Performance Analysis of Machine Learning Models for eCommerce Purchase Prediction

Lena Moroz | S3063766
Machine Learning
School of Computing, Engineering & Digital Technologies, Teesside University, United Kingdom

Abstract

The surge in global e-commerce sales underscores the importance of understanding online customer behaviour and intentions to bolster marketing strategies and drive sales. This article delves into the field of analysing online shoppers' purchase intentions, exploring performance of five ML algorithms. The study conducts extensive data exploration, feature selection, preprocessing and optimisation. Subsequently, ML models are trained and evaluated, with Random Forests and Decision Trees emerging as top performers. Hyperparameter tuning further enhances model performance, particularly in the case of the non-linear Support Vector Classifier. Feature importance analysis sheds light on critical predictors of purchase intent. The culmination of these efforts showcases the efficacy of ML models in predicting purchase intentions, offering valuable insights for e-commerce marketing strategies and data-driven management. Future work may involve exploring real-time prediction capabilities to further optimize model performance.

Keywords: *E-commerce, online shopping, purchase intention prediction, predictive analytics, hyperparameter tuning, Random Forest, SVC, Decision Tree, KNN, Logistic Regression*

1. INTRODUCTION

According to the latest statistics in 2023, global retail e-commerce sales reached an estimated 5.8 trillion U.S. dollars [1].

Comprehending online customers' behaviour and intentions is crucial for marketing endeavours, aiming to enhance customer experiences and consequently boost sales. Despite the substantial rise in e-commerce usage, the conversion rate hasn't seen a proportional increase, highlighting the necessity for tailored promotions.

Consumer conversion is a significant challenge in online sales, with the proportion of sessions culminating in purchases being negligible in comparison to the total number of visits to a website [2,3,4].

Consequently, analysing online shoppers' purchase intentions from empirical shopper data has emerged as a burgeoning field of research.

There were plenty of studies written on purchase prediction methods in the last few years [5,6]. For example, one research investigates three methods, naïve Bayes classifier, C4.5 decision tree and random forest, for the best performance [7]. However, most of these approaches operate offline, aiming to forecast purchases after the shopper has already exited the website, thereby shaping subsequent actions. Consequently, these methods prove ineffective for predicting purchases early during an ongoing session.

Thus, real-time online shopper behaviour prediction raises more interest, for example as the described analysis system with two modules where one predicts visitor shopping intent, and the other forecasts website abandonment likelihood [8]. Another interesting approach was used in the work, where the authors developed an Early Purchase Prediction framework (EPP) to predict buying intention in an ongoing session to be able to give real-time offers and discounts before the session ends [9].

2. DATA EXPLORATION AND FEATURES SELECTION

2.1. Dataset Description

This project is based on Online Shopper Intention Dataset downloaded from UCI's Machine Learning Library. It comprises feature vectors from 12,330 sessions. Among these, 84.5% (10,422 sessions) represent the negative class samples, indicating sessions that did not culminate in shopping. The remaining 15.5% (1,908 sessions) are positive class samples, denoting sessions that concluded with shopping. Each session is attributed to a distinct user over a span of one year, which aims to mitigate biases associated with campaigns, special occasions, user profiles, or time periods. The table 1 of description of each column is provided in the appendix: 14 of 18 columns were numerical, 2 categorical and 2 binaries.

2.2. Data Cleaning and Pre-processing

We dropped three columns 'Administrative', 'Informational', and 'ProductRelated' because their information is already captured by another column "PageValues." Therefore, these columns were considered redundant and not useful for the analysis. Then we deleted 173 duplicated values. There were no missing values in the dataset.

Then we encoded categorical or Boolean values into numerical (‘Weekend’, ‘Revenue’, ‘Month’ columns).

We checked distribution of the numerical features by revenue and their influence on revenue before performing scaling. We got interesting results which show the highest positive influence of traffic type 2 for the revenue and the disproportion of traffic types skewed to the side of types 1-4.

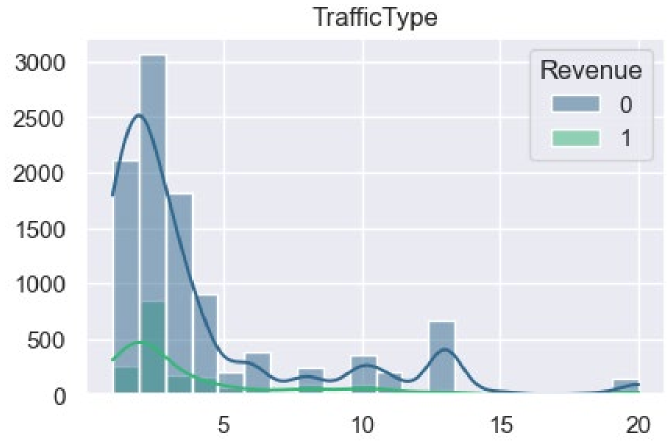


Figure 1. Distribution of Traffic Type by Revenue

Same disproportion was noticed for the Operating systems and Browser.

Then we performed scaling the date with the MinMaxScaler to standardize the scale of all features, preventing any single feature from disproportionately influencing distance-based algorithms within a model.

2.3. Exploratory Data Analysis

We built a bar chart to evaluate the monthly visitor trends, which showed the prevalence of March, April, November and December for making purchases.

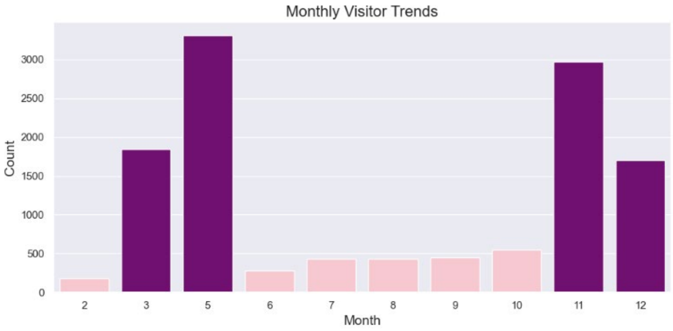


Figure 2. Bar Chart of Monthly Visitor Trends

The picture below shows the distribution of Visitor Types by Revenue, with significant imbalance between returning and new visitors.

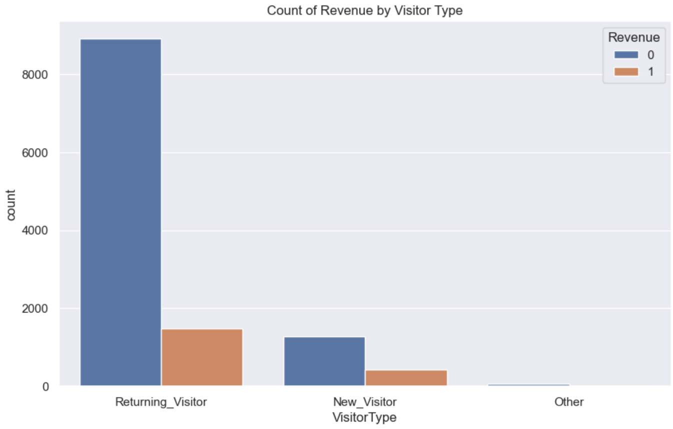


Figure 3. Bar Chart of Revenue by Visitor Type

Next chart shows the great correlation between monthly visitors and monthly purchases.

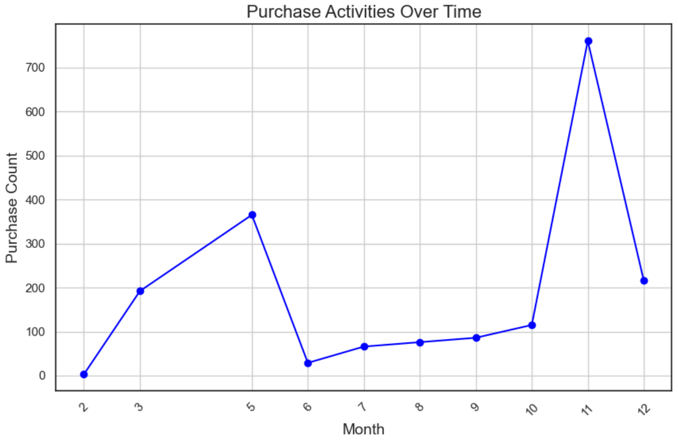


Figure 4. Purchase activities over time

For the next chart we grouped data by traffic type and calculated average and total revenue by traffic type.

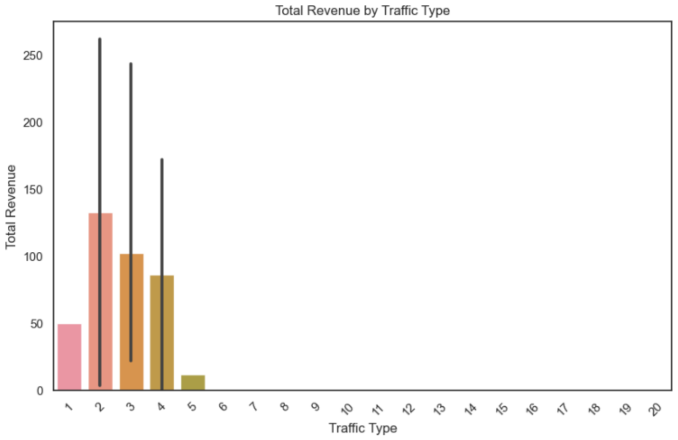


Figure 5. Total revenue by Traffic Type

The chart revealing the influence of special days on revenue is quite intriguing. Surprisingly, the chart suggests that 'normal' days, which aren't associated with any holidays or events, generate higher revenue compared to other days. However, this might be attributed to the sheer frequency of 'normal' days rather than any inherent difference in purchasing behaviour.

Although there doesn't seem to be any significant imbalance in purchase intent on specific days, aggregating the revenue from special days, accounting for all coefficients, yields the highest revenue overall.

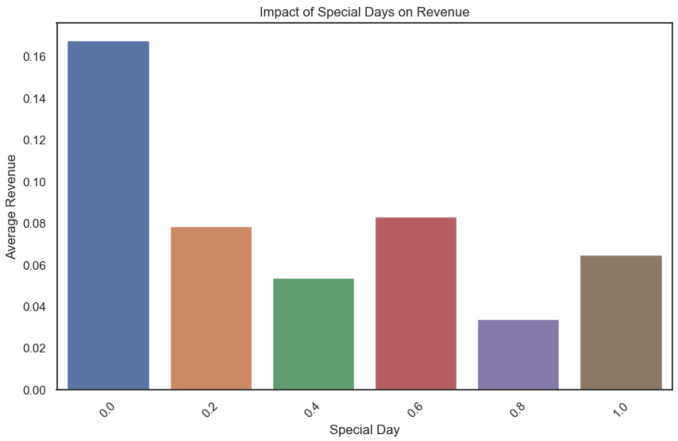


Figure 6. Impact of Special Days on Revenue

The correlation Matrix (Fig.7 in the appendix) shows high positive correlation between ExitRates and

BounceRates, Page Value and Revenue. High bounce might lead to a high exit rate, indicating that users are not finding the content engaging enough to explore further. A strong correlation between Page Value and Revenue is expected, as pages with higher values are likely to contribute more to overall revenue.

2.4. Handling class imbalance

The next visualization aids in grasping the spread of the target variable, specifically the revenue class. In binary classification scenarios such as this, maintaining a balanced distribution of classes is crucial. Here, we notice an imbalance in classes, which may result in a biased model that underperforms on the less represented class.

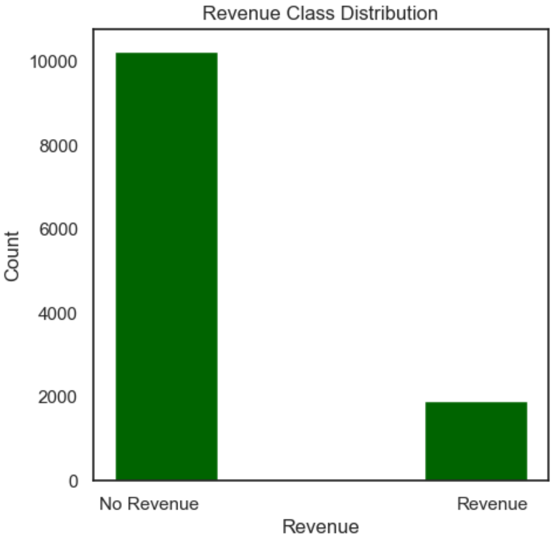


Figure 8. Revenue Class Distribution

We tackled the imbalance in the target variable using Synthetic Minority Over-sampling Technique (SMOTE). This method generates new instances of the minority class by interpolating between existing instances, effectively boosting the representation of the minority class in the dataset.

3. DEVELOPING AND EVALUATING MODELS

3.1. Models' Implementation

After implementation of SMOTE and splitting the data, we got 8199 instances of both positive and negative revenue values for the training dataset and 2050 of each for the testing one.

We employed the LazyPredict library to assess various machine learning models on our dataset. This tool automates the process of constructing, training, and assessing multiple models, offering insights into their performance on the data.

We decided to pick 5 top performing machine learning algorithms for classification tasks, which were covered in this semester. They are:

1. RandomForestClassifier: Achieved an accuracy of 0.94 with a relatively short training time of 2.07 seconds.
2. SVC (Support Vector Classifier (non-linear)): Achieved an accuracy of 0.89 with a longer training time of 7.53 seconds.
3. DecisionTreeClassifier: Achieved an accuracy of 0.90 with a short training time of 0.11 seconds.
4. KNeighborsClassifier: Achieved an accuracy of 0.88 with a short training time of 0.29 seconds.
5. LogisticRegression: Achieved an accuracy of 0.87 with a very short training time of 0.06 seconds.

3.2. Models' Evaluation

After selecting our machine learning algorithms, we assessed their performance using various evaluation metrics. These metrics encompass accuracy, precision, recall, and F1 score.

ML Model/Metrics	Accuracy	Precision	Recall	F1
RandomForest-Classifier	0.93	0.91	0.95	0.93
Support Vector Classifier (non-linear)	0.70	0.67	0.78	0.72
DecisionTree-Classifier	0.87	0.87	0.88	0.87
KNeighbors-Classifier	0.82	0.77	0.90	0.83
Logistic-Regression	0.83	0.85	0.79	0.82

Table 2. Performance of ML models

The Random Forest Classifier demonstrates high accuracy (93%) and balanced precision (91%) and recall (95%), indicating robust performance across the dataset. The F1 score is also high at 93%. The non-linear SVC exhibits moderate performance, with accuracy of 70%, precision of 67%, recall of 78%, and F1 score of 72%. The Decision Tree Classifier performs well, with accuracy of 87%, precision of 87%, recall of 88%, and F1 score of 87%. The KNN achieves a relatively high accuracy of 82% but demonstrates slightly lower precision (77%) compared to recall (90%), resulting in a slightly lower F1 score (83%). The Logistic Regression model shows balanced accuracy of 83%, precision of 85%, recall of 79%, and F1 score of 82%, with slightly higher precision compared to recall.

3.3. Models' optimization

We implemented hyperparameter tuning for all our models by employing GridSearchCV to search through the parameter grid and find the optimal combination of hyperparameters. Accuracy was chosen as the scoring metric. The table below shows the enhanced predictive performance:

Model/Metrics	RF	SVC	DT	KNN	LR
Accuracy before tuning	0.93	0.70	0.87	0.82	0.83
Accuracy after tuning	0.93	0.85	0.90	0.86	0.86

Table 3. Performance after hyperparameter tuning.

Random Forests and Decision Trees continue to exhibit the most favourable performance among the algorithms evaluated. While there were marginal enhancements observed with KNN, Decision Trees, and Logistic Regression, the non-linear SVC model notably displayed an accuracy 15% higher than before the optimisation.

3.4. Feature Importance

We implemented the feature importance analysis for the best performing model - Random Forest Classifier.

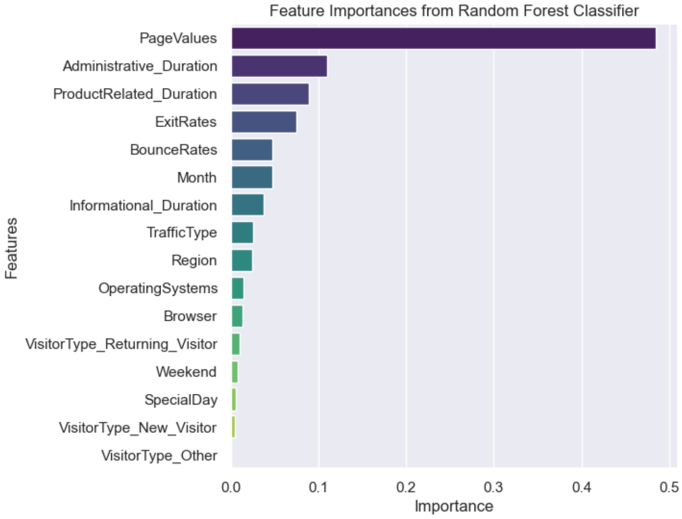


Figure 9. Feature Importance from Random Forest Classifier

According to the graph, we can observe the following findings:

- 'PageValues' emerges as the most critical factor for categorizing types based on the Randon Forest Classifier.
- Following 'ProductRelated_Duration', 'ExitRates' and 'Month' are significant, alongside 'Administrative_Duration' and 'BounceRates'. These metrics are key indicators for revenue conversion, reflecting user exits and product engagement.
- 'VisitorRype_Other', 'SpecialDay' and 'Weekend' are deemed the least influential variables in predicting outcomes according to this model.

4. RESULTS AND DISCUSSION

The model with the best performance is Random Forest Classifier with the best hyperparameters, determined through a systematic search using GridSearchCV, to be a maximum depth of None and 100 trees in the forest. This combination of hyperparameters resulted in an accuracy of 92.95% on the test dataset, indicating that the model performs well in correctly classifying the purchase intentions in the e-commerce marketplace.

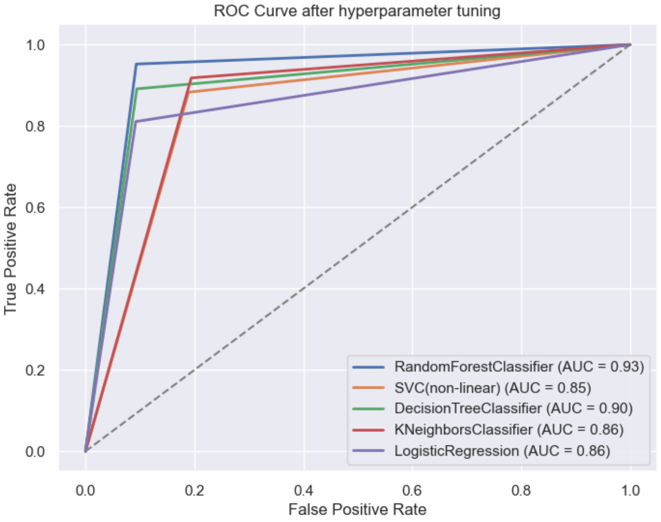


Figure 10. Comparison of ROC Curve of 5 models after hyperparameter tuning

From the ROC Curve displayed in fig.10 above, it is shown that Random Forest is having the biggest area under the curve (ROC-AUC=0.93) which is the indicator of the excellent performance for the given classification task. Second best model is Decision Trees with the result of 0.87, which is considered very good performance.

The robust performance of RF is due to several factors. Firstly, this algorithm is an ensemble learning method which can capture complex non-linear relationships between features and patterns in the purchase intention data.

In contrast, while tuning the non-linear SVC model led to a notable improvement in accuracy by 15%, achieving an accuracy of 0.85, it remained the lowest-performing model among those tested. This outcome may be attributed to several factors specific to the SVC algorithm. It can be sensitive to the choice of hyperparameters and may require more intensive tuning compared to other algorithms. Additionally, SVC may struggle with large datasets or datasets with high dimensionality, which could have an impact on its performance in our e-commerce dataset.

Evaluating the second-best performing model (with 90% accuracy) – Decision Trees – it's important to outline that RF often outperforms a single decision tree due to its ensemble nature and ability to mitigate overfitting. Moreover, DT perform better on small and medium-sized datasets with relatively few features, which is not our case.

5. CONCLUSION AND FUTURE WORK

The proposed work is driven by the necessity to develop a model that predicts visitors' shopping intent upon their initial visit to an e-commerce website. This aims to mitigate the risk of retention associated with each website visit. The primary challenge involved identifying the most suitable ML model for this task. Five classification techniques — namely, Random Forests, non-linear SVC, Decision Trees, K-Nearest Neighbours, Logistic Regression — were explored to address this issue. Additionally, oversampling was employed to enhance the performance of each classifier.

While all models developed in this project have shown promising results, RF model stands out as particularly noteworthy, demonstrating superior performance compared to the others. However, there are opportunities for improvement in future work. This could involve acquiring more data with additional features or exploring additional features that may be relevant to the purchase prediction task, especially at the earliest stages in real time.

Considering potential legal and ethical issues, legal implications may arise if the ML models exhibit bias against certain demographic groups, leading to such discriminatory outcomes as for example, applying different marketing approaches. It's essential to mitigate bias and ensure fairness in model predictions among gender, age and other groups to comply with anti-discrimination laws. E-commerce platforms should obtain informed consent from customers before collecting and utilizing their data for predictive analytics. Data professionals must ensure the responsible and ethical use of predictive analytics in e-commerce by addressing biases, safeguarding data privacy, and mitigating potential harms to customers.

In conclusion, the objective of this research was to evaluate the performance of selected algorithms for predicting purchase intentions in e-commerce. The findings suggest that ML models serve as reliable tools for e-commerce establishments to enhance their marketing strategies and make well-informed decisions.

REFERENCES

1. *Global retail e-commerce sales 2014-2027* (no date) *Statista*. Available at: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/ (Accessed: May 6, 2024).

2. Liu, X., Lee, D., & Srinivasan, K. (2019). *Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning*. Journal of Marketing Research, 56(6), 918–943. Available at: https://doi.org/10.1177/0022243719866690 (Accessed: May 6, 2024).

3. Zhou, Y., Mishra, S., Gligorijevic, J., Bhatia, T., & Bhamidipati, N. (2019). *Understanding consumer journey using attention based recurrent neural networks*. In Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Available at: https://doi.org/10.1145/3292500.3330753 (pp. 3102–3111). (Accessed: May 6, 2024).

4. Behera, R.K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P.K. (2020). *Personalized digital marketing recommender engine*. Journal of Retailing and Consumer Services, 53, 101799. Available at: https://doi.org/10.1016/j.jretconser.2019.03.026. (Accessed: May 6, 2024).

5. Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2020). *Using Word2Vec recommendation for improved purchase prediction*. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). Available at: https://ieeexplore.ieee.org/document/9206871 (Accessed: May 6, 2024).

6. Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). *A machine learning framework for customer purchase prediction in the non-contractual setting*. European Journal of Operational Research, 281(3), 588–596. Available at: https://doi.org/10.1016/j.ejor.2018.04.034. (Accessed: May 6, 2024).

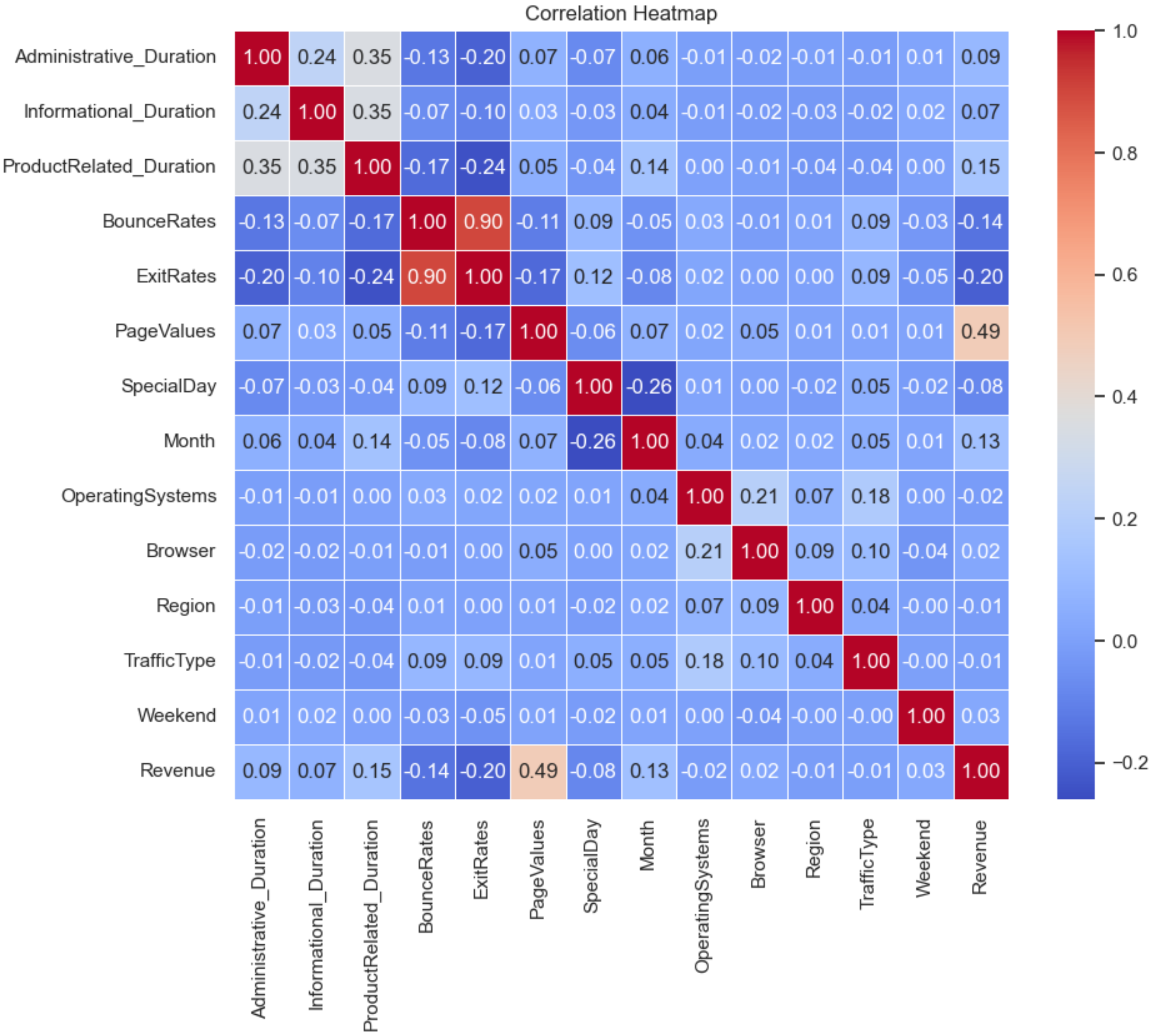
7. Baati K., Mohsil M. (2020). *Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest*. Available at: https://link.springer.com/chapter/10.1007/978-3-030-49161-1_4 (Accessed: May 6, 2024).

8.Sakar,C.O.,Polat,S.O.,Katircioglu,M.,Kastro, Y. (2019). *Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks*. Neural Comput. Appl. 31(10), 6893–6908. Available at: https://doi.org/10.1007/s00521-018-3523-0 (Accessed: May 6, 2024).

9. Esmeli R., Bader-El-Den M., Abdullahi H. (2020). *Towards early purchase intention prediction in online session-based retailing systems*. Available at: https://link.springer.com/article/10.1007/s12525-020-00448-x (Accessed: May 6, 2024).

Administrative	the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories
Administrative_Duration	
Informational	
Informational_Duration	
ProductRelated	
ProductRelated_Duration	
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 for a web page that a user visited before completing an e-commerce transaction
SpecialDay	represents the closeness of the browsing date to special days or holidays in which the sessions are more likely to be finalized with transaction
Month	month of the year
OperatingSystems	operating system
Browser	browser
Region	region
TrafficType	traffic type
VisitorType	returning or new visitor
Weekend	weekend or not
Revenue	revenue or not

Table 1. Description of the dataset columns



	Administrative_Duration	Informational_Duration	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	Weekend	Revenue
Administrative_Duration	1.00	0.24	0.35	-0.13	-0.20	0.07	-0.07	0.06	-0.01	-0.02	-0.01	-0.01	0.01	0.09
Informational_Duration	0.24	1.00	0.35	-0.07	-0.10	0.03	-0.03	0.04	-0.01	-0.02	-0.03	-0.02	0.02	0.07
ProductRelated_Duration	0.35	0.35	1.00	-0.17	-0.24	0.05	-0.04	0.14	0.00	-0.01	-0.04	-0.04	0.00	0.15
BounceRates	-0.13	-0.07	-0.17	1.00	0.90	-0.11	0.09	-0.05	0.03	-0.01	0.01	0.09	-0.03	-0.14
ExitRates	-0.20	-0.10	-0.24	0.90	1.00	-0.17	0.12	-0.08	0.02	0.00	0.00	0.09	-0.05	-0.20
PageValues	0.07	0.03	0.05	-0.11	-0.17	1.00	-0.06	0.07	0.02	0.05	0.01	0.01	0.01	0.49
SpecialDay	-0.07	-0.03	-0.04	0.09	0.12	-0.06	1.00	-0.26	0.01	0.00	-0.02	0.05	-0.02	-0.08
Month	0.06	0.04	0.14	-0.05	-0.08	0.07	-0.26	1.00	0.04	0.02	0.02	0.05	0.01	0.13
OperatingSystems	-0.01	-0.01	0.00	0.03	0.02	0.02	0.01	0.04	1.00	0.21	0.07	0.18	0.00	-0.02
Browser	-0.02	-0.02	-0.01	-0.01	0.00	0.05	0.00	0.02	0.21	1.00	0.09	0.10	-0.04	0.02
Region	-0.01	-0.03	-0.04	0.01	0.00	0.01	-0.02	0.02	0.07	0.09	1.00	0.04	-0.00	-0.01
TrafficType	-0.01	-0.02	-0.04	0.09	0.09	0.01	0.05	0.05	0.18	0.10	0.04	1.00	-0.00	-0.01
Weekend	0.01	0.02	0.00	-0.03	-0.05	0.01	-0.02	0.01	0.00	-0.04	-0.00	-0.00	1.00	0.03
Revenue	0.09	0.07	0.15	-0.14	-0.20	0.49	-0.08	0.13	-0.02	0.02	-0.01	-0.01	0.03	1.00

Figure 7. Correlation Matrix