Feature Importance Analysis for Predicting Online Shoppers' Purchasing Intent

Kathryn AuClair, Kristina Donders, and Kira Liang

Colab

I. Introduction

Online shopping is a vital component of modern commerce, having significantly transformed the retail landscape. Despite the flourishing online shopping industry, the proportion of online users who complete a purchase remains quite low. This persistent issue is on the minds of business owners: how to predict whether an online session will lead to a purchase. A significant part of the challenge in accurately forecasting whether a user intends to complete a purchase is identifying the aspects of online shopper behaviour that are most impactful. In this paper, we aim to find which features are the most important when it comes to predicting shopper intent.

For our research, we largely followed the procedure of Sakar et. al in *'A novel attribute selection approach in online shoppers' purchasing intention analysis'*. Our study tests whether correcting class imbalance helps improve the prediction of purchase intent in the dataset. Pre-processing includes label encoding of categorical fields and standardization of continuous features. We train two multi-layer perceptron classifiers: one on the original imbalanced data and one on a SMOTE-oversampled training set that synthetically increases minority-class examples so the model can learn patterns of successful purchasing sessions. Additionally, we train a simpler and more interpretable logistic regression model to use as a baseline for model comparison. The models are evaluated with PR-AUC and F1, performance metrics for imbalanced datasets, and with ROC-AUC for comparison. Finally, we conducted an ablation study to assess key influential features, such as PageValues, ExitRates, and BounceRates, to test their impact and trained a logistic regression model to get more interpretable coefficient results.

II. Background

A. Research Question

In this paper, we aim to answer the following question: **Which features in this dataset are most important for predicting purchasing intent?** We will determine which predictors are best for the prediction of shopper intent and determine whether adding or removing variables affects our model's performance. Accurately predicting the intent of users allows businesses to improve their marketing efforts, make better personalized experiences, and reduce abandonment rates, all of which have an impact on revenue and customer satisfaction. Having the ability to analyze and identify high-intent user sessions in real time would allow for targeted interventions like special offers.

Once we determine which variables are most influential, we will briefly explore ways that business owners can optimize or decrease the effects of these features to encourage user purchasing behaviour.

B. Related Work

The *Online Shoppers Purchasing Intention Dataset* was originally analyzed by Sakar, Polat, Katircioglu, and Kastro (2019), who proposed a two-module system for real-time predictions of user behaviour, purchase intention, and abandonment likelihood. The two-module system utilizes two multi-layer perceptrons, long short-term memory, and recurrent neural networks, feeding these models with compiled pageview data and session information. The main goal is to identify users with purchase intent who are more likely to abandon their cart, and enable the system of website to be able to offer timely offers or customized content to reduce abandonment likelihood. Sakar et al. emphasize the importance of feature selection and dimensionality reduction to improve future classification accuracy[1]. We aim to recreate their findings by applying various machine learning techniques and focusing on class imbalance when evaluating the effectiveness of SMOTE oversampling.

C. Dataset & Features

The *Online Shoppers Purchasing Intention Dataset* contains 18 values, including Revenue as the target value, and has data collected from 12,330 sessions. Sessions pertain to a single user in 1 year to avoid overlap with recurring campaigns, special days, or other user preferences. The dataset contains 11 integer features: Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, PageValues, SpecialDay, OperatingSystems, Browser, Region, and TrafficType. 5 continuous features: ProductRelated_Duration, BounceRates, ExitRates, Month, and VisitorType. There are also 2 binary values, Weekend as a feature, and Revenue as the target.

The variables Administrative, Informational, and ProductRelated, as well as their corresponding Duration feature, quantify how users navigate the site. These factors depict the number of pages a user viewed, total time spent on these page categories. These metrics are generated in real time from URL activity as users move between pages. The dataset also includes key Google Analytics metrics, which together offer a detailed view of user engagement: BounceRates is the proportion of single interaction sessions, ExitRates measures the percentage of pageviews that serve as the session's endpoint, and PageValues, which represents the average value of a webpage that a user visited before completing a transaction. Additionally, the SpecialDay feature quantifies the closeness of visiting the site to a specific special day (e.g., Valentine's Day, Mother's Day, etc.). It also takes into account the duration between the order and delivery date.

Overall, these factors collectively allow us to explore user behaviors, engagement, and transactional likelihood across browsing patterns, devices, and seasonal influences.

[1] Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). A novel attribute selection approach in online shoppers purchasing intention analysis. *Neural Computing and Applications*, 31(12), 6893–6908. https://doi.org/10Analyticsserve,,.1007/s00521-018-3523-0

III. Preliminary Studies

A. Feature Extraction with One-Hot Encoding

We began by performing feature selection via one-hot encoding. Feature extraction allows us to improve a machine learning model's performance by increasing accuracy, reducing training time, and enhancing the interpretability of this model[2]. We did this via one-hot encoding, which is the process of converting categorical variables into a numeric format that machine learning algorithms can more easily process. We used one-hot encoding on all of the categorical variables in the dataset, including: OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend, and Month. Using one-hot encoding will help prevent false ordinal relationships, whilst expanding feature space, ultimately improving flexibility and model interpretability.

B. Standardizing Numeric Features via Scaling

Next, we standardized numeric features in the dataset to ensure that they had equal weighting in model training. This prevented features with larger scales, such as PageValues, from having disproportionate influence on training the model. We scaled the variables using z-score standardization, allowing us to transform the data into a standard normal distribution. This method of standardization expresses each data point as the number of standard deviations it is away from the mean[3].

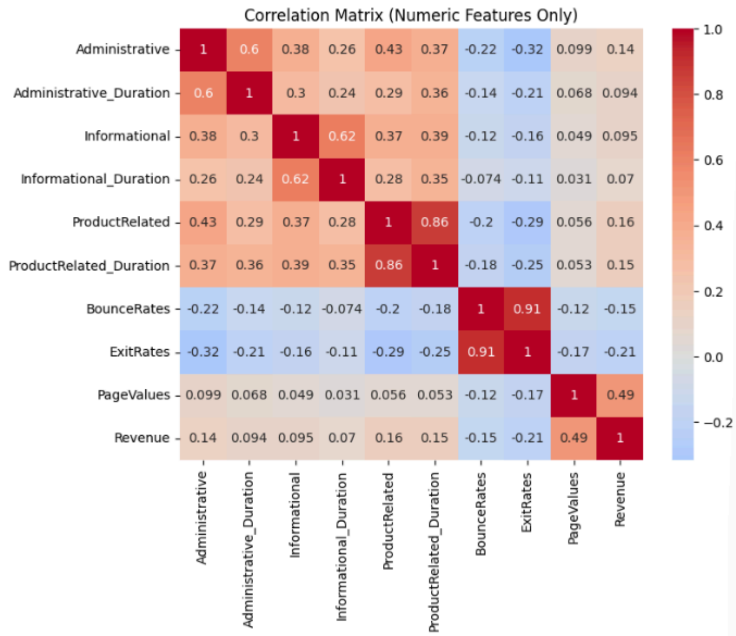
Z-score standardization, in tandem with feature extraction via one-hot encoding, prepares our dataset to be trained in a machine learning model to determine what features are most influential in determining an online shopper's purchasing intent.

C. Correlation Heatmap

Now that our dataset is standardized, we created a correlation heatmap among numeric features in the dataset to determine how related certain features are. This helps us visualize the data and illustrate multicollinearity among features.

[2] GeeksforGeeks. "What Is Feature Extraction?" *GeeksforGeeks*, 23 May 2024, www.geeksforgeeks.org/machine-learning/what-is-feature-extraction/.

[3] Mcleod, Saul. "Z-Score: Definition, Calculation and Interpretation." *Simplypsychology.org*, Simply Psychology, 6 Oct. 2023, www.simplypsychology.org/z-score.html.



This correlation matrix shows strong positive relationships between features that measure similar browser activities, like ExitRates and BounceRates (cor=0.91). This makes sense because both of these variables correspond to users leaving a webpage.

Additionally, from the correlation matrix, we can see that ProductRelated and ProductRelated_Duration were strongly correlated (cor=0.86). This makes sense- the more product-related pages a person has viewed (ProductRelated), the longer they have spent on product-related pages in that session (ProductRelated_Duration). This analysis informed our ablation study by highlighting PageValues, ExitRates, and BounceRates as candidates for systematic removal.

IV. Methods & Results

A. Training Multi-Layer Perceptron Models

For our research, we trained two multi-layer perceptron machine learning models to determine if an online shopper was likely to make a purchase based on variables measured in the dataset. A multi-layer perceptron (MLP) is a type of artificial neural network that consists of multiple layers of neurons. These neurons are generally nonlinear activation functions, which allow the model to learn complex patterns in the data[4]. We chose to use this type of model as it was the model that was used in the research paper that was used to collect this data - *Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks* (Sakar et al., 2019).

[4] Jaiswal, Sejal. "Multilayer Perceptrons in Machine Learning: A Comprehensive Guide." *Datacamp.com*, DataCamp, 7 Feb. 2024, www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning.

Our MLPs were built using scikit-learn's MLPClassifier. They each had one hidden layer with 10 neurons, making our models relatively small and simple MLPs, making them have faster computing times. This was helpful as our GPU is limited by what is available on Google Colaboratory. We used the ReLU activation function, the standard choice for neural networks, which helps our models learn nonlinear relationships that may be present in our dataset. We used an Adam Optimizer, which is a common choice for small datasets such as the *Online Shoppers Dataset,* which allows our models to adjust learning rates adaptively. We enabled early stopping, which means that, during training, the model checks performance on a validation set. If the model stops improving, training ends early, preventing overfitting and, again, saving computing time. Finally, all accuracy measures were computed on the held-out test portion of a 70/30 training/validation data split (random_state = 42). To ensure our results weren't dependent on a single split, this training was repeated with a second random seed (random_state = 1) and observed similar performance. This means that our conclusions are stable across variations in the data split.

We trained two models to answer the following question: Does fixing the class imbalance improve model performance? We identified that a class imbalance exists in this data, as about 15% of the data defines Revenue as True. Our first model, which we will call Model A, is tested on the pre-processed training dataset. Because this data is imbalanced, this model has the risk of learning to always predict 'no purchase,' as that guess is correct most of the time.

Our second model, Model B, is tested on a balanced training set using the synthetic minority over-sampling technique (SMOTE). SMOTE is a commonly used oversampling method that is used to solve the imbalance problem[5]. Its goal is to balance class distribution by randomly increasing minority class examples by replicating them. This artificially increases the size of the minority sample data.

B. Training & Interpreting a Logistic Regression Model

Additionally, we trained a logistic regression model to use as a baseline to compare our MLP's to. We did this because, unlike linear or tree-based models, MLP's are often criticized for being a 'black box,' and hard to interpret. Logistic regression models are simpler and more interpretable than our MLPs. Our goal with this model is to rank the top variables for determining shopper purchasing intent, and then to explicitly connect these key features to concrete business actions. We trained this model on the same preprocessed data on a 70/30 test-train split as we did for the MLPs.

We can also use this model to answer our research's guiding question: which features in this dataset are most important for determining users' purchasing intent? According to our model, the top key features, determined by which coefficients had the largest absolute values, are PageValues, Month_Feb, and Browser_12. A one-unit increase in purchasing intent corresponds to a logarithmic change in these key variables by the following values:

	Coefficient

[5] GeeksforGeeks. "ML | Handling Imbalanced Data with SMOTE and near Miss Algorithm in Python." *GeeksforGeeks*, 28 June 2019, www.geeksforgeeks.org/machine-learning/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/.

PageValues	1.528
Month_Feb	-1.334
Browser_12	1.303

Translating these findings into concrete business actions, if a website has high PageValues, businesses should display targeted ads and promotions or offer limited-time discounts. PageValues is the strongest determinant of purchasing intent, and businesses should strive to make their websites as engaging as possible to increase this value. During February, sessions are less likely to convert to a purchase- likely because this is right after the holiday season (Month_Feb). To alleviate this, businesses should encourage users to make purchases during this month by offering special sales and discounts to users. Finally, as Browser_12 is a significant variable, this shows that the browser type can have a significant effect on user behaviour, which may be due to differences in layout or browser performance. Browser coefficients range from -0.822 to 1.303. If Browser_12 has a high positive coefficient, it suggests that users on this browser convert at higher rates. Businesses should investigate what aspects of the site perform particularly well on this browser, such as load speed, layout, or compatibility, and apply those same optimizations to other browsers to improve conversion rates.

C. Comparing Model Performance

To evaluate our models, we split the data into training and test sets and reported performance using three metrics: F1, PR-AUC, and ROC-AUC. Additionally, to make these scores more interpretable, we compared them to the baseline logistic regression model, which is simpler and more interpretable:

F1 score: a performance metric that measures the balance between precision and recall for the positive purchase class[6]. The closer to 1 it is, the better the model.

Precision-Recall Area-Under-Curve (PR-AUC): an evaluation metric that quantifies how well a model can differentiate between classes. This is important as purchases only make up 15% of sessions. The closer to 1 it is, the better the model[7].

Receiver Operating Characteristic AUC (ROC-AUC): a metric that measures how well the model can produce scores that distinguish between positive or negative sessions[8].

	Model A (imbalanced)	Model B (SMOTE)	Baseline Model
Accuracy	0.891	0.860	0.880
F1	0.595	0.611	0.930

[6] GeeksforGeeks. "F1 Score in Machine Learning." *GeeksforGeeks*, 27 Dec. 2023, www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/.
[7] "Ultimate Guide to PR-AUC." *Coralogix*, 2025, coralogix.com/ai-blog/ultimate-guide-to-pr-auc-calculations-uses-and-limitations/.
[8] Evidently AI Team. "How to Explain the ROC AUC Score and ROC Curve?" *Www.evidentlyai.com*, 2024, www.evidentlyai.com/classification-metrics/explain-roc-curve.

PR-AUC	0.655	0.636	0.626
ROC-AUC	0.908	0.892	0.890

* highlighted scores correspond to the superior model per row

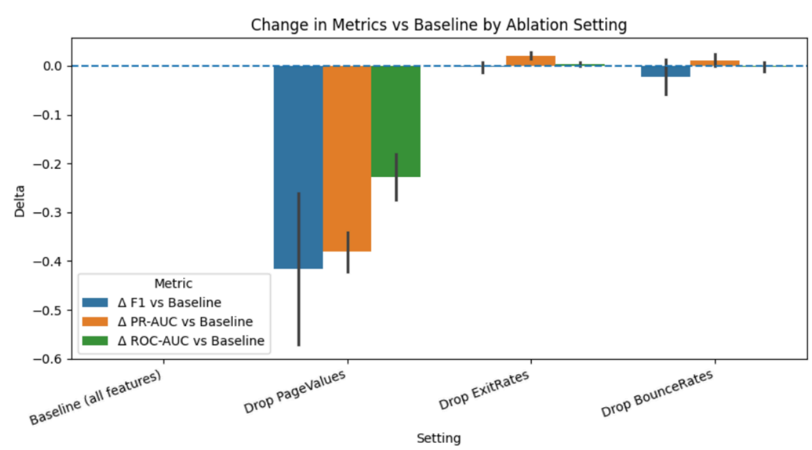
Nearly across the board, our accuracy, F1, PR-AUC, and ROC-AUC scores are very similar. The only score that significantly deviates across our models is the F1 score, in which the logistic regression model performs much better than our MLPs. We believe that this is because this dataset consists mostly of linear relationships between predictors, which logistic regressions are good at identifying. Additionally, we were surprised to see that Model A (the imbalanced MLP) scored higher PR-AUC and ROC-AUC metrics. We expected Model B (the SMOTE MLP) to have better scores than Model A across the board, as we assumed that, because the vast majority of samples in the raw dataset defined revenue as false, it would assume that all shopper behaviour ends without a purchase, as this is true most of the time.

Ultimately, we determine that Model A is the best model for determining if a user is going to make a purchase.

D. Ablation Study

In an attempt to further understand our multi-layer perceptron models, we performed an ablation analysis, a form of feature selection, on our MLPs to test how predictive certain variables are in finding purchasing intent. Conducting an ablation study enables us to determine whether destroying specific parts of the neural network will improve/hurt the predictive power of the system. Finally, we will interpret the real-life business implications of these features.

Following our correlation matrix in the Preliminary Studies portion of this paper, we determined that the BounceRates, ExitRates, and PageValues features of this dataset. To see how each variable affected our MLP's predictive power, we compared the F1, PR-AUC, and ROC-AUC scores before and after dropping each one. This is visualized in the following graph, which shows the change in these metrics vs. the baseline metrics by ablation analysis.



From this visualization, it's very clear that dropping PageValues would be a significant detriment to the model. We are hoping for larger F1, PR-AUC, and ROC-AUC scores - the closer to 1, the better. However, when PageValues is dropped, the model's predictive power significantly decreases. Its F1 score decreases by about 0.4, PR-AUC by about 0.37, and ROC-AUC by about 0.25.

Dropping ExitRates or BounceRates does not significantly change the model.

Connecting these findings to the real world, if a page has high PageValues, businesses should display targeted ads and promotions, offer discounts for a limited time, and recommend similar or complementary products on these pages. This will help maximize users' propensity to purchase items while they are on these pages, as it will excite them to make purchases quickly. If a page has high exit rates or bounce rates, businesses can try simplifying the checkout process, improving page load times, or making the page more engaging to look at. This would be an opportunity for the business to try new things on their websites to see what makes users make purchases.

V. Conclusion & Discussion

This study examined the *Online Shoppers Purchasing Intention Dataset* to determine the best model to predict whether or not a user is likely to make a purchase, and to find which features are most significant in making this prediction. Our findings indicate that a multi-layer perceptron trained on imbalanced data is best for determining shopper purchasing intent, as compared to a SMOTE-trained MLP or logistic regression model. Additionally, the month, browser type, and average number of pages a user visits in a day are the most important features in determining whether or not a user is going to make a purchase.

Finally, we concluded our paper with an ablation analysis, which determined that the average number of webpages a user visits per day is the single most influential predictor in shopper purchasing intent in this dataset.

These findings carry meaningful implications for real-world e-commerce settings, where identifying high-quality sessions can support more effective recommendations, targeted promotions, and strategies to encourage users to complete purchases. Future research may benefit from exploring different resampling techniques, trying different balancing methods, simpler alternative models, and basic ways of incorporating the order of user actions to see whether they help improve prediction accuracy.

VI. References

Coralogix. (2025). *Ultimate guide to PR-AUC: Calculations, uses, and limitations*. https://coralogix.com/ai-blog/ultimate-guide-to-pr-auc-calculations-uses-and-limitations/

Evidently, AI Team. (2024). *How to explain the ROC AUC score and ROC curve?* https://www.evidentlyai.com/classification-metrics/explain-roc-curve

GeeksforGeeks. (2019, June 28). *ML | Handling imbalanced data with SMOTE and the Near Miss algorithm in Python*. https://www.geeksforgeeks.org/machine-learning/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/

GeeksforGeeks. (2023, December 27). *F1 score in machine learning*. https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/

GeeksforGeeks. (2024, May 23). *What is feature extraction?* https://www.geeksforgeeks.org/machine-learning/what-is-feature-extraction/

Jaiswal, S. (2024, February 7). *Multilayer perceptrons in machine learning: A comprehensive guide*. DataCamp. https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning

Kargin, V. (2025). *Statistical Learning* [Lecture notes]. MATH 457: Statistical Learning, Binghamton University.

Mcleod, S. (2023, October 6). *Z-score: Definition, calculation, and interpretation*. Simply Psychology. https://www.simplypsychology.org/z-score.html

OpenAI. (2025). *Python code debugging assistance and brainstorming for code structure* [Large language model]. ChatGPT. https://chat.openai.com/

Quoraishee, S. (2024, January 7). *Ablation testing neural networks: The compensatory masquerade*. Towards Data Science. https://towardsdatascience.com/ablation-testing-neural-networks-the-compensatory-masquerade-ba27d0037a88/

Sakar, C. & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). A novel attribute selection approach in online shoppers' purchasing intention analysis. *Neural Computing and Applications, 31*(12), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0