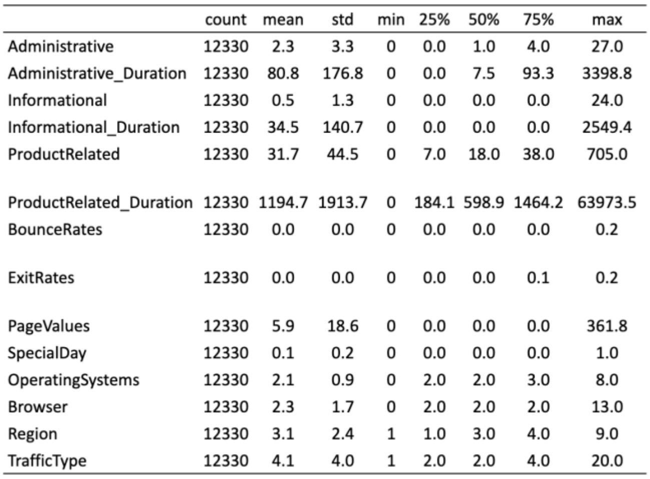
TABLE I. DATASET INFORMATION



Administrative	This is the number of pages of this type (administrative) that the user visited.
Administrative_Duration	This is the amount of time spent in this category of pages.
Informational	This is the number of pages of this type (informational) that the user visited.
Informational_Duration	This is the amount of time spent in this category of pages.
ProductRelated	This is the number of pages of this type (product related) that the user visited.
ProductRelated_Duration	This is the amount of time spent in this category of pages.
BounceRates	The percentage of visitors who enter the website through that page and exit without triggering any additional tasks.
ExitRates	The percentage of pageviews on the website that end at that specific page.
PageValues	The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction.
SpecialDay	This value represents the closeness of the browsing date to special days or holidays (eg Mother's Day or Valentine's day) in which the transaction is more likely to be finalized.
Month	Contains the month the pageview occurred, in string form.
OperatingSystems	An integer value representing the operating system that the user was on when viewing the page.
Browser	An integer value representing the browser that the user was using to view the page.
Region	An integer value representing which region the user is located in.
TrafficType	An integer value representing what type of traffic the user is categorized into.
VisitorType	A string representing whether a visitor is New Visitor, Returning Visitor, or Other.
Weekend	A boolean representing whether the session is on a weekend.
Revenue	A boolean representing whether or not the user completed the purchase.

Behavioural metrics enable us to compare and understand the profiles of purchased and not-purchased customers (question 2), and user attributes will allow us to compare purchase rates and find the most profitable subgroup to target (question 3). The column 'Month' enables a temporal trend investigation (question 1), and the column 'ProductRelated_Duration' enables us to evaluate the relationship between purchase and page browsing (question 4). All these features can be used to build a predictive model (question 5).

TABLE II. DATASET INFORMATION



	count	mean	std	min	25%	50%	75%	max
Administrative	12330	2.3	3.3	0	0.0	1.0	4.0	27.0
Administrative_Duration	12330	80.8	176.8	0	0.0	7.5	93.3	3398.8
Informational	12330	0.5	1.3	0	0.0	0.0	0.0	24.0
Informational_Duration	12330	34.5	140.7	0	0.0	0.0	0.0	2549.4
ProductRelated	12330	31.7	44.5	0	7.0	18.0	38.0	705.0
ProductRelated_Duration	12330	1194.7	1913.7	0	184.1	598.9	1464.2	63973.5
BounceRates	12330	0.0	0.0	0	0.0	0.0	0.0	0.2
ExitRates	12330	0.0	0.0	0	0.0	0.0	0.1	0.2
PageValues	12330	5.9	18.6	0	0.0	0.0	0.0	361.8
SpecialDay	12330	0.1	0.2	0	0.0	0.0	0.0	1.0
OperatingSystems	12330	2.1	0.9	0	2.0	2.0	3.0	8.0
Browser	12330	2.3	1.7	0	2.0	2.0	2.0	13.0
Region	12330	3.1	2.4	1	1.0	3.0	4.0	9.0
TrafficType	12330	4.1	4.0	1	2.0	2.0	4.0	20.0

The assumptions of this study are as follows: 1. Purchase rate peaks in both the summer and Christmas seasons. 2. Purchased and not-purchased customers have different behavioural patterns. 3. There are significant differences in purchase rates among subgroups. 4. Longer page browsing time is associated with a higher purchase rate. 5. We can build a classification model that performs significantly better than random guessing based on these features.

IV. ANALYSIS

The analytical steps are aligned with the research questions and consist of three blocks: data preparation, data analysis, and model training. The detailed steps are as follows:

1. Data overview and data quality check.

In this step, the shape of the dataframe, column names, missing values and data type of each column was first checked to have an overall understanding of the dataset and investigate data quality. Then descriptive statistics were performed for numerical columns (using count, mean, std, and quantiles) and categorical variables (using frequency), respectively. Following this, the distribution of each column was investigated using histograms. Finally, outliers were checked using visualisation (strip plots) and between-group customer attributes comparison. However, outliers will not be removed from the dataset as we don't clearly know what these outliers mean and why they exist. They might carry some meaningful information.

2. Data recode and temporal trend analysis.

In this step, inappropriate data types were converted to suitable ones for further analysis. Specifically, the 'Month' column (in the format of 'Jan', 'Feb') was converted into integer numbers (e.g. 1, 2). Then, subgroups with a small sample size (n < 50) that we found from the previous step were grouped together as a new category labelled with 0. This helps prevent misleading results from too small sample sizes when conducting subgroup analysis. After the 'Month' column was converted to numerical type, we can group dataframe by month and calculate the mean purchased rate of each, then plot the temporal trend using a line plot.

3. Behavioural profile comparison.

In this step, we compared the differences in behavioural metrics (e.g. Administrative_Duration, ProductRelated and ExitRates) between purchased and not-purchased customers using visual design (parallel coordinates graph). This method allows us to compare the differences in behavioural patterns at a glance. The 'color' parameter of parallel coordinates graph cannot take boolean data type directly, so we need to do data derivation and create a new temporary column called "Revenue_int" (dtype = 'int') to serve as the segmentation variable.

4. Subgroup comparison of the purchase rate.

The main goal of this step is to answer question 3. We first selected the categorical variables for comparison (i.e., OperatingSystems, VisitorType, Browser, Region), and then we calculated the mean and standard deviation of purchase rate and sample size of each group. Then, we create visualisation (bar plots) to view the between-group differences better. On top of these descriptive comparisons, inferential statistics (Chi-squared test was used to evaluate whether the between-group difference is statistically significant). The Chi-square test is a commonly used non-parametric statistic for assessing associations between categorical variables when the dependent variable is also a categorical variable. Like all non-parametric statistics, the Chi-square doesn't require a specific data distribution and is robust with non-normal distribution independent variables [4].

5. Correlation analysis

This step aims to answer research question 4. The metric for 'browsing time' is ProductRelated_Duration, which indicates the amount of time spent on product-related pages. Before correlation analysis, a scatterplot was first implemented to check how 'ProductRelated_Duration' is associated with 'Revenue'. Because 'Revenue' is a binary variable, correlation methods such as Pearson and Spearman would not be appropriate as Pearson assumes both variables to be continuous, and Spearman assumes both variables to be ordinal. Point biserial correlation was used for this analysis. "Point biserial correlation is the value of Pearson's product moment correlation when one of the variables is dichotomous and the other variable is metric". Similar to the Pearson correlation, the Point biserial correlation values range from -1 to 1. 1 indicates a perfect positive correlation, 0 indicates no association at all, and −1 indicates a perfect negative correlation [5]. A p-value lower than 0.05 suggests the correlation is statistically significant.

6. Modelling.

This step aims to build a classification machine learning model to predict purchase intention using online behavioural features. Before building the model, the overall correlation between features and labels was quickly checked using a correlation matrix and heatmap.

Following this was the feature engineering process where all the categorical variables were encoded. The two most commonly used encoding techniques are one-hot encoding and ordinal encoding. The former is where the original variable is removed, and a set of new binary variables is added for each unique value in the original variable, with 1 indicating the presence of the category and 0 indicating the absence of that category. The advantage of this method is it does not require any inherent order within the variable. The disadvantage is it can lead to an infeasibly large number of features when being applied to high-dimensional data. The latter is where an integer value is assigned to each existing category and does not remove or add any new columns to the data. This method is suitable for variables with order. However, in the case of high cardinality, this technique leads to an infeasibly large number of features [6].

Here, we used one-hot encoding techniques because the dataset has a manageable number of features and all of them are nominal variables without inherent order (e.g. traffic type, operating system).

After feature engineering is model training and model validation, this study used a random forest algorithm to train the model, following the training-tuning-evaluating process. The hyperparameters ‘n_estimators’ (number of decision trees in the forest) and ‘max_depth’ (maximum depth of each decision tree in the forest) were tuned using random search techniques, and cross-validation technique was used during hyperparameter tuning. For model evaluation, the Confusion Matrix, accuracy, precision, recall, and F1 score were used to assess model performance on the test set.

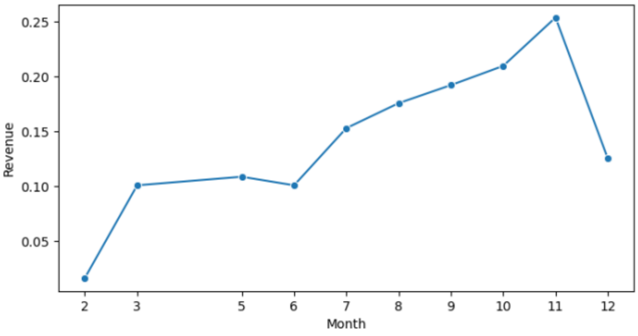
Finally, we extracted the feature importance from the best-trained model for further insights.

V. Findings, reflections and further work

1. Is there any temporal trend in terms of purchase rate?

We found a clear temporal trend from this dataset (Fig. 1). However, instead of as we expected (a dual peak pattern with one peak in summer and one during the Christmas season), there was only one single peak in November and then the purchase rate dropped sharply in December. Due to data incompleteness, we couldn’t know the purchase rate in January and April.

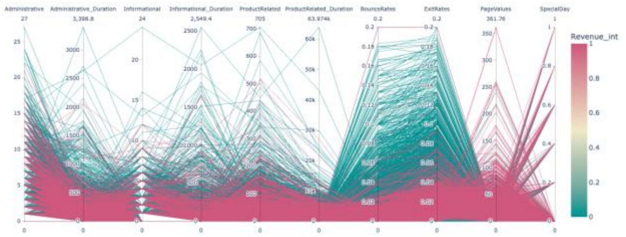
Fig. 1. Temporal trend of purchase rate.



2. What is the difference between the profiles of purchased and not purchased visiting?

The parallel coordinates plot (Fig. 2) shows very distinct profiles between purchased and not-purchased customers. The most notable differences are the higher bounce rates and exit rates of not-purchased customers and the higher page values and percentage of special days of purchased customers. These are probably the most important features for distinguishing successful and failed purchase conversions. Besides, we also noticed that the purchased group has a lower variance in most behavioural metrics.

Fig. 2. Behavioural profiles of purchased and not-purchased customers.



3. What is the difference in purchase rate between groups?

Fig. 3 illustrates the results of the between-group comparison regarding the purchase conversion rate. Operation system 3, new visitors, browser 3, and regions 2,7,8 have the relatively highest purchase rate, while operating system 1, returning customers, and browser 4 have the lowest purchase conversion rate. Further Chi-square test found that except for ‘Region’, the between-group differences in all the other groups were all statistically significant (TABLE Ⅲ), indicating the difference is unlikely to be attributed to random chance, but is likely a result of a genuine effect or relationship within the variables.

Fig. 3. Difference in purchase rate between groups.

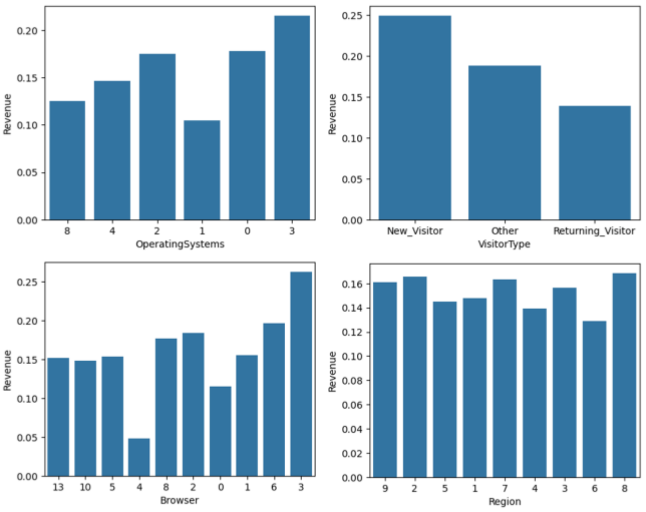


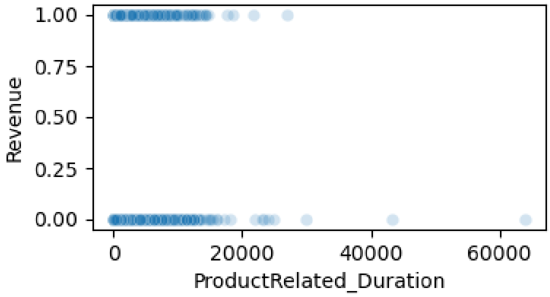
TABLE III. Results of the Chi-square test

	Operating Systems	Visitor Type	Browser	Region	Weekend
Chi-square	74.87	135.24	25.53	9.25	10.40
p	0.0000	0.0000	0.0024	0.3214	0.0013
degrees of freedom	5	2	9	8	1

4. How does browsing influence purchase conversion rate?

From the initial visualisation investigation, we could hardly find a clear association between product-related page viewing and purchase outcome (figure 4). however, further Point biserial correlation analysis showed that there is a significantly weak correlation between ProductRelated_Duration and purchase (correlation coefficient = 0.15, p = 0.0000). The not-purchased group's high variance in product-related page viewing might explain this visually not-noticeable but statistically significant result. It is not fair to conclude there is a clear association between product page browsing and purchase.

Fig. 4. Scatter plot of page viewing and purchase rate.



5. The effectiveness of the predictive model.

The overall correlation between variables is shown in Figure 5. From this heatmap, we can see there is some multicollinearity among behavioural features. The random forest algorithm is not sensitive to the multicollinearity of features, so this is acceptable for model building. The confusion matrix (figure 6) shows the result of our random forest model. After hyperparameter tuning, our model generates a good prediction result on the test set (accuracy score = 0.896, precision = 0.746, recall = 0.506, F1 score = 0.603).

Fig. 5. Heatmap of the correlation matrix.

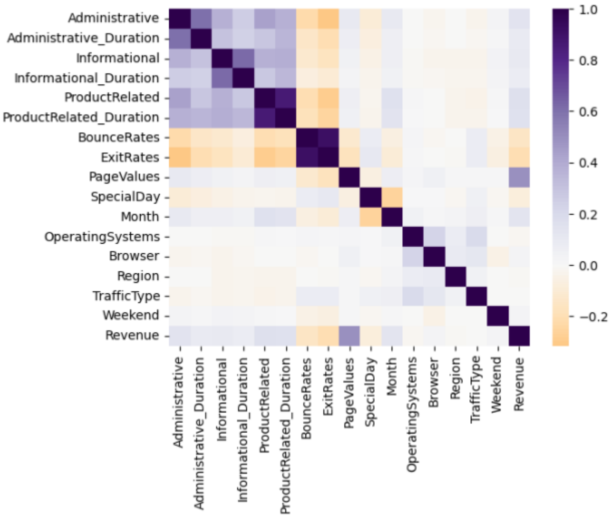


Fig. 6. Confusion matrix of model performance on test set.

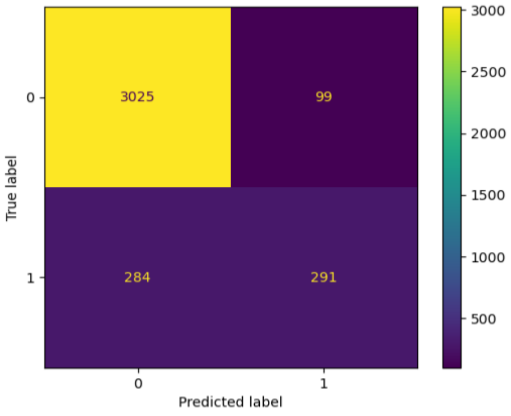
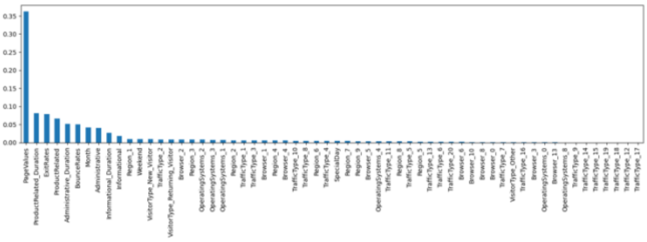


Fig. 7 shows that the the‘PageValues’ is the most important feature for predicting purchases.

Fig. 7. Feature importance.



This study explored online shopping behavioural and attribute patterns and how these patterns are associated with purchase outcomes. The dataset contains non-normal distributed continuous variables and nominal variables that are suitable for our research questions. The findings help gain insight into customers and optimise business decisions. Specifically, temporal trends investigation helps find better timing for marketing campaigns, subgroup comparison helps determine which group should be targeted with high priorities, and the predictive model enables a more timely response to prevent drop-out.

One of the limitations is the lack of comprehensive customer segmentation. We only segmented customers using single variables. Advanced methods such as clustering analysis can provide more understanding of this dataset. Another limitation is the low recall score of our model due to imbalanced labels. Further studies can consider using oversampling techniques to improve the True-negative and True-positive rates [7].

References

[1] L. Guo, L. Hua, R. Jia, B. Zhao, X. Wang, and B. Cui, ‘Buying or Browsing?: Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior’, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, Jul. 2019, pp. 1984–1992. doi: 10.1145/3292500.3330670.

[2] J. Yeo, S. Kim, E. Koh, S. Hwang, and N. Lipka, ‘Predicting Online Purchase Conversion for Retargeting’, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Cambridge United Kingdom: ACM, Feb. 2017, pp. 591–600. doi: 10.1145/3018661.3018715.

[3] Y. K. C. Sakar, ‘Online Shoppers Purchasing Intention Dataset’. UCI Machine Learning Repository, 2018. doi: 10.24432/C5F88Q.

[4] M. McHugh, ‘The Chi-square test of independence’, *Biochem. Medica*, vol. 23, pp. 143–9, Jun. 2013, doi: 10.11613/BM.2013.018.

[5] D. Kornbrot, ‘Point Biserial Correlation’, in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2014. doi: 10.1002/9781118445112.stat06227.

[6] K. Potdar, T. Pardawala, and C. Pai, ‘A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers’, *Int. J. Comput. Appl.*, vol. 175, pp. 7–9, Oct. 2017, doi: 10.5120/ijca2017915495.

[7] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, ‘Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks’, *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6893–6908, Oct. 2019, doi: 10.1007/s00521-018-3523-0.